

# Lecture Slides for Programming in C++ [The C++ Language, Libraries, Tools, and Other Topics] (Version: 2019-09-01-SENG475)

Current with the C++17 Standard



Michael D. Adams

Department of Electrical and Computer Engineering  
University of Victoria  
Victoria, British Columbia, Canada

To obtain the *most recent version* of these lecture slides or for additional information and resources related to these slides (including errata and *lecture videos*), please visit:

<http://www.ece.uvic.ca/~mdadams/cppbook>



youtube.com/iamcanadian1867



github.com/mdadams



@mdadams16

NEXT SLIDE: Algorithms

The author has taken care in the preparation of this document, but makes no expressed or implied warranty of any kind and assumes no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

Copyright © 2015, 2016, 2017, 2018, 2019 Michael D. Adams

This document is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0) License. A copy of this license can be found on page iii of this document. For a simple explanation of the rights granted by this license, see:

<http://creativecommons.org/licenses/by-nc-nd/3.0/>

UNIX and X Window System are registered trademarks of The Open Group. Windows is a registered trademark of Microsoft Corporation. Fedora is a registered trademark of Red Hat, Inc. Ubuntu is a registered trademark of Canonical Ltd. MATLAB is a registered trademark of The MathWorks, Inc. OpenGL is a registered trademark of Hewlett Packard Enterprise. The YouTube logo is a registered trademark of Google, Inc. The GitHub logo is a registered trademark of GitHub, Inc. The Twitter logo is a registered trademark of Twitter, Inc.

This document was typeset with L<sup>A</sup>T<sub>E</sub>X.

Creative Commons Legal Code

Attribution-NonCommercial-NoDerivs 3.0 Unported

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

## License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

### 1. Definitions

- a. "Adaptation" means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this license. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.
- b. "Collection" means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed

in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined above) for the purposes of this License.

- c. "Distribute" means to make available to the public the original and copies of the Work through sale or other transfer of ownership.
- d. "Licensor" means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.
- e. "Original Author" means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.
- f. "Work" means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.
- g. "You" means an individual or entity exercising rights under this



License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

- h. "Publicly Perform" means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.
- i. "Reproduce" means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

- a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections; and,
- b. to Distribute and Publicly Perform the Work including as incorporated in Collections.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats, but otherwise you have no rights to make Adaptations. Subject to 8(f), all rights not expressly granted by Licensor

are hereby reserved, including but not limited to the rights set forth in Section 4(d).

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

- a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested.
- b. You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file-sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.
- c. If You Distribute, or Publicly Perform the Work or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice,

terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work. The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Collection, at a minimum such credit will appear, if a credit for all contributing authors of Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

d. For the avoidance of doubt:

- i. Non-waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;
- ii. Waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License if Your exercise of such rights is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(b) and otherwise waives the right to collect royalties through any statutory or compulsory licensing scheme; and,
- iii. Voluntary License Schemes. The Licensor reserves the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License that is for a purpose or use which is otherwise than noncommercial as permitted

under Section 4(b).

- e. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation.

## 5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

## 7. Termination

- a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.
- b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this

License will continue in full force and effect unless terminated as stated above.

## 8. Miscellaneous

- a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.
- b. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
- c. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
- d. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.
- e. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of this License.

Creative Commons may be contacted at <http://creativecommons.org/>.

- 1 M. D. Adams, *Multiresolution Signal and Geometry Processing: Filter Banks, Wavelets, and Subdivision (Version 2013-09-26)*, University of Victoria, Victoria, BC, Canada, Sept. 2013, xxxviii + 538 pages, ISBN 978-1-55058-507-0 (print), ISBN 978-1-55058-508-7 (PDF). Available from Google Books, Google Play Books, University of Victoria Bookstore, and author's web site

<http://www.ece.uvic.ca/~mdadams/waveletbook>.

- 2 M. D. Adams, *Lecture Slides for Multiresolution Signal and Geometry Processing (Version 2015-02-03)*, University of Victoria, Victoria, BC, Canada, Feb. 2015, xi + 587 slides, ISBN 978-1-55058-535-3 (print), ISBN 978-1-55058-536-0 (PDF). Available from Google Books, Google Play Books, University of Victoria Bookstore, and author's web site

<http://www.ece.uvic.ca/~mdadams/waveletbook>.

- 3 M. D. Adams, *Continuous-Time Signals and Systems (Version 2013-09-11)*, University of Victoria, Victoria, BC, Canada, Sept. 2013, xxx + 308 pages, ISBN 978-1-55058-495-0 (print), ISBN 978-1-55058-506-3 (PDF). Available from Google Books, Google Play Books, University of Victoria Bookstore, and author's web site  
<http://www.ece.uvic.ca/~mdadams/sigsysbook>.
- 4 M. D. Adams, *Lecture Slides for Continuous-Time Signals and Systems (Version 2013-09-11)*, University of Victoria, Victoria, BC, Canada, Dec. 2013, 286 slides, ISBN 978-1-55058-517-9 (print), ISBN 978-1-55058-518-6 (PDF). Available from Google Books, Google Play Books, University of Victoria Bookstore, and author's web site  
<http://www.ece.uvic.ca/~mdadams/sigsysbook>.



- 5 M. D. Adams, *Lecture Slides for Signals and Systems (Version 2016-01-25)*, University of Victoria, Victoria, BC, Canada, Jan. 2016, xvi + 481 slides, ISBN 978-1-55058-584-1 (print), ISBN 978-1-55058-585-8 (PDF). Available from Google Books, Google Play Books, University of Victoria Bookstore, and author's web site  
<http://www.ece.uvic.ca/~mdadams/sigsysbook>.

## Part 0

### Preface

# About These Lecture Slides

- This document constitutes a detailed set of lecture slides on the C++ programming language and is current with the **C++17** standard. [\[C++17 \(full\)\]](#)
- Many aspects of the C++ language are covered from introductory to more advanced.
- Some aspects of the C++ standard library are also introduced.
- In addition, various general programming-related topics are considered.

- The author would like to thank Robert Leahy for reviewing various drafts of many of these slides and providing many useful comments that allowed the quality of these materials to be improved significantly.

- Many code examples are included throughout these slides.
- Often, in order to make an example short enough to fit on a slide, compromises had to be made in terms of good programming style.
- These deviations from good style include (but are not limited to) such things as:
  - 1 frequently formatting source code in unusual ways to conserve vertical space in listings;
  - 2 not fully documenting source code with comments;
  - 3 using short meaningless identifier names;
  - 4 omitting include guards from headers; and
  - 5 engaging in various other evil behavior such as: using many global variables and employing constructs like “**using namespace** std;”.

- In a definition, the term being defined is often typeset **like this**.
- To emphasize particular text, the text is typeset *like this*.
- To show that particular text is associated with a hyperlink to an internal target, the text is typeset like this.
- To show that particular text is associated with a hyperlink to an external document, the text is typeset like this.
- URLs are typeset like `http://www.ece.uvic.ca/~mdadams`.

- The author of the lecture slides maintains a companion web site for the slides.
- The most recent version of the slides can be downloaded from this site.
- Additional information related to the slides is also available from this site, including:
  - errata for the slides; and
  - information on the companion web site, companion Git repository, and companion YouTube channel for the slides.
- The URL of this web site is:
  - <http://www.ece.uvic.ca/~mdadams/cppbook>

- These lecture slides have a companion Git repository.
- Numerous code examples and exercises are available from this repository.
- This repository is hosted by GitHub.
- The URL of the main repository page on GitHub is:
  - [https://github.com/mdadams/cppbook\\_companion](https://github.com/mdadams/cppbook_companion)
- The URL of the actual repository itself is:
  - [https://github.com/mdadams/cppbook\\_companion.git](https://github.com/mdadams/cppbook_companion.git)



- Video lectures for some of the material covered by these lecture slides can be found on the author's YouTube channel.
- The URL of the author's YouTube channel is:
  - <https://www.youtube.com/user/iamcanadian1867>

# Software Development Environment (SDE)

- The Software Development Environment (SDE) is a collection of tools that can be used to provide a basic up-to-date environment for C++ code development.
- The SDE should work with most Linux distributions, provided that the necessary software dependencies are installed.
- Amongst other things, the SDE software provides a collection of scripts for installing packages like:
  - Boost, Catch2, CGAL, Clang, CMake, GCC, Gcovr, GDB, GSL, Lcov, Libcxx, TeX Live, Vim, Vim LSP, and YCM
- The SDE software has a Git repository hosted by GitHub.
- The URL of the main repository page on GitHub is:
  - <https://github.com/mdadams/sde>
- The URL of the actual repository itself is:
  - <https://github.com/mdadams/sde.git>
- For more information about the SDE, refer to the main repository page on GitHub.

# Part 1

## Software

# Why Is Software Important?

- almost all electronic devices run some software
- automobile engine control system, implantable medical devices, remote controls, office machines (e.g., photocopiers), appliances (e.g., televisions, refrigerators, washers/dryers, dishwashers, air conditioner), power tools, toys, mobile phones, media players, computers, printers, photocopies, disk drives, scanners, webcams, MRI machines

# Why Software-Based Solutions?

- more cost effective to implement functionality in software than hardware
- software bugs easy to fix, give customer new software upgrade
- hardware bugs extremely costly to repair, customer sends in old device and manufacturer sends replacement
- systems increasingly complex, bugs unavoidable
- allows new features to be added later
- implement only absolute minimal functionality in hardware, do the rest in software

- many more software jobs than hardware jobs
- relatively small team of hardware designers produce platform like iPhone
- thousands of companies develop applications for platform
- only implement directly in hardware when absolutely necessary (e.g., for performance reasons)

- created by Dennis Ritchie, AT&T Bell Labs in 1970s
- international standard ISO/IEC 9899:2018 (informally known as “C18”)
- available on wide range of platforms, from microcontrollers to supercomputers; very few platforms for which C compiler not available
- procedural, provides language constructs that map efficiently to machine instructions
- does not directly support object-oriented or generic programming
- application domains: system software, device drivers, embedded applications, application software
- greatly influenced development of C++
- when something lasts in computer industry for more than 40 years (outliving its creator), must be good

- created by Bjarne Stroustrup, Bell Labs
- originally C with Classes, renamed as C++ in 1983
- most recent specification of language in ISO/IEC 14882:2017 (informally known as “C++17”)
- procedural
- loosely speaking is superset of C
- directly supports object-oriented and generic programming
- maintains efficiency of C
- application domains: systems software, application software, device drivers, embedded software, high-performance server and client applications, entertainment software such as video games, native code for Android applications
- greatly influenced development of C# and Java



- developed in 1990s by James Gosling at Sun Microsystems (later bought by Oracle Corporation)
- de facto standard but not international standard
- usually less efficient than C and C++
- simplified memory management (with garbage collection)
- direct support for object-oriented programming
- application domains: web applications, Android applications

- designed by John Backus, IBM, in 1950s
- international standard ISO/IEC 1539-1:2010 (informally known as "Fortran 2008")
- application domain: scientific and engineering applications, intensive supercomputing tasks such as weather and climate modelling, finite element analysis, computational fluid dynamics, computational physics, computational chemistry

- developed by Microsoft, team led by Anders Hejlsberg
- ECMA-334 and ISO/IEC 23270:2006
- most recent language specifications not standardized by ECMA or ISO/IEC
- intellectual property concerns over Microsoft patents
- object oriented

- developed by Tom Love and Brad Cox of Stepstone (later bought by NeXT and subsequently Apple)
- used primarily on Apple Mac OS X and iOS
- strict superset of C
- no official standard that describes Objective C
- authoritative manual on Objective-C 2.0 available from Apple

- proprietary language, developed by The MathWorks
- not general-purpose programming language
- application domain: numerical computing
- used to design and simulate systems
- not used to implement real-world systems

# Why Learn C++?

- vendor neutral
- international standard
- general purpose
- powerful yet efficient
- loosely speaking, includes C as subset; so can learn two languages (C++ and C) for price of one
- easy to move from C++ to other languages but often not in other direction
- many other popular languages inspired by C++
- popular language
- consistently ranks amongst top languages in TIOBE Software Programming Community Index  
(<https://www.tiobe.com/tiobe-index/>)

## Part 2

### C++

## Section 2.1

### History of C++



# Motivation

- developed by Bjarne Stroustrup starting in 1979 at Computing Science Research Center of Bell Laboratories, Murray Hill, NJ, USA
- doctoral work in Computing Laboratory of University of Cambridge, Cambridge, UK
- study alternatives for organization of system software for distributed systems
- required development of relatively large and detailed simulator
- dissertation:
  - B. Stroustrup. *Communication and Control in Distributed Computer Systems*.  
PhD thesis, University of Cambridge, Cambridge, UK, 1979.
- in 1979, joined Bell Laboratories after having finished doctorate
- work started with attempt to analyze UNIX kernel to determine to what extent it could be distributed over network of computers connected by LAN
- needed way to model module structure of system and pattern of communication between modules
- no suitable tools available

# Objectives

- had bad experiences writing simulator during Ph.D. studies; originally used Simula for simulator; later forced to rewrite in BCPL for speed; more low level than C; BCPL was horrible to use
- notion of what properties good tool would have motivated by these experiences
- suitable tool for projects like simulator, operating system, other systems programming tasks should:
  - support for effective program organization (like in Simula) (i.e., classes, some form of class hierarchies, some form of support for concurrency, strong checking of type system based on classes)
  - produce programs that run fast (like with BCPL)
  - be able to easily combine separately compilable units into program (like with BCPL)
  - have simple linkage convention, essential for combining units written in languages such as C, Algol68, Fortran, BCPL, assembler into single program
  - allow highly portable implementations (only very limited ties to operating system)

# Timeline for C with Classes (1979–1983) I

**May 1979** work on C with Classes starts

**Oct 1979** initial version of Cpre, preprocessor that added Simula-like classes to C; language accepted by preprocessor later started being referred to as C with Classes

**Mar 1980** Cpre supported one real project and several experiments (used on about 16 systems)

**Apr 1980** first internal Bell Labs paper on C with Classes published (later to appear in ACM SIGPLAN Notices in Jan. 1982)

B. Stroustrup. **Classes: An abstract data type facility for the C language.**

Bell Laboratories Computer Science Technical Report CSTR-84, Apr. 1980.

# Timeline for C with Classes (1979–1983) II

1980 initial 1980 implementation had following features:

- classes
- derived classes
- public/private access control
- constructors and destructors
- call and return functions (call function implicitly called before every call of every member function; return function implicitly called after every return from every member function; can be used for synchronization)
- friend classes
- type checking and conversion of function arguments

1981 in 1981, added:

- inline functions
- default arguments
- overloading of assignment operator

Jan 1982 first external paper on C with Classes published

# Timeline for C with Classes (1979–1983) III

B. Stroustrup. [Classes: An abstract data type facility for the C language.](#)

*ACM SIGPLAN Notices*, 17(1):42–51, Jan. 1982.

**Feb 1983** more detailed paper on C with Classes published

B. Stroustrup. [Adding classes to the C language: An exercise in language evolution.](#)

*Software: Practice and Experience*, 13(2):139–161, Feb. 1983.

- C with Classes proved very successful; generated considerable interest
- first real application of C with Classes was network simulators

# Timeline for C84 to C++98 (1982–1998) I

- started to work on cleaned up and extended successor to C with Classes, initially called C84 and later renamed C++

**Spring 1982** started work on Cfront compiler front-end for C84; initially written in C with Classes and then transcribed to C84; traditional compiler front-end performing complete check of syntax and semantics of language, building internal representation of input, analyzing and rearranging representation, and finally producing output for some code generator; generated C code as output; difficult to bootstrap on machine without C84 compiler; Cfront software included special “half-processed” version of C code resulting from compiling Cfront, which could be compiled with native C compiler and resulting executable then used to compile Cfront

# Timeline for C84 to C++98 (1982–1998) II

**Dec 1983** C84 (C with Classes) renamed C++;  
name used in following paper prepared in Dec. 1983

B. Stroustrup. [Data abstraction in C.](#)

*Bell Labs Technical Journal*, 63(8):1701–1732, Oct. 1984.

(name C++ suggested by Rick Mascitti)

**1983** virtual functions added

**Note:** going from C with Classes to C84 added: virtual functions,  
function name and operator overloading, references, constants  
(**const**), user-controlled free-store memory control, improved  
type checking

**Jan 1984** first C++ manual

B. Stroustrup. [The C++ reference manual.](#)

AT&T Bell Labs Computer Science Technical Report No.  
108, Jan. 1984.

# Timeline for C84 to C++98 (1982–1998) III

Sep 1984 paper describing operator overloading published

B. Stroustrup. *Operator overloading in C++*.

In *Proc. IFIP WG2.4 Conference on System Implementation Languages: Experience & Assessment*, Sept. 1984.

1984 stream I/O library first implemented and later presented in

B. Stroustrup. *An extensible I/O facility for C++*.

In *Proc. of Summer 1985 USENIX Conference*, pages 57–70, June 1985.

Feb 1985 Cfront Release E (first external release); “E” for “Educational”; available to universities

Oct 1985 Cfront Release 1.0 (first commercial release)

Oct 1985 first edition of C++PL written

B. Stroustrup. *The C++ Programming Language*.

Addison Wesley, 1986.



# Timeline for C84 to C++98 (1982–1998) IV

(Cfront Release 1.0 corresponded to language as defined in this book)

**Oct 1985** tutorial paper on C++

B. Stroustrup. [A C++ tutorial](#).

In *Proceedings of the ACM annual conference on the range of computing: mid-80's perspective*, pages 56–64, Oct. 1985.

**Jun 1986** Cfront Release 1.1; mainly bug fix release

**Aug 1986** first exposition of set of techniques for which C++ was aiming to provide support (rather than what features are already implemented and in use)

B. Stroustrup. [What is object-oriented programming?](#)

In *Proc. of 14th Association of Simula Users Conference*, Stockholm, Sweden, Aug. 1986.

# Timeline for C84 to C++98 (1982–1998) V

- Sep 1986 first Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA) conference (start of OO hype centered on Smalltalk)
- Nov 1986 first commercial Cfront PC port (Cfront 1.1, Glockenspiel [in Ireland])
- Feb 1987 Cfront Release 1.2; primarily bug fixes but also added:
  - pointers to members
  - protected members
- Nov 1987 first conference devoted to C++:  
USENIX C++ conference (Santa Fe, NM, USA)
- Dec 1987 first GNU C++ release (1.13)
- Jan 1988 first Oregon Software (a.k.a. TauMetric) C++ release
- Jun 1988 first Zortech C++ release
- Oct 1988 first presented templates at USENIX C++ conference (Denver, CO, USA) in paper:

# Timeline for C84 to C++98 (1982–1998) VI

B. Stroustrup. [Parameterized types for C++](#).

In *Proc. of USENIX C++ Conference*, pages 1–18, Denver, CO, USA, Oct. 1988.

**Oct 1988** first USENIX C++ implementers workshop (Estes Park, CO, USA)

**Jan 1989** first C++ journal “The C++ Report” (from SIGS publications) started publishing

**Jun 1989** Cfront Release 2.0 major cleanup; new features included:

- multiple inheritance
- type-safe linkage
- better resolution of overloaded functions
- recursive definition of assignment and initialization
- better facilities for user-defined memory management
- abstract classes
- static member functions
- const member functions

# Timeline for C84 to C++98 (1982–1998) VII

- protected member functions (first provided in release 1.2)
- overloading of operator ->
- pointers to members (first provided in release 1.2)

1989 main features of Cfront 2.0 summarized in

B. Stroustrup. [The evolution of C++: 1985–1989.](#)  
*USENIX Computer Systems*, 2(3), Summer 1989.

first presented in

B. Stroustrup. [The evolution of C++: 1985–1987.](#)  
In *Proc. of USENIX C++ Conference*, pages 1–22, Santa Fe, NM, USA, Nov. 1987.

Nov 1989 paper describing exceptions published

A. Koenig and B. Stroustrup. [Exception handling for C++.](#)  
In *Proc. of “C++ at Work” Conference*, Nov. 1989.

followed up by

# Timeline for C84 to C++98 (1982–1998) VIII

A. Koenig and B. Stroustrup. [Exception handling for C++](#).  
In *Proc. of USENIX C++ Conference*, Apr. 1990.

**Dec 1989** ANSI X3J16 organizational meeting (Washington, DC, USA)

**Mar 1990** first ANSI X3J16 technical meeting (Somerset, NJ, USA)

**Apr 1990** Cfront Release 2.1; bug fix release to bring Cfront mostly into line with ARM

**May 1990** annotated reference manual (ARM) published

M. A. Ellis and B. Stroustrup. [The Annotated C++ Reference Manual](#).

Addison Wesley, May 1990.

(formed basis for ANSI standardization)

**May 1990** first Borland C++ release

**Jul 1990** templates accepted (Seattle, WA, USA)

**Nov 1990** exceptions accepted (Palo Alto, CA, USA)

# Timeline for C84 to C++98 (1982–1998) IX

Jun 1991 second edition of C++PL published

B. Stroustrup. *The C++ Programming Language*.  
Addison Wesley, 2nd edition, June 1991.

Jun 1991 first ISO WG21 meeting (Lund, Sweden)

Sep 1991 Cfront Release 3.0; added templates (as specified in ARM)

Oct 1991 estimated number of C++ users 400,000

Feb 1992 first DEC C++ release (including templates and exceptions)

Mar 1992 run-time type identification (RTTI) described in

B. Stroustrup and D. Lenkov. *Run-time type identification for C++*.

*The C++ Report*, Mar. 1992.

(RTTI in C++ based on this paper)

Mar 1992 first Microsoft C++ release (did not support templates or exceptions)

# Timeline for C84 to C++98 (1982–1998) X

- May 1992 first IBM C++ release (including templates and exceptions)
- Mar 1993 RTTI accepted (Portland, OR, USA)
  - Jul 1993 namespaces accepted (Munich, Germany)
  - 1993 further work on Cfront Release 4.0 abandoned after failed attempt to add exception support
- Aug 1994 ANSI/ISO Committee Draft registered
- Aug 1994 Standard Template Library (STL) accepted (Waterloo, ON, CA); described in
  - A. Stepanov and M. Lee. [The standard template library](#). Technical Report HPL-94-34 (R.1), HP Labs, Aug. 1994.
- Aug 1996 **export** accepted (Stockholm, Sweden)
  - 1997 third edition of C++PL published

# Timeline for C84 to C++98 (1982–1998) XI

B. Stroustrup. *The C++ Programming Language*.

Addison Wesley Longman, Reading, MA, USA, 3rd edition, 1997.

**Nov 1997** final committee vote on complete standard (Morristown, NJ, USA)

**Jul 1998** Microsoft releases VC++ 6.0, first Microsoft compiler to provide close-to-complete set of ISO C++

**Sep 1998** ISO/IEC 14882:1998 (informally known as C++98) published  
ISO/IEC 14882:1998 — programming languages — C++,  
Sept. 1998.

**1998** Beman Dawes starts Boost (provides peer-reviewed portable C++ source libraries)

**Feb 2000** special edition of C++PL published

B. Stroustrup. *The C++ Programming Language*.

Addison Wesley, Reading, MA, USA, special edition, Feb. 2000.



# Timeline After C++98 (1998–Present) I

- Apr 2001** motion passed to request new work item: technical report on libraries (Copenhagen, Denmark); later to become ISO/IEC TR 19768:2007
- Oct 2003** ISO/IEC 14882:2003 (informally known as C++03) published; essentially bug fix release; no changes to language from programmer's point of view
  - ISO/IEC 14882:2003 — programming languages — C++, Oct. 2003.
- 2003** work on C++0x (now known as C++11) starts
- Oct 2004** estimated number of C++ users 3,270,000
- Apr 2005** first votes on features for C++0x (Lillehammer, Norway)
  - 2005** **auto**, **static\_assert**, and rvalue references accepted in principle
- Apr 2006** first full committee (official) votes on features for C++0x (Berlin, Germany)

# Timeline After C++98 (1998–Present) II

- Sep 2006** performance technical report (TR 18015) published:  
ISO/IEC TR 18015:2006 — information technology —  
programming languages, their environments and system  
software interfaces — technical report on C++ performance,  
Sept. 2006.
- work spurred by earlier proposal to standardize subset of C++  
for embedded systems called Embedded C++ (or just EC++);  
EC++ motivated by performance concerns
- Apr 2006** decision to move special mathematical functions to separate ISO  
standard (Berlin, Germany); deemed too specialized for most  
programmers
- Nov 2007** ISO/IEC TR 19768:2007 (informally known as C++TR1)  
published;

ISO/IEC TR 19768:2007 — information technology — programming languages — technical report on C++ library extensions, Nov. 2007.

specifies series of library extensions to be considered for adoption later in C++

**2009** another particularly notable book on C++ published

B. Stroustrup. *Programming: Principles and Practice Using C++*.

Addison Wesley, Upper Saddle River, NJ, USA, 2009.

**Aug 2011** ISO/IEC 14882:2011 (informally known as C++11) ratified  
ISO/IEC 14882:2011 — information technology — programming languages — C++, Sept. 2011.

**2013** fourth edition of C++PL published

B. Stroustrup. *The C++ Programming Language*.  
Addison Wesley, 4th edition, 2013.

**2014** ISO/IEC 14882:2014 (informally known as C++14) ratified  
ISO/IEC 14882:2014 — information technology —  
programming languages — C++, Dec. 2014.

**2017** ISO/IEC 14882:2017 (informally known as C++17) ratified  
ISO/IEC 14882:2017 — information technology —  
programming languages — C++, Dec. 2017.

- reasons for using C as starting point:
  - flexibility (can be used for most application areas)
  - efficiency
  - availability (C compilers available for most platforms)
  - portability (source code relatively portable from one platform to another)
- main sources for ideas for C++ (aside from C) were Simula, Algol68, BCPL, Ada, Clu, ML; in particular:
  - Simula gave classes
  - Algol68 gave operator overloading, references, ability to declare variables anywhere in block
  - BCPL gave `//` comments
  - exceptions influenced by ML
  - templates influenced by generics in Ada and parameterized modules in Clu

# C++ User Population

Time	Estimated Number of Users
Oct 1979	1
Oct 1980	16
Oct 1981	38
Oct 1982	85
Oct 1983	??+2 (no Cpre count)
Oct 1984	??+50 (no Cpre count)
Oct 1985	500
Oct 1986	2,000
Oct 1987	4,000
Oct 1988	15,000
Oct 1989	50,000
Oct 1990	150,000
Oct 1991	400,000
Oct 2004	over 3,270,000

- above numbers are conservative
- 1979 to 1991: C++ user population doubled approximately every 7.5 months
- stable growth thereafter

# Success of C++

- C++ very successful programming language
- not luck or solely because based on C
- efficient, provides low-level access to hardware, but also supports abstraction
- non-proprietary: in 1989, all rights to language transferred to standards bodies (first ANSI and later ISO) from AT&T
- multi-paradigm language, supporting procedural, object-oriented, generic, and functional (e.g., lambda functions) programming
- does not force particular programming style
- reasonably portable
- has continued to evolve, incorporating new ideas (e.g., templates, exceptions, STL)
- stable: high degree of compatibility with earlier versions of language
- very strong bias towards providing general-purpose facilities rather than more application-specific ones

# Application Areas

- banking and financial (funds transfer, financial modelling, teller machines)
- classical systems programming (compilers, operating systems, device drivers, network layers, editors, database systems)
- small business applications (inventory systems)
- desktop publishing (document viewers/editors, image editing)
- embedded systems (cameras, cell phones, airplanes, medical systems, appliances, space technologies)
- entertainment (games)
- graphics programming
- hardware design and verification
- scientific and numeric computation (physics, engineering, simulations, data analysis, geometry processing)
- servers (web servers, billing systems)
- telecommunication systems (phones, networking, monitoring, billing, operations systems)
- middleware



## Section 2.1.1

### **References**

# Evolution of C++

- B. Stroustrup. *A history of C++: 1979–1991*.  
In *Proc. of ACM History of Programming Languages Conference*, pages 271–298, Mar. 1993
- B. Stroustrup. *The Design and Evolution of C++*.  
Addison Wesley, Mar. 1994.
- B. Stroustrup. *Evolving a language in and for the real world: C++ 1991–2006*.  
In *Proc. of the ACM SIGPLAN Conference on History of Programming Languages*, pages 4–1–4–59, 2007.
- Cfront software available from Computer History Museum's Software Preservation Group <http://www.softwarepreservation.org>.  
(See [http://www.softwarepreservation.org/projects/c\\_plus\\_plus/cfront](http://www.softwarepreservation.org/projects/c_plus_plus/cfront)).
- ISO JTC1/SC22/WG21 web site.  
<http://www.open-std.org/jtc1/sc22/wg21/>.

- ISO/IEC 14882:1998 — programming languages — C++, Sept. 1998.
- ISO/IEC 14882:2003 — programming languages — C++, Oct. 2003.
- ISO/IEC TR 18015:2006 — information technology — programming languages, their environments and system software interfaces — technical report on C++ performance, Sept. 2006.
- ISO/IEC TR 19768:2007 — information technology — programming languages — technical report on C++ library extensions, Nov. 2007.
- ISO/IEC 29124:2010 — information technology — programming languages, their environments and system software interfaces — extensions to the C++ library to support mathematical special functions, Sept. 2010.
- ISO/IEC TR 24733:2011 — information technology — programming languages, their environments and system software interfaces — extensions for the programming language C++ to support decimal floating-point arithmetic, Nov. 2011.

- ISO/IEC 14882:2011 — information technology — programming languages — C++, Sept. 2011.
- ISO/IEC 14882:2014 — information technology — programming languages — C++, Dec. 2014.
- ISO/IEC TS 18822:2015 — programming languages — C++ — file system technical specification, July 2015.
- ISO/IEC TS 19570:2015 — programming languages — technical specification for C++ extensions for parallelism, July 2015.
- ISO/IEC TS 19841:2015 — technical specification for C++ extensions for transactional memory, Oct. 2015.
- ISO/IEC TS 19568:2015 — programming languages — C++ extensions for library fundamentals, Oct. 2015.
- ISO/IEC TS 19217:2015 — programming languages — C++ extensions for concepts, Nov. 2015.

- ISO/IEC TS 19571:2016 — programming languages — technical specification for C++ extensions for concurrency, Feb. 2016.
- ISO/IEC TS 19568:2017 — programming languages — C++ extensions for library fundamentals, Mar. 2017.
- ISO/IEC TS 21425:2017 — programming languages — C++ extensions for ranges, Nov. 2017.
- ISO/IEC TS 22277:2017 — technical specification — C++ extensions for coroutines, Nov. 2017.
- ISO/IEC 14882:2017 — information technology — programming languages — C++, Dec. 2017.
- ISO/IEC TS 19216:2018 — programming languages — C++ extensions for networking, Apr. 2018.
- ISO/IEC TS 21544:2018 — programming languages — extensions to C++ for modules, May 2018.
- ISO JTC1/SC22/WG21 web site.

<http://www.open-std.org/jtc1/sc22/wg21/>.

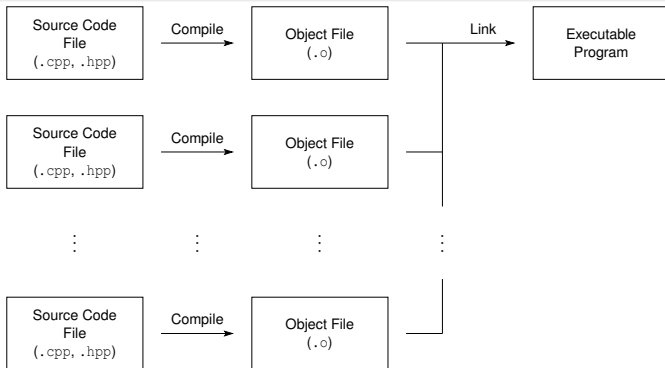
## Section 2.2

### Getting Started

```
1  #include <iostream>
2
3  int main()
4  {
5      std::cout << "Hello, world!\n";
6  }
```

- program prints message “Hello, world!” and then exits
- starting point for execution of C++ program is function called `main`; every C++ program must define function called `main`
- `#include` preprocessor directive to include complete contents of file
- `iostream` standard header file that defines various types and variables related to I/O
- `std::cout` is standard output stream (defaults to user’s terminal)
- operator `<<` is used for output

# Software Build Process



- start with C++ source code files (.cpp, .hpp)
- compile: convert source code to object code
- object code stored in object file (.o)
- link: combine contents of one or more object files (and possibly some libraries) to produce executable program
- executable program can then be run directly



- `g++` command provides both compiling and linking functionality
- command-line usage:

`g++ [options] input_file ...`

- many command-line options are supported
- some particularly useful command-line options listed on next slide
- compile C++ source file *file.cpp* to produce object code file *file.o*:

`g++ -c file.cpp`

- link object files *file\_1.o*, *file\_2.o*, ... to produce executable file *executable*:

`g++ -o executable file_1.o file_2.o ...`

- web site:

<http://www.gnu.org/software/gcc>

- C++ standards support in GCC:

<https://gcc.gnu.org/projects/cxx-status.html>

# Common g++ Command-Line Options

- `-c`
  - compile only (i.e., do not link)
- `-o file`
  - use file *file* for output
- `-g`
  - include debugging information
- `-On`
  - set optimization level to *n* (0 almost none; 3 full)
- `-std=c++17`
  - conform to C++17 standard
- `-Idir`
  - specify additional directory *dir* to search for include files
- `-Ldir`
  - specify additional directory *dir* to search for libraries
- `-llib`
  - link with library *lib*

# Common g++ Command-Line Options (Continued 1)

- `-pthread`
  - enable concurrency support (via pthreads library)
- `-pedantic-errors`
  - strictly enforce compliance with standard
- `-Wall`
  - enable most warning messages
- `-Wextra`
  - enable some extra warning messages not enabled by `-Wall`
- `-Wpedantic`
  - warn about deviations from strict standard compliance
- `-Werror`
  - treat all warnings as errors
- `-fno-elide-constructors`
  - in contexts where standard allows (but does not require) optimization that omits creation of temporary, do not attempt to perform this optimization

- `-fconstexpr-loop-limit=n`
  - set maximum number of iterations for loop in constexpr functions to *n*
- `-fconstexpr-depth=n`
  - set maximum nested evaluation depth for constexpr functions to *n*

- `clang++` command provides both compiling and linking functionality
- command-line usage:

`clang++ [options] input_file ...`

- many command-line options are supported
- command-line interface is largely compatible with that of GCC `g++` command
- web site:

<http://clang.llvm.org>

- C++ standards support in Clang:

[http://clang.llvm.org/cxx\\_status.html](http://clang.llvm.org/cxx_status.html)

# Common clang++ Command-Line Options

- many of more frequently used command-line options for clang++ identical to those for g++
- consequently, only small number of clang++ options given below
- `-fconstexpr-steps=n`
  - sets maximum number of computation steps in constexpr functions to *n*
- `-fconstexpr-depth=n`
  - sets maximum nested evaluation depth for constexpr functions to *n*

# Manually Building `hello` Program

- numerous ways in which `hello` program could be built
- often advantageous to compile each source file separately
- can compile and link as follows:
  - 1 compile source code file `hello.cpp` to produce object file `hello.o`:

```
g++ -c hello.cpp
```
  - 2 link object file `hello.o` to produce executable program `hello`:

```
g++ -o hello hello.o
```
- generally, manual building of program is quite tedious, especially when program consists of multiple source files and additional compiler options need to be specified
- in practice, we use tools to automate build process (e.g., CMake and Make)

## Section 2.3

### **C++ Basics**



# The C++ Programming Language

- created by Bjarne Stroustrup of Bell Labs
- originally known as C with Classes; renamed as C++ in 1983
- most recent specification of language in ISO/IEC 14882:2017 (informally known as “C++17”)
- next version of standard expected in approximately 2020 (informally known as “C++20”)
- procedural
- loosely speaking is superset of C
- directly supports object-oriented and generic programming
- maintains efficiency of C
- application domains: systems software, application software, device drivers, embedded software, high-performance server and client applications, entertainment software such as video games, native code for Android applications
- greatly influenced development of C# and Java

- two styles of comments provided
- comment starts with `//` and proceeds to end of line
- comment starts with `/*` and proceeds to first `*/`

```
// This is an example of a comment.  
/* This is another example of a comment. */  
/* This is an example of a comment that  
   spans  
   multiple lines. */
```

- comments of `/*...*/` style *do not nest*

```
/*  
/* This sentence is part of a comment. */  
This sentence is not part of any comment and  
will probably cause a compile error.  
*/
```

# Identifiers

- identifiers used to name entities such as: types, objects (i.e., variables), and functions
- valid identifier is sequence of one or more letters, digits, and underscore characters that does not begin with a digit [C++17 5.10/1]
- identifiers that begin with underscore (in many cases) or contain double underscores are reserved for use by C++ implementation and should be avoided [C++17 5.10/3]
- examples of valid identifiers:
  - `event_counter`
  - `eventCounter`
  - `sqrt_2`
  - `f_o_o_b_a_r_4_2`
- identifiers are case sensitive (e.g., `counter` and `cOuNtEr` are distinct identifiers)
- identifiers cannot be any of reserved keywords (see next slide)
- **scope** of identifier is context in which identifier is valid (e.g., block, function, global)

# Reserved Keywords

alignas	default	noexcept	this
alignof	delete	not	thread_local
and	do	not_eq	throw
and_eq	double	nullptr	true
asm	dynamic_cast	operator	try
auto	else	or	typedef
bitand	enum	or_eq	typeid
bitor	explicit	private	typename
bool	export	protected	union
break	extern	public	unsigned
case	false	register	using
catch	float	reinterpret_cast	virtual
char	for	return	void
char16_t	friend	short	volatile
char32_t	goto	signed	wchar_t
class	if	sizeof	while
compl	inline	static	xor
const	int	static_assert	xor_eq
constexpr	long	static_cast	override*
const_cast	mutable	struct	final*
continue	namespace	switch	
decltype	new	template	

\* Note: context sensitive

## Section 2.3.1

# **Preprocessor**

# The Preprocessor

- prior to compilation, source code transformed by preprocessor
- preprocessor output then passed to compiler for compilation
- preprocessor behavior can be controlled by preprocessor directives
- preprocessor directive occupies single line and consists of:
  - 1 hash character (i.e., “#”)
  - 2 preprocessor instruction (i.e., define, undef, include, if, ifdef, ifndef, else, elif, endif, line, error, **and** pragma)
  - 3 arguments (depending on instruction)
  - 4 line break
- preprocessor can be used to:
  - conditionally compile parts of source file
  - define macros and perform macro expansion
  - include other files
  - force error to be generated

# Source-File Inclusion

- can include contents of another file in source using preprocessor **#include** directive

- syntax:

```
#include <path_specifier>
```

or

```
#include "path_specifier"
```

- *path\_specifier* is pathname (which may include directory) identifying file whose content is to be substituted in place of include directive
- typically, angle brackets used for system header files and double quotes used otherwise
- example:

```
#include <iostream>  
#include <boost/tokenizer.hpp>  
#include "my_header_file.hpp"  
#include "some_directory/my_header_file.hpp"
```

# Defining Macros

- can define macros using **#define** directive
- syntax:

**#define** *name value*

- *name* is name of macro and *value* is value of macro
- example:

```
#define DEBUG_LEVEL 10
```

- macros can also take arguments
- generally, macros should be avoided when possible (i.e., when other better mechanisms are available to achieve desired effect)
- for example, although macros can be used as way to accomplish inlining of functions, such usage should be avoided since language mechanism exists for specifying inline functions



# Conditional Compilation

- can conditionally include code through use of if-elif-else construct
- conditional preprocessing block consists of following (in order)
  - 1 **#if**, **#ifdef**, or **#ifndef** directive
  - 2 optionally any number of **#elif** directives
  - 3 at most one **#else** directive
  - 4 **#endif** directive
- code in taken branch of if-elif-else construct passed to compiler, while code in other branches discarded
- example:

```
#if DEBUG_LEVEL == 1
// ...
#elif DEBUG_LEVEL == 2
// ...
#else
// ...
#endif
```

# Preprocessor Predicate `__has_include`

- preprocessor predicate `__has_include` can be used in expressions for preprocessor to test for existence of header files
- example:

```
#ifndef __has_include
#   if __has_include(<optional>)
#       include <optional>
#       define have_optional 1
#   elif __has_include(<experimental/optional>)
#       include <experimental/optional>
#       define have_optional 1
#       define experimental_optional
#   else
#       define have_optional 0
#   endif
#endif
```

## Section 2.3.2

# **Objects, Types, and Values**

# Fundamental Types

- boolean type: `bool`
- character types:
  - `char` (may be signed or unsigned)
  - `signed char`
  - `unsigned char`
  - `char16_t`
  - `char32_t`
  - `wchar_t`
- `char` is distinct type from `signed char` and `unsigned char`
- standard signed integer types: [\[C++17 6.9.1/2\]](#)
  - `signed char`
  - `signed short int`
  - `signed int`
  - `signed long int`
  - `signed long long int`
- standard unsigned integer types:
  - `unsigned char`
  - `unsigned short int`
  - `unsigned int`
  - `unsigned long int`
  - `unsigned long long int`

# Fundamental Types (Continued)

- “**int**” may be omitted from names of (non-character) integer types (e.g., “**unsigned**” equivalent to “**unsigned int**” and “**signed**” equivalent to “**signed int**”)
- “**signed**” may be omitted from names of signed integer types, excluding **signed char** (e.g., “**int**” equivalent to “**signed int**”)
- boolean, character, and (signed and unsigned) integer types collectively called **integral types**
- integral types must use binary positional representation; two’s complement, one’s complement, and sign magnitude representations permitted [\[C++17 6.9.1/7\]](#)
- floating-point types:
  - **float**
  - **double**
  - **long double**
- void (i.e., incomplete/valueless) type: **void**
- null pointer type: `std::nullptr_t` (defined in header file `cstdint`)

- **literal** (a.k.a. literal constant) is value written exactly as it is meant to be interpreted
- examples of literals:

`"Hello, world"`

`"Bjarne"`

`'a'`

`'A'`

`123`

`123U`

`1'000'000'000`

`3.1415`

`1.0L`

`1.23456789e-10`

# Character Literals

- character literal consists of optional prefix followed by one or more characters enclosed in single quotes
- type of character literal determined by prefix (or lack thereof) as follows:

Prefix	Literal	Type
None	ordinary	normally <b>char</b> (in special cases <b>int</b> )
u8	UTF-8	<b>char</b>
u	UCS-2	<b>char16_t</b>
U	UCS-4	<b>char32_t</b>
L	wide	<b>wchar_t</b>

- special characters can be represented by escape sequence:

Character	Escape Sequence
newline (LF)	<code>\n</code>
horizontal tab (HT)	<code>\t</code>
vertical tab (VT)	<code>\v</code>
backspace (BS)	<code>\b</code>
carriage return (CR)	<code>\r</code>
form feed (FF)	<code>\f</code>
alert (BEL)	<code>\a</code>
backslash (\)	<code>\\</code>

Character	Escape Sequence
question mark (?)	<code>\?</code>
single quote (')	<code>\'</code>
double quote (")	<code>\"</code>
octal number ooo	<code>\ooo</code>
hex number hhh	<code>\xhhh</code>
code point nnnn	<code>\unnnn</code>
code point nnnnnnnn	<code>\Unnnnnnnn</code>

- examples of character literals:

`'a'`   `'1'`   `'!'`   `'\n'`   `u'a'`   `U'a'`   `L'a'`   `u8'a'`

## Character Literals (Continued)

- decimal digit characters guaranteed to be consecutive in value (e.g., `'1'` must equal `'0' + 1`) [\[C++17 5.3/3\]](#)
- in case of ordinary character literals, alphabetic characters are *not* guaranteed to be consecutive in value (e.g., `'b'` is not necessarily `'a' + 1`)



# String Literals

- (non-raw) string literal consists of optional prefix followed by zero or more characters enclosed in double quotes
- string literal has character array type
- type of string literal determined by prefix (or lack thereof) as follows:

Prefix	Literal	Type
None	narrow	<b>const char[]</b>
u8	UTF-8	<b>const char[]</b>
u	UTF-16	<b>const char16_t[]</b>
U	UTF-32	<b>const char32_t[]</b>
L	wide	<b>const wchar_t[]</b>

- examples of string literals:

"Hello, World!\n"

"123"

"ABCDEFGH"

- adjacent string literals are concatenated (e.g., "Hel" "lo" equivalent to "Hello")
- string literals implicitly terminated by null character (i.e., '\0')
- so, for example, "Hi" means 'H' followed by 'i' followed by '\0'

# Raw String Literals

- interpretation of escape sequences (e.g., “\n”) inside string literal can be avoided by using raw literal
- raw literal has form:
  - *prefix* `R"delimiter (raw_characters) delimiter"`
- optional *prefix* is string-literal prefix (e.g., `u8`)
- optional *delimiter* is sequence of characters used to assist in delimiting string
- *raw\_characters* is sequence of characters comprising string
- escape sequences not processed inside raw literal
- raw literal can also contain newline characters
- examples of raw string literals:

```
R"(He said, "No.") "  
u8R"(He said, "No.") "  
R"foo(The answer is 42.)foo"  
R"((+|-)?[[:digit:]]+)"
```

# Integer Literals

- can be specified in decimal, binary, hexadecimal, and octal
- number base indicated by prefix (or lack thereof) as follows:

Prefix	Number Base
None	decimal
Leading 0	octal
0b or 0B	binary
0x or 0X	hexadecimal

- various suffixes can be specified to control type of literal:

- `u` or `U`
- `l` or `L`
- both `u` or `U` and `l` or `L`
- `ll` or `LL`
- both `u` or `U` and `ll` or `LL`

- can use single quote as digit separator (e.g., `1'000'000`)

- examples of integer literals:

`42`

`1'000'000'000'000ULL`

`0xdeadU`

- integer literal always nonnegative; so, for example, `-1` is integer literal `1` with negation operation applied [\[C++17 5.13.2\]](#) [\[C++17 5.13.4\]](#)

# Integer Literals (Continued)

Suffix	Decimal Literal	Non-Decimal Literal
None	<code>int</code> <code>long int</code> <code>long long int</code>	<code>int</code> <code>unsigned int</code> <code>long int</code> <code>unsigned long int</code> <code>long long int</code> <code>unsigned long long int</code>
<code>u</code> or <code>U</code>	<code>unsigned int</code> <code>unsigned long int</code> <code>unsigned long long int</code>	<code>unsigned int</code> <code>unsigned long int</code> <code>unsigned long long int</code>
<code>l</code> or <code>L</code>	<code>long int</code> <code>long long int</code>	<code>long int</code> <code>unsigned long int</code> <code>long long int</code> <code>unsigned long long int</code>
Both <code>u</code> or <code>U</code> and <code>l</code> or <code>L</code>	<code>unsigned long int</code> <code>unsigned long long int</code>	<code>unsigned long int</code> <code>unsigned long long int</code>
<code>ll</code> or <code>LL</code>	<code>long long int</code>	<code>long long int</code> <code>unsigned long long int</code>
Both <code>u</code> or <code>U</code> and <code>ll</code> or <code>LL</code>	<code>unsigned long long int</code>	<code>unsigned long long int</code>

# Floating-Point Literals

- type of literal indicated by suffix (or lack thereof) as follows:

Suffix	Type
None	<b>double</b>
f or F	<b>float</b>
l or L	<b>long double</b>

- examples of **double** literals:

1.414

1.25e-8

- examples of **float** literals:

1.414f

1.25e-8f

- examples of **long double** literals:

1.5L

1.25e-20L

- floating-point literals always nonnegative; so, for example, `-1.0` is literal `1.0` with negation operator applied

# Hexadecimal Floating-Point Literals

- hexadecimal floating-point literal has general form:

- 1 prefix `0x` or `0X`
- 2 hexadecimal digits for integer part of number (optional if at least one digit after radix point)
- 3 period character (i.e., radix point)
- 4 hexadecimal digits for fractional part of number (optional if at least one digit before radix point)
- 5 `p` character (which designates exponent to follow)
- 6 one or more decimal digits for base-16 exponent
- 7 optional floating-point literal suffix (e.g., `f` or `l`)

- examples of hexadecimal floating-point literals:

Literal	Type	Value (Decimal)
<code>0x.8p0</code>	<b>double</b>	0.5
<code>0x10.cp0</code>	<b>double</b>	16.75
<code>0x.8p0f</code>	<b>float</b>	0.5
<code>0xf.fp0f</code>	<b>float</b>	15.9375
<code>0x1p10L</code>	<b>long double</b>	1024

# Boolean and Pointer Literals

- boolean literals:

**true**

**false**

- pointer literal:

**nullptr**

# Declarations and Definitions

- **declaration** introduces identifier for type, object (i.e., variable), or function (without necessarily providing full information about identifier)
  - in case of object, specifies type (of object)
  - in case of function, specifies number of parameters, type of each parameter, and type of return value (if not automatically deduced)
- each identifier must be declared before it can be used (i.e., referenced)
- **definition** provides full information about identifier and causes entity associated with identifier (if any) to be created
  - in case of type, provides full details about type
  - in case of object, causes storage to be allocated for object and object to be created
  - in case of function, provides code for function body
- in case of objects, in most (but not all) contexts, declaring object also defines it
- can declare identifier multiple times but can define only once
- above terminology often abused, with “declaration” and “definition” being used interchangeably



# Examples of Declarations and Definitions

```
int count; // declare and define count
extern double alpha; // (only) declare alpha

void func() { // declare and define func
    int n; // declare and define n
    double x = 1.0; // declare and define x
    // ...
}

bool isOdd(int); // declare isOdd
bool isOdd(int x); // declare isOdd (x ignored)

bool isOdd(int x) { // declare and define isOdd
    return x % 2;
}

struct Thing; // declare Thing

struct Vector2 { // declare and define Vector2
    double x;
    double y;
};
```

# Variable Declarations and Definitions

- **variable declaration** (a.k.a. object declaration) introduces identifier that names object and specifies type of object
- **variable definition** (a.k.a. object definition) provides all information included in variable declaration and also causes object to be created (e.g., storage allocated for object)
- example:

```
int count;  
    // declare and define count  
double alpha;  
    // declare and define alpha  
extern double gamma;  
    // declare (but do not define) gamma
```

# Arrays

- **array** is collection of one or more objects of *same* type that are stored *contiguously* in memory
- each element in array identified by (unique) integer index, with indices starting from *zero*
- array denoted by []

- example:

```
double x[10]; // array of 10 doubles
int data[512][512]; // 512 by 512 array of ints
```

- elements of array accessed using subscripting operator []

- example:

```
int x[10];
// elements of arrays are x[0], x[1], ..., x[9]
```

- often preferable to use user-defined type for representing array instead of array type
- for example, `std::array` and `std::vector` types (to be discussed later) have numerous practical advantages over array types

# Array Example

- code:

```
int a[4] = {1, 2, 3, 4};
```

- assumptions (for some completely fictitious C++ language implementation):

- **sizeof(int)** is 4
- array **a** starts at address 1000

- memory layout:

Address		Name
1000	1	a[0]
1004	2	a[1]
1008	3	a[2]
1012	4	a[3]

# Pointers

- **pointer** is object whose value is address in memory where another object is stored
- pointer to object of type  $T$  denoted by  $T^*$
- **null pointer** is special pointer value that does not refer to any valid memory location
- null pointer value provided by **nullptr** keyword
- accessing object to which pointer refers called **dereferencing**
- dereferencing pointer performed by **indirection operator** (i.e., “\*”)
- if  $p$  is pointer,  $*p$  is object to which pointer refers
- if  $x$  is object of type  $T$ ,  $\&x$  is (normally) address of object, which has type  $T^*$
- example:

```
char c;  
char* cp = nullptr; // cp is pointer to char  
char* cp2 = &c; // cp2 is pointer to char
```

# Pointer Example

## ■ code:

```
int i = 42;  
int* p = &i;  
assert(*p == 42);
```

## ■ assumptions (for some completely fictitious C++ language implementation):

- `sizeof(int)` is 4
- `sizeof(int*)` is 4
- `&i` is `((int*)1000)`
- `&p` is `((int*)1004)`

## ■ memory layout:

Address		Name
1000	42	i
1004	1000	p

# References

- **reference** is alias (i.e., nickname) for *already existing* object
- two kinds of references:
  - 1 lvalue reference
  - 2 rvalue reference
- lvalue reference to object of type `T` denoted by `T&`
- rvalue reference to object of type `T` denoted by `T&&`
- initializing reference called **reference binding**
- lvalue and rvalue references differ in their binding properties (i.e., to what kinds of objects reference can be bound)
- in most contexts, lvalue references usually needed
- rvalue references used in context of move constructors and move assignment operators (to be discussed later)
- example:

```
int x;  
int& y = x; // y is lvalue reference to int  
int&& tmp = 3; // tmp is rvalue reference to int
```

# References Example

- code:

```
int i = 42;  
int& j = i;  
assert(j == 42);
```

- assumptions (for some completely fictitious C++ language implementation):

- `sizeof(int)` is 4
- `&i` is `((int*)1000)`

- memory layout:

Address		Name
1000	42	i, j



# References Versus Pointers

- references and pointers similar in that both can be used to refer to some other entity (e.g., object or function)
- two key differences between references and pointers:
  - 1 reference must refer to something, while pointer can have null value (**`nullptr`**)
  - 2 references cannot be rebound, while pointers can be changed to point to different entity
- references have cleaner syntax than pointers, since pointers must be dereferenced upon each use (and dereference operations tend to clutter code)
- use of pointers often implies need for memory management (i.e., memory allocation, deallocation, etc.), and memory management can introduce numerous kinds of bugs when done incorrectly
- often faced with decision of using pointer or reference in code
- generally advisable to prefer use of references over use of pointers unless compelling reason to do otherwise, such as:
  - must be able to handle case of referring to nothing
  - must be able to change entity being referenced

# Unscoped Enumerations

- **enumerated type** provides way to describe range of values that are represented by named constants called **enumerators**
- object of enumerated type can take any one of enumerators as value
- enumerator values represented by some *integral type*
- enumerator can be assigned specific value (which may be negative)
- if enumerator not assigned specific value, value defaults to zero if first enumerator in enumeration and one greater than value for previous enumerator otherwise
- example:

```
enum Suit {  
    Clubs, Diamonds, Hearts, Spades  
};  
  
Suit suit = Clubs;
```

- example:

```
enum Suit {  
    Clubs = 1, Diamonds = 2, Hearts = 4, Spades = 8  
};
```

# Scoped Enumerations

- scoped enumeration similar to unscoped enumeration, except:
  - all enumerators are placed in scope of enumeration itself
  - integral type used to hold enumerator values can be explicitly specified
  - conversions involving scoped enumerations are stricter (i.e., more type safe)
- **class** or **struct** added after **enum** keyword to make enumeration scoped
- scope resolution operator (i.e., “:”) used to access enumerators
- scoped enumerations should probably be preferred to unscoped ones
- example:

```
enum struct Season {  
    spring, summer, fall, winter  
};  
  
enum struct Suit : unsigned char {  
    clubs, diamonds, hearts, spades  
};  
  
Season season = Season::summer;  
Suit suit = Suit::spades;
```

# Type Aliases with **typedef** Keyword

- **typedef** keyword used to create alias for existing type
- example:

```
typedef long long BigInt;  
BigInt i; // i has type long long
```

```
typedef char* CharPtr;  
CharPtr p; // p has type char*
```

# Type Aliases with **using** Statement

- **using** statement can be used to create alias for existing type
- probably preferable to use **using** statement over **typedef**
- example:

```
using BigInt = long long;  
BigInt i; // i has type long long
```

```
using CharPtr = char*;  
CharPtr p; // p has type char*
```

# The **extern** Keyword

- **translation unit**: basic unit of compilation in C++ (i.e., single source code file plus all of its directly and indirectly included header files)
- **extern** keyword used to declare object/function in separate translation unit
- example:

```
extern int evil_global_variable;  
    // declaration only  
    // actual definition in another file
```

# The `const` Qualifier

SKIP SLIDE

- `const` qualifier specifies that object has value that is *constant* (i.e., cannot be changed)
- qualifier that applies to object itself said to be *top level*
- following defines `x` as `int` with value 42 that cannot be modified:

```
const int x = 42;
```

- example:

```
const int x = 42;  
x = 13; // ERROR: x is const  
const int& x1 = x; // OK  
const int* p1 = &x; // OK  
int& x2 = x; // ERROR: x const, x2 not const  
int* p2 = &x; // ERROR: x const, *p2 not const
```

- example:

```
int x = 0;  
const int& y = x;  
x = 42; // OK  
// y also changed to 42 since y refers to x  
// y cannot be used to change x, however  
// i.e., the following would cause compile error:  
// y = 24; // ERROR: y is const
```

# The **const** Qualifier and Non-Pointer/Non-Reference Types

- with types that are not pointer or reference types, **const** can only be applied to object itself (i.e., top level)
- that is, object itself may be const or non-const
- example:

```
int i = 0; // object i is modifiable  
i = 42; // OK: i can be modified  
const int ci = 0; // object ci is not modifiable  
ci = 42; // ERROR: ci cannot be modified
```



# Example: **const** Qualifier and Non-Pointer/Non-Reference types

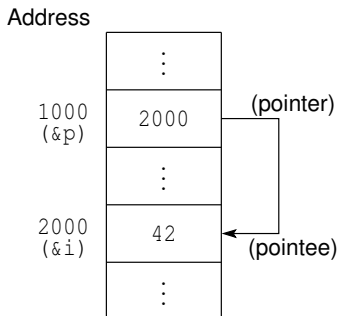
SKIP SLIDE

```
1 // with types that are not pointer or reference types, const
2 // can only be applied to object itself (i.e., top level)
3 // object itself may be const or non-const
4
5 int i = 0; // non-const int object
6 const int ci = 0; // const int object
7
8 i = 42; // OK: can modify non-const object
9 ci = 42; // ERROR: cannot modify const object
10
11 i = ci; // OK: can modify non-const object
12 ci = i; // ERROR: cannot modify const object
```

# The **const** Qualifier and Pointer Types

- every pointer is associated with two objects: pointer itself and pointee (i.e., object to which pointer points)
- **const** qualifier can be applied to each of pointer (i.e., top-level qualifier) and pointee

```
int i = 42; // pointee  
  
// p is pointer to int i  
// for example:  
// int* p = &i;  
// const int* p = &i;  
// int* const p = &i;  
// const int* const p = &i;
```



# Example: **const** Qualifier and Pointer Types

SKIP SLIDE

```
1  // with pointer types, const can be applied to each of:
2  // pointer and pointee
3  // pointer itself may be const or non-const (top-level)
4  // pointee may be const or non-const
5
6  int i = 0;
7  int j = 0;
8
9  int* pi = &i; // non-const pointer to a non-const int
10 pi = &j; // OK: can modify non-const pointer
11 *pi = 42; // OK: can modify non-const pointee
12
13 const int* pci = &i; // non-const pointer to a const int
14 // equivalently: int const* pci = &i;
15 pci = &j; // OK: can modify non-const pointer
16 *pci = 42; // ERROR: cannot modify const pointee
17
18 int* const cpi = &i; // const pointer to a non-const int
19 cpi = &j; // ERROR: cannot modify const pointer
20 *cpi = 42; // OK: can modify non-const pointee
21
22 const int* const cpci = &i; // const pointer to a const int
23 // equivalently: int const* const cpci = &i;
24 cpci = &j; // ERROR: cannot modify const pointer
25 *cpci = 42; // ERROR: cannot modify const pointee
26
27 pci = pi; // OK: adds const to pointee
28 pi = pci; // ERROR: discards const from pointee
```

# The **const** Qualifier and Reference Types

- reference is name that refers to object (i.e., referee)
- in principle, **const** qualifier can be applied to reference itself (i.e., top-level qualifier) or referee
- since reference cannot be rebound, reference itself is effectively always constant
- for this reason, does not make sense to explicitly apply **const** as top-level qualifier for reference type and language disallows this
- **const** qualifier can only be applied to referee
- example:

```
int j = 0;
int k = 42;
int& i = j;
    // i is reference; j is referee
    // referee is modifiable
const int& ci = j;
    // ci is reference; j is referee
    // referee is not modifiable
const int& ci = k; // ERROR: cannot redefine/rebind
int& const r = j;
    // ERROR: reference itself cannot be specified as const
```

# Example: **const** Qualifier and Reference Types

SKIP SLIDE

```
1  // with reference types, const can only be applied to referee
2  // reference itself cannot be rebound (i.e., is constant)
3  // referee may be const or non-const
4
5  int i = 0; const int ci = 0;
6  int il = 0; const int cil = 0;
7
8  // reference to non-const int
9  int& ri = i;
10 ri = ci; // OK: can modify non-const referee
11 int& ri = il; // ERROR: cannot redefine/rebind reference
12
13 // reference to const int
14 const int& rci = ci;
15 rci = i; // ERROR: cannot modify const referee
16 const int& rci = cil;
17     // ERROR: cannot redefine/rebind reference
18
19 // ERROR: reference itself cannot be const qualified
20 int& const cri = i; // ERROR: invalid const qualifier
21
22 // ERROR: reference itself cannot be const qualified
23 const int& const crci = ci; // ERROR: invalid const qualifier
24 // also: int const& const crci = ci; // ERROR
25
26 const int& rl = ci; // OK: adds const to referee
27 int& r2 = ci; // ERROR: discards const from referee
```

NEXT SLIDE: **constexpr** for variables

# The **const** Qualifier and Pointer-to-Pointer Types

- for given type **T**, cannot implicitly convert **T\*\*** to **const T\*\***
- although such conversion looks okay at first glance, actually would create backdoor for changing **const** objects
- can, however, implicitly convert **T\*\*** to **const T\* const\***
- for example, code like that shown below could be used to change **const** objects if **T\*\*** to **const T\*\*** were valid conversion:

```
const int i = 42;
int* p;
const int** q = &p;
    // Fortunately, this line is not valid code.
    // ERROR: cannot convert int** to const int**
*q = &i;
    // Change p (to which q points) to point to i.
    // OK: *q is not const (only **q is const)
*p = 0;
    // Set i (to which p points) to 0.
    // OK: *p is not const
    // This line would change i, which is const.
```

# The **volatile** Qualifier

- **volatile** qualifier used to indicate that object can change due to agent *external to program* (e.g., memory-mapped device, signal handler)
- compiler cannot optimize away read and write operations on **volatile** objects (e.g., repeated reads without intervening writes cannot be optimized away)
- **volatile** qualifier typically used when object:
  - corresponds to register of memory-mapped device
  - may be modified by signal handler (namely, object of type **volatile** `std::sig_atomic_t`)
- example:

```
volatile int x;  
volatile unsigned char* deviceStatus;
```

# The **auto** Keyword

- in various contexts, **auto** keyword can be used as place holder for type
- in such contexts, implication is that compiler must deduce type
- example:

```
auto i = 3; // i has type int  
auto j = i; // j has type int  
auto& k = i; // k has type int&  
const auto& n = i; // n has type const int&  
auto x = 3.14; // x has type double
```

- very useful in generic programming (covered later) when types not always easy to determine
- can potentially save typing long type names
- can lead to more readable code (if well used)
- if overused, can lead to bugs (sometimes very subtle ones) and difficult to read code



- **inline variable**: variable that may be defined in multiple translation units as long as all definitions are identical
- potential for multiple definitions avoided by having linker simply choose one of identical definitions and discard others (if more than one exists)
- can request that variable be made inline by including **inline** qualifier in variable declaration
- inline variable must have static storage duration (e.g., static class member or namespace-scope variable)
- inline variable typically used to allow definition of variable to be placed in header file without danger of multiple definitions
- inline variable has same address in all translation units

# Inline Variable: Example

inline\_variable\_1\_1.hpp

```
1 inline int magic = 42;
```

main.cpp

```
1 #include <iostream>
2 #include "inline_variable_1_1.hpp"
3 int main() {
4     std::cout << magic << "\n";
5 }
```

other.cpp

```
1 #include "inline_variable_1_1.hpp"
2 void func() { /* ... */ }
```

## Section 2.3.3

# **Operators and Expressions**

## Arithmetic Operators

Operator Name	Syntax
addition	<code>a + b</code>
subtraction	<code>a - b</code>
unary plus	<code>+a</code>
unary minus	<code>-a</code>
multiplication	<code>a * b</code>
division	<code>a / b</code>
modulo (i.e., remainder)	<code>a % b</code>
pre-increment	<code>++a</code>
post-increment	<code>a++</code>
pre-decrement	<code>--a</code>
post-decrement	<code>a--</code>

## Bitwise Operators

Operator Name	Syntax
bitwise NOT	<code>~a</code>
bitwise AND	<code>a &amp; b</code>
bitwise OR	<code>a   b</code>
bitwise XOR	<code>a ^ b</code>
arithmetic left shift	<code>a &lt;&lt; b</code>
arithmetic right shift	<code>a &gt;&gt; b</code>

## Assignment and Compound-Assignment Operators

Operator Name	Syntax
assignment	<code>a = b</code>
addition assignment	<code>a += b</code>
subtraction assignment	<code>a -= b</code>
multiplication assignment	<code>a *= b</code>
division assignment	<code>a /= b</code>
modulo assignment	<code>a %= b</code>
bitwise AND assignment	<code>a &amp;= b</code>
bitwise OR assignment	<code>a  = b</code>
bitwise XOR assignment	<code>a ^= b</code>
arithmetic left shift assignment	<code>a &lt;&lt;= b</code>
arithmetic right shift assignment	<code>a &gt;&gt;= b</code>

# Operators (Continued 2)

## Logical/Relational Operators

Operator Name	Syntax
equal	<code>a == b</code>
not equal	<code>a != b</code>
greater than	<code>a &gt; b</code>
less than	<code>a &lt; b</code>
greater than or equal	<code>a &gt;= b</code>
less than or equal	<code>a &lt;= b</code>
logical negation	<code>!a</code>
logical AND	<code>a &amp;&amp; b</code>
logical OR	<code>a    b</code>

## Member and Pointer Operators

Operator Name	Syntax
array subscript	<code>a[b]</code>
indirection	<code>*a</code>
address of	<code>&amp;a</code>
member selection	<code>a.b</code>
member selection	<code>a-&gt;b</code>
member selection	<code>a.*b</code>
member selection	<code>a-&gt;*b</code>

## Other Operators

Operator Name	Syntax
function call	<code>a(...)</code>
comma	<code>a, b</code>
ternary conditional	<code>a ? b : c</code>
scope resolution	<code>a::b</code>
sizeof	<b>sizeof</b> (a)
parameter-pack sizeof	<b>sizeof...</b> (a)
alignof	<b>alignof</b> (T)
allocate storage	<b>new</b> T
allocate storage (array)	<b>new</b> T[a]
deallocate storage	<b>delete</b> a
deallocate storage (array)	<b>delete</b> [] a

## Other Operators (Continued)

Operator Name	Syntax
type ID	<b>typeid</b> (a)
type cast	(T) a
const cast	<b>const_cast</b> <T>(a)
static cast	<b>static_cast</b> <T>(a)
dynamic cast	<b>dynamic_cast</b> <T>(a)
reinterpret cast	<b>reinterpret_cast</b> <T>(a)
throw	<b>throw</b> a
noexcept	<b>noexcept</b> (e)



# Operator Precedence

Precedence	Operator	Name	Associativity
1	::	scope resolution	none
2	. -> [] () ++ --	member selection (object) member selection (pointer) subscripting function call post-increment post-decrement	left to right

# Operator Precedence (Continued 1)

Precedence	Operator	Name	Associativity
3	<b>sizeof</b> ++ -- ~ ! - + & * <b>new</b> <b>new</b> [] <b>delete</b> <b>delete</b> [] ( )	size of object/type pre-increment pre-decrement bitwise NOT logical NOT unary minus unary plus address of indirection allocate storage allocate storage (array) deallocate storage deallocate storage (array) cast	right to left

## Operator Precedence (Continued 2)

Precedence	Operator	Name	Associativity
4	. * -> *	member selection (objects) member selection (pointers)	left to right
5	* / %	multiplication division modulus	left to right
6	+ -	addition subtraction	left to right
7	<< >>	left shift right shift	left to right
8	< <= > >=	less than less than or equal greater than greater than or equal	left to right
9	== !=	equality inequality	left to right

## Operator Precedence (Continued 3)

Precedence	Operator	Name	Associativity
10	&	bitwise AND	left to right
11	^	bitwise XOR	left to right
12		bitwise OR	left to right
13	&&	logical AND	left to right
14		logical OR	left to right
15	? :	ternary conditional	right to left

## Operator Precedence (Continued 4)

Precedence	Operator	Name	Associativity
16	= *= /= %= += -= <<= >>= &=  = ^=	assignment multiplication assignment division assignment modulus assignment addition assignment subtraction assignment left shift assignment right shift assignment bitwise AND assignment bitwise OR assignment bitwise XOR assignment	right to left
17	<b>throw</b>	throw exception	right to left
18	,	comma	left to right

# Alternative Tokens

Alternative	Primary
<b>and</b>	<code>&amp; &amp;</code>
<b>bitor</b>	<code> </code>
<b>or</b>	<code>  </code>
<b>xor</b>	<code>^</code>
<b>compl</b>	<code>~</code>
<b>bitand</b>	<code>&amp;</code>
<b>and_eq</b>	<code>&amp;=</code>
<b>or_eq</b>	<code> =</code>
<b>xor_eq</b>	<code>^=</code>
<b>not</b>	<code>!</code>
<b>not_eq</b>	<code>!=</code>

- alternative tokens above probably best avoided as they lead to more verbose code

# Expressions

- An **expression** is a sequence of operators and operands that specifies a computation. [\[C++17 8/1\]](#)
- An expression has a type and, if the type is not **void**, a value. [\[C++17 6.9.1/9\]](#)
- A **constant expression** is an expression that can be evaluated at compile time (e.g., `1 + 1`).
- Example:

```
int x = 0;
int y = 0;
int* p = &x;
double d = 0.0;
// Evaluate some
// expressions here.
```

Expression	Type	Value
<code>x</code>	<b>int</b>	0
<code>y = x</code>	<b>int&amp;</b>	reference to y
<code>x + 1</code>	<b>int</b>	1
<code>x * x + 2 * x</code>	<b>int</b>	0
<code>y = x * x</code>	<b>int&amp;</b>	reference to y
<code>x == 42</code>	<b>bool</b>	<b>false</b>
<code>*p</code>	<b>int&amp;</b>	reference to x
<code>p == &amp;x</code>	<b>bool</b>	<b>true</b>
<code>x &gt; 2 * y</code>	<b>bool</b>	<b>false</b>
<code>std::sin(d)</code>	<b>double</b>	0.0

# Operator Precedence/Associativity Example

Expression	Fully-Parenthesized Expression
<code>a + b + c</code>	<code>((a + b) + c)</code>
<code>a = b = c</code>	<code>(a = (b = c))</code>
<code>c = a + b</code>	<code>(c = (a + b))</code>
<code>d = a &amp;&amp; !b    c</code>	<code>(d = ((a &amp;&amp; (!b))    c))</code>
<code>++*p++</code>	<code>((++(*p++)))</code>
<code>a   ~b &amp; c ^ d</code>	<code>(a   (((~b) &amp; c) ^ d))</code>
<code>a[0]++ + a[1]++</code>	<code>((a[0]++) + ((a[1]++)))</code>
<code>a + b * c / d % - g</code>	<code>(a + (((b * c) / d) % (-g)))</code>
<code>++p[i]</code>	<code>((++p[i]))</code>
<code>--*++p</code>	<code>((--(*++p)))</code>
<code>a += b += c += d</code>	<code>(a += (b += (c += d)))</code>
<code>z = a == b ? ++c : --d</code>	<code>(z = ((a == b) ? (++c) : (--d)))</code>



# Division/Modulus Operator and Negative Numbers

- for integral operands, division operator yields algebraic quotient with any fractional part discarded (i.e., round towards zero)
- if quotient  $a / b$  is representable in type of result,  
 $(a / b) * b + a \% b$  is equal to  $a$
- so, assuming  $b$  is not zero and no overflow,  $a \% b$  equals  
 $a - (a / b) * b$
- result of modulus operator not necessarily nonnegative
- example:

```
1  static_assert (5 % 3 == 2);  
2  static_assert (5 % (-3) == 2);  
3  static_assert ((-5) % 3 == -2);  
4  static_assert ((-5) % (-3) == -2);
```

# Short-Circuit Evaluation

- logical-and operator (i.e., `&&`): [\[C++17 8.14\]](#)
  - groups left-to-right
  - result true if both operands are true, and false otherwise
  - second operand is *not evaluated* if first operand is false (in case of built-in logical-and operator)
- logical-or operator (i.e., `||`): [\[C++17 8.15\]](#)
  - groups left-to-right
  - result is true if either operand is true, and false otherwise
  - second operand is *not evaluated* if first operand is true (in case of built-in logical-or operator)
- example:

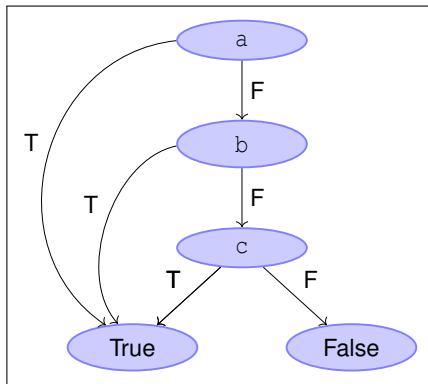
```
int x = 0;
bool b = (x == 0 || ++x == 1);
// b equals true; x equals 0
b = (x != 0 && ++x == 1);
// b equals false; x equals 0
```

- above behavior referred to as short circuit evaluation

# Short-Circuit Evaluation Example: `a || b || c`

- for three values `a`, `b`, `c` of type **bool**, consider evaluation of expression `a || b || c`
- code showing short-circuit evaluation and associated control-flow graph given below

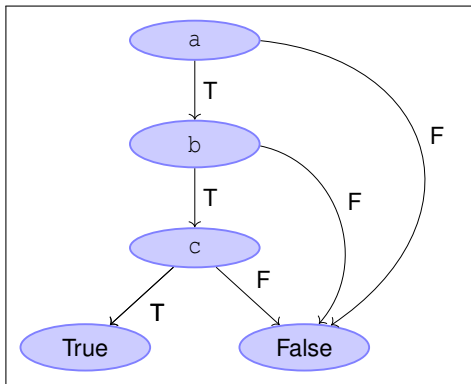
```
bool _result;  
if (a)  
    goto _true;  
if (b)  
    goto _true;  
if (c)  
    goto _true;  
_result = false;  
goto done;  
_true:  
_result = true;  
done:
```



# Short-Circuit Evaluation Example: `a && b && c`

- for three values `a`, `b`, `c` of type **bool**, consider evaluation of expression `a && b && c`
- code showing short-circuit evaluation and associated control-flow graph given below

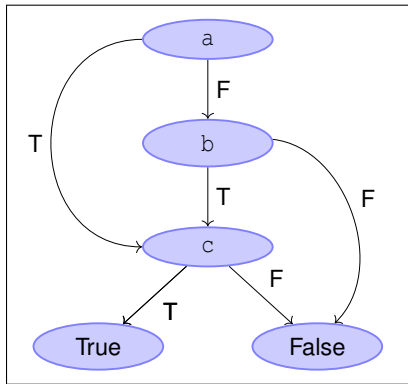
```
bool _result;  
if (!a)  
    goto _false;  
if (!b)  
    goto _false;  
if (!c)  
    goto _false;  
_result = true;  
goto done;  
_false:  
_result = false;  
done;
```



# Short-Circuit Evaluation Example: (a || b) && c

- for three values a, b, c of type **bool**, consider evaluation of expression (a || b) && c
- code showing short-circuit evaluation and associated control-flow graph given below ([example of compiler-generated assembly code](#))

```
bool _result;  
if (a)  
    goto _second;  
if (!b)  
    goto _false;  
_second:  
if (!c)  
    goto _false;  
_result = true;  
goto done;  
_false:  
_result = false;  
done:
```



# The `static_assert` Statement

- **`static_assert`** allows testing of boolean condition at compile time
- used to test sanity of code or test validity of assumptions made by code
- **`static_assert`** has two arguments:
  - 1 boolean constant expression (condition to test)
  - 2 string literal for error message to print if boolean expression not true
- second argument is optional
- failed static assertion results in compile error
- example:

```
static_assert(sizeof(int) >= 4, "int is too small");  
static_assert(1 + 1 == 2, "compiler is buggy");
```

# The **sizeof** Operator

- **sizeof** operator is used to query size of object or object type (i.e., amount of storage required)
- for object type **T**, **sizeof(T)** yields size of **T** in bytes (e.g., **sizeof(int)**, **sizeof(int[10])**)
- for expression **e**, **sizeof e** yields size of object required to hold result of **e** in bytes (e.g., **sizeof(&x)** where **x** is some object)
- **sizeof(char)**, **sizeof(signed char)**, and **sizeof(unsigned char)** guaranteed to be 1
- byte is at least 8 bits (usually exactly 8 bits except on more exotic platforms)

# The **constexpr** Qualifier for Variables

- **constexpr** qualifier indicates object has value that is *constant expression* (i.e., can be evaluated at compile time)
- **constexpr** implies **const** (but converse not necessarily true)
- following defines x as constant expression with type **const int** and value 42:

```
constexpr int x = 42;
```

- example:

```
constexpr int x = 42;  
int y = 1;  
x = 0; // ERROR: x is const  
const int& x1 = x; // OK  
const int* p1 = &x; // OK  
int& x2 = x; // ERROR: x const, x2 not const  
int* p2 = &x; // ERROR: x const, *p2 not const  
int a1[x]; // OK: x is constexpr  
int a2[y]; // ERROR: y is not constexpr
```



## Section 2.3.4

### **Control-Flow Constructs: Selection and Looping**

# The **if** Statement

- allows conditional execution of code

- syntax has form:

```
if (expression)  
    statement1  
else  
    statement2
```

- if expression *expression* is true, execute statement *statement*<sub>1</sub>; otherwise, execute statement *statement*<sub>2</sub>

- **else** clause can be omitted leading to simpler form:

```
if (expression)  
    statement1
```

- conditional execution based on more than one condition can be achieved using construct like:

```
if (expression1)  
    statement1  
else if (expression2)  
    statement2  
...  
else  
    statementn
```

# The **if** Statement (Continued 1)

- to include multiple statements in branch of **if**, *must group statements* into single statement using brace brackets

```
if (expression) {  
    statement1,1  
    statement1,2  
    statement1,3  
    ...  
} else {  
    statement2,1  
    statement2,2  
    statement2,3  
    ...  
}
```

- advisable to *always include brace brackets* even when not necessary, as this avoids potential bugs caused by forgetting to include brackets later when more statements added to branch of **if**

# The **if** Statement (Continued 2)

- if statement may include initializer:

```
if (initializer; expression)  
    statement1;  
else  
    statement2;
```

- above construct equivalent to:

```
{  
    initializer;  
    if (expression)  
        statement1;  
    else  
        statement2;  
}
```

- if condition in if statement is constant expression, **constexpr** keyword can be added after **if** keyword to yield what is called constexpr-if statement
- constexpr-if statement is evaluated at compile time and branch of if statement that is not taken is discarded

# The **if** Statement: Example

- example with **else** clause:

```
int x = someValue;
if (x % 2 == 0) {
    std::cout << "x is even\n";
} else {
    std::cout << "x is odd\n";
}
```

- example without **else** clause:

```
int x = someValue;
if (x % 2 == 0) {
    std::cout << "x is divisible by 2\n";
}
```

- example that tests for more than one condition:

```
int x = someValue;
if (x > 0) {
    std::cout << "x is positive\n";
} else if (x < 0) {
    std::cout << "x is negative\n";
} else {
    std::cout << "x is zero\n";
}
```

# The **if** Statement: Example

## ■ example with initializer:

```
int execute_command();  
if (int ret = execute_command(); ret == 0) {  
    std::cout << "command successful\n";  
} else {  
    std::cout << "command failed with status " <<  
        ret << '\n';  
}
```

## ■ example constexpr-if statement:

```
constexpr int x = 10;  
if constexpr (x < 0) {  
    std::cout << "negative\n";  
} else if constexpr (x > 0) {  
    std::cout << "positive\n";  
} else {  
    std::cout << "zero\n";  
}
```

# The **switch** Statement

- allows conditional execution of code based on integral/enumeration value
- syntax has form:

```
switch (expression) {  
  case const_expr1:  
    statements1  
  case const_expr2:  
    statements2  
  ...  
  case const_exprn:  
    statementsn  
  default:  
    statements  
}
```

- *expression* is expression of integral or enumeration type or implicitly convertible to such type; *const\_expr*<sub>*i*</sub> is constant expression of same type as *expression* after conversions/promotions
- if expression *expression* equals *const\_expr*<sub>*i*</sub>, jump to beginning of statements *statements*<sub>*i*</sub>; if expression *expr* does not equal *const\_expr*<sub>*i*</sub> for any *i*, jump to beginning of statements *statements*
- then, continue executing statements until **break** statement is encountered

# The **switch** Statement (Continued)

- switch statement can also include initializer:

```
switch (initializer; expression)  
    statement
```

- above construct equivalent to:

```
{  
    initializer;  
    switch (expression)  
        statement  
}
```

- remember that, in absence of break statement, execution in switch statement falls through from one case to next; if fall through not considered, bugs will result, such as in following code:

```
1  unsigned int x = 0;  
2  switch (x & 1) {  
3  case 0:  
4      std::cout << "x is even\n";  
5      // BUG: missing break statement  
6  case 1:  
7      std::cout << "x is odd\n";  
8      break;  
9  }
```



# The **switch** Statement: Example

## ■ example without initializer:

```
int x = someValue;
switch (x) {
case 0:
    // Note that there is no break here.
case 1:
    std::cout << "x is 0 or 1\n";
    break;
case 2:
    std::cout << "x is 2\n";
    break;
default:
    std::cout << "x is not 0, 1, or 2\n";
    break;
}
```

# The **switch** Statement: Example (Continued)

## ■ example with initializer:

```
int get_value();  
switch (int x = get_value(); x) {  
  case 0:  
  case 1:  
    std::cout << "x is 0 or 1\n";  
    break;  
  case 2:  
    std::cout << "x is 2\n";  
    break;  
  default:  
    std::cout << "x is not 0, 1, or 2\n";  
    break;  
}
```

# The **while** Statement

- looping construct
- syntax has form:

```
while (expression)  
    statement
```

- if expression *expression* is true, statement *statement* is executed; this process repeats until expression *expression* becomes false
- to allow multiple statements to be executed in loop body, **must group multiple statements** into single statement with brace brackets

```
while (expression) {  
    statement1  
    statement2  
    statement3  
    ...  
}
```

- advisable to **always use brace brackets**, even when loop body consists of only one statement

# The **while** Statement: Example

```
// print hello 10 times  
int n = 10;  
while (n > 0) {  
    std::cout << "hello\n";  
    --n;  
}
```

```
// loop forever, printing hello  
while (true) {  
    std::cout << "hello\n";  
}
```

# The **for** Statement

- looping construct
- has following syntax:

```
for (statement1; expression; statement2)  
    statement3
```

- first, execute statement *statement*<sub>1</sub>; then, while expression *expression* is true, execute statement *statement*<sub>3</sub> followed by statement *statement*<sub>2</sub>
- *statement*<sub>1</sub> and *statement*<sub>2</sub> may be omitted; *expression* treated as **true** if omitted
- to include multiple statements in loop body, *must group multiple statements* into single statement using brace brackets; advisable to *always use brace brackets*, even when loop body consists of only one statement:

```
for (statement1; expression; statement2) {  
    statement3,1  
    statement3,2  
    ...  
}
```

- any objects declared in *statement*<sub>1</sub> go out of scope as soon as **for** loop ends

# The **for** Statement (Continued)

- consider **for** loop:

```
for (statement1; expression; statement2)  
    statement3
```

- above **for** loop can be equivalently expressed in terms of **while** loop as follows (except for behavior of **continue** statement, yet to be discussed):

```
{  
    statement1;  
    while (expression) {  
        statement3  
        statement2;  
    }  
}
```

# The **for** Statement: Example

- example with single statement in loop body:

```
// Print the integers from 0 to 9 inclusive.  
for (int i = 0; i < 10; ++i)  
    std::cout << i << '\n';
```

- example with multiple statements in loop body:

```
int values[10];  
// ...  
int sum = 0;  
for (int i = 0; i < 10; ++i) {  
    // Stop if value is negative.  
    if (values[i] < 0) {  
        break;  
    }  
    sum += values[i];  
}
```

- example with error in assumption about scoping rules:

```
for (int i = 0; i < 10; ++i) {  
    std::cout << i << '\n';  
}  
++i; // ERROR: i no longer exists
```

# Range-Based **for** Statement

- variant of for loop for iterating over elements in range
- example:

```
int array[4] = {1, 2, 3, 4};  
// Triple the value of each element in the array.  
for (auto&& x : array) {  
    x *= 3;  
}
```

- range-based for loop nice in that it clearly expresses programmer intent (i.e., iterate over each element of collection)



# The **do** Statement

- looping construct
- has following general syntax:

```
do
    statement
while (expression);
```

- statement *statement* executed;  
then, expression *expression* evaluated;  
if expression *expression* is true, entire process repeats from beginning
- to execute multiple statements in body of loop, must group multiple statements into single statement using brace brackets

```
do {
    statement1
    statement2
    ...
} while (expression);
```

- advisable to *always use brace brackets*, even when loop body consists of only one statement

# The **do** Statement: Example

- example with single statement in loop body:

```
// delay by looping 10000 times  
int n = 0;  
do  
    ++n;  
while (n < 10000);
```

- example with multiple statements in loop body:

```
// print integers from 0 to 9 inclusive  
int n = 0;  
do {  
    std::cout << n << '\n';  
    ++n;  
} while (n < 10);
```

# The **break** Statement

- **break** statement causes enclosing loop or switch to be terminated immediately
- example:

```
// Read integers from standard input until an  
// error or end-of-file is encountered or a  
// negative integer is read.  
int x;  
while (std::cin >> x) {  
    if (x < 0) {  
        break;  
    }  
    std::cout << x << '\n';  
}
```

# The **continue** Statement

- **continue** statement causes next iteration of enclosing loop to be started immediately
- example:

```
int values[10];  
...  
// Print the nonzero elements of the array.  
for (int i = 0; i < 10; ++i) {  
    if (values[i] == 0) {  
        // Skip over zero elements.  
        continue;  
    }  
    // Print the (nonzero) element.  
    std::cout << values[i] << '\n';  
}
```

# The goto Statement

- **goto** statement transfers control to another statement specified by label
- should generally try to *avoid use of goto statement*
- well written code rarely has legitimate use for **goto** statement
- example:

```
int i = 0;
loop: // label for goto statement
do {
    if (i == 3) {
        ++i;
        goto loop;
    }
    std::cout << i << '\n';
    ++i;
} while (i < 10);
```

- some restrictions on use of **goto** (e.g., cannot jump over initialization in same block as **goto**)

```
goto skip; // ERROR
int i = 0;
skip:
++i;
```

## Section 2.3.5

# **Functions**

# Function Parameters, Arguments, and Return Values

- **argument** (a.k.a. **actual parameter**): argument is value supplied to function by caller; appears in parentheses of function-call operator [\[C++17 3.2\]](#)
- **parameter** (a.k.a. **formal parameter**): parameter is object/reference declared as part of function that acquires value on entry to function; appears in function definition/declaration [\[C++17 3.16\]](#)
- although abuse of terminology, parameter and argument often used interchangeably
- **return value**: result passed from function back to caller
- example:

```
int square(int i) { // i is parameter  
    return i * i; // return value is i * i  
}
```

```
void compute() {  
    int i = 3;  
    int j = square(i); // i is argument  
}
```

# Function Declarations and Definitions

- **function declaration** introduces identifier that names function and specifies following properties of function:
  - number of parameters
  - type of each parameter
  - type of return value (if not automatically deduced)

- example:

```
bool isOdd(int); // declare isOdd  
bool isOdd(int x); // declare isOdd (x ignored)
```

- **function definition** provides all information included in function declaration as well as code for body of function

- example:

```
bool isOdd(int x) { // declare and define isOdd  
    return x % 2;  
}
```



# Basic Syntax (Leading Return Type)

- most basic syntax for function declarations and definitions places return type at start (i.e., leading return-type syntax)
- basic syntax for function declaration:

*return\_type function\_name (parameter\_declarations) ;*

- examples of function declarations:

```
int min(int, int);  
double square(double);
```

- basic syntax for function definition:

```
return_type function_name (parameter_declarations)  
{  
    statements  
}
```

- examples of function definitions:

```
int min(int x, int y) {return x < y ? x : y;}  
double square(double x) {return x * x;}
```

# Trailing Return-Type Syntax

- with trailing return-type syntax, return type comes after parameter declarations and **auto** used as placeholder for where return type would normally be placed
- trailing return-type syntax for function declaration:

```
auto function_name (parameter_declarations) -> return_type;
```

- examples of function declarations:

```
auto min(int, int) -> int;  
auto square(double) -> double;
```

- trailing return-type syntax for function definition:

```
auto function_name (parameter_declarations) -> return_type  
{  
    statements  
}
```

- examples of function definitions:

```
auto min(int x, int y) -> int  
    {return x < y ? x : y;}  
auto square(double x) -> double {return x * x;}
```

# The **return** Statement

- **return** statement used to exit function, passing specified return value (if any) back to caller
- code in function executes until **return** statement is reached or execution falls off end of function
- if function return type is not **void**, **return** statement takes single parameter indicating value to be returned
- if function return type is **void**, function does not return any value and **return** statement takes either no parameter or expression of type **void**
- falling off end of function equivalent to executing **return** statement with no value
- example:

```
double unit_step(double x) {  
    if (x >= 0.0) {  
        return 1.0; // exit with return value 1.0  
    }  
    return 0.0; // exit with return value 0.0  
}
```

# Automatic Return-Type Deduction

- with both leading and trailing return-type syntax, can specify return type as **auto**
- in this case, return type of function will be automatically deduced
- if function definition has no **return** statement, return type deduced to be **void**
- otherwise, return type deduced to match type in expression of **return** statement or, if **return** statement has no expression, as **void**
- if multiple return statements, must use same type for all **return** expressions
- when return-type deduction used, function definition must be visible in order to call function (since return type cannot be determined otherwise)
- example:

```
auto square(double x) {  
    return x * x;  
    // x * x has type double  
    // deduced return type is double  
}
```

# The main Function

- entry point to program is always function called `main`
- has return type of `int`
- can be declared to take either no arguments or two arguments as follows (although other possibilities may also be supported by implementation):

[C++17 6.6.1/2]

```
int main();  
int main(int argc, char* argv[]);
```

- two-argument variant allows arbitrary number of C-style strings to be passed to program from environment in which program run
- `argc`: number of C-style strings provided to program
- `argv`: array of pointers to C-style strings
- `argv[0]` is name by which program invoked
- `argv[argc]` is guaranteed to be 0 (i.e., null pointer)
- `argv[1], argv[2], ..., argv[argc - 1]` typically correspond to command line options

# The main Function (Continued)

- suppose that following command line given to shell:

```
program one two three
```

- main function would be invoked as follows:

```
int argc = 4;
char* argv[] = {
    "program", "one", "two", "three", 0
};
main(argc, argv);
```

- return value of `main` typically passed back to operating system
- can also use function **void** `exit(int)` to terminate program, passing integer return value back to operating system
- return statement in `main` is optional [\[C++17 6.6.1/5\]](#)
- if control reaches end of `main` without encountering return statement, effect is that of executing **“return 0;”**

- **lifetime** of object is period of time in which object exists (e.g., block, function, global)

```
int x;
```

```
void wasteTime()
```

```
{
```

```
    int j = 10000;
```

```
    while (j > 0) {
```

```
        --j;
```

```
    }
```

```
    for (int i = 0; i < 10000; ++i) {
```

```
    }
```

```
}
```

- in above example: *x* global scope and lifetime; *j* function scope and lifetime; *i* block scope and lifetime

# Parameter Passing

- function parameter can be passed by value or by reference
- **pass by value**: function given copy of object from caller
- **pass by reference**: function given reference to object from caller
- to pass parameter by reference, use *reference type* for parameter
- example:

```
void increment(int& x)
    // x is passed by reference
{
    ++x;
}
```

```
double square(double x)
    // x is passed by value
{
    return x * x;
}
```



# Pass-By-Value Versus Pass-By-Reference

- if function needs to *change value of object in caller*, must pass by reference
- for example:

```
void increment(int& x)
    // x refers to object in caller
{
    ++x;
}
```

- if object being passed to function is *expensive to copy* (e.g., a very large data type), always faster to pass by reference
- for example:

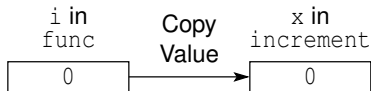
```
double compute(const std::vector<double>& x)
    // x refers to object in caller
    // object is not copied
{
    double result;
    // ... (initialize result with value computed from x)
    return result;
}
```

# Increment Example: Incorrectly Using Pass By Value

- consider code:

```
1 void increment(int x) {  
2     ++x;  
3 }  
4  
5 void func() {  
6     int i = 0;  
7     increment(i); // i is not modified  
8     // i is still 0  
9 }
```

- when func calls increment, parameter passing copies value of i in func to local variable x in increment:



- when code in increment executes, local variable x is incremented (not i in func):



# Increment Example: Correctly Using Pass By Reference

- consider code:

```
1 void increment(int& x) {  
2     ++x;  
3 }  
4  
5 void func() {  
6     int i = 0;  
7     increment(i); // i is incremented  
8     // i is now 1  
9 }
```

- when func calls increment, reference x in increment is bound to object i in func (i.e., x becomes alias for i):

i in func  
and  
x in increment

0

- when code in increment executes, x is incremented, which is alias for i in func:

i in func  
and  
x in increment

1

# The **const** Qualifier and Functions

- **const** qualifier can be used in function declaration to make promises about what non-local objects will not be modified by function
- for function parameter of pointer type, const-ness of pointed-to object (i.e., pointee) extremely important
- if pointee is const, function promises not to change pointee; for example:  

```
int strlen(const char*); // get string length
```
- for function parameter of reference type, const-ness of referred-to object (i.e., referee) extremely important
- if referee is const, function promises not to change referee; for example:  

```
std::complex<double>  
square(const std::complex<double>&);  
// compute square of number
```
- not making appropriate choice of const-ness for pointed-to or referred-to object will result in *fundamentally incorrect* code
- if function will never modify pointee/referee associated with function parameter, parameter type should be made pointer/reference to const object

# String Length Example: Not Const Correct

```
1  // ERROR: parameter type should be const char*
2  int string_length(char* s) {
3      int n = 0;
4      while (*s++ != '\0') {++n;}
5      return n;
6  }
7
8  int main() {
9      char buf[] = "Goodbye";
10     const char* const m1 = "Hello";
11     char* const m2 = &buf[0];
12     int n1 = string_length(m1);
13     // must copy argument m1 to parameter s:
14     //   char* s = m1;
15     // convert from const char* const to char*
16     // ERROR: must discard const from pointee
17     int n2 = string_length(m2);
18     // must copy argument m2 to parameter s:
19     //   char* s = m2;
20     // convert from char* const to char*
21     // OK: constness of pointee unchanged
22 }
```

# String Length Example: Const Correct

```
1  // OK: pointee is const
2  int string_length(const char* s) {
3      int n = 0;
4      while (*s++ != '\0') {++n;}
5      return n;
6  }
7
8  int main() {
9      char buf[] = "Goodbye";
10     const char* const m1 = "Hello";
11     char* const m2 = &buf[0];
12     int n1 = string_length(m1);
13     // must copy argument m1 to parameter s:
14     //   const char* s = m1;
15     // convert from const char* const to const char*
16     // OK: constness of pointee unchanged
17     int n2 = string_length(m2);
18     // must copy argument m2 to parameter s:
19     //   const char* s = m2;
20     // convert from char* const to const char*
21     // OK: can add const to pointee
22 }
```

# Square Example: Not Const Correct

```
1  #include <complex>
2
3  using Complex = std::complex<long double>;
4
5  // ERROR: parameter type should be reference to const
6  Complex square(Complex& z) {
7      return z * z;
8  }
9
10 int main() {
11     const Complex c1(1.0, 2.0);
12     Complex c2(1.0, 2.0);
13     Complex r1 = square(c1);
14     // must bind parameter z to argument c1
15     //     Complex& z = c1;
16     // convert from const Complex to Complex&
17     // ERROR: must discard const from referee
18     Complex r2 = square(c2);
19     // must bind parameter z to argument c2
20     //     Complex& z = c2;
21     // convert from Complex to Complex&
22     // OK: constness of referee unchanged
23 }
```

# Square Example: Const Correct

```
1  #include <complex>
2
3  using Complex = std::complex<long double>;
4
5  // OK: parameter type is reference to const
6  Complex square(const Complex& z) {
7      return z * z;
8  }
9
10 int main() {
11     const Complex c1(1.0, 2.0);
12     Complex c2(1.0, 2.0);
13     Complex r1 = square(c1);
14     // must bind parameter z to argument c1
15     //     const Complex& z = c1;
16     // convert from const Complex to const Complex&
17     // OK: constness of referee not discarded
18     Complex r2 = square(c2);
19     // must bind parameter z to argument c2
20     //     const Complex& z = c2;
21     // convert from Complex to const Complex&
22     // OK: can add const to referee
23 }
```



# Function Types and the **const** Qualifier

```
1  // top-level qualifiers of parameter types are
2  // not part of function type and should be omitted
3  // from function declaration
4
5  // BAD: const not part of function type
6  // (nothing here to which const can refer)
7  bool is_even(const unsigned int);
8
9  // OK
10 bool is_odd(unsigned int);
11
12 // OK: parameter with top-level const qualifier
13 // is ok in function definition
14 bool is_even(const unsigned int x) {
15     // cannot change x in function
16     return x % 2 == 0;
17 }
18
19 // OK
20 bool is_odd(unsigned int x) {
21     // x can be changed if desired
22     return x % 2 != 0;
23 }
```

# Inline Functions

- in general programming sense, **inline function** is function for which compiler copies code from function definition directly into code of calling function rather than creating separate set of instructions in memory
- since code copied directly into calling function, no need to transfer control to separate piece of code and back again to caller, *eliminating performance overhead* of function call
- inline typically used for *very short functions* (where overhead of calling function is large relative to cost of executing code within function itself)
- can request function be made inline by including **inline** qualifier along with function return type (but *compiler may ignore request*)
- inline function must be defined in each translation unit in which function is used and all definitions must be identical; this is exception to one-definition rule [C++17 10.1.6/6]
- example:

```
inline bool isEven(int x) {  
    return x % 2 == 0;  
}
```

# Inlining of a Function

- inlining of `isEven` function transforms code fragment 1 into code fragment 2

- Code fragment 1:

```
inline bool isEven(int x) {  
    return x % 2 == 0;  
}
```

```
void myFunction() {  
    int i = 3;  
    bool result = isEven(i);  
}
```

- Code fragment 2:

```
void myFunction() {  
    int i = 3;  
    bool result = (i % 2 == 0);  
}
```

# The `constexpr` Qualifier for Functions

- `constexpr` qualifier indicates return value of function is constant expression (i.e., can be evaluated at compile time) *provided that all arguments to function are constant expressions*
- `constexpr` function required to be evaluated at compile time if all arguments are constant expressions and return value *used in constant expression*
- `constexpr` functions are implicitly inline [C++17 10.1.5/1]
- `constexpr` function very restricted in what it can do (e.g., no external state, can only call `constexpr` functions, variables must be initialized)
- example:

```
constexpr int factorial(int n)
{ return n >= 2 ? (n * factorial(n - 1)) : 1; }

int u[factorial(5)];
// OK: factorial(5) is constant expression

int x = 5;
int v[factorial(x)];
// ERROR: factorial(x) is not constant
// expression
```

# Constexpr Example

```
1  constexpr int square(int i) {
2      return i * i;
3  }
4
5  constexpr int func(int n) {
6      int sum = 0;
7      for (int i = 1; i <= n; ++i) {
8          sum += square(i);
9      }
10     return sum;
11 }
12
13 int main() {
14     // at compile time, compute sum of the squares of
15     // 1, 2, 3 (i.e., 14)
16     constexpr int result = func(3);
17     static_assert(result == 14);
18 }
```

# Constexpr Function Example: square

SKIP SLIDE

```
1  #include <iostream>
2
3  constexpr double square(double x) {
4      return x * x;
5  }
6
7  int main() {
8      constexpr double a = square(2.0);
9      // must be computed at compile time
10
11     double b = square(0.5);
12     // might be computed at compile time
13
14     double t;
15     if (!(std::cin >> t)) {
16         return 1;
17     }
18     const double c = square(t);
19     // must be computed at run time
20
21     std::cout << a << ' ' << b << ' ' << c << '\n';
22 }
```

# Constexpr Function Example: power\_int (Recursive)

SKIP SLIDE

```
1  #include <iostream>
2
3  constexpr double power_int_helper(double x, int n) {
4      return (n > 0) ? x * power_int_helper(x, n - 1) : 1;
5  }
6
7  constexpr double power_int(double x, int n) {
8      return (n < 0) ? power_int_helper(1.0 / x, -n) :
9          power_int_helper(x, n);
10 }
11
12 int main() {
13     constexpr double a = power_int(0.5, 8);
14     // must be computed at compile time
15
16     double b = power_int(0.5, 8);
17     // might be computed at compile time
18
19     double x;
20     if (!(std::cin >> x)) {return 1;}
21     const double c = power_int(x, 2);
22     // must be computed at run time
23
24     std::cout << a << ' ' << b << ' ' << c << '\n';
25 }
```

# Constexpr Function Example: power\_int (Iterative)

```
1  #include <iostream>
2
3  constexpr double power_int(double x, int n) {
4      double result = 1.0;
5      if (n < 0) {
6          x = 1.0 / x;
7          n = -n;
8      }
9      while (--n >= 0) {
10         result *= x;
11     }
12     return result;
13 }
14
15 int main() {
16     constexpr double a = power_int(0.5, 8);
17     // must be computed at compile time
18
19     double b = power_int(0.5, 8);
20     // might be computed at compile time
21
22     double x;
23     if (!(std::cin >> x)) {return 1;}
24     const double c = power_int(x, 2);
25     // must be computed at run time
26
27     std::cout << a << ' ' << b << ' ' << c << '\n';
28 }
```



# Compile-Time Versus Run-Time Computation

- constexpr variables and constexpr functions provide mechanism for moving computation from run time to compile time
- benefits of compile-time computation include:
  - 1 no execution-time cost at run-time
  - 2 can facilitate compiler optimization (e.g., eliminate conditional branch if condition always true/false)
  - 3 can reduce code size since code used only for compile-time computation does not need to be included in executable
  - 4 can find errors at compile-time and link-time instead of at run time
  - 5 no concerns about order of initialization (which is not necessarily true for const objects)
  - 6 no synchronization concerns (e.g., multiple threads trying to initialize object)
- when floating point is involved, compile-time and run-time computations can yield different results, due to differences in such things as
  - rounding mode in effect
  - processor architecture used for computation (when cross compiling)

# Function Overloading

- **function overloading**: multiple functions can have same name as long as they differ in number/type of their arguments
- example:

```
void print(int x) {  
    std::cout << "int has value " << x << '\n';  
}  
  
void print(double x) {  
    std::cout << "double has value " << x << '\n';  
}  
  
void demo() {  
    int i = 5;  
    double d = 1.414;  
    print(i); // calls print(int)  
    print(d); // calls print(double)  
    print(42); // calls print(int)  
    print(3.14); // calls print(double)  
}
```

# Default Arguments

- can specify default values for arguments to functions
- example:

```
// Compute log base b of x.
double logarithm(double x, double b) {
    return std::log(x) / std::log(b);
}

// Declaration of logarithm with a default argument.
double logarithm(double, double = 10.0);

void demo() {
    double x =
        logarithm(100.0); // calls logarithm(100.0, 10.0)
    double y =
        logarithm(4.0, 2.0); // calls logarithm(4.0, 2.0)
}
```

# Argument Matching

- call of given function name chooses function that best matches actual arguments
- consider all functions in scope for which set of conversions exists so function could possibly be called
- best match is intersection of sets of functions that best match on each argument
- matches attempted in following order: [C++17 16.3.3.1]
  - 1 exact match with zero or more trivial conversions (e.g., `T` to `T&`, `T&` to `T`, adding **const** and/or **volatile**); of these, those that do not add **const** and/or **volatile** to pointer/reference better than those that do
  - 2 match with promotions (e.g., **int** to **long**, **float** to **double**)
  - 3 match with standard conversions (e.g., **float** to **int**, **double** to **int**)
  - 4 match with user-defined conversions
  - 5 match with ellipsis
- if set of best matches contains exactly one element, this element chosen as function to call
- if set of best matches is either empty or contains more than one element, function call is invalid (since either no matches found or multiple equally-good matches found)

# Argument Matching: Example

```
1  int max(int, int);
2  double max(double, double);
3
4  int main() {
5      int i, j, k;
6      double a, b, c;
7      // ...
8      k = max(i, j);
9          // viable functions: max(int, int), max(double, double)
10         // best match on first argument: max(int, int)
11         // best match on second argument: max(int, int)
12         // best viable function: max(int, int)
13         // OK: calls max(int, int)
14     c = max(a, b);
15         // viable functions: max(int, int), max(double, double)
16         // best match on first argument: max(double, double)
17         // best match on second argument: max(double, double)
18         // best viable function: max(double, double)
19         // OK: calls max(double, double)
20     c = max(i, b);
21         // viable functions: max(int, int), max(double, double)
22         // best match on first argument: max(int, int)
23         // best match on second argument: max(double, double)
24         // no best viable function
25         // ERROR: ambiguous function call
26 }
```

# The assert Macro

- `assert` macro allows testing of boolean condition at run time
- typically used to test sanity of code (e.g., test preconditions, postconditions, or other invariants) or test validity of assumptions made by code
- defined in header file `cassert`
- macro takes single argument: boolean expression
- if assertion fails, program is terminated by calling `std::abort`
- if `NDEBUG` preprocessor symbol is defined at time `cassert` header file included, all assertions are disabled (i.e., not checked)
- `assert(expr)` is constant expression if `expr` is constant expression that evaluates to **true** or `NDEBUG` is defined [\[C++17 22.3.2/1\]](#)
- example:

```
#include <cassert>

double sqrt(double x) {
    assert(x >= 0);
    // ...
}
```

## Section 2.3.6

### **Input/Output (I/O)**

- relevant declarations and such in header file `iostream`
- `std::istream`: stream from which characters/data can be read (i.e., input stream)
- `std::ostream`: stream to which characters/data can be written (i.e., output stream)
- `std::istream` `std::cin` standard input stream
- `std::ostream` `std::cout` standard output stream
- `std::ostream` `std::cerr` standard error stream
- in most environments, above three streams refer to user's terminal by default
- output operator (inserter) `<<`
- input operator (extractor) `>>`
- stream can be used as **bool** expression; converts to **true** if stream has not encountered any errors and **false** otherwise (e.g., if invalid data read or I/O error occurred)



# Basic I/O Example

```
1  #include <iostream>
2
3  int main() {
4      std::cout << "Enter an integer: ";
5      int x;
6      std::cin >> x;
7      if (std::cin) {
8          std::cout << "The integer entered was "
9              << x << ".\n";
10     } else {
11         std::cerr <<
12             "End-of-file reached or I/O error.\n";
13     }
14 }
```

- manipulators provide way to control formatting of data values written to streams as well as parsing of data values read from streams
- declarations related information for manipulators can be found in header files: `ios`, `iomanip`, `istream`, and `ostream`
- most manipulators used to control output formatting
- focus here on manipulators as they pertain to output
- manipulator may have *immediate* effect (e.g., `endl`), only affect *next* data value output (e.g., `setw`), or affect *all* subsequent data values output (e.g., `setprecision`)

# I/O Manipulators (Continued)

Name	Description
<code>setw</code>	set field width
<code>setfill</code>	set fill character
<code>endl</code>	insert newline and flush
<code>flush</code>	flush stream
<code>dec</code>	use decimal
<code>hex</code>	use hexadecimal
<code>oct</code>	use octal
<code>showpos</code>	show positive sign
<code>noshowpos</code>	do not show positive sign
<code>left</code>	left align
<code>right</code>	right align
<code>fixed</code>	write floating-point values in fixed-point notation
<code>scientific</code>	write floating-point values in scientific notation
<code>setprecision</code>	for default notation, specify maximum number of meaningful digits to display before and after decimal point; for fixed and scientific notations, specify exactly how many digits to display after decimal point (padding with trailing zeros if necessary)

# I/O Manipulators Example

```
1  #include <iostream>
2  #include <ios>
3  #include <iomanip>
4
5  int main() {
6      constexpr double pi = 3.1415926535;
7      constexpr double big = 123456789.0;
8      // default notation
9      std::cout << pi << ' ' << big << '\n';
10     // fixed-point notation
11     std::cout << std::fixed << pi << ' ' << big << '\n';
12     // scientific notation
13     std::cout << std::scientific << pi << ' ' << big << '\n';
14     // fixed-point notation with 7 digits after decimal point
15     std::cout << std::fixed << std::setprecision(7) << pi << ' '
16         << big << '\n';
17     // fixed-point notation with precision and width specified
18     std::cout << std::setw(8) << std::fixed << std::setprecision(2)
19         << pi << ' ' << std::setw(20) << big << '\n';
20     // fixed-point notation with precision, width, and fill specified
21     std::cout << std::setw(8) << std::setfill('x') << std::fixed
22         << std::setprecision(2) << pi << ' ' << std::setw(20) << big << '\n';
23 }
24
25 /* This program produces the following output:
26 3.14159 1.23457e+08
27 3.141593 123456789.000000
28 3.141593e+00 1.234568e+08
29 3.1415927 123456789.0000000
30 3.14      123456789.00
31 xxxx3.14 xxxxxxxx123456789.00
32 */
```

# Use of `std::istream::eof`

- do not use `std::istream::eof` to determine if earlier input operation has failed, as this will not always work
- `eof` simply returns end-of-file (EOF) flag for stream
- EOF flag for stream can be set during *successful* input operation (when input operation takes places just before end of file)
- when stream extractors (i.e., **operator>>**) used, fields normally delimited by whitespace
- to read all data in whitespace-delimited field, must read *one character beyond* field in order to know that end of field has been reached
- if field followed immediately by EOF without any intervening whitespace characters, reading one character beyond field will cause EOF to be encountered and EOF bit for stream to be set
- in preceding case, however, EOF being set does not mean that input operation failed, only that stream data ended immediately after field that was read

# Example: Incorrect Use of eof

- example of *incorrect* use of eof:

```
1  #include <iostream>
2
3  int main() {
4      while (true) {
5          int x;
6          std::cin >> x;
7          // std::cin may not be in a failed state.
8          if (std::cin.eof()) {
9              // Above input operation may have succeeded.
10             std::cout << "EOF encountered\n";
11             break;
12         }
13         std::cout << x << '\n';
14     }
15 }
```

- code incorrectly assumes that eof will only return true if preceding input operation has failed
- last field in stream will be incorrectly ignored if not followed by at least one whitespace character; for example, if input stream consists of three character sequence '1', space, '2', program will output:

```
1
EOF encountered
```

## Example: Correct Use of eof

- to determine if input operation failed, simply check if stream in failed state
- if stream *already known to be in failed state* and need to determine specifically if failure due to EOF being encountered, then use eof
- example of correct use of eof:

```
1  #include <iostream>
2
3  int main() {
4      int x;
5      // Loop while std::cin not in a failed state.
6      while (std::cin >> x) {
7          std::cout << x << '\n';
8      }
9      // Now std::cin must be in a failed state.
10     // Use eof to determine the specific reason
11     // for failure.
12     if (std::cin.eof()) {
13         std::cout << "EOF encountered\n";
14     } else {
15         std::cout << "input error (excluding EOF)\n";
16     }
17 }
```

- `std::endl` is not some kind of string constant
- `std::endl` is stream manipulator and declared as  
`std::ostream& std::endl(std::ostream&)`
- inserting `endl` to stream always (regardless of operating system) equivalent to outputting single newline character '`\n`' followed by flushing stream
- flushing of stream can incur very substantial overhead; so only flush when strictly necessary



## Use of `std::endl` (Continued)

- some operating systems terminate lines with single linefeed character (i.e., `'\n'`), while other operating systems use carriage-return and linefeed pair (i.e., `'\r'` plus `'\n'`)
- existence of `endl` has nothing to do with dealing with handling new lines in operating-system independent manner
- when stream opened in text mode, translation between newline characters and whatever character(s) operating system uses to terminate lines is performed automatically (both for input and output)
- above translation done for all characters input and output and has nothing to do with `endl`

# Stream Extraction Failure

- for built-in types, if stream extraction fails, value of target for stream extraction depends on reason for failure [\[C++17 25.4.2.1.2/3 \(Stage 3\)\]](#)
- in following example, what is value of `x` if stream extraction fails:

```
int x;  
std::cin >> x;  
if (!std::cin) {  
    // what is value of x?  
}
```

- in above example, `x` may be *uninitialized* upon stream extraction failure
- if failure due to I/O error or EOF, target of extraction is *not modified*
- if failure due to badly formatted data, target of extraction is zero
- if failure due to overflow, target of extraction is closest machine-representable value
- **common error:** incorrectly assume that target of extraction will always be initialized if extraction fails
- for class types, also dangerous to assume target of extraction always written upon failure

## Stream Extraction Failure (Continued)

[illegible]

# Testing Failure State of Streams

- consider `istream` or `ostream` object `s`
- `!s` is equivalent to `s.fail()`
- `bool(s)` is not equivalent to `s.good()` [\[C++17 30.5.5.4/1\]](#) [\[C++17 30.5.5.4/7\]](#)
- `s.good()` is not the same as `!s.fail()`
- do not use `good` as opposite of `fail` since this is wrong

## Section 2.3.7

### **Miscellany**

# Namespaces

- **namespace** is region that provides scope for identifiers declared inside
- namespace provides mechanism for reducing likelihood of naming conflicts
- syntax for namespace has general form:

```
namespace name {  
    body  
}
```

- *name*: identifier that names namespace
- *body*: body of namespace (i.e., code)
- all identifiers (e.g., names of variables, functions, and types) declared in *body* made to belong to scope associated with namespace *name*
- same identifier can be re-used in different namespaces, since each namespace is separate scope
- scope-resolution operator (i.e., `::`) can be used to explicitly specify namespace to which particular identifier belongs
- **using** statement can be used to bring identifiers from other namespaces into current scope

# Namespaces: Example

```
1  #include <iostream>
2
3  using std::cout; // bring std::cout into current scope
4
5  namespace mike {
6      int someValue;
7      void initialize() {
8          cout << "mike::initialize called\n";
9          someValue = 0;
10     }
11 }
12
13 namespace fred {
14     double someValue;
15     void initialize() {
16         cout << "fred::initialize called\n";
17         someValue = 1.0;
18     }
19 }
20
21 void func() {
22     mike::initialize(); // call initialize in namespace mike
23     fred::initialize(); // call initialize in namespace fred
24     using mike::initialize;
25     // bring mike::initialize into current scope
26     initialize(); // call mike::initialize
27 }
```

# Nested Namespace Definitions

- name given in namespace declaration can be qualified name in order to succinctly specify nested namespace
- consider following namespace declaration:

```
namespace foo {  
    namespace bar {  
        namespace impl {  
            // ...  
        }  
    }  
}
```

- preceding declaration can be written more succinctly as:

```
namespace foo::bar::impl {  
    // ...  
}
```



# Namespace Aliases

- identifier can be introduced as alias for namespace
- syntax has following form:

**namespace** *alias\_name* = *ns\_name*;

- identifier *alias\_name* is alias for namespace *ns\_name*
- namespace aliases particularly useful for creating short names for deeply-nested namespaces or namespaces with long names
- example:

```
1  #include <iostream>
2
3  namespace foobar {
4      namespace miscellany {
5          namespace experimental {
6              int get_meaning_of_life() {return 42;}
7              void greet() {std::cout << "hello\n";}
8          }
9      }
10 }
11
12 int main() {
13     namespace n = foobar::miscellany::experimental;
14     n::greet();
15     std::cout << n::get_meaning_of_life() << '\n';
16 }
```

# Inline Namespaces

- namespace can be made inline, in which case all identifiers in namespace also visible in enclosing namespace
- inline namespaces useful, for example, for library versioning
- example:

```
1  #include <cassert>
2
3  // some awesome library
4  namespace awesome {
5      // version 1
6      namespace v1 {
7          int meaning_of_life() {return 41;}
8      }
9      // new and improved version 2
10     // which should be default for library users
11     inline namespace v2 {
12         int meaning_of_life() {return 42;}
13     }
14 }
15
16 int main() {
17     assert(awesome::v1::meaning_of_life() == 41);
18     assert(awesome::v2::meaning_of_life() == 42);
19     assert(awesome::meaning_of_life() == 42);
20 }
```

# Unnamed Namespaces

- can create unnamed namespace (i.e., namespace without name)
- unnamed namespace often referred to as anonymous namespace
- each translation unit may contain its own unique unnamed namespace
- entities defined in unnamed namespace only visible in its associated translation unit (i.e., has internal linkage)
- example:

```
1  #include <iostream>
2
3  namespace {
4  const int forty_two = 42;
5  int x;
6  }
7
8  int main() {
9      x = forty_two;
10     std::cout << x << '\n';
11 }
```

# Memory Allocation: **new** and **delete**

- to allocate memory, use **new** statement
- to deallocate memory allocated with **new** statement, use **delete** statement
- similar to `malloc` and `free` in C
- two forms of allocation: 1) single object (i.e., nonarray case) and 2) array of objects
- array version of `new/delete` distinguished by `[]`
- example:

```
char* buffer = new char[64]; // allocate  
                        // array of 64 chars  
delete [] buffer; // deallocate array  
double* x = new double; // allocate single double  
delete x; // deallocate single object
```

- important to match nonarray and array versions of **new** and **delete**:

```
char* buffer = new char[64]; // allocate  
delete buffer; // ERROR: nonarray delete to  
                // delete array  
                // may compile fine, but crash
```

# User-Defined Literals

- C++ has several categories of literals (e.g., character, integer, floating-point, string, boolean, and pointer)
- can define additional literals based on these categories
- identifier used as suffix for user-defined literal must begin with underscore
- suffixes that do not begin with underscore are reserved for use by standard library
- example:

```
1  #include <iostream>
2  #include <complex>
3
4  std::complex<long double> operator "" _i(long double d) {
5      return std::complex<long double>(0.0, d);
6  }
7
8  int main() {
9      auto z = 3.14_i;
10     std::cout << z << '\n';
11 }
12
13 // Program output:
14 // (0,3.14)
```

- attributes provide unified syntax for implementation-defined language extensions
- attribute can be used almost anywhere in source code and can be applied to almost anything (e.g., types, variables, functions, names, code blocks, and translation units)
- specific types of entities to which attribute can be applied depends on particular attribute in question
- attribute specifiers start with two consecutive left brackets and continue to two consecutive right brackets
- example:

```
[[deprecated]]  
void some_very_old_function() { /* ... */};
```

# Some Standard Attributes

Name	Description
<code>noreturn</code>	function does not return
<code>deprecated</code>	use of entity is deprecated (i.e., allowed but discouraged)
<code>fallthrough</code>	fall through in switch statement is deliberate
<code>maybe_unused</code>	entity (e.g., variable) may be unused
<code>nodiscard</code>	used to indicate that return value of function should not be ignored

# Some GCC and Clang Attributes

## GCC C++ Compiler

Name	Description
<code>gnu::noinline</code>	do not inline function
<code>gnu::no_sanitize_address</code>	do not instrument function for address sanitizer
<code>gnu::no_sanitize_undefined</code>	do not instrument function for undefined-behavior sanitizer

## Clang C++ Compiler

Name	Description
<code>gnu::noinline</code>	do not inline function
<code>clang::no_sanitize</code>	do not instrument function for sanitizer



## Section 2.3.8

### **References**

- 1 D. Saks. [Placing const in declarations.](#)  
*Embedded Systems Programming*, pages 19–20, June 1998.
- 2 D. Saks. [What const really means.](#)  
*Embedded Systems Programming*, pages 11–14, Aug. 1998.
- 3 D. Saks. [const T vs. T const.](#)  
*Embedded Systems Programming*, pages 13–16, Feb. 1999.
- 4 D. Saks. [Top-level cv-qualifiers in function parameters.](#)  
*Embedded Systems Programming*, pages 63–65, Feb. 2000.

## Section 2.4

# Classes

- since fundamental types provided by language are quite limiting, language provides mechanism for defining new (i.e., user-defined) types
- **class** is user-defined type
- class specifies:
  - 1 how objects of class are *represented*
  - 2 *operations* that can be performed on objects of class
- not all parts of class are directly accessible to all code
- **interface** is part of class that is directly accessible to its users
- **implementation** is part of class that its users access only indirectly through interface

## Section 2.4.1

### **Members and Access Specifiers**

- class consists of *zero or more members*
- three basic kinds of members (excluding enumerators):
  - 1 data member
  - 2 function member
  - 3 type member
- **data members** define representation of class object
- **function members** (also called member functions) provide operations on such objects
- **type members** specify any types associated with class

- can control *level of access* that users of class have to its members
- three levels of access:
  - 1 public
  - 2 protected
  - 3 private
- **public**: member can be accessed by any code
- **private**: member can only be accessed by other members of class and friends of class (to be discussed shortly)
- **protected**: relates to inheritance (discussion deferred until later)
- public members constitute class interface
- private members constitute class implementation

- class typically has form:

```
class Widget // The class is named Widget.
{
public:
    // public members
    // (i.e., the interface to users)
    // usually functions and types (but not data)
private:
    // private members
    // (i.e., the implementation details only
    // accessible by members of class)
    // usually functions, types, and data
};
```



# Default Member Access

- class members are private by default
- two code examples below are exactly equivalent:

```
class Widget {  
    // ...  
};
```

```
class Widget {  
private:  
    // ...  
};
```

# The **struct** Keyword

- **struct** is class where members public by default
- two code examples below are exactly equivalent:

```
struct Widget {  
    // ...  
};
```

```
class Widget {  
public:  
    // ...  
};
```

- class example:

```
class Vector_2 { // Two-dimensional vector class.  
public:  
    double x; // The x component of the vector.  
    double y; // The y component of the vector.  
};  
  
void func() {  
    Vector_2 v;  
    v.x = 1.0; // Set data member x to 1.0  
    v.y = 2.0; // Set data member y to 2.0  
}
```

- above class has data members `x` and `y`
- members accessed by *member-selection operator* (i.e., “.”)

# Function Members

- class example:

```
class Vector_2 { // Two-dimensional vector class.
public:
    void initialize(double newX, double newY);
    double x; // The x component of the vector.
    double y; // The y component of the vector.
};

void Vector_2::initialize(double newX, double newY) {
    x = newX; // "x" means "this->x"
    y = newY; // "y" means "this->y"
}

void func() {
    Vector_2 v; // Create Vector_2 called v.
    v.initialize(1.0, 2.0); // Initialize v to (1.0, 2.0).
}
```

- above class has member function initialize
- to refer to member of class outside of class body must use *scope-resolution operator* (i.e., ::)
- for example, in case of initialize function, we use `Vector_2::initialize`
- member function always has *implicit parameter* referring to class object

# The **this** Keyword

- member function always has *implicit parameter* referring to class object
- implicit parameter accessible inside member function via **this** keyword
- **this** is pointer to object for which member function is being invoked
- data members can be accessed through **this** pointer
- since data members can also be referred to directly by their names, explicit use of **this** often not needed and normally avoided
- example:

```
class Widget {  
public:  
    int updateValue(int newValue) {  
        int oldValue = value; // "value" means "this->value"  
        value = newValue; // "value" means "this->value"  
        return oldValue;  
    }  
private:  
    int value;  
};  
  
void func() {  
    Widget x;  
    x.updateValue(5);  
    // in Widget::updateValue, variable this equals &x  
}
```

# const Member Functions

- member function has reference to object of class as implicit parameter (i.e., object pointed to by **this**)
- need way to indicate if member function can change value of object
- **const** member function cannot change value of object

```
1  class Counter {
2  public:
3      int getCount() const
4          {return count;} // count means this->count
5      void setCount(int newCount)
6          {count = newCount;} // count means this->count
7      void incrementCount()
8          {++count;} // count means this->count
9  private:
10     int count; // counter value
11 };
12
13 void func() {
14     Counter ctr;
15     ctr.setCount(0);
16     int count = ctr.getCount();
17     const Counter& ctr2 = ctr;
18     count = ctr2.getCount(); // getCount better be const!
19 }
```

# Definition of Function Members in Class Body

- member function whose definition is provided in body of class is automatically **inline**
- two code examples below are exactly equivalent:

```
class MyInteger {  
public:  
    // Set the value of the integer and return the old value.  
    int setValue(int newValue) {  
        int oldValue = value;  
        value = newValue;  
        return oldValue;  
    }  
private:  
    int value;  
};
```

```
class MyInteger {  
public:  
    // Set the value of the integer and return the old value.  
    int setValue(int newValue);  
private:  
    int value;  
};  
  
inline int MyInteger::setValue(int newValue) {  
    int oldValue = value;  
    value = newValue;  
    return oldValue;  
}
```

- example:

```
class Point_2 { // Two-dimensional point class.
public:
    using Coordinate = double; // Coordinate type.
    Coordinate x; // The x coordinate of the point.
    Coordinate y; // The y coordinate of the point.
};

void func() {
    Point_2 p;
    // ...
    Point_2::Coordinate x = p.x;
    // Point_2::Coordinate same as double
}
```

- above class has type member `Coordinate`
- to refer to type member outside of class body, we must use *scope-resolution operator* (i.e., `::`)



- normally, only class has access to its private members
- sometimes, necessary to allow another class or function to have access to private members of class
- friend of class is function/class that is allowed to access private members of class
- to make function or class friend of another class, use **friend** statement

- example:

```
class Gadget; // forward declaration of Gadget
```

```
class Widget {  
    // ...  
    friend void myFunc();  
    // function myFunc is friend of Widget  
    friend class Gadget;  
    // class Gadget is friend of Widget  
    // ...  
};
```

- generally, use of friends should be avoided except when absolutely necessary

# Class Example

```
1  class Widget {
2  public:
3      int setValue(int newValue) { // member function
4          int oldValue = value; // save old value
5          value = newValue; // change value to new value
6          return oldValue; // return old value
7      }
8  private:
9      friend void wasteTime();
10     void doNothing() {}
11     int value; // data member
12 };
13
14 void wasteTime() {
15     Widget x;
16     x.doNothing(); // OK: friend
17     x.value = 5; // OK: friend
18 }
19
20 void func() {
21     Widget x; // x is object of type Widget
22     x.setValue(5); // call Widget's setValue member
23     // sets x.value to 5
24     x.value = 5; // ERROR: value is private
25     x.doNothing(); // ERROR: doNothing is private
26 }
```

## Section 2.4.2

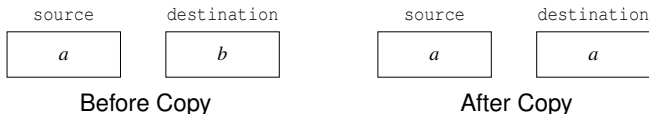
# **Constructors and Destructors**

# Propagating Values: Copying and Moving

- Suppose that we have two objects of the same type and we want to propagate the value of one object (i.e., the source) to the other object (i.e., the destination).
- This can be accomplished in one of two ways: 1) copying or 2) moving.
- **Copying** propagates the value of the source object to the destination object *without modifying the source object*.
- **Moving** propagates the value of the source object to the destination object and is *permitted to modify the source object*.
- Moving is always at least as efficient as copying, and for many types, moving is *more efficient* than copying.
- For some types, *copying does not make sense*, while moving does (e.g., `std::ostream`, `std::istream`).

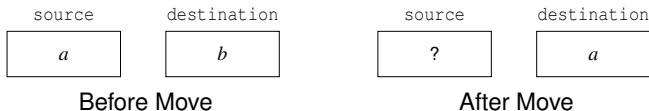
# Copying and Moving

- **Copy operation.** Propagating the value of the source object `source` to the destination object `destination` by copying.



- A copy operation *does not modify* the value of the source object.

- 
- **Move operation.** Propagating the value of the source object `source` to the destination object `destination` by moving.



- A move operation is *not guaranteed to preserve* the value of the source object. After the move operation, the source object has a value that is *valid* but *typically unspecified*.

# Constructors

- when new object created usually desirable to immediately initialize it to some known state
- prevents object from accidentally being used before it is initialized
- **constructor** is member function that is *called automatically* when object created in order to *initialize* its value
- constructor has *same name as class* (i.e., constructor for class `T` is function `T::T`)
- constructor has *no return type* (not even `void`)
- constructor *cannot be called directly* (although placement new provides mechanism for achieving similar effect, in rare cases when needed)
- constructor *can be overloaded*
- before constructor body is entered, all data members of class type are first constructed in order of declaration in class definition
- in certain circumstances, constructors may be automatically provided
- sometimes, automatically provided constructors *will not* have correct behavior

- constructor that can be called with no arguments known as **default constructor** [C++17 15.1/4]
- example:

```
class Vector { // Two-dimensional vector class.  
public:  
    Vector() // Default constructor.  
        {x_ = 0.0; y_ = 0.0;}  
        // ...  
private:  
    double x_; // The x component of the vector.  
    double y_; // The y component of the vector.  
};  
  
Vector v; // calls Vector::Vector(); v set to (0,0)  
Vector x(); // declares function x that returns Vector
```

# Defaulted Default Constructor

- **defaulted default constructor** for class `T` performs similar initialization as constructor defined as [\[C++17 15.1/7\]](#)

```
T::T() {}
```

- if class has no default member initializers, this corresponds to default constructing each data member of class type and leaving data members of built-in type uninitialized
- defaulted default constructor automatically provided (i.e., implicitly declared) as public member if no user-declared constructors [\[C++17 15.1/4\]](#)
- example:

```
#include <string>

// class has implicitly-defined defaulted
// default constructor
struct Widget {
    void foo() {}
    std::string s;
};
```



- for class `T`, constructor taking lvalue reference to `T` as first parameter that can be called with one argument known as **copy constructor** [C++17 15.8.1/1]
- used to *create* object by *copying* from already-existing object
- copy constructor for class `T` typically is of form `T(const T&)`
- example:

```
class Vector { // Two-dimensional vector class.
public:
    Vector() {x_ = 0.0; y_ = 0.0;} // Default constructor
    Vector(const Vector& v) // Copy constructor.
        {x_ = v.x_; y_ = v.y_;}
    // ...
private:
    double x_; // The x component of the vector.
    double y_; // The y component of the vector.
};

Vector v;
Vector w(v); // calls Vector::Vector(const Vector&)
Vector u = v; // calls Vector::Vector(const Vector&)
```

# Defaulted Copy Constructor

- **defaulted copy constructor** performs memberwise copy of its data members (and bases), where copy performed using: [\[C++17 15.8.1/14\]](#)
  - copy constructor for class types
  - bitwise copy for built-in types
- defaulted copy constructor automatically provided (i.e., implicitly defined) as public member if none of following user declared: [\[C++17 15.8.1/6\]](#)
  - move constructor
  - move assignment operator
  - copy assignment operator (if not relying on deprecated behavior)
  - destructor (if not relying on deprecated behavior)
- example:

```
// class has defaulted copy constructor
class Widget {
public:
    Widget(int i) {i_ = i;}
    int get() const {return i_;}
private:
    int i_;
};
```

# Move Constructor

- for class `T`, constructor taking rvalue reference to `T` as first parameter that can be called with one argument known as **move constructor** [C++17 15.8.1/2]
- used to *create* object by *moving* from already-existing object
- move constructor for class `T` typically is of form `T(T&&)`
- example:

```
class Vector { // Two-dimensional vector class.
public:
    Vector() {x_ = 0.0; y_ = 0.0;} // Default constructor
    Vector(Vector&& v) {x_ = v.x_; y_ = v.y_;} // Move constructor.
    // ...
private:
    double x_; // The x component of the vector.
    double y_; // The y component of the vector.
};

#include <utility>
Vector v;
Vector w(std::move(v)); // calls Vector::Vector(Vector&&)
Vector x = std::move(w); // calls Vector::Vector(Vector&&)
```

# Defaulted Move Constructor

- **defaulted move constructor** performs memberwise move of its data members (and bases) using: [\[C++17 15.8.1/14\]](#)
  - move constructor if available and copy constructor otherwise in case of class type
  - bitwise copy in case of built-in type
- defaulted move constructor automatically provided (i.e., implicitly defined) as public member if none of following user declared: [\[C++17 15.8.1/8\]](#)
  - copy constructor
  - copy assignment operator
  - move assignment operator
  - destructor

- example:

```
// class has defaulted move constructor
struct Widget {
    Widget();
    void foo();
};
```

# Constructor Example

```
1  class Vector { // Two-dimensional vector class.
2  public:
3      // Default constructor.
4      Vector() {x_ = 0.0; y_ = 0.0;}
5      // Copy constructor.
6      Vector(const Vector& v) {x_ = v.x_; y_ = v.y_;}
7      // Move constructor.
8      Vector(Vector&& v) {x_ = v.x_; y_ = v.y_;}
9      // Another constructor.
10     Vector(double x, double y) {x_ = x; y_ = y;}
11     // ...
12 private:
13     double x_; // The x component of the vector.
14     double y_; // The y component of the vector.
15 };
```

- four constructors provided

# Constructor Example (Continued 1)

```
1  // include definition of Vector class here
2
3  int main() {
4      Vector u;
5          // calls default constructor
6      Vector v(1.0, 2.0);
7          // calls Vector::Vector(double, double)
8      Vector w(v);
9          // calls copy constructor
10     Vector x = u;
11         // calls copy constructor
12     Vector y = Vector(1.0, 0.0);
13         // guaranteed copy/move elision
14         // calls Vector::Vector(double, double), directly
15         // constructing new object in y
16         // does not call move constructor
17     Vector z{Vector()};
18         // guaranteed copy/move elision
19         // calls default constructor, directly constructing
20         // new object in z
21         // does not call move constructor
22     Vector f();
23         // declares function f that returns Vector
24 }
```

# Constructor Example (Continued 2)

```
1  #include <utility>
2  #include <cstdlib>
3  // include definition of Vector class here
4
5  // named RVO not possible
6  Vector func1() {
7      Vector a(1.0, 0.0);
8      Vector b(0.0, 1.0);
9      if (std::rand() % 2) {return a;}
10     else {return b;}
11 }
12
13 // RVO required
14 Vector func2() {return Vector(1.0, 1.0);}
15
16 int main() {
17     Vector u(1.0, 1.0);
18     Vector v(std::move(u));
19     // move constructor invoked to propagate value from u
20     // to v
21     Vector w = func1();
22     // move constructor invoked to propagate value of object
23     // in return statement of func1 to object w in main
24     // (named RVO not possible)
25     Vector x = func2();
26     // move constructor not invoked, due to guaranteed
27     // copy/move elision (return value of func2 directly
28     // constructed in object x in main)
29 }
```

# Constructor Initializer Lists

- in constructor of class, often we want to control which constructor is used to initialize each data member
- since all data members are constructed *before* body of constructor is entered, this cannot be controlled inside body of constructor
- to allow control over which constructors are used to initialize individual data members, mechanism called **initializer lists** provided
- initializer list forces specific constructors to be used to initialize individual data members before body of constructor is entered
- data members always initialized in *order of declaration*, regardless of order in initializer list



# Constructor\_INITIALIZER List Example

```
1  class ArrayDouble { // array of doubles class
2  public:
3      ArrayDouble(); // create empty array
4      ArrayDouble(int size); // create array of specified size
5      // ...
6  private:
7      // ...
8  };
9
10 class Vector { // n-dimensional real vector class
11 public:
12     Vector(int size) : data_(size) {}
13     // force data_ to be constructed with
14     // ArrayDouble::ArrayDouble(int)
15     // ...
16 private:
17     ArrayDouble data_; // elements of vector
18 };
```

# Default Member Initializers

- can provide default values with which to initialize data members
- if initializer for data member not given in constructor initializer list, default member initializer used if specified
- example:

```
1  #include <string>
2
3  struct Widget {
4      Widget() {}
5          // constructor behaves as if it had initializer
6          // list:
7          // answer(42), message("hello")
8      int answer = 42;
9      std::string message = "hello";
10 };
```

# Member Initialization Order

- recall that data members initialized in order of declaration in class definition
- failing to consider this fact can easily leads to bugs in code
- for example, consider following code:

```
1  #include <cassert>
2
3  class Widget {
4  public:
5      Widget() : y_(42), x_(y_ + 1) {assert(x_ == 43);}
6      int x_;
7      int y_;
8  };
9
10 int main() {
11     Widget w;
12 }
```

- in Widget's default constructor, `x_` initialized before `y_`, which results in use of `y_` before its initialization
- therefore, above code has undefined behavior
- in practice, likely `x_` will simply have garbage value when body of constructor executes and assertion will fail

# Destructors

- when object reaches end of lifetime, typically some cleanup required before object passes out of existence
- **destructor** is member function that is *automatically called* when object reaches end of lifetime in order to perform any necessary cleanup
- often object may have allocated resources associated with it (e.g., memory, files, devices, network connections, processes/threads)
- when object destroyed, must ensure that any resources associated with object are released
- destructors often serve to *release resources* associated with object
- destructor for class `T` always has *name* `T::~~T`
- destructor has *no return type* (not even `void`)
- destructor *cannot be overloaded*
- destructor always takes *no parameters*

- **defaulted destructor** performs no clean-up action, except to destroy each of its data members (and bases)
- defaulted destructor automatically provided (i.e., implicitly defined) if no user-declared destructor [C++17 15.4/4]
- for classes that require additional clean-up, defaulted destructor will not yield correct behavior

# Destructor Example

- example:

```
class Widget {  
  public:  
    Widget(int bufferSize) { // Constructor.  
      // allocate some memory for buffer  
      bufferPtr_ = new char[bufferSize];  
    }  
    ~Widget() { // Destructor.  
      // free memory previously allocated  
      delete [] bufferPtr_;  
    }  
    // copy constructor, assignment operator, ...  
  private:  
    char* bufferPtr_; // pointer to start of buffer  
};
```

- if defaulted destructor were used, memory associated with `bufferPtr_` would not be freed

## Section 2.4.3

# **Operator Overloading**

# Operator Overloading

- can specify meaning of operator whose operands are one or more user-defined types through process known as **operator overloading**
- operators that can be overloaded:

arithmetic	+ - * / %
bitwise	^ &   ~ << >>
logical	! &&
relational	< > <= >= == !=
assignment	=
compound assignment	+= -= *= /= %= ^= &=  = <<= >>=
increment/decrement	++ --
subscript	[ ]
function call	( )
address, indirection	& *
others	->* , -> <b>new delete</b>

- not possible to change precedence/associativity or syntax of operators
- meaning of operator specified by specially named function



# Operator Overloading (Continued 1)

- operator @ overloaded via special function named `operator@`
- with some exceptions, operator can be overloaded as member function or nonmember function
- if operator overloaded as member function, first operand provided as **\*this** and remaining operands, if any, provided as function parameters
- if operator overloaded as nonmember function, all operands provided as function parameters
- postfix unary (increment/decrement) operators take additional dummy parameter of type **int** in order to distinguish from prefix case
- expressions involving overloaded operators interpreted as follows:

Type	Expression	Interpretation As	
		Member Function	Nonmember Function
Binary	<code>a@b</code>	<code>a.operator@(b)</code>	<code>operator@(a, b)</code>
Prefix unary	<code>@a</code>	<code>a.operator@()</code>	<code>operator@(a)</code>
Postfix unary	<code>a@</code>	<code>a.operator@(i)</code>	<code>operator@(a, i)</code>

i is dummy parameter of type **int**

# Operator Overloading (Continued 2)

- assignment, function-call, subscript, and member-selection operators must be overloaded as member functions [\[C++17 16.5.3/1\]](#) [\[C++17 16.5.4/1\]](#) [\[C++17 16.5.5/1\]](#)  
[\[C++17 16.5.6/1\]](#)
- if member and nonmember functions both defined, argument matching rules determine which is called
- if first operand of overloaded operator not object of class type, must use nonmember function
- for most part, operators can be defined quite arbitrarily for user-defined types
- for example, no requirement that “++x”, “x += 1”, and “x = x + 1” be equivalent
- of course, probably not advisable to define operators in very counterintuitive ways, as will inevitably lead to bugs in code

# Operator Overloading (Continued 3)

- some examples showing how expressions translated into function calls are as follows:

Expression	Member Function	Nonmember Function
<code>y = x</code>	<code>y.operator=(x)</code>	—
<code>y += x</code>	<code>y.operator+=(x)</code>	<code>operator+=(y, x)</code>
<code>x + y</code>	<code>x.operator+(y)</code>	<code>operator+(x, y)</code>
<code>++x</code>	<code>x.operator++()</code>	<code>operator++(x)</code>
<code>x++</code>	<code>x.operator++(int)</code>	<code>operator++(x, int)</code>
<code>x == y</code>	<code>x.operator==(y)</code>	<code>operator==(x, y)</code>
<code>x &lt; y</code>	<code>x.operator&lt;(y)</code>	<code>operator&lt;(x, y)</code>

# Operator Overloading Example: Vector

```
1  class Vector { // Two-dimensional vector class
2  public:
3      Vector() : x_(0.0), y_(0.0) {}
4      Vector(double x, double y) : x_(x), y_(y) {}
5      double x() const { return x_; }
6      double y() const { return y_; }
7  private:
8      double x_; // The x component
9      double y_; // The y component
10 };
11
12 // Vector addition
13 Vector operator+(const Vector& u, const Vector& v)
14     {return Vector(u.x() + v.x(), u.y() + v.y());}
15
16 // Dot product
17 double operator*(const Vector& u, const Vector& v)
18     {return u.x() * v.x() + u.y() * v.y();}
19
20 void func() {
21     Vector u(1.0, 2.0);
22     Vector v(u);
23     Vector w;
24     w = u + v; // w.operator=(operator+(u, v))
25     double c = u * v; // calls operator*(u, v)
26     // since c is built-in type, assignment operator
27     // does not require function call
28 }
```

# Operator Overloading Example: Array10

```
1  class Array10 { // Ten-element real array class
2  public:
3      Array10() {
4          for (int i = 0; i < 10; ++i) { // Zero array
5              data_[i] = 0;
6          }
7      }
8      const double& operator[](int index) const {
9          return data_[index];
10     }
11     double& operator[](int index) {
12         return data_[index];
13     }
14 private:
15     double data_[10]; // array data
16 };
17
18 void func() {
19     Array10 v;
20     v[1] = 3.5; // calls Array10::operator[](int)
21     double c = v[1]; // calls Array10::operator[](int)
22     const Array10 u;
23     u[1] = 2.5; // ERROR: u[1] is const
24     double d = u[1]; // calls Array10::operator[](int) const
25 }
```

# Operator Overloading: Member vs. Nonmember Functions

- in most cases, operator can be overloaded as either member or nonmember function
- some considerations that factor into decision of whether to use member or nonmember function given below
- if access to private members is required, using member function may be preferable to having nonmember friend function
- if first operand of operator is of non-class type, must use nonmember function; otherwise, either member or nonmember could be used
- if conversions for first argument to operator are desired, must use nonmember function; if such conversions not desired, must use member function

# Overloading as Member vs. Nonmember: Example

```
1  class Complex { // Complex number type.
2  public:
3      Complex(double x, double y) : x_(x), y_(y) {}
4      double real() const {return x_;}
5      double imag() const {return y_;}
6      // Alternatively, overload operator+ as a member function.
7      // Complex operator+(double b) const
8      // {return Complex(real() + b, imag());}
9  private:
10     double x_; // The real part.
11     double y_; // The imaginary part.
12 };
13
14 // Overload as a nonmember function.
15 // (A member function could instead be used. See above.)
16 Complex operator+(const Complex& a, double b)
17     {return Complex(a.real() + b, a.imag());}
18
19 // This can only be accomplished with a nonmember function.
20 Complex operator+(double b, const Complex& a)
21     {return Complex(b + a.real(), a.imag());}
22
23 void myFunc() {
24     Complex a(1.0, 2.0);
25     Complex b(1.0, -2.0);
26     double r = 2.0;
27     Complex c = a + r; /* could use nonmember or member function
28         operator+(a, r) or a.operator+(r) */
29     Complex d = r + a; /* must use nonmember function
30         operator+(r, a), since r.operator+(a) will not work */
31 }
```

# Overloading as Member vs. Nonmember: Example

```
1  #include <string_view>
2
3  class Widget {
4  public:
5      Widget();
6      Widget(std::string_view); // converting constructor
7      operator std::string_view() const; // conversion operator
8      // ...
9  };
10
11  // overload as nonmember function
12  Widget operator+(Widget, std::string_view);
13
14  int main() {
15      Widget w;
16      std::string_view sv("hello");
17      Widget a = w + sv;
18      /* OK: operator+(Widget, std::string_view) called
19      with no conversions necessary */
20      Widget b = sv + w;
21      /* OK: operator+(Widget, std::string_view) called, where
22      first argument implicitly converted to Widget by
23      Widget's converting constructor and second argument
24      implicitly converted to std::string_view by
25      Widget's conversion operator; if operator+ were
26      overloaded as member of Widget class, compiler error
27      would result as overload resolution would fail to
28      yield any viable function to call */
29  }
```



# Copy Assignment Operator

- for class `T`, `T::operator=` having exactly one parameter that is lvalue reference to `T` known as **copy assignment operator** [C++17 15.8.2/1]
- used to assign, to already-existing object, value of another object by *copying*
- copy assignment operator for class `T` typically is of form  
`T& operator=(const T&) (returning reference to *this)`
- copy assignment operator returns (nonconstant) reference in order to allow for statements like following to be valid (where `x`, `y`, and `z` are of type `T` and `T::modify` is a non-const member function):  

```
x = y = z; // x.operator=(y.operator=(z))  
(x = y) = z; // (x.operator=(y)).operator=(z)  
(x = y).modify(); // (x.operator=(y)).modify()
```
- must be careful to correctly consider case of self-assignment

# Defaulted Copy Assignment Operator

- **defaulted copy assignment operator** performs memberwise copy of its data members (and bases), where copy performed using: [C++17 15.8.2/12]
  - copy assignment operator for class types
  - bitwise copy for built-in types
- defaulted copy assignment operator automatically provided (i.e., implicitly defined) as public member if none of following user declared: [C++17 15.8.2/2]
  - move constructor
  - move assignment operator
  - copy constructor (if not relying on deprecated behavior)
  - destructor (if not relying on deprecated behavior)

- example:

```
// class has implicitly-defined defaulted  
// copy-assignment operator  
class Widget {  
public:  
    Widget(int i) {i_ = i;}  
    int get() const {return i_;}  
private:  
    int i_;  
};
```

# Self-Assignment Example

- in practice, self assignment typically occurs when references (or pointers) are involved
- example:

```
void doSomething(SomeType& x, SomeType& y) {  
    x = y; // self assignment if &x == &y  
    // ...  
}  
  
void myFunc() {  
    SomeType z;  
    // ...  
    doSomething(z, z); // results in self assignment  
    // ...  
}
```

# Move Assignment Operator

- for class `T`, `T::operator=` having exactly one parameter that is rvalue reference to `T` known as **move assignment operator** [C++17 15.8.2/3]
- used to assign, to already-existing object, value of another object by *moving*
- move assignment operator for class `T` typically is of form  
`T& operator=(T&&)` (returning reference to `*this`)
- move assignment operator returns (nonconstant) reference for same reason as in case of copy assignment operator
- in case of move, self-assignment should probably not occur, but might be prudent to test for this with assertion in order to protect against “insane” code
- standard library effectively forbids self-assignment for move [C++17 20.5.4.9/(1.3)]

# Defaulted Move Assignment Operator

- **defaulted move assignment operator** performs memberwise move of its data members (and bases) where move performed using: [C++17 15.8.2/12]
  - if class type: move assignment operator if available and copy assignment operator otherwise
  - if built-in type: bitwise copy
- defaulted move assignment operator automatically provided (i.e., implicitly defined) as public member if none of following user declared: [C++17 15.8.2/4]
  - copy constructor
  - move constructor
  - copy assignment operator
  - destructor

- **example:**

```
#include <vector>

// class has implicitly-defined defaulted
// move-assignment operator
struct Widget {
    Widget();
    std::vector<int> v;
};
```

# Copy/Move Assignment Operator Example: Complex

```
1  class Complex {
2  public:
3      Complex(double x = 0.0, double y = 0.0) :
4          x_(x), y_(y) {}
5      Complex(const Complex& a) : x_(a.x_), y_(a.y_) {}
6      Complex(Complex&& a) : x_(a.x_), y_(a.y_) {}
7      Complex& operator=(const Complex& a) { // Copy assign
8          if (this != &a) {
9              x_ = a.x_; y_ = a.y_;
10         }
11         return *this;
12     }
13     Complex& operator=(Complex&& a) { // Move assign
14         x_ = a.x_; y_ = a.y_;
15         return *this;
16     }
17 private:
18     double x_; // The real part.
19     double y_; // The imaginary part.
20 };
21
22 int main() {
23     Complex z(1.0, 2.0);
24     Complex v(1.5, 2.5);
25     v = z; // v.operator=(z)
26     v = Complex(0.0, 1.0); // v.operator=(Complex(0.0, 1.0))
27 }
```

# Assignment Operator Example: Buffer

```
1  class Buffer { // Character buffer class.
2  public:
3      Buffer(int bufferSize) { // Constructor.
4          bufSize_ = bufferSize;
5          bufPtr_ = new char[bufferSize];
6      }
7      Buffer(const Buffer& buffer) { // Copy constructor.
8          bufSize_ = buffer.bufSize_;
9          bufPtr_ = new char[bufSize_];
10         for (int i = 0; i < bufSize_; ++i)
11             bufPtr_[i] = buffer.bufPtr_[i];
12     }
13     ~Buffer() { // Destructor.
14         delete [] bufPtr_;
15     }
16     Buffer& operator=(const Buffer& buffer) { // Copy assignment operator.
17         if (this != &buffer) {
18             delete [] bufPtr_;
19             bufSize_ = buffer.bufSize_;
20             bufPtr_ = new char[bufSize_];
21             for (int i = 0; i < bufSize_; ++i)
22                 bufPtr_[i] = buffer.bufPtr_[i];
23         }
24         return *this;
25     }
26     // ...
27 private:
28     int bufSize_; // buffer size
29     char* bufPtr_; // pointer to start of buffer
30 };
```

- without explicitly-provided assignment operator (i.e., with defaulted assignment operator), memory leaks and memory corruption would result

## Section 2.4.4

### **Miscellany**



# std::initializer\_list Class Template

- class template `std::initializer_list` provides lightweight list type
- in order to use `initializer_list`, need to include header file `initializer_list`
- declaration:  

```
template <class T> initializer_list;
```
- T is type of elements in list
- `initializer_list` is very lightweight
- can query number of elements in list and obtain iterators to access these elements
- `initializer_list` often useful as parameter type for constructor

# std::initializer\_list Example

```
1  #include <iostream>
2  #include <vector>
3
4  class Sequence {
5  public:
6      Sequence(std::initializer_list<int> list) {
7          for (std::initializer_list<int>::const_iterator i =
8              list.begin(); i != list.end(); ++i)
9              elements_.push_back(*i);
10     }
11     void print() const {
12         for (std::vector<int>::const_iterator i =
13             elements_.begin(); i != elements_.end(); ++i)
14             std::cout << *i << '\n';
15     }
16 private:
17     std::vector<int> elements_;
18 };
19
20 int main() {
21     Sequence seq = {1, 2, 3, 4, 5, 6};
22     seq.print();
23 }
```

# Converting Constructors

- constructor that is not declared with **explicit** specifier is called **converting constructor**
- converting constructor can be used for *implicit* conversions
- example:

```
1  #include <string>
2  using namespace std::literals;
3
4  class Widget {
5  public:
6      Widget(const std::string&); // converting constructor
7      Widget(const char*, int); // converting constructor
8      // ...
9  };
10
11 int main() {
12     Widget v = "hello"s;
13     // invokes Widget::Widget(const std::string&)
14     Widget w = {"goodbye", 4};
15     // invokes Widget::Widget(const char*, int)
16     v = "bonjour"s; // invokes Widget::Widget(const std::string&)
17     w = {"au revoir", 2};
18     // invokes Widget::Widget(const char*, int)
19 }
```

# Explicit Constructors

- converting constructor can be used in implicit conversions (e.g., when attempting to obtain matching type for function parameter in function call)
- often, desirable to prevent constructor from being used for implicit conversions
- to accommodate this, constructor can be marked as explicit
- **explicit constructor** is constructor that cannot be used for performing implicit conversions or copy initialization
- prefixing constructor declaration with **explicit** keyword makes constructor explicit
- example:

```
class Widget {  
public:  
    explicit Widget(int); // explicit constructor  
    // ...  
};
```

# Example Without Explicit Constructor

```
1  #include <cstdlib>
2
3  // one-dimensional integer array class
4  class IntArray {
5  public:
6      // create array of int with size elements
7      IntArray(std::size_t size) { /* ... */;
8      // ...
9  };
10
11 void processArray(const IntArray& x) {
12     // ...
13 }
14
15 int main() {
16     // following lines of code almost certain to be
17     // incorrect, but valid due to implicit type
18     // conversion provided by
19     // IntArray::IntArray(std::size_t)
20     IntArray a = 42;
21     // probably incorrect
22     // implicit conversion effectively yields code:
23     // IntArray a = IntArray(42);
24     processArray(42);
25     // probably incorrect
26     // implicit conversion effectively yields code:
27     // processArray(IntArray(42));
28 }
```

# Example With Explicit Constructor

```
1  #include <cstdlib>
2
3  // one-dimensional integer array class
4  class IntArray {
5  public:
6      // create array of int with size elements
7      explicit IntArray(std::size_t size) { /* ... */ };
8      // ...
9  };
10
11 void processArray(const IntArray& x) {
12     // ...
13 }
14
15 int main() {
16     IntArray a = 42; // ERROR: cannot convert
17     processArray(42); // ERROR: cannot convert
18 }
```

# Conversion Operators

- conversion operator enables implicit or explicit conversion from class type to another type
- conversion operator can be either explicit or non-explicit
- if conversion operator is explicit, can only be used for explicit conversions; otherwise, can be used for both implicit and explicit conversions
- conversion operator to convert from class `T` to another type must be provided as (nonstatic) member function of `T`
- member function takes no parameters (except implicit **this** parameter) and has no explicit return type
- **explicit** keyword can be placed before name in declaration to make conversion operator explicit
- requiring conversion operator to be member function likely motivated by desire to allow code for performing conversions to be more easily identified (i.e., only need to examine conversion operators and converting constructors for at most two classes)

# Conversion Operator Example

```
1  #include <iostream>
2  #include <string>
3  #include <cassert>
4  using namespace std::literals;
5
6  class Widget {
7  public:
8      explicit operator int() const {return 42;}
9      operator std::string() const {return "Widget"s;}
10     // ...
11 };
12
13 int main() {
14     Widget w;
15     int i(w);
16     // direct initialization can use explicit conversion operator;
17     // uses conversion operator to convert Widget to int
18     assert(i == 42);
19     // int j = w;
20     // ERROR: copy initialization requires implicit conversion and
21     // conversion operator that converts Widget to int is explicit
22     int j = static_cast<int>(w);
23     // uses (explicit) conversion operator to convert Widget to int
24     std::string s(w);
25     // uses conversion operator to convert Widget to std::string
26     assert(s == "Widget"s);
27     std::string t = w;
28     // uses conversion operator to convert Widget to std::string
29     assert(t == "Widget"s);
30 }
```



# Explicitly Deleted/Defaulted Special Member Functions

- can explicitly default or delete special member functions (i.e., default constructor, copy constructor, move constructor, destructor, copy assignment operator, and move assignment operator)
- can also delete non-special member functions
- example:

```
class Thing {  
public:  
    Thing() = default;  
  
    // Prevent copying.  
    Thing(const Thing&) = delete;  
    Thing& operator=(const Thing&) = delete;  
  
    Thing(Thing&&) = default;  
    Thing& operator=(Thing&&) = default;  
    ~Thing() = default;  
    // ...  
};  
  
// Thing is movable but not copyable.
```

# Delegating Constructors

- sometimes, one constructor of class needs to performs all work of another constructor followed by some additional work
- rather than duplicate common code in both constructors, one constructor can use its initializer list to invoke other constructor (which must be only one in initializer list)
- constructor that invokes another constructor via initializer list called **delegating constructor**
- example:

```
1  class Widget {
2  public:
3      Widget(char c, int i) : c_(c), i_(i) {}
4      Widget(int i) : Widget('a', i) {}
5          // delegating constructor
6          // ...
7  private:
8      char c_;
9      int i_;
10 };
11
12 int main() {
13     Widget w('A', 42);
14     Widget v(42);
15 }
```

# Static Data Members

- sometimes want to have object that is shared by all objects of class
- data member that is shared by all objects of class is called **static data member**
- to make data member static, declare using **static** qualifier
- static data member must (in most cases) be defined outside body of class
- example:

```
1  class Widget {  
2  public:  
3      Widget() {++count_;}  
4      Widget(const Widget&) {++count_;}  
5      Widget(Widget&&) {++count_;}  
6      ~Widget() {--count_;}  
7      // ...  
8  private:  
9      static int count_;  
10     // total number of Widget objects in existence  
11 };  
12  
13 // Define (and initialize) count member.  
14 int Widget::count_ = 0;
```

# Static Member Functions

- sometimes want to have member function that does not operate on objects of class
- member function of class that does not operate on object of class (i.e., has no **this** variable) called **static member function**
- to make member function static, declare using **static** qualifier
- example:

```
1  class Widget {  
2  public:  
3      // ...  
4      // convert degrees to radians  
5      static double degToRad(double deg)  
6          {return (M_PI / 180.0) * deg;}  
7  private:  
8      // ...  
9  };  
10  
11 void func() {  
12     Widget x; double rad;  
13     rad = Widget::degToRad(45.0);  
14     rad = x.degToRad(45.0); // x is ignored  
15 }
```

- like non-member functions, member functions can also be qualified as **constexpr** to indicate function can be computed *at compile time* provided that all arguments to function are constant expressions
- some additional restrictions on constexpr member functions relative to nonmember case (e.g., cannot be virtual)
- constexpr member function *implicitly inline*
- constexpr member function *not implicitly const* (as of C++14)

- constructors can also be qualified as **constexpr** to indicate object construction can be performed *at compile time* provided that all arguments to constructor are constant expressions
- many restrictions on what types can have constexpr constructors and what such constructors can do (e.g., no virtual base classes)
- constexpr constructor *implicitly inline*

# Example: Constexpr Constructors and Member Functions

```
1  #include <iostream>
2
3  // Two-dimensional vector class.
4  class Vector {
5  public:
6      constexpr Vector() : x_(0), y_(0) {}
7      constexpr Vector(double x, double y) : x_(x), y_(y) {}
8      constexpr Vector(const Vector& v) : x_(v.x_), y_(v.y_) {}
9      constexpr Vector& operator=(const Vector& v)
10         {x_ = v.x_; y_ = v.y_; return *this;}
11      constexpr double x() const {return x_;}
12      constexpr double y() const {return y_;}
13      constexpr double squared_norm() const
14         {return x_ * x_ + y_ * y_;}
15         // ...
16 private:
17     double x_; // The x component of the vector.
18     double y_; // The y component of the vector.
19 };
20
21 int main() {
22     constexpr Vector v(3.0, 4.0);
23     static_assert(v.x() == 3.0 && v.y() == 4.0);
24     constexpr double d = v.squared_norm();
25     std::cout << d << '\n';
26 }
```

# Why Constexpr Member Functions Are Not Implicitly Const

```
1  class Widget {
2  public:
3      constexpr Widget() : i_(42) {}
4      constexpr const int& get() const {return i_;}
5      constexpr int& get() /* what if implicitly const? */
6          {return i_;}
7      // ...
8  private:
9      int i_;
10 };
11
12 constexpr Widget w;
13 static_assert(w.get() == 42);
14     // invokes const member function
15 constexpr int i = ++Widget().get();
16     // invokes non-const member function
17 static_assert(i == 43);
```

- in above code example, we want to have const and non-const overloads of `get` member function that can each be used in constant expressions
- so both overloads of `get` need to be `constexpr`
- if `constexpr` member functions were implicitly `const`, it would be impossible to overload on `const` in manner we wish to do here, since second overload of `get` would automatically become `const` member function (resulting in multiple conflicting definitions of `const` member function `get`)



- type for nonstatic data member can be qualified as **mutable** meaning that member does not affect externally visible state of class object
- mutable data member can be modified in const member function
- **mutable** qualifier often used for mutexes, condition variables, cached values, statistical information for performance analysis or debugging

# Example: Mutable Qualifier for Statistical Information

SKIP SLIDE

```
1  #include <iostream>
2  #include <string>
3
4  class Employee {
5  public:
6      Employee(int id, std::string& name, double salary) :
7          id_(id), name_(name), salary_(salary), accessCount_(0) {}
8      int getId() const {
9          ++accessCount_; return id_;
10     }
11     std::string getName() const {
12         ++accessCount_; return name_;
13     }
14     double getSalary() const {
15         ++accessCount_; return salary_;
16     }
17     // ...
18     // for debugging
19     void outputDebugInfo(std::ostream& out) const {
20         out << accessCount_ << '\n';
21     }
22 private:
23     int id_; // employee ID
24     std::string name_; // employee name
25     double salary_; // employee salary
26     mutable unsigned long accessCount_; // for debugging
27 };
```

# Pointers to Members

- pointer to member provides means to reference particular *nonstatic* member of class, independent of any class object instance
- pointer to member can only be formed for *nonstatic* (data or function) member of class
- can obtain pointer to member that references nonstatic member *m* in class *T* by applying address-of operator to *T::m* (i.e., using expression *&T::m*)
- special value **nullptr** can be given to pointer to member to indicate that pointer to member does not refer to any member
- pointer to member of class *T* written as *T::\**
- type of pointer to member embodies type of class and type of member within class
- example:
  - **int** *Widget::\** is pointer to member of *Widget* class having type **int**
  - **const int** *Widget::\** is pointer to member of *Widget* class having type **const int**
  - **float** (*Gadget::\**) (**int**) **const** is pointer to const member function of *Gadget* class that takes single **int** parameter and has return type of **float**

## Pointers to Members (Continued)

- since pointer to member is not associated with any class object instance, dereferencing pointer to member requires object (or pointer to object) to be specified
- given object `x` of type `T`, can access member through pointer to member `ptm` by applying member-selection operator `.*` to `x` using expression `x.*ptm`
- given pointer `p` to object of type `T`, can access member through pointer to member `ptm` by applying member-selection operator `->*` to `p` using expression `p->*ptm`

# Pointers to Members for Data Members

- conceptually, pointer to member for (nonstatic) data member can be thought of as offset (in memory) from start of class object to start of data member (i.e., location of data member *relative to start of class object*)
- since pointer to member does not identify particular object instance (i.e., value for **this**), pointer to member alone not sufficient to specify *particular instance* of member in object
- consequently, when dereferencing pointer to member, must always specify object (or pointer to object)
- example:

```
1  struct Widget {
2      int i;
3      inline static int j;
4  };
5
6  int main(){
7      Widget w, v;
8      int Widget::* ptm = &Widget::i; // pointer to member
9      int* jp = &Widget::j;
10     // address of static member is ordinary pointer
11     w.*ptm = 42; // w.*ptm references w.i
12     v.*ptm = 42; // v.*ptm references v.i
13     *jp = 42; // references Widget::j
14 }
```

# Pointers to Members and Const Example

SKIP SLIDE

```
1  #include <type_traits>
2
3  template <class T1, class T2, class T3>
4  struct triplet {
5      triplet(T1 first_, T2 second_, T3 third_) :
6          first(first_), second(second_), third(third_) {}
7      T1 first;
8      T2 second;
9      T3 third;
10 };
11
12 int main() {
13     using widget = triplet<const int, int, double>;
14     widget w(1, 1, 1.0);
15     static_assert(std::is_same_v<decltype(&widget::first),
16                   const int widget::*>);
17     static_assert(std::is_same_v<decltype(&widget::second),
18                   int widget::*>);
19     static_assert(std::is_same_v<decltype(&widget::third),
20                   double widget::*>);
21     const int widget::* cp = nullptr;
22     int widget::* p = nullptr;
23     cp = &widget::first; // OK: constness of pointee same
24     cp = &widget::second; // OK: adds const to pointee
25     // p = &widget::first; // ERROR: discards const from pointee
26     p = &widget::second; // OK: constness of pointee same
27 }
```

# Pointers to Members for Function Members

SKIP SLIDE

- pointer to member for (nonstatic) member function simply identifies particular member function of class (independent of any object instance)
- since pointer to member does not identify particular object instance (i.e., value for **this**), pointer to member alone not sufficient to invoke member function
- consequently, when dereferencing pointer to member, must always specify object (or pointer to object) so that **this** parameter can be set appropriately
- example:

```
1  struct Widget {
2      void func() { /* ... */ }
3      static void set_verbosity(int level) { /* ... */ }
4  };
5
6  int main() {
7      Widget w, v;
8      void (Widget::* ptm)() = &Widget::func; // pointer to member
9      void (*pf)(int) = &Widget::set_verbosity;
10     // address of static member is ordinary pointer
11     (w.*ptm)(); // calls w.func()
12     (v.*ptm)(); // calls v.func()
13     (*pf)(42); // calls Widget::set_verbosity()
14 }
```

# Pointers to Members: Example

SKIP SLIDE

```
1  #include <string>
2  #include <cassert>
3
4  struct Widget {
5      std::string s;
6      int i = 0;
7      void clear() {i = 0; s = "";}
8  };
9
10 int main() {
11     Widget w;
12     Widget* wp = &w;
13
14     // pointer to member of Widget of type int
15     int Widget::* iptm = nullptr;
16     // w.*iptm = 42; // ERROR: null pointer to member
17     iptm = &Widget::i; // iptm references i member of Widget
18     w.*iptm = 42; // w.*iptm references w.i
19     assert(w.i == 42);
20
21     // pointer to member of Widget of type std::string
22     std::string Widget::* sptm = &Widget::s;
23     wp->*sptm = "hello"; // wp->*sptm references w.s
24     assert(w.s == "hello");
25
26     // pointer to member of Widget that is function that takes
27     // no parameters and returns void
28     void (Widget::* fptm)() = &Widget::clear;
29     (w.*fptm)(); // w.*fptm references w.clear
30     assert(w.i == 0 && w.s == "");
31 }
```



# Pointers to Members: Example

SKIP SLIDE

```
1  #include <iostream>
2
3  class Widget {
4  public:
5      Widget(bool flag) {
6          op_ = flag ? &Widget::op_2 : &Widget::op_1;
7      }
8      void modify() {
9          // ...
10         (this->*op_ )(); // invoke member function
11         // ...
12     }
13     // ...
14 private:
15     void op_1() {std::cout << "op_1 called\n";}
16     void op_2() {std::cout << "op_2 called\n";}
17     void (Widget::*op_ )();
18     // pointer to member function of Widget class that
19     // takes no parameters and returns no value
20     // ...
21 };
22
23 int main() {
24     Widget u(false);
25     Widget v(true);
26     u.modify(); // modify invokes op_1
27     v.modify(); // modify invokes op_2
28 }
```

# Pointers to Members Example: Accumulate

```
1  #include <iostream>
2  #include <iterator>
3
4  struct Point {double x; double y;};
5  struct Thing {int i; float f;};
6
7  template <auto P, class Iter, class T>
8  T accumulate(Iter first, Iter last, T init_sum) {
9      for (auto i = first; i != last; ++i)
10         {init_sum += i->*P;}
11     return init_sum;
12 }
13
14 int main() {
15     constexpr Point p[]{{1.0, 21.0}, {0.5, 21.0}, {0.5, 0.0}};
16     constexpr Thing t[]{{1, 0.1f}, {2, 0.1f}, {3, 0.1f}};
17     std::cout
18         << accumulate<&Point::x>(std::begin(p), std::end(p), 0.0) << ' '
19         << accumulate<&Point::y>(std::begin(p), std::end(p), 0.0) << '\n';
20     std::cout
21         << accumulate<&Thing::i>(std::begin(t), std::end(t), 0) << ' '
22         << accumulate<&Thing::f>(std::begin(t), std::end(t), 0.0f) << '\n';
23 }
```

# Pointers to Members Example: Statistics Calculation

```
1  #include <iostream>
2
3  template <auto Count, auto Sum, class T, class Value>
4  void update_statistics(T& stats, Value value) {
5      ++(stats.*Count); // adjust count of values
6      stats.*Sum += value; // adjust sum of values
7  }
8
9  struct Widget {
10     int count = 0; // count
11     double sum = 0; // sum
12     short int si;
13 };
14
15 struct Gadget {
16     int n = 0; // count
17     double d;
18     double sigma = 0; // sum
19 };
20
21 int main() {
22     Widget w;
23     Gadget g;
24     for (auto&& x : {0.5, 1.5, 2.5}) {
25         update_statistics<&Widget::count, &Widget::sum>(w, x);
26         update_statistics<&Gadget::n, &Gadget::sigma>(g, x);
27     }
28     std::cout << w.sum / static_cast<double>(w.count) << '\n';
29     std::cout << g.sigma / static_cast<double>(g.n) << '\n';
30 }
```

- stream inserters write data to output stream

- overload **operator<<**

- have general form

`std::ostream& operator<<(std::ostream&, T)` where type T is typically const lvalue reference type

- example:

```
std::ostream& operator<<(std::ostream& outStream,
    const Complex& a)
{
    outStream << a.real() << ' ' << a.imag();
    return outStream;
}
```

- inserter and extractor should use *compatible formats* (i.e., what is written by inserter should be readable by extractor)

- stream extractors read data from input stream

- overload **operator>>**

- have general form

`std::istream& operator>>(std::istream&, T)` where type T is typically non-const lvalue reference type

- example:

```
std::istream& operator>>(std::istream& inStream,
    Complex& a)
{
    double real = 0.0;
    double imag = 0.0;
    inStream >> real >> imag;
    a = Complex(real, imag);
    return inStream;
}
```

# Structured Bindings

- structured bindings allow, with single statement, multiple variables to be declared and initialized with values from pair, tuple, array, or struct
- declaration uses auto keyword
- variables enclosed in brackets
- multiple variables separated by commas

# Structured Bindings Example

```
1  #include <tuple>
2  #include <array>
3  #include <cassert>
4
5  int main() {
6      int a[3] = {1, 2, 3};
7      auto [a0, a1, a2] = a;
8      assert(a0 == a[0] && a1 == a[1] && a2 == a[2]);
9
10     int b[3] = {0, 2, 3};
11     auto& [b0, b1, b2] = b;
12     ++b0;
13     assert(b[0] == 1);
14
15     std::array<int, 3> c = {1, 2, 3};
16     auto [c0, c1, c2] = c;
17     assert(c0 == c[0] && c1 == c[1] && c2 == c[2]);
18
19     auto t = std::tuple(true, 42, 'A');
20     auto [tb, ti, tc] = t;
21     assert(tb == true && ti == 42 && tc == 'A');
22 }
```

# Structured Bindings Example

```
1  #include <map>
2  #include <string>
3  #include <iostream>
4
5  int main() {
6      std::map<std::string, int> m = {
7          {"apple", 1},
8          {"banana", 2},
9          {"orange", 3},
10     };
11     for (auto&& [key, value] : m) {
12         std::cout << key << ' ' << value << '\n';
13     }
14 }
```



- each of following types said to be **literal type**: [C++17 6.9/10]
  - **void**
  - scalar type (e.g., integral, floating point, pointer, enumeration, pointer to member)
  - reference type
  - class type that has all of following properties:
    - *has trivial destructor*
    - is either: aggregate type; or type with at least one constexpr constructor that is not copy or move constructor; or closure type
    - all nonstatic data members and base classes are of nonvolatile literal types
  - array of literal type
- examples of literal types:
  - **int**, **double**[16], and `std::complex<double>`
- examples of types that are not literal types:
  - `std::vector<int>` and `std::string`
- literal types important in context of constexpr variables, functions, and constructors

# Example: Literal Types

```
1  // literal type
2  class Widget {
3  public:
4      constexpr Widget(int i = 0) : i_(i) {}
5      ~Widget() = default; // trivial destructor
6  private:
7      int i_;
8  };
9
10 // not literal type
11 class Gadget {
12 public:
13     constexpr Gadget() {}
14     ~Gadget() {} // non-trivial destructor
15 };
16
17 // not literal type
18 // no constexpr constructor, excluding copy/move constructor
19 class Foo {
20 public:
21     Foo() {}
22     ~Foo() = default; // trivial destructor
23 };
```

- constexpr variable must satisfy following requirements: [C++17 10.1.5/9]
  - its type must be literal type
  - it must be immediately initialized
  - full expression of its initialization must be constant expression (including all implicit conversions and constructor calls)

# Example: constexpr Variable Requirement Violations

```
1  #include <string>
2
3  constexpr std::string s("hello");
4      // ERROR: not literal type
5
6  constexpr int i;
7      // ERROR: not initialized
8
9  float func();
10 constexpr float f = 2.0 * func();
11      // ERROR: initializer expression not constant expression
```

# Constexpr Function Requirements

- constexpr function must satisfy following requirements: [C++17 10.1.5/3]

[C++17 10.1.5/5]

- must not be virtual
- its return type must be literal type
- each of its parameters must be of literal type
- there exists at least one set of argument values such that invocation of function could be evaluated expression of core constant expression
- function body must be either deleted or defaulted or contain any statements except:
  - asm declaration
  - goto statement
  - statement with label other than **case** and **default**
  - try block
  - definition of variable of non-literal type
  - definition of variable of static or thread storage duration
  - definition of variable for which no initialization is performed
- if function is defaulted copy/move assignment, class of which it is member must not have mutable member

# Example: constexpr Function Requirement Violations

```
1  #include <vector>
2  #include <string>
3  #include <iostream>
4
5  // ERROR: return type not literal type
6  constexpr std::vector<int> get_value()
7      {return std::vector<int>{1, 2, 3};}
8
9  // ERROR: parameter type not literal type
10 constexpr void foo(std::string s) { /* ... */ }
11
12 // ERROR: no argument exists such that function can be used
13 // in constant expression
14 constexpr void output(int i) {std::cout << i << '\n';}
15
16 constexpr void func() {
17     int i; // ERROR: variable not initialized
18     std::vector<int> v{1, 2, 3};
19     // ERROR: definition of variable of non-literal type
20     // ...
21 }
22
23 constexpr int count() {
24     static unsigned int i = 0;
25     // ERROR: definition of variable with static storage
26     // duration
27     return i++;
28 }
```

# Constexpr Constructor Requirements

- constexpr constructor must satisfy following requirements: [\[C++17 10.1.5/4\]](#)

[\[C++17 10.1.5/5\]](#)

- each of its parameters must be of literal type
- class must not have any virtual base classes
- constructor must not have function try block
- constructor body must be either deleted or defaulted or satisfy following constraints:
  - compound statement of constructor body must satisfy constraints for body of constexpr function
  - every base class sub-object and every non-static data member must be initialized
  - every constructor selected to initialize non-static members and base class must be constexpr constructor
- if constructor is defaulted copy/move constructor, class of which it is member must not have mutable member

# Example: Constexpr Constructor Requirement Violations

```
1  #include <string>
2
3  class Widget {
4  public:
5      constexpr Widget() {}
6          // ERROR: i_ not initialized
7      constexpr Widget(std::string s);
8          // ERROR: parameter type not literal type
9          // ...
10 private:
11     int i_;
12 };
13
14 // OK
15 class Base {
16 public:
17     Base(int i) : i_(i) {}
18 private:
19     int i_;
20 };
21
22 class Derived : public Base {
23 public:
24     constexpr Derived() : Base(42) {}
25         // ERROR: Base constructor not constexpr
26         // ...
27 };
```



# Example: Addresses and Constexpr

```
1  #include <cassert>
2
3  constexpr void func_1() {
4      char c = 'A';
5      // c has automatic storage (i.e., on stack)
6      const char* p = &c;
7      // OK: address of c is well defined
8      assert(*p == 'A'); // OK
9      // constexpr const char* q = &c;
10     /* ERROR: &c not constant expression;
11        address of automatic object can be different
12        for each invocation of func_1 */
13     // above results also same if c is const or constexpr
14 }
15
16 static char sc = 'A';
17 // sc has static storage (i.e., in program image)
18 constexpr void func_2() {
19     const char* p = &sc;
20     // OK: address of sc is fixed and known at compile time
21     constexpr const char* q = &sc;
22     /* OK: address of static object sc is fixed and known
23        at compile time */
24     // above results also same if sc is const or constexpr
25 }
```

# Example: Pointers/References to Constexpr Objects

```
1  class Buffer {
2  public:
3      constexpr Buffer() : data_() {}
4      constexpr const char& operator[](unsigned int i) const
5          {return data_[i];}
6      constexpr char& operator[](unsigned int i)
7          {return data_[i];}
8      constexpr const char* data() const {return data_;}
9      // ...
10 private:
11     char data_[256];
12 };
13
14 int main() {
15     constexpr Buffer b; // OK
16     constexpr Buffer a = b; // OK
17     constexpr char c = b[0]; // OK
18     // constexpr const Buffer& br = b;
19     // ERROR: reference to b is not a constant expression
20     // constexpr const char& cr = b[0];
21     // ERROR: reference to subobject of b is not constant
22     // expression
23     // constexpr const char* cp = b.data();
24     // ERROR: pointer to subobject of b is not constant
25     // expression
26     // constexpr const Buffer* bp = &b;
27     // ERROR: pointer to b is not constant expression
28 }
```

# Example: Constexpr and Accessing External State

```
1  // ERROR: static object not allowed in constexpr function
2  // (since function would have state the persists across
3  // invocations)
4  constexpr unsigned int get_count() {
5      static int count = 0;
6      return count++;
7  }
8
9  int global_count = 0;
10 // ERROR: constexpr function cannot modify state outside
11 // that function that might be used at run time
12 constexpr int get_global_count() {return global_count++;}
13
14 double alpha = 2.0;
15 // ERROR: constexpr function cannot access state outside
16 // that function that can be modified at run time.
17 constexpr double foo(double x) {return alpha * x + 3.0;}
```

# Example: Constexpr and Immediate Initialization

```
1  #include <iostream>
2  #include <cstdint>
3
4  template <class T, std::size_t Size>
5  class Buffer {
6  public:
7      // ERROR: data_ data member is constructed without
8      // being (immediately) initialized
9      // constexpr Buffer() {
10     //   for (std::size_t i = 0; i < size(); ++i)
11     //     {data_[i] = 0;}
12     // }
13     // OK: all data members are initialized when constructed
14     // (array data_ is initialized to all zero)
15     constexpr Buffer() : data_{} {}
16     static constexpr std::size_t size() {return Size;}
17     const T& operator[](std::size_t i) const {return data_[i];}
18     // ...
19 private:
20     T data_[size()];
21 };
22
23 int main() {
24     constexpr Buffer<char, 16> b;
25     for (int i = 0; i < 16; ++i)
26         {std::cout << static_cast<int>(b[i]) << '\n';}
27 }
```

# Debugging Constexpr Functions

- debugging constexpr code can often be somewhat tricky
- cannot generate output to assist in debugging (e.g., by writing to standard output/error stream) since cannot perform I/O at compile time
- use of source-level debugger not practical, since compiler would need to be run in debugger
- could first debug code without constexpr qualifier and then add constexpr qualifier after code is working, but this may not be practical if code must fundamentally execute at compile time (e.g., due to return value of function being assigned to constexpr variable)
- can use `assert` to test for conditions indicative of bugs (since `assert(expr)` is constant expression if *expr* is **true**)
- can throw exception if condition indicative of bug is detected (since still constant expression as long as throw statement not executed)

# Example: Debugging Strategies for Constexpr Functions

```
1  #include <stdexcept>
2  #include <cassert>
3
4  constexpr double sqrt(double x) {
5      // if assertion fails, sqrt function will not yield
6      // constant expression
7      assert(x >= 0.0);
8      double result = 0.0;
9      // ... (correctly initialize result)
10     return result;
11 }
12
13 constexpr int foo(unsigned x) {
14     unsigned i = 0;
15     // ... (code that changes i)
16     // assume odd i indicative of bug
17     // if i is odd (which would result in exception
18     // being thrown), foo function will not yield
19     // constant expression
20     if (i & 1) throw std::logic_error("i is odd");
21     return 0;
22 }
```

- proxy class provides modified interface to another class
- classic example of proxy class is type returned by nonconst overload of subscript operator in `std::vector<bool>` class
- in this case, proxy type serves as stand-in for single `bool` element in vector

# Proxy Class Example: BoolVector

- in this example, we consider simple container class called `BoolVector` that provides dynamically-sized array of boolean values, where booleans are *packed* into bytes
- want to provide only very basic functionality for class:
  - member function for querying size of container
  - subscript operator for accessing elements in container
- return type of nonconst overload of subscript operator is proxy type (called `Proxy`) in order to handle fact that bits are packed into bytes
- `BoolVector` is essentially greatly simplified version of `std::vector<bool>`



# Proxy Class Example: BoolVector.hpp

```
1  #include <cstdint>
2  #include <utility>
3
4  class BoolVector;
5
6  class Proxy {
7  public:
8      ~Proxy() = default;
9      Proxy& operator=(const Proxy&);
10     Proxy& operator=(bool b);
11     operator bool() const;
12 private:
13     friend class BoolVector;
14     Proxy(const Proxy&) = default;
15     Proxy(BoolVector* v, std::size_t i) : v_(v), i_(i) {}
16     BoolVector* v_;
17     std::size_t i_;
18 };
19
20 class BoolVector {
21 public:
22     BoolVector(std::size_t n) : n_(n), d_(new unsigned char[(n + 7) / 8]) {std::fill_n(d_, (n + 7) / 8, 0);}
23     ~BoolVector() {delete [] d_;}
24     std::size_t size() const {return n_;}
25     bool operator[](std::size_t i) const {return getElem(i);}
26     Proxy operator[](std::size_t i) {return Proxy(this, i);}
27 private:
28     friend class Proxy;
29     bool getElem(std::size_t i) const {return (d_[i / 8] >> (i % 8)) & 1;}
30     void setElem(std::size_t i, bool b) {(d_[i / 8] &= ~(1 << (i % 8))) |= (b << (i % 8));}
31     std::size_t n_;
32     unsigned char* d_;
33 };
34
35 inline Proxy& Proxy::operator=(const Proxy& other) {v_->setElem(i_, other); return *this;}
36 inline Proxy& Proxy::operator=(bool b) {v_->setElem(i_, b); return *this;}
37 inline Proxy::operator bool() const {return v_->getElem(i_);}
```

# Proxy Class Example: BoolVector.cpp

```
1  #include <cassert>
2  #include <iostream>
3  #include "BoolVector.hpp"
4
5  int main() {
6      constexpr int bits[] = {0, 0, 1, 1, 0, 1, 0, 1};
7      constexpr int n = sizeof(bits) / sizeof(int);
8      BoolVector v(n);
9      BoolVector w(n);
10     assert(v.size() == n && w.size() == n);
11     for (int i = 0; i < n; ++i) {
12         w[i] = v[i] = bits[i];
13     }
14     const BoolVector& cv = v;
15     for (int i = 0; i < n; ++i) {
16         assert(v[i] == bits[i]);
17         assert(w[i] == bits[i]);
18         assert(cv[i] == bits[i]);
19         std::cout << (v[i] ? '1' : '0');
20     }
21     std::cout << '\n';
22 }
```

# Implementing Postfix Increment/Decrement Operator

- often, good idea to implement postfix increment/decrement operator in terms of prefix increment/decrement operator
- ensures that prefix and postfix versions of operator always consistent
- example:

```
1  class Counter {  
2  public:  
3      Counter(int count = 0) : count_(count) {}  
4      Counter& operator++() {  
5          ++count_;  
6          return *this;  
7      }  
8      Counter operator++(int) {  
9          Counter old(*this);  
10         ++(*this);  
11         return old;  
12     }  
13     // similarly for prefix/postfix decrement  
14 private:  
15     int count_;  
16 };
```

# Pointer-to-Implementation (Pimpl) Idiom

- **pointer to implementation (pimpl)** idiom splits interface and implementation across two classes, namely, handle class and implementation class
- all implementation details placed in implementation class
- handle class provides only interface functions which simply forward calls through to implementation class
- handle object has pointer that owns implementation object
- only handle class is exposed to client
- consequently, changes to implementation class do not require client code to be recompiled (since interface has not changed)
- thus, pimpl idiom useful for reducing compile-time dependencies (which can facilitate faster compiles) and maintaining stable class ABIs

# Pimpl and `std::experimental::propagate_const`

```
1  #include <experimental/propagate_const>
2  #include <memory>
3  #include <iostream>
4
5  class WidgetImpl {
6  public:
7      void foo() {std::cout << "WidgetImpl::foo()\n";}
8      void foo() const {std::cout << "WidgetImpl::foo() const\n";}
9  };
10
11 class Widget {
12 public:
13     void foo() {p_->foo();}
14     void foo() const {p_->foo();}
15 private:
16     std::experimental::propagate_const<std::unique_ptr<WidgetImpl>> p_;
17     /* const or non-const member functions of WidgetImpl invoked
18     as appropriate based on constness of *this; using
19     std::unique_ptr<const WidgetImpl> would cause only const member
20     functions of WidgetImpl to be invoked; using
21     std::unique_ptr<WidgetImpl> would cause only non-const member
22     functions of WidgetImpl to be invoked */
23 };
24
25 int main() {
26     Widget w;
27     const Widget cw;
28     w.foo(); // calls WidgetImpl::foo()
29     cw.foo(); // calls WidgetImpl::foo() const
30 }
```

## Section 2.4.5

### **Functors**

# Functors

- **function object** (also known as **functor**) is object that can be invoked or called as if it were ordinary function
- class that provides member function that overloads **operator ()** is called **functor class** and object of that class is **functor**
- functors more flexible than functions as functors are objects and can therefore carry arbitrary state information
- when ordinary function used, function often invoked through pointer whose value cannot be determined at compile time, which makes inlining impossible
- when functor used, function to be called is fixed and always known at compile time (namely, function-call operator for functor class)
- moreover, definition of function-call operator very likely to be visible at point of use, especially if functor created from lambda expression
- consequently, functors often more amenable to inlining
- functors are extremely useful, especially in generic programming
- as we will see later, standard library makes heavy use of functors

# Functor Example: Less Than

```
1  struct LessThan { // Functor class
2      bool operator()(double x, double y) const {
3          return x < y;
4      }
5  };
6
7  void myFunc() {
8      double a = 1.0;
9      double b = 2.0;
10     LessThan lessThan; // Functor
11     bool result = lessThan(a, b);
12     // calls LessThan::operator()(double, double)
13     // lessThan is functor, not function
14     // result == true
15 }
```



# Functor Example With State

```
1  class IsGreater { // Functor class
2  public:
3      IsGreater(int threshold) : threshold_(threshold) {}
4      bool operator()(int x) const {
5          return x > threshold_;
6      }
7  private:
8      // state information for functor
9      int threshold_; // threshold for comparison
10 };
11
12 void myFunc() {
13     IsGreater isGreater(5); // functor
14     int x = 3;
15     bool result = isGreater(x);
16     // calls IsGreater::operator()(int)
17     // result == false
18 }
```

# Ordering Relations

- often, need arises to impose some ordering on data
- for example, ordering relation needed for any sorting algorithm or ordered container (such as ordered set, multiset, map, or multimap)
- to define ordering relation, sufficient to specify either less-than (i.e., “precedes”) relation or greater-than (i.e., “follows”) relation
- typically, in C++ (such as in standard library) less-than relation is used to define all other relational operators
- $\text{less}(x, y)$  is true if  $x$  precedes  $y$  in sorted order and false otherwise
- can synthesize all other relational operators from less as follows:
  - $\text{greater}(x, y) = \text{less}(y, x)$
  - $\text{equal}(x, y) = \neg \text{less}(x, y) \wedge \neg \text{less}(y, x)$
  - $\text{notEqual}(x, y) = \text{less}(x, y) \vee \text{less}(y, x)$
  - $\text{lessEqual}(x, y) = \neg \text{less}(y, x)$
  - $\text{greaterEqual}(x, y) = \neg \text{less}(x, y)$
- note: “ $\neg$ ” denotes logical NOT, “ $\wedge$ ” denotes logical AND, and “ $\vee$ ” denotes logical OR

# Selection Sort Example

```
1  #include <algorithm>
2  #include <cassert>
3  #include <forward_list>
4  #include <functional>
5
6  // reverse digits in decimal representation of integer
7  constexpr unsigned int reverse(unsigned int x) {
8      unsigned int y = 0;
9      for (; x; x /= 10)
10         {auto d = x % 10; x -= d; y = 10 * y + d;}
11     return y;
12 }
13
14 constexpr bool rev_less(unsigned int x, unsigned int y)
15     {return reverse(x) < reverse(y);}
16
17 template <class ForwardIterator, class Compare>
18 void selection_sort(ForwardIterator first, ForwardIterator last, Compare less) {
19     for (auto i = first; i != last; ++i)
20         {std::iter_swap(i, std::min_element(i, last, less));}
21 }
22
23 int main() {
24     std::forward_list<unsigned int> values{12, 21, 123, 321, 1234, 4321};
25     selection_sort(values.begin(), values.end(), std::greater<unsigned int>());
26     assert((values == std::forward_list<unsigned int>{
27         4321, 1234, 321, 123, 21, 12}));
28     selection_sort(values.begin(), values.end(), rev_less);
29     assert((values == std::forward_list<unsigned int>{
30         21, 12, 321, 123, 4321, 1234}));
31 }
```

# Bubble Sort Example

```
1  #include <algorithm>
2  #include <cassert>
3  #include <forward_list>
4  #include <functional>
5
6  template <class ForwardIterator, class Compare>
7  void bubble_sort(ForwardIterator first, ForwardIterator last, Compare less) {
8      for (auto sorted = first; first != last; last = sorted) {
9          sorted = first;
10         for (auto cur = first, prev = first; ++cur != last; ++prev) {
11             if (less(*cur, *prev)) {
12                 std::iter_swap(cur, prev);
13                 sorted = cur;
14             }
15         }
16     }
17 }
18
19 int main() {
20     std::forward_list<int> values{7, 0, 6, 1, 5, 2, 4, 3};
21     bubble_sort(values.begin(), values.end(), std::less<int>());
22     assert((values == std::forward_list<int>{0, 1, 2, 3, 4, 5, 6, 7}));
23     bubble_sort(values.begin(), values.end(), std::greater<int>());
24     assert((values == std::forward_list<int>{7, 6, 5, 4, 3, 2, 1, 0}));
25 }
```

# Comparison Object Example

```
1  #include <cassert>
2  #include <algorithm>
3  #include <set>
4
5  template <class T> class compare {
6  public:
7      constexpr compare(bool less = true) : less_(less) {}
8      constexpr bool operator()(const T& x, const T& y) const
9          {return less_ ? (x < y) : (x > y);}
10 private:
11     bool less_;
12 };
13
14 constexpr bool even_then_odd(int x, int y)
15     {if ((x % 2) != (y % 2)) {return !(x % 2);} else {return x < y;}}
```

```
16
17 int main() {
18     constexpr int values[] = {0, 7, 6, 1, 2, 5, 3, 4};
19     std::set<int, compare<int>> s1(std::begin(values), std::end(values));
20     constexpr int d1[] = {0, 1, 2, 3, 4, 5, 6, 7};
21     assert(std::equal(s1.begin(), s1.end(), std::begin(d1)));
22     std::set<int, compare<int>> s2(std::begin(values), std::end(values),
23         compare<int>(false));
24     constexpr int d2[] = {7, 6, 5, 4, 3, 2, 1, 0};
25     assert(std::equal(s2.begin(), s2.end(), std::begin(d2)));
26     std::set<int, bool (*)(int, int)> s3(std::begin(values), std::end(values),
27         even_then_odd);
28     constexpr int d3[] = {0, 2, 4, 6, 1, 3, 5, 7};
29     assert(std::equal(s3.begin(), s3.end(), std::begin(d3)));
30 }
```

# Comparison Object Propagation

- invariant of ordered container: elements of container always sorted with respect to ordering relation defined by comparison object
- thus, state for ordered container (ignoring possible allocator) consists of:
  - 1 elements in container; and
  - 2 comparison object that determines order of those elements
- consider propagating value of one container to another (via copy or move)
- when propagating value of container, two choices possible:
  - 1 propagate comparison object
  - 2 do not propagate comparison object
- if comparison object not propagated and source and destination comparison objects differ, must re-sort elements (to be consistent with destination comparison object) to avoid violating container invariant
- if no equality/inequality operator provided by comparison-object type, must assume worst (i.e., not equal) and always re-sort
- if comparison object propagated, never any need to re-sort elements
- for efficiency, prefer solution of always propagating comparison object

# Comparison Object Propagation Example

```
1  #include <algorithm>
2  #include <cassert>
3  #include <set>
4  #include <utility>
5
6  template <class T> class compare {
7  public:
8      compare(bool less = true) : less_(less) {}
9      bool operator()(const T& x, const T& y) const
10     {return less_ ? (x < y) : (x > y);}
11     bool less() const {return less_;}
12 private:
13     bool less_;
14 };
15
16 int main() {
17     constexpr int values[] = {0, 7, 6, 1, 2, 5, 3, 4};
18     std::set<int, compare<int>> s3(values, std::end(values));
19     std::set<int, compare<int>> s1(std::move(s3)); // move construct
20     assert(s1.key_comp().less()); // comparison object was moved
21     std::set<int, compare<int>> s2(s1); // copy construct
22     assert(s2.key_comp().less()); // comparison object was copied
23     s3 = std::set<int, compare<int>>(values, std::end(values),
24         compare<int>(false));
25     assert(s1.key_comp().less() && !s3.key_comp().less());
26     s1 = std::move(s3); // move assign
27     assert(!s1.key_comp().less()); // comparison object was moved
28     assert(s2.key_comp().less() && !s1.key_comp().less());
29     s2 = s1; // copy assign
30     assert(!s2.key_comp().less()); // comparison object was copied
31 }
```

## Section 2.4.6

### **References**



- 1 Arthur O'Dwyer. Return Value Optimization: Harder Than It Looks. CppCon, Bellevue, WA, USA, Sept. 25, 2018. Available online at <https://youtu.be/hA1WNtNyNbo>.
- 2 Jon Kalb. Copy Elision. C++Now, Aspen, CO, USA, May 9, 2018. Available online at <https://youtu.be/fSB57PiXpRw>.
- 3 Jon Kalb. Copy Elision. CppCon, Bellevue, WA, USA, Sept. 23–28, 2018. Available online at [https://youtu.be/IZbL-RGr\\_mk](https://youtu.be/IZbL-RGr_mk).
- 4 Roger Orr. Nothing Is Better Than Copy Or Move. ACCU Conference, Bristol, UK, Apr. 11, 2018. Available online at <https://youtu.be/-dc5vqt2tgA>.
- 5 Scott Schurr. constexpr: Introduction. CppCon, Bellevue, WA, USA, Sept 19–25, 2015. Available online at <https://youtu.be/fZjYCQ8dzTc>.
- 6 Scott Schurr. constexpr: Applications. CppCon, Bellevue, WA, USA, Sept 19–25, 2015. Available online at <https://youtu.be/q0-9yiA0Qqc>.

## Section 2.5

# Templates

- **generic programming**: algorithms written in terms of types to be specified later (i.e., algorithms are generic in sense of being applicable to any type that meets only some very basic constraints)
- templates facilitate generic programming
- extremely important language feature
- avoids code duplication
- leads to highly efficient and customizable code
- promotes code reuse
- C++ standard library makes very heavy use of templates (actually, most of standard library consists of templates)
- many other libraries make heavy use of templates (e.g., CGAL, Boost)

## Section 2.5.1

# **Function Templates**

# Motivation for Function Templates

- consider following functions:

```
int max(int x, int y)
{return x > y ? x : y;}
```

```
double max(double x, double y)
{return x > y ? x : y;}
```

*// more similar-looking max functions...*

- each of above functions has *same general form*; that is, for some type  $T$ , we have:

```
T max(T x, T y)
{return x > y ? x : y;}
```

- would be nice if we did not have to repeatedly type, debug, test, and maintain nearly identical code
- in effect, would like code to be parameterized on type  $T$

# Function Templates

- **function template** is family of functions parameterized by one or more parameters
- each template parameter can be: non-type (e.g., integral constant), type, template, or parameter pack (in case of variadic template)
- syntax for template function has general form:

**template** <parameter\_list> *function\_declaration*

- *parameter\_list*: parameters on which template function depends
- *function\_declaration*: function declaration or definition
- type parameter designated by **class** or **typename** keyword
- template parameter designated by **template** keyword
- non-type parameter designated by its type (e.g., **bool**, **int**)
- example:

```
// declaration of function template
template <class T> T max(T x, T y);

// definition of function template
template <class T> T max(T x, T y)
{ return x > y ? x : y; }
```

# Function Templates (Continued)

- to explicitly identify particular instance of template, use syntax:

*function*<parameters>

- example: for function template declaration:

```
template <class T> T max(T x, T y);
```

max<**int**> refers to **int** max(**int**, **int**)

max<**double**> refers to **double** max(**double**, **double**)

- compiler only creates code for function template when it is instantiated (i.e., used)
- therefore, definition of function template must be visible in place where it is instantiated
- consequently, function template definitions usually appear in header file
- template code only needs to pass basic syntax checks, unless actually instantiated

# Function Template Examples

```
1  // compute minimum of two values
2  template <class T>
3  T min(T x, T y) {
4      return x < y ? x : y;
5  }
6
7  // compute square of value
8  template <typename T>
9  T sqr(T x) {
10     return x * x;
11 }
12
13 // swap two values
14 template <class T>
15 void swap(T& x, T& y) {
16     T tmp = x;
17     x = y;
18     y = tmp;
19 }
20
21 // invoke function/functor multiple times
22 template <int N = 1, typename F, typename T>
23 void invoke(F func, const T& value) {
24     for (int i = 0; i < N; ++i) {
25         func(value);
26     }
27 }
```



# Template Function Overload Resolution

- overload resolution proceeds (in order) as follows:
  - 1 look for an exact match with zero or more trivial conversions on (nontemplate) functions; if found call it
  - 2 look for function template from which function that can be called with exact match with zero or more trivial conversions can be generated; if found, call it
  - 3 try ordinary overload resolution for functions; if function found, call it; otherwise, call is error
- in each step, if more than one match found, call is ambiguous and is error
- template function only used in case of exact match, unless explicitly forced
- example:

```
template <class T>
T max(T x, T y) {return x > y ? x : y;}

void func(int i, int j, double x, double y) {
    double z = max(x, y); // calls max<double>
    int k = max(i, j); // calls max<int>
    z = max(i, x); // ERROR: no match
    z = max<double>(i, x); // calls max<double>
}
```

- **qualified name** is name that specifies scope
- example:

```
#include <iostream>

int main(int argc, char** argv) {
    for (int i = 0; i < 10; ++i) {
        std::cout << "Hello, world!" << std::endl;
    }
}
```

- in above example, names `std::cout` and `std::endl` are qualified, while names `main`, `argc`, `argv`, and `i`, are not qualified

- **dependent name** is name that depends on template parameter [\[C++17 17.6.2.1/9\]](#)
- example:

```
template <class T>
void func(const T& x) {
    int i = T::magicValue;
    // ...
}
```

- name `T::magicValue` is dependent

# Qualified Dependent Names

- to avoid any potential ambiguities, compiler will automatically assume qualified dependent name does not name type unless **typename** keyword is used
- must precede qualified dependent name that names type by **typename**
- in following example, note use of **typename** keyword:

```
1  #include <vector>
2
3  template <class T>
4  void func(const T& x) {
5      std::vector<T> v(42, x);
6      // std::vector<T>::const_iterator is
7      // qualified dependent name
8      for (typename std::vector<T>::const_iterator i =
9          v.begin(); i != v.end(); ++i) {
10         // std::vector<T>::value_type is
11         // qualified dependent name
12         typename std::vector<T>::value_type x = *i;
13         // ...
14     }
15     // ...
16 }
```

# Why typename is Needed

```
1  int x = 42;
2
3  template <class T> void func() {
4      /* The compiler must be able to check the syntactic
5      correctness of this template code without knowing the
6      type T. Without knowing the type T, however, the meaning
7      of the following line of code is ambiguous, unless the
8      compiler follows some fixed rule for resolving this
9      ambiguity. In particular, is this line of code a
10     declaration of a variable x or an expression consisting
11     of a binary operator* with operands T::foo and x? */
12     T::foo* x; // Does T::foo name a type or an object?
13 }
14
15 struct ContainsType {
16     using foo = int; // foo is type
17 };
18
19 struct ContainsValue {
20     static int foo; // foo is value
21 };
22
23 int main() {
24     // Only one of the following two lines should be valid.
25     func<ContainsValue>();
26     func<ContainsType>();
27 }
```

## Section 2.5.2

# **Class Templates**

# Motivation for Class Templates

- consider almost identical complex number classes:

```
1  class ComplexDouble {
2  public:
3      ComplexDouble(double x = 0.0, double y = 0.0) : x_(x), y_(y) {}
4      double real() const { return x_; }
5      double imag() const { return y_; }
6      // ...
7  private:
8      double x_, y_; // real and imaginary parts
9  };
10
11 class ComplexFloat {
12 public:
13     ComplexFloat(float x = 0.0f, float y = 0.0f) : x_(x), y_(y) {}
14     float real() const { return x_; }
15     float imag() const { return y_; }
16     // ...
17 private:
18     float x_, y_; // real and imaginary parts
19 };
```

- both of above classes are special cases of following class parameterized on type T:

```
1  class Complex {
2  public:
3      Complex(T x = T(0), T y = T(0)) : x_(x), y_(y) {}
4      T real() const { return x_; }
5      T imag() const { return y_; }
6      // ...
7  private:
8      T x_, y_; // real and imaginary parts
9  };
```

- again, would be nice if we did not have to repeatedly type, debug, test, and maintain nearly identical code

# Class Templates

- **class template** is family of classes parameterized on one or more parameters
- each template parameter can be: non-type (e.g., integral constant), type, template, or parameter pack (in case of variadic template)
- syntax has general form:

**template** <*parameter\_list*> *class\_declaration*

- *parameter\_list*: parameter list for class
- *class\_declaration*: class/struct declaration or definition
- example:

```
// declaration of class template
template <class T, unsigned int size>
class MyArray;

// definition of class template
template <class T, unsigned int size>
class MyArray {
    // ...
    T array_[size];
};

MyArray<double, 100> x;
```



# Class Templates (Continued)

- compiler only generates code for class template when it is instantiated (i.e., used)
- since compiler only generates code for class template when it is instantiated, definition of template must be visible at point where instantiated
- consequently, class template code usually placed in header file
- template code only needs to pass basic syntax checks, unless actually instantiated
- compile errors related to class templates can often be very long and difficult to parse (especially, when template class has parameters that are template classes which, in turn, have parameters that are template classes, and so on)

# Class Template Example

```
1 // complex number class template
2 template <class T>
3 class Complex {
4 public:
5     Complex(T x = T(0), T y = T(0)) :
6         x_(x), y_(y) {}
7     T real() const {
8         return x_;
9     }
10    T imag() const {
11        return y_;
12    }
13    // ...
14 private:
15     T x_; // real part
16     T y_; // imaginary part
17 };
18
19 Complex<int> zi;
20 Complex<double> zd;
```

# Class-Template Default Parameters

- class template parameters can have *default values*
- example:

```
template <class T = int, unsigned int size = 2>
struct MyArray {
    T data[size];
};

MyArray<> a; // MyArray<int, 2>
MyArray<double> b; // MyArray<double, 2>
MyArray<double, 10> b; // MyArray<double, 10>
```

# Qualified Dependent Names Revisited

- recall, qualified dependent name assumed not to name type, unless preceded by **typename** keyword
- in following example, note use of **typename** keyword:

```
1  #include <vector>
2
3  template <class T> class Vector {
4  public:
5      using Coordinate = typename T::Coordinate;
6      using Distance = typename T::Distance;
7      Vector(const std::vector<Coordinate>& coords) :
8          coords_(coords) {}
9      Distance squaredLength() const {
10         Distance d = Distance(0);
11         for (typename
12             std::vector<Coordinate>::const_iterator i =
13             coords_.begin(); i != coords_.end(); ++i) {
14             typename std::vector<Coordinate>::value_type
15                 x = *i;
16                 d += x * x;
17         }
18         return d;
19     }
20     // ...
21 private:
22     std::vector<Coordinate> coords_;
23 };
```

# Why template is Needed

```
1  template<bool> struct Widget;  
2  
3  template<bool B> struct Gadget {  
4      static int g() {  
5          /* The compiler must be able to check the syntactic  
6          correctness of this template code without knowing the  
7          value of B. Without knowing the value of B, however,  
8          the meaning of the following line of code is ambiguous,  
9          unless the compiler follows some fixed rule for  
10         resolving this ambiguity. In particular, is this line  
11         of code using a data member called f and evaluating  
12         (f < 0 > 42) or is it calling a template member  
13         function called f with the argument 42? */  
14         return Widget<B>::f<0>(42);  
15     }  
16 };  
17  
18 template<bool B> struct Widget {  
19     template<int I> static int f(int i) {return i + I;}  
20 };  
21  
22 template<> struct Widget<false> {inline static int f = 42;};  
23  
24 int main() {  
25     // Only one of the following two lines should be valid.  
26     Gadget<true>::g();  
27     Gadget<false>::g();  
28 }
```

# Template Template Parameter Example

```
1  #include <vector>
2  #include <list>
3  #include <deque>
4  #include <memory>
5
6  template <template <class, class> class Container, class Value>
7  class Stack {
8  public:
9      // ...
10 private:
11     Container<Value, std::allocator<Value>> data_;
12 };
13
14 int main() {
15     Stack<std::vector, int> s1;
16     Stack<std::list, int> s2;
17     Stack<std::deque, int> s3;
18 }
```

# Class Template Parameter Deduction

- template parameters for class template can be deduced based on arguments passed to constructor

- example:

```
std::tuple t(42, 'A');  
    // OK: deduced as tuple<int, char>
```

- deduction only performed if no template arguments provided

- example:

```
std::tuple<int> t(1, 2);  
    // ERROR: missing template parameter, as  
    // no template parameter deduction takes place
```

# Class Template Parameter Deduction Example

```
1  #include <vector>
2  #include <tuple>
3  #include <set>
4  #include <string>
5
6  using namespace std::string_literals;
7
8  auto get_tuple() {
9      return std::tuple("Zaphod"s, 42);
10     // deduces tuple<std::string, int>
11 }
12
13 int main() {
14     std::vector v{1, 2, 3};
15     // deduces vector<int>
16     std::tuple t(true, 'A', 42);
17     // deduces tuple<bool, char, int>
18     std::pair p(42, "Hello"s);
19     // deduces pair<int, std::string>
20     std::set s{0.5, 0.25};
21     // deduces set<double>
22     //auto ptr = new std::tuple(true, 42);
23     // should deduce tuple<bool, int>?
24     // fails to compile with GCC 7.1.0
25 }
```



- can provide additional rules to be used to determine how class template parameters should be deduced when not provided
- such rules called deduction guides
- deduction guide itself can be either template or non-template
- deduction guides must be introduced in same scope as class template
- example:

```
// class definition  
template <class T> smart_ptr { /* ... */ };  
// deduction guide  
template <class T>  
smart_ptr(T*) -> smart_ptr<T>;
```

- example:

```
/// class definition  
template <class T> name { /* ... */ };  
// deduction guide  
name(const char*) -> name<std::string>;
```

# Template Deduction Guide Example

```
1  #include <string>
2  #include <type_traits>
3
4  using namespace std::string_literals;
5
6  template <class T>
7  class Name {
8  public:
9      Name(T first, T last) : first_(first), last_(last) {}
10     // ...
11 private:
12     T first_;
13     T last_;
14 };
15
16 // deduction guide
17 Name(const char*, const char*) -> Name<std::string>;
18
19 int main() {
20     Name n("Zaphod", "Beeblebrox");
21     // deduces Name<std::string> via deduction guide
22     static_assert(std::is_same_v<decltype(n), Name<std::string>>);
23     Name n2("Jane"s, "Doe"s);
24     // deduces Name<std::string> (without deduction guide)
25     static_assert(std::is_same_v<decltype(n2), Name<std::string>>);
26 }
```

# Auto Non-Type Template Parameters

- can use **auto** keyword for non-type template parameter
- in such case, type of non-type template parameter will be deduced

- example:

```
template <auto v>
struct constant {
    static constexpr decltype(v) value = v;
};
using forty_two_type = constant<42>;
// template parameter v deduced to have type int
```

- non-type template parameter type deduction probably most useful for template metaprogramming

# Example Without Auto Non-Type Template Parameter

```
1  #include <cstdlib>
2  #include <iostream>
3
4  template<class T, T v>
5  struct integral_constant {
6      using value_type = T;
7      static constexpr value_type value = v;
8      using type = integral_constant;
9      constexpr operator value_type() const noexcept
10     {return value;}
11     constexpr value_type operator() () const noexcept
12     {return value;}
13 };
14
15 using forty_two_type = integral_constant<int, 42>;
16
17 int main() {
18     constexpr forty_two_type x;
19     constexpr auto v = x.value;
20     std::cout << v << '\n';
21 }
```

# Example With Auto Non-Type Template Parameter

```
1  #include <cstdlib>
2  #include <iostream>
3
4  template<auto v>
5  struct integral_constant {
6      using value_type = decltype(v);
7      static constexpr value_type value = v;
8      using type = integral_constant;
9      constexpr operator value_type() const noexcept
10     {return value;}
11     constexpr value_type operator() () const noexcept
12     {return value;}
13 };
14
15 using forty_two_type = integral_constant<42>;
16
17 int main() {
18     constexpr forty_two_type x;
19     constexpr auto v = x.value;
20     std::cout << v << '\n';
21 }
```

## Section 2.5.3

# **Variable Templates**

# Variable Templates

- **variable template** is family of variables parameterized on one or more parameters
- each template parameter can be: non-type (e.g., integral constant), type, template, or parameter pack (in case of variadic templates)
- although less frequently used than function and class templates, variable templates quite useful in some situations
- syntax has general form:

**template** <*parameter\_list*> *variable\_declaration*

- *parameter\_list*: parameter list for variable template
- *variable\_declaration*: variable declaration or definition
- example:

```
template <class T>
T meaning_of_life = T(42);

int x = meaning_of_life<int>;
```

# Variable Template Example: pi

```
1  #include <limits>
2  #include <complex>
3  #include <iostream>
4
5  template <typename T>
6  constexpr T pi =
7      T(3.14159265358979323846264338327950288419716939937510L);
8
9  int main() {
10     std::cout.precision(
11         std::numeric_limits<long double>::max_digits10);
12     std::cout
13         << pi<int> << '\n'
14         << pi<float> << '\n'
15         << pi<double> << '\n'
16         << pi<long double> << '\n'
17         << pi<std::complex<float>> << '\n'
18         << pi<std::complex<double>> << '\n'
19         << pi<std::complex<long double>> << '\n';
20 }
```



## Section 2.5.4

### **Alias Templates**

# Alias Templates

- **alias template** is family of types parameterized on one or more parameters
- each template parameter can be: non-type (e.g., integral constant), type, template, or parameter pack (in case of variadic templates)
- syntax has general form:

**template** <*parameter\_list*> *alias\_declaration*

- *parameter\_list*: parameter list for class
- *alias\_declaration*: alias declaration (i.e., with **using**)
- example:

```
template <class Value,  
         class Alloc = std::allocator<Value>>  
using GreaterMultiSet =  
    std::multiset<Value, std::greater<Value>, Alloc>;  
  
GreaterMultiSet<int> x{4, 1, 3, 2};
```

# Alias Template Example

```
1  #include <iostream>
2  #include <set>
3
4  // alias template for set that employs std::greater for
5  // comparison
6  template <typename Value,
7          typename Alloc = std::allocator<Value>>
8  using GreaterSet = std::set<Value,
9          std::greater<Value>, Alloc>;
10
11 int main() {
12     std::set x{1, 4, 3, 2};
13     GreaterSet<int> y{1, 4, 3, 2};
14     for (auto i : x) {
15         std::cout << i << '\n';
16     }
17     std::cout << '\n';
18     for (auto i : y) {
19         std::cout << i << '\n';
20     }
21 }
```

## Section 2.5.5

### **Variadic Templates**

# Variadic Templates

- language provides ability to specify template that can take variable number of arguments
- template that can take variable number of arguments called **variadic template**
- alias templates, class templates, function templates, and variable templates may be variadic
- variable number of arguments specified by using what is called parameter pack
- parameter pack is parameter that accepts (i.e., is placeholder for) zero or more arguments (of same kind)
- parameter pack used in parameter list of template to allow to variable number of template parameters
- ellipsis (i.e., "...") is used in various contexts relating to parameter packs
- ellipsis after designator for kind of template argument in template parameter list designates argument is parameter pack
- ellipsis after parameter pack parameter expands parameter pack in context-sensitive manner

# Parameter Packs

- syntax for *non-type* template parameter pack named *Args* and containing elements of type *type* (e.g., **bool**, **int**, **unsigned int**):

*type*... *Args*

- example:

```
template <int... Is> /* ... */
```

*Is* is (non-type) template parameter pack that corresponds to zero or more (compile-time constant) values of type **int**

- syntax for *type* template parameter pack named *Args*:

**typename**... *Args*

or equivalently

**class**... *Args*

- examples:

```
template <typename... Ts> /* ... */
```

```
template <class... Ts> /* ... */
```

*Ts* is (type) template parameter pack that corresponds to zero or more types

# Parameter Packs (Continued 1)

- syntax for *template* template parameter pack named *Args*:

**template** <*parameter\_list*> **typename**... *Args*

or equivalently

**template** <*parameter\_list*> **class**... *Args*

- example:

```
template <template <class T> class... Ts>  
    /* ... */
```

Ts is (template) template parameter pack that corresponds to zero or more templates

- syntax for *function* parameter pack named *args* whose elements have types corresponding to elements of type template parameter pack *Args*:

*Args*... *args*

- example:

```
template <class... Ts> void func(Ts... args);
```

*args* is (function) parameter pack that corresponds to zero or more function parameters whose types correspond to elements of type parameter pack Ts

## Parameter Packs (Continued 2)

- in context where template arguments *cannot be deduced* (e.g., primary class templates), *only last* template parameter can be parameter pack
- in context where template arguments *can be deduced* (e.g., function templates and class template partial specializations), template parameter pack *need not be last* template parameter
- example:

```
1  template <class U, class... Ts> class C1 { /* ... */ };
2    // OK: Ts is last template parameter
3
4  template <class... Ts, class U> class C2 { /* ... */ };
5    // ERROR: Ts not last and U not deduced
6
7  template <class... Ts, class U> void f1(Ts... ts)
8    { /* ... */ } // NOT OK: Ts not last and U not deduced
9
10 template <class... Ts, class U> void f2(Ts... ts, U u)
11   { /* ... */ } // OK: Ts not last but U is deduced
12
13 int main() {
14     f1<int, int, bool>(1, 2, true);
15     // ERROR: no matching function call
16     f2<int, int>(1, 2, true); // OK
17     f2(1, 2, true); // ERROR: one argument expected
18 }
```



# Parameter Pack Expansion

- **parameter pack expansion**: expands pack into its constituent elements
- syntax for parameter pack expansion of expression *pattern*, which must contain parameter pack:

*pattern...*

- example:

```
1  template <class... Ts> void f(Ts... t) { /* ... */ }
2
3  template <class... Us> void g(Us... u) {
4      f(u...);
5      // u... is pack expansion
6      // when g is called by main,
7      // u... expands to 1, 2.0, 3.0f
8  }
9
10 int main() {
11     g(1, 2.0, 3.0f);
12 }
```

# Variadic Template Examples

```
1  #include <tuple>
2
3  // variadic alias template
4  template <class... T>
5  using My_tuple = std::tuple<bool, T...>;
6
7  // variadic class template
8  template <int... Values>
9  class Integer_sequence {
10     // ...
11 };
12
13 // variadic function template
14 template <class... Ts>
15 void print(const Ts&... values) {
16     // ...
17 }
18
19 // variadic variable template
20 template <typename T, T... Values>
21 constexpr T array[] = {Values...};
22
23 int main() {
24     Integer_sequence<1, 3, 4, 2> x;
25     auto a = array<int, 1, 2, 4, 8>;
26     My_tuple<int, double> t(true, 42, 42.0);
27     print(1'000'000, 1, 43.2, "Hello");
28 }
```

- parameter pack expansion allowed in following contexts: C++17 17.5.3/4
  - inside parentheses of function call operator
  - in template argument list
  - in function parameter list
  - in template parameter list
  - base class specifiers in class declaration
  - member initializer lists
  - braced initializer lists
  - lambda captures
  - fold expressions
  - in using declarations

# The `sizeof...` Operator

- `sizeof...` operator yields number of elements in parameter pack
- example:

```
template <int... Values>
constexpr int num_parms = sizeof...(Values);

static_assert(num_parms<1, 2, 3> == 3);
static_assert(num_parms<> == 0);
```

- example:

```
#include <cassert>

template <typename... Ts>
int number_of_arguments(const Ts&... args) {
    return sizeof...(args);
}

int main() {
    assert(number_of_arguments(1, 2, 3) == 3);
    assert(number_of_arguments() == 0);
}
```

# Variadic Function Template: sum

```
1  #include <iostream>
2  #include <string>
3
4  using namespace std::string_literals;
5
6  template <class T>
7  auto sum(T x) {
8      return x;
9  }
10
11 template <class T, class... Args>
12 auto sum(T x, Args... args) {
13     return x + sum(args...);
14 }
15
16 int main() {
17     auto x = sum(42.5, -1.0, 0.5f);
18     auto y = sum("The "s, "answer "s, "is "s);
19     std::cout << y << x << ".\n";
20     // sum(); // ERROR: no matching function call
21 }
22
23 /* Output:
24 The answer is 42.
25 */
```

# Variadic Function Template: maximum

```
1  #include <type_traits>
2  #include <string>
3  #include <cassert>
4
5  using namespace std::string_literals;
6
7  template <typename T>
8  T maximum(const T& a) {return a;}
9
10 template <typename T1, typename T2>
11 typename std::common_type_t<const T1&, const T2&>
12 maximum(const T1 &a, const T2 &b) {
13     return a > b ? a : b;
14 }
15
16 template <typename T1, typename T2, typename... Args>
17 typename std::common_type_t<const T1&, const T2&,
18     const Args&...>
19 maximum(const T1& a, const T2& b, const Args&... args) {
20     return maximum(maximum(a, b), args...);
21 }
22
23 int main() {
24     assert(maximum(1) == 1);
25     assert(maximum(1, 2, 3, 4, -1.4) == 4);
26     assert(maximum(-1'000'000L, -42L, 10, 42.42) == 42.42);
27     assert(maximum("apple"s, "zebra"s, "c++"s) == "zebra"s);
28 }
```

# Variadic Function Template With Template Template

Parameter: `print_container`

```
1  #include <iostream>
2  #include <vector>
3  #include <string>
4  #include <set>
5
6  template <template <class, class...>
7      class ContainerType, class ValueType, class... Args>
8  bool print_container(const ContainerType<ValueType, Args...>&
9      c) {
10     for (auto i = c.begin(); i != c.end(); ) {
11         std::cout << *i;
12         if (++i != c.end()) {std::cout << ' ';}
13     }
14     std::cout << '\n';
15     return bool(std::cout);
16 }
17
18 int main() {
19     using namespace std::string_literals;
20     std::vector vi{1, 2, 3, 4, 5};
21     std::set si{5, 4, 3, 2, 1};
22     std::set ss{"world"s, "hello"s};
23     print_container(vi);
24     print_container(si);
25     print_container(ss);
26 }
```

# Variadic Class Template: Integer\_sequence

```
1  #include <iostream>
2  #include <cstdlib>
3
4  template <class T, T... Values>
5  class Integer_sequence {
6  public:
7      using value_type = T;
8      using const_iterator = const T*;
9      constexpr std::size_t size() const
10         {return sizeof...(Values);}
11      constexpr T operator[](int i) const {return values_[i];}
12      constexpr const_iterator begin() const
13         {return &values_[0];}
14      constexpr const_iterator end() const
15         {return &values_[size()];}
16 private:
17     static constexpr T values_[sizeof...(Values)] =
18         {Values...};
19 };
20
21 template <class T, T... Values>
22 constexpr T
23 Integer_sequence<T, Values...>::values_[sizeof...(Values)];
24
25 int main() {
26     Integer_sequence<std::size_t, 1, 2, 4, 8> seq;
27     std::cout << seq.size() << '\n' << seq[0] << '\n';
28     for (auto i : seq) {std::cout << i << '\n';}
29 }
```



# Variadic Variable Template: `int_array`

```
1  #include <iostream>
2
3  template <int... Args>
4  constexpr int int_array[] = {Args...};
5
6  int main() {
7      for (auto i : int_array<1,2,4,8>) {
8          std::cout << i << '\n';
9      }
10 }
11
12 /* Output:
13 1
14 2
15 4
16 8
17 */
```

# Variadic Alias Template: My\_tuple

```
1  #include <iostream>
2  #include <string>
3  #include <tuple>
4
5  template <class... Ts>
6  using My_tuple = std::tuple<bool, Ts...>;
7
8  int main() {
9      My_tuple<int, std::string> t(true, 42,
10         "meaning of life");
11      std::cout << std::get<0>(t) << ' '
12         << std::get<1>(t) << ' '
13         << std::get<2>(t) << '\n';
14  }
15
16  /* Output:
17  1 42 meaning of life
18  */
```

# Fold Expressions

- may want to apply binary operator (such as  $+$ ) across all elements in parameter pack
- fold expression reduces (i.e., folds) parameter pack over binary operator
- $op$ : binary operator
- $E$ : expression that contains unexpanded parameter pack
- $I$ : expression that does not contain unexpanded parameter pack

Fold	Syntax	Expansion
unary left	$(\dots op E)$	$((E_1 op E_2) op \dots) op E_N$
unary right	$(E op \dots)$	$E_1 op (\dots op (E_{N-1} op E_N))$
binary left	$(I op \dots op E)$	$(( (I op E_1) op E_2 ) op \dots) op E_N$
binary right	$(E op \dots op I)$	$E_1 op (\dots op (E_{N-1} op (E_N op I)))$

[\[C++17 17.5.3/9\]](#)

- unary fold of empty parameter pack: [\[C++17 17.5.3/9\]](#)

Operator	Value for Empty Parameter Pack
<code>&amp;&amp;</code>	<b>true</b>
<code>  </code>	<b>false</b>
<code>,</code>	<b>void()</b>

# Sum Example Without Fold Expression

```
1  #include <iostream>
2  #include <string>
3
4  using namespace std::string_literals;
5
6  template <class T>
7  auto sum(T x) {
8      return x;
9  }
10
11 template <class T, class... Args>
12 auto sum(T x, Args... args) {
13     return x + sum(args...);
14 }
15
16 int main() {
17     auto x = sum(42.5, -1.0, 0.5f);
18     auto y = sum("The "s, "answer "s, "is "s);
19     std::cout << y << x << ".\n";
20     // sum(); // ERROR: no matching function call
21 }
22
23 /* Output:
24 The answer is 42.
25 */
```

# Sum Example With Fold Expression

```
1  #include <iostream>
2  #include <string>
3
4  using namespace std::string_literals;
5
6  template <class T, class... Args>
7  auto sum(T x, Args... args) {
8      return x + (... + args);
9  }
10
11 int main() {
12     auto x = sum(42.5, -1.0, 0.5f);
13     auto y = sum("The ", "answer ", "is ");
14     std::cout << y << x << ".\n";
15     // sum(); // ERROR: no matching function call
16 }
17
18 /* Output:
19 The answer is 42.
20 */
```

# Print Example Without Fold Expression

```
1  #include <iostream>
2  #include <string>
3
4  using namespace std::string_literals;
5
6  std::ostream& print() {return std::cout;}
7
8  template <class T>
9  std::ostream& print(const T& value) {
10     return std::cout << value;
11 }
12
13 template <class T, class... Args>
14 std::ostream& print(const T& value, const Args&... args) {
15     if (!(std::cout << value)) {
16         return std::cout;
17     }
18     return print(args...);
19 }
20
21 int main() {
22     print("The "s, "answer "s, "is "s, 42, ".\n"s);
23     print(); // OK: no-op
24 }
25
26 /* Output:
27 The answer is 42.
28 */
```

# Print Example With Fold Expression

```
1  #include <iostream>
2  #include <string>
3
4  using namespace std::string_literals;
5
6  template <class... Args>
7  std::ostream& print(const Args&... args) {
8      return (std::cout << ... << args);
9  }
10
11 int main() {
12     print("The "s, "answer "s, "is "s, 42, ".\n"s);
13     print(); // OK: no-op
14 }
15
16 /* Output:
17 The answer is 42.
18 */
```

# Fold Expression Example: All/Any/One/Even

```
1  #include <cassert>
2
3  template <class... Args>
4  bool all(Args... args)
5      {return (... && args);}
6
7  template <class... Args>
8  bool any(Args... args)
9      {return (... || args);}
10
11 template <class... Args>
12 bool one(Args... args)
13     {return (0 + ... + args) == 1;}
14
15 template <class... Args>
16 bool even(Args... args)
17     {return (1 + ... + args) % 2;}
18
19 int main() {
20     assert(all(false, true, true) == false);
21     assert(all(true, true, true) == true);
22     assert(any(false, false, true) == true);
23     assert(any(false, false, false) == false);
24     assert(one(true, false, false) == true);
25     assert(one(true, true, false) == false);
26     assert(even(true, true, false) == true);
27     assert(even(true, false, false) == false);
28     assert(even() == true && one() == false);
29 }
```



# Constexpr-Friendly Heterogeneous List Example

```
1  #include <iostream>
2  #include <tuple>
3
4  // heterogeneous list of constant values
5  template <auto... vs> class value_list {
6  public:
7      constexpr value_list() : v_(vs...) {}
8      template <int n> constexpr auto get() const
9          {return std::get<n>(v_);}
10     constexpr int size() const {return sizeof...(vs);}
11 private:
12     std::tuple<decltype(vs) ...> v_;
13 };
14
15 int main() {
16     constexpr value_list<42, true, 'A'> v;
17     constexpr auto n = v.size();
18     constexpr auto a = v.get<0>();
19     constexpr auto b = v.get<1>();
20     constexpr auto c = v.get<2>();
21     std::cout << n << ' ' << a << ' ' << b << ' ' << c << '\n';
22 }
```

# Constexpr-Friendly Homogeneous List Example

```
1  #include <iostream>
2  #include <tuple>
3
4  // homogeneous list of constant values
5  template <auto v1, decltype(v1)... vs> class value_list {
6  public:
7      constexpr value_list() : v_(v1, vs...) {}
8      template <int n> constexpr auto get() const
9          {return std::get<n>(v_);}
10     constexpr int size() const {return 1 + sizeof...(vs);}
11 private:
12     std::tuple<decltype(v1), decltype(vs)...int main() {
16     constexpr value_list<1, 2, 3> v;
17     constexpr auto n = v.size();
18     constexpr auto a = v.get<0>();
19     constexpr auto b = v.get<1>();
20     constexpr auto c = v.get<2>();
21     std::cout << n << ' ' << a << ' ' << b << ' ' << c << '\n';
22 }
```

## Section 2.5.6

# **Template Specialization**

# Template Specialization

- sometimes can be desirable to provide customized version of template for certain choices of template parameters
- customized version of templates can be specified through language feature known as **template specialization**
- two kinds of specialization: explicit and partial
- **explicit specialization** (less formally known as full specialization): customized version of template where all template parameters are fixed
- **partial specialization**: customized version of template where only some of template parameters are fixed
- class templates, function templates, and variable templates can all be specialized
- alias templates cannot be specialized
- class templates and variable templates can be partially or explicitly specialized
- function templates can only be explicitly specialized (not partially)

- syntax for explicit specialization:

**template** <> *declaration*

- *declaration*: declaration of templated entity (e.g., function, class, variable)
- example:

```
// unspecialized template
template <class T, class U>
    void func(T x, U y) { /* ... */ }

// explicit specialization of template
// (for when template parameters are bool, bool)
template <>
    void func<bool, bool>(bool x, bool y) { /* ... */ }
```

# Partial Specialization

- syntax for partial specialization of class template:

```
template <parameter_list> class_key  
    class_name <argument_list> declaration
```

- syntax for partial specialization of variable template:

```
template <parameter_list> type_name  
    variable_name <argument_list> declaration
```

- *class\_key*: class or struct keyword (for class template)
- *class\_name*: class being specialized (for class template)
- *type\_name*: type of variable (for variable template)
- *variable\_name*: variable being specialized (for variable template)
- *argument\_list*: template argument list
- *declaration*: declaration of templated entity (e.g., class, variable)
- example:

```
// unspecialized template  
template <class T, int N> class Widget { /* ... */ };  
  
// partial specialization of template  
// (for when first template parameter is bool)  
template <int N> class Widget<bool, N> { /* ... */ };
```

# Explicitly-Specialized Function Template: printPointee

```
1  #include <iostream>
2
3  // unspecialized version
4  template <class T>
5  typename std::ostream& printPointee(
6      typename std::ostream& out, const T* p)
7      {return out << *p << '\n';}
8
9  // specialization
10 template <>
11 typename std::ostream& printPointee<void>(
12     typename std::ostream& out, const void* p)
13     {return out << *static_cast<const char*>(p) << '\n';}
14
15 int main() {
16     int i = 42;
17     const int* ip = &i;
18     char c = 'A';
19     const void* vp = &c;
20     printPointee(std::cout, ip);
21     printPointee(std::cout, vp);
22 }
23
24 /* Output:
25 42
26 A
27 */
```

# Explicitly-Specialized Class Template: `is_void`

```
1  template <class T>
2  struct is_void
3      {static constexpr bool value = false;};
4
5  template <>
6  struct is_void<void>
7      {static constexpr bool value = true;};
8
9  template <>
10 struct is_void<const void>
11     {static constexpr bool value = true;};
12
13 template <>
14 struct is_void<volatile void>
15     {static constexpr bool value = true;};
16
17 template <>
18 struct is_void<const volatile void>
19     {static constexpr bool value = true;};
20
21 static_assert(is_void<int>::value == false);
22 static_assert(is_void<double*>::value == false);
23 static_assert(is_void<void>::value == true);
24 static_assert(is_void<const void>::value == true);
25 static_assert(is_void<volatile void>::value == true);
26 static_assert(is_void<const volatile void>::value == true);
27
28 int main() {}
```



# Partially-Specialized Class Template

```
1  #include <iostream>
2
3  // unspecialized version
4  template <typename T, typename V>
5  struct Widget {
6      Widget() {std::cout << "unspecialized\n";}
7  };
8
9  // partial specialization
10 template <typename T>
11 struct Widget<int, T> {
12     Widget() {std::cout << "partial\n";}
13 };
14
15 // explicit specialization
16 template <>
17 struct Widget<int, int> {
18     Widget() {std::cout << "explicit\n";}
19 };
20
21 int main() {
22     Widget<double, int> w1; // unspecialized version
23     Widget<int, double> w2; // partial specialization
24     Widget<int, int> w3; // explicit specialization
25 }
```

# Partially-Specialized Class Template: `std::vector`

- `std::vector` class employs specialization
- consider vector of elements of type `T`
- most natural way to store elements is as array of `T`
- if `T` is `bool`, such an approach makes very inefficient use of memory, since each `bool` object requires one byte of storage
- if `T` is `bool`, would be much more memory-efficient to use array of, say, `unsigned char` and pack multiple `bool` objects in each byte
- `std::vector` accomplishes this by providing (partial) specialization for case that `T` is `bool`
- declaration of base template for `std::vector` and its partial specialization for case when `T` is `bool` are as follows:

```
template <class T, class Alloc = allocator<T>>  
class vector; // unspecialized version
```

```
template <class Alloc>  
class vector<bool, Alloc>; // partial specialization
```

# Explicitly-Specialized Variable Template: `is_void_v`

```
1  template <class T>
2  constexpr bool is_void_v = false;
3
4  template <>
5  constexpr bool is_void_v<void> = true;
6
7  template <>
8  constexpr bool is_void_v<const void> = true;
9
10 template <>
11 constexpr bool is_void_v<volatile void> = true;
12
13 template <>
14 constexpr bool is_void_v<const volatile void> = true;
15
16 static_assert(is_void_v<int> == false);
17 static_assert(is_void_v<double*> == false);
18 static_assert(is_void_v<void> == true);
19 static_assert(is_void_v<const void> == true);
20 static_assert(is_void_v<volatile void> == true);
21 static_assert(is_void_v<const volatile void> == true);
22
23 int main() {}
```

# Explicitly-Specialized Variable Template: factorial

```
1  template <unsigned long long N>
2  constexpr unsigned long long
3      factorial = N * factorial<N - 1>;
4
5  template <>
6  constexpr unsigned long long
7      factorial<0> = 1;
8
9  int main() {
10     static_assert(factorial<5> == 120,
11         "factorial<5> failed");
12     static_assert(factorial<12> == 479'001'600,
13         "factorial<12> failed");
14 }
```

# Partially-Specialized Variable Template: quotient

```
1  #include <limits>
2
3  // unspecialized version
4  template <int X, int Y>
5  constexpr int quotient = X / Y;
6
7  // partial specialization (which prevents division by zero)
8  template <int X>
9  constexpr int quotient<X, 0> = (X < 0) ?
10     std::numeric_limits<int>::min() : std::numeric_limits<int>::max();
11
12 static_assert(quotient<4, 2> == 2);
13 static_assert(quotient<5, 3> == 1);
14 static_assert(quotient<4, 0> == std::numeric_limits<int>::max());
15 static_assert(quotient<-4, 0> == std::numeric_limits<int>::min());
16
17 int main() {}
```

## Section 2.5.7

### **Miscellany**

# Overload Resolution and Substitution Failure

- when creating candidate set (of functions) for overload resolution, some or all candidates of that set may be result of instantiated templates with template arguments substituted for corresponding template parameters
- process of substituting template arguments for corresponding template parameters can lead to invalid code
- if certain types of invalid code result from substitution in any of following, **substitution failure** said to occur:
  - all types used in function type (i.e., return type and types of all parameters)
  - all types used in template parameter declarations
  - all expressions used in function type
  - all expressions used in template parameter declaration
- substitution failure *not treated as error*
- instead, substitution failure simply causes overload to be *removed from candidate set*
- this behavior often referred to by term “*substitution failure is not an error (SFINAE)*”
- SFINAE behavior often exploited in template metaprogramming

# Some Kinds of Substitution Failures

- attempting to instantiate pack expansion containing multiple parameter packs of differing lengths
- attempting to create array with element type that is **void**, function type, reference type, or abstract class type
- attempting to create array with size that is zero or negative
- attempting to use type that is not class or enumeration type in qualified name
- attempting to use type in nested name specifier of qualified ID, when type does not contain specified member, or
  - specified member is not type where type is required
  - specified member is not template where template is required
  - specified member is not non-type where non-type is required
- attempting to create pointer to reference type
- attempting to create reference to **void**



## Some Kinds of Substitution Failures (Continued)

- attempting to create pointer to member of  $T$  when  $T$  is not class type
- attempting to give invalid type to non-type template parameter
- attempting to perform invalid conversion in either template argument expression, or expression used in function declaration
- attempting to create function type in which parameter has type of **void**, or in which return type is function type or array type
- attempting to create function type in which parameter type or return type is abstract class

# SFINAE Example: Truncate

```
1  class Real {
2  public:
3      using rounded_type = long long;
4      rounded_type truncate() const {
5          rounded_type result;
6          // ...
7          return result;
8      }
9      // ...
10 };
11
12 // function 1
13 template <class T>
14 typename T::rounded_type truncate(const T& x) {return x.truncate();}
15 // NOTE: example would not compile if return type specified as auto
16
17 // function 2
18 int truncate(double x) {return x;}
19
20 int main() {
21     Real r;
22     float f = 3.14f;
23     auto rounded_r = truncate(r);
24     // calls function 1 (only trivial conversions)
25     auto rounded_f = truncate(f);
26     // function 2 requires nontrivial conversions
27     // function 1 would only require trivial conversions but
28     // substitution failure occurs
29     // calls function 2 (with conversions)
30 }
```

[see [overload resolution](#)]

# SFINAE Example: Truncate Revisited

```
1  class Real {
2  public:
3      using rounded_type = long long;
4      rounded_type truncate() const {
5          rounded_type result;
6          // ...
7          return result;
8      }
9      // ...
10 };
11
12 // function 1
13 template <class T, class = typename T::rounded_type>
14 auto truncate(const T& x) {return x.truncate();}
15
16 // function 2
17 int truncate(double x) {return x;}
18
19 int main() {
20     Real r;
21     float f = 3.14f;
22     auto rounded_r = truncate(r);
23     // calls function 1 (only trivial conversions)
24     auto rounded_f = truncate(f);
25     // function 2 requires nontrivial conversions
26     // function 1 would only require trivial conversions but
27     // substitution failure occurs
28     // calls function 2 (with conversions)
29 }
```

## std::enable\_if and std::enable\_if\_t

- to make SFINAE more convenient to exploit, class template `std::enable_if` and alias template `std::enable_if_t` are provided
- declaration of class template `enable_if`:

```
template <bool B, class T = void>
struct enable_if;
```
- if B is **true**, class has member type `type` defined as T; otherwise, class has no `type` member

- possible implementation of `enable_if`:

```
1  template <bool B, class T = void>
2  struct enable_if {};
3
4  template <class T>
5  struct enable_if<true, T> {
6      using type = T;
7  };
```

- declaration of alias template `enable_if_t`:

```
template <bool B, class T = void>
using enable_if_t = typename enable_if<B, T>::type;
```

- if `enable_if_t` is used with its first parameter as **false**, substitution failure will result

# SFINAE Example: Modulo

```
1  #include <type_traits>
2  #include <cassert>
3  #include <iostream>
4
5  // ISO-Pascal modulo operator for signed integral types
6  template <class T> inline
7  std::enable_if_t<std::is_integral_v<T> && std::is_signed_v<T>, T>
8  mod(T x, T y) {
9      assert(y > 0);
10     if (x < 0) {x += (((-x) / y) + 1) * y;}
11     return x % y;
12 }
13
14 // ISO-Pascal modulo operator for unsigned integral types
15 template <class T> inline
16 std::enable_if_t<std::is_integral_v<T> && std::is_unsigned_v<T>, T>
17 mod(T x, T y)
18     {return x % y;}
19
20 int main() {
21     auto si = mod(-4, 3); // uses signed version
22     auto ui = mod(5u, 3u); // uses unsigned version
23     auto slli = mod(-5ll, 3ll); // uses signed version
24     auto ulli = mod(4ull, 3ull); // uses unsigned version
25     // auto f = mod(3.0, 4.0);
26     // ERROR: no matching function call
27     std::cout << si << ' ' << ui << ' ' << slli << ' ' << ulli << '\n';
28 }
```

# Detection Idiom Example

```
1  #include <iostream>
2  #include <experimental/type_traits>
3
4  class Widget {
5  public:
6      void foo() const {}
7      // ...
8  };
9
10 class Gadget {
11 public:
12     void foo() {}
13     // ...
14 };
15
16 // helper template for testing if class has member function called
17 // foo that can be invoked on const object with no arguments.
18 template <class T>
19 using has_usable_foo_t = decltype(std::declval<const T&().foo());
20
21 int main() {
22     std::cout
23         << "Widget "
24         << std::experimental::is_detected_v<has_usable_foo_t, Widget>
25         << '\n'
26         << "Gadget "
27         << std::experimental::is_detected_v<has_usable_foo_t, Gadget>
28         << '\n';
29 }
```

## Section 2.5.8

### **References**

- 1 D. Vandevor and N. M. Josuttis. *C++ Templates: The Complete Guide*. Addison Wesley, 2002.
- 2 P. Sommerlad. Variadic and variable templates. *Overload*, 126:14–17, Apr. 2015.  
Available online at <http://accu.org/index.php/journals/2087>.
- 3 A. Sutton. Introducing concepts. *Overload*, 129:4–8, Oct. 2015.  
Available online at <http://accu.org/index.php/journals/2157>.
- 4 A. Sutton. Defining concepts. *Overload*, 131:4–8, Feb. 2016.  
Available online at <http://accu.org/index.php/journals/2198>.



- 1 Peter Sommerlad. Variadic Templates in C++11/C++14: An Introduction. CppCon, Bellevue, WA, USA, Sept. 21, 2015. Available online at <https://youtu.be/R1G3P5SRXCw>.
- 2 Arthur O'Dwyer. Template Normal Programming. CppCon, Bellevue, WA, USA, Sept. 19, 2016. Available online at <https://youtu.be/vwrXHznaYLA> and <https://youtu.be/VIz6xBvwYd8>. (This talk is split into two parts.)
- 3 Arthur O'Dwyer. A Soupcon of SFINAE. CppCon, Bellevue, WA, USA, Sept. 27, 2017. Available online at <https://youtu.be/ybaE9qlhHvw>.
- 4 Marshall Clow. The Detection Idiom: A Better Way to SFINAE. C++Now, Aspen, CO, USA, May 19, 2017. Available online at <https://youtu.be/U3jGdnRL3KI>.

Notwithstanding the talk's title, this talk is actually about the functionality in the Library Fundamentals TS related to `is_detected`, `detected_or`, `is_detected_exact`, and `is_detected_convertible`.

- 5 Walter E. Brown. Modern Template Metaprogramming: A Compendium, Part I. CppCon, Bellevue, WA, USA, Sept. 9, 2014. Available online at <https://youtu.be/Am2is2QCvxY>.
- 6 Walter E. Brown. Modern Template Metaprogramming: A Compendium, Part II. CppCon, Bellevue, WA, USA, Sept. 9, 2014. Available online at <https://youtu.be/a0FliKwcwXE>.
- 7 Stephan T. Lavavej. Class Template Argument Deduction for Everyone. CppCon, Bellevue, WA, USA, Sept. 27, 2018. Available online at <https://youtu.be/-H-ut6j1BYU>.

## Section 2.6

# Lambda Expressions

# Motivation for Lambda Expressions

- functor classes extremely useful, especially for generic programming
- writing definitions of functor classes somewhat tedious, especially if many such classes
- functor classes all have same general structure (i.e., constructor, function-call operator, zero or more data members)
- would be nice if functor could be created without need to explicitly write functor-class definition
- lambda expressions provide compact notation for creating functors
- convenience feature (not fundamentally anything new that can be done with lambda expressions that could not already have been done without them)

# Lambda Expressions

- lambda expression consists of:
  - 1 introducer: *capture list* in square brackets
  - 2 declarator: *parameter list* in parentheses followed by *return type* using trailing return-type syntax
  - 3 *compound statement* in brace brackets
- capture list specifies objects to be captured as data members
- declarator specifies parameter list and return type of function-call operator
- compound statement specifies body of function-call operator
- if no declarator specified, defaults to ()
- if no return type specified, defaults to type of expression in return statement, or void if no return statement
- when evaluated, lambda expression yields object called **closure** (which is essentially a functor)
- examples:

```
[ ] (double x) -> int { return floor(x); }  
[ ] (int x, int y) { return x < y; }  
[ ] { std::cout << "Hello, World!\n"; }
```

# Lambda Expressions (Continued)

- closure object is unnamed (temporary object)
- closure type is unnamed
- **operator** () is always inline [\[C++17 8.1.5.1/3\]](#)
- **operator** () is const member function unless **mutable** keyword used [\[C++17 8.1.5.1/4\]](#)
- if closure type is literal type, all members of closure type automatically constexpr
- if no capture, closure type provides conversion function to pointer to function having same parameter and return types as closure type's function call operator; value returned is address of function that, when invoked, has same effect as invoking closure type's function call operator (function pointer not tied to lifetime of closure object) [\[C++17 8.1.5.1/6\]](#)
- although **operator** () in closure very similar to case of normal functor, not everything same (e.g., **operator** () member in closure type cannot access **this** pointer for closure type)

# Hello World Program Revisited

```
1  #include <iostream>
2
3  int main() {
4      []{std::cout << "Hello, World!\n";}();
5  }
```

---

```
1  #include <iostream>
2
3  struct Hello {
4      void operator()() const {
5          std::cout << "Hello, World!\n";
6      }
7  };
8
9  int main() {
10     Hello hello;
11     hello();
12 }
```

# Linear-Function Functor Example

```
1  #include <iostream>
2
3  auto make_linear_func(float a, float b)
4      {return [a, b](float x){return a * x + b;};}
5
6  int main() {
7      float a = 0.5f; float b = 1.0f;
8      auto f = make_linear_func(a, b);
9      std::cout << f(1.0f) << '\n';
10 }
```

---

```
1  #include <iostream>
2
3  class linear_func {
4  public:
5      linear_func(float a, float b) : a_(a), b_(b) {}
6      float operator()(float x) const {return a_ * x + b_;}
7  private:
8      float a_; float b_;
9  };
10
11 linear_func make_linear_func(float a, float b)
12     {return linear_func(a, b);}
13
14 int main() {
15     float a = 0.5f; float b = 1.0f;
16     linear_func f = make_linear_func(a, b);
17     std::cout << f(1.0f) << '\n';
18 }
```



# Comparison Functor Example

```
1  #include <iostream>
2  #include <algorithm>
3  #include <cstdlib>
4  #include <vector>
5
6  int main() {
7      std::vector<int> v{-3, 3, 4, 0, -2, -1, 2, 1, -4};
8      std::sort(v.begin(), v.end(),
9              [](int x, int y) {return std::abs(x) < std::abs(y);});
10     for (auto x : v) std::cout << x << '\n';
11 }
```

---

```
1  #include <iostream>
2  #include <algorithm>
3  #include <cstdlib>
4  #include <vector>
5
6  struct abs_less {
7      bool operator()(int x, int y) const
8          {return std::abs(x) < std::abs(y);}
9  };
10
11 int main() {
12     std::vector<int> v{-3, 3, 4, 0, -2, -1, 2, 1, -4};
13     std::sort(v.begin(), v.end(), abs_less());
14     for (auto x : v) std::cout << x << '\n';
15 }
```

# Capturing Objects

- locals only available if captured; non-locals always available
- can capture by value or by reference
- different locals can be captured differently
- can specify default capture mode
- can explicitly list objects to be captured or not
- might be wise to explicitly list all objects to be captured (when practical) to avoid capturing objects accidentally (e.g., due to typos)
- in member function, to capture class object by value, capture **\*this**
- in member function, can also capture **this**
- **this** must be captured by value

- (unary version of) `std::transform` applies given (unary) operator to each element in range specified by pair of iterators and writes result to location specified by another iterator
- definition of `std::transform` would typically resemble:

```
template <class InputIterator, class OutputIterator,  
         class UnaryOperator>  
OutputIterator transform(InputIterator first,  
                        InputIterator last, OutputIterator result,  
                        UnaryOperator op) {  
    while (first != last) {  
        *result = op(*first);  
        ++result;  
        ++first;  
    }  
    return result;  
}
```

# Modulus Example

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  int main() {
6      int m = 2;
7      std::vector<int> v{0, 1, 2, 3};
8      std::transform(v.begin(), v.end(), v.begin(),
9                     [m](int x){return x % m;});
10     for (auto x : v) std::cout << x << '\n';
11 }
```

---

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  class mod {
6  public:
7      mod(int m_) : m(m_) {}
8      int operator()(int x) const {return x % m;}
9  private:
10     int m;
11 };
12
13 int main() {
14     int m = 2;
15     std::vector<int> v{0, 1, 2, 3};
16     std::transform(v.begin(), v.end(), v.begin(), mod(m));
17     for (auto x : v) std::cout << x << '\n';
18 }
```

# Modulus Example: Without Lambda Expression

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  class mod {
6  public:
7      mod(int m_) : m(m_) {}
8      int operator()(int x) const {return x % m;}
9  private:
10     int m;
11 };
12
13 int main() {
14     int m = 2;
15     std::vector<int> v{0, 1, 2, 3};
16     std::transform(v.begin(), v.end(), v.begin(), mod(m));
17     for (auto x : v) std::cout << x << '\n';
18 }
```

- approximately 8.5 lines of code to generate functor

# Modulus Example: With Lambda Expression

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  int main() {
6      int m = 2;
7      std::vector<int> v{0, 1, 2, 3};
8      std::transform(v.begin(), v.end(), v.begin(),
9          [m](int x){return x % m;});
10     for (auto x : v) std::cout << x << '\n';
11 }
```

- `m` captured by value
- approximately 0.5 lines of code to generate functor

- `std::for_each` applies given function/functor to each element in range specified by pair of iterators
- definition of `std::for_each` would typically resemble:

```
template<class InputIterator, class Function>
Function for_each(InputIterator first,
                  InputIterator last, Function func) {
    while (first != last) {
        func(*first);
        ++first;
    }
    return move(func);
}
```

# Product Example

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  int main() {
6      std::vector<int> v{2, 3, 4};
7      int prod = 1;
8      std::for_each(v.begin(), v.end(),
9          [&prod](int x)->void{prod *= x;});
10     std::cout << prod << '\n';
11 }
```

---

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  class cum_prod {
6  public:
7      cum_prod(int& prod_) : prod(prod_) {}
8      void operator()(int x) const {prod *= x;}
9  private:
10     int& prod;
11 };
12
13 int main() {
14     std::vector<int> v{2, 3, 4};
15     int prod = 1;
16     std::for_each(v.begin(), v.end(), cum_prod(prod));
17     std::cout << prod << '\n';
18 }
```



# Product Example: Without Lambda Expression

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  class cum_prod {
6  public:
7      cum_prod(int& prod_) : prod(prod_) {}
8      void operator()(int x) const {prod *= x;}
9  private:
10     int& prod;
11 };
12
13 int main() {
14     std::vector<int> v{2, 3, 4};
15     int prod = 1;
16     std::for_each(v.begin(), v.end(), cum_prod(prod));
17     std::cout << prod << '\n';
18 }
```

- approximately 8.5 lines of code to generate functor

# Product Example: With Lambda Expression

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  int main() {
6      std::vector<int> v{2, 3, 4};
7      int prod = 1;
8      std::for_each(v.begin(), v.end(),
9          [&prod](int x)->void{prod *= x;});
10     std::cout << prod << '\n';
11 }
```

- `prod` captured by reference
- approximately 1 line of code to generate functor

# More Variations on Capture

```
double a = 2.14;  
double b = 3.14;  
double c = 42.0;  
  
// capture all objects by reference (i.e., a, b, and c)  
[&](double x, double y){return a * x + b * y + c;}  
  
// capture all objects by value (i.e., a, b, and c)  
[=](double x, double y){return a * x + b * y + c;}  
  
// capture all objects by value, except a  
// which is captured by reference  
[=,&a](double x, double y){return a * x + b * y + c;}  
  
// capture all objects by reference, except a  
// which is captured by value  
[&,a](double x, double y){return a * x + b * y + c;}
```

- can specify name for captured object in closure type

```
int a = 1;  
auto f = [x = a]() {return x;};
```

- can capture result of expression (e.g., to perform move instead of copy or to add arbitrary new state to closure type)

```
std::vector<int> v(1000, 1);  
auto f = [v = std::move(v)]() ->  
    const std::vector<int>& {return v;};
```

# Generalized Lambda Capture Example

```
1  #include <iostream>
2
3  int main() {
4      int x = 0;
5      int y = 1;
6      auto f = [&count = x, inc = y + 1]() {
7          return count += inc;
8      };
9      std::cout << f() << ' ';
10     std::cout << f() << '\n';
11 }
12
13 // output: 2 4
```

# Generic Lambda Expressions

- can allow compiler to deduce type of lambda function parameters
- generates closure type with templated function-call operator
- one template type parameter for each occurrence of **auto** in lambda expression's parameter declaration clause [C++17 8.1.5.1/3]

# Generic Lambda Expression Example [Generic]

```
1  #include <iostream>
2  #include <complex>
3  #include <string>
4
5  int main() {
6      using namespace std::literals;
7      auto add = [](auto x, auto y) {return x + y;};
8      std::cout << add(1, 2) << ' ' << add(1.0, 2.0) << ' '
9          << add(1.0, 2.0i) << ' ' << add("Jell"s, "o"s) << '\n';
10 }
```

---

```
1  #include <iostream>
2  #include <complex>
3  #include <string>
4
5  struct Add {
6      template <class T, class U>
7      auto operator()(T x, U y) {return x + y;};
8  };
9
10 int main() {
11     using namespace std::literals;
12     Add add;
13     std::cout << add(1, 2) << ' ' << add(1.0, 2.0) << ' '
14         << add(1.0, 2.0i) << ' ' << add("Jell"s, "o"s) << '\n';
15 }
```

# Generic Lambda Expression Example [Convenience]

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  int main() {
6      std::vector<int> v{0, 1, 2, 3, 4, 5, 6, 7};
7      // sort elements of vector in descending order
8      std::sort(v.begin(), v.end(),
9          [](auto i, auto j) {return i > j;});
10     std::for_each(v.begin(), v.end(),
11         [](auto i) {std::cout << i << '\n';});
12 }
```



# Dealing With Unnamed Types

- fact that closure types unnamed causes complications when need arises to refer to closure type
- helpful language features: **auto**, **decltype**
- helpful library features: `std::function`
- closures can be stored using **auto** or `std::function`
- closures that do not capture can be “stored” by assigning to function pointer

# Using `auto`, `decltype`, and `std::function`

```
1  #include <iostream>
2  #include <functional>
3
4  std::function<double(double)> linear(double a, double b) {
5      return [=](double x){return a * x + b;};
6  }
7
8  int main() {
9      // type of f is std::function<double(double)>
10     auto f = linear(2.0, -1.0);
11     // g has closure type
12     auto g = [](double x){return 2.0 * x - 1.0;};
13     double (*u)(double) = [](double x){return 2.0 * x - 1.0;};
14     // h has same type as g
15     decltype(g) h = g;
16     for (double x = 0.0; x < 10.0; x += 1.0) {
17         std::cout << x << ' ' << f(x) << ' ' << g(x) <<
18             ' ' << h(x) << (*u)(x) << '\n';
19     }
20 }
```

- applying function-call operator to `f` much slower than in case of `g` and `h`
- when `std::function` used, inlining of called function probably not possible
- when functor used directly (via function-call operator) inlining is very likely
- prefer `auto` over `std::function` for storing closures

# operator() as Non-const Member

```
1  #include <iostream>
2
3  int main()
4  {
5      int count = 5;
6      // Must use mutable in order to be able to
7      // modify count member.
8      auto get_count = [count]() mutable -> int {
9          return count++;
10     };
11
12     int c;
13     while ((c = get_count()) < 10) {
14         std::cout << c << '\n';
15     }
16 }
```

- **operator()** is declared as const member function unless **mutable** keyword used
- const member function cannot change (non-static) data members

# Constexpr Lambdas

```
1  #include <iostream>
2  #include <array>
3
4  template <typename T>
5  constexpr auto multiply_by(T i) {
6      return [i](auto j) {return i * j;};
7      // OK: lambda is literal type so members
8      // are automatically constexpr
9  }
10
11 int main() {
12     constexpr auto mult_by_2 = multiply_by(2);
13     std::array<int, mult_by_2(8)> a;
14     std::cout << a.size() << '\n';
15 }
```

# Comparison Functors for Containers

```
1  #include <iostream>
2  #include <vector>
3  #include <set>
4
5  int main() {
6      // The following two lines are the only important ones:
7      auto cmp = [](int* x, int* y){return *x < *y;};
8      std::set<int*, decltype(cmp)> s(cmp);
9
10     // Just for something to do:
11     // Print the elements of v in sorted order with
12     // duplicates removed.
13     std::vector<int> v = {4, 1, 3, 2, 1, 1, 1, 1};
14     for (auto& x : v) {
15         s.insert(&x);
16     }
17     for (auto x : s) {
18         std::cout << *x << '\n';
19     }
20 }
```

- note that `s` is not default constructed
- since closure types not default constructible, following would fail:  
`std::set<int*, decltype(cmp)> s;`
- note use of `decltype` in order to specify type of functor

# What Could Possibly Go Wrong?

```
1  #include <iostream>
2  #include <vector>
3  #include <functional>
4
5  std::vector<int> vec{2000, 4000, 6000, 8000, 10000};
6  std::function<int(int)> func;
7
8  void do_stuff()
9  {
10     int modulus = 10000;
11     func = [&](int x){return x % modulus;};
12     for (auto x : vec) {
13         std::cout << func(x) << '\n';
14     }
15 }
16
17 int main()
18 {
19     do_stuff();
20     for (auto x : vec) {
21         std::cout << func(x) << '\n';
22     }
23 }
```

- above code has very serious bug; what is it?

# Dangling References

- if some objects captured by reference, closure can hold dangling references
- responsibility of programmer to avoid such problems
- if will not cause performance issues, may be advisable to capture by value (to avoid problem of dangling references)
- dangling-reference example:

```
1  #include <iostream>
2  #include <functional>
3
4  std::function<double(double)> linear(double a, double b) {
5      return [&](double x){return a * x + b;};
6  }
7
8  int main() {
9      auto f = linear(2.0, -1.0);
10     // bad things will happen here
11     std::cout << f(1.0) << '\n';
12 }
```

## Section 2.6.1

### **References**



- 1 Herb Sutter. Lambdas, Lambdas Everywhere. Professional Developers Conference (PDC), Redmond, WA, USA, Oct. 27–29, 2010. Available online at <https://youtu.be/rcgRY7s0A58>.
- 2 Herb Sutter. C++0x Lambda Functions. Northwest C++ Users' Group (NWCPP), Redmond, WA, USA, May 18, 2011. Available online at <https://vimeo.com/23975522>.

## Section 2.7

# Classes and Inheritance

## Section 2.7.1

### **Derived Classes and Class Hierarchies**

# Derived Classes

- sometimes, want to express commonality between classes
- want to create new class from existing class by adding new members or replacing (i.e., hiding/overriding) existing members
- can be achieved through language feature known as *inheritance*
- generate new class with all members of already existing class, excluding special member functions (i.e., constructors, assignment operators, and destructor)
- new class called **derived class** and original class called **base class**
- derived class said to **inherit** from base class
- can add new members (not in base class) to derived class
- can hide or override member functions from base class with new version
- syntax for specifying derived class:

**class** *derived\_class* : *base\_class\_specifiers*

- *derived\_class* is name of derived class; *base\_class\_specifiers* provide base-class information

# Derived Classes (Continued)

- can more clearly express intent by explicitly identifying relationship between classes
- can facilitate code reuse by leverage existing code
- interface inheritance: allow different derived classes to be used interchangeably through interface provided by common base class
- implementation inheritance: save implementation effort by sharing capabilities provided by base class

# Person Class

```
1  #include <string>
2
3  class Person {
4  public:
5      Person(const std::string& family_name,
6             const std::string& given_name) :
7          family_name_(family_name), given_name_(given_name) {}
8      std::string family_name() const {return family_name_;}
9      std::string given_name() const {return given_name_;}
10     std::string full_name() const
11     {return family_name_ + ", " + given_name_;}
12     // ...
13 private:
14     std::string family_name_;
15     std::string given_name_;
16 };
```

# Student Class Without Inheritance

```
1  #include <string>
2
3  class Student {
4  public:
5      Student(const std::string& family_name,
6              const std::string& given_name) :
7          family_name_(family_name), given_name_(given_name) {}
8          // NEW
9      std::string family_name() const {return family_name_;}
10     std::string given_name() const {return given_name_;}
11     std::string full_name() const
12         {return family_name_ + ", " + given_name_;}
13     std::string student_id() {return student_id_;} // NEW
14 private:
15     std::string family_name_;
16     std::string given_name_;
17     std::string student_id_; // NEW
18 };
```

# Student Class With Inheritance

```
1  // include definition of Person class here
2
3  class Student : public Person {
4  public:
5      Student(const std::string& family_name,
6              const std::string& given_name,
7              const std::string& student_id) :
8          Person(family_name, given_name),
9          student_id_(student_id) {}
10     std::string student_id() {return student_id_;}
11 private:
12     std::string student_id_;
13 };
```



# Complete Inheritance Example

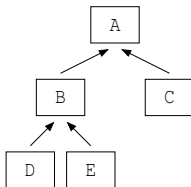
```
1  #include <string>
2
3  class Person {
4  public:
5      Person(const std::string& family_name,
6             const std::string& given_name) :
7          family_name_(family_name), given_name_(given_name) {}
8      std::string family_name() const {return family_name_;}
9      std::string given_name() const {return given_name_;}
10     std::string full_name() const
11         {return family_name_ + ", " + given_name_;}
12     // ... (including virtual destructor)
13 private:
14     std::string family_name_;
15     std::string given_name_;
16 };
17
18 class Student : public Person {
19 public:
20     Student(const std::string& family_name,
21            const std::string& given_name,
22            const std::string& student_id) :
23         Person(family_name, given_name),
24         student_id_(student_id) {}
25     std::string student_id() {return student_id_;}
26 private:
27     std::string student_id_;
28 };
```

# Class Hierarchies

- inheritance relationships between classes form what is called **class hierarchy**
- often class hierarchy represented by directed (acyclic) graph, where nodes correspond to classes and edges correspond to inheritance relationships
- class definitions:

```
class A { /* ... */ };  
class B : public A { /* ... */ };  
class C : public A { /* ... */ };  
class D : public B { /* ... */ };  
class E : public B { /* ... */ };
```

- inheritance diagram:

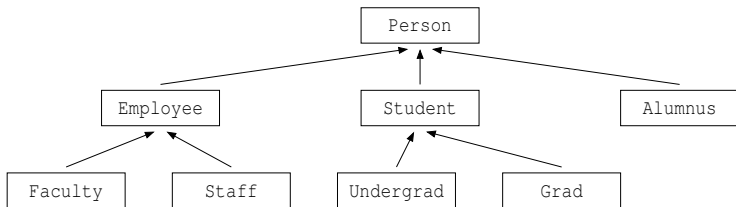


# Class Hierarchy Example

## ■ class definitions:

```
class Person { /* ... */ };  
class Employee : public Person { /* ... */ };  
class Student : public Person { /* ... */ };  
class Alumnus : public Person { /* ... */ };  
class Faculty : public Employee { /* ... */ };  
class Staff : public Employee { /* ... */ };  
class Grad : public Student { /* ... */ };  
class Undergrad : public Student { /* ... */ };
```

## ■ inheritance diagram:



- each of Employee, Student, and Alumnus is a Person; each of Faculty and Staff is an Employee; each of Undergrad and Grad is a Student

# Member Access Specifiers: **protected**

- earlier, introduced **public** and **private** access specifiers for class members
- in context of inheritance, another access specifier becomes relevant, namely, **protected**
- member declared in protected section of class can only be accessed by
  - member functions and friends of that class; and
  - by member functions and friends of *derived classes*
- protected members used to provide developers of derived classes access to some inner workings of base class without exposing such inner workings to everyone
- usually, bad idea to use protected access for data members (for similar reasons that using public access for data members is usually bad)
- protected access usually employed for function members

# Types of Inheritance

- three types of inheritance with respect to access protection: public, protected, and private
- these three types of inheritance differ in terms of accessibility, in derived class, of members inherited from base class
- private parts of base class are always inaccessible in derived class, regardless of whether public, protected, or private inheritance used
- if this were not case, all access protection could simply be bypassed by using inheritance
- access specifiers for members accessible in derived class chosen as follows:

Access Specifier in Base Class	Access Specifier in Derived Class		
	Public Inheritance	Protected Inheritance	Private Inheritance
public	public	protected	private
protected	protected	protected	private

# Types of Inheritance (Continued)

- for struct, defaults to public inheritance
- for class, defaults to private inheritance
- public and protected/private inheritance have different use cases, as we will see later

# Inheritance and Member Access Example

```
1  class Base {
2  public:
3      void f();
4  protected:
5      void g();
6  private:
7      int x;
8  };
9
10 class Derived_1 : public Base {
11     // f is public
12     // g is protected
13     // x is not accessible from Derived_1
14 };
15
16 class Derived_2 : protected Base {
17     // f is protected
18     // g is protected
19     // x is not accessible from Derived_2
20 };
21
22 class Derived_3 : private Base {
23     // f is private
24     // g is private
25     // x is not accessible from Derived_3
26 };
```

# Public Inheritance Example

```
1  class Base {
2  public:
3      void func_1();
4  protected:
5      void func_2();
6  private:
7      int x_;
8  };
9
10 class Derived : public Base {
11 public:
12     void func_3() {
13         func_1(); // OK
14         func_2(); // OK
15         x_ = 0; // ERROR: inaccessible
16     }
17 };
18
19 struct Widget : public Derived {
20     void func_4() { func_2(); } // OK
21 };
22
23 int main() {
24     Derived d;
25     d.func_1(); // OK
26     d.func_2(); // ERROR: inaccessible
27     d.x_ = 0; // ERROR: inaccessible
28 }
```



# Protected Inheritance Example

```
1  class Base {
2  public:
3      void func_1();
4  protected:
5      void func_2();
6  private:
7      int x_;
8  };
9
10 class Derived : protected Base {
11 public:
12     void func_3() {
13         func_1(); // OK
14         func_2(); // OK
15         x_ = 0; // ERROR: inaccessible
16     }
17 };
18
19 struct Widget : public Derived {
20     void func_4() { func_2(); } // OK
21 };
22
23 int main() {
24     Derived d; // OK: defaulted constructor is public
25     d.func_1(); // ERROR: inaccessible
26     d.func_2(); // ERROR: inaccessible
27     d.x_ = 0; // ERROR: inaccessible
28 }
```

# Private Inheritance Example

```
1  class Base {
2  public:
3      void func_1();
4  protected:
5      void func_2();
6  private:
7      int x_;
8  };
9
10 class Derived : private Base {
11 public:
12     void func_3() {
13         func_1(); // OK
14         func_2(); // OK
15         x_ = 0; // ERROR: inaccessible
16     }
17 };
18
19 struct Widget : public Derived {
20     void func_4() { func_2(); } // ERROR: inaccessible
21 };
22
23 int main() {
24     Derived d; // OK: defaulted constructor is public
25     d.func_1(); // ERROR: inaccessible
26     d.func_2(); // ERROR: inaccessible
27     d.x_ = 0; // ERROR: inaccessible
28 }
```

- public inheritance is inheritance in traditional object-oriented programming sense
- public inheritance models an *is-a* relationship (i.e., derived class object is a base class object)
- most common form of inheritance
- inheritance relationship visible to all code

# Public Inheritance Example

```
1  #include <string>
2
3  class Person {
4  public:
5      Person(const std::string& family_name, const std::string&
6          given_name) : family_name_(family_name),
7          given_name_(given_name) {}
8      std::string family_name() const
9          {return family_name_;}
10     std::string given_name() const
11         {return given_name_;}
12     std::string full_name() const
13         {return family_name_ + ", " + given_name_;}
14 private:
15     std::string family_name_;
16     std::string given_name_;
17 };
18
19 class Student : public Person {
20 public:
21     Student(const std::string& family_name, const std::string&
22         given_name, const std::string& student_id) :
23         Person(family_name, given_name), student_id_(student_id) {}
24     std::string student_id()
25         {return student_id_;}
26 private:
27     std::string student_id_;
28 };
```

# Protected and Private Inheritance

- protected and private inheritance not inheritance in traditional object-oriented programming sense (i.e., no is-a relationship)
- form of implementation inheritance
- *implemented-in-terms-of* relationship (i.e., derived class object implemented in terms of a base class object)
- in case of protected inheritance, inheritance relationship only seen by derived classes and their friends and class itself and its friends
- in case of private inheritance, inheritance relationship only seen by class itself and its friends (not derived classes and their friends)
- except in special circumstances, normally bad idea to use inheritance for composition
- one good use case for private/protected inheritance is in policy-based design, which exploits empty base optimization (EBO)

# Policy-Based Design Example: Inefficient Memory Usage

```
1  #include <mutex>
2
3  class ThreadSafePolicy {
4  public:
5      void lock() {mutex_.lock();}
6      void unlock() {mutex_.unlock();}
7  private:
8      std::mutex mutex_;
9  };
10
11 class ThreadUnsafePolicy {
12 public:
13     void lock() {} // no-op
14     void unlock() {} // no-op
15 };
16
17 template<class ThreadSafetyPolicy>
18 class Widget {
19     ThreadSafetyPolicy policy_;
20     // ...
21 };
22
23 int main() {
24     Widget<ThreadUnsafePolicy> w;
25     // w.policy_ has no data members, but
26     // sizeof(w.policy_) >= 1
27     // inefficient use of memory
28 }
```

# Policy-Based Design Example: Private Inheritance and EBO

```
1  #include <mutex>
2
3  class ThreadSafePolicy {
4  public:
5      void lock() {mutex_.lock();}
6      void unlock() {mutex_.unlock();}
7  private:
8      std::mutex mutex_;
9  };
10
11 class ThreadUnsafePolicy {
12 public:
13     void lock() {} // no-op
14     void unlock() {} // no-op
15 };
16
17 template<class ThreadSafetyPolicy>
18 class Widget : ThreadSafetyPolicy {
19     // ...
20 };
21
22 int main() {
23     Widget<ThreadUnsafePolicy> w;
24     // empty-base optimization (EBO) can be applied
25     // no memory overhead for no-op thread-safety policy
26 }
```

# Inheritance and Constructors

- by default, constructors not inherited
- often, derived class introduces new data members not in base class
- since base-class constructors cannot initialize derived-class data members, inheriting constructors from base class by default would be bad idea (e.g., could lead to uninitialized data members)
- in some cases, however, base-class constructors may be sufficient to initialize derived-class objects
- in such cases, can inherit all non-special base-class constructors with **using** statement
- special constructors (i.e., default, copy, and move constructors) cannot be inherited [C++14 12.9/3] [C++14 12.9/6]
- constructors to be inherited with **using** statement may still be hidden by constructors in derived class



# Inheriting Constructors Example

```
1  class Base {
2  public:
3      Base() : i_(0.0), j_(0) {}
4      Base(int i) : i_(i), j_(0) {}
5      Base(int i, int j) : i_(i), j_(j) {}
6      // ... (other non-constructor members)
7  private:
8      int i_, j_;
9  };
10
11 class Derived : public Base {
12 public:
13     // inherit non-special constructors from Base
14     // (default constructor not inherited)
15     using Base::Base;
16     // default constructor is implicitly declared and
17     // not inherited
18 };
19
20 int main() {
21     Derived a;
22     // invokes non-inherited Derived::Derived()
23     Derived b(42, 42);
24     // invokes inherited Base::Base(int, int)
25 }
```

# Inheriting Constructors Example

```
1  class Base {
2  public:
3      Base() : i_(0), j_(0), k_(0) {}
4      Base(int i, int j) : i_(i), j_(j), k_(0) {}
5      Base(int i, int j, int k) : i_(i), j_(j), k_(k) {}
6      // ... (other non-constructor members)
7  private:
8      int i_, j_, k_;
9  };
10
11 class Derived : public Base {
12 public:
13     // inherit non-special constructors from Base
14     // (default constructor not inherited)
15     using Base::Base;
16     // following constructor hides inherited constructor
17     Derived(int i, int j, int k) : Base(-i, -j, -k) {}
18     // no implicitly-generated default constructor
19 };
20
21 int main() {
22     Derived b(1, 2);
23     // invokes inherited Base::Base(int, int)
24     Derived c(1, 2, 3);
25     // invokes Derived::Derived(int, int, int)
26     // following would produce compile-time error:
27     // Derived a; // ERROR: no default constructor
28 }
```

# Inheritance, Assignment Operators, and Destructors

- by default, assignment operators not inherited (for similar reasons as in case of constructors)
- can inherit all non-special base-class assignment operators with **using** statement
- copy and move assignment operators cannot be inherited [C++14 12.8/24]
- assignment operators to be inherited with **using** statement may still be hidden by assignment operators in derived class
- cannot inherit destructor

# Inheriting Assignment Operators Example

```
1  class Base {
2  public:
3      explicit Base(int i) : i_(i) {}
4      Base& operator=(int i) {
5          i_ = i;
6          return *this;
7      }
8      // ...
9  private:
10     int i_;
11 };
12
13 class Derived : public Base {
14 public:
15     // inherit non-special constructors
16     using Base::Base;
17     // inherit non-special assignment operators
18     using Base::operator=;
19     // ...
20 };
21
22 int main() {
23     Derived d(0);
24     // invokes inherited Base::Base(int)
25     d = 42;
26     // invokes inherited Base::operator=(int)
27 }
```

# Construction and Destruction Order

## ■ order of construction:

- 1 if most-derived class in hierarchy, initialize all *virtual base class objects* in hierarchy in order of depth-first left-to-right traversal of graph of base class declarations, where left to right refers to order of appearance of base class names in class definition (virtual base classes to be discussed later)
- 2 initialize *non-virtual (direct) base class objects* in order listed in class definition
- 3 initialize *non-static data members* in order of declaration in class definition
- 4 execute *constructor body*

## ■ order of destruction is exact reverse of order of construction, namely:

- 1 execute *destructor body*
- 2 destroy *non-static data members* in reverse of construction order
- 3 destroy *non-virtual (direct) base class objects* in reverse of construction order
- 4 if most-derived class in hierarchy, destroy all *virtual base class objects* in hierarchy in reverse of construction order

# Order of Construction

```
1  #include <vector>
2  #include <string>
3
4  class Base {
5  public:
6      Base(int n) : v_(n, 0) {}
7      // ...
8  private:
9      std::vector<char> v_;
10 };
11
12 class Derived : public Base {
13 public:
14     Derived(const std::string& s) : Base(1024), s_(s)
15     { i_ = 0; }
16     // ...
17 private:
18     std::string s_;
19     int i_;
20 };
21
22 int main() {
23     Derived d("hello");
24 }
```

- construction order for Derived constructor: 1) Base class object, 2) data member s\_, 3) Derived constructor body (initializes data member i\_)

# Hiding Base-Class Member Functions in Derived Class

- can provide new versions of member functions in derived class to hide original functions in base class

```
1  #include <iostream>
2
3  class Fruit {
4  public:
5      void print() const {std::cout << "fruit\n";}
6  };
7
8  class Apple : public Fruit {
9  public:
10     void print() const {std::cout << "apple\n";}
11 };
12
13 class Banana : public Fruit {
14 public:
15     void print() const {std::cout << "banana\n";}
16 };
17
18 int main() {
19     Fruit f;
20     Apple a;
21     Banana b;
22     f.print(); // calls Fruit::print
23     a.print(); // calls Apple::print
24     b.print(); // calls Banana::print
25 }
```

# Upcasting

- derived-class object always has base-class subobject
- given reference or pointer to derived-class object, may want to find reference or pointer to corresponding base-class object
- **upcasting**: converting derived-class pointer or reference to base-class pointer or reference
- upcasting allows us to treat derived-class object as base-class object
- upcasting always safe in sense that cannot result in incorrect type (since every derived-class object is also a base-class object)
- can upcast without explicit type-cast operator as long as casted-to type is accessible; C-style cast can be used to bypass access protection (although not recommended)
- example:

```
class Base { /* ... */ };  
class Derived : public Base { /* ... */ };  
void func() {  
    Derived d;  
    Base* bp = &d;  
}
```



# Downcasting

- **downcasting**: converting base-class pointer or reference to derived-class pointer or reference
- downcasting allows us to force base-class object to be treated as derived-class object
- downcasting is not always safe (since not every base-class object is necessarily also derived-class object)
- must only downcast when known that object actually has derived type (except in case of **dynamic\_cast**)
- downcasting always requires explicit cast (e.g., **static\_cast**, **dynamic\_cast** for dynamically-checked cast in polymorphic case, or C-style cast)
- example:

```
class Base { /* ... (nonpolymorphic) */ };  
class Derived : public Base { /* ... */ };  
void func() {  
    Derived d;  
    Base* bp = &d;  
    Derived* dp = static_cast<Derived*>(bp);  
}
```

# Upcasting/Downcasting Example

```
1  class Base { /* ... (nonpolymorphic) */ };
2
3  class Derived : public Base { /* ... */ };
4
5  int main() {
6      Base b;
7      Derived d;
8      Base* bp = nullptr;
9      Derived* dp = nullptr;
10     bp = &d;
11     // OK: upcast does not require explicit cast
12     dp = bp;
13     // ERROR: downcast requires explicit cast
14     dp = static_cast<Derived*>(bp);
15     // OK: downcast with explicit cast and
16     // pointer (bp) refers to Derived object
17     Base& br = d;
18     // OK: upcast does not require explicit cast
19     Derived& dr1 = *bp;
20     // ERROR: downcast requires explicit cast
21     Derived& dr2 = *static_cast<Derived*>(bp);
22     // OK: downcast with explicit cast and
23     // object (*bp) is of Derived type
24     dp = static_cast<Derived*>(&b);
25     // BUG: pointer (&b) does not refer to Derived object
26 }
```

# Upcasting Example

```
1  class Base { /* ... */ };
2
3  class Derived : public Base { /* ... */ };
4
5  void func_1(Base& b) { /* ... */ }
6
7  void func_2(Base* b) { /* ... */ }
8
9  int main() {
10     Base b;
11     Derived d;
12     func_1(b);
13     func_1(d); // OK: Derived& upcast to Base&
14     func_2(&b);
15     func_2(&d); // OK: Derived* upcast to Base*
16 }
```

# Nonpolymorphic Behavior

```
1  #include <iostream>
2  #include <string>
3
4  class Person {
5  public:
6      Person(const std::string& family, const std::string& given) :
7          family_(family), given_(given) {}
8      void print() const {std::cout << "person: " << family_ << ',' << given_ << '\n';}
9  protected:
10     std::string family_; // family name
11     std::string given_; // given name
12 };
13
14 class Student : public Person {
15 public:
16     Student(const std::string& family, const std::string& given,
17             const std::string& id) : Person(family, given), id_(id) {}
18     void print() const {
19         std::cout << "student: " << family_ << ',' << given_ << ',' << id_ << '\n';
20     }
21 private:
22     std::string id_; // student ID
23 };
24
25 void processPerson(const Person& p) {
26     p.print(); // always calls Person::print
27     // ...
28 }
29
30 int main() {
31     Person p("Ritchie", "Dennis");
32     Student s("Doe", "John", "12345678");
33     processPerson(p); // invokes Person::print
34     processPerson(s); // invokes Person::print
35 }
```

- would be nice if processPerson called version of print that corresponds to *actual* type of object referenced by function parameter p

- **slicing**: copying or moving object of derived class to object of base class (e.g., during construction or assignment), losing part of information in so doing
- example:

```
1  class Base {
2      // ...
3      int x_;
4  };
5
6  class Derived : public Base {
7      // ...
8      int y_;
9  };
10
11 int main() {
12     Derived d1, d2;
13     Base b = d1;
14     // slicing occurs
15     Base& r = d1;
16     r = d2;
17     // more treacherous case of slicing
18     // slicing occurs
19     // d1 now contains mixture of d1 and d2
20     // (i.e., base part of d2 and derived part of d1)
21 }
```

# Inheritance and Overloading

- functions do not overload across scopes
- can employ **using** statement to bring base members into scope for overloading

# Inheritance and Overloading Example

```
1  #include <iostream>
2
3  class Base {
4  public:
5      double f(double d) const {return d;}
6      // ...
7  };
8
9  class Derived : public Base {
10 public:
11     int f(int i) const {return i;}
12     // ...
13 };
14
15 int main()
16 {
17     Derived d;
18     std::cout << d.f(0) << '\n';
19     // calls Derived::f(int) const
20     std::cout << d.f(0.5) << '\n';
21     // calls Derived::f(int) const; probably not intended
22     Derived* dp = &d;
23     std::cout << dp->f(0) << '\n';
24     // calls Derived::f(int) const
25     std::cout << dp->f(0.5) << '\n';
26     // calls Derived::f(int) const; probably not intended
27 }
```

# Using Base Members Example

```
1  #include <iostream>
2
3  class Base {
4  public:
5      double f(double d) const {return d;}
6      // ...
7  };
8
9  class Derived : public Base {
10 public:
11     using Base::f; // bring Base::f into scope
12     int f(int i) const {return i;}
13     // ...
14 };
15
16 int main()
17 {
18     Derived d;
19     std::cout << d.f(0) << '\n';
20     // calls Derived::f(int) const
21     std::cout << d.f(0.5) << '\n';
22     // calls Base::f(double) const
23     Derived* dp = &d;
24     std::cout << dp->f(0) << '\n';
25     // calls Derived::f(int) const
26     std::cout << dp->f(0.5) << '\n';
27     // calls Base::f(double) const
28 }
```



# Inheritance, Templates, and Name Lookup

- name lookup in templates takes place in two phases:
  - 1 at template definition time
  - 2 at template instantiation time
- at template definition time, compiler parses template and looks up any *nondependent* names
- result of nondependent name lookup must be *identical* in all instantiations of template (since, by definition, nondependent name does not depend on template parameter)
- at template instantiation time, compiler looks up any *dependent* names
- results of dependent name lookup can differ from one template instantiation to another (since, by definition, dependent name depends on template parameters)
- two-phase name lookup can interact with inheritance in ways that can sometimes lead to unexpected problems in code
- may need to add “**this->**” or employ **using** statement to make name dependent (when it would otherwise be nondependent)

# Name Lookup Example (Incorrect Code)

```
1  #include <iostream>
2
3  template <class T>
4  struct Base {
5      using Real = T;
6      Base(Real x_ = Real()) : x(x_) {}
7      void f() {std::cout << x << "\n";};
8      Real x;
9  };
10
11 template <class T>
12 struct Derived : Base<T> {
13     Derived(Real y_ = Real()) : y(y_) {}
14     // ERROR: Real (which is nondependent and looked up at
15     // template definition time) is assumed to be defined
16     // outside class
17     void g() {
18         x = y;
19         // ERROR: x assumed to be object outside class
20         f();
21         // ERROR: f assumed to be function outside class
22     }
23     Real y;
24 };
25
26 int main() {
27     Derived<double> w(0.0);
28     w.g();
29 }
```

# Name Lookup Example (Correct Code)

```
1  #include <iostream>
2
3  template <class T>
4  struct Base {
5      using Real = T;
6      Base(Real x_ = Real()) : x(x_) {}
7      void f() {std::cout << x << "\n";};
8      Real x;
9  };
10
11 template <class T>
12 struct Derived : Base<T> {
13     using Real = typename Base<T>::Real;
14     // OK: Base<T>::Real dependent
15     Derived(Real y_ = Real()) : y(y_) {}
16     void g() {
17         this->x = y; // OK: this->x dependent
18         this->f(); // OK: this->f dependent
19     }
20     Real y;
21 };
22
23 int main() {
24     Derived<double> w(0.0);
25     w.g();
26 }
```

## Section 2.7.2

# **Virtual Functions and Run-Time Polymorphism**

# Run-Time Polymorphism

- **polymorphism** is characteristic of being able to assign different meaning to something in different contexts
- polymorphism that occurs at run time called **run-time polymorphism** (also known as **dynamic polymorphism**)
- in context of inheritance, key type of run-time polymorphism is polymorphic function call (also known as dynamic dispatch)
- when inheritance relationship exists between two classes, type of reference or pointer to object may not correspond to actual dynamic (i.e., run-time) type of object referenced by reference or pointer
- that is, reference or pointer to type  $T$  may, in fact, refer to object of type  $D$ , where  $D$  is either directly or indirectly derived from  $T$
- when calling member function through pointer or reference, may want actual function invoked to be determined by **dynamic** type of object referenced by pointer or reference
- function call with this property said to be **polymorphic**

# Virtual Functions

- in context of class hierarchies, polymorphic function calls achieved through use of virtual functions
- **virtual function** is member function with polymorphic behavior
- when call made to virtual function through reference or pointer, actual function invoked will be determined by *dynamic* type of referenced object
- to make member function virtual, add keyword **virtual** to function declaration
- example:

```
class Base {  
public:  
    virtual void func(); // virtual function  
    // ...  
};
```

## Virtual Functions (Continued)

- once function made virtual, it will *automatically* be virtual in all derived classes, regardless of whether **virtual** keyword is used in derived classes
- therefore, not necessary to repeat **virtual** qualifier in derived classes (and perhaps preferable not to do so)
- virtual function must be defined in class where first declared unless pure virtual function (to be discussed shortly)
- derived class inherits definition of each virtual function from its base class, but may override each virtual function with new definition
- function in derived class with same name and same set of argument types as virtual function in base class overrides base class version of virtual function

# Virtual Function Example

```
1  #include <iostream>
2  #include <string>
3
4  class Person {
5  public:
6      Person(const std::string& family, const std::string& given) :
7          family_(family), given_(given) {}
8      virtual void print() const
9          {std::cout << "person: " << family_ << ',' << given_ << '\n';}
10 protected:
11     std::string family_; // family name
12     std::string given_; // given name
13 };
14
15 class Student : public Person {
16 public:
17     Student(const std::string& family, const std::string& given,
18             const std::string& id) : Person(family, given), id_(id) {}
19     void print() const {
20         std::cout << "student: " << family_ << ',' << given_ << ',' << id_ << '\n';
21     }
22 private:
23     std::string id_; // student ID
24 };
25
26 void processPerson(const Person& p) {
27     p.print(); // polymorphic function call
28     // ...
29 }
30
31 int main() {
32     Person p("Ritchie", "Dennis");
33     Student s("Doe", "John", "12345678");
34     processPerson(p); // invokes Person::print
35     processPerson(s); // invokes Student::print
36 }
```



# Override Control: The **override** Qualifier

- when looking at code for derived class, often not possible to determine if member function intended to override virtual function in base class (or one of its base classes)
- can sometimes lead to bugs where programmer expects member function to override virtual function when function not virtual
- **override** qualifier used to indicate that member function is expected to override virtual function in parent class; must come at end of function declaration
- example:

```
class Person {  
public:  
    virtual void print() const;  
    // ...  
};  
  
class Employee : public Person {  
public:  
    void print() const override; // must be virtual  
    // ...  
};
```

# Override Control: The **final** Qualifier

- sometimes, may want to prevent any further overriding of virtual function in any subsequent derived classes
- adding **final** qualifier to declaration of virtual function prevents function from being overridden in any subsequent derived classes
- preventing further overriding can sometimes allow for better optimization by compiler (e.g., via devirtualization)
- example:

```
class A {  
public:  
    virtual void doStuff();  
    // ...  
};  
  
class B : public A {  
public:  
    void doStuff() final; // prevent further overriding  
    // ...  
};  
  
class C : public B {  
public:  
    void doStuff(); // ERROR: cannot override  
    // ...  
};
```

# final Qualifier Example

```
1  class Worker {
2  public:
3      virtual void prepareEnvelope();
4      // ...
5  };
6
7  class SpecialWorker : public Worker {
8  public:
9      // prevent overriding function responsible for
10     // overall envelope preparation process
11     // but allow functions for individual steps in
12     // process to be overridden
13     void prepareEnvelope() final {
14         stuffEnvelope(); // step 1
15         lickEnvelope(); // step 2
16         sealEnvelope(); // step 3
17     }
18     virtual void stuffEnvelope();
19     virtual void lickEnvelope();
20     virtual void sealEnvelope();
21     // ...
22 };
```

# Constructors, Destructors, and Virtual Functions

- except in very rare cases, destructors in class hierarchy need to be virtual
- otherwise, invoking destructor through base-class pointer/reference would only destroy base-class part of object, leaving remainder of derived-class object untouched
- normally, bad idea to call virtual function inside constructor or destructor
- dynamic type of object changes during construction and changes again during destruction
- final overrider of virtual function will change depending where in hierarchy virtual function call is made
- when constructor/destructor being executed, object is of exactly that type, never type derived from it
- although semantics of virtual function calls during construction and destruction well defined, easy to write code where actual overrider not what expected (and might even be pure virtual)

# Problematic Code with Non-Virtual Destructor

```
1  class Base {
2  public:
3      Base() {}
4      ~Base() {} // non-virtual destructor
5      // ...
6  };
7
8  class Derived : public Base {
9  public:
10     Derived() : buffer_(new char[10'000]) {}
11     ~Derived() {delete[] buffer_;}
12     // ...
13 private:
14     char* buffer_;
15 };
16
17 void process(Base* bp) {
18     // ...
19     delete bp; // always invokes only Base::~~Base
20 }
21
22 int main() {
23     process(new Base);
24     process(new Derived); // leaks memory
25 }
```

# Corrected Code with Virtual Destructor

```
1  class Base {
2  public:
3      Base() {}
4      virtual ~Base() {} // virtual destructor
5      // ...
6  };
7
8  class Derived : public Base {
9  public:
10     Derived() : buffer_(new char[10'000]) {}
11     ~Derived() {delete[] buffer_;}
12     // ...
13 private:
14     char* buffer_;
15 };
16
17 void process(Base* bp) {
18     // ...
19     delete bp; // invokes destructor polymorphically
20 }
21
22 int main() {
23     process(new Base);
24     process(new Derived);
25 }
```

# Preventing Creation of Derived Classes

- in some situations, may want to prevent deriving from class
- language provides means for accomplishing this
- in class/struct declaration, after name of class can add keyword **final** to prevent deriving from class
- example:

```
class Widget final { /* ... */ };  
class Gadget : public Widget { /* ... */ };  
    // ERROR: cannot derive from Widget
```

- might want to prevent deriving from class with destructor that is not virtual
- preventing derivation can sometimes also facilitate better compiler optimization (e.g., via devirtualization)
- might want to prevent derivation so that objects can be copied safely without fear of slicing

# Covariant Return Type

- in some special cases, language allows relaxation of rule that type of overriding function  $f$  must be same as type of virtual function  $f$  overrides
- in particular, requirement that return type be same is relaxed
- return type of derived-class function is permitted to be type derived (directly or indirectly) from return type of base-class function
- this relaxation of return type more formally known as **covariant return type**
- *case of pointer return type*: if original return type  $B^*$ , return type of overriding function may be  $D^*$ , provided  $B$  is public base of  $D$  (i.e., may return pointer to more derived type)
- *case of reference return type*: if original return type  $B\&$  (or  $B\&\&$ ), return type of overriding function may be  $D\&$  (or  $D\&\&$ ), provided  $B$  is public base of  $D$  (i.e., may return reference to more derived type)
- covariant return type can sometimes be exploited in order to avoid need for type casts



# Covariant Return Type Example: Cloning

```
1  class Base {
2  public:
3      virtual Base* clone() const {
4          return new Base(*this);
5      }
6      // ...
7  };
8
9  class Derived : public Base {
10 public:
11     // use covariant return type
12     Derived* clone() const override {
13         return new Derived(*this);
14     }
15     // ...
16 };
17
18 int main() {
19     Derived* d = new Derived;
20     Derived* d2 = d->clone();
21     // OK: return type is Derived*
22     // without covariant return type, would need cast:
23     // Derived* d2 = static_cast<Derived*>(d->clone());
24 }
```

# Pure Virtual Functions

- sometimes desirable to require derived class to override virtual function
- **pure virtual function**: virtual function that must be overridden in every derived class
- to declare virtual function as pure, add “= 0” at end of declaration
- example:

```
class Widget {  
public:  
    virtual void doStuff() = 0; // pure virtual  
    // ...  
};
```

- pure virtual function can still be defined, although likely only useful in case of virtual destructor

- class with one or more pure virtual functions called **abstract class**
- cannot directly instantiate objects of abstract class (can only use them as base class objects)
- class that derives from abstract class need not override all of its pure virtual methods
- class that does not override all pure virtual methods of abstract base class will also be abstract
- most commonly, abstract classes have no state (i.e., data members) and used to provide interfaces, which can be inherited by other classes
- if class has no pure virtual functions and abstract class is desired, can make destructor pure virtual (but must provide definition of destructor since invoked by derived classes)

# Abstract Class Example

```
1  #include <cmath>
2
3  class Shape {
4  public:
5      virtual bool isPolygon() const = 0;
6      virtual float area() const = 0;
7      virtual ~Shape() {};
8  };
9
10 class Rectangle : public Shape {
11 public:
12     Rectangle(float w, float h) : w_(w), h_(h) {}
13     bool isPolygon() const override {return true;}
14     float area() const override {return w_ * h_;}
15 private:
16     float w_; // width of rectangle
17     float h_; // height of rectangle
18 };
19
20 class Circle : public Shape {
21 public:
22     Circle(float r) : r_(r) {}
23     float area() const override {return M_PI * r_ * r_;}
24     bool isPolygon() const override {return false;}
25 private:
26     float r_; // radius of circle
27 };
```

# Pure Virtual Destructor Example

```
1 class Abstract {  
2 public:  
3     virtual ~Abstract() = 0; // pure virtual destructor  
4     // ... (no other virtual functions)  
5 };  
6  
7 inline Abstract::~~Abstract()  
8     { /* possibly empty */ }
```

# The `dynamic_cast` Operator

- often need to upcast and downcast (as well as cast sideways) in inheritance hierarchy
- **`dynamic_cast`** can be used to safely perform type conversions on pointers and references to classes
- syntax: **`dynamic_cast`**<T> (*expr*)
- types involved must be *polymorphic* (i.e., have at least one virtual function)
- inspects run-time information about types to determine whether cast can be safely performed
- if conversion is valid (i.e., *expr* can validly be cast to T), casts *expr* to type T and returns result
- if conversion is not valid, cast fails
- if *expr* is of pointer type, **`nullptr`** is returned upon failure
- if *expr* is of reference type, `std::bad_cast` exception is thrown upon failure (where exceptions are discussed later)

# dynamic\_cast Example

```
1  #include <cassert>
2
3  class Base {
4  public:
5      virtual void doStuff() { /* ... */ };
6      // ...
7  };
8
9  class Derived1 : public Base { /* ... */ };
10 class Derived2 : public Base { /* ... */ };
11
12 bool isDerived1(Base& b) {
13     return dynamic_cast<Derived1*>(&b) != nullptr;
14 }
15
16 int main() {
17     Base b;
18     Derived1 d1;
19     Derived2 d2;
20     assert(isDerived1(b) == false);
21     assert(isDerived1(d1) == true);
22     assert(isDerived1(d2) == false);
23 }
```

# Cost of Run-Time Polymorphism

- typically, run-time polymorphism does not come without run-time cost in terms of both time and memory
- in some contexts, cost can be significant
- typically, virtual functions implemented using virtual function table
- each polymorphic class has virtual function table containing pointers to all virtual functions for class
- each polymorphic class object has pointer to virtual function table
- memory cost to store virtual function table and pointer to table in each polymorphic object
- in most cases, impossible for compiler to inline virtual function calls since function to be called cannot be known until run time
- each virtual function call is made through pointer, which adds overhead



# Curiously-Recurring Template Pattern (CRTP)

- when derived type known at compile time, may want behavior similar to virtual functions but without run-time cost (by performing binding at compile time instead of run time)
- can be achieved with technique known as **curiously-recurring template pattern (CRTP)**
- class `Derived` derives from class template instantiation using `Derived` itself as template argument
- example:

```
template <class Derived>
class Base {
    // ...
};

class Derived : public Base<Derived> {
    // ...
};
```

# CRTP Example: Static Polymorphism

```
1  #include <iostream>
2
3  template <class Derived>
4  class Base {
5  public:
6      void interface() {
7          std::cout << "Base::interface called\n";
8          static_cast<Derived*>(this)->implementation();
9      }
10     // ...
11 };
12
13 class Derived : public Base<Derived> {
14 public:
15     void implementation() {
16         std::cout << "Derived::implementation called\n";
17     }
18     // ...
19 };
20
21 int main() {
22     Derived d;
23     d.interface();
24     // calls Base::interface which, in turn, calls
25     // Derived::implementation
26     // no virtual function call, however
27
28 }
```

# CRTP Example: Static Polymorphism

```
1  class TreeNode {
2  public:
3      enum Kind {RED, BLACK}; // kinds of nodes
4      TreeNode *left(); // get left child node
5      TreeNode *right(); // get right child node
6      Kind kind(); // get kind of node
7      // ...
8  };
9
10 template <class Derived>
11 class GenericVisitor {
12 public:
13     void visit_preorder(TreeNode* node) {
14         if (node) {
15             process_node(node);
16             visit_preorder(node->left());
17             visit_preorder(node->right());
18         }
19     }
20     void visit_inorder(TreeNode* node) { /* ... */ }
21     void visit_postorder(TreeNode* node) { /* ... */ }
22     void process_red_node(TreeNode* node) { /* ... */ };
23     void process_black_node(TreeNode* node) { /* ... */ };
24 private:
25     Derived& derived() {return *static_cast<Derived*>(this);}
26     void process_node(TreeNode* node) {
27         if (node->kind() == TreeNode::RED) {
28             derived().process_red_node(node);
29         } else {
30             derived().process_black_node(node);
31         }
32     }
33 };
34
35 class SpecialVisitor : public GenericVisitor<SpecialVisitor> {
36 public:
37     void process_red_node(TreeNode* node) { /* ... */ }
38 };
39
40 int main() {SpecialVisitor v;}
```

# CRTP Example: Comparisons

```
1  #include <cassert>
2
3  template<class Derived>
4  struct Comparisons {
5      friend bool operator==(const Comparisons<Derived>& x,
6          const Comparisons<Derived>& y) {
7          const Derived& xr = static_cast<const Derived>(x);
8          const Derived& yr = static_cast<const Derived>(y);
9          return !(xr < yr) && !(yr < xr);
10     }
11     // operator!= and others
12 };
13
14 class Widget : public Comparisons<Widget> {
15 public:
16     Widget(bool b, int i) : b_(b), i_(i) {}
17     friend bool operator<(const Widget& x, const Widget& y)
18         {return x.i_ < y.i_;}
19 private:
20     bool b_;
21     int i_;
22 };
23
24 int main() {
25     Widget w1(true, 1);
26     Widget w2(false, 1);
27     assert(w1 == w2);
28 }
```

# CRTP Example: Object Counting

```
1  #include <iostream>
2  #include <cstdlib>
3
4  template <class T>
5  class Counter {
6  public:
7      Counter() {++count_;}
8      Counter(const Counter&) {++count_;}
9      ~Counter() {--count_;}
10     static std::size_t howMany() {return count_;}
11 private:
12     static std::size_t count_;
13 };
14
15 template <class T>
16 std::size_t Counter<T>::count_ = 0;
17
18 // inherit from Counter to count objects
19 class Widget: private Counter<Widget> {
20 public:
21     using Counter<Widget>::howMany;
22     // ...
23 };
24
25 int main() {
26     Widget w1; int c1 = Widget::howMany();
27     Widget w2, w3; int c2 = Widget::howMany();
28     std::cout << c1 << ' ' << c2 << '\n';
29 }
```

## Section 2.7.3

### **Multiple Inheritance and Virtual Inheritance**

# Multiple Inheritance

- language allows derived class to inherit from more than one base class
- **multiple inheritance (MI)**: deriving from more than one base class
- although multiple inheritance not best solution for most problems, does have some compelling use cases
- one compelling use case is for inheriting interfaces by deriving from abstract base classes with no data members
- when misused, multiple inheritance can lead to very convoluted code
- in multiple inheritance contexts, ambiguities in naming can arise
- for example, if class `Derived` inherits from classes `Base1` and `Base2`, each of which have member called `x`, name `x` can be ambiguous in some contexts
- scope resolution operator can be used to resolve ambiguous names

# Ambiguity Resolution Example

```
1  class Base1 {
2  public:
3      void func();
4      // ...
5  };
6
7  class Base2 {
8      void func();
9      // ...
10 };
11
12 class Derived : public Base1, public Base2 {
13 public:
14     // ...
15 };
16
17 int main() {
18     Derived d;
19     d.func(); // ERROR: ambiguous function call
20     d.Base1::func(); // OK: invokes Base1::func
21     d.Base2::func(); // OK: invokes Base2::func
22 }
```

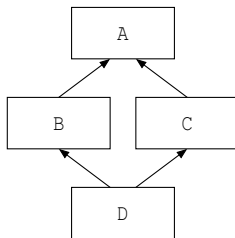


# Multiple Inheritance Example

```
1  class Input_stream {
2  public:
3      virtual ~Input_stream() {}
4      virtual int read_char() = 0;
5      virtual int read(char* buffer, int size) = 0;
6      virtual bool is_input_ready() const = 0;
7      // ... (all pure virtual, no data)
8  };
9
10 class Output_stream {
11 public:
12     virtual ~Output_stream() {}
13     virtual int write_char(char c) = 0;
14     virtual int write(char* buffer, int size) = 0;
15     virtual int flush_output() = 0;
16     // ... (all pure virtual, no data)
17 };
18
19 class Input_output_stream : public Input_stream,
20     public Output_stream {
21     // ...
22 };
```

# Dreaded Diamond Inheritance Pattern

- use of multiple inheritance can lead to so called dreaded diamond scenario
- **dreaded diamond** inheritance pattern has following form:



- class D will have *two* subobjects of class A, since class D (indirectly) inherits twice from class A
- situation like one above probably undesirable and often sign of poor design

# Dreaded Diamond Example

```
1  class Base {
2  public:
3      // ...
4  protected:
5      int data_;
6  };
7
8  class D1 : public Base { /* ... */ };
9
10 class D2 : public Base { /* ... */ };
11
12 class Join : public D1, public D2 {
13 public:
14     void method() {
15         data_ = 1; // ERROR: ambiguous
16         D1::data_ = 1; // OK: unambiguous
17     }
18 };
19
20 int main() {
21     Join* j = new Join();
22     Base* b;
23     b = j; // ERROR: ambiguous
24     b = static_cast<D1*>(j); // OK: unambiguous
25 }
```

- when using multiple inheritance, may want to ensure that only one instance of base-class object can appear in derived-class object
- **virtual base class**: base class that is only ever included once in derived class, even if derived from multiple times
- **virtual inheritance**: when derived class inherits from base class that is virtual
- virtual inheritance can be used to avoid situations like dreaded diamond pattern
- order of construction: virtual base classes constructed first in depth-first left-to-right traversal of graph of base classes, where left-to-right refers to order of appearance of base class names in class definition

# Avoiding Dreaded Diamond With Virtual Inheritance

```
1  class Base {
2  public:
3      // ...
4  protected:
5      int data_;
6  };
7
8  class D1 : public virtual Base { /* ... */ };
9
10 class D2 : public virtual Base { /* ... */ };
11
12 class Join : public D1, public D2 {
13 public:
14     void method() {
15         data_ = 1; // OK: unambiguous
16     }
17 };
18
19 int main() {
20     Join* j = new Join();
21     Base* b = j; // OK: unambiguous
22 }
```

## Section 2.7.4

### **References**

- 1 N. Meyers. The empty base C++ optimization.  
*Dr. Dobbs's Journal*, Aug. 1997.  
Available online at <http://www.cantrip.org/emptyopt.html>.
- 2 J. O. Coplien. Curiously recurring template patterns.  
*C++ Report*, pages 24–27, Feb. 1995.
- 3 S. Meyers. Counting objects in C++.  
*C++ User's Journal*, Apr. 1998.  
Available online at <http://www.drdobbs.com/cpp/counting-objects-in-c/184403484>.
- 4 A. Nasonov. Better encapsulation for the curiously recurring template pattern.  
*Overload*, 70:11–13, Dec. 2005.

## Section 2.8

# **C++ Standard Library**



- C++ standard library provides huge amount of functionality (orders of magnitude more than C standard library)
- uses `std` namespace (to avoid naming conflicts)
- well worth effort to familiarize yourself with all functionality in library in order to avoid writing code unnecessarily

- functionality can be grouped into following sublibraries:

- 1 language support library (e.g., exceptions, memory management)
- 2 diagnostics library (e.g., assertions, exceptions, error codes)
- 3 general utilities library (e.g., functors, date/time)
- 4 strings library (e.g., C++ and C-style strings)
- 5 localization library (e.g., date/time formatting and parsing, character classification)
- 6 containers library (e.g., sequence containers and associative containers)
- 7 iterators library (e.g., stream iterators)
- 8 algorithms library (e.g., searching, sorting, merging, set operations, heap operations, minimum/maximum)
- 9 numerics library (e.g., complex numbers, math functions)
- 10 input/output (I/O) library (e.g., streams)
- 11 regular expressions library (e.g., regular expression matching)
- 12 atomic operations library (e.g., atomic types, fences)
- 13 thread support library (e.g., threads, mutexes, condition variables, futures)

# Commonly-Used Header Files

## Language-Support Library

Header File	Description
<code>cstdlib</code>	run-time support, similar to <code>stdlib.h</code> from C (e.g., <code>exit</code> )
<code>limits</code>	properties of fundamental types (e.g., <code>numeric_limits</code> )
<code>exception</code>	exception handling support (e.g., <code>set_terminate</code> , <code>current_exception</code> )
<code>initializer_list</code>	<code>initializer_list</code> class template

## Diagnostics Library

Header File	Description
<code>cassert</code>	assertions (e.g., <code>assert</code> )
<code>stdexcept</code>	predefined exception types (e.g., <code>invalid_argument</code> , <code>domain_error</code> , <code>out_of_range</code> )

# Commonly-Used Header Files (Continued 1)

## General-Utilities Library

Header File	Description
<code>utility</code>	basic function and class templates (e.g., <code>swap</code> , <code>move</code> , <code>pair</code> )
<code>memory</code>	memory management (e.g., <code>unique_ptr</code> , <code>shared_ptr</code> , <code>addressof</code> )
<code>functional</code>	functors (e.g., <code>less</code> , <code>greater</code> )
<code>type_traits</code>	type traits (e.g., <code>is_integral</code> , <code>is_reference</code> )
<code>chrono</code>	clocks (e.g., <code>system_clock</code> , <code>steady_clock</code> , <code>high_resolution_clock</code> )

## Strings Library

Header File	Description
<code>string</code>	C++ string classes (e.g., <code>string</code> )
<code>cstring</code>	C-style strings, similar to <code>string.h</code> from C (e.g., <code>strlen</code> )
<code>cctype</code>	character classification, similar to <code>ctype.h</code> from C (e.g., <code>isdigit</code> , <code>isalpha</code> )

# Commonly-Used Header Files (Continued 2)

## Containers, Iterators, and Algorithms Libraries

Header File	Description
<code>array</code>	<code>array</code> class
<code>vector</code>	<code>vector</code> class
<code>deque</code>	<code>deque</code> class
<code>list</code>	<code>list</code> class
<code>set</code>	<b>set classes</b> (i.e., <code>set</code> , <code>multiset</code> )
<code>map</code>	<b>map classes</b> (i.e., <code>map</code> , <code>multimap</code> )
<code>unordered_set</code>	<b>unordered set classes</b> (i.e., <code>unordered_set</code> , <code>unordered_multiset</code> )
<code>unordered_map</code>	<b>unordered map classes</b> (i.e., <code>unordered_map</code> , <code>unordered_multimap</code> )
<code>iterator</code>	<b>iterators</b> (e.g., <code>reverse_iterator</code> , <code>back_inserter</code> )
<code>algorithm</code>	<b>algorithms</b> (e.g., <code>min</code> , <code>max</code> , <code>sort</code> )
<code>forward_list</code>	<code>forward_list</code> class

## Commonly-Used Header Files (Continued 3)

### Numerics Library

Header File	Description
<code>cmath</code>	C math library, similar to <code>math.h</code> from C (e.g., <code>sin</code> , <code>cos</code> )
<code>complex</code>	complex numbers (e.g., <code>complex</code> )
<code>numeric</code>	generalized numeric operations (e.g., <code>gcd</code> , <code>lcm</code> , <code>inner_product</code> )
<code>random</code>	random number generation (e.g., <code>uniform_int_distribution</code> , <code>uniform_real_distribution</code> , <code>normal_distribution</code> )

# Commonly-Used Header Files (Continued 4)

## I/O Library

Header File	Description
<code>iostream</code>	iostream objects (e.g., <code>cin</code> , <code>cout</code> , <code>cerr</code> )
<code>istream</code>	input streams (e.g., <code>istream</code> )
<code>ostream</code>	output streams (e.g., <code>ostream</code> )
<code>ios</code>	base classes and other declarations for streams (e.g., <code>ios_base</code> , <code>hex</code> , <code>fixed</code> )
<code>fstream</code>	file streams (e.g., <code>fstream</code> )
<code>sstream</code>	string streams (e.g., <code>stringstream</code> )
<code>iomanip</code>	manipulators (e.g., <code>setw</code> , <code>setprecision</code> )

## Regular-Expressions Library

Header File	Description
<code>regex</code>	regular expressions (e.g., <code>basic_regex</code> )

## Commonly-Used Header Files (Continued 5)

### Atomic-Operations and Thread-Support Libraries

Header File	Description
<code>atomic</code>	<b>atomics</b> (e.g., <code>atomic</code> )
<code>thread</code>	<b>threads</b> (e.g., <code>thread</code> )
<code>mutex</code>	<b>mutexes</b> (e.g., <code>mutex</code> , <code>recursive_mutex</code> , <code>timed_mutex</code> )
<code>condition_variable</code>	<b>condition variables</b> (e.g., <code>condition_variable</code> )
<code>future</code>	<b>futures</b> (e.g., <code>future</code> , <code>shared_future</code> , <code>promise</code> )



## Section 2.8.1

# **Containers, Iterators, and Algorithms**

# Standard Template Library (STL)

- large part of C++ standard library is collection of class/function templates known as standard template library (STL)
- STL comprised of three basic building blocks:
  - 1 containers
  - 2 iterators
  - 3 algorithms
- containers store elements for processing (e.g., vector)
- iterators allow access to elements for processing (which are often, but not necessarily, in containers)
- algorithms perform actual processing (e.g., search, sort)

- **container**: class that represents collection/sequence of elements
- usually container classes are template classes
- **sequence container**: collection in which every element has certain position that depends on time and place of insertion
- examples of sequence containers include:
  - `array` (fixed-size array)
  - `vector` (dynamic-size array)
  - `list` (doubly-linked list)
- **ordered/unordered associative container**: collection in which position of element in depends on its value or associated key and some predefined sorting/hashing criterion
- examples of associative containers include:
  - `set` (collection of unique keys, sorted by key)
  - `map` (collection of key-value pairs, sorted by key, keys are unique)

# Sequence Containers and Container Adapters

## Sequence Containers

Name	Description
<code>array</code>	fixed-size array
<code>vector</code>	dynamic-size array
<code>deque</code>	double-ended queue
<code>forward_list</code>	singly-linked list
<code>list</code>	doubly-linked list

## Container Adapters

Name	Description
<code>stack</code>	stack
<code>queue</code>	FIFO queue
<code>priority_queue</code>	priority queue

# Associative Containers

## Ordered Associative Containers

Name	Description
set	collection of unique keys, sorted by key
map	collection of key-value pairs, sorted by key, keys are unique
multiset	collection of keys, sorted by key, duplicate keys allowed
multimap	collection of key-value pairs, sorted by key, duplicate keys allowed

## Unordered Associative Containers

Name	Description
unordered_set	collection of unique keys, hashed by key
unordered_map	collection of key-value pairs, hashed by key, keys are unique
unordered_multiset	collection of keys, hashed by key, duplicate keys allowed)
unordered_multimap	collection of key-value pairs, hashed by key, duplicate keys allowed

# Typical Sequence Container Member Functions

- some member functions typically provided by sequence container classes listed below (where `T` denotes name of container class)

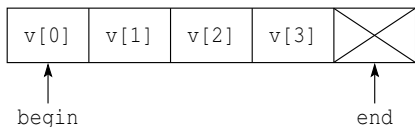
Function	Description
<code>T()</code>	create empty container (default constructor)
<code>T(const T&amp;)</code>	copy container (copy constructor)
<code>T(T&amp;&amp;)</code>	move container (move constructor)
<code>~T</code>	destroy container (including its elements)
<code>empty</code>	test if container empty
<code>size</code>	get number of elements in container
<code>push_back</code>	insert element at end of container
<code>clear</code>	remove all elements from container
<code>operator=</code>	assign all elements of one container to other
<code>operator[]</code>	access element in container

# Container Example

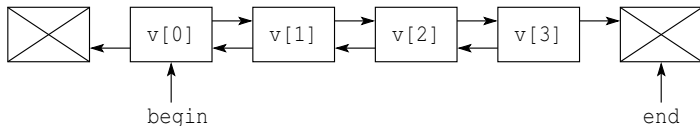
```
1  #include <iostream>
2  #include <vector>
3
4  int main() {
5      std::vector<int> values;
6
7      // append elements with values 0 to 9
8      for (int i = 0; i < 10; ++i) {
9          values.push_back(i);
10     }
11
12     // print each element followed by space
13     for (int i = 0; i < values.size(); ++i) {
14         std::cout << values[i] << ' ';
15     }
16     std::cout << '\n';
17 }
18
19 /* This program produces the following output:
20 0 1 2 3 4 5 6 7 8 9
21 */
```

# Motivation for Iterators

- different containers organize elements (of container) differently in memory
- want uniform manner in which to access elements in any arbitrary container
- organization of elements in array/vector container:



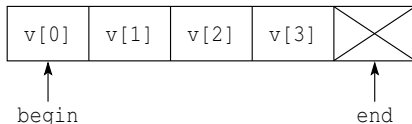
- organization of elements in doubly-linked list container:





# Motivation for Iterators (Continued)

- consider array/vector container with **int** elements:



- suppose we want to set all elements in container to zero
- we could use code like:

```
// int* begin; int* end;  
for (int* iter = begin; iter != end; ++iter)  
    *iter = 0;
```

- could we make similar-looking code work for more complicated organization like doubly-linked list?
- yes, create user-defined type that provides all pointer operations used above (e.g., dereference, increment, comparison, assignment)
- this leads to notion of iterator

- **iterator**: object that allows iteration over collection of elements, where elements are often (but not necessarily) in container
- iterators support many of same operations as pointers
- in some cases, iterator may actually be pointer; more frequently, iterator is user-defined type
- five different categories of iterators: 1) input, 2) output, 3) forward, 4) bidirectional, and 5) random access
- iterator has particular level of functionality, depending on category
- one of three possibilities of access order:
  - 1 forward (i.e., one direction only)
  - 2 forward and backward
  - 3 any order (i.e., random access)
- one of three possibilities in terms of read/write access:
  - 1 can only read referenced element (once or multiple times)
  - 2 can only write referenced element (once or multiple times)
  - 3 can read and write referenced element (once or multiple times)
- const and mutable (i.e., non-const) variants (i.e., read-only or read/write access, respectively)

# Abilities of Iterator Categories

Category	Ability	Providers
Input	Reads (once only) forward	istream (istream_iterator)
Output	Writes (once only) forward	ostream (ostream_iterator), inserter_iterator
Forward	Reads and writes forward	forward_list, unordered_set, unordered_multiset, unordered_map, unordered_multimap
Bidirectional	Reads and writes forward and backward	list, set, multiset, map, multimap
Random access	Reads and writes with random access	(built-in) array, array, vector, deque, string

Expression	Effect
<code>T(a)</code>	copies iterator (copy constructor)
<code>*a</code> <code>a-&gt;m</code>	dereference as rvalue (i.e., read only); cannot dereference at old position
<code>++a</code>	steps forward (returns new position)
<code>a++</code>	steps forward
<code>a == b</code>	test for equality
<code>a != b</code>	test for inequality

- not assignable (i.e., no assignment operator)

Expression	Effect
<code>T (a)</code>	copies iterator (copy constructor)
<code>*a</code> <code>a-&gt;m</code>	dereference as lvalue (i.e., write only); can only be dereferenced once; cannot dereference at old position
<code>++a</code>	steps forward (returns new position)
<code>a++</code>	steps forward (returns old position)

- not assignable (i.e., no assignment operator)
- no comparison operators (i.e., **operator==**, **operator!=**)

# Forward Iterators

Expression	Effect
<code>T()</code>	default constructor
<code>T(a)</code>	copy constructor
<code>a = b</code>	assignment
<code>*a</code> <code>a-&gt;m</code>	dereference
<code>++a</code>	steps forward (returns new position)
<code>a++</code>	steps forward (returns old position)
<code>a == b</code>	test for equality
<code>a != b</code>	test for inequality

- must ensure that valid to dereference iterator before doing so

- bidirectional iterators are forward iterators that provide additional functionality of being able to iterate backward over elements
- bidirectional iterators have all functionality of forward iterators as well as those listed in table below

Expression	Effect
<code>--a</code>	steps backward (returns new position)
<code>a--</code>	steps backward (returns old position)

# Random-Access Iterators

- random access iterators provide all functionality of bidirectional iterators as well as providing random access to elements
- random access iterators provide all functionality of bidirectional iterators as well as those listed in table below

Expression	Effect
<code>a[n]</code>	dereference element at index <code>n</code> (where <code>n</code> can be negative)
<code>a += n</code>	steps <code>n</code> elements forward (where <code>n</code> can be negative)
<code>a -= n</code>	steps <code>n</code> elements backward (where <code>n</code> can be negative)
<code>a + n</code>	iterator for <code>n</code> th next element
<code>n + a</code>	iterator for <code>n</code> th next element
<code>a - n</code>	iterator for <code>n</code> th previous element
<code>a - b</code>	distance from <code>a</code> to <code>b</code>
<code>a &lt; b</code>	test if <code>a</code> before <code>b</code>
<code>a &gt; b</code>	test if <code>a</code> after <code>b</code>
<code>a &lt;= b</code>	test if <code>a</code> not after <code>b</code>
<code>a &gt;= b</code>	test if <code>a</code> not before <code>b</code>

- pointers (built into language) are examples of random-access iterators



# Iterator Example

```
1  #include <iostream>
2  #include <vector>
3
4  int main() {
5      std::vector<int> values(10);
6
7      std::cout << "number of elements: " <<
8          (values.end() - values.begin()) << '\n';
9
10     // initialize elements of vector to 0, 1, 2, ...
11     for (std::vector<int>::iterator i = values.begin();
12         i != values.end(); ++i) {
13         *i = i - values.begin();
14     }
15
16     // print elements of vector
17     for (std::vector<int>::const_iterator i =
18         values.cbegin(); i != values.cend(); ++i) {
19         std::cout << ' ' << *i;
20     }
21     std::cout << '\n';
22 }
```

- do not dereference iterator unless it is known to validly reference some object
- some operations on container can *invalidate* some or all iterators referencing elements in container
- critically important to know *which operations invalidate* iterators in order to avoid using iterator that has been invalidated
- incrementing iterator *past end* of container or decrementing iterator *before beginning* of container results in undefined behavior
- input and output iterators can only be dereferenced *once* at each position

- **algorithm**: sequence of computations applied to some generic type
- algorithms use iterators to access elements involved in computation
- often pair of iterators used to specify *range* of elements on which to perform some computation
- what follows only provides brief summary of algorithms
- for more details on algorithms, see:
  - <http://www.cplusplus.com/reference/algorithm>
  - <http://en.cppreference.com/w/cpp/algorithm>

## Non-Modifying Sequence Operations

Name	Description
<code>all_of</code>	test if condition true for all elements in range
<code>any_of</code>	test if condition true for any element in range
<code>none_of</code>	test if condition true for no elements in range
<code>for_each</code>	apply function to range
<code>for_each_n</code>	apply function to first $n$ elements in sequence
<code>find</code>	find values in range
<code>find_if</code>	find element in range
<code>find_if_not</code>	find element in range (negated)
<code>find_end</code>	find last subsequence in range
<code>find_first_of</code>	find element from set in range
<code>adjacent_find</code>	find equal adjacent elements in range
<code>count</code>	count appearances of value in range
<code>count_if</code>	count number of elements in range satisfying condition
<code>mismatch</code>	get first position where two ranges differ
<code>equal</code>	test whether elements in two ranges differ
<code>search</code>	find subsequence in range
<code>search_n</code>	find succession of equal values in range

# Functions (Continued 1)

## Modifying Sequence Operations

Name	Description
<code>copy</code>	copy range of elements
<code>copy_if</code>	copy certain elements of range
<code>copy_n</code>	copy $n$ elements
<code>copy_backward</code>	copy range of elements backwards
<code>move</code>	move range of elements
<code>move_backward</code>	move range of elements backwards
<code>swap</code>	exchange values of two objects (in <code>utility</code> header)
<code>swap_ranges</code>	exchange values of two ranges
<code>iter_swap</code>	exchange values of objects referenced by two iterators
<code>transform</code>	apply function to range
<code>replace</code>	replace value in range
<code>replace_if</code>	replace values in range
<code>replace_copy</code>	copy range replacing value
<code>replace_copy_if</code>	copy range replacing value
<code>sample</code>	selects $n$ random elements from sequence

# Functions (Continued 2)

## Modifying Sequence Operations (Continued)

Name	Description
<code>fill</code>	fill range with value
<code>fill_n</code>	fill sequence with value
<code>generate</code>	generate values for range with function
<code>generate_n</code>	generate values for sequence with function
<code>remove</code>	remove value from range (by shifting elements)
<code>remove_if</code>	remove elements from range (by shifting elements)
<code>remove_copy</code>	copy range removing value
<code>remove_copy_if</code>	copy range removing values
<code>unique</code>	remove consecutive duplicates in range
<code>unique_copy</code>	copy range removing duplicates
<code>reverse</code>	reverse range
<code>reverse_copy</code>	copy range reversed
<code>rotate</code>	rotate elements in range
<code>rotate_copy</code>	copies and rotates elements in range
<code>shuffle</code>	randomly permute elements in range

## Functions (Continued 3)

### Partition Operations

Name	Description
<code>is_partitioned</code>	test if range is partitioned by predicate
<code>partition</code>	partition range in two
<code>partition_copy</code>	copies range partition in two
<code>stable_partition</code>	partition range in two (stable ordering)
<code>partition_point</code>	get partition point

### Sorting

Name	Description
<code>is_sorted</code>	test if range is sorted
<code>is_sorted_until</code>	find first unsorted element in range
<code>sort</code>	sort elements in range
<code>stable_sort</code>	sort elements in range, preserving order of equivalents
<code>partial_sort</code>	partially sort elements in range
<code>partial_sort_copy</code>	copy and partially sort range
<code>nth_element</code>	sort element in range

## Functions (Continued 4)

### Binary Search (operating on sorted ranges)

Name	Description
<code>lower_bound</code>	get iterator to lower bound
<code>upper_bound</code>	get iterator to upper bound
<code>equal_range</code>	get subrange of equal elements
<code>binary_search</code>	test if value exists in sorted range

### Set Operations (on sorted ranges)

Name	Description
<code>merge</code>	merge sorted ranges
<code>inplace_merge</code>	merge consecutive sorted ranges
<code>includes</code>	test whether sorted range includes another sorted range
<code>set_union</code>	union of two sorted ranges
<code>set_intersection</code>	intersection of two sorted ranges
<code>set_difference</code>	difference of two sorted ranges
<code>set_symmetric_difference</code>	symmetric difference of two sorted ranges



## Heap Operations

Name	Description
<code>is_heap</code>	test if range is heap
<code>is_heap_until</code>	first first element not in heap order
<code>push_heap</code>	push element into heap range
<code>pop_heap</code>	pop element from heap range
<code>make_heap</code>	make heap from range
<code>sort_heap</code>	sort elements of heap

## Functions (Continued 6)

Minimum/Maximum	
Name	Description
<code>min</code>	get minimum of given values
<code>max</code>	get maximum of given values
<code>minmax</code>	get minimum and maximum of given values
<code>min_element</code>	get smallest element in range
<code>max_element</code>	get largest element in range
<code>minmax_element</code>	get smallest and largest elements in range
<code>clamp</code>	clamp value between pair of boundary values
<code>lexicographic_compare</code>	lexicographic less-than comparison
<code>is_permutation</code>	test if range permutation of another
<code>next_permutation</code>	transform range to next permutation
<code>prev_permutation</code>	transform range to previous permutation

# Functions (Continued 7)

## Numeric Operations

Name	Description
<code>iota</code>	fill range with successive values
<code>accumulate</code>	accumulate values in range
<code>adjacent_difference</code>	compute adjacent difference of range
<code>inner_product</code>	compute inner product of range
<code>partial_sum</code>	compute partial sums of range
<code>reduce</code>	similar to <code>accumulate</code> except out of order
<code>exclusive_scan</code>	similar to <code>partial_sum</code> , excludes <i>i</i> th input element from <i>i</i> th sum
<code>inclusive_scan</code>	similar to <code>partial_sum</code> , includes <i>i</i> th input element in <i>i</i> th sum
<code>transform_reduce</code>	applies functor, then reduces out of order
<code>transform_exclusive_scan</code>	applies functor then, calculates exclusive scan
<code>transform_inclusive_scan</code>	applies functor, then calculates inclusive scan

## Other Numeric Algorithms

Name	Description
gcd	compute greatest common divisor of two integers
lcm	compute least common multiple of two integers

## Functions for Uninitialized Storage

# Algorithms Example

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4  #include <random>
5
6  int main() {
7      std::vector<int> values;
8
9      int x;
10     while (std::cin >> x) {values.push_back(x);}
11
12     std::cout << "zero count: " << std::count(
13         values.begin(), values.end(), 0) << '\n';
14
15     std::default_random_engine engine;
16     std::shuffle(values.begin(), values.end(), engine);
17     std::cout << "random order:";
18     for (auto i : values) {std::cout << ' ' << i;}
19     std::cout << '\n';
20
21     std::sort(values.begin(), values.end());
22     std::cout << "sorted order:";
23     for (auto i : values) {std::cout << ' ' << i;}
24     std::cout << '\n';
25 }
```

# Prelude to Functor Example

- consider `std::transform` function template:

```
template <class InputIterator, class OutputIterator,  
         class UnaryOperator>  
OutputIterator transform(InputIterator first,  
                        InputIterator last, OutputIterator result,  
                        UnaryOperator op);
```

- applies `op` to each element in range `[first,last)` and stores each returned value in range beginning at `result` (where ranges can overlap)
- `std::transform` might be written as:

```
template <class InputIterator, class OutputIterator,  
         class UnaryOperator>  
OutputIterator transform(InputIterator first,  
                        InputIterator last, OutputIterator result,  
                        UnaryOperator op) {  
    while (first != last) {  
        *result = op(*first);  
        ++first;  
        ++result;  
    }  
    return result;  
}
```

- `op` is entity that can be used with function call syntax (i.e., function or functor)

# Functor Example

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4
5  struct MultiplyBy { // Functor class
6      MultiplyBy(double factor) : factor_(factor) {}
7      double operator()(double x) const
8          {return factor_ * x;}
9  private:
10     double factor_; // multiplicative factor
11 };
12
13 int main() {
14     MultiplyBy mb(2.0);
15     std::vector v{1.0, 2.0, 3.0};
16     // v contains 1 2 3
17     std::transform(v.begin(), v.end(), v.begin(), mb);
18     // v contains 2 4 6
19     for (auto i : v) {std::cout << i << '\n';}
20 }
```

## Section 2.8.2

### **The `std::array` Class Template**



# The `std::array` Class Template

- one-dimensional array type, where size of array is fixed at compile time
- `array` declared as:

```
template <class T, std::size_t N>  
class array;
```

- T: type of elements in array
- N: number of elements in array
- what follows only intended to provide overview of `array`
- for additional details on `array`, see:
  - <http://en.cppreference.com/w/cpp/container/array>
  - <http://www.cplusplus.com/reference/stl/array>

# Member Types

Member Type	Description
<code>value_type</code>	T (i.e., element type)
<code>size_type</code>	type used for measuring size (i.e., <code>std::size_t</code> )
<code>difference_type</code>	type used to measure distance (i.e., <code>std::ptrdiff_t</code> )
<code>reference</code>	<code>value_type&amp;</code>
<code>const_reference</code>	<b>const</b> <code>value_type&amp;</code>
<code>pointer</code>	<code>value_type*</code>
<code>const_pointer</code>	<b>const</b> <code>value_type*</code>
<code>iterator</code>	<i>random-access</i> iterator type
<code>const_iterator</code>	const <i>random-access</i> iterator type
<code>reverse_iterator</code>	reverse iterator type (i.e., <code>reverse_iterator&lt;iterator&gt;</code> )
<code>const_reverse_iterator</code>	const reverse iterator type (i.e., <code>reverse_iterator&lt;const_iterator&gt;</code> )

# Member Functions

## Construction, Destruction, and Assignment

Member Name	Description
constructor	initializes array
destructor	destroys each element of array
<b>operator=</b>	overwrites every element of array with corresponding element of another array

## Iterators

Member Name	Description
begin	return iterator to beginning
end	return iterator to end
cbegin	return const iterator to beginning
cend	return const iterator to end
rbegin	return reverse iterator to beginning
rend	return reverse iterator to end
crbegin	return const reverse iterator to beginning
crend	return const reverse iterator to end

# Member Functions (Continued 1)

## Capacity

Member Name	Description
<code>empty</code>	test if array is empty
<code>size</code>	return size
<code>max_size</code>	return maximum size

## Element Access

Member Name	Description
<b><code>operator[]</code></b>	access element (no bounds checking)
<code>at</code>	access element (with bounds checking)
<code>front</code>	access first element
<code>back</code>	access last element
<code>data</code>	return pointer to start of element data

## Modifiers

Member Name	Description
<code>fill</code>	fill container with specified value
<code>swap</code>	swap contents of two arrays

# array Example

```
1  #include <array>
2  #include <iostream>
3  #include <algorithm>
4  #include <experimental/iterator>
5
6  int main() {
7      std::array<int, 3> a1{3, 1, 2};
8      std::array<int, 3> a2;
9      a2.fill(42);
10     for (auto i : a2) {
11         std::cout << i << '\n';
12     }
13     a2 = a1;
14     std::sort(a1.begin(), a1.end());
15     std::copy(a1.begin(), a1.end(),
16         std::experimental::make_ostream_joiner(std::cout, ", "));
17     std::cout << '\n';
18     for(auto i = a2.begin(); i != a2.end(); ++i) {
19         std::cout << *i;
20         if (i != a2.end() - 1) {std::cout << ", ";}
21     }
22     std::cout << '\n';
23 }
```

# array Example

```
1  #include <array>
2  #include <iostream>
3  #include <algorithm>
4
5  int main() {
6      // Fixed-size array with 4 elements.
7      std::array<int, 4> a{2, 4, 3, 1};
8
9      // Print elements of array.
10     for (auto i = a.cbegin(); i != a.cend(); ++i) {
11         std::cout << ' ' << *i;
12     }
13     std::cout << '\n';
14
15     // Sort elements of array.
16     std::sort(a.begin(), a.end());
17
18     // Print elements of array.
19     for (auto i = a.cbegin(); i != a.cend(); ++i) {
20         std::cout << ' ' << *i;
21     }
22     std::cout << '\n';
23 }
```

## Section 2.8.3

### **The `std::vector` Class Template**

# The `std::vector` Class Template

- dynamically-sized one-dimensional array type, where type of array elements and storage allocator specified by template parameters

- `vector` declared as:

```
template <class T, class Allocator = allocator<T>>
class vector;
```

- T: type of elements in vector
- Allocator: type of object used to handle storage allocation (unless custom storage allocator needed, use default `allocator<T>`)
- what follows only intended to provide overview of `vector`
- for additional details on `vector`, see:
  - <http://www.cplusplus.com/reference/stl/vector>
  - <http://en.cppreference.com/w/cpp/container/vector>



# Member Types

Member Type	Description
<code>value_type</code>	T (i.e., element type)
<code>allocator_type</code>	Allocator (i.e., allocator)
<code>size_type</code>	type used for measuring size (typically unsigned integral type)
<code>difference_type</code>	type used to measure distance (typically signed integral type)
<code>reference</code>	<code>value_type&amp;</code>
<code>const_reference</code>	<b>const</b> <code>value_type&amp;</code>
<code>pointer</code>	<code>allocator_traits&lt;Allocator&gt;::pointer</code>
<code>const_pointer</code>	<code>allocator_traits&lt;Allocator&gt;::const_pointer</code>
<code>iterator</code>	<i>random-access</i> iterator type
<code>const_iterator</code>	const <i>random-access</i> iterator type
<code>reverse_iterator</code>	reverse iterator type ( <code>reverse_iterator&lt;iterator&gt;</code> )
<code>const_reverse_iterator</code>	const reverse iterator type ( <code>reverse_iterator&lt;const_iterator&gt;</code> )

# Member Functions

## Construction, Destruction, and Assignment

Member Name	Description
constructor	construct vector (overloaded)
destructor	destroy vector
<b>operator=</b>	assign vector

## Iterators

Member Name	Description
begin	return iterator to beginning
end	return iterator to end
cbegin	return const iterator to beginning
cend	return const iterator to end
rbegin	return reverse iterator to beginning
rend	return reverse iterator to end
crbegin	return const reverse iterator to beginning
crend	return const reverse iterator to end

# Member Functions (Continued 1)

## Capacity

Member Name	Description
<code>empty</code>	test if vector is empty
<code>size</code>	return size
<code>max_size</code>	return maximum size
<code>capacity</code>	return allocated storage capacity
<code>reserve</code>	request change in capacity
<code>shrink_to_fit</code>	shrink to fit

## Element Access

Member Name	Description
<b><code>operator[]</code></b>	access element (no bounds checking)
<code>at</code>	access element (with bounds checking)
<code>front</code>	access first element
<code>back</code>	access last element
<code>data</code>	return pointer to start of element data

# Member Functions (Continued 2)

## Modifiers

Member Name	Description
<code>clear</code>	clear content
<code>assign</code>	assign vector content
<code>insert</code>	insert elements
<code>emplace</code>	insert element, constructing in place
<code>push_back</code>	add element at end
<code>emplace_back</code>	insert element at end, constructing in place
<code>erase</code>	erase elements
<code>pop_back</code>	delete last element
<code>resize</code>	change size
<code>swap</code>	swap content of two vectors

## Allocator

Member Name	Description
<code>get_allocator</code>	get allocator used by vector

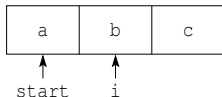
# Invalidation of References, Iterators, and Pointers

- **capacity**: total number of elements that vector could hold without requiring reallocation of memory
- any operation that causes reallocation of memory used to hold elements of vector invalidates *all* iterators, references, and pointers referring to elements in vector
- any operation that changes capacity of vector causes reallocation of memory
- any operation that adds or deletes elements can invalidate references, iterators, and pointers
- operations that can potentially invalidate references, iterators, and pointers to elements in vector include:

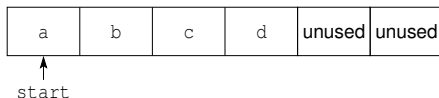
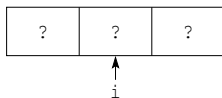
insert, erase, push\_back, pop\_back, emplace, emplace\_back,  
resize, reserve, **operator=**, assign, clear, shrink\_to\_fit, swap  
(past-the-end iterator only) [C++17 26.2.1/9]

# Iterator Invalidation Example

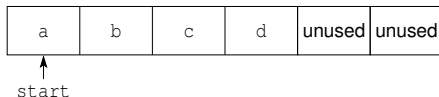
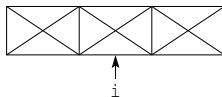
- `start` denotes pointer to first element in array holding elements of `vector`
- `i` is iterator for `vector` (e.g., `vector<T>::const_iterator` or `vector<T>::iterator`)
- initial vector has three elements and capacity of three



- `push_back(d)` invoked
- new larger array is allocated (say, twice size of original); elements in old array moved/copied to new array; then new element added



- elements in old array destroyed and memory for old array deallocated; iterator `i` is now **invalid**:



# vector Example: Constructors

```
1  std::vector<double> v0;  
2      // empty vector  
3  
4  std::vector<double> v1(10);  
5      // vector with 10 elements, each initialized to 0.0  
6      // (effectively via value initialization)  
7  
8  std::vector<double> v2(10, 5.0);  
9      // vector with 10 elements, each initialized to 5.0  
10  
11 std::vector<int> v3{1, 2, 3};  
12     // vector with 3 elements: 1, 2, 3  
13     // std::initializer_list (note brace brackets)
```

# vector Example: Iterators

```
1  #include <iostream>
2  #include <vector>
3
4  int main() {
5      std::vector v{0, 1, 2, 3};
6      for (auto& i : v) {++i;}
7      for (auto i : v) {
8          std::cout << ' ' << i;
9      }
10     std::cout << '\n';
11     for (auto i = v.begin(); i != v.end(); ++i) {
12         --(*i);
13     }
14     for (auto i = v.cbegin(); i != v.cend(); ++i) {
15         std::cout << ' ' << *i;
16     }
17     std::cout << '\n';
18     for (auto i = v.crbegin(); i != v.crend(); ++i) {
19         std::cout << ' ' << *i;
20     }
21     std::cout << '\n';
22 }
```

## ■ program output:

```
1 2 3 4
0 1 2 3
3 2 1 0
```



# vector Example

```
1  #include <iostream>
2  #include <vector>
3
4  int main() {
5      std::vector<double> values;
6      // ...
7
8      // Erase all elements and then read elements from
9      // standard input.
10     values.clear();
11     double x;
12     while (std::cin >> x) {
13         values.push_back(x);
14     }
15     std::cout << "number of values read: " <<
16         values.size() << '\n';
17
18     // Loop over all elements and print the number of
19     // negative elements found.
20     int count = 0;
21     for (auto i = values.cbegin(); i != values.cend(); ++i) {
22         if (*i < 0.0) {
23             ++count;
24         }
25     }
26     std::cout << "number of negative values: " << count <<
27         '\n';
28 }
```

# vector Example: Emplace

```
1  #include <iostream>
2  #include <vector>
3
4  int main() {
5      std::vector<std::vector<int>> v{{1, 2, 3}, {4, 5, 6}};
6      v.emplace_back(10, 0);
7      // The above use of emplace_back is more efficient than:
8      // v.push_back(std::vector<int>(10, 0));
9      for (const auto& i : v) {
10         for (const auto& j : i) {
11             std::cout << ' ' << j;
12         }
13         std::cout << '\n';
14     }
15 }
```

## ■ program output:

```
1 2 3
4 5 6
0 0 0 0 0 0 0 0 0 0
```

## Section 2.8.4

### **The `std::basic_string` Class Template**

# The `std::basic_string` Class Template

- character string type, parameterized on character type, character traits, and storage allocator
- `basic_string` declared as:

```
template <class CharT,  
          class Traits = char_traits<CharT>,  
          class Allocator = allocator<CharT>>  
          class basic_string;
```
- `CharT`: type of characters in string
- `Traits`: class that describes certain properties of `CharT` (normally, use default)
- `Allocator`: type of object used to handle storage allocation (unless custom storage allocator needed, use default)
- `string` is simply abbreviation for `basic_string<char>`
- what follows is only intended to provide overview of `basic_string` template class (and `string` class)
- for more details on `basic_string`, see:
  - [http://www.cplusplus.com/reference/string/basic\\_string](http://www.cplusplus.com/reference/string/basic_string)
  - [http://en.cppreference.com/w/cpp/string/basic\\_string](http://en.cppreference.com/w/cpp/string/basic_string)

# Member Types

Member Type	Description
<code>traits_type</code>	Traits (i.e., character traits)
<code>value_type</code>	<code>Traits::char_type</code> (i.e., character type)
<code>allocator_type</code>	Allocator
<code>size_type</code>	<code>allocator_traits&lt;Allocator&gt;::size_type</code>
<code>difference_type</code>	<code>allocator_traits&lt;Allocator&gt;::difference_type</code>
<code>reference</code>	<code>value_type&amp;</code>
<code>const_reference</code>	<b>const</b> <code>value_type&amp;</code>
<code>pointer</code>	<code>allocator_traits&lt;Allocator&gt;::pointer</code>
<code>const_pointer</code>	<code>allocator_traits&lt;Allocator&gt;::const_pointer</code>
<code>iterator</code>	<i>random-access</i> iterator type
<code>const_iterator</code>	const <i>random-access</i> iterator type
<code>reverse_iterator</code>	reverse iterator type ( <code>reverse_iterator&lt;iterator&gt;</code> )
<code>const_reverse_iterator</code>	const reverse iterator type ( <code>reverse_iterator&lt;const_iterator&gt;</code> )

## Construction, Destruction, and Assignment

Member Name	Description
constructor	construct
destructor	destroy
<b>operator=</b>	assign

## Iterators

Member Name	Description
begin	return iterator to beginning
end	return iterator to end
cbegin	return const iterator to beginning
cend	return const iterator to end
rbegin	return reverse iterator to reverse beginning
rend	return reverse iterator to reverse end
crbegin	return const reverse iterator to reverse beginning
crend	return const reverse iterator to reverse end

# Member Functions (Continued 1)

## Capacity

Member Name	Description
<code>empty</code>	test if string empty
<code>size</code>	get length of string
<code>length</code>	same as <code>size</code>
<code>max_size</code>	get maximum size of string
<code>capacity</code>	get size of allocated storage
<code>reserve</code>	change capacity
<code>shrink_to_fit</code>	shrink to fit

## Element Access

Member Name	Description
<b><code>operator[]</code></b>	access character in string (no bounds checking)
<code>at</code>	access character in string (with bounds checking)
<code>front</code>	access first character in string
<code>back</code>	access last character in string

## Member Functions (Continued 2)

### Operations

Member Name	Description
<code>clear</code>	clear string
<code>assign</code>	assign content to string
<code>insert</code>	insert into string
<code>push_back</code>	append character to string
<b><code>operator+=</code></b>	append to string
<code>append</code>	append to string
<code>erase</code>	erase characters from string
<code>pop_back</code>	delete last character from string
<code>replace</code>	replace part of string
<code>resize</code>	resize string
<code>swap</code>	swap contents with another string



# Member Functions (Continued 3)

## Operations (Continued)

Member Name	Description
<code>c_str</code>	get nonmodifiable C-string equivalent
<code>data</code>	obtain pointer to first character of string
<code>copy</code>	copy sequence of characters from string
<code>substr</code>	generate substring
<code>compare</code>	compare strings

## Search

Member Name	Description
<code>find</code>	find first occurrence of content in string
<code>rfind</code>	find last occurrence of content in string
<code>find_first_of</code>	find first occurrence of characters in string
<code>find_first_not_of</code>	find first absence of characters in string
<code>find_last_of</code>	find last occurrence of characters in string
<code>find_last_not_of</code>	find last absence of characters in string

## Member Functions (Continued 4)

### Allocator

Member Name	Description
<code>get_allocator</code>	<code>get allocator</code>

## Numeric Conversions

Name	Description
<code>stoi</code>	convert string to <b>int</b>
<code>stol</code>	convert string to <b>long</b>
<code>stoll</code>	convert string to <b>long long</b>
<code>stoul</code>	convert string to <b>unsigned long</b>
<code>stoull</code>	convert string to <b>unsigned long long</b>
<code>stof</code>	convert string to <b>float</b>
<code>stod</code>	convert string to <b>double</b>
<code>stold</code>	convert string to <b>long double</b>
<code>to_string</code>	convert integral or floating-point value to <code>string</code>
<code>to_wstring</code>	convert integral or floating-point value to <code>wstring</code>

# string Example

```
1  #include <iostream>
2  #include <string>
3
4  int main() {
5      std::string s;
6      if (!(std::cin >> s)) {
7          s.clear();
8      }
9      std::cout << "string: " << s << '\n';
10     std::cout << "length: " << s.size() << '\n';
11     std::string b;
12     for (auto i = s.cbegin(); i != s.crend(); ++i) {
13         b.push_back(*i);
14     }
15     std::cout << "backwards: " << b << '\n';
16
17     std::string msg = "Hello";
18     msg += ", World!"; // append ", World!"
19     std::cout << msg << '\n';
20
21     const char* cstr = s.c_str();
22     std::cout << "C-style string: " << cstr << '\n';
23 }
```

# Numeric/String Conversion Example

```
1  #include <iostream>
2  #include <string>
3
4  int main() {
5      double x = 42.24;
6      // Convert double to string.
7      std::string s = std::to_string(x);
8      std::cout << s << '\n';
9
10     s = "3.14";
11     // Convert string to double.
12     x = std::stod(s);
13     std::cout << x << '\n';
14
15 }
```

## Section 2.8.5

### **Other Container Classes**

# The `std::pair` Class Template

- collection of two heterogeneous objects

- `pair` declared as:

```
template <class T1, class T2>  
struct pair;
```

- T1: type of first element in `pair`
- T2: type of second element in `pair`
- first and second elements accessible via data members `first` and `second`, respectively
- elements of `pair` can also be accessed with `std::get` function template
- `pair` is effectively equivalent to `std::tuple` (to be discussed shortly) with two elements

# pair Example

```
1  #include <tuple>
2  #include <cassert>
3
4  int main() {
5      std::pair p(true, 42);
6      assert(p.first && p.second == 42);
7      assert(p.first == std::get<0>(p) &&
8             p.second == std::get<1>(p));
9      std::pair q(true, 42);
10     assert(p == q);
11     p = {false, 0};
12     assert(p != q);
13     p.swap(q);
14     auto [b, i] = p;
15     assert(b == true && i == 42);
16     assert(std::get<bool>(p) && std::get<0>(p));
17     assert(std::get<int>(p) == 42 &&
18            std::get<1>(p) == 42);
19 }
```



# The `std::tuple` Class Template

- fixed-size collection of heterogeneous values
- `tuple` is generalization of `std::pair`
- `tuple` declared as:

```
template <class... Ts>  
class tuple;
```

- `Ts`: types of elements that `tuple` holds (which may be empty)
- elements of `tuple` can be accessed with `std::get` function template

# tuple Example

```
1  #include <tuple>
2  #include <cassert>
3
4  int main() {
5      std::tuple t(true, 42, 'Z');
6      auto u = std::tuple(true, 42, 'Z');
7      assert(t == u);
8      assert(std::get<bool>(t) && std::get<0>(t));
9      assert(std::get<char>(t) == 'Z' && std::get<2>(t) == 'Z');
10     std::get<0>(t) = false;
11     assert(t != u);
12     std::tuple v(false, 0, '0');
13     u = std::tuple(true, 1, '1');
14     v.swap(u);
15     assert(std::get<0>(v));
16 }
```

# The `std::optional` Class Template

- simple container that manages optional value (i.e., value that may or may not be present)
- declaration:  

```
template <class T> class optional;
```
- T is type of optional value
- T cannot be reference type
- at any given point in time, object either contains value or does not
- object can be given value by initialization or assignment
- common use case is return value of function that can fail
- `std::bad_optional_access` exception indicates checked access to `optional` object that does not contain value
- optional value is required to be stored directly in `optional` object itself

[C++17 23.6.3/1]

## Construction, Destruction, and Assignment

Name	Description
constructor	constructs optional object
destructor	destroys optional object (and contained value)
<b>operator=</b>	assigns contents

## Observers

Name	Description
<b>operator-&gt;</b>	accesses contained value
<b>operator*</b>	accesses contained value
<b>operator bool</b>	tests if object contains value
has_value	tests if object contains value
value	returns contained value
value_or	returns contained value if available and specified default value otherwise

## Modifiers

Name	Description
swap	exchange contents
reset	clear any contained value
emplace	constructs contained value in place

## optional Example

```
1  #include <optional>
2  #include <string>
3  #include <exception>
4  #include <cassert>
5  #include <iostream>
6
7  int main() {
8      using namespace std::literals;
9      auto s = std::optional("Hello!"s);
10     assert(s && s.has_value());
11     assert(s.value() == "Hello!");
12     auto t = std::optional("Goodbye!"s);
13     s.swap(t);
14     assert(*s == "Goodbye!" && *t == "Hello!");
15     s.reset();
16     assert(!s && !s.has_value());
17     std::cout << s.value_or("Goodbye!") << '\n';
18     try {std::cout << s.value() << '\n';}
19     catch (const std::bad_optional_access&) {
20         std::cout << "caught exception\n";
21     }
22     s.emplace("Salut!");
23     std::cout << s.value() << '\n';
24 }
```

# Example: Return Type of Function That Can Fail

```
1  #include <optional>
2  #include <string>
3  #include <fstream>
4  #include <iostream>
5
6  std::optional<std::string> read_file(const char* file_name) {
7      std::ifstream in(file_name);
8      std::optional<std::string> result;
9      result.emplace(std::istreambuf_iterator<char>(in),
10         std::istreambuf_iterator<char>());
11      if (in.fail() && !in.eof()) {
12          result.reset();
13      }
14      return result;
15  }
16
17  int main(int argc, char** argv) {
18      if (argc <= 1) {return 1;}
19      auto s = read_file(argv[1]);
20      if (!s) {
21          std::cerr << "unable to read file\n";
22          return 1;
23      }
24      std::cout << *s;
25  }
```

# The `std::variant` Class Template

- simple container that corresponds to type-safe union
- can hold single value of one of set of allowable types
- declaration:  

```
template <class... Ts> class variant;
```
- `Ts` parameter pack containing all allowable types of value that can be stored in object
- container cannot hold references, arrays, or void
- can hold same type more than once and can hold differently cv-qualified versions of same type
- default initialized variant holds value of first alternative, which is default constructed
- `std::monostate` can be used as placeholder for empty type
- invalid accesses to value of `variant` object result in `std::bad_variant_access` exception being thrown



# variant Member Functions

## Construction, Destruction, and Assignment

Name	Description
constructor	constructs variant object
destructor	destroys variant object (and contained value)
<b>operator=</b>	assigns variant

## Observers

Name	Description
index	returns zero-based index of alternative held by variant
valueless_by_exception	tests if variant in invalid state

## Modifiers

Name	Description
emplace	constructs value in variant in place
swap	swaps value with another variant

# variant Example

```
1  #include <variant>
2  #include <cassert>
3  #include <iostream>
4
5  int main() {
6      std::variant<int, double> x;
7      std::variant<int, double> y;
8      x = 2;
9      assert(std::get<int>(x) == std::get<0>(x));
10     assert(!x.valueless_by_exception());
11     y = 0.5;
12     assert(std::get<double>(y) == std::get<1>(y));
13     std::cout << std::get<int>(x) << '\n';
14     std::cout << std::get<double>(y) << '\n';
15     try {std::cout << std::get<double>(x) << '\n';}
16     catch (const std::bad_variant_access&) {
17         std::cout << "bad variant access\n";
18     }
19 }
```

# The `std::any` Class

- type-safe container for single value of any type
- container may also hold no value
- declaration:  

```
class any;
```
- at any given time, object may or may not hold value
- non-member function `any_cast` provides type-safe access to contained object
- `std::bad_any_cast` exception thrown by value-returning forms of `any_cast` upon type mismatch

## Construction, Destruction, and Assignment

Name	Description
constructor	constructs any object
destructor	destroys any object
<b>operator=</b>	assigns any object

## Observers

Name	Description
has_value	tests if object holds value
type	returns typeid of contained value

## Modifiers

Name	Description
emplace	change contained object by constructing new value in place
reset	clear any contained object
swap	swaps contents of two any objects

# any Example

```
1  #include <any>
2  #include <cassert>
3  #include <string>
4  #include <iostream>
5
6  int main() {
7      std::any x{std::string("Hello")};
8      assert(x.has_value() && x.type() == typeid(std::string));
9      std::any y;
10     assert(!y.has_value());
11     x.swap(y);
12     assert(!x.has_value() && y.has_value());
13     x = y;
14     std::cout << std::any_cast<std::string>(x) << '\n';
15     y.reset();
16     assert(!y.has_value());
17     try {std::any_cast<int>(x);}
18     catch (const std::bad_any_cast&) {
19         std::cout << "any_cast failed\n";
20     }
21 }
```

## Section 2.8.6

### **Time Measurement**

- time measurement capabilities provided by part of general utilities library (of standard library)
- header file `chrono`
- identifiers in namespace `std::chrono`
- **time point**: specific point in time (measured relative to epoch)
- **duration**: time interval
- **clock**: measures time in terms of time points
- several clocks provided for measuring time
- what follows only intended to provide overview of chrono part of library
- for additional information on chrono part of library, see:
  - <http://www.cplusplus.com/reference/chrono>
  - <http://en.cppreference.com/w/cpp/chrono>

## Time Points and Intervals

Name	Description
duration	time interval
time_point	point in time

## Clocks

Name	Description
system_clock	system clock (which may be adjusted)
steady_clock	monotonic clock that ticks at constant rate
high_resolution_clock	clock with shortest tick period available



# std::chrono Example: Measuring Elapsed Time

```
1  #include <iostream>
2  #include <chrono>
3  #include <cmath>
4
5  double get_result() {
6      double sum = 0.0;
7      for (long i = 0L; i < 1000000L; ++i) {
8          sum += std::sin(i) * std::cos(i);
9      }
10     return sum;
11 }
12
13 int main() {
14     // Get the start time.
15     auto start_time =
16         std::chrono::high_resolution_clock::now();
17     // Do some computation.
18     double result = get_result();
19     // Get the end time.
20     auto end_time = std::chrono::high_resolution_clock::now();
21     // Compute elapsed time in seconds.
22     double elapsed_time = std::chrono::duration<double>(
23         end_time - start_time).count();
24     // Print result and elapsed time.
25     std::cout << "result " << result << '\n';
26     std::cout << "time (in seconds) " << elapsed_time << '\n';
27 }
```

# std::chrono Example: Determining Clock Resolution

```
1  #include <iostream>
2  #include <chrono>
3
4  // Get the granularity of a clock in seconds.
5  template <class C>
6  double granularity() {
7      return std::chrono::duration<double>(
8          typename C::duration(1)).count();
9  }
10
11 int main() {
12     std::cout << "system clock:\n" << "period "
13         << granularity<std::chrono::system_clock>() << '\n'
14         << "steady "
15         << std::chrono::system_clock::is_steady << '\n';
16     std::cout << "high resolution clock:\n" << "period "
17         << granularity<std::chrono::high_resolution_clock>()
18         << '\n' << "steady "
19         << std::chrono::high_resolution_clock::is_steady << '\n';
20     std::cout << "steady clock:\n" << "period "
21         << granularity<std::chrono::steady_clock>() << '\n'
22         << "steady "
23         << std::chrono::steady_clock::is_steady << '\n';
24 }
```

## Section 2.8.7

### **Miscellany**

# The `std::basic_string_view` Class Template

- `std::basic_string_view` class template represents constant contiguous sequence of **char**-like objects (i.e., read-only view of string)
- `basic_string_view` declared as:

```
template <class CharT,  
         class Traits = char_traits<CharT>>  
         class basic_string_view;
```

- `CharT`: type of characters in string
- `Traits`: class that describes certain properties of `CharT` (normally, use default)
- `string_view` is simply abbreviation for `basic_string_view<char>`
- for more details on `basic_string_view`, see:
  - ▢ [http://en.cppreference.com/w/cpp/string/basic\\_string\\_view](http://en.cppreference.com/w/cpp/string/basic_string_view)

# std::basic\_string\_view Example

```
1  #include <string_view>
2  #include <string>
3  #include <iostream>
4  #include <cassert>
5
6  void output(std::string_view s) {
7      std::cout << s << '\n';
8  }
9
10 int main() {
11     std::string_view hello("hello");
12     assert(!hello.empty());
13     std::string_view he = hello.substr(0, 2);
14     assert(he.size() == 2);
15     assert(he[0] == 'h' && he[1] == 'e');
16     assert(hello.find("ell") == 1);
17     assert(hello.rfind("l") == 3);
18
19     std::string goodbye("goodbye");
20     std::string_view bye(goodbye);
21     bye.remove_prefix(4);
22     std::cout << bye << '\n';
23     std::string_view good(goodbye);
24     good.remove_suffix(3);
25     std::cout << good << '\n';
26     assert(goodbye.substr(4, 3) == bye);
27     output(bye);
28 }
```

## Section 2.9

### Miscellany

# Name Lookup

- Since C++ name lookup rules are quite complicated, we only present a simplified (and therefore not fully correct) description of them here.
- **Qualified lookup.** If the name `A` is preceded by the scope-resolution operator, as in `::A` or `X::A`, then use qualified name lookup.
  - In the first case, look in the global namespace for `A`. In the second case, look up `X`, and then look inside it for `A`.
  - If `X` is a class and `A` is not a direct member, look in all of the direct bases of `X` (and then each of their bases). If `A` is found in more than one base, fail.
- **Argument-dependent lookup.** Otherwise, if the name is used as a function call, such as `A(X)`, use argument-dependent lookup.
  - Look for `A` in the namespace in which the type of `X` was declared, in the friends of `X`, and if `X` is a template instantiation, similarly for each of the arguments involved.
- **Unqualified lookup.** Start with unqualified lookup if argument-dependent lookup does not apply.
  - Start at the current scope and work outwards until the name is found.

# Argument-Dependent Lookup (ADL)

- argument-dependent lookup (ADL) applies to lookup of unqualified function name
- during ADL, other namespaces not considered during normal lookup may be searched
- in particular, namespace that declares each function argument type is included in search
- ADL also commonly referred to as Koenig lookup



# ADL Example

```
1  #include <iostream>
2
3  namespace N {
4      class C { /* ... */ };
5      void f(C x) {std::cout << "N::f\n";}
6      void g(int x) {std::cout << "N::g\n";}
7      void h(C x) {std::cout << "N::h\n";}
8  }
9
10 struct D {
11     struct E {};
12     static void p(E e) {std::cout << "D::p\n";};
13 };
14
15 void h(N::C x) {std::cout << "::h\n";}
16
17 int main() {
18     N::C x;
19     f(x);    // OK: calls N::f via ADL
20     N::f(x); // OK: calls N::f
21     g(42);   // ERROR: g not found
22     N::g(42); // OK: calls N::g
23     h(x);    // ERROR: ambiguous function call due to ADL
24     ::h(x);  // OK: calls ::h
25     N::h(x); // OK: calls N::h
26     D::E e;
27     p(e);    // ERROR: ADL only considers namespaces
28     D::p(e); // OK: calls D::p
29 }
```

# ADL Example

```
1  #include <iostream>
2
3  namespace N {
4      struct W {};
5      void f(W x) {std::cout << "N::f\n";}
6  }
7
8  struct C {
9      void f(N::W x) {std::cout << "C::f\n";}
10     void g() {
11         N::W x;
12         f(x); // calls C::f (not N::f)
13     }
14 };
15
16 int main() {
17     C c;
18     c.g();
19 }
```

# ADL Example

```
1  #include <iostream>
2  #include <string>
3
4  using namespace std::string_literals;
5
6  namespace N {
7      struct C {};
8      void f(int) {std::cout << "N::f\n";}
9      void g(C x) {std::cout << "N::g\n";}
10     void h(const std::string& x) {std::cout << "N::h\n";}
11     namespace M {
12         void f(int x) {std::cout << "N::M::f\n";}
13         // hides N::f
14         void g(int x) {std::cout << "N::M::g\n";}
15         // hides N::g
16         void h() {std::cout << "N::M::h\n";} // hides N::h
17         void u() {
18             N::C c;
19             f(42); // calls N::M::f (ADL looks nowhere)
20             g(c); // calls N::g via ADL (ADL looks in N)
21             h("hi"s); // ERROR: lookup finds N::M::h
22             // (ADL does not look in N)
23         }
24     }
25 }
26
27 int main() {N::M::u();}
```

# Swapping Values and ADL

- Consider two objects `x` and `y` of class type `T` whose values are to be swapped.
- If the class `T` provides its own `swap` function for reasons of efficiency, one would normally want to use it.
- In the absence of such a function, one would normally want to fall back on the use of `std::swap`.
- The above behavior can be achieved using code like the following:

```
using std::swap;  
swap(x, y);
```

- If the type `T` provides its own `swap` function, the name lookup on `swap` will yield this function through ADL.
- Otherwise, the name lookup will find `std::swap`.
- Thus, code like the above will result in a more efficient `swap` function being used if available, with the `std::swap` function used as a fallback.

## Part 3

### More C++

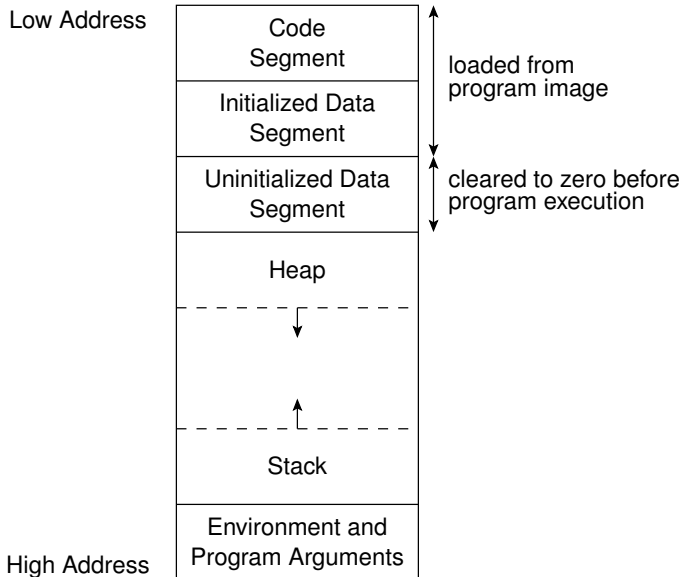
## Section 3.1

### Initialization

# Typical Memory Organization for Program

- memory organized into several regions: code segment, initialized data segment, uninitialized data segment, heap, and stack
- **code segment** (also known as text segment) contains machine code of compiled program; may be marked read only
- **initialized data segment** contains variables that are initialized to particular program-specified values upon program loading (i.e., prior to execution)
- **uninitialized data segment** (also known as BSS segment) contains variables that are not initialized to particular program-specified values upon program loading; typically cleared to zero when program loaded
- **heap** is where dynamic memory allocation takes place
- **stack** consists of stack frames used for local variables, function arguments, function return values, and caller return addresses
- stack and heap grow towards each other, with stack usually being at higher address in memory than heap

# Typical Memory Organization for Program: Diagram





# Storage Duration

- **storage duration**: how long memory for object exists
- four types of storage duration:
  - 1 automatic
  - 2 static
  - 3 dynamic
  - 4 thread
- **automatic storage duration**: storage allocated at start of enclosing code block and deallocated at end
- all local objects have automatic storage duration, except those declared with **static**, **extern**, or **thread\_local** qualifiers
- objects with automatic storage duration stored on stack
- **static storage duration**: storage is allocated at start of program and deallocated when program ends
- all objects declared at namespace scope (including global namespace) have static storage duration as well as those declared with **static** or **extern** qualifiers (e.g., static data members and static function-local variables)

## Storage Duration (Continued)

- objects with static storage duration stored in initialized and uninitialized data segments
- **dynamic storage duration**: storage is allocated and deallocated upon request using dynamic memory allocation functions (e.g., **operator new** and **operator delete**)
- objects with dynamic storage duration stored in heap
- **thread storage duration**: storage allocated when thread starts and deallocated when thread ends
- all objects declared as **thread\_local** have thread storage duration
- how objects with thread storage duration handled is very platform dependent (e.g., may involve additional segments such as .tbss and .tdata sections in ELF)

- initialization provides initial value to object at time of construction
- initialization in C++ somewhat complicated
- **initializer**: specifies initial value for object and appears in initializer section of declarator or new expression
- initializer may be one of following:

- 1 braced initializer list (i.e., possibly-empty comma-separated list of expressions and other braced initializer lists); for example:

```
std::vector<int> v{1, 2, 3};  
// initializer is {1, 2, 3}
```

- 2 comma-separated list of expressions and braced initializer lists in parentheses; for example:

```
std::vector<int> v(10, 42);  
// initializer is (10, 42)
```

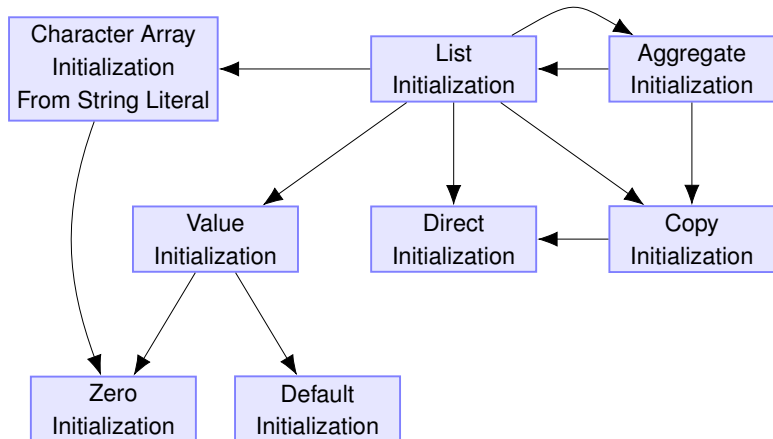
- 3 equals sign followed by expression; for example:

```
std::string s = "hello";  
// initializer is "=" plus literal "hello"
```

# Types of Initialization

- constant initialization (used to initialize objects of static storage duration from constant expressions)
- zero initialization (for initializing to zero)
- default initialization; for when no initializer is provided, such as:  
`int x; // where x is local object`
- value initialization; for when initializer is empty, such as:  
`Widget::Widget() : x() {} // where x is data member`
- direct initialization; for when initializer is explicit set of constructor arguments, such as:  
`int x(1); // where x is local object`
- copy initialization; for when initializing object from another object, such as:  
`int x = 1; // where x is local object`
- list initialization; for when initializer is braced initializer list, such as:  
`int x{1}; // where x is local object`
- aggregate initialization; for initializing aggregate type from braced initializer list (as part of list initialization), such as:  
`int a[] = {1, 2, 3}; // where a is local object`

# Initialization-Type Use Relationships



- constant initialization can use all other types of initialization (i.e., zero, default, value, direct, copy, list, and aggregate)

# Initialization Phases

- initialization performed in two phases:
  - 1 static initialization
  - 2 dynamic initialization
- **static initialization**: initialization that is conceptually performed when program loaded (i.e., before program begins execution)
- **dynamic initialization**: initialization that takes place at run time
- all static initialization happens before any dynamic initialization [\[C++17 6.6.2/2\]](#)
- for static initialization, only constant and zero initialization used directly (but constant initialization can result in other types of initialization being invoked indirectly)
- for dynamic initialization, all types of initialization other than constant initialization can be used

# How May I Initialize Thee? Let Me Count the Ways

```
1  #include <complex>
2  #include <string>
3  #include <array>
4  using Complex = std::complex<float>;
5
6  Complex gz{0, 1}; // gz is constant initialized (statically)
7  int gi; // gi is zero initialized (statically)
8  char buf[1024]; // buf is zero initialized (statically)
9
10 int main() {
11     int i; // i is default initialized (to indeterminate value)
12     Complex z1; // z1 is default initialized
13     Complex z2(); // function declaration
14     Complex z3{}; /* z3 is value initialized as part of
15        direct-list initialization */
16     Complex z4{1, -1}; /* z4 is direct initialized as part of
17        direct-list initialization */
18     Complex z5{1, -1}; // z5 is direct initialized
19     Complex z6 = {1, -1}; /* z6 is copy initialized as part of
20        copy-list initialization */
21     Complex z7 = Complex(1, -1); // z7 is copy initialized
22     static Complex u = Complex(); // u is constant initialized (statically)
23     z1 = {1, -1}; /* temporary object is direct initialized as part of
24        direct-list initialization */
25     z1 = {}; /* temporary object is value initialized as part of
26        direct-list initialization */
27     std::array<int, 3> a{1, 2, 3}; /* a is aggregate initialized as part
28        of direct-list initialization */
29     std::string s1{'h', 'i'}; // s1 is direct-list initialized
30 }
```

- must be careful about dynamic initialization order of namespace-scope objects with static storage duration (e.g., global variables)
- language makes no guarantees about order of initialization of such objects across translation units
- that is, if `x` and `y` are namespace-scope objects with static storage duration defined in *different* translation units, order in which `x` and `y` initialized is *arbitrary*
- in practice, initialization order will typically be determined by order in which linker processes corresponding object files
- language only guarantees order of initialization within translation unit, namely, initialization takes place in order of appearance of definitions
- that is, if `x` and `y` are namespace-scope objects with static storage duration defined in *same* translation unit, `x` and `y` initialized in *order of appearance* of their definitions
- initialization of namespace-scope object with static storage duration behaves as if performed prior to `main` function being called



- dynamic initialization of block-scope (i.e., local) object with static storage duration performed (in thread-safe manner) first time control passes through declaration of object [C++17 9.7/4]

# Example: Initialization Order Fiasco

util\_1.cpp

```
1  #include <vector>
2
3  std::vector<int> v{1, 2, 3, 4};
4  // invoked constructor is not constexpr; cannot use
5  // constant initialization; constructor invoked as
6  // part of dynamic initialization
```

util\_2.cpp

```
1  #include <vector>
2  extern std::vector<int> v;
3
4  std::vector<int> w{v[0], v[1]};
5  // arguments for invoked constructor not constant
6  // expressions; cannot use constant initialization;
7  // constructor invoked as part of dynamic initialization;
8  // construction of w can invoke undefined behavior
9  // since v might not yet have been constructed
```

main.cpp

```
1  int main() {
2      // ...
3  }
```

# Constant Initialization

- constant initialization is type of initialization that relates to (possibly non-const) object with static or thread storage duration being initialized with constant expression
- approximately speaking, constant initialization is performed: [C++17 6.6.2/2]
  - if object/reference with static or thread storage duration is initialized by constant expression
- effect of constant initialization is same as effect of corresponding initialization except must be performed before any dynamic initialization (i.e., as part of static initialization)
- this usually means, in practice, constant initialization performed at compile time and computed object stored as part of program image
- constant initialization provides means to initialize objects with static and thread storage duration that eliminates some potential problems due to data races and dependencies on initialization order

# Constant Initialization Example: Smart Pointers

```
1  #include <memory>
2  #include <string>
3
4  std::unique_ptr<std::string> p1; /* std::unique_ptr<std::string>
5     not of literal type, but...
6     constructor being invoked is constexpr; p1 is constant
7     initialized */
8
9  std::unique_ptr<std::string> p2(nullptr); /*
10     std::unique_ptr<std::string> not of literal type, but...
11     constructor being invoked is constexpr and argument to
12     constructor is constant expression; p2 is constant
13     initialized */
14
15  std::unique_ptr<std::string> q1; /* std::shared_ptr<std::string>
16     not of literal type, but...
17     constructor being invoked is constexpr; q1 is constant
18     initialized */
19
20  std::shared_ptr<std::string> q2(nullptr); /*
21     std::shared_ptr<std::string> not of literal type, but...
22     constructor being invoked is constexpr and argument to
23     constructor is constant expression; q2 is constant
24     initialized */
```

# Constant Initialization Example: Fiasco-Free Initialization

constant\_initialization\_2\_util.hpp

```
1 #include <complex>
2 extern std::complex<double> z0;
3 extern std::complex<double> z1;
```

util\_1.cpp

```
1 #include "constant_initialization_2_util.hpp"
2 constexpr double x{1};
3 constexpr double y{2};
4 std::complex<double> z0{x, y};
5 // invoked constructor is constexpr; all arguments to
6 // constructor are constant expressions; as part of static
7 // initialization, z0 is constant initialized to (1, 2)
```

util\_2.cpp

```
1 #include "constant_initialization_2_util.hpp"
2 std::complex<double> z1 = z0 - std::complex<double>{0, 2};
3 // as part of dynamic initialization, z1 is copy
4 // initialized to (1, 2) - (0, 2) = (1, 0); static
5 // initialization of z0 guaranteed to have been already
6 // performed
```

main.cpp

```
1 int main() {/* ... */}
```

# Zero Initialization

- **zero initialization**: initial value of object, including any padding, set to zero
- zero initialization is performed:
  - 1 for every named variable with static or thread storage duration that is not subject to constant initialization (i.e., initialized by constant expression) [C++17 6.6.2/2]
  - 2 when character array initialized with string literal that is too short to fill entire array, unfilled part of array is zero initialized [C++17 11.6.2/3]
  - 3 as part of value initialization in certain situations [C++17 11.6/8.2] [C++17 11.6/8.4] (see later slides on value initialization for details)
- every object of static storage duration is either constant or zero initialized at program startup before any other initialization takes place [C++17 11.6/10]

# Zero Initialization (Continued)

- zero initialization of object of type  $T$  has following behavior: [\[C++17 11.6/6\]](#)
  - if  $T$  is scalar type, object is initialized to value obtained by converting integer literal 0 to  $T$
  - if  $T$  is class type, each of following is zero initialized and any padding initialized to zero bits:
    - each non-static data member
    - each non-virtual base class subobject
    - if object is not base class subobject (i.e.,  $T$  is most-derived class), each virtual base class subobject anywhere in inheritance hierarchy
  - if  $T$  is union type, object's first non-static named data member is zero initialized and any padding is initialized to zero bits
  - if  $T$  is array type, each element is zero initialized
  - if  $T$  is reference type, nothing is done
- note that zero initialization not same as setting all bits of storage to zero, since some types may have zero value that is not represented as all zero bits (such as null pointer to member in case of some C++ language implementations)

# Zero Initialization Example

```
1  #include <string>
2
3  struct Point {int x; int y;};
4
5  static int ga[2]; /* ga is statically zero initialized
6    (to {0, 0}) 1 */
7  static int gb[2] = {1}; /* gb[1] is statically zero initialized
8    (to 0) as part of constant initializing gb to {1, 0} 2 */
9  char *gp; /* gp is statically zero initialized (to null
10    pointer) 3 */
11  std::string gs; /* gs is statically zero initialized
12    (to indeterminate value) and (later) dynamically
13    default initialized to empty string 4 */
14  int gi; // gi is statically zero initialized (to 0)
15
16  int main() {
17      char buf[4] = "hi"; /* buf[3] is zero initialized (to 0) as
18        part of initializing buf to {'h', 'i', '\0', 0} 5 */
19      static float f; /* f is statically zero initialized
20        (to 0.0f) 6 */
21      int i{}; /* i is zero initialized (to 0) as part of
22        list initialization 7 */
23      const Point& p = Point(); /* referenced object is
24        zero initialized (to {0, 0}) as part of
25        value initialization 8 */
26  }
```

Note: **7** indicates case *n* from earlier slide



- **default initialization**: object constructed with no initializer; for example:

```
int x; // where x is local object
```

- default initialization performed when:

- 1 variable with automatic, static, or thread storage duration declared with no initializer [\[C++17 11.6/12\]](#)
- 2 object created by new expression with no initializer [\[C++17 8.3.4/18\]](#)
- 3 non-static data member or base class not mentioned in constructor initializer list of invoked constructor [\[C++17 15.6.2/\(9.3\)\]](#)

- default initialization of object of type  $T$  has following behavior: [\[C++17 11.6/7\]](#)
  - if  $T$  is class type, default constructor invoked through overload resolution against empty argument list and called constructor provides initial value for new object
  - if  $T$  is array type, each element of array is default initialized
  - otherwise, nothing is done (which results in indeterminate value in case of object with automatic storage duration)

# Default Initialization Example

```
1  #include <string>
2
3  struct Widget {
4      Widget() {} /* w is default initialized to indeterminate
5         value 0 */
6      int w;
7  };
8
9  static std::string gs; /* gs is (statically) zero initialized
10     and then (dynamically) default initialized to
11     empty string 1 */
12
13 int main() {
14     std::string s; /* s is default initialized to
15        empty string 1 */
16     std::string* sp = new std::string; /* heap-allocated object is
17        default initialized to empty string 2 */
18     int i; /* i is default initialized to indeterminate
19        value 1 */
20     int* ip = new int; /* heap-allocated object is
21        default initialized to indeterminate value 2 */
22 }
```

- **value initialization**: object constructed with empty initializer; for example:

```
Widget::Widget() : x() {}  
    // where x is data member of Widget class
```

- value initialization performed when:

- 1 (unnamed) temporary created with initializer consisting of empty pair of parentheses or braces [C++17 11.6/11]
- 2 object created by new expression with initializer consisting of empty pair of parentheses or braces
- 3 non-static data member or base class initialized using member initializer list with empty pair of parentheses or braces [C++17 11.6/11]
- 4 named object declared with initializer consisting of empty pair of braces

# Value Initialization (Continued)

- value initialization of object of type  $T$  has following behavior: [\[C++17 11.6/8\]](#)
  - if  $T$  is class type with no default constructor or with user-provided or deleted default constructor, object is default initialized
  - if  $T$  is class type with default constructor that is neither user-provided nor deleted (i.e., class where default constructor is implicitly-defined or defaulted), object is zero initialized and then, if  $T$  has non-trivial default constructor, default initialized
  - if  $T$  is array type, each element of array is value initialized
  - otherwise, object is zero initialized
- if  $T$  is aggregate type and initializer is empty pair of braces, aggregate initialization is performed (as part of list initialization) instead of value initialization
- if  $T$  is class type with no default constructor but with constructor taking `std::initializer_list`, list initialization is performed

# Value Initialization Example

```
1  #include <string>
2
3  struct Point {int x; int y;};
4  struct Widget {
5      Widget() : x() {} // x is value initialized to 0 3
6      int x; int y{}; // y is value initialized to 0 4
7  };
8
9  int main() {
10     Point p; // p is default initialized to indeterminate value
11     Point q{}; /* q is aggregate initialized to {0, 0} as part of
12        list initialization */
13     Point* p1 = new Point(); /* heap-allocated object is
14        value initialized to {0, 0} 2 */
15     Point* p2 = new Point{}; /* heap-allocated object is aggregate
16        initialized to {0, 0} as part of list initialization */
17     std::string s{}; /* s is value initialized to empty string as part
18        of list initialization 1 */
19     const Point& pr = Point(); /* referenced object is value initialized
20        to {0, 0} 1 */
21     const Point& pr2 = Point{}; /* referenced object is aggregate
22        initialized to {0, 0} as part of list initialization */
23     const std::string& sr = std::string{}; /* referenced object is
24        value initialized to empty string as part of
25        list initialization 1 */
26     Widget w{}; /* w is value initialized to {0, 0} as part of
27        list initialization 4 */
28     int i{}; /* i is value initialized to 0 as part of
29        list initialization 4 */
30 }
```

- **direct initialization**: object initialized from explicit set of constructor arguments; for example:

```
int x(1); // where x is local object
```

- direct initialization performed when: [C++17 11.6/16]
  - 1 initializing with nonempty list of expressions in parentheses or braces [C++17 11.6/16]
  - 2 initializing object by new expression with nonempty initializer [C++17 8.3.4/(18.2)]
  - 3 initializing temporary by **static\_cast** expression [C++17 8.2.9/4]
  - 4 initializing temporary by functional-notation conversions [C++17 8.2.3/2]
  - 5 initializing non-static member or base object with constructor initializer list [C++17 15.6.2/7]
  - 6 initializing closure member from object captured by value in lambda expression [C++17 8.1.5.2/15]

# Direct Initialization (Continued)

- direct initialization of object of type `T` has following behavior:
  - if `T` is class type:
    - if initializer is prvalue expression whose type is same as `T` ignoring cv-qualification, initializer expression used to directly initialize object without materializing temporary (i.e., mandatory copy elision takes place) [\[C++17 11.6/\(17.6.1\)\]](#)
    - constructors of `T` examined, and best match obtained by overload resolution invoked to initialize object [\[C++17 11.6/\(17.6.2\)\]](#) [\[C++17 11.6/\(17.7\)\]](#)
  - if `T` is non-class type but source type is class type, conversion functions of source type and its base classes, if any, examined and best match selected by overload resolution used to convert initializer expression into object being initialized [\[C++17 11.6/\(17.6.3\)\]](#)
  - if `T` is `bool` and source type is `std::nullptr_t`, value of initialized object is **false** [\[C++17 7.14/1\]](#)
  - otherwise (neither `T` nor source type is class type), standard conversions used, if needed, to convert source value to same type as `T` and this (possibly converted) value used to set initial value of object [\[C++17 11.6/\(17.8\)\]](#)

# Direct Initialization Example

```
1  #include <string>
2  #include <vector>
3
4  struct Widget {
5      Widget() : s("hi") {} // s is direct initialized to "hi" ❶
6      explicit Widget(const std::string& s_) : s(s_) {}
7      std::string s;
8      std::string t{"bye"}; /* t is direct initialized to "bye" via
9         direct-list initialization ❷ */
10 };
11
12 int main() {
13     std::vector<int> u(2, 42); // u is direct initialized to {42, 42} ❸
14     std::vector<int> v(3); // v is direct initialized to {0, 0, 0} ❹
15     std::string s("bye"); // s is direct initialized to "bye" ❺
16     int i(1); // i is direct initialized to 1 ❻
17     int j{1}; /* j is direct initialized to 1 as part of
18        direct-list initialization */
19     double d = static_cast<double>(i); /* temporary object is
20        direct initialized to 1.0 ❼ */
21     std::string* sp = new std::string("hi"); /* heap-allocated object is
22        direct initialized to "hi" ❽ */
23     [s]() { return s.size(); }(); /* s data member in closure is
24        direct initialized to value of s in main ❹ */
25     Widget w = Widget("hi"); /* temporary object is direct initialized
26        to {"hi", "bye"} ❶ */
27 }
```



- **copy initialization**: initialize object from another object; for example:  
`int x = 1; // where x is local object`
- copy initialization performed in following situations: [\[C++17 11.6/15\]](#)
  - 1 when named variable of non-reference type declared with initializer that consist of equals sign followed by expression (including default member initializer) [\[C++17 11.6/15\]](#) [\[C++17 9.4\]](#)
  - 2 when passing argument to function by value [\[C++17 11.3.6/5\]](#) [\[C++17 11.6/15\]](#)
  - 3 when returning from function that returns by value [\[C++17 9.6.3/2\]](#)
  - 4 when throwing exception by value [\[C++17 18.1/3\]](#)
  - 5 when catching exception by value [\[C++17 18.3/15\]](#)
  - 6 as part of aggregate initialization, to initialize each element for which initializer provided (see later slides on aggregate initialization for details)  
[\[C++17 11.6.1/3\]](#)

- copy initialization of object of type  $T$  has following behavior:

[C++17 9.4][C++17 18.1][C++17 18.3][C++17 11.6.1][C++17 15.8]

- if  $T$  is class type and initializer is prvalue expression whose cv-unqualified type same as  $T$ , initializer expression itself (rather than temporary materialized therefrom) used to initialize object (i.e., mandatory copy elision takes place)
- if  $T$  is class type and cv-unqualified version of source type is  $T$  or class derived from  $T$ , non-explicit constructors of  $T$  examined and best match selected by overload resolution used to initialize object
- if  $T$  is class type and cv-unqualified version of source type is not  $T$  or derived from  $T$  or  $T$  is non-class type but source type is class type, user-defined conversion sequences that can convert from source type to  $T$  (or type derived from  $T$ ) examined and best match selected by overload resolution used to direct initialize object
- otherwise (i.e., if neither  $T$  nor source type is class type), standard conversions used, if necessary, to convert source value to cv-unqualified version of  $T$

# Copy Initialization Example

```
1  #include <string>
2  using namespace std::literals;
3
4  struct Widget {
5      std::string s = "hi"; // s is copy initialized to "hi" 1
6  };
7
8  std::string identity(std::string p) {
9      return p; // return value copy initialized from p 3
10 }
11
12 int main() {
13     std::string a[2] = {"hi", "bye"}; /*
14         as part of aggregate initialization:
15         a[0] is copy initialized to "hi" and
16         a[1] is copy initialized to "bye" 4 */
17     std::string s = "hello"s; // s is copy initialized to "hello" 1
18     std::string t = {3, 'A'}; /* t is copy initialized to "AAA" as
19         part of copy-list initialization 1 */
20     s = identity(s); // function parameter is copy initialized from s 2
21     try {
22         throw t; // exception object copy initialized from t 4
23     } catch (std::string s) {
24         // s is copy initialized from exception object 5
25     }
26     if (auto i = s.begin(); i != s.end()) {/* ... */}
27     // i is copy initialized from s.begin() 1
28 }
```

- **aggregate** is array or class with: [C++17 11.6.1/1]
  - no user-provided, explicit, or inherited constructors
  - no private or protected non-static data members
  - no virtual functions
  - no virtual, private, or protected base classes
- elements of aggregate are: [C++17 11.6.1/2]
  - for array: array elements in increasing subscript order
  - for class: direct base classes in declaration order, followed by direct non-static data members that are not members of anonymous union in declaration order
- example:

```
// aggregate type
struct Point {
    int x;
    int y;
};
```

# Aggregate Initialization

- **aggregate initialization**: initializes aggregate from braced initializer list (or string literal)
  - aggregate initialization is used:
    - when aggregate is initialized by braced initializer list (or string literal in case of initializing character array, since braces around string literal are implied)
- [C++17 11.6.1/3]
- approximately speaking, aggregate initialization has following behavior:
    - each direct public base, array element, or non-static class member in order of array subscript or appearance in class definition copy initialized from corresponding clause of initializer list
      - if initializer clause is expression, only non-narrowing implicit conversions allowed
      - if initializer clause is braced initializer list, base/element/member list initialized from clause
    - if number of initializer clauses less than number of members and bases, remaining members and bases initialized by default initializer if provided in class definition and otherwise by empty lists in accordance with usual list initialization rules (e.g., performs value initialization for non-class types and non-aggregate classes with default constructors)
    - if  $T$  is union type, only first non-static data member initialized

# Aggregate Initialization Example

```
1  struct Point {int x; int y;};
2
3  static Point gp{1, 2}; /* aggregate initialized to {1, 2} as
4     part of constant initialization */
5
6  int main() {
7     int a[3] = {1, 2, 3}; /* aggregate initialized to {1, 2, 3}
8        as part of copy-list initialization */
9     int b[3]{1, 2, 3}; /* aggregate initialized to {1, 2, 3} as
10        part of direct-list initialization */
11    int c[4]{1, 2}; /* aggregate initialized to {1, 2, 0, 0} as
12        part of direct-list initialization */
13    Point p{1, 2}; /* aggregate initialized to {1, 2} as
14        part of direct-list initialization */
15    Point q{1}; /* aggregate initialized to {1, 0} as
16        part of direct-list initialization */
17 }
```

# List Initialization

- **list initialization**: initialize object from braced initializer list
- if initializer is braced initializer list (possibly preceded by equals sign), list initialization always used (which may then, in turn, directly invoke other types of initialization, such as value, direct, copy, or aggregate initialization)
- list initialization can occur in both direct and copy initialization contexts
- **direct-list initialization**: list initialization used in direct-initialization context; for example:

```
int x{1}; // where x is local object
```

- direct-list initialization used: [C++17 11.6.4/1]
  - 1 to initialize named variable with braced initializer list
  - 2 to initialize (unnamed) temporary with braced initializer list
  - 3 to initialize object created by new expression that has braced initializer list as initializer
  - 4 for non-static data member initializer that does not use equals sign
  - 5 in constructor initializer list where braced initializer list used for initializer

# List Initialization (Continued 1)

- **copy-list initialization**: list initialization used in copy-initialization context; for example:

```
int x = {1}; // where x is local variable
```

- copy-list initialization used: [C++17 11.6.4/1]
  - 1 to initialize named variable where initializer is braced initializer list that follows equals sign
  - 2 to initialize function parameter for function call where braced initializer list used for corresponding function argument
  - 3 to initialize returned object for return statement where return expression is braced initializer list
  - 4 to initialize parameter of overloaded subscript operator in subscript expression where braced initializer list used
  - 5 to initialize parameter of overloaded assignment operator in assignment expression where right-hand side of assignment is braced initializer list
  - 6 to initialize parameter of constructor in functional cast expression where corresponding constructor argument is braced initializer list
  - 7 for non-static data member initializer that uses equals sign



## List Initialization (Continued 2)

- list initialization does not allow narrowing conversions
- direct-list initialization is allowed to use explicit constructors, whereas copy-list initialization is not
- approximately speaking, list initialization of object of aggregate type  $T$  has following behavior: [C++17 11.6.4/3]
  - if initializer list has single element of same or derived type (possibly cv-qualified), object initialized from element by copy initialization (for copy-list initialization) or direct initialization (for direct-list initialization) [C++17 11.6.4/(3.1)]
  - otherwise, if  $T$  is character array and initializer list has single element of appropriately-typed string literal, array initialized from string literal in usual manner [C++17 11.6.4/(3.2)]
  - otherwise, object is aggregate initialized [C++17 11.6.4/(3.3)]

# List Initialization (Continued 3)

- approximately speaking, list initialization of object of non-aggregate type  $T$  has following behavior: [C++17 11.6.4/3]
  - if  $T$  is class type with default constructor and braced initializer list empty, object is value initialized [C++17 11.6.4/(3.4)]
  - if  $T$  is specialization of `std::initializer_list`, object is direct initialized or copy initialized, depending on context, from braced initializer list [C++17 11.6.4/(3.5)]
  - if  $T$  is class type, constructors of  $T$  considered in two phases (first, using constructors that can be called with `std::initializer_list` as single argument; then using all constructors) [C++17 11.6.4/(3.6)]
  - if  $T$  is enumeration type, if braced initializer list has only one initializer (and some other constraints satisfied), enumeration initialized with result of converting initializer to enumeration's underlying type [C++17 11.6.4/(3.7)]
  - if  $T$  is not class type and braced initializer list has exactly one element and either  $T$  is not reference type or is reference type that is compatible with type of element, object is direct initialized (for direct-list initialization) or copy initialized (for copy-list initialization) [C++17 11.6.4/(3.8)]
  - if  $T$  is reference type that is not compatible with type of element, temporary of referenced type is list initialized and reference bound to temporary

[C++17 11.6.4/(3.9)]

# Direct-List Initialization Example

```
1  #include <initializer_list>
2  #include <vector>
3
4  struct Widget {
5      Widget() : w{1, 2, 3} {} // w is direct-list initialized 5
6      Widget(std::initializer_list<int> w_) : w{w_} {} /* w is
7          direct initialized as part of direct-list initialization 5 */
8      std::vector<int> v{3, 2, 1}; // v is direct-list initialized 4
9      std::vector<int> w;
10 };
11
12 int main() {
13     Widget w1{1, 2, 3}; // w1 is direct-list initialized 1
14     const Widget& w2 = Widget{1, 2, 3};
15     // temporary object is direct-list initialized 2
16     Widget* w3 = new Widget{1, 2, 3};
17     // heap-allocated object is direct-list initialized 3
18     for (auto&& i : {1, 2, 3}) {}
19     // temporary object is direct-list initialized 2
20 }
```

# Copy-List Initialization Example

```
1  #include <vector>
2  #include <tuple>
3  #include <initializer_list>
4
5  struct Widget {
6      Widget() : v({1, 2, 3}) {}
7      // constructor argument is copy-list initialized ❶
8      Widget(std::initializer_list<int> v_) : v{v_} {}
9      const int& operator[] (std::pair<int, int> i) const
10         {return i.first ? v[i.second] : w[i.second];}
11      std::vector<int> v;
12      std::vector<int> w = {3, 2, 1}; // w is copy-list initialized ❷
13  };
14
15  Widget func(Widget w) {
16      return {1, 2, 3}; // returned value is copy-list initialized ❸
17  }
18
19  int main() {
20      Widget w = {1, 2, 3}; // w is copy-list initialized ❹
21      w = {1, 2, 3}; // temporary object is copy-list initialized ❺
22      func({1, 2, 3}); // function argument is copy-list initialized ❻
23      Widget({1, 2, 3}); // constructor argument is copy-list initialized ❼
24      int i = w[{0, 1}];
25      // operator[] function parameter is copy-list initialized ❽
26  }
```

# Initialization of Character Array From String Literal

- special rule employed for initializing character array from string literal  
[C++17 11.6/17.3]
- array element type and character type of string literal must be compatible
- each character in string literal (including null-terminator) placed in order into successive array elements [C++17 11.6.2]
- if number of characters in initializer less than number of array elements, remaining array elements are zero initialized
- number of characters in initializer must not exceed array capacity
- special initialization rule invoked either directly or via list initialization

[C++17 11.6.4/(3.2)] [C++17 16.3.3.1.5/3]

# Character Array Initialization Example

```
1  int main() {
2      char a1[] = "hi";
3      // initialized to {'h', 'i', '\0'} from string literal
4      char a2[] = {"hi"}; /* initialized to {'h', 'i', '\0'} as
5                          part of copy-list initialization */
6      char a3[]{"hi"};
7      char a4[3>{"hi"};
8      /* each of a3 and a4 is initialized to {'h', 'i', '\0'}
9       as part of direct-list initialization */
10
11     char b1[4] = "hi";
12     char b2[4] = {"hi"};
13     char b3[4>{"hi"};
14     /* each of b1, b2, and b3 is initialized to
15      {'h', 'i', '\0', '\0'} */
16
17     char16_t c[] = u"hi"; /* initialized to
18      {u'h', u'i', u'\0'} from string literal */
19     char32_t d[] = U"hi"; /* initialized to
20      {U'h', U'i', U'\0'} from string literal */
21     wchar_t e[4]{L"hi"}; /* initialized to
22      {L'h', L'i', L'\0', L'\0'} as part of
23      direct-list initialization */
24 }
```

# Example: Aggregates and Brace Elision

```
1  struct Gadget {
2      int x;
3      int y;
4  };
5
6  struct Widget {
7      Gadget g;
8      int i;
9  };
10
11 int main() {
12     int x[2][2] = {1, 2, 3, 4};
13     // effectively initializer is {{1, 2}, {3, 4}}
14     Widget v = {1, 2, 3};
15     // effectively initializer is {{1, 2}, 3}
16     Widget w = {1, 2};
17     // effectively initializer is {{1, 2}}
18     // w initialized to {{1, 2}, 0}
19 }
```

## Example: Auto and Initialization

```
1  #include <initializer_list>
2
3  auto i1 = 42; // type of i1 deduced as int
4  auto i2(42); // type of i2 deduced as int
5  auto i3{42}; // type of i3 deduced as int
6  // auto i4{42, 42}; // ERROR: exactly one element required
7  auto i5 = {42};
8  // type of i5 deduced as std::initializer_list<int>
```



# Example: Braced Initializer Lists and Constructor Selection

```
1  #include <initializer_list>
2
3  struct Widget {
4      Widget();
5      Widget(std::initializer_list<int>);
6      Widget(int);
7  };
8
9  int main() {
10     Widget w{};
11         // invokes default constructor; for empty
12         // braced initializer list, default constructor
13         // preferred over std::initializer_list constructor
14     Widget v{42};
15         // invokes constructor taking std::initializer_list;
16         // for non-empty braced initializer list, constructor
17         // taking std::initializer_list preferred over those
18         // that do not
19 }
```

# Example: List Initialization

```
1  #include <map>
2  #include <vector>
3  #include <string>
4
5  std::map<int, std::string> m{
6      {42, "forty two"},
7      {0, "zero"}
8  }; // initialized to map with two elements
9
10 std::vector<std::string> v1{"hi", "bye"};
11 // initialized to vector with two elements
12
13 std::vector<std::string> v2{"hi", "bye"};
14 // ERROR: will try to initialize to vector with
15 // one element; invokes std::string constructor that
16 // takes two iterators as parameters; pointers
17 // to "hi" and "bye" passed as begin and end
18 // iterators; this results in undefined behavior
```

# Example: Initialization and Narrowing Conversions

```
1  struct Widget {
2      Widget(int i_) : i(i_) {}
3      int i;
4  };
5
6  int main() {
7      Widget v(42.0);
8          // OK: narrowing conversion allowed in
9          // direct initialization
10         // Widget w{42.0};
11         // ERROR: narrowing conversion not allowed in
12         // list initialization
13 }
```

## Example: Initializers and Braces/Parentheses

```
1  #include <vector>
2  #include <string>
3
4  std::vector<int> v1(3, 42);
5      // initialized to vector with elements 42, 42, 42
6  std::vector<int> v2{3, 42};
7      // initialized to vector with elements 3, 42
8
9  std::string s1(3, 'a');
10     // initialized to string consisting of 3 'a' characters
11  std::string s2{3, 'a'};
12     // initialized to string consisting of characters '\3', 'a'
```

## Example: `std::initializer_list` and Lifetime

```
1  #include <iostream>
2  #include <initializer_list>
3
4  auto f(int a, int b, int c) {
5      return std::initializer_list<int>{a, b, c};
6          // ERROR: initializer_list references elements in
7          // temporary array whose lifetime need not extend
8          // beyond lifetime of initializer_list;
9          // therefore, returned initializer_list
10         // likely references invalid data
11 }
12
13 int main() {
14     // nothing good likely to happen here
15     for (auto i : f(1, 2, 3)) {
16         std::cout << i << '\n';
17     }
18 }
```

## Section 3.1.1

### **References**

- 1 Nicolai Josuttis. The Nightmare of Initialization in C++. CppCon, Bellevue, WA, USA, Sept. 24, 2018. Available online at <https://youtu.be/7DTlWPgX6zs>.
- 2 Greg Falcon. Initialization, Shutdown, and constexpr. CppCon, Bellevue, WA, USA, Sept. 27, 2018. Available online at <https://youtu.be/6ZOygaUjzjQ>.
- 3 Timur Doumler. Initialization in Modern C++. Meeting C++, Berlin, Germany, Nov. 17, 2018. Available online at <https://youtu.be/ZfP4VAK21zc>.

## Section 3.2

# Temporary Objects



# Temporary Objects

- A **temporary object** is an unnamed object introduced by the compiler.
- Temporary objects may be used during: [\[C++17 15.2/1\]](#) [\[C++17 15.2/2\]](#)
  - evaluation of expressions
  - argument passing
  - function returns (that return by value)
  - reference initialization
- It is important to understand when temporary objects can be introduced, since the introduction of temporaries impacts performance.
- Evaluation of expression:

```
std::string s1("Hello ");  
std::string s2("World");  
std::string s;  
s = s1 + s2; // must create temporary  
    // std::string _tmp(s1 + s2);  
    // s = _tmp;
```

- Argument passing:

```
double func(const double& x);  
func(3); // must create temporary  
    // double _tmp = 3;  
    // func(_tmp);
```

# Temporary Objects (Continued)

## ■ Reference initialization:

```
int i = 2;  
const double& d = i; // must create temporary  
    // double _tmp = i;  
    // const double& d = _tmp;
```

## ■ Function return:

```
std::string getMessage();  
std::string s;  
s = getMessage(); // must create temporary  
    // std::string _tmp(getMessage());  
    // s = _tmp;
```

- In most (but not all) circumstances, a temporary object is destroyed as the last step in evaluating the full expression that contains the point where the temporary object was created. [\[C++17 15.2/4\]](#)

# Temporary Objects Example

```
1  class Complex {
2  public:
3      Complex(double re = 0.0, double im = 0.0) : re_(re),
4          im_(im) {}
5      Complex(const Complex& a) = default;
6      Complex(Complex&& a) = default;
7      Complex& operator=(const Complex& a) = default;
8      Complex& operator=(Complex&& a) = default;
9      ~Complex() = default;
10     double real() const {return re_;}
11     double imag() const {return im_;}
12 private:
13     double re_; // The real part.
14     double im_; // The imaginary part.
15 };
16
17 Complex operator+(const Complex& a, const Complex& b) {
18     return Complex(a.real() + b.real(), a.imag() + b.imag());
19 }
20
21 int main() {
22     Complex a(1.0, 2.0);
23     Complex b(1.0, 1.0);
24     Complex c;
25     // ...
26     c = a + b;
27 }
```

# Temporary Objects Example (Continued)

Original code:

```
int main() {  
    Complex a(1.0, 2.0);  
    Complex b(1.0, 1.0);  
    Complex c;  
    // ...  
    c = a + b;  
}
```

Code showing temporaries:

```
int main() {  
    Complex a(1.0, 2.0);  
    Complex b(1.0, 1.0);  
    Complex c;  
    // ...  
    Complex _tmp(a + b);  
    c = _tmp;  
}
```

# Prefix Versus Postfix Increment/Decrement

```
1  class Counter {
2  public:
3      Counter() : count_(0) {}
4      int getCount() const {return count_;}
5      Counter& operator++() { // prefix increment
6          ++count_;
7          return *this;
8      }
9      Counter operator++(int) { // postfix increment
10         Counter old(*this);
11         ++count_;
12         return old;
13     }
14 private:
15     int count_; // counter value
16 };
17
18 int main() {
19     Counter x;
20     Counter y;
21     y = ++x;
22     // no temporaries, int increment, operator=
23     y = x++;
24     // 1 temporary, 1 named, 2 constructors,
25     // 2 destructors, int increment, operator=
26 }
```

# Compound Assignment Versus Separate Assignment

SKIP SLIDE

```
1  #include <complex>
2  using std::complex;
3
4  int main() {
5      complex<double> a(1.0, 1.0);
6      complex<double> b(1.0, -1.0);
7      complex<double> z(0.0, 0.0);
8
9      // 2 temporary objects
10     // 2 constructors, 2 destructors
11     // 1 operator=, 1 operator+, 1 operator*
12     z = b * (z + a);
13
14     // no temporary objects
15     // only 1 operator+= and 1 operator*+=
16     z += a;
17     z *= b;
18 }
```

- Normally, a temporary object is destroyed as the last step in evaluating the full expression that contains point where temporary object was created.

[C++17 15.2/5] [C++17 15.2/6]

- First exception: When a default constructor with one or more default arguments is called to initialize an element of an array. [C++17 15.2/5]
- Second exception: When a *reference is bound to a temporary* (or a subobject of a temporary), the lifetime of the temporary is extended to *match the lifetime* of the reference, with following *exceptions*: [C++17 15.2/6]
  - A temporary bound to a reference member in a constructor initializer list persists until the constructor exits.
  - A temporary bound to a reference parameter in a function call persists until the completion of the full expression containing the call.
  - A temporary bound to the return value of a function in a return statement is not extended, and is destroyed at end of the full expression in the return statement.
  - A temporary bound to a reference in an initializer used in a new-expression persists until the end of the full expression containing that new-expression.

## ■ Example:

```
void func() {  
    std::string s1("Hello");  
    std::string s2(" ");  
    std::string s3("World!\n");  
    const std::string& s = s1 + s2 + s3;  
    std::cout << s; // OK?  
}
```

## ■ Example:

```
const std::string& getString() {  
    return std::string("Hello");  
}  
void func() {  
    std::cout << getString(); // OK?  
}
```



- temporary objects are created: [C++17 15.2/1]
  - when prvalue is materialized so that it can be used as xvalue
  - when needed to pass or return object of trivially-copyable type
  - when throwing exception
- materialization of temporary object is always delayed as long as possible in order to avoid creating unnecessary temporary objects [C++17 15.2/2]
- temporary objects are materialized: [C++17 15.2/2][C++17 7.4/1][C++17 8/10]
  - when binding a reference to a prvalue
  - when performing member access on a class prvalue
  - when performing array-to-pointer conversion or subscripting on array prvalue
  - when initializing object of type `std::initializer_list` from braced initializer list
  - for certain unevaluated operands
  - when prvalue appears as discarded expression

## Section 3.3

# Lvalues and Rvalues

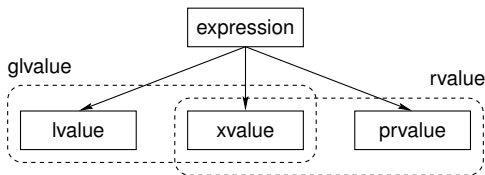
# Expressions

- An **expression** is a sequence of operators and operands that specifies a computation. [C++17 8/1]
- An expression has a type and, if the type is not void, a value.
- Example:

```
int x = 0;
int y = 0;
int* p = &x;
double d = 0.0;
// Evaluate some
// expressions here.
```

Expression	Type	Value
x	<b>int</b>	0
y = x	<b>int&amp;</b>	reference to y
x + 1	<b>int</b>	1
x * x + 2 * x	<b>int</b>	0
y = x * x	<b>int&amp;</b>	reference to y
x == 42	<b>bool</b>	<b>false</b>
*p	<b>int&amp;</b>	reference to x
p == &x	<b>bool</b>	<b>true</b>
x > 2 * y	<b>bool</b>	<b>false</b>
std::sin(d)	<b>double</b>	0.0

# Value Categories of Expressions



- Every expression can be classified into *exactly one* of three **value categories**: [C++17 6.10/1]
  - 1 lvalue
  - 2 xvalue (which stands for “expiring value”)
  - 3 prvalue (which stands for “pure rvalue”)
- An expression that is an lvalue or xvalue is called a **glvalue** (which stands for “generalized lvalue”).
- An expression that is a prvalue or an xvalue is called an **rvalue**.
- Every expression is either an lvalue or rvalue (but not both).
- Every expression is either a glvalue or prvalue (but not both).

# Value Categories of Expressions (Continued)

- How an expression can be used is dictated, in part, by its value category.
- Whether or not it is safe to move (instead of copy) depends on whether an lvalue or rvalue is involved.
- In the context of values categories, the notion of identity is important.
- An entity (such as an object or function) is said to have an **identity** if it can be distinguished from other like entities with identical attributes.
- For example, any object that is stored at a well-defined location in memory has an identity, since, given two objects with well-defined locations (and possibly identical values), one can always determine if these objects are the same object by comparing their locations.

- An **lvalue** is an expression that: [C++17 6.10/1]
  - denotes a function, object, or bitfield; and
  - has an identity.
- Since an lvalue establishes the identity of a function, object, or bitfield, an lvalue is always associated with some *well-defined location* in memory (which, in some cases, can be determined by the address-of operator).  
[C++17 8.3.1/3]
- If an lvalue expression corresponds to an object, the resources of that object are *not guaranteed* to be safe to reuse (i.e., the object is not guaranteed to be safe to use as the source for a move operation).

# Lvalues (Continued 1)

- A *named object* or *named function* is an lvalue. [\[C++17 8.1.4.1/1\]](#) [\[C++17 8.1.4.2/2\]](#)

Example:

```
int get_value();  
int i;  
int j;  
i = get_value();  
    // i and get_value are lvalues  
    // Note: get_value is not the same as get_value()  
j = i + 1; // i and j are lvalues
```

- The result of calling a function whose *return type is an lvalue reference type* is an lvalue. [\[C++17 8.2.2/11\]](#) Example:

```
int& get_value();  
++get_value(); // get_value() is an lvalue
```

- A *string literal* is an lvalue. [\[C++17 8.1.1/1\]](#) Example:

```
const char *s = "Hello"; // "Hello" is an lvalue
```

- A *named rvalue reference* is an lvalue. [\[C++17 8/7\]](#) Example:

```
int&& i = 1 + 3;  
int j = i; // i is an lvalue
```

## Lvalues (Continued 2)

- An *rvalue reference to a function* (both named and unnamed) is an lvalue. [\[C++17 8/7\]](#) Example:

```
void func();  
void (&&f)() = func;  
f(); // f is an lvalue  
std::move(func)(); // std::move(func) is an lvalue
```

- The result of each of the following built-in operators is an lvalue:
  - built-in subscripting operator (except when array rvalue involved) [\[C++17 8.2.1/1\]](#)
  - built-in indirection operator [\[C++17 8.3.1/1\]](#)
  - built-in pre-increment and pre-decrement operators [\[C++17 8.3.2/1\]](#) [\[C++17 8.3.2/2\]](#)
  - built-in assignment and compound-assignment operators [\[C++17 8.18/1\]](#)

Example:

```
char buffer[] = "Hello";  
char* s = buffer;  
*s = 'a'; // *s is an lvalue  
*(s + 1) = 'b'; // *(s + 1) is an lvalue  
++s; // ++s is an lvalue  
--s; // --s is an lvalue  
s += 2; // s += 2 is an lvalue  
s = &buffer[1];  
    // s = &buffer[1] is an lvalue  
    // buffer[1] is an lvalue
```



- Using a move (instead of a copy) is *not guaranteed to be safe* when the source is an lvalue (since other code can access the associated object by name or through a pointer or reference).
- Example:

```
void func() {  
    std::vector<int> x;  
    std::vector<int> y(x);  
    /* can we construct by moving (instead of copying)?  
       source x is lvalue; not safe to move x to y since  
       value of x might be used later */  
    y = x;  
    /* can we assign by moving (instead of copying)?  
       source x is lvalue; not safe to move x to y since  
       value of x might be used later */  
    // ...  
}
```

- A **prvalue** (i.e., pure rvalue) is an expression whose evaluation: [C++17 6.10/1]
  - computes the value of an operand of an operator; or
  - initializes an object or a bitfield.
- A prvalue never corresponds to an object (but, in some contexts, might be used to materialize a temporary object).
- A prvalue does not have an identity.
- An **xvalue** (i.e., expiring value) is an expression that: [C++17 6.10/1]
  - denotes an object or bitfield (usually near the end of its lifetime);
  - has an identity; and
  - the resources of the object/bitfield can safely be reused (i.e., is *deemed to be safe* to use as the source for a move).
- An xvalue is associated with certain kinds of expressions involving rvalue references or the materialization of a temporary object.
- An **rvalue** is an expression that is either a prvalue or an xvalue.
- Unlike an lvalue, an rvalue need not have an identity.
- Therefore, applying the (built-in) address-of operator to an rvalue (corresponding to an object) is not allowed. [C++17 8.3.1/3]

- The result of calling a function that *returns by value* (i.e., the return type is not a reference type) is a prvalue. [\[C++17 8.2.2/11\]](#) Example:

```
int get_value();  
int i = get_value();  
    // get_value() is a prvalue  
    // Note: get_value() is not the same as get_value
```

- All *literals other than string literals* are prvalues. [\[C++17 8.1.1/1\]](#) Example:

```
double pi = 3.1415; // 3.1415 is a prvalue  
int i = 42; // 42 is a prvalue  
i = 2 * i + 1; // 2 and 1 are prvalues  
char c = 'A'; // 'A' is a prvalue
```

- The *this keyword* is a prvalue. [\[C++17 8.1.2/2\]](#)

# Prvalues (Continued)

- The result of each of the following built-in operators is a prvalue:
  - built-in post-increment and post-decrement operators [\[C++17 8.2.6/1\]](#)
  - built-in arithmetic operators excluding increment and decrement operators (e.g., unary plus, unary minus, addition, subtraction, multiplication, division, and modulus) [\[C++17 8.3.1/2 unary?\]](#) [\[C++17 8.7 missing?\]](#)
  - built-in bitwise operators (e.g., bitwise NOT, bitwise AND, and bitwise OR)
  - built-in logical operators (e.g., logical NOT, logical AND, and logical OR)
  - built-in relational operators (e.g., equal, not equal, and less than)
  - built-in address-of operator [\[C++17 8.3.1/3\]](#)

Example:

```
int i;  
int j;  
i = -(3 + 5); // 3 + 5 and -(3 + 5) are prvalues  
j = i * i; // i * i is a prvalue  
j = (i == 42); // i == 42 is a prvalue  
j = (i & 7) | 2; // (i & 7) and (i & 7) | 2 are prvalues  
i = j++; // j++ is a prvalue  
int *ip = &i; // &i is a prvalue
```

- The result of calling a function whose *return type is an rvalue reference type* is an xvalue. [C++17 8/7.1] Example:

```
std::string s("Hello");
std::string t = std::move(s); // std::move(s) is xvalue
// Note: std::move returns rvalue reference type
```

- An *unnamed rvalue reference to an object* is an xvalue. [C++17 8/7] Example:

```
std::string s("Hello");
std::string t;
t = static_cast<std::string&&>(s);
// static_cast<std::string&&>(s) is xvalue
s = std::move(t); // std::move(t) is an xvalue
```

- A *temporary object* materialized from a prvalue is an xvalue. [C++17 7.4/1]

Example:

```
std::vector<int> v;
v = std::vector<int>(10, 2);
// temporary object materialized from prvalue
// std::vector<int>(10, 2) is an xvalue
std::complex<double> u;
u = std::complex<double>(1, 2);
// temporary object materialized from prvalue
// std::complex<double>(1, 2) is an xvalue
```

# Moving and Rvalues

- When the source object whose value is to be propagated is an rvalue, using a move (instead of a copy) is *always safe* (either because this must be so or the programmer has explicitly deemed this to be so).
- Example (move from temporary object):

```
void func() {  
    std::vector<int> x;  
    x = std::vector<int>(42, 0);  
    /* safe to move from temporary object materialized  
       from prvalue std::vector<int>(42, 0) since any change  
       to its value cannot be observed by other code */  
    // ...  
}
```

- Example (forced move):

```
void func() {  
    std::string s("hello");  
    std::cout << s << '\n';  
    std::string t(std::move(s));  
    /* safe to move from s to t since std::move(s) is  
       xvalue; programmer has, in effect, said "trust me, it  
       is safe to use move here"; of course, if programmer is  
       wrong, bad things will happen; programmer correct  
       in this case, since value of s not used again */  
    std::cout << t << '\n';  
}
```

- Lvalues and rvalues can be either *modifiable or nonmodifiable*.

Example:

```
int i = 0;
const int j = 2;
i = j + 3;
// i is modifiable lvalue
// j is nonmodifiable lvalue
// j + 3 is modifiable rvalue
const std::string getString();
std::string s = getString();
// getString() is nonmodifiable rvalue
```

- Class rvalues can have cv-qualified types, while non-class rvalues *always have cv-unqualified types*. [C++17 8/6] Example:

```
const int getConstInt(); // const is ignored
const std::string getConstString();
int i = getConstInt();
// getConstInt() is modifiable rvalue of type int
// (not const int)
std::string s = getConstString();
// getConstString() is nonmodifiable rvalue
```

# Moving and Lvalues/Rvalues

- With regard to propagating the value from one object to another, we can summarize the results from the earlier slides as follows:
  - If the source for a copy operation is an *lvalue*, the copy operation is *not guaranteed* to be safely replaceable by a move operation.
  - If the source for a copy operation is an *rvalue*, the copy operation is *guaranteed* to be safely replaceable by a move operation (where the guarantee may effectively come from the programmer in some cases).
- It would be highly desirable if the language would provide a mechanism that would automatically allow a move to be used in the rvalue case and a copy to be employed otherwise.
- In fact, this is exactly what the language does.
- The rules in the language for reference binding and overload resolution conspire to achieve the following final overall effect:
  - When the value of an object must be propagated to another object, a move operation is used if such an operation is *available* and the source object is an *rvalue*; otherwise, a copy operation is employed.



# Moving/Copying and Lvalues/Rvalues

- Consider the following code with respect to moving/copying:

```
1  /* Note: std::string provides copy and move constructors  
2  and copy and move assignment operators. */  
3  
4  std::string get_value();  
5  const std::string get_const_value(); // WARNING: bad idea  
6  
7  void func() {  
8      std::string s(get_value());  
9      // move elided (for reasons to be seen later)  
10     std::string t(s); // copy construction  
11     std::string u(std::move(t)); // move construction  
12     t = s; // copy assignment  
13     t = std::string("Hello"); // move assignment  
14     s = std::move(t); // move assignment  
15     t = get_const_value(); // copy assignment (not move!)  
16 }
```

- One further complication exists that is yet to be discussed.
- In some circumstances, a move/copy operation is either allowed or required to be elided (i.e., copy elision).
- This complication will be considered in detail later.

- Aside from the exceptions noted below, all of the built-in operators *require operands that are prvalues*. [\[C++17 8.2.1/1\]](#) [\[C++17 8.2.5/2\]](#)
- The operand of each of the following built-in operators must be an lvalue:
  - address of [\[C++17 8.3.1/3\]](#)
  - pre- and post-increment [\[C++17 8.2.6/1\]](#) [\[C++17 8.3.2/1\]](#)
  - pre- and post-decrement [\[C++17 8.2.6/1\]](#) [\[C++17 8.3.2/1\]](#)
- The left operand of the following built-in operators must be an lvalue:
  - assignment [\[C++17 8.18/1\]](#)
  - compound assignment [\[C++17 8.18/1\]](#)

- Whether an operator for a *class type* requires operands that are lvalues or rvalues or yield lvalues or rvalues is determined by the parameter types and return type of the operator function.
- The member selection operator may yield an lvalue or rvalue, depending on the particular manner in which the operator is used. (The behavior is fairly intuitive.) [\[C++17 8.2.5/4\]](#)
- The value category and type of the result produced by the ternary conditional operator depends on the particular manner in which the operator is employed. [\[C++17 8.16/5\]](#) [\[C++17 8.16/6\]](#)

# Implicit Lvalue-to-Rvalue Conversion

- The **lvalue-to-rvalue conversion** (which would be more accurately called the glvalue-to-prvalue conversion) is an implicit conversion from lvalues and xvalues to prvalues, which can be used in numerous circumstances.

[C++17 7.1/1]

- For non-class types, the above conversion also removes cv-qualifiers.
- Whenever an lvalue/xvalue appears as an operand of an operator that expects a prvalue operand, the lvalue-to-rvalue conversion is applied to convert the expression to a prvalue. [C++17 8/9] Example:

```
int i = 1;
int j = 2;
int k = i + j;
/* since built-in binary addition operator requires
   prvalue operands, i and j implicitly converted from
   lvalues to prvalues */
```

- The lvalue-to-rvalue conversion is not used for reference binding.

[C++17 11.6.3/5]

## Section 3.4

# Copy Elision and Implicit Moving

# Copy Elision

- normally, compiler forbidden from applying optimizations to code that would change its observable behavior (i.e., so called “as if” rule)
- one important exception to as-if rule is copy elision
- **copy elision** is code transformation that omits copy/move operation by constructing object in place to which it would later be copied/moved
- copy elision allows copy/move operations to be eliminated, thus avoiding cost of copy/move constructors
- copy elision may also eliminate need for some temporary objects, which avoids cost of constructing and destroying those objects
- copy elision either allowed or required in several contexts:
  - initialization
  - returning by value
  - passing by value
  - throwing by value
  - catching by value
- in cases where copy elision is mandatory, copy/move constructors need not be accessible or even provided at all, which provides more flexibility in dealing with non-movable non-copyable types

# Copy Elision and Returning by Value

- in return statement of function with class return type, when expression is name of non-volatile automatic object (other than function or catch-clause parameter) with same cv-unqualified type as function return type, automatic object can be constructed directly in function's return value

[C++17 15.8.3/(1.1)]

- copy elision required if (allowed as per above and) return expression is prvalue (i.e., placeholder for temporary object) [C++17 11.6/(17.6.1)]
- example:

```
1  struct Widget {
2      Widget();
3      Widget(const Widget&);
4      Widget(Widget&&);
5      // ...
6  };
7
8  Widget func1() {return Widget();}
9      // returns prvalue (i.e., placeholder for temporary object)
10 Widget func2() {Widget w; return w;} // returns named object
11
12 int main() {
13     Widget w(func1());
14     // required copy elision (not named object returned)
15     Widget x(func2());
16     // possible copy elision (named object returned)
17 }
```

# Copy Elision and Returning by Value (Continued)

SKIP SLIDE

- in context of returning by value, two forms of copy elision known by special names
- code transformation that eliminates copy from unnamed object associated with prvalue return expression to returned value in caller known as **return value optimization (RVO)**
- that is, RVO is copy elision in case that return expression is prvalue
- as seen previously, RVO is mandatory
- code transformation that eliminates copy from named object specified by return expression to returned value in caller known as **named return value optimization (NRVO)**
- that is, NRVO is copy elision in case that return expression is named object (i.e., not prvalue)
- as seen previously, NRVO is allowed but not required
- terms RVO and NRVO frequently used when discussing copy elision in context of returning by value

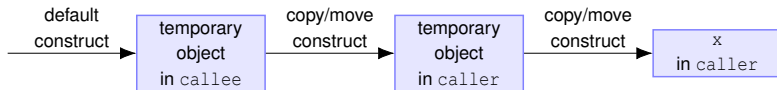


# Return-By-Value Example 1: Summary

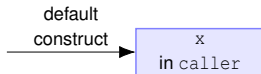
- consider following code (where type `T` is default constructible):

```
1 T callee() {return T();}  
2 void caller() {T x(callee()); /* ... */}
```

- goal is to construct object `x` in `caller` with value corresponding to `T()` (i.e., default constructed `T`)
- without copy elision, this would be achieved by:



- by maximally utilizing copy elision (and delaying creation of any temporary objects as long as possible), this can be reduced to single step:



- not only were two copy/move operations eliminated, need for any temporary objects also eliminated

# Return-By-Value Example 1: Without Copy Elision

SKIP SLIDE

- again, consider following code (where type `T` is default constructible):

```
T callee() {return T();}  
void caller() {T x(callee()); /* ... */}
```

- consider what happens without copy elision (in violation of standard in this case)
- executing body of `caller` proceeds as follows:
  - 1 storage for temporary object allocated in `caller` to hold return value of `callee`
  - 2 `caller` invokes `callee`
  - 3 expression in return statement of `callee` evaluated, resulting in construction of temporary object in `callee` (via default constructor) to hold return value
  - 4 before `callee` returns, value of temporary object in `callee` propagated to temporary object in `caller` (via move/copy construction)
  - 5 `callee` returns (after destroying its temporary object holding return value)
  - 6 value of temporary object in `caller` propagated to `x` (via move/copy construction); then temporary object destroyed
- overhead: two temporary objects created (two constructor and destructor invocations); must propagate value into and out of temporary objects

- again, consider following code (where type `T` is default constructible):

```
T callee() {return T();}  
void caller() {T x(callee()); /* ... */}
```

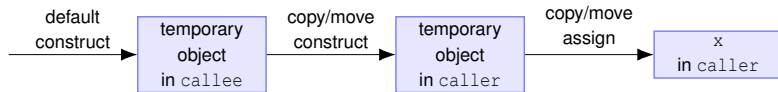
- consider what happens with copy elision (which is required by standard in this case)
- executing body of `caller` proceeds as follows:
  - 1 `caller` invokes `callee`
  - 2 expression in return statement of `callee` evaluated, resulting in return value of `callee` being constructed directly in `x` in `caller` (via default constructor)
- no overhead: no temporary objects created and therefore no need to propagate values into or out of temporary objects

# Return-By-Value Example 2: Summary

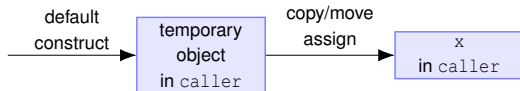
- consider following code (where type `T` is default constructible and copyable/movable):

```
1 T callee() {return T();}  
2 void caller() {T x; x = callee(); /* ... */}
```

- goal is to assign value corresponding to `T()` to `x` in `caller`
- without copy elision, this would be achieved by:



- with copy elision, this can be reduced to:



- able to eliminate one move/copy operation and one temporary object
- unlike in case of previous example, cannot eliminate temporary object in `caller` since temporary object must be materialized in order to perform assignment to `x` in `caller`

## Return-By-Value Example 2: Without Copy Elision

SKIP SLIDE

again, consider following code (where type `T` is default constructible and copyable/movable):

```
T callee() {return T();}  
void caller() {T x; x = callee(); /* ... */}
```

- consider what happens without copy elision (in violation of standard in this case)
- executing body of `caller` proceeds as follows:
  - 1 storage for temporary object allocated in `caller` to hold return value of `callee`
  - 2 `caller` invokes `callee`
  - 3 expression in return statement of `callee` evaluated, resulting in construction of temporary object in `callee` (via default constructor) to hold return value
  - 4 before `callee` returns, value of temporary object in `callee` propagated to temporary object in `caller` (via move/copy construction)
  - 5 `callee` returns (after destroying its temporary object holding return value)
  - 6 value of temporary object in `caller` propagated to `x` (via move/copy assignment); then temporary object destroyed
- overhead: two temporary objects created; must propagate value from temporary object in `callee` to temporary object in `caller`

# Return-By-Value Example 2: With Copy Elision

SKIP SLIDE

- again, consider following code (where type `T` is default constructible and copyable/movable):

```
T callee() {return T();}  
void caller() {T x; x = callee(); /* ... */}
```

- consider what happens with copy elision (which is required by standard in this case)
- executing body of `caller` proceeds as follows:
  - 1 storage for temporary object allocated in `caller` to hold return value of `callee`
  - 2 `caller` invokes `callee`
  - 3 expression in return statement of `callee` evaluated, resulting in return value of `callee` being constructed directly in temporary object in `caller` (via default constructor)
  - 4 `callee` returns
  - 5 value of temporary object in `caller` propagated to `x` (via move/copy assignment); then temporary object destroyed
- overhead: one temporary object created; but no need to propagate value from temporary object in `callee` to temporary object in `caller`

# Example Where Copy Elision Allowed But Likely Impossible

```
1  class Widget {
2  public:
3      Widget(int) { /* ... */ }
4      Widget(const Widget&) = default;
5      Widget(Widget&&) = default;
6      // ...
7  };
8
9  bool get_flag();
10
11  // eliding copy of return value is not possible
12  Widget func() {
13      Widget w(0);
14      // w must be constructed before it is known if
15      // w will be returned; so cannot know whether to
16      // construct w in returned value
17      Widget v(42);
18      // v must be constructed before it is known if
19      // v will be returned; so cannot know whether to
20      // construct v in returned value
21      if (get_flag()) {return w;}
22      else {return v;}
23  }
24
25  int main() {
26      Widget w(func());
27  }
```

# Copy Elision and Passing by Value

- in function call, when temporary class object not bound to reference would be copied/moved to class object with same cv-unqualified type, temporary object can be constructed directly in target of omitted copy/move

[C++17 11.6/15][C++17 8.2.2/4]

- copy elision always required if allowed (as per above) [C++17 11.6/(17.6.1)]
- example:

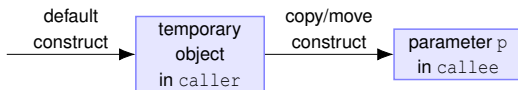
```
1  struct Widget {
2      Widget();
3      Widget(const Widget&);
4      Widget(Widget&&);
5      // ...
6  };
7
8  void func(Widget w) { /* ... */ }
9
10 int main() {
11     func(Widget()); // required copy elision
12     func(std::move(Widget())); /* BAD IDEA:
13         copy elision not allowed; move performed */
14 }
```



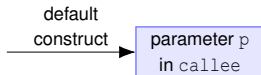
# Pass-By-Value Example: Summary

- consider following code (where type `T` is default constructible):

```
1 void callee(T p) { /* ... */ }
2 void caller() { callee(T()); }
```
- goal is to invoke `callee` with its function parameter `p` having value corresponding to `T()` (i.e., default constructed `T`)
- without copy elision, this would be accomplished by:



- with copy elision, this can be reduced to:



- by using copy elision, not only was one copy/move operation eliminated, but temporary object also eliminated

# Pass-By-Value Example: Without Copy Elision

SKIP SLIDE

- again, consider following code (where type `T` is default constructible):

```
void callee(T p) { /* ... */ }  
void caller() { callee(T()); }
```

- consider what happens without copy elision (in violation of standard in this case)
- executing body of `caller` proceeds as follows:
  - 1 temporary object created in `caller` with value corresponding to `T()` (via default constructor)
  - 2 storage for `callee`'s function parameter `p` allocated (on stack)
  - 3 value of temporary object in `caller` propagated (via move/copy constructor) to `callee`'s function parameter `p`
  - 4 `caller` transfers control to `callee`
  - 5 `callee` returns, resulting in its function parameter being destroyed (and deallocated)
  - 6 temporary object in `caller` destroyed (and deallocated)
- overhead: one temporary object created (constructor and destructor invocations); one move/copy required to propagate value from temporary object elsewhere

- again, consider following code (where type `T` is default constructible):

```
void callee(T p) { /* ... */ }  
void caller() { callee(T()); }
```

- consider what happens with copy elision (as required by standard in this case)
- executing body of `caller` proceeds as follows:
  - 1 function parameter `p` constructed (on stack) with value corresponding to `T()` (via default constructor)
  - 2 `caller` transfers control to `callee`
  - 3 `callee` returns, resulting in its function parameter being destroyed (and deallocated)
- no overhead: no temporary objects created and therefore no need to propagate values into or out of temporary objects

# Copy Elision and Throwing by Value

SKIP SLIDE

- in throw expression, when operand is name of non-volatile automatic object (other than function or catch-clause parameter) whose scope does not extend beyond end of innermost enclosing try block (if there is one), copy/move operation from operand to exception object can be omitted by constructing automatic object directly into exception object [\[C++17 15.8.3/\(1.2\)\]](#)
- copy elision required if (allowed as per above and) throw expression is prvalue [\[C++17 11.6/\(17.6.1\)\]](#)

## ■ example:

```
1  struct Widget {  
2      Widget();  
3      Widget(const Widget &);  
4      Widget(Widget&&);  
5      // ...  
6  };  
7  
8  void func_1() {  
9      throw Widget(); // required copy elision (prvalue)  
10 }  
11  
12 void func_2() {  
13     Widget w; throw w; // possible copy elision (not prvalue)  
14 }
```

# Copy Elision and Catching by Value

SKIP SLIDE

- when exception declaration of exception handler declares object of same type (except for cv-qualification) as exception object, copy/move operation can be omitted by treating exception declaration as alias for exception object if meaning of program will be unchanged except for execution of constructors and destructors for object declared by exception declaration

[C++17 15.8.3/(1.3)]

- in this context, copy elision never required
- example:

```
1  struct Widget {  
2      Widget();  
3      Widget(const Widget &);  
4      Widget(Widget&&);  
5      // ...  
6  };  
7  
8  int main() {  
9      try {throw Widget();}  
10     catch (Widget foo) { // possible copy elision  
11         /* ... (foo not modified) */  
12     }  
13 }
```

# Copy Elision and Initialization

- if prvalue used as initializer of object with same type (except for cv-qualification), object must be initialized directly [\[C++17 11.6/\(17.6.1\)\]](#)

- example:

```
1  class Widget {
2  public:
3      Widget();
4      Widget(const Widget&);
5      Widget(Widget&&);
6      // ...
7  };
8
9  Widget func();
10
11 int main() {
12     Widget w = Widget(func());
13     // copy elision required for initialization;
14     // no copy/move in main function;
15     // returned value from func directly constructed in w;
16     // func may need copy/move to propagate return value
17     // out of func, if not elided
18     Widget u{Widget()};
19     // copy elision required for initialization;
20     // no copy/move; new Widget object is default
21     // constructed directly in u
22 }
```

# Mandatory Copy Elision Example: Factory Function

```
1  class Widget {
2  public:
3      Widget() { /* ... */ }
4      // not copyable
5      Widget(const Widget&) = delete;
6      Widget& operator=(const Widget&) = delete;
7      // not movable
8      Widget(Widget&&) = delete;
9      Widget& operator=(Widget&&) = delete;
10     // ...
11 };
12
13 Widget make_widget() {
14     return Widget();
15 }
16
17 int main() {
18     Widget w(make_widget());
19     // OK: copy elision required
20     Widget v{Widget()};
21     // OK: copy elision required
22     Widget u(Widget());
23     // function declaration
24 }
```

# Maximally Delayed Materialization of Temporary Objects

SKIP SLIDE

```
1  #include <cassert>
2
3  class Widget {
4  public:
5      Widget() : c_(0) {}
6      Widget(const Widget& other) : c_(other.c_ + 1) {}
7      Widget& operator=(const Widget& other) {c_ = other.c_ + 1; return *this;}
8      int count() const {return c_;}
9  private:
10     int c_;
11 };
12
13 Widget widget_1() {return Widget();}
14 // mandatory copy elision for return value
15 Widget widget_2() {return widget_1();}
16 // mandatory copy elision for return value
17 Widget widget_3() {return widget_2();}
18 // mandatory copy elision for return value
19
20 int main() {
21     Widget w{widget_3()}; // no temporary object; no copy/move
22     assert(w.count() == 0);
23     w = widget_3();
24     // widget_1 directly constructs return value into
25     // temporary object in main
26     assert(w.count() == 1);
27     Widget v{Widget(Widget(Widget(Widget(Widget(Widget()))))));}
28     // default constructs directly into v; no copy/move
29     assert(v.count() == 0);
30 }
```



- A copy operation associated with a return statement may be (elided or) converted to a move operation if an automatic storage duration variable is returned. [\[C++17 9.6.3/2\]](#) [\[C++17 18.8.3/\(3.1\)\]](#)
- Overload resolution to select the *constructor* for the copy is first performed as if the object were designated by an rvalue. [\[C++17 15.8.3\]](#)
- If the first overload resolution fails or was not performed, or if the type of the first parameter of the selected constructor is not an rvalue reference to the object's type (possibly cv-qualified), overload resolution is performed again, considering the object as an lvalue. [\[C++17 15.8.3\]](#)
- In a future version of the C++ standard, the condition on a constructor being selected seems likely to be relaxed to cover the cases of conversion operators and slicing.

# Example: Return Statements and Moving/Copying

```
1  #include <string>
2
3  class Widget {
4  public:
5      Widget(const std::string&) {/* ... */}
6      Widget(const Widget&) = default;
7      Widget(Widget&&) = default;
8      // ...
9  };
10
11 Widget get_value_1() {
12     Widget w("goodbye");
13     return w;
14     // copy elision is allowed, but not required;
15     // if move/copy not elided:
16     // since w is local object, w first treated as if rvalue,
17     // resulting in move constructor being selected
18     // to propagate return value to caller
19 }
20
21 Widget get_value_2() {
22     std::string s("hello");
23     return s;
24     // effectively: return Widget(s);
25     // copy elision required (since returning temporary)
26 }
27
28 int main() {
29     Widget w(get_value_1());
30     Widget v(get_value_2());
31 }
```

# Use of `std::move` in Return Expressions

- consider function that returns by value
- for such function, almost never desirable to use result of invocation of `std::move` as return expression, since this prohibits copy elision
- example (bad use of `std::move`):

```
Widget make_widget() {  
    return std::move(Widget());  
    // BAD: move performed instead of eliding move  
}
```

- only use result of invoking `std::move` as return expression if:
  - 1 copy elision cannot be used; and
  - 2 copy would be performed instead of move
- example (potentially good use of `std::move`):

```
Buffer read(Buffer&& buffer) {  
    // ... (read data into buffer)  
    return std::move(buffer);  
    // OK: not eligible for copy elision and  
    // would have copied (not moved)  
}
```

# Example: Incorrect Use of `std::move`

```
1  #include <string>
2
3  class Widget {
4  public:
5      Widget(const std::string&) { /* ... */}
6      Widget(const Widget&) = default;
7      Widget(Widget&&) = default;
8      // ...
9  };
10
11 Widget get_value_1() {
12     return Widget("hello");
13     // copy elision required
14 }
15
16 // Note: This type of usage of std::move is highly undesirable,
17 // as it prevents copy elision.
18 Widget get_value_2() {
19     return std::move(Widget("hello"));
20     // since cv-unqualified return-expression type and
21     // cv-unqualified return type differ, copy elision is not
22     // allowed; move constructor is selected to propagate return
23     // value to caller; if std::move had not been used, copy/move
24     // would have been completely eliminated (due to required
25     // copy elision)
26 }
27
28 int main() {
29     Widget w(get_value_1());
30     Widget v(get_value_2());
31 }
```

# Copy/Move/Elide Example: Widget (1)

```
1  #include <utility>
2
3  class Widget {
4  public:
5      Widget();
6      Widget(const Widget&);
7      Widget(Widget&&);
8      Widget& operator=(const Widget&);
9      Widget& operator=(Widget&&);
10     // ...
11 };
12
13 Widget func_0(Widget w) {
14     // copy elision for return value not allowed since object being
15     // returned is function parameter; treating return expression as
16     // rvalue results in move being performed
17     return w;
18 }
19
20 Widget func_1() {
21     try {
22         // copy elision for exception object is required
23         throw Widget();
24     } catch (Widget w) {
25         // copy elision for return value not allowed since object
26         // being returned is catch-clause parameter; treating return
27         // expression as rvalue results in move being performed
28         return w;
29     }
30 }
```

## Copy/Move/Elide Example: Widget (2)

```
32 Widget func_2(Widget& w) {  
33     // copy elision for return value not allowed for several reasons  
34     // (e.g., function parameter, not automatic, cv-unqualified type  
35     // mismatch); return expression cannot be treated as rvalue since  
36     // w is not automatic object; so copy is performed  
37     return w;  
38 }  
39  
40 Widget func_3() {  
41     static Widget w;  
42     // copy elision for return value not allowed since w is not  
43     // automatic object; return expression cannot be treated as  
44     // rvalue since w is not automatic object; so copy is performed  
45     return w;  
46 }  
47  
48 Widget g;  
49 Widget func_4() {  
50     // copy elision for return value not allowed since g is not  
51     // automatic; cannot treat w as rvalue since not automatic  
52     // object; so copy is performed  
53     return g;  
54 }  
55  
56 void func_5() {  
57     Widget w;  
58     // copy elision for exception object is allowed; if copy not  
59     // elided, move performed  
60     throw w;  
61 }
```

## Copy/Move/Elide Example: Widget (3)

```
63 void func_6(Widget w) {  
64     // copy elision is not allowed since object in throw expression  
65     // is function parameter; copy performed in C++17, but may become  
66     // move in future (P1155)  
67     throw w;  
68 }  
69  
70 Widget func_7(Widget&& w) {  
71     // copy elision is not allowed for several reasons (e.g.,  
72     // cv-unqualified return type does not match cv-unqualified  
73     // return-expression type, not automatic object); return  
74     // expression cannot be treated as rvalue since w is not  
75     // automatic object; copy performed in C++17, but may become  
76     // move in future (P0527)  
77     return w;  
78 }  
79  
80 Widget func_8(Widget& w) {  
81     Widget&& x = std::move(w);  
82     // copy elision is not allowed for several reasons (e.g.,  
83     // cv-unqualified return type does not match cv-unqualified  
84     // return-expression type and not automatic object); copy  
85     // performed in C++17, but may become move in future (P0527)  
86     return x;  
87 }
```

# Copy/Move/Elide Example: Conversion (1)

```
1  class Widget {
2  public:
3      Widget();
4      Widget(const Widget&);
5      Widget(Widget&&);
6      // ...
7  };
8
9  class Gadget {
10 public:
11     Gadget();
12     Gadget(const Gadget&);
13     Gadget(Gadget&&);
14     Gadget(const Widget& w); // copying converting constructor
15     Gadget(Widget&& w); // moving converting constructor
16     // ...
17 };
18
19 class Doodad {
20 public:
21     Doodad();
22     Doodad(const Doodad&);
23     Doodad(Doodad&&);
24     operator Widget() const& // copying conversion operator
25     operator Widget() && // moving conversion operator
26     // ...
27 };
```



## Copy/Move/Elide Example: Conversion (2)

```
29 Gadget func_1() {
30     Widget w;
31     // copy elision is not permitted since cv-unqualified return type
32     // and cv-unqualified return-expression type do not match;
33     // when w treated as rvalue, moving converting constructor found;
34     // performs move via moving converting constructor
35     return w;
36 }
37
38 Widget func_2() {
39     Doodad t;
40     // copy elision is not permitted since cv-unqualified return type
41     // does not match cv-unqualified return-expression type;
42     // when t treated as rvalue, no constructor is found;
43     // performs copy via copying conversion operator in C++17,
44     // but may change to move in future (P1155) since some compilers
45     // do this anyways (in violation of standard)
46     return t;
47 }
```

# Copy/Move/Elide Example: Slicing

```
1  #include <iostream>
2
3  class Base {
4  public:
5      Base();
6      Base(const Base&);
7      Base(Base&&);
8      Base& operator=(const Base&);
9      Base& operator=(Base&&);
10     // ...
11 };
12
13 class Derived : public Base {
14 public:
15     Derived();
16     Derived(const Derived&);
17     Derived(Derived&&);
18     Derived& operator=(const Derived&);
19     Derived& operator=(Derived&&);
20     // ...
21 };
22
23 Base base_1() {
24     Derived x;
25     // copy elision is not permitted since cv-unqualified
26     // return-expression type does not match cv-unqualified return
27     // type; copy is performed in C++17, but may become move in
28     // future (P1155)
29     return x;
30 }
```

# Copy/Move/Elide Example: Other Operators (1)

```
1  class BigInt {
2  public:
3      BigInt();
4      BigInt(const BigInt&);
5      BigInt(BigInt&&);
6      BigInt& operator=(const BigInt&);
7      BigInt& operator=(BigInt&&);
8      BigInt& operator+=(int);
9      BigInt& operator++();
10     // ...
11 };
12
13 BigInt func_1(BigInt c) {
14     c += 1;
15     // copy elision for return value not allowed since return
16     // expression is function parameter; move is performed
17     return c;
18 }
19
20 BigInt func_2(BigInt c) {
21     // copy elision for return value not allowed since
22     // cv-unqualified return-expression type does not match
23     // cv-unqualified return type; copy is performed
24     return c += 1;
25 }
```

## Copy/Move/Elide Example: Other Operators (2)

```
27 BigInt func_3(const BigInt& c) {  
28     // copy elision for return value not allowed for several  
29     // reasons (e.g., function parameter, not automatic,  
30     // cv-unqualified type mismatch); copy is performed  
31     return BigInt(c) += 1;  
32 }
```

# Copy/Move/Elide Example: Parentheses

```
1  class Widget {
2  public:
3      Widget();
4      Widget(const Widget&);
5      Widget(Widget&&);
6      // ...
7  };
8
9  Widget func_0() {
10     Widget w;
11     // copy elision is permitted
12     // if not elided, move is performed
13     return w;
14 }
15
16 Widget func_1() {
17     Widget w;
18     // standard seems to suggest copy elision not
19     // permitted since (w) is not name of object, but:
20     // Clang 7.0.0 (-std=c++17 -O2) elides copy;
21     // MSVC 19.15 (/std:c++17 /O2) elides copy;
22     // GCC 8.2 (-std=c++17 -O2) performs move
23     return (w);
24 }
```

## Section 3.5

### **Rvalue References**

## Section 3.5.1

### **Introduction**

# Motivation Behind Rvalue References

- Rvalue references were added to the language in C++11 in order to provide support for:
  - 1 move operations; and
  - 2 perfect forwarding.
- A move operation is used to propagate the value from one object to another, much like a copy operation, except that a move operation makes fewer guarantees, allowing for greater efficiency and flexibility in many situations.
- Perfect forwarding relates to being able to pass function arguments from a template function through to another function (called by the template function) while preserving certain properties of those arguments.



# Terminology: Named and Cv-Qualified

- A type that includes one or both of the qualifiers **const** and **volatile** is called a **cv-qualified type**.
- A type that is not cv-qualified is called **cv-unqualified**. [\[C++17 6.9.3/1\]](#)
- Example:  
The types **const int** and **volatile char** are cv-qualified.  
The types **int** and **char** are cv-unqualified.
- An object or function that is named by an identifier is said to be **named**.
- An object or function that cannot be referred to by name is said to be **unnamed**.
- Example:

```
std::vector<int> v = {1, 2, 3, 4};  
std::vector<int> w;  
w = v; // w and v are named  
w = std::vector<int>(2, 0);  
    // w is named  
    // std::vector<int>(2, 0) is unnamed
```

## Section 3.5.2

### **Copying and Moving**

# Propagating Values: Copying and Moving

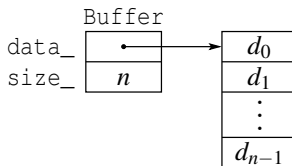
- Suppose that we have two objects of the same type and we want to propagate the value of one object (i.e., the source) to the other object (i.e., the destination).
- This can be accomplished in one of two ways:
  - 1 copying; or
  - 2 moving.
- **Copying** propagates the value of the source object to the destination object *without modifying the source object*.
- **Moving** propagates the value of the source object to the destination object and is *permitted to modify the source object*.
- Moving is always at least as efficient as copying, and for many types, moving is *more efficient* than copying.
- For some types, *copying does not make sense*, while moving does (e.g., `std::ostream` and `std::istream`).

# Buffer Example: Moving Versus Copying

- Consider a class that represents a character buffer (whose size is fixed at run time).

```
class Buffer {  
public:  
    // ...  
private:  
    char* data_; // pointer to buffer data  
    std::size_t size_; // buffer size (in characters)  
};
```

- Pictorially, the data structure looks like the following:



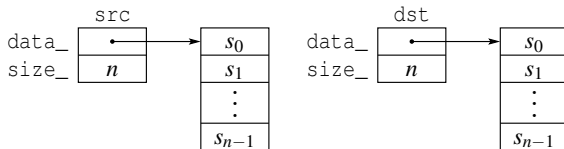
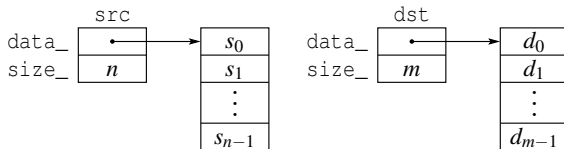
- How would copying be implemented?
- How would moving be implemented?

# Buffer Example: Copying

- code for copying from source `src` to destination `dst` (not self assignment):

```
delete [] dst.data_;  
dst.data_ = new T[src.size_];  
dst.size_ = src.size_;  
std::copy_n(src.data_, src.size_, dst.data_);
```

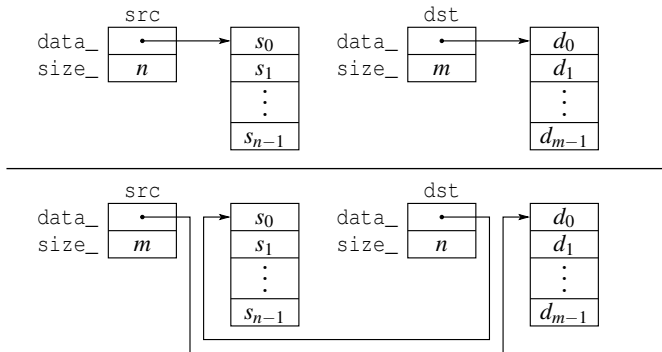
- copying requires: one array delete, one array new, copying of element data, and updating `data_` and `size_` data members
- copying proceeds as follows:



# Buffer Example: Moving

- code for moving from source `src` to destination `dst`:

```
std::swap(src.data_, dst.data_);  
std::swap(src.size_, dst.size_);
```
- moving only requires updating `data_` and `size_` data members
- although not considered here, could also free data array associated with `src` if desirable to release memory as soon as possible
- moving proceeds as follows:



# Moving Versus Copying

- Moving is usually more efficient than copying, often by very large margin.
- So, we should prefer moving to copying.
- We can safely replace a copy by a move when subsequent code does not depend on the value of source object.
- It would be convenient if the language could provide a mechanism for automatically using a move (instead of a copy) in situations where doing so is always guaranteed to be safe.
- For reasons of efficiency, it would also be desirable for the language to provide a mechanism whereby the programmer can override the normal behavior and force a move (instead of a copy) in situations where such a transformation is known to be safe only due to some special additional knowledge about program behavior.
- Rvalue references (in concert with the rules for reference binding and overload resolution) provide the above mechanisms.

## Section 3.5.3

### **References, Reference Binding, and Overload Resolution**



- A **reference** is an alias (i.e., nickname) for an already existing object.
- The language has two kinds of references:
  - 1 lvalue references
  - 2 rvalue references
- An **lvalue reference** is denoted by `&` (often read as “ref”).

```
int i = 5;  
int& j = i; // j is lvalue reference to int  
const int& k = i; k is lvalue reference to const int
```

- An **rvalue reference** is denoted by `&&` (often read as “ref ref”).

```
int&& i = 5; // i is rvalue reference to int  
const int&& j = 17; // j is rvalue reference to const int
```

- The act of initializing a reference is known as **reference binding**.
- Lvalue and rvalue references differ only in their properties relating to:
  - reference binding; and
  - overload resolution.

# References: Binding and Overload Resolution

- The kinds of expressions, to which lvalue and rvalue references can *bind*, differ.
- For a nonreference type `T` (such as `int` or `const int`), what kinds of expressions can validly be placed in each of the boxes in the example below?

```
T& r =  ;  
T&& r =  ;
```

- Lvalue and rvalue references also behave differently with respect to *overload resolution*.
- Let `T` be a cv-unqualified nonreference type. Which overloads of `func` will be called in the example below?

```
T operator+(const T&, const T&);  
void func(const T&);  
void func(T&&);  
T x;  
func(x); // calls which version of func?  
func(x + x); // calls which version of func?
```

# Reference Binding

- Implicit lvalue-to-rvalue conversion is disabled when binding to references. [\[C++17 11.6.3?\]\[C++17 16.3.3.1.4/1\]\[C++17 7.1\]](#)
- An lvalue reference can bind to an lvalue as long as doing so would not result in the *loss* of any cv qualifiers.

```
const int i = 0;
int& r1 = i; // ERROR: drops const
const int& r2 = i; // OK
const volatile int& r3 = i; // OK
```

- The loss of cv qualifiers must be avoided for *const and volatile correctness*.
- Similarly, an rvalue reference can bind to an rvalue as long as doing so would not result in the *loss* of any cv qualifiers.

```
const std::string getValue();
std::string&& r1 = getValue(); // ERROR: drops const
const std::string&& r2 = getValue(); // OK
```

- Again, the loss of cv qualifiers must be avoided for *const and volatile correctness*.

# Reference Binding (Continued)

- An lvalue reference can be bound to an rvalue only if doing so would not result in the *loss* of any cv qualifier and the lvalue reference is *const*.

```
const std::string getConstValue();  
std::string& r1 = getConstValue(); // ERROR: drops const  
const std::string& r2 = getValue(); // OK  
int& ri1 = 42; // ERROR: not const reference  
const int& ri2 = 42; // OK
```

- The requirement that the lvalue reference be *const* is to prevent temporary objects from being modified in a very uncontrolled manner, which can lead to subtle bugs.
- An rvalue reference can *never* be bound to an lvalue.

```
int i = 0;  
int&& r1 = i; // ERROR: cannot bind to lvalue  
int&& r2 = 42; // OK
```

- Allowing rvalue reference to bind to lvalues would violate the principle of type-safe overloading, which can lead to subtle bugs.

# Why Rvalue References Cannot Bind to Lvalues

- In effect, rvalue references were introduced into the language to allow a function to know if one of its reference parameters is bound to an object whose value is safe to change without impacting other code, namely, an rvalue (i.e., a temporary object or xvalue).
- Since an rvalue reference can only bind to an rvalue, any rvalue reference parameter to a function is *guaranteed* to be bound to a temporary object or xvalue.
- Example:

```
class Thing {  
public:  
    // Move constructor  
    // parameter x known to be safe to use as source for move  
    Thing(Thing&& x);  
    // Move assignment operator  
    // parameter x known to be safe to use as source for move  
    Thing& operator=(Thing&& x);  
    // ...  
};  
// parameter x known to be safe to modify  
void func(Thing&& x);
```

- If rvalue references could bind to lvalues, the above guarantee could not be made, as an rvalue reference could then refer to an object whose value cannot be changed safely, namely, an lvalue.

# Why Non-Const Lvalue References Cannot Bind to Rvalues

- If non-const lvalue references could bind to rvalues, temporary objects could be modified in many undesirable circumstances.

```
void func(int& x) {  
    // ...  
}  
  
int main() {  
    int i = 1;  
    int j = 2;  
    func(i + j);  
    // ERROR: cannot bind non-const lvalue  
    // reference to rvalue  
    // What would be consequence if allowed?  
}
```

# Reference Binding Summary

	Rvalue				Lvalue			
	T	const T	volatile T	const volatile T	T	const T	volatile T	const volatile T
T&&	✓	C	V	C,V	✗	✗	✗	✗
const T&&	✓	✓	V	V	✗	✗	✗	✗
volatile T&&	✓	✗	✓	C	✗	✗	✗	✗
const volatile T&&	✓	✓	✓	✓	✗	✗	✗	✗
T&	✗	✗	✗	✗	✓	C	V	C,V
const T&	✓	✓	V	V	✓	✓	V	V
volatile T&	✗	✗	✗	✗	✓	C	✓	C
const volatile T&	✗	✗	✗	✗	✓	✓	✓	✓

✓: allowed    C: strips **const**    V: strips **volatile**    ✗: other

# Reference Binding Example

```
1  #include <string>
2  using std::string;
3
4  string value() {
5      return string("Hello");
6  }
7
8  const string constValue() {
9      return string("World");
10 }
11
12 int main() {
13     string i("mutable");
14     const string j("const");
15
16     string& r01 = i;
17     string& r02 = j; // ERROR: drops const
18     string& r03 = value(); // ERROR: non-const lvalue reference from rvalue
19     string& r04 = constValue(); // ERROR: non-const lvalue reference from rvalue
20
21     const string& r05 = i;
22     const string& r06 = j;
23     const string& r07 = value();
24     const string& r08 = constValue();
25
26     string&& r09 = i; // ERROR: rvalue reference from lvalue
27     string&& r10 = j; // ERROR: rvalue reference from lvalue
28     string&& r11 = value();
29     string&& r12 = constValue(); // ERROR: drops const
30
31     const string&& r13 = i; // ERROR: rvalue reference from lvalue
32     const string&& r14 = j; // ERROR: rvalue reference from lvalue
33     const string&& r15 = value();
34     const string&& r16 = constValue();
35 }
```



- Lvalues strongly prefer binding to lvalue references.
- Rvalues strongly prefer binding to rvalue references.
- Modifiable lvalues and rvalues weakly prefer binding to non-const references.

# Overload Resolution Summary

	Priority							
	Rvalue				Lvalue			
	T	const T	volatile T	const volatile T	T	const T	volatile T	const volatile T
T&&	1							
const T&&	2	1						
volatile T&&	2		1					
const volatile T&&	3	2	2	1				
T&					1			
const T&	4	3			2	1		
volatile T&					2		1	
const volatile T&					3	2	2	1

# Overloading Example 1

```
1  #include <iostream>
2  #include <string>
3
4  void func(std::string& x) {
5      std::cout << "func(std::string&) called\n";
6  }
7
8  void func(const std::string& x) {
9      std::cout << "func(const std::string&) called\n";
10 }
11
12 void func(std::string&& x) {
13     std::cout << "func(std::string&&) called\n";
14 }
15
16 void func(const std::string&& x) {
17     std::cout << "func(const std::string&&) called\n";
18 }
19
20 const std::string&& constValue(const std::string&& x) {
21     return static_cast<const std::string&&>(x);
22 }
23
24 int main() {
25     const std::string cs("hello");
26     std::string s("world");
27     func(s);
28     func(cs);
29     func(cs + s);
30     func(constValue(cs + s));
31 }
32
33 /* Output:
34 func(std::string&) called
35 func(const std::string&) called
36 func(std::string&&) called
37 func(const std::string&&) called
38 */
```

# Overloading Example 2

```
1  #include <iostream>
2  #include <string>
3
4  void func(const std::string& x) {
5      std::cout << "func(const std::string&) called\n";
6  }
7
8  void func(std::string&& x) {
9      std::cout << "func(std::string&&) called\n";
10 }
11
12 const std::string&& constValue(const std::string&& x) {
13     return static_cast<const std::string&&>(x);
14 }
15
16 int main() {
17     const std::string cs("hello");
18     std::string s("world");
19     func(s);
20     func(cs);
21     func(cs + s);
22     func(constValue(cs + s));
23 }
24
25 /* Output:
26 func(const std::string&) called
27 func(const std::string&) called
28 func(std::string&&) called
29 func(const std::string&) called
30 */
```

# Why Rvalue References Cannot Bind to Lvalues (Revisited)

- If an rvalue reference could bind to an lvalue, this would violate the principle of type-safe overloading.

```
1  #include <iostream>
2  #include <string>
3
4  template <class T>
5  class Container {
6  public:
7      // ...
8      // Forget to provide the following function:
9      // void push_back(const T& value); // Copy semantics
10     void push_back(T&& value); // Move semantics
11 private:
12     // ...
13 };
14
15 int main() {
16     std::string s("Hello");
17     Container<std::string> c;
18     // What would happen here if lvalues
19     // could bind to rvalue references?
20     c.push_back(s);
21     std::cout << s << '\n';
22 }
```

## Section 3.5.4

### **Moving**

- A non-template constructor for class `T` is a **move constructor** if it can be called with one parameter that is of type `T&&`, **const** `T&&`, **volatile** `T&&`, or **const volatile** `T&&`. [\[C++17 15.8.1/2\]](#)
- Example:

```
class T {  
public:  
    T();  
    T(const T&); // copy constructor  
    T(T&&); // move constructor  
    // ...  
};  
  
T a;  
T b(std::move(a)); // calls T::T(T&&)  
T c(b); // calls T::T(const T&)
```

# Move Assignment Operators

- A **move assignment operator** `T::operator=` is a non-static non-template member function of class `T` with exactly one parameter of type `T&&`, **const** `T&&`, **volatile** `T&&`, or **const volatile** `T&&`.

[C++17 15.8.2/3]

- Example:

```
class T {  
public:  
    T();  
    T(const T&); // copy constructor  
    T(T&&); // move constructor  
    T& operator=(const T&); // copy assignment operator  
    T& operator=(T&&); // move assignment operator  
    // ...  
};  
  
T func(int);  
  
T a;  
T b;  
a = func(1); // calls T::operator=(T&&)  
b = a; // calls T::operator=(const T&)
```

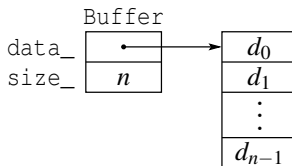


# Buffer Example Revisited

- Recall the class from earlier that represents a character buffer (whose size is fixed at run time).

```
class Buffer {  
public:  
    // ...  
private:  
    char* data_; // pointer to buffer data  
    std::size_t size_; // buffer size (in characters)  
};
```

- Pictorially, the data structure looks like the following:



# Example Without Move Construction/Assignment

```
1  #include <algorithm>
2  #include <cstdint>
3
4  class Buffer {
5  public:
6      Buffer(std::size_t size, char value = 0) :
7          size_(size), data_(new char[size])
8          {std::fill_n(data_, size, value);}
9      Buffer(const Buffer& b) : size_(b.size_), data_(new char[b.size_])
10         {std::copy_n(b.data_, b.size_, data_);}
11      Buffer& operator=(const Buffer& b) {
12          if (this != &b) {
13              delete[] data_;
14              size_ = b.size_; data_ = new char[b.size_];
15              std::copy_n(b.data_, b.size_, data_);
16          }
17          return *this;
18      }
19      ~Buffer() {delete[] data_;}
20  private:
21      char* data_; // pointer to buffer data
22      std::size_t size_; // buffer size (in characters)
23  };
24
25  Buffer getBuffer() {return Buffer(65536, 'A');}
26
27  int main() {
28      Buffer x(0);
29      Buffer y = getBuffer(); // construct from temporary object
30      x = Buffer(32768, 'B'); // assign from temporary object
31  }
```

# Example With Move Construction/Assignment

```
1  #include <algorithm>
2  #include <cstdint>
3  #include <utility>
4
5  class Buffer {
6  public:
7      Buffer(std::size_t size, char value = 0) :
8          size_(size), data_(new char[size])
9          {std::fill_n(data_, size, value);}
10     Buffer(const Buffer& b) : size_(b.size_), data_(new char[b.size_])
11         {std::copy_n(b.data_, b.size_, data_);}
12     Buffer& operator=(const Buffer& b) {
13         if (this != &b) {
14             delete[] data_;
15             size_ = b.size_; data_ = new char[b.size_];
16             std::copy_n(b.data_, b.size_, data_);
17         }
18         return *this;
19     }
20     // Move constructor
21     Buffer(Buffer&& b) : size_(b.size_), data_(b.data_)
22         {b.size_ = 0; b.data_ = nullptr;}
23     // Move assignment operator
24     Buffer& operator=(Buffer&& b) {
25         std::swap(size_, b.size_);
26         std::swap(data_, b.data_);
27         return *this;
28     }
29     ~Buffer() {delete[] data_;}
30 private:
31     char* data_; // pointer to buffer data
32     std::size_t size_; // buffer size (in characters)
33 };
34
35 Buffer getBuffer() {return Buffer(65536, 'A');}
36
37 int main() {
38     Buffer x(0);
39     Buffer y = getBuffer(); // construct from temporary object
40     x = Buffer(32768, 'B'); // assign from temporary object
41 }
```

- As we have seen, a reference parameter of a function that is bound to modifiable rvalue can be modified safely (i.e., no observable change in behavior outside of function).
- Sometimes may want to allow a move to be used instead of a copy, when this would not normally be permitted.
- We can allow moves by casting to a non-const rvalue reference.
- This casting can be accomplished by `std::move`, which is declared (in the header file `utility`) as:

```
template <class T>
constexpr typename std::remove_reference<T>::type&&
    move(T&&) noexcept;
```

- For an object `x` of type `T`, `std::move(x)` is similar to `static_cast<T&&>(x)` but saves typing and still works correctly when `T` is a reference type (a technicality yet to be discussed).

- Prior to C++11, a swap function (such as `std::swap`) would typically look like this:

```
1  template <class T>
2  void swap(T& x, T& y) {
3      T tmp(x); // copy x to tmp
4      x = y;    // copy y to a
5      y = tmp;  // copy tmp to y
6  }
```

- In the above code, a swap requires three *copy* operations (namely, one copy constructor call and two copy assignment operator calls).
- For many types `T`, this use of copying is *very inefficient*.
- Furthermore, the above code requires that `T` *must be copyable* (i.e., `T` has a copy constructor and copy assignment operator).
- In C++11, we can write a much better swap function.

- As of C++11, a swap function would typically look like this:

```
1  template <class T>
2  void swap(T& x, T& y) {
3      T tmp(std::move(x)); // move x to tmp
4      x = std::move(y); // move y to x
5      y = std::move(tmp); // move tmp to y
6  }
```

- The function `std::move` casts its argument to an rvalue reference.
- Assuming that `T` provides a move constructor and move assignment operator, a swap requires three *move* operations (i.e., one move constructor call and two move assignment operator calls) and *no copying*.
- The use of `std::move` above is essential in order for copying to be avoided.

# Implication of Rvalue-Reference Type Function Parameters

- Due to the properties of rvalue references, a function parameter of rvalue-reference type may be regarded as being bound to an object whose value will not be relied upon in the caller.
- Therefore, an object associated with a function parameter of rvalue-reference type can always be safely modified (i.e., without fear of adversely affecting the caller).
- This fact can often be exploited in order to obtain more efficient code.
- Consider the code for a function with the following declaration:

```
void func(std::vector<double>&& x);
```

- Since `x` is of rvalue-reference type, we are guaranteed that the caller will not rely upon the value of the object referenced by `x`.
- If obliterating the value of `x` would allow us to more efficiently implement `func`, we can safely do so.
- For example, we could safely modify `x` in place or move from it, without fear of adversely affecting the caller.

# Reference-Qualified Member Functions

- every nonstatic member function has implicit parameter **\*this**
- possible to provide reference qualifiers for implicit parameter
- allows overloading member functions on lvalueness/rvalueness of **\*this**
- cannot mix reference qualifiers and non-reference qualifiers in single overload set
- provides mechanism for treating lvalue and rvalue cases differently
- useful for facilitating move semantics or preventing operations not appropriate for lvalues or rvalues



# Reference-Qualified Member Functions Example

```
1  #include <iostream>
2
3  class Widget {
4  public:
5      void func() const &
6          {std::cout << "const lvalue\n";}
7      void func() &
8          {std::cout << "non-const lvalue\n";}
9      void func() const &&
10         {std::cout << "const rvalue\n";}
11     void func() &&
12         {std::cout << "non-const rvalue\n";}
13 };
14
15 const Widget getConstWidget() {return Widget();}
16
17 int main(){
18     Widget w;
19     const Widget cw;
20     w.func(); // non-const lvalue
21     cw.func(); // const lvalue
22     Widget().func(); // non-const rvalue
23     getConstWidget().func(); // const rvalue
24 }
```

# Lvalueness/Rvalueness and the `*this` Parameter

```
1  class Int {
2  public:
3      Int(int x = 0) : value_(x) {}
4      // only allow prefix increment for lvalues
5      Int& operator++() & {++value_; return *this;}
6      // The following allows prefix increment for rvalues:
7      // Int& operator++() {++value_; return *this;}
8      // ...
9  private:
10     int value_;
11 };
12
13 int one() {return 1;}
14
15 int main() {
16     int i = 0;
17     int j = ++i; // OK
18     // int k = ++one(); // ERROR (not lvalue)
19     Int x(0);
20     Int y = ++x; // OK
21     // Int z = ++Int(1); // ERROR (not lvalue)
22 }
```

# Move Semantics and the `*this` Parameter

```
1  #include <iostream>
2  #include <vector>
3  #include <utility>
4
5  class Buffer {
6  public:
7      Buffer(char value = 0) : data_(1024, value) {}
8      void data(std::vector<char>& x) const &
9          {x = data_;}
10     void data(std::vector<char>& x) &&
11         {x = std::move(data_);}
12     // ...
13 private:
14     std::vector<char> data_;
15 };
16
17 Buffer getBuffer() {return Buffer(42);}
18
19 int main() {
20     std::vector<char> d;
21     Buffer buffer;
22     buffer.data(d); // copy into d
23     getBuffer().data(d); // move into d
24 }
```

## Section 3.5.5

### **Reference Collapsing and Forwarding References**

# References to References

- A reference to a reference is not allowed, since such a construct clearly makes no sense. [C++17 11.3.2/5]

```
int i = 0;  
int& & j = i; // ILLEGAL: reference to reference
```

- Although one cannot directly create a reference to a reference, a reference to a reference can arise indirectly in several contexts.
- Typedef name:

```
typedef int& RefToInt;  
typedef RefToInt& T; // reference to reference
```

- Template function parameters:

```
template <class T> T func(const T& x) {return x;}  
int x = 1;  
func<int&>(x); // reference to reference
```

- Decltype specifier:

```
int i = 1;  
decltype((i))& j = i; // reference to reference
```

# References to References (Continued)

- Auto specifier:

```
int i = 0;  
auto&& j = i; // reference to reference
```

- Class templates:

```
template <class T>  
struct Thing {  
    void func(T&&) {} // reference to reference  
                        // if T is reference type  
};  
Thing<int&> x;
```

- If, during type analysis, a reference to a reference type is obtained, the reference to reference is converted to a simple reference via a process called **reference collapsing**.

# Reference Collapsing Rules

- Let `TR` denote a type that is a reference to type `T` (where `T` may be cv qualified).
- The effect of reference collapsing is summarized below. [\[C++17 11.3.2/6\]](#).

Before Collapse	After Collapse
<code>TR&amp;</code>	<code>T&amp;</code>
<code>const TR&amp;</code>	<code>T&amp;</code>
<code>volatile TR&amp;</code>	<code>T&amp;</code>
<code>const volatile TR&amp;</code>	<code>T&amp;</code>
<code>TR&amp;&amp;</code>	<code>TR</code>
<code>const TR&amp;&amp;</code>	<code>TR</code>
<code>volatile TR&amp;&amp;</code>	<code>TR</code>
<code>const volatile TR&amp;&amp;</code>	<code>TR</code>

- In other words:
  - An lvalue reference to any reference yields an lvalue reference.
  - An rvalue reference to an lvalue reference yields an lvalue reference.
  - An rvalue reference to an rvalue reference yields rvalue reference.
  - Any cv qualifiers applied to a reference type are discarded (since cv qualifiers cannot be applied to a reference).

# Reference Collapsing Examples

- Due to reference collapsing, T&& syntax may not always be an rvalue reference. Example:

```
using IntRef = int&;  
int i = 0;  
IntRef&& r = i; // r is int& (i.e., lvalue reference)
```

- Example:

```
using IntRef = int&;  
using IntRefRef = int&&;  
using ConstIntRefRef = const int&&;  
using ConstIntRef = const int&;  
using T1 = const IntRef&; // T1 is int&  
using T2 = const IntRefRef&; // T2 is int&  
using T3 = IntRefRef&&; // T3 is int&&  
using T4 = ConstIntRef&&; // T4 is const int&  
using T5 = ConstIntRefRef&&; // T5 is const int&&
```

- Example:

```
int i = 0;  
int& j = i;  
auto&& k = j;  
    // j cannot be inferred to have type int  
    // since rvalue reference cannot be bound to lvalue  
    // j inferred to have type int&  
    // reference collapsing of int& && yields int&
```



# Forwarding References

- A *cv-unqualified* rvalue reference that appears in a type-deducing context for template parameters is called a **forwarding reference**. [\[C++17 17.8.2.1/3\]](#)
- Type deduction for template parameters of template functions is defined in such a way as to facilitate perfect forwarding.
- Consider the following template-parameter type-deduction scenario:

```
template<class T>  
void f(T&& p);
```

```
f(expr); // invoke f
```

- Let  $E$  denote the type of the expression  $expr$ . The type  $T$  is then deduced as follows:
  - 1 If  $expr$  is an *lvalue*,  $T$  is deduced as  $E\&$ , in which case the type of  $p$  yielded by reference collapsing is  $E\&$ .
  - 2 If  $expr$  is an *rvalue*,  $T$  is deduced as  $E$ , in which case  $p$  will have the type  $E\&\&$ .
- Thus, the type  $T\&\&$  will be an lvalue reference type if  $expr$  is an lvalue, and an rvalue reference type if  $expr$  is an rvalue.
- Therefore, the lvalue/rvalue-ness of  $expr$  can be determined *inside*  $f$  based on whether  $T\&\&$  is an lvalue reference type or rvalue reference type.

# Forwarding References Example

```
1  #include <utility>
2
3  template <class T> void f(T&& p);
4  int main() {
5      int i = 42;
6      const int ci = i;
7      const int& rci = i;
8      f(i);
9          // i is lvalue with type int
10         // T is int&
11         // p has type int&
12     f(ci);
13         // ci is lvalue with type const int
14         // T is const int&
15         // p has type const int&
16     f(rci);
17         // rci is lvalue with type const int&
18         // T is const int&
19         // p has type const int&
20     f(2);
21         // 2 is rvalue with type int
22         // T is int
23         // p has type int&&
24     f(std::move(i));
25         // std::move(i) is rvalue with type int&&
26         // T is int
27         // p has type int&&
28 }
```

## Section 3.5.6

### **Perfect Forwarding**

- **Perfect forwarding** is the act of passing a template function's arguments to another function:
  - without rejecting any arguments that can be passed to that other function
  - without losing any information about the arguments' cv-qualifications or lvalue/rvalue-ness; and
  - without requiring overloading.
- In C++03, for example, the best approximations of perfect forwarding turn all rvalues into lvalues and require at least two (and often more) overloads.

# Perfect-Forwarding Example

- Consider a *template* function `wrapper` and another function `func`, each of which takes one argument.
- Suppose that we want to perfectly forward the argument of `wrapper` to `func`.
- The function `wrapper` is to do nothing other than simply call `func`.
- In doing so, `wrapper` must pass its actual argument through to `func`.
- This must be done in such a way that the argument to `wrapper` and argument to `func` have *identical properties* (i.e., match in terms of cv-qualifiers and lvalue/rvalue-ness).
- In other words, the following two function calls must have *identical behavior*, where *expr* denotes an arbitrary expression:  

```
wrapper(expr) ;  
func(expr) ;
```
- The solution to a perfect-forwarding problem, such as this one, turns out to be more difficult than it might first seem.

# Perfect-Forwarding Example: First Failed Attempt

- For our first attempt, we propose the following code for the (template) function `wrapper`:

```
template <class T>
void wrapper(T p) {
    func(p);
}
```

- If `func` takes its parameter by reference, calls to `wrapper` and `func` (with the same argument) can have different behaviors.
- Suppose, for example, that we have the following declarations:

```
void func(int&); // uses pass by reference
int i;
```

- Then, the following two function calls are *not equivalent*:

```
wrapper(i);
// T is deduced as int
// copy of i passed to func
// wrapper cannot change i

func(i);
// i passed by reference
// func can change i
```

- Problem: The original and forwarded arguments are *distinct objects*.

# Perfect-Forwarding Example: Second Failed Attempt

- For our second attempt, we propose the following code for the (template) function `wrapper`:

```
template <class T>
void wrapper(T& p) {
    func(p);
}
```

- If, for example, the function argument is an rvalue (such as a non-string literal or temporary object), calls to `wrapper` and `func` (with the same argument) can have different behaviors.
- Suppose, for example, that we have the following declaration:

```
void func(int); // uses pass by value
```

- Then, the following two function calls are *not equivalent*:

```
wrapper(42);
// T is deduced as int
// ERROR: cannot bind rvalue to
//   nonconst lvalue reference

func(42);
// OK
```

- Problem: The original and forwarded arguments do not match in terms of *lvalue/rvalue-ness*.

# Perfect-Forwarding Example: Third Failed Attempt

- For our third attempt, we propose the following code for the (template) function wrapper:

```
template <class T>
void wrapper(const T& p) {
    func(p);
}
```

- If, for example, the function argument is a non-const object, calls to wrapper and func (with the same argument) will have different behaviors.
- Suppose, for example, that we have the following declaration:

```
void func(int&);
int i;
```

- Then, the following two function calls are *not equivalent*:

```
wrapper(i);
// ERROR: wrapper cannot call func, as this
// would discard const qualifier

func(i);
// OK
```

- Problem: The original and forwarded arguments do not match in terms of *cv-qualifiers*.



# Perfect-Forwarding Example: Solution

- Finally, we propose the following code for the (template) function `wrapper`:

```
template <class T>
void wrapper(T&& p) {
    func(static_cast<T&&>(p));
}
```

- Consider now, for example, the following scenario:

```
int i = 42;
const int ci = i;
int& ri = i;
const int& rci = i;
wrapper(expr); // invoke wrapper
```

- The parameter `p` is an alias for the object yielded by the expression `expr`.
- The argument `expr` and argument to `func` match in terms of cv-qualifiers and lvalue/rvalue-ness.

<i>expr</i>	<i>expr</i>		T	argument to <i>func</i>	
	Type	Category		Type (T&&)	Category
<code>i</code>	<b>int</b>	lvalue	<b>int&amp;</b>	<b>int&amp;</b>	lvalue
<code>ci</code>	<b>const int</b>	lvalue	<b>const int&amp;</b>	<b>const int&amp;</b>	lvalue
<code>ri</code>	<b>int&amp;</b>	lvalue	<b>int&amp;</b>	<b>int&amp;</b>	lvalue
<code>rci</code>	<b>const int&amp;</b>	lvalue	<b>const int&amp;</b>	<b>const int&amp;</b>	lvalue
<code>42</code>	<b>int</b>	rvalue	<b>int</b>	<b>int&amp;&amp;</b>	rvalue

## Perfect-Forwarding Example: Solution (Continued)

- Although we only considered one specific scenario on the previous slide, the solution works in general.
- That is, the `wrapper` function from the previous slide will perfectly forward its single argument, regardless of what the argument happens to be (or which overload of `func` is involved).
- Thus, we have a general solution to the perfect-forwarding problem in the single-argument case.
- This solution is easily extended to an arbitrary number of arguments.

# The `std::forward` Template Function

- To avoid the need for an explicit type-cast operation when forwarding an argument, the standard library provides the `std::forward` function specifically for performing such a type conversion.
- The template function `forward` is defined as:

```
template<class T>
T&& forward(typename std::remove_reference<T>::type& x)
    noexcept {
    return static_cast<T&&>(x);
}
```

- A typical usage of `forward` might look something like:

```
template <class T1, class T2>
void wrapper(T1&& x1, T2&& x2) {
    func(std::forward<T1>(x1), std::forward<T2>(x2));
}
```

- The expression `forward<T>(a)` is an lvalue if `T` is an lvalue reference type and an rvalue otherwise.
- The use of `std::forward` instead of an explicit type cast improves code readability by making the programmer's intent clear.

# Perfect-Forwarding Example Revisited

- We now revisit the perfect-forwarding example from earlier.
- In the earlier example, perfect forwarding was performed by the following function:

```
template <class T>
void wrapper(T&& e) {
    func(static_cast<T&&>(e));
}
```

- The above code can be made more readable, however, by rewriting it to make use of `std::forward` as follows:

```
template <class T>
void wrapper(T&& e) {
    func(std::forward<T>(e));
}
```

# Forwarding Example

```
1  #include <iostream>
2  #include <string>
3  #include <utility>
4
5  void func(std::string& s) {
6      std::cout << "func(std::string&) called\n";
7  }
8
9  void func(std::string&& s) {
10     std::cout << "func(std::string&&) called\n";
11 }
12
13 template <class T>
14 void wrapper(T&& x) {
15     func(std::forward<T>(x));
16 }
17
18 template <class T>
19 void buggy_wrapper(T x) {func(x);}
20
21 int main() {
22     using namespace std::literals;
23     std::string s("Hi"s);
24     wrapper(s);           // which overload of func called?
25     buggy_wrapper(s);     // which overload of func called?
26     wrapper("Hi"s);       // which overload of func called?
27     buggy_wrapper("Hi"s); // which overload of func called?
28 }
```

# Perfect-Forwarding Use Case: Wrapper Functions

- A **wrapper function** is simply a function used to invoke another function, possibly with some additional processing.
- Example:

```
1  #include <iostream>
2  #include <utility>
3  #include <string>
4
5  std::string emphasize(const std::string& s)
6      {return s + "!";}
7
8  std::string emphasize(std::string&& s)
9      {return s + "!!!!";}
10
11 template <class A>
12 auto wrapper(A&& arg) {
13     std::cout << "Calling with argument " << arg << '\n';
14     auto result = emphasize(std::forward<A>(arg));
15     std::cout << "Return value " << result << '\n';
16     return result;
17 }
18
19 int main() {
20     std::string s("Bonjour");
21     wrapper(s);
22     wrapper(std::string("Hello"));
23 }
```

# Perfect-Forwarding Use Case: Factory Functions

- A **factory function** is simply a function used to create objects.
- Often, perfect forwarding is used by factory functions in order to pass arguments through to a constructor, which performs the actual object creation.
- Example:

```
1  #include <iostream>
2  #include <string>
3  #include <complex>
4  #include <utility>
5  #include <memory>
6
7  // Make an object of type T.
8  template<typename T, typename Arg>
9  std::shared_ptr<T> factory(Arg&& arg) {
10     return std::shared_ptr<T>(
11         new T(std::forward<Arg>(arg)));
12 }
13
14 int main() {
15     using namespace std::literals;
16     auto s(factory<std::string>("Hello"s));
17     auto z(factory<std::complex<double>>(1.0i));
18     std::cout << *s << ' ' << *z << '\n';
19 }
```

# Perfect-Forwarding Use Case: Emplace Operations

- Many container classes provide an operation that creates a new element directly inside the container, often referred to as an **emplace operation**.
- Some or all of the arguments to a member function performing an emplace operation correspond to arguments for a constructor invocation.
- Thus, an emplace operation typically employs perfect forwarding.
- The member function performing the emplace operation forwards some or all of its arguments to the constructor responsible for actually creating the new object.
- Some examples of emplace operations in the standard library include:
  - `std::list` **class**: `emplace`, `emplace_back`, `emplace_front`
  - `std::vector` **class**: `emplace`, `emplace_back`
  - `std::set` **class**: `emplace`, `emplace_hint`
  - `std::forward_list` **class**: `emplace_front`, `emplace_after`



# Other Perfect-Forwarding Examples

- `std::thread` constructor uses forwarding to pass through arguments to thread function
- `std::packaged_task` function-call operator uses forwarding to pass through arguments to associated function
- `std::async` uses forwarding to pass through arguments to specified callable entity
- `std::make_unique` forwards arguments to `std::unique_ptr` constructor
- `std::make_shared` forwards arguments to `std::shared_ptr` constructor

## Section 3.5.7

### **References**

- 1 S. Meyers. **Universal references in C++11**. *Overload*, 111:8–12, Oct. 2012.
- 2 T. Becker, C++ Rvalue References Explained, 2013, [http://thbecker.net/articles/rvalue\\_references/section\\_01.html](http://thbecker.net/articles/rvalue_references/section_01.html)
- 3 E. Bendersky, Understanding Lvalues and Rvalues in C and C++, 2011, <http://eli.thegreenplace.net/2011/12/15/understanding-lvalues-and-rvalues-in-c-and-c>
- 4 E. Bendersky, Perfect Forwarding and Universal References in C++, 2014, <http://eli.thegreenplace.net/2014/perfect-forwarding-and-universal-references-in-c/>
- 5 M. Kilpelainen. **Lvalues and rvalues**. *Overload*, 61:12–13, June 2004.
- 6 H. E. Hinnant, Forward, ISO/IEC JTC1/SC22/WG21/N2951, Sept. 27, 2009, <http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2009/n2951.html>

- 7 H. Hinnant and D. Krugler, Proposed Wording for US 90, ISO/IEC JTC1/SC22/WG21/N3143, Oct. 15, 2010, <http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2010/n3143.html>
- 8 R. Smith, Guaranteed copy elision through simplified value categories, ISO/IEC JTC1/SC22/WG21 P0135R0, Sept. 27, 2015, <http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2015/p0135r0.html>

- 1 S. Meyers. Universal References in C++11. C++ and Beyond, Asheville, NC, USA, Aug. 5–8, 2012. Available online at <https://channel9.msdn.com/Shows/GoingDeep/Cpp-and-Beyond-2012-Scott-Meyers-Universal-References-in-Cpp11>.  
This talk discusses rvalue/forwarding references.
- 2 S. Meyers. Adventures in Perfect Forwarding, Facebook C++ Conference. Menlo Park, CA, USA, June 2, 2012. Available online at <https://www.facebook.com/Engineering/videos/10151094464083109>.  
This talk introduces perfect forwarding and discusses matters such as how to specialize forwarding templates and how to address interactions between forwarding and the pimpl idiom.
- 3 H. Hinnant. Everything You Ever Wanted to Know About Move Semantics. Inside Bloomberg, July 25, 2016. Available online at <https://youtu.be/vLinb2fgkHk>.  
This talk discusses various aspects of move semantics.

## Section 3.6

# Exceptions

## Section 3.6.1

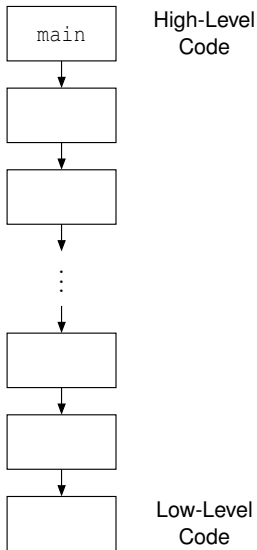
### **Introduction**

# Exceptions

- exceptions are language mechanism for handling exceptional (i.e., abnormal) situations
- exceptional situation perhaps best thought of as case when code could not do what it was asked to do and usually (but not always) corresponds to error condition
- exceptions often employed for error handling
- exceptions propagate information from point where error *detected* to point where error *handled*
- code that encounters error that it is unable to handle throws exception
- code that wants to handle error catches exception and performs processing necessary to handle error
- exceptions provide convenient way in which to *separate error detection from error handling*



# The Problem



- error detected in low-level code
- want to handle error in high-level code
- must propagate error information up call chain

# Traditional Error Handling

- if any error occurs, terminate program
  - overly draconian
- pass error code back from function (via return value, reference parameter, or global object) and have caller check error code
  - errors are ignored by default (i.e., explicit action required to check for error condition)
  - caller may forget to check error code allowing error to go undetected
  - code can become cluttered with many checks of error codes, which can adversely affect code readability and maintainability
- call error handler if error detected
  - may not be possible or practical for handler to recover from particular error (e.g., handler may not have access to all information required to recover from error)

# Example: Traditional Error Handling

```
1  #include <iostream>
2
3  bool func3() {
4      bool success = false;
5      // ...
6      return success;
7  }
8
9  bool func2() {
10     if (!func3()) {return false;}
11     // ...
12     return true;
13 }
14
15 bool func1() {
16     if (!func2()) {return false;}
17     // ...
18     return true;
19 }
20
21 int main() {
22     if (!func1()) {
23         std::cout << "failed\n";
24         return 1;
25     }
26     // ...
27 }
```

# Error Handling With Exceptions

- when error condition detected, signalled by throwing exception (with **throw** statement)
- exception is object that describes error condition
- thrown exception caught by handler (in **catch** clause of **try** statement), which takes appropriate action to handle error condition associated with exception
- handler can be in different function from where exception thrown
- error-free code path tends to be relatively simple, since no need to explicitly check for error conditions
- error condition less likely to go undetected, since uncaught exception terminates program

# Example: Exceptions

```
1  #include <iostream>
2  #include <stdexcept>
3
4  void func3() {
5      bool success = false;
6      // ...
7      if (!success) {throw std::runtime_error("Yikes!");}
8  }
9
10 void func2() {
11     func3();
12     // ...
13 }
14
15 void func1() {
16     func2();
17     // ...
18 }
19
20 int main() {
21     try {func1();}
22     catch (...) {
23         std::cout << "failed\n";
24         return 1;
25     }
26     // ...
27 }
```

## safe\_divide Example: Traditional Error Handling

```
1  #include <iostream>
2  #include <vector>
3  #include <utility>
4
5  std::pair<bool, int> safe_divide(int x, int y) {
6      if (!y) {
7          return std::pair(false, 0);
8      }
9      return std::pair(true, x / y);
10 }
11
12 int main() {
13     std::vector<std::pair<int, int>> v = {{10, 2}, {10, 0}};
14     for (auto p : v) {
15         auto result = safe_divide(p.first, p.second);
16         if (result.first) {
17             int quotient = result.second;
18             std::cout << quotient << '\n';
19         } else {
20             std::cerr << "division by zero\n";
21         }
22     }
23 }
```

# safe\_divide Example: Exceptions

```
1  #include <iostream>
2  #include <stdexcept>
3  #include <utility>
4  #include <vector>
5
6  int safe_divide(int x, int y) {
7      if (!y) {
8          throw std::overflow_error("divide by zero");
9      }
10     return x / y;
11 }
12
13 int main() {
14     std::vector<std::pair<int, int>> v = {{10, 2}, {10, 0}};
15     for (auto p : v) {
16         try {
17             std::cout << safe_divide(p.first, p.second) <<
18                 '\n';
19         }
20         catch(const std::overflow_error& e) {
21             std::cerr << "division by zero\n";
22         }
23     }
24 }
```

# safe\_add Example: Traditional Error Handling

SKIP SLIDE

```
1  #include <limits>
2  #include <vector>
3  #include <iostream>
4
5  std::pair<bool, int> safe_add(int x, int y) {
6      return ((y > 0 && x > std::numeric_limits<int>::max() - y)
7             || (y < 0 && x < std::numeric_limits<int>::min() - y)) ?
8          std::pair(false, 0) : std::pair(true, x + y);
9  }
10
11 int main() {
12     constexpr int int_min = std::numeric_limits<int>::min();
13     constexpr int int_max = std::numeric_limits<int>::max();
14     std::vector<std::pair<int, int>> v{
15         {int_max, int_max}, {1, 2}, {int_min, int_min},
16         {int_max, int_min}, {int_min, int_max}
17     };
18     for (auto x : v) {
19         auto result = safe_add(x.first, x.second);
20         if (result.first) {
21             std::cout << result.second << '\n';
22         } else {
23             std::cout << "overflow\n";
24         }
25     }
26 }
```



# safe\_add Example: Exceptions

SKIP SLIDE

```
1  #include <limits>
2  #include <vector>
3  #include <iostream>
4  #include <stdexcept>
5
6  int safe_add(int x, int y) {
7      return ((y > 0 && x > std::numeric_limits<int>::max() - y)
8              || (y < 0 && x < std::numeric_limits<int>::min() - y)) ?
9          throw std::overflow_error("addition") : x + y;
10 }
11
12 int main() {
13     constexpr int int_min = std::numeric_limits<int>::min();
14     constexpr int int_max = std::numeric_limits<int>::max();
15     std::vector<std::pair<int, int>> v{
16         {int_max, int_max}, {1, 2}, {int_min, int_min},
17         {int_max, int_min}, {int_min, int_max}
18     };
19     for (auto x : v) {
20         try {
21             int result = safe_add(x.first, x.second);
22             std::cout << result << '\n';
23         }
24         catch (const std::overflow_error&) {
25             std::cout << "overflow\n";
26         }
27     }
28 }
```

# Exceptions Versus Traditional Error Handling

## ■ advantages of exceptions:

- exceptions allow for error handling code to be easily separated from code that detects error
- exceptions can easily pass error information many levels up call chain
- passing of error information up call chain managed by language (no explicit code required)

## ■ disadvantages of exceptions:

- writing code that always behaves correctly in presence of exceptions requires great care (as we shall see)
- although possible to have no execution-time cost when exceptions not thrown, still have memory cost (to store information needed for stack unwinding for case when exception is thrown)

## Section 3.6.2

# Exceptions

# Exceptions

- exceptions are objects
- type of object used to indicate kind of error
- value of object used to provide details about particular occurrence of error
- exception object can have any type (built-in or class type)
- for convenience, standard library provides some basic exception types
- all exception classes in standard library derived (directly or indirectly) from `std::exception` class
- exception object is propagated from one part of code to another by throwing and catching
- exception processing disrupts normal control flow

# Standard Exception Classes

## Exception Classes Derived from `exception` Class

Type	Description
<code>logic_error</code>	faulty logic in program
<code>runtime_error</code>	error caused by circumstances beyond scope of program
<code>bad_typeid</code>	invalid operand for <b><code>typeid</code></b> operator
<code>bad_cast</code>	invalid expression for <b><code>dynamic_cast</code></b>
<code>bad_weak_ptr</code>	<b><code>bad</code></b> <code>weak_ptr</code> given
<code>bad_function_call</code>	function has no target
<code>bad_alloc</code>	storage allocation failure
<code>bad_exception</code>	use of invalid exception type in certain contexts
<code>bad_variant_access</code>	variant accessed in invalid way

# Standard Exception Classes (Continued 1)

## Exception Classes Derived from `bad_cast` Class

Type	Description
<code>bad_any_cast</code>	invalid cast for any

## Exception Classes Derived from `logic_error` Class

Type	Description
<code>domain_error</code>	domain error (e.g., square root of negative number)
<code>invalid_argument</code>	invalid argument
<code>length_error</code>	length too great (e.g., resize vector beyond <code>max_size</code> )
<code>out_of_range</code>	out of range argument (e.g., subscripting error in <code>vector::at</code> )
<code>future_error</code>	invalid operations on future objects
<code>bad_optional_access</code>	optional accessed in invalid way

## Standard Exception Classes (Continued 2)

### Exception Classes Derived from `runtime_error` Class

Type	Description
<code>range_error</code>	range error
<code>overflow_error</code>	arithmetic overflow error
<code>underflow_error</code>	arithmetic underflow error
<code>regex_error</code>	error in regular expressions library
<code>system_error</code>	operating-system or other low-level error

### Exception Classes Derived from `runtime_error::system_error` Class

Type	Description
<code>ios_base::failure</code>	I/O failure

## Section 3.6.3

# **Throwing and Catching Exceptions**



# Throwing Exceptions

- throwing exception accomplished by **throw** statement
- throwing exception transfers control to handler
- object is passed
- type of object determines which handlers can catch it
- handlers specified with **catch** clause of **try** block
- for example

```
throw "OMG!";
```

can be caught by handler of **const char\*** type, as in:

```
try {  
    // ...  
}  
catch (const char* p) {  
    // handle character string exceptions here  
}
```

# Throwing Exceptions (Continued)

- throw statement initializes temporary object called **exception object**
- type of exception object determined by *static* type of operand of **throw** (so slicing can occur) [\[C++17 18.1/3\]](#)
- if thrown object is class object, copy/move constructor and destructor must be accessible [\[C++17 18.1/5\]](#)
- temporary may be moved/copied several times before caught
- advisable for type of exception object to be user defined to reduce likelihood of different parts of code using type in conflicting ways

# Catching Exceptions

- exception can be caught by **catch** clause of **try-catch** block
- code that might throw exception placed in **try** block
- code to handle exception placed in **catch** block
- **try-catch** block can have multiple **catch** clauses
- **catch** clauses checked for match in order specified and only first match used
- **catch** (...) can be used to catch any exception
- example:

```
try {  
    // code that might throw exception  
}  
catch (const std::logic_error& e) {  
    // handle logic_error exception  
}  
catch (const std::runtime_error& e) {  
    // handle runtime_error exception  
}  
catch (...) {  
    // handle other exception types  
}
```

- catch exceptions by reference in order to:
  - avoid copying, which might throw
  - allow exception object to be modified and then rethrown
  - avoid slicing

# Exception During Exception: Catching By Value

SKIP SLIDE

```
1  #include <iostream>
2  #include <stdexcept>
3
4  class Error {
5  public:
6      Error(int value) : value_(value) {}
7      Error(Error&& e) : value_(e.value_) {}
8      Error(const Error&) throw std::runtime_error("copy");
9      int get() const {return value_;}
10 private:
11     int value_; // error code
12 };
13
14 void func2() {throw Error(42);} // might move
15
16 void func1() {
17     try {func2();}
18     // catch by value (copy throws)
19     catch (Error e) {
20         std::cerr << "yikes\n";
21     }
22 }
23
24 int main() {
25     try {func1();}
26     catch (...) {std::cerr << "exception\n";}
27 }
```

# Throwing Polymorphically: Failed Attempt

SKIP SLIDE

```
1  #include <iostream>
2
3  class Base {};
4  class Derived : public Base {};
5
6  void func(Base& x) {
7      throw x; // always throws Base
8  }
9
10 int main() {
11     Derived d;
12     try {func(d);}
13     catch (Derived& e) {
14         std::cout << "Derived\n";
15     }
16     catch (...) {
17         std::cout << "not Derived\n";
18     }
19 }
```

- type of exception object determined from *static* type of throw expression

# Throwing Polymorphically

SKIP SLIDE

```
1  #include <iostream>
2
3  class Base {
4  public:
5      virtual void raise() {throw *this;}
6  };
7  class Derived : public Base {
8  public:
9      virtual void raise() {throw *this;}
10 };
11
12 void func(Base& x) {
13     x.raise();
14 }
15
16 int main() {
17     Derived d;
18     try {func(d);}
19     catch (Derived& e) {
20         std::cout << "Derived\n";
21     }
22     catch (...) {
23         std::cout << "not Derived\n";
24     }
25 }
```

# Rethrowing Exceptions

- caught exception can be rethrown by **throw** statement with no operand
- example:

```
try {  
    // code that may throw exception  
}  
catch (...) {  
    throw; // rethrow caught exception  
}
```



# Rethrowing Example: Exception Dispatcher Idiom

SKIP SLIDE

```
1  void handle_exception() {
2      try {throw;}
3      catch (const exception_1& e) {
4          log_error("exception_1 occurred");
5          // ...
6      }
7      catch (const exception_2& e) {
8          log_error("exception_2 occurred");
9          // ...
10     }
11     // ...
12 }
13
14 void func() {
15     try {operation();}
16     catch (...) {handle_exception();}
17     // ...
18     try {another_operation();}
19     catch (...) {handle_exception();}
20 }
```

- allows reuse of exception handling code

# Transfer of Control from Throw Site to Handler

- when exception is thrown, control is transferred to nearest handler (in catch clause) with matching type, where “nearest” means handler for try block most recently entered (by thread) and not yet exited [\[C++17 18.1/2\]](#)
- if no matching handler found, `std::terminate()` is called [\[C++17 18.3/9\]](#)
- as control passes from throw expression to handler, destructors are invoked for all automatic objects constructed since try block entered, where automatic objects destroyed in reverse order of construction [\[C++17 18.2/2\]](#)
- process of calling destructors for automatic objects constructed on path from try block to throw expression called **stack unwinding** [\[C++17 18.2/1\]](#)
- object not deemed to be constructed if constructor exits due to exception (in which case destructor will not be invoked)
- do not throw exception in destructor since destructors called during exception processing and throwing exception during exception processing will terminate program

# Stack Unwinding Example

```
1  void func1() {
2      std::string dave("dave");
3      try {
4          std::string bye("bye");
5          func2();
6      }
7      catch (const std::runtime_error& e) { // Handler
8          std::cerr << e.what() << '\n';
9      }
10 }
11
12 void func2() {
13     std::string world("world");
14     func3(0);
15 }
16
17 void func3(int x) {
18     std::string hello("hello");
19     if (x == 0) {
20         std::string first("first");
21         std::string second("second");
22         throw std::runtime_error("yikes"); // Throw site
23     }
24 }
```

- calling func1 will result in exception being thrown in func3
- during stack unwinding, destructors called in order for second, first, hello, world, and bye (i.e., reverse order of construction); dave unaffected

# Another Stack Unwinding Example

## SKIP SLIDE

```
1 #include <iostream>
2 #include <stdexcept>
3 #include <string>
4 #include <utility>
5
6 class Widget {
7 public:
8     Widget(const std::string& s) : s_(s) {}
9     Widget(const Widget& other) : s_(other.s_) {std::cerr << "copy ctor\n";}
10    Widget(Widget&& other) : s_(std::move(other.s_)) {std::cerr << "move ctor\n";}
11    std::string get() const {return s_;}
12    ~Widget() {std::cerr << "dtor " << s_ << '\n';}
13 private:
14     std::string s_;
15 };
16
17 Widget func_1(Widget w) {
18     Widget bjarne("bjarne");
19     {
20         Widget hello("hello");
21         Widget bye("bye");
22         throw std::runtime_error("Yikes");
23         Widget bonjour("bonjour");
24     }
25     return bjarne;
26 }
27
28 int main() {
29     Widget zaphod("zaphod");
30     try {
31         Widget ford("ford");
32         Widget u = func_1(zaphod);
33         Widget arthur("arthur");
34     } catch (...) {std::cerr << "exception\n";}
35 }
```

■ objects destroyed during stack unwinding: bye, hello, bjarne, w, ford

# Function Try Blocks

- function try blocks allow entire function to be wrapped in try block
- function returns when control flow reaches end of catch block (return statement needed for non-void function)

- example:

```
1  #include <iostream>
2  #include <stdexcept>
3
4  int main()
5  try {
6      throw std::runtime_error("yikes");
7  }
8  catch (const std::runtime_error& e) {
9      std::cerr << "runtime error " << e.what() << '\n';
10 }
```

- although function try blocks can be used for any function, most important use cases are for constructors and destructors
- function try block only way to catch exceptions thrown during construction of data members or base objects (which happens before constructor body is entered) or during destruction of data members or base objects (which happens after destructor body exited)

# Exceptions and Construction/Destruction

- order of construction (assuming no virtual base classes):
  - 1 non-virtual base class objects as listed from left to right in class definition
  - 2 non-static data members as listed from top to bottom in class definition
  - 3 constructor body
- order of destruction is exact reverse of order of construction, namely:
  - 1 destructor body
  - 2 non-static data members as listed from bottom to top in class definition
  - 3 non-virtual base class objects as listed from right to left in class definition
- lifetime of object begins when constructor completes
- constructor might throw in:
  - constructor of base class object
  - constructor of data member
  - constructor body
- need to perform cleanup for constructor body
- will assume destructors do not throw (since very bad idea to throw in destructor)
- any exception caught in function try block of constructor or destructor rethrown implicitly (at end of catch block) [C++17 18.3/14](#)

# Construction/Destruction Example

```
1  #include <string>
2  #include <iostream>
3
4  struct Base {
5      Base() {}
6      ~Base() {};
7  };
8
9  class Widget : public Base {
10 public:
11     Widget() {}
12     ~Widget() {}
13     // ...
14 private:
15     std::string s_;
16     std::string t_;
17 };
18
19 int main() {
20     Widget w;
21     // ...
22 }
```

# Function Try Block Example

```
1  #include <iostream>
2  #include <stdexcept>
3
4  class Gadget {
5  public:
6      Gadget() {throw std::runtime_error("ctor");}
7      ~Gadget() {}
8  };
9
10 class Widget {
11 public:
12     // constructor uses function try block
13     Widget()
14     try {std::cerr << "ctor body\n";}
15     catch (...) {std::cerr << "exception in ctor\n";}
16     ~Widget() {std::cerr << "dtor body\n";}
17 private:
18     Gadget g_;
19 };
20
21 int main()
22 try {Widget w;}
23 catch (...) {
24     std::cerr << "terminating due to exception\n";
25     return 1;
26 }
```



# Function Try Block Example [Constructor]

SKIP SLIDE

```
1  #include <iostream>
2  #include <stdexcept>
3
4  struct Gadget {
5      Gadget(bool b) {
6          if (b) {throw std::runtime_error("yikes");}
7      }
8  };
9
10 struct Widget : public Gadget {
11     Widget(bool b) try : Gadget(b) {}
12     catch(const std::exception& e) {
13         std::cerr << e.what() << '\n';
14         // exception automatically rethrown
15     }
16 };
17
18 int main() try {
19     Widget v(true);
20 } catch (const std::exception& e) {
21     std::cerr << e.what() << '\n';
22 }
```

# Function Try Block Example [Destructor]

SKIP SLIDE

```
1  #include <iostream>
2  #include <stdexcept>
3
4  struct Gadget {
5      ~Gadget() noexcept(false)
6          {throw std::runtime_error("yikes");}
7  };
8
9  class Widget : public Gadget {
10 public:
11     ~Widget() try {}
12     catch(const std::exception& e) {
13         std::cerr << e.what() << '\n';
14         // exception automatically rethrown
15     }
16 };
17
18 int main() try {
19     Widget w;
20 } catch (const std::exception& e) {
21     std::cerr << e.what() << '\n';
22 }
```

# Function Try Block Example [Normal Function]

SKIP SLIDE

```
1  #include <iostream>
2  #include <stdexcept>
3
4  int func(int x) try {
5      throw std::runtime_error("whatever");
6  } catch(const std::exception& e) {
7      std::cerr << e.what() << '\n';
8      return x;
9      // exception not automatically rethrown
10     // function does not emit exception
11 }
12
13 int main() {
14     std::cout << func(42) << '\n';
15 }
```

## Section 3.6.4

# **Exception Specifications**

# The **noexcept** Specifier

- **noexcept** specifier in function declaration indicates whether or not function can throw exceptions
- **noexcept** specifier with **bool** constant expression argument indicates function does not throw exceptions if expression **true** (otherwise, may throw)
- **noexcept** without argument equivalent to **noexcept (true)**
- except for destructors, not providing **noexcept** specifier equivalent to **noexcept (false)**
- if **noexcept** specifier not provided for destructor, specifier identical to that of implicit declaration (which is, in practice, usually **noexcept**)

[\[C++17 15.4/3\]](#)

- example:

```
void func1(); // may throw anything
void func2() noexcept(false); // may throw anything
void func3() noexcept(true); // does not throw
void func4() noexcept; // does not throw
template <class T>
void func5(T) noexcept(sizeof(T) <= 4);
    // does not throw if sizeof(T) <= 4
```

# The **noexcept** Specifier (Continued 1)

- exception specification for function is *part of function's type*
- example:  

```
void f() noexcept;  
auto g = f; // g is noexcept
```
- exception specification for function is *not part of function's signature*
- consequently, cannot overload on noexcept specifier
- example:  

```
void f();  
void f() noexcept;  
// ERROR: both functions have same signature
```

## The **noexcept** Specifier (Continued 2)

- nontrivial **bool** expression for **noexcept** specifier often useful in templates
- example (swap function):

```
1  #include <type_traits>
2  #include <utility>
3
4  // swap two values
5  template <class T>
6  void exchange(T& a, T& b) noexcept (
7      std::is_nothrow_move_constructible_v<T> &&
8      std::is_nothrow_move_assignable_v<T>) {
9      T tmp(std::move(a)); // move construction
10     a = std::move(b); // move assignment
11     b = std::move(tmp); // move assignment
12 }
```

# The **noexcept** Specifier (Continued 3)

- if function with **noexcept (true)** specifier throws exception, `std::terminate` is called immediately

- example:

```
// This function will terminate the program.  
void die_die_die() noexcept {  
    throw 0;  
}
```

- advisable not to use **noexcept (true)** specifier unless clear that no reasonable usage of function can throw (in current or *any future* version of code)
- in practice, can often be difficult to guarantee that function will never throw exception (especially when considering *all future* versions of code)



# Exceptions and Function Calls

- for some (nonreference) class type `T` and some constant `bool` expression `expr`, consider code such as:

```
T func(T) noexcept (expr);  
T x;  
T y = func(x); // function call
```

- function call can throw exception as result of:
  - 1 parameter passing (if pass by value)
  - 2 function execution *including return statement*
- in parameter passing, construction and destruction of each parameter happens in context of *calling* function [C++17 8.2.2/4]
- consequently, invocation of `noexcept` function can still result in exception being thrown due to parameter passing
- in case of return by value, construction of temporary (if not elided) to hold return value happens in context of *called* function [C++17 9.6.3/2] [C++17 9.6.3/3]
- consequently, must exercise care not to violate `noexcept` contract if `noexcept` function returns by value

- if exception due to parameter passing must be avoided:
  - pass by reference; or
  - ensure **noexcept** move and/or copy constructor as appropriate; or
  - ensure function invoked in manner such that copy elision is guaranteed
- if exception due to return by value must be avoided:
  - ensure **noexcept** move or copy constructor as appropriate; or
  - ensure that function invoked in manner such that copy elision is guaranteed

# noexcept Operator

- **noexcept** operator takes expression and returns **bool** indicating if expression can throw exception
- does not actually evaluate expression
- in determining result, only considers **noexcept** specifications for functions involved
- example:

```
1  #include <cstdlib>
2  #include <utility>
3
4  void increment(int&) noexcept;
5  char* memAlloc(std::size_t);
6
7  // does not throw exception, but not declared noexcept
8  void doesNotThrow() {}
9
10 int main() {
11     static_assert(noexcept(1 + 1) == true);
12     static_assert(noexcept(memAlloc(0)) == false);
13     // Note: does not evaluate expression
14     static_assert(noexcept(increment(*(int*)0)) == true);
15     static_assert(noexcept(increment(std::declval<int&>())) ==
16         true);
17     // Note: only uses noexcept specifiers
18     static_assert(noexcept(doesNotThrow()) == false);
19 }
```

# noexcept Operator (Continued)

- **noexcept** operator particularly useful for templates

- example:

```
1  #include <iostream>
2  #include <type_traits>
3
4  class Int256 { /* ... */ }; // 256-bit integer
5  class BigInt { /* ... */ }; // arbitrary-precision integer
6
7  // function will not throw exception
8  Int256 operator+(const Int256& x, const Int256& y) noexcept;
9
10 // function may throw exception
11 BigInt operator+(const BigInt& x, const BigInt& y);
12
13 // whether function may throw exception depends on T
14 template <class T>
15 T add(const T& x, const T& y) noexcept(noexcept(x + y) &&
16     std::is_nothrow_move_constructible_v<T>)
17 {return x + y;}
18
19 int main() {
20     Int256 i1, i2;
21     BigInt b1, b2;
22     std::cout << "int " << noexcept(add(1, 1)) << '\n'
23         << "Int256 " << noexcept(add(i1, i2)) << '\n'
24         << "BigInt " << noexcept(add(b1, b2)) << '\n';
25 }
```

## Section 3.6.5

### **Storing and Retrieving Exceptions**

# Storing and Retrieving Exceptions

- might want to store exception and then later retrieve and rethrow it
- exception can be stored using `std::exception_ptr` type
- current exception can be retrieved with `std::current_exception`
- rethrow exception stored in `exception_ptr` object using `std::rethrow_exception`
- provides mechanism for moving exceptions between threads:
  - store exception on one thread
  - then retrieve and rethrow stored exception on another thread
- `std::make_exception_ptr` can be used to make `exception_ptr` object

# Example: Storing and Retrieving Exceptions

```
1  #include <exception>
2  #include <stdexcept>
3
4  void yikes() {
5      throw std::runtime_error("Yikes!");
6  }
7
8  std::exception_ptr getException() {
9      try {
10         yikes();
11     }
12     catch (...) {
13         return std::current_exception();
14     }
15     return nullptr;
16 }
17
18 int main() {
19     std::exception_ptr e = getException();
20     std::rethrow_exception(e);
21 }
```

## Section 3.6.6

### **Exception Safety**



- **resource**: physical or virtual component of limited availability within computer system
- examples of resources include: memory, files, devices, network connections, processes, threads, and locks
- essential that acquired resource properly released when no longer needed
- when resource not properly released when no longer needed, **resource leak** said to occur
- exceptions have important implications in terms of resource management
- must be careful to avoid resource leaks

# Resource Leak Example

```
1 void useBuffer(char* buf) { /* ... */ }
2
3 void doWork() {
4     char* buf = new char[1024];
5     useBuffer(buf);
6     delete[] buf;
7 }
```

- if `useBuffer` throws exception, code that deletes `buf` is never reached

- cleanup operations should always be performed in destructors
- following structure for code is *fundamentally flawed*:

```
void func()
{
    initialize();
    do_work();
    cleanup();
}
```

- code with preceding structure *not exception safe*
- if `do_work` throws exception, `cleanup` never called and cleanup operation not performed
- in best case, not performing cleanup will probably cause resource leak

# Exception Safety and Exception Guarantees

- in order for exception mechanism to be useful, must know what can be assumed about state of program when exception thrown
- operation said to be **exception safe** if it leaves program in valid state when operation is terminated by exception
- several levels of exception safety: basic, strong, nothrow
- **basic guarantee**: all invariants preserved and no resources leaked
- with basic guarantee, partial execution of failed operation may cause side effects
- **strong guarantee**: in addition to basic guarantee, failed operation guaranteed to have no side effects (i.e., commit semantics)
- with strong guarantee, operation can still fail causing exception to be thrown
- **nothrow guarantee**: in addition to basic guarantee, promises not to emit exception (i.e., operation guaranteed to succeed even in presence of exceptional circumstances)

# Exception Guarantees

- assume all functions throw if not known otherwise
- code must always provide basic guarantee
- nothrow guarantee should always be provided by destructors
- whenever possible, nothrow guarantee should be provided by:
  - move operations (i.e., move constructors and move assignment operators)
  - swap operations
- provide strong guarantee when natural to do so and not more costly than basic guarantee
- examples of strong guarantee:
  - `push_back` for container, subject to certain container-dependent conditions being satisfied (e.g., for `std::vector`, element type has nonthrowing move or is copyable)
  - `insert` on `std::list`
- examples of nothrow guarantee:
  - swap of two containers
  - `pop_back` for container

# Resource Acquisition Is Initialization (RAII)

- resource acquisition is initialization (RAII) is programming idiom used to *avoid resource leaks* and *provide exception safety*
- associate resource with owning object (i.e., RAII object)
- period of time over which resource held is tied to lifetime of RAII object
- resource acquired during creation of RAII object
- resource released during destruction of RAII object
- provided RAII object properly destroyed, resource leak cannot occur

# Resource Leak Example Revisited

## ■ implementation 1 (not exception safe; has memory leak):

```
1  void useBuffer(char* buf) { /* ... */ }
2
3  void doWork() {
4      char* buf = new char[1024];
5      useBuffer(buf);
6      delete[] buf;
7  }
```

## ■ implementation 2 (exception safe):

```
1  template <class T> class SmartPtr {
2      public:
3          SmartPtr(int size) : ptr_(new T[size]) {}
4          ~SmartPtr() {delete[] ptr_;}
5          SmartPtr(SmartPtr&& other) {ptr_ = other.ptr_; other.ptr_ = nullptr;}
6          SmartPtr& operator=(SmartPtr&& other) {
7              delete[] ptr_; ptr_ = other.ptr_; other.ptr_ = nullptr;
8              return *this;
9          }
10         T* get() const {return ptr_;}
11     private:
12         T* ptr_;
13 };
14
15 void useBuffer(char* buf) { /* ... */ }
16
17 void doWork() {
18     SmartPtr<char> buf(1024);
19     useBuffer(buf.get());
20 }
```

# RAII Example: Stream Formatting Flags

```
1  #include <iostream>
2  #include <ios>
3  #include <boost/io/ios_state.hpp>
4
5  // not exception safe
6  void unsafeOutput(std::ostream& out, unsigned int x) {
7      auto flags = out.flags();
8      // if exception thrown during output of x, old
9      // formatting flags will not be restored
10     out << std::hex << std::showbase << x << '\n';
11     out.flags(flags);
12 }
13
14 // exception safe
15 void safeOutput(std::ostream& out, unsigned int x) {
16     boost::io::ios_flags_saver ifs(out);
17     out << std::hex << std::showbase << x << '\n';
18 }
```

- RAII objects can be used to save and restore state



## Other RAII Examples

- `std::unique_ptr` and `std::shared_ptr` can be used to manage memory; memory is released in destructor [see [smart pointers](#)]
- `std::scoped_lock`, `std::unique_lock`, and `std::shared_lock` (and `std::lock_guard`) can be used to manage locks held on mutexes; lock is released in destructor
- `std::ifstream`, `std::ofstream`, and `std::fstream` can be used to manage files; file is closed in destructor
- `std::string` can be used to manage strings; string buffer freed in destructor
- `std::vector` can be used to manage dynamic arrays; array data freed in destructor

## Section 3.6.7

### **Exceptions: Implementation, Cost, and Usage**

# Implementation of Exception Handling

SKIP SLIDE

- standard does not specify how exception handling is to be implemented; only specifies behavior of exception handling
- consider typical implementation here
- potentially significant memory overhead for storing exception object and information required for stack unwinding
- possible to have zero time overhead if no exception thrown
- time overhead significant when exception thrown
- not practical to create exception object on stack, since object frequently needs to be propagated numerous levels up call chain
- exception objects tend to be small
- exception object can be stored in small fixed-size buffer falling back on heap if buffer not big enough

- memory required to maintain sufficient information to unwind stack when exception thrown
- two common strategies for maintaining information for stack unwinding: stack-based and table-based strategies
- stack-based strategy:
  - information for stack unwinding is saved on call stack, including list of destructors to execute and exception handlers that might catch exception
  - when exception is thrown, walk stack executing destructors until matching catch found
- table-based strategy:
  - store information to assist in stack unwinding in static tables outside stack
  - call stack used to determine which scopes entered but not exited
  - use look-up operation on static tables to determine where thrown exception will be handled and which destructors to execute
- table-based strategy uses less space on stack but potentially requires considerable storage for tables

# Appropriateness of Using Exceptions

- use of exceptions not appropriate in all circumstances
- in practice, exceptions can sometimes (depending on C++ implementation) have prohibitive memory cost for systems with *very limited memory* (e.g., some embedded systems)
- since throwing exception has significant time overhead only use for *infrequently occurring* situations (not common case)
- in code where exceptions can occur, often much more difficult to bound how long code path will take to execute
- since difficult to predict response time of code in presence of exceptions, exceptions often cannot be used in *time critical* component of real-time system (where operation must be guaranteed to complete in specific maximum time)
- considerable amount of code in existence that is *not exception safe*, especially legacy code
- cannot use exceptions in manner that would allow exceptions to propagate into code that is not exception safe

# Enforcing Invariants: Exceptions Versus Assertions

- whether invariants should be enforced by exceptions or assertions somewhat controversial
- would recommend only using exceptions for errors from which recovery is likely to be possible
- if error condition detected is indicative of serious programming error, program state may already be sufficiently invalid (e.g., stack trampled, heap corrupted) that exception handling will not work correctly anyhow
- tendency amongst novice programmers is to use exceptions in places where their use is either highly questionable or clearly inappropriate

## Section 3.6.8

### **Exception Gotchas**

# shared\_ptr Example: Not Exception Safe (Prior to C++17)

```
1  #include <memory>
2
3  class T1 { /* ... */ };
4  class T2 { /* ... */ };
5
6  void func(std::shared_ptr<T1> p, std::shared_ptr<T2> q)
7  { /* ... */ }
8
9  void doWork() {
10     // potential memory leak
11     func(std::shared_ptr<T1>(new T1),
12         std::shared_ptr<T2>(new T2));
13     // ...
14 }
```

## ■ one problematic order:

- 1 allocate memory for T1
- 2 construct T1
- 3 allocate memory for T2
- 4 construct T2
- 5 construct shared\_ptr<T1>
- 6 construct shared\_ptr<T2>
- 7 call func

## ■ if step 3 or 4 throws, memory leaked

## ■ another problematic order:

- 1 allocate memory for T1
- 2 allocate memory for T2
- 3 construct T1
- 4 construct T2
- 5 construct shared\_ptr<T1>
- 6 construct shared\_ptr<T2>
- 7 call func

## ■ if step 3 or 4 throws, memory leaked



## shared\_ptr Example: Exception Safe (Prior to C++17)

```
1  #include <memory>
2
3  class T1 { /* ... */ };
4  class T2 { /* ... */ };
5
6  void func(std::shared_ptr<T1> p, std::shared_ptr<T2> q)
7  { /* ... */ }
8
9  void doWork() {
10     func(std::make_shared<T1>(), std::make_shared<T2>());
11     // ...
12 }
```

- previously problematic line of code now does following:

- 1 perform following operations in any order:

- construct `shared_ptr<T1>` via `make_shared<T1>`
- construct `shared_ptr<T2>` via `make_shared<T2>`

- 2 call `func`

- each of `T1` and `T2` objects managed by `shared_ptr` at all times so no memory leak possible if exception thrown
- similar issue arises in context of `std::unique_ptr` and can be resolved by using `std::make_unique` in similar way as above

# Stack Example

- stack class template parameterized on element type T

```
1  template <class T>
2  class Stack
3  {
4  public:
5      // ...
6      // Pop the top element from the stack.
7      T pop() {
8          // If the stack is empty...
9          if (top_ == start_)
10             throw "stack is empty";
11         // Remove the last element and return it.
12         return * (--top_);
13     }
14 private:
15     T* start_; // start of array of stack elements
16     T* end_;   // one past end of array
17     T* top_;   // one past current top element
18 };
```

- what is potentially problematic about this code with respect to exceptions?

## Section 3.6.9

### **References**

- 1 D. Abrahams. **Exception-safety in generic components.**  
In *Lecture Notes in Computer Science*, volume 1766, pages 69–79.  
Springer, 2000.  
A good tutorial on exception safety by an expert on the subject.
- 2 T. Cargill. **Exception handling: A false sense of security.**  
*C++ Report*, 6(9), Nov. 1994.  
Available online at [http://ptgmedia.pearsoncmg.com/images/020163371x/supplements/Exception\\_Handling\\_Article.html](http://ptgmedia.pearsoncmg.com/images/020163371x/supplements/Exception_Handling_Article.html).  
An early paper that first drew attention to some of the difficulties in writing exception-safe code.
- 3 Exception-Safe Coding in C++, <http://exceptionsafecode.com>, 2014.
- 4 V. Kochhar, How a C++ Compiler Implements Exception Handling, <http://www.codeproject.com/Articles/2126/How-a-C-compiler-implements-exception-handling>, 2002.

- 5 C++ FAQ — Exceptions and Error Handling,  
<https://isocpp.org/wiki/faq/exceptions>, 2016.

- 1 Jon Kalb. Exception-Safe Code. CppCon, Bellevue, WA, USA, Sep 7–12, 2014. Available online at [https://youtu.be/W7fIy\\_54y-w](https://youtu.be/W7fIy_54y-w), <https://youtu.be/b9xMIKb1jMk>, and <https://youtu.be/MiKxfdkMJW8>. (This talk is in three parts.)
- 2 Jon Kalb. Exception-Safe Coding. C++Now, Aspen, CO, USA, May 13–18, 2012. Available online at <https://youtu.be/N9bR0ztmmEQ> and <https://youtu.be/UiZfODgB-Oc>. (This talk is in two parts.)
- 3 Peter Edwards. C++ Exception Handling — The Gory Details of an Implementation. Dublin C/C++ User Group Meetup, Feb. 2018. Available online at <https://youtu.be/XpRL7exdFL8>.

## Section 3.7

# Smart Pointers

## Section 3.7.1

### **Introduction**



# Memory Management, Ownership, and Raw Pointers

- responsibility of owner of chunk of dynamically-allocated memory to deallocate that memory when no longer needed
- so managing dynamically-allocated memory essentially reduces to problem of ownership management
- raw pointer does not have any ownership relationship with memory to which pointer refers
- consequently, raw pointer does not itself directly participate in memory management (e.g., deallocation)
- raw pointers often problematic in presence of exceptions, since such pointers do not know how to free their pointed-to memory
- raw pointers should only be used in situations where no ownership responsibility for pointees is needed (e.g., to simply observe object without managing its associated memory)

# Smart Pointers

- **smart pointer** is object that has interface similar to raw pointer (e.g., provides operations such as indirection/dereferencing and assignment) but offers some additional functionality
- smart pointers provide RAII mechanism for managing memory resource (i.e., pointed-to memory)
- unlike raw pointer, smart pointer owns its pointed-to memory
- consequently, smart pointer must provide mechanism for deallocating pointed-to memory when no longer needed
- some smart-pointer types allow only exclusive ownership, while others allow shared ownership
- destructor for smart pointer releases memory to which pointer refers if no longer needed (i.e., no other owners remain)
- smart pointers play crucial role in writing exception-safe code
- smart pointers should always be used (instead of raw pointers) when ownership of piece of memory needs to be tracked (e.g., so that it can be deallocated when no longer needed)

## Section 3.7.2

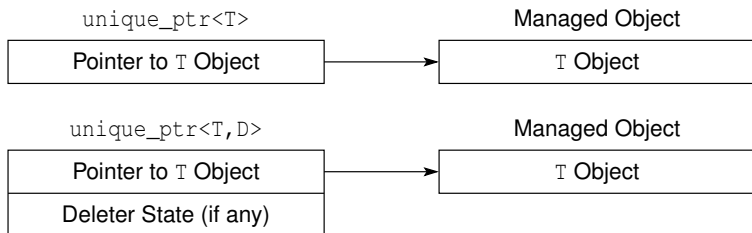
### **The `std::unique_ptr` Class Template**

# The `std::unique_ptr` Template Class

- `std::unique_ptr` is *smart pointer* that retains *exclusive* ownership of object through pointer
- declaration:

```
template <class T, class Deleter = std::default_delete<T>>  
    class unique_ptr;
```
- T is type of object to be managed (i.e., owned object)
- Deleter is callable entity used to delete owned object
- also correctly handles array types via partial specialization (e.g., T could be array of **char**)
- owned object destroyed when `unique_ptr` object goes out of scope
- no two `unique_ptr` objects can own same object
- `unique_ptr` object is *movable*; move operation transfers ownership
- `unique_ptr` object is *not copyable*, as copying would create additional owners
- `std::make_unique` template function often used to create `unique_ptr` objects (for exception-safety reasons)

# The `std::unique_ptr` Template Class (Continued)



- reasonable implementation would have zero memory cost for deleter state in case of:
  - default deleter
  - deleter of functor/closure type with no state
- if no memory cost for deleter state, `unique_ptr` has same memory cost as raw pointer

# std::unique\_ptr Member Functions

## Construction, Destruction, and Assignment

Member Name	Description
constructor	constructs new <code>unique_ptr</code>
destructor	destroys managed object (if any)
<b>operator=</b>	assigns <code>unique_ptr</code>

## Modifiers

Member Name	Description
release	returns pointer to managed object and releases ownership
reset	replaces managed object
swap	swaps managed objects

## Observers

Member Name	Description
get	returns pointer to managed object
get_deleter	returns deleter used for destruction of managed object
<b>operator bool</b>	checks if there is associated managed object

## Dereferencing/Subscripting

Member Name	Description
<b>operator*</b>	dereferences pointer to managed object
<b>operator-&gt;</b>	dereferences pointer to managed object
<b>operator[]</b>	provides indexed access to managed array

# std::unique\_ptr Example 1

```
1  #include <memory>
2  #include <cassert>
3
4  void func() {
5      auto p1(std::make_unique<int>(42));
6      assert(*p1 == 42);
7
8      // std::unique_ptr<int> p3(p1); // ERROR: not copyable
9      // p3 = p1; // ERROR: not copyable
10
11     std::unique_ptr<int> p2(std::move(p1)); // OK: movable
12     // Transfers ownership from p1 to p2, invalidating p1.
13     assert(p1.get() == nullptr && *p2 == 42);
14
15     p1 = std::move(p2); // OK: movable
16     // Transfers ownership from p2 to p1, invalidating p2.
17     assert(p2.get() == nullptr && *p1 == 42);
18
19     p1.reset();
20     // Invalidates p1.
21     assert(p1.get() == nullptr);
22 }
```



# std::unique\_ptr Example 2

SKIP SLIDE

```
1  #include <memory>
2  #include <cassert>
3
4  int main() {
5      auto p0 = std::make_unique<int>(0);
6      assert(*p0 == 0);
7      int* r0 = p0.get();
8      auto p1 = std::make_unique<int>(1);
9      assert(*p1 == 1);
10     auto r1 = p1.get();
11     p0.swap(p1);
12     assert(p0.get() == r1 && p1.get() == r0);
13     p1.swap(p0);
14     assert(p0.get() == r0 && p1.get() == r1);
15     p1.reset();
16     assert(p1.get() == nullptr);
17     assert(!p1);
18     int* ip = p1.release();
19     assert(!p1);
20     // ... Do not throw exceptions here.
21     delete ip;
22     p1.reset(new int(42));
23     assert(*p1 == 42);
24 }
```

# Example: `std::unique_ptr` with Custom Deleter

SKIP SLIDE

```
1  #include <memory>
2  #include <iostream>
3  #include <cstring>
4  #include <cstdlib>
5
6  using up = std::unique_ptr<char[], void(*) (char*)>;
7
8  char *allocate(std::size_t n) {
9      return static_cast<char*>(std::malloc(n));
10 }
11
12 void deallocate(char* p) {
13     std::cout << "deallocate called\n";
14     std::free(p);
15 }
16
17 up string_duplicate(const char *s) {
18     std::size_t len = std::strlen(s);
19     up result(allocate(len + 1), deallocate);
20     std::strcpy(result.get(), s);
21     return result;
22 }
23
24 int main() {
25     auto p = string_duplicate("Hello, World!");
26     std::cout << p.get() << '\n';
27 }
```

# TwoBufs Example With Resource Leak

SKIP SLIDE

```
1  #include <cstddef>
2  #include <limits>
3
4  class TwoBufs {
5  public:
6      TwoBufs(std::size_t aSize, std::size_t bSize) :
7          a_(nullptr), b_(nullptr) {
8          a_ = new char[aSize];
9          // If new throws, a_ will be leaked.
10         b_ = new char[bSize];
11     }
12     ~TwoBufs() {
13         delete[] a_;
14         delete[] b_;
15     }
16     // ...
17 private:
18     char* a_;
19     char* b_;
20 };
21
22 void doWork() {
23     // This may leak memory.
24     TwoBufs x(1000000,
25         std::numeric_limits<std::size_t>::max());
26     // ...
27 }
```

# TwoBufs Example Corrected With `unique_ptr`

SKIP SLIDE

```
1  #include <cstddef>
2  #include <limits>
3  #include <memory>
4
5  class TwoBufs {
6  public:
7      TwoBufs(std::size_t aSize, std::size_t bSize) :
8          a_(std::make_unique<char[]>(aSize)),
9          b_(std::make_unique<char[]>(bSize)) {}
10     ~TwoBufs() {}
11     // ...
12 private:
13     std::unique_ptr<char[]> a_;
14     std::unique_ptr<char[]> b_;
15 };
16
17 void doWork() {
18     // This will not leak memory.
19     TwoBufs x(1000000,
20         std::numeric_limits<std::size_t>::max());
21 }
```

## Section 3.7.3

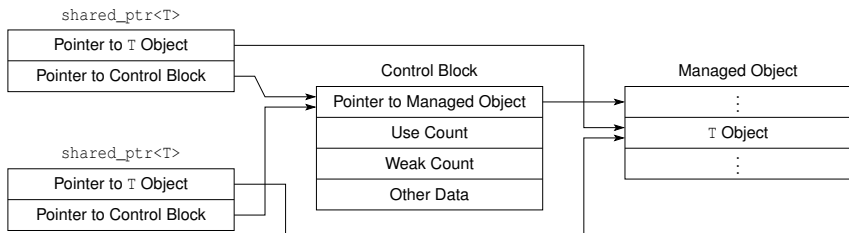
### **The `std::shared_ptr` Class Template**

# The `std::shared_ptr` Template Class

- `std::shared_ptr` is *smart pointer* that retains *shared* ownership of object through pointer
- declaration:  

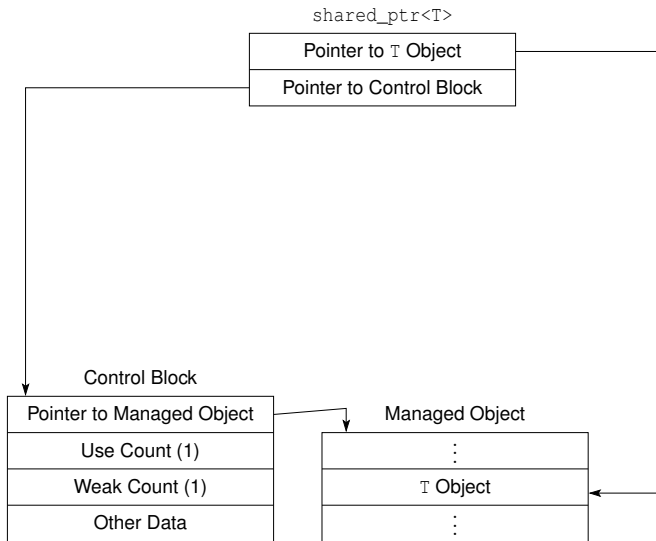
```
template <class T> class shared_ptr;
```
- T is type of pointed-to object, where pointed-to object is object being managed (i.e., owned) or subobject thereof
- multiple `shared_ptr` objects may own same object
- owned object is deleted when last remaining owning `shared_ptr` object is destroyed, assigned another pointer via assignment, or reset via `reset`
- `shared_ptr` object is *movable*, where move transfers ownership
- `shared_ptr` object is *copyable*, where copy creates additional owner
- `shared_ptr` only guarantees access to underlying control block is thread safe (e.g., no guarantee made for accesses to owned object)
- `std::make_shared` (and `std::allocate_shared`) often used to create `shared_ptr` objects (for both efficiency and exception-safety reasons)
- `shared_ptr` has more overhead than `unique_ptr` so `unique_ptr` should be preferred unless shared ownership required

# The `std::shared_ptr` Template Class (Continued)

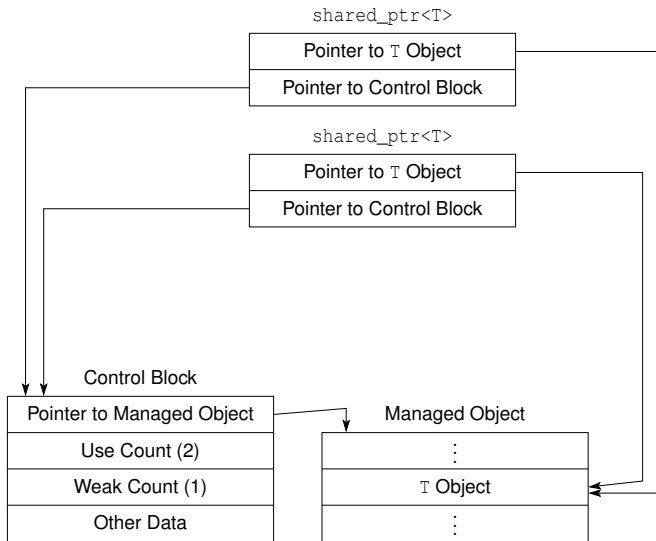


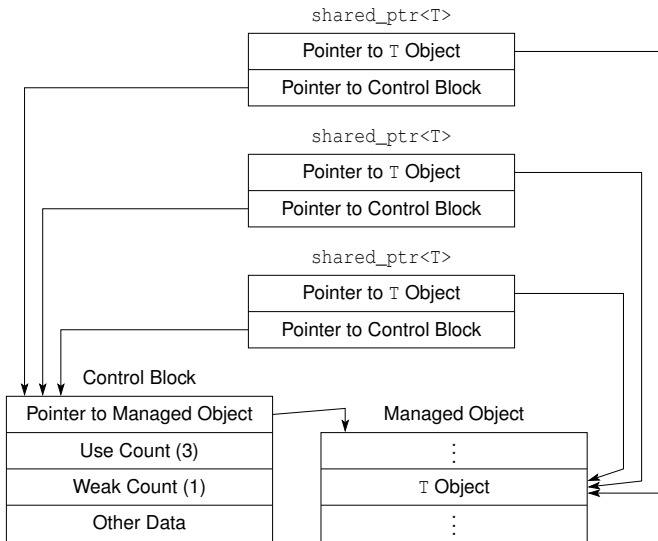
- each `shared_ptr<T>` object contains:
  - pointer to object of type `T` (i.e., managed object or subobject thereof)
  - pointer to control block
- control block contains:
  - pointer to managed object (for deletion)
  - use count: number of `shared_ptr` instances pointing to object
  - weak count: to be discussed later
  - other data (i.e., deleter and allocator)
- managed object is deleted when use count reaches zero
- `make_shared` can allow memory for control block and managed object to be allocated together in single memory allocation

# std::shared\_ptr Reference Counting Example









## Construction, Destruction, and Assignment

Member Name	Description
constructor	constructs new <code>shared_ptr</code>
destructor	destroys managed object if no other references to it remain
<b>operator=</b>	assigns <code>shared_ptr</code>

## Modifiers

Member Name	Description
reset	replaces managed object
swap	swaps values of two <code>shared_ptr</code> objects

## std::shared\_ptr Member Functions (Continued)

### Observers

Member Name	Description
get	returns pointer to pointed-to object
use_count	returns number of <code>shared_ptr</code> objects referring to same managed object
<b>operator bool</b>	checks if there is associated pointed-to/managed object
owner_before	provides owner-based ordering of shared pointers

### Dereferencing/Subscripting

Member Name	Description
<b>operator*</b>	dereferences pointer to pointed-to object
<b>operator-&gt;</b>	dereferences pointer to pointed-to object
<b>operator[]</b>	provides indexed access to pointed-to array

# Prefer Use of `std::make_shared`

- when creating `std::shared_ptr` objects, prefer to use `std::make_shared` (as opposed to explicit use of **`new`** with `shared_ptr`)
- use of `make_shared` allows for greater efficiency
- control block and owned object can be allocated together
- *one less memory-allocation operation* required
- *better cache efficiency* due to control block and owned object being placed contiguously in memory
- *better exception safety* (avoid resource leaks)

# std::shared\_ptr Example

```
1  #include <memory>
2  #include <cassert>
3
4  int main() {
5      auto p1(std::make_shared<int>(0));
6      assert(*p1 == 0 && p1.use_count() == 1);
7
8      std::shared_ptr<int> p2(p1);
9      assert(*p2 == 0 && p2.use_count() == 2);
10
11     *p2 = 42;
12     assert(*p1 == 42);
13
14     p2.reset();
15     assert(!p2);
16     assert(*p1 == 42 && p1.use_count() == 1);
17
18     int* ip = p1.get();
19     assert(*ip == 42);
20
21     ip = p2.get();
22     assert(ip == nullptr);
23 }
```

# std::shared\_ptr and const

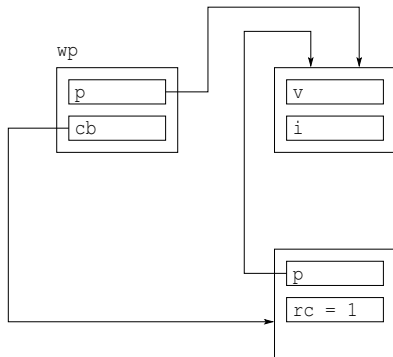
```
1  #include <memory>
2  #include <iostream>
3  #include <string>
4
5  int main() {
6      std::shared_ptr<std::string> s =
7          std::make_shared<std::string>("hello");
8
9      std::shared_ptr<const std::string> cs = s;
10
11     *s = "goodbye";
12
13     // *cs = "bonjour"; // ERROR: const
14
15     std::cout << *cs.get() << '\n';
16 }
```

# Example: Shared Pointer to Subobject of Managed Object

```
1  #include <memory>
2  #include <vector>
3  #include <cassert>
4  #include <iostream>
5
6  struct Widget {
7      Widget(const std::vector<int>& v_, int i_) :
8          v(v_), i(i_) {}
9      ~Widget() {std::cout << "destructor called\n";}
10     std::vector<int> v;
11     int i;
12 };
13
14 int main() {
15     auto wp(std::make_shared<Widget>(
16         std::vector<int>{1, 2, 3}, 42));
17     assert(wp.use_count() == 1);
18     assert(wp->i == 42 && wp->v.size() == 3);
19     std::shared_ptr<std::vector<int>> vp(wp, &wp->v);
20     assert(wp.use_count() == 2 && vp.use_count() == 2);
21     assert(vp->size() == 3);
22     wp = nullptr; // equivalently: wp.reset();
23     // managed Widget object not destroyed
24     assert(vp.use_count() == 1 && vp->size() == 3);
25     vp = nullptr; // equivalently: vp.reset();
26     // managed Widget object destroyed
27     // ...
28 }
```

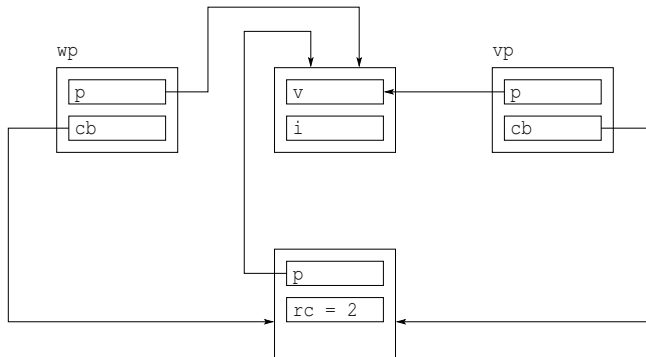


## Example: Shared Pointer to Subobject of Managed Object (Continued 1)



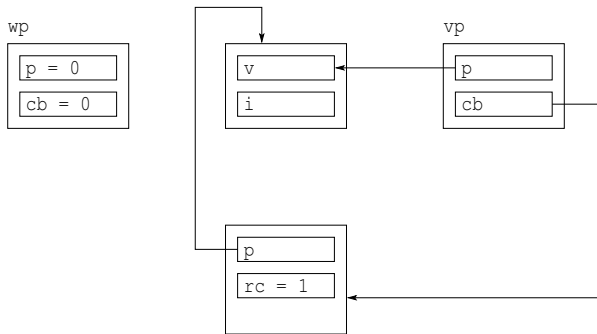
```
15  auto wp(std::make_shared<Widget>(
16      std::vector<int>{1, 2, 3}, 42));
17  assert(wp.use_count() == 1);
18  assert(wp->i == 42 && wp->v.size() == 3);
```

## Example: Shared Pointer to Subobject of Managed Object (Continued 2)



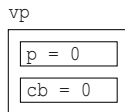
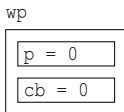
```
19 std::shared_ptr<std::vector<int>> vp(wp, &wp->v);  
20 assert(wp.use_count() == 2 && vp.use_count() == 2);  
21 assert(vp->size() == 3);
```

## Example: Shared Pointer to Subobject of Managed Object (Continued 3)



```
22 wp = nullptr; // equivalently: wp.reset();  
23 // managed Widget object not destroyed  
24 assert(vp.use_count() == 1 && vp->size() == 3);
```

## Example: Shared Pointer to Subobject of Managed Object (Continued 4)



```
25  vp = nullptr; // equivalently: vp.reset();  
26  // managed Widget object destroyed
```

# The `std::enable_shared_from_this` Class Template

SKIP SLIDE

- may want class object to be able to generate additional `shared_ptr` instances referring to itself
- requires object to have access to information in its associated `shared_ptr` control block
- access to such information obtained through use of `std::enable_shared_from_this` class template
- declaration:

```
template <class T> class enable_shared_from_this;
```

- T is type of object being managed by `shared_ptr`
- class can inherit from `enable_shared_from_this` to inherit `shared_from_this` member functions that can be used to obtain `shared_ptr` instance pointing to **\*this**
- `shared_from_this` is overloaded to provide both `const` and `non-const` versions

# enable\_shared\_from\_this Example

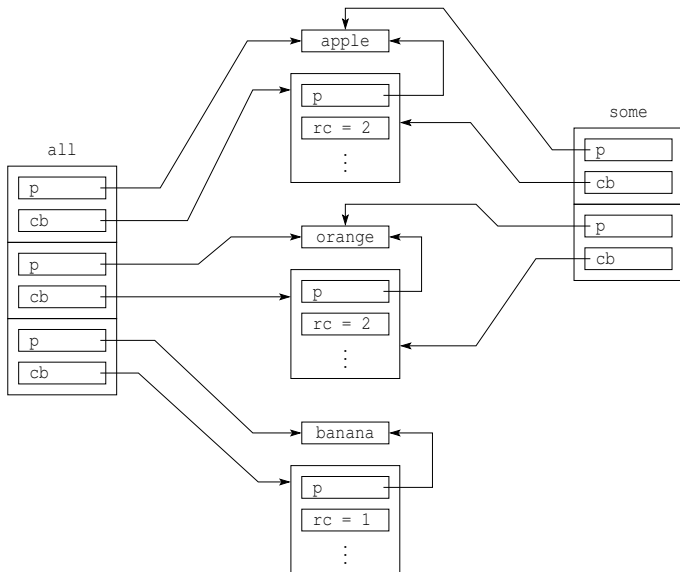
SKIP SLIDE

```
1  #include <memory>
2  #include <cassert>
3
4  // Aside: This is an example of the CRTP.
5  class Widget : public std::enable_shared_from_this<Widget>
6  {
7  public:
8      std::shared_ptr<Widget> getSharedPtr() {
9          return shared_from_this();
10     }
11     std::shared_ptr<const Widget> getSharedPtr() const {
12         return shared_from_this();
13     }
14     // ...
15 };
16
17 int main() {
18     std::shared_ptr<Widget> a(new Widget);
19     std::shared_ptr<Widget> b = a->getSharedPtr();
20     assert(b == a);
21     std::shared_ptr<const Widget> c = a->getSharedPtr();
22     assert(c == a);
23 }
```

## Example: std::shared\_ptr

```
1  #include <memory>
2  #include <array>
3  #include <string>
4  #include <iostream>
5
6  using namespace std::literals;
7
8  int main() {
9      std::array<std::shared_ptr<const std::string>, 3> all = {
10         std::make_shared<const std::string>("apple"s),
11         std::make_shared<const std::string>("orange"s),
12         std::make_shared<const std::string>("banana"s)
13     };
14     std::array<std::shared_ptr<const std::string>, 2> some =
15         {all[0], all[1]};
16
17     for (auto& x : all) {
18         std::cout << *x << ' ' << x.use_count() << '\n';
19     }
20 }
21
22 /* output:
23 apple 2
24 orange 2
25 banana 1
26 */
```

## Example: `std::shared_ptr` (Continued)





## Section 3.7.4

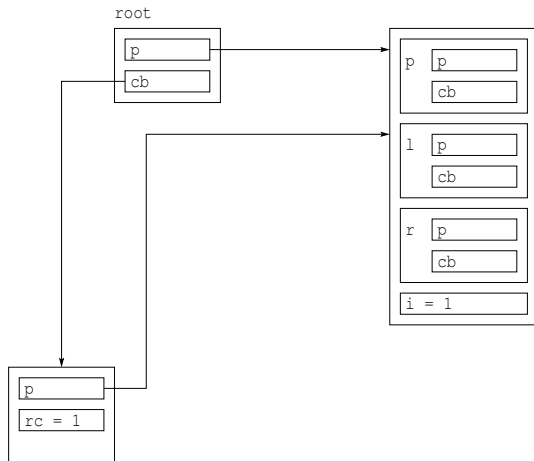
### **The `std::weak_ptr` Class Template**

- reference counting nature of `std::shared_ptr` causes it to leak memory in case of circular references
- such cycles should be broken with `std::weak_ptr` (to be discussed shortly)

# Circular Reference Example

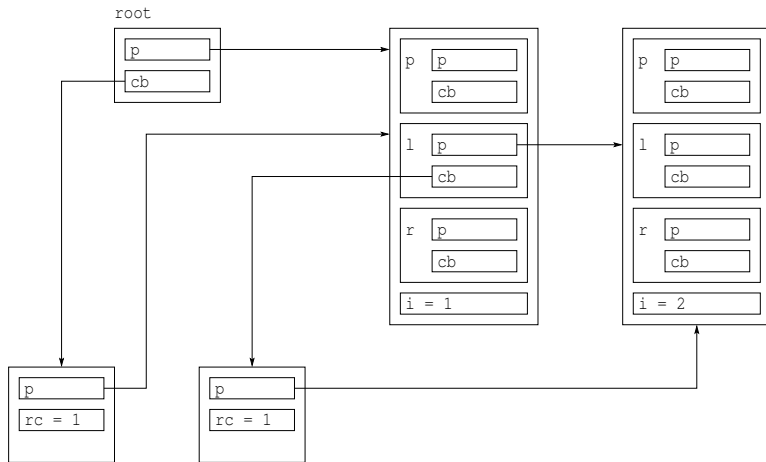
```
1  #include <memory>
2  #include <iostream>
3  #include <cassert>
4
5  struct Node {
6      Node(int id_) : id(id_) {}
7      ~Node() {std::cout << "destroying node " << id << '\n';}
8      std::shared_ptr<Node> parent;
9      std::shared_ptr<Node> left;
10     std::shared_ptr<Node> right;
11     int id;
12 };
13
14 void func() {
15     std::shared_ptr<Node> root(std::make_shared<Node>(1));
16     assert(root.use_count() == 1);
17     root->left = std::make_shared<Node>(2);
18     assert(root.use_count() == 1 &&
19            root->left.use_count() == 1);
20     root->left->parent = root;
21     assert(root.use_count() == 2 &&
22            root->left.use_count() == 1);
23     // When root is destroyed, the reference count for each
24     // of the managed Node objects does not reach zero, and
25     // no Node object is destroyed.
26     // Node::~~Node is not called here
27 }
```

# Circular Reference Example (Continued 1)



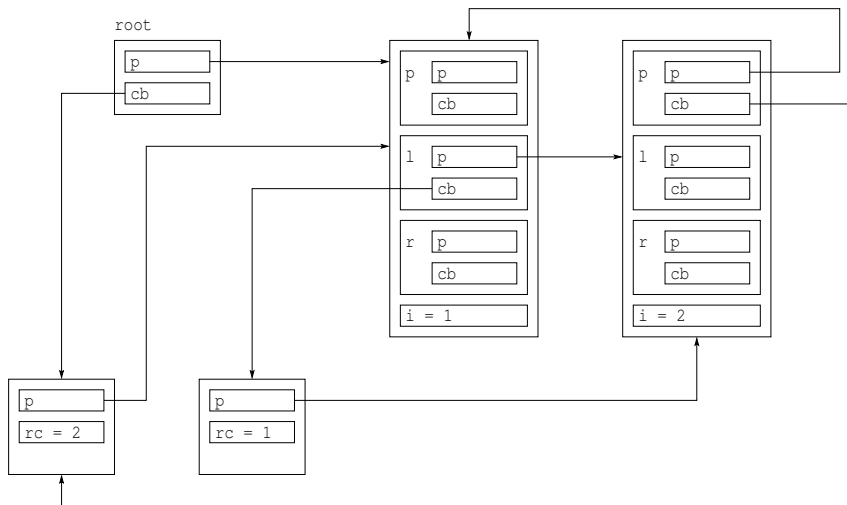
- create new node, referenced by `root`

# Circular Reference Example (Continued 2)



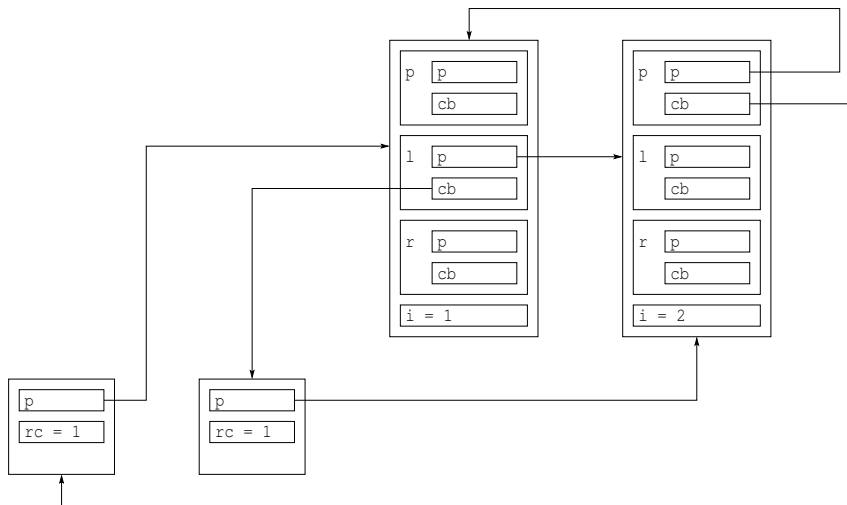
- create new node, making it left child of root node (parent link not set)

# Circular Reference Example (Continued 3)



- set parent link for left child of root node

## Circular Reference Example (Continued 4)



- after destroying `root`, neither node destroyed

# The `std::weak_ptr` Template Class

- `std::weak_ptr` is *smart pointer* that holds *non-owning* (i.e., “weak”) reference to object managed by `std::shared_ptr`
- `weak_ptr` must be converted to `std::shared_ptr` in order to access referenced object
- declaration:  

```
template <class T> class weak_ptr;
```
- T is type of referenced object
- `weak_ptr` object is *movable* and *copyable*
- `std::weak_ptr` is used to break circular references with `std::shared_ptr`



## Construction, Destruction, and Assignment

Member Name	Description
constructor	constructs new <code>weak_ptr</code>
destructor	destroys <code>weak_ptr</code>
<b>operator=</b>	assigns <code>weak_ptr</code>

## Modifiers

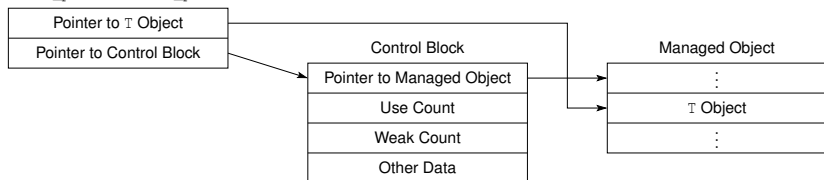
Member Name	Description
reset	releases reference to managed object
swap	swaps values of two <code>weak_ptr</code> objects

## Observers

Member Name	Description
<code>use_count</code>	returns number of <code>shared_ptr</code> objects referring to same managed object
<code>expired</code>	checks if referenced object was already deleted
<code>lock</code>	creates <code>shared_ptr</code> that manages referenced object
<code>owner_before</code>	provides owner-based ordering of weak pointers

# Typical `shared_ptr`/`weak_ptr` Implementation

`shared_ptr<T>` or `weak_ptr<T>`



- each `shared_ptr<T>` and `weak_ptr<T>` object contains:
  - pointer to object of type `T` (i.e., managed object or subobject thereof)
  - pointer to control block
- control block contains:
  - pointer to managed object (for deletion)
  - use count: number of `shared_ptr` instances pointing to object
  - weak count: number of `weak_ptr` instances pointing to object, plus one if use count is nonzero
  - other data (i.e., deleter and allocator)
- managed object is deleted when use count reaches zero
- control block is deleted when weak count reaches zero (which implies use count is also zero)

## ■ shared\_ptr destructor pseudocode:

```
decrement use count and if it reaches zero {  
    delete managed object  
    decrement weak count and if it reaches zero {  
        delete control block  
    }  
}
```

## ■ weak\_ptr destructor pseudocode:

```
decrement weak count and if it reaches zero {  
    delete control block  
}
```

## ■ must be thread safe

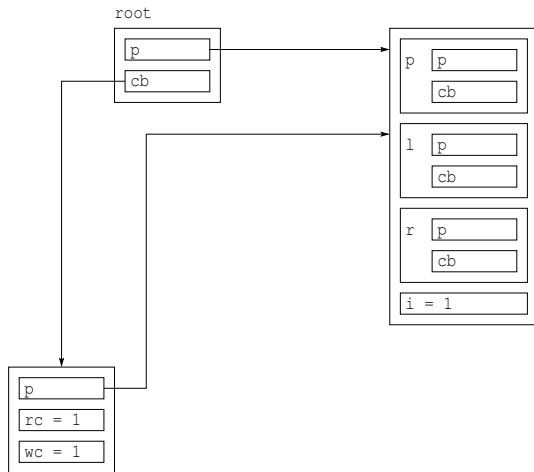
# std::weak\_ptr Example

```
1  #include <memory>
2  #include <iostream>
3  #include <cassert>
4
5  void func(std::weak_ptr<int> wp) {
6      auto sp = wp.lock();
7      if (sp) {
8          std::cout << *sp << '\n';
9      } else {
10         std::cout << "expired\n";
11     }
12 }
13
14 int main() {
15     std::weak_ptr<int> wp;
16     {
17         auto sp = std::make_shared<int>(42);
18         wp = sp;
19         assert(wp.use_count() == 1 && wp.expired() == false);
20         func(wp);
21         // When sp destroyed, wp becomes expired.
22     }
23     assert(wp.use_count() == 0 && wp.expired() == true);
24     func(wp);
25 }
```

# Avoiding Circular References With `std::weak_ptr`

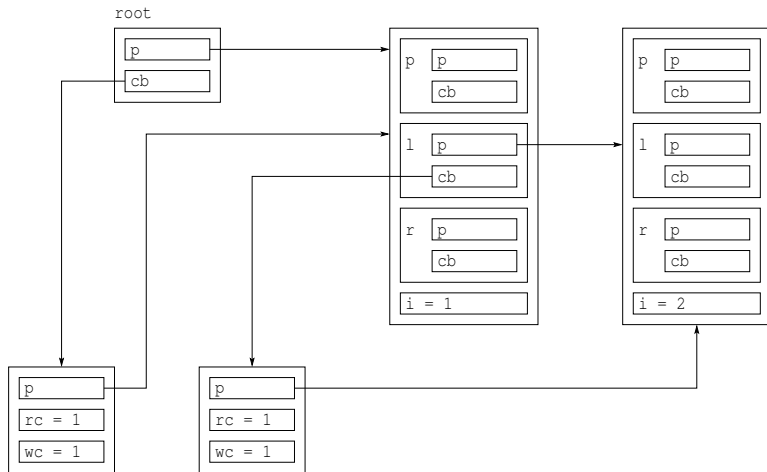
```
1  #include <memory>
2  #include <iostream>
3  #include <cassert>
4
5  struct Node {
6      Node(int id_) : id(id_) {}
7      ~Node() {std::cout << "destroying node " << id << '\n';}
8      std::weak_ptr<Node> parent;
9      std::shared_ptr<Node> left;
10     std::shared_ptr<Node> right;
11     int id;
12 };
13
14 void func() {
15     std::shared_ptr<Node> root(std::make_shared<Node>(1));
16     assert(root.use_count() == 1);
17     root->left = std::make_shared<Node>(2);
18     assert(root.use_count() == 1 &&
19            root->left.use_count() == 1);
20     root->left->parent = root;
21     assert(root.use_count() == 1 &&
22            root->left.use_count() == 1);
23     // The reference count for each of the managed Node
24     // objects reaches zero, and these objects are
25     // destroyed.
26     // Node::~~Node is called twice here
27 }
```

# Avoiding Circular References Example (Continued 1)



- created new node, referenced by `root`

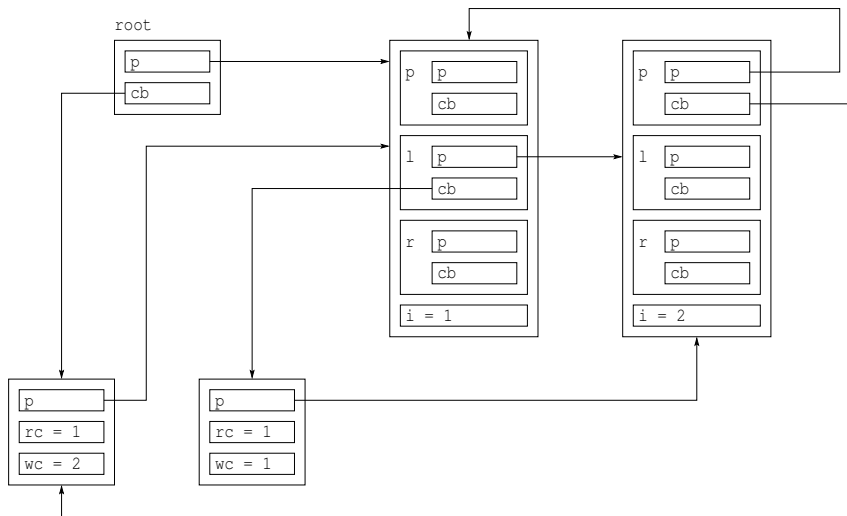
# Avoiding Circular References Example (Continued 2)



- created new node, making it left child of root node (parent link not set)

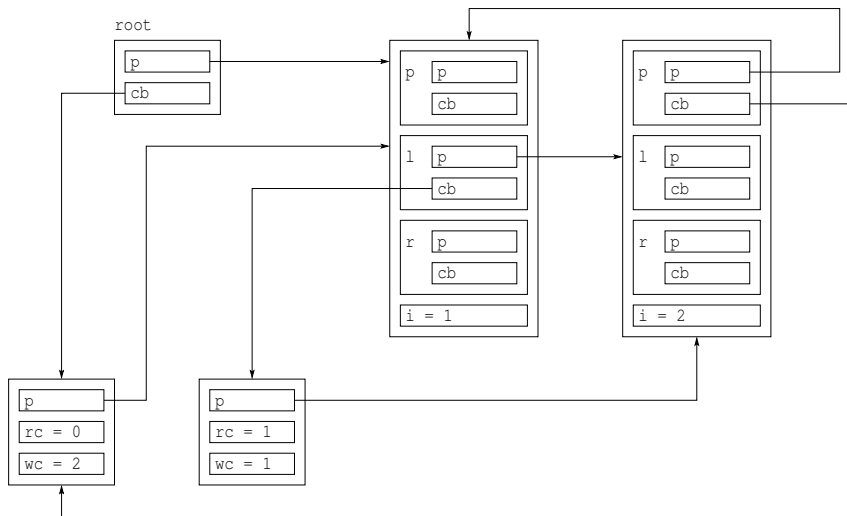


# Avoiding Circular References Example (Continued 3)



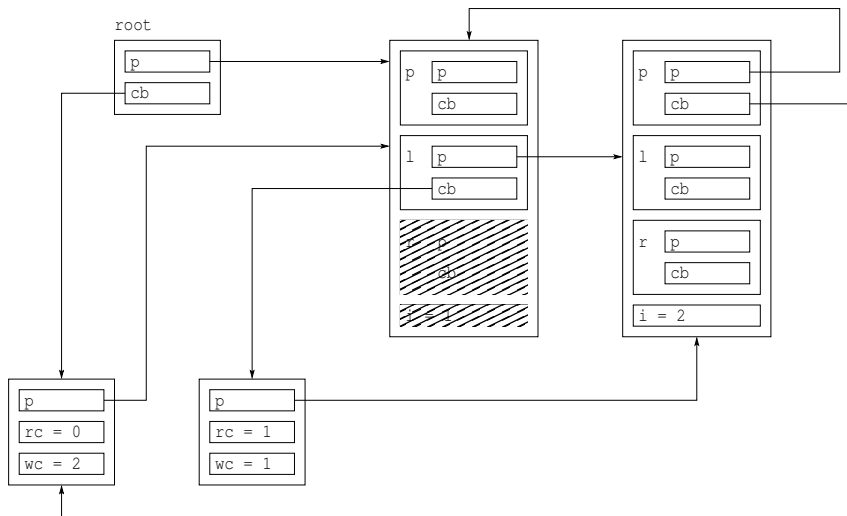
- set parent link (which is `weak_ptr`) for left child of root node

## Avoiding Circular References Example (Continued 4)



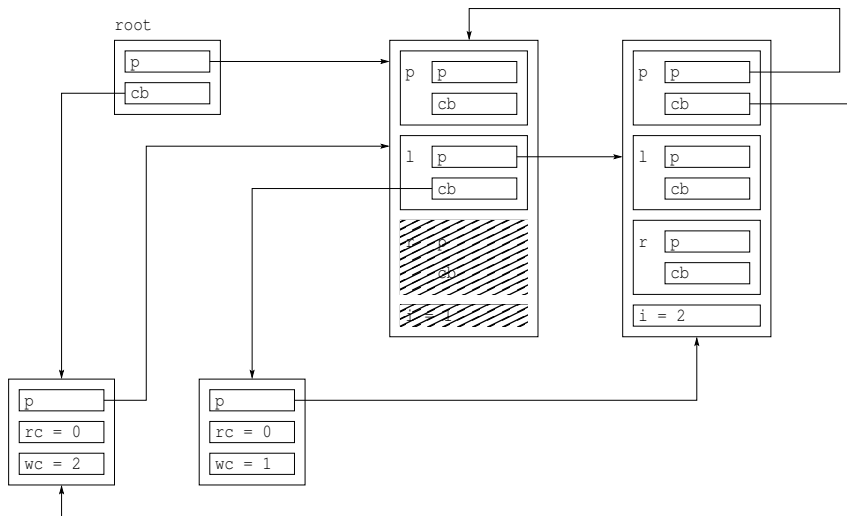
- started to destroy `root`; decremented use count, which reaches zero

# Avoiding Circular References Example (Continued 5)



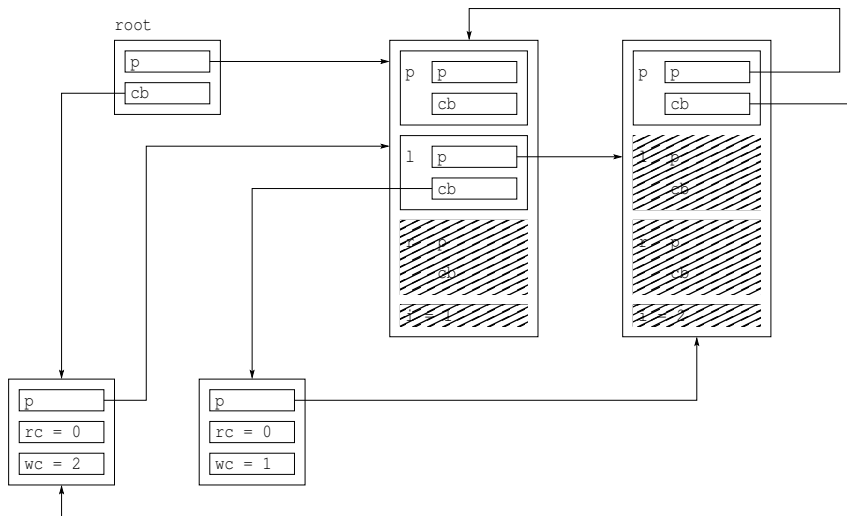
- started to destroy root node; `r` has been destroyed; about to destroy `l`

# Avoiding Circular References Example (Continued 6)



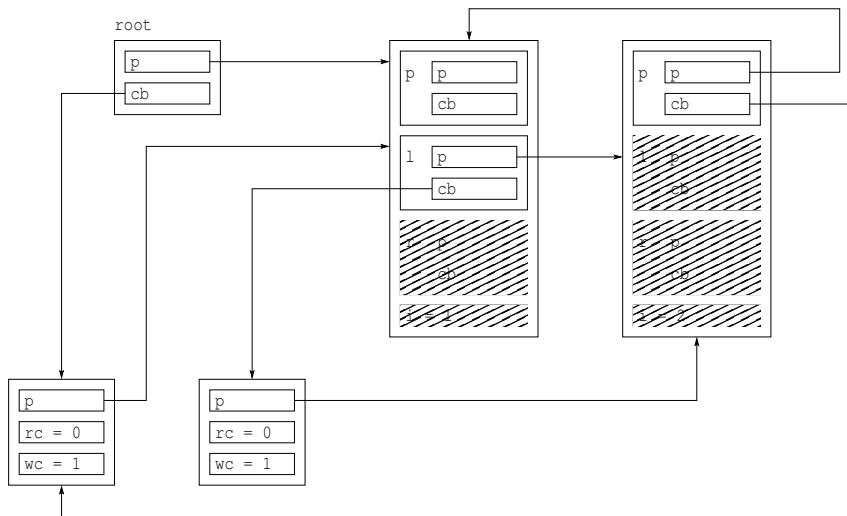
- started to destroy `l` (in root node); decremented use count, which reaches zero

# Avoiding Circular References Example (Continued 7)



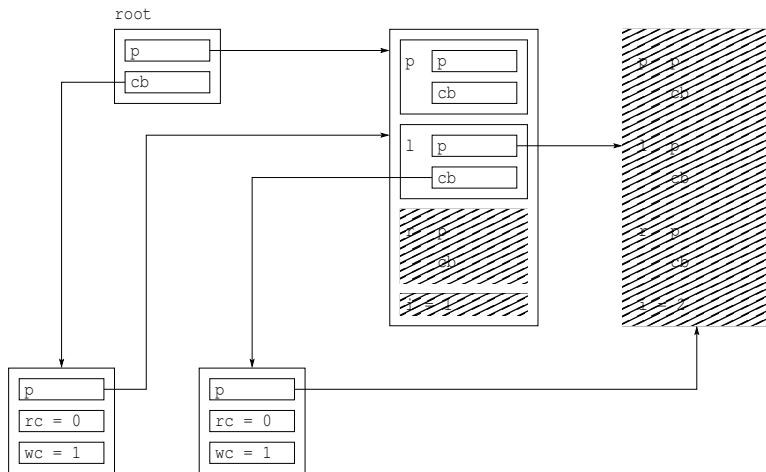
- started to destroy left node

# Avoiding Circular References Example (Continued 8)



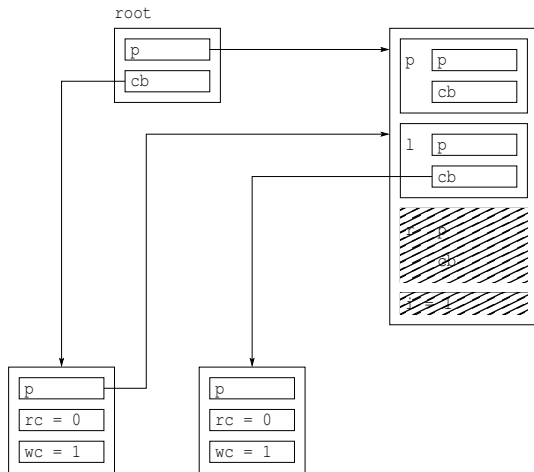
- started to destroy  $p$  in left node; decremented weak count (which is not yet zero)

# Avoiding Circular References Example (Continued 9)



- destroyed `p` in left node, and completed destruction of left node

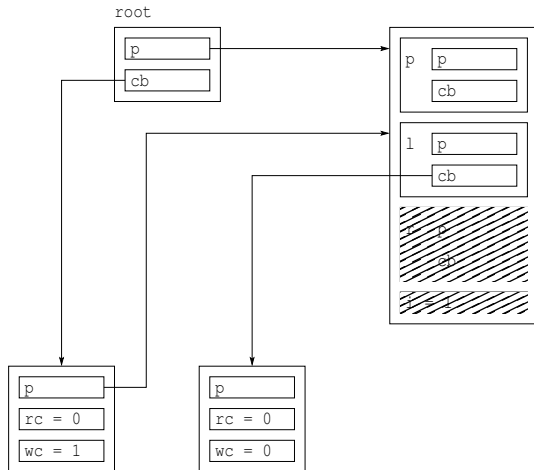
# Avoiding Circular References Example (Continued 10)



- left node has been destroyed

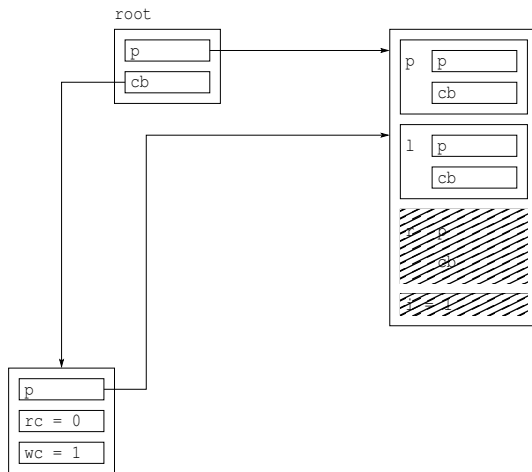


# Avoiding Circular References Example (Continued 11)



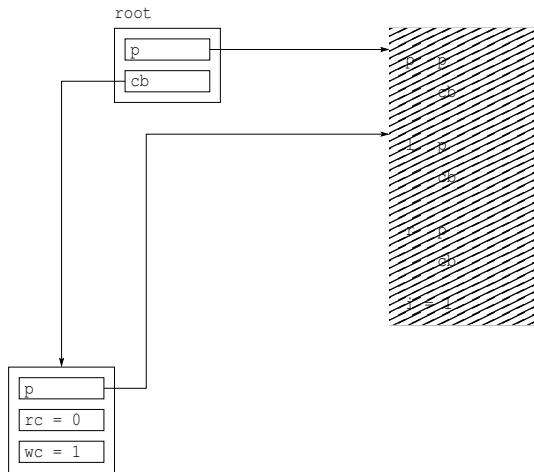
- continue destruction of `l` in root node; decrement weak count, which reaches zero

# Avoiding Circular References Example (Continued 12)



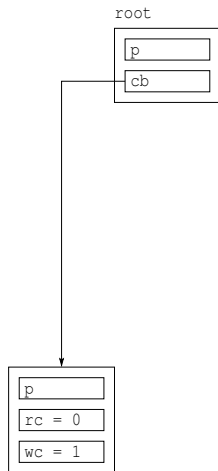
- destroyed control block for (previously destroyed) left child of root node

# Avoiding Circular References Example (Continued 13)



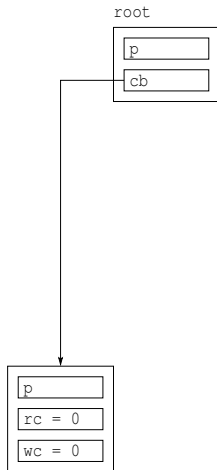
- finished destroying `l` in root node; destroyed `p` in root node; and completed destruction of root node

# Avoiding Circular References Example (Continued 14)



- root node has been destroyed

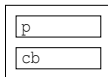
## Avoiding Circular References Example (Continued 15)



- continuing with destruction of `root`; decremented weak count, which reaches zero

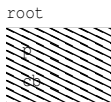
# Avoiding Circular References Example (Continued 16)

root



- destroyed control block

# Avoiding Circular References Example (Continued 17)



- `root` has been destroyed

## Section 3.7.5

### **The `boost::intrusive_ptr` Class Template**



# The `boost::intrusive_ptr` Class Template

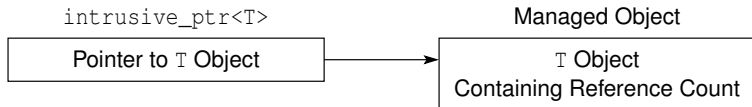
- `boost::intrusive_ptr` provides intrusive shared pointer type
- aside from being intrusive, similar to `boost::shared_ptr` (which is similar to `std::shared_ptr`)
- declaration:  

```
template <class T> class intrusive_ptr;
```
- T is type of referenced object
- new reference is added by calling user-provided function with signature:  

```
void intrusive_ptr_add_ref(T*)
```
- reference is eliminated by calling user-provided function with signature:  

```
void intrusive_ptr_release(T*)
```
- `intrusive_ptr_release` responsible for destroying underlying object when reference count reaches zero
- functions `intrusive_ptr_add_ref` and `intrusive_ptr_release` must be provided in such way as to be found when called unqualified

# The `boost::intrusive_ptr` Class Template (Continued 1)



- `intrusive_ptr` itself has same memory cost as raw pointer
- managed object (of type `T`) must provide means for reference counting, which is accessed through user-provided functions `intrusive_ptr_add_ref` and `intrusive_ptr_release`

# intrusive\_ptr Example

```
1  #include <boost/intrusive_ptr.hpp>
2  #include <iostream>
3  #include <string>
4  #include <cassert>
5
6  class Person {
7  public:
8      Person(const std::string& name) : name_(name),
9          refCount_(0) {}
10     void hold() {++refCount_;}
11     void release() {if (--refCount_ == 0) {delete this;}}
12     unsigned refCount() const {return refCount_;}
13 private:
14     ~Person() {std::cout << "dtor called\n";}
15     std::string name_;
16     unsigned refCount_; // reference count
17 };
18
19 void intrusive_ptr_add_ref(Person* p) {p->hold();}
20 void intrusive_ptr_release(Person* p) {p->release();}
21
22 int main() {
23     boost::intrusive_ptr<Person> a(new Person("Bjarne"));
24     {
25         boost::intrusive_ptr<Person> b = a;
26         assert(a->refCount() == 2);
27     }
28     assert(a->refCount() == 1);
29 }
```

## Section 3.7.6

### **Smart-Pointer Usage Examples**

# Temporary Heap-Allocated Objects

- create heap-allocated object for temporary use inside function/block
  - object will be automatically deallocated upon leaving function/block
- 

```
1  #include <memory>
2
3  void func() {
4      // ...
5      int size = /* ... */;
6      auto buffer(std::make_unique<char[]>(size));
7      // ... (use buffer)
8      // when buffer destroyed, pointee automatically
9      // freed
10 }
```

# Decoupled Has-A Relationship

- instead of making object member of class, store object outside class and make pointer to object member of class
  - might want to do this for object that:
    - is optional (e.g., is not always used or is lazily initialized)
    - has one of several base/derived types
  - pointer in class object owns decoupled object
- 

```
1  #include <memory>
2
3  class Widget {
4      // ...
5  private:
6      // ...
7      std::unique_ptr<Type> item_;
8      // decoupled object has type Type
9  };
```

# Decoupled Fixed-But-Dynamically-Sized Array

- array stored outside class object, where array size fixed but determined at run time
  - class object has pointer that owns decoupled array
- 

```
1  #include <memory>
2
3  class Widget {
4  public:
5      using Element = int;
6      Widget(std::size_t size) :
7          array_(std::make_unique<Element[]>(size),
8              size_(size)) {}
9      // ...
10 private:
11     // ...
12     std::unique_ptr<Element[]> array_;
13     std::size_t size_;
14 };
```

- with **pimpl idiom**, interface and implementation split across two classes, namely, handle class and implementation class
- handle object has pointer that owns implementation object

---

```
1  #include <experimental/propagate_const>
2  #include <memory>
3
4  class Widget {
5  public:
6      // ... (member functions that forward calls to
7      // implementation object)
8  private:
9      class WidgetImpl; // implementation class defined elsewhere
10     std::experimental::propagate_const<std::unique_ptr<
11         WidgetImpl>> impl_;
12     // incomplete type WidgetImpl is allowed
13     // ...
14 };
```



- tree, where tree owns root node and each node owns its children
  - recursive destruction of nodes may cause stack-overflow problems, especially for unbalanced trees (but such problems can be avoided by dismantling tree from bottom upwards)
- 

```
1  #include <memory>
2  #include <array>
3
4  class Tree {
5  public:
6      class Node {
7          // ...
8      private:
9          std::array<std::unique_ptr<Node>, 2> children_;
10         // owning pointers (parent owns children)
11         Node* parent_; // non-owning pointer
12         // ...
13     };
14     // ...
15 private:
16     std::unique_ptr<Node> root_;
17     // ...
18 };
```

# Doubly-Linked List

- doubly-linked list, where list owns first list node and each list node owns its successor
- recursive destruction of nodes can cause stack-overflow problems, for sufficiently large lists (but deep recursions can be avoided with extra work)

---

```
1  #include <memory>
2
3  class List {
4  public:
5      class Node {
6          // ...
7      private:
8          std::unique_ptr<Node> next_;
9          // owning pointer (node owns successor)
10         Node* prev_; // non-owning pointer
11     };
12     // ...
13 private:
14     // ...
15     std::unique_ptr<Node> head_;
16 };
```

# Tree That Provides Strong References

- tree that provides strong references to data in nodes
- tree owns root node and each node owns its children
- accessor for node data returns object having pointer that keeps node alive

```
1  #include <memory>
2  #include <array>
3
4  class Tree {
5  public:
6      using Data = /* ... */;
7      class Node {
8          // ...
9      private:
10         std::array<std::shared_ptr<Node>, 2> children_;
11         std::weak_ptr<Node> parent_;
12         Data data_;
13     };
14     std::shared_ptr<Data> find(/* ... */) {
15         std::shared_ptr<Node> sp;
16         // ...
17         return {sp, &(sp->data)};
18         // use shared_ptr aliasing constructor
19     }
20 private:
21     std::shared_ptr<Node> root_;
22 };
```

# Directed Acyclic Graph

- encapsulated directed acyclic graph (DAG), where graph owns root nodes and each node owns its children
- pointers in graph object own root nodes
- pointers in each node object owns children
- care needed for destruction, otherwise may overflow stack

---

```
1  #include <memory>
2  #include <vector>
3
4  class Dag {
5  public:
6      class Node {
7          // ...
8      private:
9          std::vector<std::shared_ptr<Node>> children_;
10         // owning pointers
11         std::vector<Node*> parents_; // non-owning pointers
12         // ...
13     };
14 private:
15     std::vector<std::shared_ptr<Node>> roots_; // owning pointers
16 };
```

# Factory Function

- factory function that returns object on heap
  - factory function should use `std::unique_ptr` if object will not be shared
  - factory function should use `std::shared_ptr` if object will be shared
  - provide factory functions using each of `std::unique_ptr` and `std::shared_ptr` if both sharing and non-sharing cases are common
- 

```
1  #include <memory>
2
3  std::unique_ptr<Widget> makeWidget() {
4      return std::make_unique<Widget>();
5  }
6
7  std::shared_ptr<Gadget> makeGadget() {
8      return std::make_shared<Gadget>();
9  }
```

# Factory Function With Cache

- cache of objects on heap
  - object in cache should only continue to live while it has external user
  - object returned to user is owning pointer
  - cache entries have non-owning pointers to corresponding objects
- 

```
1  #include <memory>
2
3  std::shared_ptr<Widget> makeWidget(int id) {
4      static std::map<int, std::weak_ptr<Widget>> cache;
5      static std::mutex mut;
6      std::scoped_lock<std::mutex> lock(mut);
7      auto sp = cache[id].lock();
8      if (!sp) {
9          sp = std::make_shared<Widget>(id);
10         cache[id] = sp;
11     }
12     return sp;
13 }
```

## Section 3.7.7

### **References**

- 1 Michael VanLoon. Lightning Talk: Anatomy of a Smart Pointer. CppCon, Bellevue, WA, USA, Sept. 9, 2014. Available online at [https://youtu.be/bxaj\\_0o4XAI](https://youtu.be/bxaj_0o4XAI).
- 2 Herb Sutter. Leak-Freedom in C++. . . By Default. CppCon, Bellevue, WA, USA, Sept. 23, 2016. Available online at <https://youtu.be/JfmTagWcqoE>.



## Section 3.8

# Memory Management

- object said to have **dynamic storage duration** if its lifetime is independent of scope in which object created (i.e., lifetime of object does not end until explicitly ended)
- often need to use objects (or arrays of objects) with dynamic storage duration
- in what follows, we consider how such objects are managed
- *new expressions* used to create objects or arrays of objects with dynamic storage duration
- *delete expressions* used to destroy such objects
- in order to handle any necessary memory allocation and deallocation, new and delete expressions in turn use (single-object and array) *operator new* and (single-object and array) *operator delete*

# Potential Problems Arising in Memory Management

- **leaked object**: object created but not destroyed when no longer needed
- leaked objects are problematic because can cause program to waste memory or exhaust all available memory
- **premature deletion** (a.k.a. **dangling references**): object is deleted when one or more references to object still exist
- premature deletion is problematic because, if object accessed after deletion, results of doing so will be unpredictable (e.g., read garbage value or overwrite other variables in program)
- **double deletion**: object is deleted twice, invoking destructor twice
- double deletion is problematic invoking destructor on nonexistent object is unpredictable and furthermore double deletion can often corrupt data structures used by memory allocator

- type can have restriction on address at which objects of that type can start, called **alignment requirement** [C++17 6.11/1]
- for given object type  $T$ , starting address for objects of type  $T$  must be integer multiple of  $N$  bytes, where integer  $N$  is called **alignment** of type
- alignment of 1 corresponds to no restriction on alignment (since starting address of object can be any address in memory)
- alignment of 2 restricts starting address of object to be even (i.e., integer multiple of 2)
- for efficiency reasons and due to restrictions imposed by hardware, alignment of particular type may be greater than 1
- for fundamental type  $T$ , not uncommon for alignment of  $T$  to equal **sizeof**( $T$ )

# The `alignof` Operator

- `alignof` operator is used to query alignment of type
- for object type `T`, `alignof(T)` yields alignment used for objects of this type
- `alignof(char)`, `alignof(signed char)`, and `alignof(unsigned char)` guaranteed to be 1
- fundamental types of size greater than 1 often have alignment greater than 1
- `std::max_align_t` is type having maximum alignment supported by implementation in all contexts
- **extended alignment** is alignment exceeding `alignof(std::max_align_t)` [\[C++17 6.11/3\]](#)
- in some contexts, may be possible to use extended alignment
- every alignment value must be nonnegative power of two [\[C++17 6.11/4\]](#)

# The **alignas** Specifier

- in some contexts, alignment can be controlled using **alignas** specifier
- when declaring variable, can specify its alignment in memory with **alignas** specifier
- when defining class, can impose stricter alignment on class with **alignas** specifier
- example:

```
1  // Widget objects have maximal alignment
2  struct alignas(std::max_align_t) Widget {
3      char c;
4  };
5  static_assert(alignof(Widget) == alignof(std::max_align_t));
6
7  alignas(4096) static char x[8192];
8  static_assert(alignof(x) == 4096);
9  // x is aligned on 4096-byte boundary
10
11 alignas(double) float f;
12 static_assert(alignof(f) == alignof(double));
13 // f has same alignment as double
```

## Section 3.8.1

# **New and Delete Expressions**

# New Expressions

- new expression used to create object or array of objects with dynamic storage duration
- new expression has one of following forms:
  - scope\_prefix* **new** *placement\_args* *type* *initializer*
  - scope\_prefix* **new** *placement\_args* (*type*) *initializer*
- *scope\_prefix*: optional unary :: operator which controls lookup of allocation function
- *placement\_args*: optional list of additional arguments for memory allocation function enclosed in parentheses
- *type*: type of object to be created which may be array type
- *initializer*: optional list of arguments used to initialize newly created object or array (e.g., constructor arguments for class type object)
- new expression where optional placement arguments provided referred to **placement new** expression
- new expression returns pointer to object created for non-array type or pointer to first element in array for array type



# New Expressions (Continued)

- examples of new expressions:

```
int* ip1 = new int;  
int* ip2 = new int(42);  
std::vector<int>* vp1 = new std::vector<int>(100, 42);  
int* aip1 = new int[256];  
std::string* asp = new std::string[64];  
int* aip2 = new (std::nothrow) int[10000];  
alignas(std::string) char buf[sizeof(std::string)];  
std::string* sp = new (static_cast<void*>(&buf))  
    std::string("Hello");  
Widget* wp = ::new Widget; // Note: Widget is class
```

- evaluating new expression performs following:

- 1 invokes allocation function to obtain address in memory where new object or array of objects should be placed
- 2 invokes constructors to create objects in storage obtained from allocation function
- 3 if constructor fails (i.e., throws), any successfully constructed objects are destroyed (in reverse order from construction) and deallocation function called to free memory in which object or array was being constructed

# Delete Expressions

- delete expression used to destroy object or array of objects created by new expression and deallocate associated memory
- delete expression has one of two forms:  
*scope\_prefix delete expr*  
*scope\_prefix delete[] expr*
- *scope\_prefix*: optional unary :: operator which controls lookup of deallocation function
- *expr*: pointer to object or array previously created by new expression or null pointer
- first form (sometimes called **single-object delete expression**) is used to dispose of single object obtained from new expression
- second form (sometimes called **array delete expression**) is used to dispose of array of objects obtained from new expression
- delete expression has void type
- if *expr* is null pointer, evaluation of delete expression effectively does nothing (i.e., no destructors called and no deallocation function called)

# Delete Expressions (Continued 1)

- single object created by new expression must be deleted with single-object delete expression
- array created by new expression must be deleted with array delete expression
- examples of delete expressions:

```
int *ip = new int(42);  
delete ip;  
std::vector<int> *vp = new std::vector<int>;  
delete vp;  
std::string* asp = new std::string[1024];  
delete[] asp;
```

- examples of incorrect delete expressions:

```
std::string* sp = new std::string;  
delete[] sp;  
    // ERROR: must use single-object delete expression  
std::string* asp = new std::string[1024];  
delete asp;  
    // ERROR: must use array delete expression
```

# Delete Expressions (Continued 2)

- evaluating single-object delete expression performs following:
  - 1 if object of class type, invokes destructor
  - 2 invokes deallocation function for object
- evaluating array delete expression performs following:
  - 1 if array element of class type (with non-trivial destructor):
    - 1 determines size of array (which is typically stored just before array element data)
    - 2 invokes destructor for each array element (in reverse order from construction, namely, backwards order)
  - 2 invokes deallocation function for array

# Typical Strategy for Determining Array Size in Array Delete

```
1  #include <cassert>
2  #include <cstdlib>
3  #include <iostream>
4
5  // A class with a nontrivial destructor.
6  struct Widget {
7      ~Widget();
8      // ...
9  };
10
11 int main(int argc, char** argv) {
12     for (std::size_t n = 1; n < 1024; ++n) {
13         /* Allocate an array of a class type with a nontrivial
14         destructor. */
15         Widget* p = new Widget[n];
16         std::cout << p << '\n';
17         /* Assert that the array size n is stored immediately
18         before the first array element. This approach is
19         not required by the C++ language standard, but some
20         C++ implementations do this. */
21         assert(reinterpret_cast<std::size_t*>(p)[-1] == n);
22         /* The array delete expression must invoke the
23         destructor for n objects before deallocating memory.
24         What is at issue here is how n is determined, since
25         it is not specified in the delete expression. */
26         delete[] p;
27     }
28 }
```

# New Expressions and Allocation

- new expression uses allocation function to determine where object (or array of objects) to be created should be placed in memory
- for non-array types, allocation function is single-object operator new (i.e., **operator new**) (discussed later), which can be overloaded
- for array types, allocation function is array operator new (i.e., **operator new[]**) (discussed later), which can be overloaded
- allocation function need not allocate memory (since placement arguments of new expression may be used to specify address at which to place new object)
- if allocation function has non-throwing exception specification, new expression returns null pointer upon failure otherwise `std::bad_alloc` exception is thrown
- for array type, requested size of memory may exceed size of actual array data (i.e., overhead to store size of array for use at deletion time)
- if new expression begins with unary `::` operator, allocation function's name looked up in global scope; otherwise, looked up in class scope if applicable and then global scope

# Allocation Function Overload Resolution

- overload resolution for (single-object and array) operator new performed using argument list consisting of:
  - 1 amount of space requested, which has type `std::size_t`
  - 2 if type has extended alignment, type's alignment, which has type `std::align_val_t`
  - 3 if placement new expression, placement arguments
- if no matching function found, alignment removed from argument list and overload resolution performed again
- expression "**new** T" results in one of following calls:

```
operator new(sizeof(T))
operator new(sizeof(T), std::align_val_t(alignof(T)))
```
- expression "**new**(42, f) T" results in one of following calls:

```
operator new(sizeof(T), 42, f)
operator new(sizeof(T), std::align_val_t(alignof(T),
42, f))
```

# Allocation Function Overload Resolution (Continued)

- expression “**new** T[7]” results in one of following calls:

```
operator new[](sizeof(T) * 7 + x)  
operator new[](sizeof(T) * 7 + x, std::align_val_t(  
    alignof(T)))
```

where  $x$  is nonnegative implementation-dependent constant representing array allocation overhead (typically,  $x > 0$  if  $T$  has nontrivial destructor)

- expression “**new** (42, f) T[7]” results in one of following calls:

```
operator new[](sizeof(T) * 7 + x, 42, f)  
operator new[](sizeof(T) * 7 + x, std::align_val_t(  
    alignof(T)), 42, f)
```

where  $x$  is nonnegative implementation-dependent constant representing array allocation overhead (typically,  $x > 0$  if  $T$  has nontrivial destructor)



# New Expressions and Deallocation

- when evaluating new expression, if allocation function succeeds but construction fails, must invoke deallocation function (as part of cleanup)
- for non-array types, deallocation function is single-object operator delete (i.e., **operator delete**) (to be discussed shortly)
- for array types, deallocation function is array operator delete (i.e., **operator delete[]**) (to be discussed shortly)
- if new expression begins with unary `::` operator, deallocation function's name looked up in global scope; otherwise, looked up in class scope and then global scope

# Delete Expressions and Deallocation

- for single-object delete expression, deallocation function is single-object operator delete (i.e., **operator delete**) (to be discussed shortly)
- for array delete expression, deallocation function is array operator delete (i.e., **operator delete []**) (to be discussed shortly)
- if delete expression prefixed by unary **::** operator, deallocation function's name looked up only at global scope; otherwise at class scope if applicable and then global scope

# Single-Object Operator New (i.e., **operator new**)

- single-object operator new (i.e., **operator new**) is operator used to determine address at which to place new object to be created
- most frequently invoked indirectly via new expression, but can be called directly
- operator new may or may not allocate memory
- operator can be overloaded as global function or (implicitly static) member function
- operator has return type **void\*** and returns address at which new object to be created should be placed
- first parameter to operator always of type `std::size_t` and specifies number of bytes of storage needed for new object to be created
- several overloads of global operator new provided by language and standard library
- `std::nothrow` is dummy variable of type **const** `std::nothrow_t` that can be used for overload disambiguation

# Single-Object Operator New Overloads

- **void\* operator new**(std::size\_t size);
  - allocates `size` bytes of storage that is suitably aligned for any object of this size not having extended alignment
  - throws `std::bad_alloc` exception upon failure
- **void\* operator new**(std::size\_t size, std::align\_val\_t align);
  - allocates `size` bytes of storage with guaranteed alignment of `align`
  - throws `std::bad_alloc` exception upon failure
- **void\* operator new**(std::size\_t size, **const** std::nothrow\_t& tag);
  - allocates `size` bytes of storage suitably aligned for any object of this size not having extended alignment
  - returns *null pointer* upon failure
- **void\* operator new**(std::size\_t size, std::align\_val\_t align, **const** std::nothrow\_t& tag);
  - allocates `size` bytes of storage with guaranteed alignment of `align`
  - returns *null pointer* upon failure

- **void\* operator new**(std::size\_t size, **void\*** ptr) **noexcept**;
  - *non-allocating*
  - simply returns ptr, assuming ptr points to storage of at least size bytes with appropriate alignment
  - cannot fail
  - not useful to invoke directly, since function effectively does nothing
  - intended to be invoked by non-allocating *placement new* expressions

# Single-Object Operator New Examples

```
1  #include <new>
2  #include <cassert>
3  #include <string>
4
5  void func_1() {
6      // allocating operator new
7      std::string* sp = static_cast<std::string*>(
8          ::operator new(sizeof(std::string)));
9      // allocation succeeded since no exception thrown
10     assert(sp);
11     // ... (deallocate memory)
12 }
13
14 void func_2() {
15     // allocating and non-throwing operator new
16     std::string* sp = static_cast<std::string*>(
17         ::operator new(sizeof(std::string), std::nothrow));
18     // sp may be null since allocation might have failed
19     // ... (deallocate memory)
20 }
21
22 void func_3() {
23     int i;
24     // non-allocating operator new
25     int* ip = static_cast<int*>(::operator new(sizeof(int),
26         static_cast<void*>(&i)));
27     assert(ip == &i);
28 }
```

# Array Operator New (i.e., **operator new**[])

- array operator new (i.e., **operator new**[]) is operator used to determine address at which to place array of objects to be created
- array operator new may or may not allocate memory
- array operator new can be overloaded as global function or (implicitly static) member function
- operator has return type **void\*** and returns address at which new array of objects to be created should be placed
- first parameter to operator always of type `std::size_t` and specifies number of bytes of storage needed for new array of objects to be created
- several overloads of global array operator new provided by language and standard library
- `std::nothrow` is dummy variable of type **const** `std::nothrow_t` that can be used for overload disambiguation

# Array Operator New Overloads

- **void\* operator new[]**(std::size\_t size);
  - allocates `size` bytes of storage that is suitably aligned for any object of this size not having extended alignment
  - throws `std::bad_alloc` exception upon failure
- **void\* operator new[]**(std::size\_t size, std::align\_val\_t align);
  - allocates `size` bytes of storage with alignment of `align`
  - throws `std::bad_alloc` exception upon failure
- **void\* operator new[]**( std::size\_t size, **const** std::nothrow\_t& tag);
  - allocates `size` bytes of storage suitably aligned for any object of this size not having extended alignment
  - returns *null pointer* upon failure
- **void\* operator new[]**( std::size\_t size, std::align\_val\_t align, **const** std::nothrow\_t& tag);
  - allocates `size` bytes of storage with guaranteed alignment of `align`
  - returns *null pointer* upon failure



- **void\* operator new[]**(std::size\_t size, void\* ptr) noexcept;
  - *non-allocating*
  - simply returns ptr, assuming ptr points to storage of at least size bytes with appropriate alignment
  - cannot fail
  - not useful to invoke directly, since function effectively does nothing
  - intended to be invoked by non-allocating (array) *placement new* expressions

# Array Operator New Examples

```
1  #include <new>
2  #include <cassert>
3  #include <string>
4
5  void func_1() {
6      // allocating array operator new
7      std::string* sp = static_cast<std::string*>(
8          ::operator new[(1000 * sizeof(std::string))];
9      // allocation succeeded since no exception thrown
10     assert(sp);
11     // ... (deallocate)
12 }
13
14 void func_2() {
15     std::string* sp = static_cast<std::string*>(
16         ::operator new[(1000 * sizeof(std::string), std::nothrow));
17     // sp may be null since allocation might have failed
18     // ... (deallocate)
19 }
20
21 void func_3() {
22     static int a[1000];
23     int* ip = static_cast<int*>(::operator new[(1000 * sizeof(int),
24         static_cast<void*>(a))];
25     assert(ip == a);
26 }
```

# Single-Object Operator Delete (i.e., **operator delete**)

- single-object operator delete (i.e., **operator delete**) is operator used to deallocate memory for object allocated with operator new
- can be invoked through delete expression or through new expression if constructor throws exception
- always has return type of **void**
- first parameter always pointer of type **void\***
- standard library deallocation functions do nothing if pointer is null
- can be overloaded as global function or (implicitly static) member function

# Single-Object Operator Delete Overloads

- **void operator delete(void\* ptr) noexcept;**  
**void operator delete(void\* ptr, std::size\_t size) noexcept;**  
**void operator delete (void\* ptr, std::align\_val\_t align) noexcept;**  
**void operator delete (void\* ptr, std::size\_t size, std::align\_val\_t align) noexcept;**
  - deallocates storage associated with object at address `ptr`, which was allocated by single-object operator `new`

# Single-Object Operator Delete Examples

```
1  #include <new>
2  #include <cassert>
3  #include <string>
4
5  void func_1() {
6      // allocating operator new
7      std::string* sp = static_cast<std::string*>(
8          ::operator new(sizeof(std::string)));
9      // allocation succeeded since no exception thrown
10     assert(sp);
11     ::operator delete(sp);
12 }
13
14 void func_2() {
15     // allocating and non-throwing operator new
16     std::string* sp = static_cast<std::string*>(
17         ::operator new(sizeof(std::string), std::nothrow));
18     // sp may be null since allocation might have failed
19     // deleting null pointer is allowed
20     ::operator delete(sp);
21 }
```

# Array Operator Delete (i.e., **operator delete[]**)

- array operator delete (i.e., **operator delete[]**) is operator used to deallocate memory for array of objects allocated with array operator new
- can be invoked through delete expression or through new expression if constructor throws exception
- always has return type of **void**
- first parameter always pointer of type **void\***
- standard library deallocation functions do nothing if pointer is null
- can be overloaded as global function or (implicitly static) member function

# Array Operator Delete Overloads

- **void operator delete[] (void\* ptr) noexcept;**  
**void operator delete[] (void\* ptr, std::size\_t size) noexcept;**  
**void operator delete[] (void\* ptr, std::align\_val\_t align) noexcept;**  
**void operator delete[] (void\* ptr, std::size\_t size, std::align\_val\_t align) noexcept;**
  - deallocates storage associated with array of objects at address `ptr`, which was allocated by array operator `new`

# Array Operator Delete Examples

```
1  #include <new>
2  #include <cassert>
3  #include <string>
4
5  void func_1() {
6      // allocating array operator new
7      std::string* sp = static_cast<std::string*>(
8          ::operator new[(1000 * sizeof(std::string))];
9      // allocation succeeded since no exception thrown
10     assert(sp);
11     ::operator delete[(sp);
12 }
13
14 void func_2() {
15     std::string* sp = static_cast<std::string*>(
16         ::operator new[(1000 * sizeof(std::string), std::nothrow));
17     // sp may be null since allocation might have failed
18     // deleting null pointer is allowed
19     ::operator delete[(sp);
20 }
```



# Example: New/Delete Expression and Overload Resolution

SKIP SLIDE

```
1  #include <cstddef>
2  #include <new>
3
4  // Gadget does not have extended alignment
5  struct Gadget {
6      int i;
7  };
8
9  // Widget has extended alignment
10 struct alignas(2 * alignof(std::max_align_t)) Widget {
11     int i;
12 };
13
14 int main() {
15     Gadget* gp = new Gadget;
16     // invokes operator new(std::size_t)
17     delete gp;
18     // invokes operator delete(void *)
19     Widget* wp = new Widget;
20     // invokes operator new(std::size_t, std::align_val_t)
21     delete wp;
22     // invokes operator delete(void *, std::align_val_t)
23 }
```

# Replacing Global Operator New/Delete

SKIP SLIDE

- some global versions of single-object and array operator new and operator delete can be replaced
- to replace function, define in single translation unit
- undefined behavior if more than one replacement provided in program or if replacement defined with inline specifier

# Example: Replacing Global Operator New/Delete

## SKIP SLIDE

```
1  #include <stdio>
2  #include <stdlib>
3  #include <new>
4  #include <string>
5
6  void* operator new(std::size_t size) {
7      auto ptr = std::malloc(size);
8      if (!ptr) {throw std::bad_alloc();}
9      std::printf("operator new(%zu) returning %p\n", size, ptr);
10     return ptr;
11 }
12
13 void operator delete(void* ptr) noexcept {
14     std::printf("operator delete(%p)\n", ptr);
15     std::free(ptr);
16 }
17
18 void* operator new[](std::size_t size) {
19     auto ptr = std::malloc(size);
20     if (!ptr) {throw std::bad_alloc();}
21     std::printf("operator new[](%zu) returning %p\n", size, ptr);
22     return ptr;
23 }
24
25 void operator delete[](void* ptr, std::size_t size) noexcept {
26     std::printf("operator delete[](%p)\n", ptr);
27     std::free(ptr);
28 }
29
30 int main() {
31     std::string* ip = new std::string;
32     delete ip;
33     std::string* ap = new std::string[10];
34     delete[] ap;
35 }
```

# Example: Overloading Operator New/Delete

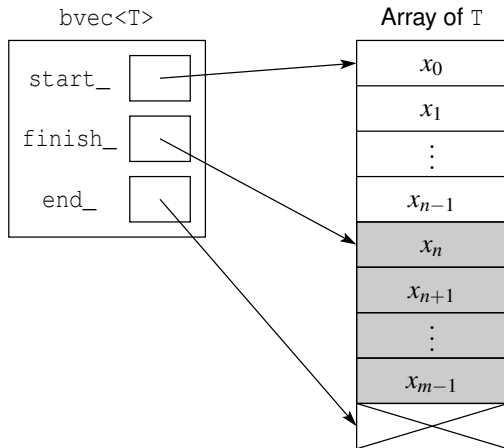
SKIP SLIDE

```
1  #include <cstdlib>
2  #include <stdio>
3  #include <stdexcept>
4
5  class Widget {
6  public:
7      void *operator new(std::size_t size) {
8          std::printf("Widget::operator new\n");
9          if (void *p = std::malloc(size); !p) throw std::bad_alloc();}
10         else {return p;}
11     }
12     void operator delete(void* p) noexcept {
13         std::printf("Widget::operator delete\n");
14         std::free(p);
15     }
16     // ...
17 };
18
19 int main() {
20     Widget* wp = new Widget; // invokes Widget::operator new
21     delete wp; // invokes Widget::operator delete
22     Widget* vp = ::new Widget; // invokes global operator new
23     ::delete vp; // invokes global operator delete
24 }
```

# Motivation for Placement New

```
1  #include <cstdint>
2
3  // heap-allocated array of bounded size
4  template <class T>
5  class bvec {
6  public:
7      // create empty vector that can hold max_size elements
8      // why is this implementation extremely inefficient?
9      bvec(std::size_t max_size) {
10         start_ = new T[max_size];
11         end_ = start_ + max_size;
12         finish_ = start_; // mark array empty
13     }
14     // why is this implementation extremely inefficient?
15     ~bvec() {
16         delete[] start_;
17     }
18     // ...
19 private:
20     T* start_; // start of storage for element data
21     T* finish_; // one past end of element data
22     T* end_; // end of storage for element data
23 };
```

# Motivation for Placement New: Diagram



# Placement New

- placement new expression is new expression that specifies one or more (optional) placement arguments
- often, placement new used for purpose of constructing object *at specific place* in memory
- this is accomplished by forcing non-allocating overload of operator new to be used (via placement arguments of new expression)

- example:

```
alignas(std::string) char buffer[sizeof(std::string)];  
std::string* sp =  
    new (static_cast<void*>(buffer)) std::string("Hello");  
assert(static_cast<void*>(sp) == buffer);  
// ... (destroy)
```

- although, in principle, placement new can also be used with new expressions for arrays, not very practically useful (since objects in array can always be created using single-object placement new expressions)

# Placement New Examples

```
1  #include <new>
2  #include <vector>
3  #include <string>
4  #include <cassert>
5  #include <utility>
6
7  void func_1() {
8      alignas(int) char buffer[sizeof(int)];
9      int* ip = ::new (static_cast<void>(buffer)) int(42);
10     assert(static_cast<void>(ip) == buffer && *ip == 42);
11 }
12
13 void func_2() {
14     alignas(std::string) char buffer[sizeof(std::string)];
15     std::string* vp = ::new (static_cast<void>(buffer)) std::string("hello");
16     assert(static_cast<void>(vp) == buffer && vp->size() == 5 &&
17         (*vp)[0] == 'h');
18     // ... (destroy)
19 }
20
21 template <class T, class... Args> T* construct_at(void* ptr, Args&&... args)
22     {return ::new (ptr) T(std::forward<Args>(args)...);}
23
24 void func_3() {
25     alignas(std::vector<int>) char buffer[sizeof(std::vector<int>)];
26     std::vector<int>* vp = construct_at<std::vector<int>>(buffer, 1000, 42);
27     assert(static_cast<void>(vp) == buffer && vp->size() == 1000 &&
28         (*vp)[0] == 42);
29     // ... (destroy vector)
30 }
```



# Direct Destructor Invocation

- can directly invoke destructor of class object
- only very special circumstances necessitate direct invocation of destructor
- used in situations where deallocation must be performed separately from destruction (in which case delete expression cannot be used as it performs both destruction and deallocation together)
- typical use case is for implementing container classes where destruction of object stored in container and deallocation of memory occupied by that object done at different points in time
- given pointer `p` to class object of type `T`, can directly invoke destructor through pointer using syntax `p->~T()` or `(*p).~T()`
- example:

```
alignas(std::vector<int>) char buf[
    sizeof(std::vector<int>)];
std::vector<int>* vp = ::new (static_cast<void*>(buf))
    std::vector<int>(1024);
vp->~vector();
```

# Pseudodestructors

- as discussed previously, given identifier  $T$  that names type and pointer  $p$  to object of that type, can directly invoke destructor with syntax  $p \rightarrow \sim T()$
- language allows for  $T$  to name built-in type [\[C++17 8.2.4/1\]](#)
- if  $T$  names built-in type, above syntax invokes placeholder for destructor called pseudodestructor, which effectively does nothing
- example:

```
void func() {  
    using T = int;  
    int i = 0;  
    int* ip = &i;  
    ip->~T();  
    // OK: pseudodestructor invocation (does nothing)  
    //ip->~int(); // ERROR: int is not identifier  
}
```

- consequently, in code that directly invokes destructor, do not need to treat built-in type as special case
- pseudodestructor functionality often extremely useful in template code that directly invokes destructors

## Section 3.8.2

### **More on Memory Management**

# std::addressof Function Template

- for memory management purposes, often necessary to obtain address of object
- if class overloads address-of operator, obtaining address of object becomes more difficult
- for convenience, standard library provides `std::addressof` function template for querying address of object, which yields correct result even if class overloads address-of operator

- declaration:

```
template <class T>
    constexpr T* addressof(T& arg) noexcept;
template <class T>
    const T* addressof(const T&&) = delete;
```

- `addressof` function should be used any time address of object is required whose class may have overloaded address-of operator

- example:

```
template <class T> foo(const T& x) {
    const T* p = std::addressof(x);
    // ...
}
```

# std::addressof Example

```
1  #include <iostream>
2  #include <cassert>
3  #include <memory>
4
5  // class that overloads address-of operator
6  class Foo {
7  public:
8      Foo(int i) : i_(i) {}
9      const Foo* operator&() const {return nullptr;}
10     Foo* operator&() {return nullptr;}
11     int get() const {return i_;}
12     // ...
13 private:
14     int i_;
15 };
16
17 int main() {
18     Foo f(42);
19     assert(&f == nullptr);
20     assert(std::addressof(f) != nullptr &&
21            std::addressof(f)->get() == 42);
22     std::cout << std::addressof(f) << '\n';
23 }
```

# The `std::aligned_storage` Class Template

- often need can arise for buffer of particular size and alignment
- for convenience, standard library provides `std::aligned_storage` class template for specifying such buffers

- declaration:

```
template <std::size_t Size, std::size_t Align =  
    __default_alignment> struct aligned_storage;
```

- Size is size of storage buffer in bytes
- Align is alignment of storage buffer (which has implementation-dependent default)
- for additional convenience, `std::aligned_storage_t` alias template also provided

- declaration:

```
template <std::size_t Size, std::size_t Align =  
    __default_alignment> using aligned_storage_t = typename  
    aligned_storage<Len, Align>::type;
```

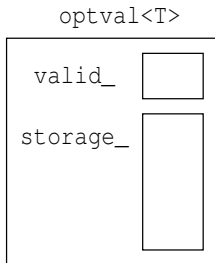
- example:

```
std::aligned_storage_t<sizeof(std::string),  
    alignof(std::string)> buffer;
```

# Optional Value Example

- consider container class template called `optval` that can hold optional value
- class templated on type `T` of optional value
- container object in one of two states:
  - 1 holding value of type `T`
  - 2 not holding any value
- can query if container is holding value, and if so, access held value
- somewhat similar in spirit to `std::optional`
- want to store object of type `T` *in `optval` object itself*
- no memory allocation required
- example demonstrates use of placement new (to construct object at particular place in memory) and direct invocation of destructor

# Optional Value Example: Diagram





# Optional Value Example: optval.hpp

```
1  #include <new>
2  #include <type_traits>
3
4  template <class T> class optval {
5  public:
6      optval() : valid_(false) {}
7      ~optval() {clear();}
8      optval(const optval&) = delete; // for simplicity
9      optval& operator=(const optval&) = delete; // for simplicity
10     bool has_value() const noexcept {return valid_;}
11     const T& get() const {return reinterpret_cast<const T*>(storage_);}
12     void clear() noexcept {
13         if (valid_) {
14             valid_ = false;
15             reinterpret_cast<T*>(&storage_)->~T();
16         }
17     }
18     void set(const T& value) {
19         clear();
20         ::new (static_cast<void*>(&storage_)) T(value);
21         valid_ = true;
22     }
23 private:
24     bool valid_; // is value valid?
25     std::aligned_storage_t<sizeof(T), alignof(T)> storage_;
26     // storage for value
27     // or alternatively: alignas(T) char storage_[sizeof(T)];
28 };
```

# Optional Value Example: User Code

```
1  #include <cassert>
2  #include <string>
3  #include <iostream>
4  #include "optional_1_util.hpp"
5
6  int main() {
7      optval<std::string> s;
8      assert(!s.has_value());
9      s.set("Hello, World");
10     assert(s.has_value());
11     std::cout << s.get() << '\n';
12     s.clear();
13     assert(!s.has_value());
14 }
```

# Handling Uninitialized Storage

- sometimes need may arise to work with uninitialized storage
- may want to construct objects in uninitialized storage (by using placement new to invoke constructor) and later destroy objects
- may want to move or copy objects into uninitialized storage (by using placement new to invoke move or copy constructor)
- code required to perform above operations is not very long, but must be written with some care to ensure that exceptions handled correctly
- standard library provides functions that perform these operations for convenience
- these functions useful for code that manages memory without using standard-compliant allocators

# Functions for Uninitialized Storage

## Operations on Uninitialized Memory

Name	Description
<code>uninitialized_copy</code>	copy range of objects to uninitialized area of memory
<code>uninitialized_copy_n</code>	copy number of objects to uninitialized area of memory
<code>uninitialized_fill</code>	copy object to uninitialized area of memory, defined by range
<code>uninitialized_fill_n</code>	copy object to uninitialized area of memory, defined by start and count
<code>uninitialized_move</code>	move range of objects to uninitialized area of memory
<code>uninitialized_move_n</code>	move number of objects to uninitialized area of memory

# Functions for Uninitialized Storage (Continued)

## Operations on Uninitialized Memory (Continued)

Name	Description
<code>uninitialized_default_construct</code>	construct objects by default initialization in uninitialized area of memory defined by range
<code>uninitialized_default_construct_n</code>	construct objects by default initialization in uninitialized area of memory defined by start and count
<code>uninitialized_value_construct</code>	construct objects by value initialization in uninitialized area of memory defined by range
<code>uninitialized_value_construct_n</code>	construct objects by value initialization in uninitialized area of memory defined by start and count
<code>destroy_at</code>	destroy object at given address
<code>destroy</code>	destroy range of objects
<code>destroy_n</code>	destroy number of objects in range

# Some Example Implementations

```
1  template<class InputIter, class ForwardIter>
2  ForwardIter uninitialized_copy(InputIter first, InputIter last,
3  ForwardIter result) {
4      using Value = typename std::iterator_traits<ForwardIter>::value_type;
5      ForwardIter current = result;
6      try {
7          for (; first != last; ++first, (void) ++current) {
8              ::new (static_cast<void*>(std::addressof(*current))) Value(*first);
9          }
10     } catch (...) {
11         for (; result != current; ++result) {
12             result->~Value();
13         }
14         throw;
15     }
16     return current;
17 }
```

---

```
1  template <class ForwardIter>
2  void destroy(ForwardIter first, ForwardIter last) {
3      for (; first != last; ++first) {
4          std::destroy_at(std::addressof(*first));
5      }
6  }
```

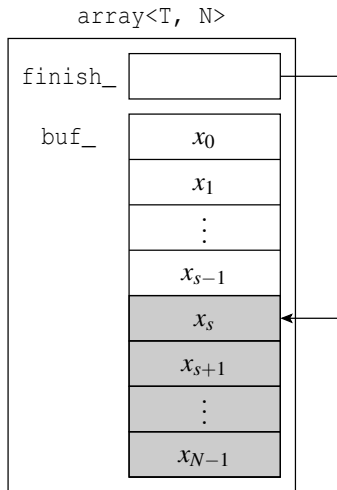
---

```
1  template <class T>
2  void destroy_at(T* p) {p->~T();}
```

# Bounded Array Example

- consider class template `array` for bounded one-dimensional array whose maximum size is compile-time constant
- class templated on element type `T` and number `N` of elements in array
- array element data is stored *in array object itself*
- no memory allocation required
- provide only basic container functionality in order to keep example to reasonable size for slides
- example demonstrates handling of uninitialized memory using standard library functions
- similar in spirit to `boost::static_vector`

# Bounded Array Example: Diagram





# Bounded Array Example: `aligned_buffer.hpp`

```
1  // type-aware aligned buffer class
2  // provides buffer suitably aligned for N elements of type T
3  template <class T, std::size_t N>
4      class aligned_buffer {
5  public:
6      const T* start() const noexcept
7          {return reinterpret_cast<const T*>(storage_);}
8      T* start() noexcept {return reinterpret_cast<T*>(storage_);}
9      const T* end() const noexcept {return start() + N;}
10     T* end() noexcept {return start() + N;}
11 private:
12     alignas(T) char storage_[N * sizeof(T)]; // aligned buffer
13 };
```

# Bounded Array Example: array.hpp (1)

```
1  #include <memory>
2  #include <algorithm>
3  #include <type_traits>
4  #include "aligned_buffer.hpp"
5
6  template <class T, std::size_t N> class array {
7  public:
8      array() : finish_(buf_.start()) {}
9      array(const array& other);
10     array(array&& other);
11     ~array() {clear();}
12     array& operator=(const array& other);
13     array& operator=(array&& other);
14     explicit array(std::size_t size);
15     array(std::size_t size, const T& value);
16     constexpr std::size_t max_size() const noexcept {return N;}
17     std::size_t size() const noexcept {return finish_ - buf_.start();}
18     T& operator[](std::size_t i) {return buf_.start()[i];}
19     const T& operator[](std::size_t i) const {return buf_.start()[i];}
20     T& back() {return finish_[-1];}
21     const T& back() const {return finish_[-1];}
22     void push_back(const T& value);
23     void pop_back();
24     void clear() noexcept;
25 private:
26     T* finish_; // one past last element in buffer
27     aligned_buffer<T, N> buf_; // buffer for array elements
28 };
```

# Bounded Array Example: array.hpp (2)

```
30  template <class T, std::size_t N>
31  array<T, N>::array(const array& other) {
32      finish_ = std::uninitialized_copy(other.buf_.start(),
33      static_cast<const T*>(other.finish_), buf_.start());
34  }
35
36  template <class T, std::size_t N>
37  array<T, N>::array(array&& other) {
38      finish_ = std::uninitialized_move(other.buf_.start(), other.finish_,
39      buf_.start());
40  }
41
42  template <class T, std::size_t N>
43  array<T, N>& array<T, N>::operator=(const array& other) {
44      if (this != &other) {
45          clear();
46          finish_ = std::uninitialized_copy(other.buf_.start(),
47          static_cast<const T*>(other.finish_), buf_.start());
48      }
49      return *this;
50  }
51
52  template <class T, std::size_t N>
53  array<T, N>& array<T, N>::operator=(array&& other) {
54      if (this != &other) { // self-assignment check of questionable value
55          clear();
56          finish_ = std::uninitialized_move(other.buf_.start(), other.finish_,
57          buf_.start());
58      }
59      return *this;
60  }
```

## Bounded Array Example: array.hpp (3)

```
62 template <class T, std::size_t N>
63 array<T, N>::array(std::size_t size) {
64     if (size > max_size()) {size = max_size();}
65     std::uninitialized_default_construct_n(buf_.start(), size);
66     finish_ = buf_.start() + size;
67 }
68
69 template <class T, std::size_t N>
70 array<T, N>::array(std::size_t size, const T& value) {
71     if (size > max_size()) {size = max_size();}
72     finish_ = std::uninitialized_fill_n(buf_.start(), size, value);
73 }
74
75 template <class T, std::size_t N>
76 void array<T, N>::push_back(const T& value) {
77     if (finish_ == buf_.end()) {return;}
78     finish_ = std::uninitialized_fill_n(finish_, 1, value);
79 }
80
81 template <class T, std::size_t N>
82 void array<T, N>::pop_back() {
83     --finish_;
84     std::destroy_at(finish_);
85 }
```

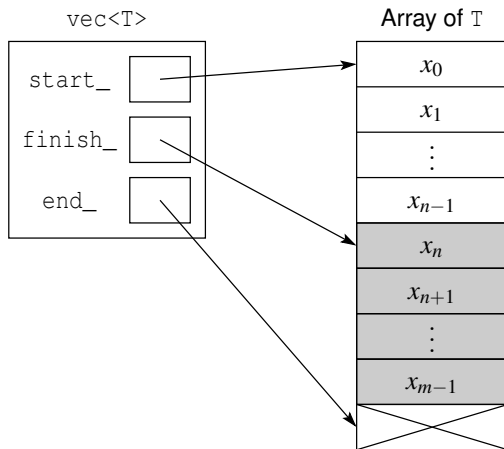
## Bounded Array Example: array.hpp (4)

```
87  template <class T, std::size_t N>
88  void array<T, N>::clear() noexcept {
89      std::destroy(buf_.start(), finish_);
90      finish_ = buf_.start();
91  }
```

# Vector Example

- consider class template `vec` that is one-dimensional dynamically-resizable array
- class templated on array element type `T`
- storage for element data allocated with *operator new*
- similar in spirit to `std::vector` but much simplified:
  - cannot specify allocator to be used (i.e., always uses operator new and operator delete for memory allocation)
  - does not provide iterators
  - only provides strong exception-safety guarantee for element insertion if move constructor for container element type does not throw

# Vector Example: Diagram



# Vector Example: vec.hpp (1)

```
1  #include <new>
2  #include <algorithm>
3  #include <type_traits>
4  #include <memory>
5
6  template <class T> class vec {
7  public:
8      vec() noexcept : start_(nullptr), finish_(nullptr), end_(nullptr) {}
9      vec(const vec& other);
10     vec(vec&& other) noexcept;
11     ~vec();
12     vec& operator=(const vec& other);
13     vec& operator=(vec&& other) noexcept;
14     explicit vec(std::size_t size);
15     vec(std::size_t n, const T& value);
16     std::size_t capacity() const noexcept {return end_ - start_;}
17     std::size_t size() const noexcept {return finish_ - start_;}
18     T& operator[](int i) {return start_[i];}
19     const T& operator[](int i) const {return start_[i];}
20     T& back() {return finish_[-1];}
21     const T& back() const {return finish_[-1];}
22     void push_back(const T& value);
23     void pop_back();
24     void clear() noexcept;
25 private:
26     void grow(std::size_t n);
27     T* start_; // start of element storage
28     T* finish_; // one past last valid element
29     T* end_; // end of element storage
30 };
```



## Vector Example: vec.hpp (2)

```
32  template <class T>
33  vec<T>::vec(const vec& other) {
34      start_ = static_cast<T*> (::operator new (other.size() * sizeof(T)));
35      end_ = start_ + other.size();
36      try {
37          finish_ = std::uninitialized_copy(other.start_, other.finish_, start_);
38      } catch (...) {
39          ::operator delete (start_);
40          throw;
41      }
42  }
43
44  template <class T>
45  vec<T>::vec(vec&& other) noexcept {
46      start_ = other.start_;
47      other.start_ = nullptr;
48      end_ = other.end_;
49      other.end_ = nullptr;
50      finish_ = other.finish_;
51      other.finish_ = nullptr;
52  }
53
54  template <class T>
55  vec<T>::~~vec() {
56      clear();
57      ::operator delete (start_);
58  }
```

## Vector Example: vec.hpp (3)

```
60  template <class T>
61  vec<T>& vec<T>::operator=(const vec& other) {
62      if (this != &other) {
63          clear();
64          if (other.size() > capacity()) {grow(other.size());}
65          finish_ = std::uninitialized_copy(other.start_, other.finish_, start_);
66      }
67      return *this;
68  }
69
70  template <class T>
71  vec<T>& vec<T>::operator=(vec&& other) noexcept {
72      if (this != &other) { // self-assignment check of questionable value
73          clear();
74          ::operator delete(start_);
75          start_ = other.start_;
76          other.start_ = nullptr;
77          finish_ = other.finish_;
78          other.finish_ = nullptr;
79          end_ = other.end_;
80          other.end_ = nullptr;
81      }
82      return *this;
83  }
```

## Vector Example: vec.hpp (4)

```
85  template <class T>
86  vec<T>::vec(std::size_t n) {
87      start_ = static_cast<T*> (::operator new(n * sizeof(T)));
88      end_ = start_ + n;
89      try {std::uninitialized_default_construct_n(start_, n);}
90      catch (...) {
91          ::operator delete(start_);
92          throw;
93      }
94      finish_ = end_;
95  }
96
97  template <class T>
98  vec<T>::vec(std::size_t n, const T& value) {
99      start_ = static_cast<T*> (::operator new(n * sizeof(T)));
100     end_ = start_ + n;
101     try {std::uninitialized_fill_n(start_, n, value);}
102     catch (...) {
103         ::operator delete(start_);
104         throw;
105     }
106     finish_ = end_;
107 }
```

## Vector Example: vec.hpp (5)

```
109 template <class T>
110 void vec<T>::push_back(const T& value) {
111     if (finish_ == end_) {
112         // might want to check for overflow here
113         grow(2 * capacity());
114     }
115     finish_ = std::uninitialized_fill_n(finish_, 1, value);
116 }
117
118 template <class T>
119 void vec<T>::pop_back() {
120     --finish_;
121     std::destroy_at(finish_);
122 }
123
124 template <class T>
125 void vec<T>::clear() noexcept {
126     if (size()) {
127         std::destroy(start_, finish_);
128         finish_ = start_;
129     }
130 }
```

## Vector Example: vec.hpp (6)

```
132 template <class T>
133 void vec<T>::grow(std::size_t n) {
134     T* new_start = static_cast<T*> (::operator new(n * sizeof(T)));
135     std::size_t old_size = size();
136     try {
137         std::uninitialized_move(start_, finish_, new_start);
138     } catch (...) {
139         ::operator delete(new_start);
140         throw;
141     }
142     ::operator delete(start_);
143     start_ = new_start;
144     finish_ = new_start + old_size;
145     end_ = new_start + n;
146 }
```

## Section 3.8.3

### **Allocators**

- allocators provide uniform interface for allocating and deallocating memory for object of particular type
- interface that allocator must provide specified in C++ standard
- each allocator type embodies particular memory allocation policy
- perform allocation, construction, destruction, and deallocation
- allocation separate from construction
- destruction separate from deallocation
- encapsulate information about allocation strategy and addressing model
- hide memory management and addressing model details from containers
- allow reuse of code implementing particular allocation strategy with any allocator-aware container

# Containers, Allocators, and the Default Allocator

- container class templates typically take allocator type as parameter
- this allows more than one memory allocation policy to be used with given container class template
- in case of standard library, many container class templates take allocator type as template parameter, including:
  - `vector`, `list`
  - `set`, `multiset`, `map`, `multimap`
  - `unordered_set`, `unordered_multiset`, `unordered_map`, `unordered_multimap`
- all container class templates in standard library that take allocator as parameter use default of `std::allocator<T>` where T must be type of element held by container
- `std::allocator` employs operator `new` and operator `delete` for memory allocation
- in many contexts, default allocator is quite adequate



# Application Use of Allocator

```
1  #include <memory>
2  #include <vector>
3  #include <map>
4  #include <cassert>
5  #include <boost/pool/pool_alloc.hpp>
6
7  int main() {
8      // use default allocator
9      std::vector<int> u;
10     u.push_back(42);
11
12     // explicitly specify default allocator
13     std::vector<int, std::allocator<int>> v;
14     static_assert(std::is_same_v<decltype(u), decltype(v)>);
15     assert(u.get_allocator() == v.get_allocator());
16     v.push_back(42);
17
18     // specify an allocator type from Boost
19     std::vector<int, boost::pool_allocator<int>> w;
20     w.push_back(42);
21
22     // explicitly specify default allocator
23     std::map<int, long, std::less<int>,
24             std::allocator<std::pair<const int, long>>> x;
25     x.insert({1, 2});
26 }
```

# Why Not Just Always Use the Default Allocator?

- custom allocators used when greater control is needed over how memory is managed
- often this greater control is desired for:
  - improved efficiency (e.g., better locality and less contention)
  - debugging
  - performance analysis (e.g., collecting statistics on memory allocation)
  - testing (e.g., forcing allocation failures)
  - security (e.g., locking and clearing memory)
- since many allocation strategies are possible, one strategy cannot be best in all situations
- some allocation strategies include:
  - stack-based allocation
  - per-container allocation
  - per-thread allocation (which avoids synchronization issues)
  - pooled allocation
  - arena allocation
- may want to handle relocatable data (e.g., shared memory)
- may want to use memory mapped files

- other examples of (standard-compliant) allocators include:
  - `std::pmr::polymorphic_allocator` (allocator whose behavior depends on memory resource with which it was constructed)
  - `boost::interprocess::allocator` (shared memory allocator)
  - `boost::pool_alloc` (pool allocator)
  - `boost::fast_pool_alloc` (pool allocator)

- allocator handles memory allocation for objects of specific type (e.g., allocator for `ints`)
- allocator normally accessed by container type through interface of traits class called `std::allocator_traits`
- container class typically use allocator for managing memory associated with container element data
- four basic types of operations provided by allocator through traits class:
  - allocate memory
  - deallocate memory
  - construct object
  - destroy object
- two allocator instances deemed *equal* if memory allocated with each instance can be deallocated with other
- allocator objects may have state

# Allocator Members

- allocator type for objects of (cv-unqualified) type `T`
- many members are optional, with `std::allocator_traits` class effectively providing defaults for omitted members
- `value_type`:
  - type `T` of object for which allocator manages (i.e., allocates and deallocates) memory
- `pointer`:
  - pointer type used to refer to storage obtained from allocator (not necessarily `T*`)
  - optional: default of `T*` provided by `allocator_traits`
- `const_pointer`:
  - `const` version of `pointer`
  - optional: default of `const T*` provided by `allocator_traits`
- `pointer allocate(size_type n)`:
  - allocate storage suitable for `n` objects of type `T`
- `void deallocate(pointer ptr, size_type n)`:
  - deallocates storage pointed to by `ptr`, where `ptr` must have been obtained by previous call to `allocate` and `n` must match value given in that call

- **void** `construct(value_type* ptr, Args&&... args):`
  - constructs object of type `T` in storage pointed to by `ptr` using specified arguments `args`
  - optional: default behavior provided by `allocator_traits` is to use placement new expression
- **void** `destroy(value_type* ptr):`
  - destroys object of type `T` in storage pointed to by `ptr`
  - optional: default behavior provided by `allocator_traits` is to directly invoke destructor

- `pointer` and `const_pointer` must satisfy requirements of random-access and contiguous iterators [\[C++17 20.5.3.5/5\]](#)
- `pointer` and `const_pointer` can be fancy pointers (i.e., smart pointers)
- fancy pointers useful, for example, in allocating storage in shared memory region

# Malloc-Based Allocator: Allocator Code

```
1  #include <cstdlib>
2  #include <limits>
3  #include <new>
4
5  template <class T> struct allocator {
6      using value_type = T;
7      allocator() noexcept {};
8      template <class U> allocator(const allocator<U>&) noexcept {}
9      T* allocate(std::size_t n) const;
10     void deallocate(T* p, std::size_t n) const noexcept;
11     template <class U> bool operator==(const allocator<U>&)
12         const noexcept {return true;}
13     template <class U> bool operator!=(const allocator<U>&)
14         const noexcept {return false;}
15 };
16
17 template <class T> T* allocator<T>::allocate(std::size_t n) const {
18     if (!n) {return nullptr;}
19     if (n > std::numeric_limits<std::size_t>::max() / sizeof(T))
20         {throw std::bad_array_new_length();}
21     void* p = std::malloc(n * sizeof(T));
22     if (!p) {throw std::bad_alloc();}
23     return static_cast<T*>(p);
24 }
25
26 template <class T> void allocator<T>::deallocate(T* p, std::size_t)
27     const noexcept
28     {std::free(p);}
```



# Malloc-Based Allocator: User Code

```
1  #include "allocator.hpp"
2  #include <cassert>
3  #include <vector>
4  #include <type_traits>
5
6  int main() {
7      std::vector<int, allocator<int>> v;
8          // uses allocator<int> for memory allocation
9      std::vector<int> w;
10         // or equivalently, std::vector<int, std::allocator<int>>
11         // uses std::allocator<int> for memory allocation
12     static_assert (!std::is_same_v<decltype(v)::allocator_type,
13                   decltype(w)::allocator_type>);
14     for (int i = 0; i < 128; ++i) {
15         v.push_back(42);
16         w.push_back(42);
17     }
18     std::vector<int, allocator<int>> x;
19     assert(v.get_allocator() == x.get_allocator());
20 }
```

# Allocator Propagation

- in certain contexts, must consider if and how allocators should be propagated between container objects
- **lateral propagation** refers to propagation of allocator when copying, moving, and swapping containers:
  - when container copy/move constructed, what allocator does new container receive?
  - when container copy/move assigned, what allocator does copied-to/moved-to container receive?
  - when containers swapped, what allocator does each container receive?
- **deep propagation** refers to propagation of allocator from parent container to its descendants in hierarchy of nested containers:
  - if container contains types which themselves require allocators, how can contained elements be made aware of container's allocator so that compatible allocator can be used?
- each allocator has its own lateral propagation properties, which can be accessed via `std::allocator_traits`
- deep allocator propagation can be controlled via `std::scoped_allocator_adaptor`

# New-Based Allocator

```
1  #include <limits>
2  #include <new>
3  #include <type_traits>
4
5  template <class T> struct allocator {
6      using value_type = T;
7      using propagate_on_container_move_assignment = std::true_type;
8      using is_always_equal = std::true_type;
9      allocator() noexcept {};
10     allocator(const allocator&) noexcept {};
11     template <class U> allocator(const allocator<U>&) noexcept {}
12     ~allocator() {}
13     T* allocate(std::size_t n);
14     void deallocate(T* p, std::size_t n) const noexcept
15         {::operator delete(p);}
16 };
17
18 template <class T> T* allocator<T>::allocate(std::size_t n) {
19     if (n > std::numeric_limits<std::size_t>::max() / sizeof(T))
20         {throw std::bad_array_new_length();}
21     return static_cast<T*> (::operator new(n * sizeof(T)));
22 }
23
24 template <class T, class U> inline bool operator==(const allocator<T>&,
25     const allocator<U>&) noexcept
26     {return true;}
27
28 template <class T, class U> inline bool operator!=(const allocator<T>&,
29     const allocator<U>&) noexcept
30     {return false;}
```

# Fixed-Size Arena Allocator: Example

- consider example of simple allocator that allocates memory from fixed-size buffer
- arena class (called `arena`) provides memory allocation from fixed-size buffer with some prescribed minimum alignment
- allocator class (called `salloc`) provides interface to particular `arena` instance
- `salloc` object holds pointer to `arena` object (so allocator is stateful)
- `arena` object makes no attempt to deallocate memory (i.e., deallocate operation does nothing)
- allocator might be used for relatively small allocations from stack (where `arena` object would be local variable)
- allocator always propagated for copy, move, and swap (i.e., POCMA, POCCA, and POCS, as defined later, all true)
- two instances of allocator not necessarily equal

# Fixed-Size Arena Allocator: Code (1)

```
1  #include <cstddef>
2  #include <limits>
3  #include <memory>
4  #include <new>
5
6  template <std::size_t N, std::size_t Align = alignof(std::max_align_t)>
7  class arena {
8  public:
9      arena() : ptr_(buf_) {}
10     arena(const arena&) = delete;
11     arena& operator=(const arena&) = delete;
12     ~arena() = default;
13     constexpr std::size_t alignment() const {return Align;}
14     constexpr std::size_t capacity() const {return N;}
15     constexpr std::size_t used() const {return ptr_ - buf_;}
16     constexpr std::size_t free() const {return N - used();}
17     template <std::size_t ReqAlign> void* allocate(std::size_t n);
18     void deallocate(void* ptr, std::size_t n) {}
19     void clear() {ptr_ = buf_;}
20 private:
21     template <std::size_t ReqAlign>
22     static char* align(char* ptr, std::size_t n, std::size_t max);
23     alignas(Align) char buf_[N]; // storage buffer
24     char* ptr_; // pointer to first unused byte
25 };
```

## Fixed-Size Arena Allocator: Code (2)

```
27 template <std::size_t N, std::size_t Align>
28 template <std::size_t ReqAlign>
29 char* arena<N, Align>::align(char* ptr, std::size_t n, std::size_t max) {
30     void* p = ptr;
31     return static_cast<char*>(std::align(ReqAlign, n, p, max));
32 }
33
34 template <std::size_t N, std::size_t Align>
35 template <std::size_t ReqAlign>
36 void* arena<N, Align>::allocate(std::size_t n) {
37     char* ptr = this->align<std::max(Align, ReqAlign)>(ptr_, n, free());
38     if (!ptr) throw std::bad_alloc();
39     ptr_ = ptr + n;
40     return ptr;
41 }
```

# Fixed-Size Arena Allocator: Code (3)

```
43 template <class T, std::size_t N, std::size_t Align = alignof(T)>
44 class salloc {
45 public:
46     using value_type = T;
47     using propagate_on_container_move_assignment = std::true_type;
48     using propagate_on_container_copy_assignment = std::true_type;
49     using propagate_on_container_swap = std::true_type;
50     using is_always_equal = std::false_type;
51     using arena_type = arena<N, Align>;
52     salloc select_on_container_copy_construction() const {return *this;}
53     template <class U> struct rebind {using other = salloc<U, N, Align>;};
54     template <class T2>
55     salloc(const salloc<T2, N, Align>& other) : a_(other.a_) {}
56     salloc(arena_type& a) : a_(&a) {}
57     ~salloc() = default;
58     salloc(const salloc&) = default;
59     salloc(salloc&& other) = default;
60     salloc& operator=(const salloc&) = default;
61     salloc& operator=(salloc&& other) = default;
62     T* allocate(std::size_t n) {
63         if (n > std::numeric_limits<std::size_t>::max() / sizeof(T))
64             {throw std::bad_alloc();}
65         return static_cast<T*>(a_->template allocate<alignof(T)>(
66             n * sizeof(T)));
67     }
68     void deallocate(T* p, std::size_t n)
69         {return a_->deallocate(p, n * sizeof(T));}
```

# Fixed-Size Arena Allocator: Code (4)

```
70 private:
71     template <class T1, std::size_t N1, std::size_t A1, class T2,
72               std::size_t N2, std::size_t A2>
73     friend bool operator==(const salloc<T1, N1, A1>&,
74                            const salloc<T2, N2, A2>&);
75     template <class, std::size_t, std::size_t> friend class salloc;
76     arena_type* a_; // arena from which to allocate storage
77 };
78
79 template <class T1, std::size_t N1, std::size_t A1, class T2, std::size_t N2,
80           std::size_t A2>
81 inline bool operator==(const salloc<T1, N1, A1>& a,
82                        const salloc<T2, N2, A2>& b)
83     {return N1 == N2 && A1 == A2 && a.a_ == b.a_;}
84
85 template <class T1, std::size_t N1, std::size_t A1, class T2, std::size_t N2,
86           std::size_t A2>
87 inline bool operator!=(const salloc<T1, N1, A1>& a,
88                        const salloc<T2, N2, A2>& b)
89     {return !(a == b);}
```



# Fixed-Size Arena Allocator: User Code

```
1  #include <vector>
2  #include <list>
3  #include <iostream>
4  #include "salloc.hpp"
5
6  int main() {
7      using alloc = salloc<int, 1024, sizeof(int)>;
8      alloc::arena_type a;
9      std::vector<int, alloc> v{{0, 1, 2, 3}, a};
10     std::vector<int, alloc> w{{0, 2, 4, 6}, a};
11     std::list<int, alloc> p{{1, 3, 5, 7}, a};
12     std::cout << a.free() << '\n';
13     v.push_back(42);
14     for (auto&& i : v) {std::cout << i << '\n';}
15     for (auto&& i : w) {std::cout << i << '\n';}
16     for (auto&& i : p) {std::cout << i << '\n';}
17     std::cout << a.free() << '\n';
18
19     // std::vector<int, alloc> x(1024);
20     // std::list<int, alloc> y;
21     // ERROR: allocator cannot be default constructed
22 }
```

# Allocator-Aware Containers

- container that uses allocator sometimes referred to as **allocator aware**
- typically much more difficult to develop allocator-aware container than container that does not use allocator
- type of pointer returned by allocator not necessarily same as pointer to element type, which sometimes complicates code somewhat
- much of complexity in implementing allocator-aware container, however, arises from issue of allocator propagation

# The `std::allocator_traits` Class Template

- allocators intended to be used via allocator user (e.g., container) indirectly through traits class `std::allocator_traits`
- declaration:  

```
template <class Alloc> struct allocator_traits;
```
- `allocator_traits` provides uniform interface to allocators used by containers
- some properties of allocator types are optional
- in cases where allocator type did not specify optional properties, `allocator_traits` provides default

# Lateral Allocator Propagation

- properties of allocator in `std::allocator_traits` used to control lateral allocator propagation
- container copy constructor obtains allocator for new container by invoking `select_on_container_copy_construction` in `allocator_traits`
- container move constructor always propagates allocator by move
- container copy assignment replaces allocator (in copied-to container) only if `propagate_on_container_copy_assignment` (POCCA) in `allocator_traits` is true
- container move assignment replaces allocator (in moved-to container) only if `propagate_on_container_move_assignment` (POCMA) in `allocator_traits` is true
- container swap will swap allocators of two containers only if `propagate_on_container_swap` (POCS) in `allocator_traits` is true
- if POCS is false, swapping two standard-library containers with unequal allocators is undefined behavior (since swap must not invalidate iterators and iterators would have to be invalidated in this case)

# Allocator-Traits Querying Example

```
1  #include <memory>
2  #include <type_traits>
3  #include <boost/interprocess/managed_shared_memory.hpp>
4  #include <boost/interprocess/allocators/allocator.hpp>
5  #include <iostream>
6
7  template <class T> void print(std::ostream& out = std::cout) {
8      out << std::is_same_v<typename T::pointer, typename T::value_type*> << ' '
9          << std::is_same_v<typename T::const_pointer,
10             const typename T::value_type*> << ' '
11             << T::is_always_equal::value << ' '
12             << T::propagate_on_container_move_assignment::value << ' '
13             << T::propagate_on_container_copy_assignment::value << ' '
14             << T::propagate_on_container_swap::value << '\n';
15  }
16
17  int main() {
18      namespace bi = boost::interprocess;
19      print<std::allocator_traits<std::allocator<int>>>());
20      print<std::allocator_traits<bi::allocator<int,
21         bi::managed_shared_memory::segment_manager>>>());
22  }
23
24  /* Output:
25  1 1 1 1 0 0
26  0 0 0 0 0 0
27  */
```

# Optional Value Example

- consider container class template called `optval` that can hold optional value
- class templated on element type `T` and allocator type
- container object in one of two states:
  - 1 holding value of type `T`
  - 2 not holding any value
- can query if container is holding value, and if so, access held value
- want to store object of type `T` in memory obtained *from allocator*
- example illustrates basic use of allocator in container class

# Optional Value Example: Code (1)

```
1  #include <memory>
2  #include <type_traits>
3  #include <utility>
4
5  template <class T, class Alloc = std::allocator<T>>
6  class optval : private Alloc {
7  public:
8      using value_type = T;
9      using allocator_type = Alloc;
10 private:
11     using traits = typename std::allocator_traits<Alloc>;
12 public:
13     using pointer = typename traits::pointer;
14     using const_pointer = typename traits::const_pointer;
15     optval(std::allocator_arg_t, const allocator_type& alloc) :
16         Alloc(alloc), value_(nullptr) {}
17     optval() : optval(std::allocator_arg, allocator_type()) {}
18     optval(std::allocator_arg_t, const allocator_type& alloc,
19         const optval& other);
20     optval(const optval& other);
21     optval(std::allocator_arg_t, const allocator_type& alloc, optval&& other)
22         noexcept;
23     optval(optval&& other) noexcept;
24     optval(std::allocator_arg_t, const allocator_type& alloc, const T& value);
25     optval(const T& value);
26     ~optval();
27     optval& operator=(const optval& other);
28     optval& operator=(optval&& other)
29         noexcept (traits::propagate_on_container_move_assignment::value);
30     void swap(optval& other) noexcept;
```

## Optional Value Example: Code (2)

```
31     allocator_type get_allocator() const {return alloc_();}  
32     bool has_value() const noexcept {return value_;}  
33     const T& get() const {return *value_;}  
34     void clear() noexcept;  
35     void set(const T& value);  
36 private:  
37     pointer copy_(allocator_type a, const value_type& value);  
38     allocator_type& alloc_() {return *this;}  
39     const allocator_type& alloc_() const {return *this;}  
40     pointer value_; // pointer to optional value  
41 };  
42  
43 template <class T, class Alloc>  
44 optval<T, Alloc>::optval(const optval& other) : optval(std::allocator_arg,  
45     traits::select_on_container_copy_construction(other.alloc_()), other) {}  
46  
47 template <class T, class Alloc>  
48 optval<T, Alloc>::optval(std::allocator_arg_t, const allocator_type& alloc,  
49     const optval& other) : Alloc(alloc), value_(nullptr) {  
50     if (other.value_) {value_ = copy_(alloc_(), *other.value_);}  
51 }  
52  
53 template <class T, class Alloc>  
54 optval<T, Alloc>::optval(optval&& other) noexcept : Alloc(std::move(other)) {  
55     value_ = other.value_;  
56     other.value_ = nullptr;  
57 }
```



## Optional Value Example: Code (3)

```
59 template <class T, class Alloc>
60 optval<T, Alloc>::optval(std::allocator_arg_t, const allocator_type& alloc,
61     optval&& other) noexcept : Alloc(alloc) {
62     value_ = other.value_;
63     other.value_ = nullptr;
64 }
65
66 template <class T, class Alloc>
67 optval<T, Alloc>::optval(std::allocator_arg_t, const allocator_type& alloc,
68     const T& value) : Alloc(alloc), value_(nullptr)
69     {value_ = copy_(alloc_(), value);}
70
71 template <class T, class Alloc>
72 optval<T, Alloc>::optval(const T& value) : optval(std::allocator_arg,
73     allocator_type(), value) {}
74
75 template <class T, class Alloc>
76 optval<T, Alloc>::~~optval()
77     {clear();}
```

## Optional Value Example: Code (4)

```
79  template <class T, class Alloc>
80  auto optval<T, Alloc>::operator=(const optval& other) -> optval& {
81      if (this != &other) {
82          if constexpr(trait::propagate_on_container_copy_assignment::value) {
83              allocator_type a = other.alloc_();
84              pointer p = other.value_ ? copy_(a, *other.value_) : nullptr;
85              clear();
86              alloc_() = other.alloc_();
87              value_ = p;
88          } else {
89              pointer p = other.value_ ? copy_(alloc_(), *other.value_) : nullptr;
90              clear();
91              value_ = p;
92          }
93      }
94      return *this;
95  }
```

## Optional Value Example: Code (5)

```
97  template <class T, class Alloc>
98  auto optval<T, Alloc>::operator=(optval&& other)
99      noexcept (traits::propagate_on_container_move_assignment::value) -> optval& {
100      using std::swap;
101      if (this != &other) {
102          if constexpr (traits::propagate_on_container_move_assignment::value) {
103              clear();
104              swap(alloc_(), other.alloc_());
105              swap(value_, other.value_);
106          } else if (alloc_() == other.alloc_()) {
107              clear();
108              swap(value_, other.value_);
109          } else {
110              pointer p = copy_(alloc_(), other.value_);
111              swap(value_, other.value_);
112              other.clear();
113              value_ = p;
114          }
115      }
116      return *this;
117  }
118
119  template <class T, class Alloc>
120  void optval<T, Alloc>::swap(optval& other) noexcept {
121      using std::swap;
122      assert(traits::propagate_on_container_swap::value ||
123          alloc_() == other.alloc_());
124      if constexpr (traits::propagate_on_container_swap::value)
125          {swap(alloc_(), other.alloc_());}
126      swap(value_, other.value_);
127  }
```

## Optional Value Example: Code (6)

```
129 template <class T, class Alloc>
130 void optval<T, Alloc>::clear() noexcept {
131     if (value_) {
132         traits::destroy(alloc_(), std::addressof(*value_));
133         traits::deallocate(alloc_(), value_, 1);
134         value_ = nullptr;
135     }
136 }
137
138 template <class T, class Alloc>
139 void optval<T, Alloc>::set(const T& value) {
140     pointer p = copy_(alloc_(), value);
141     clear();
142     value_ = p;
143 }
144
145 template <class T, class Alloc>
146 auto optval<T, Alloc>::copy_(allocator_type a, const value_type& value) ->
147     pointer {
148     pointer p = traits::allocate(alloc_(), 1);
149     try {traits::construct(a, std::addressof(*p), value);}
150     catch (...) {
151         traits::deallocate(a, p, 1);
152         throw;
153     }
154     return p;
155 }
```

# The `std::scoped_allocator_adaptor` Class Template

- when using stateful allocators with nested containers, often need to ensure that allocator state is propagated from parent container to its descendants
- `std::scoped_allocator_adaptor` can be used to address this type of allocator propagation problem (i.e., deep allocator propagation)
- declaration:  

```
template <class OuterAlloc, class... InnerAllocs>  
class scoped_allocator_adaptor : public OuterAlloc;
```
- `OuterAlloc`: allocator type for outermost container in nesting
- `InnerAllocs`: parameter pack with allocator types for each subsequent container in nesting
- if `InnerAllocs` has too few allocator types for number of nesting levels, last allocator type repeated as necessary
- `scoped_allocator_adaptor` useful when all containers in nesting must use same stateful allocator, such as typically case when using shared-memory-segment allocator

# scoped\_allocator\_adaptor Example

```
1  #include <scoped_allocator>
2  #include <vector>
3  #include <list>
4  #include <iostream>
5  #include "salloc.hpp"
6
7  int main() {
8      constexpr std::size_t align = alignof(std::max_align_t);
9      using inner_alloc = salloc<int, 1024, align>;
10     using inner = inner_alloc::value_type;
11     using outer_alloc = salloc<std::list<int, inner_alloc>, 1024,
12         align>;
13     using outer = outer_alloc::value_type;
14     using alloc = std::scoped_allocator_adaptor<outer_alloc,
15         inner_alloc>;
16
17     using container = std::vector<outer, alloc>;
18     alloc::arena_type a;
19     container v(container::allocator_type(a, a));
20     v.reserve(4);
21     std::list<inner, inner_alloc> p({1, 2, 3}, a);
22     v.push_back(p);
23     for (auto&& y : v) {
24         for (auto&& x : y) {
25             std::cout << x << '\n';
26         }
27     }
28 }
```

# scoped\_allocator\_adaptor Example

```
1  #include <vector>
2  #include <scoped_allocator>
3  #include <boost/interprocess/managed_shared_memory.hpp>
4  #include <boost/interprocess/allocators/adaptive_pool.hpp>
5
6  namespace bi = boost::interprocess;
7
8  template <class T>
9  using alloc = typename bi::adaptive_pool<T, typename
10     bi::managed_shared_memory::segment_manager>;
11
12 int main () {
13     using row = std::vector<int, alloc<int>>;
14     using matrix = std::vector<row,
15         std::scoped_allocator_adaptor<alloc<row>>>;
16     bi::managed_shared_memory s(bi::create_only, "data", 8192);
17     matrix v(s.get_segment_manager());
18     v.resize(4);
19     for (int i = 0; i < 4; ++i) {v[i].push_back(0);}
20     bi::shared_memory_object::remove("data");
21 }
```

## Section 3.8.4

### **References**



- 1 T. Koppe, A Visitor's Guide to C++ Allocators,  
[https://rawgit.com/google/cxx-std-draft/allocator-paper/allocator\\_user\\_guide.html](https://rawgit.com/google/cxx-std-draft/allocator-paper/allocator_user_guide.html).

- 1 Alisdair Meredith. Making Allocators Work. CppCon, Sept. 10, 2014. Available online at <http://youtu.be/YkiYOP3d64E> and <http://youtu.be/Q5kyiFevMJQ>. (This talk is in two parts.)
- 2 Alisdair Meredith. Allocators in C++11. C++Now, Aspen, CO, USA, May 2013. Available online at [https://youtu.be/v7B\\_8IbHjxA](https://youtu.be/v7B_8IbHjxA).
- 3 Andrei Alexandrescu. `std::allocator` is to Allocation What `std::vector` is to Vexation. CppCon, Bellevue, WA, USA, Sept. 24, 2015. Available online at <https://youtu.be/LIb3L4vKZ7U>.
- 4 Alisdair Meredith. An allocator model for `std2`. CppCon, Bellevue, WA, USA, Sept. 25, 2017. Available online at [https://youtu.be/oCi\\_QZ6K\\_qk](https://youtu.be/oCi_QZ6K_qk).  
This talk explains how allocators evolved from C++98 to C++17 and briefly how they might be further evolved in future versions of the C++ standard.

- 5 Bob Steagall. How to Write a Custom Allocator. CppCon, Bellevue, WA, USA, Sept. 28, 2017. Available online at <https://youtu.be/kSWfush1vB8>.

This talk discusses how to write allocators for C++14/C++17 and how to use such allocators in containers.

- 6 Bob Steagall. Testing the Limits of Allocator Awareness. C++Now, Aspen, CO, USA, May 18, 2017. Available online at <https://youtu.be/fmJfKm9ano8>.

This talk briefly introduces allocators and then describes a test suite for allocators and presents some results obtained with this test suite.

- 7 Pablo Halpern. Modern Allocators: The Good Parts. CppCon, Bellevue, WA, USA, Sept. 29, 2017. Available online at <https://youtu.be/v3dz-AK0VL8>.

This talk introduces polymorphic allocators and considers a simple example of a polymorphic allocator and a container that uses a polymorphic allocator.

- 8** Sergey Zubkov. From Security to Performance to GPU Programming: Exploring Modern Allocators. CppCon, Bellevue, WA, USA, Sept. 25, 2017. Available online at <https://youtu.be/HdQ4aOZyuHw>.
- 9** Stephan Lavavej. STL Features and Implementation Techniques. CppCon, Bellevue, WA, USA, 2014. Available online at <https://youtu.be/dTeKf5Oek2c>.  
This talk briefly discusses allocators in C++11 at 26:26–31:32.
- 10** Arthur O'Dwyer. An Allocator is a Handle to a Heap. C++Now, Aspen, CO, USA, May 7, 2018. Available online at <https://youtu.be/0MdSJsCTRkY>.
- 11** Michael Spencer. Alignment in C++: Use, Limitations, and Extension. C++Now, Aspen, CO, USA, May 12–17, 2013. Available online at <https://youtu.be/uSZFrmhayIM>. [The part of the talk in 00:00-34:13 is likely to be of most interest.]

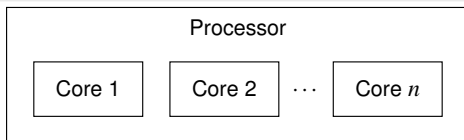
## Section 3.9

# Concurrency

## Section 3.9.1

### **Preliminaries**

# Processors



- A **core** is an independent processing unit that reads and executes program instructions, and consists of registers, an arithmetic logic unit (ALU), a control unit, and usually a cache.
- A **processor** is a computing element that consists of one or more cores, an external bus interface, and possibly a shared cache.
- A **thread** is a sequence of instructions (which can be executed by a core).
- At any given time, a core can execute one thread or, if the core supports simultaneous multithreading (such as hyperthreading), multiple threads.
- In the simultaneous multithreading case, the threads share the resources of the core.
- A processor with more than one core is said to be **multicore**.
- Most modern processors are multicore.
- Multicore processors can *simultaneously* execute *multiple* threads.

- A multicore processor said to be **homogeneous** if all of its cores are identical.
- A multicore processor said to be **heterogeneous** if its has more than one type of core.
- Different types of cores might be used in order to:
  - provide different types of functionality (e.g., CPU and GPU)
  - provide different levels of performance (e.g., high-performance CPU and energy-efficient CPU)



# Memory Hierarchy

SKIP SLIDE



- The component of a system that stores program instructions and data is called **main memory**.
- A **cache** is fast memory used to store copies of instructions and/or data from main memory.
- Main memory is **very slow** compared to the speed of a processor core.
- Due to the latency of main memory, caches are **essential** for good performance.
- Instruction and data caches may be **separate** or **unified** (i.e., combined).
- A cache may be **local** to single core or **shared** between two or more cores.
- The lowest-level (i.e., L1) cache is usually on the core and local to the core.
- The higher-level (i.e., L2, L3, . . . , LL [last level]) caches are usually shared between some or all of the cores.

- Intel Core i7-3820QM Processor (Q2 2012)
  - used in Lenovo W530 notebook
  - 64 bit, 2.7 GHz
  - 128/128 KB L1 cache, 1 MB L2 cache, 8 MB L3 cache
  - *4 cores*
  - *8 threads* (2 threads/core)
- Intel Core i7-5960X Processor Extreme Edition (Q3 2014)
  - targets desktops/notebooks
  - 64 bit, 3 GHz
  - 256/256 KB L1 cache, 2 MB L2 cache, 20 MB L3 cache
  - *8 cores*
  - *16 threads* (2 threads/core)
- Intel Xeon Processor E7-8890 v2 (Q1 2014)
  - targets servers
  - 64 bit, 2.8 GHz
  - 480/480 KB L1 cache, 3.5 MB L2 cache, 37.5 MB L3 cache
  - *15 cores*
  - *30 threads* (2 threads/core)

- Qualcomm Snapdragon 805 SoC (Q1 2014)
  - used in *Google Nexus 6*
  - 32-bit 2.7 GHz *quad-core* Qualcomm Krait 450 (ARMv7-A)
  - 16/16 KB L1 cache (per core), 2 MB L2 cache (shared)
  - 600 MHz Qualcomm Adreno 420 GPU
- Samsung Exynos 5 Octa 5433 SoC
  - used in *Samsung Galaxy Note 4*
  - high-performance 1.9 GHz *quad-core* ARM Cortex-A57 paired with energy-efficient 1.3 GHz *quad-core* ARM Cortex-A53 (big.LITTLE); both 32-bit (64-bit capable but disabled) (ARMv8-A)
  - Cortex-A57: 48/32 KB L1 cache, 512 KB to 2 MB L2 cache?
  - 700 MHz Mali-T760MP6 GPU
- Apple A8 SoC (2014)
  - used in *Apple iPhone 6*, *Apple iPhone 6 Plus*
  - 64-bit 1.4 GHz *dual-core* CPU (ARMv8-A)
  - 64/64 KB L1 cache (per core), 1 MB L2 cache (shared), 4 MB L3 cache
  - PowerVR Series 6XT GX6450 (quad-core) GPU

# Why Multicore Processors?

- in past, greater processing power obtained through *higher clock rates*
- *clock rates have stopped rising*, topping out at about 5 GHz (little change since about 2005)
- power consumption is linear in clock frequency and quadratic in voltage, but higher frequency typically requires higher voltage; so, considering effect of frequency and voltage together, power consumption grows approximately with *cube* of frequency
- greater power consumption translates into *increased heat production*
- higher clock rates would result in processors *overheating*
- transistor counts *still increasing* (Moore's law: since 1960s, transistor count has doubled approximately every 18 months)
- instead of increasing processing power by raising clock rate of processor core, simply *add more processor cores*
- $n$  cores running at clock rate  $f$  use significantly less power and generate less heat than single core at clock rate  $nf$
- going multicore allows for *greater processing power* with *lower power consumption* and *less heat production*

## Section 3.9.2

# **Multithreaded Programming**

# Concurrency

- A **thread** is a sequence of instructions that can be independently managed by the operating-system scheduler.
- A **process** provides the resources that program needs to execute (e.g., address space, files, and devices) and at least one thread of execution.
- All threads of a process share the *same* address space.
- **Concurrency** is the situation where multiple threads execute over time periods (i.e., from start of execution to end) that *overlap* (but no threads are required to run simultaneously).
- **Parallelism** refers to the situation where multiple threads execute *simultaneously*.
- Concurrency can be achieved with:
  - 1 multiple single-threaded processes; or
  - 2 a single multithreaded process.
- A single multithreaded process is usually preferable, since this approach is typically much less resource intensive and data can often be shared much more easily between threads in a single process (due to the threads having a common address space).

# Why Multithreading?

- Keep all of the processor cores busy (i.e., *fully utilize* all cores).
  - Most modern systems have multiple processor cores, due to having either multiple processors or a single processor that is multicore.
  - A single thread cannot fully utilize the computational resources available in such systems.
- Keep processes *responsive*.
  - In graphics applications, keep the GUI responsive while the application is performing slow operations such as I/O.
  - In network server applications, keep the server responsive to new connections while handling already established ones.
- *Simplify* the coding of cooperating tasks.
  - Some programs consist of several logically distinct tasks.
  - Instead of having the program manage when the computation associated with different tasks is performed, each task can be placed in a separate thread and the operating system can perform scheduling.
  - For certain types of applications, multithreading can significantly reduce the conceptual complexity of the program.

## Section 3.9.3

# **Multithreaded Programming Models**



- A **memory model** (also known as a **memory-consistency model**) is a formal specification of the effect of read and write operations on the memory system, which in effect describes how memory appears to programs.
- A memory model is essential in order for the semantics of a multithreaded program to be well defined.
- The memory model must address issues such as:
  - ordering
  - atomicity
- The memory model affects:
  - programmability (i.e., ease of programming)
  - performance
  - portability

# Sequential Consistency (SC)

- The environment in which a multithreaded program is run is said to have **sequential consistency (SC)** if the result of any execution of the program is the same as if the operations of all threads are executed in *some sequential order*, and the operations of each individual thread appear in this sequence in *the order specified by the program*.
- In other words, in a sequentially-consistent execution of a multithreaded program, threads behave as if their operations were simply *interleaved*.
- Consider the multithreaded program (with two threads) shown below, where x, y, a, and b are all integer variables and initially zero.

## Thread 1 Code

```
x = 1;  
a = y;
```

## Thread 2 Code

```
y = 1;  
b = x;
```

- Some sequentially-consistent executions of this program include:
  - x = 1; y = 1; b = x; a = y;
  - y = 1; x = 1; a = y; b = x;
  - x = 1; a = y; y = 1; b = x;
  - y = 1; b = x; x = 1; a = y;

# Sequential-Consistency (SC) Memory Model

- Since SC implies that memory must behave in a particular manner, SC implicitly defines a memory model, known as the **SC memory model**.
- In particular, SC implies that each write operation is *atomic* and becomes visible to all threads *simultaneously*.
- Thus, with the SC model, *all* threads see write operations on memory occur *atomically* in the *same* order, leading to all threads having a *consistent view* of memory.
- The SC model precludes (or makes extremely difficult) many hardware optimizations, such as:
  - store buffers
  - caches
  - out-of-order instruction execution
- The SC model also precludes many compiler optimizations, including:
  - reordering of loads and stores
- Although the SC model very is *intuitive*, it comes at a *very high cost* in terms of performance.

# Load/Store Reordering Example: Single Thread

- Consider the program with the code below, where  $x$  and  $y$  are integer variables, all initially zero.

## Original Thread 1 Code

```
x = 1;  
y = 1;  
// ...
```

- Suppose that, during optimization, the compiler transforms the preceding code to that shown below, effectively *reordering two stores*.

## Optimized Thread 1 Code

```
y = 1;  
x = 1;  
// ...
```

- The execution of the optimized code is *indistinguishable* from a sequentially-consistent execution of the original code.
- The optimized program runs *as if* it were the original program.
- In a *single-threaded* program, loads and stores can be reordered without invalidating the SC model (if data dependencies are correctly considered).

# Load/Store Reordering Example: Multiple Threads

- Consider the addition of a second thread to the program to yield the code below.

## Original Thread 1 Code

```
x = 1;  
y = 1;  
// ...
```

## Thread 2 Code

```
if (y == 1) {  
    assert(x == 1);  
}
```

- Suppose that the compiler makes the same optimization to the code for thread 1 as on the previous slide, yielding the code below.

## Optimized Thread 1 Code

```
y = 1;  
x = 1;  
// ...
```

## (Unchanged) Thread 2 Code

```
if (y == 1) {  
    assert(x == 1);  
}
```

- Thread 2 can observe `x` and `y` being modified in the wrong order (i.e., an order that is inconsistent with SC execution).
- The assertion in thread 2 can never fail in the original program, but can sometimes fail in the optimized program.
- In a *multithreaded* program, the reordering of loads and stores must be avoided *if SC is to be maintained*.

# Store-Buffer Example: Without Store Buffer

SKIP SLIDE

- Consider the program below, where  $x$ ,  $y$ ,  $a$ , and  $b$  are integer variables, all initially zero.

## Thread 1 Code

```
x = 1;  
a = y;
```

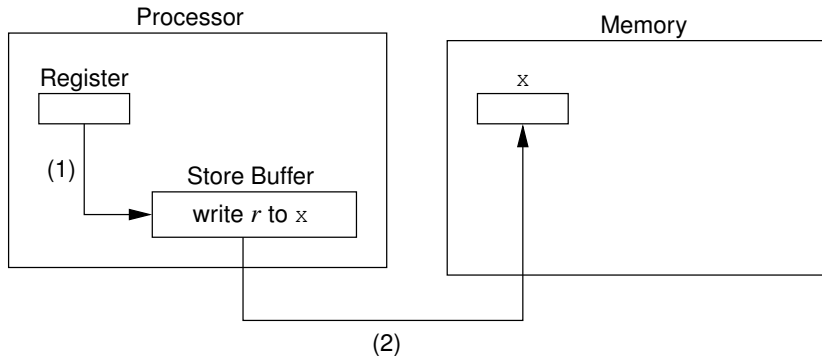
## Thread 2 Code

```
y = 1;  
b = x;
```

- Some possible sequentially-consistent executions of the program include:
  - $x = 1$ ;  $y = 1$ ;  $b = x$ ;  $a = y$ ; ( $a$  is 1,  $b$  is 1)
  - $y = 1$ ;  $x = 1$ ;  $a = y$ ;  $b = x$ ; ( $a$  is 1,  $b$  is 1)
  - $x = 1$ ;  $a = y$ ;  $y = 1$ ;  $b = x$ ; ( $a$  is 0,  $b$  is 1)
  - $y = 1$ ;  $b = x$ ;  $x = 1$ ;  $a = y$ ; ( $a$  is 1,  $b$  is 0)
- In every sequentially-consistent execution of the program, one of “ $x = 1$ ,” or “ $y = 1$ ,” must execute first.
- If “ $x = 1$ ,” executes first, then  $b$  cannot be assigned 0.
- If “ $y = 1$ ,” executes first, then  $a$  cannot be assigned 0.
- No sequentially-consistent execution can result in  $a$  and  $b$  *both* being 0.

# Store-Buffer Example: Store Buffer

SKIP SLIDE



- (1) transfer data from register to store buffer
- (2) flush store buffer to memory

# Store-Buffer Example: With Store Buffer (Not SC)

SKIP SLIDE

Core 1		Core 2		Memory	
Code	Store Buffer	Code	Store Buffer	x	y
<code>x = 1;</code>	write 1 to x pending			0	0
	no change	<code>y = 1;</code>	write 1 to y pending	0	0
<code>a = y;</code> <code>// a = 0;</code>	no change		no change	0	0
	no change	<code>b = x;</code> <code>// b = 0;</code>	no change	0	0
	write 1 to x completed		no change	1	0
			write 1 to y completed	1	1

- The execution of the program results in `a` and `b` *both* being 0, which *violates SC*.
- The program behaves as if the lines of code in each thread were *reordered* (i.e., reversed), yielding: `a = y; b = x; x = 1; y = 1;`.
- A store buffer (or cache) must be avoided, *if SC is to be maintained*.



# Atomicity of Memory Operations

- A fundamental property of SC is that all memory operations are *atomic*.
- Atomic memory operations require *synchronization* between processor cores.
- This synchronization *greatly increases the time required to access memory*, as a result of the time needed by processor cores to communicate and coordinate access to memory.
- Therefore, requiring all memory operations to be atomic is not desirable.
- Allowing non-atomic memory operations, however, would be *inconsistent with a fundamental property of SC*.

# Data Races

- If memory operations are *not all atomic*, the possibility exists for something known as a data race.
- Two memory operations are said to **conflict** if they access the *same* memory location and *at least one* of the operations is a write.
- Two conflicting memory operations form a **data race** if they are from different threads and can be executed *at the same time*.
- A program with data races usually has unpredictable behavior (e.g., due to torn reads, torn writes, or worse).
- Example (data race):
  - Consider the multithreaded program listed below, where *x*, *y*, and *z* are (nonatomic) integer variables shared between threads and are initially zero.

## Thread 1 Code

```
x = 1;  
a = y + z;
```

## Thread 2 Code

```
y = 1;  
b = x + z;
```

- The program has data races on both *x* and *y*.
- Since *z* is not modified by any thread, *z* cannot participate in a data race.

# Torn Reads

- A **torn read** is a read operation that (due to lack of atomicity) has only partially read its value when another (concurrent) write operation on the same location is performed.
- Consider a two-byte unsigned (big-endian) integer variable  $x$ , which is initially 1234 (hexadecimal).
- Suppose that the following (nonatomic) memory operations overlap in time:
  - thread 1 reads  $x$ ; and
  - thread 2 writes 5678 (hexadecimal) to  $x$ .
- Initially,  $x$  is 1234: 

Byte 0	Byte 1
12	34
- Thread 1 reads 12 from the first byte of  $x$ .
- Thread 2 writes 56 and 78 to the first and second bytes of  $x$ , respectively, yielding: 

Byte 0	Byte 1
56	78
- Thread 1 reads the second byte of  $x$  to obtain the value 78.
- The value read by thread 1 (i.e., 1278) is neither the value of  $x$  prior to the write by thread 2 (i.e., 1234) nor the value of  $x$  after the write by thread 2 (i.e., 5678).

# Torn Writes

- A **torn write** is a write operation that (due to lack of atomicity) has only partially written its value when another (concurrent) read or write operation on the same location is performed.
- Consider a two-byte unsigned (big-endian) integer variable  $x$ , which is initially 0.
- Suppose that the following (nonatomic) memory operations overlap in time:
  - thread 1 writes 1234 (hexadecimal) to  $x$ ; and
  - thread 2 writes 5678 (hexadecimal) to  $x$ .
- Initially,  $x$  is 0: 

Byte 0	Byte 1
00	00
- Thread 1 writes 12 to the first byte of  $x$ , yielding: 

Byte 0	Byte 1
12	00
- Thread 2 writes 56 and 78 to the first and second bytes of  $x$ , respectively, yielding: 

Byte 0	Byte 1
56	78
- Thread 1 writes 34 to the second byte of  $x$ , yielding: 

Byte 0	Byte 1
56	34
- The resulting value in  $x$  (i.e., 5634) is neither the value written by thread 1 (i.e., 1234) nor the value written by thread 2 (i.e., 5678).

# SC Data-Race Free (SC-DRF) Memory Model

- From a programmability standpoint, SC is extremely desirable, as it allows one to reason easily about the behavior of a multithreaded program.
- Unfortunately, as we saw earlier, SC precludes almost all useful compiler optimizations and hardware optimizations.
- As it turns out, if we drop the requirement that all memory operations be atomic and then restrict programs to be data-race free, SC can be provided while still allowing most compiler and hardware optimizations.
- This observation is the motivation behind the so called SC-DRF memory model.
- The **sequential-consistency for data-race free programs (SC-DRF) model** provides SC *only for programs that are data-race free*.
- The data-race free constraint is not overly burdensome, since data races will likely result in bugs anyhow.
- Several programming languages have used SC-DRF as the basis for their memory model, including C++, C, and Java.

- The C++ programming language employs, at its default memory model, the *SC-DRF* model.
- Again, with the SC-DRF model, a program behaves as if its execution is sequentially consistent, provided that the program is data-race free.
- Support is also provided for other (more relaxed) memory models.
- For certain memory accesses, it is possible to override the default (i.e., SC-DRF) memory model, if desired.
- The execution of a program that is not data-race free results in *undefined behavior*.

## Section 3.9.4

# **Thread Management**

# The `std::thread` Class

- `std::thread` class provides means to create new thread of execution, wait for thread to complete, and perform other operations to manage and query state of thread
- `thread` object may or may not be associated with thread (of execution)
- `thread` object that is associated with thread said to be **joinable**
- default constructor creates `thread` object that is **unjoinable**
- can also construct `thread` object by providing callable entity (e.g., function or functor) and arguments (if any), resulting in new thread invoking callable entity
- thread function provided with **copies** of arguments so must use reference wrapper class like `std::reference_wrapper` for reference semantics
- `thread` class is movable but **not copyable** [C++17 33.3.2.1/1]
- each `thread` object has ID
- IDs of joinable `thread` objects are **unique** [C++17 33.3.2.1/1]
- all unjoinable `thread` objects have **same** ID, distinct from ID of every joinable `thread` object [C++17 33.3.2.1/1]



# The `std::thread` Class (Continued)

- **join operation** waits for `thread` object's thread to complete execution and results in object becoming `unjoinable`
- **detach operation** dissociates thread from `thread` object (allowing thread to continue to execute independently) and results in object becoming `unjoinable`
- using `thread` object as source for move operation results in object becoming `unjoinable` [C++17 33.3.2.2/10]
- if `thread` object joinable when destructor called, exception is thrown
- static member function `hardware_concurrency` returns number of hardware threads that can run simultaneously (or zero if not well defined)
- thread creation and join operations establish synchronizes-with relationship (to be discussed later)

## Member Types

Member Name	Description
<code>id</code>	thread ID type
<code>native_handle_type</code>	system-dependent handle type for underlying thread entity

## Construction, Destruction, and Assignment

Member Name	Description
constructor	construct thread (overloaded)
destructor	destroy thread
<b>operator=</b>	move assign thread

## Member Functions

Member Name	Description
joinable	check if thread joinable
get_id	get ID of thread
native_handle	get native handle for thread
hardware_concurrency (static)	get number of concurrent threads supported by hardware
join	wait for thread to finish executing
detach	permit thread to execute independently
swap	swap threads

# Example: Hello World With Threads

```
1  #include <iostream>
2  #include <thread>
3
4  void hello()
5  {
6      std::cout << "Hello World!\n";
7  }
8
9  int main()
10 {
11     std::thread t(hello);
12     t.join();
13 }
```

---

```
1  #include <iostream>
2  #include <thread>
3
4  int main()
5  {
6      std::thread t([](){
7          std::cout << "Hello World!\n";
8      });
9      t.join();
10 }
```

# Example: Thread-Function Argument Passing (Copy/Move Semantics)

```
1  #include <iostream>
2  #include <vector>
3  #include <utility>
4  #include <thread>
5
6  void doWork(const std::vector<int>& v) {
7      for (auto i : v) {
8          std::cout << i << '\n';
9      }
10 }
11
12 int main() {
13     std::vector v{1, 2, 3, 4};
14
15     // copy semantics
16     std::thread t1(doWork, v);
17     t1.join();
18
19     // move semantics
20     std::thread t2(doWork, std::move(v));
21     t2.join();
22 }
```

# Example: Thread-Function Argument Passing (Reference Semantics)

```
1  #include <iostream>
2  #include <vector>
3  #include <functional>
4  #include <thread>
5
6  void doWork(const std::vector<int>& v) {
7      for (auto i : v) {
8          std::cout << i << '\n';
9      }
10 }
11
12 int main() {
13     std::vector v{1, 2, 3, 4};
14
15     // copy semantics
16     std::thread t1(doWork, v);
17     t1.join();
18
19     // reference semantics
20     std::thread t2(doWork, std::ref(v));
21     t2.join();
22 }
```

# Example: Moving Threads

```
1  #include <thread>
2  #include <iostream>
3  #include <utility>
4
5  // Return a thread that prints a greeting message.
6  std::thread makeThread() {
7      return std::thread([](){
8          std::cout << "Hello World!\n";
9      });
10 }
11
12 // Return the same thread that was passed as an argument.
13 std::thread identity(std::thread t) {
14     return t;
15 }
16
17 int main() {
18     std::thread t1(makeThread());
19     std::thread t2(std::move(t1));
20     t1 = std::move(t2);
21     t1 = identity(std::move(t1));
22     t1.join();
23 }
```

# Example: Lifetime Bug

```
1  #include <iostream>
2  #include <vector>
3  #include <algorithm>
4  #include <chrono>
5  #include <thread>
6  #include <numeric>
7
8  void threadFunc(const std::vector<int>* v) {
9      std::cout << std::accumulate(v->begin(), v->end(), 0)
10         << '\n';
11 }
12
13 void startThread() {
14     std::vector<int> v(1000000, 1);
15     std::thread t(threadFunc, &v);
16     t.detach();
17     // v is destroyed here but detached thread
18     // may still be using v
19 }
20
21 int main() {
22     startThread();
23     // Give the thread started by startThread
24     // sufficient time to complete its work.
25     std::this_thread::sleep_for(std::chrono::seconds(5));
26 }
```



# The `std::this_thread` Namespace

SKIP SLIDE

Name	Description
<code>get_id</code>	get ID of current thread
<code>yield</code>	suggest rescheduling current thread so as to allow other threads to run
<code>sleep_for</code>	blocks execution of current thread for at least specified duration
<code>sleep_until</code>	blocks execution of current thread until specified time reached

# Example: Identifying Threads

SKIP SLIDE

```
1  #include <thread>
2  #include <iostream>
3
4  // main thread ID
5  std::thread::id mainThread;
6
7  void func() {
8      if (std::this_thread::get_id() == mainThread) {
9          std::cout << "called by main thread\n";
10     } else {
11         std::cout << "called by secondary thread\n";
12     }
13 }
14
15 int main() {
16     mainThread = std::this_thread::get_id();
17     std::thread t([]){
18         // call func from secondary thread
19         func();
20     };
21     // call func from main thread
22     func();
23     t.join();
24 }
```

- **thread storage duration**: object initialized before first use in thread and, if constructed, destroyed on thread exit
- each thread has its own instance of object
- only objects declared **thread\_local** have this storage duration
- **thread\_local** implies **static** for variable of block scope
- **thread\_local** can appear together with **static** or **extern** to adjust linkage
- example:

```
thread_local int counter = 0;
static thread_local int x = 0;
thread_local int y;

void func() {
    thread_local int counter = 0;
    // equivalent to:
    // static thread_local int counter = 0;
}
```

# Example: Thread Local Storage

SKIP SLIDE

```
1  #include <iostream>
2  #include <vector>
3  #include <thread>
4
5  thread_local int counter = 0;
6
7  void doWork(int id) {
8      static const char letters[] = "abcd";
9      for (int i = 0; i < 10; ++i) {
10         std::cout << letters[id] << counter << '\n';
11         ++counter;
12     }
13 }
14
15 int main() {
16     std::vector<std::thread> workers;
17     for (int i = 1; i <= 3; ++i) {
18         // invoke doWork in new thread
19         workers.emplace_back(doWork, i);
20     }
21     // invoke doWork in main thread
22     doWork(0);
23     for (auto& t : workers) {t.join();}
24 }
```

# The `std::thread` Class and Exception Safety

- The astute reader will notice that most code examples on these lecture slides (both earlier and later) that directly employ `std::thread` are not exception safe.
- Some of the exception safety problems in these examples could be eliminated by using a RAII class to wrap `std::thread` objects.
- Unfortunately, the standard library does not provide such a RAII class.
- At a very basic level, one could provide a thread wrapper class that has similar functionality to `std::thread`, except that its destructor automatically joins with the underlying thread if the thread is still joinable at destruction time. (See next slide.)
- Although such an approach will work in some situations (such as in the case of many of the simple code examples on these lecture slides), it can potentially lead to deadlocks and other problems in more complex code.
- A more general solution would be to provide a class that allows arbitrary code to be executed just prior to thread destruction, in order to perform the appropriate (application-dependent) “clean-up” action. (For example, see `boost::scoped_thread` in the Boost Threads library.)

# The `std::thread` Class and Exception Safety (Continued)

`scoped_thread_1_1.hpp`

```
1  #include <thread>
2
3  // A minimalist inheritance-based replacement for std::thread
4  // that joins automatically in the destructor.
5  // (One must be careful not to use this type polymorphically
6  // since the destructor is not virtual.)
7  class scoped_thread : public std::thread {
8  public:
9      using std::thread::thread;
10     scoped_thread(scoped_thread&&) = default;
11     scoped_thread& operator=(scoped_thread&&) = default;
12     scoped_thread(const scoped_thread&) = delete;
13     scoped_thread& operator=(const scoped_thread&) = delete;
14     ~scoped_thread() if (joinable()) {join();}
15 };
```

`main.cpp`

```
1  #include <vector>
2  #include <unistd.h>
3  #include "scoped_thread_1_1.hpp"
4
5  void worker() {sleep(1);}
6
7  int main() {
8      std::vector<scoped_thread> threads;
9      for (int i = 0; i < 16; ++i) {threads.emplace_back(worker);}
10 }
```

## Section 3.9.5

# **Sharing Data Between Threads**

# Shared Data

- In multithreaded programs, it is often necessary to *share resources* between threads.
- Shared resources might include such things as variables, memory, files, devices, and so on.
- The sharing of resources, however, can lead to various problems when multiple threads want access to the same resource simultaneously.
- The most commonly shared resource is *variables*.
- When variables are shared between threads, the possibility exists that one thread may attempt to access a variable while another thread is modifying the same variable.
- Such *conflicting accesses* to variables can lead to data corruption and other problems.
- More generally, when any resource is shared, the potential for problems exists.
- Therefore, mechanisms are needed for ensuring that shared resources can be accessed safely.



# Race Conditions

- A **race condition** is a behavior where the outcome depends on the relative ordering of the execution of operations on two or more threads.
- Sometimes, a race condition may be benign (i.e., does not cause any problem).
- Usually, the term “race condition” used to refer to a race condition that is not benign (i.e., breaks invariants or results in undefined behavior).
- A data race is a particularly evil type of race condition.
- A **deadlock** is a situation in which two or more threads are unable to make progress due to being **blocked** waiting for resources held by each other.
- A **livelock** is a situation in which two or more threads are **not blocked** but are unable to make progress due to needing resources held by each other.
- Often, race conditions can lead to deadlocks, livelocks, crashes, and other unpredictable behavior.

- A **critical section** is a piece of code that accesses a shared resource (e.g., data structure) that must not be simultaneously accessed by more than one thread.
- A synchronization mechanism is needed at the entry to and exit from a critical section.
- The mechanism needs to provide *mutual exclusion* (i.e., prevent critical sections in multiple threads from executing simultaneously).
- Example (FIFO queue):
  - One thread is adding an element to a queue while another thread is removing an element from the same queue.
  - Since both threads modify the queue at the same time, they could corrupt the queue data structure.
  - Synchronization must be employed so that the execution of the parts of the code that add and remove elements are executed in a *mutually exclusive* manner (i.e., cannot run at the same time).

# Data-Race Example

## Shared (Global) Data

```
double balance = 100.00; // bank account balance  
double credit = 50.00; // amount to deposit  
double debit = 10.00; // amount to withdraw
```

## Thread 1 Code

```
// double tmp = balance;  
// tmp = tmp + credit;  
// balance = tmp;  
balance += credit;
```

## Thread 2 Code

```
// double tmp = balance;  
// tmp = tmp - debit;  
// balance = tmp;  
balance -= debit;
```

- above code has data race on `balance` object (i.e., more than one thread may access `balance` at same time with at least one thread writing)

# Example: Data Race (Counter)

```
1  #include <iostream>
2  #include <thread>
3
4  unsigned long long counter = 0;
5
6  void func() {
7      for (unsigned long long i = 0; i < 1'000'000; ++i) {
8          ++counter;
9      }
10 }
11
12 int main() {
13     std::thread t1(func);
14     std::thread t2(func);
15     t1.join();
16     t2.join();
17     std::cout << counter << '\n';
18 }
```

## Example: Data Race and/or Race Condition (IntSet)

```
1  #include <iostream>
2  #include <thread>
3  #include <unordered_set>
4
5  class IntSet {
6  public:
7      bool contains(int i) const
8          {return s_.find(i) != s_.end();}
9      void add(int i)
10         {s_.insert(i);}
11 private:
12     std::unordered_set<int> s_;
13 };
14
15 IntSet s;
16
17 int main() {
18     std::thread t1([](){
19         for (int i = 0; i < 10'000; ++i) {s.add(2 * i);}
20     });
21     std::thread t2([](){
22         for (int i = 0; i < 10'000; ++i) {s.add(2 * i + 1);}
23     });
24     t1.join(); t2.join();
25     std::cout << s.contains(1000) << '\n';
26 }
```

## Section 3.9.6

# **Mutexes**

# Mutexes

- A **mutex** is a locking mechanism used to synchronize access to a shared resource by providing *mutual exclusion*.
- A mutex has two basic operations:
  - **acquire**: lock (i.e., hold) the mutex
  - **release**: unlock (i.e., relinquish) the mutex
- A mutex can be held *by only one thread* at any given time.
- If a thread attempts to acquire a mutex that is already held by another thread, the operation will either block until the mutex can be acquired or fail with an error.
- A thread holding a (nonrecursive) mutex *cannot relock* the mutex.
- A thread **acquires** the mutex **before** accessing the shared resource and **releases** the mutex **when finished** accessing the resource.
- Since only one thread can hold a mutex at any given time and the shared resource is only accessed by the thread holding the mutex, mutually-exclusive access is guaranteed.

# The `std::mutex` Class

- `std::mutex` class provides mutex functionality
- *not movable* and *not copyable* [\[C++17 30.4.1.2/3\]](#)
- `lock` member function acquires mutex (blocking as necessary)
- `unlock` member function releases mutex
- thread that owns mutex should not attempt to lock mutex again [\[C++17 33.4.3.2/7\]](#)
- all prior `unlock` operations on given mutex *synchronize with* `lock` operation (on *same* mutex) (*synchronizes-with* relationship to be discussed *later*) [\[C++17 33.4.3.2/11\]](#)



## Member Types

Name	Description
<code>native_handle_type</code>	system-dependent handle type for underlying mutex entity

## Construction, Destruction, and Assignment

Name	Description
constructor	construct mutex
destructor	destroy mutex

## Other Member Functions

Name	Description
<code>lock</code>	acquire mutex, blocking if not available
<code>try_lock</code>	try to lock mutex without blocking
<code>unlock</code>	release mutex
<code>native_handle</code>	get handle for underlying mutex entity

## Example: Avoiding Data Race Using Mutex (Counter) (mutex)

```
1  #include <iostream>
2  #include <thread>
3  #include <mutex>
4
5  std::mutex m;
6  unsigned long long counter = 0;
7
8  void func() {
9      for (unsigned long long i = 0; i < 1'000'000; ++i) {
10         m.lock(); // acquire mutex
11         ++counter;
12         m.unlock(); // release mutex
13     }
14 }
15
16 int main() {
17     std::thread t1(func);
18     std::thread t2(func);
19     t1.join();
20     t2.join();
21     std::cout << counter << '\n';
22 }
```

- above code fixes data race from earlier example

# The `std::scoped_lock` Template Class

- `std::scoped_lock` is RAII class for mutexes
- declaration:  

```
template <class... Ts> class scoped_lock;
```
- parameter pack `Ts` specifies types of mutexes to be locked
- can be used with any mutex types providing necessary locking interface (e.g., `std::mutex` and `std::recursive_mutex`)
- constructor takes one or more mutexes as arguments
- mutexes acquired in constructor and released in destructor
- `scoped_lock` objects are *not movable* and *not copyable*
- using `scoped_lock` avoids problem of inadvertently failing to release mutexes (e.g., due to exception or forgetting `unlock` calls)
- in multiple mutex case, employs deadlock avoidance algorithm from `std::lock` (discussed later) when acquiring mutexes
- advisable to use `scoped_lock` instead of calling `lock` and `unlock` explicitly
- `scoped_lock` effectively replaces (and extends) `lock_guard`

## Member Types

Name	Description
<code>mutex_type</code>	underlying mutex type if only one

## Construction, Destruction, and Assignment

Name	Description
constructor	construct scoped lock, optionally locking given mutexes
destructor	destroy scoped lock, releasing underlying mutexes

## Example: Avoiding Data Race Using Mutex (Counter) (scoped\_lock)

```
1  #include <iostream>
2  #include <thread>
3  #include <mutex>
4
5  std::mutex m;
6  unsigned long long counter = 0;
7
8  void func() {
9      for (unsigned long long i = 0; i < 1'000'000; ++i) {
10         // scoped_lock constructor acquires mutex
11         std::scoped_lock lock(m);
12         ++counter;
13         // scoped_lock destructor releases mutex
14     }
15 }
16
17 int main() {
18     std::thread t1(func);
19     std::thread t2(func);
20     t1.join();
21     t2.join();
22     std::cout << counter << '\n';
23 }
```

- above code improves upon [earlier example](#) by avoiding possibility of not releasing mutex (e.g., due to exception or forgetting to call `unlock`)

# Example: Avoiding Data Race Using Mutex (IntSet) (scoped\_lock)

```
1  #include <iostream>
2  #include <mutex>
3  #include <thread>
4  #include <unordered_set>
5
6  class IntSet {
7  public:
8      bool contains(int i) const {
9          std::scoped_lock lock(m_);
10         return s_.find(i) != s_.end();
11     }
12     void add(int i) {
13         std::scoped_lock lock(m_);
14         s_.insert(i);
15     }
16 private:
17     std::unordered_set<int> s_;
18     mutable std::mutex m_;
19 };
20
21 IntSet s;
22
23 int main() {
24     std::thread t1([]() {
25         for (int i = 0; i < 10'000; ++i) {s.add(2 * i);}
26     });
27     std::thread t2([]() {
28         for (int i = 0; i < 10'000; ++i) {s.add(2 * i + 1);}
29     });
30     t1.join(); t2.join();
31     std::cout << s.contains(1000) << '\n';
32 }
```

# Acquisition of Multiple Locks

- if multiple locks must be acquired, critical that locks always acquired in same order; otherwise, deadlock can occur
- for example, consider scenario in which two mutexes must be acquired before performing some processing
- if two mutexes not acquired in consistent order, situations like following can arise:
  - 1 thread 1 acquires mutex 1
  - 2 thread 2 acquires mutex 2
  - 3 thread 1 tries to acquire mutex 2 and blocks waiting for thread 2 to release this mutex
  - 4 thread 2 tries to acquire mutex 1 and blocks waiting for thread 1 to release this mutex, resulting in *deadlock*
- if, in above example, two mutexes instead always acquired in same order, this type of deadlock cannot occur

# Example: Acquiring Two Locks for Swap (Incorrect)

```
1  #include <mutex>
2  #include <thread>
3  #include <utility>
4  #include <vector>
5
6  class BigBuf // A Big Buffer
7  {
8  public:
9      static constexpr long size() {return 16 * 1024L * 1024L;}
10     BigBuf() : data_(size()) {}
11     BigBuf& operator=(const BigBuf&) = delete;
12     BigBuf& operator=(BigBuf&&) = delete;
13     void swap(BigBuf& other) {
14         if (this == &other) {return;}
15         // acquiring the two mutexes in this way can result in deadlock
16         std::scoped_lock lock1(m_);
17         std::scoped_lock lock2(other.m_);
18         std::swap(data_, other.data_);
19     }
20     // ...
21 private:
22     std::vector<char> data_;
23     mutable std::mutex m_;
24 };
25
26 BigBuf a;
27 BigBuf b;
28
29 int main() {
30     std::thread t1([]() {
31         for (long i = 0; i < 1'000'000; ++i) a.swap(b);
32     });
33     std::thread t2([]() {
34         for (long i = 0; i < 1'000'000; ++i) b.swap(a);
35     });
36     t1.join(); t2.join();
37 }
```



# Example: Acquiring Two Locks for Swap [scoped\_lock]

```
1  #include <mutex>
2  #include <thread>
3  #include <utility>
4  #include <vector>
5
6  class BigBuf // A Big Buffer
7  {
8  public:
9      static constexpr long size() {return 16 * 1024L * 1024L;}
10     BigBuf() : data_(size()) {}
11     BigBuf& operator=(const BigBuf&) = delete;
12     BigBuf& operator=(BigBuf&&) = delete;
13     void swap(BigBuf& other) {
14         if (this == &other) {return;}
15         std::scoped_lock sl(m_, other.m_);
16         std::swap(data_, other.data_);
17     }
18     // ...
19 private:
20     std::vector<char> data_;
21     mutable std::mutex m_;
22 };
23
24 BigBuf a;
25 BigBuf b;
26
27 int main() {
28     std::thread t1([]() {
29         for (long i = 0; i < 1'000'000; ++i) a.swap(b);
30     });
31     std::thread t2([]() {
32         for (long i = 0; i < 1'000'000; ++i) b.swap(a);
33     });
34     t1.join(); t2.join();
35 }
```

# The `std::unique_lock` Template Class

- `std::unique_lock` is another RAII class for mutexes
- declaration:  

```
template <class T> class unique_lock;
```
- template parameter `T` specifies type of mutex (e.g., `std::mutex`, `std::recursive_mutex`)
- unlike case of `std::scoped_lock`, in case of `unique_lock` do not have to hold mutex over entire lifetime of RAII object
- have choice of whether to acquire mutex upon construction
- also can acquire and release mutex many times throughout lifetime of `unique_lock` object
- upon destruction, if mutex is held, it is released
- since mutex is always guaranteed to be released by destructor, cannot forget to release mutex
- `unique_lock` is used in situations where want to be able to transfer ownership of lock (e.g., return from function) or RAII object needed for mutex but do not want to hold mutex over entire lifetime of RAII object
- *movable* but *not copyable* [\[C++17 33.4.4.3\]](#)

# std::unique\_lock Members

## Member Types

Name	Description
<code>mutex_type</code>	underlying mutex type

## Construction, Destruction, and Assignment

Name	Description
constructor	construct unique lock
destructor	destroy unique lock
<b>operator=</b>	move assign

## Locking Functions

Name	Description
<code>lock</code>	acquire mutex, blocking if not available
<code>try_lock</code>	try to lock mutex without blocking
<code>try_lock_for</code>	try to lock mutex without blocking
<code>try_lock_until</code>	try to lock mutex without blocking
<code>unlock</code>	release mutex

## Observer Functions

Name	Description
<code>owns_lock</code>	tests if lock owns associated mutex
<b><code>operator bool</code></b>	tests if lock owns associated mutex

# Example: Avoiding Data Race Using Mutex (unique\_lock)

```
1  #include <iostream>
2  #include <thread>
3  #include <mutex>
4
5  unsigned long counter = 0;
6  std::mutex m;
7
8  void func() {
9      for (unsigned long i = 0; i < 1'000'000; ++i) {
10         // Create a lock object without acquiring the mutex.
11         std::unique_lock lock(m, std::defer_lock);
12         // ...
13         lock.lock(); // Acquire the mutex.
14         ++counter;
15         lock.unlock(); // Release the mutex.
16         // ...
17         lock.lock(); // Acquire the mutex.
18         ++counter;
19         lock.unlock(); // Release the mutex.
20         // ...
21         // The unique_lock destructor releases the mutex (if held).
22     }
23 }
24
25 int main() {
26     std::thread t1(func);
27     std::thread t2(func);
28     t1.join();
29     t2.join();
30     std::cout << counter << '\n';
31 }
```

# The `std::lock` Template Function

- `std::lock` variadic template function that can acquire multiple locks simultaneously without risk of deadlock (assuming the only locks involved are ones passed to `lock`)
- declaration:

```
template <class T1, class T2, class... TN>
void lock(T1&, T2&, TN& ...);
```
- takes as arguments one or more locks to be acquired

# Example: Acquiring Two Locks for Swap [unique\_lock and lock]

```
1  #include <mutex>
2  #include <thread>
3  #include <utility>
4  #include <vector>
5
6  class BigBuf // A Big Buffer
7  {
8  public:
9      static constexpr long size() {return 16 * 1024L * 1024L;}
10     BigBuf() : data_(size()) {}
11     BigBuf& operator=(const BigBuf&) = delete;
12     BigBuf& operator=(BigBuf&&) = delete;
13     void swap(BigBuf& other) {
14         if (this == &other) {return;}
15         std::unique_lock lock1(m_, std::defer_lock);
16         std::unique_lock lock2(other.m_, std::defer_lock);
17         std::lock(lock1, lock2);
18         std::swap(data_, other.data_);
19     }
20     // ...
21 private:
22     std::vector<char> data_;
23     mutable std::mutex m_;
24 };
25
26 BigBuf a;
27 BigBuf b;
28
29 int main() {
30     std::thread t1([]() {
31         for (long i = 0; i < 1'000'000; ++i) a.swap(b);
32     });
33     std::thread t2([]() {
34         for (long i = 0; i < 1'000'000; ++i) b.swap(a);
35     });
36     t1.join(); t2.join();
37 }
```

# The `std::timed_mutex` Class

- `std::timed_mutex` class provides mutex that allows timeout to be specified when acquiring mutex
- if mutex cannot be acquired in time specified, acquire operation fails (i.e., does not lock mutex) and error returned
- adds `try_lock_for` and `try_lock_until` member functions to try to lock mutex with timeout



## Example: Acquiring Mutex With Timeout (std::timed\_mutex)

```
1  #include <vector>
2  #include <iostream>
3  #include <thread>
4  #include <mutex>
5  #include <chrono>
6
7  std::timed_mutex m;
8
9  void doWork() {
10     for (int i = 0; i < 10000; ++i) {
11         std::unique_lock lock(m, std::defer_lock);
12         int count = 0;
13         while (!lock.try_lock_for(
14             std::chrono::microseconds(1))) {++count;}
15         std::cout << count << '\n';
16     }
17 }
18
19 int main() {
20     std::vector<std::thread> workers;
21     for (int i = 0; i < 16; ++i) {
22         workers.emplace_back(doWork);
23     }
24     for (auto& t : workers) {t.join();}
25 }
```

# Recursive Mutexes

- A **recursive mutex** is a mutex for which a thread may own *multiple* locks *at the same time*.
- After a mutex is first locked by thread A, thread A can acquire additional locks on the mutex (without releasing the lock already held).
- The mutex is not available to other threads until thread A releases all of its locks on the mutex.
- A recursive mutex is typically used when code that locks a mutex must call other code that locks the same mutex (in order to avoid deadlock).
- For example, a function that acquires a mutex and recursively calls itself (resulting in the mutex being relocked) would need to employ a recursive mutex.
- A recursive mutex has *more overhead* than a nonrecursive mutex.
- Code that uses recursive mutexes can often be *more difficult to understand* and therefore *more prone to bugs*.
- Consequently, the use of recursive mutexes should be *avoided if possible*.

# Recursive Mutex Classes

- recursive mutexes provided by classes `std::recursive_mutex` and `std::recursive_timed_mutex`
- `recursive_mutex` class similar to `std::mutex` class except allows relocking
- `recursive_timed_mutex` class similar to `std::timed_mutex` class except allows relocking
- implementation-defined limit to number of levels of locking allowed by recursive mutex

# Shared Mutexes

- A **shared mutex** (also known as a **multiple-reader/single-writer mutex**) is a mutex that allows both *shared and exclusive* access.
- A shared mutex has *two types of locks*: shared and exclusive.
- **Exclusive lock**:
  - *Only one* thread can hold an *exclusive* lock on a mutex.
  - While a thread holds an exclusive lock on a mutex, no other thread can hold any type of lock on the mutex.
- **Shared lock**:
  - *Any number* of threads (within implementation limits) can take a *shared* lock on a mutex.
  - While any thread holds a shared lock on a mutex, no thread may take an exclusive lock on the mutex.
- A shared mutex would typically be used to protect shared data that is seldom updated but cannot be safely updated if any thread is reading it.
- A thread takes a shared lock for reading, thus allowing *multiple readers*.
- A thread takes an exclusive lock for writing, thus allowing *only one writer with no readers*.
- A shared mutex need not be fair in its granting of locks (e.g., readers could starve writers).

# The `std::shared_mutex` Class

- `std::shared_mutex` class provides shared mutex functionality
- *not movable* and *not copyable*
- `lock` member function acquires exclusive ownership of mutex (blocking as necessary)
- `unlock` member function releases exclusive ownership
- `lock_shared` member function acquires shared ownership of mutex (blocking as necessary)
- `unlock_shared` member function releases shared ownership

## Construction, Destruction, and Assignment

Name	Description
constructor	construct mutex
destructor	destroy mutex
<b>operator=</b> [deleted]	not movable or copyable

## Exclusive Locking Functions

Name	Description
lock	acquire exclusive ownership of mutex, blocking if not available
try_lock	try to acquire exclusive ownership of mutex without blocking
unlock	release exclusive ownership of mutex

## Shared Locking Functions

Name	Description
lock_shared	acquire shared ownership of mutex, blocking if not available
try_lock_shared	try to acquire shared ownership of mutex without blocking
unlock_shared	release shared ownership of mutex

## Other Functions

Name	Description
native_handle	get handle for underlying mutex entity

# The `std::shared_lock` Template Class

- `std::shared_lock` is RAII class for shared mutexes
- declaration:  

```
template <class T> class shared_lock;
```
- template parameter T specifies type of mutex (e.g., `std::shared_mutex` or `std::shared_timed_mutex`)
- similar interface as `std::unique_lock` but uses shared locking
- constructor may optionally acquire mutex
- may acquire and release mutex many times throughout lifetime of object
- destructor releases mutex if held
- all operations mapped onto shared locking primitives (e.g., `lock` mapped to `lock_shared`, `unlock` mapped to `unlock_shared`)
- for exclusive locking with shared mutexes, `std::unique_lock` and `std::scoped_lock` (and `std::lock_guard`) can be used



# Example: `std::shared_mutex`

```
1  #include <thread>
2  #include <mutex>
3  #include <iostream>
4  #include <vector>
5  #include <shared_mutex>
6
7  std::mutex coutMutex;
8  int counter = 0;
9  std::shared_mutex counterMutex;
10
11 void writer() {
12     for (int i = 0; i < 10; ++i) {
13         {
14             std::scoped_lock lock(counterMutex);
15             ++counter;
16         }
17         std::this_thread::sleep_for(std::chrono::milliseconds(100));
18     }
19 }
20
21 void reader() {
22     for (int i = 0; i < 100; ++i) {
23         int c;
24         {
25             std::shared_lock lock(counterMutex);
26             c = counter;
27         }
28         {
29             std::scoped_lock lock(coutMutex);
30             std::cout << std::this_thread::get_id() << ' ' << c << '\n';
31         }
32         std::this_thread::sleep_for(std::chrono::milliseconds(10));
33     }
34 }
35
36 int main() {
37     std::vector<std::thread> threads;
38     threads.emplace_back(writer);
39     for (int i = 0; i < 16; ++i) threads.emplace_back(reader);
40     for (auto& t : threads) t.join();
41 }
```

# The `std::shared_timed_mutex` Class

- `std::shared_timed_mutex` class provides shared mutex
- `shared_timed_mutex` interface similar to that of `shared_mutex` but allows timeout for acquiring mutex
- adds `try_lock_for` and `try_lock_until` member functions to try to acquire exclusive ownership of mutex with timeout
- adds `try_lock_shared_for` and `try_lock_shared_until` member functions to try to acquire shared ownership of mutex with timeout

# Example: std::shared\_timed\_mutex

```
1  #include <thread>
2  #include <mutex>
3  #include <iostream>
4  #include <vector>
5  #include <shared_mutex>
6
7  std::mutex coutMutex;
8  int counter = 0;
9  std::shared_timed_mutex counterMutex;
10
11 void writer() {
12     for (int i = 0; i < 10; ++i) {
13         {
14             std::scoped_lock lock(counterMutex);
15             ++counter;
16         }
17         std::this_thread::sleep_for(std::chrono::milliseconds(100));
18     }
19 }
20
21 void reader() {
22     for (int i = 0; i < 100; ++i) {
23         int c;
24         {
25             std::shared_lock lock(counterMutex);
26             c = counter;
27         }
28         {
29             std::scoped_lock lock(coutMutex);
30             std::cout << std::this_thread::get_id() << ' ' << c << '\n';
31         }
32         std::this_thread::sleep_for(std::chrono::milliseconds(10));
33     }
34 }
35
36 int main() {
37     std::vector<std::thread> threads;
38     threads.emplace_back(writer);
39     for (int i = 0; i < 16; ++i) threads.emplace_back(reader);
40     for (auto& t : threads) t.join();
41 }
```

## `std::once_flag` and `std::call_once`

- sometimes may want to perform action only once in code executed in multiple threads
- can be achieved through use of `std::once_flag` type in conjunction with `std::call_once` template function
- `std::once_flag` class represents flag used to track if action performed
- declaration of `std::call_once`:

```
template <class Callable, class... Args>
void call_once(std::once_flag& flag, Callable&& f,
               Args&&... args);
```

- `std::call_once` invokes `f` only once based on value of `flag` object
- first invocation of `f` is guaranteed to complete before any threads return from `call_once`
- useful for one-time initialization of dynamically generated objects

# Example: One-Time Action

```
1  #include <iostream>
2  #include <vector>
3  #include <thread>
4  #include <mutex>
5
6  std::once_flag flag;
7
8  void worker(int id) {
9      std::call_once(flag, [id]() {
10         // This code will be invoked only once.
11         std::cout << "first: " << id << '\n';
12     });
13 }
14
15 int main() {
16     std::vector<std::thread> threads;
17     for (int i = 0; i < 16; ++i) {
18         threads.emplace_back(worker, i);
19     }
20     for (auto& t : threads) {
21         t.join();
22     }
23 }
```

# Example: One-Time Initialization

```
1  #include <vector>
2  #include <thread>
3  #include <mutex>
4  #include <cassert>
5  #include <memory>
6
7  std::unique_ptr<int> value;
8  std::once_flag initFlag;
9
10 void initValue() {value = std::make_unique<int>(42);}
11
12 const int& getValue() {
13     std::call_once(initFlag, initValue);
14     return *value.get();
15 }
16
17 void doWork() {
18     const int& v = getValue();
19     assert(v == 42);
20     // ...
21 }
22
23 int main() {
24     std::vector<std::thread> threads;
25     for (int i = 0; i < 4; ++i)
26         threads.emplace_back(doWork);
27     for (auto& t : threads) {t.join();}
28 }
```

# Static Local Variable Initialization and Thread Safety

- initialization of static local object is thread safe [\[C++17 9.7/4\]](#)
- object is initialized first time control passes through its declaration
- object deemed initialized upon completion of initialization
- if control enters declaration concurrently while object being initialized, concurrent execution waits for completion of initialization
- code like following is thread safe:

```
const std::string& meaningOfLife() {  
    static const std::string x("42");  
    return x;  
}
```

## Section 3.9.7

### **Condition Variables**



# Condition Variables

- In concurrent programs, the need often arises for a thread to *wait until a particular event occurs* (e.g., I/O has completed or data is available).
- Having a thread *repeatedly check* for the occurrence of an event can be *inefficient* (i.e., can waste processor resources).
- It is often better to have the thread block and then only resume execution after the event of interest has occurred.
- A **condition variable** is a synchronization primitive that allows threads to *wait (by blocking)* until a particular condition occurs.
- A condition variable corresponds to some event of interest.
- A thread that wants to wait for an event, performs a *wait operation* on the condition variable.
- A thread that wants to notify one or more waiting threads of an event performs a *signal operation* on the condition variable.
- When a signalled thread resumes, however, the signalled condition is not guaranteed to be true (and must be rechecked), since, for example, another thread may have caused the condition to change or a spurious awakening may have occurred.

# The `std::condition_variable` Class

- `std::condition_variable` class provides condition variable
- *not movable* and *not copyable*
- `wait`, `wait_for`, and `wait_until` member functions used to wait for condition
- `notify_one` and `notify_all` used to signal waiting thread(s) of condition
- must re-check condition when awaking from wait since:
  - spurious awakenings are permitted
  - between time thread is signalled and time it awakens and locks mutex, another thread could cause condition to change
- concurrent invocation is allowed for `notify_one`, `notify_all`, `wait`, `wait_for`, `wait_until` [\[C++17 33.5/2\]](#)
- each of `wait`, `wait_for`, and `wait_until` atomically releases mutex and blocks [\[C++17 33.5/3\]](#) why atomic? (diagram)
- `notify_one` and `notify_all` are atomic [\[C++17 33.5/3\]](#)

## Member Types

Name	Description
<code>native_handle_type</code>	system-dependent handle type for underlying condition variable entity

## Construction, Destruction, and Assignment

Name	Description
<code>constructor</code>	construct object
<code>destructor</code>	destroy object
<b><code>operator=</code></b> [deleted]	not movable or copyable

## Notification and Waiting Member Functions

Name	Description
<code>notify_one</code>	notify one waiting thread
<code>notify_all</code>	notify all waiting threads
<code>wait</code>	blocks current thread until notified
<code>wait_for</code>	blocks current thread until notified or specified duration passed
<code>wait_until</code>	blocks current thread until notified or specified time point reached

## Native Handle Member Functions

Name	Description
<code>native_handle</code>	get native handle associated with condition variable

# Example: Condition Variable (IntStack)

```
1  #include <iostream>
2  #include <vector>
3  #include <thread>
4  #include <mutex>
5  #include <condition_variable>
6
7  class IntStack {
8  public:
9      IntStack() {}
10     IntStack(const IntStack&) = delete;
11     IntStack& operator=(const IntStack&) = delete;
12     int pop() {
13         std::unique_lock lock(m_);
14         c_.wait(lock, [this]() {return !v_.empty();});
15         int x = v_.back();
16         v_.pop_back();
17         return x;
18     }
19     void push(int x) {
20         std::scoped_lock lock(m_);
21         v_.push_back(x);
22         c_.notify_one();
23     }
24 private:
25     std::vector<int> v_;
26     mutable std::mutex m_;
27     mutable std::condition_variable c_; // not empty
28 };
29
30 constexpr int numIters = 1000;
31 IntStack s;
32
33 int main() {
34     std::thread t1([]() {
35         for (int i = 0; i < numIters; ++i) s.push(2 * i + 1);
36     });
37     std::thread t2([]() {
38         for (int i = 0; i < numIters; ++i) std::cout << s.pop() << '\n';
39     });
40     t1.join(); t2.join();
41 }
```

- **latch**: basic one-time synchronization mechanism that allows threads to block until particular event occurs certain number of times
- latch maintains count as state
- count is initialized to some nonzero value when latch created
- can decrement count
- can block until count reaches zero
- latch can only be used once (i.e., count cannot be reset after it reaches zero)

# Latch Example: User Code

```
1  #include <functional>
2  #include <thread>
3  #include <vector>
4  #include "latch_1.hpp"
5
6  void worker(latch& ready) {
7      // ... (perform very slow initialization)
8      // wait for all threads to complete initialization
9      ready.count_down_and_wait();
10     // ... (perform real work)
11 }
12
13 int main() {
14     constexpr int num_workers = 32;
15     std::vector<std::thread> workers;
16     latch ready(num_workers);
17     for (int i = 0; i < num_workers; ++i) {
18         workers.emplace_back(worker, std::ref(ready));
19     }
20     for (auto& i : workers) {i.join();}
21 }
```

# Latch Example: latch\_1.hpp

```
1  #include <condition_variable>
2  #include <cstdint>
3  #include <mutex>
4
5  class latch {
6  public:
7      explicit latch(std::size_t count) : count_(count) {}
8      // ... (not movable, not copyable, destructor)
9      void count_down() {
10         std::scoped_lock<std::mutex> lock(mutex_);
11         if (!--count_) {ready_.notify_all();}
12     }
13     void wait() {
14         std::unique_lock<std::mutex> lock(mutex_);
15         ready_.wait(lock, [this]() {return !count_;});
16     }
17     bool try_wait() {
18         std::scoped_lock<std::mutex> lock(mutex_);
19         return !count_;
20     }
21     void count_down_and_wait() {
22         std::unique_lock<std::mutex> lock(mutex_);
23         if (!--count_) {ready_.notify_all();}
24         else {ready_.wait(lock, [this]() {return !count_;});}
25     }
26 private:
27     mutable std::mutex mutex_;
28     mutable std::condition_variable ready_; // count reached zero
29     std::size_t count_;
30 };
```



# The `std::condition_variable_any` Class

- with `std::condition_variable` class, `std::unique_lock<std::mutex>` class must be used for wait operation
- `std::condition_variable_any` class allows any mutex type (meeting certain basic requirements) to be used
- interface of `std::condition_variable_any` class similar to that of `std::condition_variable` class
- prefer `condition_variable` to `condition_variable_any` since former may be more efficient

# Thread Pools

- **thread pool** is collection of threads which stand ready to be given work (i.e., tasks to execute)
- number of threads in thread pool may be fixed or dynamic, depending on particular thread-pool variant
- each thread in thread pool can be either idle or executing task
- typically, tasks given to thread pool are placed on queue and then assigned to threads as they become available
- after done executing task, thread returns to idle state, waiting to be assigned another task
- thread pool allows overhead of many thread creation operations to be avoided
- by using thread pool can also more easily control total number of threads and number of running threads
- thread pool preferred over creating new threads for each task when there is large number of short-running tasks (rather than small number of long-running ones)

# Simple Thread Pool Interface Example

thread\_pool\_1\_1.hpp

```
1  #include <functional>
2  #include <cstddef>
3
4  class thread_pool {
5  public:
6      using size_type = std::size_t;
7      thread_pool(size_type num_threads);
8      thread_pool(const thread_pool&) = delete;
9      thread_pool& operator=(const thread_pool&) = delete;
10     ~thread_pool();
11     size_type num_threads() const;
12     void schedule(std::function<void()>&& task);
13     void shutdown();
14 private:
15     // ...
16 };
```

app.cpp

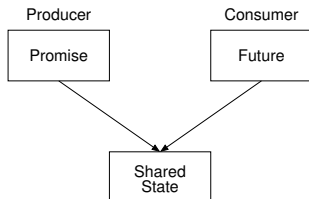
```
1  #include <iostream>
2  #include "thread_pool_1_1.hpp"
3
4  int main() {
5      thread_pool tp(8);
6      for (int i = 0; i < 10'000; ++i) {
7          tp.schedule([]() {std::cout << "Hello, World\n";});
8      }
9      tp.shutdown();
10 }
```

## Section 3.9.8

# **Promises and Futures**

# Promises and Futures

- promise and future together form *one-time* communication channel for passing result (i.e., value or exception) of computation from one thread to same or another thread
- **promise**: object associated with promised result (i.e., value or exception) to be produced
- **future**: object through which promised result later made available
- **shared state**: holds promised result for access through future object (shared by promise object and corresponding future object)
- producer of result uses promise object to store result in shared state
- consumer uses future object (corresponding to promise) to retrieve result from shared state



# Promises and Futures (Continued)

- promises and futures useful in both single-threaded and multithreaded programs
- in single-threaded programs, might be used to propagate exception to another part of program
- in multithreaded program, often need arises to do some computation asynchronously and then later get result when ready
- requires synchronization between threads producing and consuming result
- thread consuming result must *wait until result is available*
- must *avoid data races* when accessing result shared between threads
- this type of synchronization can be accomplished via promise and future

# The `std::promise` Template Class

- `std::promise` provides access to promise-future shared state for writing result
- declaration:  

```
template <class T> class promise;
```
- T is type of result associated with promise (which can be **void**)
- movable but *not copyable*
- `set_value` member function sets result to particular value
- `set_exception` member function sets result to exception
- can set result *only once*
- `get_future` member function retrieves future associated with promise
- `get_future` may be called *only once*
- if `promise` object is destroyed before its associated result is set, `std::future_error` exception will be thrown if attempt made to retrieve result from corresponding `future` object

## Construction, Destruction, and Assignment

Name	Description
constructor	construct object
destructor	destroy object
<b>operator=</b>	move assignment



## Other Functions

Name	Description
swap	swap two promise objects
get_future	get future associated with promised result
set_value	set result to specified value
set_value_at_thread_exit	set result to specified value while delivering notification only at thread exit
set_exception	set result to specified exception
set_exception_at_thread_exit	set result to specified exception while delivering notification only at thread exit

# The `std::future` Template Class

- `std::future` provides access to promise-future shared state for reading result
- declaration:  

```
template <class T> class future;
```
- T is type of result associated with future (which can be **void**)
- movable but *not copyable*
- `get` member function retrieves result, blocking if result not yet available
- `get` may be called *only once*
- `wait` member function waits for result to become available without actually retrieving result

## Construction, Destruction, and Assignment

Name	Description
constructor	construct object
destructor	destroy object
<b>operator=</b>	move assignment

## Other Functions

Name	Description
share	transfer shared state to <code>shared_future</code> object
get	get result
valid	check if <code>future</code> object refers to shared state
wait	wait for result to become available
wait_for	wait for result to become available or time duration to expire
wait_until	wait for result to become available or time point to be reached

## Example: Promises and Futures (Without `std::async`)

```
1  #include <future>
2  #include <thread>
3  #include <iostream>
4  #include <utility>
5
6  double computeValue() {
7      return 42.0;
8  }
9
10 void produce(std::promise<double> p) {
11     // write result to promise
12     p.set_value(computeValue());
13 }
14
15 int main() {
16     std::promise<double> p;
17     auto f = p.get_future(); // save future before move
18     std::thread producer(produce, std::move(p));
19     std::cout << f.get() << '\n';
20     producer.join();
21 }
```

# The `std::shared_future` Template Class

- `std::shared_future` similar to `future` except object can be copied
- `shared_future` object can be obtained by using `share` member function of `future` class to transfer contents of `future` object into `shared_future` object
- `shared_future` is *copyable* (unlike `future`)
- allows multiple threads to wait for same result (associated with `shared_future` object)
- `get` member can be called multiple times

## Example: `std::shared_future`

```
1  #include <iostream>
2  #include <vector>
3  #include <thread>
4  #include <future>
5
6  void consume(std::shared_future<int> f) {
7      std::cout << f.get() << '\n';
8  }
9
10 int main() {
11     std::promise<int> p;
12     std::shared_future f = p.get_future().share();
13     std::vector<std::thread> consumers;
14     for (int i = 0; i < 16; ++i) {
15         consumers.emplace_back(consume, f);
16     }
17     p.set_value(42);
18     for (auto& i : consumers) {
19         i.join();
20     }
21 }
```

# The `std::async` Template Function

- `std::async` template function used to launch callable entity (e.g., function or functor) asynchronously
- declaration (uses default launch policy):

```
template <class Func, class... Args>
future<typename result_of<typename decay<Func>::type(
    typename decay<Args>::type...>::type>
    async(Func&& f, Args&&... args);
```

- declaration (with launch policy parameter):

```
template <class Func, class... Args>
future<typename result_of<typename decay<Func>::type(
    typename decay<Args>::type...>::type>
    async(launch policy, Func&& f, Args&&... args);
```

- numerous launch policies supported via bitmask `std::launch`
- if `async` bit set, execute on new thread
- if `deferred` bit set, execute on calling thread when result needed
- if multiple bits set, implementation free to choose between them
- in asynchronous execution case, essentially creates promise to hold result and returns associated future; launches thread to execute function/functor and sets promise when function/functor returns

# The `std::async` Template Function (Continued)

- `future` (i.e., `future` and `shared_future`) objects created by `async` function have slightly different behavior than `future` objects created in other ways
- in case of `future` object created by `async` function: if `future` object is *last* `future` object referencing its shared state, destructor for `future` object will *block* until result associated with `future` object becomes ready



## Example: Promises and Futures (With `std::async`)

```
1  #include <future>
2  #include <iostream>
3
4  double computeValue() {
5      return 42.0;
6  }
7
8  int main() {
9      // invoke computeValue function asynchronously in
10     // separate thread
11     auto f = std::async(std::launch::async, computeValue);
12     std::cout << f.get() << '\n';
13 }
```

# Example: Futures and Exceptions

```
1  #include <iostream>
2  #include <vector>
3  #include <cmath>
4  #include <future>
5  #include <stdexcept>
6
7  double squareRoot(double x) {
8      if (x < 0.0) {
9          throw std::domain_error(
10             "square root of negative number");
11      }
12      return std::sqrt(x);
13  }
14
15  int main() {
16      std::vector values{1.0, 2.0, -1.0};
17      std::vector<std::future<double>> results;
18      for (auto x : values) {
19          results.push_back(std::async(squareRoot, x));
20      }
21      for (auto& x : results) {
22          try {
23              std::cout << x.get() << '\n';
24          } catch (const std::domain_error&) {
25              std::cout << "error\n";
26          }
27      }
28  }
```

# The `std::packaged_task` Template Class

- `std::packaged_task` template class provides wrapper for callable entity (e.g., function or functor) that makes return value available via future

- declaration:

```
template <class R, class... Args>  
    class packaged_task<R(Args...)>;
```

- template parameters `R` and `Args` specify return type and arguments for callable entity
- similar to `std::function` except return value of wrapped function made available via future
- packaged task often used as thread function
- movable but *not copyable*
- `get_future` member retrieves future associated with packaged task
- `get_future` can be called *only once*

# std::packaged\_task Members

## Construction, Destruction, and Assignment

Name	Description
constructor	construct object
destructor	destroy object
<b>operator=</b>	move assignment

## Other Functions

Name	Description
valid	check if task object currently associated with shared state
swap	swap two task objects
get_future	get future associated with promised result
<b>operator()</b>	invoke function
make_ready_at_thread_exit	invoke function ensuring result ready only once current thread exits
reset	reset shared state, abandoning any previously stored result

# Example: Packaged Task

```
1  #include <iostream>
2  #include <thread>
3  #include <future>
4  #include <utility>
5  #include <chrono>
6
7  int getMeaningOfLife() {
8      // Let the suspense build before providing the answer.
9      std::this_thread::sleep_for(std::chrono::milliseconds(
10         1000));
11     // Return the answer.
12     return 42;
13 }
14
15 int main() {
16     std::packaged_task<int()> pt(getMeaningOfLife);
17     // Save the future.
18     auto f = pt.get_future();
19     // Start a thread running the task and detach the thread.
20     std::thread t(std::move(pt));
21     t.detach();
22     // Get the result via the future.
23     int result = f.get();
24     std::cout << "The meaning of life is " << result << '\n';
25 }
```

## Example: Packaged Task With Arguments

```
1  #include <iostream>
2  #include <cmath>
3  #include <thread>
4  #include <future>
5
6  double power(double x, double y) {
7      return std::pow(x, y);
8  }
9
10 int main() {
11     // invoke task in main thread
12     std::packaged_task<double(double, double)> task(power);
13     task(0.5, 2.0);
14     std::cout << task.get_future().get() << '\n';
15     // reset shared state
16     task.reset();
17     // invoke task in new thread
18     auto f = task.get_future();
19     std::thread t(std::move(task), 2.0, 0.5);
20     t.detach();
21     std::cout << f.get() << '\n';
22 }
```

## Section 3.9.9

### **Atomics**

- To avoid data races when sharing data between threads, it is often necessary to employ *synchronization* (e.g., by using mutexes).
- Atomic types are another mechanism for providing synchronized access to data.
- An operation that is indivisible is said to be **atomic** (i.e., no parts of any other operations can interleave with any part of an atomic operation).
- Most processors support atomic memory operations via special machine instructions.
- Atomic memory operations cannot result in torn reads or torn writes.
- The standard library offers the following types in order to provide support for atomic memory operations:
  - `std::atomic_flag`
  - `std::atomic`
- These types provide a uniform interface for accessing the atomic memory operations of the underlying hardware.



- An atomic type provides guarantees regarding:
  - 1 atomicity; and
  - 2 ordering.
- An ordering guarantee specifies the manner in which memory operations can become visible to threads.
- Several memory ordering schemes are supported by atomic types.
- The default memory order is sequentially consistent (`std::memory_order_seq_cst`).
- Initially, only this default will be considered.

# The `std::atomic_flag` Class

- `std::atomic_flag` provides flag with basic atomic operations
- flag can be in one of two states: set (i.e., true) or clear (i.e., false)
- two operations for flag:
  - **test and set**: set state to true and query previous state
  - **clear**: set state to false
- default constructor initializes flag to *unspecified* state
- *not movable* and *not copyable*
- implementation-defined macro `ATOMIC_FLAG_INIT` can be used to set flag to clear state in (static or automatic) initialization using statement of the form “`std::atomic_flag f = ATOMIC_FLAG_INIT;`” [C++17 32.8/4]
- guaranteed to be *lock free*
- intended to be used as building block for higher-level synchronization primitives, such as spinlock mutex

## Member Functions

Member Name	Description
<code>constructor</code>	constructs object
<code>clear</code>	atomically sets flag to false
<code>test_and_set</code>	atomically sets flag to true and obtains its previous value

# Example: Suboptimal Spinlock Mutex

```
1  #include <iostream>
2  #include <thread>
3  #include <atomic>
4  #include <mutex>
5
6  class SpinLockMutex {
7  public:
8      SpinLockMutex() {f_.clear();}
9      void lock() {while (f_.test_and_set()) {}}
10     void unlock() {f_.clear();}
11 private:
12     std::atomic_flag f_; // true if thread holds mutex
13 };
14
15 SpinLockMutex m;
16 unsigned long long counter = 0;
17
18 void doWork() {
19     for (unsigned long long i = 0; i < 1000000ULL; ++i)
20         {std::scoped_lock lock(m); ++counter;}
21 }
22
23 int main() {
24     std::thread t1(doWork), t2(doWork);
25     t1.join(); t2.join();
26     std::cout << counter << '\n';
27 }
```

- default memory order is suboptimal (and will be revisited later)

# Example: One-Time Wait

```
1  #include <iostream>
2  #include <atomic>
3  #include <thread>
4  #include <chrono>
5
6  // notReady flag initially not set
7  std::atomic_flag notReady = ATOMIC_FLAG_INIT;
8  int result = 0;
9
10 int main() {
11     notReady.test_and_set(); // indicate result not ready
12     std::thread producer([]){
13         std::this_thread::sleep_for(std::chrono::seconds(1));
14         result = -42;
15         notReady.clear(); // indicate result ready
16     };
17     std::thread consumer([]){
18         // loop until result ready
19         while (notReady.test_and_set()) {}
20         std::cout << result << '\n';
21     };
22     producer.join();
23     consumer.join();
24 }
```

- This is *not* a particularly good use of `atomic_flag`.

# The `std::atomic` Template Class

- `std::atomic` class provides types with atomic operations
- declaration:  

```
template <class T> struct atomic;
```
- provides object of type `T` with atomic operations
- has partial specializations for integral types and pointer types
- full specializations for all fundamental types
- in order to use class type for `T`, `T` must be trivially copyable [\[C++17 32.6/1\]](#) and bitwise equality comparable
- not required to be lock free
- on most popular platforms `atomic` is lock free when `T` is built-in type
- *not move constructible* and *not copy constructible*
- assignable but assignment operator returns value not reference
- most operations have memory order argument
- default memory order is SC (`std::memory_order_seq_cst`)

## Basic

Member Name	Description
constructor	constructs object
<b>operator</b> =	atomically store value into atomic object
is_lock_free	check if atomic object is lock free
store	atomically replaces value of atomic object with given value
load	atomically reads value of atomic object
<b>operator</b> T	obtain result of load
exchange	atomically replaces value of atomic object with given value and obtain value of previous value
compare_exchange_weak	similar to <code>exchange_strong</code> but may fail spuriously
compare_exchange_strong	atomically compare value of atomic object to given value and perform <code>exchange</code> if equal or <code>load</code> otherwise

## Fetch

Member Name	Description
fetch_add	atomically adds given value to value stored in atomic object and obtains value held previously
fetch_sub	atomically subtracts given value from value stored in atomic object and obtains value held previously
fetch_and	atomically replaces value of atomic object with bitwise AND of atomic object's value and given value, and obtains value held previously
fetch_or	atomically replaces value of atomic object with bitwise OR of atomic object's value and given value, and obtains value held previously
fetch_xor	atomically replaces value of atomic object with bitwise XOR of atomic object's value and given value, and obtains value held previously



## Increment and Decrement

Member Name	Description
<b>operator++</b>	atomically increment the value of atomic object by one and obtain value after incrementing
<b>operator++(int)</b>	atomically increment the value of atomic object by one and obtain value before incrementing
<b>operator--</b>	atomically decrement the value of atomic object by one and obtain value after decrementing
<b>operator--(int)</b>	atomically decrement the value of atomic object by one and obtain value after decrementing

## Compound Assignment

Member Name	Description
<b>operator+=</b>	atomically adds given value to value stored in atomic object
<b>operator-=</b>	atomically subtracts given value from value stored in atomic object
<b>operator&amp;=</b>	atomically performs bitwise AND of given value with value stored in atomic object
<b>operator =</b>	atomically performs bitwise OR of given value with value stored in atomic object
<b>operator^=</b>	atomically performs bitwise XOR of given value with value stored in atomic object

## Constants

Member Name	Description
<code>is_always_lock_free</code>	indicates if type always lock free

# Example: Atomic Counter

```
1  #include <iostream>
2  #include <vector>
3  #include <thread>
4  #include <atomic>
5
6  class AtomicCounter {
7  public:
8      AtomicCounter() : c_(0) {}
9      int operator++() {return ++c_;}
10     int get() const {return c_.load();}
11 private:
12     std::atomic<int> c_;
13 };
14
15 AtomicCounter counter;
16
17 void doWork() {
18     for (int i = 0; i < 10000; ++i) {++counter;}
19 }
20
21 int main() {
22     std::vector<std::thread> v;
23     for (int i = 0; i < 10; ++i)
24         {v.emplace_back(doWork);}
25     for (auto& t : v) {t.join();}
26     std::cout << counter.get() << '\n';
27 }
```

# Example: Atomic Increment With Compare and Swap

```
1  #include <atomic>
2
3  template <class T>
4  void atomicIncrement(std::atomic<T>& x) {
5      T curValue = x;
6      while (!x.compare_exchange_weak(curValue,
7      curValue + 1)) {}
8  }
```

# Example: Counting Contest

```
1  #include <iostream>
2  #include <vector>
3  #include <atomic>
4  #include <thread>
5
6  constexpr int numThreads = 10;
7  std::atomic ready(false);
8  std::atomic done(false);
9  std::atomic startCount(0);
10
11 void doCounting(int id) {
12     ++startCount;
13     while (!ready) {}
14     for (volatile int i = 0; i < 20000; i++) {}
15     bool expected = false;
16     if (done.compare_exchange_strong(expected, true))
17         {std::cout << "winner: " << id << '\n';}
18 }
19
20 int main() {
21     std::vector<std::thread> threads;
22     for (int i = 0; i < numThreads; ++i)
23         {threads.emplace_back(doCounting, i);}
24     while (startCount != numThreads) {}
25     ready = true;
26     for (auto& t : threads) {t.join();}
27 }
```

# An Obligatory Note on **volatile**

- **volatile** qualifier not useful for multithreaded programming
- **volatile** qualifier makes *no guarantee of atomicity*
- can create object of **volatile**-qualified type whose size is sufficiently large that no current processor can access object atomically
- some platforms may happen to guarantee memory operations on (suitably-aligned) **int** object to be atomic, but in such cases this is normally true *even without volatile qualifier*
- **volatile** qualifier *does not adequately address issue of memory consistency*
- **volatile** qualifier does not imply use of memory barriers or other mechanisms needed for memory consistency
- optimizer and hardware might reorder operations (on non-**volatile** objects) across operations on **volatile** objects

## Section 3.9.10

# **Atomics and the Memory Model**

# Semantics of Multithreaded Programs

- To be able to reason about the behavior of a program, we must know:
  - the *order* in which the operations of the program are performed; and
  - when the effects of each operation become *visible* to other operations in the program, which may be performed in different threads.
- In a single-threaded program, the ordering of operations and when the effects of operations become visible is quite intuitive.
- In a multi-threaded program, this matter becomes *considerably more complicated*.
- In what follows, we examine the above matter more closely (which essentially relates to the memory model).



# Happens-Before Relationships

- For two operations  $A$  and  $B$  performed in the *same or different* threads,  $A$  is said to **happen before**  $B$  if the effects of  $A$  become *visible* to the thread performing  $B$  before  $B$  is performed.
- The happens-before relationship is a *much stronger condition* than “happens earlier in time”.
- If operation  $A$  happens earlier in time than operation  $B$ , this does not imply that the effects of  $A$  must be *visible* to the thread performing  $B$  before  $B$  is performed, due to the effects of caches, store buffers, and so on, which *delay* the visibility of results.
- Happening earlier in time is only a necessary but not sufficient condition for a happens-before relationship to exist.
- Happens-before relationships are *not always transitive*.
- In the absence of something known as a dependency-ordered-before relationship (to be discussed later), which arise relatively less frequently, happens-before relationships are *transitive* (i.e., if  $A$  happens before  $B$  and  $B$  happens before  $C$  then  $A$  happens before  $C$ ).

# “Earlier In Time” Versus Happens Before

- Consider the multithreaded program (with two threads) shown below, where  $x$  and  $y$  are integer variables, *initially zero*.

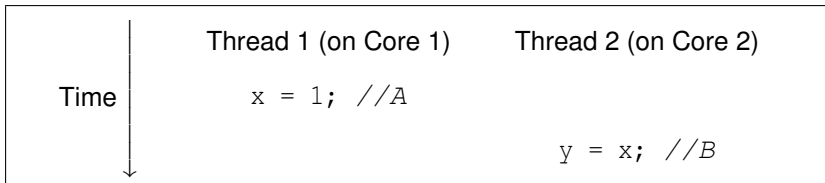
## Thread 1 Code

```
x = 1; // A
```

## Thread 2 Code

```
y = x; // B
```

- Suppose that the run-time platform is such that memory operations on  $x$  are *atomic* so the program is data-race free.
- Consider what happens when the program executes with the particular timing shown below, where *operation A occurs earlier in time than operation B*.



- The value read for  $x$  in operation  $B$  will not necessarily be 1, since the result of  $A$  *may not yet be visible* to thread 2 (e.g., due to caching).

# Sequenced-Before Relationships

- Given two operations  $A$  and  $B$  performed in the *same* thread, the operation  $A$  is **sequenced before**  $B$  if  $A$  precedes  $B$  in program order (i.e., source-code order).
- Sequenced-before relationships are *transitive* (i.e., if  $A$  is sequenced before  $B$ , and  $B$  is sequenced before  $C$ , then  $A$  is sequenced before  $C$ ).
- Example: In the code below, statement  $A$  is **sequenced before** statement  $B$ ;  $B$  is **sequenced before** statement  $C$ ; and, by transitivity,  $A$  is **sequenced before**  $C$ .

```
x = 1;    // A
y = 2;    // B
z = x + 1; // C
```

- Example:
  - Consider the line of code below, which performs (in order) the following operations: 1) multiplication, 2) addition, and 3) assignment.  
 $y = a * x + b;$  //  $(y = ((a * x) + b));$
  - Multiplication is **sequenced before** addition.
  - Addition is **sequenced before** assignment.
  - Thus, by transitivity, multiplication is **sequenced before** assignment.

## Sequenced-Before Relationships (Continued)

- For two operations  $A$  and  $B$  in the *same* thread, if  $A$  is *sequenced before*  $B$  then  $A$  *happens before*  $B$ .
- In other words, program order establishes happens-before relationships for operations *within a single thread*.
- A sequenced-before relationship is essentially an *intra-thread happens-before* relationship. (Note that “intra” means “within”.)
- Example: In the code below, statement  $A$  is sequenced before statement  $B$ . Therefore,  $A$  happens before  $B$ . Similarly,  $B$  happens before statement  $C$ , and  $A$  happens before  $C$ .

```
x = 1;    // A
y = 2;    // B
z = x + 1; // C
```

# Inter-Thread Happens-Before Relationships

- Establishing whether a happens-before relationship exists between operations in different threads is somewhat more complicated than the same-thread case.
- Inter-thread happens-before relationships establish happens-before relationships for operations in *different* threads.
- For two operations  $A$  and  $B$  in *different* threads, if  $A$  **inter-thread happens before**  $B$  then  $A$  happens before  $B$ .
- Inter-thread happens-before relationships are *transitive* (i.e., if  $A$  inter-thread happens before  $B$  and  $B$  inter-thread happens before  $C$  then  $A$  inter-thread happens before  $C$ ).
- Some form of *synchronization* is required to establish an inter-thread happens-before relationship.
- The various forms that this synchronization may take will be introduced on later slides.

# Summary of Happens-Before Relationships

- For two operations  $A$  and  $B$  in either the *same or different* threads,  $A$  happens before  $B$  if:
  - 1  $A$  and  $B$  are in the *same* thread and  $A$  is sequenced before (i.e., intra-thread happens before)  $B$ ; or
  - 2  $A$  and  $B$  are in *different* threads and  $A$  inter-thread happens before  $B$ .
- In other words,  $A$  happens before  $B$  if  $A$  either intra-thread happens before or inter-thread happens before  $B$ .
- Intra-thread happens-before (i.e., sequenced-before) relationships are *transitive*.
- Inter-thread happens-before relationships are *transitive*.
- Happens-before relationships are *mostly but not always transitive*.
- A happens-before relationship is important because it tells us if the result of one operation *can be seen* by a thread performing another operation.

# Synchronizes-With Relationships

- A variety of relationships can imply an inter-thread happens-before relationship, with one being the synchronizes-with relationship.
- For two operations *A* and *B* in *different* threads, if *A* **synchronizes with** *B* then *A* *inter-thread happens before* *B*.
- Example:

- Consider the two-threaded program shown below, with the shared variable *x* of type **int**, where *x* is initially zero.

## Thread 1 Code

```
1  x = 1;  
2  // A (call of foo)  
3  foo();
```

## Thread 2 Code

```
1  bar();  
2  // B (return from bar)  
3  assert(x == 1);
```

- Suppose that the `foo` and `bar` functions are such that:
  - 1 any calls to `bar` will block until `foo` is called for the first time; and
  - 2 if a call to `foo` is made prior to a call to `bar` returning, then the call to `foo` synchronizes with the return of the call to `bar`.
- In this example, the call to `foo` is always made prior to `bar` returning.
- Consequently, *A* synchronizes with *B*, which implies that *A* must inter-thread happen before *B*, which in turn implies that *A* *happens before* *B*.
- Therefore, the assertion in thread 2 *can never fail*.

# Examples of Synchronizes-With Relationships

- **Thread creation.** The completion of the constructor for a `thread` object  $T$  synchronizes with the start of the invocation of the thread function for  $T$ .  
[\[C++17 33.3.2.2/6\]](#)
- **Thread join.** The completion of the execution of a thread function for a `thread` object  $T$  synchronizes with (the return of) a `join` operation on  $T$ .  
[\[C++17 33.3.2.5/4\]](#)
- **Mutex unlock/lock.** All prior `unlock` operations on a mutex  $M$  synchronize with (the return of) a `lock` operation on  $M$ . [\[C++17 33.4.3.2/11\]](#)
- **Atomic.** A suitably tagged atomic write operation  $W$  on a variable  $x$  synchronizes with a suitably tagged atomic read operation on  $x$  that reads the value stored by  $W$  (where the meaning of “suitably tagged” will be discussed later).



# Synchronizes-With Relationship: Thread Create and Join

```
1  #include <thread>
2  #include <cassert>
3
4  int x = 0;
5
6  void doWork() {
7      // A1 (start of thread execution)
8      assert(x == 1); // OK: M1 synchronizes with A1
9      x = 2;
10     // A2 (end of thread execution)
11 }
12
13 int main() {
14     x = 1;
15     std::thread t(doWork); // M1 (completion of constructor)
16     t.join(); // M2 (return from join)
17     assert(x == 2); // OK: A2 synchronizes with M2
18 }
```

- since construction of thread (M1) synchronizes with start of thread function execution (A1), M1 happens before A1 implying that assertion in doWork cannot fail
- since completion of execution of thread function (A2) synchronizes with join operation (M2), A2 happens before M2 implying that assertion in main cannot fail

# Synchronizes-With Relationship: Mutex Lock/Unlock

## Shared Data

```
std::mutex m;  
int x = 0;  
int y = 0;
```

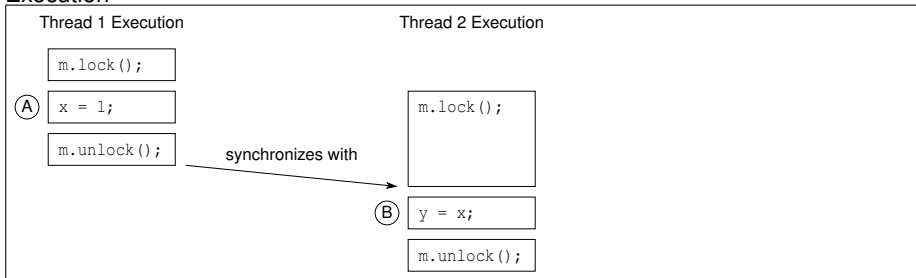
## Thread 1 Code

```
m.lock();  
x = 1;  
m.unlock();
```

## Thread 2 Code

```
m.lock();  
y = x;  
m.unlock();
```

## Execution



- since unlock synchronizes with lock, A happens before B; thus, for timing shown, B must see 1 for x

- Most operations on atomic types allow a memory order to be specified.

- Example:

```
std::atomic<int> x = 0;  
x.store(42, std::memory_order_seq_cst);  
int y = x.load(std::memory_order_seq_cst);
```

- The following memory orders are supported:

- sequentially consistent (`std::memory_order_seq_cst`)
- acquire-release (`std::memory_order_acq_rel`)
- acquire (`std::memory_order_acquire`)
- release (`std::memory_order_release`)
- consume (`std::memory_order_consume`)
- relaxed (`std::memory_order_relaxed`)

- Read operations can use the orders:

- sequentially consistent, acquire, consume, and relaxed.

- Write operations can use the orders:

- sequentially consistent, release, and relaxed.

- Read-modify-write operations can use:

- all of the orders allowed for read and write operations; and
- acquire-release.

# Memory Models

- Although several memory orders can be employed for operations on atomic types, these orders support *four basic models*:
  - 1 sequentially consistent,
  - 2 acquire release,
  - 3 consume release, and
  - 4 relaxed.
- These models differ in the guarantees that they make regarding:
  - whether all writes to all atomic objects become visible to *all* threads *simultaneously* (i.e., total order for all writes to all atomic objects); and
  - whether operations on atomic objects in different threads can establish a *synchronization* relationship (namely, a synchronizes-with or dependency-ordered-before [discussed later] relationship).
- The models listed from strongest (i.e., makes the most guarantees) to weakest (i.e., makes the least guarantees) are:
  - 1 sequentially consistent,
  - 2 acquire release,
  - 3 consume release, and
  - 4 relaxed.

# Memory Models (Continued 1)

- These models are *hierarchical* in the sense that each model makes at least all of the same guarantees as its weaker counterparts.
- As we proceed from stronger to weaker models, more guarantees are lost.
- A stronger model may require additional synchronization by hardware, which can *degrade performance*.
- A weaker model *may not provide sufficient guarantees* for the correct functioning of code.
- Using a model that fails to provide sufficient guarantees for correct code behavior will result in *bugs*.
- Also, as the model is weakened, it becomes more difficult to reason about the behavior of code, leading to *incomprehensible code* and an *increased likelihood of (often very subtle) bugs*.

# Modification Order

- All writes to a particular atomic object  $M$  (over its lifetime) occur in some particular total order, called its **modification order**.
- Each atomic object has its own well-defined modification order.
- For a particular atomic object  $M$ , *all* threads in a program are guaranteed to see  $M$  change in a manner *consistent with its modification order*.
- Essentially, this guarantee ensures that, once a given thread has seen a particular value of an atomic object, a subsequent read by that thread cannot retrieve an earlier value of the object.
- If such a guarantee were not made, the memory model would be so weak as to be impractical to use.
- Modification order is primarily a *conceptual* tool that is useful for describing memory-model behavior.
- In practice, a thread is unlikely to actually observe every change in the modification order of an object.

## Modification Order (Continued)

- For each atomic object  $M$ , each thread has its own current position in object's modification order.
- A thread's current position in the modification order of a particular atomic object need not be the same for all threads.
- A read from an atomic object  $M$  by a thread  $T$  can *optionally* move  $T$ 's current position to a later position in the modification order of  $M$  and then returns the value at the current position.
- A write to an atomic object  $M$  by a thread  $T$  appends the value to be written to the modification order of  $M$  and updates  $T$ 's current position in the modification order of  $M$  to correspond to the value written.
- An read-modify-write operation  $A$  on an atomic object  $M$  reads the *last* value in the modification order of  $M$ , modifies the value read appropriately, appends the resulting value to the modification order of  $M$ , and updates  $T$ 's current position in the modification order of  $M$  to correspond to the value written.

# Modification Order Example

- Consider an atomic object  $M$  with the modification sequence:
  - 0, 1, 2, 3, 4, 5, 6, 7, 8.
- A thread could, for example, legitimately see  $M$  undergo any of the following sequences of updates:
  - 0, 4, 8
  - 8
  - 2, 7
  - 0, 1, 2, 5, 7, 8
  - 0, 1, 2, 3, 4, 5, 6, 7, 8
- A thread would, for example, be guaranteed *never* to see  $M$  undergo any of the following sequences of updates, as all of these sequences are *inconsistent* with the modification order of  $M$ :
  - 1, 0
  - 1, 2, 1
  - 42
  - 0, 1, 2, 3, 4, 5, 6, 7, 6, 8



# Relative Ordering of Changes to Different Atomic Objects

- Although each atomic object has its own well-defined modification order, it is not necessarily the case that the modification orders for individual objects can be combined into a single total order over *all* atomic objects.
- Practically speaking, the reason for this is the delay in the visibility of results introduced by store buffers, caches, and so on.
- If a single total order for writes to all atomic objects is not guaranteed, this implies that the relative order of changes to *different* atomic objects need not appear the same to different threads.
- Ensuring the existence of a single total order over all atomic objects would require a significant amount of additional processor synchronization, which can significantly degrade performance.
- Therefore, this guarantee is not required to be made in all cases, the idea being that we only ask for the guarantee when it is needed for correct code behavior.

# Modification Order Revisited

- Consider a program with two threads and two shared integer atomic objects `x` and `y`, each having the modification order: 0, 1.
- Suppose that no requirement is imposed to guarantee the existence of a single total order on writes to *all* atomic objects.
- Thread 1 could see `x` and `y` change in the following manner, consistent with their stated modification order:

Variable	Updates to Value Seen By Thread
<code>x</code>	0 ..... 1 .....
<code>y</code>	0 ..... 1 .....

- Thread 2 could see `x` and `y` change in the following manner, consistent with their stated modification order:

Variable	Updates to Value Seen By Thread
<code>x</code>	0 ..... 1 .....
<code>y</code>	0 ..... 1 .....

- Observe that thread 1 and thread 2 do not see `x` and `y` change in the same order relative to one another (i.e., thread 1 sees `x` change before `y`, while thread 2 sees `y` change before `x`).

# Sequentially-Consistent Model

- The sequentially-consistent model simply corresponds to the default memory model for the language, namely, SC-DRF. (Since data races cannot occur on atomic objects, SC-DRF degenerates into SC for such objects.)
- For the sequentially-consistent model, all memory operations (i.e., read, write, and read-modify-write) must use the sequentially-consistent memory order (`std::memory_order_seq_cst`).
- A *total ordering* is guaranteed on all sequentially-consistent writes to *all* atomic objects.
- All sequentially-consistent writes to atomic objects must become *visible* to all threads *simultaneously*.
- A sequentially-consistent write operation  $W$  on an atomic object  $M$  (in one thread) *synchronizes with* a sequentially-consistent operation on  $M$  (in another thread) that reads the value written by  $W$ .
- This model allows for relatively *easy reasoning* about code behavior.

# Example: Sequentially-Consistent Model

- shared data:

`x` and `y` are of type `std::atomic<int>` and both are initially zero

- thread 1 code (writes `x`):

```
x.store(1, std::memory_order_seq_cst);
```

- thread 2 code (writes `y`):

```
y.store(1, std::memory_order_seq_cst);
```

- thread 3 code (reads `x` then `y`):

```
int x1 = x.load(std::memory_order_seq_cst);
```

```
int y1 = y.load(std::memory_order_seq_cst);
```

- thread 4 code (reads `y` then `x`):

```
int y2 = y.load(std::memory_order_seq_cst);
```

```
int x2 = x.load(std::memory_order_seq_cst);
```

- memory order guarantees total order for all writes to all atomic objects

- so, thread 3 and thread 4 must agree about order in which `x` and `y` are modified

- not possible to see `x1 == 1` and `y1 == 0` in thread 3 (implying `x` modified before `y`) and `x2 == 0` and `y2 == 1` in thread 4 (implying `y` modified before `x`)

# Example: Sequentially-Consistent Model

```
1  #include <atomic>
2  #include <thread>
3  #include <cassert>
4
5  std::atomic<int> x, y, c;
6
7  void w_x() {x.store(1, std::memory_order_seq_cst);}
8
9  void w_y() {y.store(1, std::memory_order_seq_cst);}
10
11 void r_xy() {
12     while (!x.load(std::memory_order_seq_cst)) {}
13     if (y.load(std::memory_order_seq_cst)) {++c;}
14 }
15
16 void r_yx() {
17     while (!y.load(std::memory_order_seq_cst)) {}
18     if (x.load(std::memory_order_seq_cst)) {++c;}
19 }
20
21 int main() {
22     x = 0; y = 0; c = 0;
23     std::thread t1(w_x), t2(w_y), t3(r_xy), t4(r_yx);
24     t1.join(); t2.join(); t3.join(); t4.join();
25     assert(c != 0); // assertion cannot fail
26 }
```

- assertion cannot fail: when **while** loop in `r_xy` terminates, all threads must see `x` as nonzero; when **while** loop in `r_yx` terminates, all threads must see `y` as nonzero; at least one of these must happen before **if**

# Acquire-Release Model

- For the acquire-release model, the memory order is chosen as follows:
  - a read operation uses the acquire order (`std::memory_order_acquire`)
  - a write operation uses the release order (`std::memory_order_release`)
  - a read-modify-write operation uses one of the orders allowed for read and write operations, or the acquire-release order (`std::memory_order_acq_rel`), which results in read acquire and write release.
- *No total ordering* exists on all writes to *all* atomic objects (unlike in the sequentially-consistent model).
- Consequently, threads do not necessarily have to agree on the *relative order* in which different atomics objects are modified.
- A write-release operation  $W$  on an atomic object  $M$  *synchronizes with* a read-acquire operation on  $M$  that reads the value written by  $W$  (or a value written by the release sequence headed by  $W$ ). [\[C++17 32.4/2\]](#)
- The acquire-release model is useful for situations that involve *pairwise synchronization* of threads, such as with mutexes.
- With the acquire-release model, it is often still possible to reason about code behavior without too much difficulty.

# Example: Acquire-Release Model

- shared data:

`x` and `y` are of type `std::atomic<int>` and both are initially zero

- thread 1 code (writes `x`):

```
x.store(1, std::memory_order_release);
```

- thread 2 code (writes `y`):

```
y.store(1, std::memory_order_release);
```

- thread 3 code (reads `x` then `y`):

```
int x1 = x.load(std::memory_order_acquire);
```

```
int y1 = y.load(std::memory_order_acquire);
```

- thread 4 code (reads `y` then `x`):

```
int y2 = y.load(std::memory_order_acquire);
```

```
int x2 = x.load(std::memory_order_acquire);
```

- no ordering relationship between stores to `x` and `y`

- so, thread 3 and thread 4 do not need to agree about order in which `x` and `y` are modified

- possible to see `x1 == 1` and `y1 == 0` in thread 3 (i.e., thread 3 sees `x` change before `y`) and `x2 == 0` and `y2 == 1` in thread 4 (i.e., thread 4 sees `y` change before `x`)

# Example: Acquire-Release Model

```
1  #include <atomic>
2  #include <thread>
3  #include <cassert>
4
5  std::atomic<int> x, y, c;
6
7  void w_x() {x.store(1, std::memory_order_release);}
8
9  void w_y() {y.store(1, std::memory_order_release);}
10
11 void r_xy() {
12     while (!x.load(std::memory_order_acquire)) {}
13     if (y.load(std::memory_order_acquire)) {++c;}
14 }
15
16 void r_yx() {
17     while (!y.load(std::memory_order_acquire)) {}
18     if (x.load(std::memory_order_acquire)) {++c;}
19 }
20
21 int main() {
22     x = 0; y = 0; c = 0;
23     std::thread t1(w_x), t2(w_y), t3(r_xy), t4(r_yx);
24     t1.join(); t2.join(); t3.join(); t4.join();
25     assert(c != 0); // assertion can fail
26 }
```

- **assertion can fail: one thread seeing  $x$  or  $y$  being nonzero does not imply other thread sees same**



## Example: Spinlock Mutex Using `std::atomic_flag`

```
1  #include <iostream>
2  #include <thread>
3  #include <atomic>
4
5  class SpinLockMutex {
6  public:
7      SpinLockMutex() {f_.clear();}
8      void lock() {
9          while (f_.test_and_set(std::memory_order_acquire)) {}
10     }
11     void unlock() {f_.clear(std::memory_order_release);}
12 private:
13     std::atomic_flag f_; // true if thread holds mutex
14 };
15
16 SpinLockMutex m;
17 unsigned long long counter = 0;
18
19 void doWork() {
20     for (unsigned long long i = 0; i < 1000000ULL; ++i)
21         {m.lock(); ++counter; m.unlock();}
22 }
23
24 int main() {
25     std::thread t1(doWork), t2(doWork);
26     t1.join(); t2.join();
27     std::cout << counter << '\n';
28 }
```

- uses acquire-release model

## Example: Spinlock Mutex and `std::scoped_lock`

```
1  #include <iostream>
2  #include <thread>
3  #include <atomic>
4  #include <mutex>
5
6  class SpinLockMutex {
7  public:
8      SpinLockMutex() {f_.clear();}
9      void lock() {
10         while (f_.test_and_set(std::memory_order_acquire)) {}
11     }
12     void unlock() {f_.clear(std::memory_order_release);}
13 private:
14     std::atomic_flag f_; // true if thread holds mutex
15 };
16
17 SpinLockMutex m;
18 unsigned long long counter = 0;
19
20 void doWork() {
21     for (unsigned long long i = 0; i < 1000000ULL; ++i)
22         {std::scoped_lock lg(m); ++counter;}
23 }
24
25 int main() {
26     std::thread t1(doWork), t2(doWork);
27     t1.join(); t2.join();
28     std::cout << counter << '\n';
29 }
```

# Carries-A-Dependency Relationships

- For two operations  $A$  and  $B$  performed in the *same* thread,  $A$  is said to **carry a dependency** to  $B$  if the result of  $A$  is used as an operand for  $B$  (ignoring some special cases). [\[C++17 4.7.1/7\]](#)
- Example: In the code below, statement  $A$  *carries a dependency* to statement  $B$  but not statement  $C$ .

```
x = 42;    // A
y = x + 1; // B
z = 0;     // C
```

- Note that “carries a dependency to” is a subset of “is sequenced before” (i.e., the former implies the latter). [\[C++17 4.7.1/7\]](#)
- The carries-a-dependency-to relationship is *transitive* (i.e., if  $A$  carries a dependency to  $B$  and  $B$  carries a dependency to  $C$  then  $A$  carries a dependency to  $C$ ).
- Example: In the code below, statement  $A$  carries a dependency to statement  $B$ ; and  $B$  carries a dependency to statement  $C$ . Therefore, transitively,  $A$  carries a dependency to  $C$ .

```
x = 42;    // A
y = x + 1; // B
z = 2 * y; // C
```

# Dependency-Ordered-Before Relationships

- Another type of synchronization relationship is known as a dependency-ordered-before relationship.
- A write-release operation  $A$  is *dependency ordered before* a read-consume operation  $B$  if  $B$  reads the value written by  $A$  (or any side effect in the release sequence headed by  $A$ ).
- For two operations  $A$  and  $B$  performed in *different* threads, if  $A$  is dependency ordered before  $B$  then  $A$  *inter-thread happens before*  $B$ .
- Thus, dependency-ordered-before relationships can also establish happens-before relationships.

# Inter-Thread Happens-Before Relationships Revisited

- The inter-thread happens before relation describes an *arbitrary concatenation* of sequenced-before, synchronizes-with, and dependency-ordered-before relations, *with two exceptions*: [C++17 4.7.1/9]
  - 1 a concatenation is not permitted to end with dependency ordered before followed by (one or more) sequenced before; and
  - 2 a concatenation is not permitted to consist entirely of sequenced-before relations.
- The first restriction is required since a dependency-ordered-before relationship synchronizes *only data dependencies*.
- The second restriction is required since inter-thread happens-before relationship must (by definition) involve operations in *different* threads.

# Consume-Release Model

- For the consume-release model, the memory order is chosen as follows:
  - a write operation uses release order (`std::memory_order_release`)
  - a read operation uses the consume order (`std::memory_order_consume`)
- The consume-release model is identical to the acquire-release model with one important difference, namely the type of synchronization relationship established.
- A write-release operation  $W$  is *dependency ordered before* a read-consume operation (in a different thread) that reads the value stored by  $W$  (or any side effect in the release sequence headed by  $W$ ).
- In other words, the consume-release model establishes a *dependency-ordered-before* relationship, whereas the acquire-release model establishes a *synchronizes-with* relationship.
- In this sense, the consume-release model is weaker than the acquire-release model (i.e., less data is synchronized).

# Example: Consume-Release Model

```
1  #include <thread>
2  #include <atomic>
3  #include <cassert>
4
5  int x = 0;
6  std::atomic y(0);
7
8  void producer() {
9      x = 42;
10     y.store(1, std::memory_order_release);
11 }
12
13 void consumer() {
14     int a;
15     while (!(a = y.load(std::memory_order_consume))) {}
16     assert(x == 42); // data race
17 }
18
19 int main() {
20     std::thread t1(producer);
21     std::thread t2(consumer);
22     t1.join();
23     t2.join();
24 }
```

- program has *data race* on x; a does not carry dependency to x so x = 42 does not necessarily happen before x used in assertion
- if consume changed to acquire, no data race and assertion cannot fail

# Example: Publishing Data Via Pointer

```
1  #include <thread>
2  #include <atomic>
3  #include <cassert>
4  #include <string>
5
6  std::atomic<std::string*> p(nullptr);
7  int x = 0;
8
9  void producer() {
10     std::string* s = new std::string("Hello");
11     x = 42;
12     p.store(s, std::memory_order_release);
13 }
14
15 void consumer() {
16     std::string* s;
17     while (!(s = p.load(std::memory_order_consume))) {}
18     assert(*s == "Hello");
19     // assert(x == 42); would result in data race
20 }
21
22 int main() {
23     std::thread t1(producer), t2(consumer);
24     t1.join(); t2.join();
25 }
```

- assertion cannot fail; store to p is dependency ordered before load and load carries dependency to \*s in assertion



# Relaxed Model

- For the relaxed model, all memory operations use the relaxed order (`std::memory_order_relaxed`).
- Like in the acquire-release model, *no total order* exists on updates to *all* atomic objects (collectively).
- Operations on the same variable *within a single thread* satisfy a happens-before relationship (i.e., within a single thread, accesses to a single atomic variable must follow program order).
- Unlike in the acquire-release model, *no inter-thread synchronization* relationship is established.
- No requirement exists on the ordering relative to other threads.
- The relaxed order is sometime suitable for updating counters (e.g., blind event counters).
- Except in very trivial cases, it can be *extremely difficult to reason* about the meaning and/or correctness of code that uses relaxed order.

# Behavior of Relaxed Model

- consider atomic memory operations with relaxed order
- for each individual atomic object, all threads have view of updates that is consistent with single modification sequence
- read operation (e.g., `load`):
  - if current position not set, return any element in sequence and set current position to that of returned element
  - otherwise, either leave current position unchanged or move later in sequence and return value at current position
- write operation (e.g., `store`):
  - append value to end of sequence
  - set current position to correspond to appended value
- read-modify-write operation (e.g., `increment`, `decrement`, `exchange`, `compare_exchange`):
  - read last value from sequence
  - modify read value as appropriate to obtain new value
  - append new value to end of sequence
  - set current position to correspond to that of appended value
- considerable flexibility in value returned by read

# Example: Relaxed Model

```
1  #include <atomic>
2  #include <thread>
3  #include <cassert>
4
5  std::atomic<int> x, y, c;
6
7  void w_x() {x.store(1, std::memory_order_relaxed);}
8
9  void w_y() {y.store(1, std::memory_order_relaxed);}
10
11 void r_xy() {
12     while (!x.load(std::memory_order_relaxed)) {}
13     if (y.load(std::memory_order_relaxed)) {++c;}
14 }
15
16 void r_yx() {
17     while (!y.load(std::memory_order_relaxed)) {}
18     if (x.load(std::memory_order_relaxed)) {++c;}
19 }
20
21 int main() {
22     x = 0; y = 0; c = 0;
23     std::thread t1(w_x), t2(w_y), t3(r_xy), t4(r_yx);
24     t1.join(); t2.join(); t3.join(); t4.join();
25     assert(c != 0); // assertion can fail
26 }
```

- **assertion can fail: one thread seeing  $x$  or  $y$  being nonzero does not imply other thread sees same**

# Example: Blind Event Counters

```
1  #include <vector>
2  #include <iostream>
3  #include <thread>
4  #include <atomic>
5
6  std::atomic<unsigned long long> counter(0);
7
8  void doWork() {
9      for (long i = 0; i < 100'000L; ++i)
10         {counter.fetch_add(1, std::memory_order_relaxed);}
11  }
12
13  int main() {
14      std::vector<std::thread> workers;
15      for (int i = 0; i < 10; ++i) {workers.emplace_back(doWork);}
16      for (auto& t : workers) {t.join();}
17      std::cout << "counter " << counter << '\n';
18  }
```

- `fetch_add` can use *relaxed* order, since only incrementing counter *blindly* (i.e., not taking action based on value of counter)
- thread join operations provide synchronization to ensure desired value read for counter when output

# Example: Done Flag

```
1  #include <atomic>
2  #include <chrono>
3  #include <thread>
4  #include <vector>
5
6  std::atomic<bool> done;
7
8  void doWork() {
9      while (!done.load(std::memory_order_relaxed)) {
10         // do something here
11     }
12 }
13
14 int main() {
15     std::vector<std::thread> workers;
16     done.store(false, std::memory_order_relaxed); // I hope? ;)
17     for (int i = 0; i < 16; ++i) {workers.emplace_back(doWork);}
18     std::this_thread::sleep_for(std::chrono::seconds(5));
19     done = true; // not relaxed
20     for (auto& t : workers) {t.join();}
21 }
```

- `done.store` can be relaxed due to synchronization from thread create
- `done.load` can be relaxed since order not important; different order as if other threads ran at different speeds
- assign to `done` must be sequentially-consistent to prevent assign from floating past join (due to single-thread optimization)

## Example: `std::shared_ptr` Reference Counting

- The copy constructor for `shared_ptr` (which increments a reference count) would look something like:

```
// ...
controlBlockPtr = other->controlBlockPtr;
controlBlockPtr->refCount.fetch_add(1,
    std::memory_order_relaxed);
// ...
```

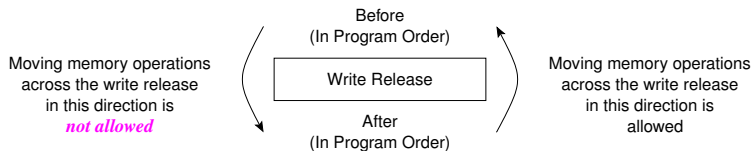
- The destructor for `shared_ptr` (which decrements a reference count) would look something like:

```
// ...
if (!controlBlockPtr->refCount.fetch_sub(1,
    std::memory_order_acq_rel)) {
    delete controlBlockPtr;
}
// ...
```

- The increment operation can use *relaxed* order, since *no action is taken* based on the reference count value.
- The decrement operation needs to use *acquire-release* order so that the decrement cannot float and the correct view of the data is seen by the thread doing the delete (all decrements form a *synchronization chain*).

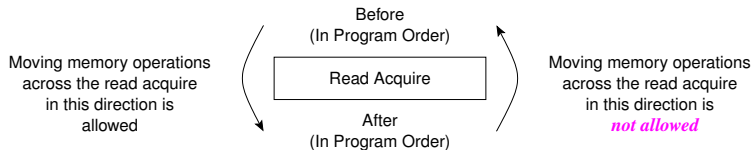
# Release Semantics for Memory Operations

- Release semantics is a property that can only apply to operations that *write to memory* (i.e., read-modify-write operations or plain writes).
- A write operation that has release semantics is called a **write release**.
- A write release operation  $W$  cannot be reordered with any read or write operation that *precedes*  $W$  in program order (i.e., memory operations cannot be moved from before  $W$  to after  $W$ ).
- The term release semantics originates from mutexes.
- In the context of mutexes, the operations prior to a mutex release operation, which correspond to operations in a critical section, must not be moved after the mutex release operation, as operations after the mutex release operation are not protected by the mutex.



# Acquire Semantics for Memory Operations

- Acquire semantics is a property that can only apply to operations that *read from memory* (i.e., read-modify-write operations or plain reads).
- A read operation that has acquire semantics is called a **read acquire**.
- A read acquire operation  $R$  cannot be reordered with any read or write operation that *follows*  $R$  in program order (i.e., memory operations cannot be moved from after  $R$  to before  $R$ ).
- The term acquire semantics originates from mutexes.
- In the context of mutexes, the operations following a mutex acquire operation, which correspond to operations in a critical section, must not be moved before the mutex acquire operation, as operations before the mutex acquire operation are not protected by the mutex.





- A **release sequence** headed by a release operation  $A$  on an atomic object  $M$  is a maximal contiguous subsequence of side effects in the modification order of  $M$ , where the first operation is  $A$ , and every subsequent operation [C++17 4.7.1/5]
  - is performed by the same thread that performed  $A$ , or
  - is an atomic read-modify-write operation.

# Release Sequence Example

```
1  #include <thread>
2  #include <atomic>
3  #include <cassert>
4
5  int x = 0;
6  std::atomic y(0);
7
8  int main() {
9      std::thread t1([]() {
10         x = 42;
11         y.store(1, std::memory_order_release); // A
12         y.store(2, std::memory_order_relaxed); // B
13     });
14     std::thread t2([]() {
15         int r;
16         while ((r = y.load(std::memory_order_acquire)) // C
17             < 2) {}
18         assert(x == 42);
19     });
20     t1.join();
21     t2.join();
22 }
```

- stores to y in A and B constitute release sequence headed by store in A
- when while loop terminates, load in C will have read value written by store in B (not store in A)
- A synchronizes with C, since C reads value in release sequence headed by A
- assertion cannot fail, since A happens before C

# Fences

- A **memory fence** (also known as a **memory barrier**) is an operation that causes the processor and compiler to enforce an *ordering constraint* on memory operations issued before and after the fence operation.
- Certain types of memory operations before a fence are guaranteed not to be reordered with certain types of memory operations after the fence.
- A fence may also introduce *synchronizes-with* relationships under certain circumstances.
- An **acquire fence** prevents the reordering of any *read or write* following the fence (in program order) with any *read* prior to the fence (in program order). (That is, a memory operation after the fence cannot be moved before any read operation before the fence.)
- A **release fence** prevents the reordering of any *read or write* prior to the fence (in program order) with any *write* following the fence (in program order). (That is, a memory operation before the fence cannot be moved after any write operation after the fence.)
- A fence is *not* a release or acquire operation. as it does not read/write memory.

- memory fences can be inserted via function

`std::atomic_thread_fence`

- declaration:

```
void atomic_thread_fence(std::memory_order order)  
    noexcept;
```

- no effect if order is `std::memory_order_relaxed`
- acquire fence if order is `std::memory_order_acquire` or `std::memory_order_consume`
- release fence if order is `std::memory_order_release`
- both acquire and release fence if order is `std::memory_order_acq_rel`
- sequentially consistent acquire and release fence if order is `std::memory_order_seq_cst`

# Fences and Synchronizes-With Relationships

- ***Release fence and acquire fence.*** A release fence  $A$  synchronizes with an acquire fence  $B$  if there exist atomic operations  $X$  and  $Y$ , both operating on some atomic object  $M$ , such that  $A$  is sequenced before  $X$ ,  $X$  modifies  $M$ ,  $Y$  is sequenced before  $B$ , and  $Y$  reads the value written by  $X$  or a value written by any side effect in the hypothetical release sequence  $X$  would head if it were a release operation. [\[C++17 32.9/2\]](#)
- ***Release fence and acquire operation.*** A release fence  $A$  synchronizes with an atomic operation  $B$  that performs an acquire operation on an atomic object  $M$  if there exists an atomic operation  $X$  such that  $A$  is sequenced before  $X$ ,  $X$  modifies  $M$ , and  $B$  reads the value written by  $X$  or a value written by any side effect in the hypothetical release sequence  $X$  would head if it were a release operation. [\[C++17 32.9/3\]](#)
- ***Release operation and acquire fence.*** An atomic operation  $A$  that is a release operation on an atomic object  $M$  synchronizes with an acquire fence  $B$  if there exists some atomic operation  $X$  on  $M$  such that  $X$  is sequenced before  $B$  and reads the value written by  $A$  or a value written by any side effect in the release sequence headed by  $A$ . [\[C++17 32.9/4\]](#)

# Example: Incorrect Code Without Fence

```
1  #include <thread>
2  #include <atomic>
3  #include <iostream>
4
5  std::atomic ready(false);
6  int data = 0;
7
8  void produce() {
9      data = 42; // write to data can move after store in A
10     // release fence needed here
11     ready.store(true, std::memory_order_relaxed); // A
12 }
13
14 void consume() {
15     while (!ready.load(std::memory_order_relaxed)) {} // B
16     // acquire fence needed here
17     std::cout << data << '\n';
18     // read of data can move before load in B
19 }
20
21 int main() {
22     std::thread t1(produce);
23     std::thread t2(consume);
24     t1.join(); t2.join();
25 }
```

- atomic store (to ready) does not synchronize with atomic load (of ready), due to relaxed order; results in race on data

# Example: Correct Code With Fence

```
1  #include <thread>
2  #include <atomic>
3  #include <iostream>
4
5  std::atomic ready(false);
6  int data = 0;
7
8  void produce() {
9      data = 42;
10     std::atomic_thread_fence(std::memory_order_release);
11     ready.store(true, std::memory_order_relaxed);
12 }
13
14 void consume() {
15     while (!ready.load(std::memory_order_relaxed)) {}
16     std::atomic_thread_fence(std::memory_order_acquire);
17     std::cout << data << '\n';
18 }
19
20 int main() {
21     std::thread t1(produce);
22     std::thread t2(consume);
23     t1.join(); t2.join();
24 }
```

- release fence synchronizes with acquire fence, due to atomic load (of ready) reading from result of atomic store (to ready)

# Memory Orders: The Bottom Line

- Use sequentially-consistent order unless there is a compelling case to do otherwise.
- In situations where semantics dictate a clear pairwise synchronization between threads, consider the use of acquire-release order if it can be easily seen to yield correct code.
- Only consider relaxed order in situations where the performance penalty of using a stronger order would be unacceptable.
- *Be very wary of using relaxed order.* Even world experts on the C++ memory model acknowledge that this can be tricky.
- Always have any code using relaxed order thoroughly reviewed by people who are extremely knowledgeable about memory models.



## Section 3.9.11

### **References**

- 1 A. Williams. *C++ Concurrency in Action*.

Manning Publications, Shelter Island, NY, USA, 2012.

This is a fairly comprehensive book on concurrency and multithreaded programming in C++. It is arguably the best book available for those who want to learn how to write multithreaded code using C++.

- 2 M. J. Batty. *The C11 and C++11 Concurrency Model*.

PhD thesis, University of Cambridge, Cambridge, UK, Nov. 2014.

This very well written Ph.D. thesis introduces the C++11/C11 memory model and presents work in mathematically formalizing, refining, and validating this model.

- 3 M. Batty. *Multicore Programming: C++0x*.

University of Cambridge, Cambridge, UK, Nov. 2010.

This set of slides appear to have been used for part of a course on multicore programming at the University of Cambridge.

- 4 M. Herlihy and N. Shavit. *The Art of Multiprocessor Programming*. Morgan Kaufmann, Burlington, MA, USA, 2008.  
A good reference for concurrent programming.
- 5 S. V. Adve and K. Gharachorloo. Shared memory consistency models: A tutorial.  
*IEEE Computer*, 29(12):66–76, Dec. 1996.
- 6 S. V. Adve and H.-J. Boehm. Memory models: A case for rethinking parallel languages and hardware.  
*Communications of the ACM*, 53(8):90–101, Aug. 2010.
- 7 H.-J. Boehm and S. V. Adve. You don't know jack about shared variables or memory models.  
*Communications of the ACM*, 55(2):48–54, Feb. 2012.
- 8 H.-J. Boehm, Memory Model Rationale, ISO/IEC JTC1/SC22/WG14/N1479, May 2010.  
<http://www.open-std.org/jtc1/sc22/wg14/www/docs/n1479.htm>

- 1 Herb Sutter. `atomic<>` Weapons: The C++11 Memory Model and Modern Hardware. C++ and Beyond, Asheville, NC, USA, Aug. 5–8, 2012. Available online at <https://channel9.msdn.com/Shows/Going+Deep/Cpp-and-Beyond-2012-Herb-Sutter-atomic-Weapons-1-of-2> and <https://channel9.msdn.com/Shows/Going+Deep/Cpp-and-Beyond-2012-Herb-Sutter-atomic-Weapons-2-of-2>. (This talk is in two parts.)
- 2 Herb Sutter. C++ Concurrency. C++ and Beyond, Asheville, NC, USA, Aug. 5–8, 2012. Available online at <https://channel9.msdn.com/Shows/Going+Deep/C-and-Beyond-2012-Herb-Sutter-Concurrency-and-Parallelism>.
- 3 Herb Sutter. Lock-Free Programming (Or, Juggling Razor Blades). CppCon, 2014. Available online at <https://youtu.be/c1gO9aB9nbs> and <https://youtu.be/CmxkPChOcvw>. (This talk is in two parts.)

- 4 Hans-J. Boehm. Threads and Shared Variables in C++11. Going Native, Redmond, WA, USA, Feb. 2–3, 2012. Available online at <https://channel9.msdn.com/Events/GoingNative/GoingNative-2012/Threads-and-Shared-Variables-in-C-11>.
- 5 Mike Long. Introducing the C++ Memory Model. Norwegian Developers Conference, Oslo, Norway, Jun. 15–19, 2014. Available online at <https://vimeo.com/97419179>.
- 6 Herb Sutter. Machine Architecture and You: Things Your Programming Language Never Told You. Northwest C++ Users' Group, Redmond, WA, USA, Sept. 19, 2007. Available online at <http://nwcpp.org/september-2007.html>.
- 7 Pablo Halpern. Overview of Parallel Programming in C++. CppCon, Bellevue, WA, USA, Sept. 8, 2014. Available online at <https://youtu.be/y0GSc5fKt18>.

- 8 Valentin Ziegler. C++ Memory Model. Meeting C++, Berlin, Germany, Dec. 6, 2014. Available online at <https://youtu.be/gpsz8sc6mNU>.
- 9 Jeff Preshing. How Ubisoft Montreal Develops Games for Multicore — Before and After C++11. CppCon, Bellevue, WA, USA, Sept. 11, 2014. Available online at <https://youtu.be/X1T3IQ4N-3g>.
- 10 Sean Parent. Better Code: Concurrency. code::dive, Wrocław, Poland, Nov. 15–16, 2016. Available online at <https://youtu.be/QIHy8pXbneI>.

## Section 3.10

# Compilers and Linkers

# Application Binary Interfaces (ABIs)

- application binary interface (ABI) specifies how data structures or subroutines accessed at machine-code level
- in other words, ABI is like API, but at level of machine code
- ABI typically covers details such as:
  - processor instruction set
  - sizes, layouts, and alignments of data types
  - calling conventions, which include details such as:
    - how parameters passed to functions
    - how return values passed back from functions
    - what state (e.g., registers) needs to be preserved by caller/callee
  - how exceptions handled
  - how source code identifiers should be mapped to identifiers in object code (i.e., name mangling)
  - how calls made to operating system
- compilers need to adhere to ABI to allow object-code compatibility between different compilers



# Name Mangling

- name mangling is process that maps set of names to another set of names
- typically name mangling used to map names in source code (which may be overloaded) to names in object code
- to keep linkers from becoming overly complex, do not want linkers to have to support overloaded identifiers
- thus, each overloaded function must have distinct name in object code
- unique names in object code obtained by applying name mangling to names from source code
- name mangling rules not specified by C++ standard
- typically, name mangling rules addressed by ABI standard, such as Itanium ABI
- some examples of name mangling results with Itanium ABI as follows:

Declaration	Mangled Name
<b>bool</b> func( <b>int</b> )	<code>_Z4funci</code>
<b>bool</b> func( <b>double</b> , <b>double</b> )	<code>_Z4funcdd</code>

## Section 3.10.1

### **Itanium C++ ABI**

- one ABI followed by number of popular C++ compilers is Itanium C++ ABI
- Itanium C++ ABI builds on various other ABI documents (some of which are related to either System V or Intel Itanium architecture)
- amongst other things, Itanium C++ ABI specifies:
  - some implementation details for constructors and destructors
  - name mangling rules
- web site:
  - <https://itanium-cxx-abi.github.io/cxx-abi/abi.html>
- developed jointly by informal industry coalition consisting of CodeSourcery, Compaq, EDG, HP, IBM, Intel, Red Hat, and SGI
- Itanium C++ ABI followed by GCC and Clang on most platforms (with Microsoft Windows being notable exception)

# Constructors

- each constructor written by programmer nominally results in compiler emitting code for *multiple* functions, which together implement several variations on construction
- each constructor for class  $T$  nominally associated with three functions:
  - 1 **base object constructor**: function that creates non-virtual (direct) base class objects and non-static data members of  $T$  and initializes  $T$  object itself
  - 2 **complete object constructor**: function that creates all virtual base class objects of  $T$  and then additionally performs all work of base object constructor
  - 3 **allocating object constructor**: function that performs all work of complete object constructor after obtaining storage for new  $T$  object from allocation function (i.e., **operator new**)
- if no virtual base classes, complete object constructor and base object constructor are same (e.g., one function is alias for other)
- ABI specification seems to suggest that allocating constructor only required if class has virtual destructor
- this said, however, GCC appears to never emit code for allocating object constructor

# Destructors

- each destructor written by programmer nominally results in compiler emitting code for *multiple* functions, which together implement several variations on destruction
- destructor for class  $\mathbb{T}$  nominally associated with three functions:
  - 1 **base object destructor**: function that performs any clean-up action for  $\mathbb{T}$  object itself and destroys non-static data members and non-virtual (direct) base class objects of  $\mathbb{T}$
  - 2 **complete object destructor**: function that, in addition to actions of base object destructor, destroys all virtual base class objects of  $\mathbb{T}$
  - 3 **deleting destructor**: function that, in addition to actions of complete object destructor, invokes appropriate deallocation function for  $\mathbb{T}$
- complete object destructor in charge of destroying virtual base class objects, whereas base object destructor is not
- if no virtual base classes, base object destructor and complete object destructor are same (e.g., one is alias for other)
- deleting destructor must be emitted when  $\mathbb{T}$  has virtual destructor; otherwise, may be emitted but not required

# Name Mangling

- many rules for name mangling; only a few mentioned here
- all mangled symbols begin with `_Z`
- nested name delimited by `N` and `E`
- `St` denotes `std` namespace; `Sa` denotes `std::allocator`
- template arguments delimited by `I` and `E`
- letter codes for some built-in types:

Letter Code	Built-In Type
v	<b>void</b>
b	<b>bool</b>
c	<b>char</b>
h	<b>unsigned char</b>
i	<b>int</b>

Letter Code	Built-In Type
j	<b>unsigned int</b>
l	<b>long</b>
m	<b>unsigned long</b>
f	<b>float</b>
d	<b>double</b>

- letter codes for cv-qualifiers:

Letter Code	Qualifier
V	<b>volatile</b>
K	<b>const</b>

# Name Mangling (Continued)

- letter codes for type qualifiers:

Letter Code	Qualifier
P	pointer
R	lvalue reference
O	rvalue reference

- codes for constructors and destructors:

Code	Description
C1	complete object constructor
C2	base object constructor
C3	allocating constructor
D0	deleting destructor
D1	complete object destructor
D2	base object destructor

# Name Mangling Examples

```
1  #include <iostream>
2  #include <vector>
3
4  struct Widget {
5      Widget();
6      Widget(int);
7      ~Widget();
8      int grog() const;
9  };
10 Widget::Widget() {}
11     // Widget::Widget() -> _ZN6WidgetC1Ev, _ZN6WidgetC2Ev
12 Widget::Widget(int i) {}
13     // Widget::Widget(int) -> _ZN6WidgetC1Ei, _ZN6WidgetC2Ei
14 Widget::~~Widget() {}
15     // Widget::~~Widget() -> _ZN6WidgetD1Ev, _ZN6WidgetD2Ev
16 int Widget::grog() const {return 42;}
17     // Widget::grog() const -> _ZNK6Widget4grogEv
18
19 namespace foo {
20 void func() {}
21     // foo::func() -> _ZN3foo4funcEv
22 }
23
24 void squander(const std::vector<int>& v) {}
25     // squander(const std::vector<int>& v) ->
26     // _Z8squanderRKSt6vectorIiSaIiEE
27
28 int main() {
29     std::cout << "Hello, World!\n"; // std::cout -> _ZSt4cout
30 }
```



## ■ C++filt:

- `c++filt` program, which is part of GNU Binary Utilities
- filter that copies character stream from standard input to standard output replacing any mangled names with their unmangled forms
- web site: <https://www.gnu.org/software/binutils>

## ■ LLVM Cxxfilt:

- `llvm-cxxfilt` program, which is part of LLVM software
- filter that copies character stream from standard input to standard output replacing any mangled names with their unmangled forms (in similar fashion as `c++filt`)
- LLVM web site: <https://llvm.org>

## ■ online name demangler:

- <http://demangler.com>

## Section 3.10.2

### **References**

- 1 J. R. Levine. *Linker and Loaders*.  
Morgan Kaufmann, San Francisco, CA, USA, 2000.
- 2 W. von Hagen. *The Definitive Guide to GCC*.  
Apress, 2nd edition, 2006.
- 3 R. M. Stallman and the GCC Developer Community, GNU Compiler Collection Internals, Free Software Foundation, 2018. Available online at <https://gcc.gnu.org/onlinedocs> (direct link: <https://gcc.gnu.org/onlinedocs/gccint.pdf>).

- 1 Dave Watson. C++ Exceptions and Stack Unwinding. CppCon, Bellevue, WA, USA, Sept. 24–29, 2017. Available online at [https://youtu.be/\\_Ivd3qzgT7U](https://youtu.be/_Ivd3qzgT7U).
- 2 Peter Edwards. C++ Exception Handling: The Gory Details of an Implementation. Dublin C/C++ Meetup, Feb. 2018. Available online at <https://youtu.be/XpRL7exdFL8>.
- 3 Matt Godbolt. The Bits Between the Bits: How We Get to main(). CppCon, Bellevue, WA, USA, Sept. 28, 2018. Available online at <https://youtu.be/dOfucXtyEsU>.
- 4 Michael Spencer. My Little Object File: How Linkers Implement C++. CppCon, Bellevue, WA, USA, Sept. 29, 2017. Available online at <https://youtu.be/a5L66zguFe4>.

## Part 4

### Even More C++

## Section 4.1

### Undefined Behavior and Other Evil Stuff

# Undefined, Unspecified, and Implementation-Defined Behavior

- **undefined behavior**: behavior for which standard imposes no requirements (i.e., anything could happen) [\[C++17 3.27\]](#)
- **unspecified behavior**: behavior, for a well-formed program construct and correct data, that depends on the implementation; implementation is not required to document which behavior occurs; range of possible behaviors usually specified in standard [\[C++17 3.28\]](#)
- **implementation-defined behavior**: behavior, for a well-formed program construct and correct data, that depends on the implementation and that each implementation documents (i.e., only know what will happen for a particular implementation) [\[C++17 3.12\]](#)
- *always avoid undefined behavior* and *do not rely on unspecified behavior*; otherwise cannot guarantee correct behavior of program
- *try to avoid relying on implementation-defined behavior*; otherwise cannot guarantee correct behavior of program across all language implementations (i.e., code will not be portable)

# Examples of Undefined Behavior

- dereferencing a null pointer; for example:

```
char* p = nullptr;  
char c = *p; // undefined behavior
```

- attempting to modify a string literal or any other const object (excluding mutable data members):

```
const int x = 0;  
const_cast<int&>(x) = 42; // undefined behavior
```

- signed integer overflow

- evaluating an expression that is not mathematically defined; for example:

```
double z = 0.0;  
double x = 1.0 / z; // undefined behavior
```

- not returning a value from a value-returning function (other than `main`)

```
int get_value() {  
    // undefined behavior  
}
```

- multiple definitions of the same entity



# Examples of Undefined Behavior (Continued)

- performing pointer arithmetic that yields a result before start of or after end (i.e., one past last element) of an array; for example:

```
int v[10];  
int* p = &v[0];  
--p; // undefined behavior
```

- using pointers to objects whose lifetime has ended
- left-shifting values by a negative amount; for example:

```
int i = 1;  
i <<= (-3); // undefined behavior
```

- shifting values by an amount greater than or equal to the number of bits in the number; for example:

```
int i = 42;  
i >>= 10000; // undefined behavior
```

- using an automatic variable whose value has not been initialized; for example:

```
void func() {  
    int i;  
    ++i; // undefined behavior  
}
```

# Examples of Unspecified Behavior

- order in which arguments to a function are evaluated; for example:

```
1  #include <iostream>
2
3  int count() {
4      static int c = 0;
5      return c++;
6  }
7
8  void func(int x, int y) {
9      std::cout << x << ' ' << y << '\n';
10 }
11
12 int main() {
13     func(count(), count());
14     // what values are passed to func?
15     // 0, 1; or 1, 0?
16 }
```

# Examples of Implementation-Defined Behavior

- meaning of **#pragma** directive
- nesting limit for **#include** directives
- search locations for `" "` and `<>` headers
- sequence of places searched for header
- signedness of `char`
- `sizeof` built-in types other than **`char`**, **`signed char`**, **`unsigned char`**
- type of `size_t`, `ptrdiff_t`
- parameters to `main` function
- alignment (i.e., restrictions on the addresses at which an object of a particular type can be placed)
- result of right shift of negative value
- precise types used in various parts of C++ standard library (e.g., actual type named by `vector<T>::iterator` [\[C++17 26.3.11.1/2\]](#))
- meaning of **`asm`** declaration
- for more examples, see “Index of implementation-defined behavior” section in C++11 standard

# Private Member Access Without Friends (Legal But Evil)

```
1  #include <iostream>
2
3  template <typename Tag>
4  typename Tag::type saved_private_v;
5
6  template <typename Tag, typename Tag::type x>
7  bool save_private_v = (saved_private_v<Tag> = x);
8
9  class Widget {
10 public:
11     Widget(int i) : i_(i) {}
12 private:
13     int i_;
14     int f_() const {return i_;}
15 };
16
17 struct Widget_i_ {using type = int Widget::*;};
18 struct Widget_f_ {using type = int (Widget::*)() const;};
19
20 template bool save_private_v<Widget_i_, &Widget::i_>;
21 template bool save_private_v<Widget_f_, &Widget::f_>;
22
23 int main() {
24     Widget w(42);
25     std::cout << w.*saved_private_v<Widget_i_> << '\n';
26     std::cout << (w.*saved_private_v<Widget_f_>()) << '\n';
27 }
```

## Section 4.2

### **C++ Compatibility**

- many changes have been made to C++ language and standard library during evolution of C++ from C++98 to present
- some changes resulted in incompatibilities between different versions of C++ standard
- subsequent slides list some reference material that discusses how C++ standard changed from one version to next
- knowing such changes helps to understand incompatibilities between different versions

- 1 Leor Zolman. An Overview of C++11/14. CppCon, Bellevue, WA, USA, Sept 8, 2014. (This talk is in two parts.) Available online at <https://youtu.be/Gycxew-hztI> and <https://youtu.be/pBI0tS2yfjw>.
- 2 Alisdair Meredith. A Quick Tour of C++14. CppCon, Bellevue, WA, USA, Sept. 11, 2014. Available online at [https://youtu.be/fBU1R7jp\\_TE](https://youtu.be/fBU1R7jp_TE).
- 3 Alisdair Meredith. C++17 in Breadth. CppCon, Bellevue, WA, USA, Sept. 19, 2016. (This talk is in two parts.) Available online at <https://youtu.be/22jIHfvelZk> and <https://youtu.be/-rIixnNJM4k>.
- 4 Nicolai Josuttis. C++17: The Language Features. Norwegian Developers Conference, London, UK, Jan. 16–20 2017. Available online at <https://youtu.be/pEzV32yRu4U>.

- 5 Nicolai Josuttis. C++17: The Library Features. Norwegian Developers Conference, London, UK, Jan. 16–20 2017. Available online at <https://youtu.be/ELwTKHiKZS4>.
- 6 Bryce Lebach. C++17 Features. C++Now, Aspen, CO, USA, May 16, 2017. Available online at <https://youtu.be/LvwXJjRQfHk>.



## Section 4.3

### **C Compatibility**

- Although C++ attempted to maintain compatibility with C where possible, numerous incompatibilities between the languages exist.
- Unfortunately, as C++ and C continue to evolve, the number of incompatibilities between these languages continues to grow.
- In practice, many C programs are valid C++ programs and can therefore be compiled with a C++ compiler.
- Some C programs, however, may require a significant number of changes to be valid C++.
- Moreover, in some cases, the semantics of the source code may change, depending on whether it is interpreted as C++ or C.
- A few examples of incompatibilities between C++ and C are given on the subsequent slides.
- More detailed information on incompatibilities between C++ and C can be found at:
  - <http://david.tribble.com/text/cdiffs.htm>
  - [https://en.wikipedia.org/wiki/Compatibility\\_of\\_C\\_and\\_C++](https://en.wikipedia.org/wiki/Compatibility_of_C_and_C++)

# Conflicts with New Keywords

```
1  #include <stdio.h>
2  #include <unistd.h>
3
4  /* Delete a file. */
5  int delete(const char* filename) { /* note function name */
6      return unlink(filename);
7  }
8
9  int main(int argc, char** argv) {
10     if (argc >= 2) {
11         if (delete(argv[1])) {
12             printf("cannot delete file\n");
13             return 1;
14         }
15     }
16     return 0;
17 }
```

- 
- C++ introduces many new keywords.
  - Some C programs might use some of these keywords as identifiers (e.g., `new`, `delete`).

# Function Declarations Without Arguments

```
1  #include <stdio.h>
2
3  int plusOne(); /* no arguments specified */
4
5  int main(int argc, char** argv) {
6      printf("%d\n", plusOne(0));
7      return 0;
8  }
9
10 int plusOne(int i) {
11     return i + 1;
12 }
```

- 
- In C, a function declaration without arguments implies that the arguments are unspecified.
  - In C++, a function declaration without arguments implies that the function takes no arguments.

# Implicit Return Type

```
1  #include <stdio.h>
2
3  myfunc() { /* implicit return type */
4      return 3;
5  }
6
7  int main(int argc, char **argv) {
8      int i;
9      i = myfunc();
10     printf("%d\n", i);
11     return 0;
12 }
```

- 
- In C, if the return type of a function is not specified, it is treated as **int**.
  - In C++, the return type of a function must always be explicitly specified.

# More Restrictive Conversions Involving `void*`

```
1  int main(int argc, char** argv) {  
2      int i;  
3      int* ip;  
4      void* vp;  
5      ip = &i;  
6      vp = ip;  
7      ip = vp; /* problematic */  
8      return 0;  
9  }
```

- 
- C provides an implicit conversion from `void*` to any pointer type, while C++ does not.

# Scoping Rules for Nested Structs

```
1  struct outer {
2      struct inner {
3          int i;
4      };
5      int j;
6  };
7
8  struct inner a = {1}; /* inner vs. outer::inner */
9
10 int main(int argc, char** argv) {
11     return 0;
12 }
```

- 
- C and C++ both allow nested **struct** types, but the scoping rules differ.

# Type of Character Literal

```
1  #include <stdio.h>
2
3  int main() {
4      printf("%d\n", sizeof('a'));
5  }
```

- 
- A character literal (such as 'A') is of type **char** in C++, but type **int** in C.
  - Consequently, the above program will print a value of 1 when compiled as C++ and a value greater than 1 (namely, the value of **sizeof(int)**) when compiled as C.
  - Thus, the same source code can have *different semantics*, depending on whether it is interpreted as C++ or C.



## Part 5

# Libraries

## Section 5.1

### **Boost Libraries**

## Section 5.1.1

### **Introduction**

- Boost libraries are collection of free peer-reviewed portable C++ source libraries
- license encourages both commercial and non-commercial use
- often Boost libraries later adopted by C++ standard
- web site: <http://www.boost.org>

# Some Boost Libraries

## Containers and Data Structures

Library	Description
Bimap	bidirectional maps (i.e., associative containers in which both types stored in map can be used as key)
Container	standard library containers and extensions
Heap	priority queue data structures
Intrusive	intrusive containers and algorithms
Multi-Array	generic N-dimensional array
Multi-Index	containers that maintain one or more indices with different sorting and access semantics

## Iterators

Library	Description
Iterator	concepts that extend C++ standard iterator requirements and components for building iterators based on these concepts; includes several iterator adaptors

# Some Boost Libraries (Continued 1)

## Math and Numerics

Library	Description
Interval	interval arithmetic
Math	various numeric types and math functions
Multiprecision	extended precision arithmetic types for floating-point, integer, and rational arithmetic
Rational	rational number class

## String and Text Processing

Library	Description
Lexical Cast	general literal text conversions, such as converting <code>int</code> to <code>std::string</code> or vice versa
Tokenizer	break a string or other character sequence into a series of tokens

# Some Boost Libraries (Continued 2)

## Image and Geometry Processing

Library	Description
Geometry	geometric algorithms, primitives, and spatial index
GIL	generic image library
Graph	graph types and algorithms

## Input/Output

Library	Description
I/O State Savers	classes for saving/restoring state associated with I/O streams

## Miscellaneous

Library	Description
Program Options	process program options via command line or configuration file

## Some Boost Libraries (Continued 3)

### Concurrent Programming

Library	Description
Fiber	userland threads library
Compute	parallel/GPU computing library
Lockfree	lock-free containers (e.g., stacks and queues)



## Section 5.1.2

### **Boost Container Library**

- Boost Container library provides support for numerous *nonintrusive* containers
- containers provided by library include:
  - enhanced versions of several containers from standard library
  - several non-standard containers

## Standard Container Types

Type	Description
<code>vector</code>	similar to <code>std::vector</code>
<code>list</code>	similar to <code>std::list</code>
<code>deque</code>	similar to <code>std::deque</code>
<code>set</code>	similar to <code>std::set</code>
<code>multiset</code>	similar to <code>std::multiset</code>
<code>map</code>	similar to <code>std::map</code>
<code>multimap</code>	similar to <code>std::multimap</code>

# Container Types (Continued)

## Non-Standard Container Types

Type	Description
<code>stable_vector</code>	vector with non-contiguous elements and stable element references
<code>flat_set</code>	set based on sorted vector
<code>flat_multiset</code>	multiset based on sorted vector
<code>flat_map</code>	map based on sorted vector
<code>flat_multimap</code>	multimap based on sorted vector
<code>slist</code>	singly-linked list
<code>static_vector</code>	vector of bounded size with storage for elements that is contiguous and statically allocated
<code>small_vector</code>	vector-like container optimized for case of containing few elements

# flat\_set Example

```
1  #include <iostream>
2  #include <cassert>
3  #include <boost/container/flat_set.hpp>
4
5  int main() {
6      namespace bc = boost::container;
7      bc::flat_set<std::string> c;
8      c.reserve(4);
9      c.insert("hi");
10     c.insert("apple");
11     c.insert("bye");
12     c.insert("foo");
13     for (auto&& i : c) {
14         std::cout << i << '\n';
15     }
16     std::cout << '\n';
17     auto j = c.find("foo");
18     assert(j != c.end() && *j == "foo");
19     c.erase(j);
20     c.shrink_to_fit();
21     for (auto&& i : c) {
22         std::cout << i << '\n';
23     }
24 }
```

## Section 5.1.3

### **Boost Intrusive Library**

- Boost Intrusive library provides support for numerous *intrusive* and *semi-intrusive* containers
- containers provided by library include those based on:
  - linked lists
  - trees
  - hash tables

## Intrusive Container Types

Type	Description
slist	singly-linked list
list	doubly-linked list
set	set/map based on red-black tree
multiset	multiset/multimap based on red-black tree
rbtree	red-black tree
avl_set	set/map based on AVL tree
avl_multiset	multiset/multimap based on AVL tree
avltree	AVL tree
splay_set	set/map based on splay tree
splay_multiset	multiset/multimap based on splay tree
splaytree	splay tree
sg_set	set/map based on scapegoat tree
sg_multiset	multiset/multimap based on scapegoat tree
sgtree	scapegoat tree



# Container Types (Continued)

## Semi-Intrusive Container Types

Type	Description
<code>unordered_set</code>	unordered set/map based on hash table
<code>unordered_multiset</code>	unordered multiset/multimap based on hash table

# Base and Member Hooks

- **hook** is class object that must be added to a user's class in order for user's class to be usable with intrusive container
- hook used to encapsulate data used to manage nodes in container (e.g., successor and predecessor links for doubly-linked list)
- two kinds of hooks:
  - 1 base hook
  - 2 member hook
- base hook is included in user's class as base class object using *public* inheritance
- member hook included in user's class as *public* data member

# slist With Base Hook

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/slist.hpp>
4
5  namespace bi = boost::intrusive;
6
7  struct Widget : public bi::slist_base_hook<> {
8      Widget(int i_) : i(i_) {}
9      int i;
10 };
11
12 using WidgetList = bi::slist<Widget>;
13
14 int main() {
15     std::vector<Widget> buffer;
16     for (int i = 0; i < 10; ++i) {buffer.push_back(Widget(i));}
17     WidgetList a;
18     for (auto&& i : buffer) {a.push_front(i);}
19     for (auto i = a.begin(); i != a.end(); ++i) {
20         if (i != a.begin()) {std::cout << ' ';}
21         std::cout << i->i;
22     }
23     std::cout << '\n';
24     while (!a.empty()) {a.erase_after(a.before_begin());}
25 }
```

# slist With Member Hook

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/slist.hpp>
4
5  namespace bi = boost::intrusive;
6
7  struct Widget {
8      Widget(int i_) : i(i_) {}
9      int i;
10     bi::slist_member_hook<> hook;
11 };
12
13 using WidgetList = bi::slist<Widget, bi::member_hook<Widget,
14     bi::slist_member_hook<>, &Widget::hook>>;
15
16 int main() {
17     std::vector<Widget> buffer;
18     for (int i = 0; i < 10; ++i) {buffer.push_back(Widget(i));}
19     WidgetList a;
20     for (auto&& i : buffer) {a.push_front(i);}
21     for (auto i = a.begin(); i != a.end(); ++i) {
22         if (i != a.begin()) {std::cout << ' ';}
23         std::cout << i->i;
24     }
25     std::cout << '\n';
26     while (!a.empty())
27         {a.erase_after(a.before_begin());}
28 }
```

# list With Base Hook

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/list.hpp>
4
5  namespace bi = boost::intrusive;
6
7  struct Widget : public bi::list_base_hook<> {
8      Widget(int i_) : i(i_) {}
9      int i;
10 };
11
12 using WidgetList = bi::list<Widget>;
13
14 int main() {
15     std::vector<Widget> buffer;
16     for (int i = 0; i < 10; ++i) {buffer.push_back(Widget(i));}
17     WidgetList a;
18     for (auto&& i : buffer) {a.push_back(i);}
19     for (auto i = a.begin(); i != a.end(); ++i) {
20         if (i != a.begin()) {std::cout << ' ';}
21         std::cout << i->i;
22     }
23     std::cout << '\n';
24     while (!a.empty()) {a.erase(a.begin());}
25 }
```

# list With Member Hook

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/list.hpp>
4
5  namespace bi = boost::intrusive;
6
7  struct Widget {
8      Widget(int i_) : i(i_) {}
9      int i;
10     bi::list_member_hook<> hook;
11 };
12
13 using WidgetList = bi::list<Widget, bi::member_hook<Widget,
14     bi::list_member_hook<>, &Widget::hook>>;
15
16 int main() {
17     std::vector<Widget> buffer;
18     for (int i = 0; i < 10; ++i) {buffer.push_back(Widget(i));}
19     WidgetList a;
20     for (auto& i : buffer) {a.push_back(i);}
21     for (auto i = a.begin(); i != a.end(); ++i) {
22         if (i != a.begin()) {std::cout << ' ';}
23         std::cout << i->i;
24     }
25     std::cout << '\n';
26     while (!a.empty()) {a.erase(a.begin());}
27 }
```

# list With Multiple Base Hooks

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/list.hpp>
4
5  namespace bi = boost::intrusive;
6
7  struct Alpha {};
8  struct Beta {};
9  struct Widget : public bi::list_base_hook<bi::tag<Alpha>>,
10     public bi::list_base_hook<bi::tag<Beta>> {
11     Widget(int i_) : i(i_) {}
12     int i;
13 };
14
15 int main() {
16     std::vector<Widget> buffer;
17     for (int i = 0; i < 10; ++i) {buffer.push_back(Widget(i));}
18     bi::list<Widget, bi::base_hook<bi::list_base_hook<bi::tag<Alpha>>>>
19     a;
20     bi::list<Widget, bi::base_hook<bi::list_base_hook<bi::tag<Beta>>>>
21     b;
22     for (auto&& i : buffer)
23         {a.push_back(i); b.push_front(i);}
24     for (auto&& w : a) {std::cout << w.i << '\n';}
25     std::cout << '\n';
26     for (auto&& w : b) {std::cout << w.i << '\n';}
27     while (!a.empty()) {a.erase(a.begin());}
28     while (!b.empty()) {b.erase(b.begin());}
29 }
```

# set With Base Hook

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/set.hpp>
4
5  namespace bi = boost::intrusive;
6
7  struct Widget : public bi::set_base_hook<> {
8      Widget(int i_) : i(i_) {}
9      bool operator<(const Widget& other) const {return i < other.i;}
10     int i;
11 };
12
13 int main() {
14     int values[] = {1, 3, 5, 7, 9, 0, 2, 4, 6, 8};
15     std::vector<Widget> buffer;
16     for (auto i : values) {buffer.push_back(Widget(i));}
17     bi::set<Widget> a;
18     for (auto&& i : buffer) {a.insert(a.end(), i);}
19     for (auto&& w : a) {std::cout << w.i << '\n';}
20     if (a.find(7) != a.end()) {std::cout << "7 is in set\n";}
21     while (!a.empty())
22         {a.erase(a.begin());}
23 }
```



# set and list With Base Hooks

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/set.hpp>
4  #include <boost/intrusive/list.hpp>
5
6  namespace bi = boost::intrusive;
7
8  struct Widget : public bi::set_base_hook<>,
9      public bi::list_base_hook<> {
10      Widget(int i_) : i(i_) {}
11      bool operator<(const Widget& other) const {return i < other.i;}
12      int i;
13  };
14
15  int main() {
16      int values[] = {1, 3, 5, 7, 9, 0, 2, 4, 6, 8};
17      std::vector<Widget> buffer;
18      for (auto i : values) {buffer.push_back(Widget(i));}
19      bi::set<Widget> a;
20      bi::list<Widget> b;
21      for (auto&& i : buffer)
22          {a.insert(a.end(), i); b.push_back(i);}
23      if (a.find(7) != a.end()) {std::cout << "found 7\n\n";}
24      for (auto&& w : a) {std::cout << w.i << '\n';}
25      std::cout << '\n';
26      for (auto&& w : b) {std::cout << w.i << '\n';}
27      while (!a.empty()) {a.erase(a.begin());}
28      while (!b.empty()) {b.erase(b.begin());}
29  }
```

# Achieving Map Functionality With set

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/set.hpp>
4
5  namespace bi = boost::intrusive;
6
7  struct Widget : public bi::set_base_hook<> {
8      Widget(int i_, int j_) : i(i_), j(j_) {}
9      int i;
10     int j;
11 };
12
13 struct Is_key {
14     using type = int;
15     const type& operator() (const Widget& w) const
16     {return w.i;}
17 };
18
19 int main() {
20     int values[] = {1, 3, 5, 7, 9, 0, 2, 4, 6, 8};
21     std::vector<Widget> buffer;
22     for (auto i : values) {buffer.push_back(Widget(i, -i));}
23     bi::set<Widget, bi::key_of_value<Is_key>> a;
24     for (auto&& i : buffer) {a.insert(a.end(), i);}
25     for (auto&& w : a) {std::cout << w.i << ' ' << w.j << '\n';}
26     auto i = a.find(5);
27     std::cout << i->j << '\n';
28     while (!a.empty()) {a.erase(a.begin());}
29 }
```

# unordered\_set With Base Hook

```
1  #include <iostream>
2  #include <vector>
3  #include <boost/intrusive/unordered_set.hpp>
4
5  namespace bi = boost::intrusive;
6
7  class Widget : public bi::unordered_set_base_hook<> {
8  public:
9      Widget(int value = 0) : value_(value) {}
10     int get_value() const {return value_;}
11 private:
12     int value_;
13 };
14
15 bool operator==(const Widget& a, const Widget& b)
16     {return a.get_value() == b.get_value();}
17
18 std::size_t hash_value(const Widget& a) {return std::size_t(a.get_value());}
19
20 int main() {
21     std::vector<Widget> widgets;
22     for (int i = 0; i < 10; ++i) {widgets.push_back(Widget(i));}
23     using bucket_type = bi::unordered_set<Widget>::bucket_type;
24     using bucket_traits = bi::unordered_set<Widget>::bucket_traits;
25     bucket_type buckets[100];
26     bi::unordered_set<Widget> s(bucket_traits(buckets, 100));
27     for (auto&& w : widgets) {s.insert(w);}
28     for (auto&& i : s) {std::cout << i.get_value() << '\n';}
29     if (s.find(7) != s.end()) {std::cout << "found 7\n";}
30     return 0;
31 }
```

## Section 5.1.4

### **Boost Iterator Library**

- Boost iterator library consists of two parts:
  - 1 system of concepts which extend C++ standard iterator requirements
  - 2 framework of components for building iterators based on these concepts
- tricky to write standard-conforming iterators
- by using Boost Iterator library, can often significantly reduce amount of code needed to implement standard-conforming iterators

# Forward Iterator Example: Iterator Class Without Boost (1)

```
1  #include <type_traits>
2  #include <iterator>
3  #include <cstddef>
4
5  // singly-linked list node base (for intrusive container)
6  template <class T> struct slist_node_base {
7      slist_node_base(T* next_) : next(next_) {}
8      T* next; // pointer to next node in list
9  };
10
11 // single-linked list iterator (const and non-const)
12 template <class T> class slist_iter {
13 public:
14     using iterator_category = std::forward_iterator_tag;
15     using value_type = typename std::remove_const_t<T>;
16     using difference_type = std::ptrdiff_t;
17     using reference = T&;
18     using pointer = T*;
19     slist_iter(T* node = nullptr) : node_(node) {}
20     template <class OtherT, class =
21         std::enable_if_t<std::is_convertible_v<OtherT*, T*>>>
22         slist_iter(const slist_iter<OtherT>& other) : node_(other.node_) {}
23     reference operator*() {return *node_;}
24     pointer operator->() {return node_;}
```

[see [SFINAE](#)]

## Forward Iterator Example: Iterator Class Without Boost (2)

```
25  slist_iter& operator++() {
26      node_ = node_->next;
27      return *this;
28  }
29  slist_iter operator++(int) {
30      slist_iter old(*this);
31      node_ = node_->next;
32      return old;
33  }
34  template <class OtherT> bool operator==(const slist_iter<OtherT>& other)
35      const {return node_ == other.node_;}
36  template <class OtherT> bool operator!=(const slist_iter<OtherT>& other)
37      const {return !(*this == other);}
38  private:
39      template <class> friend class slist_iter;
40      T* node_; // pointer to list node
41  };
```

# Forward Iterator Example: Iterator Class With Boost

```
1  #include <type_traits>
2  #include <boost/iterator/iterator_facade.hpp>
3
4  template <class T> struct slist_node_base {
5      slist_node_base(T* next_) : next(next_) {}
6      T* next; // pointer to next node in list
7  };
8
9  template <class T> class slist_iter : public boost::iterator_facade<
10     slist_iter<T>, T, boost::forward_traversal_tag> {
11 public:
12     using base = typename boost::iterator_facade<slist_iter<T>, T,
13         boost::forward_traversal_tag>;
14     using typename base::reference;
15     using typename base::value_type;
16     slist_iter(T* node = nullptr) : node_(node) {}
17     template <class OtherT, class =
18         std::enable_if_t<std::is_convertible_v<OtherT*, T*>>>
19         slist_iter(const slist_iter<OtherT>& other) : node_(other.node_) {}
20 private:
21     reference dereference() const {return *node_;}
22     template <class OtherT> bool equal(const slist_iter<OtherT>& other) const
23         {return node_ == other.node_;}
24     void increment() {node_ = node_->next;}
25     template <class> friend class slist_iter;
26     friend class boost::iterator_core_access;
27     T* node_; // pointer to list node
28 };
```



# Forward Iterator Example: User Code

```
1  #include <iostream>
2  #include <vector>
3  #include "iterator_facade_2.hpp"
4
5  struct Node : public slist_node_base<Node> {
6      Node(Node* next_, int value_) : slist_node_base<Node>(next_),
7          value(value_) {}
8      int value;
9  };
10
11 int main() {
12     constexpr int num_nodes = 10;
13     std::vector<Node> nodes; nodes.reserve(num_nodes);
14     for (int i = 0; i < num_nodes - 1; ++i)
15         {nodes.push_back(Node(&nodes[i + 1], i));}
16     nodes.push_back(Node(nullptr, num_nodes - 1));
17     slist_iter<Node> begin(&nodes[0]);
18     slist_iter<Node> end;
19     slist_iter<const Node> cbegin(begin);
20     slist_iter<const Node> cend(end);
21     for (auto i = cbegin; i != cend; ++i) {std::cout << i->value << '\n';}
22     slist_iter<Node> i(begin);
23     slist_iter<const Node> ci(cbegin);
24     // slist_iter<Node> j(cbegin); // ERROR
25     i = begin;
26     // i = ci; // ERROR
27     ci = cbegin;
28     ci = i;
29 }
```

# Random-Access Iterator Example: Iterator Class Without Boost (1)

```
1  #include <type_traits>
2  #include <iterator>
3
4  // array element iterator
5  template <class T> class array_iter {
6  public:
7      using iterator_category = typename std::random_access_iterator_tag;
8      using value_type = std::remove_const_t<T>;
9      using reference = T&;
10     using pointer = T*;
11     using difference_type = std::ptrdiff_t;
12     array_iter(T* ptr = nullptr) : ptr_(ptr) {}
13     template <class OtherT, class =
14         std::enable_if_t<std::is_convertible_v<OtherT*, T*>>>
15         array_iter(const array_iter<OtherT>& other) : ptr_(other.ptr_) {}
16     reference operator*() const {return *ptr_;}
17     pointer operator->() const {return ptr_;}
18     array_iter& operator++() {
19         ++ptr_;
20         return *this;
21     }
22     array_iter operator++(int) {
23         array_iter old(*this);
24         ++ptr_;
25         return old;
26     }
27     array_iter& operator--() {
28         --ptr_;
29         return *this;
30     }
```

## Random-Access Iterator Example: Iterator Class Without Boost (2)

```
31     array_iter operator--(int) {
32         array_iter old(*this);
33         --ptr_;
34         return old;
35     }
36     array_iter& operator+=(difference_type n) {
37         ptr_ += n;
38         return *this;
39     }
40     array_iter& operator-=(difference_type n) {
41         ptr_ -= n;
42         return *this;
43     }
44     reference operator[](difference_type n) const {return ptr_[n];}
45     array_iter operator+(difference_type n) const
46     {return array_iter(ptr_ + n);}
47     difference_type operator-(const array_iter& other) const
48     {return ptr_ - other.ptr_;}
49     array_iter operator-(difference_type n) const
50     {return array_iter(ptr_ - n);}
51     template <class OtherT> bool operator==(const array_iter<OtherT>& other)
52     const {return ptr_ == other.ptr_;}
53     template <class OtherT> bool operator!=(const array_iter<OtherT>& other)
54     const {return ptr_ != other.ptr_;}
55     template <class OtherT> bool operator<(const array_iter<OtherT>& other)
56     const {return ptr_ < other.ptr_;}
57     template <class OtherT> bool operator>(const array_iter<OtherT>& other)
58     const {return ptr_ > other.ptr_;}
59     template <class OtherT> bool operator<=(const array_iter<OtherT>& other)
60     const {return ptr_ <= other.ptr_;}
```

## Random-Access Iterator Example: Iterator Class Without Boost (3)

```
61     template <class OtherT> bool operator>=(const array_iter<OtherT>& other)
62         const {return ptr_ >= other.ptr_;}
63 private:
64     template <class> friend class array_iter;
65     T* ptr_; // pointer to array element
66 };
67
68 template <class T>
69 array_iter<T> operator+(typename array_iter<T>::difference_type n,
70     const array_iter<T>& iter) {return array_iter<T>(iter) += n;}
```

# Random-Access Iterator Example: Iterator Class With Boost

```
1  #include <boost/iterator/iterator_facade.hpp>
2  #include <type_traits>
3
4  // array element iterator
5  template <class T> class array_iter : public boost::iterator_facade<
6      array_iter<T>, T, boost::random_access_traversal_tag> {
7  public:
8      using typename boost::iterator_facade<array_iter<T>, T,
9          boost::random_access_traversal_tag>::reference;
10     using typename boost::iterator_facade<array_iter<T>, T,
11         boost::random_access_traversal_tag>::difference_type;
12     array_iter(T* ptr = nullptr) : ptr_(ptr) {}
13     template <class OtherT, class =
14         std::enable_if_t<std::is_convertible_v<OtherT*, T*>>>
15     array_iter(const array_iter<OtherT>& other) : ptr_(other.ptr_) {}
16 private:
17     reference dereference() const {return *ptr_;}
18     template <class OtherT> bool equal(const array_iter<OtherT>& other) const
19     {return ptr_ == other.ptr_;}
20     void increment() {++ptr_;}
21     void decrement() {--ptr_;}
22     void advance(difference_type n) {ptr_ += n;}
23     difference_type distance_to(const array_iter& other) const
24     {return other.ptr_ - ptr_;}
25     template <class> friend class array_iter;
26     friend class boost::iterator_core_access;
27     T* ptr_; // pointer to array element
28 }
```

# Random-Access Iterator Example: User Code

```
1  #include <iostream>
2  #include <cassert>
3  #include "iterator_facade_1.hpp"
4
5  int main() {
6      char buffer[] = "Hello, World!\n";
7      std::size_t length = sizeof(buffer) - 1;
8      array_iter<char> begin(buffer);
9      array_iter<char> end(buffer + length);
10     array_iter<const char> cbegin = begin;
11     array_iter<const char> cend = end;
12     assert(begin + length == end);
13     assert(cbegin + length == end);
14     for (auto i = cbegin; i != cend; ++i)
15         {std::cout << *i << '\n';}
16     array_iter<char> i(begin);
17     array_iter<const char> ci(cbegin);
18     // array_iter<char> j(cbegin); // ERROR
19     i = begin;
20     // i = ci; // ERROR
21     ci = cbegin;
22     ci = i;
23 }
```

## Section 5.1.5

### **Miscellaneous Examples**

# Math Constants $\pi$ and $e$

```
1  #include <iostream>
2  #include <limits>
3  #include <boost/math/constants/constants.hpp>
4
5  int main() {
6      namespace bmc = boost::math::constants;
7
8      std::cout.precision(std::numeric_limits<
9          long double>::max_digits10);
10
11     constexpr auto f_pi = bmc::pi<float>();
12     constexpr auto d_pi = bmc::pi<double>();
13     constexpr auto ld_pi = bmc::pi<long double>();
14
15     std::cout << f_pi << '\n';
16     std::cout << d_pi << '\n';
17     std::cout << ld_pi << '\n';
18
19     constexpr auto f_e = bmc::e<float>();
20     constexpr auto d_e = bmc::e<double>();
21     constexpr auto ld_e = bmc::e<long double>();
22
23     std::cout << f_e << '\n';
24     std::cout << d_e << '\n';
25     std::cout << ld_e << '\n';
26 }
```



```
1  #include <iostream>
2  #include <boost/math/constants/constants.hpp>
3
4  template <class Real>
5  Real area_of_circle(Real r) {
6      namespace bmc = boost::math::constants;
7      return bmc::pi<Real>() * r * r;
8  }
9
10 int main() {
11     double r;
12     while (std::cin >> r) {
13         std::cout << area_of_circle(r) << '\n';
14     }
15 }
```

# Computing Factorials With Arbitrary Precision

```
1  #include <cmath>
2  #include <boost/multiprecision/gmp.hpp>
3  #include <iostream>
4
5  using boost::multiprecision::mpz_int;
6
7  mpz_int factorial(const mpz_int& n) {
8      mpz_int result = 1;
9      for (mpz_int i = n; i >= 2; --i) {
10         result *= i;
11     }
12     return result;
13 }
14
15 int main() {
16     std::cout << factorial(200) << '\n';
17 }
18
19 /* Output:
20 788657867364790503552363213932185062295135977687173263294
21 742533244359449963403342920304284011984623904177212138919
22 638830257642790242637105061926624952829931113462857270763
23 317237396988943922445621451664240254033291864131227428294
24 853277524242407573903240321257405579568660226031904170324
25 062351700858796178922222789623703897374720000000000000000
26 0000000000000000000000000000000000000000000000000000000
27 */
```

# multi\_array Example

```
1  #include <boost/multi_array.hpp>
2  #include <iostream>
3  #include <iomanip>
4
5  int main() {
6      using Array2 = boost::multi_array<int, 2>;
7      int num_rows = 5;
8      int num_cols = 7;
9      Array2 a(boost::extents[num_rows][num_cols]);
10     for (int row = 0; row < num_rows; ++row) {
11         for (int col = 0; col < num_cols; ++col) {
12             a[row][col] = num_cols * row + col;
13         }
14     }
15     Array2 b(a);
16     assert(b.shape()[0] == num_rows && b.shape()[1] == num_cols);
17     Array2 c;
18     c.resize(boost::extents[b.shape()[0]][b.shape()[1]]);
19     c = b;
20     for (int row = 0; row < num_rows; ++row) {
21         for (int col = 0; col < num_cols; ++col) {
22             if (col) {std::cout << ' ';}
23             std::cout << std::setw(2) << c[row][col];
24         }
25         std::cout << '\n';
26     }
27 }
```

## 2-D Array Class With multi\_array

```
1  #include <boost/multi_array.hpp>
2  #include <iostream>
3  #include <iomanip>
4
5  template <class T>
6  class array2 {
7  public:
8      using value_type = T;
9      array2(int num_rows = 0, int num_cols = 0) :
10         a_(boost::extents[num_rows][num_cols]) {}
11      array2(const array2& other) : a_(other.a_) {}
12      array2& operator=(const array2& other) {
13         if (this != &other) {
14             a_.resize(boost::extents[other.a_.shape()[0]][
15                 other.a_.shape()[1]]);
16             a_ = other.a_;
17         }
18         return *this;
19     }
20     int num_rows() const {return a_.shape()[0];}
21     int num_cols() const {return a_.shape()[1];}
22     const value_type& operator()(int row, int col) const
23         {return a_[row][col];}
24     value_type& operator()(int row, int col) {return a_[row][col];}
25 private:
26     using array = boost::multi_array<T, 2>;
27     array a_;
28 };
```

## 2-D Array Class With multi\_array (Continued)

```
30 template <class T>
31 std::ostream& operator<<(std::ostream& out, const array2<T>& a) {
32     auto width = out.width();
33     for (int row = 0; row < a.num_rows(); ++row) {
34         for (int col = 0; col < a.num_cols(); ++col) {
35             if (col) {out << ' ';}
36             out << std::setw(width) << a(row, col);
37         }
38         out << '\n';
39     }
40     return out;
41 }
42
43 int main() {
44     array2<int> a(5, 7);
45     for (int row = 0; row < a.num_rows(); ++row) {
46         for (int col = 0; col < a.num_cols(); ++col) {
47             a(row, col) = a.num_cols() * row + col;
48         }
49     }
50     array2<int> b(a);
51     std::cout << "a:\n" << std::setw(2) << a;
52     std::cout << "b:\n" << std::setw(2) << b;
53 }
```

# multi\_array Array Layout Example

```
1  #include <algorithm>
2  #include <iostream>
3  #include <boost/multi_array.hpp>
4
5  template <class T> std::ostream& operator<<(std::ostream& out,
6      const boost::multi_array_ref<T, 2>& x) {
7      const int m = x.shape()[0];
8      const int n = x.shape()[1];
9      for (int i = 0; i < m; ++i) {
10         for (int j = 0; j < n; ++j)
11             {out << ((j != 0) ? " " : "") << x[i][j];}
12         out << '\n';
13     }
14     return out;
15 }
16
17 int main() {
18     constexpr int n = 3;
19     float data[n * n] = {
20         1, 2, 3,
21         4, 5, 6,
22         7, 8, 9
23     };
24     int row_major[] = {1, 0};
25     bool ascending[] = {false, true};
26     // use column-major order
27     boost::multi_array_ref<float, 2> x(data, boost::extents[n][n],
28         boost::general_storage_order<2>(boost::fortran_storage_order()));
29     // use row-major order but with rows in reverse order
30     boost::multi_array_ref<float, 2> y(data, boost::extents[n][n],
31         boost::general_storage_order<2>(row_major, ascending));
32     // use row-major order
33     boost::multi_array<float, 2> z(boost::extents[n][n]);
34     std::copy_n(data, n * n, z.data());
35     std::cout << x << '\n' << y << '\n' << z << '\n';
36 }
```

# Program Options Example

```
1  #include <iostream>
2  #include <string>
3  #include <boost/program_options.hpp>
4
5  int main(int argc, char** argv) {
6      namespace po = boost::program_options;
7      po::options_description desc("Allowed options");
8      desc.add_options()
9          ("help,h", "Print help information.")
10         ("count,c", po::value<int>()->default_value(1), "Specify count.")
11         ("file,f", po::value<std::string>(), "Specify file name.");
12      po::variables_map vm;
13      try {
14          po::store(po::parse_command_line(argc, argv, desc), vm);
15          po::notify(vm);
16      } catch (po::error& e) {
17          std::cerr << "usage:\n" << desc << '\n';
18          return 1;
19      }
20      if (vm.count("help")) {std::cout << desc << "\n"; return 1;}
21      if (vm.count("file")) {
22          std::cout << "file: " << vm["file"].as<std::string>() << '\n';
23      }
24      if (vm.count("count")) {
25          std::cout << "count: " << vm["count"].as<int>() << '\n';
26      }
27      return 0;
28  }
```

# Rational Numbers Example

```
1  #include <iostream>
2  #include <cassert>
3  #include <boost/rational.hpp>
4  #include <exception>
5
6  int main() {
7      using boost::rational;
8      const rational<int> zero;
9      rational<int> three(3);
10     rational<int> ninth(1, 9);
11     rational<int> third(1, 3);
12     auto result = three * ninth;
13     assert(result == third);
14     try {
15         std::cout << three / zero << '\n';
16     } catch (const boost::bad_rational& e) {
17         std::cout << "bad rational " << e.what() << '\n';
18     }
19     // rational<int> x(1.5); // ERROR: no matching call
20     // result = 3.0; // ERROR: no matching call
21     result = 42;
22     assert(result == rational<int>(42));
23     std::cout << result << '\n';
24 }
```



## Section 5.1.6

### **References**

- 1 Boost C++ Libraries Web Site, <http://www.boost.org>.
- 2 Boost Library Incubator Web Site, <http://www.blincubator.com>.
- 3 B. Schaling, The Boost C++ Libraries,  
<http://theboostcpplibraries.com>. [This is an online version of  
Schaling's book on Boost.]

- 1 Boris Schaling. Containers in Boost. C++ Now, 2013. Available online at <https://youtu.be/FM-fUjhoCp0>.
- 2 Boris Schaling. Boost.Graph for Beginners. C++ Now, 2013. Available online at <https://youtu.be/uYvBH7TZlFk>.
- 3 Nat Goodspeed. The Fiber Library. C++ Now, 2016. Available online at <https://youtu.be/gcNphOWuUb0>.
- 4 Kyle Lutz. Boost.Compute: A library for GPU/parallel computing. C++ Now, 2015. Available online at <https://youtu.be/q7oCb1CtTT8>.

## Section 5.2

# Computational Geometry Algorithms Library (CGAL)

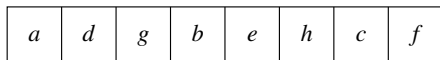
# Computational Geometry Algorithms Library (CGAL)

- very powerful open-source C++ library for geometric computation
- used by many commercial organizations, such as: British Telecom, Boeing, France Telecom, GE Health Care, The MathWorks
- very well documented (extensive manual, more than 4000 pages)
- provides data types for representing various geometric objects, such as:
  - points, lines, planes, polygons
  - Voronoi diagrams
  - 2D, 3D and  $d$ D triangulations
  - polygon meshes
  - kinetic data structures
- provides algorithms for manipulating these data types
- available for Microsoft Windows and Unix/Linux platforms
- some Linux distributions already have packages for CGAL (e.g., Fedora packages: CGAL, CGAL-devel, CGAL-demos-source)
- web site: <http://www.cgal.org>
- online manual (latest version): <http://www.cgal.org/Manual/latest>

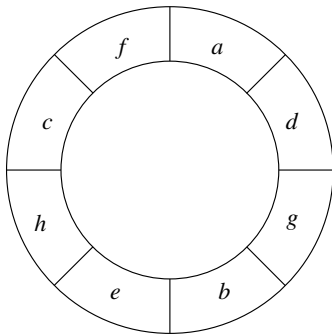
- provides support for polygon meshes
- can read/write polygon mesh data in various common formats
- built-in support for several subdivision schemes
- by using CGAL, can greatly simplify amount of effort required to implement methods using subdivision surfaces or wavelet transforms for polygon meshes
- in CGAL manual, most relevant material is that pertaining to:
  - 2D and 3D linear geometry kernels
  - 3D polyhedral surfaces
  - 3D surface subdivision methods

- **handle**: object used to reference element stored in some data structure (i.e., object can be dereferenced to obtain access to element)
- for data structure storing elements of type  $T$ , handle type might be:
  - simple pointer (i.e.,  $T^*$ )
  - smart pointer (i.e., user-defined type that behaves like pointer)
- examples of handle types:
  - types used to access vertices, facets, halfedges of polygon mesh

# Linear Sequences Versus Circular Sequences



Linear Sequence



Circular Sequence

- linear sequence:
  - has well defined first and last element
  - fits well with iterator model
- circular sequence:
  - does not have well defined first and last element
  - does not fit well with iterator model



- *iterators* are very useful, but intended for use with linear sequences of elements (i.e., sequences with well-defined first and last element)
- often want iterator-like functionality for circular sequences of elements
- **circulator**: object that allows iteration over elements in circular sequence of elements
- examples of circulator types:
  - type to allow iteration over all halfedges incident on vertex in polygon mesh
  - type to allow iteration over all halfedges incident on facet in polygon mesh
- circulators come in const and mutable (i.e., non-const) forms
- mutable circulator can be used to modify referenced element, while const circulator cannot

## Section 5.2.1

### **Geometry Kernels**

# Real Number Types

- **float**: single-precision floating point type
- **double**: double-precision floating point type
- `Interval_nt`: interval-arithmetic type
- `MP_Float`: arbitrary-precision floating-point type

# MP\_Float Class

- `MP_Float` is arbitrary-precision floating-point type
- additions, subtractions, and multiplications computed exactly
- does not provide division or square root (which is not typically problematic as division rarely needed and square root almost always avoided in geometric computation)
- no roundoff error
- no overflow error unless astronomically large numbers involved (arbitrary length mantissa; integral-valued double exponent can overflow, but extremely unlikely)
- very slow, can require considerable memory (unbounded)
- default constructor does not initialize to particular value
- stream inserter (i.e., **`operator<<`**) for `MP_Float` first converts `MP_Float` to **`double`** and then outputs result
- stream extractor (i.e., **`operator>>`**) for `MP_Float` first reads **`double`** and then converts to `MP_Float`

# MP\_Float Example

```
1  #include <iostream>
2  #include <CGAL/MP_Float.h>
3
4  int main() {
5      CGAL::MP_Float x;
6      CGAL::MP_Float y;
7      if (!(std::cin >> x >> y)) {return 1;}
8      if (x < y) {
9          std::cout << x << " is less than " << y << '\n';
10     }
11     CGAL::MP_Float z = -(x + y) * (x - y) + x;
12     std::cout << z << '\n';
13 }
```

# Interval\_nt Class

- declared as: `template <bool M = true> Interval_nt<M>`
- `M` indicates if safe rounding mode enabled
- if safe rounding mode enabled, rounding mode always restored to round towards zero (required by C++); must be careful if safe rounding mode not used
- when safe rounding mode not used, faster but need to worry about things like compiler options like `-frounding-math`
- `using Interval_nt_advanced = Interval_nt<false>;` (i.e., `Interval_nt_advanced` is `Interval_nt` with safe rounding mode disabled)
- interval-arithmetic number type (internally uses floating-point type)
- represents interval  $[a, b]$
- every arithmetic operation performed twice, once while rounding towards  $-\infty$  to produce result  $a'$  and once while rounding towards  $+\infty$  to produce result  $b'$
- true answer must lie on interval  $[a', b']$
- approximately twice of time cost of built-in floating-point type

- represent geometric objects (e.g., point, line, line segment, ray, plane, triangle, circle, )
- points in 2 or 3 dimensions
- provide operations on geometric objects (e.g., intersection, composition)
- allow certain conditions to be tested involving geometric objects (e.g., collinear, coplanar, equality)

# Point Representation

- Cartesian kernels: coordinates represented in Cartesian form
- homogeneous kernels: coordinates represented in homogeneous form



# Simple\_cartesian and Cartesian Classes

- geometry kernel that represents coordinates in Cartesian form

- declaration:

```
template <class F> Simple_cartesian<F>
```

- declaration:

```
template <class F> Cartesian<F>
```

- F field number type (used to represent coordinates)
- F often chosen as **double**
- Cartesian is reference counted version of Simple\_cartesian, which allows more efficient copying of objects
- Cartesian probably preferred if frequent copying occurs

# Simple\_homogeneous and Homogeneous Classes

- geometry kernel that represents coordinates in homogeneous form
- declaration:  

```
template <class R> Simple_homogeneous<R>
```
- declaration:  

```
template <class R> Homogeneous<R>
```
- R ring number type used for representing numerator and denominator of rational coordinates
- Homogeneous is reference counted version of Simple\_homogeneous, which allows more efficient copying of objects
- Homogeneous probably preferred if frequent copying occurs

# Constructions

- produces new geometric object from other objects
- result is not one of a small number of enumerable values
- result is numerical (e.g., involves real numbers)
- create line segment from two points
- create triangle from three points
- create plane from three (non-coplanar) points
- create circle from three (non-collinear) points
- find intersection of line and plane
- exact construction: any newly created geometric objects resulting from construction are exactly represented (i.e., no roundoff/overflow error)
- inexact construction: newly created geometric objects are not guaranteed to be exactly represented (e.g., due to roundoff error)
- extremely important to be aware of whether kernel being used provides exact constructions; affects how you write code!!!

# Predicates

- does not involve any newly computed numerical data
- result is one of very small set of values, such as boolean or enumerated type
- typically used to make decisions (i.e., affect control flow)
- are three points collinear (true or false)
- are four points coplanar (true or false)
- what is position of point relative to oriented line (left of, right of, or on)
- what is position of point relative to oriented circle (inside, outside, or on)
- exact predicate: result of test is guaranteed to be correct (i.e., result determined as if by exact computation)
- inexact predicate: result of test may be incorrect (e.g., due to roundoff/overflow error)
- extremely important to be aware of whether kernel being used provides exact predicates; affects how you write code!!!

# Kernel Member Types: Basic Types

Member Type	Description
FT	field number type (e.g., <b>double</b> )
RT	ring number type (e.g., <b>int</b> )
Boolean	boolean type ( <b>bool</b> or <code>Uncertain&lt;bool&gt;</code> )
Sign	sign ( <code>Sign</code> or <code>Uncertain&lt;Sign&gt;</code> )
Comparison_result	comparison result ( <code>Comparison_result</code> or <code>Uncertain&lt;Comparison_result&gt;</code> )
Orientation	orientation ( <code>Orientation</code> or <code>Uncertain&lt;Orientation&gt;</code> )
Oriented_side	oriented side ( <code>Oriented_side</code> or <code>Uncertain&lt;Oriented_side&gt;</code> )
Bounded_side	bounded side ( <code>Bounded_side</code> or <code>Uncertain&lt;Bounded_side&gt;</code> )
Angle	angle ( <code>Angle</code> or <code>Uncertain&lt;Angle&gt;</code> )

# Kernel Member Types: Geometric Objects in Two Dimensions

Member Type	Description
Point_2	point in two dimensions
Vector_2	vector in two dimensions
Direction_2	direction in two dimensions
Line_2	line in two dimensions
Ray_2	ray in two dimensions
Segment_2	line segment in two dimensions
Triangle_2	triangle in two dimensions
Iso_rectangle_2	axis-aligned rectangle in two dimensions
Circle_2	circle in two dimensions

## Kernel Member Types: Geometric Objects in Three Dimensions

Member Type	Description
Point_3	point in three dimensions
Vector_3	vector in three dimensions
Direction_3	direction in three dimensions
Iso_cuboid_3	axis-aligned cuboid in three dimensions
Line_3	line in three dimensions
Ray_3	ray in three dimensions
Circle_3	circle in three dimensions
Sphere_3	sphere in three dimensions
Segment_3	line segment in three dimensions
Plane_3	plane in three dimensions
Triangle_3	triangle in three dimensions
Tetrahedron_3	tetrahedron in three dimensions

- coordinate representation
- exact or inexact constructions
- exact or inexact predicates
- in practice, almost always require exact predicates
- if code well designed, need for exact constructions can usually be avoided
- for  $T$  chosen as any numeric type that has roundoff/overflow error (e.g., **float**, **double**, **long double**), the following kernels do not provide exact constructions or exact predicates:

```
Simple_cartesian<T>  
Cartesian<T>  
Simple_homogeneous<T>  
Homogeneous<T>
```



# Filtered\_kernel Class

- class to convert kernel with inexact predicates into one with exact predicates
- declared as:  

```
template <class K> Filtered_kernel<K>
```
- K is kernel from which to make filtered kernel
- predicates of K replaced by predicates using numeric type `Interval_nt`
- if interval arithmetic can yield reliable answer, result used
- otherwise, exception thrown and caught by class and predicate using `MP_Float` used
- for exact predicates with `Simple_cartesian<double>`, use:  
`Filtered_kernel<Simple_cartesian<double>>` or equivalently  
`Exact_predicates_inexact_constructions_kernel`
- `Exact_predicates_inexact_constructions_kernel` very commonly used

- exact predicate cannot at any point rely on a computation that is not exact
- no floating point arithmetic (since it has roundoff error)
- no integer arithmetic that might overflow
- no inexact constructions
- no inexact predicates
- `Filtered_predicate` may be helpful

# Filtered\_predicate Class

- adapter for predicate functors for producing efficient exact predicates
- declared as:

```
template <class EP, class FP, class CE, class CF>  
Filtered_predicate<EP, FP, CE, CF>
```

- EP is exact predicate (typically uses arbitrary-precision type such as MP\_Float)
- FP is filtering predicate (typically uses interval-arithmetic type like Interval\_nt)
- CE and CF are function objects for converting arguments of unfiltered predicate to types used by exact and filtering predicates
- must be careful about operation used in unfiltered predicate being plugged into `Filtered_kernel`
- for kernel ring number type RT, can safely use addition, subtraction, multiplication
- can also safely use `sign`

- execution of code for filtered predicate functor proceeds as follows:
  - 1 invoke unfiltered (i.e., original) predicate functor for numeric type  
`CGAL::Interval_nt<false>`  
if any operation on interval arithmetic type yields uncertain result (e.g., `CGAL::sign`), exception is thrown, with thrown exception being caught by filtered predicate functor
  - 2 if no exception thrown (so that unfiltered functor returns normally), return return value of unfiltered functor (and we are done); otherwise, continue
  - 3 invoke unfiltered predicate functor for numeric type `CGAL::MP_Float`
  - 4 return return value of unfiltered functor

# Filtered Predicate Example

```
1  #include <CGAL/Cartesian.h>
2  #include <CGAL/MP_Float.h>
3  #include <CGAL/Interval_nt.h>
4  #include <CGAL/Filtered_predicate.h>
5  #include <CGAL/Cartesian_converter.h>
6
7  template <class K>
8  struct Test_orientation_2 {
9      using RT = typename K::RT;
10     using Point_2 = typename K::Point_2;
11     using result_type = typename K::Orientation;
12     result_type operator()(const Point_2& p, const Point_2& q,
13         const Point_2& r) const {
14         RT prx = p.x() - r.x();
15         RT pry = p.y() - r.y();
16         RT qrx = q.x() - r.x();
17         RT qry = q.y() - r.y();
18         return CGAL::sign(prx * qry - qrx * pry);
19     }
20 };
21
22 using Kernel = CGAL::Cartesian<double>;
23 using Ia_kernel = CGAL::Cartesian<CGAL::Interval_nt<false>>;
24 using Exact_kernel = CGAL::Cartesian<CGAL::MP_Float>;
25 using Test_orientation = CGAL::Filtered_predicate<
26     Test_orientation_2<Exact_kernel>,
27     Test_orientation_2<Ia_kernel>,
28     CGAL::Cartesian_converter<Kernel, Exact_kernel>,
29     CGAL::Cartesian_converter<Kernel, Ia_kernel>
30 >;
31
32 int main() {
33     double big = 1e50;
34     Kernel::Point_2 p(0.0, 0.0), q(1.0, 1.0), r(2.0 * big, 2.0 * big);
35     Test_orientation orientation;
36     std::cout << orientation(p, q, r) << "\n";
37 }
```

## Filtered Predicate Example (Continued)

- for example on previous slide, execution of filtered predicate functor proceeds as follows:
  - 1 invoke  
`Test_orientation_2<Cartesian<CGAL::Interval_nt<false>>>`  
functor with points  $([0,0],[0,0])$ ,  $([1,1],[1,1])$ ,  
 $([2 \cdot 10^{50}, 2 \cdot 10^{50}], [2 \cdot 10^{50}, 2 \cdot 10^{50}])$
  - 2 `CGAL::sign` called for  $[-1.55414 \cdot 10^{85}, 1.55414 \cdot 10^{85}]$ , which results in exception being thrown
  - 3 exception caught by filtered predicate code
  - 4 invoke `Test_orientation_2<Cartesian<CGAL::MP_Float>>` functor with points  $(0,0)$ ,  $(1,1)$ ,  $(2 \cdot 10^{50}, 2 \cdot 10^{50})$
  - 5 `CGAL::sign` called for 0, resulting in return value of 0
  - 6 filtered predicate returns 0
- critically important that `RT` used for all arithmetic operations and not **double** (or **float**); otherwise, arithmetic computation done using wrong numeric type, which will prevent predicate from being correct (i.e., exact)

## Section 5.2.2

# **Polygon Meshes**

# Polyhedron\_3 Class

- represents polyhedral surface (i.e., polygon mesh), which consists of vertices, edges, and facets and incidence relationship amongst them
- each edge represented by pair of halfedges
- declaration for `Polyhedron_3` class:

```
template <class Kernel,  
         class PolyhedronItems = CGAL::Polyhedron_items_3,  
         template <class T, class I>  
         class HalfedgeDS = CGAL::HalfedgeDS_default,  
         class Alloc = CGAL_ALLOCATOR(int)>  
class Polyhedron_3;
```

- `Kernel` is geometry kernel, which specifies such things as how points are represented and provides basic geometric operations/predicates (e.g., `CGAL::Cartesian<double>` and `CGAL::Filtered_kernel<CGAL::Cartesian<double>>`)
- `PolyhedronItems` specifies data types for representing vertices and facets (in many cases, default will suffice)
- `HalfedgeDS` specifies halfedge data structure for representing polygon mesh and `Alloc` specifies allocator (defaults should suffice)



## Basic Types

Type	Description
Vertex	vertex type
Halfedge	halfedge type
Facet	facet type
Point_3	point type (for vertices)

## Handles

Type	Description
Vertex_const_handle	const handle to vertex
Vertex_handle	handle to vertex
Halfedge_const_handle	const handle to halfedge
Halfedge_handle	handle to halfedge
Facet_const_handle	const handle to facet
Facet_handle	handle to facet

## Iterators

Type	Description
<code>Vertex_const_iterator</code>	const iterator over all vertices
<code>Vertex_iterator</code>	iterator over all vertices
<code>Halfedge_const_iterator</code>	const iterator over all halfedges
<code>Halfedge_iterator</code>	iterator over all halfedges
<code>Facet_const_iterator</code>	const iterator over all facets
<code>Facet_iterator</code>	iterator over all facets
<code>Edge_const_iterator</code>	const iterator over all edges (every other halfedge)
<code>Edge_iterator</code>	iterator over all edges (every other halfedge)

## Circulators

Type	Description
<code>Halfedge_around_vertex_const_circulator</code>	const circulator of halfedges around vertex (CW)
<code>Halfedge_around_vertex_circulator</code>	circulator of halfedges around vertex (CW)
<code>Halfedge_around_facet_const_circulator</code>	const circulator of halfedges around facet (CCW)
<code>Halfedge_around_facet_circulator</code>	circulator of halfedges around facet (CCW)

# Polyhedron\_3 Function Members

## Size

Name	Description
size_of_vertices	get number of vertices
size_of_halfedges	get number of halfedges
size_of_facets	get number of facets

## Iterators

Name	Description
vertices_begin	iterator for first vertex in mesh
vertices_end	past-the-end vertex iterator
halfedges_begin	iterator for first halfedge in mesh
halfedges_end	past-the-end halfedge iterator
facets_begin	iterator for first facet in mesh
facets_end	past-the-end facet iterator
edges_begin	iterator for first edge in mesh
edges_end	past-the-end edge iterator

## Combinatorial Predicates

Name	Description
<code>is_closed</code>	true if no border edges (no boundary)
<code>is_pure_triangle</code>	true if all facets are triangles
<code>is_pure_quad</code>	true if all facets are quadrilaterals

## Border Halfedges

Name	Description
<code>normalized_border_is_valid</code>	true if border is normalized
<code>normalize_border</code>	sort halfedges such that non-border edges precede border edges (i.e., normalize border)
<code>size_of_border_halfedges</code>	get number of border halfedges (border must be normalized)
<code>size_of_border_edges</code>	get number of border edges (border must be normalized)
<code>border_halfedges_begin</code>	halfedge iterator starting with border edges (border must be normalized)
<code>border_edges_begin</code>	edge iterator starting with border edges (border must be normalized)

- `Facet` type represents facet (i.e., face) in polyhedral surface
- actual class type to which `Facet` corresponds depends on choice of `PolyhedronItems` template parameter for `Polyhedron_3` class
- depending on actual class type to which `Facet` refers, level of functionality offered by `Facet` class may differ (e.g., available function members may differ)
- `Facet` class may contain following optional information:
  - plane equation (corresponding to plane containing facet)
  - handle for halfedge that is incident on facet
- some member functions in `Facet` class provide access to halfedge-around-facet circulator
- halfedge-around-facet circulator may be either forward or bidirectional

## Operations Available If Facet Plane Supported

Name	Description
plane	get plane equation

## Operations Available If Facet Halfedge Supported

Name	Description
halfedge	get halfedge incident on facet
facet_begin	get circulator of halfedges around facet (CCW)
set_halfedge	set incident halfedge
facet_degree	get degree of facet (i.e., number of edges on boundary of facet)
is_triangle	true if facet is triangle
is_quad	true if facet is quadrilateral



- `Vertex` type represents vertex in polyhedral surface
- actual class type to which `Vertex` corresponds depends on choice of `PolyhedronItems` template parameter for `Polyhedron_3` class
- depending on actual class type to which `Vertex` refers, level of functionality offered by `Vertex` class may differ (e.g., available function members may differ)
- `Vertex` class may contain following optional information:
  - point (corresponding to vertex position)
  - handle for halfedge that is incident on vertex
- some member functions in `Vertex` class provide access to halfedge-around-vertex circulator
- halfedge-around-vertex circulator may be either forward or bidirectional

# Vertex Function Members

## Operations Available If Vertex Point Supported

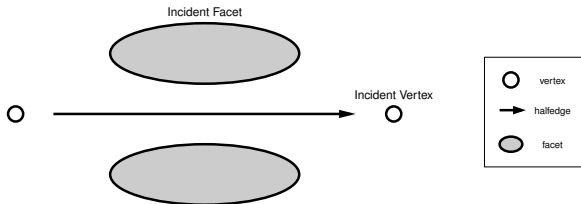
Name	Description
<code>point</code>	get point associated with vertex

## Operations Available If Vertex Halfedge Supported

Name	Description
<code>halfedge</code>	get halfedge incident on vertex
<code>vertex_begin</code>	circulator of halfedges around vertex (CW)
<code>set_halfedge</code>	set incident halfedge
<code>vertex_degree</code>	get valence of vertex
<code>is_bivalent</code>	true if vertex has valence two
<code>is_trivalent</code>	true if vertex has valence three

# Polyhedron\_3::Halfedge

- `Halfedge` type represents halfedge in polyhedral surface
- actual class type to which `Halfedge` corresponds depends on choice of `PolyhedronItems` template parameter for `Polyhedron_3` class
- depending on actual class type to which `Halfedge` refers, level of functionality offered by `Halfedge` class may differ (e.g., available function members may differ)
- each halfedge directly associated with one vertex and one facet, referred to as incident vertex and incident facet, respectively
- **incident vertex** is vertex at *terminal end* of halfedge
- **incident facet** is facet to *left* of halfedge



- halfedge contains:
  - handle for next halfedge around incident facet in CCW direction
  - handle for opposite halfedge
- together, these two handles allow for efficient iteration around:
  - halfedges incident on facet in CCW direction *only*; and
  - halfedges incident on vertex in CW direction *only*
- halfedge may optionally contain:
  - handle for *previous* halfedge around incident facet in CCW direction
- addition of this optional handle allows for efficient iteration around:
  - halfedges incident on facet in *both* (CW and CCW) directions; and
  - halfedges incident on vertex in *both* (CW and CCW) directions
- halfedge may also contain following optional information:
  - handle for incident vertex
  - handle for incident facet
- if halfedge class provides `prev` member function, halfedge-around-vertex and halfedge-around-facet circulators are bidirectional; otherwise, they are forward only

# Halfedge Function Members

## Adjacency Queries

Name	Description
opposite	get opposite halfedge
next	get next halfedge incident on same facet in CCW order
prev	get previous halfedge incident on same facet in CCW order
next_on_vertex	get next halfedge incident on same vertex in CW order
prev_on_vertex	get previous halfedge incident on same vertex in CW order

## Circulators

Name	Description
vertex_begin	get halfedge-around-vertex circulator for incident vertex (CW order)
facet_begin	get halfedge-around-facet circulator for incident facet (CCW order)

# Halfedge Function Members (Continued 1)

## Border Queries

Name	Description
<code>is_border</code>	true if border halfedge
<code>is_border_edge</code>	true if associated edge on border

## Vertex Valence Queries

Name	Description
<code>vertex_degree</code>	get valence of incident vertex
<code>is_bivalent</code>	true if incident vertex has valence two
<code>is_trivalent</code>	true if incident vertex has valence three

## Facet Degree Queries

Name	Description
<code>facet_degree</code>	get degree of incident facet
<code>is_triangle</code>	true if incident facet is triangle
<code>is_quad</code>	true if incident facet is quadrilateral

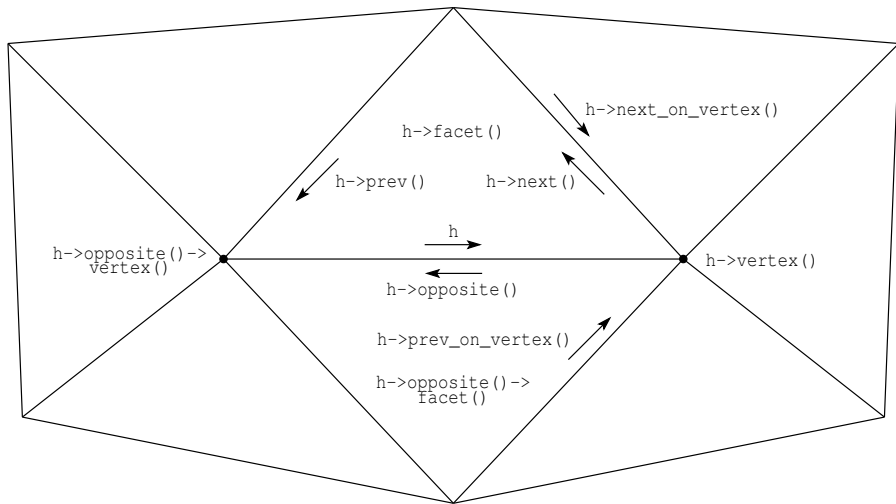
## Operations Available If Halfedge Vertex Supported

Name	Description
vertex	get handle for incident vertex of halfedge

## Operations Available If Halfedge Facet Supported

Name	Description
facet	get handle for incident facet of halfedge

# Adjacency Example





- **operator<<** and **operator>>** are overloaded for I/O
- read and write polygon mesh data in OFF format

- be mindful of operations on `Polyhedron_3` that may invalidate handles, iterators, or circulators
- halfedge-around-vertex circulators and halfedge-around-facet circulators iterate in opposite directions (i.e., CCW versus CW)
- be careful about const correctness (e.g., const versus mutable handles/iterators/circulators)
- some `Polyhedron_3` operations only valid if border normalized (e.g., `size_of_border_halfedges`, `size_of_border_edges`)
- exactly *one* of two halfedges associated with border edge is border halfedge

## Section 5.2.3

### **Surface Subdivision Methods**

- several functions provided for performing subdivision of polygon meshes (represented by `Polyhedron_3`)
- generic subdivision functions apply specific topologic refinement rule but allow arbitrary geometric refinement rule
- specific subdivision functions apply specific subdivision method
- contained in `CGAL::Subdivision_method_3` namespace

# Subdivision Functions

## Generic Subdivision Methods

Function	Description
PQQ	perform primal quadrilateral quadrisection with arbitrary geometric refinement rule
PTQ	perform primal triangle quadrisection with arbitrary geometric refinement rule
DQQ	perform dual quadrilateral quadrisection with arbitrary geometric refinement rule
Sqrt3	perform $\sqrt{3}$ topologic refinement with arbitrary geometric refinement rule

## Specific Subdivision Methods

Function	Description
CatmullClark_subdivision	perform Catmull-Clark subdivision
Loop_subdivision	perform Loop subdivision
DooSabin_subdivision	perform Doo-Sabin subdivision
Sqrt3_subdivision	perform Kobbelt $\sqrt{3}$ subdivision

## Section 5.2.4

### **Example Programs**

- This program generates a simple triangle mesh corresponding to a tetrahedron.
- First, a polygon mesh corresponding to a tetrahedron is constructed.
- Then, the resulting mesh is written to standard output in Object File Format (OFF).

- This program extracts some basic information from a polygon mesh.
- First, a polygon mesh is read from standard input in Object File Format (OFF).
- Then, various information is extracted from the mesh, including:
  - the type of mesh (e.g., triangle, quadrilateral, or general)
  - the number of vertices, edges, faces, and halfedges in the mesh
  - the minimum, maximum, and average valence of vertices in the mesh
  - the number of nonplanar faces in the mesh
- The above information is printed to standard output.



- This program performs subdivision on a polygon mesh.
- First, a mesh is read from standard input in Object File Format (OFF).
- Next, the mesh is refined using the given number of iterations of the specified subdivision method.
- Finally, the refined mesh is written to standard output in OFF.
- Several subdivision schemes are supported, including: Loop, Catmull-Clark, Doo-Sabin, and Kobbelt  $\sqrt{3}$ .

## Section 5.3

# OpenGL Utility Toolkit (GLUT)

# OpenGL Utility Toolkit (GLUT)

- simple windowing API for OpenGL
- intended to be used with small to medium sized OpenGL programs
- language binding for C
- window-system independent
- supports most mainstream operating systems (Microsoft Windows, Linux/Unix)
- provides window management functionality (e.g., creating/destroying windows, displaying/resizing windows, and querying/setting window attributes)
- allows for user input (e.g., via keyboard, mouse)
- routines for drawing common wireframe/solid 3-D objects such as sphere, torus, and well-known teapot model
- register callback functions to handle various types of events (e.g., display, resize, keyboard, special keyboard, mouse, timer, idle) and then loop processing events
- open-source implementation of GLUT called Freeglut is available from <http://sourceforge.net/projects/freeglut>

- event-driven model: flow of program determined by events (e.g., mouse clicks, key presses)
- application making use of event-driven model performs some initialization and then enters an event-processing loop for duration of execution
- each iteration of event-processing loop does following:
  - 1 wait for event
  - 2 process event
- many libraries for building graphical user interfaces (GUIs) employ event-driven model
- GLUT uses event-driven model

# Structure of GLUT Application

- 1 initialize GLUT library by calling `glutInit`
- 2 set display mode (via `glutInitDisplay`)
- 3 perform any additional initialization such as:
  - create windows (via `glutCreateWindow`)
  - register callback functions for handling various types of events (e.g., via `glutDisplayFunc`, `glutReshapeFunc`, `glutKeyboardFunc`)
  - setup initial OpenGL state (e.g., depth buffering, shading, lighting, clear color)
- 4 enter main event-processing loop by calling `glutMainLoop` [Note that `glutMainLoop` never returns.]

- OpenGL and GLUT header files in GL (or GLUT) directory
- to use GLUT, need to include `glut.h` in GL (or GLUT) directory
- header file `glut.h` also includes all necessary OpenGL header files (e.g., `gl.h`, `glu.h`, `glext.h`)

# Event Types

Event Type	Description
<i>display</i>	window contents needs to be displayed
overlay display	overlay plane contents needs to be displayed
<i>reshape</i>	window has been resized
<i>keyboard</i>	key has been pressed
<i>mouse</i>	mouse button has been pressed or released
motion	mouse moved within window while one or more buttons pressed
passive motion	mouse moved within window while no buttons pressed
visibility	visibility of window has changed (covered versus uncovered)
entry	mouse has left or entered window
<i>special keyboard</i>	special key has been pressed (e.g., arrow keys, function keys)

## Event Types (Continued)

Event Type	Description
spaceball motion	spaceball translation has occurred
spaceball rotate	spaceball rotation has occurred
spaceball button	spaceball button has been pressed or released
button box	button box activity has occurred
dials	dial activity has occurred
tablet motion	tablet motion has occurred
tablet button	table button has been pressed or released
menu status	menu status change
<i>idle</i>	no event activity has occurred
<i>timer</i>	timer has expired



# Functions

## Initialization

Function	Description
glutInit	initialize GLUT library
glutInitWindowSize	set initial window size for glutCreateWindow
glutInitWindowPosition	set initial window position for glutCreateWindow
glutInitDisplayMode	set initial display mode

## Beginning Event Processing

Function	Description
glutMainLoop	enter GLUT event-processing loop

# Functions (Continued 1)

## Window Management

Function	Description
<code>glutCreateWindow</code>	create top-level window
<code>glutCreateSubWindow</code>	create subwindow
<code>glutSetWindow</code>	set current window
<code>glutGetWindow</code>	get current window
<code>glutDestroyWindow</code>	destroys specified window
<code>glutPostRedisplay</code>	mark current window as needing to be redisplayed
<code>glutSwapBuffers</code>	swaps buffers of current window if double buffered (flushes graphics output via <code>glFlush</code> )
<code>glutPositionWindow</code>	request change to position of current window
<code>glutReshapeWindow</code>	request change to size of current window
<code>glutFullScreen</code>	request current window to be made full screen
<code>glutSetWindowTitle</code>	set title of current top-level window
<code>glutSetIconTitle</code>	set title of icon for current top-level window
<code>glutSetCursor</code>	set cursor image for current window

## Functions (Continued 2)

### Menu Management

Function	Description
<code>glutCreateMenu</code>	create new pop-up menu
<code>glutSetMenu</code>	set current menu
<code>glutGetMenu</code>	get current menu
<code>glutDestroyMenu</code>	destroy specified menu
<code>glutAddMenuEntry</code>	add menu entry to bottom of current menu
<code>glutAddSubMenu</code>	add submenu trigger to bottom of current menu
<code>glutChangeToMenuEntry</code>	change specified menu item in current menu into menu entry
<code>glutChangeToSubMenu</code>	change specified menu item in current menu into submenu trigger
<code>glutRemoveMenuItem</code>	remove specified menu item
<code>glutAttachMenu</code>	attach mouse button for current window to current menu
<code>glutDetachMenu</code>	detach attached mouse button from current window

# Functions (Continued 3)

## Callback Registration

Function	Description
<code>glutDisplayFunc</code>	sets display callback for current window
<code>glutReshapeFunc</code>	sets reshape callback for current window
<code>glutKeyboardFunc</code>	sets keyboard callback for current window
<code>glutMouseFunc</code>	sets mouse callback for current window
<code>glutMotionFunc</code>	set motion callback for current window
<code>glutPassiveMotionFunc</code>	set passive motion callback for current window
<code>glutVisibilityFunc</code>	set visibility callback for current window
<code>glutEntryFunc</code>	set mouse enter/leave callback for current window
<code>glutSpecialFunc</code>	sets special keyboard callback for current window
<code>glutIdleFunc</code>	set global idle callback
<code>glutTimerFunc</code>	registers timer callback to be triggered in specified number of milliseconds

## Functions (Continued 4)

### State Retrieval

Function	Description
<code>glutGet</code>	retrieves simple GLUT state (e.g., size or position of current window)
<code>glutDeviceGet</code>	retrieves GLUT device information (e.g., keyboard, mouse, spaceball, tablet)
<code>glutGetModifiers</code>	retrieve modifier key state when certain callbacks generated (i.e., state of shift, control, and alt keys)

### Font Rendering

Function	Description
<code>glutBitmapCharacter</code>	renders bitmap character using OpenGL
<code>glutBitmapWidth</code>	get width of bitmap character
<code>glutStrokeCharacter</code>	renders stroke character using OpenGL
<code>glutStrokeWidth</code>	get width of stroke character

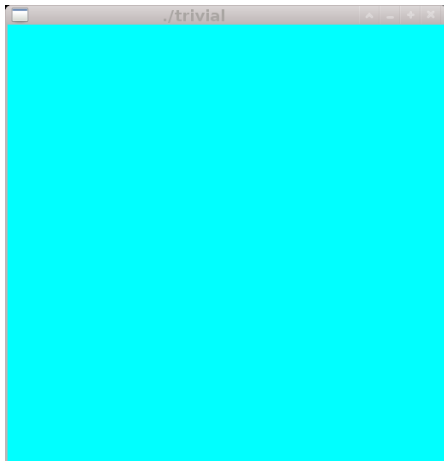
# Functions (Continued 5)

## Geometric Object Rendering

Function	Description
glutSolidSphere	render solid sphere
glutWireSphere	render wireframe sphere
glutSolidCube	render solid cube
glutWireCube	render wireframe cube
glutSolidCone	render solid cone
glutWireCone	render wireframe cone
glutSolidTorus	render solid torus
glutWireTorus	render wireframe torus
glutSolidOctahedron	render solid octahedron
glutWireOctahedron	render wireframe octahedron
glutSolidTetrahedron	render solid tetrahedron
glutWireTetrahedron	render wireframe tetrahedron
glutSolidTeapot	render solid teapot
glutWireTeapot	render wireframe teapot

# Minimalist GLUT Program

- minimalist program using GLUT
- create window that is cleared to particular color



# Minimalist GLUT Program: Source Code

```
1  // Create a window that is cleared to a particular color
2  // when drawn.
3
4  #include <GL/glut.h>
5
6  void display() {
7      glClearColor(0.0, 1.0, 1.0, 0.0);
8      glClear(GL_COLOR_BUFFER_BIT | GL_DEPTH_BUFFER_BIT);
9      glutSwapBuffers();
10 }
11
12 int main(int argc, char** argv) {
13     glutInit(&argc, argv);
14     glutInitDisplayMode(GLUT_DOUBLE | GLUT_RGB);
15     glutInitWindowSize(512, 512);
16     glutCreateWindow(argv[0]);
17     glutDisplayFunc(display);
18     glutMainLoop();
19     return 0;
20 }
```



- 1 M. J. Kilgard. *The OpenGL Utility Toolkit (GLUT): Programming Interface (API Version 3)*, Nov. 1996.  
Available from <http://www.opengl.org/resources/libraries/glut/glut-3.spec.pdf>.
- 2 R. S. Wright, B. Lipchak, and N. Haemel. *OpenGL SuperBible*. Addison-Wesley, Upper Saddle River, NJ, USA, 4th edition, 2007.
- 3 GLUT home page:  
<http://www.opengl.org/resources/libraries/glut>
- 4 GLUT manual (HTML format):  
<http://www.opengl.org/resources/libraries/glut/spec3/spec3.html>

## Section 5.4

### OpenGL Framework (GLFW) Library

# OpenGL Framework (GLFW) Library

- lightweight open-source windowing API for OpenGL, OpenGL ES, and Vulkan
- language binding for C
- window-system independent
- supports most mainstream operating systems (e.g., Microsoft Windows, OS X, and Linux/Unix)
- provides window management functionality (e.g., creating/destroying windows, displaying/resizing windows, and querying/setting window attributes)
- allows for user input (e.g., via keyboard, mouse, and joystick)
- allows application to register callback functions to handle various types of events (e.g., window refresh, window resize, keyboard, and mouse) and then loop processing events
- web site: <http://www.glfw.org>

- GLFW and modern GLUT (e.g., FreeGLUT) offer somewhat similar functionality
- GLFW allows greater control over event processing loop
- GLFW has clipboard support
- GLFW supports dragging and dropping of files/directories in window
- GLUT has much longer history than GLFW (which can make code examples and tutorials using GLUT relatively easier to find)
- GLUT has built-in support for rendering text and some basic geometric objects
- GLUT has primitive support for menus

# Event-Driven Model

- event-driven model: flow of program determined by events (e.g., mouse clicks and key presses)
- application making use of event-driven model performs some initialization and then enters event-processing loop for duration of execution
- each iteration of event-processing loop does following:
  - 1 wait for event
  - 2 process event
- many libraries for building graphical user interfaces (GUIs) employ event-driven model
- GLFW uses event-driven model

# Structure of GLFW Application

- 1 initialize GLFW library by calling `glfwInit`
- 2 perform any additional initialization such as:
  - select type of OpenGL (or OpenGL ES) context to be used for subsequently created windows (via `glfwWindowHint`)
  - create windows (via `glfwCreateWindow`)
  - register callback functions for handling various types of events (e.g., via `glfwRefreshCallback`, `glfwSetWindowSizeCallback`, `glfwSetCharCallback`)
  - configure initial OpenGL state (e.g., depth buffering and clear color) and shaders
- 3 enter main event-processing loop, which repeatedly calls `glfwWaitEvents`, `glfwPollEvents`, or other similar functions
- 4 cleanup GLFW library by calling `glfwTerminate`

- GLFW header files in directory `GLFW`
- to use GLFW, need to include `glfw3.h`:  
**#include** <GLFW/glfw3.h>
- header file `glfw3.h` also includes all necessary OpenGL header files (e.g., `gl.h`, `glu.h`, `glext.h`)
- if using OpenGL extension loading library (such as GLEW), header for this library should be included before `glfw3.h`

## Keyboard, Mouse, and Joystick Events

Event Type	Description
<i>key</i>	key has been pressed, released, or repeated
<i>character</i>	character has been typed without modifiers
character with modifiers	character has been typed with modifiers
<i>mouse button</i>	mouse button has been pressed or released
cursor position	cursor has moved
cursor enter	cursor has entered or left client area of window
scroll	scrolling device has been used (e.g., mouse wheel or touchpad scrolling area)
joystick	joystick has been connected or disconnected
drop	files/directories have been dropped on window



# Event Types (Continued 1)

## Framebuffer, Window, and Monitor Events

Event Type	Description
framebuffer size	framebuffer has been resized
<i>window close</i>	window has been closed
<i>window refresh</i>	window contents need to be redrawn
<i>window size</i>	window size has changed
window position	window position has changed
window iconify	window has been iconified or deiconified
window focus	window focus has changed (i.e., been gained or lost)
monitor	monitor has been connected or disconnected

## Other Events

Event Type	Description
error	error has occurred in GLFW library

# Functions

## Initialization and Termination

Function	Description
<code>glfwInit</code>	initialize GLFW library
<code>glfwTerminate</code>	cleanup GLFW library

## Version

Function	Description
<code>glfwGetVersion</code>	get version of GLFW library
<code>glfwGetVersionString</code>	get version string of GLFW library

## Window Creation and Destruction

Function	Description
<code>glfwCreateWindow</code>	create window and its associated OpenGL or OpenGL ES context
<code>glfwDestroyWindow</code>	destroy window and its associated context
<code>glfwDefaultWindowHints</code>	reset all window hints to their default values
<code>glfwWindowHint</code>	set window hints for subsequently created windows

# Functions (Continued 1)

## Setting and Querying Window Attributes

Function	Description
<code>glfwWindowShouldClose</code>	get close flag for specified window
<code>glfwSetWindowShouldClose</code>	set close flag for specified window
<code>glfwSetWindowTitle</code>	set title of specified window
<code>glfwSetWindowIcon</code>	set icon for specified window
<code>glfwGetWindowPos</code>	get position of client area of specified window
<code>glfwSetWindowPos</code>	set position of client area of specified window
<code>glfwGetWindowSize</code>	get size of client area of specified window
<code>glfwSetWindowSize</code>	set size of client area of specified window
<code>glfwSetWindowSizeLimits</code>	set size limits of client area of specified window
<code>glfwSetWindowAspectRatio</code>	set required aspect ratio of client area of specified window

## Functions (Continued 2)

### Setting and Querying Window Attributes (Continued)

Function	Description
<code>glfwGetFramebufferSize</code>	get size of framebuffer of specified window
<code>glfwGetWindowFrameSize</code>	get size of frame of window
<code>glfwGetWindowMonitor</code>	get monitor that specified window uses for full-screen mode
<code>glfwSetWindowMonitor</code>	set monitor that specified window uses for full-screen mode
<code>glfwGetWindowAttrib</code>	get attribute of specified window
<code>glfwGetWindowUserPointer</code>	get user pointer of specified window
<code>glfwSetWindowUserPointer</code>	set user pointer of specified window

## Functions (Continued 3)

### Window Management

Function	Description
<code>glfwIconifyWindow</code>	iconifies specified window
<code>glfwRestoreWindow</code>	restores (i.e., deiconifies) specified window
<code>glfwMaximizeWindow</code>	maximizes specified window
<code>glfwShowWindow</code>	make specified window visible
<code>glfwHideWindow</code>	hide specified window
<code>glfwFocusWindow</code>	bring specified window to front and give it input focus
<code>glfwSwapBuffers</code>	swap front and back buffers of specified window when rendering with OpenGL or OpenGL ES
<code>glfwSwapInterval</code>	set swap interval for current OpenGL or OpenGL ES context

## Functions (Continued 4)

### Callback Registration

Function	Description
<code>glfwSetErrorCallback</code>	sets error callback function
<code>glfwSetWindowPosCallback</code>	sets window-position callback function for specified window
<code>glfwSetWindowSizeCallback</code>	sets window-size callback function for specified window
<code>glfwSetWindowCloseCallback</code>	sets window-close callback function for specified window
<code>glfwSetWindowRefreshCallback</code>	sets window-refresh callback function for specified window
<code>glfwSetWindowFocusCallback</code>	sets window-focus callback function for specified window

# Functions (Continued 5)

## Callback Registration (Continued 1)

Function	Description
<code>glfwSetWindowIconifyCallback</code>	sets window-iconify callback function for specified window
<code>glfwSetFramebufferSizeCallback</code>	sets callback function for framebuffer size event
<code>glfwSetKeyCallback</code>	sets (physical) key callback function for specified window
<code>glfwSetCharCallback</code>	sets character callback function for specified window
<code>glfwSetCharModsCallback</code>	sets character-with-modifiers callback function for specified window
<code>glfwSetMouseButtonCallback</code>	sets mouse-button callback function for specified window
<code>glfwGetMonitorCallback</code>	set monitor configuration callback function

## Functions (Continued 6)

Callback Registration (Continued 2)

Function	Description
<code>glfwSetCursorPosCallback</code>	sets cursor-position callback function for specified window
<code>glfwSetCursorEnterCallback</code>	sets cursor-boundary-crossing callback function for specified window
<code>glfwSetScrollCallback</code>	sets scroll callback function for specified window
<code>glfwSetDropCallback</code>	sets file-drop callback function for specified window
<code>glfwSetJoystickCallback</code>	sets joystick-configuration callback function



# Functions (Continued 7)

## Event Handling

Function	Description
<code>glfwPostEmptyEvent</code>	post empty event to event queue
<code>glfwPollEvents</code>	process any pending events and return immediately
<code>glfwWaitEvents</code>	wait until at least one event is pending, then process all pending events and return
<code>glfwWaitEventsTimeout</code>	wait until at least one event pending or timeout expires, then process any pending events and return

## Timing

Function	Description
<code>glfwGetTime</code>	get value of timer in seconds
<code>glfwSetTime</code>	set value of timer
<code>glfwGetTimerValue</code>	get value of timer in clock ticks
<code>glfwGetTimerFrequency</code>	get frequency of clock tick

## Functions (Continued 8)

### Keyboard, Mouse, Joystick, and Cursor

Function	Description
<code>glfwGetInputMode</code>	get value of input option for specified window (e.g., cursor, sticky keys/buttons)
<code>glfwSetInputMode</code>	set input option for specified window
<code>glfwGetKeyName</code>	get localized name of specified printable key
<code>glfwGetKey</code>	get last reported state of keyboard key for specified window
<code>glfwGetMouseButton</code>	get last reported state of mouse button for specified window
<code>glfwGetCursorPos</code>	get position of cursor relative to client area of specified window
<code>glfwSetCursorPos</code>	set position of cursor relative to client area of specified window

# Functions (Continued 9)

## Keyboard, Mouse, Joystick, and Cursor (Continued)

Function	Description
<code>glfwCreateCursor</code>	create custom cursor
<code>glfwCreateStandardCursor</code>	creates cursor with standard shape
<code>glfwDestroyCursor</code>	destroys cursor
<code>glfwSetCursor</code>	set cursor for use in specified window
<code>glfwJoystickPresent</code>	test if joystick is present
<code>glfwGetJoystickAxes</code>	get values of all axes of specified joystick
<code>glfwGetJoystickButtons</code>	get state of all buttons of specified joystick
<code>glfwGetJoystickName</code>	get name of specified joystick

## Clipboard

Function	Description
<code>glfwGetClipboardString</code>	gets contents of clipboard as string
<code>glfwSetClipboardString</code>	sets clipboard to specified string

# Functions (Continued 10)

## Monitor Management

Function	Description
<code>glfwGetMonitors</code>	get currently connected monitors
<code>glfwGetPrimaryMonitor</code>	get primary monitor
<code>glfwGetMonitorPos</code>	get position of specified monitor's viewport on virtual screen
<code>glfwGetMonitorPhysicalSize</code>	get physical size of specified monitor
<code>glfwGetMonitorName</code>	get name of specified monitor
<code>glfwGetVideoModes</code>	get available video modes for specified monitor
<code>glfwGetVideoMode</code>	get current video mode of specified monitor
<code>glfwSetGamma</code>	set gamma for specified monitor
<code>glfwGetGammaRamp</code>	get current gamma ramp for specified monitor
<code>glfwSetGammaRamp</code>	set current gamma ramp for specified monitor

# Functions (Continued 11)

## Contexts and Extensions

Function	Description
<code>glfwMakeContextCurrent</code>	make context of specified window current for calling thread
<code>glfwGetCurrentContext</code>	get window whose context is current on calling thread
<code>glfwExtensionSupported</code>	tests if specified API extension is supported by current OpenGL or OpenGL ES context
<code>glfwGetProcAddress</code>	get address of specified OpenGL or OpenGL ES core or extension function (if supported) for current context

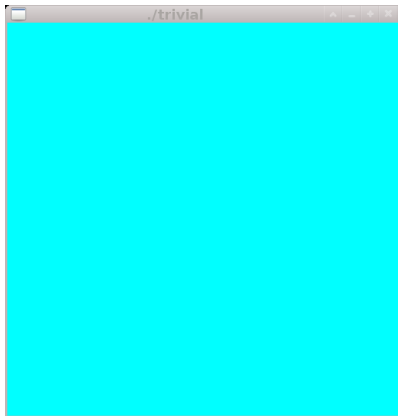
## Functions (Continued 12)

### Vulkan

Function	Description
<code>glfwVulkanSupported</code>	tests if Vulkan loader has been found
<code>glfwGetRequiredInstanceExtensions</code>	get Vulkan instance extensions required by GLFW
<code>glfwGetInstanceProcAddress</code>	get address of specified Vulkan instance function
<code>glfwGetPhysicalDevicePresentationSupport</code>	test if specified queue family can present images
<code>glfwCreateWindowSurface</code>	create Vulkan surface for specified Window

# Minimalist GLFW Program

- minimalist program using GLFW
- create window that is cleared to particular color



# Minimalist GLFW Program: Source Code

```
1  #include <cstdlib>
2  #include <GLFW/glfw3.h>
3
4  void display(GLFWwindow* window) {
5      glClearColor(0.0, 1.0, 1.0, 0.0);
6      glClear(GL_COLOR_BUFFER_BIT | GL_DEPTH_BUFFER_BIT);
7      glfwSwapBuffers(window);
8  }
9
10 int main(int argc, char** argv) {
11     if (!glfwInit()) {return EXIT_FAILURE;}
12     glfwSwapInterval(1);
13     GLFWwindow* window = glfwCreateWindow(512, 512, argv[0],
14         nullptr, nullptr);
15     if (!window) {
16         glfwTerminate();
17         return EXIT_FAILURE;
18     }
19     glfwMakeContextCurrent(window);
20     glfwSetWindowRefreshCallback(window, display);
21     while (!glfwWindowShouldClose(window)) {
22         glfwWaitEvents();
23     }
24     glfwTerminate();
25     return EXIT_SUCCESS;
26 }
```



- 1 GLFW Reference Manual, <http://www.glfw.org/docs/latest>

## Section 5.5

# OpenGL Mathematics (GLM) Library

- open-source mathematics library for graphics software based on OpenGL Shading Language (GLSL)
- intended for use with OpenGL
- written in C++
- developed by Christophe Riccio
- provides classes and functions with similar naming conventions and functionality as in GLSL
- web site: <http://glm.g-truc.net>

- library has numerous header files
- header files under `glm` directory
- all header files for core GLM functionality can be included by including header file `glm.hpp`
- for matrix transformation functionality, include `gtc/matrix_transform.hpp`
- for string conversion functionality, include `gtx/string_cast.hpp`
- for type value functionality, include `gtc/type_ptr.hpp`
- all identifiers placed in namespace `glm`

- provides vector and matrix types similar to GLSL

- vector types:

- `vec2`, `vec3`, `vec4`
- `bvec2`, `bvec3`, `bvec4`
- `ivec2`, `ivec3`, `ivec4`
- `uvec2`, `uvec3`, `uvec4`
- `dvec2`, `dvec3`, `dvec4`

- matrix types:

- `mat2x2`, `mat2x3`, `mat2x4`, `mat2`,  
`mat3x2`, `mat3x3`, `mat3x4`, `mat3`,  
`mat4x2`, `mat4x3`, `mat4x4`, `mat4`
- `dmat2x2`, `dmat2x3`, `dmat2x4`, `dmat2`,  
`dmat3x2`, `dmat3x3`, `dmat3x4`, `dmat3`,  
`dmat4x2`, `dmat4x3`, `dmat4x4`, `dmat4`

- provides GLSL functions (e.g., `inverse` and `transpose`)
- provides functions that offer functionality similar to legacy OpenGL/GLU functions (e.g., `rotate`, `scale`, `translate`, `frustum`, `ortho`, `lookAt`, `perspective`, `pickMatrix`, `project`, and `unProject`)

# Code Example: Basic Usage

```
1  #include <iostream>
2  #include <glm/glm.hpp>
3  #include <glm/gtc/matrix_transform.hpp>
4  #include <glm/gtx/string_cast.hpp>
5  #include <cmath>
6
7  int main() {
8      glm::mat4 mv(1.0f);
9      mv = mv * glm::lookAt(glm::vec3(0.0f, 0.0f, 0.0f),
10         glm::vec3(1.0f, 0.0f, 0.0f),
11         glm::vec3(0.0f, 0.0f, 1.0f));
12      mv = mv * glm::translate(mv, glm::vec3(1.0f, 1.0f, 1.0f));
13      mv = mv * glm::rotate(mv, glm::radians(90.0f),
14         glm::vec3(0.0f, 0.0f, 1.0f));
15      mv = mv * glm::scale(mv, glm::vec3(1.0f, 1.0f, 2.0f));
16      glm::mat4 p = glm::perspective(glm::radians(90.0f), 1.0f,
17         1.0f, 2.0f);
18      glm::mat4 mvp = p * mv;
19      glm::vec4 v(1.0f, -1.0f, -1.0f, 1.0f);
20      std::cout << glm::to_string(glm::vec3(mv * v)) << '\n';
21      std::cout << glm::to_string(glm::vec3(mvp * v)) << '\n';
22      std::cout << glm::radians(180.0f) << '\n';
23      std::cout << glm::degrees(M_PI) << '\n';
24  }
```

## Code Example: value\_ptr

```
1  #include <GL/glew.h>
2  #include <GL/gl.h>
3  #include <glm/glm.hpp>
4  #include <glm/gtc/type_ptr.hpp>
5
6  void setUniform(GLint loc) {
7      glm::mat4 m(1.0f);
8      // ...
9      glUniform4fv(loc, 4, glm::value_ptr(m));
10 }
```



## Section 5.6

# Open Graphics Library (OpenGL)

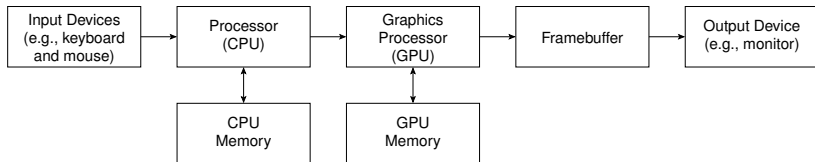
# Open Graphics Library (OpenGL)

- application programming interface (API) for high-performance high-quality 2-D and 3-D graphics rendering
- most widely adopted 2-D and 3-D graphics API in industry
- bindings for numerous programming languages (i.e., C, Java, and Fortran)
- focus exclusively on C language binding herein
- window-system and operating-system independent
- available on all mainstream systems (e.g., Microsoft Windows, OS X, and Linux/Unix)
- vendor-neutral, controlled by independent consortium with many organizations as members (including companies such as Intel, NVIDIA, and AMD)
- official web site: <http://www.opengl.org>
- OpenGL ES provides (simplified) subset of OpenGL API for embedded systems (e.g., mobile phones, game consoles, personal navigation devices, personal media players, automotive systems, settop boxes)

# OpenGL Functionality

- geometric primitives include points, line segments, and triangles
- arrange geometric primitives in 3-D space and select desired vantage point for viewing composed scene
- calculate colors of objects (e.g., by explicit assignment, lighting, texture mapping, or combination thereof)
- convert mathematical description of objects to pixels on screen (i.e., rasterization)
- can eliminate hidden parts of objects (via depth buffering), perform antialiasing, and so on
- some functionality relies on shaders provided by application program
- only concerned with rendering
- no mechanism provided for creating windows or obtaining user input (e.g., via mouse or keyboard)
- another library must be used in conjunction with OpenGL in order to manage windows and handle user input

# Modern OpenGL



- main responsibility of application is to provide graphics data to GPU
- application program running on CPU sends graphics data to GPU
- programs running on GPU called **shaders** control rendering
- GPU performs all rendering
- high performance achieved by offloading rendering work to GPU, with GPU being highly specialized for rendering
- image formed and stored in framebuffer
- shaders written in OpenGL Shading Language (GLSL)
- application program uses OpenGL to compile and link shader source code to yield executable shader program that runs on GPU

- OpenGL is state machine
- OpenGL functions can be roughly classified into two categories:
  - 1 primitive generation
  - 2 state management
- primitive-generation functions:
  - produce graphics output if primitive is visible
  - how vertices are processed and appearance of primitive controlled by OpenGL state
- state-management functions:
  - enabling/disabling OpenGL functionality (e.g., depth buffering)
  - configuring shader programs
  - setting/querying shader variables

# Contexts and Profiles

- feature that may be removed in future version of OpenGL is said to be **deprecated**
- **profile** defines subset of OpenGL functionality targeted to specific application domains
- two profiles: core and compatibility
- **core profile** provides functionality mandated by particular version of OpenGL (which does not include deprecated and removed features)
- **compatibility profile** restores support for all functionality that has been removed from OpenGL
- all OpenGL implementations must support core profile, but are not required to support compatibility profile
- for given profile, two types of contexts: full or forward compatible
- **forward compatible context** does not support deprecated features from profile
- **full context** supports deprecated features from profile

- header files for OpenGL located in `GL` (or `OpenGL`) directory
- definitions necessary for OpenGL can be found in header file `gl.h`
- above header file provides definitions of all constants and data types (e.g., `GLint` and `GLfloat`) and function declarations for OpenGL
- on some platforms, in order to access newer OpenGL functionality, may need to include `glew.h` (typically in `GL` directory) before `gl.h`
- normally, OpenGL used in conjunction with another helper library such as `GLFW` or `GLUT`
- other helper libraries also have header files of their own that must be included
- often header files for helper libraries include `gl.h`

Type	Description
GLboolean	boolean
GLbyte	8-bit signed two's complement integer
GLubyte	8-bit unsigned integer
GLchar	8-bit character
GLshort	16-bit signed two's complement integer
GLushort	16-bit unsigned integer
GLint	32-bit signed two's complement integer
GLuint	32-bit unsigned integer
GLfloat	single-precision floating-point value
GLdouble	double-precision floating-point value

- OpenGL types do not necessarily correspond to similarly named C types (e.g., GLint is not necessarily **int**)



# Function Naming Conventions

- all OpenGL functions begin with `gl`
- some OpenGL commands have numerous variants that differ in number and type of parameters
- such commands are named using following pattern:

*generic\_name N T V*

where *generic\_name* is generic name of function, *N* is digit (i.e., 2, 3, 4) indicating number of components, *T* is one or two letters indicating data type of components, *V* is either nothing or letter *v* to indicate component data specified as individual values or as vector (i.e., pointer to array), respectively

Number <i>N</i>	
2	( <i>x, y</i> )
3	( <i>x, y, z</i> )
4	( <i>x, y, z, w</i> )

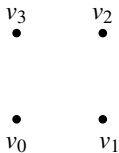
Data Type <i>T</i>			
b	GLbyte	ub	GLubyte
s	GLshort	us	GLushort
i	GLint	ui	GLuint
f	GLfloat	d	GLdouble

- `glUniform3f`: specific version of generic `glUniform` function that takes data in form of three `GLfloat` parameters
- `glUniform3fv`: specific version of generic `glUniform` function that takes data in form of pointer to array of triples of `GLfloat` values

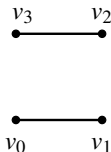
# Representing Geometric Objects

- geometric objects represented using vertices
- each vertex has variety of attributes, such as:
  - positional coordinates
  - color
  - texture coordinates
  - surface normal
  - any other data associated with point in space
- position represented using homogeneous coordinates
- vertex data must be stored in vertex buffer objects (VBOs)
- VBOs must be associated with vertex array objects (VAOs)
- VAOs/VBOs allow application program to transfer data to GPU *once* and then select between different data on GPU by activating different VAOs/VBOs

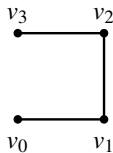
# Geometric Primitives



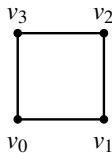
points



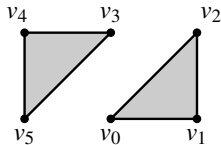
lines



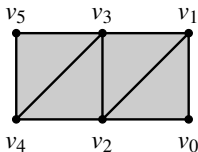
line strip



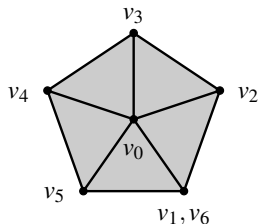
line loop



triangles



triangle strip



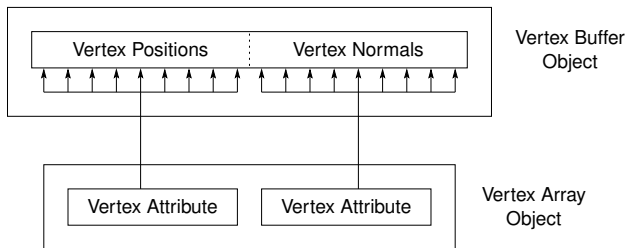
triangle fan

- all primitives specified by vertices

# Provoking Vertex

- each primitive has provoking vertex
- one of two conventions can be used to determine provoking vertex: first vertex or last vertex
- for example, with last vertex convention, provoking vertex for triangle is third (i.e., last) vertex of triangle
- convention defaults to last vertex
- convention can be set with `glProvokingVertex`
- provoking vertex becomes important, for example, when using flat interpolation

# Vertex Array and Vertex Buffer Objects (VAOs and VBOs)



- vertex buffer objects (VBOs) store vertex attributes (e.g., positions, normals, colors, and texture coordinates)
- storage for VBOs resides in GPU memory
- vertex array objects (VAOs) allow data stored in VBOs to be associated with vertex attributes for vertex shader
- VAOs specify layout (e.g., offset and stride) and format (e.g., type) of data in VBOs
- to render primitives need VAO (which, in turn, is associated with one or more VBOs)

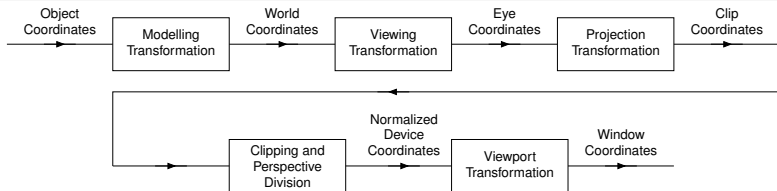
# Vertex Array Objects (VAOs)

- VAOs store data for geometric object
- VAO identified by name, which is integer of type `GLuint`
- create one or more VAOs by generating VAO names via `glGenVertexArrays`
- VAO initialized as follows:
  - 1 bind specific VAO for initialization via `glBindVertexArray`
  - 2 update VBOs associated with VAO, and specify layout and format of VBO data and its correspondence with vertex attributes for rendering via `glVertexAttribPointer`
- data in VAO rendered as follows:
  - 1 bind VAO for use in rendering via `glBindVertexArray`
  - 2 draw content of currently enabled arrays via `glDrawArrays`
- only enabled attributes will be used for rendering (where attributes are enabled with `glEnableVertexAttribArray`)

# Vertex Buffer Objects (VBOs)

- vertex buffer objects (VBOs) provide means to transfer data to GPU memory
- vertex data must be stored in VBO associated with VAO
- each VBO associated with name, which is integer of type GLuint
- generate VBO names via `glGenBuffers`
- bind specific VBO for initialization via `glBindBuffer` (after first binding associated VAO)
- allocate underlying storage for VBO (and optionally load data into VBO) via `glBufferData`
- load data into VBO via `glBufferSubData`

# Coordinate Systems

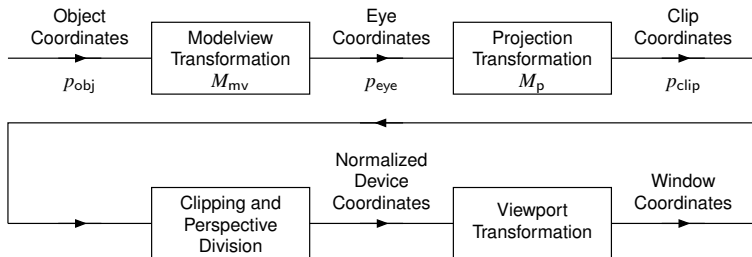


- **object coordinates**: coordinates of object relative to its local origin
- **world coordinates**: coordinates of three-dimensional environment (i.e., world) being rendered
- **eye coordinates**: coordinates relative to camera from which world is being viewed
- **clip coordinates**: coordinates normalized such that viewing volume falls in  $[-1, 1] \times [-1, 1] \times [-1, 1]$
- **normalized device coordinates**: result of converting clip coordinates to Cartesian coordinates by perspective division (i.e., dividing by  $w$  coordinate)
- **window coordinates**: coordinates relative to graphics window



- appearance of rendered scene determined by camera position, orientation, and viewing volume
- camera positioned at origin
- camera oriented to point in negative  $z$  direction with positive  $y$  axis pointing up
- orthographic projection in direction of  $z$  axis with clipping planes  $x = -1$ ,  $x = 1$ ,  $y = -1$ ,  $y = 1$ ,  $z = -1$ , and  $z = 1$
- viewing volume is  $[-1, 1] \times [-1, 1] \times [-1, 1]$  (i.e., cube centered at origin with sides of length 2)
- different camera position, orientation, and viewing volume can be achieved by employing transformations
- perspective projection accomplished by applying transformation that warps viewing volume into frustum

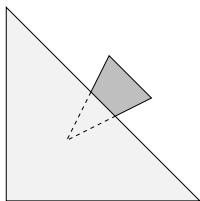
# Transformations



- often modelling and viewing transformations combined into single transformation called modelview transformation
- $p_{eye} = M_{mv}p_{obj}$
- $p_{clip} = M_p p_{eye} = M_p M_{mv} p_{obj}$
- clip coordinates and normalized device coordinates still retain depth (i.e.,  $z$ ) information in order to facilitate depth buffering

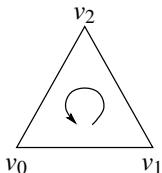
- viewport transformation determines drawable region within window
- viewport transformation set via `glViewport`

# Depth Buffering

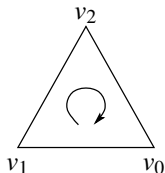


- in above figure, darker triangle is partially occluded by lighter triangle from vantage point of camera
- in OpenGL, camera always pointing in direction of negative  $z$  axis
- therefore,  $z$  coordinate can be used to determine distance of fragment from eye, with lesser value (i.e., closer to  $-\infty$ ) corresponding to greater distance
- if depth buffering enabled, fragment not drawn if its  $z$  coordinate less than  $z$  coordinate of previously drawn pixel

# Face Culling



Counterclockwise (CCW)  
Winding Order



Clockwise (CW)  
Winding Order

- winding order used to distinguish front and back sides of triangles
- which winding order corresponds to front side of triangle specified via `glFrontFace`
- which side (or sides) of triangle should be culled specified via `glCullFace`
- if face culling enabled, culled side of triangles not rendered

- `glEnable` and `glDisable` used to enable and disable specific functionality

Value	Meaning
<code>GL_CULL_FACE</code>	if enabled, cull polygons based on their winding in window coordinates (e.g., do not render backs of faces)
<code>GL_DEPTH_TEST</code>	if enabled, do depth comparisons and update depth buffer
<code>GL_LINE_SMOOTH</code>	if enabled, draw lines with antialiasing

Function	Description
<code>glClear</code>	clear buffer to preset values
<code>glClearColor</code>	specify clear values for color buffers

- program typically consists of steps like following:
  - 1 create window associated with OpenGL context
  - 2 initialize shaders (e.g., compile and link) and other OpenGL state (e.g., depth buffering and clear color)
  - 3 initialize data to be drawn
  - 4 register callback functions to process events
  - 5 enter main event-processing loop, which repeatedly waits for event of interest and then handles it by invoking appropriate callback function
- events of interest typically include such things as:
  - request to redraw window
  - window-resize notification
  - keyboard input
  - mouse-button press/release

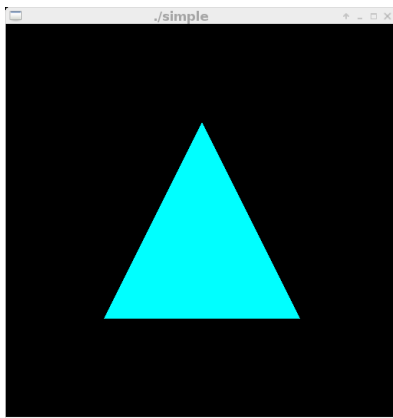


## Section 5.6.1

### **Simple OpenGL Program**

# OpenGL Application Program Example

- consider very simple OpenGL application program (which utilizes GLFW)
- draws triangle in window
- rendered output shown below



```
1  #include <cstdlib>
2  #include <string>
3  #include <GL/glew.h>
4  #include <GLFW/glfw3.h>
```

# Main Function

```
6   GLuint vao = 0;

103 void fatalError() {
104     glfwTerminate();
105     std::exit(EXIT_FAILURE);
106 }

108 int main(int argc, char** argv) {
109     if (!glfwInit()) {return EXIT_FAILURE;}
110     GLFWwindow* window = makeWindow(512, 512, argv[0]);
111     if (!window) {fatalError();}
112     glfwMakeContextCurrent(window);
113     glewExperimental = GL_TRUE;
114     if (glewInit() != GLEW_OK) {fatalError();}
115     GLuint program = makeProgram(vShaderSource,
116     fShaderSource);
117     if (!program) {fatalError();}
118     glUseProgram(program);
119     glClearColor(0.0, 0.0, 0.0, 0.0);
120     GLuint vbo;
121     makeVao(program, vao, vbo);
122     glfwSetWindowRefreshCallback(window, refresh);
123     while (!glfwWindowShouldClose(window))
124         {glfwWaitEvents();}
125     glfwTerminate();
126     return EXIT_SUCCESS;
127 }
```

# Make Window

```
84 GLFWwindow* makeWindow(int width, int height,  
85     const std::string& title) {  
86     glfwWindowHint(GLFW_CONTEXT_VERSION_MAJOR, 3);  
87     glfwWindowHint(GLFW_CONTEXT_VERSION_MINOR, 3);  
88     glfwWindowHint(GLFW_OPENGL_FORWARD_COMPAT, GL_TRUE);  
89     glfwWindowHint(GLFW_OPENGL_PROFILE,  
90         GLFW_OPENGL_CORE_PROFILE);  
91     GLFWwindow* window = glfwCreateWindow(width, height,  
92         title.c_str(), nullptr, nullptr);  
93     return window;  
94 }
```

# Vertex and Fragment Shaders

```
8  const std::basic_string<GLchar> vShaderSource = R"(
9  #version 330
10 in vec3 aPosition;
11 void main() {
12     gl_Position = vec4(aPosition, 1.0);
13 }
14 )";
15
16 const std::basic_string<GLchar> fShaderSource = R"(
17 #version 330
18 out vec4 fColor;
19 void main() {
20     fColor = vec4(0.0, 1.0, 1.0, 1.0);
21 }
22 )";
```

# Compiling Shaders

```
24 GLuint compileShader(GLuint type,
25     const std::basic_string<GLchar>& source) {
26     GLuint shader = glCreateShader(type);
27     if (!shader) {return 0;}
28     const GLchar* cp = &source[0];
29     GLint len = source.size();
30     glShaderSource(shader, 1, &cp, &len);
31     glCompileShader(shader);
32     GLint status = GL_FALSE;
33     glGetShaderiv(shader, GL_COMPILE_STATUS, &status);
34     if (status != GL_TRUE)
35         {glDeleteShader(shader); return 0;}
36     return shader;
37 }
```

# Linking Shader Program

```
39 GLuint makeProgram(  
40     const std::basic_string<GLchar>& vShaderSource,  
41     const std::basic_string<GLchar>& fShaderSource) {  
42     GLuint vShader = compileShader(GL_VERTEX_SHADER,  
43         vShaderSource);  
44     if (!vShader) {return 0;}  
45     GLuint fShader = compileShader(GL_FRAGMENT_SHADER,  
46         fShaderSource);  
47     if (!fShader) {glDeleteShader(vShader); return 0;}  
48     GLuint program = glCreateProgram();  
49     GLint status = GL_FALSE;  
50     if (program) {  
51         glAttachShader(program, vShader);  
52         glAttachShader(program, fShader);  
53         glLinkProgram(program);  
54         glGetProgramiv(program, GL_LINK_STATUS, &status);  
55     }  
56     glDeleteShader(vShader);  
57     glDeleteShader(fShader);  
58     if (!program) {return 0;}  
59     if (status != GL_TRUE)  
60         {glDeleteProgram(program); return 0;}  
61     return program;  
62 }
```



# Initialize Vertex Array Object (VAO)

```
64 void makeVao(GLuint program, GLuint& vao,  
65             GLuint& vbo) {  
66     static const GLfloat vertices[][3] = {  
67         {-0.50, -0.50, 0.0},  
68         { 0.50, -0.50, 0.0},  
69         { 0.00, 0.50, 0.0}  
70     };  
71     glGenVertexArrays(1, &vao);  
72     glGenBuffers(1, &vbo);  
73     glBindVertexArray(vao);  
74     glBindBuffer(GL_ARRAY_BUFFER, vbo);  
75     glBufferData(GL_ARRAY_BUFFER, sizeof(vertices),  
76                 vertices, GL_STATIC_DRAW);  
77     GLuint aPosition = glGetAttribLocation(program,  
78         "aPosition");  
79     glVertexAttribPointer(aPosition, 3, GL_FLOAT, GL_FALSE,  
80         0, 0);  
81     glEnableVertexAttribArray(aPosition);  
82 }
```

# Window Refresh Callback

```
6  GLuint vao = 0;

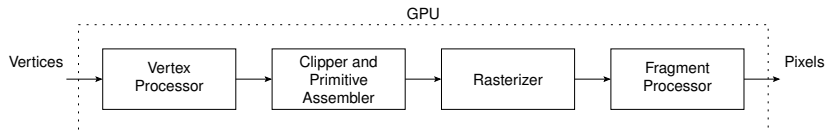
96 void refresh(GLFWwindow* window) {
97     glClear(GL_COLOR_BUFFER_BIT | GL_DEPTH_BUFFER_BIT);
98     glBindVertexArray(vao);
99     glDrawArrays(GL_TRIANGLES, 0, 3);
100    glfwSwapBuffers(window);
101 }
```

## Section 5.6.2

### **Shaders**

- shader is user-defined program that runs on GPU and provides functionality associated with some particular stage of rendering pipeline
- shaders written in OpenGL Shading Language (GLSL)
- as of OpenGL 3.1, application program must provide shaders as no default shaders provided (in core profile)
- several types of shaders:
  - vertex shader
  - tessellation control shader
  - tessellation evaluation shader
  - geometry shader
  - fragment shader
  - compute shader
- each type of shader performs specific type of task on GPU

# Rendering Pipeline and Shaders



- each type of shader performs distinct task within rendering pipeline
- vertex shader (which is associated with vertex processor block) provides any last geometric transformation of vertices before being fed to remainder of rendering pipeline
- geometry shader (which is associated with vertex processor block) generates actual primitives to be rendered based on primitives received from previous pipeline stage
- fragment shader (which is associated with fragment processor block) provides color to each pixel in framebuffer

- shaders written in GLSL
- GLSL is portable multiplatform C-like language
- GLSL borrows heavily from C syntax
- provides simplified subset of C language with numerous modifications:
  - adds new data types, such as matrix and vector types
  - adds overloaded operators and constructors
- supports C and C++ style comments
- GLSL keywords cannot be used as identifiers
- names beginning with “gl\_” prefix reserved by GLSL

# Reserved Keywords

attribute	inout	mat4x4
const	float	vec2
uniform	int	vec3
varying	void	vec4
layout	bool	ivec2
centroid	true	ivec3
flat	false	ivec4
smooth	invariant	bvec2
noperspective	discard	bvec3
break	return	bvec4
continue	mat2	uint
do	mat3	uvec2
for	mat4	uvec3
while	mat2x2	uvec4
switch	mat2x3	lowp
case	mat2x4	mediump
default	mat3x2	highp
if	mat3x3	precision
else	mat3x4	sampler1D
in	mat4x2	sampler2D
out	mat4x3	sampler3D

## Reserved Keywords (Continued)

samplerCube	usamplerCube
sampler1DShadow	usampler1DArray
sampler2DShadow	usampler2DArray
samplerCubeShadow	sampler2DRect
sampler1DArray	sampler2DRectShadow
sampler2DArray	isampler2DRect
sampler1DArrayShadow	usampler2DRect
sampler2DArrayShadow	samplerBuffer
isampler1D	isamplerBuffer
isampler2D	usamplerBuffer
isampler3D	sampler2DMS
isamplerCube	isampler2DMS
isampler1DArray	usampler2DMS
isampler2DArray	sampler2DMSArray
usampler1D	isampler2DMSArray
usampler2D	usampler2DMSArray
usampler3D	struct

plus other keywords added since OpenGL 3.3



# The #version Directive

- `#version` directive specifies which version of GLSL should be used to compile/link shader
- if `#version` directive specified, must be first statement in source
- if no `#version` directive given, version 1.10 is assumed
- `#version` directive takes two parameters (with second being optional):
  - 1 integer specifying GLSL version (scaled by a factor of 100)
  - 2 profile name, which can be either `core` or `compatibility` with `core` being default
- for OpenGL 3.3 and above, corresponding GLSL version matches OpenGL version (e.g., OpenGL 4.1 uses GLSL 4.1); for earlier OpenGL versions, relationship between OpenGL and GLSL versions as follows:

OpenGL Version	GLSL Version
2.0	1.10
2.1	1.20
3.0	1.30
3.1	1.40
3.2	1.50

- for example, to specify use of GLSL 3.30 with core profile:  
`#version 330`

## Scalar and Void Types

Type	Description
<b>void</b>	dummy type for functions without return value
<b>bool</b>	boolean type
<b>int</b>	signed integer type
<b>uint</b>	unsigned integer type
<b>float</b>	single-precision floating-point type

## Vector of **float** Types

Type	Description
<b>vec2</b>	two-component vector of <b>float</b>
<b>vec3</b>	three-component vector of <b>float</b>
<b>vec4</b>	four-component vector of <b>float</b>

## Vector of **bool** Types

Type	Description
<b>bvec2</b>	two-component vector of <b>bool</b>
<b>bvec3</b>	three-component vector of <b>bool</b>
<b>bvec4</b>	four-component vector of <b>bool</b>

# Basic Types (Continued 1)

## Vector of **int** Types

Type	Description
<b>ivec2</b>	two-component vector of <b>int</b>
<b>ivec3</b>	three-component vector of <b>int</b>
<b>ivec4</b>	four-component vector of <b>int</b>

## Vector of **uint** Types

Type	Description
<b>uvec2</b>	two-component vector of <b>uint</b>
<b>uvec3</b>	three-component vector of <b>uint</b>
<b>uvec4</b>	four-component vector of <b>uint</b>

## Matrix of **float** Types

Type	Description	Type	Description
<b>mat2</b>	$2 \times 2$ matrix of <b>float</b>	<b>mat3x2</b>	$3 \times 2$ matrix of <b>float</b>
<b>mat3</b>	$3 \times 3$ matrix of <b>float</b>	<b>mat3x3</b>	same as <b>mat3</b>
<b>mat4</b>	$4 \times 4$ matrix of <b>float</b>	<b>mat3x4</b>	$3 \times 4$ matrix of <b>float</b>
<b>mat2x2</b>	same as <b>mat2</b>	<b>mat4x2</b>	$4 \times 2$ matrix of <b>float</b>
<b>mat2x3</b>	$2 \times 3$ matrix of <b>float</b>	<b>mat4x3</b>	$4 \times 3$ matrix of <b>float</b>
<b>mat2x4</b>	$2 \times 4$ matrix of <b>float</b>	<b>mat4x4</b>	same as <b>mat4</b>

## Basic Types (Continued 2)

- numerous sampler types
- numerous other types added since OpenGL 3.3
- matrix types stored in column-major order
- no pointer types
- **const** qualifier similar to C
- **struct** can be used to construct user-defined types

- standard C/C++ arithmetic and logical operators
- operators overloaded for matrix and vector types
- for two operands of vector type, multiplication operator performs component-wise multiplication
- for two operands of matrix type or one operand of matrix type and one of vector type, multiplication operator performs standard matrix/vector multiplication
- example:

```
mat4 a; mat4 b; mat4 c;  
vec4 u; vec4 v; vec4 w;  
// ...  
v = a * u; // standard matrix-vector multiplication  
c = a * b; // standard matrix-vector multiplication  
w = u * v; // component-wise multiplication
```

# Operators (Continued 1)

- first, second, third, and fourth components of vector (if they exist) can be selected by:
  - subscripting operator with subscripts 0, 1, 2, and 3, respectively; or
  - selection operator with *x*, *y*, *z*, and *w*, respectively; or
  - selection operator with *r*, *g*, *b*, and *a*, respectively; or
  - selection operator with *s*, *t*, *p*, and *q*, respectively

- example:

```
vec3 v;  
// ...  
float x = v.x;  
float y = v.y;  
float z = v.z;
```

- components of matrices can be accessed by subscripting operator
- single subscripting on matrix results in column of matrix
- double subscripting on matrix results in element of matrix

- example:

```
mat2 a;  
// ...  
vec2 v = a[0];  
float f = a[0][0];
```

- can also form vectors by selecting multiple elements from vector (e.g., swizzling and smearing)
- example:

```
vec4 v; vec4 u;  
vec3 a;  
// ...  
u = v.wzyx; // vec4(v.w, v.z, v.y, v.x)  
u = v.xxyy; // vec4(v.x, v.x, v.y, v.y)  
a = v.xyz;  // vec3(v.x, v.y, v.z)  
u = a.xxxx; // vec4(a.x, a.x, a.x, a.x)
```

- selection statements

- **if**
- **if-else**
- ternary operator
- **switch**

- looping statements

- **for**
- **while**
- **do-while**

- also has **break** and **continue**

- no **goto** statement

- only in fragment shader: **discard** statement



- numerous built-in functions provided (e.g., `abs`, `sin`, `cos`, `sqrt`)
- user-defined functions are supported
- recursion not allowed
- function overloading supported (including for user-defined functions)
- **return** statement to return from function

- constructor is function with same name as type
- used to create value of type named by function
- constructor parameters for matrix types specified in column-major order
- example:

```
vec3 v3 = vec3(1.0, 2.0, 3.0);  
mat2 m2 = mat2(1.0, 2.0, 3.0, 4.0);  
    // first column of m2 is 1.0, 2.0  
    // second column of m2 is 3.0, 4.0  
mat4 m4 = mat4(1.0); // identity matrix  
vec4 v4 = vec4(0.0); // zero vector  
const int lut[3] = int[3](1, 2, 4);  
vec2 va[2] =  
    vec2[(vec2(1.0, 2.0), vec2(3.0, 4.0))];  
bool b = bool(1);
```

# Conversions

- number of implicit conversions allowed, some of which identified below
- integer types (e.g., **int** and **uint**) can be implicitly converted to **float**
- each integer vector type (e.g., **ivec4**) can be implicitly converted to floating-point vector type of same dimension (e.g., **vec4**)
- floating-point type *cannot* be implicitly converted to integer type
- unsigned integer type (e.g., **uint**) *cannot* be implicitly converted to signed integer type (e.g., **int**)
- example:

```
int i; uint ui; float f; vec4 v4; ivec4 iv4;
// ...
f = i; // OK
// i = f; // ERROR: no implicit conversion
i = int(f); // OK
// iv4 = v4; // ERROR: no implicit conversion
iv4 = ivec4(v4); // OK
// i = 0u; // ERROR: no implicit conversion
// i = ui; // ERROR: no implicit conversion
i = int(ui); // OK
```

## Angle and Trigonometric Functions

Function	Description
<code>radians</code>	convert from degrees to radians
<code>degrees</code>	convert from radians to degrees
<code>sin</code>	sine function
<code>cos</code>	cosine function
<code>tan</code>	tangent function
<code>asin</code>	arcsine function
<code>acos</code>	arccosine function
<code>atan</code>	arctangent function

# Built-In Functions (Continued 1)

## Exponential Functions

Function	Description
<code>pow</code>	exponentiation function
<code>exp</code>	base- $e$ exponentiation function
<code>log</code>	natural logarithm function
<code>exp2</code>	base-2 exponentiation function
<code>log2</code>	base-2 logarithm function
<code>sqrt</code>	square-root function
<code>inversesqrt</code>	reciprocal of square-root function

# Built-In Functions (Continued 2)

Common Functions

Function	Description
abs	absolute-value function
sign	signum function
floor	floor function
ceil	ceiling function
fract	fractional-part function
mod	modulo function
min	minimum of two values
max	maximum of two values
clamp	clamp value to specified range
mix	affine combination of two values
step	step function
smoothstep	smooth step function

## Built-In Functions (Continued 3)

### Geometric Functions

Function	Description
length	length of vector
distance	distance between two points
dot	dot product
cross	cross product
normalize	get vector of unit length
faceforward	get vector that points in same direction as reference vector
reflect	get vector that points in direction of reflection
refract	get vector that points in direction of refraction

## Built-In Functions (Continued 4)

### Fragment Processing Functions

Function	Description
dFdx	partial derivative of argument with respect to x
dFdy	partial derivative of argument with respect to y
fwidth	sum of absolute value of derivatives in x and y

### Matrix Functions

Function	Description
matrixCompMult	multiply matrices component-wise

### Texture Lookup

Function	Description
texture2D	perform 2D texture lookup
textureCube	perform cubemap texture lookup



## Built-In Functions (Continued 5)

Vector Relational Functions

Function	Description
<code>lessThan</code>	component-wise less-than comparison
<code>lessThanEqual</code>	component-wise less-than-or-equal comparison
<code>greaterThan</code>	component-wise greater-than comparison
<code>greaterThanEqual</code>	component-wise greater-than-or-equal comparison
<code>equal</code>	component-wise equality comparison
<code>notEqual</code>	component-wise inequality comparison
<code>any</code>	any component is true
<code>all</code>	all components are true
<code>not</code>	component-wise logical complement operation

# The **in** and **out** Qualifiers

- shader parameters (i.e., input and output variables of shaders) and function parameters can be qualified with **in** and **out** qualifiers
- parameter declared **in**:
  - value given to parameter will be copied into parameter when function called
  - function may then modify parameter but changes will not affect caller
  - essentially pass-by-value semantics
- parameter declared **out**:
  - parameter will not have its value initialized by caller so initial value of parameter at start of function is undefined
  - function must modify parameter
  - after function's execution is complete, value of parameter will be copied into variable that user specified when calling function
- default qualifier is **in**
- example:

```
float foo(float x, int i, out int n);  
float calculate(in float x, in float y, in int n);
```

# The **in** and **out** Qualifiers (Continued)

## ■ example:

```
void calc(float x, int i, out float y, out int j) {  
    // at this point, y and j are undefined  
    y = ++x;  
    j = ++i;  
}  
  
void func() {  
    float a = 0.0;  
    int b = 0;  
    float c = 0.0;  
    int d = 0;  
    calc(a, b, c, d);  
    // a and b are unchanged by function call  
    // c is 1.0, d is 1  
}
```

# The **uniform** Qualifier

- global variables and interface blocks can be declared with **uniform** qualifier
- **uniform** qualifier indicates that value of variable does not change across multiple shader invocations during rendering of single primitive (i.e., during `glDraw*` call)
- uniform variables form linkage between shader and application program
- used to declare variables shared between shader and application program (e.g., projection matrix, light source position, material color)
- uniform variable cannot be modified in shader
- uniform variable can only be modified by application program
- uniform variable can be used in multiple shaders (e.g., vertex and fragment shaders)
- if used in multiple shaders, must have identical declaration in each
- example:

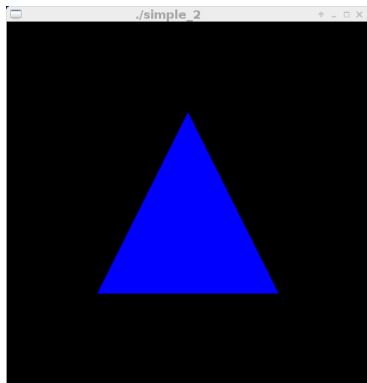
```
uniform mat4 projectionMatrix;  
uniform mat4 modelViewMatrix;
```

# Interpolation Qualifiers

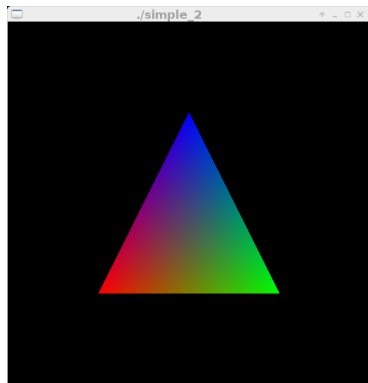
- outputs from and inputs to shader can be qualified with interpolation qualifier
- interpolation qualifier controls how value of particular variable is interpolated
- interpolation qualifiers: **smooth**, **noperspective**, **flat**
- **smooth** qualifier: perspective-correct interpolation is performed
- **noperspective** qualifier: linear interpolation is performed
- **flat** qualifier: no interpolation is performed (i.e., value taken from provoking vertex of primitive)
- default qualifier is **smooth**
- example:  
`flat out vec4 color;`

# Interpolation Example

- single triangle rendered with vertices having color attributes of red, green, and blue, with provoking vertex being last vertex



Without Interpolation (I.e., Flat)



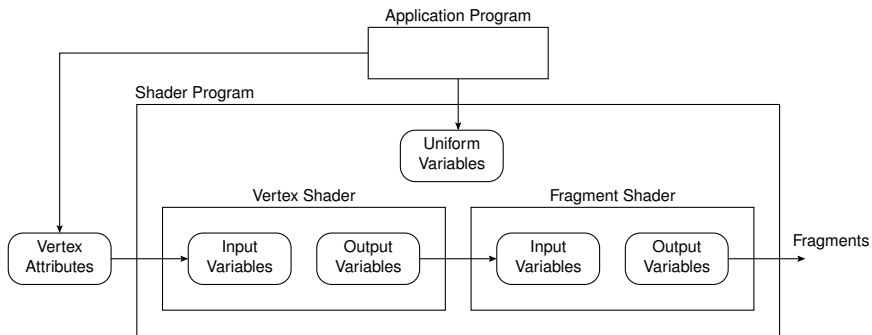
With Interpolation (E.g., Smooth)

- layout qualifiers used to specify how storage for variable allocated amongst other things
- layout qualifiers (e.g., `location`) provided by using **layout** keyword
- `location` layout qualifier can be used to specify location associated with variable
- vertex shaders allow input layout qualifiers on input variable declarations
- example: following will establish vertex shader input `vPosition` to be copied in from location number 1:  

```
layout(location = 1) in vec4 vPosition;
```
- example: following will establish vertex shader input `colors` copied in from location numbers 6, 7, and 8:  

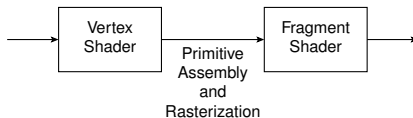
```
layout(location = 6) in vec4 colors[3];
```

# Configuration with Vertex and Fragment Shaders

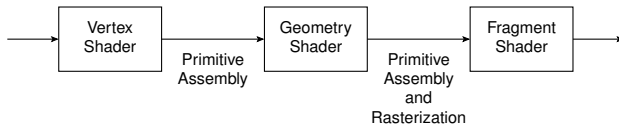




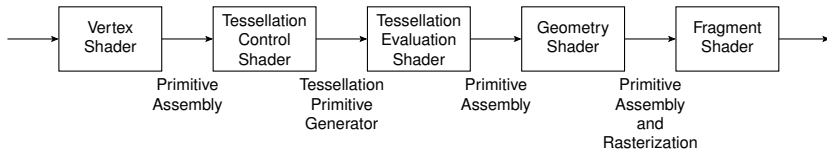
# Various Configurations of Shaders



## Vertex and Fragment Shaders



## Vertex, Geometry, and Fragment Shaders



## Vertex, Tessellation, Geometry, and Fragment Shaders

- vertex shader is programmable shader stage in rendering pipeline that handles processing of individual vertices
- vertex shader provided with vertex attribute data (e.g., position, normal, color, and texture coordinates) from VAO from drawing command
- for each vertex in input vertex stream, produces one vertex for output vertex stream
- must be one to one correspondence between input vertices and output vertices
- processes each vertex independently
- some uses of vertex shaders include:
  - vertex position transformation using modelview and projection matrices
  - normal transformation and (if needed) normalization
  - texture coordinate generation and transformation
  - per-vertex lighting
  - color computation

# Vertex Shader Inputs and Outputs

- built-in input variables:

- **int** `gl_VertexID`: index of vertex currently being processed
- **int** `gl_InstanceID`: index of current instance when doing some form of instanced rendering

- other inputs associated with vertex attributes from VAO/VBO

- built-in output variables:

- **vec4** `gl_Position`: clip-space output position of current vertex
- **float** `gl_PointSize`: pixel width/height of point being rasterized; only has meaning for point primitives
- **float** `gl_ClipDistance[]`: distance from vertex to each user-defined clipping half-space

- vertex shader must set `gl_Position`

# Vertex Shader Example

```
1  // use version 3.30 of GLSL (core profile)
2  #version 330
3
4  // input attribute variable for vertex position
5  in vec4 aPosition;
6
7  // uniform variable for modelview-projection matrix
8  uniform mat4 uModelViewProjMatrix;
9
10 void main() {
11     // set output position for vertex
12     gl_Position = uModelViewProjMatrix * aPosition;
13 }
```

- fragment shader is programmable shader stage that processes fragment generated by rasterization into set of colors and single depth value
- for each sample of pixels covered by primitive, fragment is generated
- each fragment has window space position, some other values, and all of interpolated per-vertex output values from last vertex processing stage
- takes single fragment as input and produces single fragment as output
- some uses of fragment shaders include:
  - per-fragment lighting
  - computing colors and texture coordinates per fragment
  - texture application (texture and bump mapping)
  - environment mapping
  - fog computation

# Fragment Shader Inputs and Outputs

- built-in input variables:

- **vec4** `gl_FragCoord`: location of fragment in window space
- **bool** `gl_FrontFacing`: indicates if fragment was generated by front face of primitive (only triangles can have back face)
- **int vec2** `gl_PointCoord`: location within point primitive that defines position of fragment relative to side of point

- other input variables correspond to outputs of previous shader stage

- built-in output variables:

- **float** `gl_fragDepth`: depth of fragment which defaults to `gl_FragCoord.z`

- **vec4** *output variable for fragment color*

# Fragment Shader Example

```
1  // use version 3.30 of GLSL (core profile)
2  #version 330
3
4  // output variable for color
5  out vec4 fColor;
6
7  void main() {
8      // set output color to white
9      fColor = vec4(1.0, 1.0, 1.0, 1.0);
10 }
```

- controls processing of primitives between vertex shader (or optional tessellation stage) and fixed-function vertex post-processing stage
- use of geometry shader optional
- takes single primitive as input and outputs zero or more primitives
- some uses of geometry shaders include:
  - layered rendering
  - transform feedback



# Geometry Shader Inputs

- one input primitive per geometry shader invocation
- type of input primitives specified by layout qualifier, which is one of:  
points, lines, lines\_adjacency, triangles,  
triangles\_adjacency
- number of input vertices determined by input primitive type (e.g., three for triangles)
- per-vertex inputs available as members of elements in array `gl_in`:
  - **vec4** `gl_Position`: vertex position
  - **float** `gl_PointSize`: pixel width/height of point being rasterized; only used for point primitive
  - **float** `gl_ClipDistance[]`: distance to clipping planes
- `gl_in` contains  $N$  elements (with indices starting from 0), where  $N$  is number of vertices in input primitive
- each shader input produced by previous pipeline stage is always array with one element per vertex
- per-primitive inputs:
  - `gl_PrimitiveIDIn`: current input primitive's ID
  - `gl_InvocationID`: current instance

# Geometry Shader Outputs

- type of output primitive generated specified by layout qualifier, which is one of: `points`, `line_strip`, `triangle_strip`
- can generate zero or more output primitives
- maximum number of vertices that can be generated specified by `max_vertices` layout qualifier
- per-vertex outputs:
  - **vec4** `gl_Position`: vertex position
  - **float** `gl_PointSize`: pixel width/height of point being rasterized; only used for point primitive
  - **float** `gl_ClipDistance[]`: distance to clipping planes
- per-primitive outputs:
  - **vec4** `gl_PrimitiveID`: primitive ID to pass to fragment shader
- `EmitVertex` called to process vertex outputs after all per-vertex outputs set
- after `EmitVertex` called, output variables have undefined values
- `EndPrimitive` called to signal end of primitive in order to start next output primitive
- not required to call `EndPrimitive` after last output primitive

# Geometry Shader Example (Passthrough)

```
1  // use version 3.30 of GLSL (core profile)
2  #version 330
3
4  // input primitives are triangles
5  layout(triangles) in;
6
7  // input variable for color
8  in vec3 vColor[];
9
10 // output primitives are triangle strips
11 // at most three vertices will be generated
12 layout(triangle_strip, max_vertices = 3) out;
13
14 // output variable for color
15 out vec3 gColor;
16
17 void main() {
18     // for each vertex of input triangle...
19     for (int i = 0; i < 3; ++i) {
20         // set position and color of output vertex
21         gl_Position = gl_in[i].gl_Position;
22         gColor = vColor[i];
23         // mark vertex as finished
24         EmitVertex();
25     }
26     EndPrimitive(); // optional
27 }
```

# Using Shader Programs

- shaders need to be compiled and linked to yield executable shader program
- OpenGL provides compiler and linker
- normally, program should have vertex and fragment shaders
- to generate executable shader program:
  - 1 create program via `glCreateProgram`
  - 2 for each shader in program:
    - 1 create shader via `glCreateShader`
    - 2 load shader source via `glShaderSource`
    - 3 compile shader source to object code via `glCompileShader` and check status of compile via `glGetShaderiv`
    - 4 attach shader object code to program via `glAttachShader`
  - 3 link program `glLinkProgram` and check status of link via `glGetProgramiv`
- shader program currently in use selected via `glUseProgram`
- shader and program can be deleted when no longer needed via `glDeleteShader` and `glDeleteProgram`

# Identifying Shader Variables in Application

- application program needs means to refer to attribute and uniform variables in shaders (e.g., in order to associate data with such variables)
- each attribute and uniform variable has integer identifier known as **location**
- location used as means to unambiguously name shader variable
- GLSL provides mechanism to force variable to have particular location via `location` layout qualifier
- location of variable can be queried by name (which is most useful when `location` layout qualifier not employed)
- can force attribute variable to use particular location via `glBindAttribLocation` prior to linking shader program

# Identifying Shader Variables in Application (Continued)

- get location of shader variable via `glGetAttribLocation`
- example: query location of attribute variable `aPosition`:

```
GLuint program; // shader program ID
// ...
GLint loc = glGetAttribLocation(program,
    "aPosition");
```

- get location of uniform variable via `glGetUniformLocation`
- example: query location of uniform variable `uModelViewProjMatrix`:

```
GLuint program; // shader program ID
// ...
GLint loc = glGetUniformLocation(program,
    "uModelViewProjMatrix");
```

# Associating Data in VAO with Attribute Variable

- application program needs to be able to associate shader attribute variable with data source (namely, data in VBO of VAO)
- to associate data in (VBO of) VAO with attribute variable in vertex shader, call `glVertexAttribPointer` when VAO/VBO containing attribute data is bound
- invocation of `glVertexAttribPointer` specifies:
  - location of vertex attribute variable
  - number of components per vertex attribute (e.g., 1, 2, 3, or 4)
  - type of each component (e.g., `GL_FLOAT` or `GL_DOUBLE`)
  - whether fixed-point values should be normalized (e.g., to  $[-1, 1]$  for signed values and  $[0, 1]$  for unsigned values)
  - stride (i.e., byte offset) between consecutive vertex attributes in array
  - offset of first component of first vertex attribute in array
- to enable use of attribute data associated with VAO, call `glEnableVertexAttribArray` when VAO containing attribute data is bound

# Example: Associating Data in VAO with Attribute Variable

## Part of Vertex Shader

```
1  in vec3 aPosition;
```

## Part of Application Program

```
1  GLuint program; // program ID
2  GLuint vao; // VAO ID
3  GLuint vbo; // VBO ID
4  GLuint offset; // offset of data in VBO
5  GLsizei stride; // stride of data in VBO
6  // ...
7  GLint loc = glGetAttribLocation(program,
8    "aPosition");
9  glBindVertexArray(vao);
10 glBindBuffer(GL_ARRAY_BUFFER, vbo);
11 glVertexAttribPointer(loc, 3, GL_FLOAT, GL_FALSE,
12   stride, reinterpret_cast<GLvoid*>(offset));
13 glEnableVertexAttribArray(loc);
```



# Accessing Uniform Variables from Application Program

- application program needs to be able to access uniform variables in shader
- application can only write to uniform variables since data flows in one direction only (i.e., from application to shader)
- uniform variable identified by location
- to modify uniform variable, must know its location
- modify uniform variable via `glUniform*` (which identifies variable to change by its location)
- example:

## Part of Shader

```
uniform float uTime;
```

## Part of Application Program

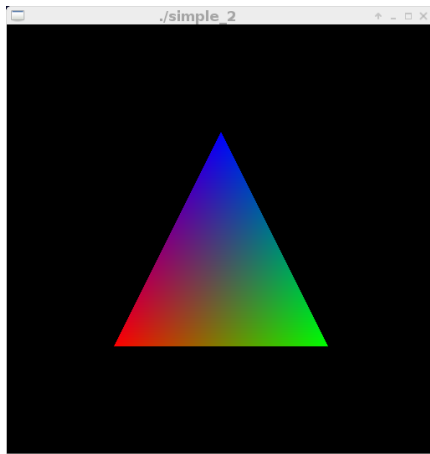
```
GLuint program; // shader program ID  
// ...  
GLint loc = glGetUniformLocation(program, "uTime");  
glUniform1f(loc, 1.5f);
```

## Section 5.6.3

### **Shader Examples**

# Simple: Shader Example

- vertex shader provided with two attributes per vertex (position and color)
- want smooth interpolation of color across faces
- rendering output shown below for mesh consisting of single triangle



# Simple: Vertex and Fragment Shaders

## Vertex Shader

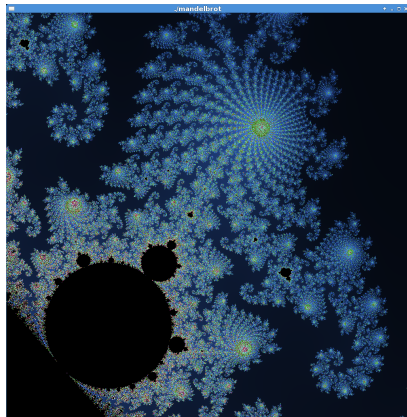
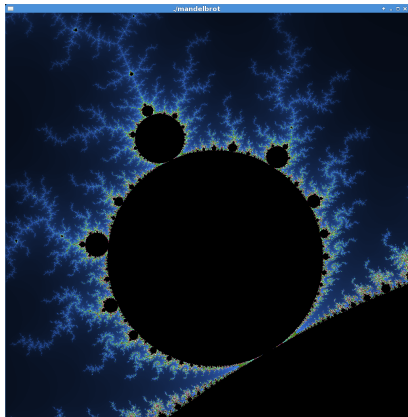
```
1  #version 330
2
3  in vec4 aPosition; // input vertex position attribute
4  in vec4 aColor; // input vertex color attribute
5
6  out vec4 vColor; // output vertex color (interpolated)
7
8  // uniform variable for modelview-projection
9  // matrix product
10 uniform mat4 uModelViewProjMatrix;
11
12 void main() {
13     vColor = aColor;
14     gl_Position = uModelViewProjMatrix * aPosition;
15 }
```

## Fragment Shader

```
1  #version 330
2
3  in vec4 vColor; // input color (interpolated)
4
5  out vec4 fColor; // output fragment color
6
7  void main() {
8     fColor = vColor;
9  }
```

# Mandelbrot: Shader Example

- render triangles to cover entire drawing area and texture map Mandelbrot set onto triangles using fragment shader
- some examples of rendering results shown below

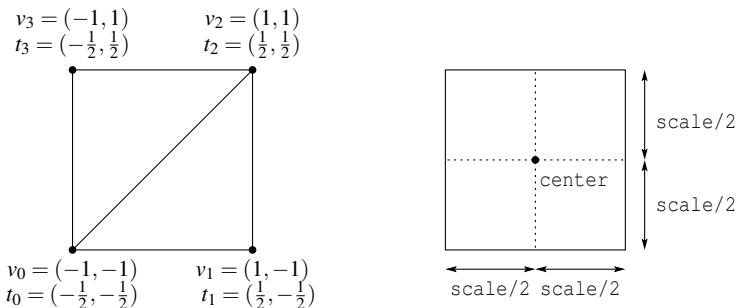


- **Mandelbrot set:** set of all complex numbers  $c$  such that sequence  $z_0, z_1, z_2, \dots$  does not tend toward infinity, where

$$z_n = \begin{cases} z_{n-1}^2 + c & \text{if } n \geq 1 \\ c & \text{if } n = 0 \end{cases}$$

- associate rectangular region in complex plane with graphics viewport
- for point corresponding to each pixel in viewport, determine number of steps in above iterative process for which result does not become too large (i.e., tending towards infinity)
- assign color to each pixel depending on obtained iteration count

# Mandelbrot: Application Program



- application program simply renders two triangles that cover full extent of viewport ( $\{v_k\}$  are positional coordinates;  $\{t_k\}$  are texture coordinates)
- texture coordinate region  $[-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$  corresponds to full viewport
- square region in complex plane of width/height `scale` centered at point `center` is mapped onto region  $[-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$  in texture coordinates

# Mandelbrot: Vertex Shader

```
1  #version 330
2
3  in vec3 aPosition; // position vertex attribute
4  in vec3 aTexCoord; // texture-coordinate vertex attribute
5
6  out vec3 vTexCoord; // texture coordinate (interpolated)
7
8  void main() {
9      vTexCoord = aTexCoord;
10     gl_Position = vec4(aPosition, 1.0);
11 }
```

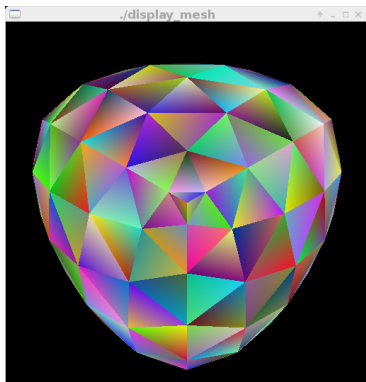


# Mandelbrot: Fragment Shader

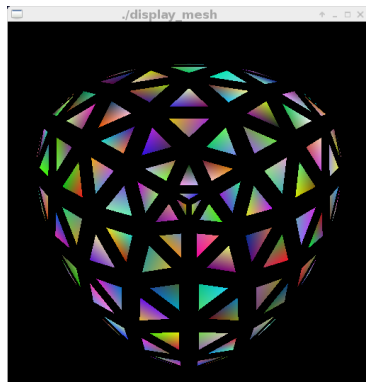
```
1  #version 330
2
3  in vec3 vTexCoord; // texture coordinates
4
5  out vec4 fColor; // vertex color
6
7  uniform vec2 center; // center of viewing region
8  uniform float scale; // width/height of viewing region
9  uniform int maxIters; // maximum iteration count
10
11 int mandelbrot(vec2 c) {
12     vec2 z = vec2(0.0, 0.0);
13     int i;
14     for (i = 0; i < maxIters; ++i) {
15         z = vec2(z.x * z.x - z.y * z.y + c.x, 2.0 * z.x * z.y + c.y);
16         if (length(z) > 2.0) {break;}
17     }
18     return i;
19 }
20
21 float lookup(float x, float c) {return c * mod(x, 1.0 / c);}
22
23 void main() {
24     int i = mandelbrot(vec2(scale * vTexCoord.x + center.x,
25                             scale * vTexCoord.y + center.y));
26     float t = float(i) / maxIters;
27     fColor = vec4(lookup(t, 2.0), lookup(t, 4.0), lookup(t, 8.0), 1.0);
28 }
```

# ShrinkFace: Shader Example

- use geometry shader to shrink triangles sent to rendering pipeline
- triangles contracted towards their centroid so that triangles that were originally touching now have gap between them
- example rendering results are shown below

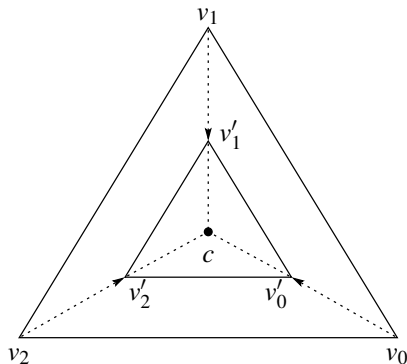


Rendered Normally

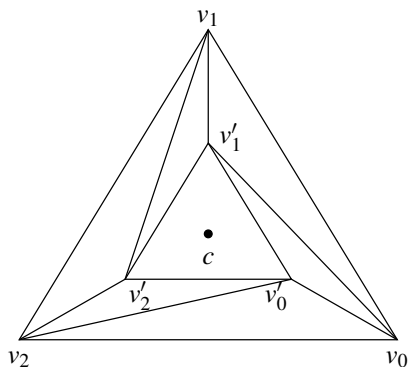


Rendered with Shrunk Faces

# ShrinkFace: Triangle Shrinking



- each vertex  $v_k$  moved in direction of centroid  $c$  to new position  $v'_k = \frac{1}{2}(v_k + c)$  (i.e., midpoint of  $v_k$  and  $c$ )
- gap formed by shrinking of triangle is filled with new triangles drawn in black



- gap can be filled with triangle strip with vertices:  $v_2, v'_2, v_1, v'_1, v_0, v'_0, v_2, v'_2$

# ShrinkFace: Vertex Shader

```
1  #version 330
2
3  in vec3 aPosition; // position vertex attribute
4  in vec3 aColor; // color vertex attribute
5
6  out vec3 vColor; // color (interpolated)
7
8  void main() {
9      gl_Position = vec4(aPosition, 1.0);
10     vColor = aColor;
11 }
```

# ShrinkFace: Geometry Shader

```
1  #version 330
2
3  layout(triangles) in; // triangle primitives as input
4  in vec3 vColor[]; // input vertex colors
5
6  layout(triangle_strip, max_vertices=11) out;
7  // triangle strips as output; at most 11 vertices
8  out vec3 gColor; // output color (interpolated)
9
10 uniform mat4 uModelViewProjMatrix;
11 // modelview-projection matrix product
```

# ShrinkFace: Geometry Shader (Continued)

```
13 void main() {
14     vec3 v[6];
15     for (int i = 0; i < 3; ++i) {v[i] = gl_in[i].gl_Position.xyz;}
16
17     // compute centroid of triangle
18     vec3 c = (v[0] + v[1] + v[2]) / 3.0;
19
20     // compute vertices of shrunk triangle and generate
21     // triangle strip consisting only of shrunk triangle
22     for (int i = 0; i < 3; ++i) {
23         v[i + 3] = c + 0.5 * (v[i] - c);
24         gl_Position = uModelViewProjMatrix * vec4(v[i + 3], 1.0);
25         gColor = vColor[i];
26         EmitVertex();
27     }
28     EndPrimitive();
29
30     // generate triangle strip to fill gap between triangles
31     // introduced by shrinking
32     const int lut[] = int[](2, 5, 1, 4, 0, 3, 2, 5);
33     for (int i = 0; i < 8; ++i) {
34         gl_Position = uModelViewProjMatrix * vec4(v[lut[i]], 1.0);
35         gColor = vec3(0.0, 0.0, 0.0);
36         EmitVertex();
37     }
38 }
```

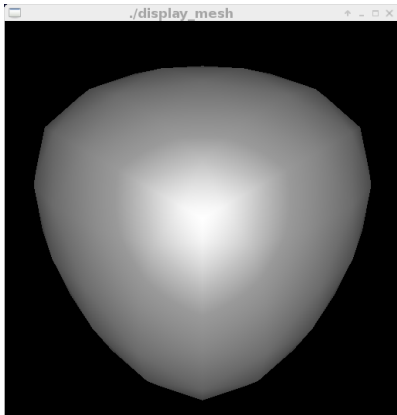
# ShrinkFace: Fragment Shader

```
1  #version 330
2
3  in vec3 gColor; // input color
4
5  out vec4 fColor; // output color
6
7  void main() {
8      fColor = vec4(gColor, 1.0);
9  }
```

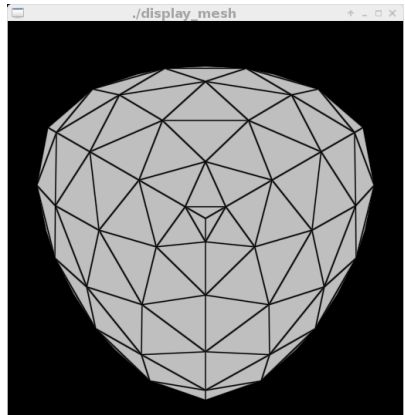


# Wireframe: Shader Example

- use geometry shader to assist in superimposing wireframe on rendered surface
- example rendering output shown below

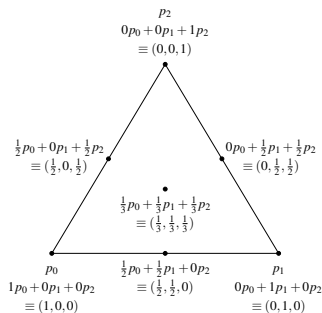


Without Edges Shown



Edges Shown Using Shader

# Wireframe: General Approach



- points on edge of triangle must have exactly one or two barycentric coordinates equal to zero, while points in the interior must have three nonzero coordinates
- if at least one of barycentric coordinates is small, must be in vicinity of edge
- if in vicinity of edge, use different color

# Wireframe: Vertex Shader

```
1  #version 330
2
3  in vec3 aPosition; // position vertex attribute
4  in vec3 aColor; // color vertex attribute
5
6  out vec3 vColor; // output color (interpolated)
7
8  uniform mat4 uModelViewProjMatrix;
9  // modelview-projection matrix product
10
11 void main() {
12     gl_Position = uModelViewProjMatrix * vec4(aPosition, 1.0);
13     vColor = aColor;
14 }
```

# Wireframe: Geometry Shader

```
1  #version 330
2
3  layout(triangles) in; // triangles as input
4  in vec3 vColor[]; // vertex colors
5
6  layout(triangle_strip, max_vertices=3) out;
7  // triangle strips as output; at most 3 vertices
8  out vec3 gColor; // output color
9  noperspective out vec3 gBaryCoord;
10 // output barycentric coordinates (interpolated)
11
12 void main() {
13     const vec3 lut[3] = vec3[3](
14         vec3(1.0, 0.0, 0.0),
15         vec3(0.0, 1.0, 0.0),
16         vec3(0.0, 0.0, 1.0));
17     for (int i = 0; i < 3; ++i) {
18         gl_Position = gl_in[i].gl_Position;
19         gBaryCoord = lut[i];
20         gColor = vColor[i];
21         EmitVertex();
22     }
23 }
```

# Wireframe: Fragment Shader

```
1  #version 330
2
3  in vec3 gColor; // input color
4  noperspective in vec3 gBaryCoord;
5  // input barycentric coordinates
6
7  out vec4 fColor; // output color
8
9  void main() {
10     const vec3 edgeColor = vec3(0.0, 0.0, 0.0);
11     const float edgeWidth = 1.0;
12     vec3 d = fwidth(gBaryCoord);
13     vec3 a3 = smoothstep(vec3(0.0), d * edgeWidth, gBaryCoord);
14     float v = min(min(a3.x, a3.y), a3.z);
15     fColor = vec4(mix(edgeColor, gColor, v), 1.0);
16 }
```

- upper threshold for `smoothstep` chosen relative to approximate gradient magnitude so thickness of edges in wireframe same regardless of triangle size
- simpler code for calculating `a3` shown below would cause thickness of edges in wireframe to depend on triangle size, which would be less aesthetically pleasing:

```
vec3 a3 = smoothstep(vec3(0.0), vec3(0.02), gBaryCoord);
```

# Ambient-Diffuse-Specular (ADS) Lighting Model

## ■ light properties:

- $\ell_a$ : ambient component of light source
- $\ell_d$ : diffuse component of light source
- $\ell_s$ : specular component of light source

## ■ material properties:

- $k_a$ : ambient reflection constant
- $k_d$ : diffuse reflection constant
- $k_s$ : specular reflection constant
- $\alpha$ : shininess constant

## ■ vectors:

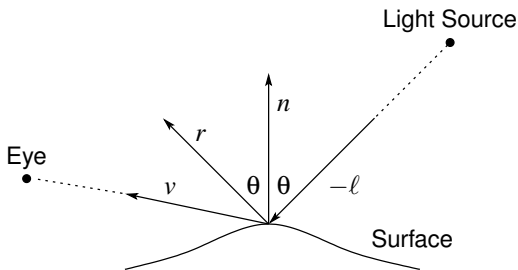
- $\ell$ : unit vector vector in direction from point on surface to light source
- $n$ : unit normal at point on surface
- $v$ : unit vector in direction from point on surface to viewer
- $r$ : unit vector in direction that perfectly reflected light ray would take from this point on surface (i.e.,  $r = 2(\ell \cdot n)n - \ell$ )

## ■ illumination $i$ of point on surface given by:

$$i = k_a \ell_a + \max\{(\ell \cdot n), 0\} k_d \ell_d + \max\{(r \cdot v)^\alpha, 0\} u(\ell \cdot n) k_s \ell_s$$

where  $u$  is unit-step function

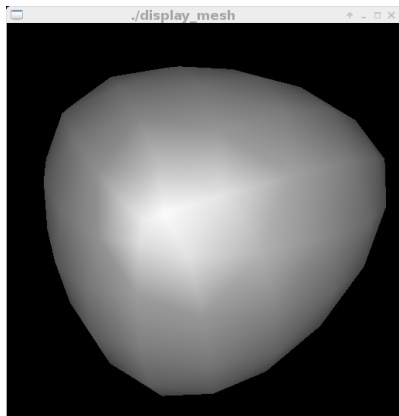
# ADS Lighting Model: Diagram



- $\ell$ : unit vector in direction from point on surface to light source
- $n$ : unit normal at point on surface
- $v$ : unit vector in direction from point on surface to viewer
- $r$ : unit vector in direction that perfectly reflected light ray would take from this point on surface

# Per-Vertex Lighting: Shader Example

- per-vertex lighting using ambient-diffuse-specular (ADS) model
- example rendering result shown below





# Per-Vertex Lighting: Vertex Shader

```
1  #version 330
2
3  in vec3 aPosition; // position vertex attribute
4  in vec3 aNormal; // normal vertex attribute
5
6  out vec3 vColor; // output color (interpolated)
7
8  uniform mat4 uModelViewMatrix; // modelview matrix
9  uniform mat3 uNormalMatrix; // normal transformation matrix
10 uniform mat4 uModelViewProjMatrix;
11     // modelview-projection matrix product
12
13 struct LightSourceParams {
14     vec4 position; // position
15     vec3 ambient; // ambient component
16     vec3 diffuse; // diffuse component
17     vec3 specular; // specular component
18 };
19 uniform LightSourceParams uLight; // light parameters
20
21 struct MaterialParams {
22     vec3 ambient; // ambient reflectance
23     vec3 diffuse; // diffuse reflectance
24     vec3 specular; // specular reflectance
25     float shininess; // specular exponent
26 };
27 uniform MaterialParams uMaterial; // material parameters
```

# Per-Vertex Lighting: Vertex Shader (Continued)

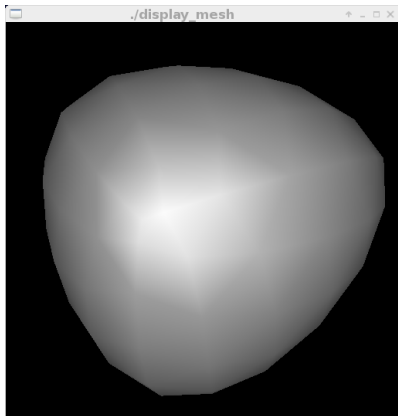
```
29  vec3 ads(vec4 position, vec3 normal) {
30      vec3 s = normalize(vec3(uLight.position - position));
31      vec3 v = normalize(-position.xyz);
32      vec3 r = reflect(-s, normal);
33      float sn = dot(s, normal);
34      vec3 ambient = uLight.ambient * uMaterial.ambient;
35      vec3 diffuse = uLight.diffuse * uMaterial.diffuse *
36          max(sn, 0.0);
37      diffuse = clamp(diffuse, 0.0, 1.0);
38      vec3 specular = (sn > 0.0) ? (uLight.specular *
39          uMaterial.specular * pow(max(dot(r, v), 0.0),
40          uMaterial.shininess)) : vec3(0.0);
41      specular = clamp(specular, 0.0, 1.0);
42      return clamp(ambient + diffuse + specular, 0.0, 1.0);
43  }
44
45  void main() {
46      vec3 eyeNorm = normalize(uNormalMatrix * aNormal);
47      vec4 eyePos = uModelViewMatrix * vec4(aPosition, 1.0);
48      vColor = ads(eyePos, eyeNorm);
49      gl_Position = uModelViewProjMatrix *
50          vec4(aPosition, 1.0);
51  }
```

# Per-Vertex Lighting: Fragment Shader

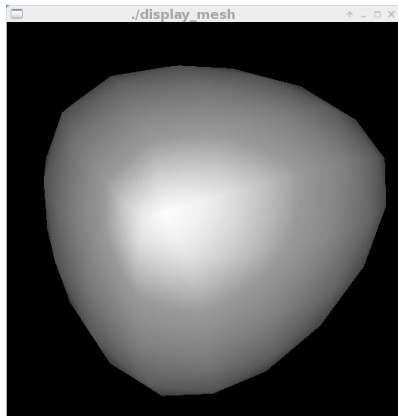
```
1 #version 330
2
3 in vec3 vColor; // input color
4
5 out vec4 fColor; // output color
6
7 void main() {
8     fColor = vec4(vColor, 1.0);
9 }
```

# Per-Fragment Lighting: Shader Example

- per-fragment lighting using ambient-diffuse-specular (ADS) model
- example rendering result shown along with per-vertex lighting result for comparison



Per-Vertex Lighting



Per-Fragment Lighting

# Per-Fragment Lighting: Vertex Shader

```
1  #version 330
2
3  in vec3 aPosition; // position vertex attribute
4  in vec3 aNormal; // normal vertex attribute
5
6  out vec3 vPosition; // output position (interpolated)
7  out vec3 vNormal; // output normal (interpolated)
8
9  uniform mat4 uModelViewMatrix; // modelview matrix
10 uniform mat3 uNormalMatrix; // normal transformation matrix
11 uniform mat4 uModelViewProjMatrix;
12     // modelview-projection matrix product
13
14 void main() {
15     vNormal = normalize(uNormalMatrix * aNormal);
16     vPosition = vec3(uModelViewMatrix * vec4(aPosition, 1.0));
17     gl_Position = uModelViewProjMatrix * vec4(aPosition, 1.0);
18 }
```

# Per-Fragment Lighting: Fragment Shader

```
1  #version 330
2
3  in vec3 vNormal; // input normal
4  in vec3 vPosition; // input position
5
6  out vec4 fColor; // output color
7
8  struct LightSourceParams {
9      vec4 position; // position
10     vec3 ambient; // ambient component
11     vec3 diffuse; // diffuse component
12     vec3 specular; // specular component
13 };
14 uniform LightSourceParams uLight; // light parameters
15
16 struct MaterialParams {
17     vec3 ambient; // ambient reflectance
18     vec3 diffuse; // diffuse reflectance
19     vec3 specular; // specular reflectance
20     float shininess; // specular exponent
21 };
22 uniform MaterialParams uMaterial; // material parameters
```

# Per-Fragment Lighting: Fragment Shader (Continued)

```
24  vec3 ads(vec4 position, vec3 normal) {
25      vec3 s = normalize(vec3(uLight.position - position));
26      vec3 v = normalize(-position.xyz);
27      vec3 r = reflect(-s, normal);
28      float sn = dot(s, normal);
29      vec3 ambient = uLight.ambient * uMaterial.ambient;
30      vec3 diffuse = uLight.diffuse * uMaterial.diffuse *
31          max(sn, 0.0);
32      diffuse = clamp(diffuse, 0.0, 1.0);
33      vec3 specular = (sn > 0.0) ? uLight.specular *
34          uMaterial.specular * pow(max(dot(r, v), 0.0),
35          uMaterial.shininess) : vec3(0.0);
36      specular = clamp(specular, 0.0, 1.0);
37      return clamp(ambient + diffuse + specular, 0.0, 1.0);
38  }
39
40  void main() {
41      fColor = vec4(ads(vec4(vPosition, 1.0), vNormal), 1.0);
42  }
```

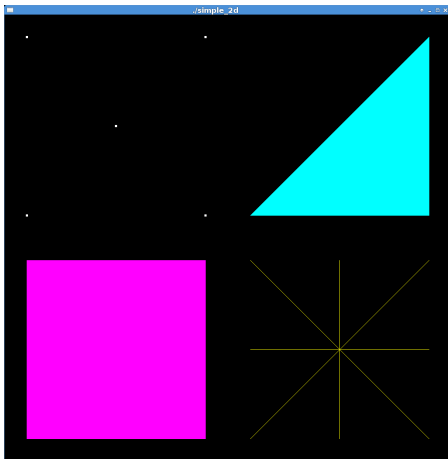
## Section 5.6.4

### **OpenGL Example Programs**



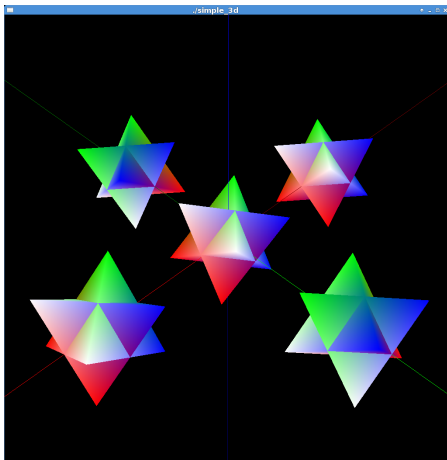
# OpenGL Example Program: simple\_2d

- simple 2-D graphics
- draws points, lines, triangle, and quadrilateral



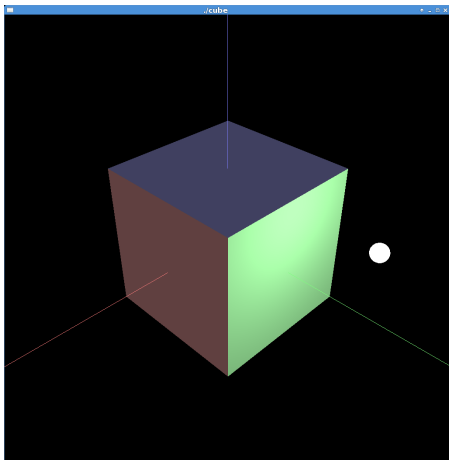
# OpenGL Example Program: simple\_3d

- simple 3-D graphics
- draws and animates several simple polyhedra



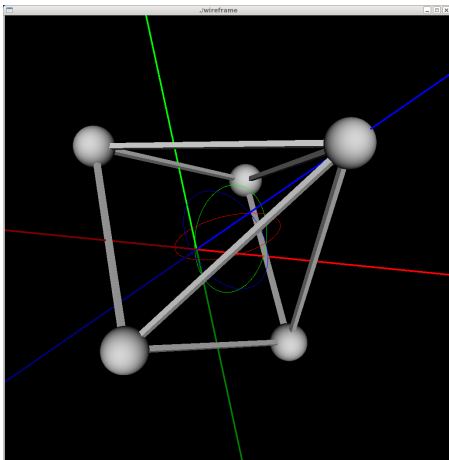
# OpenGL Example Program: cube

- 3-D graphics with lighting
- draws cube with lighting



# OpenGL/CGAL Example Program: wireframe

- wireframe mesh viewer
- allows polygon mesh to viewed as wireframe



## Section 5.6.5

### **References**

- 1 D. Shreiner, G. Sellers, J. Kessenich, and B. Licea-Kane. *OpenGL Programming Guide*. Addison-Wesley, Upper Saddle River, NJ, USA, 8th edition, 2013.
- 2 R. S. Wright Jr., N. Haemel, G. Sellers, and B. Lipchak. *OpenGL Superbible*. Addison-Wesley, Upper Saddle River, NJ, USA, 5th edition, 2011.
- 3 E. Angel and D. Shreiner. *Interactive Compute Graphics — A Top-Down Approach with Shader-Based OpenGL*. Addison-Wesley, Boston, MA, USA, 6th edition, 2012.
- 4 M. Bailey and S. Cunningham. *Graphics Shaders — Theory and Practice*. CRC Press, Boca Raton, FL, USA, 2nd edition, 2012.
- 5 R. J. Rost. *OpenGL Shading Language*. Addison-Wesley, Boston, MA, USA, 2nd edition, 2006.

- 6 D. Wolff. *OpenGL 4.0 Shading Language Cookbook*. Packt Publishing, Birmingham, UK, 2011.
- 7 The OpenGL graphics system: A specification (version 4.4 (core profile)), Mar. 2014.
- 8 The OpenGL shading language — language version 4.40, June 2014.
- 9 OpenGL Web Site, <http://www.opengl.org>.
- 10 OpenGL Software Development Kit (SDK), <https://www.opengl.org/sdk> (full documentation on each OpenGL function can be found at <http://www.opengl.org/sdk/docs/man>).
- 11 Khronos Group on YouTube, <https://www.youtube.com/user/khronosgroup>.

- 1 Ed Angel and Dave Shreiner. An Introduction to OpenGL Programming. SIGGRAPH 2013, Available online at <https://youtu.be/6-9XFm7XAT8>.



- OpenGL Extension Wrangler Library (GLEW)

<http://glew.sourceforge.net>

<http://www.opengl.org/sdk/libs/GLEW>

- OpenGL FrameWork (GLFW) Library

<http://www.glfw.org>

- OpenGL Utility Toolkit (GLUT) Library

<http://sourceforge.net/projects/freeglut>

- OpenGL Mathematics (GLM) Library

<http://glm.g-truc.net>

- Qt Library

<http://www.qt.io>

<http://www.qt.io/developers>

## Section 5.7

### Other Libraries

## ■ Eigen

- C++ library for linear algebra
- web site: <http://eigen.tuxfamily.org>

## ■ Lapack++

- C++ library for high-performance linear-algebra computations
- C++ wrapper for LAPACK and BLAS
- web site: <http://lapackpp.sourceforge.net>

## ■ Armadillo

- C++ library for linear algebra
- web site: <http://arma.sourceforge.net>

## ■ GNU Scientific Library

- C library for numerical analysis
- web site: <http://www.gnu.org/software/gsl>

## ■ GNU Multiprecision Library

- C library for arbitrary-precision arithmetic
- web site: <http://gmplib.org>

## ■ Boost.uBLAS

- C++ library for numerical computation
- web site:

<http://www.boost.org/doc/libs/release/libs/numeric/ublas>

## ■ Boost.Rational

- C++ rational number library
- web site: [www.boost.org/doc/libs/release/libs/rational](http://www.boost.org/doc/libs/release/libs/rational)

## ■ Boost.Interval

- C++ interval arithmetic library
- web site: [www.boost.org/doc/libs/release/libs/numeric/interval/doc/interval.htm](http://www.boost.org/doc/libs/release/libs/numeric/interval/doc/interval.htm)

## ■ Boost.Math

- C++ library
- provides math constants, GCD, LCM, quaternions, and more
- web site: <http://www.boost.org/doc/libs/release/libs/math>

## ■ Linear Algebra Package (LAPACK)

- Fortran library for numerical computing

- web site: <http://www.netlib.org/lapack>

## ■ Basic Linear Algebra Subprograms (BLAS)

- de facto API for publishing libraries to perform basic linear algebra operations
- written in Fortran
- web site: <http://www.netlib.org/blas>

## Part 6

# Programming

## Section 6.1

### **Good Programming Practices**

# Formatting, Naming, Documenting

- Be consistent with the *formatting* of the source code (e.g., indentation strategy, tabs versus spaces, spacing, brackets/parentheses).
- Avoid a formatting style that runs against common practices.
- Be consistent in the *naming conventions* used for identifiers (e.g., names of objects, functions, namespaces, types) and files.
- Avoid bizarre naming conventions that run against common practices.
- *Comment* your code. If code is well documented, it should be possible to quickly ascertain what the code is doing without any prior knowledge of the code.
- Use *meaningful names* for identifiers (e.g., names of objects, functions, types, etc.). This improves the readability of code.
- Avoid *magic literal constants*. Define a constant object and give it a meaningful name.

```
constexpr double miles_per_kilometer = 0.621371;
```



- If a program requires that certain *constraints on user input* be satisfied in order to work correctly, do not assume that these constraints will be satisfied. Instead, always check them.
- Always handle errors *gracefully*.
- Provide *useful* error messages.
- Always *check return codes*. Even if the operation/function theoretically cannot fail (under the assumption of bug-free code), in practice it may fail due to a bug.
- If an operation is performed that can fail, check the *status of the operation* to ensure that it did not fail (even if you think that it should not fail). For example, check for error conditions on streams.
- If a function can fail, always check its *return value*.

- Do not *unnecessarily complicate* code. Use the simplest solution that will meet the needs of the problem at hand.
- Do not impose *bogus limitations*. If a more general case can be handled without complicating the code and this more general case is likely to be helpful to handle, then handle this case.
- Do not *unnecessarily optimize* code. Highly optimized code is often much less readable. Also, highly optimized code is often more difficult to write correctly (i.e., without bugs). Do not write grossly inefficient code that is obviously going to cause performance problems, but do not optimize things beyond avoiding gross inefficiencies that you know will cause performance problems.

- Avoid *duplication* of code. If similar code is needed in more than one place, put the code in a function. Also, utilize templates to avoid code duplication.
- The avoidance of code duplication has many advantages.
  - 1 It simplifies code understanding. (Understand once, instead of  $n$  times.)
  - 2 It simplifies testing. (Test once, instead of  $n$  times.)
  - 3 It simplifies debugging. (Fix bugs in one place, instead of  $n$  places.)
  - 4 It simplifies code maintenance. (Change code in one place, instead of  $n$  places.)
- Make good use of the available *libraries*. Do not reinvent the wheel. If a library provides code with the needed functionality, use the code in the library.

- Avoid *multiple returns paths* (i.e., multiple points of exit) in functions when they serve to complicate (rather than simplify) code structure.
- Whenever possible, avoid the use of *global state* (i.e., global variables).
- Ensure that the code is *const correct*.
- If an object does not need to change, make it `const` (or `constexpr`). This improves the readability of code. This also helps to ensure const correctness of code.
- Whenever possible, avoid placing identifiers (such as the names of variables, functions, and types) in the global namespace.
- Avoid bringing many unknown identifiers into scope. For example, avoid constructs like:

```
using namespace std;
```

Only bring identifiers into scope if they are truly needed.

- Whenever possible, avoid relying on behavior that is *not promised by the language*, as this can lead to brittle code. Avoid invoking undefined behavior. Whenever possible, avoid relying on unspecified or implementation-defined behavior.
- Whenever possible, avoid relying on *undocumented features* of libraries, as this can lead to brittle code.
- Enable *compiler warning messages*. Pay attention to warning messages issued by the compiler.
- Learn how to use a *source-level debugger*. There will be times when you will absolutely need it.
- Be careful to avoid using references, pointers, iterators that do not reference valid data. Always be clear about which operations invalidate references, pointers, and iterators.

# Testing: Preconditions and Postconditions

- **precondition**: condition that must be true before function is called
- for example, precondition for function that computes square root of  $x$ :  
 $x \geq 0$
- **postcondition**: condition that must be true after function is called
- for example, postcondition for function that removes entry from table of size  $n$ : new size of table  $n - 1$
- whenever feasible, check for violations of preconditions and postconditions for functions
- if precondition or postcondition is violated at run time, terminate program immediately in order to help in localizing bug (e.g., by using `assert` or calling `std::abort` or `std::terminate`)

- The single most important thing when writing code is that it does the job it was intended to do *correctly*. That is, there should not be any bugs.
- *Test* your code. If you do not spend as much time testing your code as you do writing it, you are likely not doing enough testing.
- Tests should exercise as much of the code as possible (i.e., provide good *code coverage*).
- Design and structure your code so that it is easy to test. In other words, testing should be considered *during design*.
- Your code will have bugs. Design your code so that it will help you to isolate bugs. Use *assertions*. Use *preconditions* and *postconditions*.
- Design your code so that is modular and can be written and tested *in pieces*. The first testing of the software should never be testing the entire software as a whole.
- Often in order to adequately test code, one has to write separate *specialized test code*.

- subscribing operator for 1-D array class:

```
template <class T>
const T& Array_1<T>::operator[](int i) const {
    // Precondition: index is in allowable range
    assert(i >= 0 && i < data_.size());
    return data_[i];
}
```

- function taking pointer parameter:

```
int stringLength(const char* ptr) {
    // Precondition: pointer is not null
    assert(ptr);
    // Code to compute and return string length.
    // ...
}
```

- function that modifies highly complicated data structure:

```
void modifyDataStructure(Type& dataStructure) {
    // Precondition: data structure is in valid state
    assert(isDataStructureValid(dataStructure));
    // Complicated code to update data structure.
    // ...
    // Postcondition: data structure is in valid state
    assert(isDataStructureValid(dataStructure));
}
```



## Section 6.2

# Algorithms

# Software Performance

- two most basic performance measures, which are often of most interest:
  - 1 time complexity
  - 2 space complexity
- **time complexity**: amount of time required to execute code
- **space complexity**: amount of memory needed for code execution
- normally must consider both time and space complexities, since one type of complexity can often be traded off for other
- from practical standpoint, real-world time and memory usage are what matter most (as opposed to some approximate theoretical measures of code complexity)
- need techniques that can provide guidance when designing software so that more likely that later implementation (of design) will have acceptable performance
- many factors can potentially impact performance, including:
  - CPU instruction count
  - cache efficiency
  - degree of parallelism and concurrency
  - resource utilization (e.g., memory, disk, and network)

# Random-Access Machine (RAM) Model

- algorithms can be measured in machine-independent way using random-access machine (RAM) model
- model assumes single processor
- instructions executed sequentially with no concurrent operations
- elementary types: integer and floating point numbers
- each elementary operation takes one time unit
- elementary operations include:
  - arithmetic operations (e.g., addition, subtraction, multiplication, division) on elementary types
  - loads and stores of elementary types
  - branch operations (e.g., conditional branch, jump)
  - subroutine call
- loops and subroutines are not considered elementary operations, but rather as composition of numerous elementary operations
- each memory access takes one time unit
- unbounded amount of memory available

# Worst-Case, Average, and Amortized Complexity

- complexity expressed as function of input problem size
- **worst-case complexity**: gives upper bound on complexity of algorithm for any input of given size
- **average complexity**: gives average complexity of algorithm in statistical sense if probability measure assigned to all inputs of given size
- often algorithm may only approach worst-case complexity for very small fraction of possible inputs, in which case average complexity might be more practically useful than worst-case complexity
- sometimes algorithm may be invoked many times and cost of single invocation difficult to determine in isolation (e.g., time complexity of `push_back` member function of `std::vector`)
- **amortized complexity**: complexity per invocation of algorithm evaluated over sequence of invocations
- amortized complexity makes guarantee about total expense of sequence of invocations of algorithm, rather than single invocation (e.g., `push_back` member function of `std::vector` takes amortized constant time)

# Asymptotic Analysis of Algorithms

- asymptotic analysis deals with behavior of algorithm as problem size becomes arbitrarily large
- **asymptotic complexity**: complexity of algorithm in limit as problem size becomes *arbitrarily large*
- often interested in:
  - asymptotic time complexity
  - asymptotic space complexity
- asymptotic time and space complexities of algorithm often much *easier to determine* than exact running time and memory usage
- often (but not always!) algorithm that is asymptotically more efficient will be best choice for all but very small inputs
- asymptotic notation (to be discussed next) provides way to describe functions that is very useful for asymptotic analysis

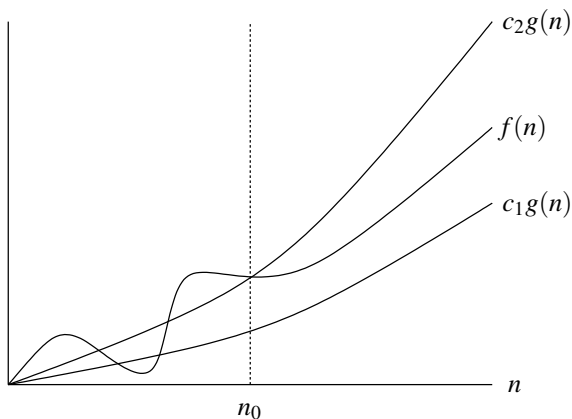
# Big-Theta ( $\Theta$ ) Notation

- **big-theta ( $\Theta$ ) notation**: for function  $g$ ,  $\Theta(g)$  denotes set of all functions  $f$  for which positive constants  $c_1$ ,  $c_2$ , and  $n_0$  exist such that

$$0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n) \quad \text{for all } n \geq n_0$$

- functions in  $\Theta(g)$  grow asymptotically at **same** rate as  $g$  (to within constant factor)
- effectively,  $f(n)$  is sandwiched between  $c_1 g(n)$  and  $c_2 g(n)$  for sufficiently large  $n$  (i.e.,  $n \geq n_0$ )
- used to provide (asymptotic) **lower and upper bounds** on function, each to within constant factor (provides asymptotically tight bound)
- if  $f \in \Theta(g)$ , then for sufficiently large  $n$ ,  $f(n)$  equals  $g(n)$  to within constant factor
- examples:
  - $f(n) = an^2 + bn + c$  where  $a, b, c$  are constants and  $a > 0$ ;  
 $f \in \Theta(n^2)$  but  $f \notin \Theta(n)$  and  $f \notin \Theta(n^3)$
  - $f(n) = \sum_{i=0}^d a_i n^i$  where  $\{a_i\}$  are constants and  $a_d > 0$ ;  
 $f \in \Theta(n^d)$  but  $f \notin \Theta(n^{d+1})$  and  $f \notin \Theta(n^{d-1})$

## Big-Theta ( $\Theta$ ) Notation (Continued)



- $f \in \Theta(g)$
- for  $n \geq n_0$ ,  $f(n)$  is *lower bounded* by  $c_1g(n)$  and *upper bounded* by  $c_2g(n)$
- asymptotically,  $f$  grows at *same* rate as  $g$  to within constant factor

# Big-Oh ( $O$ ) Notation

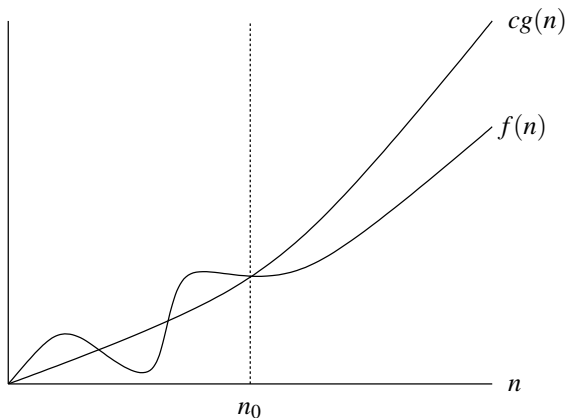
- **big-oh ( $O$ ) notation**: for function  $g$ ,  $O(g)$  denotes set of all functions  $f$  for which positive constants  $c$  and  $n_0$  exist such that

$$0 \leq f(n) \leq cg(n) \quad \text{for all } n \geq n_0$$

- functions in  $O(g)$  grow asymptotically at rate **at most** that of  $g$  (to within constant factor)
- used to provide (asymptotic) **upper bound** on function to within constant factor
- if  $f \in O(g)$ , then for sufficiently large  $n$ ,  $f(n)$  is less than or equal to  $g(n)$  to within constant factor
- since  $\Theta(g(n)) \subset O(g(n))$ ,  $f(n) \in \Theta(g(n))$  implies  $f(n) \in O(g(n))$
- often used to bound worst-case running time of algorithm
- examples:
  - $f(n) = 3n^2 + 2n + 1$ ;  $f \in O(n^2)$  and  $f \in O(n^3)$  but  $f \notin O(n)$
  - $f(n) = 5n + 42$ ;  $f \in O(n)$  and  $f \in O(n^2)$  but  $f \notin O(1)$
  - $f(n) = \sum_{i=0}^d a_i n^i$  where  $\{a_i\}$  are constants and  $a_d > 0$ ;  
 $f \in O(n^d)$  and  $f \in O(n^{d+1})$  but  $f \notin O(n^{d-1})$



# Big-Oh ( $O$ ) Notation (Continued)



- $f \in O(g)$
- for  $n \geq n_0$ ,  $f(n)$  is *upper bounded* by  $cg(n)$
- asymptotically,  $f$  grows at rate *no greater than* that of  $g$  to within constant factor

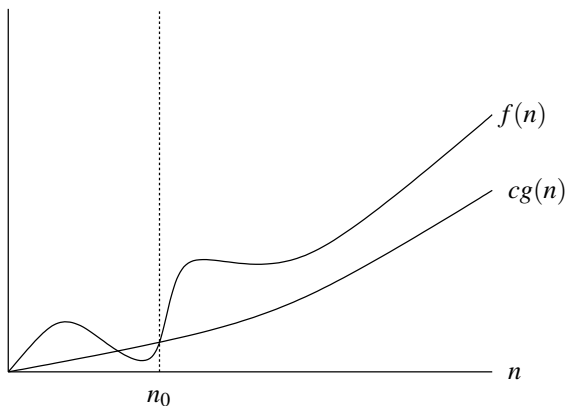
# Big-Omega ( $\Omega$ ) Notation

- **big-omega ( $\Omega$ ) notation:** for function  $g$ ,  $\Omega(g)$  denotes set of all functions  $f$  for which positive constants  $c$  and  $n_0$  exist such that

$$0 \leq cg(n) \leq f(n) \quad \text{for all } n \geq n_0$$

- functions in  $\Omega(g)$  grow asymptotically at rate **at least** that of  $g$  (to within constant factor)
- used to provide (asymptotic) **lower bound** on function to within constant factor
- if  $f \in \Omega(g)$ , then for sufficiently large  $n$ ,  $f(n)$  is greater than or equal to  $g(n)$  to within constant factor
- since  $\Theta(g(n)) \subset \Omega(g(n))$ ,  $f(n) \in \Theta(g(n))$  implies  $f(n) \in \Omega(g(n))$
- examples:
  - $f(n) = 5n^3 + n$ ;  $f \in \Omega(n^3)$  and  $f \in \Omega(n^2)$  but  $f \notin \Omega(n^4)$
  - $f(n) = an^2 + bn + c$  where  $a, b, c$  are constants and  $a > 0$ ;  
 $f \in \Omega(n^2)$  and  $f \in \Omega(n)$  but  $f \notin \Omega(n^3)$
  - $f(n) = \sum_{i=0}^d a_i n^i$  where  $\{a_i\}$  are constants and  $a_d > 0$ ;  
 $f \in \Omega(n^d)$  and  $f \in \Omega(n^{d-1})$  but  $f \notin \Omega(n^{d+1})$

# Big-Omega ( $\Omega$ ) Notation (Continued)



- $f \in \Omega(g)$
- for  $n \geq n_0$ ,  $f(n)$  **lower bounded** by  $cg(n)$
- asymptotically,  $f$  grows at rate **no less than** that of  $g$  to within constant factor

- **small-oh ( $o$ ) notation**: for function  $g$ ,  $o(g)$  denotes set of all functions  $f$  such that, for any positive constant  $c$ , positive constant  $n_0$  exists such that

$$0 \leq f(n) < cg(n) \quad \text{for all } n \geq n_0$$

- functions in  $o(g)$  grow asymptotically at *strictly lesser* rate than  $g$  (to within constant factor)
- used to provide *upper bound* on function that is *not asymptotically tight*
- $f \in o(g)$  implies that  $f(n)$  becomes insignificant relative to  $g(n)$  as  $n$  becomes arbitrarily large (i.e.,  $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$ )
- examples:
  - $f(n) = 3n^3 + 2n + 1$ ;  $f \in o(n^5)$  and  $f \in o(n^4)$  but  $f \notin o(n^3)$
  - $f(n) = 2n^2$ ;  $f \notin o(n^2)$  but  $f \in O(n^2)$
  - $f(n) = \sum_{i=0}^d a_i n^i$  where  $\{a_i\}$  are constants and  $a_d > 0$ ;  
 $f \in o(n^{d+1})$  and  $f \in o(n^{d+2})$  but  $f \notin o(n^d)$  and  $f \notin o(n^{d-1})$

# Small-Omega ( $\omega$ ) Notation

SKIP SLIDE

- **small-omega ( $\omega$ ) notation**: for function  $g$ ,  $\omega(g)$  denotes set of all functions  $f$  such that, for any positive constant  $c$ , positive constant  $n_0$  exists such that

$$0 \leq cg(n) < f(n) \quad \text{for all } n \geq n_0$$

- functions in  $\omega(g)$  grow asymptotically at *strictly greater* rate than  $g$  (to within constant factor)
- used to provide *lower bound* on function that is *not asymptotically tight*
- $f \in \omega(g)$  implies that  $f(n)$  becomes arbitrarily large relative to  $g(n)$  as  $n$  becomes arbitrarily large (i.e.,  $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = \infty$ )
- examples:
  - $f(n) = 3n^2$ ;  $f \in \omega(n)$  but  $f \notin \omega(n^2)$
  - $f(n) = an^2 + bn + c$  where  $a, b, c$  are constants and  $a > 0$ ;  
 $f \in \omega(n)$  and  $f \in \omega(1)$  but  $f \notin \omega(n^2)$  and  $f \notin \omega(n^3)$
  - $f(n) = \sum_{i=0}^d a_i n^i$  where  $\{a_i\}$  are constants and  $a_d > 0$ ;  
 $f \in \omega(n^{d-1})$  but  $f \notin \omega(n^d)$  and  $f \notin \omega(n^{d+1})$

# Asymptotic Notation in Equations and Inequalities

- when asymptotic notation stands alone on right-hand side of equation, equal sign means set membership
- for example:
  - $f(n) = \Theta(g(n))$  means  $f(n) \in \Theta(g(n))$
- more generally, when asymptotic notation appears in formula, interpreted as placeholder for some anonymous function
- for example:
  - $3n^2 + 2n + 1 = 3n^2 + \Theta(n)$  means  $3n^2 + 2n + 1 = 3n^2 + f(n)$  where  $f(n)$  is some function in  $\Theta(n)$  (i.e.,  $f(n) = 2n + 1 \in \Theta(n)$ )
- using asymptotic notation in this way can help to reduce clutter in formulas

# Properties of $\Theta$ , $O$ , and $\Omega$

## ■ sum of functions:

- if  $f_1 \in \Theta(g)$  and  $f_2 \in \Theta(g)$ , then  $f_1 + f_2 \in \Theta(g)$
- if  $f_1 \in O(g)$  and  $f_2 \in O(g)$ , then  $f_1 + f_2 \in O(g)$
- if  $f_1 \in \Omega(g)$  and  $f_2 \in \Omega(g)$ , then  $f_1 + f_2 \in \Omega(g)$

## ■ multiplication by constant:

- for all positive functions  $f$  and all positive constants  $a$ ,  $af \in \Theta(f)$ ,  $af \in O(f)$ , and  $af \in \Omega(f)$

## ■ product of functions:

- for all positive functions  $f_1, f_2, g_1, g_2$ , if  $f_1 \in \Theta(g_1)$  and  $f_2 \in \Theta(g_2)$ , then  $f_1 f_2 \in \Theta(g_1 g_2)$
- for all positive functions  $f_1, f_2, g_1, g_2$ , if  $f_1 \in O(g_1)$  and  $f_2 \in O(g_2)$ , then  $f_1 f_2 \in O(g_1 g_2)$
- for all positive functions  $f_1, f_2, g_1, g_2$ , if  $f_1 \in \Omega(g_1)$  and  $f_2 \in \Omega(g_2)$ , then  $f_1 f_2 \in \Omega(g_1 g_2)$

## ■ examples:

- if  $f \in \Theta(n)$ , then  $nf(n) \in \Theta(n^2)$
- if  $f$  and  $g$  are positive functions in  $\Theta(1)$ , then  $f + g \in \Theta(1)$

- $\log_2 n \in \Theta(\log_b n)$  for all  $b > 1$  (i.e., base of logarithm does not impact asymptotic analysis)



# Remarks on Asymptotic Complexity

- one must be careful in interpreting results of asymptotic complexity analysis
- asymptotic complexity only considers algorithm behavior when problem size becomes *arbitrarily large*
- for example: for problems of size  $n < 10^{10}$ , algorithm  $A$  with time complexity  $f(n) = (\frac{1}{10^{10}})n^2$  will take less time than Algorithm  $B$  with time complexity  $g(n) = n$ , in spite of fact that  $f(n) = \Theta(n^2)$  and  $g(n) = \Theta(n)$  (i.e., algorithm  $A$  has greater asymptotic complexity than algorithm  $B$ )
- asymptotic complexity *hides constant factors*
- for example: for problems of size  $n$ , algorithm  $A$  with time complexity  $f(n) = n$  is clearly preferable to algorithm  $B$  with time complexity  $g(n) = 1000n$ , but both  $f$  and  $g$  are in  $\Theta(n)$  (i.e., both algorithms have same asymptotic complexity)
- asymptotic complexities can be used for guidance but should not be followed blindly

# Some Common Complexities

Name	Complexity
constant	$O(1)$
logarithmic	$O(\log n)$
fractional power	$O(n^c)$ , $c \in (0, 1)$
linear	$O(n)$
log-linear	$O(n \log n)$
quadratic	$O(n^2)$
cubic	$O(n^3)$
exponential	$O(a^n)$
factorial	$O(n!)$
double exponential	$O(a^{b^n})$

- above complexities listed in order of *increasing* (asymptotic) growth rate
- that is, for sufficiently large  $n$ ,  
$$1 < \log n < \sqrt{n} < n < n \log n < n^2 < n^3 < \dots < 2^n < n! < 2^{2^n}$$

# Recurrence Relations

- **recurrence relation** is equation that implicitly defines sequence in terms of itself
- for example, Fibonacci number sequence  $f$  is solution to recurrence relation:

$$f(n) = \begin{cases} f(n-1) + f(n-2) & n \geq 2 \\ 1 & n \in \{0, 1\} \end{cases}$$

- recurrence relations often arise when trying to determine complexity of algorithm that employs recursion
- for example, consider time complexity of recursive Fibonacci algorithm:

```
1  unsigned long long fibonacci(unsigned int n) {  
2      if (n <= 2) {  
3          return 1;  
4      } else {  
5          return fibonacci(n - 1) + fibonacci(n - 2);  
6      }  
7  }
```

- time complexity  $T$  of above algorithm leads to recurrence relation  
 $T(n) = c + T(n-1) + T(n-2)$

# Solving Recurrence Relations

- no known general technique for solving recurrence relations
- solving recurrence relations somewhat of an art
- linear constant coefficient difference equations can be solved using  $z$  transform
- Master theorem can be used to solve some recurrence relations of form:

$$f(n) = g(n) + af(n/b)$$

- Akra-Bazzi theorem can be used to solve some recurrence relations of form:

$$f(n) = g(n) + \sum_{i=0}^{L-1} a_i f(b_i n + h_i(n))$$

- need to be careful about non-integer sequence indices arising in recurrence relations like:

$$T(n) = \sum_{i=0}^{L-1} a_i T(n/b_i) + f(n)$$

- preceding formula does not make sense if  $n/b_i$  is not integer
- in many cases, if this issue ignored, correct asymptotic bound still obtained, although without being correctly justified
- numerous software tools available for solving recurrence relations, such as WolframAlpha and PURRS

# Solutions for Some Common Recurrence Relations

Recurrence Relation	Solution
$f(n) = \begin{cases} b + f(n-1) & n \geq 2 \\ a & n = 1 \end{cases}$	$f(n) = b(n-1) + a \in \Theta(n)$
$f(n) = \begin{cases} bn + f(n-1) & n \geq 2 \\ a & n = 1 \end{cases}$	$f(n) = \frac{1}{2}bn(n+1) + b - a \in \Theta(n^2)$
$f(n) = \begin{cases} b + f(\lfloor n/2 \rfloor) & n \geq 2 \\ a & n = 1 \end{cases}$	$f(n) \in \Theta(\log n)$
$f(n) = \begin{cases} b + f(\lceil n/2 \rceil) & n \geq 2 \\ a & n = 1 \end{cases}$	$f(n) \in \Theta(\log n)$
$f(n) = \begin{cases} b + f(\lfloor n/2 \rfloor) + f(\lceil n/2 \rceil) & n \geq 2 \\ a & n = 1 \end{cases}$	$f(n) \in \Theta(n)$
$f(n) = \begin{cases} bn + f(\lfloor n/2 \rfloor) + f(\lceil n/2 \rceil) & n \geq 2 \\ a & n = 1 \end{cases}$	$f(n) \in \Theta(n \log n)$
$f(n) = \begin{cases} c + f(n-1) + f(n-2) & n \geq 3 \\ b & n = 2 \\ a & n = 1 \end{cases}$	$f \in \Theta(2^n)$

# Matrix Multiplication Algorithm: Time Complexity

SKIP SLIDE

- consider algorithm for multiplying  $m \times n$  matrix by  $n \times p$  matrix:

```
1  template <class T, int m, int n, int p>
2  void multiply(const T (&a)[m][n], const T (&b)[n][p],
3      T (&c)[m][p]) {
4      for (int i = 0; i < m; ++i) {
5          for (int j = 0; j < p; ++j) {
6              T sum = T(0);
7              for (int k = 0; k < n; ++k) {
8                  sum += a[i][k] * b[k][j];
9              }
10             c[i][j] = sum;
11         }
12     }
13 }
```

- total time cost per line (assuming basic operations on  $T$  are  $O(1)$ ):

Line	Total Time Cost
4	$c_{4,1}m + c_{4,2}$
5	$m(c_{5,1}p + c_{5,2})$
6	$mp(c_6)$
7	$mp(c_{7,1}n + c_{7,2})$
8	$mpn(c_8)$
10	$mp(c_{10})$

- asymptotic time complexity is  $a_1mnp + a_2mp + a_3m + a_4 = \Theta(mnp)$

# Matrix Multiplication Algorithm: Space Complexity

SKIP SLIDE

- again, consider algorithm for multiplying  $m \times n$  matrix by  $n \times p$  matrix:

```
1  template <class T, int m, int n, int p>
2  void multiply(const T (&a)[m][n], const T (&b)[n][p],
3      T (&c)[m][p]) {
4      for (int i = 0; i < m; ++i) {
5          for (int j = 0; j < p; ++j) {
6              T sum = T(0);
7              for (int k = 0; k < n; ++k) {
8                  sum += a[i][k] * b[k][j];
9              }
10             c[i][j] = sum;
11         }
12     }
13 }
```

- $a$ ,  $b$ , and  $c$  are references and each effectively incur memory cost of pointer
- $m$ ,  $n$ , and  $p$  are constant expressions and require no storage
- assuming object of type  $T$  requires  $O(1)$  space, each of  $a$ ,  $b$ ,  $c$ ,  $i$ ,  $j$ ,  $k$ , and  $sum$ , requires  $\Theta(1)$  space
- asymptotic space complexity is  $\Theta(1)$

# Iterative Fibonacci Algorithm: Time Complexity

- consider iterative algorithm for computing  $n$ th Fibonacci number:

```
1  unsigned long long fibonacci(unsigned int n) {  
2      unsigned long long a[3] = {1, 1, 1};  
3      for (int i = 3; i <= n; ++i) {  
4          a[0] = a[1];  
5          a[1] = a[2];  
6          a[2] = a[0] + a[1];  
7      }  
8      return a[2];  
9  }
```

- total time cost per line (assuming  $n \geq 2$ ):

Line	Total Time Cost
2	$c_1$
3	$(n-2)c_{3,1} + c_{3,2}$
4	$(n-2)c_4$
5	$(n-2)c_5$
6	$(n-2)c_6$
8	$c_8$

- asymptotic time complexity is

$$(c_{3,1} + c_4 + c_5 + c_6)n + (c_1 - 2c_{3,1} + c_{3,2} - 2c_4 - 2c_5 - 2c_6 + c_8) = a_1n + a_2 = \Theta(n)$$



# Iterative Fibonacci Algorithm: Space Complexity

- again, consider iterative algorithm for computing  $n$ th Fibonacci number:

```
1  unsigned long long fibonacci(unsigned int n) {
2      unsigned long long a[3] = {1, 1, 1};
3      for (int i = 3; i <= n; ++i) {
4          a[0] = a[1];
5          a[1] = a[2];
6          a[2] = a[0] + a[1];
7      }
8      return a[2];
9  }
```

- storage cost per variable:

Variable	Storage Cost
$n$	$c_1$
$a$	$c_2$
$i$	$c_3$

- asymptotic space complexity is  $c_1 + c_2 + c_3 = a_1 = \Theta(1)$

# Recursive Fibonacci Algorithm: Time Complexity

- consider recursive algorithm for computing  $n$ th Fibonacci number:

```
1  unsigned long long fibonacci(unsigned int n) {  
2      if (n <= 2) {  
3          return 1;  
4      } else {  
5          return fibonacci(n - 1) + fibonacci(n - 2);  
6      }  
7  }
```

- time cost  $T(n)$  satisfies recurrence relation:

$$T(n) = \begin{cases} T(n-1) + T(n-2) + c_1 & n \geq 3 \\ c_2 & n \in \{1, 2\} \end{cases}$$

- asymptotic time complexity is  $\Theta(2^n)$

# Recursive Fibonacci Algorithm: Space Complexity

- again, consider recursive algorithm for computing  $n$ th Fibonacci number:

```
1  unsigned long long fibonacci(unsigned int n) {  
2      if (n <= 2) {  
3          return 1;  
4      } else {  
5          return fibonacci(n - 1) + fibonacci(n - 2);  
6      }  
7  }
```

- during recursion, function calls nest to depth of at most  $n - 2 = \Theta(n)$
- each invocation of function incurs memory cost for local variable  $n$
- each function call also incurs space on stack for return address and possibly other saved state
- asymptotic space complexity is  $(n - 2)c_1 + c_0 = a_1n + a_0 = \Theta(n)$

# Amdahl's Law

- may want to determine overall speedup that can be achieved by introducing speedup into some part of task
- overall speedup  $s_o$  of whole task given by

$$s_o = \frac{1}{(1 - f_e) + \frac{f_e}{s_e}},$$

where  $s_e$  is speedup of part of task that benefits from enhancement and  $f_e$  is fraction of time consumed by part of task benefitting from enhancement

- preceding result known as **Amdahl's law**
- overall speedup is limited by fraction of time that enhancement can be exploited:

$$s_o \leq \frac{1}{1 - f_e} \quad \text{and} \quad \lim_{s_e \rightarrow \infty} s_o = \frac{1}{1 - f_e}$$

- for example, if  $f_e = 25\%$  and  $s_e = 2$ , then  $s_o = 1.1429$

## Section 6.2.1

### **References**

- 1 WolframAlpha Recurrence Relation Solver,  
<https://www.wolframalpha.com/examples/Recurrences.html>.
- 2 Parma University's Recurrence Relation Solver (PURRS),  
<http://www.cs.unipr.it/purrs>.
- 3 T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*.  
MIT Press, Cambridge, MA, USA, 3rd edition, 2009.
- 4 A. Levitin. *Introduction to the Design and Analysis of Algorithms*.  
Pearson, Boston, MA, USA, 3rd edition, 2012.
- 5 J. Erickson. *Algorithms*, Jan. 2015.  
Available online from  
<http://www.cs.illinois.edu/~jeffe/teaching/algorithms/>.
- 6 M. Akra and L. Bazzi. *On the solution of linear recurrence equations*.  
*Computational Optimization and Applications*, 10(2):195–210, May 1998.

- 7 M. Drmota and W. Szpankowski. [A master theorem for discrete divide and conquer recurrences.](#)  
In *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pages 342–361, 2011.
- 8 S. Roura. [Improved master theorems for divide-and-conquer recurrences.](#)  
*Communications of the ACM*, 48(2):170–205, Mar. 2001.

## Section 6.3

# Data Structures



# Abstract Data Types (ADTs)

- **abstract data type (ADT)** is model for data type where behavior specified from point of view of user of type (i.e., with implementation details hidden)
- ADT specifies:
  - general nature of entity represented by type
  - set of allowable states/values that type can assume
  - set of operations that can be performed on type
  - any preconditions or postconditions for operations
- often, ADT also provides complexity guarantees (e.g., time or space complexity guarantees for various operations)
- for example, (generic) integer type is ADT:
  - can assume integer values
  - provides basic arithmetic operations, relational operations, and so on
  - particular representation used for integers not specified by ADT
- in contrast to ADT, concrete (i.e., non-abstract) data type provides very specific details as to how type is implemented

# Container ADTs

- **container ADT** (also called **collection ADT**): stores collection of objects, organized in way that follows some specific access rules
- operations for container ADT often include:
  - clear: remove all elements from container
  - is empty: test if container is empty (i.e., contains no elements)
  - size: query number of elements in container
  - insert: insert element in container
  - remove: remove element from container
  - find: locate element in container if present
- often container ADT provides means to traverse elements in container (e.g., via iterator ADT)
- if elements in container consist of key-value pairs where key used to find corresponding value in container, container said to be **associative**
- if elements in container have well-defined order, container said to be **ordered**; otherwise, **unordered**
- if all elements stored in container of same type, container said to be **homogeneous**; otherwise, **heterogeneous**

## ■ examples of realizations of container ADTs:

- `std::array`, `std::vector`, `std::list`, `std::forward_list`
- `std::set`, `std::multiset`, `std::map`, `std::multimap`
- `std::unordered_set`, `std::unordered_multiset`,  
`std::unordered_map`, `std::unordered_multimap`
- `boost::intrusive::slist`, `boost::intrusive::list`

## ■ container ADTs can differ in many ways:

- number of elements container can store (e.g., one versus multiple)
- whether values stored by container must be unique
- associative versus non-associative
- ordered versus unordered
- homogeneous versus heterogeneous
- intrusive versus nonintrusive
- concurrency properties (e.g., not thread safe, thread safe, lock free)

# Iterator ADTs

- **iterator ADT** is ADT used to traverse collection of elements, which are often stored in container
- typically iterator ADT provided as part of container ADT
- operations provided by iterator ADT may include:
  - dereference: access element to which iterator refers
  - next: go to next element
  - previous: go to previous element
  - advance: advance by  $n$  elements (where  $n$  can be negative for backwards direction)
- iterator specifies order in which elements can be accessed; for example:
  - forward, bidirectional (i.e., forward and backward), random access
- iterator may only permit certain types of element access; for example:
  - read only (const), read and write (non-const), write only (output)
  - one dereference per element or multiple dereferences per element
- examples of realizations of iterator ADT:
  - `iterator` and `const_iterator` types in numerous C++ standard library containers, such as `std::vector` and `std::set`

# Container and Iterator Considerations

- are elements in container stored contiguously in memory?
- what is fixed storage overhead of container (if any)?
- what is per-element storage overhead of container (if any)?
- is container limited in size (e.g., container based on fixed size array)?
- is container dynamic (i.e., can it be changed once created) or static?
- can element be inserted at start, end, or arbitrary position in container in worst-case or amortized  $O(1)$  time?
- can element be removed at start, end, or arbitrary position in container in  $O(1)$  time?
- can element be accessed at start, end, or arbitrary position in  $O(1)$  time?
- can element be located in container efficiently (e.g.,  $O(\log n)$  time or better)?

# Container and Iterator Considerations (Continued)

- can container be traversed (e.g., via iterator) efficiently?
- what is storage cost of iterator (e.g., 1 pointer)?
- in what order can iterator access elements (e.g., forward, bidirectional, random access)?
- what circumstances result in element references being invalidated?
- what circumstances result in iterators being invalidated?
- what is per-element and amortized time cost of traversing elements in container?

## Section 6.3.1

### **Lists, Stacks, and Queues**

- **list ADT** is ADT that stores countable number of ordered values, where same value may occur more than once
- operations for list ADT include:
  - clear: remove all elements from list
  - is empty: test if list empty
  - size: query number of elements in list
  - insert: insert element in list
  - remove: remove element from list
- operations for traversing elements in list (which are often provided via iterator ADT) include:
  - successor: get next element in list
  - predecessor (optional): get previous element in list
- examples of realizations of list ADT:
  - `std::vector`, `std::forward_list`, **and** `std::list`
  - `boost::intrusive::slist` **and** `boost::intrusive::list`



# Array-Based Lists

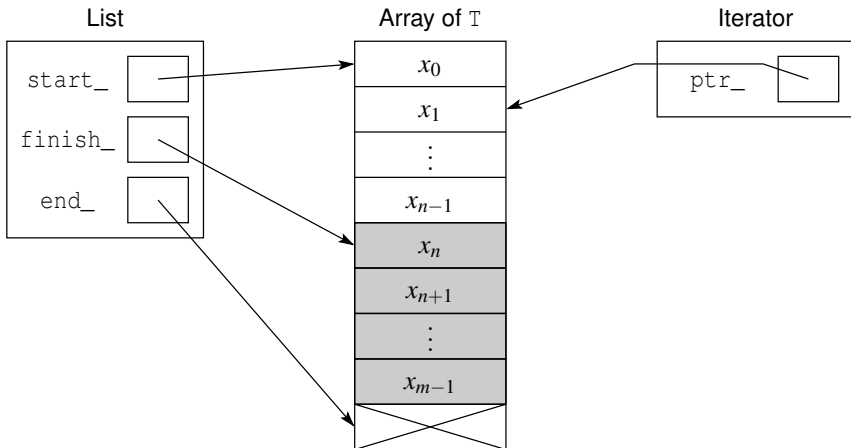
- can represent list with array

- code example:

```
1  template <class T> class Iterator {  
2      // ...  
3      T* ptr_; // pointer to referenced element  
4  };  
5  
6  template <class T> class List {  
7      // ...  
8      T* start_; // pointer to start of element data  
9      T* finish_; // pointer to end of element data  
10     T* end_; // pointer to end of allocated storage  
11 };
```

- array capacity (i.e., allocated size) is `end_ - start_`
- array size (i.e., number of elements) is `finish_ - start_`

# Array-Based Lists: Diagram



# Remarks on Array-Based Lists

## ■ advantages:

- elements stored contiguously in memory (which is cache friendly)
- no per-element storage overhead
- can insert at end of list in amortized  $O(1)$  time
- can remove at end of list in  $O(1)$  time
- can access element in any position in  $O(1)$  time
- (random-access) iterator has storage cost of one pointer

## ■ disadvantages:

- cannot insert or remove at start or arbitrary position in  $O(1)$  time
- if capacity of array exceeded, memory reallocation and copying required
- if array can be reallocated, insert at end can only at best guarantee amortized (not worst-case)  $O(1)$  time
- if array reallocated, element references invalidated

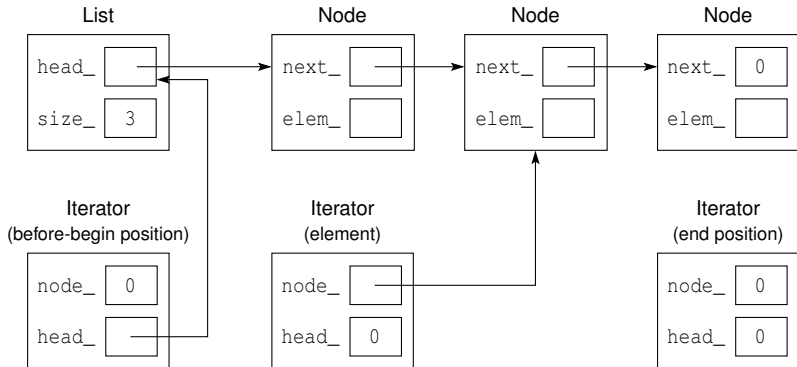
## ■ useful when insertion and removal only performed at end of list and stable references to elements not needed

- **singly-linked list** is node-based implementation of list where each node tracks its successor (but not predecessor)
- null pointer used as sentinel value to denote “no such node”; for example, null pointer used to indicate:
  - no successor node for last node in list
  - no head (i.e., first) node for empty list
- for singly-linked list, insertion and removal normally defined to take place at position *after* that specified by iterator
- to specify insertion or removal at start of list requires “before-begin” iterator

# Singly-Linked Lists: Code

```
1  // list node
2  template <class T> struct Node {
3      Node* next_; // pointer to next node in list
4      T elem_; // element data
5  };
6
7  // list
8  template <class T> class List {
9      // ...
10     Node<T>* head_; // pointer to first node in list
11     std::size_t size_; // number of elements in list
12 };
13
14 // iterator
15 template <class T> class Iterator {
16     // ...
17     Node<T>* node_; // pointer to node with referenced element
18     Node<T>** head_; // pointer to list head pointer
19     // one of three possibilities:
20     // 1) iterator refers to before-begin position:
21     //     head_ points to list head pointer and node_ is null
22     // 2) iterator refers to end position:
23     //     head_ and node_ both null
24     // 3) iterator refers to element in list:
25     //     head_ is null and node_ points to referenced node
26 };
```

# Singly-Linked List: Diagram



# Remarks on Singly-Linked Lists

## ■ advantages:

- can insert element after (but not before) particular position in  $O(1)$  time
- can remove element at start of list in  $O(1)$  time
- no capacity exceeded problem like with array
- reduced memory cost relative to doubly-linked list as consequence of node not tracking predecessor
- element references are stable
- can find successor in list in  $O(1)$  time

## ■ disadvantages:

- element data not contiguous in memory
- has per-element storage overhead (1 pointer for successor)
- cannot insert element before particular position in  $O(1)$  time
- cannot remove element at arbitrary position in  $O(1)$  time
- cannot efficiently iterate backwards over elements in list
- cannot find predecessor in list in  $O(1)$  time
- (forward) iterator requires two pointers for state (due to need for “before-begin” iterator)

- typically useful when insertions and removals always performed at start of list

# Singly-Linked List With Header Node

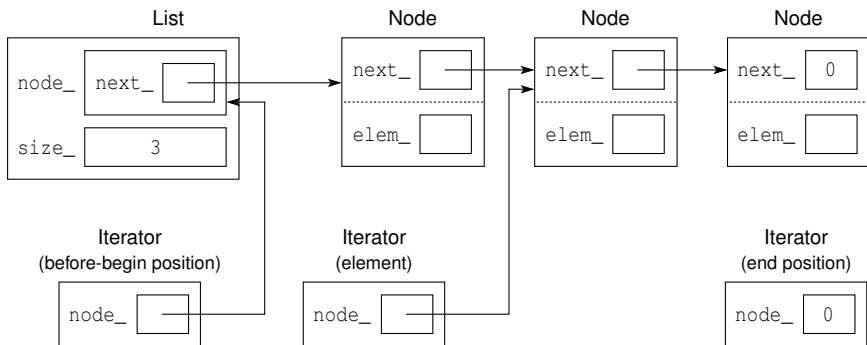
- **singly-linked list with header node** is node-based implementation of list where each node tracks its successor (but not predecessor)
- null pointer used as sentinel value to denote “no such node”; for example, null pointer used to indicate:
  - no successor for last node in list
  - no head node for empty list
- header node used as placeholder for one-before start of list (i.e., “before-begin” position)



# Singly-Linked List With Header Node: Code

```
1  // list node base class
2  struct node_base {
3      // ...
4      node_base* next_;
5  };
6
7  // list node derived class (with list element)
8  template <class T> struct node : public node_base {
9      T elem_;
10 };
11
12 // list iterator class
13 template <class T> class slist_iter {
14     // ...
15     node_base* node_;
16     // one of three possibilities:
17     // 1) iterator refers to before-begin position:
18     //     node_ points to header node
19     // 2) iterator refers to end position:
20     //     node_ is null
21     // 3) iterator refers to element in list:
22     //     node_ points to referenced element's node
23 };
24
25 // list class
26 template <class T> class list {
27     // ...
28     node_base node_;
29     std::size_t size_;
30 };
```

# Singly-Linked List With Header Node: Diagram



# Remarks on Singly-Linked List With Header Node

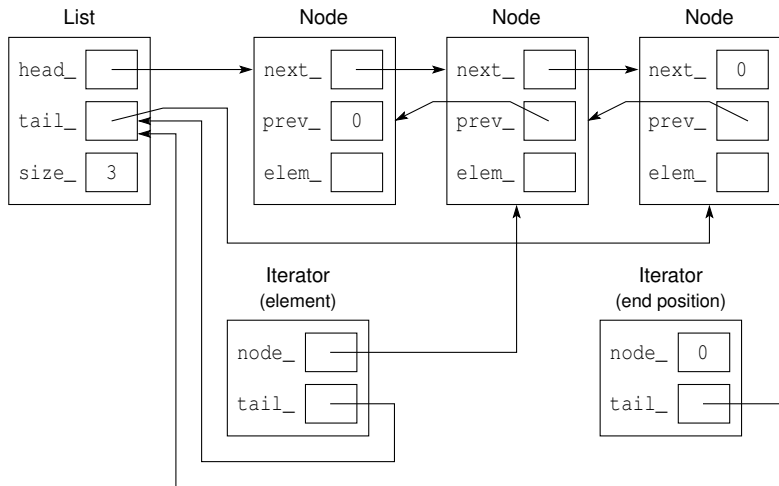
- advantages and disadvantages mostly similar to those of classic singly-linked list
- effectively *no memory cost* for header node over standard singly-linked list
- in absence of header node, special representation of before-begin iterator needed, which causes problems for efficient implementation of forward iterator
- use of header node *facilitates more efficient iterator type*
- (forward) iterator can be implemented with single pointer as state
- typically, singly-linked list with header node used to implement `std::forward_list`

- **doubly-linked list**: node-based implementation of list where each node tracks both its successor and predecessor
- null pointer used as sentinel value to indicate “no such node”; for example, null pointer used to indicate:
  - no successor for last node in list
  - no predecessor for first node in list
  - no head or tail node for empty list
- normally, for doubly-linked list, insertion defined to take place at position *before* that specified by iterator and removal defined to take place *at* position specified by iterator

# Doubly-Linked Lists: Code

```
1 // list node class
2 template <class T> struct Node {
3     Node* next_; // pointer to next node in list
4     Node* prev_; // pointer to previous node in list
5     T elem_; // element
6 };
7
8 // iterator class
9 template <class T> class Iterator {
10     // ...
11     Node<T>* node_; // node of referenced element
12     Node<T>** tail_; // pointer to tail pointer of list
13     // tail_ always points to tail_ pointer in list object
14     // one of two possibilities:
15     // 1) iterator refers to end position:
16     //     node_ is null
17     // 2) iterator refers to element in list:
18     //     node_ points to referenced element's node
19 };
20
21 // list class
22 template <class T> class List {
23     // ...
24     Node<T>* head_; // pointer to first node in list
25     Node<T>* tail_; // pointer to last node in list
26     std::size_t size_; // number of elements in list
27 };
```

# Doubly-Linked List: Diagram



# Remarks on Doubly-Linked Lists

## ■ advantages:

- no capacity-exceeded problem like in array case
- stable references to elements
- can insert or remove at arbitrary position in  $O(1)$  time
- can find successor and predecessor in  $O(1)$  time
- can efficiently iterate both forwards and backwards over elements in list

## ■ disadvantages:

- elements not stored contiguously in memory
- per-element storage overhead (2 pointers)
- relative to singly-linked list, has greater per-element storage overhead (1 additional pointer for predecessor)
- iterator storage cost is more than single pointer (i.e., 2 pointers)

## ■ most useful for lists where insertion and removal can happen anywhere in list

# Doubly-Linked List With Sentinel Node

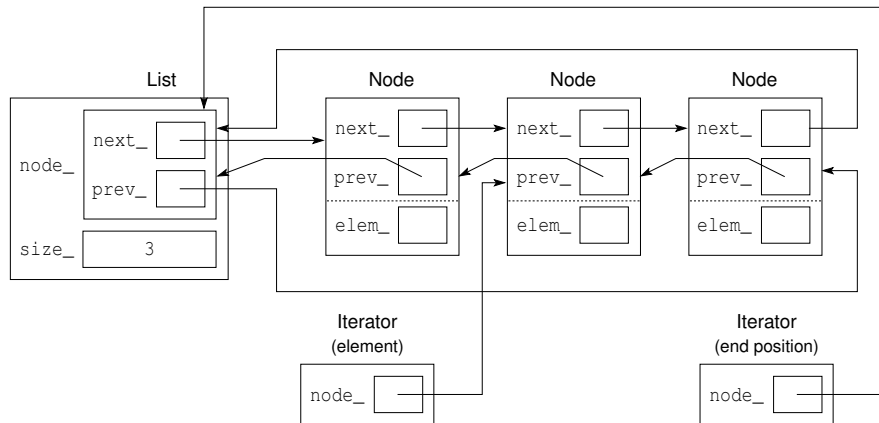
- list has one dummy node called **sentinel node** and zero or more regular (i.e., non-sentinel) nodes
- list object itself has sentinel node as member
- each regular node is associated with list element
- sentinel node is not associated with any list element
- each (regular and sentinel) node has pointer to its successor and predecessor
- if list not empty, successor of sentinel node is node corresponding to first element in list; otherwise, successor is sentinel node itself
- if list not empty, predecessor of sentinel node is node corresponding to last element in list; otherwise, predecessor is sentinel node itself
- thus, sentinel and regular nodes effectively form augmented list that is *circular*
- augmented list never empty, since always contains sentinel node
- augmented list has no beginning or end, since circular
- using sentinel node eliminates many special cases for insertion and removal, which leads to simpler and more efficient code



# Doubly-Linked List With Sentinel Node: Code

```
1  // list node base class (which does not have element data)
2  struct Node_base {
3      Node_base* next_; // pointer to next node in list
4      Node_base* prev_; // pointer to previous node in list
5  };
6
7  // list node (which has element data)
8  template <class T> struct Node : public Node_base {
9      T elem_; // element data
10 };
11
12 // list
13 template <class T> class List {
14     // ...
15     Node_base node_; // sentinel node
16 };
17
18 // list iterator
19 template <class T> class Iterator {
20     // ...
21     Node_base* node_; // pointer to referenced node
22     // one of two possibilities:
23     // 1) iterator refers to end position:
24     //     node_ points to sentinel node
25     // 2) iterator refers to element in list:
26     //     node_ points to referenced element's node
27 };
```

# Doubly-Linked List With Sentinel Node: Diagram



# Remarks on Doubly-Linked Lists With Sentinel Node

- advantages and disadvantages mostly similar to those of classic doubly-linked list
- effectively *no memory cost* for sentinel node over standard doubly-linked list
- sentinel node effectively makes list circular and always nonempty
- sentinel node *eliminates special cases* caused by empty list and insertion and removal at start and end of list (simplifying code)
- in absence of sentinel node, special representation needed for end iterator, which causes problems for efficient implementation of bidirectional iterator (namely, consider predecessor operation for iterator that refers to end of list)
- use of sentinel node *facilitates more efficient iterator type*
- (bidirectional) iterator can be implemented with single pointer as state
- typically, doubly-linked list with sentinel node used to implement `std::list`

- **stack ADT** is ADT for container where elements can only be inserted or removed in last-in first-out (LIFO) order
- can only insert and remove elements at top of stack
- operations provided by stack ADT:
  - clear: remove all elements from stack
  - is empty: test if stack is empty
  - top: access element at top of stack (without removing)
  - push: add element to top of stack
  - pop: remove element from top of stack
- **stack overflow**: attempting to perform push operation when insufficient space available for element being added
- **stack underflow**: attempting to perform pop operation when stack empty
- example realizations of stack ADT:
  - `std::stack`
  - `boost::lockfree::stack`

# Array Implementation of Stack

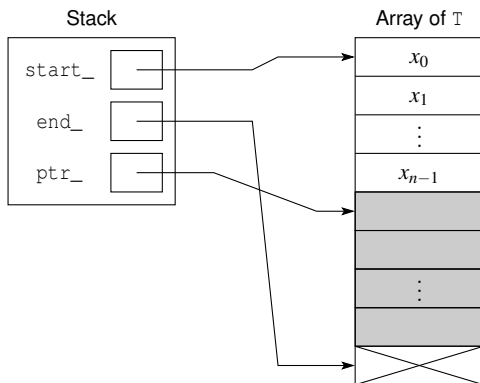
- stack can be efficiently implemented using array

- code example:

```
1  template <class T> class Stack {  
2      // ...  
3      T* start_; // pointer to start of element storage  
4      T* end_;   // pointer to end of element storage  
5      T* ptr_;   // pointer to next free slot on stack  
6  };
```

- stack empty if `ptr_ equals start_`
- stack has reached capacity if `ptr_ equals end_`
- push operation stores element at `*ptr_` and then increments `ptr_`
- pop operation decrements `ptr_`
- top operation provides access to `ptr_[-1]`
- due to possibility of exceeding array capacity, cannot guarantee each push operation takes constant time; can only hope for amortized (not worst-case)  $O(1)$  time
- memory efficient: only per-element storage cost is element data itself
- cache-efficient: element data is contiguous in memory

# Array Implementation of Stack: Diagram



# Remarks on Array Implementation of Stack

## ■ advantages:

- elements stored contiguously in memory
- no per-element storage overhead

## ■ disadvantages:

- if capacity of array exceeded, must reallocate and copy
- if array grown, can only guarantee amortized (not worst-case)  $O(1)$  time for push
- if array reallocated, elements references are invalidated

# Node-Based Implementation of Stack

- stack can be efficiently implemented using node-based *singly-linked* list

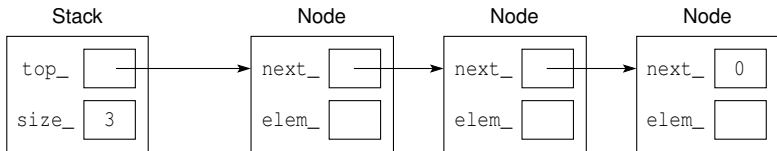
- code example:

```
1 // stack node
2 template <class T> struct Node {
3     Node* next_; // pointer to next node in stack
4     T elem_; // element data
5 };
6
7 // stack
8 template <class T> class Stack {
9     // ...
10    Node<T>* top_; // pointer to node at top of stack
11 };
```

- only need list to be singly linked (as opposed to doubly linked), since all insertions and removals performed at start of list (i.e., top of stack)



# Node-Based Implementation of Stack: Diagram



## ■ advantages:

- no capacity-exceeded problem as in array case
- can perform push operation in  $O(1)$  time in worst case
- element references are stable

## ■ disadvantages:

- element data not contiguous in memory
- has per-element storage overhead (i.e., 1 pointer for successor)
- relative to array-based implementation, requires more space

- **queue ADT** is container where elements can only be inserted and removed in first-in first-out (FIFO) order
- elements removed from front (a.k.a. head) of queue
- elements inserted at back (a.k.a. tail) of queue
- operations for queue ADT include:
  - clear: remove all elements from queue
  - is empty: test if queue is empty
  - front: access element at front of queue (without removing)
  - enqueue: insert element at back of queue
  - dequeue: remove element from front of queue
- examples of realizations of queue ADT:
  - `std::queue`
  - `boost::lockfree::queue`
- double-ended queue ADT is similar to queue ADT except allows elements to be inserted or removed at either front or back

# Array Implementation of Queue

- array implementation of bounded queue

- code example:

```
1 // bounded queue
2 template <class T> Queue {
3     // ...
4     T* start_; // start of array for queue elements
5     T* end_; // end of array for queue elements
6     T* head_; // pointer to element at front of queue
7     T* tail_; // pointer to back of queue
8     std::size_t size_; // number of entries in queue
9 };
```

- array used in circular fashion

- queue is empty if `size_` is zero

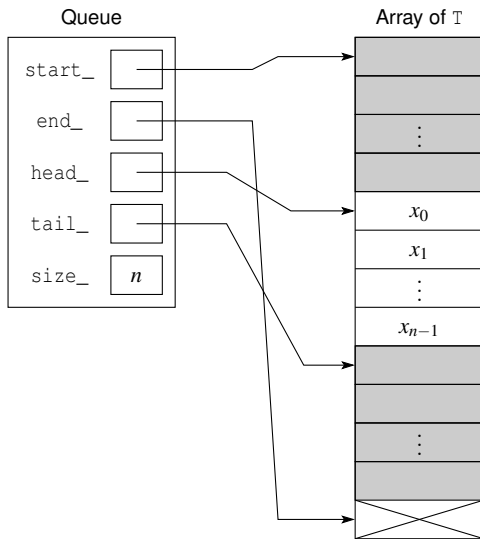
- queue is full if `size_ equals` `max_size`

- if queue not full, enqueue operation places element at `tail_` and then increments `tail_` with wraparound and increments `size_`

- if queue not empty, dequeue operation increments `head_` with wraparound and decrements `size_`

- front operation provides access to `*head_`

# Array Implementation of Queue: Diagram



# Remarks on Array Implementation of Queue

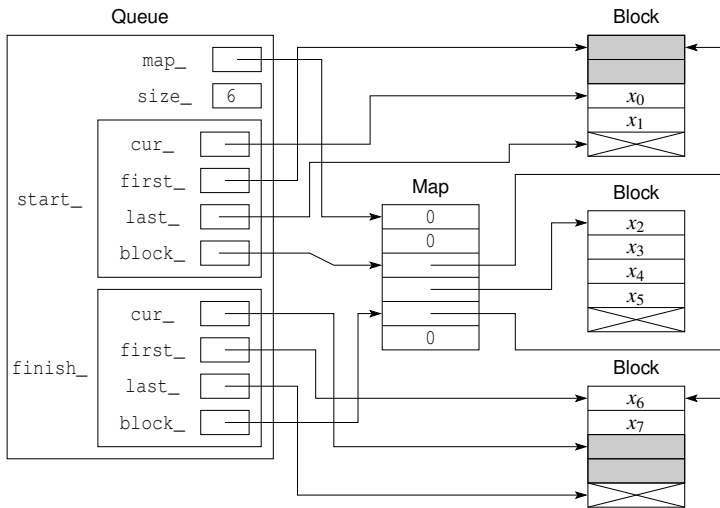
- although only consider queue of bounded size, could extend to unbounded case by using dynamically-resizable array
- advantages:
  - elements stored in contiguous buffer, occupying at most two contiguous regions of memory (i.e., contiguous region with potential hole in middle)
  - can insert and remove in  $O(1)$  time
  - can access front element in  $O(1)$  time
- disadvantages:
  - queue must be of bounded size
  - relaxing restriction of bounded size raises other issues associated with reallocation of array when capacity exceeded (e.g., worst case enqueue time not  $O(1)$ , element references not stable)

# Array of Arrays Implementation of Queue

- array of arrays can be used to implement (unbounded) queue
- code example:

```
1  // how many Ts held in each block?
2  template <class T> constexpr std::size_t block_size
3      = sizeof(T) < 512 ? 512 / sizeof(T) : 1;
4
5  template <class T> class Iterator {
6      // ...
7      T* cur_; // pointer to referenced element
8      T* first_; // pointer to first element in block
9      T* last_; // pointer to end element in block
10     T** node_; // pointer to current block
11 };
12
13 template <class T> class Queue {
14     // ...
15     T** map_; // array of block pointers
16     std::size_t size_; // size of map array
17     Iterator start_; // iterator for first element in queue
18     Iterator finish_; // iterator for end element in queue
19 };
```

# Array of Arrays Implementation of Queue: Diagram





# Remarks on Array of Arrays Implementation of Queue

- advantages:
  - elements never change their location so pointers and references to elements are stable
- disadvantages:
  - although each individual block holding element data is contiguous, blocks not contiguous
  - although elements are never relocated by insertions and removals, iterators can be invalidated
- similar data structure used in some implementations of `std::deque`

# Node-Based Implementation of Queue

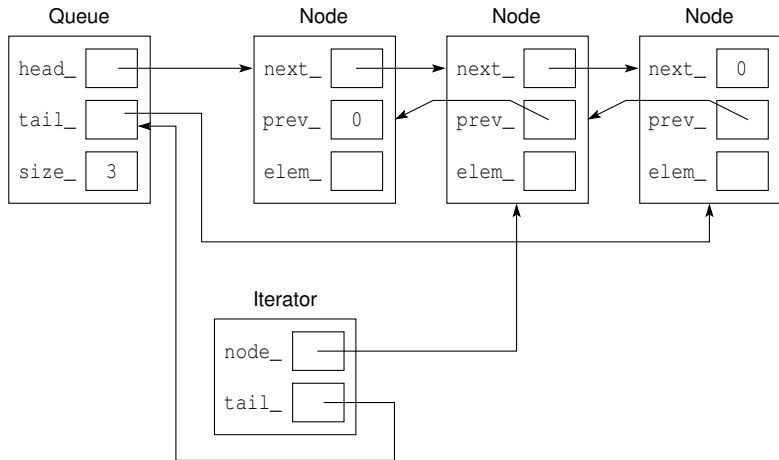
- doubly-linked list implementation of queue

- code example:

```
1  // queue node
2  template <class T> struct Node {
3      Node* next_; // pointer to next entry in queue
4      Node* prev_; // pointer to previous entry in queue
5      T elem_; // element data
6  };
7
8  template <class T> class Queue {
9      // ...
10     Node<T>* first_; // first entry in queue
11     Node<T>* last_; // last entry in queue
12     std::size_t size_; // number of queued elements
13 };
```

- enqueue operation uses insert operation of linked list to insert element at end of list
- dequeue operation uses remove operation of linked list to remove element at head of list
- front operation provides access to element at head of list

# Node-Based Implementation of Queue: Diagram



# Remarks on Node-Based Implementation of Queue

- advantages:
  - enqueue and dequeue operations can be performed in  $O(1)$  time
  - stable element references
- disadvantages:
  - elements not stored contiguously in memory
- could use doubly-linked list with sentinel node in order to facilitate more efficient iterator

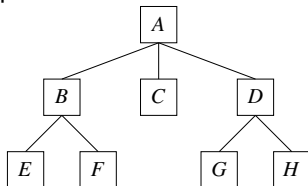
## Section 6.3.2

### **Multiway and Binary Trees**

# Trees

- tree is non-linear hierarchical data type
- tree consists of zero or more nodes
- except root, each node has parent
- each node has zero or more children
- tree containing no nodes is empty
- node  $q$  said to be **parent** of node  $n$  if  $n$  is child of  $q$
- **root node**: node in tree with no parent
- node  $q$  said to be **sibling** of node  $n$  if  $q$  and  $n$  have same parent
- tree said to be **ordered** if linear ordering of children of each node

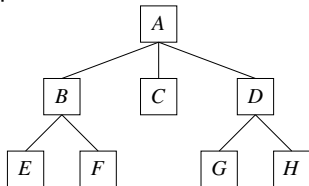
- example:



- $A$  is root node
- $B$  is child of  $A$
- $A$  is parent of  $B$
- $C$  and  $D$  are siblings of  $B$

- **path** of length  $k$  in tree is sequence of  $k + 1$  nodes  $n_0, n_1, \dots, n_k$  where  $n_i$  is parent of  $n_{i+1}$
- node  $q$  said to be **ancestor** of node  $n$  if  $q$  is on path from root node to  $n$
- node  $q$  is said to be **descendant** of node  $n$  if  $q$  on path from  $n$  to leaf
- every node is both ancestor and descendant of itself
- node  $q$  said to be **proper ancestor** of  $n$  if ancestor of, and distinct from,  $n$
- node  $q$  is said to be **proper descendant** of  $n$  if  $q$  is descendant of, and distinct from,  $n$

- example:



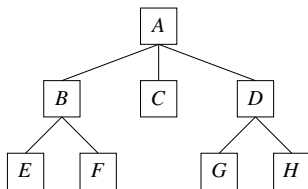
- $A, B, F$  is path of length 2
- $A$  and  $B$  are proper ancestors of  $E$
- $E$  and  $F$  are proper descendants of  $B$
- $B$  is ancestor and descendant of  $B$

# Tree Terminology (Continued 1)

- **subtree** rooted at node  $n$  is tree that consists of  $n$  and all of its descendants (e.g., subtree of root is entire tree)
- **degree** of node is number of children
- **degree** of tree is maximum node degree taken over all nodes in tree
- **internal node** is node that has at least one child
- **external node** (also called **leaf node**) is node that does not have any children

---

■ example:



- tree consisting of nodes  $B$ ,  $E$ , and  $F$  is subtree associated with node  $B$
- degree of node  $B$  is 2
- degree of tree is 3
- $A$ ,  $B$ , and  $D$  are internal nodes
- $C$ ,  $E$ ,  $F$ ,  $G$ , and  $H$  are leaf nodes

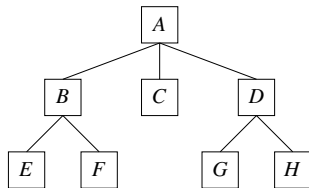


## Tree Terminology (Continued 2)

- **depth** of node (also called **level**) is length of path from root to node (or equivalently, number of proper ancestors of node) (e.g., root node has depth of zero)
- **$d$ th level** of tree is all nodes at depth  $d$  in tree
- **height** of node is length of longest path from node to any leaf (e.g., leaf node has height of zero)
- **height** of tree is maximum node height taken over all nodes in tree (i.e., height of root) if tree is nonempty; otherwise, defined to be  $-1$

---

■ example:

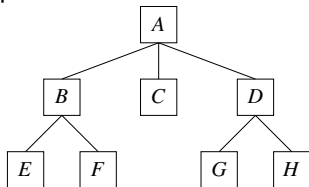


- depths of nodes  $C$  and  $E$  are 1 and 2, respectively
- nodes  $B$ ,  $C$ , and  $D$  are at level 1
- height of node  $D$  is 1
- height of tree is 2

# Tree Terminology (Continued 3)

- **weight** of node  $n$  is number of descendant leaf nodes possessed by  $n$
- **weight** of tree is number of leaf nodes in tree (i.e., weight of root node)

- example:



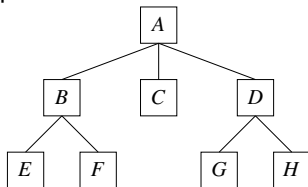
- weights of nodes  $B$  and  $C$  are 2 and 0, respectively
- weight of tree is 5

# Tree Traversal

- **preorder traversal**: node visited before its descendants (i.e., parent before children)
- **postorder traversal**: node visited after its descendants (i.e., children before parent)
- preorder traversal might be used, for example, to print hierarchical document, where nodes correspond to sections in document
- postorder traversal might be used, for example, to compute space used by files in directory and its subdirectories

---

- example:



- preorder traversal visits nodes in order: *A, B, E, F, C, D, G, H*
- postorder traversal visits nodes in order: *E, F, B, C, G, H, D, A*

- representing directory tree in hierarchical file system
  - each internal node corresponds to directory
  - each leaf node corresponds to file (or empty directory)
- representing arithmetic expressions
  - each internal node corresponds to operator
  - each leaf node corresponds to operand
- representing decision-making process
  - each internal node corresponds to question with yes/no answer
  - each leaf node corresponds to final outcome of decision-making process
- searching for elements in collection
  - nodes correspond to elements in collection
  - nodes positioned in tree based on element keys

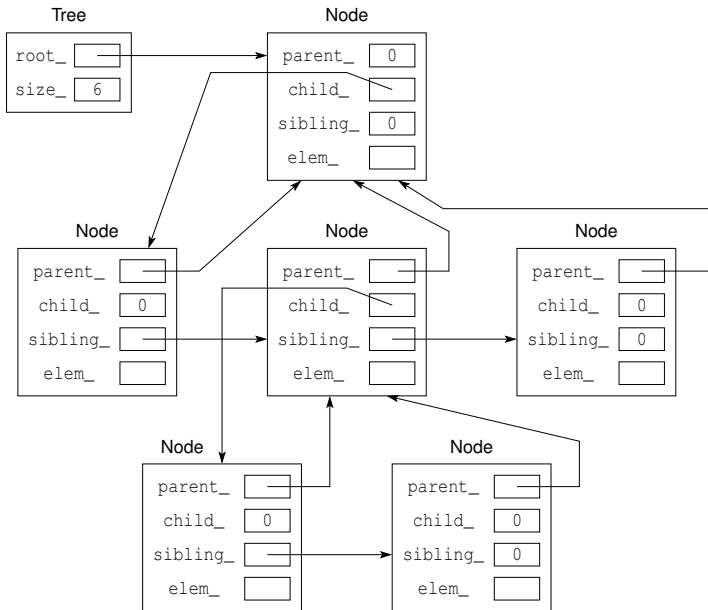
- **tree ADT** provides abstraction of tree data type
- operations provided by tree ADT include:
  - clear: remove all nodes from tree
  - size: get number of nodes in tree
  - is empty: test if tree is empty (i.e., contains no nodes)
  - root: get root node of tree
  - parent: get parent of node (which is not root)
  - children: get children of node
  - is internal: test if node is internal node
  - is external: test if node is external node
  - is root: test if node is root
  - replace: replace element in node
- may provide iterator ADT for traversing tree
- often tree ADT by itself is not particularly useful
- instead, tree ADT typically used to build other more task-specific ADTs (e.g., set ADT, multiset ADT, and so on)

# Node-Based Tree Implementation

- node-based implementation of tree
- each node has pointer to first child and next sibling
- subsequent children can be accessed by following sibling pointers from first child
- allows size of node data structure to be constant (i.e., independent of maximum number of children)
- code example:

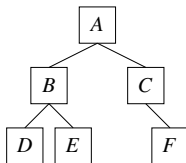
```
1  template <class T> struct Node {  
2      Node* parent_; // pointer to parent  
3      Node* child_; // pointer to first child  
4      Node *sibling_; // pointer to next sibling  
5      T elem_;  
6  };  
7  
8  template <class T> class Tree {  
9      // ...  
10     Node<T>* root_; // pointer to root node  
11     std::size_t size_; // number of nodes in tree  
12 };
```

# Node-Based Tree Implementation: Diagram



- each internal node has at most two children
  - each node, excluding root node, labelled as either left or right child
  - **left subtree** is tree rooted at left child
  - **right subtree** is tree rooted at right child
- 

- example:

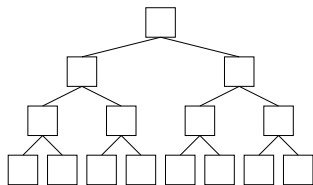


- root node is *A*
- left child of *A* is *B*
- right child of *A* is *C*
- left subtree of *A* is tree consisting of nodes *B*, *D*, and *E*
- right subtree of *A* is tree consisting of nodes *C* and *F*

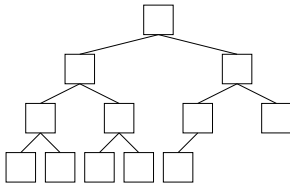


# Perfect and Complete Trees

- binary tree said to be **perfect** (or **full**) if each internal node has exactly two children (which results in all leaves being at same level)
- binary tree said to be **complete** if perfect except possibly for deepest level which must be filled from left to right
- perfect implies complete



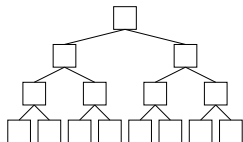
Perfect



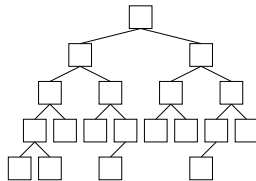
Complete

# Balanced Binary Trees

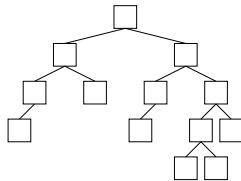
- binary tree said to be **perfectly balanced** if left and right subtrees of each (interior) node contain same number of nodes (i.e., perfect tree)
- binary tree said to be **strictly balanced** if can be formed by discarding zero or more leaf nodes from perfect tree
- binary tree said to be **height balanced** if height of left and right subtrees of each (interior) node differ by at most one
- perfectly balanced implies strictly balanced (but converse does not hold)
- strictly balanced implies height balanced (but converse does not hold)



Perfectly Balanced



Strictly Balanced



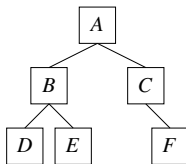
Height Balanced

# Binary Tree Traversal

- **preorder traversal**: visit node, then left subtree, then right subtree
- **postorder traversal**: visit left subtree, then right subtree, then node
- **level-order traversal**: visit nodes from left to right within level from top downwards
- one additional traversal order for binary trees: in order
- **in-order traversal**: visit left subtree, then node, then right subtree

---

■ example:



- preorder traversal order: *A, B, D, E, C, F*
- postorder traversal order: *D, E, B, F, C, A*
- inorder traversal order: *D, B, E, A, C, F*
- level order traversal order: *A, B, C, D, E, F*

- **binary tree ADT** provides abstraction of binary tree
- operations provided by binary tree ADT that are common to general (i.e., *m*-ary) tree ADT include:
  - create: make empty tree
  - root: get root node
  - parent: get parent of node
  - is internal: test if node is internal (i.e., non-leaf)
  - is external: test if node is external (i.e., leaf)
  - is root: test if node is root of tree
  - is empty: test if binary tree is empty (i.e., contains no nodes)
  - size: get number of nodes in tree
  - clear: remove all nodes from tree
  - replace: replace element in node
  - add root: add root node (to empty tree)

- other operations provided by binary tree ADT include:
  - left child: get left child of node
  - right child: get right child of node
  - has left child: test if node has left child
  - has right child: test if node has right child
  - insert left: insert node as left child of node (which must be leaf)
  - insert right: insert node as right child of node (which must be leaf)
  - remove: remove node (which must be leaf)
- may provide iterator ADT for traversing tree

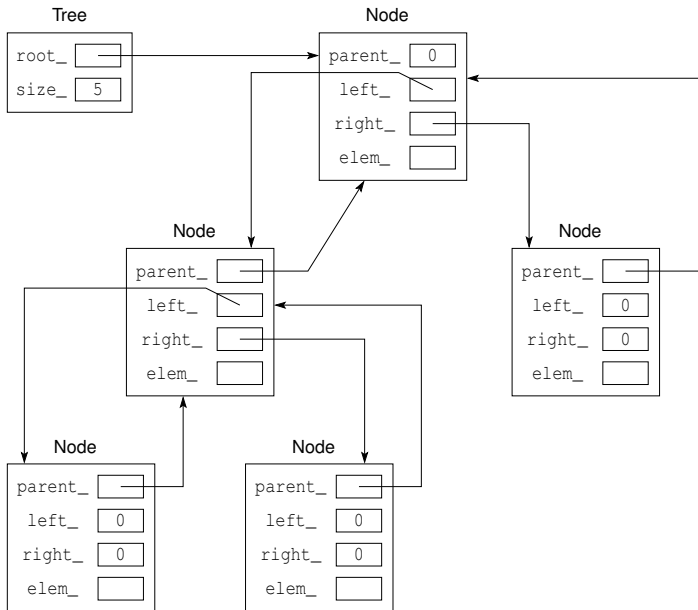
# Node-Based Binary Tree

- node-based implementation of binary tree
- node data structure and tree data structure
- null pointer used as sentinel to indicate “no such node” (e.g., child of leaf, parent of root, etc.)
- code example:

```
1  template <class T> struct Node {  
2      Node* parent_; // pointer to parent  
3      Node* left_; // pointer to left child  
4      Node* right_; // pointer to right child  
5      T elem_; // element data  
6  };  
7  
8  template <class T> class Tree {  
9      // ...  
10     Node<T>* root_; // pointer to root node  
11     std::size_t size_; // number of nodes in tree  
12 };
```

- node-based implementation preferred for trees that are not complete
- in practice, sentinel node often preferred when iterator functionality must be provided

# Node-Based Binary Tree: Diagram



## ■ advantages:

- can handle case of tree that is not complete without gross memory inefficiency
- can provide stable element references

## ■ disadvantages:

- has per-element storage overhead (3 pointers: 1 for parent, 1 for first child, and 1 for second child or next sibling)
- element data not contiguous



# Array-Based Binary Tree

- complete binary tree can be implemented using array
- position in array determines position in tree
- let  $\text{index}(n)$  denote index of node  $n$
- let  $\text{parent}(n)$ ,  $\text{left}(n)$ , and  $\text{right}(n)$  denote parent, left child, and right child of node  $n$
- root node has index 0; and

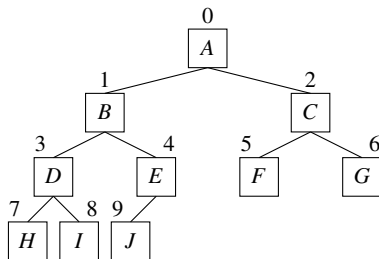
$$\begin{aligned}\text{index}(\text{left}(n)) &= 2\text{index}(n) + 1 \\ \text{index}(\text{right}(n)) &= 2\text{index}(n) + 2 \\ \text{index}(\text{parent}(n)) &= \lfloor (\text{index}(n) - 1) / 2 \rfloor\end{aligned}$$

- code example:

```
1  template <class T> class Tree {  
2      // ...  
3      T* start_; // start of element data  
4      T* end_;   // end of element data  
5      std::size_t size_; // allocated size of data  
6  };
```

# Array-Based Binary Tree: Diagram

- example of complete tree with nodes labelled with corresponding array indices:



# Remarks on Array-Based Binary Tree

## ■ advantages:

- memory efficient: no per-element storage overhead (i.e., no memory cost for representing connectivity of nodes in tree)
- cache efficient: element data stored contiguously in memory

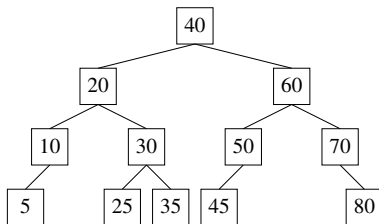
## ■ disadvantages:

- can only handle complete trees
- although could generalize this approach to handle non-complete tree, would be grossly inefficient in terms of memory usage
- if array capacity exceeded, costly reallocation and copy required
- if array reallocation occurs, cannot provide stable references to elements

- array implementation should be preferred for complete trees (unless inability to guarantee stable element references is problematic)

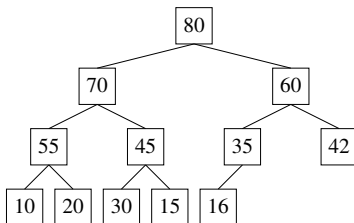
# Binary Search Trees

- binary tree is said to have **binary search tree** property if, for each node  $n$  with key  $k$ , following holds:
  - every key in left subtree of  $n$  is less than or equal to  $k$ ; and
  - every key in right subtree of  $n$  is greater than or equal to  $k$
- for tree of height  $h$ , can find element in  $O(h)$  time
- example of binary search tree:



# Heaps

- tree said to have **heap property** if, for each node  $n$  in tree, following holds:
  - key of  $n$  is greater than or equal to key of each descendant of  $n$
- **heap** is tree that satisfies heap property
- inserting or removing node can be done in  $O(\log n)$  time without breaking heap property (but may need rearrangement of some nodes)
- example of heap:



## Section 6.3.3

### **Hash Tables**

# Basic Idea Behind Hash Tables

- rather than navigating through search tree comparing search key to element key, hashing tries to reference element directly in table based on key
- effectively, hashing transforms key into address in table
- basic operations provided by hash table:
  - insert: add element to hash table
  - remove: remove element from hash table
  - find: search for element in hash table based on key
- want above operations to take  $O(1)$  time on average
- order of elements in table unimportant

- hash table of size  $m$  consists of  $m$  **slots** (also called **buckets**) numbered from 0 to  $m - 1$  (inclusive)
- each element stored in hash table has key and possibly some associated value
- each slot can be empty or contain element data
- slot in which element stored determined by applying hash function to key
- **load factor** is ratio of number of elements in hash table to hash table size
- **collision** said to occur when two distinct keys map to same index in hash table
- often, choosing table size as prime number helps to ensure more uniform distribution of elements over slots



# Hash Table Example

- collection of 10-digit employee numbers
- employee number is key
- hash function yields last four digits of employee number

Index	Slot
0000	
0001	0019910001
0002	5919870002
0003	
	⋮
9997	1212009997
9998	
9999	1122339999

# Hash Functions

- **hash function**: maps key  $k$  of given type to integer in  $\{0, 1, \dots, m - 1\}$
- hash function usually specified as composition of two functions:
  - 1 hash code map
  - 2 compression map
- **hash code map**: maps key to integer
- **compression map**: maps integer to integer in  $\{0, 1, \dots, m - 1\}$
- first hash-code map  $h_1$  applied to key  $k$  and then compression map  $h_2$  applied to result to yield hash function  $h(k) = h_2(h_1(k))$
- hash function is typically many-to-one mapping (which can therefore result in collisions)
- goal of hash function is to distribute keys uniformly across elements of  $\{0, 1, \dots, m - 1\}$ , which will reduce likelihood of collisions
- hash function should be fast to compute

- various strategies can be used to generate hash-code map
- integer cast:
  - reinterpret bits of key as integer
- component sum:
  - partition key into integers of fixed size
  - then, sum these integers ignoring overflow
- polynomial accumulation:
  - partition bits of key into sequence of components of fixed length  $a_0, a_1, \dots, a_{n-1}$
  - then, evaluate polynomial  $p(z) = \sum_{i=0}^{n-1} a_i z_i^i$  for some fixed value of  $z$ , ignoring overflow

- various strategies can be used to generate compression map
- let  $m$  denote size of hash table
- $m$  usually chosen to be prime in order to better distribute keys over hash values
- division:
  - $h_2(i) = i \bmod m$
- multiply, add, and divide:
  - $h_2(i) = (ai + b) \bmod m$ , where  $a$  and  $b$  nonnegative integers and  $a \bmod m \neq 0$

# Collision Resolution by Chaining

- chaining also called closed addressing
- with chaining, collisions handled by allowing multiple elements to be placed in single slot
- elements in slot stored in linked list
- **simple uniform hashing**: keys equally likely to hash into any of slots
- for load factor  $\alpha$ , successful and unsuccessful searches take average-case time  $\Theta(1 + \alpha)$  under assumption of simple uniform hashing
- if insertion of elements already in hash table not allowed, insert operation has worst-case  $O(1)$  time
- removal of element has worst-case  $O(1)$  time
- can support insert, remove, and search all in  $O(1)$  time on average (under assumption of simple uniform hashing)
- hash table cannot fill since each slot can potentially hold any number of elements

# Collision Resolution by Open Addressing

- with open addressing, only one element allowed to be stored per slot in table so in case of collision alternate choice must be made for slot to store element
- sequence of indices to consider (in order) when inserting (or searching for) element with given key called **probe sequence**
- examine table at each position in probe sequence until slot for element is found (e.g., empty slot for insertion)
- for each possible key  $k$ , probe sequence should be permutation of  $\{0, 1, \dots, m-1\}$  so that all slots are reachable
- many possible choices for probe sequence (e.g., linear, quadratic, double hashing, and random hashing)
- load factor  $\alpha$  must satisfy  $\alpha \leq 1$  (since only one element stored per slot)
- **uniform hashing**: probe sequence of each key equally likely to be any of  $m!$  permutations of  $\{0, 1, \dots, m-1\}$
- number of probes in unsuccessful search is at most  $\frac{1}{1-\alpha}$ , assuming uniform hashing

- with linear probing, probe sequence starts at hash value of key and then proceeds as necessary sequentially, wrapping around to beginning of table when end of table reached
- $i$ th value in probe sequence for key  $k$  given by  $h(k, i) = (h'(k) + i) \bmod m$ , where  $h'$  is hash function
- suffers from primary clustering, where colliding elements clump together causing future collisions to generate longer sequence of probes
- expected number of probes for insertion or unsuccessful search is  $\frac{1}{2} \left( 1 + \frac{1}{(1-\alpha)^2} \right)$
- expected number of probes for successful search is  $\frac{1}{2} \left( 1 + \frac{1}{1-\alpha} \right)$

# Linear Probing Example

- integer key; hash function  $h(k) = k \bmod 13$
- insert 18, 15, 23, 31, 44, and 9 (in order):

0	1	2	3	4	5	6	7	8	9	10	11	12
					18							

0	1	2	3	4	5	6	7	8	9	10	11	12
		15			18							

0	1	2	3	4	5	6	7	8	9	10	11	12
		15			18					23		

0	1	2	3	4	5	6	7	8	9	10	11	12
		15			18	31				23		



0	1	2	3	4	5	6	7	8	9	10	11	12
		15			18	31	44			23		



0	1	2	3	4	5	6	7	8	9	10	11	12
		15			18	31	44		9	23		

$k$	$h(k)$
18	5
15	2
23	10
31	5
44	5
9	9

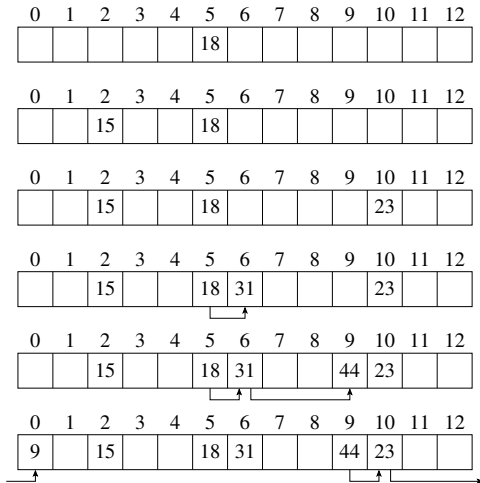


# Quadratic Probing

- with quadratic probing, distance between probes is determined by quadratic polynomial
- $i$ th value in probe sequence for key  $k$  given by  $h(k, i) = (h'(k) + c_1i + c_2i^2) \bmod m$ , where  $h'$  is hash function and  $c_1$  and  $c_2$  are nonnegative integer constants
- $c_1$ ,  $c_2$ , and  $m$  must be carefully chosen to guarantee successful insertion is possible
- most often  $c_1 = 0$  and  $c_2 = 1$  and  $m$  prime
- must ensure that loading factor  $\alpha \leq \frac{1}{2}$  in order to guarantee successful insertion
- eliminates primary clustering, but suffers from secondary clustering

# Quadratic Probing Example

- integer key; hash function  $h(k) = k \bmod 13$ ;  $c_1 = 0$  and  $c_2 = 1$
- insert 18, 15, 23, 31, 44, and 9 (in order):



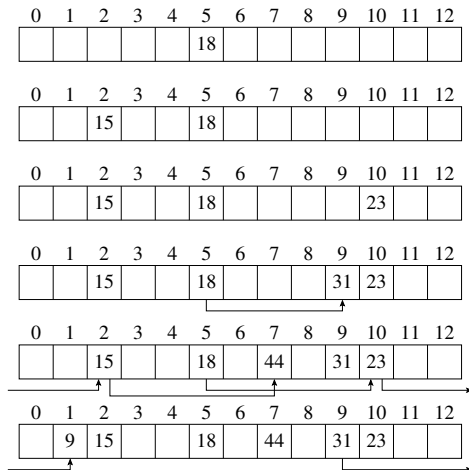
$k$	$h(k)$
18	5
15	2
23	10
31	5
44	5
9	9

# Double Hashing

- with double hashing, distance between successive probes determined by secondary hash function
- $i$ th value in probe sequence for key  $k$  given by  $h(k, i) = (h_1(k) + ih_2(k)) \bmod m$ , where  $h_1$  is (primary) hash function and  $h_2$  is secondary hash function
- $h_2$  must never be zero
- $h_2$  must be coprime with  $m$  for entire hash table to be searched
- for example, could let  $m$  be prime and have  $h_2$  always yield strictly positive integer less than  $m$

# Double Hashing Example

- integer key; hash function  $h(k) = k \bmod 13$ ; secondary hash function  $d(k) = 7 - k \bmod 7$
- insert 18, 15, 23, 31, 44, and 9 (in order):



$k$	$h(k)$	$d(k)$
18	5	3
15	2	6
23	10	5
31	5	4
44	5	5
9	9	5

- with random hashing, probe sequence generated by output of pseudorandom number generator seeded by key
- random number generation is relatively expensive
- in practice, double hashing tends to work about as well

# Open Addressing: Insertion, Removal, and Search

- with open addressing, removal of elements can be problematic; in earlier linear probing example, consider removal of element with key 31 followed by search for element with key 44
- simplest solution is to distinguish between slot that has always been empty and slot from which element was deleted
- to perform search:
  - probe until:
    - element with query key is found; or
    - empty slot is found; or
    - all slots have been unsuccessfully probed
- to insert element (assuming not already in table):
  - examine successive slots in probe sequence until
    - slot found that is empty or “deleted”
  - if all slots have been unsuccessfully probed, error
  - store element in located slot
- to remove element:
  - remove element and mark occupied slot with special “deleted” marker

- if keep adding elements to hash table, eventually size of table will need to be increased, due to loading factor becoming too large (for good performance or correct behavior)
- **rehashing**: rebuilding hash table with different number of slots
- typical threshold for load factor  $\alpha$  for rehashing:
  - 1 for chaining
  - $\frac{1}{2}$  for open addressing

# Some Applications of Hash Tables

- dictionary searches (e.g., spelling checkers, natural language understanding)
- accessing tree or graph nodes by name (e.g., city names on geographical maps)
- symbol tables in compilers
- transposition tables used in some games (e.g., chess)



## Section 6.3.4

### **Sets, Multisets, Maps, and Multimaps**

# Set and Multiset ADTs

- **set ADT** is container that stores collection of *unique* values
- set can be ordered (i.e., elements have well-defined order) or unordered
- operations provided by set ADT include:
  - clear: remove all elements from set
  - is empty: test if set is empty
  - size: query cardinality of set (i.e., number of elements in set)
  - insert: insert value in set
  - remove: remove value from set
  - find: locate value in set if present (i.e., for testing set membership)
- **multiset ADT** similar to set ADT except that *duplicate values allowed*
- example realizations of set/multiset ADT:
  - `std::set` and `std::multiset`
  - `std::unordered_set` and `std::unordered_multiset`
  - `boost::intrusive::unordered_set` and `boost::intrusive::unordered_multiset`
  - `boost::intrusive::set` and `boost::intrusive::multiset`

# Map and Multimap ADTs

- **map (or associative array) ADT** is container that stores pairs each consisting of key and value, where keys are *unique*
- each element in map consists of key and value
- operations provided by map ADT include:
  - clear: remove all elements from map
  - is empty: test if map is empty
  - size: query number of elements in map
  - insert: insert element in map
  - remove: remove element from map
  - find: locate element in map if present based on its key
- **multimap ADT** similar to map ADT except that keys *need not be unique*
- example realizations of map/multimap ADT:
  - `std::map` **and** `std::multimap`
  - `std::unordered_map` **and** `std::unordered_multimap`
  - `boost::intrusive::set` **and** `boost::intrusive::multiset`

# Remarks on Implementation of Sets and Maps

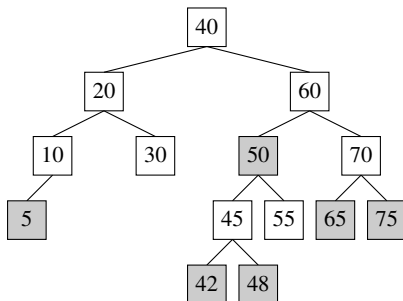
- ordered sets, multisets, maps, and multimaps typically implemented using balanced binary search tree [see [binary search trees](#)]
- unordered sets, multisets, maps, and multimaps typically implemented using hash table

# Red-Black Trees

- red-black trees first proposed by Bayer (1972)
- red-black tree is approximately height-balanced binary search tree
- requires one additional field per node, namely, color (i.e., red or black)
- binary search tree with following invariants:
  - each node is either red or black
  - root node is black
  - if node is red, then both of its children are black
  - every path from given node to any of its descendant nil nodes (i.e., null pointer) contains same number of black nodes
- invariants guarantee approximate height balancing
- path from root to farthest leaf no more than twice as long as path from root to nearest leaf
- height  $h$  of tree with  $n$  nodes is bounded by  $h \leq 2\log_2(n + 1)$
- invariants maintained by rotation and color flipping operations
- memory cost only 1 additional bit per node (for color), relative to classic binary tree

# Red-Black Trees (Continued)

- some C++ standard library implementations use red-black trees for types that provide binary search tree functionality (e.g., `std::set` and `std::map`)
- example realizations of red-black trees:
  - `boost::intrusive::rbtree`, `boost::intrusive::set`, and `boost::intrusive::multiset`
- example of red-black tree (where red nodes are shaded gray):



- AVL trees first proposed by Adelson-Velsky and Landis (1962)
- AVL tree is height-balanced binary search tree
- balance factor  $b$  of node  $n$  is defined as  $b = r - \ell$ , where  $\ell$  and  $r$  are heights of left and right subtrees of  $n$ , respectively
- AVL tree is binary search tree such that, for every node  $n$ , balance factor  $b$  of  $n$  satisfies  $b \in \{-1, 0, 1\}$  (i.e., for each node in tree, height of left and right subtrees differ by at most one)
- need to store balance factor in each node
- AVL trees more rigidly balanced than red-black trees
- height  $h$  of tree with  $n$  nodes is bounded by

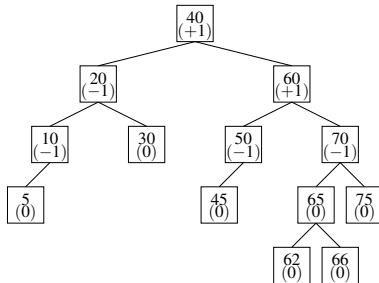
$$h \leq c \log_2(n + d) + b \approx 1.440 \log_2(n + 1.065) - 0.328,$$

where  $c = \frac{1}{\log_2 \phi}$ ,  $b = \frac{c}{2} \log_2 5 - 2$ ,  $d = 1 + \frac{1}{\phi^4 \sqrt{5}}$ , and  $\phi = \frac{1+\sqrt{5}}{2}$

- memory cost is 2 bits per node (for balance factor), relative to classic binary tree
- rebalancing achieved by rotation operations

# AVL Trees (Continued)

- since AVL trees more rigidly balanced than red-black trees, search operations typically faster in AVL tree
- insertion and removal operations typically slower in AVL tree than in red-black tree, due to more work being required for tree re-balancing
- example realizations of AVL trees:
  - `boost::intrusive::avl_tree`, `boost::intrusive::avl_set`, and `boost::intrusive::avl_multiset`
- example of AVL tree:





- treap is combination of binary search tree and heap
- each node has key and priority
- nodes arranged to form binary search tree with respect to key
- nodes also arranged to form heap with respect to priority
- if priorities chosen randomly, tree will be well balanced with high probability
- treaps provide benefits of balanced search trees, but rebalancing (which is driven by heap property) is less complicated than with some other types of balanced search trees
- example realizations of treaps:
  - `boost::intrusive::treap`, `boost::intrusive::treap_set`, and `boost::intrusive::treap_multiset`

- splay tree is self-adjusting binary search tree with property that *searches for more frequently accessed elements can be performed more quickly*
- splay tree keeps more recently accessed elements closer to root
- caching effect comes at cost of tree rebalancing being required each time search is performed
- significant disadvantage of splay tree is that height of tree can become linear in number of elements
- in worst case, insertion, removal, and search operations take amortized  $O(\log n)$  time
- example realizations of splay trees:
  - `boost::intrusive::splay_tree`, `boost::intrusive::splay_set`,  
and `boost::intrusive::splay_multiset`

# Scapegoat Trees

- scapegoat tree is self-balancing binary search tree
- provides worst-case  $O(\log n)$  search time
- provides insertion and removal in amortized  $O(\log n)$  time
- unlike other self-balancing binary search trees that provide worst-case  $O(\log n)$  search time, scapegoat trees have *no additional per-node overhead* compared to regular binary search tree
- rebalancing can potentially be very expensive, although only infrequently
- consequently, insertion and removal operations have worst-case  $O(n)$  time
- example realizations of scapegoat trees:
  - `boost::intrusive::sgtree`, `boost::intrusive::sg_set`, and `boost::intrusive::sg_multiset`

## Section 6.3.5

### **Priority Queues**

# Priority Queue ADT

- **priority queue ADT** is ADT similar to queue except that each element on queue also has corresponding priority
- element at front of queue is always element with highest priority
- operations provided by priority queue ADT include:
  - front: access element at front of queue (i.e., element with highest priority)
  - insert: insert element in queue with specified priority
  - remove: remove element from front of queue (i.e., element with highest priority)
  - update priority (optional): update priority of element in queue
- if priority queue has **stability property**, elements with equal priority will be removed in FIFO order
- examples of realization of priority queue ADT:
  - `std::priority_queue`,
  - `boost::heap::priority_queue` and `boost::heap::fibonacci_heap`

# Remarks on Priority Queue Implementations

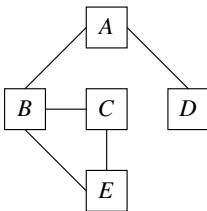
- priority queues typically implemented with heaps [see [heaps](#)]
- heaps can always be constructed to be complete trees
- consequently, can reasonably choose to implement priority queue with either array-based or node-based tree
- in practice, stability often ensured by augmenting priority with integer sequence number, which is incremented with each insertion
- array-based implementation more memory efficient but does not have stable element references (if underlying array can be reallocated)
- node-based implementation can offer stable element references

## Section 6.3.6

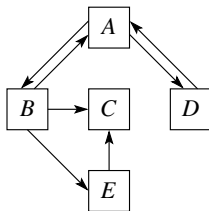
# Graphs

# Graphs

- graph is concept from discrete mathematics
- collection of nodes and edges
- nodes can be connected by edges
- directed graph: edges are directed (i.e., have direction)
- undirected graph: edges are undirected
- examples of graphs:



Undirected



Directed



- **graph ADT** is abstraction of mathematical notion of graph
- operations for graph ADT include:
  - adjacent: tests if edge from one vertex to another
  - neighbours: list all vertices that have edge to another vertex
  - insert vertex: add vertex to graph
  - remove vertex: remove vertex from graph
  - insert edge: add edge from one vertex to another
  - remove edge: remove edge from one vertex to another
  - get vertex value: get value associated with vertex
  - set vertex value: set value associated with vertex
  - get edge value: get value associated with edge
  - set edge value: set value associated with edge
- examples of realization of graph ADT:
  - `boost::adjacency_list` and `boost::adjacency_matrix`

# Adjacency Matrix Implementation of Graph

- adjacency (i.e., connectivity) of  $n$  nodes can be represented using  $n \times n$  binary matrix called **adjacency matrix**
- $(i, j)$ th element of adjacency matrix is 1 if graph has edge from node  $i$  to node  $j$  and 0 otherwise
- if graph is undirected, adjacency matrix is symmetric and only lower triangular part of matrix need be stored
- if graph is directed, adjacency matrix is not necessarily symmetric, and entire matrix must be stored
- can test adjacency of two nodes (which requires examining element in matrix) in  $O(1)$  time
- identifying all neighbours of given node takes  $O(n)$  time
- iterating over all edges is slow
- storage cost of adjacency matrix is  $\Theta(n^2)$
- high storage cost easier to justify if graph has large number of edges

# Adjacency List Implementation of Graph

- adjacency (i.e., connectivity) of  $v$  nodes can be represented using  $v$  linked lists
- $i$ th list contains node  $j$  if and only if graph has edge from node  $i$  to node  $j$
- can test adjacency of two nodes (which requires traversing linked list) in worst-case  $O(d)$  time, where  $d$  is largest valence of nodes in graph
- identifying all neighbours of given node very cheap
- storage cost of adjacency list is  $O(v + e)$ , where  $v$  and  $e$  is number of node and edges, respectively

## Section 6.3.7

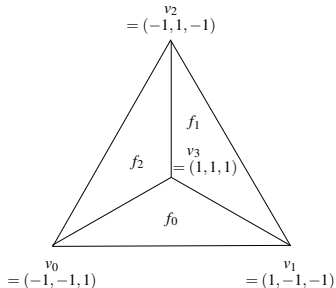
# **Data Structures and File Formats for Polygon Meshes**

# Naive Triangle-Mesh Data Structure

```
struct Face {  
    int vertexIndexes[3]; // indexes of vertices of triangle  
};  
  
Vertex vertices[numVertices]; // vertex array  
Face triangles[numTriangles]; // triangle array
```

- edges not explicitly represented
- some adjacency information not readily accessible
- to find neighboring face, must scan through face array looking for face with two vertices in common, which takes  $O(n)$  time (where  $n$  is number of vertices)
- need data structures that allow efficient access to adjacency information

# Naive Triangle-Mesh Data Structure Example



Vertices

Array Index	Array Element
0	$(-1, -1, 1)$
1	$(1, -1, -1)$
2	$(-1, 1, -1)$
3	$(1, 1, 1)$

Faces

Array Index	Array Element
0	0, 1, 3
1	1, 2, 3
2	0, 3, 2

# Half-Edge Data Structure

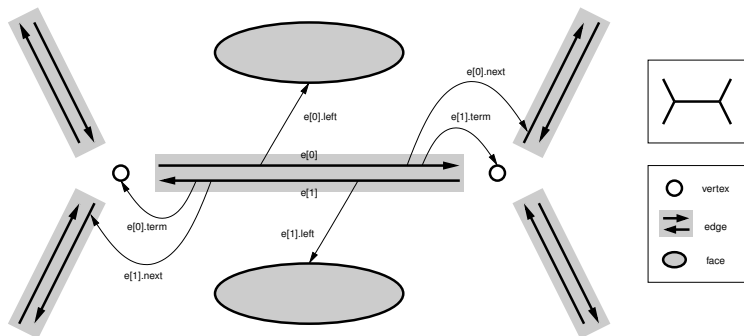
- described in:

K. Weiler. [Edge-based data structures for solid modeling in curved-surface environments.](#)

*IEEE Computer Graphics and Applications*, 5(1):21–40, Jan. 1985.

- every edge represented as pair of directed edges, each called **half-edge**
- 6 pointers plus 2 bits (i.e., 2 one-bit integers) per edge
- used in Computational Geometry Algorithms Library (CGAL)
- representing edges in terms of *directed* line segments often advantageous in algorithms

# Half-Edge Data Structure (Continued)



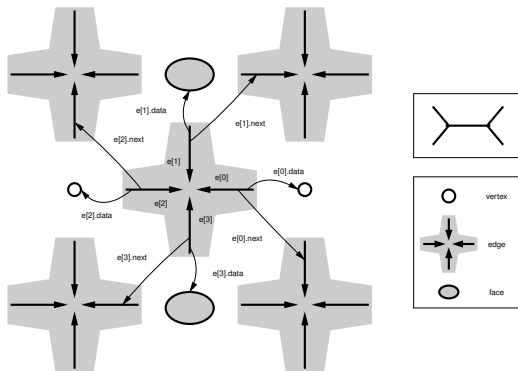
```
struct HalfEdge {  
    int index; // index of half-edge in parent edge  
    HalfEdge* next; // next CCW half-edge around left face  
    Vertex* term; // terminal vertex  
    Face* left; // left face  
};  
  
struct Edge {  
    HalfEdge e[2]; // pair of symmetric half-edges  
};
```



- proposed in:
  - L. Guibas and J. Stolfi. [Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams.](#)  
*ACM Transactions on Graphics*, 4(2):74–123, Apr. 1985.
- simultaneously represents graph and its *dual*
- each edge belongs to four circular singly-linked lists corresponding to two vertices and two faces incident to edge
- vertex/face represented by ring of quad-edges
- 8 pointers plus 4 two-bit integers per edge
- used in various research software available on Internet (e.g., Scape terrain-simplification software, Dani Lischinski's constrained DT software)

# Quad-Edge Data Structure (Continued)

SKIP SLIDE

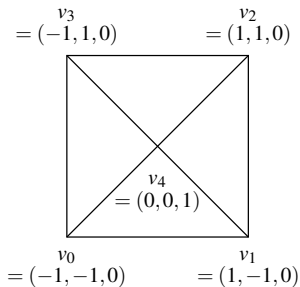


```
struct QuadEdge {  
    int index; // index of quad-edge in parent edge  
    QuadEdge* next; // next CCW quad-edge with same origin  
    void* data; // face or vertex  
};  
  
struct Edge {  
    QuadEdge* e[4]; // four quad-edges of edge  
};
```

# Object File Format (OFF)

- simple scheme for encoding the geometry and topology of a polygon mesh
- also has provisions for including color and normal information

# OFF Example (Triangle Mesh)



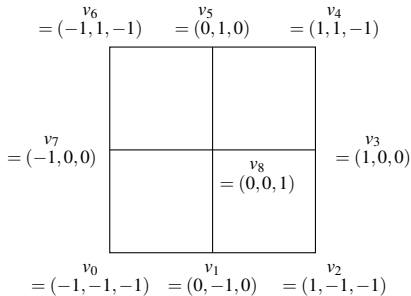
Mesh

```
OFF
5 4 0
-1 -1 0
1 -1 0
1 1 0
-1 1 0
0 0 1
3 0 1 4
3 1 2 4
3 2 3 4
3 0 4 3
```

Corresponding  
OFF File

# OFF Example (Quad Mesh)

SKIP SLIDE



Mesh

```
OFF
9 4 0
-1 -1 -1
0 -1 0
1 -1 -1
1 0 0
1 1 -1
0 1 0
-1 1 -1
-1 0 0
0 0 1
4 0 1 8 7
4 1 2 3 8
4 8 3 4 5
4 7 8 5 6
```

Corresponding  
OFF File

## Section 6.3.8

### **Intrusive Containers**

# Intrusive Containers

- container said to be **intrusive** if it requires help from elements it intends to store in order to store them
- intrusive container **directly** places user's objects in container (**not copies** of user's objects)
- node pointers exposed to user of container, which allows some operations to be performed more efficiently
- intrusive container **does not own** elements it stores
- lifetime of stored object not bound to or managed by container (i.e., lifetime of stored objects managed by user)
- can store element in multiple intrusive containers simultaneously (which is not possible with nonintrusive containers)
- more coupling between code for container and code using container

# Shortcomings of Non-Intrusive Containers

- object can only belong to one container
- only copies of objects stored in nonintrusive containers
- creating copies of values can become bottleneck (due to memory allocation and copying)
- noncopyable and nonmovable objects cannot be stored in nonintrusive containers (unless objects can be directly constructed inside container and are guaranteed not to be copied/moved subsequently)
- cannot store derived object in nonintrusive container and retain original type (i.e., copying derived object into container would result in slicing)



# Advantages of Intrusive Containers

- same object can be placed in multiple intrusive containers simultaneously
- intrusive containers do not invoke memory management operations since do not own stored elements
- complexity of inserting and removing elements in intrusive containers more predictable since no memory allocation involved
- intrusive containers tend to allow stronger complexity guarantees (since no memory allocation or copying performed)
- intrusive containers offer better exception safety guarantees (since do not need to make copy of element to place in container)
- for intrusive container, computation of iterator from pointer or reference to element is  $O(1)$  time operation, which is often not true for nonintrusive containers (e.g., for nonintrusive list `std::list`, this operation takes  $O(n)$  time)

# Disadvantages of Intrusive Containers

- in order to use type with intrusive container, must change definition of type
- each type stored in intrusive container needs additional memory to hold information for container
- intrusive containers unavoidably expose some implementation details of container to user
- since some implementation details are exposed, easier to break invariants of container; for example:
  - changing key of element in map
  - corrupting pointers used to link nodes in container
- user must assume responsibility for memory management (since container does not)
- user must manage lifetime of objects placed in container independent from lifetime of container itself, which can be error prone:
  - when destroying container before object, must be careful to avoid resource leaks
  - destroying object while in container, likely to be disastrous since container uses part of object to implement container

# Disadvantages of Intrusive Containers (Continued)

- typically, intrusive containers not copyable (and often not movable as well) since such containers do not directly perform memory allocation
- analyzing thread safety of program using intrusive containers often more difficult since container contents can be modified without going through container interface

- node-based implementation of list where each node tracks both its successor and predecessor
- value type (which stores user data) and node type (which is used to maintain list) are *same*
- null pointer used as sentinel value to indicate “no such node” (e.g., no successor/predecessor node or no head/tail node)
- in order for elements of type  $T$  to be used with container,  $T$  must include special type as data member (which encapsulates next/previous pointers for linked list)
- uses pointer to member to identify member that holds list node state [see [pointers to members](#)]

# Intrusive Doubly-Linked List: Code

SKIP SLIDE

```
1  // type encapsulating links for list
2  // (i.e., part of list node)
3  template <class T> struct list_hook {
4      // ...
5      T* next_; // pointer to next node in list
6      T* prev_; // pointer to previous node in list
7  };
8
9  // iterator class
10 // (constness of T determines if const_iterator)
11 template <class T, list_hook<T> T::* P> class list_iterator {
12     // ...
13     // node_ptr is T* (where T may be const qualified)
14     node_ptr node_; // pointer to node of referenced element
15     node_ptr const* tail_; // pointer to list tail node pointer
16 };
17
18 template <class T, list_hook<T> T::* P> class list {
19     // ...
20     T* head_; // pointer to first node in list
21     T* tail_; // pointer to last node in list
22     std::size_t size_; // number of elements in list
23 };
```

# Intrusive Doubly-Linked List: Code (Continued)

SKIP SLIDE

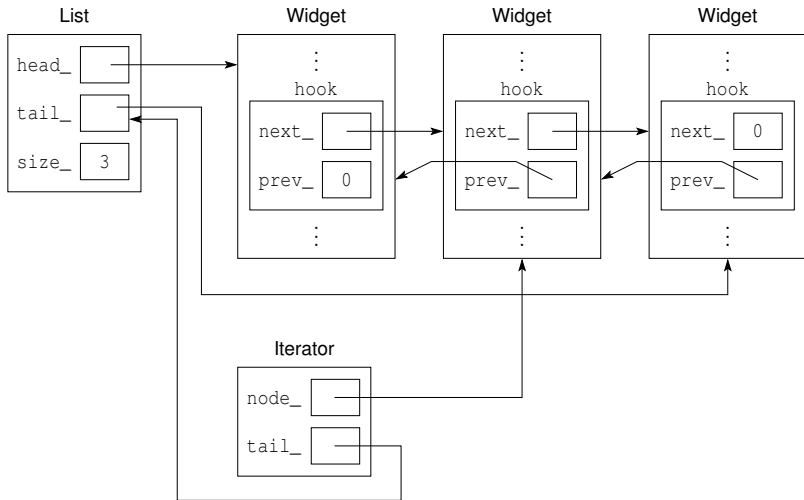
```
1 // list node with user data
2 struct Widget {
3     // ...
4     list_hook<Widget> hook; // public
5     // ...
6 };
7
8 // type for list of Widget objects
9 using Widget_list = list<Widget, &Widget::hook>;
```

---

```
1 // list node with user data
2 // (which can be placed on two lists simultaneously)
3 struct Gadget {
4     // ...
5     list_hook<Gadget> first_hook; // public
6     list_hook<Gadget> second_hook; // public
7     // ...
8 };
9
10 // type for list of Gadget objects using first_hook
11 using First_gadget_list = list<Gadget, &Gadget::first_hook>;
12
13 // type for list of Gadget objects using second_hook
14 using Second_gadget_list = list<Gadget, &Gadget::second_hook>;
```

# Intrusive Doubly-Linked List: Diagram

SKIP SLIDE



# Remarks on Intrusive Doubly-Linked List

SKIP SLIDE

- node pointer and value pointer are equivalent (i.e., pointers to next and previous nodes have type  $T^*$ )
- storage cost of iterator is two pointers (but one pointer would be more desirable)
- iterator state requires pointer to list tail pointer in order to handle case of decrementing end iterator (which has null node pointer)
- implementation does not require any non-portable constructs



# Intrusive Doubly-Linked List With Sentinel Node

- **intrusive doubly-linked list with sentinel node** is circular doubly-linked list with dummy node that serves as sentinel (instead of using null pointer)
- value type (which stores user data) and node type (which is used to maintain list) are *distinct*
- in particular, value type contains node type as data member
- effectively sentinel node makes list circular
- in order for elements of type  $T$  to be used with container,  $T$  must include special type as data member (which encapsulates next/previous pointers for linked list)

# Intrusive Doubly-Linked List With Sentinel Node: Code

```
1  // list node class
2  struct list_hook {
3      // ...
4      list_hook* next_; // pointer to next node in list
5      list_hook* prev_; // pointer to previous node in list
6  };
7
8  // list traits class (no data members)
9  template <class T, list_hook<T> T::* P> class list_traits {
10     // functions for mapping between object and node pointers
11 };
12
13 // list iterator class
14 // (constness of T determines if const_iterator)
15 template <class T, list_hook<T> T::* P> class list_iterator :
16     list_traits<T, P>{
17     // ...
18     list_hook* node_; // pointer to node of referenced element
19 };
20
21 // list
22 template <class T, list_hook<T> T::* P> class list :
23     list_traits<T, P> {
24     // ...
25     list_hook node_; // sentinel node
26     std::size_t size_; // number of elements in list
27 };
```

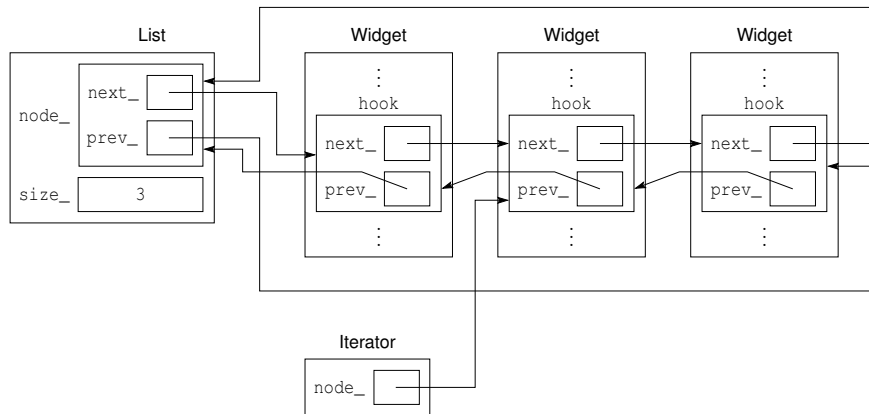
# Intrusive Doubly-Linked List With Sentinel Node: Code (Continued)

```
1 // list node with user data
2 struct Widget {
3     // ...
4     list_hook hook; // public
5     // ...
6 };
7
8 // type for list of Widget objects
9 using Widget_list = list<Widget, &Widget::hook>;
```

---

```
1 // list node with user data
2 // (which can be placed on two lists simultaneously)
3 struct Gadget {
4     // ...
5     list_hook first_hook; // public
6     list_hook second_hook; // public
7     // ...
8 };
9
10 // type for list of Gadget objects using first_hook
11 using First_gadget_list = list<Gadget, &Gadget::first_hook>;
12
13 // type for list of Gadget objects using second_hook
14 using Second_gadget_list = list<Gadget, &Gadget::second_hook>;
```

# Intrusive Doubly-Linked List With Sentinel Node: Diagram



# Remarks on Intrusive Doubly-Linked List With Sentinel Node

- circular list avoids many special cases in implementation of list class (since circular list never empty and list has no beginning or end)
- node and value types are distinct (i.e., node pointers are of type `list_hook`, not `T`)
- storage cost of iterator is one pointer
- implementation requires non-portable construct to determine value pointer from node pointer
- determining value pointer from node pointer cannot work in all cases (in particular, if value type uses virtual inheritance)
- limitations on what types can be placed in container
- another variation on this intrusive list approach can be obtained by using inheritance to add required list state to user's type (instead of adding by data members), which has some advantages

# Examples of Intrusive Containers

- as of C++17, all container classes in standard library are nonintrusive
- Boost library has good selection of intrusive containers, which includes (amongst others):
  - `boost::intrusive::slist` (intrusive singly-linked list)
  - `boost::intrusive::list` (intrusive doubly-linked list)
  - `boost::intrusive::set` (intrusive set/map)
  - `boost::intrusive::multiset` (intrusive multiset/multimap)
  - `boost::intrusive::unordered_set` (intrusive unordered set/map)
  - `boost::intrusive::unordered_multiset` (intrusive unordered multiset/multimap)
  - `boost::intrusive_ptr` (intrusive reference-counted smart pointer)

## Section 6.3.9

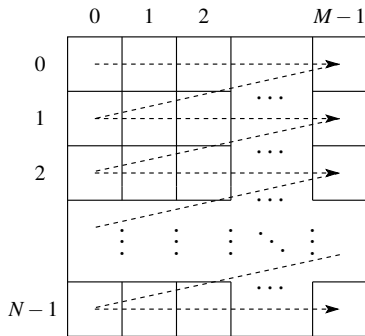
### **Miscellany**

# Memory Management for Containers

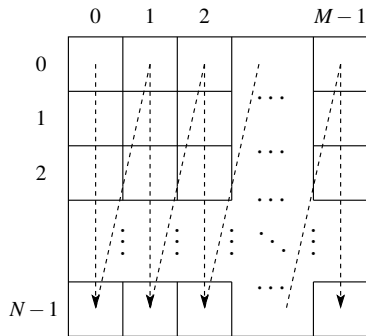
- for reasons of efficiency or functionality (or even correctness), often necessary to:
  - separate memory allocation from construction
  - separate memory deallocation from destruction
- operator new can be used to perform only memory allocation (without construction)
- placement new can be used to perform only construction (without memory allocation)
- operator delete can be used to perform only memory deallocation (without destruction)
- direct invocation of destructor can be used to perform only destruction (without memory deallocation)
- allocator type provides interface that decouples allocation/deallocation and construction/destruction
- numerous convenience functions provided by standard library for dealing with uninitialized storage



# Row-Major Versus Column-Major Order



Row-Major Order



Column-Major Order

## Section 6.3.10

### **References**

- 1 T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, USA, 3rd edition, 2009.
- 2 N. Dale. *C++ Plus Data Structures*. Jones and Bartlett, Sudbury, MA, USA, 3rd edition, 2003.
- 3 M. A. Weiss. *Data Structures and Algorithm Analysis in C++*. Pearson, Boston, MA, USA, 4th edition, 2014.
- 4 F. M. Carrano and J. J. Prichard. *Data Abstraction and Problem Solving With C++: Walls and Mirrors*. Addison Wesley, 3rd edition, 2001.
- 5 A. Drozdek. *Data Structures and Algorithms in C++*. Brooks/Cole, Pacific Grove, CA, USA, 2nd edition, 2001.

- 6 G. Adelson-Velskii and E. M. Landis. *An algorithm for the organization of information*.  
*Proc. of the USSR Academy of Sciences*, 146:263–266, 1962.  
In Russian.
- 7 R. Bayer. *Symmetric binary B-trees: Data structure and maintenance algorithms*.  
*Acta Informatica*, 1(4):290–306, 1972.
- 8 Boost C++ Libraries Web Site, <http://www.boost.org>.
- 9 M. Austern, The Standard Librarian: Defining Iterators and Const Iterators, Dr. Dobb's Journal, Jan. 2001, Available online at  
<http://www.drdobbs.com/the-standard-librarian-defining-iterato/184401331>.
- 10 R. Mattethat, Implementing Splay Trees in C++, Dr. Dobb's Journal, Sept. 2005, Available online at <http://www.drdobbs.com/cpp/implementing-splay-trees-in-c/184402007>.

## Section 6.4

### **Finite-Precision Arithmetic**

- What do each of the following functions output when executed?

```
void func1() {  
    double x = 0.1;  
    double y = 0.3;  
    double z = 0.4;  
    if (x + y == z)  
        {std::cout << "true\n";}  
    else  
        {std::cout << "false\n";}  
}
```

```
void func2() {  
    double x = 1e50;  
    double y = -1e50;  
    double z = 1.0;  
    if (x + y + z == z + y + x)  
        {std::cout << "true\n";}  
    else  
        {std::cout << "false\n";}  
}
```

```
void func3() {  
    for (double x = 0.0; x != 1.0; x += 0.1)  
        {std::cout << "hello\n";}  
}
```

# Example: Controlling Precision of Output

```
1  #include <iostream>
2  #include <iomanip>
3  #include <limits>
4  #include <boost/io/ios_state.hpp>
5
6  template <class T> void func(T x) {
7      boost::io::ios_flags_saver saver(std::cout);
8      std::cout << x << ' ' <<
9          std::setprecision(std::numeric_limits<T>::digits10) << x << ' ' <<
10         std::setprecision(std::numeric_limits<T>::max_digits10) << x << ' ' <<
11         std::hexfloat << x << '\n';
12 }
13
14 int main() {
15     func(0.1f);
16     func(0.1);
17     func(0.1L);
18 }
19
20 /* example output:
21 0.1 0.1 0.1000000001 0x1.999999ap-4
22 0.1 0.1 0.10000000000000000001 0x1.9999999999999999ap-4
23 0.1 0.1 0.1000000000000000000001 0xc.ccccccccccccccdp-7
24 */
```

# Example: Determining Next/Previous Representable Value

```
1  #include <cmath>
2  #include <iostream>
3  #include <limits>
4  #include <boost/io/ios_state.hpp>
5
6  template <class T> void func(T x) {
7      boost::io::ios_flags_saver saver(std::cout);
8      T prev = std::nextafter(x, -INFINITY);
9      T next = std::nextafter(x, INFINITY);
10     std::cout.precision(std::numeric_limits<T>::max_digits10);
11     std::cout << prev << ' ' << x << ' ' << next << '\n';
12     std::cout << std::hexfloat;
13     std::cout << prev << ' ' << x << ' ' << next << '\n';
14 }
15
16 int main() {
17     func(0.0f);
18     func(0.0);
19     func(0.0L);
20 }
21
22 /* example output:
23 -1.40129846e-45 0 1.40129846e-45
24 -0x1p-149 0x0p+0 0x1p-149
25 -4.9406564584124654e-324 0 4.9406564584124654e-324
26 -0x0.000000000000001p-1022 0x0p+0 0x0.000000000000001p-1022
27 -3.64519953188247460253e-4951 0 3.64519953188247460253e-4951
28 -0x0.0000000000000001p-16385 0x0p+0 0x0.0000000000000001p-16385
29 */
```



# Number Representations Using Different Radixes

- Note: All numbers are base 10, unless explicitly indicated otherwise.
- What is the representation of  $\frac{1}{3}$  in base 3?  
 $\frac{1}{3} = 0.\overline{3} = 0.1_3$
- What is the representation of  $\frac{1}{10}$  in base 2?  
 $\frac{1}{10} = 0.1 = 0.0001\overline{1}_2$
- A number may have a representation with a finite number of non-zero digits in one particular number base but not in another.
- Therefore, when a value must be represented with a limited number of significant digits, the number base matters (i.e., affects the approximation error).
- For example, in base 2,  $\frac{1}{10}$  cannot be represented exactly using only a finite number of significant digits.  
 $0.00011_2 = 0.09375$   
 $0.000110011_2 = 0.099609375$   
...

# Finite-Precision Number Representations

- finite-precision number representation only capable of representing small fixed number of digits
- due to limited number of digits, many values cannot be represented exactly
- in cases that desired value cannot be represented exactly, choose nearest representable value (i.e., round to nearest representable value)
- finite-precision representations can suffer from error due to roundoff, underflow, and overflow
- two general classes of finite-precision representations:
  - 1 fixed-point representations
  - 2 floating-point representations

# Fixed-Point Number Representations

- **fixed-point representation**: radix point remains fixed at same position in number
- if radix point fixed to right of least significant digit position, integer format results

Integer Format	$a_{n-1} a_{n-2} \cdots a_1 a_0 \blacksquare$
----------------	---

- if radix point fixed to left of most significant digit position, purely fractional format results

Fractional Format	$\blacksquare a_{n-1} a_{n-2} \cdots a_1 a_0$
-------------------	---

- fixed-point representations *quite limited in range* of values that can be represented
- numbers that vary greatly in magnitude cannot be represented easily using fixed-point representations
- one solution to range problem would be for programmer to maintain scaling factor for each fixed-point number, but this is clumsy and error prone

# Floating-Point Number Representations

- **floating-point representation**: radix point is not fixed at particular position within number; instead radix point allowed to move and scaling factor automatically maintained to track position of radix point
- in general, floating-point value represents number  $x$  of form

$$x = sr^e,$$

- $s$  is signed integer with fixed number of digits, and called **significand**
- $e$  is signed integer with fixed number of digits, and called **exponent**
- $r$  is integer satisfying  $r \geq 2$ , and called **radix**
- in practice,  $r$  typically 2
- for fixed  $r$ , representation of particular  $x$  not unique if no constraints placed on  $s$  and  $e$  (e.g.,  $5 \cdot 10^0 = 0.5 \cdot 10^1 = 0.05 \cdot 10^2$ )

# Floating-Point Number Representations (Continued)

- to maximize number of significant digits in significand,  $s$  and  $e$  usually chosen such that first nonzero digit in significand is to immediate left of radix point (i.e.,  $1 \leq |s| < r$ ); number in this form called **normalized**; otherwise called **denormalized**
- other definitions of normalized/denormalized sometimes used but above one consistent with IEEE 754 standard
- Example:

$$\begin{aligned}0.75 &= 0.11_2 = 1.1_2 \cdot 2^{-1} \\1.25 &= 1.01_2 = 1.01_2 \cdot 2^0 \\-0.5 &= -0.1_2 = -1.0_2 \cdot 2^{-1}\end{aligned}$$

# IEEE 754 Standard (IEEE Std. 754-1985)

- most widely used standard for (binary) floating-point arithmetic
- specifies four floating-point formats: single, double, single extended, and double extended
- single and double formats called basic formats
- radix 2
- three integer parameters determine values representable in given format:
  - number  $p$  of significand bits (i.e., precision)
  - maximum exponent  $E_{\max}$
  - minimum exponent  $E_{\min}$
- parameters for four formats are as follows:

Parameter	Single	Single Extended	Double	Double Extended
$p$	24	$\geq 32$	53	$\geq 64$
$E_{\max}$	127	$> 1023$	1023	$\geq 16383$
$E_{\min}$	-126	$\leq -1022$	-1022	$\leq -16382$
Exponent bias	127	unspecified	1023	unspecified

- with each format, numbers of following form can be represented

$$(-1)^s 2^E (b_0.b_1b_2 \cdots b_{p-1})$$

where  $s \in \{0, 1\}$ ,  $E$  is integer satisfying  $E_{\min} \leq E \leq E_{\max}$ , and  $b_i \in \{0, 1\}$

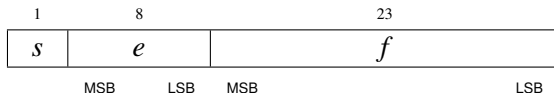
- in addition, can represent four special values:  $+\infty$ ,  $-\infty$ , signaling NaN, and quiet NaN
- NaNs produced by:
  - operations with at least one NaN operand
  - operations yielding indeterminate forms, such as  $0/0$ ,  $(\pm\infty)/(\pm\infty)$ ,  $0 \cdot (\pm\infty)$ ,  $(\pm\infty) \cdot 0$ ,  $(+\infty) + (-\infty)$ , and  $(-\infty) + (\infty)$
  - real operations that yield complex results, such as square root of negative number, logarithm of negative number, inverse sine/cosine of number that lies outside  $[-1, 1]$

- always represent number in normalized form whenever possible; in such cases,  $b_0 = 1$  and  $b_0$  need not be stored explicitly as part of significand
- bit patterns with reserved exponent values (i.e., exponent values that lie outside the range  $E_{\min} \leq E \leq E_{\max}$ ) used to represent  $\pm 0$ ,  $\pm \infty$ , denormalized numbers, and NaNs
- each of (basic) formats consist of three fields:
  - a sign bit,  $s$
  - a biased exponent,  $e = E + \text{bias}$
  - a fraction,  $f = .b_1b_2 \cdots b_{p-1}$
- only difference between formats is size of biased exponent and fraction fields
- value represented by basic format number related to its sign, exponent, and fraction field, but relationship is complicated by the presence of zeros, infinities, and NaNs
- “strange” combination of biased and sign-magnitude formats used to encode floating-point value chosen so that nonnegative floating-point values ordered in same way as integers, allowing integer comparison to compare floating-point numbers

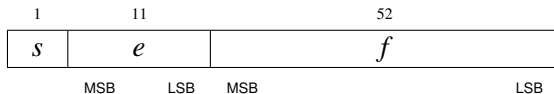


# IEEE 754 Basic Formats (Continued)

## ■ single format:



## ■ double format:



## ■ summary of encodings:

Case	Exponent	Fraction	Value
Normal	$E_{\min} \leq E \leq E_{\max}$	—	$(-1)^s 2^E (1 + f)$
Denormal	$E = E_{\min} - 1$	$f \neq 0$	$(-1)^s 2^{E_{\min}} f$
Zero	$E = E_{\min} - 1$	$f = 0$	$(-1)^s 0$
Infinity	$E = E_{\max} + 1$	$f = 0$	$(-1)^s \infty$
NaN	$E = E_{\max} + 1$	$f \neq 0$	NaN

# IEEE 754 Encoding Examples

- How would the number  $5.25_{10}$  be represented in single format?
- $5.25_{10} = 101.01_2 \cdot 2^0 = 1.0101_2 \cdot 2^2$
- Therefore,  $s = 0$ ,  $e = 2_{10} + 127_{10} = 129_{10} = 10000001_2$ , and  $f = 0101000 \dots 0$ , resulting in the word:

0	10000001	010100000000000000000000
$s$	$e$	$f$

- How would the number  $-9.125_{10}$  be represented in double format?
- $-9.125_{10} = -1001.001_2 \cdot 2^0 = -1.001001_2 \cdot 2^3$
- Therefore,  $s = 1$ ,  $e = 3_{10} + 1023_{10} = 1026_{10} = 10000000010_2$ , and  $f = 001001000 \cdots 0$ , resulting in the word:

[illegible]

- Understand the impact of using finite-precision arithmetic.
- Do not make invalid assumptions about the set of values that can be represented by a particular fixed-point or floating-point type.
- Integer arithmetic can *overflow*. Be careful to avoid overflow.
- Floating-point arithmetic can *overflow and underflow*.
- Perhaps, more importantly, however, floating-point arithmetic has *roundoff error*. If you are not deeply troubled by the presence of roundoff error, you should be as it can cause major problems in many situations.

- 1 D. Goldberg. [What every computer scientist should know about floating-point arithmetic.](#)  
*ACM Computing Surveys*, 23(1):5–48, Mar. 1991
- 2 IEEE Std. 754-1985 — IEEE standard for binary floating-point arithmetic, 1985
- 3 IEEE Std. 754-2008 — IEEE standard for floating-point arithmetic, 2008

- 1 John Farrier. Demystifying Floating Point. CppCon, Bellevue, WA, USA, Sept. 24, 2015. Available online at <https://youtu.be/k12BJGSc2Nc>.

## Section 6.5

# Interval Arithmetic

# Interval Arithmetic

- interval arithmetic is technique for placing bounds on error in numerical computation
- often values provided as input to numerical computation not known exactly, rather only known to within certain tolerance
- uncertainty may be due to measurement error or other factors
- consider numerous measured quantities that are provided as input to some numerical computation
- since measured quantity never known exactly (as measurement always introduces uncertainty), more natural to represent quantity by range
- therefore, would be convenient to have form of arithmetic that operates on values that correspond to ranges
- this is essentially what interval arithmetic does
- interval arithmetic represents each value as range of possibilities and defines set of rules for performing arithmetic on these ranges

- rounding error analysis in numerical algorithms
- filtered robust geometric predicates
- robustly finding intersection of curves and surfaces
- more robust root finding in ray tracing
- computing optimal solutions to geometric matching problems under bounded error
- finding polygonal approximations of implicit curves
- computer-assisted mathematical proofs



- in real interval arithmetic, each value is represented as real interval:

$$[a_1, a_2] = \{x \in \mathbb{R} \mid a_1 \leq x \leq a_2\}$$

- addition, subtraction, and multiplication defined as:

$$A + B = \{a + b \mid a \in A \wedge b \in B\}$$

$$A - B = \{a - b \mid a \in A \wedge b \in B\}$$

$$A \cdot B = \{a \cdot b \mid a \in A \wedge b \in B\}$$

- assuming division by interval containing 0 is not allowed, division defined as:

$$A/B = \{a/b \mid a \in A \wedge b \in B\}$$

- addition:

$$A + B = [a_1, a_2] + [b_1, b_2] = [a_1 + b_1, a_2 + b_2]$$

- negation:

$$-B = -[b_1, b_2] = [-b_2, -b_1]$$

- formula for negation follows from fact that:

- $x \geq b_1 \Rightarrow -x \leq -b_1$  and
- $x \leq b_2 \Rightarrow -x \geq -b_2$

- subtraction:

$$A - B = [a_1, a_2] - [b_1, b_2] = [a_1 - b_2, a_2 - b_1]$$

- formula for subtraction follows from combining addition and negation

# Multiplication and Division

- multiplication:

$$A \cdot B = [a_1, a_2] \cdot [b_1, b_2] = [\min\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}, \max\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}]$$

(e.g.,  $[a_1, a_2] \cdot [b_1, b_2] = [a_1b_1, a_2b_2]$  if  $0 \leq a_1 \leq a_2$  and  $0 \leq b_1 \leq b_2$ )

- reciprocal (assuming division by interval containing 0 not allowed):

$$1/B = 1/[b_1, b_2] = [1/b_2, 1/b_1]$$

- formula for reciprocal follows from fact that, since  $0 \notin [b_1, b_2]$ ,  $x \in [b_1, b_2]$ ,  $b_1, b_2$  all have same sign (implying  $b_1x > 0$  and  $b_2x > 0$ ) and consequently:

$$\square x \geq b_1 \Rightarrow \frac{x}{b_1x} \geq \frac{b_1}{b_1x} \Rightarrow 1/b_1 \geq 1/x \Rightarrow 1/x \leq 1/b_1$$

$$\square x \leq b_2 \Rightarrow \frac{x}{b_2x} \leq \frac{b_2}{b_2x} \Rightarrow 1/b_2 \leq 1/x \Rightarrow 1/x \geq 1/b_2$$

- division (assuming division by interval containing 0 not allowed):

$$A/B = [a_1, a_2]/[b_1, b_2] = [\min\{a_1/b_1, a_1/b_2, a_2/b_1, a_2/b_2\}, \max\{a_1/b_1, a_1/b_2, a_2/b_1, a_2/b_2\}]$$

- formula for division follows from fact that division is simply multiplication by reciprocal

# Allowing Division By Interval Containing Zero

- consider implications of allowing division by interval containing zero
- reciprocal, if  $0 \in [b_1, b_2]$ :

$$1/B = 1/[b_1, b_2] = \begin{cases} (-\infty, 1/b_1] & b_1 \neq 0, b_2 = 0 \\ [1/b_2, +\infty) & b_1 = 0, b_2 \neq 0 \\ (-\infty, 1/b_1] \cup [1/b_2, +\infty) & b_1 \neq 0, b_2 \neq 0 \\ \emptyset & b_1 = b_2 = 0 \end{cases}$$

- thus, if division by interval containing 0 is allowed, result cannot always be represented by interval of form

$$[a_1, a_2] = \{x \in \mathbb{R} \mid a_1 \leq x \leq a_2\}$$

- in particular, arithmetic can yield result that corresponds to:
  - interval unbounded at one end
  - empty set
  - union of two separate intervals

## Allowing Division By Interval Containing Zero (Continued)

- to accommodate division by interval containing zero, represent sets of following forms:

$$\begin{aligned}[a_1, a_2] &= \{x \in \mathbb{R} \mid a_1 \leq x \leq a_2\} \\ [a_1, +\infty) &= \{x \in \mathbb{R} \mid x \geq a_1\} \\ (-\infty, a_2] &= \{x \in \mathbb{R} \mid x \leq a_2\} \\ (-\infty, +\infty) \\ \emptyset\end{aligned}$$

- for sake of simplicity, result of form  $(-\infty, \beta_1] \cup [\beta_2, +\infty)$  (where  $\beta_1 < \beta_2$ ) is mapped to  $(-\infty, +\infty)$

# Floating-Point Interval Arithmetic

- in case of floating-point interval arithmetic, interval bounds are floating-point values
- that is, represent intervals of following form, where  $F$  is set of machine-representable real numbers:

$$[a_1, a_2] = \{x \in F \mid a_1 \leq x \leq a_2\}$$

- since floating-point value can only represent finite number of real numbers, some real numbers cannot be represented exactly
- when arithmetic operation performed, result must always be rounded to machine-representable value
- processor typically allows for control over how rounding performed by supporting several rounding modes, such as:
  - round to nearest
  - round towards zero (i.e., truncate)
  - round upwards (i.e., towards  $+\infty$ )
  - round downwards (i.e., towards  $-\infty$ )

## Floating-Point Interval Arithmetic (Continued)

- must ensure that rounding does not cause interval to no longer bracket result that would be obtained by (exact) real interval arithmetic
- need to select shortest interval that contains result that would be obtained from (exact) real interval arithmetic
- lower bound of result must be computed with rounding downwards
- upper bound of result must be computed with rounding upwards
- using rounding in this way ensures that resulting interval will bracket idealized (exact real) interval

# Floating-Point Interval Arithmetic Operations

$$A + B = [a_1, a_2] + [b_1, b_2] = [a_1 \nabla b_1, a_2 \triangle b_2]$$

$$A - B = [a_1, a_2] - [b_1, b_2] = [a_1 \nabla b_2, a_2 \triangle b_1]$$

$$A \cdot B = [a_1, a_2] \cdot [b_1, b_2]$$

	$b_2 \leq 0$	$b_1 < 0 < b_2$	$b_1 \geq 0$
$a_2 \leq 0$	$[a_2 \nabla b_2, a_1 \triangle b_1]$	$[a_1 \nabla b_2, a_1 \triangle b_1]$	$[a_1 \nabla b_2, a_2 \triangle b_1]$
$a_1 < 0 < a_2$	$[a_2 \nabla b_1, a_1 \triangle b_1]$	$[\min\{a_1 \nabla b_2, a_2 \nabla b_1\}, \max\{a_1 \triangle b_1, a_2 \triangle b_2\}]$	$[a_1 \nabla b_2, a_2 \triangle b_2]$
$a_1 \geq 0$	$[a_2 \nabla b_1, a_1 \triangle b_2]$	$[a_2 \nabla b_1, a_2 \triangle b_2]$	$[a_1 \nabla b_1, a_2 \triangle b_2]$

$$A/B = [a_1, a_2]/[b_1, b_2] \text{ where } 0 \notin [b_1, b_2]$$

	$b_2 < 0$	$b_1 > 0$
$a_2 \leq 0$	$[a_2 \nabla b_1, a_1 \triangle b_2]$	$[a_1 \nabla b_1, a_2 \triangle b_2]$
$a_1 < 0 < a_2$	$[a_2 \nabla b_2, a_1 \triangle b_2]$	$[a_1 \nabla b_1, a_2 \triangle b_1]$
$a_1 \geq 0$	$[a_2 \nabla b_2, a_1 \triangle b_1]$	$[a_1 \nabla b_2, a_2 \triangle b_1]$



- to accommodate division by intervals containing zero, represent intervals of form

$$\begin{aligned}[a_1, a_2] &= \{x \in \mathbb{R} \mid a_1 \leq x \leq a_2\} \\ [a_1, +\infty) &= \{x \in \mathbb{R} \mid x \geq a_1\} \\ (-\infty, a_2] &= \{x \in \mathbb{R} \mid x \leq a_2\} \\ (-\infty, +\infty) & \\ \emptyset &\end{aligned}$$

- arithmetic operations as defined on subsequent slides
- if any operand is  $\emptyset$ , result of operation is also  $\emptyset$

# Addition and Subtraction

SKIP SLIDE

$$A + B$$

	$(-\infty, b_2]$	$[b_1, b_2]$	$[b_1, +\infty)$	$(-\infty, +\infty)$
$(-\infty, a_2]$	$(-\infty, a_2 \triangle b_2]$	$(-\infty, a_2 \triangle b_2]$	$(-\infty, +\infty)$	$(-\infty, +\infty)$
$[a_1, a_2]$	$(-\infty, a_2 \triangle b_2]$	$[a_1 \nabla b_1, a_2 \triangle b_2]$	$[a_1 \nabla b_1, +\infty)$	$(-\infty, +\infty)$
$[a_1, +\infty)$	$(-\infty, +\infty)$	$[a_1 \nabla b_1, +\infty)$	$[a_1 \nabla b_1, +\infty)$	$(-\infty, +\infty)$
$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$

$$A - B$$

	$(-\infty, b_2]$	$[b_1, b_2]$	$[b_1, +\infty)$	$(-\infty, +\infty)$
$(-\infty, a_2]$	$(-\infty, +\infty)$	$(-\infty, a_2 \triangle b_1]$	$(-\infty, a_2 \triangle b_1]$	$(-\infty, +\infty)$
$[a_1, a_2]$	$[a_1 \nabla b_2, +\infty)$	$[a_1 \nabla b_2, a_2 \triangle b_1]$	$(-\infty, a_2 \triangle b_1]$	$(-\infty, +\infty)$
$[a_1, +\infty)$	$[a_1 \nabla b_2, +\infty)$	$[a_1 \nabla b_2, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$
$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$

$$A \cdot B$$

	$[b_1, b_2]$ $b_2 \leq 0$	$[b_1, b_2]$ $b_1 < 0 < b_2$	$[b_1, b_2]$ $b_1 \geq 0$	$[0, 0]$
$[a_1, a_2]$ $a_2 \leq 0$	$[a_2 \nabla b_2, a_1 \triangle b_1]$	$[a_1 \nabla b_2, a_1 \triangle b_1]$	$[a_1 \nabla b_2, a_2 \triangle b_1]$	$[0, 0]$
$[a_1, a_2]$ $a_1 < 0 < a_2$	$[a_2 \nabla b_1, a_1 \triangle b_1]$	$[\min\{a_1 \nabla b_2, a_2 \nabla b_1\}, \max\{a_1 \triangle b_1, a_2 \triangle b_2\}]$	$[a_1 \nabla b_2, a_2 \triangle b_2]$	$[0, 0]$
$[a_1, a_2]$ $a_1 \geq 0$	$[a_2 \nabla b_1, a_1 \triangle b_2]$	$[a_2 \nabla b_1, a_2 \triangle b_2]$	$[a_1 \nabla b_1, a_2 \triangle b_2]$	$[0, 0]$
$[0, 0]$	$[0, 0]$	$[0, 0]$	$[0, 0]$	$[0, 0]$
$(-\infty, a_2]$ $a_2 \leq 0$	$[a_2 \nabla b_2, +\infty)$	$(-\infty, +\infty)$	$(-\infty, a_2 \triangle b_1]$	$[0, 0]$
$(-\infty, a_2]$ $a_2 \geq 0$	$[a_2 \nabla b_1, +\infty)$	$(-\infty, +\infty)$	$(-\infty, a_2 \triangle b_2]$	$[0, 0]$
$[a_1, +\infty)$ $a_1 \leq 0$	$(-\infty, a_1 \triangle b_1]$	$(-\infty, +\infty)$	$[a_1 \nabla b_2, +\infty)$	$[0, 0]$
$[a_1, +\infty)$ $a_1 \geq 0$	$(-\infty, a_1 \triangle b_2]$	$(-\infty, +\infty)$	$[a_1 \nabla b_1, +\infty)$	$[0, 0]$
$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$[0, 0]$

# Multiplication (Continued)

SKIP SLIDE

$A \cdot B$

	$(-\infty, b_2]$ $b_2 \leq 0$	$(-\infty, b_2]$ $b_2 \geq 0$	$[b_1, +\infty)$ $b_1 \leq 0$	$[b_1, +\infty)$ $b_1 \geq 0$	$(-\infty, +\infty)$
$[a_1, a_2]$ $a_2 \leq 0$	$[a_2 \nabla b_2, +\infty)$	$[a_1 \nabla b_2, +\infty)$	$(-\infty, a_1 \triangle b_1]$	$(-\infty, a_2 \triangle b_1]$	$(-\infty, +\infty)$
$[a_1, a_2]$ $a_1 < 0 < a_2$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$
$[a_1, a_2]$ $a_1 \geq 0$	$(-\infty, a_1 \triangle b_2]$	$(-\infty, a_2 \triangle b_2]$	$[a_2 \nabla b_1, +\infty)$	$[a_1 \nabla b_1, +\infty)$	$(-\infty, +\infty)$
$[0, 0]$	$[0, 0]$	$[0, 0]$	$[0, 0]$	$[0, 0]$	$[0, 0]$
$(-\infty, a_2]$ $a_2 \leq 0$	$[a_2 \nabla b_2, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, a_2 \triangle b_1]$	$(-\infty, +\infty)$
$(-\infty, a_2]$ $a_2 \geq 0$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$
$[a_1, +\infty)$ $a_1 \leq 0$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$
$[a_1, +\infty)$ $a_1 \geq 0$	$(-\infty, a_1 \triangle b_2]$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$[a_1 \nabla b_1, +\infty)$	$(-\infty, +\infty)$
$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$

$$A/B, 0 \notin B$$

	$\begin{matrix} [b_1, b_2] \\ b_2 < 0 \end{matrix}$	$\begin{matrix} [b_1, b_2] \\ b_1 > 0 \end{matrix}$	$\begin{matrix} (-\infty, b_2] \\ b_2 < 0 \end{matrix}$	$\begin{matrix} [b_1, +\infty) \\ b_1 > 0 \end{matrix}$
$\begin{matrix} [a_1, a_2] \\ a_2 \leq 0 \end{matrix}$	$[a_2 \nabla b_1, a_1 \triangle b_2]$	$[a_2 \nabla b_1, a_2 \triangle b_2]$	$[0, a_1 \triangle b_2]$	$[a_1 \nabla b_1, 0]$
$\begin{matrix} [a_1, a_2] \\ a_1 < 0 < a_2 \end{matrix}$	$[a_2 \nabla b_2, a_1 \triangle b_2]$	$[a_1 \nabla b_1, a_2 \triangle b_1]$	$[a_2 \nabla b_2, a_1 \triangle b_2]$	$[a_1 \nabla b_1, a_2 \triangle b_1]$
$\begin{matrix} [a_1, a_2] \\ a_1 \geq 0 \end{matrix}$	$[a_2 \nabla b_2, a_1 \triangle b_1]$	$[a_1 \nabla b_2, a_2 \triangle b_1]$	$[a_2 \nabla b_2, 0]$	$[0, a_2 \triangle b_1]$
$[0, 0]$	$[0, 0]$	$[0, 0]$	$[0, 0]$	$[0, 0]$
$\begin{matrix} (-\infty, a_2] \\ a_2 \leq 0 \end{matrix}$	$[a_2 \nabla b_1, +\infty)$	$(-\infty, a_2 \triangle b_2]$	$[0, +\infty)$	$(-\infty, 0]$
$\begin{matrix} (-\infty, a_2] \\ a_2 \geq 0 \end{matrix}$	$[a_2 \nabla b_2, +\infty)$	$(-\infty, a_2 \triangle b_1]$	$[a_2 \nabla b_2, +\infty)$	$(-\infty, a_2 \triangle b_1]$
$\begin{matrix} [a_1, +\infty) \\ a_1 \leq 0 \end{matrix}$	$(-\infty, a_1 \triangle b_2]$	$[a_1 \nabla b_1, +\infty)$	$(-\infty, a_1 \triangle b_2]$	$[a_1 \nabla b_1, +\infty)$
$\begin{matrix} [a_1, +\infty) \\ a_1 \geq 0 \end{matrix}$	$(-\infty, a_1 \triangle b_1]$	$(a_1 \nabla b_2, +\infty)$	$(-\infty, 0]$	$[0, +\infty)$
$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$

# Division (Continued)

SKIP SLIDE

$$A/B, 0 \in B$$

	$[0, 0]$	$\begin{matrix} [b_1, b_2] \\ b_1 < b_2 = 0 \end{matrix}$	$\begin{matrix} [b_1, b_2] \\ 0 = b_1 < b_2 \end{matrix}$	$\begin{matrix} (-\infty, b_2] \\ b_2 = 0 \end{matrix}$	$\begin{matrix} [b_1, +\infty) \\ b_1 = 0 \end{matrix}$	$(-\infty, +\infty)$
$\begin{matrix} [a_1, a_2] \\ a_2 < 0 \end{matrix}$	$\emptyset$	$[a_2 \bigvee b_1, +\infty)$	$(-\infty, a_2 \bigtriangleup b_2]$	$[0, +\infty)$	$(-\infty, 0]$	$(-\infty, +\infty)$
$\begin{matrix} [a_1, a_2] \\ a_1 \leq 0 \leq a_2 \end{matrix}$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$
$\begin{matrix} [a_1, a_2] \\ a_1 > 0 \end{matrix}$	$\emptyset$	$(-\infty, a_1 \bigtriangleup b_1]$	$[a_1 \bigvee b_2, +\infty)$	$(-\infty, 0]$	$[0, +\infty)$	$(-\infty, +\infty)$
$\begin{matrix} (-\infty, a_2] \\ a_2 < 0 \end{matrix}$	$\emptyset$	$[a_2 \bigvee b_1, +\infty)$	$(-\infty, a_2 \bigtriangleup b_2]$	$[0, +\infty)$	$(-\infty, 0]$	$(-\infty, +\infty)$
$\begin{matrix} (-\infty, a_2] \\ a_2 > 0 \end{matrix}$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$
$\begin{matrix} [a_1, +\infty) \\ a_1 < 0 \end{matrix}$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$
$\begin{matrix} (a_1, +\infty) \\ a_1 > 0 \end{matrix}$	$\emptyset$	$(-\infty, a_1 \bigtriangleup b_1]$	$[a_1 \bigvee b_2, +\infty)$	$(-\infty, 0]$	$[0, +\infty)$	$(-\infty, +\infty)$
$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$

# Comparisons

- definition of comparison operations introduces some complications
- many ways in which comparison operations might be defined
- for comparison operator  $\circ$  (i.e., equality, inequality, less than, greater than, less than or equal, greater than or equal), one possible way to define  $[a_1, a_2] \circ [b_1, b_2]$  would be as follows:
  - yields true if  $x \circ y$  is satisfied for all  $x \in [a_1, a_2]$  and all  $y \in [b_1, b_2]$ ,
  - yields false if  $x \circ y$  is violated for all  $x \in [a_1, a_2]$  and all  $y \in [b_1, b_2]$ ,
  - yields indeterminate (or throws exception) otherwise
- for example, with preceding definition:
  - $[0, 1] \leq [1, 2]$  would be true
  - $[0, 1] \leq [-2, -1]$  would be false
  - $[0, 2] \leq [1, 3]$  would be indeterminate
  - $[0, 1] = [0, 0]$  would be indeterminate
- above definition of comparison operations is particularly useful in number applications

# Setting and Querying Rounding Mode

- header file `cfenv` contains various information relevant to floating-point environment
- defines macros (which expand to nonnegative integral constants) for following rounding modes:
  - `FE_TOWARDZERO`: round towards zero
  - `FE_TONEAREST`: round to nearest representable value
  - `FE_UPWARD`: round towards positive infinity
  - `FE_DOWNWARD`: round towards negative infinity
- current rounding mode can be set with `std::fesetround`
- `int fesetround(int round)`
  - attempts to set current rounding mode to `round`
  - returns 0 upon success non-zero value otherwise
- current rounding mode can be queried with `std::fegetround`
- `int fegetround()`
  - returns value of current rounding mode
- floating-point environment access and modification only meaningful when `#pragma STDC FENV_ACCESS` is supported and set to ON



# Impact of Current Rounding Mode

- current rounding mode *affects*:
  - results of floating-point arithmetic operations outside of constant expressions
  - results of standard library mathematical functions (e.g., `sin`, `cos`, `tan`, `exp`, `log`, and `sqrt`)
  - floating-point to floating-point implicit conversion and casts
  - string conversions (e.g., `strtod`)
  - library rounding functions `nearbyint`, `rint`, and `lrint`
- current rounding mode *does not affect*:
  - floating-point to integer implicit conversions and casts (which are always towards zero)
  - results of floating-point arithmetic operations in constant expressions (which are always to nearest)
  - library functions `round`, `lround`, `ceil`, `floor`, and `trunc`
- behavior of many things affected by current rounding mode
- since some algorithms may rely on use of particular rounding mode, one must be careful to always restore previous rounding mode

# Rounding Mode Example

```
1  #include <iostream>
2  #include <cmath>
3  #include <cfenv>
4  #include <limits>
5
6  #pragma STDC FENV_ACCESS ON
7
8  int main() {
9      std::cout.precision(std::numeric_limits<double>::max_digits10);
10     int old_mode = std::fegetround();
11     int modes[] = {FE_TONEAREST, FE_TOWARDZERO, FE_UPWARD, FE_DOWNWARD};
12     for (auto mode : modes) {
13         if (std::fesetround(mode)) {abort();}
14         std::cout << std::sqrt(2.0) << '\n';
15     }
16     if (std::fesetround(old_mode)) {abort();}
17     std::cout << std::sqrt(2.0) << '\n';
18 }
19
20 /* Example output:
21 1.4142135623730951
22 1.4142135623730951
23 1.4142135623730952
24 1.4142135623730951
25 1.4142135623730951
26 */
```

## Section 6.5.1

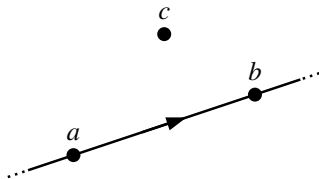
# **Applications in Geometry Processing**

- interval arithmetic frequently employed in geometry processing
- one application of interval arithmetic is for efficient implementation of exact geometric predicates
- geometric predicate tests for one of small number of possibilities involving geometric objects such as points, lines, and planes
- some basic geometric predicates include tests for such things as:
  - on which side of oriented line point located (i.e., 2-dimensional orientation test)
  - on which side of oriented plane point located (i.e., 3-dimensional orientation test)
  - on which side of circle point located (i.e., in-circle test)
  - on which side of sphere point located (i.e., in-sphere test)
- geometric predicates like those above essential in many geometric algorithms
- exact predicate is one that must always yield correct result

# Filtered Geometric Predicates

- determining result of geometric predicate involves arithmetic computation
- if arithmetic used for computation not exact, predicate may yield incorrect result
- vast majority of algorithms cannot tolerate incorrect results from predicates
- unfortunately, using exact arithmetic extremely costly
- use interval arithmetic to quickly determine bound on numerical results of interest
- if bound obtained from interval arithmetic sufficient to make determination of predicate result, high cost of using exact arithmetic avoided
- only if bound insufficient, recompute result using exact arithmetic
- in practice, interval arithmetic often sufficient to determine predicate result, leading to great increase in efficiency

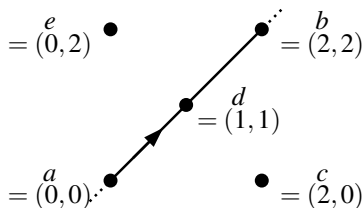
# Two-Dimensional Orientation Test



- given three points  $a = (a_x, a_y)$ ,  $b = (b_x, b_y)$ , and  $c = (c_x, c_y)$  in  $\mathbb{R}^2$ , determine to which side of directed line through  $a$  and  $b$  point  $c$  lies
- can be determined from sign of determinant of  $2 \times 2$  matrix
- $\text{orient2d}(a, b, c) = \det \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} a_x - c_x & b_x - c_x \\ a_y - c_y & b_y - c_y \end{bmatrix}$
- if  $\text{orient2d}(a, b, c)$  is positive, negative, or zero, then  $c$  is respectively to left of, to right of, or collinear with directed line through  $a$  and  $b$

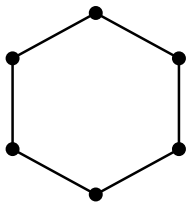
# Example: Two-Dimensional Orientation Test

SKIP SLIDE

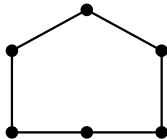


- $\text{orient2d}(a,b,c) = \det \begin{bmatrix} 0 & 2 & 2 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix} = -4 < 0$ ;  $c$  is right of oriented line through  $a$  and  $b$
- $\text{orient2d}(a,b,d) = \det \begin{bmatrix} 0 & 2 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} = 0$ ;  $d$  is on oriented line through  $a$  and  $b$
- $\text{orient2d}(a,b,e) = \det \begin{bmatrix} 0 & 2 & 0 \\ 0 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} 0 & 2 \\ -2 & 0 \end{bmatrix} = 4 > 0$ ;  $e$  is left of oriented line through  $a$  and  $b$

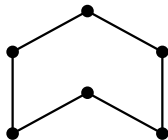
# Convex Polygons



Strictly Convex



Convex But Not Strictly  
Convex

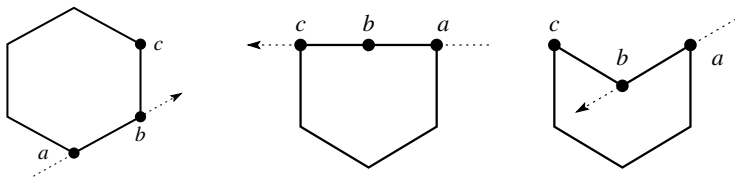


Nonconvex

- A polygon is said to be **convex** if every line segment between two points on the polygon's boundary is contained strictly inside or on the boundary of the polygon.
- A polygon is said to be **strictly convex** if it is convex and all of its interior angles are strictly less than 180 degrees.

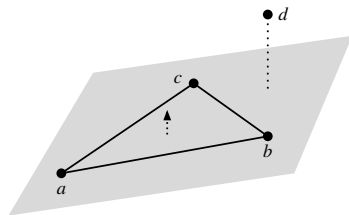


# Polygon Convexity Test



- let  $a$ ,  $b$ , and  $c$  be three consecutive vertices of polygon in counterclockwise (CCW) order
- polygon is *strictly convex* if and only if, for every choice of  $a, b, c$ ,  $c$  is to left of directed line through  $ab$  (i.e.,  $\text{orient2d}(a, b, c) > 0$ )
- polygon is *convex* if and only if, for every choice of  $a, b$ , and  $c$ ,  $c$  is to left of or collinear with directed line through  $ab$  (i.e.,  $\text{orient2d}(a, b, c) \geq 0$ )

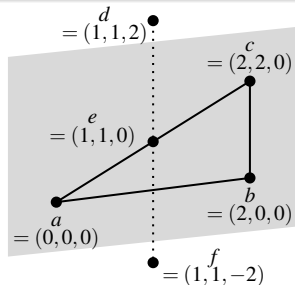
# Three-Dimensional Orientation Test



- given four points  $a = (a_x, a_y, a_z)$ ,  $b = (b_x, b_y, b_z)$ ,  $c = (c_x, c_y, c_z)$ , and  $d = (d_x, d_y, d_z)$  in  $\mathbb{R}^3$ , determine to which side of oriented plane through  $a$ ,  $b$ , and  $c$  point  $d$  lies; above side of oriented plane is side from which  $a$ ,  $b$ , and  $c$  appear in CCW order
- can be determined from sign of determinant of  $3 \times 3$  matrix
- $\text{orient3d}(a, b, c, d) = \det \begin{bmatrix} a_x & b_x & c_x & d_x \\ a_y & b_y & c_y & d_y \\ a_z & b_z & c_z & d_z \\ 1 & 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} a_x - d_x & b_x - d_x & c_x - d_x \\ a_y - d_y & b_y - d_y & c_y - d_y \\ a_z - d_z & b_z - d_z & c_z - d_z \end{bmatrix}$
- if  $\text{orient3d}(a, b, c, d)$  is positive, negative, or zero, then  $d$  is respectively below, above, or on oriented plane through  $a$ ,  $b$ , and  $c$

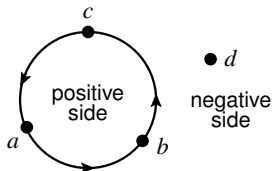
# Example: Three-Dimensional Orientation Test

SKIP SLIDE



- $\text{orient3d}(a,b,c,d) = \det \begin{bmatrix} 0 & 2 & 2 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} -1 & 1 & 1 \\ -1 & -1 & 1 \\ -2 & -2 & -2 \end{bmatrix} = -8 < 0$ ;  $d$  above oriented plane through  $a$ ,  $b$ , and  $c$
- $\text{orient3d}(a,b,c,e) = \det \begin{bmatrix} 0 & 2 & 2 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} -1 & 1 & 1 \\ -1 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix} = 0$ ;  $e$  lies in oriented plane through  $a$ ,  $b$ , and  $c$
- $\text{orient3d}(a,b,c,f) = \det \begin{bmatrix} 0 & 2 & 2 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & -2 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} -1 & 1 & 1 \\ -1 & -1 & 1 \\ 2 & 2 & 2 \end{bmatrix} = 8 > 0$ ;  $f$  below oriented plane through  $a$ ,  $b$ , and  $c$

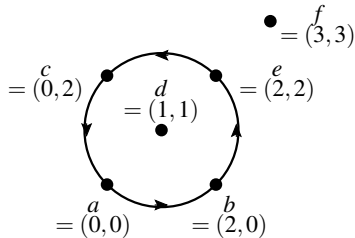
# Side-of-Oriented-Circle Test



- given four points  $a = (a_x, a_y)$ ,  $b = (b_x, b_y)$ ,  $c = (c_x, c_y)$ , and  $d = (d_x, d_y)$  in  $\mathbb{R}^2$ , determine whether  $d$  is on positive side, negative side, or boundary of oriented circle through  $a$ ,  $b$ , and  $c$ ; positive side is side to left of oriented path through  $a$ ,  $b$ , and  $c$
- can be determined from result of 3-dimensional orientation test
- project  $a$ ,  $b$ ,  $c$ , and  $d$  (in third dimension) onto paraboloid  $f(x, y) = x^2 + y^2$  and perform orientation test on resulting four points
- $\text{inCircle}(a, b, c, d) = \text{orient3d}(a', b', c', d')$ , where  $a' = (a_x, a_y, a_x^2 + a_y^2)$ ,  $b' = (b_x, b_y, b_x^2 + b_y^2)$ ,  $c' = (c_x, c_y, c_x^2 + c_y^2)$ , and  $d' = (d_x, d_y, d_x^2 + d_y^2)$
- if  $\text{inCircle}(a, b, c, d)$  is positive, negative, or zero, then  $d$  lies respectively on positive side, negative side, or boundary of oriented circle through  $a$ ,  $b$ , and  $c$

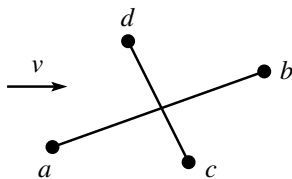
# Example: Side-of-Oriented-Circle Test

SKIP SLIDE



- $\text{inCircle}(a, b, c, d) = \det \begin{bmatrix} 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ 0^2+0^2 & 2^2+0^2 & 0^2+2^2 & 1^2+1^2 \\ 1 & 1 & 1 & 1 \end{bmatrix} = 8 > 0$ ;  $d$  on positive side of oriented circle through  $a$ ,  $b$ , and  $c$  (i.e.,  $d$  inside circle)
- $\text{inCircle}(a, b, c, e) = \det \begin{bmatrix} 0 & 2 & 0 & 2 \\ 0 & 0 & 2 & 2 \\ 0^2+0^2 & 2^2+0^2 & 0^2+2^2 & 2^2+2^2 \\ 1 & 1 & 1 & 1 \end{bmatrix} = 0$ ;  $e$  on oriented circle through  $a$ ,  $b$ , and  $c$
- $\text{inCircle}(a, b, c, f) = \det \begin{bmatrix} 0 & 2 & 0 & 3 \\ 0 & 0 & 2 & 3 \\ 0^2+0^2 & 2^2+0^2 & 0^2+2^2 & 3^2+3^2 \\ 1 & 1 & 1 & 1 \end{bmatrix} = -24 < 0$ ;  $f$  on negative side of oriented circle through  $a$ ,  $b$ , and  $c$  (i.e.,  $f$  outside circle)

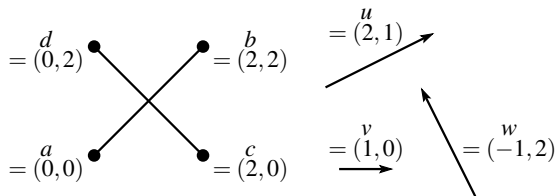
# Preferred-Direction Test



- given two line segments  $ab$  and  $cd$  and vector  $v$ , determine if, compared to orientation of  $cd$ , orientation of  $ab$  is more close, less close, or equally close to the orientation of  $v$
- can be determined from result of computation involving dot products
- $\text{prefDir}(a, b, c, d, v) = |d - c|^2 ((b - a) \cdot v)^2 - |b - a|^2 ((d - c) \cdot v)^2$
- if  $\text{prefDir}(a, b, c, d, v)$  is positive, negative, or zero, then compared to orientation of  $cd$ , orientation of  $ab$  is more close, less close, or equally close to orientation of  $v$ , respectively

# Example: Preferred-Direction Test

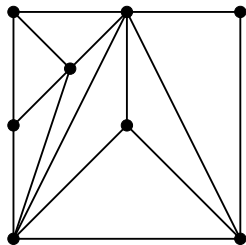
SKIP SLIDE



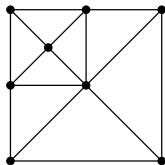
- $\text{prefDir}(a, b, c, d, u) = |(-2, 2)|^2 [(2, 2) \cdot (2, 1)]^2 - |(2, 2)|^2 [(-2, 2) \cdot (2, 1)]^2 = 256 > 0$ ;  $ab$  closer than  $cd$  to direction of  $u$
- $\text{prefDir}(a, b, c, d, v) = |(-2, 2)|^2 [(2, 2) \cdot (1, 0)]^2 - |(2, 2)|^2 [(-2, 2) \cdot (1, 0)]^2 = 0$ ;  $ab$  and  $cd$  equally close to direction of  $v$
- $\text{prefDir}(a, b, c, d, w) = |(-2, 2)|^2 [(2, 2) \cdot (-1, 2)]^2 - |(2, 2)|^2 [(-2, 2) \cdot (-1, 2)]^2 = -256 < 0$ ;  $cd$  closer than  $ab$  to direction of  $w$

# Triangulations

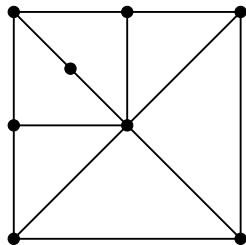
- A **triangulation** of a set  $V$  of vertices is a set  $T$  of triangles such that:
- the union of the vertices of all triangles in  $T$  is  $V$ ;
  - the interiors of any two triangles in  $T$  are disjoint; and
  - the union of the triangles in  $T$  is the convex hull of  $V$ .



Triangulation



Triangulation

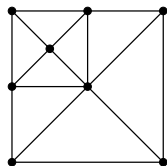


Invalid Triangulation

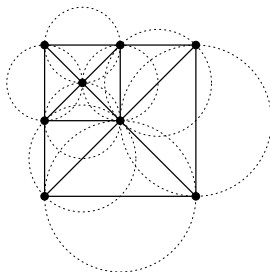


# Delaunay Triangulations

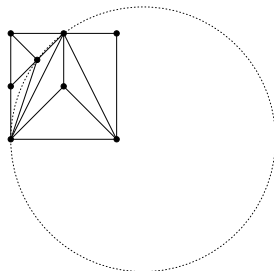
- A triangulation is said to be **Delaunay** if each triangle in the triangulation is such that the interior of its *circumcircle contains no vertices*.



Delaunay  
Triangulation

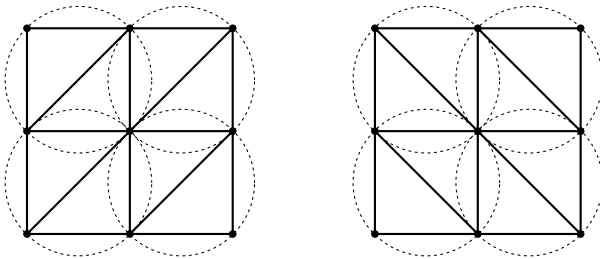


Delaunay Triangulation  
Showing Circumcircles



Non-Delaunay  
Triangulation Showing  
Violation of Circumcircle  
Condition

# Nonuniqueness of Delaunay Triangulations



# Comments on Delaunay Triangulations

- Delaunay triangulation maximizes minimum interior angle of all triangles in triangulation
- avoids long-thin triangles to whatever extent is possible
- long-thin triangles often undesirable for interpolation purposes; can lead to large discretization error and large errors in derivatives
- Delaunay triangulation only guaranteed to be unique if no four points are cocircular
- when not unique, schemes exist for making unique choice from set of all possible Delaunay triangulations, such as one proposed in:

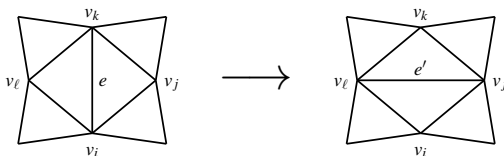
C. Dyken and M. S. Floater. [Preferred directions for resolving the non-uniqueness of Delaunay triangulations.](#)

*Computational Geometry—Theory and Applications*, 34:96–101, 2006.

- dual graph of Delaunay triangulation is Voronoi diagram (circumcircle centers become vertices, original vertices become faces)

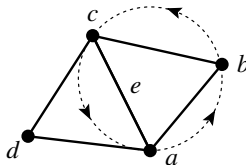
# Edge Flips

- edge  $e$  in triangulation said to be **flippable** if  $e$  has two incident faces (i.e., is not on triangulation boundary) and union of these two faces is strictly convex quadrilateral  $q$ .
- if  $e$  is flippable, valid triangulation obtained if  $e$  deleted from triangulation and replaced by other diagonal  $e'$  of quadrilateral  $q$
- such transformation known as **edge flip**
- edge-flip example:



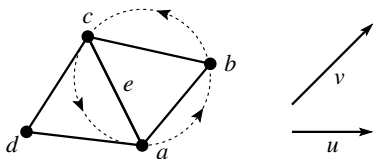
- number of different triangulations of  $n$  vertices upper bounded by  $\binom{n-2}{2} = \frac{n^2-n}{2}$ , which is  $O(n^2)$
- all triangulations of set of vertices have same number of edges
- every triangulation reachable from every other triangulation by edge flips

# Locally-Delaunay Test



- given flippable edge  $e$  in triangulation with incident faces  $abc$  and  $acd$  whose vertices are specified in CCW order (and whose union is strictly convex quadrilateral), determine if  $e$  is locally Delaunay
- result can be determined using side-of-oriented-circle test
- $$\text{localDelaunay}(a, b, c, d) = \begin{cases} 1 & \text{inCircle}(a, b, c, d) \leq 0 \\ 0 & \text{inCircle}(a, b, c, d) > 0 \end{cases}$$
- if  $\text{localDelaunay}(a, b, c, d) \neq 0$ , edge  $e$  is locally Delaunay
- if every flippable edge in triangulation is locally Delaunay, triangulation is Delaunay

# Locally Preferred-Directions Delaunay Test



- given flippable edge  $e$  in triangulation with incident faces  $abc$  and  $acd$  whose vertices are specified in CCW order (and whose union is strictly convex quadrilateral), determine if  $e$  is locally preferred-directions Delaunay with first and second direction vectors  $u$  and  $v$ , respectively (where  $u$  and  $v$  are nonzero and neither parallel nor orthogonal)
- result can be determined using side-of-oriented-circle and preferred-direction tests
- $\alpha(a, b, c, d, u, v)$   
$$= \begin{cases} 1 & \text{prefDir}(a, c, b, d, u) > 0 \\ 1 & \text{prefDir}(a, c, b, d, u) = 0 \text{ and } \text{prefDir}(a, c, b, d, v) > 0 \\ 0 & \text{otherwise} \end{cases}$$

# Locally Preferred-Directions Delaunay Test (Continued)

- $\text{localPrefDirDelaunay}(a, b, c, d, u, v)$   
$$= \begin{cases} 1 & \text{inCircle}(a, b, c, d) < 0 \\ 0 & \text{inCircle}(a, b, c, d) > 0 \\ \alpha(a, b, c, d, u, v) & \text{otherwise} \end{cases}$$
- if (and only if)  $\text{localPrefDirDelaunay}(a, b, c, d, u, v) \neq 0$ , edge  $e$  is locally preferred-directions Delaunay with first and second direction vectors  $u$  and  $v$ , respectively
- if every flippable edge in triangulation is locally preferred-directions Delaunay, triangulation is preferred-directions Delaunay

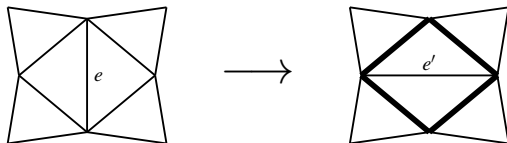
# Lawson Local Optimization Procedure (LOP)

- Lawson local optimization procedure (LOP) finds optimal triangulation of set of points via edge flips
- flippable edge said to be **optimal** if:
  - 1 it is not flippable; or
  - 2 it is flippable and satisfies some optimality criterion, such as the locally-Delaunay or preferred-directions locally-Delaunay condition
- edge said to be **suspect** if its optimality is currently uncertain
- initially, all flippable edges are marked as suspect
- while at least one suspect edge remains, perform following:
  - select suspect edge  $e$
  - if edge  $e$  is optimal, mark  $e$  as not suspect; otherwise, flip  $e$  to obtain edge  $e'$ , mark  $e'$  as not suspect, and mark any edges whose optimality might be affected by flip of  $e$  as suspect
- essentially, LOP simply keeps flipping (flippable) edges that are not optimal until all edges are optimal



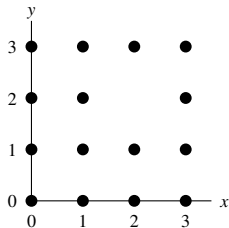
# Finding Delaunay Triangulations with Lawson LOP

- given any triangulation of set  $P$  of points, can compute Delaunay triangulation of  $P$  using Lawson LOP
- select optimality criterion as locally-Delaunay or preferred-directions locally-Delaunay condition
- when edge flipped, which edges can have their optimality affected?
- let  $e$  denote edge being flipped
- let  $q$  denote quadrilateral formed by union of two faces incident on  $e$
- let  $e'$  denote edge obtained by applying edge flip to  $e$
- edges that should be marked as suspect are all flippable edges belonging to  $q$
- for example, if edge  $e'$  was produced by flipping edge  $e$ , would need to mark all edges drawn with thicker line (as shown below) as suspect

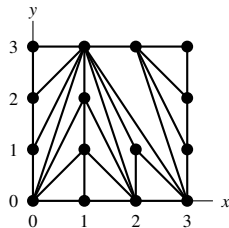


# Preferred-Directions Delaunay Triangulation Example

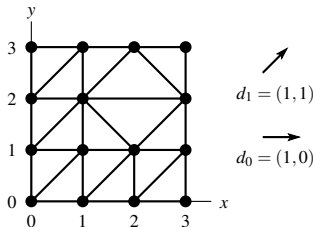
SKIP SLIDE



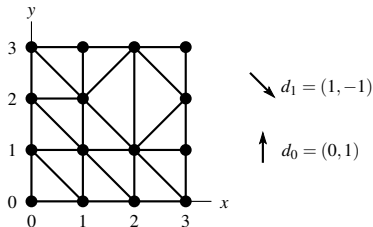
Point Set



Non-Delaunay Triangulation



Delaunay Triangulation With Preferred Directions  $(1, 0)$  and  $(1, 1)$



Delaunay Triangulation With Preferred Directions  $(0, 1)$  and  $(1, -1)$

## Section 6.5.2

### **References**

- 1 U. W. Kulisch. [Complete interval arithmetic and its implementation on the computer.](#)  
In A. Cuyt, W. Kramer, W. Luther, and P. Markstein, editors, *Numerical Validation in Current Hardware Architectures*, pages 7–26.  
Springer-Verlag, Berlin, Germany, 2009.
- 2 G. Bohlender, U. Kulisch, and R. Lohner. [Definition of the arithmetic operations and comparison relations for an interval arithmetic standard,](#) Nov. 2008.
- 3 IEEE Std. 1788-2015 — IEEE standard for interval arithmetic, 2015.
- 4 H. Bronnimann, C. Burnikel, and S. Pion. [Interval arithmetic yields efficient dynamic filters for computational geometry.](#)  
*Discrete Applied Mathematics*, 109:25–47, 2001.

- 5 C. Dyken and M. S. Floater. Preferred directions for resolving the non-uniqueness of Delaunay triangulations.  
*Computational Geometry—Theory and Applications*, 34:96–101, 2006.
- 6 C. L. Lawson. Software for  $C^1$  surface interpolation.  
In J. R. Rice, editor, *Mathematical Software III*, pages 161–194. Academic Press, New York, NY, USA, 1977.
- 7 B. Delaunay. Sur la sphere vide.  
*Bulletin of the Academy of Sciences of the USSR, Classe des Sciences Mathematiques et Naturelle*, 7(6):793–800, 1934.
- 8 J. R. Shewchuk. Adaptive precision floating-point arithmetic and fast robust geometric predicates.  
*Discrete & Computational Geometry*, 18:305–363, 1997.

- 9 S. Fortune and C. J. Van Wyk. [Efficient exact arithmetic for computational geometry.](#)  
In *Proc. of Symposium on Computational Geometry*, pages 163–172, 1993.
- 10 V. Lefevre. [Correctly rounded arbitrary-precision floating-point summation.](#)  
*IEEE Trans. on Computers*, 2017.  
[To appear.](#)

## Section 6.6

### Cache-Efficient Code

# The Memory Latency Problem

- over time, processors have continued to become faster
- speed improvements in memory, however, have not kept pace with processors
- compared to speed of processor, main memory is *very slow*
- consequently, bottlenecks in algorithms can often be due to memory speed
- very substantial amount of complexity in modern processors devoted to reducing impact of memory latency
- particularly important feature for hiding memory latency is cache
- effective utilization of cache often critical to writing high-performance code



## Section 6.6.1

# **Memory Hierarchy and Caches**

# Principle of Locality

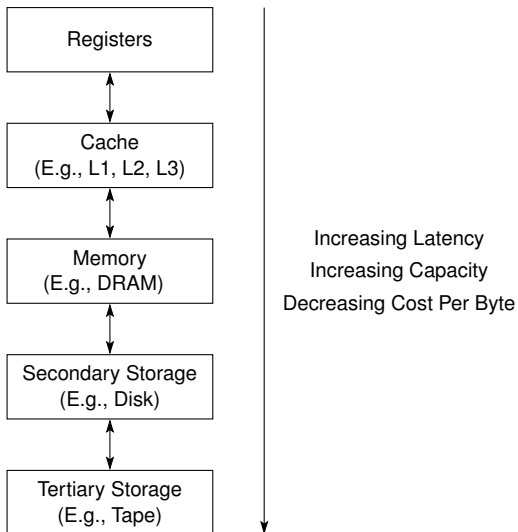
- **locality of reference**: programs do not access all code or data uniformly
- two basic types of locality:
  - 1 temporal
  - 2 spatial
- **temporal locality**: tendency to reuse same information stored in memory within relatively small time interval (e.g., code in loops, top of stack)
- example (where accesses to `i` and `sum` have good temporal locality, due to their repeated use in loop):

```
int func(int);  
int sum = 0;  
for (int i = 0; i < 10000; ++i) {sum += func(i);}
```

- **spatial locality**: tendency to use information stored in nearby locations in memory together (e.g., sequential code, neighbouring elements in array)
- example (where accesses to neighbouring elements of `a` have good spatial locality):

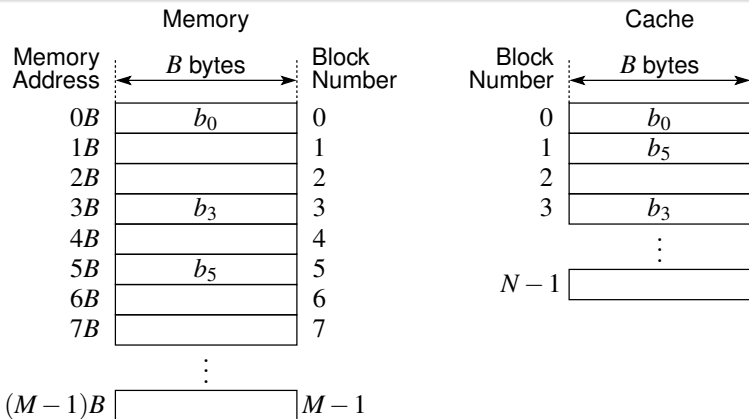
```
int a[1024];  
// ...  
a[42] = a[43] * a[44] + a[45];
```
- to exploit locality, memory hierarchy is employed

# Memory Hierarchy



- **cache**: fast (but relatively small) memory
- **data cache** (a.k.a. D cache): cache that holds only data
- **instruction cache** (a.k.a., I cache): cache that holds only instructions
- **unified cache**: cache that holds both instructions and data
- **translation lookaside buffer (TLB)**: memory cache that stores recent translations of virtual to physical addresses
- may be several levels to cache hierarchy
- **level-1 (L1) cache** closest to processor, while **last-level (LL) cache** farthest
- when processor needs to read or write location, checks cache
- when data needed is available in cache, **cache hit** said to occur
- when data needed cannot be supplied by cache, **cache miss** said to occur
- cache may be local to single core or shared between multiple cores
- L1 cache usually on core and local to core, while higher-level caches often shared between some or all cores

# Memory and Cache



- memory partitioned into blocks of  $B$  bytes (where  $B$  is typically power of two)
- memory comprised of  $M$  blocks for total memory size of  $BM$  bytes
- cache can hold  $N$  blocks for total cache size of  $BN$  bytes
- size of cache much less than size of memory (i.e.,  $BN \ll BM$ )

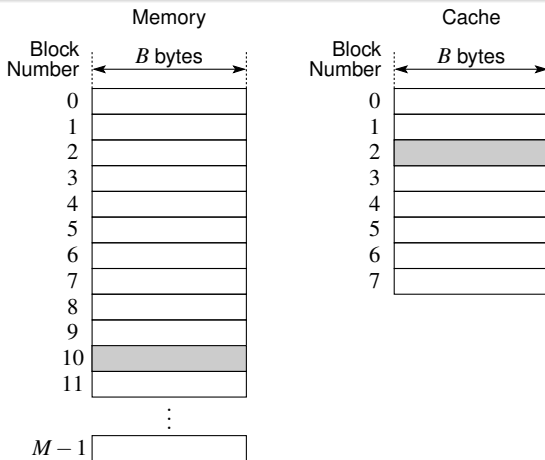
# Block Placement

- **block placement policy**: strategy used to determine where block can be placed in cache
- three basic block placement policies:
  - 1 direct mapped
  - 2 set associative
  - 3 fully associative
- **direct mapped**: each block has only one place it can appear in cache
- typically, memory block  $i$  mapped to cache block  $\text{mod}(i, n)$ , where  $n$  is number of blocks in cache
- **set associative**: block can be placed in restricted number of places in cache; block first mapped to group of blocks in cache called set, and then block can be placed anywhere within that set
- typically, memory block  $i$  can be placed in any cache block in set  $\text{mod}(i, S)$ , where  $S$  is number of sets in cache
- if each set contains  $k$  blocks, called  **$k$ -way set associative**
- **fully associative**: block can be placed anywhere in cache

## Block Placement (Continued)

- strictly speaking, set associative includes direct mapped and fully associative as special cases
- direct mapped equivalent to 1-way set associative
- fully associative equivalent to  $N$ -way set associative, where  $N$  is total number of blocks in cache
- block placement policies typically employ expressions of form  $\text{mod}(n, m)$  where  $m = 2^k$ , since result is simply given by  $k$  least significant bits (LSBs) of  $n$
- for example:
  - $\text{mod}(10, 4) = \text{mod}(1010_2, 2^2) = 10_2 = 2$
  - $\text{mod}(42, 16) = \text{mod}(101010_2, 2^4) = 1010_2 = 10$

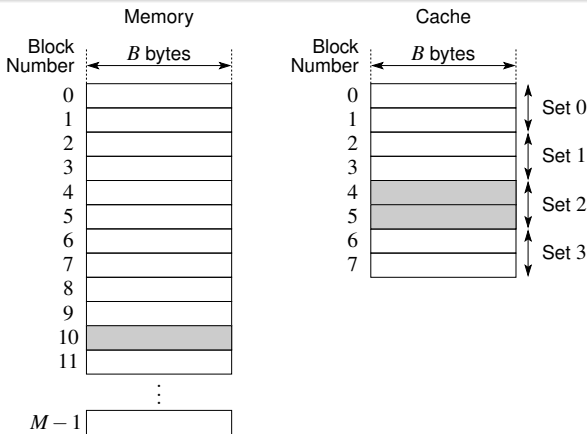
# Direct-Mapped Cache Example



- memory block  $i$  can only be placed in cache block  $\text{mod}(i, N)$ , where  $N$  is number of blocks in cache
- for example, if  $N = 8$  (as above), memory block 10 can only be placed in cache block  $\text{mod}(10, 8) = 2$  [recall:  $\text{mod}(1010_2, 2^3) = 010_2 = 2$ ]

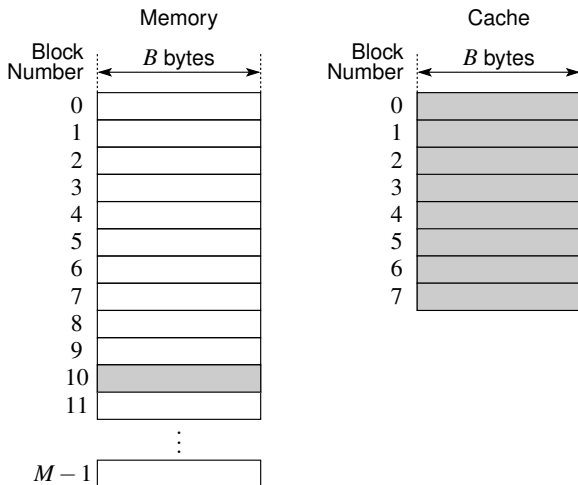


# $K$ -Way Set-Associative Cache Example



- memory block  $i$  can be placed in any of  $K$  cache blocks in set  $\text{mod}(i, S)$ , where  $S$  is number of sets
- for example, if  $S = 4$  and  $K = 2$  (as above), memory block 10 can be placed in any of cache blocks in set  $\text{mod}(10, 4) = 2$  [recall:  $\text{mod}(1010_2, 2^2) = 10_2 = 2$ ]

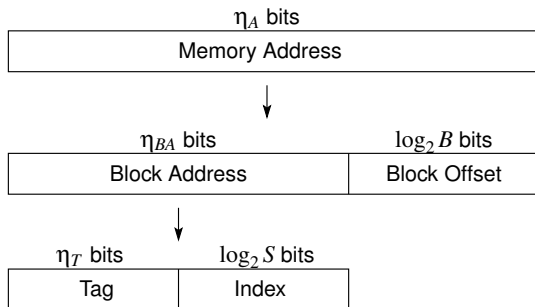
# Fully-Associative Cache Example



- any memory block can be placed in any cache block
- for example, memory block 10 could be placed in any cache block

- **block identification strategy**: method used to find block if in cache
- when address is referenced, need to determine if associated data in cache, and if it is, find it
- without loss of generality, we can consider case of  $K$ -way set associative cache
- memory block  $i$  can be mapped to any block in set  $s = \text{mod}(i, S)$ , where  $S$  is number of sets in cache
- each cache entry is associated with one particular set in cache and contains:
  - valid bit to indicate if cache entry in use
  - tag to identify which block is in cache entry (if block is valid)
  - data for block (if block is valid)

# Decomposition of Memory Address



- $B$  is cache block size,  $N$  is number of blocks in cache,  $K$  is cache associativity, and  $S$  is number of sets (where  $S = N/K$ )
- memory address decomposed into block address and block offset
- block address then decomposed into tag and index
- in fully associative case (i.e.,  $S = 1$ ), index not present
- index  $s$  identifies set in which block  $i$  can be placed (i.e.,  $s = \text{mod}(i, S)$ )

## Cache Entries for $i$ th Set (for $K$ -Way Set Associative Cache)

Valid	Tag	Data
$v_{i,0}$	$t_{i,0}$	$d_{i,0}$
$v_{i,1}$	$t_{i,1}$	$d_{i,1}$
$\vdots$	$\vdots$	$\vdots$
$v_{i,K-1}$	$t_{i,K-1}$	$d_{i,K-1}$

- need to determine if any entry matches tag and (if not fully associative) index
- first determine set in which block can be placed:
  - if not fully associative, determined by index
  - otherwise, cache only has one set
- then look in this set for matching tag
- if match found, cache hit; otherwise, cache miss

# Block Replacement

- **block replacement policy**: strategy used to determine which block should be replaced (i.e., evicted) upon miss when no unused cache entry available
- in case of direct mapped cache, only one choice for block to replace so no freedom in choice of replacement policy
- in case of set-associative or fully-associative cache, have some choice in block to replace
- some commonly-used replacement policies include:
  - 1 random
  - 2 least recently used (LRU)
  - 3 first-in first-out (FIFO)
  - 4 approximate LRU
- **random**: block to be replaced is randomly chosen (often using pseudorandom number generator)
- **least-recently used (LRU)**: block that has not been used for longest time is replaced
- **first-in first-out (FIFO)**: block that has been in cache longest is replaced

# Write Policy

- **write policy**: strategy used to handle writes to memory
- two aspects to write policy:
  - 1 cache-hit policy (i.e., how to handle cache hit)
  - 2 cache-miss policy (i.e., how to handle cache miss)
- two basic write-hit policies:
  - 1 **write through**: information written to both block in cache and block in lower-level memory
  - 2 **write back**: information written only to block in cache; modified cache block written to main memory only when replaced
- two basic write-miss policies:
  - 1 **write allocate** (a.k.a. fetch on write): write miss brings block into cache, followed by write-hit action
  - 2 **no write allocate** (a.k.a. write around): write miss only updates lower-level memory, leaving cache unchanged
- usually, write through used with no write allocate, and write back used with write allocate
- write through always combined with write buffer to avoid always having to wait for lower-level memory

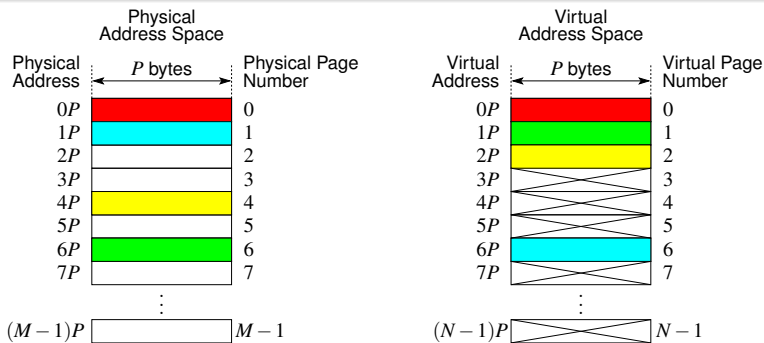
# Cache Misses

- **compulsory miss** (a.k.a. cold miss): miss due to address being accessed for first time (impossible to avoid; misses even with infinite sized cache)
- **capacity miss**: miss due to cache not being large enough (i.e., program working set is much larger than cache capacity resulting in block being evicted from cache and later accessed again)
- **conflict miss**: miss due to limited associativity (i.e., miss that would have been avoided with fully associative cache); occurs when too many blocks mapped to same set resulting in memory locations being mapped to same cache entry
- **coherence miss**: miss due to cache flushes to keep multiple caches consistent (i.e., coherent) in multiprocessor system
- **true sharing miss**: coherence miss that is due to multiple threads sharing same data in cache block
- **false sharing miss**: coherence miss that is due to threads accessing different data that happens to reside in same cache block (i.e., cache block is shared between threads but not data within cache block)



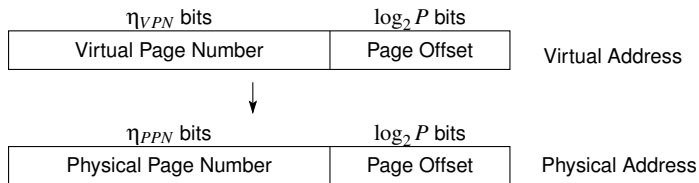
- virtual memory is memory management technique that maps addresses called virtual address into physical addresses in computer memory
- allows amount of memory used by system to exceed that which is physically available
- allows processes to share memory
- provides memory protection
- each process has its own virtual address space
- programs access memory using virtual addresses
- memory management unit (MMU) translates virtual addresses to physical addresses

# Virtual Address Space



- memory partitioned into pages of size  $P$  bytes (where  $P$  is typically power of two)
- physical address space comprised of  $M$  pages
- virtual address space comprised of  $N$  pages
- virtual address space typically at least as large as physical address space (i.e.,  $PN \geq PM$ )
- can arbitrarily map pages in virtual address space to physical pages

# Address Translation



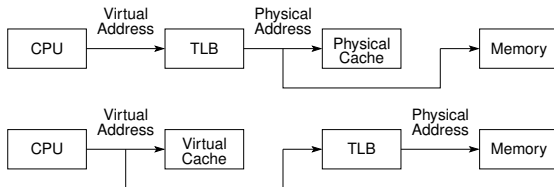
- $P$  is page size
- virtual address and physical address both decomposed into page number and page offset
- address translation only changes page number part of address
- when virtual address translated to physical address, page offset does not change

# Translation Lookaside Buffer (TLB)

- address translation is slow process
- to reduce translation time, use cache called translation lookaside buffer (TLB)
- TLB caches information for address-translation mappings

# Virtual and Physical Caches

- if virtual memory employed, question arises as to whether memory caches should use virtual or physical addressing
- cache that employs physical addressing called **physical cache** (or physically-addressed cache)
- cache that employs virtual addressing called **virtual cache** (or virtually-addressed cache)
- key difference between use of virtual and physical cache is where address translation takes place:



- in case of accessing physical cache, always require address translation
- in case of accessing virtual cache, only need address translation on cache miss

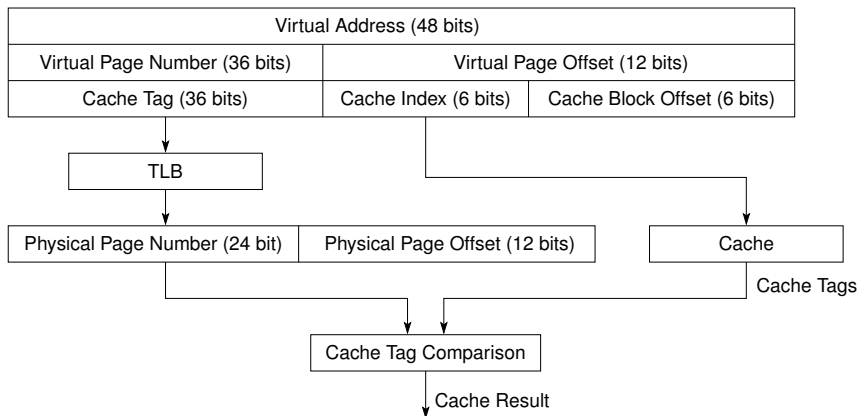
# Virtual Versus Physical Caches

- virtual cache has advantage of eliminating address translation time for cache hit
- virtual cache has disadvantage of introducing numerous complications:
  - same virtual address (in different processes) can refer to distinct physical addresses (which is typically resolved by adding process ID to virtual address instead of flushing cache on each context switch)
  - two distinct virtual addresses can refer to same physical address, which is called aliasing (aliasing typically resolved, in case of direct-mapped cache, by restricting address mapping such that aliases map to same cache set)

# Virtually-Indexed Physically-Tagged (VIPT) Caches

- cache accesses require tag and index
- in case of virtually-indexed physically-tagged cache, index derived from virtual address and tag derived from physical address
- virtually-indexed physically-tagged cache tries to achieve simplicity of physical cache with speed closer to that of virtual cache
- recall that page offset is unaffected by address translation
- use page offset part of virtual address (which is unaffected by address translation) to determine index for cache (i.e., select set in cache)
- doing this allow us to overlap reading of tags and performing address translation
- this approach faster, but imposes some restrictions on cache parameters
- in particular, number of sets in cache cannot exceed number of cache blocks per page (without additional complications)
- L1 cache often virtually indexed and physically tagged

# VIPT Cache Example



- 48-bit virtual address, 36-bit physical address
- 64-byte cache block
- 4 KB page size
- L1 data cache: 32 KB, 8-way set associative, 64 entries per set



- **hit rate**: fraction of memory access that hit in cache
- **miss rate**: fraction of memory access that miss in cache ( $1 - \text{hit rate}$ )
- **miss penalty**: time to replace block from lower level in memory hierarchy to cache
- **hit time**: time to access cache memory (including tag comparison)
- **average memory access time (AMAT)**:
  - $\text{AMAT} = \text{hit time} + \text{miss rate} \cdot \text{miss penalty}$

# Intel Core i7

- 64-bit processor, x86-64 instruction set
- 36-bit physical addresses and 48-bit virtual addresses
- three-level cache hierarchy; all levels use 64-byte block size; two-level TLB
- L1 cache:
  - I cache: 32 KB 4-way set associative; D cache: 32 KB 8-way set associative; per core, pseudo LRU replacement, virtually indexed and physically tagged
- L2 cache:
  - 256 KB, 8-way set associative, per core, pseudo-LRU replacement, physically indexed (and tagged)
- L3 cache:
  - 2 MB per core, 16-way set associative, pseudo-LRU replacement (with ordered selection algorithm), physically indexed (and tagged)
- first-level TLB:
  - I TLB: 128 entries, 4-way set associative, pseudo-LRU replacement; D TLB: 64 entries, 4-way set associative, pseudo-LRU replacement
- second-level TLB:
  - 512 entries, 4-way set associative, pseudo-LRU replacement, 4 KB page size

- 32-bit processor, ARM v7 instruction set
- 32-bit physical and virtual addresses
- two-level cache hierarchy; both levels use 64-byte block size
- L1 cache:
  - separate I and D caches; 16 KB or 32 KB 4-way set associative using way prediction and random replacement; virtually indexed and physically tagged
- optional L2 cache:
  - 8-way set associative, 128 KB to 1 MB; physically indexed and physically tagged
- TLB:
  - pair of TLBs (I and D), each of which fully associative with 32 entries and variable page size (4 KB, 16 KB, 64 KB, 1 MB, 16 MB); replacements done by round robin
  - TLB misses handled in hardware, which walks page table structure in memory

## Section 6.6.2

### **Cache-Efficient Algorithms**

- to effectively exploit cache, need to maximize locality
- various transformations can be applied to code in order to increase locality
- algorithm may be either cache aware or cache oblivious
- **cache aware**: has knowledge of memory hierarchy such as cache parameters (e.g., cache size, cache block size)
- **cache oblivious**: has no knowledge of particulars of memory hierarchy

# Code Transformations to Improve Cache Efficiency

- numerous transformations can be applied to code in order to improve spatial and/or temporal locality
- **merging arrays**: improve spatial locality by using array of aggregate type instead of multiple arrays
- **loop interchange**: change nesting of loops to access data in order stored in memory
- **loop fusion**: combine two or more independent loops that have same looping and some variables overlap
- **blocking**: improve temporal locality by accessing blocks of data repeatedly

# Array Merging Example

- before array merging:

```
constexpr int num_points = 32'768;  
static double x[num_points]; // x coordinates  
static double y[num_points]; // y coordinates  
static double z[num_points]; // z coordinates
```

- after array merging:

```
constexpr int num_points = 32'768;  
struct Point {  
    double x; // x coordinate  
    double y; // y coordinate  
    double z; // z coordinate  
};  
static Point p[num_points];
```

- x, y, and z coordinates of particular point likely to be accessed together
- use array of aggregate type instead of three separate arrays in order to improve spatial locality and reduce potential conflicts

# Loop Interchange Example

- before loop interchange:

```
constexpr int n = 2'048;
static double a[n][n];
// ...
for (int j = 0; j < n; ++j) {
    for (int i = 0; i < n; ++i) {
        a[i][j] *= 2.0;
    }
}
```

- after loop interchange:

```
constexpr int n = 2'048;
static double a[n][n];
// ...
for (int i = 0; i < n; ++i) {
    for (int j = 0; j < n; ++j) {
        a[i][j] *= 2.0;
    }
}
```

- interchange loops so that array elements accessed consecutively instead of with large stride in order to improve locality and reduce potential conflicts



# Loop Fusion Example

- before loop fusion:

```
constexpr int n = 2'048;
static float a[n][n], b[n][n], c[n][n], d[n][n];
// ...
for (int i = 0; i < n; ++i) {
    for (int j = 0; j < n; ++j)
        {a[i][j] = b[i][j] * c[i][j];}
}
for (int i = 0; i < n; ++i) {
    for (int j = 0; j < n; ++j)
        {d[i][j] = a[i][j] + c[i][j];}
}
```

- after loop fusion:

```
constexpr int n = 2'048;
static float a[n][n], b[n][n], c[n][n], d[n][n];
// ...
for (int i = 0; i < n; ++i) {
    for (int j = 0; j < n; ++j) {
        a[i][j] = b[i][j] * c[i][j];
        d[i][j] = a[i][j] + c[i][j];
    }
}
```

- merge loops in order to improve temporal locality (due to reuse of `a[i][j]` and `c[i][j]` in each innermost loop iteration)

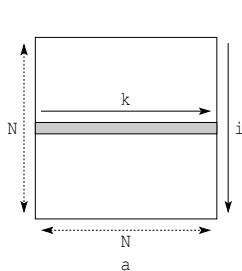
# Blocking Example

- before blocking:

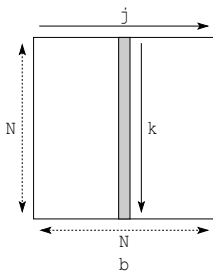
```
1  // compute  $c := c + a b$ , where  $a, b, c$  are  $N$ -by- $N$ 
2  // matrices
3  template <class T, int N>
4  void naive_multiply(const T (&a)[N][N], const T (&b)[N][N],
5    T (&c)[N][N]) {
6      for (int i = 0; i < N; ++i) {
7          for (int j = 0; j < N; ++j) {
8              double s = 0;
9              for (int k = 0; k < N; ++k)
10                 {s += a[i][k] * b[k][j];}
11                 c[i][j] += s;
12             }
13         }
14     }
```

- want to partition computation into blocks of size  $B \times B$ , where  $B$  chosen so that each block fits in cache

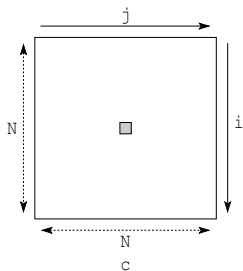
# Blocking Example (Continued 0.5)



use each row of  $a$   $N$  times  
(once for each column of  $b$ )



for each row of  $a$ ,  
use each of  $N$  columns of  $b$   
in succession



update elements in  $c$  in  
left-to-right top-to-bottom order

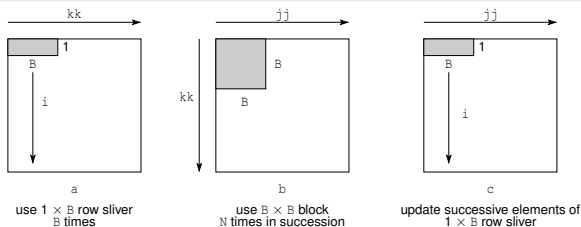
# Blocking Example (Continued 1)

- after blocking (with blocking factor B):

```
1  // compute  $c := c + a b$ , where  $a$ ,  $b$ ,  $c$  are  $N$ -by- $N$ 
2  // matrices
3  template <int B, class T, int N>
4  void blocked_multiply(const T (&a)[N][N], const T (&b)[N][N],
5      T (&c)[N][N]) {
6      for (int kk = 0; kk < N; kk += B) {
7          for (int jj = 0; jj < N; jj += B) {
8              for (int i = 0; i < N; ++i) {
9                  for (int j = jj; j < std::min(jj + B, N); ++j) {
10                     double s = 0;
11                     for (int k = kk; k < std::min(kk + B, N); ++k)
12                         {s += a[i][k] * b[k][j];}
13                     c[i][j] += s;
14                 }
15             }
16         }
17     }
18 }
```

- performing computation using blocking significantly improves locality
- potentially many fewer cache misses
- unfortunately, code using blocking much less readable (i.e., more difficult to understand)

## Blocking Example (Continued 2)



- key idea is that block of  $b$  brought into cache, fully utilized, then discarded
- innermost loop pair (i.e., for  $j$  and  $k$ ) multiplies  $1 \times B$  sliver of  $a$  by  $B \times B$  block of  $b$  and accumulates result in  $1 \times B$  sliver of  $c$
- references to  $a$  have: good spatial locality, since elements accessed consecutively in loop for  $k$ ; and good temporal locality, since each sliver accessed  $B$  times in succession in loop for  $j$
- references to  $b$  have good temporal locality, since entire block accessed  $N$  times in succession in loop for  $i$
- references to  $c$  have good spatial locality since each element of sliver written in succession in loop for  $j$

# Cache-Aware Versus Cache-Oblivious Algorithms

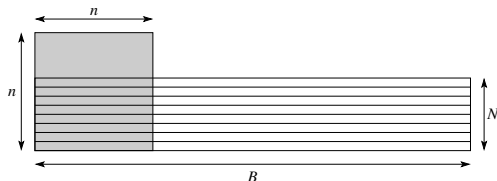
- cache-aware approaches require knowledge of memory hierarchy and caches (e.g., cache size and cache block size for each level of cache) in order to choose key tuning parameters
- often, such knowledge of memory hierarchy difficult to obtain in reliable manner
- furthermore, effective cache size may differ significantly from true cache size, if multiple threads using cache (which reduces effective cache size)
- if tuning parameters not well chosen, performance can potentially be very poor
- in contrast, cache oblivious approaches:
  - require no knowledge of memory hierarchy and caches
  - require no “magical” tuning parameters
  - effectively autotune
  - handle multilevel caches automatically
  - well accommodate multiprogrammed environments

## Section 6.6.3

# **Cache-Oblivious Algorithms**

# Tall Caches

- suppose that cache has size  $M$  with block size  $B$  and  $N = M/B$  entries
- cache is said to be **tall** if  $N > c'B$  for some sufficiently large constant  $c' \geq 1$ ; otherwise, said to be **short**
- essentially, tall property ensures that  $N$  exceeds  $B$  by large enough margin that any (possibly non-contiguous) data of size  $D$  is guaranteed to fit in cache if  $D \leq M$
- that is, if size of some data does not exceed cache size, then that data must fit in cache
- this is not the case for short caches
- for example,  $n \times n$  block of elements inside larger array stored in **row-major order** with  $n^2 < M$  will not necessarily fit in cache if cache is short





# Idealized Cache Model

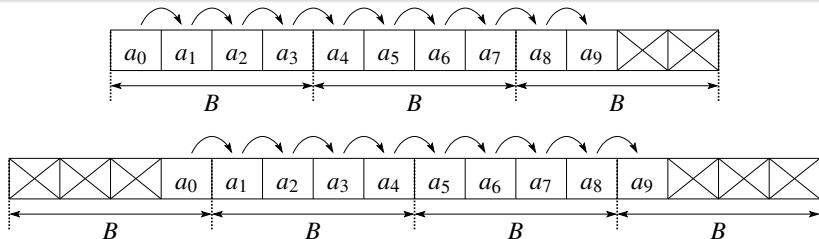
- idealized cache model employs two-level memory hierarchy (i.e., cache and main memory)
- assumptions of idealized cache model:
  - fully associative
  - optimal replacement policy (i.e., evict cache block whose next access will be furthest in future)
  - tall cache
- idealized model only crude approximation to real-world caches
- real-world caches usually not fully associative and never employ optimal replacement policy (which requires noncausal hardware)
- real-world caches, however, usually tend to be tall

# Remarks on Assumption of Optimal-Replacement Policy

- reasonable to question validity of assumption of optimal-replacement policy in idealized cache model
- Sleator and Tarjan (1985) have shown that amortized cost of LRU replacement policy within constant factor of optimal replacement policy
- suppose that algorithm that incurs  $Q$  cache misses on ideal cache of size  $M$
- then, on fully-associative cache of size  $2M$  that uses LRU replacement policy, at most  $2Q$  cache misses
- therefore, *to within constant factor, LRU replacement as good as optimal replacement (for fully-associative cache)*
- implication is that for asymptotic analysis can assume optimal or LRU replacement as convenient
- in this sense, assumption of optimal-replacement policy is quite reasonable

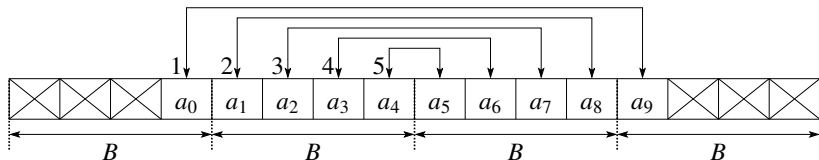
- when analyzing algorithms with respect to idealized cache model typically we are interested in
  - amount of work  $W$  (ordinary running time)
  - number of cache misses  $Q$
- cache oblivious algorithms often based on divide and conquer

# Scanning



- cache block holds  $B$  array elements
- consider scanning  $N$  elements of array in order (e.g., to compute sum or minimum/maximum)
- requires  $\Theta(N)$  work (assuming work per element is  $O(1)$ )
- scanning  $N$  elements stored contiguously in memory incurs either  $\lceil N/B \rceil + 1$  or  $\lceil N/B \rceil$  cache misses (i.e.,  $\Theta(N/B)$  cache misses)
- may require one more than  $\lceil N/B \rceil$  cache misses due to arbitrary alignment
- cache oblivious and optimal (i.e., incurs only minimum number of cache misses)

# Array Reversal



- cache block holds  $B$  array elements
- consider reversing elements of  $N$ -element array  $a$
- use two parallel scans, one from each end of array, and each step swaps two corresponding elements
- for  $i$  in  $0, 1, \dots, \lfloor N/2 \rfloor - 1$ , swap  $a[i]$  and  $a[N - 1 - i]$
- requires  $\Theta(N)$  work (i.e.,  $\lfloor N/2 \rfloor$  swap operations)
- assuming at least two blocks fit in cache, incurs either  $\lceil N/B \rceil + 1$  or  $\lceil N/B \rceil$  cache misses (i.e.,  $\Theta(N/B)$  cache misses)
- cache oblivious and optimal

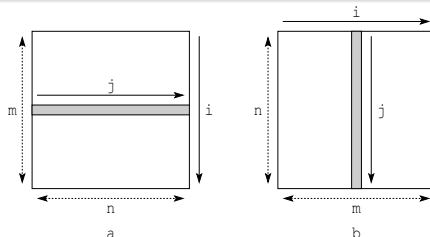
# Naive Matrix Transposition

- naive matrix transpose code has following form:

```
1  template <class T, int m, int n>
2  void transpose(const T (&a)[m][n], T (&b)[n][m]) {
3      for (int i = 0; i < m; ++i) {
4          for (int j = 0; j < n; ++j) {
5              b[j][i] = a[i][j];
6          }
7      }
8  }
```

- arrays stored in row-major order
- although data in `a` being accessed sequentially, data in `b` being accessed with large stride
- many unnecessary cache misses on accesses to `b` if number of rows in `b` sufficiently large

# Naive Matrix Transposition: Performance



- cache block holds  $L$  matrix elements
- requires  $\Theta(mn)$  work (which is optimal)
- in innermost loop, accesses to  $b$  use potentially large stride
- strided access to  $b$  can potentially result in large number of cache misses
- if all blocks for entire column of  $b$  cannot be kept resident in cache simultaneously, every access to  $b$  will miss
- in this case, at most  $\lceil mn/L \rceil + 1 + mn$  cache misses
- any matrix-transpose algorithm must access all  $mn$  elements of  $a$  and all  $mn$  elements of  $b$ , which incurs at most  $2(\lceil mn/L \rceil + 1)$  cache misses
- naive approach incurs about  $(\frac{L-1}{L})mn$  more cache misses than this

# Cache-Oblivious Matrix Transposition

- consider Rec-Transpose algorithm for matrix transposition from page 4:7 of:

M. Frigo, C. E. Leiserson, H. Prokop, and S. Ramachandran.

[Cache-oblivious algorithms.](#)

*ACM Transactions on Algorithms*, 8(1):4:1–4:22, Jan. 2012.

- given  $m \times n$  matrix  $A$  and  $n \times m$  matrix  $B$ , place  $A^T$  into  $B$
- $A$  and  $B$  assumed to correspond to distinct objects (i.e., not in-place transposition)
- based on divide and conquer strategy
- algorithm halves largest of dimensions  $m$  and  $n$ , and recurs
- two cases to consider (i.e.,  $m$  or  $n$  largest)
- if more than one case applies (i.e.,  $m = n$ ), choose one case arbitrarily



# Cache-Oblivious Matrix Transposition (Continued)

- case 1. if  $n = \max\{m, n\}$  (i.e., number of columns in  $A$  and rows in  $B$  largest):
  - decompose problem as follows:

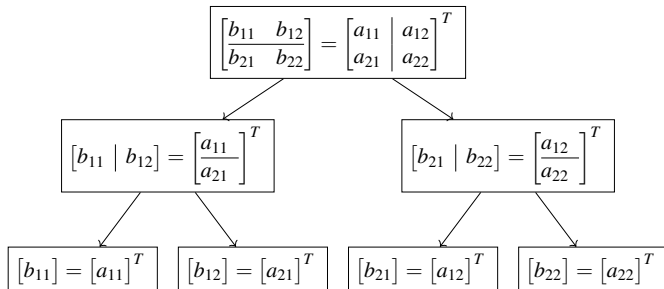
$$\text{let } A = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \text{ and } B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}; \text{ so } \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} A_1^T & A_2^T \end{bmatrix}$$

- recurse to solve  $B_1 = A_1^T$  and  $B_2 = A_2^T$
- case 2. if  $m = \max\{m, n\}$  (i.e., number of rows in  $A$  and columns in  $B$  largest):
  - decompose problem as follows:

$$\text{let } A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \text{ and } B = \begin{bmatrix} B_1 & B_2 \end{bmatrix}; \text{ so } \begin{bmatrix} B_1 & B_2 \end{bmatrix} = \begin{bmatrix} A_1^T \\ A_2^T \end{bmatrix}$$

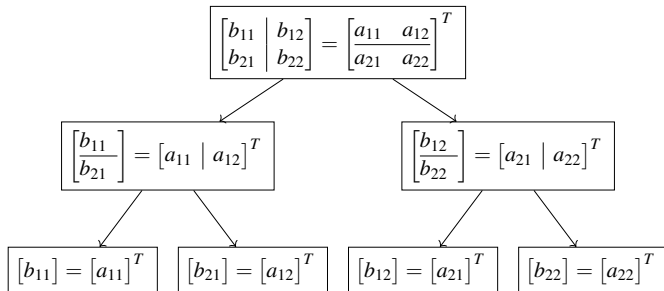
- recurse to solve  $B_1 = A_1^T$  and  $B_2 = A_2^T$
- conceptually, base case for recursion occurs when  $m = n = 1$
- in practice, stop recursion earlier

# Cache-Oblivious Matrix Transposition Example 1A

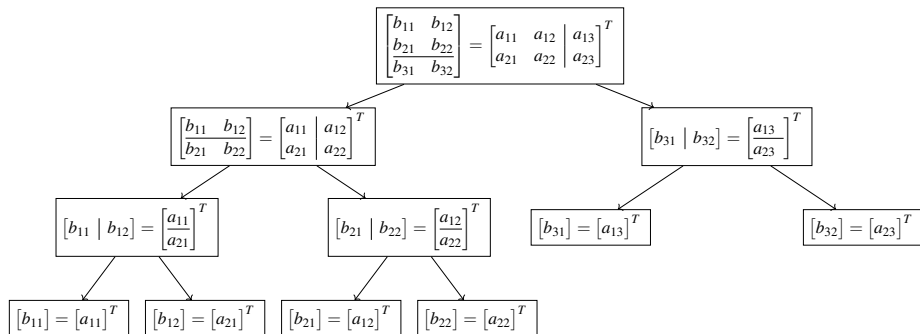


# Cache-Oblivious Matrix Transposition Example 1B

SKIP SLIDE



# Cache-Oblivious Matrix Transposition Example 2



# Cache-Oblivious Matrix Transposition: Performance

- let  $L$  denote number of array elements per cache block
- for  $m \times n$  matrix, cache-oblivious matrix-transposition algorithm:
  - requires  $\Theta(mn)$  work
  - incurs  $\Theta(1 + mn/L)$  cache misses, assuming idealized cache model
- any matrix-transposition algorithm must write to  $mn$  distinct elements, which occupy at least  $\lceil mn/L \rceil = \Omega(1 + mn/L)$  cache lines
- therefore, cache-oblivious algorithm is asymptotically optimal

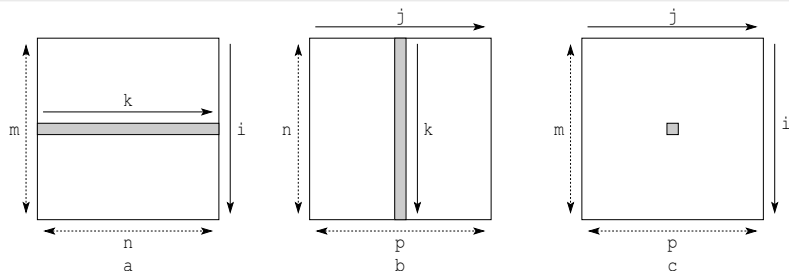
# Naive Matrix Multiplication

- naive matrix multiply code has following form:

```
1  template <class T, int m, int n, int p>
2  void multiply(const T (&a)[m][n], const T (&b)[n][p],
3      T (&c)[m][p]) {
4      for (int i = 0; i < m; ++i) {
5          for (int j = 0; j < p; ++j) {
6              T sum = T(0);
7              for (int k = 0; k < n; ++k) {
8                  sum += a[i][k] * b[k][j];
9              }
10             c[i][j] = sum;
11         }
12     }
13 }
```

- arrays stored in row-major order
- in innermost loop, `b` accessed with potentially large stride, which is problematic
- in second innermost loop, row of `a` is accessed `p` times in succession, which is problematic if row does not fit in cache
- many unnecessary cache misses likely to result in case of larger matrices

# Naive Matrix Multiplication: Performance



- cache block holds  $L$  matrix elements
- innermost loop (in which  $k$  varies) computes dot product of  $i$ th row of  $a$  with  $k$ th column of  $b$  to yield  $(i, j)$ th element of  $c$
- second innermost loop (over  $j$ ) changes column of  $b$  to use in dot product with  $i$ th row of  $a$  (reusing  $i$ th row of  $a$   $p$  times)
- requires  $\Theta(mnp)$  work, which is  $\Theta(n^3)$  in case of square matrices
- assuming that row of  $a$  and column of  $b$  do not fit in cache simultaneously, algorithm incurs  $\Theta(mnp/L + mnp + mp/L)$  cache misses, which is  $\Theta(n^3)$  in case of square matrices

# Cache-Oblivious Matrix Multiplication

- consider Rec-Mult algorithm for matrix multiplication from pages 4:4–4:5 of:

M. Frigo, C. E. Leiserson, H. Prokop, and S. Ramachandran.

[Cache-oblivious algorithms.](#)

*ACM Transactions on Algorithms*, 8(1):4:1–4:22, Jan. 2012.

- given  $m \times n$  matrix  $A$  and  $n \times p$  matrix  $B$ , compute  $m \times p$  matrix  $C$ , where  $C := C + AB$
- if  $C$  initialized to 0, computation  $C := C + AB$  yields  $C = AB$
- based on divide and conquer strategy
- algorithm halves largest of three dimensions  $m$ ,  $n$ , and  $p$ , and recurs
- three cases to consider (i.e.,  $m$ ,  $n$ , or  $p$  largest)
- if more than one of three cases applies (e.g., if  $m = n = p$ ), choose one case arbitrarily



# Cache-Oblivious Matrix Multiplication (Continued 1)

- case 1. if  $m = \max\{m, n, p\}$  (i.e., number of rows in  $A$  and  $C$  largest):
  - decompose problem as follows:

$$\text{let } C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \text{ and } A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}; \quad \text{so } AB = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} B = \begin{bmatrix} A_1 B \\ A_2 B \end{bmatrix}$$

- recurse to compute  $C_1 := C_1 + A_1 B$  and  $C_2 := C_2 + A_2 B$
- case 2. if  $n = \max\{m, n, p\}$  (i.e., number of columns in  $A$  and rows in  $B$  largest):
  - decompose problem as follows:

$$\text{let } A = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \text{ and } B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}; \quad \text{so } AB = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = A_1 B_1 + A_2 B_2$$

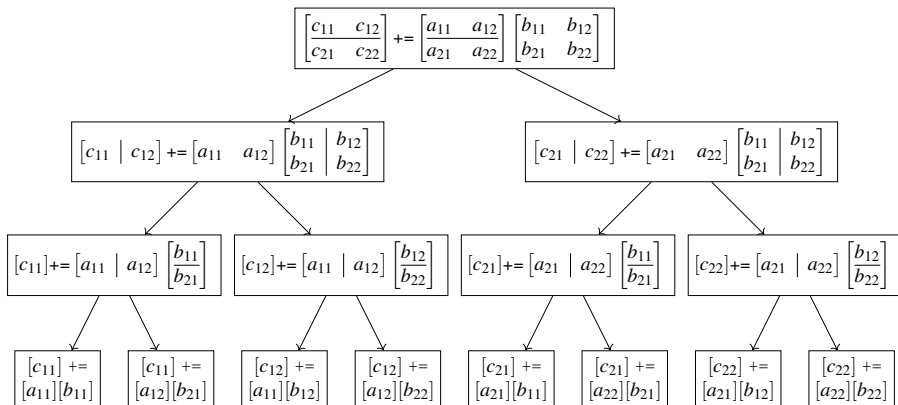
- recurse to compute  $C := C + A_1 B_1$  and then  $C := C + A_2 B_2$
- case 3. if  $p = \max\{m, n, p\}$  (i.e., number of columns in  $B$  and  $C$  largest):
  - decompose problem as follows:

$$\text{let } C = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \text{ and } B = \begin{bmatrix} B_1 & B_2 \end{bmatrix}; \quad \text{so } AB = A \begin{bmatrix} B_1 & B_2 \end{bmatrix} = \begin{bmatrix} AB_1 & AB_2 \end{bmatrix}$$

- recurse to compute  $C_1 := C_1 + AB_1$  and  $C_2 := C_2 + AB_2$

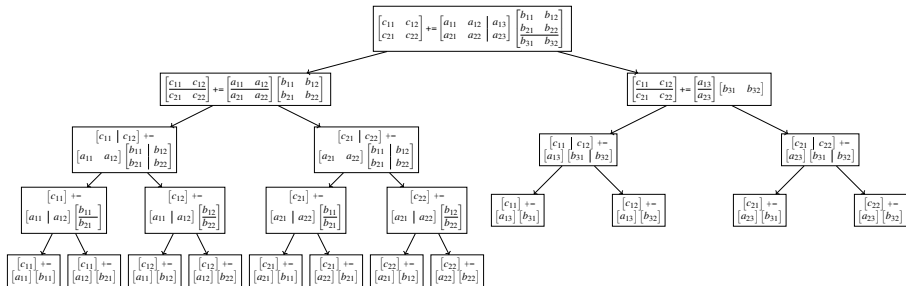
- conceptually, base case for recursion occurs when  $m = n = p = 1$ , in which case two elements multiplied and added into result matrix
- in practice, however, stop recursion earlier

# Cache-Oblivious Matrix Multiplication Example 1



# Cache-Oblivious Matrix Multiplication Example 2

SKIP SLIDE



# Cache-Oblivious Matrix Multiplication: Performance

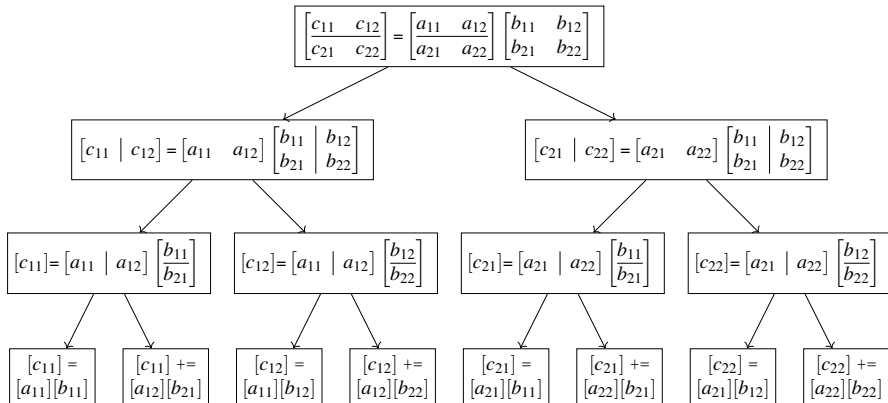
- cache block holds  $L$  matrix elements
- cache size  $M$  (in matrix elements)
- to multiply  $m \times n$  matrix by  $n \times p$  matrix:
  - requires  $\Theta(mnp)$  work
  - incurs  $\Theta\left(m + n + p + \frac{1}{L}(mn + np + mp) + \frac{1}{LM^{1/2}}mnp\right)$  cache misses, assuming idealized cache model
- to multiply two square matrices (i.e.,  $m = n = p$ ):
  - requires  $\Theta(n^3)$  work
  - incurs  $\Theta\left(\frac{1}{LM^{1/2}}n^3\right)$  cache misses, assuming idealized cache model
- Hong and Kung (1981) have shown this to be optimal bound for cache misses for matrix multiplication
- therefore, cache-oblivious algorithm is optimal

# Cache-Oblivious Matrix Multiplication Revisited

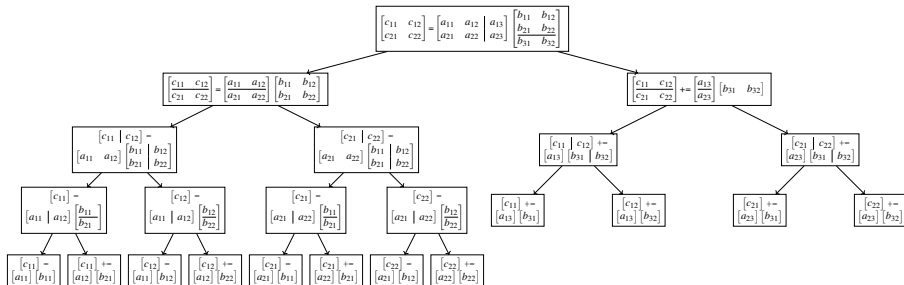
- strictly speaking, Rec-Mult algorithm performs matrix multiply and accumulate
- can zero matrix that holds result to achieve effect of matrix multiply alone
- alternatively, can use additional state during recursion to handle whether accumulation done
- add accumulate flag to recursion
- when accumulate flag is set, result is added to destination; otherwise, result overwrites destination
- always perform call for left child in recursion tree first; then perform call for right child
- at start of recursion, accumulate flag set to false
- for cases 1 and 3, accumulate flag passed unmodified in call for both children in recursion tree
- for case 2, accumulate flag passed unmodified in call for left child in recursion tree and set to true in call for right child in recursion tree

# Cache-Oblivious Matrix Multiplication Revisited Example 1

SKIP SLIDE



# Cache-Oblivious Matrix Multiplication Revisited Example 2





# Strassen's Algorithm for Matrix Multiplication

SKIP SLIDE

- given two  $n \times n$  matrices  $A$  and  $B$  where  $n$  is power of two, compute  $C = AB$
- approach based on divide and conquer
- partition  $A$ ,  $B$ , and  $C$  into equally sized block matrices:

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}, \quad B = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}, \quad C = \begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix}$$

- define (using only 7 matrix multiplications instead of 8):

$$\begin{aligned} M_1 &= (A_{1,1} + A_{2,2})(B_{1,1} + B_{2,2}), & M_2 &= (A_{2,1} + A_{2,2})B_{1,1}, \\ M_3 &= A_{1,1}(B_{1,2} - B_{2,2}), & M_4 &= A_{2,2}(B_{2,1} - B_{1,1}), \\ M_5 &= (A_{1,1} + A_{1,2})B_{2,2}, & M_6 &= (A_{2,1} - A_{1,1})(B_{1,1} + B_{1,2}), \\ M_7 &= (A_{1,2} - A_{2,2})(B_{2,1} + B_{2,2}) \end{aligned}$$

- can compute  $C$  as follows:

$$\begin{aligned} C_{1,1} &= M_1 + M_4 - M_5 + M_7, & C_{1,2} &= M_3 + M_5, \\ C_{2,1} &= M_2 + M_4, & C_{2,2} &= M_1 - M_2 + M_3 + M_6 \end{aligned}$$

- Strassen's matrix multiplication algorithm optimal in cache-oblivious sense

# Discrete Fourier Transform (DFT)

- discrete Fourier transform (DFT) of vector  $x$  of  $n$  complex numbers is vector  $y$  (of  $n$  complex numbers) given by

$$y(i) = \sum_{j=0}^{n-1} x(j) \omega_n^{-ij} \quad \text{where } \omega_n = e^{2\pi\sqrt{-1}/n}$$

- for any factorization  $n = n_1 n_2$  of  $n$ , we have

$$y(i_1 + i_2 n_1) = \sum_{j_2=0}^{n_2-1} \left[ \left( \sum_{j_1=0}^{n_1-1} x(j_1 n_2 + j_2) \omega_{n_1}^{-i_1 j_1} \right) \omega_n^{-i_1 j_2} \right] \omega_{n_2}^{-i_2 j_2}$$

- in preceding equation, inner and outer summations are DFTs
- operationally, computation specified in above equation can be performed by:
  - 1 computing  $n_2$  DFTs of size  $n_1$  (i.e., inner summation)
  - 2 multiplying result by factors  $\omega_n^{-i_1 j_2}$  (called twiddle factors)
  - 3 computing  $n_1$  DFTs of size  $n_2$  (i.e., outer summation)

# Cache-Oblivious Fast Fourier Transform (FFT)

- “six-step” variant of Cooley-Tukey FFT algorithm
- want to compute (one-dimensional) FFT of  $n$  element array  $x$ , where  $n$  is composite and preferably power of two
- FFT is computed in place (i.e., output in  $x$ )
- algorithm consists of following steps (in order):
  - 1 factor  $n$  as  $n = n_1 n_2$ , where  $n_1$  is as close to  $\sqrt{n}$  as possible
  - 2 treat input vector  $x$  as row-major  $n_1 \times n_2$  matrix  $A$ , and use cache-oblivious transpose algorithm to transpose  $A$  in place (by writing transpose of  $A$  to auxiliary array  $B$  and then copying  $B$  back to  $A$ )
  - 3 for each of  $n_2$  rows of  $A$ , replace row with its  $n_1$ -point DFT, where each DFT is computed recursively
  - 4 multiply  $A$  by twiddle factors
  - 5 transpose  $A$  in place (so that inputs to next stage placed in contiguous locations)
  - 6 for each of  $n_1$  rows of  $A$ , replace row with its  $n_2$ -point DFT, where each DFT is computed recursively
  - 7 transpose  $A$  in place to yield output array  $x$  with elements in correct order

## Example: Four-Point DFT

- Note: The DFT of  $[x_0 \ x_1]$  is easily shown to be  $[x_0 + x_1 \ x_0 - x_1]$ .
- Consider computing the DFT of the following sequence of length  $n = 4$ :

$$[x_0 \ x_1 \ x_2 \ x_3].$$

- Factor  $n$  as  $n_1 n_2$ , where  $n_1 = n_2 = 2$ .
- Treat the one-dimensional array of size 4 as the following  $2 \times 2$  array stored in row-major order:

$$\begin{bmatrix} x_0 & x_1 \\ x_2 & x_3 \end{bmatrix}.$$

- Transpose the matrix to obtain:

$$\begin{bmatrix} x_0 & x_2 \\ x_1 & x_3 \end{bmatrix}.$$

- Replace each row of the matrix by its two-point DFT to yield:

$$\begin{bmatrix} x_0 + x_2 & x_0 - x_2 \\ x_1 + x_3 & x_1 - x_3 \end{bmatrix}.$$

## Example: Four-Point DFT (Continued 1)

- Apply the twiddle factors to the matrix to obtain (where  $\omega_4 = e^{2\pi\sqrt{-1}/4} = \sqrt{-1}$ ):

$$\begin{bmatrix} \omega_4^{-(0.0)}(x_0 + x_2) & \omega_4^{-(1.0)}(x_0 - x_2) \\ \omega_4^{-(0.1)}(x_1 + x_3) & \omega_4^{-(1.1)}(x_1 - x_3) \end{bmatrix} = \begin{bmatrix} x_0 + x_2 & x_0 - x_2 \\ x_1 + x_3 & \sqrt{-1}(-x_1 + x_3) \end{bmatrix}.$$

- Transpose the matrix to yield:

$$\begin{bmatrix} x_0 + x_2 & x_1 + x_3 \\ x_0 - x_2 & \sqrt{-1}(-x_1 + x_3) \end{bmatrix}.$$

- Replace each row of the matrix by its two-point DFT to yield:

$$\begin{bmatrix} x_0 + x_1 + x_2 + x_3 & x_0 - x_1 + x_2 - x_3 \\ x_0 - x_2 + \sqrt{-1}(-x_1 + x_3) & x_0 - x_2 + \sqrt{-1}(x_1 - x_3) \end{bmatrix}.$$

- Transpose the matrix to obtain:

$$\begin{bmatrix} x_0 + x_1 + x_2 + x_3 & x_0 - x_2 + \sqrt{-1}(-x_1 + x_3) \\ x_0 - x_1 + x_2 - x_3 & x_0 - x_2 + \sqrt{-1}(x_1 - x_3) \end{bmatrix}.$$

## Example: Four-Point DFT (Continued 2)

- Treat the  $2 \times 2$  matrix stored in row-major order as a one-dimensional array to yield:

$$\begin{bmatrix} x_0 + x_1 + x_2 + x_3 \\ x_0 - x_2 + \sqrt{-1}(-x_1 + x_3) \\ x_0 - x_1 + x_2 - x_3 \\ x_0 - x_2 + \sqrt{-1}(x_1 - x_3) \end{bmatrix}^T.$$

# Cache-Oblivious FFT: Performance

- can be proven by induction that algorithm requires  $O(n \log_2 n)$  work
- cache block holds  $L$  elements of array
- $Z$  cache size in units of array element size
- can be shown that algorithm incurs  $O(1 + (n/L)(1 + \log_Z n))$  cache misses
- preceding cache miss result asymptotically optimal for Cooley-Tukey algorithm, matching lower bound by Hong and Kung when  $n$  is exact power of two

## Section 6.6.4

### **References**



- 1 H. Prokop. [Cache-oblivious algorithms](#).  
Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, June 1999.
- 2 E. D. Demaine. [Cache-oblivious algorithms and data structures](#).  
In *Lecture Notes from the EEF Summer School on Massive Data Sets*, BRICS, University of Aarhus, Denmark, June 2002.
- 3 M. Frigo, C. E. Leiserson, H. Prokop, and S. Ramachandran.  
[Cache-oblivious algorithms](#).  
*ACM Transactions on Algorithms*, 8(1):4:1–4:22, Jan. 2012.
- 4 J. Hennessy and D. A. Patterson. [Computer Architecture — A Quantitative Approach](#).  
Morgan-Kaufmann, San Francisco, CA, USA, 4th edition, 2007.

- 5 D. H. Bailey. [FFTs in external or hierarchical memory.](#)  
*Journal of Supercomputing*, 4:23–35, 1990.
- 6 J. S. Vitter and E. A. M. Shriver. [Algorithms for parallel memory, II: Hierarchical multilevel memories.](#)  
*Algorithmica*, 12:148–169, 1994.
- 7 P. Kumar. [Cache-oblivious algorithms.](#)  
In *Algorithms for Memory Hierarchies*, pages 192–212. Springer Verlag, 2003.
- 8 V. Strassen. [Gaussian elimination is not optimal.](#)  
*Numerische Mathematik*, 13(4):354–356, Aug. 1969.
- 9 D. D. Sleator and R. E. Tarjan. [Amortized efficiency of list update and paging rules.](#)  
*Communications of the ACM*, 28(2):202–208, Feb. 1985.

- 10 J.-W. Hong and H. T. Kung. [I/O complexity: the red-blue pebbling game.](#)  
In *Proc. of ACM Symposium on Theory of Computing*, pages 326–333,  
Milwaukee, WI, USA, 1981.
- 11 B. Jacob and T. Mudge. [Virtual memory in contemporary microprocessors.](#)  
*IEEE Micro*, 18(4):60–75, July 1998.
- 12 S. Chatterjee and S. Sen. [Cache-efficient matrix transposition.](#)  
In *Proc. of IEEE International Symposium on High-Performance Computer Architecture*, Toulouse, France, 2000.
- 13 J. W. Cooley and J. W. Tukey. [An algorithm for the machine calculation of complex Fourier series.](#)  
*Mathematics of Computation*, 19:297–301, 1965.

## Section 6.7

# Vectorization

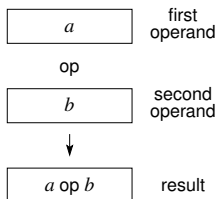
## Section 6.7.1

# **Vector Processing**

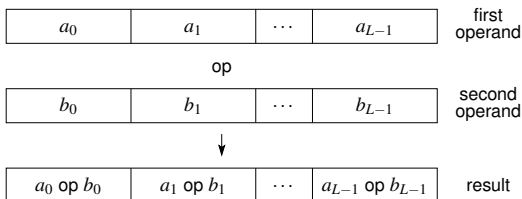
- vector processor has instruction set that can operate on one-dimensional arrays of data called vectors
- vector processing has its roots in early supercomputers
- approach has been refined significantly over the years
- attempts to exploit data-level parallelism
- most modern processors provide some level of vector processing functionality

# Scalar Versus Vector Instructions

Scalar Operation  
With Two Operands



Vector Operation  
With Two Operands



- each operand of scalar instruction is single value
- each operand of vector instruction is set of  $L$  values known as **vector**
- $L$  called **vector length**
- same operation applied to each of  $L$  elements of vector
- operation might, for example, be: load/store, arithmetic operation, logical operation, comparison, conversion operation, or shuffle operation

# Vector-Memory and Vector-Register Architectures

- two basic approaches to vector processing:

- 1 vector-memory architecture
- 2 vector-register architecture

- vector-memory architecture:

- for all vector operations, operands fetched from main memory and results written back to main memory
- includes early vector machines through mid 1980s
- no longer used much (if at all) in modern processors
- large startup time for vector operations

- vector-register architecture:

- for all vector operations except loads and stores, operands read from and written to vector registers
- used by most modern processors that support vector operations



# Vector-Register Architectures

- vector register is collection of  $N$  elements of same type, where each element is  $M$  bits in size
- $N$  called vector length
- vector register size  $NM$  typically 128 to 512
- advantages of vector processing:
  - potential speedup by factor of  $N$
  - often more energy efficient relative to other approaches for increasing performance (such as wider superscalar or higher clock rate)
  - potentially smaller code size, since single instruction can perform multiple operations

# Vector Extensions

- modern high-performance CPU architectures have specialized instructions to exploit parallelism in loops
- commonly referred as single-instruction multiple-data (SIMD) extensions
- operate on multiple elements of wide vector register simultaneously
- reduces runtime trip count of loop by vectorization factor
- requires sophisticated analysis and heuristics in order to make good decisions about vectorization safety and profitability
- widen each operation in loop from scalar type to vector type
- applies same operation in parallel to number of data items packed into large register (e.g., 64, 128, 256, 512 bits)
- particularly useful for algorithms with high degree of data-level parallelism, such as those often found in multimedia systems, graphics, and image/video/audio processing

- Streaming SIMD Extensions (SSE) is family of vector extensions to Intel x86/x86-64 instruction set architecture (namely, SSE, SSE2, SSE3, SSSE3, SSE4.1, and SSE 4.2)
- collectively, SSE family added:
  - in case of x86: *8 128-bit vector registers*, known as XMM0 to XMM7
  - in case of x86-64: *16 128-bit vector registers*, known as XMM0 through XMM15
- each vector register can be used to hold:
  - 16 8-bit bytes
  - 8 16-bit integers
  - 4 32-bit integers
  - 2 64-bit integers
  - 4 32-bit single-precision floating-point numbers
  - 2 64-bit double-precision floating-point numbers

# Intel x86/x86-64 Advanced Vector Extensions (AVX)

- Advanced Vector Extensions (AVX) is family of vector extensions to Intel x86/x86-64 instruction set architecture (namely, AVX, AVX2, and AVX-512) that builds upon SSE
- AVX extends 16 vector registers of SSE from 128 to *256 bits*
- renames vector registers as YMM0 to YMM7 for x86 and YMM0 to YMM15 for x86-64
- each 256-bit vector register can be used to hold:
  - 32 8-bit bytes
  - 16 16-bit integers
  - 8 32-bit integers
  - 4 64-bit integers
  - 8 32-bit single-precision floating-point numbers
  - 4 64-bit double-precision floating-point numbers
  - 8 32-bit single-precision floating-point numbers
  - 4 64-bit double-precision floating-point numbers
- AVX-512 extends vector registers to *512 bits*

- NEON is vector extension to ARM Cortex-A series and Cortex-R52 processors
- *16 128-bit vector registers*
- NEON instructions perform same operations in all lanes of vectors
- vector registers can hold:
  - 16 8-bit character
  - 8 16-bit integer
  - 4 32-bit integer
  - 2 64-bit integer
  - 8 16-bit floating-point (only in Armv8.2-A)
  - 4 32-bit floating-point
  - 2 64-bit floating-point (only in Armv8-A/R)

# Checking for Processor Vector Support on Linux

- on Linux systems, information on processor can be found in `/proc/cpuinfo`
- level of processor support for vector operations can be determined by checking for various processor flags/features in this file
- on Intel x86/x86-64 systems, look for flags/features:
  - `mmx`, `sse`, `sse2`, `ssse3`, `sse4_1`, `sse4_2`, `avx`, `avx2`
- on ARM systems, look for flags/features:
  - `neon`

## Section 6.7.2

### **Code Vectorization**

- consider loop in function:

```
void axpy(float a, float* x, float y, int n) {  
    for (int i = 0; i < n; ++i) {  
        x[i] = a * x[i] + y;  
    }  
}
```

- loop vectorization: scalar computations in body of above loop could be grouped to allow use of vector operations
- consider code in basic block:

```
a = b + c * d;  
e = f + g * h;  
i = j + k * l;  
m = n + o * p;
```

- basic-block vectorization: four statements in preceding code follow similar pattern and could be grouped together to allow vector operations to be used



# Conceptualizing Loop Vectorization

- can think of loop vectorization in terms of loop unrolling
- consider following loop where, for simplicity, we assume  $n$  multiple of 4:  

```
for (int i = 0; i < n; ++i) {c[i] = a[i] + b[i];}
```

- can partially unroll loop to obtain following, where each iteration of new loop corresponds to 4 iterations of original loop:

```
for (int i = 0; i < n; i += 4) {  
    c[i + 0] = a[i + 0] + b[i + 0]; // iteration i  
    c[i + 1] = a[i + 1] + b[i + 1]; // iteration i + 1  
    c[i + 2] = a[i + 2] + b[i + 2]; // iteration i + 2  
    c[i + 3] = a[i + 3] + b[i + 3]; // iteration i + 3  
}
```

- code in body of new loop can be mapped to vector operations of length 4 on vector registers v0, v1, and v2:
  - 1 load a[i] to a[i + 3] into v0
  - 2 load b[i] to b[i + 3] into v1
  - 3 add v0 and v1, writing result into v2
  - 4 store v2 into c[i] to c[i + 3]
- using non-standard C++ syntax, vectorized loop can be expressed as:

```
for (int i = 0; i < n; i += 4)  
    {c[i : i + 3] = a[i : i + 3] + b[i : i + 3];}
```

# Approaches to Vectorization

- several approaches to vectorization can be taken:

- 1 auto-vectorization

- compiler automatically vectorizes code when deemed both safe and profitable

- 2 auto-vectorization with compiler hints

- annotations added to source code to guide auto-vectorization

- 3 explicit directives

- special directives added to source code to exercise control over vectorization (e.g., OpenMP, Cilk Plus)

- 4 computation using vector data types

- use special vector types provided by compiler (e.g., `__m128` for GCC on x86, `int8x8_t` for GCC on ARM, `__m128` for MSVC on x86)

- 5 compiler intrinsics

- use special low-level functions provided by compiler (e.g., `_mm_add_epi16` for GCC on x86, `vaddq_s16` for GCC on ARM, `_arm_sadd16` for MSVC on ARM)

- 6 inline assembly language

- use SIMD instructions directly by using assembly language

- above approaches listed in order of decreasing ease of use and increasing degree of programmer control

- easiest way to vectorize code is to have compiler do this automatically
- called auto-vectorization
- most compilers have support for auto-vectorization
- advantages of auto-vectorization:
  - easy to use
  - less error prone (no bugs, unless compiler has bug)
  - sometimes compiler may be able to make better judgement as to whether vectorization would be beneficial
- compiler, however, must be very conservative when vectorizing code
- compiler cannot transform code in way that changes its behavior
- unfortunately, compiler often does not have sufficient knowledge of code behavior to perform vectorization well (or at all)

- GCC supports auto-vectorization
- GCC has two vectorizers:
  - 1 loop vectorizer
  - 2 basic-block vectorizer
- both vectorizers enabled by default for optimization level of at least 3 (where optimization level specified with `-O` option)
- GCC fully supports OpenMP 4.5 for C/C++ (but not Fortran) as of GCC 6.1 and fully supports OpenMP 4.0 as of GCC 4.9.1

# GCC Compiler Options Related to Vectorization

- `-ftree-vectorize` and `-fno-tree-vectorize`
  - enable and disable all vectorization, respectively
- `-ftree-loop-vectorize` and `-fno-tree-loop-vectorize`
  - enable and disable loop vectorizer, respectively
- `-ftree-slp-vectorize` and `-fno-tree-slp-vectorize`
  - enable and disable basic-block vectorizer, respectively
- `-fopt-info-vec-optimized`
  - enable remarks that identify places in code where vectorization successfully applied
- `-fopt-info-vec-missed`
  - enable remarks that identify places in code where vectorization could not be applied
- `-march=native`
  - use instructions supported by local CPU
  - to see which flags are enabled with `-march=native`, use:  
`g++ -march=native -Q --help=target`

- `-fopenmp`
  - enable OpenMP support (which requires GOMP library)
- `-fopenmp-simd`
  - enable OpenMP SIMD support (which does not require run-time library)
- `-S`
  - produce assembly language output only (instead of object code)
- `-fverbose-asm`
  - enable generation of more verbose assembly language output (e.g., compiler version and command-line options, source-code lines associated with assembly instructions, hints on which high-level expressions correspond to various assembly instruction operands)

- Clang supports auto-vectorization
- Clang has two vectorizers:
  - 1 loop vectorizer
  - 2 superword-level parallelism (SLP) vectorizer
- loop vectorizer widens instructions in loops to operate on multiple consecutive iterations (i.e., performs loop vectorization)
- SLP vectorizer combines similar independent scalar instructions into vector instructions
- both loop and SLP vectorizers enabled by default for optimization level of at least 1 (where optimization level specified by `-O` option)
- Clang supports all non-offloading features of OpenMP 4.5 as of Clang 3.9

# Clang Compiler Options Related to Vectorization

- `-fvectorize` and `-fno-vectorize`
  - enable and disable loop vectorizer, respectively
- `-fslp-vectorize` and `-no-fslp-vectorize`
  - enable and disable SLP vectorizer, respectively
- `-fslp-vectorize-aggressive`
  - enable more aggressive vectorization in SLP vectorizer
- `-Rpass=loop-vectorize`
  - enable remarks that identify loops that were successfully vectorized
- `-Rpass-missed=loop-vectorize`
  - enable remarks that identify loops that failed vectorization and indicate if vectorization specified
- `-Rpass-analysis=loop-vectorize`
  - enable remarks that identify statements that caused vectorization to fail
- `-fopenmp`
  - enable OpenMP support (which requires OMP library)
- `-S`
  - produce assembly language output only (instead of object code)



# Assessing Quality of Vectorized Code

- to assess quality of vectorized code generated by compiler, often very helpful to view assembly code generated by compiler
- quick inspection of assembly code can often give clear indication as to how well particular part of code was vectorized
- most compilers provide option to generate assembly source as compilation output (instead of object code)
- to assist in locating assembly source corresponding to particular part of C++/C source code (such as loop) can inject comments into assembly code using **asm**
- example:

```
1  float innerprod(float* a, float* b, int n) {  
2      float result = 0.0f;  
3      asm volatile ("# loop start");  
4      for (int i = 0; i < n; ++i) {result += a[i] * b[i];}  
5      asm volatile ("# loop end");  
6      return result;  
7  }
```

# Assessing Quality of Vectorized Code (Continued)

```
1      .file   "inner_product_1.cpp"
2      .text
3      .globl _Z9innerprodPfS_i
4      .type   _Z9innerprodPfS_i, @function
5      _Z9innerprodPfS_i:
6      .LFB0:
7      .cfi_startproc
8      #APP
9      # 3 "inner_product_1.cpp" 1
10     # loop start
11     # 0 "" 2
12     #NO_APP
13     xorl    %eax, %eax
14     vxorps  %xmm0, %xmm0, %xmm0
15     .L3:
16     cmpl    %eax, %edx
17     jle     .L2
18     vmovss  (%rdi,%rax,4), %xmm1
19     vfmadd231ss (%rsi,%rax,4), %xmm1, %xmm0
20     incq    %rax
21     jmp     .L3
22     .L2:
23     #APP
24     # 5 "inner_product_1.cpp" 1
25     # loop end
26     # 0 "" 2
27     #NO_APP
28     ret
29     .cfi_endproc
```

# Auto-Vectorization with Hints

- in order to allow compiler to perform auto-vectorization more effectively, can provide hints to compiler
- place annotations in code to provide compiler with additional information to guide vectorization
- annotations typically provide information that compiler could not reasonably deduce on its own but is important in making decisions regarding vectorization
- approach is relatively easy to use since compiler still does most of work
- must be careful to provide correct information to compiler, however; otherwise, compiler may generate incorrect code

- numerous obstacles to vectorization:
  - data dependencies
  - control-flow dependencies
  - aliasing
  - noncontiguous memory accesses
  - misaligned data
- by eliminating such obstacles, compiler can perform auto-vectorization more effectively

# Data Dependencies and Vectorization

- vectorization changes order of computation compared to sequential case
- changing order of computation may yield different result
- cannot replace sequential loop with vectorized version if this would change result of computation
- need to consider independence of unrolled loop operations, which depends on vectorization factor
- three types of data dependencies:
  - 1 flow dependency (read after write)
  - 2 output dependency (write after write)
  - 3 antidependency (write after read)
- flow and output dependencies are of most concern for vectorization

- **flow dependency** (also called **read-after-write dependency**) is type of data dependency that occurs when variable is written in one iteration of loop and read in subsequent iteration
- **dependency distance** is difference in iteration number in which read and write of variable occur
- example of flow dependency with dependency distance of 1:  

```
for (int i = 1; i < n; ++i)  
    {a[i] = a[i - 1] + 1;}
```
- if dependency distance less than vectorization factor, vectorized loop cannot be guaranteed to yield same result as sequential version

# Flow Dependence Example 1

- consider vectorization of following loop with vectorization factor of 4:

```
for (int i = 1; i < n; ++i)
    {a[i] = a[i - 1] + b[i];}
```

- loop exhibits flow dependence (i.e., read after write) on `a[i-1]` (dependence distance 1)
- loop in partially unrolled form (assuming number of iterations multiple of 4):

```
for (int i = 1; i < n; i += 4) {
    a[i + 0] = a[i - 1] + b[i + 0];
    a[i + 1] = a[i + 0] + b[i + 1];
    a[i + 2] = a[i + 1] + b[i + 2];
    a[i + 3] = a[i + 2] + b[i + 3];
}
```

- loop in vectorized form (assuming number of iterations multiple of 4):

```
for (int i = 1; i < n; i += 4)
    {a[i : i + 3] = a[i - 1 : i + 2] + b[i : i + 3];}
```

- vectorized loop will not always produce same results as sequential loop (due to flow dependence with dependence distance 1)
- therefore, with vectorization factor of 4, loop not legal to vectorize

# Flow Dependence Example 1: Sequential Loop

- suppose that:

```
constexpr int n = 5;  
int a_data[n] = {-1, -2, -3, -4, -5};  
int b_data[n] = {0, 1, 2, 3, 4};  
int* a = a_data;  
int* b = b_data;
```

- sequential loop:

```
for (int i = 1; i < n; ++i) {  
    a[i] = a[i - 1] + b[i]  
}
```

- computation for loop iteration:

i	a[i - 1]	b[i]	a[i]
1	-1	1	0
2	0	2	2
3	2	3	5
4	5	4	9

- upon loop termination, array pointed to by a contains:  
{-1, 0, 2, 5, 9}



# Flow Dependence Example 1: Vectorized Loop

- again, suppose that:

```
constexpr int n = 5;  
int a_data[n] = {-1, -2, -3, -4, -5};  
int b_data[n] = {0, 1, 2, 3, 4};  
int* a = a_data;  
int* b = b_data;
```

- vectorized loop:

```
for (int i = 1; i < n; i += 4) {  
    a[i : i + 3] = a[i - 1 : i + 2] + b[i : i + 3];  
}
```

- computation for loop iteration:

i	a[i - 1 : i + 2]	b[i : i + 3]	a[i : i + 3]
1	{-1, -2, -3, -4}	{1, 2, 3, 4}	{0, 0, 0, 0}

- upon loop termination, array pointed to by a contains:

{-1, 0, 0, 0, 0}

# Flow Dependence Example 2

- consider vectorizing following loop using vectorization factor of 4:

```
for (int i = 5; i < n; ++i)
    {a[i] = a[i - 5] + b[i];}
```

- loop exhibits flow dependence (i.e., read after write) on `a[i-5]` (dependence distance 5)
- loop in partially unrolled form (assuming number of iterations multiple of 4):

```
for (int i = 5; i < n; i += 4) {
    a[i + 0] = a[i - 5] + b[i + 0];
    a[i + 1] = a[i - 4] + b[i + 1];
    a[i + 2] = a[i - 3] + b[i + 2];
    a[i + 3] = a[i - 2] + b[i + 3];
}
```

- loop in vectorized form (assuming number of iterations multiple of 4):

```
for (int i = 5; i < n; i += 4)
    {a[i : i + 3] = a[i - 5 : i - 2] + b[i : i + 3];}
```

- vectorized loop will always yield same result as sequential loop since no flow dependence occurs within single iteration of vectorized loop
- with vectorization factor of 4, loop legal to vectorize

- **output dependency** (also called **write-after-write dependency**) is type of data dependency that occurs when same variable is written in more than one iteration
- example of output dependency:  

```
for (int i = 0; i < n; ++i)  
    {a[i % 2] = b[i] + c[i];}
```
- generally unsafe to perform vectorization of loops with output dependencies

# Control-Flow Dependencies and Vectorization

- control-flow dependencies can lead to different operations for elements in vector

- consider loop in following function:

```
void func(float* a, float* b, int n) {  
    for (int i = 0; i < n; ++i) {  
        a[i] = (a[i] > 1.0) ? a[i] / b[i] : a[i];  
    }  
}
```

- code has control-flow dependence on `a[i]` (code behavior depends on condition `a[i] > 1.0`)
- good compiler might be able to vectorize above function
- when control-flow dependencies become more complex, however, vectorization extremely difficult or impossible to perform
- therefore, control-flow dependencies are best avoided

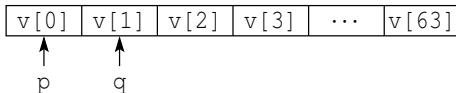
# Aliasing

- when same memory location can be accessed through different names, **aliasing** said to occur
- example of aliasing:

- code:

```
float v[64];  
float* p = &v[0];  
float* q = &v[1];  
// p and q can be used to access same memory  
// e.g., p[1] and q[0] refer to same object
```

- memory layout:



- aliasing often limits ability of compiler to perform optimization
- in effect, aliasing can introduce new data dependencies that would not otherwise exist
- failing to take aliasing into account could lead to illegal optimizations (i.e., optimizations that change code behavior)

# Aliasing and Optimization: An Example

- consider code:

```
1  void func(int* a, int* b, int* c) {  
2      *a = 42;  
3      *b = 0;  
4      *c = *a;  
5  }
```

- at first glance, might seem that code can be optimized to yield:

```
1  void func(int* a, int* b, int* c) {  
2      *a = 42;  
3      *b = 0;  
4      *c = 42;  
5  }
```

- above optimized code is incorrect, since `a` might equal `b`, in which case `*c` should be assigned 0, not 42

# Aliasing and Vectorization: An Example

- consider code:

```
1 void add(float* a, float* b, float* c) {  
2     for (int i = 0; i < 1024; ++i) {  
3         a[i] = b[i] + c[i];  
4     }  
5 }
```

- if only this code visible to compiler, simply vectorizing loop in this function is not legal
- a could be aliased to b or c (i.e., storage pointed to by a, b, and c could overlap); for example, if a equals b + 1, then b[i] same as a[i - 1]
- in this case, sequential and parallel execution of loop would yield different results
- best compiler could do might be to:
  - generate two different versions of code for loop, one without vectorization for aliasing case and one with vectorization for case of no aliasing
  - emit runtime aliasing check that decides which version of code for loop to execute
- this solution less than ideal as it incurs cost of runtime check and results in increased code size

# The `__restrict__` Keyword

- sometimes highly beneficial to have means to indicate to compiler that aliasing cannot occur (so that compiler can better optimize code)
- although not part of C++ standard, some compilers support special keyword for this purpose; for example:
  - GCC and Clang support `__restrict__` keyword
  - MSVC supports `__restrict` keyword
- keyword can be applied to pointer or reference
- during execution of block in which restricted pointer/reference `p` is declared, if some object that is accessible through `p` (directly or indirectly) is modified by any means, then all access to that object in that block must occur through `p` (directly or indirectly)
- important only to use `__restrict__` if certain that no aliasing can occur; otherwise, code behavior likely to be incorrect
- example:

```
void func(int* __restrict__ p, int* __restrict__ q) {  
    // compiler can safely assume that any data modified through p  
    // will only be accessed through p; and similarly for q  
    // thus, data pointed to by p and q cannot overlap  
    // ... (code modifies data pointed to by p and q)  
}
```



# Noncontiguous Memory Accesses

- vector load/store operation typically reads/writes contiguous block of memory (that is appropriately aligned)
- noncontiguous data typically needs multiple instructions to be read/written
- example of code with noncontiguous memory accesses:

```
// in loop, array elements accessed with stride 2
for (int i = 0; i < n; i += 2) {
    c[i] = a[i] + b[i];
}
```

- sometimes noncontiguous memory access problem can be addressed by choosing different layout for data in memory (e.g., struct of arrays instead of array of structs)
- other times, problem may be resolvable by restructuring code to perform computations in different order

- for reasons of performance, vector load and store operations often impose restrictions on data alignment
- typically, target address for vector load or store of  $n$ -byte register needs to be aligned on  $n$ -byte boundary
- for some architectures, such alignment is strict requirement (i.e., code will not work if data misaligned)
- for other architectures, such alignment is not strictly required, but substantial performance penalty may be incurred in case of misaligned data
- for this reason, important to align data appropriately whenever possible
- also, to allow compiler to vectorize in most effective manner possible, important to let compiler know when data is appropriately aligned

# Handling Misaligned Data

- sometimes not possible or practical to avoid misaligned data
- in such cases, can still partially vectorize
- peel first few iterations of loop where data is misaligned and process data using scalar operations
- peel last few iterations (as necessary) where insufficient data to fill vector register and process data using scalar operations
- use vector operations for remainder of iterations
- compared to case of properly aligned data that is multiple of vector size, above approach likely to be slower and have larger code size
- alternatively, could add padding before and/or after data to ensure data with padding is appropriately aligned and multiple of vector length, but this approach often not practical

# Controlling Alignment of Data

- for non-heap allocation, can use **alignas** qualifier to control alignment of object
- for heap allocation, can use `std::aligned_alloc` to allocate memory with particular alignment
- `std::free` can be used to free memory allocated by `std::aligned_alloc`
- example:

```
1  #include <cassert>
2  #include <cstdlib>
3  #include <cstdint>
4
5  int main() {
6      alignas(4096) static char buffer[65536];
7      static_assert(alignof(buffer) == 4096);
8      float* fp = static_cast<float*>(
9          std::aligned_alloc(4096, sizeof(float)));
10     if (!fp) {return 1;}
11     assert(!(reinterpret_cast<intptr_t>(fp) % 4096));
12     std::free(fp);
13 }
```

# Informing Compiler of Data Alignment

- to facilitate more effective vectorization by compiler, important to be able to indicate data alignment in code
- unfortunately, C++ standard does not provide mechanism for doing this
- some compilers (such as GCC and Clang) support intrinsic function called `__builtin_assume_aligned` that can be used to indicate alignment
- `__builtin_assume_aligned` declared as:  

```
void* __builtin_assume_aligned(const void *p, size_t align, ...);
```
- this function simply returns its first argument `p` and allows compiler to assume that returned pointer is at least `align` bytes aligned (when invoked with two arguments)
- example:  

```
void func(float* a, float* b, int n) {  
    // *a and *b can be assumed aligned to 64-byte boundary  
    a = static_cast<float*>(__builtin_assume_aligned(a, 64));  
    b = static_cast<float*>(__builtin_assume_aligned(b, 64));  
    for (int i = 0; i < n; ++i) { /* ... */ }  
}
```
- in case of compilers that do not support `__builtin_assume_aligned`, another approach would need to be found

# Profitability of Vectorization

- vectorization can often provide significant speedup (in some cases linear with vectorization factor), but costs need to be considered
- vector loop bodies can be larger than their scalar forms, as more complex operations may be needed, increasing code size
- vector loop may have increased startup costs to prepare for vectorized execution
- if aliasing is potential problem, require overhead of runtime aliasing check
- vector instructions may take more cycles

# Vectorization Example: Version 1

- source code:

```
1  #include <cstdint>
2
3  template <std::size_t n, class T>
4  void add(const T (&a)[n], T (&b)[n]) {
5      for (int i = 0; i < n; ++i) {
6          b[i] += a[i];
7      }
8  }
```

- since `a` and `b` may be aliased, compiler must generate code that correctly handles aliased case (as well as non-aliased case)
- often, will generate code that tests for aliasing at run time and uses result to decide between code for aliased case or non aliased case
- since compiler does not know alignment of `a` and `b`, must generate code that handles any valid alignment

- source code:

```
1  #include <cstdint>
2
3  template <std::size_t n, class T>
4  void add(const T (&__restrict__ a)[n],
5         T (&__restrict__ b)[n]) {
6      for (int i = 0; i < n; ++i) {
7          b[i] += a[i];
8      }
9  }
```

- compiler can assume no aliasing (due to use of `__restrict__`)
- since compiler does not know alignment of `a` and `b`, must generate code that handles any valid alignment



# Vectorization Example: Version 3

## ■ source code:

```
1  #include <cstdint>
2
3  template <std::size_t n, std::size_t align, class T>
4  void add(const T (& __restrict__ a)[n],
5         T (& __restrict__ b)[n]) {
6      const T* ap = static_cast<const T*>(
7          __builtin_assume_aligned(&a, align));
8      T* bp = static_cast<T*>(
9          __builtin_assume_aligned(&b, align));
10     for (int i = 0; i < n; ++i) {
11         bp[i] += ap[i];
12     }
13 }
```

- compiler can assume no aliasing (due to use of `__restrict__`) and align-byte alignment (due to use of `__builtin_assume_aligned`)
- code generated for vectorized loop in case of

`add<65536, 16 * alignof(float), float>:`

```
12  .L2:
13      vmovaps (%rsi,%rax), %ymm0
14      vaddps (%rdi,%rax), %ymm0, %ymm0
15      vmovaps %ymm0, (%rsi,%rax)
16      addq    $32, %rax
17      cmpq    $262144, %rax
18      jne     .L2
```

# Vectorization Example: Invoking add Function

- when calling add function, must be careful to ensure that assumptions about aliasing and data alignment are not violated

- source code:

```
1  #include <cstdint>
2  #include <iostream>
3  #include <algorithm>
4  #include <numeric>
5  #include "example4_util.hpp"
6
7  int main() {
8      constexpr std::size_t n = 65536;
9      constexpr std::size_t align = 16 * alignof(float);
10     alignas(align) static float a[n];
11     alignas(align) static float b[n];
12     std::iota(&a[0], &a[n], 1);
13     std::fill(&b[0], &b[n], -1);
14     add<n, align>(a, b);
15     for (auto i : b) {std::cout << i << '\n';}
16 }
```

- if add function invoked in manner that violates assumptions about aliasing or data alignment, code not guaranteed to work correctly (e.g., may crash or yield incorrect results)

# Basic Requirements for Vectorizable Loops

- requirements imposed on vectorizable loops vary from compiler to compiler
- so whether given loop can be vectorized will in many cases depend on particular compiler being used
- consult compiler documentation for specific requirements
- some typical examples of requirements imposed on vectorizable loops include:
  - loop is countable (i.e., number of loop iterations known at run time upon entry to loop); this implies, for example, no conditional termination of loop
  - straight-line code in loop body (i.e., no control flow); no switch statements; if statements only allowable when can be implemented as masked assignments
  - must be innermost loop if nested
  - no function calls, except some basic math functions (such as `std::pow`, `std::sqrt`, and `std::sin`) and some inline functions

- OpenMP is industry standard API for parallel computing
- supports C++, C, and Fortran
- OpenMP 4.0 added constructs for expressing SIMD data-level parallelism
- although OpenMP offers large amount of functionality, we only focus on SIMD-related functionality here
- use pragmas to control vectorization
- `simd` pragma allows explicit control of vectorization of for loops
- `declare simd` pragma instructs compiler to generate vectorized version of function (which can be used to vectorize loops containing function calls)

# OpenMP `simd` Pragma

- vectorized loop can be achieved with OpenMP `simd` pragma

- syntax:

```
#pragma omp simd [clause...]  
/* for statement in canonical form */
```

- `simd` pragma must be immediately followed by for loop in canonical form
- optional clauses may be specified to affect behavior of pragma (i.e., `safelen`, `linear`, `aligned`, `private`, `lastprivate`, `reduction`, and `collapse`)
- amongst other things, canonical form of for loop implies:
  - induction variable has integer, pointer, or random-access iterator type
  - limited test and increment/decrement for induction variable
  - iteration count known before execution of loop
- can target inner or outer loops
- loop must be suitable for vectorization (e.g., no data-dependence problems)
- example:

```
#pragma omp simd  
for (int i = 0; i < n; ++i) {c[i] = a[i] + b[i];}
```

# OpenMP declare simd Pragma

- can generate vectorized versions of functions with `declare simd pragma`
- syntax:

```
#pragma omp declare simd [clause...]  
/* function declaration/definition */
```

- optional clauses may be specified to affect behavior of pragma (i.e., `simdlen`, `linear`, `aligned`, `uniform`, `inbranch`, and `notinbranch`)
- example:

```
#pragma omp declare simd  
float foo(float a, float b, float c) {  
    return a * b + c;  
}
```

# OpenMP SIMD-Related Pragma Clauses

- **safelen**(*length*)
  - specifies *length* as maximum number of iterations that can be run concurrently in safe manner (i.e., without data-dependence problems)
- **collapse**(*n*)
  - specifies how many (nested) loops to associate with loop construct (i.e., how many nested loops to combine)
- **simdlen**(*length*)
  - specifies *length* as preferred length of vector registers used
- **aligned**(*argument-list[:alignment]*)
  - specifies items in *argument-list* as having given alignment (e.g., *alignment*)
- **uniform**(*argument-list*)
  - indicates each argument in *argument-list* has constant value between iterations of given loop (i.e., constant value across all SIMD lanes)
- **inbranch**
  - specifies that function will always be called from inside conditional statement of SIMD loop
- **notinbranch**
  - specifies that function will never be called from inside conditional statement of SIMD loop

- `linear(list[:linear-step])`
  - specifies that, for every iteration of original scalar loop, each variable in *list* is incremented by particular step *step* (i.e., variable is incremented by *step* times vector length for vectorized loop)
- `private(list)`
  - declares variables in *list* to be private to each iteration
- `lastprivate(list)`
  - declares variables in *list* to be private to each iteration, and last value is copied out from last iteration instance
- `reduction(operator:list)`
  - specifies variables in *list* are reduction variables for operator *operator*



# Example: Vectorized Loop

```
1  #include <cstdint>
2  #include <iostream>
3  #include <numeric>
4
5  template <std::size_t align, std::size_t n, class T>
6  [[ gnu::noinline ]]
7  void multiply(const T (&a)[n], const T (&b)[n], T (&c)[n]) {
8      #pragma omp simd aligned(a, b, c : align)
9      for (int i = 0; i < n; ++i) {
10         c[i] = a[i] * b[i];
11     }
12 }
13
14 int main() {
15     constexpr std::size_t n = 65536;
16     constexpr std::size_t align = 16 * alignof(float);
17     alignas(align) static float a[n];
18     alignas(align) static float b[n];
19     alignas(align) static float c[n];
20     std::iota(a, &a[n], 0);
21     std::iota(b, &b[n], 0);
22     multiply<align>(a, b, c);
23     for (auto x : c) {
24         std::cout << x << '\n';
25     }
26 }
```

# Example: Vectorized Loop and Function

```
1  #include <cstdint>
2  #include <iostream>
3  #include <numeric>
4
5  #pragma omp declare simd notinbranch
6  float func(float a, float b) {
7      return a * a + b * b;
8  }
9
10 int main() {
11     constexpr std::size_t n = 65536;
12     constexpr std::size_t align = 16 * alignof(float);
13     alignas(align) static float a[n];
14     alignas(align) static float b[n];
15     alignas(align) static float c[n];
16     std::iota(a, &a[n], 0);
17     std::iota(b, &b[n], 0);
18     #pragma omp simd aligned(a, b, c : align)
19     for (int i = 0; i < n; ++i) {
20         c[i] = func(a[i], b[i]);
21     }
22     for (auto x : c) {
23         std::cout << x << '\n';
24     }
25 }
```

## Section 6.7.3

### **References**

- 1 Pablo Halpern. Introduction to Vector Parallelism. CppCon, Bellevue, WA, USA, Sept. 21, 2016. Available online at <https://youtu.be/h6Q-5Q2N5ck>.
- 2 Georg Zitzlsberger. C++ SIMD parallelism with Intel Cilk Plus and OpenMP 4.0. Meeting C++, Berlin, Germany, Dec. 5–6, 2014. Available online at <https://youtu.be/6oKRL7jz2LY>.

- 1 OpenMP web site, <http://www.openmp.org>.
- 2 Clang OpenMP page, <http://openmp.llvm.org>.
- 3 Cilk Plus web site, <http://www.cilkplus.org>.
- 4 A. Peleg, S. Wilkie, and U. Weiser. [Intel MMX for multimedia PCs](#). *Communications of the ACM*, 40(1):25–38, Jan. 1997.

## Section 6.8

# Documentation for Software Development

# Documentation for Software Development

- documentation plays essential role in software development process
- many benefits to formalizing in writing various aspects of software at different points in development process
- consider two types of documents:
  - 1 software requirements specification
  - 2 software design description
- software requirements specification (SRS): describes what software should do (from external viewpoint)
- software design description (SDD): describes how software works internally

# Software Requirements Specification (SRS)

- establishes agreement between consumer and contractors on what software is expected to do as well as what it is not expected to do
- can be thought of as contract between customer and contractor
- functionality: what does software do? (what problem does it solve?)
- external interfaces: how does software interact with external agents, such as humans, hardware, and software (e.g., command-line interface, graphical user interface, application program interface)
- performance: speed, availability, response time, recovery time of various functions
- attributes: considerations regarding reliability, availability, maintainability, portability, security
- design constraints imposed on implementation: implementation language, resource limits, operating environments
- assumptions upon which requirements are based



- distinguish classes of requirements:
  - essential: software will be unacceptable unless requirement met
  - conditional: would enhance software if requirement met, but not unacceptable if requirement not met
  - optional: class of functionality that may or may not be worthwhile
- should not leave details of software requirements to be determined
- only focus on what the software needs to do, not how done (i.e., should not describe any design or implementation details)
- typical use cases
- constraints imposed on software:
  - time constraints
  - memory constraints
- software limitations:
  - restrictions on input data
  - allowable ranges for parameters of methods
  - dependencies on other software (e.g., other programs needed to function)

# External Interfaces

- external interfaces: how software interacts with external agents, such as humans, hardware, and software
- command line interface (CLI) (for program)
  - options (e.g., required versus optional, default settings)
  - standard input, output, error
  - exit status
- graphical user interface (GUI) (for program)
  - window layout
  - user interaction (e.g., mouse/keyboard actions)
- application program interface (API) (for library)
  - constants
  - types, classes/methods
  - functions
  - namespaces
- format of all data used by software

- establishes basis for agreement between customer and contractors
- reduces development effort by thoroughly considering all requirements before starting design
- provides basis for estimating costs and schedules
- provides baseline for validation and verification
- facilitates transfer of software product to new users or machines
- serves as basis for enhancement

# SRS Example: Sorting Program

- single program that performs sorting
- given records as input, program sorts records and outputs records in sorted order
- record data format (for input and output):
  - records delimited by single newline character
  - each record consists of one or more fields, separated by one or more whitespace characters
- restrictions/constraints:
  - may assume sufficient memory to buffer all records
  - software must work without any modification to source code on any platform with C++ compiler compliant with C++11 standard
- records read from standard input
- sorted records written to standard output
- any error/warning messages written to standard error
- sorts records using  $n$ th field in record as key
- can sort in ascending or descending order
- sort key may be numeric or string

# SRS Example: Sorting Program (Continued)

- command-line interface:

```
sort [-r] [-k $n] [-n]
```

- supported command-line options:

Option	Description
-k \$n	Sort using <i>n</i> th field in record; if not specified, <i>n</i> defaults to 1.
-n	Treat key as real number (instead of string) for sorting purposes; if not specified, key treated as string.
-r	Sort in descending (instead of ascending) order; if not specified, defaults to ascending order.

- give examples illustrating expected use cases

# Software Design Description (SDD)

- high-level design: overview of entire system, identifying all its components at some level of abstraction (i.e., overall software architecture)
- detailed design (a.k.a. low-level design): full details of system and its components (e.g., types, functions, APIs, pseudocode, etc.)
- describes high-level and detailed design of software
- some context regarding functionality provided by software
- how design is recursively structured into constituent parts and role of those parts
- types and interfaces (e.g., classes and public members)
- data structures used to represent information to be processed
- internal interfaces (and external interfaces not described in SRS)
- interaction amongst entities
- algorithms

- describe overall structure of software
- carefully consider choice of data structures used to represent information being processed, as choice will almost always have performance implications
- specify any data formats used internally by software
- provide pseudocode for key parts of software
- state any potentially limiting assumptions made

# Benefits of SDD

- encourages better planning by forcing design ideas to be more carefully considered and organized
- allows greater scrutiny of design
- captures important design decisions, such as rationale for particular design choices
- allows newcomers to development team to become acquainted with software more easily
- provides point of reference to be used throughout project
- promotes reuse of code (since well documented code more likely to be reused)
- facilitates better software testing (since certain types of testing benefit from understanding of software design)



# SDD Example: Sorting Program

- Key alias for type that represents sort key (alias for `std::string`)
- Compare functor class for comparing Key objects
- Dataset class represents collection of all records
- specify all class interfaces (i.e., public members)
- Dataset class provides:
  - constructor that creates dataset by reading all records from input stream
  - function to output all records in sorted order to output stream
- Dataset class to use `std::multimap<Key, std::string, Compare>`
- allows  $n$  records to be sorted in  $O(n \log n)$  time [ $n$  insertions, each requiring  $O(\log n)$  time]
- handling  $n$  records requires  $O(n)$  memory
- only uses C++ standard library

# Requirements/Design Document for Degree Project

- document is combination of SRS and SDD with some added information about testing strategies
- briefly introduce problem being addressed by software
- describe each program and library to be developed
- identify parts of any external software (e.g., programs or libraries) that will be used
- describe user interface (e.g., CLI, GUI) for each program
- fully specify all data formats used
- describe overall structure of each program and library
- identify all key data structures and algorithms to be used
- provide pseudocode for key parts of the software
- state any potentially limiting assumptions made by software
- indicate how programs and library code will be tested
- offer any other information that may be helpful (since above list is not exhaustive)
- provide sufficient detail for other people to understand how software is to be structured and how it will be implemented and tested

- 1 IEEE Std. 1016-2009 — IEEE standard for information technology — systems design — software design descriptions, July 2009.
- 2 IEEE Std. 830-1998 — IEEE recommended practice for software requirements specifications, Oct. 1998.

## Section 6.9

# Software Testing

- **software testing** is process of checking to ensure that software (e.g., program, library, or system):
  - meets formal requirements (e.g., functional specification)
  - works as expected
- this involves such things as checking that software:
  - responds correctly to all inputs
  - has acceptable resource consumption (e.g., time and memory)
  - is sufficiently usable
  - works properly in all intended environments
  - does what is supposed to do
  - does not have undesired side effects (i.e., does not do things that it is not supposed to do)

Triangle Program

# Examples of Well-Known Software Failures

## ■ Therac-25 Radiation Overdoses

- from 1985 to 1987, at least six incidents in which patients being treated with Therac-25 radiation-therapy machine received massive overdoses of radiation (as much as 100 times intended dosage); at least three patients died
- failure due to race condition in control software

## ■ Northeast Blackout of 2003

- in 2003, widespread blackout occurred affecting several states/provinces in USA and Canada (i.e., New York, New Jersey, Maryland, Connecticut, Massachusetts, Ohio, and Ontario) with outage lasting from hours to days
- at time, world's second most widespread blackout in history
- blackout contributed to almost 100 deaths
- triggered by local outage that went undetected due to race condition in monitoring software

## ■ Ariane 5 Rocket Explosion

- in 1996, unmanned Ariane 5 launched by European Space Agency exploded 37 seconds after lift off
- cause of failure was numerical overflow in inertial guidance system software

# Examples of Well-Known Software Failures (Continued)

## ■ Toyota Electronic Throttle Control System

- in some 2009–2011 Toyota vehicles, electronic throttle control system (ETCS) caused sudden unintended acceleration
- at least 89 people were killed as result
- problem likely caused by stack overflow in control software

## ■ NASA Mars Climate Orbiter

- in 1998, NASA Mars Climate Orbiter approached Mars at wrong angle entering upper atmosphere and disintegrating
- failure due to software using incorrect units in calculations controlling thrusters

## ■ NASA Mars Polar Lander

- in 1999, when attempting to land on Mars, NASA Mars Polar Lander prematurely terminated firing of engine causing spacecraft to crash into planet's surface
- flight-control software mistook vibrations of stowed legs as surface touchdown

- **error**: mistake made by programmer when coding
- **defect** (also known as **fault** or **bug**): flaw in source code resulting from error
- **failure**: occurs when code corresponding to defect executes (i.e., defect causes system to behave incorrectly)
- not all defects result in failures, since code corresponding to defect might not execute (e.g., defect may be in unreachable code)
- **incident**: instance in which code does not behave as expected
- **test case**: specification of set of inputs and expected results associated with checking particular program behavior
- **test**: act of exercising software with test cases



# Some Types of Faults

## ■ input/output faults

- correct input not accepted; incorrect input accepted; wrong format; cosmetic; incomplete or missing result

## ■ interface faults

- parameter mismatch; incompatible types; call to wrong procedure; call to nonexistent procedure

## ■ logic faults

- missing or duplicate cases; missing or extraneous conditions; incorrect loop iteration; incorrect operator; infinite recursion

## ■ data faults

- using uninitialized variable; dereferencing null pointer; incorrect initialization; wrong variable used; wrong data reference; incorrect subscript; incorrect data scope; inconsistent data; scaling or units error

## ■ computation faults

- incorrect algorithm; missing computation; incorrect operation; arithmetic overflow or underflow; insufficient precision

## ■ multithreading faults

- race condition; deadlock

- testing methods can be classified into two categories:
  - 1 **static**: testing without executing code
  - 2 **dynamic**: testing by executing code
- examples of static testing methods:
  - code reviews
  - static analysis tools
- examples of dynamic testing methods:
  - black box
  - white box
  - gray box
  - dynamic analysis tools

# Code Reviews

- **code review** is process whereby source code examined by one or more people in effort to identify bugs or other problems
- variety of terms used to refer to code reviews (e.g., code inspections and code walkthroughs) but no widely accepted meaning for such terms
- code reviews can be applied at almost any stage of software development
- code review typically has several participants with original author often being one
- code review meeting might last couple of hours
- may involve some preparatory work prior to meeting (e.g., examining source code)
- code review process can take many forms, such as:
  - read code line by line looking for errors
  - select some test cases and then, for each test case, step through code as if being executed by computer
- may be formal written documentation to complete as part of review
- code reviews often find actual bug as opposed to failure resulting from bug
- code reviews can be effective at finding certain types of errors

- static analysis tools examine code without executing it
- most frequently analyze source code
- can be used to find certain types of code defects
- examples of types of defects that can be found by static analysis tools include:
  - use of uninitialized variable
  - unused variable
  - unreachable code (i.e., code that never executes) or dead code (i.e., code that has no effect)
  - coding standards violation
  - security vulnerability
  - syntax violation
  - inconsistent interfaces between modules and components

# Examples of Static Analysis Tools

- Clang Tidy
- Clang Static Analyzer
- CppCheck (<http://cppcheck.sourceforge.net>)
- CppLint (<http://github.com/cpplint/cpplint>)
- many commercial products also available, such as:
  - Coverity Scan (<http://scan.coverity.com>), which is free for use in open-source projects
  - CppDepend (<http://www.cppdepend.com>)
  - Klocwork  
(<http://www.roguewave.com/products-services/klocwork>)
  - PVS-Studio Analyzer (<http://www.viva64.com/en/pvs-studio>)

# Black-Box Testing (a.k.a. Functional Testing)

- **black-box testing** (also known as **functional testing**) is testing method that checks if software meets its specifications *without using knowledge of its internal structure*
- concentrates on finding circumstances in which software does not behave according to its specifications
- test cases derived solely from specifications
- advantages:
  - tester does not need to know internal structure of program
  - test cases can be created as soon as functional specifications complete
  - tester needs no knowledge of implementation, including specific programming languages
  - tests will be performed from user's point of view
  - testing can help to expose ambiguities or inconsistencies in specifications
  - as tester and developer are independent of each other, testing balanced and unprejudiced
- disadvantages:
  - may leave many program paths untested
  - high probability of repeating tests already performed by developer

# White-Box Testing (a.k.a. Structural Testing)

- **white-box testing** (also known as **structural testing**) is testing method that employs *knowledge of internal structure of software* under test
- tester has access to source code and possibly documentation describing software internals, such data structures and algorithms used
- derive testing strategy based on examination of internal structure of software
- typically choose tests to exercise as much of code and control flow as possible
- advantages:
  - more thorough testing possible than with black-box testing
  - forces developer to reason more carefully about implementation
- disadvantages:
  - as details of code are considered, skilled testers required with knowledge of programming
  - tests more sensitive to changes in underlying implementation (and may be rendered useless due to implementation changes)

- **gray-box testing** is combination of black-box (i.e., functional) and white-box (i.e., structural) testing
- in gray-box testing, tester has partial knowledge of internal structure of code to be tested
- for example, tester may have access to documentation describing internal data structures and algorithms used
- tester does not have access to source code
- advantages:
  - testing performed from point of view of users rather than developer
  - allows tester to prioritize tests based on partial knowledge of underlying program structure
- disadvantages:
  - some tests may be redundant since developer may have performed some similar tests



# Equivalence-Class Testing (a.k.a. Equivalence Partitioning)

- **equivalence-class testing** is testing method in which input domain (i.e., set of all possible inputs) for software being tested partitioned into equivalence classes, which then drive testing process
- particular partitioning of input domain determined by equivalence relation (i.e., elements in same equivalence class if and only if equivalent as determined by equivalence relation)
- each possible input appears in exactly one equivalence class
- union of all equivalence classes yields set of all possible inputs
- equivalence relation chosen so that inputs in same equivalence class expected to cause software to behave similarly
- consequently, if testing one input in equivalence class would detect defect, testing all other inputs in same equivalence class would be expected to find same defect
- challenge in equivalence-class testing is making good choice of equivalence relation (which determines equivalence classes) and good choice is *crucial*

- often equivalence relation chosen based on:
  - ranges of input variables
  - expected output
  - type of computation needed
- with regard to type of inputs considered, equivalence-class testing said to be:
  - **normal** if considers only valid inputs
  - **robust** if considers both valid and invalid inputs
- robust testing only possible in situations where invalid inputs required to yield well-defined behavior

## ■ advantages:

- equivalence classes allow many inputs to be tested using one representative element from equivalence class, greatly reducing number of test cases
- since equivalence classes are disjoint, can eliminate/reduce redundancy in tests

## ■ disadvantages:

- just because members of equivalence class should in theory behave similarly does not mean that they actually will in practice (e.g., due to unanticipated bugs)

# Equivalence-Class Testing Example [Volume Discount]

- consider program that calculates volume discount, where customer can order from 1 to 299 units:

Quantity	1-9	10-49	50-99	100-199	200-299
Discount (%)	0	1	2	3	5

- suppose that we partition into equivalence classes on basis of type of calculation performed (i.e., type of discount calculation)
- for normal testing, leads to 5 equivalence classes:

Class	1	2	3	4	5
Quantity	1-9	10-49	50-99	100-199	200-299

- for robust testing, add 2 equivalence classes:

Class	6	7
Quantity	0 or less	300 or greater

- select one test case for each equivalence class

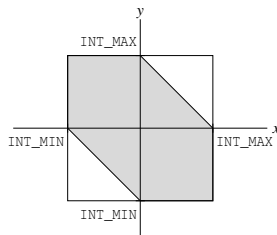
# Equivalence-Class Testing Example [Safe Add]

- consider function that computes sum of two integers safely (i.e., ensuring that overflow does not occur) and has declaration:

```
std::pair<int, bool> safe_add(int x, int y);
```

- one possible choice for equivalence classes:

- 1 overflow in positive direction:  $x \geq 0$  and  $y \geq 0$  and  $y > \text{INT\_MAX} - x$
- 2 overflow in negative direction:  $x < 0$  and  $y < 0$  and  $y < \text{INT\_MIN} - x$
- 3 no overflow:
  - $x \geq 0$  and  $y < 0$ ; or  $x < 0$  and  $y \geq 0$ ; or
  - $x \geq 0$  and  $y \geq 0$  and  $y \leq \text{INT\_MAX} - x$ ; or
  - $x < 0$  and  $y < 0$  and  $y \geq \text{INT\_MIN} - x$



- could then use three test cases, one from each equivalence class

# Boundary-Value Testing

- **boundary-value testing** is testing method that involves testing inputs at boundaries of equivalence classes
- rationale behind boundary-value testing is that errors tend to occur near extreme values of inputs (which typically correspond to boundaries of equivalence classes)
- for example, off-by-one errors common (e.g., iterate one too many or few times or array index one too great or little)
- with regard to types of inputs allowed, boundary-value testing said to be:
  - **normal** if only valid inputs considered
  - **robust** if considers both valid and invalid inputs
- robust testing only possible in situations where invalid inputs required to yield well-defined behavior

- again, consider program that calculates volume discount, where customer can order from 1 to 299 units:

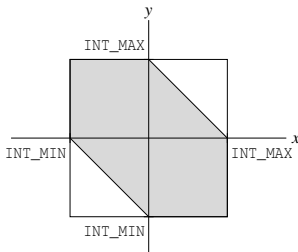
Quantity	Discount (%)
1-9	0
10-49	1
50-99	2
100-199	3
200-299	5

- for normal testing, could use values:
  - 1,2,8,9
  - 10,11,48,49
  - 50,51,98,99
  - 100,101,198,199
  - 200,201,298,299
- for robust testing, could add additional values:
  - -1,0
  - 300,301

# Boundary-Value Testing Example [Safe Add]

- again, consider function that computes sum of two integers safely (i.e., ensuring that overflow does not occur) and has declaration:  

```
std::pair<int, bool> safe_add(int x, int y);
```
- as shown in figure, can partition input domain into three equivalence classes (i.e., overflow in positive direction, overflow in negative direction, and no overflow):



- could choose test cases to correspond to points at or near boundaries between equivalence classes:
  - $x \geq 0$  and  $y \geq 0$  and  $y = \text{INT\_MAX} - x + \Delta$ , where  $\Delta \in \{-1, 0, 1, 2\}$
  - $x < 0$  and  $y < 0$  and  $y = \text{INT\_MIN} - x - \Delta$ , where  $\Delta \in \{-1, 0, 1, 2\}$



- **special-value testing** is type of testing where tester uses domain knowledge, experience with similar programs, and information about soft spots to devise test cases
- effectively tester guesses likely errors and tests for them
- form of experience-based testing
- perhaps most widely used form of functional testing
- most intuitive and probably least uniform (i.e., most ad hoc)
- effectiveness of special-value testing heavily dependent on skill of tester
- example: some test cases for sorting subroutine might include:
  - input list is empty
  - input list contains exactly one element
  - all entries in list have same value
  - input list is already sorted

# Special-Value Testing (Continued)

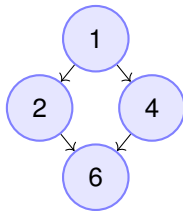
- example: some test cases for binary search in container (e.g., array) might include:
  - exactly one element in container being searched
  - container size is power of 2
  - container size is one greater and one less than power of 2
  - query element in container
  - query element not in container
- example: to test algorithm that calculates sinc function, might use knowledge that handling cases of computing  $\text{sinc}(x) = \sin(x)/x$  for  $x = 0$  and  $x \neq 0$  likely to be performed differently, leading to at least two test cases:
  - test for both zero and nonzero values

# Random Testing

- instead of choosing input variables deterministically, generate values randomly
- black-box testing technique
- has advantage of avoiding possible bias in testing
- may be difficult to determine how much testing is sufficient
- can be quite beneficial, for example, in testing algorithms that have very many possible edge/corner cases, which may be impossible to reliably identify and enumerate (in full)
- for example, might use random testing for algorithm that inserts point in triangulation (since certain bugs may only manifest themselves for very particular inputs that would be difficult to identify manually in reliable manner)
- particular test case that fails during random testing might also be codified as non-random test to ensure that failed test case always included in future testing; in this sense, random testing can also be used as means to generate non-random tests

- **control-flow graph** is directed graph that shows paths of execution in code
- nodes correspond to statement fragments or statements
- given two nodes  $i$  and  $j$  in control-flow graph, edge from  $i$  to  $j$  exists if and only if statement fragment corresponding to node  $j$  can be executed immediately after statement fragment corresponding to node  $i$
- program can be viewed in terms of control-flow graph
- control-flow graphs are particularly useful for purposes of considering structural coverage in software testing

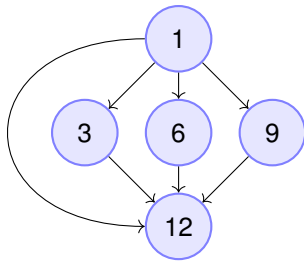
# Control-Flow Graph Examples: If-Else and Switch



declarations:

```
bool c; double x; double y;
```

```
1  if (c) {  
2      y = x * x;  
3  } else {  
4      y = x;  
5  }  
6  // ...
```

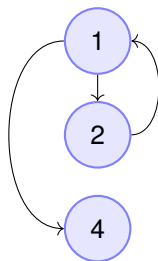


declarations:

```
int n; double x; double y;
```

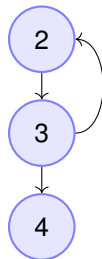
```
1  switch (n) {  
2  case 0:  
3      y = 0.0;  
4      break;  
5  case 1:  
6      y = 2.0 * x;  
7      break;  
8  case 2:  
9      y = 0.5 * x * x;  
10     break;  
11 }  
12 // ...
```

# Control-Flow Graph Examples: While and Do-While Loops



declarations:

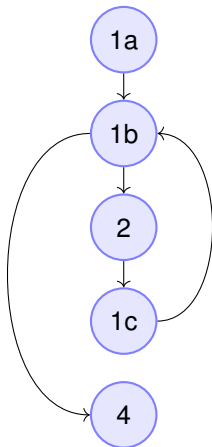
```
int n;  
1 while (n > 0) {  
2     --n;  
3 }  
4 // ...
```



declarations:

```
int n;  
1 do {  
2     --n;  
3 while (n > 0);  
4 // ...
```

# Control-Flow Graph Examples: For Loop



declarations:

```
int a[1024];  
1 for (int i = 0; i < 1024; ++i) {  
2     a[i] = 0;  
3 }  
4 // ...
```

# Structural Coverage Analysis

- structural coverage analysis used to determine which code structure exercised by test procedures
- such analysis can be helpful in order to:
  - create additional test cases to increase coverage
  - eliminate redundant test cases
  - detect unreachable code
  - determine quantitative measure of coverage, which can serve as indirect measure of quality
- structural coverage criteria can be classified into one of two types:
  - 1 control flow
  - 2 data flow
- control flow criteria measure flow of control between statements and sequences of statements
- data flow criteria measure flow of data between variable assignments and references to variables



# Examples of Structural-Coverage Criteria

- some examples of coverage criteria based on control flow include:
  - function coverage
  - statement coverage
  - decision coverage
  - condition coverage
  - condition/decision coverage
  - modified condition/decision coverage
  - multiple condition coverage
  - all paths coverage
- some examples of coverage criteria based on data flow include:
  - all-definitions coverage
  - all C-uses coverage
  - all P-uses coverage
  - all-uses coverage
  - all DU-paths coverage

- **function coverage** ensures that every function invoked at least once
- level of function coverage can be measured as number of functions invoked divided by total number of functions in program
- function coverage does not imply that all statements execute
- with only function coverage, very large fraction of code may never be executed
- function coverage is extremely weak coverage criterion and therefore not very useful by itself

# Statement Coverage

- **statement coverage** ensures that every statement executed at least once
- statement coverage ensures that each node in control-flow graph visited during testing (consequently, also called **node coverage**)
- level of statement coverage can be measured as number of (distinct) statements executed divided by total number of (distinct) statements in program
- statement coverage alone not particularly strong coverage criterion
- some weaknesses of statement coverage include:
  - does not usually test all branch outcomes
  - predicates may be tested for only one value
  - loop bodies may only be executed once
- statement coverage for following code fragment can be achieved with single test case (e.g.,  $(a,b) = (1,1)$ ):

```
if (a > 0)
    ++x;
if (b > 0)
    --x;
```
- in above example, statement coverage achieved with single test case does not exercise all possible branch outcomes

# Conditions and Decisions

- **condition**: boolean expression containing no boolean operators (i.e., boolean expression that cannot be decomposed into boolean subexpressions)
- **decision**: boolean expression composed of conditions and zero or more boolean operators (or integral expression in case of **switch** statement)
- decisions associated with branching (i.e., control-flow) constructs (e.g., **if**, **switch**, **for**, **while**, **do**, and ternary operator)
- example:
  - consider code fragment:

```
if (x > 0 && x < 1024) { /* ... */
```
  - `x > 0` is condition
  - `x < 1024` is condition
  - `x > 0 && x < 1024` is decision but not condition since can be decomposed further
- short-circuit evaluation of boolean expressions has implications in terms of what constitutes decision, as short-circuit evaluation introduces additional control flow

# Decision Coverage (a.k.a. Branch Coverage)

- **decision coverage** (also known as **branch coverage**) ensures that:
  - 1 every point of entry invoked at least once
  - 2 every decision assumes every possible outcome (e.g., true or false) at least once (i.e., every branch direction traversed at least once)
- decision coverage ensures that each edge in control-flow graph traversed during testing (consequently, also called **edge coverage**)
- decision coverage includes statement coverage (so decision coverage stronger criterion than statement coverage)
- some weaknesses of decision coverage include:
  - does not necessarily ensure that each condition upon which decision based assumes all possible values (e.g., true and false)
- decision coverage for following code fragment requires at least two test cases (e.g.,  $(a,b) \in \{(0,0), (1,1)\}$ ):

```
if (a > 0)
    ++x;
if (b > 0)
    --x;
```

# Decision Coverage Example [(a || b) && c]

- consider following code fragment (with partial control-flow graph shown **earlier**):

```
if ((a || b) && c)
    ++x;
```

- assuming no short-circuiting logic:
  - **one decision** ((a || b) && c) with three conditions, namely, a, b, and c
  - decision coverage requires at least two test cases (i.e., one test case for each possible outcome of single binary decision), such as:

#	a	b	c	((a    b) && c)
1	F	F	F	F
2	T	T	T	T

- assuming short-circuiting logic:
  - effectively, **three decisions**: one with condition a, one with condition b, and one with condition c
  - decision coverage requires at least three test cases, such as:

#	a	b	c	((a    b) && c)
1	F	F	X	F
2	F	T	F	F
3	T	X	T	T

- **condition coverage** ensures that:
  - 1 every point of entry invoked at least once
  - 2 every condition in every decision assumes each possible outcome at least once
- condition coverage often does not yield decision coverage
- condition coverage alone is very weak criterion so only useful when combined with other criteria
- for following code fragment, condition coverage can be achieved (without statement or decision coverage) with only two test cases (e.g.,  $(a,b) \in \{(0,1), (1,0)\}$ ):

```
    if (a > 0 && b > 0)
        ++x;
```
- in above example, note that condition coverage is achieved without achieving either decision or statement coverage

# Condition/Decision Coverage

- **condition/decision coverage** ensures that:
  - 1 every point of entry in program invoked at least once
  - 2 every decision assumes each possible outcome (e.g., true or false) at least once
  - 3 every condition in each decision assumes each possible outcome at least once
- that is, condition/decision coverage simply combines requirements for decision coverage with those for condition coverage
- condition/decision coverage includes statement, decision, and condition coverages
- for following code fragment, condition/decision coverage can be achieved with two test cases (e.g.,  $(a,b) \in \{(0,0), (1,1)\}$ ):

```
if (a > 0 && b > 0)
    ++x;
```



# Condition/Decision Coverage Example [(a || b) && c]

- consider following code fragment (with partial control-flow graph shown **earlier**):

```
if ((a || b) && c)
    ++x;
```

- assuming no short-circuiting logic:
  - **one decision** ((a || b) && c) with three conditions, namely, a, b, and c
  - condition/decision coverage requires at least two test cases, such as:

#	a	b	c	((a    b) && c)
1	F	F	F	F
2	T	T	T	T

- assuming short-circuiting logic:
  - effectively, **three decisions**: one with condition a, one with condition b, and one with condition c
  - condition/decision coverage requires at least three test cases, such as:

#	a	b	c	((a    b) && c)
1	F	F	X	F
2	F	T	F	F
3	T	X	T	T

# Modified Condition/Decision Coverage (MCDC)

- **modified condition/decision coverage (MCDC)** ensures that:
  - every point of entry invoked at least once
  - every condition in each decision in program assumes each possible outcome at least once
  - each condition in decision shown to independently affect that decision's outcome
- several different variants of MCDC, each differing in how independence shown (i.e., last condition above)
- **unique-cause MCDC**: condition shown to independently affect decision's outcome by varying just that condition while holding fixed all other possible conditions
- **masking MCDC**: condition is shown to independently affect decision's outcome by applying principle of boolean logic to assure that no other condition influences output (even though more than one condition in decision may change value)

- unique-cause MCDC cannot handle decisions with strongly coupled conditions
- for decision with  $n$  uncoupled (i.e., independent) conditions, need at least  $n + 1$  test cases
- MCDC includes statement, decision, condition, and condition/decision coverages

- consider following code fragment:

```
if (a || b)
    x = 42;
else
    x = 0;
```

- assuming no short-circuiting logic, unique-cause MCDC requires at least 3 test cases, such as:

#	a	b	(a    b)
1	F	F	F
2	T	F	T
3	F	T	T

	Independence Pairs
a	(1, 2)
b	(1, 3)

- assuming short-circuiting logic, masking MCDC requires at least 3 test cases, such as:

#	a	b	(a    b)
1	F	F	F
2	T	X	T
3	F	T	T

	Independence Pairs
a	(1, 2)
b	(1, 3)

# MCDC Example [a && b && c && d]

- consider following code fragment:

```
if (a && b && c && d)
    x = 42;
else
    x = 0;
```

- assuming no short-circuiting logic, unique-cause MCDC requires at least 5 test cases, such as:

#	a	b	c	d	a && b && c && d
1	T	T	T	T	T
2	F	T	T	T	F
3	T	F	T	T	F
4	T	T	F	T	F
5	T	T	T	F	F

	Independence Pairs
a	(1, 2)
b	(1, 3)
c	(1, 4)
d	(1, 5)

- assuming short-circuiting logic, masking MCDC requires at least 5 test cases, such as:

#	a	b	c	d	a && b && c && d
1	T	T	T	T	T
2	F	X	X	X	F
3	T	F	X	X	F
4	T	T	F	X	F
5	T	T	T	F	F

	Independence Pairs
a	(1, 2)
b	(1, 3)
c	(1, 4)
d	(1, 5)

# MCDC Example [(a || b) && c]

- consider following code fragment (with partial control-flow graph shown earlier):

```
if ((a || b) && c)
    ++x;
```

- assuming no short-circuiting logic, unique-cause MCDC requires at least 4 test cases, such as:

#	a	b	c	(a    b) && c
1	F	F	T	F
2	T	F	T	T
3	F	T	T	T
4	T	F	F	F

Independence Pairs	
a	(1, 2)
b	(1, 3)
c	(2, 4)

- assuming short-circuiting logic, masking MCDC requires at least 4 test cases, such as:

#	a	b	c	(a    b) && c
1	F	F	X	F
2	T	X	T	T
3	F	T	T	T
4	T	X	F	F

Independence Pairs	
a	(1, 2)
b	(1, 3)
c	(2, 4)

- as seen earlier, condition/decision coverage only requires 3 test cases

# Multiple-Condition Coverage

- **multiple-condition coverage** ensures that:
  - 1 every point of entry invoked at least once
  - 2 each possible combination of condition outcomes in each decision occurs at least once
- that is, multiple-condition coverage requires exhaustive testing of all possible combinations of condition outcomes
- multiple-condition coverage is very strong criterion as it includes statement, decision, condition, and condition/decision, and MCDC coverages
- for decision with  $n$  uncoupled (i.e., independent) conditions and no short-circuiting logic, need at least  $2^n$  test cases
- often, multiple-condition coverage is impractical as it can result in very large number of test cases being required

# Multiple-Condition Coverage Example [(a || b) && c]

- consider following code fragment (with partial control-flow graph shown **earlier**):

```
if ((a || b) && c)
    ++x;
```

- assuming no short-circuiting logic, multiple-condition coverage requires at least  $2^3 = 8$  test cases (i.e., 1 test case for each possible combination of outcomes from 3 conditions):

#	a	b	c	(a    b) && c
1	F	F	F	F
2	T	F	F	F
3	F	T	F	F
4	T	T	F	F

#	a	b	c	(a    b) && c
5	F	F	T	F
6	T	F	T	T
7	F	T	T	T
8	T	T	T	T

- assuming short-circuiting logic, multiple-condition coverage requires at least 5 test cases, such as:

#	a	b	c	(a    b) && c
1	F	F	X	F
2	T	X	F	F
3	F	T	F	F
4	T	X	T	T
5	F	T	T	T



# All-Paths Coverage

- **all-paths coverage** ensures that every path in program taken at least once
- in other words, all-paths coverage ensures that every possible path from source node to sink node in control-flow graph is taken during testing
- all-paths coverage includes function, statement, decision, condition, condition/decision, MCDC, and multiple-condition coverages
- except in case of trivial programs, number of possible paths astronomically large
- testing every path infeasible except in very trivial programs
- consequently, all-paths coverage of little practical utility
- moreover, even if every path could be tested, not sufficient to test every path through code, since:
  - testing every path may not guarantee that program meets all specifications
  - defects in code could be due to missing paths and non-existent paths cannot be tested
  - whether bug manifests itself on particular path through code may depend on program state at each step (i.e., data sensitivity)

# All-Paths Coverage Examples

- consider following code fragment:

```
if (a)
    x = 42;
if (b)
    y = 42;
```

- all-paths coverage would require at least 4 test cases, such as:

#	a	b
1	F	F
2	T	F
3	F	T
4	T	T

- 
- consider following code fragment, where `n` is of type **unsigned int** and **unsigned int** is 64-bit integer:

```
for (unsigned int i = 0; i < n; ++i)
    sum += a[i];
```

- all-paths coverage would require at least  $2^{64} \approx 1.84 \cdot 10^{19}$  (i.e., about 18.4 quintillion) test cases

# Comparison of Various Structural-Coverage Types

Coverage Criterion	Statement Coverage	Decision Coverage	Condition Coverage	Condition/Decision Coverage	MCDC Coverage	Multiple Condition Coverage
every point of entry and exit invoked at least once		✓	✓	✓	✓	✓
every statement executed at least once	✓	✓*		✓*	✓*	✓*
every decision has taken all possible outcomes at least once		✓		✓	✓	✓
every condition in each decision has taken all possible outcomes at least once			✓	✓	✓	✓
every condition in each decision shown to independently affect that decision's outcome					✓	✓
every combination of condition outcomes within each decision invoked at least once						✓

\*not explicitly required in coverage definition but always implicitly satisfied

- data-flow coverage considers how data modified and accessed (i.e., writes to and reads from variables)
- data-flow analysis particularly well suited to detecting problems such as:
  - variable used before being initialized
  - variable initialized but never used
  - variable initialized more than once before its use
- data-flow coverage criteria not used as much in practice as control-flow counterparts

# Variable Definitions and Uses

- each occurrence of variable in program classified as being either definition or use
- **definition**: occurrence of variable where its value is set (i.e., written)  

```
std::string s; // definition of s (via construction)
i = 0; // definition of i
*p = 42; // definition of *p
```
- **use**: occurrence of variable where its value is used (i.e., read)
- each use can be classified as either being predicate use or computation use
- **predicate use (P-use)**: use of variable in predicate (i.e., condition) of branch statement; for example:  

```
if (x > 0) { /* ... */ } // P-use of x
while (n != 0) { /* ... */ } // P-use of n
```
- **computation use (C-use)**: use of variable that is not P-use; for example:  

```
y = x + 1; // C-use of x
y = sin(x); // C-use of x (pass by value parameter)
```
- variable can be both used and re-defined in same statement; for example:  

```
x = 3 * x + 2; // x read and then written
increment(x); // x incremented (via pass by reference)
```

# Definition-Clear and Definition-Use Paths

- C-use associated with *node* in graph
- P-use associated with *edge* in graph
- **definition-clear (DC) path** with respect to variable  $x$  from node  $i$  to *node*  $j$  is path  $(i, n_1, \dots, n_m, j)$ , where  $m \geq 0$ , that contains no definitions of  $x$  in nodes  $n_1, \dots, n_m$
- **definition-clear (DC) path** with respect to variable  $x$  from node  $i$  to *edge*  $(j, k)$  is path  $(i, n_1, \dots, n_m, j, k)$ , where  $m \geq 0$ , that contains no definitions of  $x$  in nodes  $n_1, \dots, n_m, j$
- **definition-use (DU) path** with respect to variable  $x$  is path  $(n_1, \dots, n_j, n_k)$  such that  $n_1$  has definition of  $x$  and either:
  - node  $n_k$  has C-use of  $x$  and  $(n_1, \dots, n_j, n_k)$  is DC simple path with respect to  $x$ ; or
  - edge  $(n_j, n_k)$  has P-use of  $x$  and  $(n_1, \dots, n_j)$  is DC loop-free path with respect to  $x$

# Define/Use Coverage Criteria

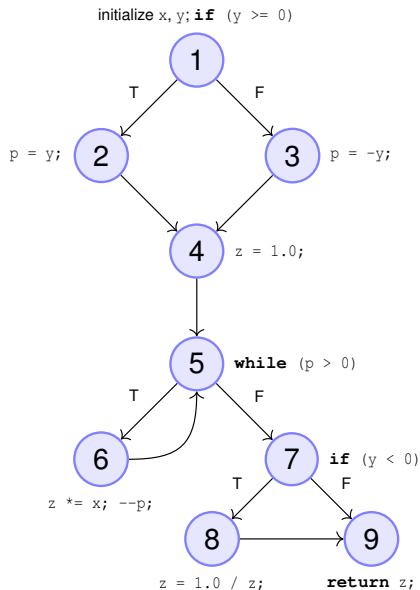
- **all definitions coverage**: for every program variable  $x$ , at least one DC path from every definition of  $x$  to at least one C-use or one P-use of  $x$  must be covered
- **all C-uses coverage**: for every program variable  $x$ , at least one DC path from every definition of  $x$  to every C-use of  $x$  must be covered
- **all P-uses coverage**: for every program variable  $x$ , at least one DC path from every definition of  $x$  to every P-use (including all outgoing edges of predicate statement) of  $x$  must be covered
- **all uses coverage**: for every program variable  $x$ , at least one DC path from every definition of  $x$  to every C-use and P-use (including all outgoing edges of predicate statement) of  $x$  must be covered
- **all DU-paths coverage**: for every program variable  $x$ , every DU path from every definition of  $x$  to every C-use and P-use of  $x$  must be covered
- all-definitions coverage subsumed by all-uses coverage subsumed by all DU-paths coverage

# Power Example: Source Code

```
1  double power(double x, int y) {  
2      int p;  
3      if (y >= 0) {  
4          p = y;  
5      } else {  
6          p = -y;  
7      }  
8      double z = 1.0;  
9      while (p > 0) {  
10         z *= x;  
11         --p;  
12     }  
13     if (y < 0) {  
14         z = 1.0 / z;  
15     }  
16     return z;  
17 }
```



# Power Example: Control-Flow Graph, Definitions, C-Uses, and P-Uses



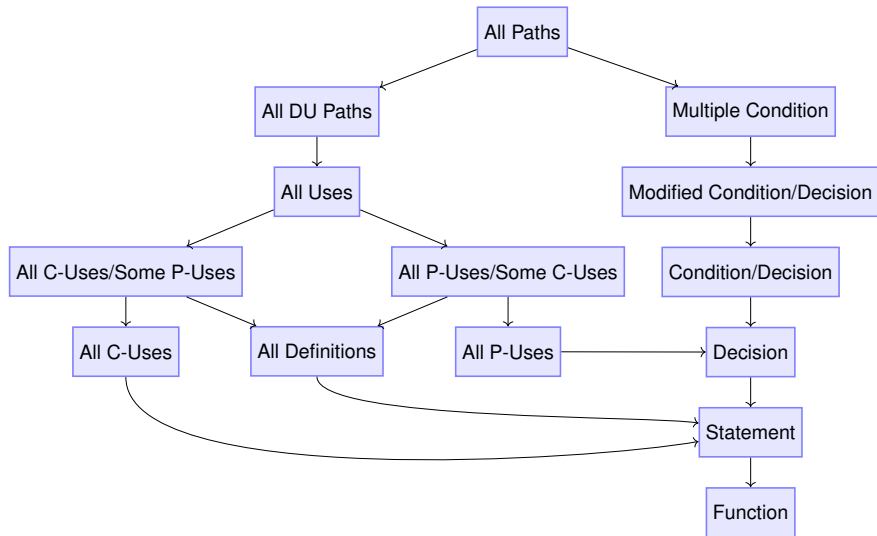
Node	Definitions	C Uses
1	x, y	
2	p	y
3	p	y
4	z	
5		
6	z, p	x, z, p
7		
8	z	z
9		z

Edge	P Uses
(1, 2)	y
(1, 3)	y
(5, 6)	p
(5, 7)	p
(7, 8)	y
(7, 9)	y

# Difficulties With Data-Flow Coverage

- often does not scale well to large real-world programs
- how to handle arrays
- aliasing poses challenges
- underestimation of aliases may fail to consider some definitions or uses
- overestimation of aliases may introduce infeasible test obligations
- can be more difficult to design test cases

# Structural Coverage Criteria Subsumption Hierarchy



# Structural Coverage in Various Standards

- DO-178C — Software Considerations in Airborne Systems and Equipment Certification
  - primary document used by certification authorities (such as FAA, EASA, and Transport Canada) to approve all commercial software-based aerospace systems
  - code that could result in catastrophic failure (e.g., crash) must have statement, decision, and MCDC coverages
  - code that could result in hazardous failure (e.g., passenger fatality) must have decision and statement coverages
  - code that could result in major failure (e.g., passenger injury) must have statement coverage
- ISO 26262 — Road vehicles — Functional safety
  - standard for functional safety of electrical and electronic systems in production automobiles
  - for most critical systems, highly recommends decision and MCDC coverages and recommends statement coverage
  - for least critical systems, highly recommends statement coverage and recommends decision and MCDC coverages

# Structural Coverage in Various Standards (Continued)

- IEC 61508 — Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems
  - safety standard that applies to many industries
  - for most critical systems, highly recommends statement, decision, and MCDC coverages
  - for least critical systems, recommends statement, decision, and MCDC coverages
- EN 50128 — Railway Applications — Communication, Signalling and Processing Systems (a.k.a. IEC 62279)
  - pertains to programmable electronic systems used in railway control and protection applications
  - for most critical systems, highly recommends statement, decision, and MCDC coverage
- IEEE 1008-1987 — IEEE Standard for Software Unit Testing
  - defines integrated approach to systematic and documented unit testing
  - specifies statement coverage as completeness requirement
  - recommends decision coverage for code that is critical or has inadequate requirements specification

# Structural Coverage in Practice

- for software in safety-critical systems (i.e., systems in which failures could result in death or serious injury), desirable to have 100% coverage with respect to statement, decision, condition/decision, and MCDC criteria
- for software that is not safety/mission critical, probably want to achieve 80% coverage or better with respect to each of statement, decision, and condition/decision coverage,
- many organizations require at least 85% statement coverage for all projects
- some people believe that anything less than 100% statement, decision, or condition/decision coverage does not ensure sufficient quality, but achieving 100% coverage often requires very substantial effort

- dynamic analysis tools examine code by executing it
- dynamic analysis tools can be used to:
  - find certain types of code defects
  - measure structural coverage
- some types of code defects that can be found by dynamic analysis tools include:
  - invalid memory accesses
  - memory leaks
  - data races
  - use of uninitialized variable
  - signed integer overflow
  - use of memory after deallocation
  - double deallocation

# Examples of Dynamic Analysis Tools

- Address Sanitizer (ASan)
- Undefined-Behavior Sanitizer (UBSan)
- Thread Sanitizer (TSan)
- Memory Sanitizer (MSan)
- Leak Sanitizer (LSan)
- Valgrind (Callgrind, Cachegrind, Helgrind, Memcheck, and Massif) and KCacheGrind
- Gcov, Lcov, Gcovr, and LLVM Cov
- Kcov
- Control Flow Integrity (CFI) Sanitizer  
(<https://clang.llvm.org/docs/ControlFlowIntegrity.html>)
- many commercial products also available, such as:
  - Testwell CTC++ (<http://www.testwell.fi/ctcdesc.html>)
  - Bullseye Coverage (<http://www.bullseye.com>)
  - PurifyPlus (<http://teambblue.unicomsi.com/products/purifyplus>)



# Testing Levels



- software testing levels are stages of testing performed as part of software-development lifecycle
- four levels of testing (in order performed) are:
  - 1 unit testing
  - 2 integration testing
  - 3 system testing
  - 4 acceptance testing

# Unit Testing

- **unit testing** is stage of software testing in which individual components (i.e., units) tested
- first level of software testing; happens before integration testing
- unit being tested typically corresponds to:
  - class or class method
  - function
- goal is to isolate unit and validate its correctness
- only ensures that unit works correctly in isolation
- conducted by developer
- usually uses white-box testing but may also employ gray-box and black-box testing
- can be done manually but is usually automated
- unit testing allows problems to be found early in development cycle
- unit testing might include (amongst other things): static code analysis, running test cases in conjunction with structural-coverage analysis, and code reviews

# Integration Testing

- **integration testing** is stage of software testing in which individual components combined and tested as group
- second level of testing; occurs after unit testing and before system testing
- components may be combined together incrementally or all at once (i.e., big bang)
- usually prefer to combine components incrementally as this typically allows bugs to be isolated more quickly
- purpose is to expose defects in interfaces and interaction between integrated components
- progressively larger groups of components integrated until software works as system
- employs black-box or white-box testing
- performed by tester (or in some cases developer)

# System Testing

- **system testing** is testing of complete software product (i.e., consisting of all integrated components) against specifications
- third level of testing; occurs after integration testing and before acceptance testing
- performed before completed software product introduced to market
- goal is to ensure that system meets its requirements
- conducted by tester
- employs black-box testing
- examples of different types of testing that may be included in system testing:
  - graphical user interface testing
  - usability testing
  - compatibility testing
  - performance testing
  - regression testing
  - installation testing
  - recovery testing

- **acceptance testing** is stage of testing that evaluates compliance with business requirements and assesses whether acceptable for delivery
- fourth (and last) level of testing; after system testing
- usually uses black-box testing and often ad hoc
- internal acceptance testing performed by members of organization that developed software but are not directly involved in project
- external acceptance testing performed by people outside organization that developed software

- **regression testing** involves re-running tests to ensure that software still works correctly after change or addition
- **regression**: defect (where software has reverted to exhibiting incorrect behavior) due to code change
- changes that may require regression testing include:
  - applying bug fixes
  - adding enhancements or new features
  - making configuration changes (e.g., changes to configuration files)
- may re-run full test suite or subset thereof, depending on particular situation

- **performance testing** checks how software will behave and perform under various workloads
- may consider such factors as:
  - speed
  - resource usage (e.g., memory, disk, and network)
  - reliability
  - scalability
  - responsiveness
  - throughput
- **load testing** checks how software behaves under anticipated workloads
- **stress testing** checks how software behaves under extreme workloads
- **soak testing** (also known as **endurance testing**) checks how well software can handle expected workload over long periods of time

- **fuzz testing** (also known as **fuzzing**) is automated technique used to find bugs in code by providing massive amounts of random data as input to test subject in effort to induce failure
- tool used to perform fuzz testing called **fuzzer**
- particularly effective for code that processes streams of characters (e.g., file-format processing and parsers)
- often exploit genetic algorithms in order to mutate input data
- may be aware of input structure
- may be aware of program structure (e.g., coverage-guided fuzzing)



## ■ American Fuzzy Lop (AFL)

- security-oriented fuzzer that uses compiler-time instrumentation and genetic algorithms
- web site: <http://lcamtuf.coredump.cx/afl>

## ■ libFuzzer

- coverage-guided fuzzer in LLVM
- web site: <https://llvm.org/docs/LibFuzzer.html>

## ■ OSS-Fuzz

- service provided by Google for continuous fuzzing of open-source software with large user base
- employs libFuzzer fuzzing engine
- web site: <https://github.com/google/oss-fuzz>

- **unit-test framework** is tool that provides infrastructure for facilitating unit testing
- by using unit-test framework, can avoid developing code that runs tests and instead focus on only on code for tests themselves
- **test fixture**: provides properly initialized context in which to run tests (which is typically used by *multiple* tests)
- **mock object**: simulated object that mimics behavior of real objects in controlled ways
- unit-test frameworks vary greatly in features offered, and may include support for things such as:
  - test fixtures (to reduce duplication of boiler-plate code)
  - mock objects
  - mechanisms for performing floating-point comparisons with tolerances
  - variety of output formats for reporting test results
  - multithreaded code

# Examples of Unit-Test Frameworks

- Catch2
- Google Test (<http://github.com/google/googletest>)
- Boost Test (<http://www.boost.org>)
- many commercial products also available, such as:
  - Parasoft C/C++ Test (<https://www.parasoft.com/products/ctest>)
  - Testwell CTA++ (<http://www.testwell.fi/ctadesc.html>)

# Bugs in Concurrent Programs

- multi-threaded programs can have all same types of bugs as single-threaded programs plus many more
- some types of bugs specific to multi-threaded programs include:
  - race conditions
  - data races
  - deadlocks
  - livelocks
  - starvation
  - atomicity violation (i.e., assumption that code block executes atomically is violated)
  - ordering violation (i.e., assumption regarding relative order in which operations performed is violated)
  - unintended sharing

# Strategies for Testing Concurrent Programs

- conduct detailed code inspections, looking for problems such as:
  - accessing shared data without taking lock
  - taking multiple locks in inconsistent orders
  - making invalid assumptions about happens-before relationships
- test code in manner that varies code timing as much as possible, since many bugs will only show with very particular code timing (as many bugs caused by race conditions)
- as in testing of any code, ensure:
  - good code coverage
  - boundary conditions encountered
- perform load (i.e., stress) testing, since many types of bugs more likely to show themselves under extreme conditions
- allow test code to run for longer periods of time in hopes that less frequently occurring timings that may expose bugs will eventually be encountered

# Strategies for Testing Concurrent Programs (Continued)

- utilize data-race detection tools such as:
  - Thread Sanitizer (TSan)
  - Valgrind/Helgrind
- test code on all platforms of interest (e.g., different processor architectures and operating systems)
- since implementations of synchronization primitives may differ across operating systems, some bugs may show themselves only under certain operating systems
- some implementations of synchronization primitives (particularly those on processor architectures with relatively strong memory models) may provide more guarantees than what is requested by code, which can mask bugs
- suppose, for example, that code requires particular aligned 32-bit memory access to be atomic in order to work correctly, but does not specifically request access to be atomic, which is bug
- preceding bug would never show itself on processor architecture where all aligned 32-bit memory accesses are atomic

## Section 6.9.1

### **References**

- 1 P. C. Jorgensen. *Software Testing — A Craftsman's Approach*. CRC Press, Boca Raton, FL, USA, 4th edition, 2014.
- 2 G. J. Myers, C. Sandler, and T. Badgett. *The Art of Software Testing*. Wiley, Hoboken, NJ, USA, 3rd edition, 2012.
- 3 A. Zeller. *Why Programs Fail — A Guide to Systematic Debugging*. Morgan Kaufmann, Burlington, MA, USA, 2nd edition, 2009.
- 4 R. Black. *Advanced Software Testing — Volume 1 — Guide to the ISTQB Advanced Certification as an Advanced Test Analyst*. Rocky Nook, Santa Barbara, CA, USA, 2009.
- 5 R. Black. *Advanced Software Testing — Volume 2 — Guide to the ISTQB Advanced Certification as an Advanced Test Manager*. Rocky Nook, Santa Barbara, CA, USA, 2nd edition, 2014.



- 6 J. L. Mitchell and R. Black. *Advanced Software Testing — Volume 3 — Guide to the ISTQB Advanced Certification as an Advanced Technical Test Analyst*.  
Rocky Nook, Santa Barbara, CA, USA, 2nd edition, 2015.
- 7 K. J. Hayhurst, D. S. Veerhusen, J. J. Chilenski, and L. K. Rierson. [A practical tutorial on modified condition/decision coverage](#).  
Technical Report NASA/TM-2001-210876, NASA, May 2001.
- 8 J. J. Chilenski and S. P. Miller. [Applicability of modified condition/decision coverage to software testing](#).  
*Software Engineering Journal*, 9(5):193–200, Sept. 1994.
- 9 J. J. Chilenski. [An investigation of three forms of the modified condition decision coverage \(MCDC\) criterion](#).  
Technical Report DOT/FAA/AR-01/18, Federal Aviation Administration, Apr. 2001.

- 10 L. A. Clarke, A. Podgurski, D. J. Richardson, and S. J. Zeil. [A formal evaluation of data flow path selection criteria.](#)  
*IEEE Trans. on Software Engineering*, 15(11):1318–1332, Nov. 1989.
- 11 S. Rapps and E. J. Weyuker. [Selecting software test data using data flow information.](#)  
*IEEE Trans. on Software Engineering*, 11(4):367–375, Apr. 1985.
- 12 T. Su, K. Wu, W. Miao, G. Pu, J. He, Y. Chen, and Z. Su. [A survey on data-flow testing.](#)  
*ACM Computing Surveys*, 50(1):5:1–5:35, Mar. 2017.
- 13 H. Zhu. [A formal analysis of the subsume relation between software test adequacy criteria.](#)  
*IEEE Trans. on Software Engineering*, 22(4):248–255, Apr. 1996.

- 14 T. K. Paul and M. F. Lau. A systematic literature review on modified condition and decision coverage.  
In *Proc. of ACM Symposium on Applied Computing*, pages 1301–1308, Mar. 2014.
- 15 T. Su, A bibliography of papers and tools on data-flow testing,  
<https://tingsu.github.io/files/dftbib.html>, 2017.

- 1 Marshall Clow. Making Your Library More Reliable with Fuzzing. C++Now, Aspen, CO, USA, May 10, 2018. Available online at <https://youtu.be/LlLJRHToyUk>.
- 2 Craig Young. Fuzz Smarter, Not Harder: An AFL Fuzz Primer. BSidesSF, San Francisco, CA, USA, Feb. 28–29, 2016. Available online at <https://youtu.be/29Rb05bftwo>.
- 3 Daniel Marjamaki. Cppcheck: Static Analysis of C++ Code. SwedenCpp::Stockholm, Sundbyberg, Sweden, Dec. 11, 2018. Available online at <https://youtu.be/ATkTqBqjYBY>.

## Part 7

# Debugging and Testing Tools

## Section 7.1

# Debuggers

# Source-Level Debuggers

- unfortunately, software does not always work as intended due to errors in code (i.e., bugs)
- how does one go about fixing bugs in time-efficient manner?
- source-level debugger is essential tool
- single stepping: step through execution of code, one source-code line at a time
- breakpoints: pause execution at particular points in code
- watchpoints: pause execution when the value of variable is changed
- print values of variables

- GNU Debugger (GDB) is powerful source-level debugger
- home page: <http://www.gnu.org/software/gdb>
- available on most platforms (e.g., Unix, Microsoft Windows)
- most popular source-level debugger on Unix systems
- allows one to see what is happening inside program as it executes or what a program was doing at the moment it crashed
- has all of the standard functionality of a source-level debugger (e.g., breakpoints, watchpoints, single-stepping)
- `gdb` command
- command-line usage:  
`gdb [options] executable`



`help`

Print help information.

`quit`

Exit debugger.

`run` [*arglist*]

Start the program (with *arglist* if specified).

`print` *expr*

Display the value of the expression *expr*.

`bt`

Display a stack backtrace.

`list`

Type the source code lines in the vicinity of where the program is currently stopped.

`break` *function*

Set a breakpoint at the function *function*.

`watch` *expr*

Set a watchpoint for the expression *expr*.

`c`

Continue running the program (e.g., after stopping at a breakpoint).

`next`

Execute the next program line, stepping over any function calls in the line.

`step`

Execute the next program line, stepping into any function calls in the line.

# GNU Data Display Debugger (DDD)

- graphical front-end to command-line debuggers such as GDB
- has some fancy graphical data display functionality
- all `gdb` commands available in text window, but can use graphical interface to enter commands as well
- home page: <http://www.gnu.org/software/ddd>
- `ddd` command

## Section 7.2

# Code Sanitizers

- **code sanitizer**: tool for automatically performing variety of run-time checks on code
- typically requires compiler to instrument code
- may also need library for run-time support
- several code sanitizers supported by Clang and/or GCC, including:
  - Address Sanitizer (**ASan**)
  - Thread Sanitizer (**TSan**)
  - Memory Sanitizer (**MSan**)
  - Undefined-Behavior Sanitizer (**UBSan**)
  - Leak Sanitizer (**LSan**)
- sanitizers easy to use
- can easily catch many bugs
- overhead of code sanitizer typically much less than that of other competing approaches for detecting similar types of bugs

# Address Sanitizer (ASan)

- Address Sanitizer (ASan) can be used to detect numerous errors related to memory addressing, such as:
  - out of bounds accesses to heap, stack, and globals
  - heap use after free
  - stack use after return
  - stack use after scope
  - double or invalid free
  - memory leaks
  - initialization order problems
- supported by both Clang and GCC
- compiler instruments all loads/stores and inserts redzones around stack and global variables
- run-time library provides malloc replacement (with redzone and quarantine functionality) and bookkeeping for error messages
- typically introduces about 2 times slowdown
- about 1.5 to 3 times memory overhead

# Using Address Sanitizer

- need to enable address sanitizer at compile and link time using `-fsanitize=address` option for Clang and GCC
- environment variable `ASAN_OPTIONS` can be set to whitespace-separated list of options to control some sanitizer behavior at run time
- some sanitizer options include:
  - `strip_path_prefix`
  - `verbosity`
  - `detect_leaks`
  - `allocator_may_return_null`
  - `check_initialization_order`
  - `detect_stack_use_after_return`
  - `new_delete_type_mismatch`
  - `exitcode`
- to enable checking for initialization order problems, use `ASAN_OPTIONS="check_initialization_order=1"`

# Out-of-Bounds Access to Globals

global\_buffer\_overflow.cpp

```
1  #include <iostream>
2  int a[4] = {1, 2, 3, 4};
3  int main() {
4      for (int i = 0; i <= 4; ++i) {
5          std::cout << a[i] << '\n';
6      }
7  }
```

program output (truncated):

```
=====
==3359==ERROR: AddressSanitizer: global-buffer-overflow on address 0
    x0000006020d0 at pc 0x000000400d31 bp 0x7ffeb78d2350 sp 0
    x7ffeb78d2348
READ of size 4 at 0x0000006020d0 thread T0
    #0 0x400d30 in main global_buffer_overflow.cpp:5
    #1 0x7f83da8d4fdf in __libc_start_main (/lib64/libc.so.6+0x1ffdf)
    #2 0x400bf8 (global_buffer_overflow+0x400bf8)
0x0000006020d0 is located 0 bytes to the right of global variable 'a'
    defined in 'global_buffer_overflow.cpp:2:5' (0x6020c0) of size 16
SUMMARY: AddressSanitizer: global-buffer-overflow
    global_buffer_overflow.cpp:5 in main
Shadow bytes around the buggy address:
```



# Out-of-Bounds Access to Stack

stack\_buffer\_overflow.cpp

```
1  #include <iostream>
2  int main() {
3      int a[4] = {1, 2, 3, 4};
4      for (int i = 0; i <= 4; ++i)
5          {std::cout << a[i] << '\n';}
6  }
```

program output (truncated):

```
=====
==3364==ERROR: AddressSanitizer: stack-buffer-overflow on address 0x7ffc3e811cf0
    at pc 0x000000400e53 bp 0x7ffc3e811c70 sp 0x7ffc3e811c68
READ of size 4 at 0x7ffc3e811cf0 thread T0
    #0 0x400e52 in main stack_buffer_overflow.cpp:5
    #1 0x7f10c7a7df in __libc_start_main (/lib64/libc.so.6+0x1ffdf)
    #2 0x400c48 (stack_buffer_overflow+0x400c48)
Address 0x7ffc3e811cf0 is located in stack of thread T0 at offset 112 in frame
    #0 0x400d06 in main stack_buffer_overflow.cpp:2
    This frame has 2 object(s):
        [32, 33) 'c'
        [96, 112) 'a' <== Memory access at offset 112 overflows this variable
HINT: this may be a false positive if your program uses some custom stack unwind
      mechanism or swapcontext
      (longjmp and C++ exceptions *are* supported)
SUMMARY: AddressSanitizer: stack-buffer-overflow stack_buffer_overflow.cpp:5 in
      main
Shadow bytes around the buggy address:
```

# Out-of-Bounds Access to Heap

heap\_buffer\_overflow.cpp

```
1  #include <iostream>
2  #include <cstring>
3  int main() {
4      char* p = new char[5];
5      std::strcpy(p, "Hello");
6      std::cout << p << '\n';
7      delete[] p;
8  }
```

program output (truncated):

```
=====
==3360==ERROR: AddressSanitizer: heap-buffer-overflow on address 0x602000000015
    at pc 0x7f7497932399 bp 0x7ffd8defc240 sp 0x7ffd8defb9f0
WRITE of size 6 at 0x602000000015 thread T0
    #0 0x7f7497932398 in __interceptor_memcpy ../../../../src/libsanitizer/asan/
        asan_interceptors.cc:456
    #1 0x400dd4 in main heap_buffer_overflow.cpp:5
    #2 0x7f7496c7dfdf in __libc_start_main (/lib64/libc.so.6+0x1ffdf)
    #3 0x400ca8 (heap_buffer_overflow+0x400ca8)
0x602000000015 is located 0 bytes to the right of 5-byte region [0x602000000010,0
    x602000000015)
allocated by thread T0 here:
    #0 0x7f7497997170 in operator new[](unsigned long) ../../../../src/
        libsanitizer/asan/asan_new_delete.cc:82
    #1 0x400dbf in main heap_buffer_overflow.cpp:4
SUMMARY: AddressSanitizer: heap-buffer-overflow ../../../../src/libsanitizer/asan
    /asan_interceptors.cc:456 in __interceptor_memcpy
Shadow bytes around the buggy address:
```

# Use After Free

use\_after\_free.cpp

```
1  int main() {  
2      int* p = new int[16];  
3      delete[] p;  
4      *p = 42;  
5  }
```

program output (truncated):

```
=====  
==3366==ERROR: AddressSanitizer: heap-use-after-free on address 0x606000000020 at  
    pc 0x000000400836 bp 0x7ffc752b5c20 sp 0x7ffc752b5c18  
WRITE of size 4 at 0x606000000020 thread T0  
    #0 0x400835 in main use_after_free.cpp:4  
    #1 0x7f0b6dab5fdf in __libc_start_main (/lib64/libc.so.6+0x1ffdf)  
    #2 0x400738 (use_after_free+0x400738)  
0x606000000020 is located 0 bytes inside of 64-byte region [0x606000000020,0  
    x606000000060)  
freed by thread T0 here:  
    #0 0x7f0b6e7cfe70 in operator delete[](void*) ../../../../src/libsanitizer/  
    asan/asan_new_delete.cc:128  
    #1 0x400801 in main use_after_free.cpp:3  
previously allocated by thread T0 here:  
    #0 0x7f0b6e7cf170 in operator new[](unsigned long) ../../../../src/  
    libsanitizer/asan/asan_new_delete.cc:82  
    #1 0x4007f1 in main use_after_free.cpp:2  
SUMMARY: AddressSanitizer: heap-use-after-free use_after_free.cpp:4 in main  
Shadow bytes around the buggy address:
```

# Stack Use After Return

stack\_use\_after\_return.cpp

```
1  int* g = nullptr;
2  void foobar() {int i = 42; g = &i;}
3  int main() {
4      foobar();
5      return *g;
6  }
```

program output (truncated) (with ASAN\_OPTIONS=detect\_stack\_use\_after\_return=1):

```
=====
==3365==ERROR: AddressSanitizer: stack-use-after-return on address 0x7f74e7500020
    at pc 0x000000400a88 bp 0x7ffdbd534e20 sp 0x7ffdbd534e18
READ of size 4 at 0x7f74e7500020 thread T0
    #0 0x400a87 in main stack_use_after_return.cpp:5
    #1 0x7f74eblddfdf in __libc_start_main (/lib64/libc.so.6+0x1ffdf)
    #2 0x4008b8 (stack_use_after_return+0x4008b8)
Address 0x7f74e7500020 is located in stack of thread T0 at offset 32 in frame
    #0 0x400976 in foobar() stack_use_after_return.cpp:2
    This frame has 1 object(s):
    [32, 36) 'i' <== Memory access at offset 32 is inside this variable
HINT: this may be a false positive if your program uses some custom stack unwind
      mechanism or swapcontext
      (longjmp and C++ exceptions *are* supported)
SUMMARY: AddressSanitizer: stack-use-after-return stack_use_after_return.cpp:5 in
      main
Shadow bytes around the buggy address:
```

# Stack Use After Scope

use\_after\_scope.cpp

```
1  #include <iostream>
2  int main() {
3      int* p;
4      {int x = 0; p = &x;}
5      std::cout << *p << '\n';
6  }
```

program output (truncated):

```
=====
==3367==ERROR: AddressSanitizer: stack-use-after-scope on address 0x7ffeld6a6c40
    at pc 0x000000400b6b bp 0x7ffeld6a6c10 sp 0x7ffeld6a6c08
READ of size 4 at 0x7ffeld6a6c40 thread T0
    #0 0x400b6a in main use_after_scope.cpp:5
    #1 0x7fea2e596fdf in __libc_start_main (/lib64/libc.so.6+0x1ffdf)
    #2 0x400a58 (use_after_scope+0x400a58)
Address 0x7ffeld6a6c40 is located in stack of thread T0 at offset 32 in frame
    #0 0x400b16 in main use_after_scope.cpp:2
This frame has 1 object(s):
    [32, 36) 'x' <== Memory access at offset 32 is inside this variable
HINT: this may be a false positive if your program uses some custom stack unwind
      mechanism or swapcontext
      (longjmp and C++ exceptions *are* supported)
SUMMARY: AddressSanitizer: stack-use-after-scope use_after_scope.cpp:5 in main
Shadow bytes around the buggy address:
```

# Double Free

## double\_free.cpp

```
1  int main() {  
2      int* p = new int[16];  
3      delete[] p;  
4      delete[] p;  
5  }
```

## program output (truncated):

```
=====  
==3358==ERROR: AddressSanitizer: attempting double-free on 0x606000000020 in  
thread T0:  
#0 0x7fdc05ed8e70 in operator delete[](void*) ../../../../src/libsanitizer/  
asan/asan_new_delete.cc:128  
#1 0x4007b9 in main double_free.cpp:4  
#2 0x7fdc051befdf in __libc_start_main (/lib64/libc.so.6+0x1ffdf)  
#3 0x4006e8 (double_free+0x4006e8)  
0x606000000020 is located 0 bytes inside of 64-byte region [0x606000000020,0  
x606000000060)  
freed by thread T0 here:  
#0 0x7fdc05ed8e70 in operator delete[](void*) ../../../../src/libsanitizer/  
asan/asan_new_delete.cc:128  
#1 0x4007b1 in main double_free.cpp:3  
previously allocated by thread T0 here:  
#0 0x7fdc05ed8170 in operator new[](unsigned long) ../../../../src/  
libsanitizer/asan/asan_new_delete.cc:82  
#1 0x4007a1 in main double_free.cpp:2  
SUMMARY: AddressSanitizer: double-free ../../../../src/libsanitizer/asan/  
asan_new_delete.cc:128 in operator delete[](void*)
```

# Memory Leaks

memory\_leak.cpp

```
1  #include <iostream>
2  #include <cstring>
3  int main() {
4      char* p = new char[1024];
5      std::strcpy(p, "Hello, World!\n");
6      std::cout << p;
7  }
```

program output (truncated):

```
Hello, World!
=====
==3362==ERROR: LeakSanitizer: detected memory leaks
Direct leak of 1024 byte(s) in 1 object(s) allocated from:
    #0 0x7f7413651170 in operator new[](unsigned long) ../../../../src/
        libsanitizer/asan/asan_new_delete.cc:82
    #1 0x400b51 in main memory_leak.cpp:4
SUMMARY: AddressSanitizer: 1024 byte(s) leaked in 1 allocation(s).
```

# Initialization Order Problems

## init\_order\_main.cpp

```
1  #include <iostream>
2  extern int B;
3  int A = B;
4  int main()
5  {std::cout << A << '\n';}
```

## init\_order\_other.cpp

```
1  #include <cstdlib>
2  int B = std::atoi("42");
```

## program output (truncated) (with ASAN\_OPTIONS=check\_initialization\_order=1):

```
=====
==3361==ERROR: AddressSanitizer: initialization-order-fiasco on address 0
x000000602440 at pc 0x000000400f14 bp 0x7fff92151540 sp 0x7fff92151538
READ of size 4 at 0x000000602440 thread T0
#0 0x400f13 in __static_initialization_and_destruction_0 init_order_main.cpp
:3
#1 0x400f13 in _GLOBAL__sub_I_A init_order_main.cpp:5
#2 0x40103c in __libc_csu_init (init_order+0x40103c)
#3 0x7f933e2e7f6e in __libc_start_main (/lib64/libc.so.6+0x1ff6e)
#4 0x400c98 (init_order+0x400c98)
0x000000602440 is located 0 bytes inside of global variable 'B' defined in '
init_order_other.cpp:2:5' (0x602440) of size 4
registered at:
#0 0x7f933ef5b7c8 in __asan_register_globals ../../../../src/libsanitizer/
asan/asan_globals.cc:317
#1 0x400fe9 in _GLOBAL__sub_I_00099_1_B (init_order+0x400fe9)
SUMMARY: AddressSanitizer: initialization-order-fiasco init_order_main.cpp:3 in
__static_initialization_and_destruction_0
Shadow bytes around the buggy address:
```



# Thread Sanitizer (TSan)

- Thread Sanitizer (TSan) detects data races and deadlocks
- supported by Clang and GCC
- compiler instruments code to intercept all loads/stores
- run-time library provides malloc replacement, intercepts all synchronization, and handles loads/stores
- does not instrument prebuilt libraries and inline assembly
- about 4 to 10 times slower
- about 5 to 8 times more memory
- only supported on 64-bit Linux

# Using Thread Sanitizer

- need to enable sanitizer at compile and link time using `-fsanitize=thread` option for Clang and GCC
- environment variable `TSAN_OPTIONS` can be set to whitespace-separated list of options to control some sanitizer behavior at run time
- some sanitizer options include:
  - `strip_path_prefix`
  - `verbosity`
  - `report_bugs`
  - `history_size`
  - `suppressions`
  - `exitcode`
- for example, to set per-thread history size value to 7, use `TSAN_OPTIONS="history_size=7"`
- at least some versions of TSan do not detect potential deadlock if it actually happens (although arguably if deadlock happens, probably it will be noticed)

# Data Race

## data\_race.cpp

```
1  #include <thread>
2  int x = 0;
3  int main() {
4      std::thread t([&]{x = 42;});
5      x = 43;
6      t.join();
7  }
```

## program output (truncated):

```
=====
WARNING: ThreadSanitizer: data race (pid=10305)
Write of size 4 at 0x000001553848 by thread T1:
#0 main::~$0::operator()() const data_race.cpp:4:22 (data_race+0x4bfc81)
#1 void std::_Bind_simple<main::~$0 ()>::_M_invoke<(std::_Index_tuple<>) /usr/lib/gcc/x86_64-redhat-
linux/4.9.2/../../../../include/c++/4.9.2/functional:1699:18 (data_race+0x4bfbf8)
#2 std::_Bind_simple<main::~$0 ()>::operator()() /usr/lib/gcc/x86_64-redhat-linux/4.9.2/../../../../
include/c++/4.9.2/functional:1688:16 (data_race+0x4bfb98)
#3 std::thread::_Impl<std::_Bind_simple<main::~$0 ()> >::_M_run() /usr/lib/gcc/x86_64-redhat-linux
/4.9.2/../../../../include/c++/4.9.2/thread:115:13 (data_race+0x4bf94c)
#4 execute_native_thread_routine_compat /gcc-7.1.0/build/x86_64-pc-linux-gnu/libstdc++-v3/src/c
++11/../../../../src/libstdc++-v3/src/c++11/thread.cc:110 (libstdc++.so.6+0xba46f)
Previous write of size 4 at 0x000001553848 by main thread:
#0 main data_race.cpp:5:4 (data_race+0x4be2ce)
Location is global 'x' of size 4 at 0x000001553848 (data_race+0x000001553848)
Thread T1 (tid=10310, running) created by main thread at:
#0 pthread_create /llvm-clang-4.0.0/src/projects/compiler-rt/lib/tsan/rtl/tsan_interceptors.cc:897 (
data_race+0x44f89b)
#1 __gthread_create /gcc-7.1.0/build/x86_64-pc-linux-gnu/libstdc++-v3/include/x86_64-pc-linux-gnu/bits/
gthr-default.h:662 (libstdc++.so.6+0xba5b2)
#2 std::thread::_M_start_thread(std::shared_ptr<std::thread::_Impl_base>, void (*)()) /gcc-7.1.0/build/
x86_64-pc-linux-gnu/libstdc++-v3/src/c++11/../../../../src/libstdc++-v3/src/c++11/thread.cc:191
(libstdc++.so.6+0xba5b2)
#3 main data_race.cpp:4:14 (data_race+0x4be2bd)
SUMMARY: ThreadSanitizer: data race data_race.cpp:4:22 in main::~$0::operator()() const
```

# Deadlock

deadlock.cpp

```
1  #include <iostream>
2  #include <thread>
3  #include <mutex>
4  std::mutex m0;
5  std::mutex m1;
6  void func1(int n) {
7      for (auto i = n; i > 0; --i) {
8          std::scoped_lock<std::mutex> l0(m0);
9          std::scoped_lock<std::mutex> l1(m1);
10         std::cout << "a\n";
11     }
12 }
13 void func2(int n) {
14     for (auto i = n; i > 0; --i) {
15         std::scoped_lock<std::mutex> l1(m1);
16         std::scoped_lock<std::mutex> l0(m0);
17         std::cout << "b\n";
18     }
19 }
20 int main() {
21     std::thread t1([]{func1(10'000);});
22     std::thread t2([]{func2(10'000);});
23     t1.join(); t2.join();
24 }
```

# Deadlock (Continued)

## program output (truncated):

```
=====
WARNING: ThreadSanitizer: lock-order-inversion (potential deadlock) (pid=29188)
  Cycle in lock order graph: M9 (0x000000404240) => M10 (0x000000404200) => M9
  Mutex M10 acquired here while holding mutex M9 in thread T1:
    #0 pthread_mutex_lock ../../../../src/libsanitizer/sanitizer_common/sanitizer_common_interceptors.inc
      :3799 (libtsan.so.0+0x407ab)
    #1 __gthread_mutex_lock /include/c++/8.2.0/x86_64-pc-linux-gnu/bits/gthr-default.h:748 (deadlock+0x4016e6)
    #2 std::mutex::lock() /include/c++/8.2.0/bits/std_mutex.h:103 (deadlock+0x4016e6)
    #3 std::scoped_lock<std::mutex>::scoped_lock(std::mutex&) /include/c++/8.2.0/mutex:610 (deadlock+0x4016e6)
    #4 func1(int) deadlock.cpp:9 (deadlock+0x4016e6)
  [text deleted]
  Hint: use TSAN_OPTIONS=second_deadlock_stack=1 to get more informative warning message
  Mutex M9 acquired here while holding mutex M10 in thread T2:
    #0 pthread_mutex_lock ../../../../src/libsanitizer/sanitizer_common/sanitizer_common_interceptors.inc
      :3799 (libtsan.so.0+0x407ab)
    #1 __gthread_mutex_lock /include/c++/8.2.0/x86_64-pc-linux-gnu/bits/gthr-default.h:748 (deadlock+0x401786)
    #2 std::mutex::lock() /include/c++/8.2.0/bits/std_mutex.h:103 (deadlock+0x401786)
    #3 std::scoped_lock<std::mutex>::scoped_lock(std::mutex&) /include/c++/8.2.0/mutex:610 (deadlock+0x401786)
    #4 func2(int) deadlock.cpp:16 (deadlock+0x401786)
  [text deleted]
SUMMARY: ThreadSanitizer: lock-order-inversion (potential deadlock) /include/c++/8.2.0/x86_64-pc-linux-gnu/
  bits/gthr-default.h:748 in __gthread_mutex_lock
=====
a
b
ThreadSanitizer: reported 1 warnings
```

# Memory Sanitizer (MSan)

- Memory Sanitizer (MSan) detects reads from uninitialized memory
- in contrast, ASan cannot detect uninitialized reads
- currently, MSan only supported by Clang (not GCC)
- compiler instruments all loads/stores
- uses bit to bit shadow mapping
- if not all code instrumented (so that not all stores are observed), false positives can result
- about 3 to 6 times slowdown
- about 2 to 3 times memory overhead

# Using Memory Sanitizer

- need to enable sanitizer at compile and link time using `-fsanitize=memory` option for Clang
- environment variable `MSAN_OPTIONS` can be set to whitespace-separated list of options to control some sanitizer behavior at run time
- some sanitizer options include:
  - `strip_path_prefix`
  - `verbosity`
- for example, to set verbosity level to 2, use `MSAN_OPTIONS="verbosity=2"`

# Read From Uninitialized Memory

uninitialized\_1.cpp

```
1  int main(int argc, char** argv) {  
2      int x[2];  
3      x[0] = 1;  
4      if (x[argc % 2]) {  
5          return 1;  
6      }  
7  }
```

program output (truncated):

```
==22595==WARNING: MemorySanitizer: use-of-uninitialized-value  
#0 0x4a46c3 in main uninitialized_1.cpp:4:6  
#1 0x7f5d3908ffdf in __libc_start_main (/lib64/libc.so.6+0x1ffdf)  
#2 0x41a77e in _start (uninitialized_1+0x41a77e)  
SUMMARY: MemorySanitizer: use-of-uninitialized-value uninitialized_1.  
      cpp:4:6 in main  
Exiting
```



# Undefined-Behavior Sanitizer (UBSan)

- Undefined-Behavior Sanitizer (UBSan) detects code that results in various types of undefined behavior
- some types of problems detected include:
  - using misaligned or null pointer
  - signed integer overflow
  - conversion to, from, or between floating-point types which would overflow destination
  - reaching end of value-returning function with returning value
  - out of bounds array indexing where array bound can be statically determined
- compiler instruments code with extra checks
- supported by Clang and GCC
- slowdown varies between 0% and 50%

# Using Undefined-Behavior Sanitizer

- need to enable sanitizer at compile and link time using `-fsanitize=undefined` option for Clang and GCC
- environment variable `UBSAN_OPTIONS` can be set to whitespace-separated list of options to control some sanitizer behavior at run time
- some sanitizer options include:
  - `suppressions`
  - `strip_path_prefix`
  - `verbosity`

# Signed Integer Overflow

signed\_integer\_overflow.cpp

```
1  #include <iostream>
2  #include <limits>
3  int main() {
4      int x = std::numeric_limits<int>::max();
5      int y = x + 1;
6      std::cout << y << '\n';
7  }
```

program output:

```
signed_integer_overflow.cpp:5:14: runtime error: signed integer
      overflow: 2147483647 + 1 cannot be represented in type 'int'
-2147483648
```

# Invalid Shift

invalid\_shift.cpp

```
1  #include <iostream>
2  int main() {
3      int x = 32678;
4      int y = 1 << x;
5      std::cout << y << '\n';
6  }
```

program output:

```
invalid_shift.cpp:4:12: runtime error: shift exponent 32678 is too
    large for 32-bit type 'int'
0
```

# Leak Sanitizer (LSan)

- Leak Sanitizer (LSan) detects memory leaks
- supported by Clang and GCC
- adds almost no performance overhead until end of program, at which point extra leak-detection checks performed
- need to enable sanitizer at compile and link time using `-fsanitize=leak` option for Clang and GCC (or by using ASan, which includes LSan functionality)
- environment variable `LSAN_OPTIONS` can be set to whitespace-separated list of options to control some sanitizer behavior at run time
- some sanitizer options include:
  - `strip_path_prefix`
  - `verbosity`

# Memory Leak

heap\_buffer\_overflow.cpp

```
1  #include <iostream>
2  #include <cstring>
3  int main() {
4      char* p = new char[1024];
5      std::strcpy(p, "Hello, World!\n");
6      std::cout << p;
7  }
```

program output:

```
Hello, World!
=====
==10786==ERROR: LeakSanitizer: detected memory leaks
Direct leak of 1024 byte(s) in 1 object(s) allocated from:
    #0 0x7faa5e0a7436 in operator new[](unsigned long) ../../../../src/
        libsanitizer/lsan/lsan_interceptors.cc:164
    #1 0x400894 in main memory_leak.cpp:4
SUMMARY: LeakSanitizer: 1024 byte(s) leaked in 1 allocation(s).
```

## Section 7.2.1

### **References**

- 1 Kostya Serebryany. Sanitize Your C++ Code. CppCon, 2014. Available online at [https://youtu.be/V2\\_80g0eOMc](https://youtu.be/V2_80g0eOMc).
- 2 Kostya Serebryany. Beyond Sanitizers. CppCon, 2015. Available online at <https://youtu.be/qTkYDA0En6U>.



## Section 7.3

# Clang Static Analyzer

# Clang Static Analyzer

- Clang Static Analyzer is static analysis tool for C/C++, which is part of Clang
- supports many checks (e.g., new-delete mismatch, assigning uninitialized values, and so on)
- interface provided mainly through two programs:
  - 1 `scan-build`, which is used to invoke static analyzer when building code
  - 2 `scan-view`, which is used to view results from static analyzer
- can be used with most build processes (e.g., CMake, Make, or direct compiler invocation)
- incurs cost of processing code with static analyzer in addition to compilation
- static analysis can be much slower than compilation since detailed analysis of code can incur significant computational cost
- static analyzer can sometimes yield false positives
- web site: <https://clang-analyzer.llvm.org>

# Some Supported Checks

- uninitialized arguments
- dereferencing null pointers
- division by zero
- address of stack memory that escapes function
- undefined result of binary operator
- uninitialized array subscript
- assigning uninitialized values
- uninitialized branch condition
- blocks that capture uninitialized values
- uninitialized value being returned from function
- new-delete mismatch

# scan-build Program

- scan-build program allows user to run static analyzer over codebase as part of performing regular build
- during project build, as source files compiled, also processed by static analyzer
- upon completion of build, results can be viewed in web browser
- scan-build is used to invoke command that builds code (typically CMake or Make)
- only files that are compiled are analyzed
- command line has following form:

```
scan-build [options] command [command_options]
```

- some commonly-used options include:

Option	Description
-h	print help information and exit
-v	increase verbosity of output
-o <i>target_dir</i>	set output directory to <i>target_dir</i>
-V	view analysis results in web browser
--force-analyze-debug-code	force assertions to be enabled

- enabling assertions can help to reduce false positives
- so advisable to either analyze project in its debug configuration or use `--force-analyze-debug-code` **flag of scan-build** to force assertions to be enabled
- can use verbose output to debug `scan-build`

# Mechanics of scan-build Program

- `scan-build` sets `CXX` and `CC` environment variables to Clang Static Analyzer program and then invokes user-specified command
- when Clang Static Analyzer invoked, performs static analysis and then invokes real compiler to compile code
- internally, uses `CCC_CXX` and `CCC_CC` environment variables (set by `scan-build`) to determine real compiler programs to use
- can enable verbosity of static analyzer with environment variable `CCC_ANALYZER_VERBOSE`

# Using scan-build With CMake

- must run CMake configure operation and build operation with `scan-build`
- all should work fine as long as `CMAKE_CXX_COMPILER` not set by CMakeLists file or on command line
- for example:

```
scan-build cmake -S$SOURCE_DIR -B$BINARY_DIR
scan-build cmake --build $BINARY_DIR
```

# Uninitialized Value: Code

```
1  #include <iostream>
2
3  void func(int i) {
4      for (int j; j < i; ++j) {
5          std::cout << "hello\n";
6      }
7  }
```

- in above code, variable `j` is not initialized



# Uninitialized Value: Static Analyzer Output

[Summary](#) > Report 1dae28

## Bug Summary

File: uninitialized\_value.cpp

Warning: [line 4, column 16](#)

The left operand of '<' is a garbage value

[Report Bug](#)

## Annotated Source Code

[?]

```
1  #include <iostream>
2
3  void func(int i) {
4      for (int j; j < i; ++j) {
5
6      }
7  }
```

1 'j' declared without an initial value →

2 ← The left operand of '<' is a garbage value

```
std::cout << "hello\n";
```

# New-Delete Mismatch: Code

```
1 void func() {  
2     int *ip = new int[1024];  
3     // ...  
4     delete ip;  
5 }
```

- in above code, array new not paired with array delete

# New-Delete Mismatch: Static Analyzer Output

[Summary](#) > Report 131b66

## Bug Summary

**File:** new\_delete\_mismatch.cpp

**Warning:** [line 4, column 2](#)

Memory allocated by 'new[]' should be deallocated by 'delete[]', not 'delete'

[Report Bug](#)

## Annotated Source Code

[?]

```
1 void func() {  
2     int *ip = new int[1024];
```

1 Memory is allocated →

```
3     // ...  
4     delete ip;
```

2 ← Memory allocated by 'new[]' should be deallocated by 'delete[]', not 'delete'

```
5 }
```

## Section 7.3.1

### **References**

- 1 Gabor Horvath. Make Friends with the Clang Static Analysis Tools. CppCon, Bellevue, WA, USA, Sept. 18–23, 2016. Available online at <https://youtu.be/AQF6hjLKsnM>.

## Section 7.4

### Clang-Tidy

- Clang-Tidy is static analysis tool for C/C++, which is part of Clang
- supports many checks, which consider such things as:
  - correctness
  - efficiency
  - readability
  - modern style
- can automatically fix code in many cases
- by default, only small subset of checks enabled
- probably not advisable to enable all checks, since many benign warnings may result, obscuring warnings that indicate serious problems
- web site: <https://clang.llvm.org/extra/clang-tidy>

# The clang-tidy Command

- command line has following form:

```
clang-tidy [options] [$source_file]...
```

- some options include:

Option	Description
-checks= <i>string</i>	specify check to include/exclude
-p <i>build_path</i>	set build path to <i>build_path</i>
-version	print version information and exit
-help	print help information and exit
-list-checks	list all enabled checks and exit
-fix	apply suggested fixes
-fix-errors	apply suggested fixes even if compilation errors found
-warnings-as-errors= <i>string</i>	treat specified warnings as errors



# Division By Zero

divide\_by\_zero.cpp

```
1  int func(int x) {  
2      if (!x) {  
3          return 1024 / x;  
4      } else {  
5          return x;  
6      }  
7  }
```

clang-tidy output:

```
divide_by_zero.cpp:3:15: warning: Division by zero [clang-analyzer-core  
    .DivideZero]  
        return 1024 / x;  
                      ^  
divide_by_zero.cpp:2:6: note: Assuming 'x' is 0  
    if (!x) {  
        ^  
divide_by_zero.cpp:2:2: note: Taking true branch  
    if (!x) {  
        ^  
divide_by_zero.cpp:3:15: note: Division by zero  
        return 1024 / x;  
                      ^
```

# New-Delete Mismatch

new\_delete\_mismatch.cpp

```
1  int main() {  
2      char* p = new char[1024];  
3      delete p;  
4  }
```

clang-tidy output:

```
new_delete_mismatch.cpp:3:2: warning: 'delete' applied to a pointer that was  
    allocated with 'new[]'; did you mean 'delete[]'? [clang-diagnostic-  
    mismatched-new-delete]  
    delete p;  
    ^  
    []  
new_delete_mismatch.cpp:2:12: note: allocated with 'new[]' here  
    char* p = new char[1024];  
              ^  
new_delete_mismatch.cpp:3:2: warning: Memory allocated by 'new[]' should be  
    deallocated by 'delete[]', not 'delete' [clang-analyzer-unix.  
    MismatchedDeallocator]  
    delete p;  
    ^  
new_delete_mismatch.cpp:2:12: note: Memory is allocated  
    char* p = new char[1024];  
              ^  
new_delete_mismatch.cpp:3:2: note: Memory allocated by 'new[]' should be  
    deallocated by 'delete[]', not 'delete'  
    delete p;  
    ^
```

# Missing Return Statement

no\_return.cpp

```
1  int func(int x) {  
2      if (x >= 0) {  
3          return 1;  
4      }  
5  }
```

clang-tidy output:

```
no_return.cpp:5:1: warning: control may reach end of non-void function  
    [clang-diagnostic-return-type]  
}  
^
```

# Stack Address Escapes Function

stack\_address\_escape.cpp

```
1  int* p;  
2  
3  void test() {  
4      int x = 42;  
5      p = &x;  
6  }
```

clang-tidy output:

```
stack_address_escape.cpp:6:1: warning: Address of stack memory  
    associated with local variable 'x' is still referred to by the  
    global variable 'p' upon returning to the caller. This will be a  
    dangling reference [clang-analyzer-core.StackAddressEscape]  
}  
^  
  
stack_address_escape.cpp:6:1: note: Address of stack memory associated  
    with local variable 'x' is still referred to by the global variable  
    'p' upon returning to the caller. This will be a dangling  
    reference  
}  
^
```

# Undefined Operand

undefined\_operand.cpp

```
1  int test() {  
2      int x;  
3      return x + 1;  
4  }
```

clang-tidy output:

```
undefined_operand.cpp:3:11: warning: The left operand of '+' is a  
    garbage value [clang-analyzer-core.UndefinedBinaryOperatorResult]  
    return x + 1;  
           ^  
undefined_operand.cpp:2:2: note: 'x' declared without an initial value  
    int x;  
    ^  
undefined_operand.cpp:3:11: note: The left operand of '+' is a garbage  
    value  
    return x + 1;  
           ^
```

# Clang Tidy With CMake

- Clang Tidy relatively easy to use with CMake
- CMake provides `CXX_CLANG_TIDY` property for targets that can be set to command to invoke for performing code checking with Clang Tidy
- consider example to illustrate use of Clang Tidy with CMake
- want each source file to be processed by Clang Tidy in addition to being compiled
- project consists of three files:
  - 1 `CMakeLists.txt`
  - 2 `app.cpp`
  - 3 `lib.cpp`

# Clang Tidy CMake Example: CMakeLists File

## CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.6 FATAL_ERROR)
2  project(clang_tidy LANGUAGES CXX)
3
4  find_program(CLANG_TIDY_PROGRAM NAMES "clang-tidy"
5    DOC "Path to clang-tidy executable")
6  if (CLANG_TIDY_PROGRAM)
7    set(CLANG_TIDY_OPTIONS "-warnings-as-errors=*")
8    set(RUN_CLANG_TIDY "${CLANG_TIDY_PROGRAM}" "${CLANG_TIDY_OPTIONS}")
9  endif()
10
11  add_library(lib lib.cpp)
12  add_executable(app app.cpp lib)
13  set(targets lib app)
14
15  if (CLANG_TIDY_PROGRAM)
16    set_target_properties(${targets} PROPERTIES CXX_CLANG_TIDY
17      "${RUN_CLANG_TIDY}")
18  endif()
```

# Clang Tidy CMake Example: Source Code

lib.cpp

```
1  int func() {}  
2  
3  char* foobar() {  
4      char c;  
5      return &c;  
6  }
```

app.cpp

```
1  int func();  
2  char* foobar();  
3  
4  int main() {  
5      auto x = 1 / 2;  
6      double y = 1 / x;  
7      char* cp = new char[1024];  
8      delete cp;  
9      func();  
10     foobar();  
11     int i;  
12     return i + 1;  
13 }
```



## Section 7.4.1

### **References**

- 1 Daniel Jasper. Keep Your Code Sane With Clang Tidy. Meeting C++, Berlin, Germany, Dec. 4–5, 2015. Available online at <https://youtu.be/nzCLcfH3pb0>.

## Section 7.5

### Valgrind

- Valgrind is software framework that provides number of tools for dynamic analysis
- tools provided by Valgrind include:
  - **Memcheck** (memory error detector)
  - Cachegrind (cache and branch-prediction profiler)
  - **Callgrind** (call-graph generating cache and branch-prediction profiler)
  - **Helgrind** (thread error detector)
  - DRD (thread error detector)
  - **Massif** (heap profiler)
  - DHAT (dynamic heap analysis tool)
- above tools can be used to:
  - detect many memory management and threading bugs
  - profile programs in detail
- access to various Valgrind tools provided via `valgrind` program
- `valkyrie` program provides GUI for Memcheck and Helgrind tools
- Valgrind software is open source
- web site: <http://www.valgrind.org>

- Valgrind works by simulating processor in software
- Valgrind does not fully support IEEE 754 (floating-point arithmetic standard)
- consequently, code that uses floating-point arithmetic in particular ways may not behave correctly
- do not use Valgrind with ASan, TSan, or MSan code sanitizers

# Support for Floating-Point Arithmetic in Valgrind

- unfortunately, Valgrind does not fully support IEEE 754 floating-point standard (at time of this writing, at least)
- lacks support for floating-point exceptions
- lacks full support for floating-point rounding modes (e.g., some instructions always use round to nearest)
- consequently, code that relies on control over rounding mode will not behave correctly (e.g., code using interval arithmetic)
- code using CGAL library, for example, is often problematic, due to heavy use of interval arithmetic (which needs rounding-mode control) for efficient exact geometric predicates
- does not support extended floating-point formats used by some architectures
- consequently, can sometimes obtain less accurate results from floating-point arithmetic (which in extreme cases might cause numerical instability)

- Helgrind is Valgrind tool for detecting synchronization errors in programs that use POSIX pthreads threading primitives
- Helgrind can detect:
  - misuses of POSIX pthreads API (e.g., thread exits while holding locks, recursively locking non-recursive mutex, and so on)
  - potential deadlocks arising from lock ordering problems
  - data races (i.e., accessing memory without adequate locking or synchronization)
- run-time overhead of Helgrind can be quite significant (e.g., slowdown of 100 times not unusual)
- to run program *program* using Helgrind, use command like:  
`valgrind --tool=helgrind program`
- Helgrind described in Valgrind documentation:
  - <http://valgrind.org/docs/manual/hg-manual.html>

# Data Race Example

data\_race\_1\_0.cpp

```
1  #include <thread>
2  #include <iostream>
3
4  unsigned long count = 0;
5
6  void func() {
7      for (unsigned long i = 0; i < 1'000'000; ++i) {
8          ++count;
9      }
10 }
11
12 int main() {
13     std::thread t1(func);
14     std::thread t2(func);
15     t1.join();
16     t2.join();
17     std::cout << count << '\n';
18 }
```



# Helgrind Output for Data Race Example

output for command: `valgrind --tool=helgrind ./data_race_1_0`

```
[text deleted]
==5517== Thread #3 was created
[text deleted]
==5517==   by 0x400D1E: main (data_race_1_0.cpp:14)
[text deleted]
==5517== Thread #2 was created
[text deleted]
==5517==   by 0x400D0D: main (data_race_1_0.cpp:13)
[text deleted]
==5517== Possible data race during read of size 8 at 0x603250 by thread #3
==5517== Locks held: none
==5517==   at 0x400CD8: func() (data_race_1_0.cpp:8)
[text deleted]
==5517== This conflicts with a previous write of size 8 by thread #2
==5517== Locks held: none
==5517==   at 0x400CE3: func() (data_race_1_0.cpp:8)
[text deleted]
==5517== Address 0x603250 is 0 bytes inside data symbol "count"
[text deleted]
==5517== Possible data race during write of size 8 at 0x603250 by thread #3
==5517== Locks held: none
==5517==   at 0x400CE3: func() (data_race_1_0.cpp:8)
[text deleted]
==5517==
==5517== This conflicts with a previous write of size 8 by thread #2
==5517== Locks held: none
==5517==   at 0x400CE3: func() (data_race_1_0.cpp:8)
[text deleted]
==5517== Address 0x603250 is 0 bytes inside data symbol "count"
[text deleted]
==5517== For counts of detected and suppressed errors, rerun with: -v
==5517== Use --history-level=approx or =none to gain increased speed, at
==5517== the cost of reduced accuracy of conflicting-access information
==5517== ERROR SUMMARY: 2 errors from 2 contexts (suppressed: 0 from 0)
```

- Memcheck is Valgrind tool for memory error detection
- can detect certain types of invalid memory accesses:
  - overrunning/underrunning heap blocks
  - overrunning top of stack
  - accessing memory after being freed
- can detect certain other types of memory-related problems, such as:
  - using uninitialized variables
  - incorrect freeing of heap memory (e.g., double freeing heap blocks and mismatched use of `malloc/new/new[]` and `free/delete/delete[]`)
  - overlapping source and destination buffers in `memcpy` and other related functions
  - passing suspicious (e.g., excessively large) value to size parameter of memory allocation function
  - memory leaks
- web page: <http://valgrind.org/docs/manual/mc-manual.html>

# The valgrind Program With Memcheck

- command line interface has following form:  
`valgrind [options] program [program_options]`
- to use Memcheck tool, must specify `--tool=memcheck` option
- some Memcheck-specific options include:

Option	Description
<code>--leak-check=<i>mode</i></code>	specifies mode for checking for memory leaks, where <i>mode</i> is <code>no</code> (disabled), <code>summary</code> (summary information), or <code>yes/full</code> (detailed information)

# Example: Source Code

## memory\_leak.cpp

```
1  int main() {  
2      char* buf = new char[1024];  
3      double* dp = new double;  
4      // ... (no delete or delete[])  
5  }
```

## new\_delete\_mismatch.cpp

```
1  int main() {  
2      char* buf = new char[1024];  
3      // ...  
4      delete buf;  
5  }
```

## uninitialized\_memory.cpp

```
1  #include <iostream>  
2  
3  int main() {  
4      int i;  
5      std::cout << i << '\n';  
6  }
```

# Memcheck Output for memory\_leak Program

command:

```
valgrind --tool=memcheck -q --leak-check=yes ./memory_leak
```

output:

```
==15265== 8 bytes in 1 blocks are definitely lost in loss record 1 of 2
==15265==   at 0x4C2E1FC: operator new(unsigned long) (vg_replace_malloc.c:334)
==15265==   by 0x400611: main (memory_leak.cpp:3)
==15265==
==15265== 1,024 bytes in 1 blocks are definitely lost in loss record 2 of 2
==15265==   at 0x4C2E8E9: operator new[](unsigned long) (vg_replace_malloc.c:423)
==15265==   by 0x400603: main (memory_leak.cpp:2)
==15265==
```

# Memcheck Output for new\_delete\_mismatch Program

## command:

```
valgrind -q --tool=memcheck ./new_delete_mismatch
```

## output:

```
==15267== Mismatched free() / delete / delete []
==15267==    at 0x4C2F21A: operator delete(void*) (vg_replace_malloc.c:576)
==15267==    by 0x400638: main (new_delete_mismatch.cpp:4)
==15267== Address 0x5ab4c80 is 0 bytes inside a block of size 1,024 alloc'd
==15267==    at 0x4C2E8E9: operator new[](unsigned long) (vg_replace_malloc.c:423)
==15267==    by 0x400623: main (new_delete_mismatch.cpp:2)
==15267==
```

# Memcheck Output for uninitialized\_memory Program

command:

```
valgrind -q --tool=memcheck ./uninitialized_memory
```

output (truncated):

```
==15266== Conditional jump or move depends on uninitialised value(s)
==15266== at 0x4F3B8CB: std::ostreambuf_iterator<char, std::char_traits<char> > std::num_put<char, std::
ostreambuf_iterator<char, std::char_traits<char> > >::_M_insert_int<long>(std::ostreambuf_iterator<
char, std::char_traits<char> >, std::ios_base&, char, long) const (locale_facets.tcc:874)
==15266== by 0x4F47234: put (locale_facets.h:2371)
==15266== by 0x4F47234: std::ostream& std::ostream::_M_insert<long>(long) (ostream.tcc:73)
==15266== by 0x400798: main (uninitialized_memory.cpp:5)
==15266==
==15266== Use of uninitialised value of size 8
==15266== at 0x4F3B3DE: int std::_int_to_char<char, unsigned long>(char*, unsigned long, char const*, std
::_Ios_Fmtflags, bool) (locale_facets.tcc:803)
==15266== by 0x4F3B8F4: std::ostreambuf_iterator<char, std::char_traits<char> > std::num_put<char, std::
ostreambuf_iterator<char, std::char_traits<char> > >::_M_insert_int<long>(std::ostreambuf_iterator<
char, std::char_traits<char> >, std::ios_base&, char, long) const (locale_facets.tcc:876)
==15266== by 0x4F47234: put (locale_facets.h:2371)
==15266== by 0x4F47234: std::ostream& std::ostream::_M_insert<long>(long) (ostream.tcc:73)
==15266== by 0x400798: main (uninitialized_memory.cpp:5)
==15266==
==15266== Conditional jump or move depends on uninitialised value(s)
==15266== at 0x4F3B3EB: int std::_int_to_char<char, unsigned long>(char*, unsigned long, char const*, std
::_Ios_Fmtflags, bool) (locale_facets.tcc:806)
==15266== by 0x4F3B8F4: std::ostreambuf_iterator<char, std::char_traits<char> > std::num_put<char, std::
ostreambuf_iterator<char, std::char_traits<char> > >::_M_insert_int<long>(std::ostreambuf_iterator<
char, std::char_traits<char> >, std::ios_base&, char, long) const (locale_facets.tcc:876)
==15266== by 0x4F47234: put (locale_facets.h:2371)
==15266== by 0x4F47234: std::ostream& std::ostream::_M_insert<long>(long) (ostream.tcc:73)
==15266== by 0x400798: main (uninitialized_memory.cpp:5)
==15266==
```

- Massif is Valgrind tool for performing heap profiling
- measures how much heap memory program uses, including both useful space and extra bytes allocated for bookkeeping and alignment
- can optionally measure size of program stack
- can use to determine how memory consumption evolves over lifetime of program as well as peak memory usage
- Massif tool of Valgrind accessed via `valgrind` program
- running program with Massif tool generates output file containing heap profiling information
- output file can be analyzed with various tools, such as `ms_print` and Massif Visualizer
- web page: <http://valgrind.org/docs/manual/ms-manual.html>



# The valgrind Command with Massif

- command line interface has following form:  
`valgrind [options] program [program_options]`
- to use Massif tool, must specify `--tool=massif` option
- some Massif-specific options include:

Option	Description
<code>--massif-out-file=<i>file</i></code>	sets output file to <i>file</i> ; defaults to <code>massif.out-\$pid</code> where <code>\$pid</code> is process ID
<code>--stacks=<i>flag</i></code>	specifies whether stack profiling should be performed, where <i>flag</i> is <code>true</code> or <code>false</code>
<code>--time-unit=<i>unit</i></code>	specifies time unit for profiling, where <i>unit</i> is <code>i</code> (instructions), <code>ms</code> (time), or <code>B</code> (bytes)

# Analyzing Massif Data Files

- Valgrind software includes program called `ms_print` for analyzing Massif data files
- `ms_print` output is text based
- GUI-based tools also available for analyzing Massif data files, such as Massif Visualizer
- Massif Visualizer is accessible via `massif-visualizer` program
- Massif Visualizer web site:

<https://github.com/KDE/massif-visualizer>

# Massif Example

source code for heap\_1 program:

```
1  #include <algorithm>
2  #include <cstdlib>
3  #include <iostream>
4  #include <numeric>
5  #include <vector>
6
7  int main() {
8      std::size_t size = 1;
9      while (size < 1'000'000) {
10         std::vector<float> v;
11         v.reserve(size);
12         std::generate_n(std::back_inserter(v), size,
13             []() -> float {return drand48();});
14         auto sum = std::accumulate(v.begin(), v.end(), 0.0f);
15         std::cout << sum / size << '\n';
16         size <= 1;
17     }
18 }
```

command to invoke Massif tool for heap\_1 program:

```
valgrind --tool=massif --massif-out-file=profile.massif --stacks=yes
./heap_1
```

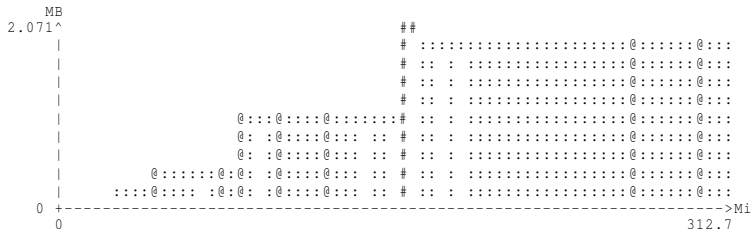
## Massif Example: ms\_print Output

command:

```
ms_print --x=70 --y=10 profile.massif
```

output (truncated):

```
Command: ./heap_1
Massif arguments: --massif-out-file=profile.massif --stacks=yes
ms_print arguments: -x=70 -y=10 profile.massif
```



Number of snapshots: 71

Detailed snapshots: [8, 14, 16, 19, 24, 31 (peak), 55, 65]

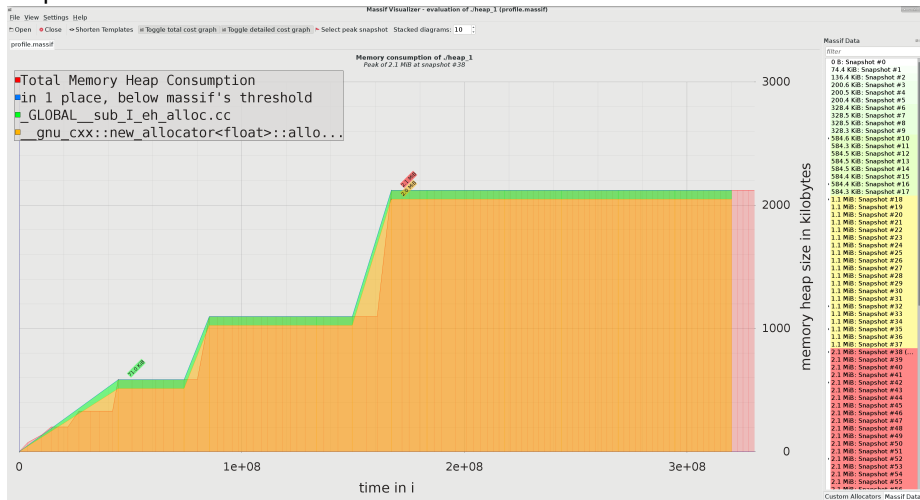
n	time (i)	total (B)	useful-heap (B)	extra-heap (B)	stacks (B)
0	0	0	0	0	0
1	5,653,795	90,704	90,112	24	568
2	9,989,271	139,760	139,264	24	472
3	15,104,837	205,296	204,800	24	472
4	19,380,693	205,296	204,800	24	472
5	25,946,699	336,296	335,872	24	400
6	30,435,811	336,296	335,872	24	400

# Massif-Visualizer Example: Tool Output

command:

```
massif-visualizer profile.massif
```

output:



## Section 7.5.1

### **References**

- 1 P. Floyd. [Valgrind part 1 — introduction](#).  
*Overload*, 108:14–15, Apr. 2012.
- 2 P. Floyd. [Valgrind part 2 — basic memcheck](#).  
*Overload*, 109:24–28, June 2012.
- 3 P. Floyd. [Valgrind part 3 — advanced memcheck](#).  
*Overload*, 110:4–7, Aug. 2012.
- 4 P. Floyd. [Valgrind part 4 — cachegrind and callgrind](#).  
*Overload*, 111:4–7, Oct. 2012.
- 5 P. Floyd. [Valgrind part 5 — massif](#).  
*Overload*, 112:20–24, Dec. 2012.

## Section 7.6

### Gcov and LLVM Cov



- Gcov is code coverage analysis tool, which is part of GCC software
- intended to be used in conjunction with GCC compiler
- supports measurement of function, statement, and decision (i.e., branch) coverage
- can be used to assess test coverage and possibly as guide for improving efficiency of code
- in order to generate data for Gcov, program being run must be properly instrumented
- compiler **instruments** code to count number of times each basic block (i.e., node in control-flow graph) and each branch (i.e., edge in control-flow graph) executes
- when program run, coverage data files generated
- coverage data files can then be processed and displayed with Gcov
- web site: <https://gcc.gnu.org/onlinedocs/gcc/Gcov.html>

# The gcov Program

- can examine data files produced by execution of compiler-instrumented code by using `gcov` program
- command line interface for `gcov` has following form:

```
gcov [options] [$file]...
```

- `$file` is source file for which coverage information desired or corresponding object file
- some commonly-used options include:

Option	Description
-a	write individual execution counts for every basic block
-b	write branch frequencies/counts to output file
-c	use branch counts instead of frequencies
-m	use demangled function names in output
-h	print help information and exit

- 1 build each program for which code coverage information is desired
  - must compile and link with `--coverage` option with GCC
  - one block-graph description file (with `.gcno` extension) generated for each object file produced during compilation and placed in same directory as corresponding object file
- 2 run each program for which coverage information is desired one or more times
  - when program exits, one count file (with `.gda` extension) generated for each object file associated with program
  - if output file does not yet exist, file created
  - if output file already exists, statistics are added to those already there
  - that is, statistics maintained in data files are cumulative
- 3 run `gcov` program to format data for display
  - for each source file with corresponding coverage data file, `gcov` generates *source\_file.gcov* (or transformed version of this name) for *source\_file.cpp*

# Using Gcov: Practical Considerations

- optimization and inline functions can cause strange behaviors in coverage statistics
- for example, optimization can cause multiple lines of code to be merged together, which will lead to unusual results for affected lines
- for above reasons, desirable to disable optimization (including function inlining) when performing coverage analysis
- exceptions can introduce many additional conditional branches, which may be undesirable when performing branch-coverage analysis
- consequently, when performing branch-coverage analysis, may be desirable to disable exceptions if not used

# Signum Example: Source Code

app.cpp

```
1  #include <cstdlib>
2  #include <iostream>
3  #include "utility.hpp"
4  int main(int argc, char** argv) {
5      if (argc < 2) {
6          return 1;
7      }
8      double x = std::atof(argv[1]);
9      std::cout << signum(x) << '\n';
10 }
```

utility.hpp

```
1  int signum(double x);
```

signum.cpp

```
1  #include "utility.hpp"
2  int signum(double x) {
3      if (x > 0) {
4          return 1;
5      } else if (x < 0) {
6          return -1;
7      } else {
8          return 0;
9      }
10 }
```

# Signum Example: Using Gcov

- build program `app` using GCC, ensuring that `--coverage` option is used for both compiling and linking; for example, use command sequence like:

```
g++ -O0 --coverage -c app.cpp
g++ -O0 --coverage -c signum.cpp
g++ -O0 --coverage -o app app.o signum.o
```

- run `app` program twice as follows:

```
./app 0
./app 1
```

- since `app` program run twice, statistics are accumulated from both runs of program
- run Gcov; for example, use command like:  
`gcov -b -c -m app.o signum.o`
- view resulting `.gcov` files (examples of which are given on following slides)

# Signum Example: Gcov Output [app.cpp.gcov]

```

-:      0:Source:app.cpp
-:      0:Programs:2
-:      1:#include <cstdlib>
-:      2:#include <iostream>
-:      3:#include "utility.hpp"
2:      4:int main(int argc, char** argv) {
2:      5:    if (argc < 2) {
branch 0 taken 0 (fallthrough)
branch 1 taken 2
      #####:      6:        return 1;
-:      7:    }
2:      8:    double x = std::atof(argv[1]);
2:      9:    std::cout << signum(x) << '\n';
call      0 returned 2
call      1 returned 2
call      2 returned 2
2:     10:}
-----
main:
function main called 2 returned 100% blocks executed 86%
-----
main:
function main called 2 returned 100% blocks executed 86%
-----
```

# Signum Example: Gcov Output [signum.cpp.gcov]

```
-:      0:Source:signum.cpp
-:      0:Programs:2
-:      1:#include "utility.hpp"
function signum(double) called 2 returned 100% blocks executed 83%
  2:      2:int signum(double x) {
  2:      3:      if (x > 0) {
branch 0 taken 1 (fallthrough)
branch 1 taken 1
  1:      4:          return 1;
  1:      5:      } else if (x < 0) {
branch 0 taken 0 (fallthrough)
branch 1 taken 1
    #####:      6:          return -1;
    -:      7:      } else {
    1:      8:          return 0;
    -:      9:      }
    -:     10:}
```



# LLVM Cov with GCC-Style Coverage

- LLVM Cov is code coverage analysis tool, which is part of LLVM software, intended for use with Clang compiler
- supports measurement of function, statement, and decision (i.e., branch) coverage
- manner in which coverage information collected similar to case of Gcov
- LLVM Cov provided as program called `llvm-cov`
- coverage data files produced by compiler and program execution processed with `gcov` subcommand of `llvm-cov`
- `gcov` subcommand of `llvm-cov` has very similar interface as `gcov` program (of Gcov)
- web site: <http://llvm.org/docs/CommandGuide/llvm-cov.html>

# Signum Example (Revisited): Using LLVM Cov

- build program `app` using Clang, ensuring that `--coverage` option is used for both compiling and linking; for example, use command sequence like:

```
clang++ -O0 --coverage -c app.cpp
clang++ -O0 --coverage -c signum.cpp
clang++ -O0 --coverage -o app app.o signum.o
```

- run `app` program twice as follows:

```
./app 0
./app 1
```

- since `app` program run twice, statistics are accumulated from both runs of program
- run LLVM Cov; for example, use command like:  

```
llvm-cov gcov -b -c app.o signum.o
```
- view resulting `.gcov` files (which look similar to ones produced by Gcov; see earlier Gcov example)

- Lcov is graphical front end for Gcov
- collects Gcov data from multiple source files and creates HTML pages containing source code annotated with coverage information
- also provides overview pages for easy navigation
- Lcov supports statement, function, and decision (i.e., branch) coverage measurement
- web site: <http://ltp.sourceforge.net/coverage/lcov.php>

# The l<sub>cov</sub> Program: Capture

- to capture (i.e., import) coverage data from data files generated by executing instrumented code, l<sub>cov</sub> has following command-line interface:

```
lcov -c [options]
```

- some commonly-used options include:

Option	Description
--gcov-tool <i>path</i>	set location of Gcov program to <i>path</i>
--rc <i>keyword=value</i>	set configuration parameter <i>keyword</i> to <i>value</i>
-o <i>file</i>	write trace data to file <i>file</i>
--no-external	do not capture coverage data from external source files
-d <i>dir</i>	search for .da files in directory <i>dir</i>
-h	print help information and exit
-q	do not print progress messages

# The lcov Program: Capture (Continued)

- Lcov relies on external program to process information in coverage data files
- external program must have similar command-line interface and behavior as `gcov`
- external program used for above functionality controlled by `--gcov-tool` option of `lcov`
- if using GCC, external program can simply be `gcov`
- if using Clang, however, need to invoke `gcov` subcommand of `llvm-cov` through wrapper program to provide expected interface
- this can be accomplished by creating file containing following shell script and then specifying this file to `lcov` via `--gcov-tool` option:

```
#!/usr/bin/env bash
exec llvm-cov gcov "$@"
```

# The l<sub>cov</sub> Program: Summary

- to display summary information from tracefile, l<sub>cov</sub> has following command-line interface:

```
lcov --summary [options]
```

- some commonly-used options include:

Option	Description
<code>--rc <i>keyword=value</i></code>	set configuration parameter <i>keyword</i> to <i>value</i>

# The genhtml Program

- coverage report can be generated in HTML format from one or more trace files with `genhtml` program, which has following command-line interface:

```
genhtml [options] $trace_file...
```

- some commonly-used options include:

Option	Description
<code>--legend</code>	include legend in report
<code>--branch-coverage</code>	include branch coverage information in report
<code>-o <i>dir</i></code>	set output directory for HTML document to <i>dir</i>
<code>-h</code>	print help information and exit

# Using Lcov

- 1 build project with GCC or Clang, ensuring that `--coverage` option is used for compiling and linking (as in earlier Gcov example)
- 2 run program (one or more times) to collect coverage data (e.g., as in earlier Gcov or LLVM Cov examples)
- 3 process coverage data with Lcov; for example, use command like following (where coverage data files under current directory):

```
lcov -c --gcov-tool $gcov_tool -d . -o coverage.lcov \
    --rc lcov_branch_coverage=1
```

where `$gcov_tool` is `gcov` for GCC or pathname of **wrapper script** shown earlier for Clang

- 4 if HTML output desired, generate output using `genhtml`; for example, use command like following:

```
genhtml --branch-coverage --legend -o output coverage.lcov
```

(main HTML document accessible via `output/index.html`)

- 5 if plaintext output desired, brief summary can be generated from tracefile using command like:

```
lcov --summary coverage.lcov --rc lcov_branch_coverage=1
```



# Signum Example (Revisited): Lcov Summary Output

```
Reading tracefile coverage.lcov
Summary coverage rate:
  lines.....: 83.3% (10 of 12 lines)
  functions...: 100.0% (2 of 2 functions)
  branches....: 66.7% (4 of 6 branches)
```

## ***LCOV - code coverage report***

**Current view:** [top level](#) - [signum](#)

**Test:** [coverage.info](#)

**Date:** [2018-09-19 14:58:31](#)

**Legend:** Rating: low: < 75 % medium: >= 75 % high: >= 90 %

	Hit	Total	Coverage
<b>Lines:</b>	<b>10</b>	<b>12</b>	<b>83.3 %</b>
<b>Functions:</b>	<b>2</b>	<b>2</b>	<b>100.0 %</b>
<b>Branches:</b>	<b>4</b>	<b>6</b>	<b>66.7 %</b>

Filename	Line Coverage ↕			Functions ↕		Branches ↕	
app.cpp	<div><div></div></div>	<b>83.3 %</b>	5 / 6	<b>100.0 %</b>	1 / 1	<b>50.0 %</b>	1 / 2
signum.cpp	<div><div></div></div>	<b>83.3 %</b>	5 / 6	<b>100.0 %</b>	1 / 1	<b>75.0 %</b>	3 / 4

Generated by: [LCOV version 1.13-15-g94eac0e](#)

## LCOV - code coverage report

Current view: [top level](#) - [signum](#) - [app.cpp](#) ([source](#) / [functions](#))

Test: [coverage.info](#)

Date: 2018-09-19 14:58:31

Hit Total Coverage

Lines: 5 6 83.3 %

Functions: 1 1 100.0 %

Legend: Lines: hit not hit | Branches: + taken - not taken # not executed

Branches: 1 2 50.0 %

Branch data	Line data	Source code
1	:	: #include <cstdlib>
2	:	: #include <iostream>
3	:	: #include "utility.hpp"
4	2	: int main(int argc, char** argv) {
5	[ - + ]: 2	: if (argc < 2) {
6	: 0	: return 1;
7	:	: }
8	2	: double x = std::atof(argv[1]);
9	2	: std::cout << signum(x) << '\n';
10	2	: }

Generated by: [LCOV version 1.13-15-g94eac0e](#)

## LCOV - code coverage report

Current view: [top level](#) - [signum](#) - [signum.cpp](#) ([source](#) / [functions](#))

Test: [coverage.info](#)

Date: 2018-09-19 14:58:31

Legend: Lines:   hit   not hit | Branches:   + taken   - not taken   # not executed

	Hit	Total	Coverage
Lines:	5	6	83.3 %
Functions:	1	1	100.0 %
Branches:	3	4	75.0 %

	Branch data	Line data	Source code
1		:	: #include "utility.hpp"
2		:	: int signum(double x) {
3	[ + + ]	2	: if (x > 0) {
4		:	: return 1;
5	[ - + ]	1	: } else if (x < 0) {
6		0	: return -1;
7		:	: } else {
8		1	: return 0;
9		:	: }
10		:	: }

Generated by: [LCOV version 1.13-15-g94eac0e](#)

- some potentially unexpected behaviors may be observed when using Lcov
- branches may be shown to be present at what might be unexpected places in code, due to implicit control flow associated with short-circuit evaluation or exception handling
- constructors or destructors that have been called may be reported as never having been called, due to compiler implementing constructors and destructors in manner similar to that specified in Itanium C++ ABI
- counts for different instantiations of same template are aggregated
- some functions associated with static initialization that have been called are reported as never having been called (probably due to bug)

# Templates Example: Source Code

## main.cpp

```
1  #include "utility.hpp"
2
3  int main() {
4      Widget<int> wi(42);
5      wi.get();
6      Widget<double> wd(42.0);
7      wd.get();
8      abs(42);
9      abs(42.0);
10 }
```

## utility.hpp

```
1  template <class T> class Widget {
2  public:
3      Widget (T x) : x_(x) {}
4      T get() const {
5          return x_;
6      }
7  private:
8      T x_;
9  };
10
11 template <class T> T abs(T x) {
12     if (x < 0) {
13         return -x;
14     } else {
15         return x;
16     }
17 }
```

# Templates Example: Gcov Output [main.cpp.gcov]

```
-:      0:Source:main.cpp
-:      0:Graph:main.gcno
-:      0:Data:main.gcda
-:      0:Runs:1
-:      0:Programs:1
-:      1:#include "utility.hpp"
-:      2:
function main called 1 returned 100% blocks executed 100%
1:      3:int main() {
1:      4:      Widget<int> wi(42);
call    0 returned 1
1:      5:      wi.get();
call    0 returned 1
1:      6:      Widget<double> wd(42.0);
call    0 returned 1
1:      7:      wd.get();
call    0 returned 1
1:      8:      abs(42);
call    0 returned 1
1:      9:      abs(42.0);
call    0 returned 1
1:     10:}
```

# Templates Example: Gcov Output [utility.hpp.gcov]

```
-:      0:Source:utility.hpp
-:      0:Graph:main.gcno
-:      0:Data:main.gcda
-:      0:Runs:1
-:      0:Programs:1
-:      1:template <class T> class Widget {
-:      2:public:
----- 2:      3:      Widget (T x) : x_(x) {}

Widget<double>::Widget(double):
function Widget<double>::Widget(double) called 1 returned 100% blocks executed 100%
----- 1:      3:      Widget (T x) : x_(x) {}

Widget<int>::Widget(int):
function Widget<int>::Widget(int) called 1 returned 100% blocks executed 100%
----- 1:      3:      Widget (T x) : x_(x) {}

      2:      4:      T get() const {
      2:      5:          return x_;
-:      6:      }

Widget<double>::get() const:
function Widget<double>::get() const called 1 returned 100% blocks executed 100%
      1:      4:      T get() const {
      1:      5:          return x_;
-:      6:      }

Widget<int>::get() const:
function Widget<int>::get() const called 1 returned 100% blocks executed 100%
      1:      4:      T get() const {
      1:      5:          return x_;
-:      6:      }

-----
-:      7:private:
-:      8:      T x_;
-:      9:};
-:     10:
      2:     11:template <class T> T abs(T x) {
      2:     12:     if (x < 0) {
#####      13:         return -x;
-:     14:     } else {
      2:     15:         return x;
-:     16:     }
-:     17:}
```



## ***LCOV - code coverage report***

**Current view:** [top level](#) - [templates\\_1](#) - [main.cpp](#) ([source](#) / [functions](#))

**Test:** [total.lcov](#)

**Date:** 2018-11-22 19:18:28

**Legend:** Lines:   hit   not hit

	Hit	Total	Coverage
<b>Lines:</b>	<b>8</b>	<b>8</b>	<b>100.0 %</b>
<b>Functions:</b>	<b>1</b>	<b>1</b>	<b>100.0 %</b>

Line data	Source code
1	: #include "utility.hpp"
2	:
3	1 : int main() {
4	1 :     Widget<int> wi(42);
5	1 :     wi.get();
6	1 :     Widget<double> wd(42.0);
7	1 :     wd.get();
8	1 :     abs(42);
9	1 :     abs(42.0);
10	1 : }

Generated by: [LCOV version 1.13-16-ge675080](#)

## LCOV - code coverage report

Current view: [top level](#) - [templates\\_1](#) - [utility.hpp](#) ([source](#) / [functions](#))

Test: [total.lcov](#)

Date: **2018-11-22 19:18:28**

	Hit	Total	Coverage
Lines:	10	11	90.9 %
Functions:	4	4	100.0 %

Legend: Lines: hit not hit

Line data	Source code
1	: template <class T> class Widget {
2	: public:
3	2 :     Widget (T x) : x_(x) {}
4	2 :     T get() const {
5	2 :         return x_;
6	:     }
7	1 : private:
8	1 :     T x_;
9	:     };
10	1 :
11	1 : template <class T> T abs(T x) {
12	:     if (x < 0) {
13	:         return -x;
14	:     } else {
15	:         return x;
16	:     }
17	2 : }

Generated by: [LCOV version 1.13-16-ge675080](#)

# Exceptions Example: Source Code

## main.cpp

```
1  #include <vector>
2  #include "utility.hpp"
3
4  void increment_3(int& i) noexcept {++i;}
5  void increment_4(int& i) {++i;}
6
7  int main() {
8      std::vector<int> u;
9      std::vector<int> v{1, 2, 3, 4};
10     u = v;
11     increment_1(u[0]);
12     increment_2(u[0]);
13     increment_3(u[0]);
14     increment_4(u[0]);
15 }
```

## utility.hpp

```
1  void increment_1(int&) noexcept;
2  void increment_2(int&);
```

## utility.cpp

```
1  #include "utility.hpp"
2
3  void increment_1(int& i) noexcept {++i;}
4  void increment_2(int& i) {++i;}
```

# Exceptions Example: Gcov Output [main.cpp.gcov]

```
0:Source:main.cpp
0:Programs:2
1:#include <vector>
2:#include "utility.hpp"
3:
4:void increment_3(int& i) noexcept {++i;}
-----
increment_3(int&):
function increment_3(int&) called 1 returned 100% blocks executed 100%
increment_3(int&):
function increment_3(int&) called 1 returned 100% blocks executed 100%
-----
5:void increment_4(int& i) {++i;}
-----
increment_4(int&):
function increment_4(int&) called 1 returned 100% blocks executed 100%
increment_4(int&):
function increment_4(int&) called 1 returned 100% blocks executed 100%
-----
6:
7:int main() {
8:   std::vector<int> u;
call 0 returned 1
call 1 returned 1
call 2 never executed
9:   std::vector<int> v{1, 2, 3, 4};
call 0 returned 1
call 1 returned 1
branch 2 taken 1 (fallthrough)
branch 3 taken 0 (throw)
call 4 returned 1
call 5 never executed
10:   u = v;
call 0 returned 1
branch 1 taken 1 (fallthrough)
branch 2 taken 0 (throw)
11:   increment_1(u[0]);
call 0 returned 1
call 1 returned 1
12:   increment_2(u[0]);
call 0 returned 1
call 1 returned 1
branch 2 taken 1 (fallthrough)
branch 3 taken 0 (throw)
13:   increment_3(u[0]);
call 0 returned 1
call 1 returned 1
14:   increment_4(u[0]);
call 0 returned 1
call 1 returned 1
-----
15:}
-----
main:
function main called 1 returned 100% blocks executed 72%
main:
function main called 1 returned 100% blocks executed 72%
-----
```

# Exceptions Example: Gcov Output [utility.cpp.gcov]

```
-:      0:Source:utility.cpp
-:      0:Programs:2
-:      1:#include "utility.hpp"
-:      2:
function increment_1(int&) called 1 returned 100% blocks executed 100%
  1:      3:void increment_1(int& i) noexcept {++i;}
function increment_2(int&) called 1 returned 100% blocks executed 100%
  1:      4:void increment_2(int& i) {++i;}
```

## ***LCOV - code coverage report***

**Current view:** [top level - exceptions\\_1](#)

**Test:** [coverage.info](#)

**Date:** 2018-09-20 18:48:59

**Legend:** Rating: low: < 75 % medium: >= 75 % high: >= 90 %

	Hit	Total	Coverage
<b>Lines:</b>	13	13	100.0 %
<b>Functions:</b>	5	5	100.0 %
<b>Branches:</b>	3	6	50.0 %

Filename	Line Coverage ↕			Functions ↕		Branches ↕	
<a href="#">main.cpp</a>	<div style="width: 100%; background-color: green;"></div>	100.0 %	11 / 11	100.0 %	3 / 3	50.0 %	3 / 6
<a href="#">utility.cpp</a>	<div style="width: 100%; background-color: green;"></div>	100.0 %	2 / 2	100.0 %	2 / 2	-	0 / 0

Generated by: [LCOV version 1.13-15-g94eac0e](#)

## LCOV - code coverage report

Current view: [top level - exceptions\\_1 - main.cpp](#) (source / functions)

Test: [coverage.info](#)

Date: 2018-09-20 18:48:59

	Hit	Total	Coverage
--	-----	-------	----------

Lines:	11	11	100.0 %
--------	----	----	---------

Functions:	3	3	100.0 %
------------	---	---	---------

Legend: Lines:   hit   not hit | Branches:   + taken   - not taken   # not executed

Branches:	3	6	50.0 %
-----------	---	---	--------

Branch data

Line data

Source code

1	:	:	#include <vector>
2	:	:	#include "utility.hpp"
3	:	:	
4	:	1	void increment_3(int& i) noexcept {++i;}
5	:	1	void increment_4(int& i) {++i;}
6	:	:	
7	:	1	int main() {
8	:	2	std::vector<int> u;
9	[ + - ]:	2	std::vector<int> v{1, 2, 3, 4};
10	[ + - ]:	1	u = v;
11	:	1	increment_1(u[0]);
12	[ + - ]:	1	increment_2(u[0]);
13	:	1	increment_3(u[0]);
14	:	1	increment_4(u[0]);
15	:	1	}

Generated by: [LCOV version 1.13-15-g94eac0e](#)

## LCOV - code coverage report

Current view: [top level](#) - [exceptions\\_1](#) - [utility.cpp](#) ([source](#) / [functions](#))

Test: [coverage.info](#)

Date: 2018-09-20 18:48:59

	Hit	Total	Coverage
--	-----	-------	----------

Lines:	2	2	100.0 %
--------	---	---	---------

Functions:	2	2	100.0 %
------------	---	---	---------

Branches:	0	0	-
-----------	---	---	---

Legend: Lines: hit not hit | Branches: + taken - not taken # not executed

Branch data

Line data

Source code

1	:	:	#include "utility.hpp"
2	:	:	
3	:	1	void increment_1(int& i) noexcept {++i;}
4	:	1	void increment_2(int& i) {++i;}

Generated by: [LCOV version 1.13-15-g94eac0e](#)



# Short-Circuiting Example: Source Code

utility.hpp

```
1  class Flag {
2  public:
3      explicit Flag(bool b = false) noexcept : b_(b) {}
4      explicit operator bool() const noexcept {return b_;}
5      Flag operator&&(const Flag& other) noexcept {
6          return Flag(b_ && other.b_);
7      }
8      Flag operator||(const Flag& other) noexcept {
9          return Flag(b_ || other.b_);
10     }
11 private:
12     bool b_;
13 };
```

# Short-Circuiting Example: Source Code

main.cpp

```
1  #include <tuple>
2  #include "utility.hpp"
3
4  bool func(bool a, bool b, bool c) noexcept {
5      return (a || b) && c;
6  }
7
8  Flag func(Flag a, Flag b, Flag c) noexcept {
9      return (a || b) && c;
10 }
11
12 int main() {
13     std::tuple<bool, bool, bool> tests[] = {{0, 0, 0}, {1, 1, 1}};
14     for (auto&& [a, b, c] : tests) {
15         func(a, b, c);
16         func(Flag(a), Flag(b), Flag(c));
17     }
18 }
```

short-circuit evaluation example

## LCOV - code coverage report

Current view: [top level](#) - [decision](#) - [main.cpp](#) ([source](#) / [functions](#))

Test: [coverage.info](#)

Date: 2018-09-20 20:16:20

Hit Total Coverage

Lines: 10 10 100.0 %

Functions: 3 3 100.0 %

Branches: 6 8 75.0 %

Legend: Lines: hit not hit | Branches: + taken - not taken # not executed

	Branch data	Line data	Source code
1		:	: #include <tuple>
2		:	: #include "utility.hpp"
3		:	:
4		:	:
5	[ + + ] [ - + ]	2 :	bool func(bool a, bool b, bool c) noexcept {
	[ + - ]	2 :	return (a    b) && c;
6		:	}
7		:	:
8		2 :	Flag func(Flag a, Flag b, Flag c) noexcept {
9		2 :	return (a    b) && c;
10		:	}
11		:	:
12		1 :	int main() {
13		1 :	std::tuple<bool, bool, bool> tests[] = {{0, 0, 0}, {1, 1, 1}};
14	[ + + ]	3 :	for (auto&& [a, b, c] : tests) {
15		2 :	func(a, b, c);
16		2 :	func(Flag(a), Flag(b), Flag(c));
17		:	}
18		1 :	}

Generated by: [LCOV version 1.13-15-g94eac0e](#)

## LCOV - code coverage report

Current view: [top level](#) - [decision](#) - [utility.hpp](#) ([source](#) / [functions](#))

Test: [coverage.info](#)

Date: 2018-09-20 20:16:20

Legend: Lines:   hit   not hit | Branches:   +   -   # not taken   # not executed

	Hit	Total	Coverage
Lines:	5	5	100.0 %
Functions:	3	3	100.0 %
Branches:	6	8	75.0 %

	Branch data	Line data	Source code
1		:	: class Flag {
2		:	: public:
3		10	: explicit Flag(bool b = false) noexcept : b(b) {}
4		:	: explicit operator bool() const noexcept {return b_;}
5		2	: Flag operator&&(const Flag& other) noexcept {
6	[ + + ] [ + - ]	2	: return Flag(b_ && other.b_);
7		:	: }
8		2	: Flag operator  (const Flag& other) noexcept {
9	[ + + ] [ - + ]	2	: return Flag(b_    other.b_);
10		:	: }
11		:	: private:
12		:	: bool b_;
13		:	: };

Generated by: [LCOV version 1.13-15-g94eac0e](#)

# Virtual Destructor Example 1A: Source Code

example\_1\_1.cpp

```
1  struct Gadget {  
2      Gadget();  
3      virtual ~Gadget();  
4  };  
5  Gadget::Gadget() {}  
6  Gadget::~~Gadget() {}  
7  
8  struct Widget : Gadget {  
9      Widget();  
10     virtual ~Widget();  
11 };  
12 Widget::Widget() {}  
13 Widget::~~Widget() {}  
14  
15 int main() {  
16     Widget w;  
17 }
```

## LCOV - code coverage report

Current view: [top level](#) - [example\\_1](#) - [example\\_1\\_1.cpp](#) ([source](#) / [functions](#))

Test: **total.lcov**

Date: **2018-11-22 15:23:25**

Hit Total Coverage

Lines: **7** **7** **100.0 %**

Functions: **5** **7** **71.4 %**

Legend: Lines: hit not hit

Function Name ↕	Hit count
<a href="#">_ZN6GadgetD0Ev</a>	<b>0</b>
<a href="#">_ZN6WidgetD0Ev</a>	<b>0</b>
<a href="#">_ZN6GadgetC2Ev</a>	<b>1</b>
<a href="#">_ZN6GadgetD2Ev</a>	<b>1</b>
<a href="#">_ZN6WidgetC2Ev</a>	<b>1</b>
<a href="#">_ZN6WidgetD2Ev</a>	<b>1</b>
<a href="#">main</a>	<b>1</b>

Generated by: [LCOV version 1.13-16-ge675080](#)

- `_ZN6GadgetD0Ev` is mangled name for deleting destructor variant (in Itanium C++ ABI) of `Gadget::~~Gadget()`
- `_ZN6WidgetD0Ev` is mangled name for deleting destructor variant (in Itanium C++ ABI) of `Widget::~~Widget()`

# Virtual Destructor Example 1B: Source Code

example\_1\_2.cpp

```
1  struct Gadget {
2      Gadget();
3      virtual ~Gadget();
4  };
5  Gadget::Gadget() {}
6  Gadget::~~Gadget() {}
7
8  struct Widget : Gadget {
9      Widget();
10     virtual ~Widget();
11 };
12 Widget::Widget() {}
13 Widget::~~Widget() {}
14
15 int main() {
16     Widget w;
17     Widget* wp = new Widget;
18     delete wp;
19     // invoke deleting destructor (Itanium C++ ABI)
20     Gadget* gp = new Gadget;
21     delete gp;
22     // invoke deleting destructor (Itanium C++ ABI)
23 }
```

## ***LCOV - code coverage report***

**Current view:** [top level](#) - [example\\_1](#) - [example\\_1\\_2.cpp](#) ([source](#) / [functions](#))

**Test:** [total.lcov](#)

**Date:** 2018-11-22 15:23:27

Hit Total Coverage

**Lines:** 11 11 100.0 %

**Functions:** 7 7 100.0 %

**Legend:** Lines: hit not hit

Function Name ↕	Hit count
<a href="#">_ZN6GadgetD0Ev</a>	1
<a href="#">_ZN6WidgetD0Ev</a>	1
<a href="#">main</a>	1
<a href="#">_ZN6WidgetC2Ev</a>	2
<a href="#">_ZN6WidgetD2Ev</a>	2
<a href="#">_ZN6GadgetC2Ev</a>	3
<a href="#">_ZN6GadgetD2Ev</a>	3

Generated by: [LCOV version 1.13-16-ge675080](#)

- [\\_ZN6GadgetD0Ev](#) is now called (unlike previously)
- [\\_ZN6WidgetD0Ev](#) is now called (unlike previously)



- Gcov and Lcov somewhat tricky to use with CMake
- one very convenient way to collect coverage information with CMake is by using `CodeCoverage` module developed by Lars Bilke, which can be obtained from:
  - <https://github.com/bilke/cmake-modules/blob/master/CodeCoverage.cmake>
- slightly modified version of above module can be found at:
  - <https://github.com/mdadams/cmake-modules/blob/master/CodeCoverage.cmake>

# Gcov/Lcov CMake Example: CMakeLists File

## CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
2  project(coverage_example LANGUAGES CXX C)
3  option(ENABLE_COVERAGE "Enable coverage" false)
4  set(CMAKE_VERBOSE_MAKEFILE true)
5  set(CMAKE_EXPORT_COMPILE_COMMANDS true)
6
7  if (ENABLE_COVERAGE)
8      set(CMAKE_BUILD_TYPE "Debug" CACHE STRING
9          "Set the build type." FORCE)
10     include(CodeCoverage.cmake)
11     append_coverage_compiler_flags()
12 endif()
13
14 add_executable(app app.cpp)
15
16 if (ENABLE_COVERAGE)
17     setup_target_for_coverage_lcov(
18         NAME coverage
19         EXECUTABLE ${CMAKE_CURRENT_SOURCE_DIR}/run_tests
20         EXECUTABLE_ARGS ${CMAKE_CURRENT_BINARY_DIR}
21         LCOV_ARGS -rc lcov_branch_coverage=1
22         GENHTML_ARGS --branch-coverage
23         DEPENDENCIES app)
24 endif()
```

# Gcov/Lcov CMake Example: Source File and Test Script

app.cpp

```
1  #include <iostream>
2
3  void func(int x) {
4      if (x % 2) {
5          std::cout << "odd\n";
6      } else {
7          std::cout << "even\n";
8      }
9  }
10
11 int main() {
12     for (int i = 0; i < 100; ++i) {
13         func(i);
14     }
15 }
```

run\_tests

```
1  #! /usr/bin/env bash
2
3  BINARY_DIR="$1"
4
5  $BINARY_DIR/app
```

- Gcovr is another tool for processing GCC coverage data
- similar in functionality to Lcov
- can generate XML output that can be used with Cobertura plugin of Jenkins continuous integration server
- web site: <https://gcovr.com>
- GitHub web page: <https://github.com/gcovr/gcovr>

## Section 7.6.1

### **References**

- 1 B. J. Gough, An Introduction to GCC, Network Theory Limited, UK, 2004.
- 2 GNU Compiler Collection (GCC), <https://gcc.gnu.org>.
- 3 LLVM Clang, <https://clang.llvm.org>.
- 4 C. Lattner and V. Adve. **LLVM: A compilation framework for lifelong program analysis and transformation.**  
In *Proc. of International Symposium on Code Generation and Optimization*, Mar. 2004.

## Section 7.7

### LLVM XRay

- LLVM XRay is function call tracing system which combines compiler-inserted instrumentation points and run-time library that can enable and disable instrumentation
- XRay is part of LLVM software and needs to be used in conjunction with Clang compiler
- allows collection of trace of function calls and information about stack frames for those calls
- very minimal overhead when collection disabled
- moderate overhead when collection enabled
- consists of three main parts:
  - 1 compiler-inserted instrumentation points
  - 2 runtime library for enabling/disabling tracing at runtime
  - 3 suite of tools for analyzing traces
- web site: <https://llvm.org/docs/XRay.html>



# Instrumenting Application

- when compiling with Clang, must enable instrumentation of code
- Clang has several options that relate to instrumentation of code for XRay:
  - `-fxray-instrument`
    - enable instrumentation for XRay
  - `-fxray-instruction-threshold n`
    - specify that functions with at least *n* instructions should be instrumented
  - `-fxray-always-instrument`
    - treat each function as if marked with `always-instrument` attribute
  - `-fxray-never-instrument`
    - treat each function as if marked with `never-instrument` attribute
- in order to allow any instrumentation to be performed by compiler, must provide `-fxray-instrument` option
- other options may also be used to assert greater control over how instrumentation performed
- must link with XRay Runtime Library (to be discussed shortly)
- providing `-fxray-instrument` option to Clang when linking, will automatically link with XRay Runtime Library

- Clang compiler attributes for XRay include:
  - `clang::xray_always_instrument`
    - force (free or member) function to always be instrumented for XRay
  - `clang::xray_never_instrument`
    - inhibit (free or member) function from being instrumented for XRay
  - `clang::xray_log_args(n)`
    - preserve *n* arguments of function in logging (as of Clang 7.0.0, only *n* = 1 is supported)
- above attributes can be used to assert finer control over XRay instrumentation (if necessary)
- some examples of using attributes follow:

```
[[clang::xray_always_instrument]] void func_1() { /* ... */ }  
[[clang::xray_never_instrument]] void func_2() { /* ... */ }
```

- XRay Runtime Library is part of LLVM compiler-rt project
- environment variable `XRAY_OPTIONS` contains whitespace-delimited list of options that control behavior of runtime library at application start up
- some supported options include:
  - `patch_premain`
    - if instrumentation points should be enabled prior to entering `main`
    - valid values: `true` and `false`; default value: `false`
  - `xray_mode`
    - default mode to install and initialize before `main`
    - valid values: `xray-basic` (basic) and `xray-fdr` (flight data recorder)
  - `xray_logfile_base`
    - base name of XRay log file
    - default value: `xray-log.`
  - `verbosity`
    - runtime verbosity level
    - integer value; default value: 0
- application can also control settings of XRay Runtime Library via API

- XRay supports two modes of operation:
  - 1 basic
  - 2 flight data recorder
- in basic mode, traces of application's execution periodically appended to single log file
- environment variable `XRAY_BASIC_OPTIONS` can be used to provide additional options for basic mode
- in flight data recorder (FDR) mode, tracing events written to fixed-size circular buffer in memory and traces can be dumped on demand through triggering API
- environment variable `XRAY_FDR_OPTIONS` can be used to provide additional options for FDR mode

- trace log files can be analyzed with `llvm-xray` program

- command-line interface for `llvm-xray`:

```
llvm-xray [options] subcommand [subcommand_options]
```

- supports several subcommands:

- `account`
  - performs basic function call accounting statistics with various options for sorting and output formats
- `convert`
  - converts XRay log file from one format to another
- `extract`
  - extracts instrumentation map from binary and returns as YAML
- `graph`
  - generates DOT graph of function call relationships between functions in XRay trace
- `stack`
  - reconstructs function call stacks from timeline of function calls in XRay trace

# Code Example

example\_1.cpp

```
1  #include <iostream>
2  #include <thread>
3  #include <chrono>
4
5  [[gnu::noinline, clang::xray_always_instrument]]
6  void delay() {
7      using namespace std::chrono_literals;
8      std::this_thread::sleep_for(1ms);
9  }
10
11 [[gnu::noinline, clang::xray_always_instrument,
12    clang::xray_log_args(1)]]
13 void func(int i, std::size_t n) {
14     for (std::size_t j = 0; j < n; ++j) {
15         std::cout << (i ? 'X' : 'O') << std::flush;
16         delay();
17     }
18 }
19
20 [[clang::xray_never_instrument]]
21 int main() {
22     std::thread t1([]() { func(0, 500); });
23     std::thread t2([]() { func(1, 500); });
24     t1.join(); t2.join();
25     std::cout << '\n';
26 }
```

- build instrumented code with command:

```
clang++ -std=c++17 -fxray-instrument -O1 example_1.cpp -o example_1
```

- run application with command:

```
XRAY_OPTIONS="patch_premain=true xray_mode=xray-basic" ./example_1
```

# Account Subcommand Example

command:

```
llvm-xray account -k -m ./example_1 xray-log.example_1.5fNAQv
```

output produced:

```
Functions with latencies: 3
funcid  count [    min,      med,      90p,      99p,      max]    sum function
  1      1000 [ 0.000782, 0.000827, 0.000835, 0.000899, 0.000919] 0.825670 <invalid>:0:0: delay()
  2         1 [ 0.420710, 0.420710, 0.420710, 0.420710, 0.420710] 0.420710 <invalid>:0:0: func(int,
    unsigned long)
  3      1000 [ 0.000781, 0.000825, 0.000831, 0.000896, 0.000917] 0.822853 <invalid>:0:0: void std::
    this_thread::sleep_for<long, std::ratio<11, 10001> >(std::chrono::duration<long, std::ratio<1
    1, 10001> > const&)
```



# Stack Subcommand Example

## command:

```
llvm-xray stack -m ./example_1 xray-log.example_1.5fNAQv
```

## output produced:

```
Unique Stacks: 2
Top 10 Stacks by leaf sum:

Sum: 1439155104
lvl  function                                count      sum
#0  func(int, unsigned long)                  1      1475472614
#1  delay()                                   500     1446943616
#2  void std::this_thread::sleep_for<long, std::ratio<1l, 100... 499     1439155104

Sum: 1437712794
lvl  function                                count      sum
#0  func(int, unsigned long)                  1      1472486054
#1  delay()                                   500     1442899912
#2  void std::this_thread::sleep_for<long, std::ratio<1l, 100... 500     1437712794

Top 10 Stacks by leaf count:

Count: 500
lvl  function                                count      sum
#0  func(int, unsigned long)                  1      1472486054
#1  delay()                                   500     1442899912
#2  void std::this_thread::sleep_for<long, std::ratio<1l, 100... 500     1437712794

Count: 499
lvl  function                                count      sum
#0  func(int, unsigned long)                  1      1475472614
#1  delay()                                   500     1446943616
#2  void std::this_thread::sleep_for<long, std::ratio<1l, 100... 499     1439155104
```

# Trace Subcommand Example

## command:

```
llvm-xray convert -f yaml -y -m ./example_1 xray-log.example_1.5fNAQv
```

## output produced:

```
---
header:
  version:      3
  type:         0
  constant-tsc: true
  nonstop-tsc:  true
  cycle-frequency: 3500000000
records:
- { type: 0, func-id: 2, function: 'func(int, unsigned long)', args: [ 0 ], cpu: 5, thread: 29992,
  process: 29991, kind: function-enter-arg, tsc: 1027021087397834 }
- { type: 0, func-id: 2, function: 'func(int, unsigned long)', args: [ 1 ], cpu: 4, thread: 29993,
  process: 29991, kind: function-enter-arg, tsc: 1027021087460270 }
- { type: 0, func-id: 1, function: 'delay()', cpu: 5, thread: 29992, process: 29991, kind: function-enter
  , tsc: 1027021087525002 }
- { type: 0, func-id: 3, function: 'void std::this_thread::sleep_for<long, std::ratio<11, 10001> >(std::
  chrono::duration<long, std::ratio<11, 10001> > const&)', cpu: 5, thread: 29992, process: 29991,
  kind: function-enter, tsc: 1027021087526060 }
- { type: 0, func-id: 1, function: 'delay()', cpu: 4, thread: 29993, process: 29991, kind: function-enter
  , tsc: 1027021087532070 }
- { type: 0, func-id: 3, function: 'void std::this_thread::sleep_for<long, std::ratio<11, 10001> >(std::
  chrono::duration<long, std::ratio<11, 10001> > const&)', cpu: 4, thread: 29993, process: 29991,
  kind: function-enter, tsc: 1027021087533716 }
- { type: 0, func-id: 3, function: 'void std::this_thread::sleep_for<long, std::ratio<11, 10001> >(std::
  chrono::duration<long, std::ratio<11, 10001> > const&)', cpu: 4, thread: 29993, process: 29991,
  kind: function-exit, tsc: 1027021090288388 }
```

- 1 Dean Berris. Debugging With LLVM XRay. Pacific++, Christchurch, NZ, Oct. 26, 2017. Available online at [https://youtu.be/cBc\\_MxbzqhY](https://youtu.be/cBc_MxbzqhY). [This talk gives an introduction to the LLVM XRay tool, including what it is, what it does, and how it works.]
- 2 Dean Michael Berris. XRay in LLVM: Function Call Tracing and Analysis. LLVM Developers Meeting, San Jose, CA, USA, Oct. 18–19, 2017. Available online at [https://youtu.be/jyL-\\_\\_zOGcU](https://youtu.be/jyL-__zOGcU).

## Section 7.8

### Miscellaneous Tools

- Kcov is tool for measuring code coverage
- unlike Gcov/Lcov, Kcov does not require instrumentation of code
- Kcov only supports measurement of statement coverage
- based on breakpoints
- can generate HTML reports similar in appearance to Lcov
- web site: <https://github.com/SimonKagstrom/kcov>

- 1 Simon Kagstrom. Kcov — A Single-Step Code Coverage Tool. SwedenCpp::Stockholm, Stockholm, Sweden, Sept. 20, 2018. Available online at <https://youtu.be/1QMHbp5LUKg>.

## Section 7.9

### Catch2

- Catch2 (originally known as Catch) is multiparadigm test framework for C++
- Catch2 stands for “C++ automated test cases in a header”
- primarily distributed as single header library
- open source; released under Boost Software License
- written by Phil Nash
- official Git repository: <http://github.com/catchorg/Catch2>
- Google group: <http://groups.google.com/group/catch-forum>



# Counter Class Example: counter.hpp

```
1  #include <limits>
2  #include <stdexcept>
3
4  class counter {
5  public:
6      using count_type = std::size_t;
7      static constexpr count_type max_count() {
8          return std::numeric_limits<count_type>::max();
9      }
10     counter(count_type count = 0) : count_(count) {}
11     count_type get_count() const {
12         return count_;
13     }
14     void increment() {
15         if (count_ == max_count()) {
16             throw std::overflow_error("counter overflow");
17         }
18         ++count_;
19     }
20 private:
21     count_type count_;
22 };
```

# Counter Class Example: Test Code

```
1  #define CATCH_CONFIG_MAIN
2  #include <catch2/catch.hpp>
3  #include "counter.hpp"
4
5  TEST_CASE("constructor", "[counter]") {
6      counter x;
7      CHECK(x.get_count() == 0);
8      counter y(1);
9      CHECK(y.get_count() == 1);
10 }
11
12 TEST_CASE("maximum count", "[counter]") {
13     CHECK(counter::max_count() == std::numeric_limits<
14         counter::count_type>::max());
15 }
16
17 TEST_CASE("increment (no overflow)", "[counter]") {
18     counter x(0);
19     REQUIRE(x.get_count() == 0);
20     x.increment();
21     CHECK(x.get_count() == 1);
22 }
23
24 TEST_CASE("increment (overflow)", "[counter]") {
25     counter x(counter::max_count());
26     CHECK_THROWS_AS(x.increment(), std::overflow_error);
27 }
```

# Section Example

```
1  #define CATCH_CONFIG_MAIN
2  #include <catch2/catch.hpp>
3  #include <cstdlib>
4  #include <vector>
5
6  TEST_CASE("Check resize", "[vector]") {
7      constexpr std::size_t size = 128;
8      std::vector<int> x(size);
9      REQUIRE(x.size() == size);
10     REQUIRE(x.capacity() >= size);
11     SECTION("Increase size") {
12         std::size_t n = size * 16;
13         x.resize(n);
14         CHECK(x.size() == n);
15         CHECK(x.capacity() >= n);
16     }
17     SECTION("Decrease size") {
18         std::size_t n = size / 16;
19         x.resize(n);
20         CHECK(x.size() == n);
21         CHECK(x.capacity() >= n);
22     }
23     SECTION("Zero size") {
24         x.resize(0);
25         CHECK(x.size() == 0);
26         CHECK(x.capacity() >= size);
27     }
28 }
```

# Approximate Comparison Example

```
1  #define CATCH_CONFIG_MAIN
2  #include <catch2/catch.hpp>
3
4  TEST_CASE("addition") {
5      float x = 0.0f;
6      for (int i = 0; i < 10; ++i) {x += 0.1f;}
7      CHECK_NOFAIL(x == 1.0f);
8      // condition may be false due to roundoff error
9      CHECK(x == Approx(1.0f));
10     // should pass
11     CHECK(x == Approx(1.0f).margin(0.01f));
12     // should pass (absolute tolerance 0.01)
13     CHECK(x == Approx(1.0f).epsilon(0.01f));
14     // should pass (relative tolerance 1%)
15 }
```

# Type-Parameterized Test Example: `stack.hpp`

```
1  #include <cstddef>
2  #include <vector>
3
4  // Note: T is not allowed to be bool.
5  template <class T>
6  class Stack {
7  public:
8      bool empty() const {return s_.empty();}
9      std::size_t size() const {return s_.size();}
10     const T& top() const {return s_.back();}
11     void push(const T& x) {s_.push_back(x);}
12     void pop() {s_.pop_back();}
13 private:
14     std::vector<T> s_;
15 };
```

# Type-Parameterized Test Example: Test Code

```
1  #define CATCH_CONFIG_MAIN
2  #include <catch2/catch.hpp>
3  #include <complex>
4  #include "stack.hpp"
5
6  TEMPLATE_TEST_CASE("Check default constructor", "[constructors]",
7      int, double, std::complex<double>) {
8      Stack<TestType> s;
9      CHECK(s.empty());
10     CHECK(s.size() == 0);
11 }
12
13 TEMPLATE_TEST_CASE("Check push and pop", "[modifiers]",
14     int, double, std::complex<double>) {
15     std::size_t size = 0;
16     Stack<TestType> s;
17     REQUIRE(s.size() == size);
18     while (size < 3) {
19         ++size; s.push(TestType(size));
20         REQUIRE(s.size() == size);
21         REQUIRE(s.top() == TestType(size));
22     }
23     while (size > 0) {
24         s.pop(); --size;
25         REQUIRE(s.size() == size);
26     }
27 }
```

# Test Fixture Example

```
1  #define CATCH_CONFIG_MAIN
2  #include <catch2/catch.hpp>
3  #include <deque>
4  #include <stdexcept>
5
6  class TestFixture {
7  public:
8      TestFixture() : q0(), q1{1, 2, 3} {
9      }
10 protected:
11     std::deque<int> q0;
12     std::deque<int> q1;
13 };
14
15 TEST_CASE_METHOD(TestFixture, "Check deque at", "[deque]") {
16     CHECK_THROWS_AS(q0.at(0), std::out_of_range);
17     CHECK(q1.at(0) == 1);
18     CHECK_THROWS_AS(q1.at(3), std::out_of_range);
19 }
20
21 TEST_CASE_METHOD(TestFixture, "Check deque clear", "[deque]") {
22     q0.clear();
23     CHECK(q0.size() == 0);
24     q1.clear();
25     CHECK(q1.size() == 0);
26 }
```

- 1 Phil Nash. Modern C++ Testing with Catch2. CppCon, Bellevue, WA, USA, Sept. 24, 2018. Available online at [https://youtu.be/Ob5\\_XZrFQH0](https://youtu.be/Ob5_XZrFQH0).
- 2 Phil Nash. Modern C++ Testing with Catch2. Meeting C++, Berlin, Germany, Nov. 9, 2017. Available online at <https://youtu.be/3tIE6X5FjDE>.
- 3 Phil Nash. Modern C++ Testing with Catch2. C++ Edinburgh, Edinburgh, UK, Aug. 14, 2017. Available online at <https://youtu.be/grC0S6ZK59U>.
- 4 Phil Nash. Test Driven C++ with Catch. CppCon, Bellevue, WA, USA, Sept. 22, 2015. Available online at <https://youtu.be/gdzP3pAC6UI>.
- 5 Phil Nash. Testdriven C++ with Catch. Meeting C++, Berlin, Germany, Dec. 5–6, 2014. Available online at <https://youtu.be/C2LcIp56i-8>.



## Part 8

### Performance Analysis Tools

## Section 8.1

### Perf

# Linux Kernel Perf Event Interface

- Linux kernel provides Perf Event (i.e., `perf_event`) interface for performance monitoring
- `perf_event_open` system call returns file descriptor that can then be used to collect performance information
- collection of performance data started and stopped with `ioctl` system call
- performance data accessed either via `read` or `mmap` system call
- Perf Event interface used by numerous performance analysis tools and libraries on Linux systems (e.g., Perf and PAPI)
- supports many profiling/tracing features, including:
  - CPU performance monitoring counters
  - statically defined tracepoints
  - user and kernel dynamic tracepoints
- good documentation on Perf Event interface is scarce

- open-source profiling tool
- can collect aggregated counts of events during code execution
- can perform event-driven sample-based profiling
- uses Perf Event interface of Linux kernel
- noninvasive (i.e., no code instrumentation required)
- low overhead (i.e., code runs close to native speed)
- sample-based profiling can collect stack traces in addition to instruction pointer
- does not provide call counts for functions
- web site: <https://perf.wiki.kernel.org>

- hardware event:
  - event measurable by performance monitoring unit (PMU) of processor
  - examples: CPU cycles (`cycles`) and cache misses (`cache-misses`)
- hardware cache event:
  - event measurable by PMU of processor
  - examples: L1 data cache load misses (`L1-dcache-load-misses`) and data translation-lookaside-buffer load misses (`dTLB-load-misses`)
- software event:
  - low-level events based on kernel counters
  - examples: CPU clock (`cpu-clock`) and page fault (`page-faults`)
- kernel tracepoint event:
  - predefined static instrumentation points in kernel code where trace information can be collected
  - examples: entering `open` system call (`syscalls:sys_enter_open`) and context switch (`sched:sched_switch`)
- probe event:
  - user-defined events dynamically inserted into kernel
  - created using uprobes or kprobes

# Some Events

Event	Description
cache-misses	cache misses
cache-references	cache accesses
cycles	CPU cycles
cpu-clock	CPU wall-time clock
instructions	CPU instructions
cs	context switches
faults	page faults

- stack trace is list of stack frames

# Event-Based Sampling

- with event-based sampling, sampling process driven by one or more types of events
- sample is taken upon occurrence of every  $n$ th event, where  $n$  is either:
  - directly specified by user; or
  - dynamically chosen by kernel in order to (approximately) meet average sampling rate specified by user
- default sampling event is cycles with average sampling rate that depends on Perf version (typically 1000 Hz to 4000 Hz)
- cycles event does not necessarily have constant relationship with time, due to CPU frequency scaling
- each sample captures:
  - timestamp
  - CPU number, process ID (PID), and thread ID (TID)
  - instruction pointer
  - stack trace (optional)
- can perform sampling:
  - system wide, per processor, per program, or per thread



# Event Specifiers

- event specifier consists of event name, optionally followed by colon and then one of more event modifiers
- list of event modifiers as follows:

Modifier	Description
u	user-space counting
k	kernel counting
h	hypervisor counting
i	non-idle counting
G	guest counting (in KVM guests)
H	host counting (not in KVM guests)
p	preciseness level (i.e., amount of skid)
S	read sample value
D	pin event to PMU

# Event Specifiers (Continued)

- number  $n$  of p's in modifier influences preciseness of event measurement as follows:

$n$	Description
0	can have arbitrary skid
1	must have constant skid
2	requested (but not required) to have zero skid
3	must have zero skid

- if zero skid required but not supported, error will be generated
- some examples of event specifiers are as follows:

Event Specifier	Meaning
<code>cycles:u</code>	clock cycles in user space
<code>cache-misses:u</code>	cache misses in user space
<code>cache-misses:k</code>	cache misses in kernel
<code>cache-misses:uppp</code>	cache misses in user space with zero skid

- measurements involving hardware counters typically employ interrupts
- when hardware counter for event overflows, interrupt occurs
- when overflow interrupt occurs, takes CPU some amount of time to stop processor and pinpoint exactly which instruction was active at time of overflow
- due to this delay, can often be offset in execution flow between instruction claimed to be active at time of overflow and instruction that actually was active
- this offset known as **skid**
- in some cases, for example, skid could result in caller function event being recorded in callee function
- due to skid, some care must be taken when interpreting profiling results

# The perf Program

- functionality of Perf software provided by `perf` program
- command line interface has following form:

```
perf [options] command [args]
```

- some common commands include:

Command	Description
<code>list</code>	list all symbolic event types
<code>stat</code>	run command and gather performance count statistics
<code>record</code>	run command and record its profile into Perf data file
<code>report</code>	read Perf data (created by Perf record) and display profile
<code>script</code>	read Perf data (created by Perf record) and display trace output
<code>annotate</code>	read Perf data (created by Perf record) and display annotated code

- some common options include:

Option	Description
<code>--help</code>	print help information and exit
<code>--version</code>	print version information and exit

# Perf List Command

- list all symbolic event types
- command line interface has following form:

```
perf list [event_type]
```

- event types include:

Event Type	Description
hw	hardware
sw	software
cache	cache
tracepoint	tracepoint
pmu	PMU
<i>glob_expr</i>	any event matching glob expression <i>glob_expr</i>

- only lists event types available to invoking user
- some events only available to root user

# Perf List Example

```
$ perf list
List of pre-defined events (to be used in -e):
branch-instructions OR branches      [Hardware event]
branch-misses                        [Hardware event]
bus-cycles                           [Hardware event]
cache-misses                         [Hardware event]
cache-references                     [Hardware event]
cpu-cycles OR cycles                 [Hardware event]
instructions                         [Hardware event]
ref-cycles                           [Hardware event]
[text deleted]
alignment-faults                     [Software event]
context-switches OR cs               [Software event]
cpu-clock                            [Software event]
cpu-migrations OR migrations         [Software event]
[text deleted]
L1-dcache-load-misses                [Hardware cache event]
L1-dcache-loads                      [Hardware cache event]
L1-dcache-prefetch-misses            [Hardware cache event]
L1-dcache-store-misses               [Hardware cache event]
L1-dcache-stores                     [Hardware cache event]
L1-icache-load-misses                [Hardware cache event]
[text deleted]
cache-misses OR cpu/cache-misses/    [Kernel PMU event]
cache-references OR cpu/cache-references/ [Kernel PMU event]
cpu-cycles OR cpu/cpu-cycles/        [Kernel PMU event]
instructions OR cpu/instructions/    [Kernel PMU event]
mem-loads OR cpu/mem-loads/          [Kernel PMU event]
mem-stores OR cpu/mem-stores/        [Kernel PMU event]
[text deleted]
```

- run command and gather performance count statistics
- command line interface has following form:

```
perf stat [options] command [args]
```

- some common options include:

Option	Description
-e <i>event</i>	specify event for which to gather statistics
-p <i>pid</i>	consider events on existing process ID
-t <i>tid</i>	consider events on existing thread ID
-a	consider all processors (i.e., system wide)
-r <i>n</i>	repeat command <i>n</i> times and print averages and standard deviations
-C <i>cpu</i>	consider only CPUs specified by <i>cpu</i>

# Perf Stat Example

```
$ perf stat dd if=/dev/urandom of=/dev/null bs=1K count=32K status=none
Performance counter stats for
'dd if=/dev/urandom of=/dev/null bs=1K count=32K status=none':

    1727.055828      task-clock (msec)      #    0.999 CPUs utilized
           1        context-switches      #    0.001 K/sec
          13        cpu-migrations        #    0.008 K/sec
          60        page-faults           #    0.035 K/sec
  5,805,261,702      cycles                #    3.361 GHz
  2,115,865,103      stalled-cycles-frontend #   36.45% frontend cycles idle
<not supported>    stalled-cycles-backend
 12,108,757,065      instructions          #    2.09  insns per cycle
                                     #    0.17  stalled cycles per insn
 254,471,634        branches                # 147.344 M/sec
   257,282          branch-misses           #    0.10% of all branches

1.728232622 seconds time elapsed
```



# Perf Record Command

- run command and record its profile into Perf data file
- command line interface has following form:

```
perf record [options] command [args]
```

- some common options include:

Option	Description
-e <i>event</i>	specify event name
-a	collect data from all processors
-p <i>pid</i>	collect data from existing process ID <i>pid</i>
-t <i>tid</i>	collect data from existing thread ID <i>tid</i>
-C <i>cpu</i>	collect data from CPUs <i>cpu</i>
-c <i>count</i>	set event count between samples to <i>count</i>
-o <i>file</i>	set output file to <i>file</i>
-F <i>freq</i>	set sampling frequency to approximately <i>freq</i>
-g	enable call graph (i.e., stack trace) recording

- output file defaults to `perf.data`
- by default, uses `cycles` event with sampling frequency set to version-dependent value (typically, 1000 Hz to 4000 Hz)

# Perf Record Example

```
$ perf record -g -F 99 -o perf.data dd if=/dev/urandom of=/dev/null \
  bs=1K count=3200K status=none
[ perf record: Woken up 9 times to write data ]
[ perf record: Captured and wrote 2.121 MB perf.data (16246 samples) ]
$ ls
perf.data
```

# Perf Report Command

- read Perf data (created by Perf record) and display profile
- command line interface has following form:

```
perf report [options]
```

- some common options include:

Option	Description
<code>-i <i>file</i></code>	set input file to <i>file</i>
<code>-v</code>	increase verbosity level
<code>-n</code>	show number of samples for each symbol
<code>-C <i>cpu</i></code>	only show events for CPU <i>cpu</i>
<code>--pid <i>pid</i></code>	only show events for process ID <i>pid</i>
<code>--tid <i>tid</i></code>	only show events for thread ID <i>tid</i>
<code>-d <i>dsos</i></code>	only consider symbols in DSO/object files <i>dsos</i>
<code>-S <i>syms</i></code>	only consider symbols <i>syms</i>
<code>-s <i>key</i></code>	sort data by key <i>key</i> (such as PID)
<code>--stdio</code>	use stdio interface
<code>-U</code>	only display entries that resolve to symbol
<code>-D</code>	dump raw trace data

- input file defaults to `perf.data`

# Perf Report Example

```
$ perf record -g -e cycles:u -F 13000 -o perf.data ./array_sum
1
1
$ perf report -i perf.data -d array_sum --stdio
# To display the perf.data header info, please use --header/--header-only options.
#
# dso: array_sum
# Samples: 1K of event 'cycles:u'
# Event count (approx.): 158559166
#
# Children      Self    Command      Symbol
# .....      .....      .....      .....
#
# 79.94%       79.94%   array_sum    [.] naive_sum
#           |
#           |--naive_sum
#             main
#             __libc_start_main
#             0x48e258d4c544155
#
# 7.56%       7.56%   array_sum    [.] improved_sum
#           |
#           ---improved_sum
#             main
#             __libc_start_main
#             0x48e258d4c544155
#
#
# (For a higher level overview, try: perf report --sort comm,dso)
#
```

- read Perf data (created by Perf record) and display trace output
- command line interface has following form:

```
perf script [options]
```

- some common options include:

Option	Description
<code>-i <i>file</i></code>	set input file to <i>file</i>
<code>--pid <i>pid</i></code>	only show events for process ID <i>pid</i>
<code>--tid <i>tid</i></code>	only show events for thread ID <i>tid</i>
<code>-C <i>cpu</i></code>	only show events for CPU <i>cpu</i>

- input file defaults to `perf.data`

# Perf Script Example

```
$ perf record -g -e cycles:u -F 13000 -o perf.data ./array_sum
1
1
$ perf script -i perf.data
array_sum 15602 2408817.214222:          1 cycles:u:
                cf0 _start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214230:          1 cycles:u:
                cf0 _start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214234:          2 cycles:u:
                cf0 _start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214237:          7 cycles:u:
                cf0 _start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214241:         25 cycles:u:
                cf0 _start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214245:         88 cycles:u:
                cf0 _start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214248:        308 cycles:u:
                cf0 _start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214253:       1081 cycles:u:
ffffffffff8179bef0 page_fault ([kernel.kallsyms])
                cf0 _start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214270:        3147 cycles:u:
4980 _dl_start (/usr/lib64/ld-2.20.so)

array_sum 15602 2408817.214274:        4536 cycles:u:
4b8f _dl_start (/usr/lib64/ld-2.20.so)
cf8 _dl_start_user (/usr/lib64/ld-2.20.so)

[text deleted]
```

# Perf Annotate Command

- read Perf data (created by Perf record) and display annotated code
- command line interface has following form:

```
perf annotate [options]
```

- some common options include:

Option	Description
-i <i>file</i>	set input file to <i>file</i>
-s <i>sym</i>	annotate symbol <i>sym</i>
-d <i>dsos</i>	only consider symbols in DSO/object files <i>dsos</i>
-v	increase verbosity level
-l	print matching source lines
-P	do not shorten displayed pathnames
-k <i>file</i>	set vmlinux pathname to <i>file</i>
--stdio	use stdio interface
--no-source	disable displaying of source code

- input file defaults to `perf.data`

# Perf Annotate Example

```
$ perf record -g -e cycles:u -F 13000 -o perf.data ./array_sum
```

```
[text deleted]
```

```
$ perf annotate -i perf.data -s naive_sum -l --stdio
```

```
[text deleted]
```

```
      : double naive_sum(const double a[][N]) {
0.00 :      400807:      push    %rbp
0.00 :      400808:      mov     %rsp,%rbp
0.00 :      40080b:      lea     0x4000(%rdi),%rcx
      :      double sum = 0.0;
0.00 :      400812:      pxor    %xmm0,%xmm0
0.00 :      400816:      lea     0x2000000(%rdi),%rdx
0.00 :      40081d:      mov     %rdi,%rax
      :      for (int j = 0; j < N; ++j) {
      :          for (int i = 0; i < M; ++i) {
      :              sum += a[i][j];
0.00 :      400820:      addsd   (%rax),%xmm0
array_sum.cpp:11 100.00 :      400824:      add     $0x4000,%rax
[text deleted]
      : double naive_sum(const double a[][N]) {
      :     double sum = 0.0;
      :     for (int j = 0; j < N; ++j) {
      :         for (int i = 0; i < M; ++i) {
0.00 :      40082a:      cmp     %rdx,%rax
0.00 :      40082d:      jne     400820 <naive_sum(double const (*) [2048])+0x19>
0.00 :      40082f:      add     $0x8,%rdi
[text deleted]
0.00 :      400833:      cmp     %rcx,%rdi
0.00 :      400836:      jne     400816 <naive_sum(double const (*) [2048])+0xf>
[text deleted]
      :     }
      :     }
      :     return sum;
      : }
0.00 :      400838:      pop     %rbp
0.00 :      400839:      retq
```



# Example: Source Code

```
1  #include <iostream>
2  #include <algorithm>
3
4  constexpr int M = 4096;
5  constexpr int N = 4096;
6
7  [[gnu::noinline]]
8  double naive_sum(const double a[][N]) {
9      double sum = 0.0;
10     for (int j = 0; j < N; ++j) {
11         for (int i = 0; i < M; ++i) {
12             sum += a[i][j];
13         }
14     }
15     return sum;
16 }
17
18 [[gnu::noinline]]
19 double improved_sum(const double a[][N]) {
20     double sum = 0.0;
21     for (int i = 0; i < M; ++i) {
22         for (int j = 0; j < N; ++j) {
23             sum += a[i][j];
24         }
25     }
26     return sum;
27 }
```

## Example: Source Code (Continued)

```
29 int main() {
30     for (int i = 0; i < 16; ++i) {
31         static double a[M][N];
32         static double b[M][N];
33         std::fill_n(&a[0][0], M * N, 1.0 / (M * N));
34         std::fill_n(&b[0][0], M * N, 1.0 / (M * N));
35         std::cout << naive_sum(a) << ' ';
36         std::cout << improved_sum(b) << '\n';
37     }
38 }
```

# Profile of Cycles

```
# To display the perf.data header info, please use --header/--header-only options.
#
# dso: array_sum
# Samples: 16K of event 'cycles:u'
# Event count (approx.): 14049539983
#
# Children      Self    Command      Symbol
# .....      .....      .....      .....
#
# 99.97%        0.00%   array_sum   [.] __libc_start_main
#           |
#           |--__libc_start_main
#             0x46e258d4c544155
#
# 99.97%        10.92%   array_sum   [.] main
#           |
#           ---main
#             __libc_start_main
#             0x46e258d4c544155
#
# 82.97%        82.97%   array_sum   [.] naive_sum
#           |
#           ---naive_sum
#             main
#             __libc_start_main
#             0x46e258d4c544155
#
# 5.90%         5.90%   array_sum   [.] improved_sum
#           |
#           ---improved_sum
#             main
#             __libc_start_main
#             0x46e258d4c544155
#
# [text deleted]
```

# Cycles for naive\_sum

```
    : 0000000000400807 <naive_sum(double const (*) [4096])>:
    : _Z9naive_sumPA4096_Kd():
[text deleted]
    : [[gnu::noinline]]
    : double naive_sum(const double a[][N]) {
0.00 : 400807: push %rbp
0.00 : 400808: mov %rsp,%rbp
0.00 : 40080b: lea 0x8000(%rdi),%rcx
    : double sum = 0.0;
0.00 : 400812: pxor %xmm0,%xmm0
0.00 : 400816: lea 0x8000000(%rdi),%rdx
0.00 : 40081d: mov %rdi,%rax
    : for (int j = 0; j < N; ++j) {
    : for (int i = 0; i < M; ++i) {
    : sum += a[i][j];
0.00 : 400820: addsd (%rax),%xmm0
array_sum.cpp:12 99.93 : 400824: add $0x8000,%rax
[text deleted]
    : for (int j = 0; j < N; ++j) {
    : for (int i = 0; i < M; ++i) {
0.07 : 40082a: cmp %rdx,%rax
0.00 : 40082d: jne 400820 <naive_sum(double const (*) [4096])+0x19>
0.00 : 40082f: add $0x8,%rdi
[text deleted]
    : [[gnu::noinline]]
    : double naive_sum(const double a[][N]) {
    : double sum = 0.0;
    : for (int j = 0; j < N; ++j) {
0.00 : 400833: cmp %rcx,%rdi
0.00 : 400836: jne 400816 <naive_sum(double const (*) [4096])+0xf>
    : for (int i = 0; i < M; ++i) {
    : sum += a[i][j];
    : }
    : }
    : return sum;
    : }
0.00 : 400838: pop %rbp
0.00 : 400839: retq
```

# Cycles for improved\_sum

```
      : 000000000040083a <improved_sum(double const (*) [4096])>:
      : _Z12improved_sumPA4096_Kd():
[text deleted]
      : [[gnu::noinline]]
      : double improved_sum(const double a[][N]) {
0.00 : 40083a: push    %rbp
0.00 : 40083b: mov     %rsp,%rbp
0.00 : 40083e: lea     0x8000000(%rdi),%rdx
      :     double sum = 0.0;
0.00 : 400845: pxor    %xmm0,%xmm0
0.00 : 400849: lea     0x8000(%rdi),%rax
      :     for (int i = 0; i < M; ++i) {
      :         for (int j = 0; j < N; ++j) {
      :             sum += a[i][j];
0.00 : 400850: addsd   (%rdi),%xmm0
array_sum.cpp:23 99.70 : 400854: add     $0x8,%rdi
[text deleted]
      :     for (int i = 0; i < M; ++i) {
      :         for (int j = 0; j < N; ++j) {
0.30 : 400858: cmp     %rax,%rdi
0.00 : 40085b: jne     400850 <improved_sum(double const (*) [4096])+0x16>
      :     }
      :
      : [[gnu::noinline]]
      : double improved_sum(const double a[][N]) {
      :     double sum = 0.0;
      :     for (int i = 0; i < M; ++i) {
0.00 : 40085d: cmp     %rdx,%rdi
0.00 : 400860: jne     400849 <improved_sum(double const (*) [4096])+0xf>
      :         for (int j = 0; j < N; ++j) {
      :             sum += a[i][j];
      :         }
      :     }
      :     return sum;
      : }
0.00 : 400862: pop     %rbp
0.00 : 400863: retq
```

# Profile of Cache Misses

```
# To display the perf.data header info, please use --header/--header-only options.
#
# dso: array_sum
# Samples: 25K of event 'cache-misses:u'
# Event count (approx.): 256620000
#
# Children      Self    Command      Symbol
# .....      .....      .....      .....
#
# 99.99%       0.00%   array_sum   [.] __libc_start_main
#           |
#           |--__libc_start_main
#             0x46e258d4c544155
#
# 99.99%       3.67%   array_sum   [.] main
#           |
#           ---main
#             __libc_start_main
#             0x46e258d4c544155
#
# 93.74%      93.73%   array_sum   [.] naive_sum
#           |
#           ---naive_sum
#             main
#             __libc_start_main
#             0x46e258d4c544155
#
#  2.58%       2.58%   array_sum   [.] improved_sum
#           |
#           ---improved_sum
#             main
#             __libc_start_main
#             0x46e258d4c544155
#
# [text deleted]
```

# Cache Misses for naive\_sum

```
      : 0000000000400807 <naive_sum(double const (*) [4096])>:
      : _Z9naive_sumPA4096_Kd():
[text deleted]
      : [[gnu::noinline]]
      : double naive_sum(const double a[][N]) {
0.00 : 400807: push    %rbp
0.00 : 400808: mov     %rsp,%rbp
0.00 : 40080b: lea     0x8000(%rdi),%rcx
      : double sum = 0.0;
0.00 : 400812: pxor    %xmm0,%xmm0
0.00 : 400816: lea     0x8000000(%rdi),%rdx
0.00 : 40081d: mov     %rdi,%rax
      : for (int j = 0; j < N; ++j) {
      :     for (int i = 0; i < M; ++i) {
      :         sum += a[i][j];
0.00 : 400820: addsd   (%rax),%xmm0
array_sum.cpp:12 99.93 : 400824: add     $0x8000,%rax
[text deleted]
      :     for (int j = 0; j < N; ++j) {
      :         for (int i = 0; i < M; ++i) {
0.07 : 40082a: cmp     %rdx,%rax
0.00 : 40082d: jne     400820 <naive_sum(double const (*) [4096])+0x19>
0.00 : 40082f: add     $0x8,%rdi
[text deleted]
      : [[gnu::noinline]]
      : double naive_sum(const double a[][N]) {
      :     double sum = 0.0;
      :     for (int j = 0; j < N; ++j) {
0.00 : 400833: cmp     %rcx,%rdi
0.00 : 400836: jne     400816 <naive_sum(double const (*) [4096])+0xf>
      :         for (int i = 0; i < M; ++i) {
      :             sum += a[i][j];
      :         }
      :     }
      :     return sum;
      : }
0.00 : 400838: pop     %rbp
0.00 : 400839: retq
```

# Cache Misses for improved\_sum

```
: 000000000040083a <improved_sum(double const (*) [4096])>:
: _Z12improved_sumPA4096_Kd():
[text deleted]
: [[gnu::noinline]]
: double improved_sum(const double a[][N]) {
0.00 : 40083a: push    %rbp
0.00 : 40083b: mov     %rsp,%rbp
0.00 : 40083e: lea     0x8000000(%rdi),%rdx
:         double sum = 0.0;
0.00 : 400845: pxor    %xmm0,%xmm0
0.00 : 400849: lea     0x8000(%rdi),%rax
:         for (int i = 0; i < M; ++i) {
:             for (int j = 0; j < N; ++j) {
:                 sum += a[i][j];
0.00 : 400850: addsd   (%rdi),%xmm0
array_sum.cpp:23 99.70 : 400854: add     $0x8,%rdi
[text deleted]
:         for (int i = 0; i < M; ++i) {
:             for (int j = 0; j < N; ++j) {
0.30 : 400858: cmp     %rax,%rdi
0.00 : 40085b: jne     400850 <improved_sum(double const (*) [4096])+0x16>
:         }
:     }
: [[gnu::noinline]]
: double improved_sum(const double a[][N]) {
:     double sum = 0.0;
:     for (int i = 0; i < M; ++i) {
0.00 : 40085d: cmp     %rdx,%rdi
0.00 : 400860: jne     400849 <improved_sum(double const (*) [4096])+0xf>
:         for (int j = 0; j < N; ++j) {
:             sum += a[i][j];
:         }
:     }
:     return sum;
: }
0.00 : 400862: pop     %rbp
0.00 : 400863: retq
```

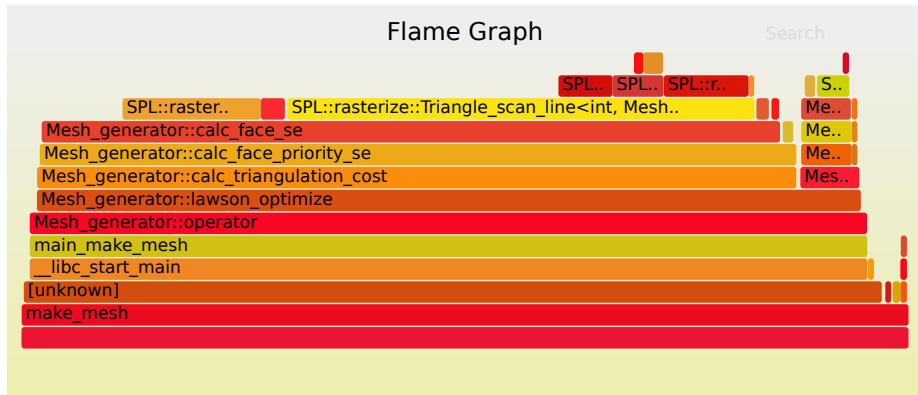


- avoid sampling in lockstep with periodic behavior exhibited by programs (e.g., caused by timeouts or loops)
- since programmers often choose timeout (and other timing related) values to be “nice” numbers, such as integer multiples of 0.01 s, may be beneficial to choose sampling frequency of 99 Hz instead of 100 Hz or 999 Hz instead of 1000 Hz
- sample-based profiling only provides meaningful results if sufficient number of samples collected
- can use `taskset` command to pin process for particular CPU
- might want to force single-threaded program to run on fixed CPU so that migration does not impact measurements (e.g., due to caching effects)

# Flame Graphs

- flame graph provides way to visualize collection of stack traces
- useful for visualizing output of profiler that collects stack traces using sampling (e.g., Perf)
- stack trace represented as column of boxes, with each box corresponding to function in stack trace
- function executing at time of stack trace shown at top of column
- vertical direction corresponds to stack depth
- horizontal direction spans stack trace collection (does not represent time)
- left to right ordering has no special meaning
- when identical function boxes horizontally adjacent, merged
- width of each function box shows frequency with which function present in part of stack trace ancestry
- functions with wider boxes more frequent in stack traces than those with narrower boxes

# Flame Graph Example



# Generating Flame Graphs

- can generate flamegraphs from Perf data by using software available from
  - <https://github.com/brendangregg/FlameGraph>
- need to use `stackcollapse-perf.pl` and `flamegraph.pl` programs
- convert Perf data from binary to text format via Perf script command; for example:

```
perf script -i perf.data > tmp.perf
```
- fold stack samples into single lines via `stackcollapse-perf.pl` command; for example:

```
stackcollapse-perf.pl tmp.perf > tmp.folded
```
- generate flame graph in SVG format via `flamegraph.pl` command; for example:

```
flamegraph.pl tmp.folded > flamegraph.svg
```

## Section 8.1.1

### **References**

- 1 Brendan Gregg. Linux Profiling at Netflix. Southern California Linux Expo (SCaLE), Los Angeles, CA, USA, Feb. 27, 2015. Available online at [https://youtu.be/\\_Ik8oiQvWgo](https://youtu.be/_Ik8oiQvWgo).
- 2 Mans Rullgard. Performance Analysis Using the perf Suite. Embedded Linux Conference, March 2015, San Jose, CA, USA. Available online at <https://youtu.be/kWnx6eOGVYo>.
- 3 Chandler Carruth. Tuning C++: Benchmarks, and CPUs, and Compilers! Oh My!. CppCon, Bellevue, WA, USA, Sept. 24, 2015. Available online at <https://youtu.be/nXaxk27zwlk>.
- 4 Brendan Gregg. Blazing Performance with Flame Graphs. Large Installation System Administration Conference (LISA), Washington, DC, USA, Nov. 2013. Available online at <https://youtu.be/nZfNehCzGdw>.

- 1 B. Gregg, The Flame Graph. ACM Queue, March 2016, pages 1–28.

## Section 8.2

### **Performance API (PAPI) Software**



- often easy to identify in general terms which parts of code are slow
- sometimes more difficult to pinpoint precise reason why code is slow (i.e., what is precise cause of bottleneck)
- often need to consider factors such as:
  - cache behavior
  - memory and resource contention
  - floating-point efficiency
  - branch behavior
- often, processor itself in best position to provide information related to above factors

# Hardware Performance Counters

- hardware performance counters are specialized registers used to measure various aspects of processor performance
- hardware counters can provide insight into:
  - timing
  - cache behaviors (e.g., cache misses and cache coherence protocol events)
  - branch behaviors (e.g., incorrect branch predictions)
  - pipeline behavior (e.g., stalls)
  - memory and resource access patterns
  - floating-point efficiency
  - instructions per cycle
- hardware counter information can be obtained with:
  - subroutine or basic block resolution
  - process or thread attribution
- provide low-level information that often cannot be obtained easily through other means
- useful for performance analysis and tuning (e.g., identifying bottlenecks in code)
- use of hardware performance counters has no or little overhead

# Performance API (PAPI) Software

- Performance API (PAPI) software provides portable and efficient API for accessing hardware performance counters found on modern processors
- more generally allows monitoring of system information on range of components, such as CPUs, network interface cards, and power monitors
- consists of library and several utility programs
- open source
- written in C
- supports most mainstream Unix-based operating systems (e.g., Linux, OS X, and other Unix variants); older versions support Microsoft Windows
- supports most modern processors (e.g., Intel and AMD 32- and 64-bit x86, ARM, MIPS, Intel Itanium II, UltraSparc I, II, and III, and IBM Power 4, 5, 6, and 7)
- web site: <http://icl.utk.edu/papi>

- event is simply some action that can be counted
- **native event**: event that is specified in platform-dependent manner and directly corresponds to particular hardware counter
- which native events are available will depend on underlying hardware
- **preset event**: event that is specified in platform-independent manner, which is then mapped to appropriate native event(s) (e.g., `PAPI_TOT_INS`)
- **derived event**: preset event derived from multiple native events
- if hardware does not directly support counting of particular event, event count can sometimes be computed by using combination of native events
- for example, `PAPI_L1_TCM` might be derived from L1 data misses plus L1 instruction misses
- preset events usually available for most processors, where derived events used in cases where no corresponding native event exists

## Events (Continued)

- which events supported and which combinations of supported events can be used together depends on hardware
- hardware will typically have some upper limit on number of events that can be monitored simultaneously
- some events often cannot be used with others (even if upper limit on number of events not exceeded)
- `papi_avail` or `papi_native_avail` utility (discussed later) can be used to determine number of hardware counters available
- `papi_avail` utility (discussed later) can be used to determine which preset events are supported
- `papi_native_avail` utility (discussed later) can be used to determine which native events are supported
- `papi_event_chooser` utility (discussed later) can be used to determine which events can be used with which other events

- must include header file `papi.h`
- library initialized with function `PAPI_library_init`
- depending on which functions used, may need to explicitly initialize library

- calls low-level API
- easier to use than low-level API
- usually enough for more basic measurements
- for preset events only
- high-level interface functions will initialize library if needed (so `PAPI_library_init` need not be explicitly called)

# Functions in PAPI High-Level Interface

Function	Description
PAPI_accum_counters	add current counts to array and reset counters
PAPI_flips	get floating-point instruction rate and real and processor time
PAPI_flops	get floating-point operation rate and real and processor time
PAPI_ipc	get instructions per cycle and real and processor time
PAPI_num_counters	get number of hardware counters available on system
PAPI_read_counters	copy current counts to array and reset counters
PAPI_start_counters	start counting hardware events
PAPI_stop_counters	stop counters and return current counts



# Some Commonly-Used Preset Events

## Instruction Mix

Name	Description
PAPI_LD_INS	number of load instructions
PAPI_SR_INS	number of store instructions
PAPI_LST_INS	number of load/store instructions
PAPI_BR_INS	number of branch instructions
PAPI_INT_INS	number of integer instructions
PAPI_FP_INS	number of floating-point instructions
PAPI_VEC_INS	number of vector/SIMD instructions
PAPI_VEC_SP	number of single-precision vector/SIMD instructions
PAPI_VEC_DP	number of double-precision vector/SIMD instructions
PAPI_TOT_INS	number of instructions in total

# Some Commonly-Used Preset Events (Continued 1)

## Clock Cycles

Name	Description
PAPI_TOT_CYC	total number of clock cycles

## FLOPS

Name	Description
PAPI_FP_OPS	number of floating-point operations
PAPI_SP_OPS	number of floating-point operations executed, optimized to count scaled single-precision vector operations
PAPI_DP_OPS	number of floating-point operations executed, optimized to count scaled double-precision vector operations

## Translation Lookaside Buffer (TLB)

Name	Description
PAPI_TLB_DM	number of data TLB misses
PAPI_TLB_IM	number of instruction TLB misses
PAPI_TLB_TL	number of TLB misses (in total)

# Some Commonly-Used Preset Events (Continued 2)

## L1 Cache Behavior

Name	Description
PAPI_L1_DCA	number of L1 data cache accesses
PAPI_L1_DCH	number of L1 data cache hits
PAPI_L1_DCM	number of L1 data cache misses
PAPI_L1_DCR	number of L1 data cache reads
PAPI_L1_DCW	number of L1 data cache writes
PAPI_L1_ICA	number of L1 instruction cache accesses
PAPI_L1_ICH	number of L1 instruction cache hits
PAPI_L1_ICM	number of L1 instruction cache misses
PAPI_L1_ICR	number of L1 instruction cache reads
PAPI_L1_ICW	number of L1 instruction cache writes
PAPI_L1_LDM	number of L1 load misses
PAPI_L1_STM	number of L1 store misses
PAPI_L1_TCA	number of L1 cache accesses (in total)
PAPI_L1_TCH	number of L1 cache hits (in total)
PAPI_L1_TCM	number of L1 cache misses (in total)
PAPI_L1_TCR	number of L1 cache reads (in total)
PAPI_L1_TCW	number of L1 cache writes (in total)

# Some Commonly-Used Preset Events (Continued 3)

## L2 and L3 Cache Behavior

Name	Description
PAPI_L2_LDM	number of L2 load misses
PAPI_L2_STM	number of L2 store misses
PAPI_L2_TCA	number of L2 cache accesses (in total)
PAPI_L2_TCH	number of L2 cache hits (in total)
PAPI_L2_TCM	number of L2 cache misses (in total)
PAPI_L2_TCR	number of L2 cache reads (in total)
PAPI_L2_TCW	number of L2 cache writes (in total)
PAPI_L3_LDM	number of L3 load misses
PAPI_L3_STM	number of L3 store misses
PAPI_L3_TCA	number of L3 cache accesses (in total)
PAPI_L3_TCH	number of L3 cache hits (in total)
PAPI_L3_TCM	number of L3 cache misses (in total)
PAPI_L3_TCR	number of L3 cache reads (in total)
PAPI_L3_TCW	number of L3 cache writes (in total)

- most frequently used events are often those related to cache behavior
- instructions per cycle could be computed from events:
  - `PAPI_TOT_CYC` and `PAPI_TOT_INS`
- L1 cache data miss rate could be computed from events:
  - `PAPI_L1_DCM` and `PAPI_L1_DCA`; or
  - `PAPI_L1_DCM` and `PAPI_L1_DCH`; or
  - `PAPI_L1_DCM`, `PAPI_LD_INS`, and `PAPI_SR_INS`
- L2 cache (total) miss rate could be computed from events:
  - `PAPI_L2_TCM` and `PAPI_L2_TCA`; or
  - `PAPI_L2_TCM` and `PAPI_L2_TCH`; or
  - `PAPI_L2_TCM`, `PAPI_LD_INS`, and `PAPI_SR_INS`

# Code Example Using PAPI High-Level Interface

```
1  #include <iostream>
2  #include <papi.h>
3
4  void do_work() {for (volatile auto i = 1'000'000; i > 0; --i) {}}
5
6  int main() {
7      constexpr int num_events = 2;
8      int events[num_events] = {PAPI_TOT_INS, PAPI_TOT_CYC};
9      long long values[num_events];
10     if (PAPI_start_counters(events, num_events) != PAPI_OK)
11         {std::cerr << "cannot start counters\n"; return 1;}
12     do_work();
13     if (PAPI_stop_counters(values, num_events) != PAPI_OK)
14         {std::cerr << "cannot stop counters\n"; return 1;}
15     for (auto i : values) {std::cout << i << '\n';}
16 }
```

# PAPI Low-Level Interface

- several dozen functions available in low-level API
- provides increased efficiency and functionality
- can obtain more detailed information about hardware
- low-level interface works with event sets
- **event set**: set of events to be monitored
- some low-level API functions listed on next slide
- low-level interface functions do not initialize library (so `PAPI_library_init` must be called explicitly)

# Some Functions in PAPI Low-Level Interface

Function	Description
PAPI_library_init	initialize PAPI library
PAPI_shutdown	cleanup PAPI library
PAPI_create_eventset	create event set
PAPI_destroy_eventset	destroys empty event set
PAPI_cleanup_eventset	removes all events from event set
PAPI_add_event	add preset or native hardware event to event set
PAPI_add_events	add multiple preset or native hardware events to event set
PAPI_start	start counting hardware events in event set
PAPI_read	read hardware counters from event set
PAPI_reset	reset hardware event counts in event set
PAPI_accum	adds hardware counters from event set to elements in array and resets counters
PAPI_stop	stop counting hardware events in event set



# Code Example Using PAPI Low-Level Interface

```
1  #include <iostream>
2  #include <papi.h>
3
4  void do_work() {for (volatile auto i = 1'000'000; i > 0; --i) {}}
5
6  int main() {
7      constexpr int num_events = 2;
8      int event_set = PAPI_NULL;
9      int events[num_events] = {PAPI_TOT_INS, PAPI_TOT_CYC};
10     long long values[num_events];
11     if (PAPI_library_init(PAPI_VER_CURRENT) != PAPI_VER_CURRENT)
12         {std::cerr << "cannot initialize\n"; return 1;}
13     if (PAPI_create_eventset(&event_set) != PAPI_OK)
14         {std::cerr << "cannot create event set\n"; return 1;}
15     if (PAPI_add_events(event_set, events, num_events) != PAPI_OK)
16         {std::cerr << "cannot add events\n"; return 1;}
17     if (PAPI_start(event_set) != PAPI_OK)
18         {std::cerr << "cannot start\n"; return 1;}
19     do_work();
20     if (PAPI_stop(event_set, values) != PAPI_OK)
21         {std::cerr << "cannot stop\n"; return 1;}
22     if (PAPI_cleanup_eventset(event_set) != PAPI_OK)
23         {std::cerr << "cannot cleanup event set\n"; return 1;}
24     if (PAPI_destroy_eventset(&event_set) != PAPI_OK)
25         {std::cerr << "cannot destroy event set\n"; return 1;}
26     for (auto i : values) {std::cout << i << '\n';}
27 }
```

# PAPI Utilities

Name	Description
<code>papi_avail</code>	provides availability and detail information for PAPI preset events
<code>papi_clockres</code>	measures and reports clock latency and resolution for PAPI timers
<code>papi_cost</code>	computes execution time costs for basic PAPI operations
<code>papi_command_line</code>	executes PAPI preset or native events from command line
<code>papi_decode</code>	provides availability and detail information for PAPI preset events
<code>papi_event_chooser</code>	given list of named events, lists other events that can be counted with them
<code>papi_mem_info</code>	provides information on memory architecture of current processor
<code>papi_native_avail</code>	provides detailed information for PAPI native events
<code>papi_version</code>	provides version information for PAPI

# Example papi\_avail Output

Available events and hardware information.

```
-----
PAPI Version           : 5.3.2.0
Vendor string and code : GenuineIntel (1)
Model string and code  : Intel(R) Core(TM) i7-3820QM CPU @ 2.70GHz (58)
CPU Revision          : 9.0000000
CPUTID Info           : Family: 6 Model: 58 Stepping: 9
CPU Max Megahertz     : 3700
CPU Min Megahertz     : 1200
Hdw Threads per core  : 2
Cores per Socket      : 4
Sockets               : 1
NUMA Nodes            : 1
CPUs per Node         : 8
Total CPUs            : 8
Running in a VM        : no
Number Hardware Counters : 11
Max Multiplex Counters : 64
-----
```

Name	Code	Avail	Deriv	Description (Note)
PAPI_L1_DCM	0x80000000	Yes	No	Level 1 data cache misses
PAPI_L1_ICM	0x80000001	Yes	No	Level 1 instruction cache misses
PAPI_L2_DCM	0x80000002	Yes	Yes	Level 2 data cache misses
PAPI_L2_ICM	0x80000003	Yes	No	Level 2 instruction cache misses
PAPI_L3_DCM	0x80000004	No	No	Level 3 data cache misses
PAPI_L3_ICM	0x80000005	No	No	Level 3 instruction cache misses
PAPI_L1_TCM	0x80000006	No	Yes	Level 1 cache misses
PAPI_L2_TCM	0x80000007	Yes	No	Level 2 cache misses
PAPI_L3_TCM	0x80000008	Yes	No	Level 3 cache misses

[99 lines deleted]

-----  
Of 108 possible events, 43 are available, of which 14 are derived.

avail.c

PASSED

# Example papi\_mem\_info Output

Memory Cache and TLB Hierarchy Information.

-----  
TLB Information.

There may be multiple descriptors for each level of TLB  
if multiple page sizes are supported.

L1 Data TLB:

Page Size:	4 KB
Number of Entries:	64
Associativity:	4

[other TLB information deleted]

Cache Information.

L1 Data Cache:

Total size:	32 KB
Line size:	64 B
Number of Lines:	512
Associativity:	8

L1 Instruction Cache:

Total size:	32 KB
Line size:	64 B
Number of Lines:	512
Associativity:	8

L2 Unified Cache:

Total size:	256 KB
Line size:	64 B
Number of Lines:	4096
Associativity:	8

[information for L3 Unified Cache deleted]

mem\_info.c

PASSED

# Example papi\_native\_avail Output

Available native events and hardware information.

```
-----
PAPI Version           : 5.3.2.0
Vendor string and code : GenuineIntel (1)
Model string and code  : Intel(R) Core(TM) i7-3820QM CPU @ 2.70GHz (58)
CPU Revision           : 9.000000
CPUID Info             : Family: 6 Model: 58 Stepping: 9
CPU Max Megahertz      : 3700
CPU Min Megahertz      : 1200
Hdw Threads per core   : 2
Cores per Socket       : 4
Sockets                : 1
NUMA Nodes             : 1
CPUs per Node          : 8
Total CPUs             : 8
Running in a VM        : no
Number Hardware Counters : 11
Max Multiplex Counters : 64
-----
```

=====
Native Events in Component: perf\_event
=====

[lines deleted]

```
-----
| perf::L1-DCACHE-LOADS                                     |
|           L1 cache load accesses                         |
|-----|
```

[lines deleted]

=====
Native Events in Component: coretemp
=====

```
-----
| coretemp::hwmon0:temp1_input                             |
|           degrees C, acpitz module, label ?             |
|-----|
```

[lines deleted]

Total events reported: 322

native\_avail.c

PASSED

## Section 8.2.1

### **References**

- 1 B. Sprunt. The basics of performance-monitoring hardware. *IEEE Micro*, 22(4):64–71, July 2002.
- 2 P. Mucci, Performance Monitoring with PAPI, Dr. Dobbs's Journal, June 2005. Available online at <http://www.drdobbs.com/tools/performance-monitoring-with-papi/184406109>.

## Section 8.3

### Gprof



- open-source tool for code-execution profiling
- can be used to collect statistics from program run, including:
  - amount of time spent in each function
  - how many times each function called
  - callers and callees of each function (i.e., call graph information)
- based on compiler instrumentation of code and sampling
- works with GCC and Clang compilers
- instrumentation added to code gathers function call information used to generate call graphs and function call counts
- timing of code execution accomplished by statistical sampling at run time
- program counter probed at regular intervals by interrupting program
- typical sampling period 100 or 1000 samples/second

- since sampling is statistical process, timing measurements not exact (i.e., only statistical approximation)
- if *too few samples taken* (e.g., in case of short-running program), timing measurements very inaccurate
- overhead caused by instrumentation can be quite high (about 30% to 260%)
- *overhead distorts timing measurements* (e.g., instrumentation added to code changes code timing) so timing of code with and without profiling can potentially be quite different
- may not correctly handle multi-threaded applications
- cannot profile shared libraries
- cannot measure time spent in kernel mode (e.g., system calls); only user-space code profiled
- has difficulties with call graphs containing non-trivial cycles (e.g., mutual recursion)

# The gprof Command

- command line interface has following form:

```
gprof [options] [executable_file] [profile_file...]
```

- `executable_file` defaults to `a.out`
- `profile_file` defaults to `gmon.out`
- some common options include:

Option	Description
<code>-b</code>	omit explanations of meaning of all fields in output
<code>-I <i>dirs</i></code>	add directories <i>dirs</i> to search path for source files
<code>-p</code>	show flat profile
<code>-q</code>	show call graph
<code>-h</code>	print help information and exit
<code>-s</code>	summarize information in profile data files and write to <code>gmon.sum</code>

- compile and link program with `-pg` option; for example:

```
g++ -c -pg -g -O example_1.cpp
```

```
g++ -c -pg -g -O example_2.cpp
```

```
g++ -pg -g -O -o example example_1.o example_2.o
```

- run program which will produce profiling data file `gmon.out`; for example:

```
example
```

- run `gprof` to analyze profiling data; for example:

```
gprof example
```

- several `gmon` files can be combined with `gprof -s` to accumulate data over several runs of program
- `gprof2dot` can be used to convert call graph to graphical form

- output can be generated in following forms:
  - flat profile
  - call graph
- flat profile reports:
  - how much of total execution time spent in each function
  - how many times each function called
  - output sorted by percentage
- call graph reports:
  - for each function, which functions called it, which other functions it called, and how many times
  - estimate of how much time spent in subroutines of each function
- flat profile useful to identify most expensive functions
- call graph useful to identify places where function calls could be eliminated

# Example: Source Code

```
1  #include <algorithm>
2
3  constexpr int M = 1024;
4  constexpr int N = 1024;
5  constexpr int P = 1024;
6
7  // c += a * b
8  void naive_matmul(const double a[][N], const double b[][P],
9  double c[][P]) {
10     for (int i = 0; i < M; ++i) {
11         for (int j = 0; j < N; ++j) {
12             for (int k = 0; k < P; ++k)
13                 {c[i][j] += a[i][k] * b[k][j];}
14         }
15     }
16 }
17
18 // c += a * b
19 void improved_matmul(const double a[][N], const double b[][P],
20 double c[][P]) {
21     for (int i = 0; i < M; ++i) {
22         for (int k = 0; k < P; ++k) {
23             for (int j = 0; j < N; ++j)
24                 {c[i][j] += a[i][k] * b[k][j];}
25         }
26     }
27 }
```

## Example: Source Code (Continued)

```
29  int main(int argc, char** argv) {
30      static double a[M][N];
31      static double b[N][P];
32      static double c0[M][P];
33      static double c1[M][P];
34      std::fill_n(&a[0][0], M * N, 1.0);
35      std::fill_n(&b[0][0], N * P, 1.0);
36      std::fill_n(&c0[0][0], M * P, 0.0);
37      naive_matmul(a, b, c0);
38      std::fill_n(&a[0][0], M * N, 1.0);
39      std::fill_n(&b[0][0], N * P, 1.0);
40      std::fill_n(&c1[0][0], M * P, 0.0);
41      improved_matmul(a, b, c1);
42  }
```

- each row in table corresponds to function
- columns in table have following meanings:
  - % time: percentage of total running time of program used by this function
  - cumulative seconds: running sum of number of seconds accounted for by this function and those listed above it
  - self seconds: number of seconds accounted for by this function alone (i.e., excluding descendants)
  - calls: number of times this function was invoked if function is profiled, blank otherwise
  - self ms/call: average number of milliseconds spent in this function per call (excluding descendants) if function is profiled, blank otherwise
  - total ms/call: average number of milliseconds spent in this function and its descendants per call if function is profiled, blank otherwise
  - name: name of function
- entries in table sorted first by self seconds and then by function name



# Example: Flat Profile

Flat profile:

Each sample counts as 0.01 seconds.

% time	cumulative seconds	self seconds	calls	self s/call	total s/call	name
89.48	7.55	7.55	1	7.55	7.55	naive_matmul( double const (*) [1024], double const (*) [1024], double (*) [1024])
11.23	8.50	0.95	1	0.95	0.95	improved_matmul( double const (*) [1024], double const (*) [1024], double (*) [1024])
0.12	8.51	0.01				main
0.00	8.51	0.00	1	0.00	0.00	

\_\_GLOBAL\_\_sub\_I\_\_Z12naive\_matmulPA1024\_KdS1\_PA1024\_d

# Call Graph Information

- table describes call graph of program
- one multi-line entry in table per function, with each entry containing information for function and its callers and callees
- line with index number in left margin lists current function
- lines above current function list its callers
- lines below current function list its callees
- for current function, fields have following meanings:
  - index: unique integer index for this function
  - % time: percentage of total time spent in this function and its children
  - self: total amount of time spent in this function
  - children: total amount of time propagated into this function by its children
  - called: number of times this function called nonrecursively (possibly followed by “+” and number of recursive calls)
  - name: name of this function (with index printed after it)

# Call Graph Information (Continued)

- for each parent of current function, fields have following meanings:
  - self: amount of time propagated directly from function into this parent
  - children: amount of time propagated from function's children into this parent
  - called: number of times parent called function, followed by "/", followed by total number of times function called
  - name: name of this parent (with its index printed after name)
- if parents of current function cannot be determined, "<spontaneous>" is printed in name field
- for each child of current function, fields have following meanings:
  - self: amount of time propagated directly from child to current function
  - children: amount of time propagated from child's children to current function
  - called: number of times current function called child, followed by "/", followed by total number of times child called
  - name: name of function (followed by its index)

# Example: Call Graph

## Call graph

granularity: each sample hit covers 2 byte(s) for 0.12% of 8.51 seconds

index	% time	self	children	called	name
					<spontaneous>
[1]	100.0	0.01	8.50		main [1]
		7.55	0.00	1/1	naive_matmul(double const (*) [1024], double const (*) [1024], double (*) [1024]) [2]
		0.95	0.00	1/1	improved_matmul(double const (*) [1024], double const (*) [1024], double (*) [1024]) [3]
-----					
		7.55	0.00	1/1	main [1]
[2]	88.7	7.55	0.00	1	naive_matmul(double const (*) [1024], double const (*) [1024], double (*) [1024]) [2]
-----					
		0.95	0.00	1/1	main [1]
[3]	11.1	0.95	0.00	1	improved_matmul(double const (*) [1024], double const (*) [1024], double (*) [1024]) [3]
-----					
		0.00	0.00	1/1	__libc_csu_init [16]
[10]	0.0	0.00	0.00	1	
					_GLOBAL__sub_I__Z12naive_matmulPA1024_KdS1_PA1024_d [10]
-----					

## Index by function name

```
[10] _GLOBAL__sub_I__Z12naive_matmulPA1024_KdS1_PA1024_d (matmul_array.cpp) [3]
    improved_matmul(double const (*) [1024], double const (*) [1024], double (*) [1024])
[2] naive_matmul(double const (*) [1024], double const (*) [1024], double (*) [1024])
[1] main
```

## Section 8.3.1

### **References**

- 1 S. L. Graham, P. B. Kessler, and M. K. McKusick. [gprof: A call graph execution profiler](#).  
*ACM SIGPLAN Notices*, 17(6):120–126, June 1982.
- 2 S. L. Graham, P. B. Kessler, and M. K. McKusick. [gprof: A call graph execution profiler](#).  
*ACM SIGPLAN Notices*, 39(4):49–57, Apr. 2004.

## Section 8.4

### **Valgrind/Callgrind**

- Callgrind is **Valgrind** tool that collects function call graph information and measures number of instructions executed and cache behavior for program
- does not measure execution time per se; but provides sufficient information to make clock cycle estimates (as is done in KCachegrind)
- can be used to determine cache hit/miss counts and miss rate on program wide, per function, and per source-code line basis
- simulates L1 instruction/data cache and L2 cache
- parameters for each cache can be specified (i.e., size, associativity, and line size) but default to values taken from machine's cache
- simplistic cache model only approximates real cache
- handles code in shared libraries
- typically 15 to 100 times slower (depending on whether cache and branch simulation enabled)



# The valgrind Command with Callgrind Tool

- command line interface has following form:

```
valgrind [options] program [program_options]
```

- to use Callgrind tool, must specify `--tool=callgrind` option
- some tool-independent options include:

Option	Description
<code>--help</code>	print help information and exit
<code>--log-file=</code> <i>file</i>	set file for log information to <i>file</i>

- some Callgrind-specific options include:

Option	Description
<code>--callgrind-out-file=</code> <i>file</i>	sets output file to <i>file</i> ; defaults to <code>callgrind.out-<math>\\$pid</math></code> where $\$pid$ is process ID
<code>--cache-sim=</code> <i>b</i>	specifies if information on cache use should be collected, where <i>b</i> is <code>yes</code> or <code>no</code>
<code>--branch-sim=</code> <i>b</i>	specifies if branching information should be collected, where <i>b</i> is <code>yes</code> or <code>no</code>

# The callgrind\_annotate Command

- command line interface has following form:

```
callgrind_annotate [options] $callgrind_out_file
```

- some options include:

Option	Description
<code>--help</code>	print help information and exit
<code>--auto=<i>b</i></code>	specifies if all source files should be annotated, where <i>b</i> is <code>yes</code> or <code>no</code>

- build code as one would normally (no compile-time instrumentation is needed); for example:

```
g++ -g -O -o array_sum array_sum.cpp
```

- run program using `valgrind` with `Callgrind` tool; for example:

```
valgrind --tool=callgrind --cache-sim=yes \  
  --log-file=callgrind.log \  
  --callgrind-out-file=callgrind.out \  
  ./array_sum
```

- display results with `callgrind_annotate`; for example:

```
callgrind_annotate --auto=yes callgrind.out
```

- alternatively, display results in graphical form with tool like `KCachegrind` (discussed later)

# Example: Source Code

```
1  #include <iostream>
2  #include <algorithm>
3
4  constexpr int M = 2048;
5  constexpr int N = 2048;
6
7  double naive_sum(const double a[][N]) {
8      double sum = 0.0;
9      for (int j = 0; j < N; ++j) {
10         for (int i = 0; i < M; ++i)
11             {sum += a[i][j];}
12     }
13     return sum;
14 }
15
16 double improved_sum(const double a[][N]) {
17     double sum = 0.0;
18     for (int i = 0; i < M; ++i) {
19         for (int j = 0; j < N; ++j)
20             {sum += a[i][j];}
21     }
22     return sum;
23 }
```

## Example: Source Code (Continued)

```
25 int main() {  
26     static double a[M][N];  
27     std::fill_n(&a[0][0], M * N, 1.0 / (M * N));  
28     std::cout << naive_sum(a) << '\n';  
29     static double b[M][N];  
30     std::fill_n(&b[0][0], M * N, 1.0 / (M * N));  
31     std::cout << improved_sum(b) << '\n';  
32 }
```

# Example: Callgrind

```
$ valgrind --tool=callgrind --cache-sim=yes --branch-sim=yes --log-file=callgrind.log --
  callgrind-out-file=callgrind.out ./array_sum
$ cat callgrind.log
==23469== Callgrind, a call-graph generating cache profiler
==23469== Copyright (C) 2002-2013, and GNU GPL'd, by Josef Weidendorfer et al.
==23469== Using Valgrind-3.10.1 and LibVEX; rerun with -h for copyright info
==23469== Command: ./array_sum
==23469== Parent PID: 23449
==23469==
--23469-- warning: L3 cache found, using its data for the LL simulation.
==23469== For interactive control, run 'callgrind_control -h'.
==23469==
==23469== Events      : Ir Dr Dw Ilmr Dlmr Dlmw LLMr DLMr DLMw Bc Bcm Bi Bim
==23469== Collected : 70339139 9142838 8663373 1601 4738282 1051026 1585 4728422 1050172
  17247597 30398 4923 423
==23469==
==23469== I    refs:      70,339,139
==23469== I1  misses:      1,601
==23469== L1i misses:      1,585
==23469== I1  miss rate:      0.0%
==23469== L1i miss rate:      0.0%
==23469==
==23469== D    refs:      17,806,211 ( 9,142,838 rd + 8,663,373 wr)
==23469== D1  misses:      5,789,308 ( 4,738,282 rd + 1,051,026 wr)
==23469== L1d misses:      5,778,594 ( 4,728,422 rd + 1,050,172 wr)
==23469== D1  miss rate:      32.5% ( 51.8% + 12.1% )
==23469== L1d miss rate:      32.4% ( 51.7% + 12.1% )
==23469==
==23469== LL refs:      5,790,909 ( 4,739,883 rd + 1,051,026 wr)
==23469== LL misses:      5,780,179 ( 4,730,007 rd + 1,050,172 wr)
==23469== LL miss rate:      6.5% ( 5.9% + 12.1% )
==23469==
==23469== Branches:      17,252,520 (17,247,597 cond + 4,923 ind)
==23469== Mispredicts:      30,821 ( 30,398 cond + 423 ind)
==23469== Mispred rate:      0.1% ( 0.1% + 8.5% )
```

.....

Ir	Dr	Dw	Ilmr	Dlmr	Dlmw	Ilmr	Dlmr	Dlmw	Bc	Bcm	Bi	Bim	
.	.	.	.	.	.	.	.	.	.	.	.	.	#include <iostream>
.	.	.	.	.	.	.	.	.	.	.	.	.	#include <algorithm>
.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	constexpr int M = 2048;
.	.	.	.	.	.	.	.	.	.	.	.	.	constexpr int N = 2048;
1	0	0	1	0	0	1	.	.	.	.	.	.	
4,097	.	.	.	.	.	.	.	.	.	.	.	.	double naive_sum(const double a[][N]) {
4,096	0	0	0	0	0	0	0	0	2,048	3	.	.	double sum = 0.0;
8,390,656	0	0	0	0	0	0	0	0	4,194,304	2,063	.	.	for (int j = 0; j < N; ++j) {
8,388,608	4,194,304	0	0	4,194,304	0	0	4,194,304	.	.	.	.	.	for (int i = 0; i < M; ++i)
.	.	.	.	.	.	.	.	.	.	.	.	.	{sum += a[i][j];}
.	.	.	.	.	.	.	.	.	.	.	.	.	}
1	1	0	0	1	0	0	1	.	.	.	.	.	return sum;
.	.	.	.	.	.	.	.	.	.	.	.	.	}
1	.	.	.	.	.	.	.	.	.	.	.	.	
2,049	.	.	.	.	.	.	.	.	.	.	.	.	double improved_sum(const double a[][N]) {
4,096	0	0	0	0	0	0	0	0	2,048	3	.	.	double sum = 0.0;
8,388,608	0	0	0	0	0	0	0	0	4,194,304	2,058	.	.	for (int i = 0; i < M; ++i) {
8,388,608	4,194,304	0	0	524,288	0	0	524,288	.	.	.	.	.	for (int j = 0; j < N; ++j)
.	.	.	.	.	.	.	.	.	.	.	.	.	{sum += a[i][j];}
.	.	.	.	.	.	.	.	.	.	.	.	.	}
1	1	0	0	1	0	0	1	.	.	.	.	.	return sum;
2	0	0	1	0	0	1	.	.	.	.	.	.	}
.	.	.	.	.	.	.	.	.	.	.	.	.	
2	0	1	0	0	1	0	0	1	.	.	.	.	int main() {
.	.	.	.	.	.	.	.	.	.	.	.	.	static double a[M][N];
16,787,459	4,194,305	0	1	4,194,305	0	1	4,194,305	0	4,196,352	2,066	.	.	std::fill_n(a[0][0], M * N, 1.0 / (M * N));
.	.	.	.	.	.	.	.	.	.	.	.	.	std::cout << naive_sum(a) << '\n';
2	0	1	0	0	1	0	0	1	.	.	.	.	std::cout << naive_sum(a) << '\n';
.	.	.	.	.	.	.	.	.	.	.	.	.	=> array_sum.cpp:naive_sum(double const (*) [2048]) (1x)
2	0	1	0	0	1	0	0	1	.	.	.	.	static double b[M][N];
16,783,363	4,194,305	0	0	524,289	0	0	524,289	0	4,196,352	2,061	.	.	std::fill_n(b[0][0], M * N, 1.0 / (M * N));
6	2	0	1	2	0	1	1	.	.	.	.	.	std::cout << improved_sum(b) << '\n';
.	.	.	.	.	.	.	.	.	.	.	.	.	=> array_sum.cpp:improved_sum(double const (*) [2048]) (1x)
.	.	.	.	.	.	.	.	.	.	.	.	.	}
-----													
Ir	Dr	Dw	Ilmr	Dlmr	Dlmw	Ilmr	Dlmr	Dlmw	Bc	Bcm	Bi	Bim	
48	92	0	0	100	0	0	100	0	49	14	0	0	percentage of events annotated

- open-source call-graph profiling-data visualization tool
- part of K Desktop Environment (KDE)
- supports Callgrind data
- allows graphical visualization of:
  - call-graph relationships between functions (e.g., callers and callees)
  - function costs/counts
  - annotated source/assembly with costs/counts
- allows much easier interpretation of Callgrind data (relative to `callgrind_annotate`)
- to allow annotation of assembly, add `--dump-instr=yes` option to `valgrind` command for Callgrind
- use command of form:  
`kcachegrind $callgrind_out_file`
- web site: <https://kcachegrind.github.io>



# KCachegrind Example: Source/Assembly Annotation

callgrind.out [./array\_sum]

File View Go Settings Help

Open Back Forward Up % Relative Cycle Detection Relative to Parent Shorten Templates Cycle Estimation

Flat Profile

Search: (No Grouping)

Incl.	Self	Called	Function
100.00	0.00	(0)	0x0000000000000000
99.35	0.00	1	0x0000000000000040
99.35	0.00	1	(below main)
99.31	21.07	1	main
67.68	67.68	1	naive_sum(double const (*) [2048])
10.54	10.54	1	improved_sum(double const (*) [2048])
0.62	0.00	1	_dl_start
0.62	0.00	1	_dl_sysdep_start
0.62	0.00	1	_dl_main
0.59	0.11	7	_dl_relocate_object
0.50	0.13	1 808	_dl_lookup_symbol
0.37	0.23	1 808	_do_lookup_x
0.14	0.07	10 732	check_match.isra
0.07	0.03	6 221	strcmp
0.04	0.04	32 594	strcmp2
0.04	0.00	92	_dl_runtime_resolve
0.04	0.00	92	_dl_fixup
0.03	0.00	1	_libc_csu_init
0.03	0.00	1	_GLOBAL__sub_I
0.03	0.00	1	std::ios_base::Init
0.03	0.00	1	_dl_init
0.03	0.00	7	call_init.part.0
0.03	0.00	1	_GLOBAL__sub_I
0.03	0.00	1	malloc
0.03	0.00	1	malloc_hook_init
0.03	0.00	1	ptmalloc_init.part.0
0.02	0.02	1	_dl_addr
0.02	0.00	22	std::locale::locale
0.02	0.00	22	std::locale::_S_in
0.02	0.00	1	std::locale::_S_in
0.02	0.00	1	std::locale::_Impl
0.01	0.00	2	std::ostream&::operator<<
0.01	0.00	12	_dl_catch_error
0.01	0.00	12	_dl_map_object

main

Types Callers All Callers Callee Map Source Code

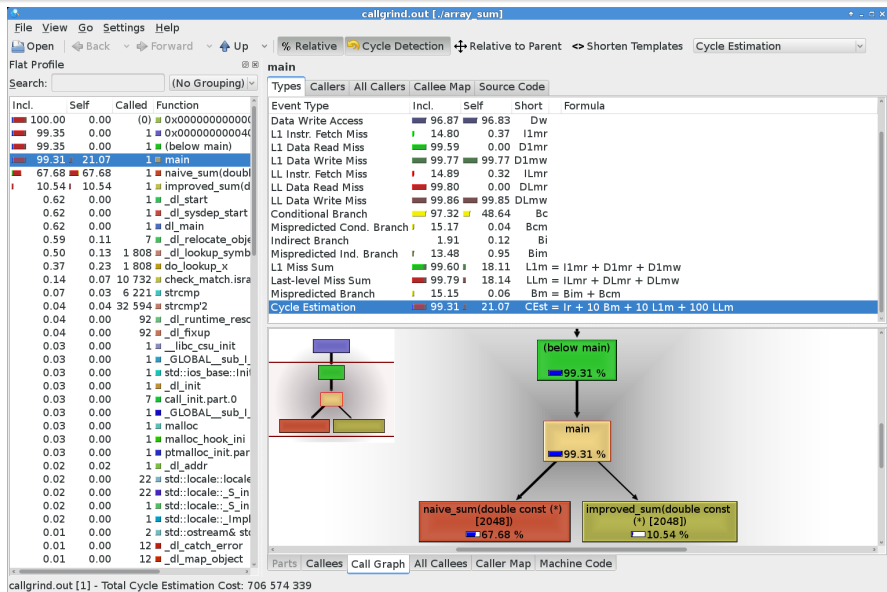
```
# CEst Source ('/home/mdadams/work/cpp/slides/software/callgrind/array_sum.cpp')
22      return sum;
23  }
24
25  0.00 int main() {
26      static double a[M][N];
27      std::fill_n(a[0][0], M * N, 1.0 / (M * N));
28      std::cout << naive_sum(a) << '\n';
29      static double b[M][N];
30      std::fill_n(b[0][0], M * N, 1.0 / (M * N));
31      std::cout << improved_sum(b) << '\n';
32  }
33
34  --- Inlined from '/usr/local/sde-2.15.0/packages/gcc-7.1.0/include/c++/7.1.0/bits/stl_algorithms.h' ---
35
0
```

#	CEst	Hex	Assembly Instructions	Source Position
40 0853		75 f3	jne 400848 <improved_sum(...)	
40 0855		48 39 d7	cmp %rdx,%rdi	
40 0858		75 e7	jne 400841 <improved_sum(...)	
40 085A		f3 c3	repz retq	
40 085C	0.00	48 83 ec 18	sub \$0x18,%rsp	array_sum.cpp:25
40 0860	0.00	b8 80 11 60 02	mov \$0x2601180,%eax	array_sum.cpp:25
40 0865	0.00	f2 0f 10 05 4b 01 00	movsd 0x14b(%rip),%xmm0	... stl_algorithms.h:754
40 086C		00		
40 086D	0.00	48 8d 90 00 00 00 02	lea 0x2000000(%rax),%rdx	stl_algorithms.h:753
40 0874	8.76	f2 0f 11 00	movsd %xmm0,(%rax)	stl_algorithms.h:754
40 0878	0.59	48 83 c0 08	add \$0x8,%rax	stl_algorithms.h:752
40 087C	0.59	48 39 d0	cmp %rdx,%rax	stl_algorithms.h:753
40 087F	0.59	75 f3	jne 400874 <main+0x18>	stl_algorithms.h:753
40 0881	0.00	bf 80 11 60 02	mov \$0x2601180,%edi	array_sum.cpp:28
40 0886	0.00	e8 7c ff ff ff	callq 400807 <naive_sum(double const (*) [2048])>	array_sum.cpp:28
40 0888	67.68	bf 60 10 60 00	mov \$0x601060,%edi	ostream:221

Parts Callee Call Graph All Callee Caller Map Machine Code

callgrind.out [1] - Total Cycle Estimation Cost: 706 574 339

# KCachegrind Example



## Section 8.4.1

### **References**

- 1 P. Floyd. [Valgrind part 4 — cachegrind and callgrind](#). *Overload*, 111:4–7, Oct. 2012.

## Part 9

### **Build Tools**

## Section 9.1

### **Build Tools**

- Build tools are programs that automate the creation of executable programs, libraries, and other artifacts from source code.
- Build tools also typically provide some basic facilities for testing and packaging the artifacts generated by the build process.
- Building software requires careful tracking of:
  - what items need to be built, and
  - the dependencies between these items.
- Dependency tracking is necessary to:
  - determine the order in which items must be built, and
  - minimize the number of items that need to be re-built when a change is made to the code.
- In the case of very small projects, it may be feasible to perform the build process manually.
- For larger projects, however, the build process is far too complex to manage by hand, and build tools are therefore needed.

- Some examples of build tools include:
  - CMake (a cross-platform tool)
  - GNU Build System (also known as Autotools) (for Unix)
  - Make (for Unix)
  - MSBuild (for Microsoft Visual Studio under Microsoft Windows)
  - Xcodebuild (for Apple Xcode)



## Section 9.2

**Make**

- `make` command
- controls generation of executables and/or other non-source files from program's source files
- extremely popular tool for automating build process
- available on many platforms (e.g., Unix, Microsoft Windows, Mac OS X); used extensively on Unix systems
- very flexible
- can handle building multiple programs consisting of hundreds of source files or single program consisting of only one source file
- can be used to build almost anything (i.e., need not be a program)
- for example, all materials for this course typeset using  $\text{\LaTeX}$  (e.g., coursepack, slides, handouts, exams), and `make` utility used to compile  $\text{\LaTeX}$  source code into PDF documents
- one of most popular implementations of `make` is GNU Make
- GNU Make web site: <http://www.gnu.org/software/make>

# The make Command

- target is something that can be built, typically (but not necessarily) file such as executable file or object file
- `make` command driven by data file called `makefile`
- `makefile` usually named `Makefile` or `makefile`
- command-line usage:

`make [options] [targets]`

- *targets*: zero or more targets to be built
- *options*: zero or more options
- by default, looks for `makefile` called `makefile` and then `Makefile`
- if no targets are specified, will build first target specified in `makefile`
- only builds files that are out of date
- most common command-line options include:
  - n                show commands that would be executed but do not actually execute them
  - f *makefile*    use `makefile` *makefile*

- comment starts at hash character (i.e., “#”) and continues until end of line; example:

*# This comment continues until the end of the line.*

- supports variables
- some important variables used by built-in rules:

Name	Description
<b>CXX</b>	C++ compiler command
<b>CXXFLAGS</b>	C++ compiler options
<b>LDFLAGS</b>	linker options

- to assign value to variable, use equal sign; example:

**CXX** = g++

- to substitute value of variable, use dollar sign followed by variable name in parentheses; example:

\$ (**CXX**)

# Makefiles (Continued 1)

- makefile specifies targets and rules for building targets
- each rule in makefile has following form:

```
targets : prerequisites  
_____commands  
_____...
```

- indentation shown above must be with tab character and not spaces
- *targets*: list of one or more targets
- *prerequisites*: files on which targets depend (i.e., files used to produce targets)
- *commands*: actions that must be carried out to produce target from its prerequisites

## Makefiles (Continued 2)

- normally, each target associated with file of same name (and building target will create this file)
- phony target: target that is not associated with any file
- to identify target as phony make it prerequisite of special target called **“.PHONY”**; example (specify `all` as phony target):

```
.PHONY: all
```

- some special built-in variables that can be used in rules:

Name	Description
<code>\$@</code>	target
<code>\$&lt;</code>	name of first prerequisite
<code>\$^</code>	names of all of prerequisites separated by spaces

# Makefile for hello Program

```
1  CXX = g++          # The C++ compiler command.
2  CXXFLAGS = -g -O # The C++ compiler options.
3  LDFLAGS =          # The linker options (if any).
4
5  # The all target builds all of the programs handled by
6  # the makefile.
7  # This target has the dependency chain:
8  #   all -> hello -> hello.o -> hello.cpp
9  all: hello
10
11 # The clean target removes all of the executable files
12 # and object files produced by the build process.
13 clean:
14  ___rm -f hello *.o
15
16 # The hello target builds the hello executable.
17 hello: hello.o
18  ___$(CXX) $(CXXFLAGS) -o $@ $^ $(LDFLAGS)
19
20 # Indicate that the all and clean targets do not
21 # correspond to actual files.
22 .PHONY: all clean
23
24 # The following rule is effectively built into make and
25 # therefore need not be explicitly specified:
26 # hello.o: hello.cpp
27 # ___$(CXX) $(CXXFLAGS) -c $<
```

# Commentary on Makefile for `hello` Program

- `all` target: builds all of the programs handled by the makefile (e.g., `hello`)
- `clean` target: removes all of the executable files and object files produced by build process (e.g., `hello`, `hello.o`)
- although `all` and `clean` have no special meaning to make, very common practice to provide targets with these particular names in all makefiles
- `hello` target: compiles and links the `hello` program
- chain of dependencies for `all` target:  
`all`  $\rightarrow$  `hello`  $\rightarrow$  `hello.o`  $\rightarrow$  `hello.cpp`
- `all` and `clean` examples of phony targets



## Section 9.2.1

### **References**

- 1 S. I. Feldman. [Make](#) — a program for maintaining computer programs. *Software: Practice and Experience*, 9(4):255–265, Apr. 1979.

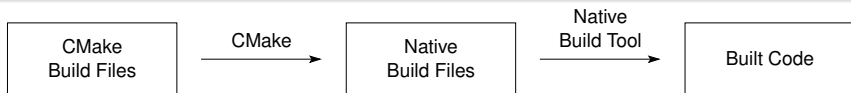
## Section 9.3

### CMake

- CMake is open-source cross-platform family of tools designed to build, test, and package software
- controls software build process (e.g., compiling and linking) using simple platform- and compiler-independent configuration files
- used in conjunction with native build environments
- generates files appropriate for whatever build environment being used
- supports native build environments such as Unix Make, Apple Xcode, and Microsoft Visual Studio
- automatically generates dependency information for source files
- supports parallel builds
- created by Kitware (<http://www.kitware.com>)
- web site: <https://cmake.org>

- CMake has very large user base and is employed in many open-source and commercial projects
- some users of CMake include:
  - Blender (<https://www.blender.org>)
  - Clang (<http://clang.llvm.org>)
  - Computational Geometry Algorithms Library (CGAL) (<http://www.cgal.org>)
  - Jasper Image Processing/Coding Tool Kit (<http://www.ece.uvic.ca/~mdadams/jasper>)
  - K Desktop Environment (KDE) (<https://www.kde.org>)
  - MySQL (<https://www.mysql.com>)
  - Netflix (<https://www.netflix.com>)
  - OpenCV (<http://opencv.org>)
  - Qt (<https://www.qt.io>)
  - Second Life (<http://secondlife.com>)

# Build Process



- **CMake build files:** files used by CMake to describe build process for software project (i.e., `CMakeLists.txt` and other build files it references)
- **native build tool:** program used to build code for particular software development environment being employed (e.g., `make` for Unix, `MSBuild` for Microsoft Visual Studio, and `xcodebuild` for Apple Xcode)
- **native build files:** files used by native build tool to control build process (e.g., `makefiles` for Unix, `project/solution` files for Microsoft Visual Studio, and `project` files for Apple Xcode)
- build process consists of two steps:
  - 1 CMake used, with CMake build files as input, to produce native build files
  - 2 native build tool invoked to build code using native build files generated by CMake
- strictly speaking, CMake does not itself build code, but rather produces build files that can be used by native build tool to build code

- **source directory**: top-level directory of source tree for code to be built
- **binary directory**: directory under which all files generated by build process will be placed
- source directory must contain `CMakeLists.txt` file which is used to describe build process
- **cache**: file where CMake stores values of variables used for configuration of build process (i.e., `CMakeCache.txt` in binary directory)
- **build-system generator**: entity within CMake that produces native build files (i.e., build files targeting particular native build tool)
- CMake provides numerous generators (e.g., generators for Unix Make, Apple Xcode, and Microsoft Visual Studio)
- **build configuration**: description of build to be performed with particular set of parameters (e.g., optimized or debug version)
- some generators support multiple configurations using single build
- for generators that support only single configuration, need to specify which configuration to build

# In-Source Versus Out-of-Source Builds

- **in-source build**: when binary directory chosen to be inside source tree (e.g., same as source directory)
- **out-of-source build**: when binary directory chosen to be outside source tree
- when out-of-source build used, contents of source directory not modified in any way by build process
- in contrast, when in-source build used, build process can generate many new files under source directory
- out-of-source builds have numerous advantages over in-source builds; in particular, out-of-source builds:
  - avoid cluttering source tree with many files generated by build process, which can cause numerous difficulties (e.g., interacting poorly with version control systems)
  - facilitate easy removal of all files generated by build process without risk of accidentally removing source files
  - allow for multiple builds from single source tree (e.g., debug and release builds)
- for above reasons, in-source builds should generally be avoided



# The `cmake` Command for Configuring

- To generate build files for a native build tool (i.e., configure), use a command of the form:

```
cmake [options] [$srcdir]
```

- The source directory `$srcdir` may be optionally specified.
- The source and binary directories default to the current directory (resulting in an in-source build), but may *both* be set by using the `-S` and `-B` options.
- Some options include:

Option	Description
<code>-S <i>srcdir</i></code>	Set the source directory to <i>srcdir</i> .
<code>-B <i>bindir</i></code>	Set the binary directory to <i>bindir</i> .
<code>-D <i>var=val</i></code>	Set the CMake variable <i>var</i> to <i>val</i> .
<code>-G <i>gen</i></code>	Set the build-system generator to <i>gen</i> .
<code>--version</code>	Print name/version banner and exit.
<code>--help</code>	Print usage information and exit.
<code>--debug-output</code>	Enable debugging output.
<code>--trace</code>	Enable tracing output.

# The `cmake` Command for Configuring (Continued 1)

- Some supported generators include:

Name	Description
Unix_Makefiles	makefiles for Unix Make
Xcode	project files for Apple Xcode
Visual_Studio_15_2017	project files for Microsoft Visual Studio 15
Ninja	build files for Ninja build system

- Some environment variables used by `cmake` include:

Option	Description
<code>CMAKE_CONFIG_TYPE</code>	The default build configuration.
<code>CMAKE_BUILD_PARALLEL_LEVEL</code>	The maximum number of concurrent processes to use when building.
<code>CC</code>	The command for compiling C source.
<code>CFLAGS</code>	The default compilation flags for compiling C source.
<code>CXX</code>	The command for compiling C++ source.
<code>CXXFLAGS</code>	The default compilation flags for compiling C++ source.
<code>LDFLAGS</code>	The default linker flags.

- To invoke the native build tool in a platform-independent manner for the build files in the binary directory `$bindir`, use a command of the form:

```
cmake --build $bindir [$options]
```

- Some options include:

Option	Description
<code>--target <i>target</i></code>	Build the target <i>target</i> instead of the default targets.
<code>--config <i>config</i></code>	For a multi-configuration generator, select the build configuration <i>config</i> . For a single-configuration generator, this option is ignored.
<code>--clean-first</code>	Build the “clean” target first.
<code>-j <i>n</i></code>	Set the maximum number of concurrent processes to <i>n</i> .
<code>--</code>	Pass the remaining options to the native build tool.

# Hello World Example

- source directory `$SOURCE_DIR` contains two files:
  - 1 `CMakeLists.txt`
  - 2 `hello.cpp`
- commands to build with binary directory `$BINARY_DIR`:

```
cmake -S$SOURCE_DIR -B$BINARY_DIR
cmake --build $BINARY_DIR
```

`$SOURCE_DIR/hello.cpp`

```
1 #include <iostream>
2 int main() {std::cout << "Hello, World!\n";}
```

`$SOURCE_DIR/CMakeLists.txt`

```
1 # Specify minimum required version of CMake.
2 cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
3
4 # Specify project and identify languages used.
5 project(hello LANGUAGES CXX)
6
7 # Add program target called hello.
8 add_executable(hello hello.cpp)
```

## Section 9.3.1

### **CMakeLists Files**

# Projects, Targets, and Build Configurations

- **project**: collection of source code to be built using CMake
- **target**: something to be built by build process, such as executable or library
- target typically associated with one or more source files
- target has numerous properties (e.g., compiler flags and linker flags)
- target names cannot contain whitespace
- by default, following build configurations are supported:

Name	Description
Debug	basic debugging code/information enabled
Release	basic optimization enabled
RelWithDebInfo	optimized build with debugging code/information enabled as well
MinSizeRel	smallest (but not necessarily fastest) code

# Comments and Commands

- comment starts with hash character (i.e., “#”) and continues until end of line
- file consists of sequence of commands
- command consists of following (in order):

- 1 command name
- 2 opening parenthesis
- 3 whitespace-separated arguments
- 4 closing parenthesis

- command example:

```
cmake_minimum_required(VERSION 3.1)
```

- command names are case insensitive
- anything in double quotes treated as single argument; for example, as in:

```
message("Hello World")
```

- backslash character can be used to escape character such as double quote; for example, as in:

```
message("\${X} is not a variable expansion")
```

- variable name is sequence of one or more letters, digits, and underscore characters that does not begin with digit (e.g., `MATH_LIBRARY`, `i`)
- variable names are case sensitive
- value of variable can be treated as string or list of strings
- value of variable `X` is accessed as `${X}`
- boolean tests are case insensitive
- all of following considered false: `OFF`, `0`, `NO`, `FALSE`, `NOTFOUND`, `*-NOTFOUND`, `IGNORE`
- variable can be internal or cache
- cache variable persists across separate invocations of `cmake` while internal variable does not
- internal variable take precedence over cache variable



- module is file containing re-usable piece of CMakeLists code
- normally use “.cmake” file name extension
- most modules can be classified into one of following categories:
  - find
  - system introspection
  - utility
- find module:
  - determines location of software elements such as header files and libraries
  - often module name starts with prefix “Find”
  - examples: FindBoost and FindOpenGL
- system introspection module:
  - provides information about target system or compiler (e.g., size of various types, availability of header files, compiler version)
  - often module name starts with prefix “Test” or “Check”
  - examples: CheckCXXSourceCompiles and CheckIncludeFile
- utility module:
  - provides additional functions/macros for convenience
  - example: ExternalProject

- module can be accessed via `include` command
- find module normally accessed via `find_package` command (instead of directly using `include` command)

## ■ Source and binary directories:

- `CMAKE_BINARY_DIR`. The full path to the top-level directory of the current CMake build tree. For an in-source build, this is the same as `CMAKE_SOURCE_DIR`.
- `CMAKE_SOURCE_DIR`. The full path to the top-level directory of the current CMake source tree. For an in-source build, this is the same as `CMAKE_BINARY_DIR`.
- `CMAKE_CURRENT_SOURCE_DIR`. The full path to the source directory that is currently being processed by cmake.
- `CMAKE_CURRENT_BINARY_DIR`. The full path to the binary directory that is currently being processed by cmake.

## ■ Build type:

- `CMAKE_BUILD_TYPE`. In the case of single-configuration generators, specifies the build type (e.g., Release, Debug, RelWithDebInfo, MinSizeRel). In the case of multi-configuration generators, unused.
- `BUILD_SHARED_LIBS`. Specifies if all libraries created should default to shared (instead of static).
- `BUILD_TESTING`. Specifies if testing is enabled (when the `CTest` module is used).

# Commonly-Used Variables (Continued 1)

## ■ C++ compiler:

- ❑ `CMAKE_CXX_COMPILER_ID`. The C++ compiler in use (e.g., Clang, GNU, Intel, MSVC).
- ❑ `CMAKE_CXX_STANDARD`. Used to initialize the `CXX_STANDARD` property on all targets, which selects version of C++ standard (e.g., 98, 11, and 14).
- ❑ `CMAKE_CXX_STANDARD_REQUIRED`. Used to initialize the `CXX_STANDARD_REQUIRED` property of all targets. This property determines whether the specified version of C++ standard is required.
- ❑ `CMAKE_CXX_COMPILER`. The compiler command used for C++ source code.
- ❑ `CMAKE_CXX_FLAGS`. The compiler flags for compiling C++ source code.
- ❑ `CMAKE_CXX_FLAGS_DEBUG`. The compiler flags for compiling C++ source code for a debug build.
- ❑ `CMAKE_CXX_FLAGS_RELEASE`. The compiler flags for compiling C++ source code for a release build.
- ❑ `CMAKE_CXX_FLAGS_RELWITHDEBINFO`. The compiler flags for compiling C++ source code for a release build with debug flags.
- ❑ `CMAKE_CXX_FLAGS_MINSIZEREL`. The compiler flags for compiling C++ source code for a release build with minimum code size.

# Commonly-Used Variables (Continued 2)

## ■ Linker:

- `CMAKE_EXE_LINKER_FLAGS`. The linker flags used to create executables. This variable also has configuration-specific variants, such as `CMAKE_EXE_LINKER_FLAGS_RELEASE`.
- `CMAKE_SHARED_LINKER_FLAGS`. The linker flags used to create shared libraries. This variable also has configuration-specific variants, such as `CMAKE_SHARED_LINKER_FLAGS_RELEASE`.
- `CMAKE_STATIC_LINKER_FLAGS`. The linker flags used to create static libraries. This variable also has configuration-specific variants, such as `CMAKE_STATIC_LINKER_FLAGS_RELEASE`.

## ■ Target OS:

- `CMAKE_SYSTEM_NAME`. The name of the target system's OS (e.g., Linux, Windows, Darwin).
- `UNIX`. Specifies if the target system's OS is UNIX (or UNIX-like).
- `APPLE`. Specifies if the target system's OS is Mac OS X.
- `WIN32`. Specifies if the target system's OS is Microsoft Windows (32- or 64-bit).

# Commonly-Used Variables (Continued 3)

## ■ Makefile builds:

- `CMAKE_VERBOSE_MAKEFILE`. Enable/disable verbose output from Makefile builds.
- `CMAKE_RULE_MESSAGES`. Specify if a progress message should be reported by each makefile rule.

## ■ Other:

- `CMAKE_MODULE_PATH`. The list of directories to search for CMake modules. (This is used by commands like `include` and `find_package`.)
- `CMAKE_PREFIX_PATH`. The list of directories specifying installation prefixes to be searched by the `find_package`, `find_program`, `find_library`, and `find_file` commands.
- `CMAKE_PROJECT_NAME`. The name of the current project.
- `CMAKE_CURRENT_LIST_DIR`. The directory of the listfile currently being processed. (The values of `CMAKE_CURRENT_SOURCE_DIR` and `CMAKE_CURRENT_LIST_DIR` can differ, for example, when a listfile outside the current source directory is included.)
- `CMAKE_EXPORT_COMPILE_COMMANDS`. Enable/disable output of compile commands during generation in file `compile_commands.json`.

## ■ Initialization:

- `cmake_minimum_required`. Set the minimum required version of Cmake for a project.
- `cmake_policy`. Manage CMake policy settings. (This is used to select between old and new behaviors in CMake.)
- `project`. Set a name, version, and enable languages for the entire project. (If no languages specified, defaults to C and C++.)
- `option`. Provide an option that the user can (optionally) select.

## ■ Adding targets:

- `add_executable`. Add a program target.
- `add_library`. Add a library target.
- `add_test`. Add a test target. (This is used in conjunction with the module CTest.)
- `add_custom_target`. Add a target with no output file that is always out of date.
- `add_custom_command`. Add a custom build rule to the generated build system.

## ■ Setting properties for a specific target:

- `target_compile_definitions`. Add compile definitions to a target.
- `target_compile_options`. Add compile options to a target
- `target_include_directories`. Add include directories to a target.
- `target_link_libraries`. Add libraries to the list of libraries to be used for linking a target. (May be used multiple times for the same target.)
- `set_target_properties`. Set properties for a target. (Some properties include: `OUTPUT_NAME`, `SOVERSION`, and `VERSION`.)



### ■ Setting properties for all targets (which is usually best avoided):

- `add_compile_options`. Adds options to the compilation of source files in the current directory and below. (This command should precede an `add_executable` or `add_library` command.)
- `add_definitions`. Adds `-D` define flags to the compilation of source files in the current directory and below.
- `include_directories`. Add directories to the list of include directories used for compiling programs.
- `link_libraries`. Add libraries to the list of libraries used for linking programs. (This command appends to the list, each time it is invoked.)
- `link_directories`. Specify directories in which the linker is to look for libraries.

### ■ Processing other files or directories:

- `add_subdirectory`. Add a subdirectory to the build.
- `include`. Load and run CMake code from a file or module.

# Commonly-Used Commands (Continued 3)

## ■ Querying external packages and programs:

- `find_package`. Load settings for an external software package (e.g., Doxygen, Threads, Boost, OpenGL, GLEW, GLUT, CGAL, PkgConfig).
- `find_library`. Find an external library.
- `find_program`. Find an external program.

## ■ Assignment, control flow, functions, and macros:

- `set`. Set a CMake, cache, or environment variable to a given value.
- `if`, `elseif`, `else`, and `endif`. Conditionally execute a group of commands.
- `foreach` and `endforeach`. Evaluate a group of commands for each value in a list.
- `while` and `endwhile`. Evaluate a group of commands while a condition is true.
- `function` and `endfunction`. Record a function for later invocation as a command.
- `macro` and `endmacro`. Record a macro for later invocation as a command.

## ■ String and list processing:

- `list`. Perform operations on lists.
- `string`. Perform operations on strings.

## ■ Other:

- `message`. Display a message to the user.
- `configure_file`. Copy a file to another location and modify its contents.
- `install`. Specify rules to run at install time (e.g., rules to install programs, libraries, and header files).
- `math`. Evaluate mathematical expressions.
- `file`. Manipulate files.
- `enable_language`. Enable a language.

# Commonly-Used Modules

- **CheckIncludeFiles** module, which provides:
  - `check_include_files`. Check if the specified files can be included.
- **CheckCXXSourceCompiles** module, which provides:
  - `check_cxx_source_compiles`. Check if the specified C++ source code compiles and links to produce an executable.
- **CheckFunctionExists** module, which provides:
  - `check_function_exists`. Check if the specified C function is provided by libraries on the system.
- **CTest** module:
  - Configure a project for testing with CTest/CDash.
- **CPack** module:
  - Configure a project to use CPack to build binary and source package installers.
- **PkgConfig** module, which requires `pkg-config` tool to be available and provides:
  - `pkg_search_module`. Finds a package via `pkg-config`.

# Commonly-Used Modules (Continued 1)

- **ExternalProject module**, which provides:
  - `externalproject_add`. Create custom targets to build projects in external trees.
- **GNUInstallDirs module**:
  - Define GNU standard installation directories (e.g., `CMAKE_INSTALL_INCLUDEDIR`, `CMAKE_INSTALL_LIBDIR`, and `CMAKE_INSTALL_MANDIR`).
- **GenerateExportHeader module**, which provides:
  - `generate_export_header`. Generate a header file containing export macros to be used for a shared library.
- **CMakePackageConfigHelpers module**, which provides:
  - `configure_package_config_file`. Create a package configuration file for installing a project or library. (This should be used instead of `configure_file`.)
  - `write_basic_package_version_file`. Write a package version file.

# Some Find and Pkg-Config Modules

## ■ Boost

- <https://cmake.org/cmake/help/v3.10/module/FindBoost.html>
- **variables:** Boost\_FOUND, Boost\_INCLUDE\_DIRS, Boost\_LIBRARY\_DIRS, Boost\_LIBRARIES
- **imported targets:** Boost::boost, Boost::component

## ■ CGAL (Computational Geometry Algorithms Library)

- **variables:** CGAL\_INCLUDE\_DIRS, CGAL\_LIBRARY, GMP\_LIBRARIES

## ■ Doxygen

- <https://cmake.org/cmake/help/v3.10/module/FindDoxygen.html>
- **variables:** DOXYGEN\_FOUND, DOXYGEN\_EXECUTABLE
- **imported targets:** Doxygen::doxygen, Doxygen::dot

## ■ GLEW (OpenGL Extension Wrangler Library)

- <https://cmake.org/cmake/help/v3.10/module/FindGLEW.html>
- **variables:** GLEW\_FOUND, GLEW\_INCLUDE\_DIRS, GLEW\_LIBRARIES
- **imported targets:** GLEW::GLEW

## ■ GLFW (OpenGL Helper Library) [pkg-config module]

- **variables:** GLFW\_FOUND, GLFW\_INCLUDE\_DIRS, GLFW\_LIBRARIES

## ■ GLUT (OpenGL Utility Toolkit)

- <https://cmake.org/cmake/help/v3.10/module/FindGLUT.html>
- **variables:** GLUT\_FOUND, GLUT\_INCLUDE\_DIR, GLUT\_LIBRARIES
- **imported targets:** GLUT::GLUT

## ■ OpenGL (Open Graphics Library)

- <https://cmake.org/cmake/help/v3.10/module/FindOpenGL.html>
- **variables:** OPENGL\_FOUND, OPENGL\_INCLUDE\_DIR, OPENGL\_LIBRARIES
- **imported targets:** OpenGL::GL, OpenGL::GLU, OpenGL::GLX

## ■ SPL (Signal/Geometry Processing Library)

- **variables:** SPL\_FOUND, SPL\_INCLUDE\_DIRS, SPL\_LIBRARY\_DIRS, SPL\_LIBRARIES, SNDFILE\_INCLUDE\_DIRS, SNDFILE\_LIBRARIES

## ■ Threads

- <https://cmake.org/cmake/help/v3.10/module/FindThreads.html>
- **variables:** CMAKE\_THREAD\_LIBS\_INIT
- **imported targets:** Threads::Threads

# Using Per-Target Versus Global Settings

- can set compiler options, compiler definitions, include directories, and link libraries in two ways:
  - 1 per target (e.g., using `target_compile_options`, `target_compile_definitions`, `target_include_directories`, and `target_link_libraries`)
  - 2 globally (e.g., using `add_compile_options`, `add_definitions`, `include_directories`, and `link_libraries`)
- per-target approach allows properties to be specified with finer granularity than global approach
- finer-granularity control over properties often necessary, especially when building more complex projects
- if executable targets in project do not all use same set of libraries, global specification of include directories and link libraries can introduce artificial dependencies on some libraries
- per-target specification of link libraries allows automatic propagation of library dependencies when hierarchies of libraries used (which, for example, may avoid need to link against same library multiple times)



## Section 9.3.2

### **Examples**

# Hello World Example Revisited

hello.cpp

```
1  #include <iostream>
2
3  int main() {std::cout << "Hello, World!\n";}
```

CMakeLists.txt

```
1  # Specify minimum required version of CMake.
2  cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
3
4  # Specify project and identify languages used.
5  project(hello LANGUAGES CXX)
6
7  # Print message indicating detected OS.
8  if (UNIX)
9      set(platform "Unix")
10 elseif (WIN32)
11     set(platform "Microsoft Windows")
12 else()
13     set(platform "Unknown")
14 endif()
15 message("OS is ${platform}")
16
17 # Add program target called hello.
18 add_executable(hello hello.cpp)
```

# Test Example

- want to build and test hello-world program
- code written in C++
- files in project:
  - CMakeLists.txt
  - hello.cpp
  - test\_wrapper.in
  - run\_test
- project has:
  - executable target `hello`
  - test target `run_test`

# Test Example: Source Code (Including Some Scripts)

hello.cpp

```
1 #include <iostream>
2
3 int main() {std::cout << "Hello, World!\n";}
```

test\_wrapper.in (with execute permission set)

```
1 #!/bin/sh
2 # Initialize the environment for the command being invoked.
3 export CMAKE_SOURCE_DIR="@CMAKE_SOURCE_DIR@"
4 export CMAKE_BINARY_DIR="@CMAKE_BINARY_DIR@"
5 "$@"
```

run\_test

```
1 #!/bin/sh
2 # Test if the hello program produces the desired output.
3 ($CMAKE_BINARY_DIR/hello | grep "^Hello, World!$") || \
4     exit 1
```

# Test Example: CMakeLists File

## CMakeLists.txt

```
1  # Specify minimum required version of CMake.
2  cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
3
4  # Specify project and identify languages used.
5  project(hello LANGUAGES CXX)
6
7  # Include the CTest module for testing.
8  include(CTest)
9
10 # Find the Bourne shell.
11 find_program(sh SH_COMMAND)
12
13 # Add program target called hello.
14 add_executable(hello hello.cpp)
15
16 # Create a wrapper script that initializes the environment
17 # for any test scripts.
18 configure_file(${CMAKE_SOURCE_DIR}/test_wrapper.in
19               ${CMAKE_BINARY_DIR}/test_wrapper @ONLY)
20
21 # Add a test that invokes run_test via a wrapper script.
22 add_test(run_test ${SH_COMMAND}
23          ${CMAKE_BINARY_DIR}/test_wrapper
24          ${CMAKE_SOURCE_DIR}/run_test)
```

# Threads Example

- want to build simple multithreaded program using pthread library
- code written in C++
- files in project:
  - CMakeLists.txt
  - hello.cpp
- project has:
  - executable target `hello`

# Threads Example: Source and CMakeLists Files [With Imported Targets]

hello.cpp

```
1  #include <iostream>
2  #include <thread>
3
4  int main() {
5      std::thread t1([]() {std::cout << "Hello, World!\n";});
6      t1.join();
7  }
```

CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.4 FATAL_ERROR)
2  project(threads_example LANGUAGES CXX)
3
4  # Require compliance with C++11 standard.
5  set(CMAKE_CXX_STANDARD 11)
6  set(CMAKE_CXX_STANDARD_REQUIRED TRUE)
7
8  # Find the threads library, indicating a preference for the
9  # pthread library.
10 set(THREADS_PREFER_PTHREAD_FLAG ON)
11 find_package(Threads REQUIRED)
12
13 # Define a program target called hello.
14 add_executable(hello hello.cpp)
15
16 # Set the libraries for the hello target.
17 target_link_libraries(hello Threads::Threads)
```

# Boost Log Example

- want to build simple program using Boost Log
- code written in C++
- uses Log component of Boost library
- files in project:
  - CMakeLists.txt
  - main.cpp
- project has:
  - executable target `my_app`



# Boost Log Example: Source Code

main.cpp

```
1  #include <boost/log/trivial.hpp>
2
3  int main() {
4      BOOST_LOG_TRIVIAL(warning)
5          << "A warning severity message";
6      BOOST_LOG_TRIVIAL(error)
7          << "An error severity message";
8      BOOST_LOG_TRIVIAL(fatal)
9          << "A fatal severity message";
10 }
```

## CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.4 FATAL_ERROR)
2  project(boost_example LANGUAGES CXX)
3
4  # Find the required libraries (i.e., POSIX threads and Boost).
5  set(Boost_USE_MULTITHREADED ON)
6  find_package(Threads REQUIRED)
7  find_package(Boost 1.54.0 REQUIRED COMPONENTS log)
8
9  # Define a program target called my_app.
10 add_executable(my_app main.cpp)
11
12 # Set the includes, defines, and libraries for the my_app target.
13 target_include_directories(my_app PUBLIC ${Boost_INCLUDE_DIRS})
14 target_compile_definitions(my_app PUBLIC "-DBOOST_LOG_DYN_LINK")
15 target_link_libraries(my_app ${Boost_LIBRARIES}
16     ${CMAKE_THREAD_LIBS_INIT})
```

## CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.5 FATAL_ERROR)
2  project(boost_example LANGUAGES CXX)
3
4  # Find the required libraries (i.e., POSIX threads and Boost).
5  set(Boost_USE_MULTITHREADED ON)
6  find_package(Threads REQUIRED)
7  find_package(Boost 1.54.0 REQUIRED COMPONENTS log)
8
9  # Define a program target called my_app.
10 add_executable(my_app main.cpp)
11
12 # Set the defines, includes, and libraries for the my_app target.
13 target_compile_definitions(my_app PUBLIC "-DBOOST_LOG_DYN_LINK")
14 target_link_libraries(my_app Boost::log Threads::Threads)
```

- want to build simple OpenGL/GLFW application
- code written in C++
- uses OpenGL and GLFW libraries (as well as GLEW library)
- files in project:
  - CMakeLists.txt
  - trivial.cpp
- project has:
  - executable target `trivial`

# OpenGL/GLFW Example: Source Code

trivial.cpp

```
1  #include <cstdlib>
2  #include <GLFW/glfw3.h>
3
4  void display(GLFWwindow* window) {
5      glfwMakeContextCurrent(window);
6      glClearColor(0.0, 1.0, 1.0, 0.0);
7      glClear(GL_COLOR_BUFFER_BIT | GL_DEPTH_BUFFER_BIT);
8      glfwSwapBuffers(window);
9  }
10
11 int main(int argc, char** argv) {
12     if (!glfwInit()) {return EXIT_FAILURE;}
13     glfwSwapInterval(1);
14     GLFWwindow* window = glfwCreateWindow(512, 512, argv[0],
15         nullptr, nullptr);
16     if (!window) {
17         glfwTerminate();
18         return EXIT_FAILURE;
19     }
20     glfwSetWindowRefreshCallback(window, display);
21     while (!glfwWindowShouldClose(window))
22         {glfwWaitEvents();}
23     glfwTerminate();
24     return EXIT_SUCCESS;
25 }
```

## CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.2 FATAL_ERROR)
2  project(opengl_example LANGUAGES CXX)
3  set(CMAKE_CXX_STANDARD 11)
4  set(CMAKE_CXX_STANDARD_REQUIRED TRUE)
5
6  # Find the required libraries (i.e., OpenGL, GLEW, and GLFW).
7  find_package(OpenGL REQUIRED)
8  find_package(GLEW REQUIRED)
9  find_package(PkgConfig REQUIRED)
10 pkg_search_module(GLFW REQUIRED glfw3)
11
12 # Define a program target called trivial.
13 add_executable(trivial trivial.cpp)
14
15 # Set the includes and libraries for the trivial target.
16 target_include_directories(trivial PUBLIC ${GLFW_INCLUDE_DIRS}
17     ${GLEW_INCLUDE_DIRS} ${OPENGL_INCLUDE_DIR})
18 target_link_libraries(trivial ${GLFW_LIBRARIES} ${GLEW_LIBRARIES}
19     ${OPENGL_LIBRARIES})
```

## CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.8 FATAL_ERROR)
2  project(opengl_example LANGUAGES CXX)
3  set(CMAKE_CXX_STANDARD 11)
4  set(CMAKE_CXX_STANDARD_REQUIRED TRUE)
5
6  # Find the required libraries (i.e., OpenGL, GLEW, and GLFW).
7  find_package(OpenGL REQUIRED)
8  find_package(GLEW REQUIRED)
9  find_package(PkgConfig REQUIRED)
10 pkg_search_module(GLFW REQUIRED glfw3)
11
12 # Define a program target called trivial.
13 add_executable(trivial trivial.cpp)
14
15 # Set the includes and libraries for the trivial target.
16 target_include_directories(trivial PUBLIC ${GLFW_INCLUDE_DIRS})
17 target_link_libraries(trivial ${GLFW_LIBRARIES} GLEW::GLEW
18     OpenGL::GL)
```

- want to build simple OpenGL/GLUT application
- code written in C++
- uses OpenGL and GLUT libraries
- files in project:
  - `CMakeLists.txt`
  - `trivial.cpp`
- project has:
  - executable target `trivial`



trivial.cpp

```
1  #include <GL/glut.h>
2
3  void display() {
4      glClearColor(0.0, 1.0, 1.0, 0.0);
5      glClear(GL_COLOR_BUFFER_BIT | GL_DEPTH_BUFFER_BIT);
6      glutSwapBuffers();
7  }
8
9  int main(int argc, char** argv) {
10     glutInit(&argc, argv);
11     glutInitDisplayMode(GLUT_DOUBLE | GLUT_RGB);
12     glutInitWindowSize(512, 512);
13     glutCreateWindow(argv[0]);
14     glutDisplayFunc(display);
15     glutMainLoop();
16 }
```

## CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.9 FATAL_ERROR)
2  project(opengl_example LANGUAGES CXX)
3
4  # Find the required libraries (i.e., OpenGL and GLUT).
5  find_package(OpenGL REQUIRED)
6  find_package(GLUT REQUIRED)
7
8  # Define a program target called trivial.
9  add_executable(trivial trivial.cpp)
10
11 # Set the includes and libraries for the trivial target.
12 target_include_directories(trivial PUBLIC ${GLUT_INCLUDE_DIR}
13     ${OPENGL_INCLUDE_DIR})
14 target_link_libraries(trivial ${GLUT_LIBRARIES} ${OPENGL_LIBRARIES})
```

## CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.8 FATAL_ERROR)
2  project(opengl_example LANGUAGES CXX)
3
4  # Find the required libraries (i.e., OpenGL and GLUT).
5  find_package(OpenGL REQUIRED)
6  find_package(GLUT REQUIRED)
7
8  # Define a program target called trivial.
9  add_executable(trivial trivial.cpp)
10
11 # Set the includes and libraries for the trivial target.
12 target_link_libraries(trivial GLUT::GLUT OpenGL::GL)
```

- want to build simple CGAL application
- code written in C++
- uses CGAL library (as well as GMP library)
- files in project:
  - CMakeLists.txt
  - orient\_test.cpp
- project has:
  - executable target `orient_test`

# CGAL Example: Source Code

orient\_test.cpp

```
1  #include <iostream>
2  #include <string>
3  #include <CGAL/Cartesian.h>
4  #include <CGAL/Filtered_kernel.h>
5
6  std::string toString(CGAL::Orientation orient) {
7      switch (orient) {
8          case CGAL::LEFT_TURN:
9              return "left turn";
10         case CGAL::RIGHT_TURN:
11             return "right turn";
12         case CGAL::COLLINEAR:
13             return "collinear";
14     }
15 }
16
17 int main(int argc, char** argv) {
18     using Point = CGAL::Point_2<CGAL::Filtered_kernel<
19         CGAL::Cartesian<double>>>;
20     Point a, b, q;
21     while (std::cin >> a >> b >> q) {
22         auto orient = CGAL::orientation(a, b, q);
23         std::cout << toString(orient) << '\n';
24     }
25 }
```

# CGAL Example: CMakeLists File

## CMakeLists.txt

```
1  # Specify minimum required version of CMake.
2  cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
3
4  # Specify project and enable the C++ language.
5  project(cgal_demo LANGUAGES CXX)
6
7  # Find the required CGAL package.
8  find_package(CGAL REQUIRED)
9
10 # On some systems, GCC may need the -frounding-math option.
11 if (CMAKE_CXX_COMPILER_ID MATCHES GNU)
12     add_compile_options("-frounding-math")
13 endif()
14
15 # Add a program target called orient_test.
16 add_executable(orient_test orient_test.cpp)
17
18 # Specify the includes and libraries for the orient_test target.
19 target_include_directories(orient_test PUBLIC ${CGAL_INCLUDE_DIRS})
20 target_link_libraries(orient_test ${CGAL_LIBRARY} ${GMP_LIBRARIES})
```

# HG2G Example: Overview

- want to be able to build and install HG2G library and application that uses library
- code written in C++
- files in project:
  - CMakeLists.txt
  - app/CMakeLists.txt
  - app/answer.cpp
  - hg2g/CMakeLists.txt
  - hg2g/answer.cpp
  - hg2g/question.cpp
  - hg2g/include/hg2g/answer.hpp
  - hg2g/include/hg2g/config.hpp.in
- project has:
  - **library target** hg2g
  - **executable target** answer
  - option HG2G\_ZAPHOD (which takes a boolean value)

# HG2G Example: Library Source Code

hg2g/include/hg2g/config.hpp.in

```
1 #ifndef HG2G_CONFIG_H
2 #define HG2G_CONFIG_H
3 #define HG2G_VERSION "@HG2G_VERSION@"
4 #cmakedefine HG2G_ZAPHOD
5 #endif
```

hg2g/include/hg2g/answer.hpp

```
1 #include <string>
2 namespace hg2g {
3     std::string answer_to_ultimate_question();
4     std::string ultimate_question();
5 }
```

hg2g/answer.cpp

```
1 #include "hg2g/answer.hpp"
2 namespace hg2g {
3     std::string answer_to_ultimate_question() {return "42";}
4 }
```

hg2g/question.cpp

```
1 #include "hg2g/answer.hpp"
2 namespace hg2g {
3     std::string ultimate_question() {throw 42;}
4 }
```



# HG2G Example: Application Source Code

app/answer.cpp

```
1  #include <iostream>
2  #include <hg2g/config.hpp>
3  #include <hg2g/answer.hpp>
4
5  int main() {
6  #ifdef HG2G_ZAPHOD
7      std::cout << "HG2G_ZAPHOD is defined\n";
8  #endif
9      std::cout << "According to version " << HG2G_VERSION <<
10         " of the HG2G library:\n";
11      std::cout <<
12         "The answer to the ultimate question is " <<
13         hg2g::answer_to_ultimate_question() << "...\n";
14  }
```

# HG2G Example: CMakeLists Files

## CMakeLists.txt

```
1 cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
2 project(hg2g_example LANGUAGES CXX)
3 option(HG2G_ZAPHOD "Define HG2G_ZAPHOD" FALSE)
4
5 # Set the version number and name.
6 set(HG2G_VERSION_MAJOR 1)
7 set(HG2G_VERSION_MINOR 42)
8 set(HG2G_VERSION_PATCH 0)
9 string(CONCAT HG2G_VERSION "${HG2G_VERSION_MAJOR}"
10      "${HG2G_VERSION_MINOR}" "${HG2G_VERSION_PATCH}")
11
12 # Process the subdirectories hg2g and app.
13 add_subdirectory(hg2g)
14 add_subdirectory(app)
```

## app/CMakeLists.txt

```
1 # Add a program target called answer.
2 add_executable(answer answer.cpp)
3
4 # Link the answer program against the hg2g library.
5 target_link_libraries(answer hg2g)
6
7 # Install the answer program in the bin directory.
8 install(TARGETS answer DESTINATION bin)
```

# HG2G Example: CMakeLists Files (Continued 1)

hg2g/CMakeLists.txt

```
1  # Place the names of the header and source files into
2  # variables (for convenience).
3  set(hg2g_headers include/hg2g/answer.hpp
4      "${CMAKE_CURRENT_BINARY_DIR}/include/hg2g/config.hpp")
5  set(hg2g_sources answer.cpp question.cpp)
6
7  # Add a library target called hg2g.
8  add_library(hg2g ${hg2g_sources} ${hg2g_headers})
9
10 # Specify the include directories for library.
11 target_include_directories(hg2g PUBLIC
12     include
13     "${CMAKE_CURRENT_BINARY_DIR}/include")
14
15 # Create a header file containing the config information.
16 configure_file(
17     include/hg2g/config.hpp.in
18     "${CMAKE_CURRENT_BINARY_DIR}/include/hg2g/config.hpp")
19
20 # Install the library in the lib directory.
21 install(TARGETS hg2g DESTINATION lib)
22
23 # Install the header files in the include/hg2g directory.
24 install(FILES ${hg2g_headers} DESTINATION include/hg2g)
```

# External Project Example

hello, hg2g, and example\_100 are subdirectories containing CMake projects

CMakeLists.txt

```
1  cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
2  # Specify the project and do not enable any languages.
3  project(examples LANGUAGES CXX)
4  # Include the module for external project functionality.
5  include(ExternalProject)
6  # Create a list of the subdirectories containing
7  # CMake projects to be built.
8  list(APPEND dirs hello hg2g "example 100")
9  # Add each project as an external project.
10 foreach(dir IN LISTS dirs)
11     # Set target name to directory name with any
12     # spaces changed to underscores.
13     string(REPLACE " " "_" target "${dir}")
14     # Add external project.
15     externalproject_add("${target}"
16         SOURCE_DIR "${CMAKE_SOURCE_DIR}/${dir}"
17         BINARY_DIR "${CMAKE_BINARY_DIR}/${dir}"
18         CMAKE_ARGS
19         "-DCMAKE_CXX_COMPILER=${CMAKE_CXX_COMPILER}"
20         INSTALL_COMMAND "")
21 endforeach()
```

- want to build L<sup>A</sup>T<sub>E</sub>X document (i.e., produce PDF document from L<sup>A</sup>T<sub>E</sub>X source)
- files in project:
  - CMakeLists.txt
  - main.tex
  - bib.bib
  - cmake\_modules/UseLATEX.cmake

main.tex

```
1 \documentclass{article}
2 \usepackage{graphicx}
3 \author{John Doe}
4 \title{Why I Like C++}
5 \begin{document}
6 \maketitle
7 \section{Why I Like C++}
8 What can I say?
9 C++~\cite{TCPL4} is a great language!\newline
10 \includegraphics[width=1in,height=1in,keepaspectratio]
11     {cpp.png}
12 \bibliographystyle{plain}
13 \bibliography{bib}
14 \end{document}
```

bib.bib

```
1 @book{
2     TCPL4,
3     author = "B. Stroustrup",
4     title = "The {C++} Programming Language",
5     edition = "4th",
6     publisher = "Addison Wesley",
7     year = 2013
8 }
```

## CMakeLists.txt

```
1 cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
2
3 # Specify the project name and indicate that no languages
4 # should be enabled.
5 project(my_doc NONE)
6
7 # Add the cmake_modules directory to the module search path.
8 set(CMAKE_MODULE_PATH ${CMAKE_MODULE_PATH}
9     ${CMAKE_SOURCE_DIR}/cmake_modules)
10
11 # Include the UseLATEX module.
12 include(UseLATEX)
13
14 # Specify the properties of the LaTeX document such as its
15 # constituent source files (e.g., LaTeX, BibTeX, images,
16 # figures, etc.)
17 add_latex_document(main.tex IMAGES cpp.png BIBFILES bib.bib)
```

## cmake\_modules/UseLATEX.cmake

This file is taken from <https://cmake.org/Wiki/images/8/80/UseLATEX.cmake>.

# Code Profiling Example: Overview

- want to provide file that can be included in CMakeLists file that defines option called `ENABLE_PROFILING` for building code with profiling enabled
- for sake of simplicity, only consider cases of using GCC and Clang C++ compilers
- place definition of option in file named `profiling.cmake`
- include `profiling.cmake` in `CMakeLists.txt`



# Code Profiling Example: profiling.cmake

## profiling.cmake

```
1  # Define an option for enabling code profiling, which is disabled
2  # by default.
3  option(ENABLE_PROFILING "Enable code profiling with gprof." false)
4
5  if (ENABLE_PROFILING)
6      if (CMAKE_CXX_COMPILER_ID MATCHES GNU OR
7          CMAKE_CXX_COMPILER_ID MATCHES Clang)
8          # The GCC or Clang C++ compiler is being used.
9          # Add the -pg option to the flags used for compiling.
10         set(CMAKE_CXX_FLAGS "${CMAKE_CXX_FLAGS} -pg")
11         # Add the -pg option to the linker flags used for creating
12         # executables.
13         set(CMAKE_EXE_LINKER_FLAGS "${CMAKE_EXE_LINKER_FLAGS} -pg")
14         # Add the -pg option to the linker flags used for creating
15         # shared libraries.
16         set(CMAKE_SHARED_LINKER_FLAGS
17             "${CMAKE_SHARED_LINKER_FLAGS} -pg")
18     else()
19         # Handle the case of unsupported compilers.
20         message(FATAL_ERROR
21             "Only GCC and Clang are currently supported.")
22     endif()
23 endif()
```

# Code Profiling Example: CMakeLists.txt

CMakeLists.txt (showing how one might use profiling.cmake)

```
1  cmake_minimum_required(VERSION 3.1 FATAL_ERROR)
2  project(gprof LANGUAGES CXX)
3
4  # Include the file containing the profiling option definition.
5  # Note: To enable profiling, one can add the command line option
6  # -DENABLE_PROFILING=true to cmake.
7  include(profiling.cmake)
8
9  add_executable(app app.cpp)
```

## Section 9.3.3

### **References**

- 1 Ken Martin and Bill Hoffman, Mastering CMake — A Cross-Platform Build System — CMake 3.1, Kitware, 2015. ISBN 978-1-930934-31-3.
- 2 CMake Tutorial (excerpt from the book “Mastering CMake”),  
<https://cmake.org/cmake-tutorial>.
- 3 CMake FAQ, [https://cmake.org/Wiki/CMake\\_FAQ](https://cmake.org/Wiki/CMake_FAQ).
- 4 CMake Wiki, <https://cmake.org/Wiki/CMake>.
- 5 Kenneth Moreland, “UseLATEX.cmake: L<sup>A</sup>T<sub>E</sub>X Document Building Made Easy,” Version 2.4.0. Available online at  
<https://cmake.org/Wiki/images/d/d7/UseLATEX.pdf>.
- 6 CMakeUserUseLATEX  
<https://cmake.org/Wiki/CMakeUserUseLATEX>
- 7 UseLATEX GitHub Site <https://github.com/kmorel/UseLATEX>

- 1 Bill Hoffman. Google Tech Talk — Building Science With CMake. October 8, 2015. Available online at <https://youtu.be/TqjtN8NGtl4>.  
A very basic introduction to CMake.
- 2 Daniel Pfeifer. Effective CMake. C++ Now, May 19, 2017, Aspen, CO, USA. Available online at <https://youtu.be/bsXLMQ6WgIk>.
- 3 Florent Castelli. Introduction to CMake. SwedenCpp::Stockholm 0xC, Sundbyberg, Sweden, Apr. 26, 2018. Available online at <https://youtu.be/jt3meXdP-QI>.
- 4 Deniz Bahadir. More Modern CMake — Working With CMake 3.12 and Later. Meeting C++, Berlin, Germany, Nov. 16, 2018. Available online at <https://youtu.be/TsddSCzYiRs>.

## Part 10

# Version Control Systems

## Section 10.1

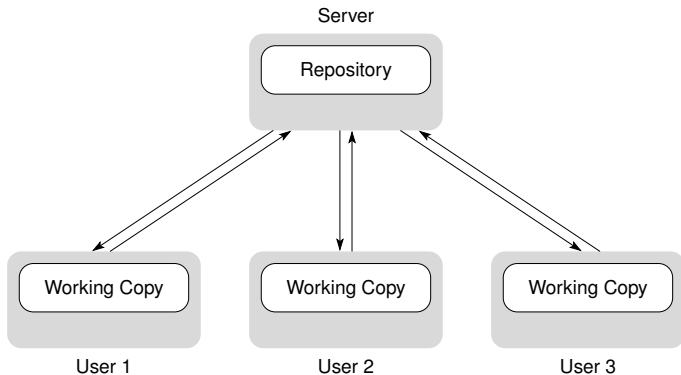
# Version Control Systems

# Version Control Systems

- **Version control** (also known as **revision control**) is the management of changes to programs, documents, and other collections of information.
- In practice, multiple versions of the same software will often be in existence at any given time.
- For the purposes of locating and fixing a bug, it is critically important to be able to access different versions of the software, since only certain versions of the software may have the bug.
- When concurrently developing multiple versions of some software, it is necessary to be able to keep track of what information belongs to which versions.
- Having developers manually maintain version information themselves is impractical, as this would be very error prone.
- Therefore, a version control system (VCS) is used to manage changes in a systematic manner.
- Some examples of VCSes include: Source Code Control System (SCCS), Revision Control System (RCS), Concurrent Versions System (CVS), Subversion, Mercurial, and Git.

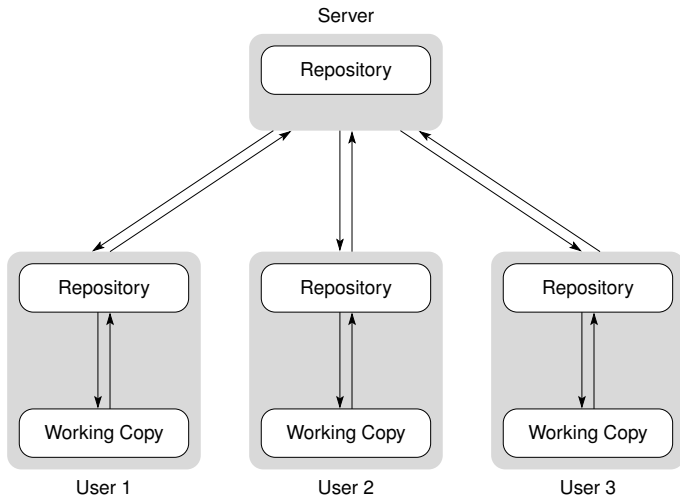


# Centralized Version Control



- repository resides only on server
- users do not have their own local copy of repository
- examples: CVS and Subversion

# Distributed Version Control



- each user has their own local copy of repository
- examples: Git and Mercurial

# Pros and Cons of Distributed Version Control

- advantages of distributed (over centralized) version control:
  - most operations (namely, ones that do not synchronize with other repositories) are local and extremely fast
  - all operations, except those that synchronize with other repositories, can be performed without network connection
  - more robust (e.g., data loss less likely due to replication of information across repositories, less reliance on network/server connectivity)
  - committing new changesets can be done locally without anyone else seeing them
  - easier to share changes with only, say, one or two people before showing changes to everyone
- disadvantages of distributed (relative to centralized) version control:
  - since repository is stored locally, more local disk space is required
  - if repository becomes large, downloading it can require considerable amount of time

## Section 10.2

### Git

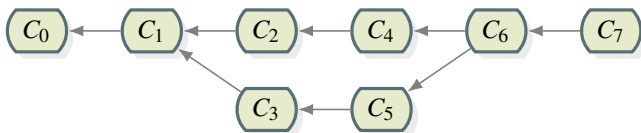
- Git is open-source distributed (i.e., decentralized) version control system
- created by Linus Torvalds
- development started in 2005 with first release made later in same year
- designed to support projects varying in size from very small to very large with speed and efficiency
- can efficiently handle very large numbers of files
- can efficiently handle large numbers of parallel branches
- revision history of file modelled as directed acyclic graph (DAG)
- official web site: <https://git-scm.com>

- Git has a very large user base and is employed heavily in industry
- some organizations using Git include:
  - Apple (<https://github.com/apple>)
  - eBay (<https://github.com/eBay>)
  - Facebook (<https://github.com/facebook>)
  - Google (<https://github.com/google>)
  - Intel (<https://github.com/intel>)
  - Microsoft (<https://github.com/Microsoft>)
  - NVIDIA (<https://github.com/NVIDIA>)
  - Twitter (<https://github.com/twitter>)
- some projects using Git include:
  - Linux Kernel (<https://github.com/torvalds/linux>)
  - Android (<https://android-review.googlesource.com>)
  - Qt (<http://code.qt.io>)
  - Gnome (<https://git.gnome.org>)
  - Eclipse (<https://git.eclipse.org>)
  - KDE (<https://github.com/KDE>)
  - FreeDesktop (<https://cgit.freedesktop.org>)

- A **repository** is effectively a database that records the information for all of the versions of all of the files in the directory tree under version control.
- A **commit** is simply a record (i.e., snapshot) of all of the files that comprise a particular version in the repository.
- This record includes, for each file, the location of the file in the directory tree as well as the contents of the file.
- For each version of the directory tree, the repository has a corresponding commit (i.e., snapshot).

# Revision History and Directed Acyclic Graphs

- The revision history can be represented as a directed acyclic graph (DAG).
- Each node in the graph corresponds to a commit in the repository.
- Each edge in the graph points to the immediately preceding commit in the revision history.
- Example of DAG:



$C_1$  based on changing content from  $C_0$ ;

$C_2$  based on changing content from  $C_1$ ;

$C_4$  based on changing content from  $C_2$ ;

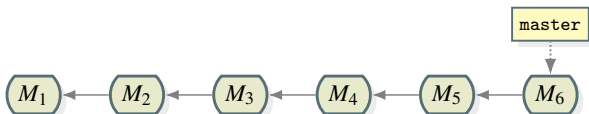
$C_6$  based on *merging* content from  $C_4$  and  $C_5$ ;

...

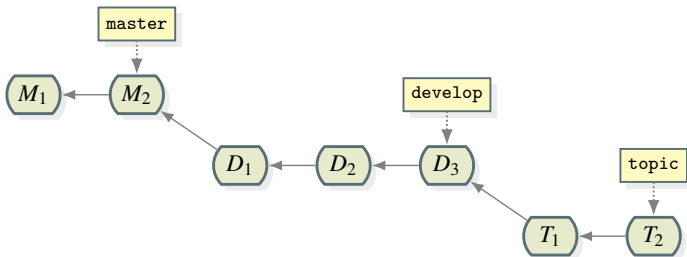


# Branching Workflows

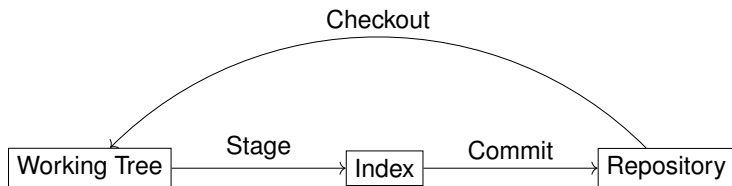
- single (master) branch:



- master, development, and topic branches:



- master branch: used for releases (highly stable, well tested)
- development branch: used for development work (possibly unstable)
- topic branch: used for highly experimental work



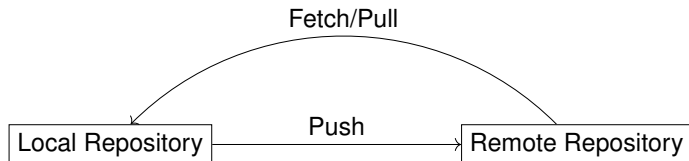
## ■ three distinct types of data:

- **working tree**: directory hierarchy containing files on which user is working
- **index** (also known as **staging area**): place where changes that are tentatively marked to be committed are stored
- **repository**: database used to store all versions of data and associated metadata

## ■ three basic local operations on data:

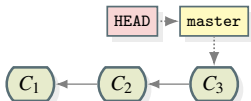
- **checking out**: populates working tree with particular version of data from repository
- **staging**: applies changes in working tree to index
- **committing**: applies changes in index to repository

# Local and Remote Picture



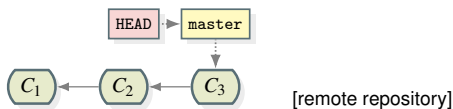
- **clone**: creates local repository that is copy of remote repository
- three basic operations for propagating changes between repositories:
  - **push**: propagate changes from local repository to remote repository
  - **fetch**: propagate changes from remote repository to local repository without merging changes
  - **pull**: propagate changes from remote repository to local repository and merge changes

- The name `HEAD` is a reference to the current working revision (i.e., a branch or commit) in the repository.
- Normally (i.e., except in the case of a detached `HEAD`), `HEAD` refers to the current branch.
- In the case of a detached `HEAD`, `HEAD` refers to a specific commit (not a branch).
- Consider a repository having a single `master` branch and three commits  $C_1$ ,  $C_2$ , and  $C_3$  (with  $C_3$  being the most recent), where the current branch is `master`. This would appear as follows:

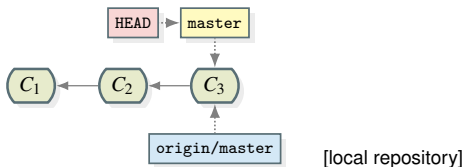


# Remote-Tracking Branches

- Consider a remote repository whose commit history is as shown below, with a single branch `master`.



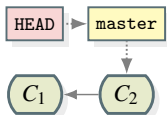
- Cloning the above repository will produce a new local repository whose commit history is as shown below, with a (local) branch `master` and a remote-tracking branch `origin/master`.



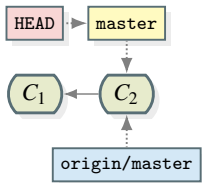
- A branch fetched from a remote repository is called a **remote-tracking branch**.
- A remote-tracking branch is a reference to a commit in the remote repository and is used for operations like pushing and fetching/pulling.

# Commit History Example I

- 1 Consider the following remote repository with a single branch `master`:

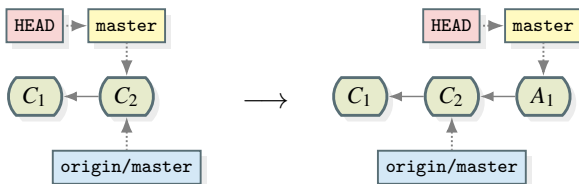


- 2 Cloning the remote repository yields a new local repository that is identical to the remote repository but with a remote-tracking branch `origin/master` added as follows:

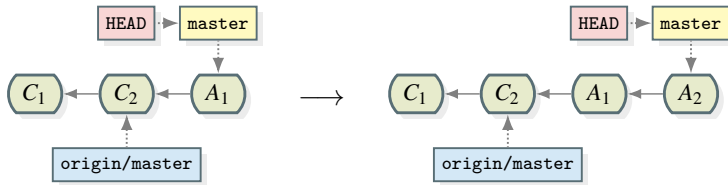


# Commit History Example II

- 3 Committing change  $A_1$  to the `master` branch of the local repository transforms the local repository as follows:

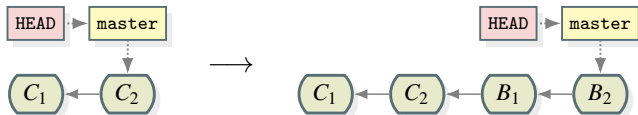


- 4 Committing change  $A_2$  to the `master` branch of the local repository transforms the local repository as follows:

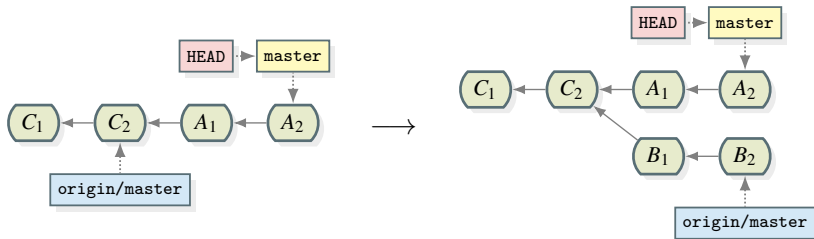


# Commit History Example III

- 5 Another user committing changes  $B_1$  and  $B_2$  to the remote repository transforms the remote repository as follows:



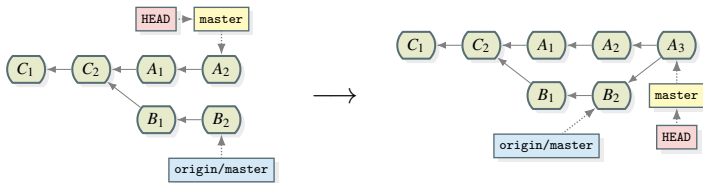
- 6 Fetching (to the local repository) from the remote repository transforms the local repository as follows:



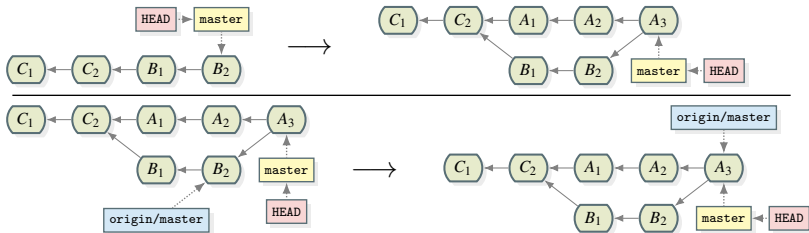


# Commit History Example IV

- 7 Merging the `origin/master` branch into the `master` branch (in the local repository) transforms the local repository as follows:



- 8 Pushing (from the local repository) to the remote repository transforms the remote and local repositories, respectively, as follows:



# Commit History Example: Commands to Setup Remote

```
1  TOP_DIR=`pwd`
2
3  cd $TOP_DIR
4  mkdir remote
5  cd remote
6  git init
7  printf "apple\n" >> fruits.txt
8  git add fruits.txt
9  git commit -m "Added file fruits.txt" # Commit C1
10 printf "banana\n" >> fruits.txt
11 git add fruits.txt
12 git commit -m "Added banana to fruits.txt" # Commit C2
13 git init --bare .git
14 mv .git $TOP_DIR/remote.git
15 cd $TOP_DIR
16 rm -rf remote
```

# Commit History Example: Remaining Commands

```
1  cd $TOP_DIR
2  git clone remote.git local-1
3
4  cd $TOP_DIR/local-1
5  printf "grape\n" >> fruits.txt
6  git add fruits.txt
7  git commit -m "Added grape to fruits.txt" # Commit A1
8  printf "orange\n" >> fruits.txt
9  git add fruits.txt
10 git commit -m "Added orange to fruits.txt" # Commit A2
11
12 cd $TOP_DIR
13 git clone remote.git local-2
14 cd local-2
15 printf "red\n" >> colors.txt
16 git add colors.txt
17 git commit -m "Added file colors.txt" # Commit B1
18 printf "green\n" >> colors.txt
19 git add colors.txt
20 git commit -m "Added green to colors.txt" # Commit B2
21 git push
22
23 cd $TOP_DIR/local-1
24 git push # ERROR: local repository not up to date
25 git fetch
26 git merge -m "Merged changes." # Commit A3
27 git push
```

- Git employs three levels of configuration settings, which in order of decreasing priority are as follows:
  - 1 local (i.e., per repository)
  - 2 global (i.e., per user)
  - 3 system (i.e., system wide)
- Configuring system settings may require administrator privileges.
- On Linux systems, the global settings are typically stored in the file `$HOME/.gitconfig`.

- Configure user information and clone the repository:

```
git config --global user.name "John Doe"  
git config --global user.email jdoe@gmail.com  
git clone $repository $directory
```

- Edit the working tree and stage changes as appropriate for the local repository:

```
git add $path_to_add  
git mv $source_path $destination_path  
git rm $file_to_remove  
git rm -r $directory_to_remove
```

- Check what changes are staged and then commit these changes to the local repository:

```
git status  
git commit
```

- Push changes from the local repository to the remote repository:

```
git push
```

- As needed, retrieve changes from the remote repository and merge them locally (e.g., if a push failed due to being out of date):

```
git pull
```

## Section 10.2.1

### **Basic Commands**

# Determining the Version of Git

- To query the version of the Git software, type:

```
git --version
```

- The original release dates for a few versions of Git are as follows:

Version	Original Release Date
1.0	2005-12-21
1.7	2010-02-13
1.8	2012-10-21
1.9	2014-02-14
2.0	2014-05-28
2.3	2015-02-05
2.8	2016-03-28
2.12	2017-02-24
2.16	2018-01-17

# Obtaining Help on the `git` Command

- To obtain general help for the `git` command, use a command of the form:

```
git help [options]
```

- To obtain detailed information for the `git` command or guide `$item`, use a command of the form:

```
git help [options] $item
```

- Some commonly-used options include:

Option	Description
<code>-a</code>	list all commands for which help available
<code>-g</code>	list all available help guides
<code>-w</code>	display in HTML format using a web browser
<code>-m</code>	display in man (i.e., manual page) format

- To obtain detailed help on the `commit` command with the information displayed in HTML format in a web browser, type:

```
git help -w commit
```

- To list all of the help guides available, type:

```
git help -g
```



# Configuring Git

- To set the variable `$name` to the value `$value`, use a command of the form:

```
git config [options] $name $value
```

- To unset the variable `$name`, use a command of the form:

```
git config [options] --unset $name
```

- To list all of the current variables settings, use a command of the form:

```
git config [options] -l
```

- Some commonly-used options include:

Option	Description
<code>--system</code>	consider only the system-wide settings
<code>--global</code>	consider only the global (i.e., per-user) settings
<code>--local</code>	consider only the local (i.e., per-repository) settings

# Some Commonly-Used Git Variables

Variable	Description
<code>core.askPass</code>	program for entering user name and password credentials
<code>core.editor</code>	program for editing
<code>core.pager</code>	program for paging output
<code>credential.helper</code>	external program to be called when a user name or password credential is needed
<code>user.name</code>	user's full name
<code>user.email</code>	user's email address
<code>web.browser</code>	program for browsing web

# Configuring User Information

- To globally set the user name to “John Doe”, type:

```
git config --global user.name "John Doe"
```

- To globally set the email address to “jdoe@gmail.com”, type:

```
git config --global user.email jdoe@gmail.com
```

- To list all system, global, and local variables, type:

```
git config -l
```

- To list only the global variables, type:

```
git config --global -l
```

- To list only the local (i.e., per-repository) variables for the current repository, type:

```
git config --local -l
```

# Configuring User-Credential-Related Information

- To enable the global caching of user credentials for 1 hour (i.e., 3600 seconds), type:

```
git config --global \
    credential.helper 'cache --timeout=3600'
```

- To disable all caching of user credentials (i.e., at the system, global, and repository levels) and purge any cached values, type:

```
git config --unset credential.helper
git config --global --unset credential.helper
git config --system --unset credential.helper
git credential-cache exit
```

- To ensure that prompting for user credentials employs standard input/output (as opposed to, say, a pop-up window), type:

```
git config --unset core.askPass
git config --global --unset core.askPass
git config --system --unset core.askPass
unset GIT_ASKPASS
unset SSH_ASKPASS
```

# Creating an Empty Repository

- To create an empty repository, use a command of the form:

```
git init [$directory]
```

- If `$directory` is not specified, it defaults to the current directory.
- The repository is created in the directory `$directory`.
- If `$directory` already contains a repository, the repository is re-initialized (in a non-destructive manner).
- All of the information used internally by Git to maintain the state of the repository is stored in a directory named `.git` at the top-level directory in the working tree.
- To create a new repository in the directory `hello` (which does not currently exist), type:  

```
git init hello
```

# Cloning a Repository

- To clone a repository `$repository`, use a command of the form:  
`git clone [options] $repository [$directory]`
- If `$directory` is not specified, it defaults to a value derived from `$repository`.
- The cloned repository is placed in the directory `$directory`.
- The repository specifier `$repository` can be a URL (for a repository accessed through a network server) or a directory (for a repository accessed through the local file system).
- Some commonly-used options include:

Option	Description
<code>-b \$branch</code>	after cloning, checkout the branch <code>\$branch</code>

- To clone the repository associated with the URL `https://github.com/uvic-aurora/hello-world.git` to the directory `hello-world`, type:  

```
git clone \  
https://github.com/uvic-aurora/hello-world.git \  
hello-world
```

# Adding Files/Directories to the Index

- To add the files/directories `$path...` to the index (i.e., mark for committing later), use a command of the form::

```
git add $path...
```

- The contents of `$path...` *at the time that the `git add` command is run* are staged; subsequent changes to these contents are not automatically staged.
- When a directory is staged, all directories and files that are contained under it are staged (i.e., staging is recursive).
- To prevent certain files/directories in a directory from being staged, they can be listed in a `.gitignore` file in that directory.
- To add the files `README` and `LICENSE` to the index, type:

```
git add README LICENSE
```

- To add the directory `src` (and everything contained under it) to the index, type:

```
git add src
```

# Removing Files/Directories from the Index

- To remove all changes from the index, type:  
`git reset`
- To remove the files/directories `$path...` from the index, use a command of the form:  
`git reset $path...`
- To undo the effects of the command “`git add README LICENSE`”, type:  
`git reset README LICENSE`
- To undo the effects of the command “`git add README`”, type:  
`git reset README`



# Renaming Files

- To move the file/directory `$source` to `$destination`, use a command of the form:

```
git mv [options] $source $destination
```

- To move multiple files `$s_1`, `$s_2`, ..., `$s_n` to the directory `$destination_directory`, use a command of the form:

```
git mv [options] $s_1 $s_2 ... $s_n \  
$destination_directory
```

- Some commonly-used options include:

Option	Description
<code>-f</code>	force moving even if the target exists

- To rename the file `README` to `README.old`, type:

```
git mv README README.old
```

- To move the files `hello.cpp` and `goodbye.cpp` to the directory `src`, type:

```
git mv hello.cpp goodbye.cpp src
```

# Removing Files

- To remove the files/directories `$path...` from the working tree and the index, use a command of the form:

```
git rm [options] $path...
```

- Some commonly-used options include:

Option	Description
<code>-f</code>	override the up-to-date check
<code>-r</code>	if the given path is a directory, recursively remove files below it

- To remove the directory `src` and all files and directories beneath it from the working tree and index, type:

```
git rm -r src
```

- To remove the files `README` and `LICENSE` from the working tree and index, type:

```
git rm README LICENSE
```

# Committing Changes

- To commit all staged changes, use a command of the form:

```
git commit [options]
```

- Some commonly-used options include:

Option	Description
-a	automatically stage any files that have been modified or deleted
-m \$message	set message to \$message

- To commit all staged changes with the message “Fixed overflow bug”, type:

```
git commit -m "Fixed overflow bug"
```

- To commit all staged changes with the message “Fixed overflow bug”, automatically staging any files that have been modified or deleted, type:

```
git commit -a -m "Fixed overflow bug"
```

# Checking the Status of the Working Tree

- To display the status of the working tree, use a command of the form:

```
git [options] status
```

- Some commonly-used options include:

Option	Description
--long	give the output in long format (default)
--short	give the output in short format

- The information displayed by this command includes:

- paths (i.e., files and directories) that have differences between the index and the current HEAD commit, (i.e., what would be committed by running `git commit`)
- paths that have differences between the working tree and index as well as paths that are not tracked by Git (i.e., what could be committed by running `git add` before `git commit`)

- To display the status of the working tree in long form, type:

```
git status
```

# Showing Commit Logs

- To show the commit history (which can be limited to a particular revision range `$revision_range` or files/directories `$path...`), use a command of the form:

```
git log [options] [$revision_range] [--] $path...
```

- Some commonly-used options include:

Option	Description
<code>--since \$date</code>	select commits more recent than date <code>\$date</code>
<code>--until \$date</code>	select commits older than date <code>\$date</code>
<code>-\$n</code>	select last <code>\$n</code> commits
<code>-S \$pattern</code>	select commits adding/removing string matching pattern <code>\$pattern</code>
<code>--graph</code>	draw text-based graph of commit history
<code>--all</code>	consider all branches/remotes/tags
<code>--grep \$pattern</code>	select only commits with message matching <code>\$pattern</code>

## Showing Commit Logs (Continued)

- To show the commit history for the file `README` since `2016-01-01`, type:  

```
git log --since 2016-01-01 README
```
- To show the commit history for all files between `2014-01-01` and `2014-12-31`, type:  

```
git log --since 2014-01-01 --until 2014-12-31
```
- To show the commit history for all files in all branches with a text-based graph, type:  

```
git log --all --graph
```
- To show the commit history for all commits made since `v1.0` until and including `v2.0` (assuming that `v1.0` and `v2.0` exist), type:  

```
git log v1.0..v2.0
```

# Showing Changes

- To show changes between the working tree and the index (i.e., what could be staged but has not yet been) for files/directories `$path...` (which defaults to all files/directories), use a command of the form:

```
git diff [options] [$path...]
```

- To show changes between the index and the named commit `$commit` (which defaults to `HEAD`) for the files/directories `$path...` (which defaults to all files/directories), use a command of the form:

```
git diff [options] --cached [$commit] -- \
[$path...]
```

- To show changes between the working tree and the named commit `$commit` (which defaults to `HEAD`) for the files/directories `$path...` (which defaults to all files/directories), use a command of the form:

```
git diff [options] [$commit] -- [$path...]
```

- To show changes between two arbitrary commits `$commit1` and `$commit2` for the files/directories `$path...`, use a command of the form:

```
git diff [options] $commit1 $commit2 -- \
[$path...]
```

## Showing Changes (Continued)

- Some commonly-used options include:

Option	Description
<code>-b</code>	ignore changes in amount of whitespace
<code>-w</code>	ignore all whitespace
<code>--ignore-blank-lines</code>	ignore blank lines

- To show all changes between the working tree and the index for all files/directories, type:  
`git diff`
- To show all differences between the working tree and the index for the file `README`, ignoring changes in amount of whitespace, type:  
`git diff -b README`



# Finding Lines Matching a Pattern

- To find all lines of text in the files `$path...` (which defaults to all files) in the working tree that satisfy the condition specified by the `p_options`, use a command of the form:

```
git grep [options] [p_options] -- [$path...]
```

- Some commonly-used options include:

Option	Description
<code>-l</code>	print only names of files with matches
<code>-i</code>	ignore case
<code>--max-depth \$depth</code>	descend at most <code>\$depth</code> levels of directories
<code>-v</code>	select non-matching lines
<code>-F</code>	patterns are fixed strings
<code>-E</code>	patterns are extended POSIX regular expressions
<code>-e \$pattern</code>	specify pattern <code>\$pattern</code>
<code>--and</code>	logical and
<code>--or</code>	logical or
<code>--not</code>	logical not
<code>(</code>	for grouping logical operations
<code>)</code>	for grouping logical operations
<code>--cache</code>	search in the index instead of the working tree

# Finding Lines Matching a Pattern (Continued)

- To search for the text “hello” in a case insensitive manner in all of the files in the working tree, type:

```
git grep -i -e hello
```

- To print only the names of the files that match the pattern specified in the preceding example, type:

```
git grep -i -e hello -l
```

- To find all of the files in the working tree with suffixes “.cpp” or “.hpp” that have lines containing either “#include <vector>” or “#include <list>”, type:

```
git grep -e '#include <vector>' --or \  
-e '#include <list>' -- '*.cpp' '*.hpp'
```

- To perform the same search as the preceding example but in the index rather than the working tree, type:

```
git grep --cache -e '#include <vector>' --or \  
-e '#include <list>' -- '*.cpp' '*.hpp'
```

# Removing Untracked Files and Directories

- To remove all untracked files in the working tree, use a command of the form:

```
git clean [options]
```

- Some commonly-used options include:

Option	Description
-d	remove untracked directories in addition to untracked files
-f	force removal of files
-i	enable interactive mode
-n	show what would be done without actually doing anything
-x	do not use standard ignore rules (such as those specified in <code>.gitignore</code> files)

- To remove all untracked files and directories in the working tree excluding those ignored by Git, type:

```
git clean -d -f
```

- To remove all untracked files and directories in the working tree including those ignored by Git, type:

```
git clean -d -f -x
```

# .gitignore Files

- A `.gitignore` file specifies which files and directories are intentionally untracked and should be ignored by Git.
- The purpose of a `.gitignore` file is to ensure that certain files not tracked by Git remain untracked.
- The `.gitignore` file lists patterns specifying files that should be ignored by Git.
- A “!” prefix negates a pattern.
- A leading slash matches the directory containing the `.gitignore` file. For example, `/hello.cpp` matches `hello.cpp` but not `some/subdirectory/hello.cpp`.
- The patterns in a `.gitignore` file apply to the directory containing the file as well as all directories below the file in the working tree.
- The patterns in a `.gitignore` file at a higher level in the tree are overridden by patterns in a `.gitignore` file at a lower level.
- A `.gitignore` file in the root directory of the working tree can be used to establish ignore defaults for the whole tree.

# .gitignore File Example

```
# ignore all object files
*.o
# ignore all library files
*.a
# ignore foobaz only in this directory
/foobaz
# ignore foo only in directory example
/example/foo
```

- A `.gitattributes` file is used to specify attributes for files/directories.
- For example, the determination of whether a file employs a binary or text format can be controlled via a `.gitattributes` file.
- An example of a `.gitattributes` file is as follows:

```
# Consider all PNM files to be binary.  
*.pnm binary
```
- The settings in a `.gitattributes` file apply to the directory containing the file as well as all directories below the file in the working tree.
- The settings in a `.gitattributes` file at a higher level in the tree are overridden by settings in a `.gitattributes` file at a lower level.
- A `.gitattributes` file in the root directory of the working tree can be used to establish attribute defaults for the whole tree.

# Tracking Empty Directories

- The current implementation of `git` does not allow empty directories to be tracked.
- The best workaround for this problem is to create a `.gitignore` file in the directory that ignores all files except the `.gitignore` file itself.
- Such a `.gitignore` file might look like the following:

```
# First, ignore everything in this directory.  
*  
  
# Now, override the preceding rule and force  
# the .gitignore file not to be ignored.  
!.gitignore
```

- This is not a perfect solution as it requires that the “empty” directory contain one file (namely, a `.gitignore` file) and this file be committed to the repository.

## Section 10.2.2

### **Remote-Related Commands**



# Listing, Adding, and Removing Remotes

- To list the remotes, use a command of the form:

```
git remote [general_options]
```

- Some general options include:

Option	Description
-v	enable verbose mode

- To add the remote `$remote` with the associated URL `$url`, use a command of the form:

```
git remote add $remote $url
```

- To remove the remote `$remote`, use a command of the form:

```
git remote rm $remote
```

- To rename a remote from `$old` to `$new`, use a command of the form:

```
git remote rename $old $new
```

- To show detailed information on the remote `$remote`, use a command of the form:

```
git remote [general_options] show $remote
```

# Fetching Changes from Another Repository

- To fetch changes from the remote `$remote` (which normally defaults to `origin`), use a command of the form:  

```
git fetch [$remote]
```
- A fetch operation gathers any commits from the target branch that do not exist in current branch, and stores them in the local repository.
- It is always safe to perform a fetch in sense that no conflicts can arise, since no merge is attempted.
- By fetching frequently, one can keep their local repository up to date without being forced to merge.

# Pushing Changes to Another Repository

- To push changes to the branch `$branch` (which normally defaults to the current branch) of the remote `$remote` (which normally defaults to `origin`), use a command of the form:

```
git push [options] [$remote [$branch]]
```

- When pushing a new local branch to a remote, the `-u` option should be specified.
- To delete the branch `$branch` on the remote `$remote` only, use a command of the form:

```
git push --delete origin $branch
```

- The preceding command is useful if one wants to delete a branch that exists on the remote but not in the local repository.
- To delete the tag `$tag` on the remote `$remote` only, use a command of the form:

```
git push --delete origin $tag
```

- To push to the default remote and branch, type:

```
git push
```

# Pulling Changes from Another Repository

- To pull changes from the branch `$branch` of the remote `$remote`, use a command of the form:

```
git pull [$remote [$branch]]
```

- To pull from the default remote and branch, type:

```
git pull
```

- A pull is approximately a fetch followed by merge.
- A pull automatically merges commits without letting them be reviewed first.
- For this reason, some people suggest that it is better to use fetch and merge separately instead of performing a pull.
- Also, the use of pull operations can, in some cases, result in unnecessary merge commits.

- To merge changes from the branch `$branch` (which normally defaults to the upstream branch for the current branch) into the current branch, use a command of the form:

```
git merge [$branch]
```

- To merge from the default branch, type:

```
git merge
```

- Note that the merge direction is from the branch `$branch` into the current branch.
- It is advisable to ensure that any outstanding changes are committed before running `git merge` in order to reduce the likelihood of major difficulties in the case of a conflict.
- If a conflict arises, no commit will be performed and manual intervention is required to resolve the conflict.

## Section 10.2.3

### **Branch-Related Commands**

# Listing, Creating, and Deleting Branches

- To list all of the branches, use a command of the form:

```
git branch [options]
```

- To create a branch `$branch`, use a command of the form:

```
git branch [options] $branch
```

- To delete the (local) branch `$branch`, use a command of the form:

```
git branch [options] -d $branch
```

- Some commonly-used options include:

Option	Description
-a	list both remote and local branches
-r	list or delete remote branches
-v	enable verbose mode for listing (use twice for extra verbose)

- When a new branch is created with `git branch`, this does not automatically checkout (i.e., switch to using) the new branch.

# Checking Out a Branch

- To checkout (i.e., switch to) the branch `$branch`, use a command of the form:  

```
git checkout $branch
```
- Checking out a branch changes the files/directories of the working tree to match that branch.
- If you have local modifications to one or more files that are different between the current branch and the branch to which you are switching, the command refuses to switch branches in order to preserve your modifications in context.



## Section 10.2.4

### **Tag-Related Commands**

# Listing, Creating, and Deleting Tags

- To list all tags, type:

```
git tag
```

- To tag a commit `$commit` (which defaults to `HEAD`) with the name `$name`, use a command of the form:

```
git tag [options] $name [$commit]
```

- To delete the (local) tags with names `$name...`, use a command of the form:

```
git tag -d $name...
```

- Some commonly-used options include:

Option	Description
-a	make an unsigned annotated tag

- To create an annotated tag `version-1.0` for the most recent commit on the `master` branch, type:

```
git tag -a version-1.0 master
```

- To delete the tag `version-1.0`, type:

```
git tag -d version-1.0
```

- To push a tag `$tag` to the remote `$remote`, use a command of the form:  
`git push $remote $tag`
- To push the tag `v1.0` to the remote `origin`, type:  
`git push origin v1.0`

## Section 10.2.5

### **Miscellany**

# Duplicating a Repository

- can create exact duplicate of entire Git repository (including all tags and local branches) by using bare-clone and mirror-push operations
- to copy repository `$source_repo` to (already existing) remote repository `$destination_repo` (overwriting contents of repository), use command sequence:

```
# Create a bare clone of the repository.  
git clone --bare $source_repo $bare_dir  
# Mirror push to the destination repository.  
git -C $bare_dir push --mirror $destination_repo  
# Remove the temporary local bare repository.  
rm -rf $bare_dir
```

- to copy repository `$source_repo` to local repository directory `$destination_dir`, simply perform bare clone operation using command:  

```
git clone --bare $source_repo $destination_dir
```

# Avoiding Repeated Passphrase Entry for SSH Authentication

- if SSH used to access repository, SSH passphrase often needs to be provided
- to avoid having to enter SSH passphrase every time it is needed, can use SSH Agent to cache passphrase and provide it as required
- to start SSH Agent and provide it with passphrase to cache for private key file `$key_file`, use command sequence:

```
# Start SSH Agent
eval `ssh-agent`
# Provide passphrase for particular key.
ssh-add $key_file
```
- on Unix systems, SSH key information typically stored in directory `$HOME/.ssh`

- A file is said to be **derived** if it is generated from one or more other files (e.g., an object file is derived from its corresponding source code file, a PDF or PostScript file is derived from its corresponding  $\text{\LaTeX}$  source files).
- Do not place derived files under version control, as such files are completely redundant and can often lead to a significant increase in repository size.
- Do not place large unchanging datasets under version control, as this will greatly increase repository size without any tangible benefit (i.e., since the datasets are not changing, there will never be multiple versions of them to manage).
- Avoid placing sensitive information (e.g., passwords) under version control.
- Remember that deleting a file from a particular commit does not delete that file from the repository, since the file will still exist in other commits.

- Gitg. A GNOME GUI client for viewing Git repositories.  
<https://wiki.gnome.org/Apps/Gitg>.
- Meld. A visual diff and merge tool. <http://meldmerge.org>.
- Hub. A command-line wrapper for Git that facilitates easier use of GitHub.  
<https://hub.github.com>.



## Section 10.2.6

### **References**

- 1 Official Git Web Site. <https://git-scm.com>, August 2016.

This web site has many excellent resources related to Git, including:

- 1 Git Downloads: <https://git-scm.com/downloads>.

This web page has the Git software for various platforms, including Linux, Mac OS X, and Windows.

- 2 Git Book: Scott Chacon and Ben Straub, Pro Git,  
<http://git-scm.com/book>.

This online book can also be downloaded in several formats (including PDF).

- 3 Git Videos: <https://git-scm.com/videos>.

This web page has several short videos on various aspects of Git.

- 2 Good Resources for Learning Git and GitHub, <https://help.github.com/articles/good-resources-for-learning-git-and-github>  
August 2016.

This web page has a list of many excellent resources for learning both Git and GitHub.

- 3 TryGit Tutorial, <https://try.github.com>, August 2016.  
This online Git tutorial allows the user to try Git in their web browser.
- 4 J. Loeliger. *Version Control with Git*.  
O'Reilly, Sebastopol, CA, USA, 2009.

- 1 Linus Torvalds. Google Tech Talk: Linus Torvalds on git — Git: Source code control the way it was meant to be!. May 2007. Available online at <https://youtu.be/4XpnKHJAok8>.  
Linus Torvalds shares his thoughts on Git, the source control management system he created.
- 2 Matthew McCullough. The Basics of Git and GitHub. July 2013. Available online at <https://youtu.be/U8GBXvdmHT4>.  
This is an excellent introduction to using Git.
- 3 Scott Chacon. Introduction to Git with Scott Chacon of GitHub. June 2011. Available online at <https://youtu.be/ZDR433b0HJY>.  
This is another popular introduction to using Git.
- 4 Matthew McCullough. Advanced Git: Graphs, Hashes, and Compression, Oh My!. Sept. 2012. Available online at <https://youtu.be/ig5E8CcdM9g>.  
This is a very good more advanced talk on Git.

## Part 11

### Miscellaneous Tools

## Section 11.1

### Online C++ Compilers

# Online C++ Compilers

- numerous online (i.e., web-based) C++ compilers available
- typically provide means to upload, edit, build, and run code
- usually provide some control over compiler options and libraries
- often support multiple compilers and/or compiler versions
- some examples of online C++ compilers include:
  - Wandbox
    - <http://melpon.org/wandbox>
  - Ideone
    - <https://ideone.com>
  - Coliru
    - <http://coliru.stacked-crooked.com>
  - Repl.it
    - <https://repl.it/repls/SmallIvorySyntax>
  - [Compiler Explorer](#) (discussed in more detail shortly)
  - [C++ Insights](#) (discussed in more detail shortly)

# Online C++ Compiler Example

[Donate](#) [Coliru](#) [Restore defaults](#) [Help](#) [Feedback](#)  
[Editor](#) [Command](#) [Q&A](#) [Read](#) [Write](#)

```
1 #include <iostream>
2 #include <string>
3 #include <vector>
4
5 template<typename T>
6 std::ostream& operator<<(std::ostream& os, const std::vector<T>& vec)
7 {
8     for (auto& el : vec)
9     {
10         os << el << ' ';
11     }
12     return os;
13 }
14
15 int main()
16 {
17     std::vector<std::string> vec = {
18         "Hello", "from", "GCC", __VERSION__, "!"
19     };
20     std::cout << vec << std::endl;
21 }
22
```

g++ -std=c++17 -O2 -Wall -pedantic -pthread main.cpp && ./a.out

Compile, link and run...

Share!



# Compiler Explorer

- Compiler Explorer is interactive web-based compiler that supports C++, C, and numerous other languages
- can load, save, and edit source code
- source code is automatically compiled as edited
- can simultaneously display source code and corresponding assembly code
- can select language, compiler, and compiler options
- source code can utilize numerous libraries (such as Boost)
- extremely useful for understanding what assembly code is generated by compiler
- also helpful for quickly testing compilation of small code snippets with variety of compilers
- created by Matt Godbolt
- web site running Compiler Explorer: <https://godbolt.org>
- GitHub page:  
<https://github.com/mattgodbolt/compiler-explorer>

# Some Compiler Explorer Functionalities

- uses color coding to show correspondence between lines of source and assembly code
- can hover over instruction name in assembly code for information on instruction
- display of assembly code can be controlled in various ways:
  - show or hide machine op codes
  - use mangled versus demangled names
  - show or hide debugging information
  - trim extra whitespace
  - filter unused labels or extra assembler directives
- can display assembly code from multiple compilers at same time
- can draw control flow graph corresponding to assembly code
- can be used to test whether different compilers can successfully compile particular source code
- can share code and Compiler Explorer session via URL
- does not support execution of code (at least as of October 2018)

# Compiler Explorer Example



Editor Diff View More ▾

C++ Insights shows how compilers see your code ✕

Share Other ▾ Policies ▾

C++ source #1 ✕

A ▾ Save/Load + Add new... ▾

C++ ▾

```
1 #include <iostream>
2
3 int get_answer() {
4     return 42;
5 }
6
7 int main() {
8     std::cout << get_answer() << '\n';
9 }
10
```

x86-64 gcc 8.2 (Editor #1, Compiler #1) C++ ✕

x86-64 gcc 8.2 ▾

-std=c++17 -O3

A ▾ 11010 ☒ LX0: ☒ text ☒ // ☐ \s+ ☒ Intel ☒ Demangle Libraries + Add new... ▾

```
1 get_answer():
2     mov     eax, 42
3     ret
4 main:
5     sub     rsp, 24
6     mov     esi, 42
7     mov     edi, OFFSET FLAT:_ZSt4cout
8     call    std::basic_ostream<char, std::char_traits<char>>::operator<< (std::basic_ostream<char, std::char_traits<char>>*, const char*)@plt
9     lea     rsi, [rsp+15]
10    mov     edx, 1
11    mov     BYTE PTR [rsp+15], 10
12    mov     rdi, rax
13    call    std::basic_ostream<char, std::char_traits<char>>::operator<< (std::basic_ostream<char, std::char_traits<char>>*, const char*)@plt
14    xor     eax, eax
15    add     rsp, 24
16    ret
17 _GLOBAL__sub_I_get_answer():
18    sub     rsp, 8
19    mov     edi, OFFSET FLAT:_ZStL8__ioinit
20    call    std::ios_base::Init::Init() [complete object]
21    mov     edx, OFFSET FLAT:__dso_handle
22    mov     esi, OFFSET FLAT:_ZStL8__ioinit
23    mov     edi, OFFSET FLAT:_ZNSt8ios_base4InitD1Ev
24    add     rsp, 8
25    jmp     __cxa_atexit
```


Output (0/0) x86-64 gcc 8.2 ⓘ -cached (3336788)

Compiler Explorer uses cookies & related technologies. [Check our cookie policy](#)

Do NOT allow nonessential cookies

Allow cookies

# Compiler Explorer Example

Editor Diff View More ▾Share Other ▾ Policies ▾

C++ source #1 x

Save/Load + Add new... ▾

C++ ▾

```
1 #include <iostream>
2
3 int get_answer() {
4     return 42;
5 }
6
7 int main() {
8     std::cout << get_answer() << '\n';
9 }
```

x86-64 gcc 8.2 (Editor #1, Compiler #1) C++ x

x86-64 gcc 8.2 ▾

✓ -O3

A ▾

☐ 11010 ☒ .LX0: ☒ .text ☒ // ☐ \s+ ☒ Intel ☒ Demangle

Libraries + Add new... ▾

```
1 get_answer():
2     mov     eax, 42
3     ret
4 main:
5     sub     rsp, 24
6     mov     esi, 42
7     mov     edi, OFFSET FLAT:_ZSt4cout
8     call    std::basic_ostream<char,
9           __mbstate_t>::operator<<(const
10          int&, const char*)<
11          [rsp+15], 10
```

Output (0/0) x86-64 gcc 8.2 i - cached (1196188)

Transfers program control to a return address located on the top of the stack. The address is usually placed on the stack by a CALL instruction, and the return is made to the instruction that follows the CALL instruction. More information available in the context menu.

```
16     ret
17 _GLOBAL__sub_I_get_answer():
18     sub     rsp, 8
19     mov     edi, OFFSET FLAT:_ZStL8
20     call    std::ios_base::Init::In
21     mov     edx, OFFSET FLAT:___dso_
22     mov     esi, OFFSET FLAT:_ZStL8
23     mov     edi, OFFSET FLAT:_ZNSt8
24     add     rsp, 8
```

Compiler Explorer uses cookies & related technologies.  
[Check our cookie policy](#)

Do NOT allow nonessential cookies

Allow cookies

# Example Showing GCC Code-Coverage Instrumentation


The screenshot displays the Compiler Explorer interface. On the left, the C++ source code for a function `unitstep(double x)` is shown. The function returns 1.0 if `x >= 0.0`, and 0.0 otherwise. The middle pane shows the assembly code for the function, with some lines highlighted in green. The right pane shows the assembly code with GCC code-coverage instrumentation, including labels like `__gcov0.unitstep(double)(%rip)` and `__gcov0.unitstep(double)+8(%rip)`, and a `__gcov__` section at the bottom.

```
1 double unitstep(double x) {
2     if (x >= 0.0) {
3         return 1.0;
4     } else {
5         return 0.0;
6     }
7 }
8
```

```
1 unitstep(double):
2     pushq   %rbp
3     movq    %rsp, %rbp
4     movsd   %xmm0, -8(%rbp)
5
6     movsd   -8(%rbp), %xmm0
7     pxor    %xmm1, %xmm1
8     comisd   %xmm1, %xmm0
9     jnb     .L6
10
11     movsd   .LC1(%rip), %xmm0
12     jmp     .L4
13
14 .L6:
15     pxor    %xmm0, %xmm0
16
17 .L4:
18     popq    %rbp
19     ret
20
```

```
1 unitstep(double):
2     pushq   %rbp
3     movq    %rsp, %rbp
4     movsd   %xmm0, -8(%rbp)
5 +     movq    __gcov0.unitstep(double)(%rip), %rax
6 +     addq    $1, %rax
7 +     movq    %rax, __gcov0.unitstep(double)(%rip)
8     movsd   -8(%rbp), %xmm0
9     pxor    %xmm1, %xmm1
10    comisd   %xmm1, %xmm0
11    jnb     .L6
12 +     movq    __gcov0.unitstep(double)+8(%rip), %rax
13 +     addq    $1, %rax
14 +     movq    %rax, __gcov0.unitstep(double)+8(%rip)
15    movsd   .LC1(%rip), %xmm0
16    jmp     .L4
17
18 .L6:
19     pxor    %xmm0, %xmm0
20
21 .L4:
22     popq    %rbp
23     ret
24
25 + __gcov__.unitstep(double):
26 +     .quad   .LPBX0
27 +     .long   273018407
28 +     .long   559119906
29 +     .long   -1032003044
30 +     .zero   4
31 +     .long   2
32 +     .zero   4
33 +     .quad   __gcov0.unitstep(double)
34 + .LC2:
35 +     .string "/tmp/compiler-explorer-compiler1181018-56"
36 + .LPBX0:
```

# Example Showing Short-Circuit Evaluation

 COMPILER EXPLORER

Add... More

Share Other Policies

C++ source #1

x86-64 gcc 8.2 (Editor #1, Compiler #1) C++

A Save/Load Add new... C++

```
1 bool func(bool a, bool b, bool c)
2 {
3     if ((a || b) && c) {
4         return true;
5     } else {
6         return false;
7     }
8 }
```

x86-64 gcc 8.2 -O0 -fverbose-asm

A 11010 .LX0: .text /s+ Intel Demangle Libraries

Add new... Add tool...

```
3 func(bool, bool, bool):
4     pushq    %rbp    #
5     movq    %rsp, %rbp    #,
6     movl    %edi, %eax    # a, tmp89
7     movl    %esi, %ecx    # b, tmp91
8     movb    %al, -4(%rbp)    # tmp90, a
9     movl    %ecx, %eax    # tmp91, tmp92
10    movb    %al, -8(%rbp)    # tmp92, b
11    movl    %edx, %eax    # tmp93, tmp94
12    movb    %al, -12(%rbp)    # tmp94, c
13    cmpb    $0, -4(%rbp)    #, a
14    jne     .L2    #,
15    cmpb    $0, -8(%rbp)    #, b
16    je      .L3    #,
17.L2:
18    cmpb    $0, -12(%rbp)    #, c
19    je      .L3    #,
20    movl    $1, %eax    #, _1
21    jmp     .L4    #
22.L3:
23    movl    $0, %eax    #, _1
24.L4:
25    popq    %rbp    #
26    ret
```

Output (0/0) x86-64 gcc 8.2 - cached (6428B)

- C++ Insights is Clang-based tool that makes visible numerous code transformations performed by compiler
- created by Andreas Fertig
- for example, tool allows for things such as following to be seen:
  - operator invocations
  - expansion of range-based for loop into “classic” for loop
  - classes generated from lambda expressions
  - code generation associated with structured bindings
- C++ Insights page (which allows tool to be accessed via web browser):
  - <https://cppinsights.io>
- GitHub repository page:
  - <https://github.com/andreasfertig/cppinsights>

# Example: Structured Bindings and Operators



Source:

```
1 #include <iostream>
2
3 struct Point {
4     int x;
5     int y;
6 };
7
8 Point get_point()
9 {
10     return {1, 2};
11 }
12
13 int main()
14 {
15     Point p;
16     auto [a, b] = get_point();
17     std::cout << a << '\n';
18     std::cout << b << '\n';
19 }
20
```

Console:

Insights exited with result code: 0



Issues

About

Made by [Andreas Fertig](#)

Powered by [Flask](#) and [CodeMirror](#)

Insight:

```
1 #include <iostream>
2
3 struct Point {
4     int x;
5     int y;
6 /* public: inline Point() noexcept; */
7 /* public: inline constexpr Point(const Point &); */
8 /* public: inline constexpr Point(Point &&); */
9 /* public: inline ~Point() noexcept; */
10 };
11
12 Point get_point()
13 {
14     return {1, 2};
15 }
16
17
18 int main()
19 {
20     Point p = Point();
21     Point __get_point16 = get_point();
22     int& a = __get_point16.x;
23     int& b = __get_point16.y;
24     std::operator<<(std::cout.operator<<(a), '\n');
25     std::operator<<(std::cout.operator<<(b), '\n');
26 }
27
28
```



# Example: Lambda Expressions

[Issues](#)[About](#)

Source:

```
1 #include <iostream>
2
3 int main() {
4     auto f = [](){std::cout << "Hello, World!\n";};
5     f();
6 }
7
```

Made by [Andreas Fertig](#)

Powered by [Flask](#) and [CodeMirror](#)

Insight:

```
1 #include <iostream>
2
3 int main()
4 {
5
6     class __lambda_4_14
7     {
8     public: inline /*constexpr */ void operator()() const
9     {
10         std::operator<<(std::cout, "Hello, World!\n");
11     }
12
13 };
14
15 __lambda_4_14 f = __lambda_4_14{};
16 f.operator()();
17 }
18
19
```

Console:

Insights exited with result code: 0

- 1 Matt Godbolt. Compiler Explorer — Part 1 of 2. March 9, 2018. Available online at [https://youtu.be/4\\_HL3PH4wDg](https://youtu.be/4_HL3PH4wDg).
- 2 Matt Godbolt. Compiler Explorer — Part 2 of 2. March 16, 2018. Available online at [https://youtu.be/1u\\_ku\\_OJPDg](https://youtu.be/1u_ku_OJPDg).

## Section 11.2

### Clang Format

# Clang Format

- Clang Format is tool for performing automatic formatting of source code for C++ (and other languages in C family)
- accessible via `clang-format` program
- command-line interface for `clang-format`:

```
clang-format [options] $file...
```

- some common options include:

Option	Description
<code>-help</code>	print usage information and exit
<code>-i</code>	perform formatting in place
<code>-style <i>style</i></code>	use coding style <i>style</i> (supported styles include: LLVM, Google, Chromium, Mozilla, WebKit)

- can be used as standalone tool or integrated with various editors (e.g., Vim or Emacs) or IDEs
- online documentation can be found at:

- <http://clang.llvm.org/docs/ClangFormat.html>

- 1 Daniel Jasper. clang-format: Automatic formatting for C++. European LLVM Conference, Paris, France, Apr. 29–30, 2013. Available online at [https://youtu.be/s7JmdCfI\\_\\_c](https://youtu.be/s7JmdCfI__c).
- 2 Daniel Jasper and Manuel Klimek. An update on Clang-based C++ Tooling. LLVM Developers' Meeting, San Jose, CA, USA, Oct. 30, 2015. Available online at <https://youtu.be/1S2A0VWGOWs>.

## Section 11.3

# Language Server Protocol (LSP) Clients and Servers

# Language Server Protocol (LSP)

- Language Server Protocol (LSP) is protocol used between editor or IDE and server that provides language features such as completion, identifier queries (e.g., go to definition and find all references), formatting, and identifier renaming
- LSP is language agnostic; servers available for many languages including C++ and C
- originally proposed by Microsoft
- allows language-aware component of software to be separated from editing component
- server has intimate knowledge of particular programming language
- client is typically editor or IDE
- base protocol consists of header and content part (similar to HTTP)
- content part of messages uses JSON-RPC to describe requests, responses, and notifications
- web site: <https://microsoft.github.io/language-server-protocol/specification>

- Clangd is LSP server developed as part of LLVM project
- supports C++ and C languages
- has basic LSP support (e.g., completion, diagnostics, fix-its, signature help, local rename, go to definition, and formatting)
- has support for global completion (e.g., completion still knows about `std::vector` even if `vector` header not included in translation unit)
- has support for project-wide index
- LSP server provided by `clangd` program
- `clangd` not normally run directly by user; instead, invoked indirectly by editor or IDE
- web site: <https://clang.llvm.org/extra/clangd.html>



- Cquery is highly-scalable low-latency LSP server for C, C++ and Objective C
- designed for large code bases
- fast but has large memory overhead for caching
- web site: <https://github.com/cquery-project/cquery>

- growing number of IDEs and editors providing support for LSP
- some editors and IDEs with LSP support include:
  - Microsoft Visual Studio IDE (<https://visualstudio.microsoft.com/vs>)
  - Eclipse IDE (<https://www.eclipse.org/ide>)
  - Vim (<https://www.vim.org>) and Neovim (<https://neovim.io>)
    - Vim LSP (<https://github.com/prabirshrestha/vim-lsp>)
    - LanguageClient-Neovim (<https://github.com/autozimu/LanguageClient-neovim>)
    - Vim-LSC (<https://github.com/natebosch/vim-lsc>)
    - Asynchronous Lint Engine (ALE) (<https://github.com/w0rp/ale>)
  - Emacs (<https://www.gnu.org/software/emacs>)
    - Emacs LSP Mode (<https://github.com/emacs-lsp/lsp-mode>)
    - Emacs Polyglot (<https://github.com/joaotavora/eglot>)

- 1 Ilya Biryukov. Global Code Completion and Architecture of clang. European LLVM Developers Meeting, Bristol, England, April 16–17, 2018. Available online at <https://youtu.be/Bvjrz3QioBI>.
- 2 Marc-Andre Laperle. Advancing Clangd: Bringing Persisted Indexing to Clang Tooling. LLVM Developers Meeting, San Jose, CA, USA, Oct. 18–19, 2017. Available online at <https://youtu.be/Y9JB3hlAWeA>.
- 3 Ilya Biryukov. Clangd: Architecture of a Scalable C++ Language Server. CppCon, Bellevue, WA, USA, Sept. 28, 2018. Available online at <https://youtu.be/5HIyAXj1YNQ>

## Section 11.4

### **YouCompleteMe (YCM)**

- YouCompleteMe (YCM) is code completion engine for Vim editor
- some features of YCM include:
  - semantic code completion
  - real-time diagnostic display for fast syntax checking
  - find declaration, definitions, usages of identifiers
  - display type information for classes, variables, functions
  - fix coding errors by applying fix-it hints
  - display documentation for identifiers
- supports C++, C, and other languages
- uses libclang to perform semantic analysis of code
- web site: <http://valloric.github.io/YouCompleteMe>

# YCM With Vim: Completion Example

```
#include <iostream>

int main() {
    std::cout << "Hello, World!\n";
    return !std::c
}
~      cin
~      cend( const _Container &__cont )
~      cerr
~      clog
~      copy( _II __first, _II __last, _OI __result )
~      cout
~      crend( const _Container &__cont )
~      ctype<typename _CharT>
~      calloc( size_t __nmemb, size_t __size )
~      cbegin( const _Container &__cont )
~      codecvt<typename _InternT, typename _ExternT, typename _StateT
~      collate<typename _CharT>
~      crbegin( const _Container &__cont )
~      clearerr( FILE *__stream )
~      conditional<bool, typename, typename>
~      ctype_base
~      char_traits<class _CharT>
~      common_type<typename _Tp>

-- INSERT --
```

# Specifying Compiler Commands

- to perform semantic analysis, YCM uses library interface to Clang compiler (libclang)
- libclang must be provided with set of compile flags in order to parse code
- YCM allows compile flags to be specified in two ways:
  - 1 via compilation database (i.e., `compile_commands.json` file)
  - 2 via (user-provided) Python script (i.e., `.ycm_extra_conf.py`)
- use of compilation database is preferred approach, since database can often be automatically generated by build tools
- YCM searches for file named `compile_commands.json`, starting in directory of opened file and then moving up directory hierarchy
- YCM also searches for file name `.ycm_extra_conf.py`, starting in directory of opened file and then moving up directory hierarchy
- full functionality of YCM requires that compile-flag information be specified

# Specifying Compiler Commands with CMake

- if using CMake, `compile_commands.json` file can be automatically generated (in build directory) by setting `CMAKE_EXPORT_COMPILE_COMMANDS` to `true` when configuring with CMake
- then, symlink (or copy) `compile_commands.json` in build directory to directory searched by YCM (if needed)
- can set `CMAKE_EXPORT_COMPILE_COMMANDS` when configuring with `cmake` command via command line option  
“`-DCMAKE_EXPORT_COMPILE_COMMANDS=ON`”
- alternatively, can set `CMAKE_EXPORT_COMPILE_COMMANDS` via line in `CMakeLists.txt` like:  

```
set(CMAKE_EXPORT_COMPILE_COMMANDS ON)
```



# YCM Commands

- `:YcmRestartServer`
  - restarts `ycmd` completion server (e.g., if server exits for some reason)
- `:YcmForceCompileAndDiagnostics`
  - forces YCM to immediately recompile file and display any new diagnostics encountered
- `:YcmDiags`
  - populates Vim's locationlist with errors or warnings if any were detected in file and then opens locationlist
- `:YcmShowDetailedDiagnostic`
  - shows full diagnostic text when user's cursor is on line with diagnostic
- `:YcmDebugInfo`
  - displays various debug information for current file
- `:YcmToggleLogs`
  - shows list of logfiles created by YCM, `ycmd` server, and semantic engine server for current filetype, if any
- `:YcmCompleter`
  - gives access to numerous additional IDE-like features like go to definition and apply fix-it

# Some YcmCompleter Subcommands

- `GoTo`
  - performs context-sensitive goto operation based on current cursor position (e.g., `GoToInclude`, `GoToDeclaration`, `GoToDefinition`, etc.)
- `GoToInclude`
  - looks for header file on current line and jumps to it
- `GoToDeclaration`
  - looks up identifier at current cursor position and jumps to its declaration
- `GoToDefinition`
  - looks up identifier at current cursor position and jumps to its definition
- `GoToType`
  - looks up identifier at current cursor position and jumps to definition of its type
- `GetType`
  - displays type of identifier at current cursor position
- `FixIt`
  - attempts to make changes to buffer to correct diagnostics on current line

# Vim-Related Comments

- enable file type detection in `.vimrc` file with `“:filetype on”`
- typically choose to map commonly used YCM commands to Vim keyboard shortcuts
- to map `“:YcmForceCompileAndDiagnostics”` to F5 key, add line like following to `.vimrc` file:

```
nnoremap <F5> :YcmForceCompileAndDiagnostics<CR>
```
- to map `“:YcmCompleter GoTo”` to key sequence `<leader>jd`, add line like following to `.vimrc` file:

```
nnoremap <leader>jd :YcmCompleter GoTo<CR>
```
- leader key (i.e., `<leader>`) can be queried with `“:let mapleader”` (often set to comma on Linux systems)
- leader key can be set to comma with `“:let mapleader=“, ””`
- for more information on Vim, refer to Vim web site:
  - <http://www.vim.org>

## Part 12

### Miscellany

## Section 12.1

### Miscellany

# The abs Function

- Consider a program with the following source listing:

```
#include <iostream>
#include <cstdlib>
int main() {std::cout << abs(-1.5) << '\n';}
```

- The C++ implementation is permitted (but not required) to place the `C abs` function in the global namespace.
- If the implementation does not do this, the above program will fail to compile (avoiding the more troubling problem discussed next).
- If, however, the C++ implementation does do this (which is not uncommon in practice), the above program will compile successfully, but behave unexpectedly when run.
- In particular, the program will output a value of 1, instead of the value of 1.5 that was likely expected by the programmer.
- Since the `C abs` function is declared as `int abs(int)`, the use of this function will introduce a conversion from `double` to `int`, leading to the unexpected result.

# The abs Function (Continued)

- The problems of the previous slide can be easily avoided as follows.
- First, include the header `cmath`, which provides overloads of `std::abs` for various built-in types, including **double**.
- Then, invoke the function `std::abs` (instead of `::abs`).
- For example, the following code will behave as expected, outputting the value of 1.5:

```
#include <iostream>
#include <cmath>
int main() {std::cout << std::abs(-1.5) << '\n';}
```

# Sizeof Class Versus Sum of Member Sizes

- compilers can (and do) add padding to classes/structs

- Example:

```
1  #include <iostream>
2
3  class Widget {
4  // ...
5  private:
6      char c;
7      int i;
8  };
9
10 int main() {
11     // two numbers printed not necessarily the same
12     std::cout << sizeof(char) + sizeof(int) << ' ' <<
13         sizeof(Widget) << '\n';
14     std::cout << alignof(int) << ' ' <<
15         alignof(Widget) << '\n';
16 }
```

- many processors place alignment restrictions on data (e.g., data type of size  $n$  must be aligned to start on address that is multiple of  $n$ )
- other factors can also add to size of class/struct (e.g., virtual function table pointer)

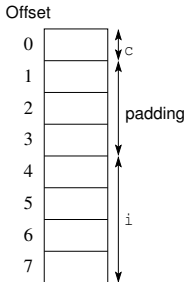


# Sizeof Class Versus Sum of Member Sizes (Continued)

- consider following type:

```
struct Widget {  
    char c;  
    int i;  
};
```

- suppose that **sizeof(int)** is 4 and **alignof(int)** is 4
- compiler adds padding to structure so that **int** data member is suitably aligned (i.e., offset is multiple of 4)
- memory layout for `Widget` object:



- Returning an rvalue reference to an rvalue reference parameter can potentially lead to very subtle bugs.
- Example:

```
std::string&& join(std::string&& s, const char* p) {  
    return std::move(s.append(" ").append(p));  
}  
  
std::string getMessage() {return "Hello";}  
  
void func() {  
    const string& r = join(getMessage(), " World");  
    // lifetime of temporary returned by getMessage  
    // not extended to lifetime of r since not  
    // directly bound to r  
    // r now refers to destroyed temporary  
}
```

- Fix:

```
std::string join(std::string&& s, const char* p) {  
    return std::move(s.append(" ").append(p));  
}
```

- Returning by rvalue reference should probably be avoided, except in very special circumstances (such as `std::forward` and `std::move`).

# Be Careful When Mixing Signed and Unsigned Types

```
1  #include <cassert>
2
3  int main() {
4      short ss = -1;
5      int si = -1;
6      long sl = -1;
7      long long sll = -1;
8      unsigned short us = 0;
9      unsigned int ui = 0;
10     unsigned long ul = 0;
11     unsigned long long ull = 0;
12     // comparison between signed and unsigned types
13     assert(ss < ui); // FAILS: ss becomes UINT_MAX
14     // comparison between signed and unsigned types
15     assert(si < ui); // FAILS: si becomes UINT_MAX
16     // comparison between signed and unsigned types
17     assert(sl < ul); // FAILS: sl becomes ULONG_MAX
18     // comparison between signed and unsigned types
19     assert(sll < ull); // FAILS: sll becomes ULLONG_MAX
20 }
```

- be aware of rules for promotions and conversions involving integral types
- if these rules not considered, code may not behave in manner expected

## Section 12.2

### Exercises

## [Q.1] What Is Wrong With This Code?

main.cpp

```
1  #include <complex>
2  #include <iostream>
3  #include <vector>
4
5  std::complex<long double>
6  square(std::complex<long double>& x) {
7      return x * x;
8  }
9
10 void do_stuff(std::vector<int>& v) {
11     int* ip = v.data();
12     std::cout << ip << '\n';
13     for (std::vector<int>::iterator i = v.begin();
14         i != v.end(); ++i) {
15         std::cout << (i - v.begin()) << ' ' << *i << '\n';
16     }
17 }
18
19 int main() {
20     std::vector<int> v{1, 2, 4, 8};
21     do_stuff(v);
22     std::complex<long double> fortytwo{42};
23     std::cout << square(fortytwo) << '\n';
24 }
```

## [Q.1] Solution: Use Const Qualifier Correctly

main.cpp

```
1  #include <complex>
2  #include <iostream>
3  #include <vector>
4
5  std::complex<long double>
6  square(const std::complex<long double>& x) {
7      return x * x;
8  }
9
10 void do_stuff(const std::vector<int>& v) {
11     const int* ip = v.data();
12     std::cout << ip << '\n';
13     for (std::vector<int>::const_iterator i = v.begin();
14          i != v.end(); ++i) {
15         std::cout << (i - v.begin()) << ' ' << *i << '\n';
16     }
17 }
18
19 int main() {
20     const std::vector<int> v{1, 2, 4, 8};
21     do_stuff(v);
22     const std::complex<long double> fortytwo{42};
23     std::cout << square(fortytwo) << '\n';
24 }
```

## [Q.2] What Is Wrong With This Code?

main.cpp

```
1  #include <iostream>
2
3  class Counter {
4  public:
5      Counter(unsigned c = 0) : c_(c) {}
6      Counter(Counter& other) : c_(other.c_) {}
7      Counter(const Counter&& other) : c_(other.c_) {}
8      ~Counter() {}
9      Counter& operator=(Counter& other)
10         {c_ = other.c_; return *this;}
11      Counter& operator=(const Counter&& other)
12         {c_ = other.c_; return *this;}
13      void clear() {c_ = 0;}
14      const unsigned get_count() {return c_;}
15      const Counter& operator++() {++c_; return *this;}
16      operator bool() {return c_ != 0;}
17 private:
18     unsigned c_; // counter value
19 };
20
21 int main() {
22     Counter zero;
23     Counter c{zero};
24     ++c;
25     std::cout << c.get_count() << '\n';
26     std::cout << zero.get_count() << '\n';
27 }
```

## [Q.2] Solution: Use Const Qualifier Correctly

main.cpp

```
1  #include <iostream>
2
3  class Counter {
4  public:
5      Counter(unsigned c = 0) : c_(c) {}
6      Counter(const Counter& other) : c_(other.c_) {}
7      Counter(Counter&& other) : c_(other.c_) {}
8      ~Counter() {}
9      Counter& operator=(const Counter& other)
10         {c_ = other.c_; return *this;}
11      Counter& operator=(Counter&& other)
12         {c_ = other.c_; return *this;}
13      void clear() {c_ = 0;}
14      unsigned get_count() const {return c_;}
15      Counter& operator++() {++c_; return *this;}
16      operator bool() const {return c_ != 0;}
17 private:
18     unsigned c_; // counter value
19 };
20
21 int main() {
22     const Counter zero;
23     Counter c{zero};
24     ++c;
25     std::cout << c.get_count() << '\n';
26     std::cout << zero.get_count() << '\n';
27 }
```



## [Q.3] What Is Wrong With This Code?

foo.hpp

```
1  #ifndef foo_hpp
2  #define foo_hpp
3  namespace foo {
4  bool is_odd(int x) {return (x % 2) != 0;}
5  bool is_even(int x) {return (x % 2) == 0;}
6  }
7  #endif
```

main.cpp

```
1  #include <iostream>
2  #include "foo.hpp"
3  int main() {
4      std::cout << foo::is_odd(42) << ' ' <<
5          foo::is_even(42) << '\n';
6  }
```

other.cpp

```
1  #include "foo.hpp"
2  // ...
```

## [Q.3] Solution: Functions Should Be Inline

foo.hpp

```
1  #ifndef foo_hpp
2  #define foo_hpp
3  namespace foo {
4  inline bool is_odd(int x) {return (x % 2) != 0;}
5  inline bool is_even(int x) {return (x % 2) == 0;}
6  }
7  #endif
```

main.cpp

```
1  #include <iostream>
2  #include "foo.hpp"
3  int main() {
4      std::cout << foo::is_odd(42) << ' ' <<
5          foo::is_even(42) << '\n';
6  }
```

other.cpp

```
1  #include "foo.hpp"
2  // ...
```

## [Q.4] What Is Wrong With This Code?

foo.hpp

```
1  #ifndef foo_hpp
2  #define foo_hpp
3  namespace foo {
4  inline bool is_odd(int x);
5  inline bool is_even(int x);
6  }
7  #endif
```

foo.cpp

```
1  #include "foo.hpp"
2  namespace foo {
3  bool is_odd(int x) {return (x % 2) != 0;}
4  bool is_even(int x) {return (x % 2) == 0;}
5  }
```

app.cpp

```
1  #include <iostream>
2  #include "foo.hpp"
3  int main() {
4      std::cout << foo::is_odd(42) << ' ' <<
5      foo::is_even(42) << '\n';
6  }
```

## [Q.4] Solution: Place Inline Function Definitions in Header File

foo.hpp

```
1  #ifndef foo_hpp
2  #define foo_hpp
3  namespace foo {
4  inline bool is_odd(int x) {return (x % 2) != 0;}
5  inline bool is_even(int x) {return (x % 2) == 0;}
6  }
7  #endif
```

app.cpp

```
1  #include <iostream>
2  #include "foo.hpp"
3  int main() {
4      std::cout << foo::is_odd(42) << ' ' <<
5          foo::is_even(42) << '\n';
6  }
```

## [Q.5] What Is Wrong With This Code?

foo.hpp

```
1  #ifndef foo_hpp
2  #define foo_hpp
3  namespace foo {
4  template <typename T> T abs(const T& x);
5  }
6  #endif
```

foo.cpp

```
1  #include "foo.hpp"
2  namespace foo {
3  template <typename T> T abs(const T& x)
4  { return (x < 0) ? (-x) : x; }
5  }
```

app.cpp

```
1  #include <iostream>
2  #include "foo.hpp"
3  int main() {
4      std::cout << foo::abs(-42) << ' ' <<
5      foo::abs(-3.14) << '\n';
6  }
```

## [Q.5] Solution 1: Explicit Template Instantiation

foo.hpp

```
1  #ifndef foo_hpp
2  #define foo_hpp
3  namespace foo {
4  template <typename T> T abs(const T& x);
5  }
6  #endif
```

foo.cpp

```
1  #include "foo.hpp"
2  namespace foo {
3  template <typename T> T abs(const T& x)
4  { return (x < 0) ? (-x) : x; }
5  template int abs<int>(const int&);
6  template double abs<double>(const double&);
7  }
```

app.cpp

```
1  #include <iostream>
2  #include "foo.hpp"
3  int main() {
4      std::cout << foo::abs(-42) << ' ' <<
5          foo::abs(-3.14) << '\n';
6  }
```

## [Q.5] Solution 2: Define Function Template in Header File

foo.hpp

```
1  #ifndef foo_hpp
2  #define foo_hpp
3  namespace foo {
4  template <typename T> T abs(const T& x)
5      {return (x < 0) ? (-x) : x;}
6  }
7  #endif
```

app.cpp

```
1  #include <iostream>
2  #include "foo.hpp"
3  int main() {
4      std::cout << foo::abs(-42) << ' ' <<
5          foo::abs(-3.14) << '\n';
6  }
```

# Remarks on Headers Files and Function Declarations

- Every function (whether it be inline or non-inline, or template or non-template) must be *declared* before being used.
- Consequently, functions that are part of an interface should normally be declared in a *header file* so that users of the interface can obtain the declarations needed for the interface by simply including the header file.
- An inline function should always be *defined* before being used.
- Consequently, an inline function that is declared in a header file should normally also be *defined* in the file.
- A template function must be *defined* at its point of use in order for the template to be implicitly instantiated.
- Consequently, a template function that is declared in a header file should normally also be *defined* in the file.
- A function must not be defined more than once.
- Consequently, unless a function is inline or a template, it should not be defined in a header file, as this will result in multiple definitions if the header file is included by more than one source file.



## [Q.6] What Is Wrong With This Code?

foo.hpp

```
1  #ifndef foo_hpp
2  #define foo_hpp
3  #include <cmath>
4  namespace foo {
5  double log(double x, double b);
6  }
7  #endif
```

foo.cpp

```
1  #include <cmath>
2  #include "foo.hpp"
3  namespace foo {
4  double log(double x, double b = 10.0)
5      {return std::log(x) / std::log(b);}
6  }
```

app.cpp

```
1  #include <iostream>
2  #include "foo.hpp"
3  int main() {
4      std::cout << foo::log(16.0, 2.0) << ' ' <<
5          foo::log(10.0) << '\n';
6  }
```

## [Q.6] Solution: Place Default Arguments in Header File

foo.hpp

```
1 #ifndef foo_hpp
2 #define foo_hpp
3 #include <cmath>
4 namespace foo {
5 double log(double x, double b = 10.0);
6 }
7 #endif
```

foo.cpp

```
1 #include <cmath>
2 #include "foo.hpp"
3 namespace foo {
4 double log(double x, double b)
5 {return std::log(x) / std::log(b);}
6 }
```

app.cpp

```
1 #include <iostream>
2 #include "foo.hpp"
3 int main() {
4     std::cout << foo::log(16.0, 2.0) << ' ' <<
5     foo::log(10.0) << '\n';
6 }
```

## [Q.7] What Is Wrong With This Code?

foo.hpp

```
1 namespace foo {  
2 constexpr int abs(int x);  
3 }
```

foo.cpp

```
1 #include "foo.hpp"  
2  
3 namespace foo {  
4 constexpr int abs(int x) {return (x < 0) ? (-x) : x;}  
5 }
```

app.cpp

```
1 #include <iostream>  
2 #include "foo.hpp"  
3  
4 int main() {  
5     constexpr auto a = foo::abs(-42);  
6     std::cout << a << '\n';  
7 }
```

## [Q.7] Solution: Define Constexpr Function in Header

foo.hpp

```
1 namespace foo {  
2 constexpr int abs(int x) {return (x < 0) ? (-x) : x;}  
3 }
```

app.cpp

```
1 #include <iostream>  
2 #include "foo.hpp"  
3  
4 int main() {  
5     constexpr auto a = foo::abs(-42);  
6     std::cout << a << '\n';  
7 }
```

## [Q.8] What Is Wrong With This Code?

```
1  #include <cmath>
2  #include <cassert>
3
4  constexpr double func(double x) {
5      assert(x >= 0.0);
6      return std::sqrt(x) + std::sin(x) * std::cos(x);
7  }
```

## [Q.8] Answer: Invalid Constexpr Function

```
1  #include <cmath>
2  #include <cassert>
3
4  constexpr double func(double x) {
5      assert(x >= 0.0);
6      return std::sqrt(x) + std::sin(x) * std::cos(x);
7  }
```

- `std::sqrt`, `std::sin`, and `std::cos` are not constexpr functions (at least, in fully compliant C++17 implementation) [\[C++17 20.5.5.6/1\]](#)
- consequently, `func` is invalid constexpr function, since no arguments exist for which function can be evaluated as constant expression
- invocation of `assert` is okay, since `assert` can be used in constant expression

## [Q.9] What Is Wrong With This Code?

```
1  #include <iostream>
2
3  constexpr unsigned int factorial(unsigned int x) {
4      unsigned int result;
5      result = 1;
6      for (unsigned int i = 2; i <= x; ++i) {
7          result *= i;
8      }
9      return result;
10 }
11
12 int main() {
13     constexpr auto x = factorial(4);
14     std::cout << x << '\n';
15 }
```

## [Q.9] Solution: Initialize Constexpr Function Variables

```
1  #include <iostream>
2
3  constexpr unsigned int factorial(unsigned int x) {
4      unsigned int result = 1;
5      for (unsigned int i = 2; i <= x; ++i) {
6          result *= i;
7      }
8      return result;
9  }
10
11 int main() {
12     constexpr auto x = factorial(4);
13     std::cout << x << '\n';
14 }
```



## [Q.10] What Is Wrong With This Code?

```
1  #include <vector>
2  #include <iostream>
3
4  constexpr std::vector<float> get_value(float a) {
5      return {a * 1.0f, a * 2.0f, a * 3.0f};
6  }
7
8  int main() {
9      constexpr auto v = get_value(2.0);
10     for (auto x : v) {
11         std::cout << x << '\n';
12     }
13 }
```

## [Q.10] Solution: constexpr Requires Literal Types

```
1  #include <array>
2  #include <iostream>
3
4  constexpr std::array<float, 3> get_value(float a) {
5      return {a * 1.0f, a * 2.0f, a * 3.0f};
6  }
7
8  int main() {
9      constexpr auto v = get_value(2.0);
10     for (auto x : v) {
11         std::cout << x << '\n';
12     }
13 }
```

## [Q.11] What Is Wrong With This Code?

```
1  #include <algorithm>
2  #include <cstdint>
3
4  template <class T>
5  void in_place_transpose(T* a, std::size_t m, std::size_t n) {
6      T tmp[m][n];
7      std::copy_n(a, m * n, &tmp[0][0]);
8      for (std::size_t i = 0; i < m; ++i) {
9          for (std::size_t j = 0; j < n; ++j) {
10             a[j * m + i] = tmp[i][j];
11         }
12     }
13 }
```

- ignore possibility of integer overflow
- ignore potential cache-efficiency issues

## [Q.11] Answer: Code Has At Least Two Serious Problems

```
1  #include <algorithm>
2  #include <cstdlib>
3
4  template <class T>
5  void in_place_transpose(T* a, std::size_t m, std::size_t n) {
6      T tmp[m][n];
7      std::copy_n(a, m * n, &tmp[0][0]);
8      for (std::size_t i = 0; i < m; ++i) {
9          for (std::size_t j = 0; j < n; ++j) {
10             a[j * m + i] = tmp[i][j];
11         }
12     }
13 }
```

- code uses variable length array (VLA) (i.e., `tmp`)
- VLAs are forbidden in (standard-compliant) C++ (i.e., array sizes must be constant expressions)
- even with compilers that allow VLAs in violation of C++ standard (such as GCC), object of unknown (and potentially very large) size being created on stack
- can easily result in stack overflow

## [Q.11] Solution: Avoid VLAs and Creating Large Stack Objects

```
1  #include <algorithm>
2  #include <vector>
3  #include <cstdlib>
4
5  template <class T>
6  void in_place_transpose(T* a, std::size_t m, std::size_t n) {
7      std::vector<T> tmp(a, a + m * n);
8      for (std::size_t i = 0; i < m; ++i) {
9          for (std::size_t j = 0; j < n; ++j) {
10             a[j * m + i] = tmp[i * n + j];
11         }
12     }
13 }
```

- use of `std::vector` for temporary array avoids need for VLA
- since `std::vector` allocates its underlying array storage from heap, potential stack overflow problem avoided

## [Q.FCC1] How to Invoke `func` at Compile Time?

```
1  int func() {  
2      int i = 0;  
3      // ... (valid code for constexpr function)  
4      return i;  
5  }  
6  
7  int main() {  
8      return func();  
9  }
```

```
1  constexpr int func() {  
2      int i = 0;  
3      // ... (valid code for constexpr function)  
4      return i;  
5  }  
6  
7  int main() {  
8      constexpr int i = func();  
9      return i;  
10 }
```

## [Q.FCC2] How to Invoke `func` at Compile Time?

```
1  void func() {  
2      // ... (valid code for constexpr function)  
3  }  
4  
5  int main() {  
6      func();  
7  }
```



```
1  constexpr void func() {  
2      // ... (valid code for constexpr function)  
3  }  
4  
5  static_assert((func(), true));  
6  
7  int main() {  
8  }
```

### ■ memory and cache parameters:

- maximum memory size is 1024 bytes
- cache size is 64 bytes
- 4-byte cache block size
- 4-way set associative cache

### ■ determine:

#### 1 number of bits in address

- $\log_2 1024 = 10$  (i.e., log base 2 of number of addresses)

#### 2 number of blocks in memory

- $1024/4 = 256$  (i.e., memory size divided by block size)

#### 3 number of blocks in cache

- $64/4 = 16$  (i.e., cache size divided by block size)

#### 4 number of sets in cache

- $16/4 = 4$  (i.e., number of cache blocks divided by set associativity)

#### 5 number of bits in block offset

- $\log_2 4 = 2$  (i.e., log base 2 of cache block size)

#### 6 number of bits in cache index

- $\log_2 4 = 2$  (i.e., log base 2 of number of sets)

#### 7 number of bits in cache tag

- $10 - (2 + 2) = 6$  (i.e., number of address bits minus number of cache index and block offset bits)

- same memory and cache parameters as on previous slide
- valid entries in cache:

Tag	Index	Data			
		Byte 0	Byte 1	Byte 2	Byte 3
001010 <sub>2</sub>	10 <sub>2</sub>	11 <sub>16</sub>	12 <sub>16</sub>	13 <sub>16</sub>	14 <sub>16</sub>
001010 <sub>2</sub>	11 <sub>2</sub>	21 <sub>16</sub>	22 <sub>16</sub>	23 <sub>16</sub>	24 <sub>16</sub>
001001 <sub>2</sub>	00 <sub>2</sub>	31 <sub>16</sub>	32 <sub>16</sub>	33 <sub>16</sub>	34 <sub>16</sub>
001111 <sub>2</sub>	01 <sub>2</sub>	41 <sub>16</sub>	42 <sub>16</sub>	43 <sub>16</sub>	44 <sub>16</sub>

- does memory access to each of following addresses result in cache hit or miss, and if hit what value for address in cache?
  - 1 0010100011<sub>2</sub>
    - tag 001010<sub>2</sub>, index 00<sub>2</sub>, block offset 11<sub>2</sub>; cache miss
  - 2 0010101101<sub>2</sub>
    - tag 001010<sub>2</sub>, index 11<sub>2</sub>, block offset 01<sub>2</sub>; cache hit; data 22<sub>16</sub>

- virtual and physical memory parameters:

- 16-bit virtual address
- 12-bit physical address
- page size is 256 bytes

- determine:

- 1 number of virtual pages

- $2^{16}/2^8 = 2^8 = 256$  (i.e., number of virtual addresses divided by page size)

- 2 number of physical pages

- $2^{12}/2^8 = 2^4 = 16$  (i.e., number of physical addresses divided by page size)

- 3 number of bits in page offset (PO)

- $\log_2 256 = 8$  (i.e., log base 2 of page size)

- 4 number of bits in virtual page number (VPN)

- $16 - 8 = 8$  (i.e., virtual address size in bits minus number of bits in PO)

- 5 number of bits in physical page number (PPN)

- $12 - 8 = 4$  (i.e., physical address size in bits minus number of bits in PO)

## [Q.C2] Virtual Memory Exercise (Continued)

- same virtual memory parameters as on previous slide
- page table contains following address translation information:

VPN	PPN	Flags
00000000 <sub>2</sub>	1111 <sub>2</sub>	present, readable, writable, not executable
00000001 <sub>2</sub>	1010 <sub>2</sub>	present, readable, writable, not executable
10000000 <sub>2</sub>	0000 <sub>2</sub>	present, readable, not writable, executable

- determine result of translation and protection check for accesses to following addresses:
  - 1 data read 1111111100000000<sub>2</sub>
    - VPN 11111111<sub>2</sub>, PO 00000000<sub>2</sub> access violation (no address mapping)
  - 2 instruction fetch 0000000000000000<sub>2</sub>
    - VPN 00000000<sub>2</sub>, PO 00000000<sub>2</sub>; access violation (physical address 111100000000<sub>2</sub>)
  - 3 data write 0000000100000000<sub>2</sub>
    - VPN 00000001<sub>2</sub>, PO 00000000<sub>2</sub>; physical address 101000000000<sub>2</sub>
- could virtual address xxxxxxxx00000000<sub>2</sub> map to physical address yyyy00001111<sub>2</sub>?
  - no (page offsets differ)

## [Q.A1] What Is Wrong With This Code?

```
1  #include <cstdint>
2
3  // write elements of source array src to destination
4  // array dst in reverse order
5  template <class T, std::size_t N>
6  void reverse(const T (&src)[N], T (&dst)[N]) {
7      for (std::size_t i = 0; i < N; ++i) {
8          dst[N - 1 - i] = src[i];
9      }
10 }
```

## [Q.A1] Answer: Aliasing Case Not Handled Correctly

```
1  #include <iostream>
2
3  // include definition of reverse
4  #include "aliasing_1_1.hpp"
5
6  int main() {
7      int a[4] = {1, 2, 3, 4};
8      reverse(a, a);
9      for (auto&& x : a) {
10         std::cout << x << '\n';
11     }
12 }
```

- reverse will not work correctly if `src` and `dst` parameters refer to same array
- consider above usage of `reverse`, which is problematic

```
1  #include <cstddef>
2
3  template <class T, std::size_t N>
4  void reverse(const T (&src)[N], T (&dst)[N]) {
5      // Check for aliasing case.
6      if (src == dst) {
7          for (std::size_t i = 0; i < N / 2; ++i) {
8              using std::swap;
9              swap(dst[i], dst[N - 1 - i]);
10         }
11     } else {
12         for (std::size_t i = 0; i < N; ++i) {
13             dst[N - 1 - i] = src[i];
14         }
15     }
16 }
```



## [Q.A2] What Is Wrong With This Code?

```
1  // Append the string pointed to by src to the string
2  // pointed to by dst.
3  // Suppose that no restrictions are placed on src and dst.
4  // What is wrong with this code?
5  char* string_concat(char* dst, const char* src) {
6      while (*dst != '\0') {
7          ++dst;
8      }
9      while (*src != '\0') {
10         *dst++ = *src++;
11     }
12     *dst = '\0';
13 }
```

```
1  #include <iostream>
2
3  char* string_concat(char *, const char*);
4
5  int main() {
6      char a[1024] = "bye";
7      string_concat(a, a);
8      std::cout << a << '\n';
9  }
```

- `string_concat` will not work correctly if source and destination strings overlap
- consider above code, which will not work correctly (and probably result in segmentation fault)
- best solution is to change interface so that overlap of source and destination forbidden (since detecting aliasing cannot be done efficiently)

## [Q.A3] What Is Wrong With This Code?

```
1  class Complex {
2  public:
3      Complex(double r, double i) : r_(r), i_(i) {}
4      double real() const {return r_;}
5      double imag() const {return i_;}
6      // What is wrong with the following function?
7      Complex& operator*=(const Complex& other) {
8          auto r = r_;
9          auto i = i_;
10         r_ = r * other.r_ - i * other.i_;
11         i_ = r * other.i_ + i * other.r_;
12         return *this;
13     }
14     // ...
15 private:
16     double r_; // real part
17     double i_; // imaginary part
18 };
```

## [Q.A3] Answer: Does Not Correctly Handle Aliasing

```
1  #include <iostream>
2  #include "aliasing_3_1.hpp"
3
4  int main() {
5      Complex a(1.0, 1.0);
6      Complex b(a);
7      b *= a;
8      a *= a;
9      // The value of b is correct.
10     std::cout << b.real() << ' ' << b.imag() << '\n';
11     // The value of a is not correct.
12     std::cout << a.real() << ' ' << a.imag() << '\n';
13 }
```

- `Complex::operator*=` will not correctly handle case that `*this` and other are aliased
- above code will not work correctly

```
1  class Complex {
2  public:
3      Complex(double r, double i) : r_(r), i_(i) {}
4      double real() const {return r_;}
5      double imag() const {return i_;}
6      Complex& operator*=(const Complex& other) {
7          auto r = r_;
8          auto i = i_;
9          auto other_r = other.r_;
10         auto other_i = other.i_;
11         r_ = r * other_r - i * other_i;
12         i_ = r * other_i + i * other_r;
13         return *this;
14     }
15     // ...
16 private:
17     double r_; // real part
18     double i_; // imaginary part
19 };
```

- consider simple triangle classification program
- program performs following:
  - 1 read (from standard input stream) three integers  $a$ ,  $b$ , and  $c$  that represent lengths of sides of triangle
  - 2 classify triangle with sides of lengths  $a$ ,  $b$ , and  $c$  as one of:
    - 1 equilateral (i.e., all three sides equal in length)
    - 2 isosceles (i.e., exactly two sides equal in length)
    - 3 scalene (i.e., all sides have distinct lengths)
  - 3 write (to standard output stream) triangle classification
- what set of test cases would be effective for testing above program?

- problem somewhat ill-posed in sense that following important information not stated:
  - how critical is correct program function
  - what assumptions can be made about input data
- as starting point, might consider test cases that cover:
  - equilateral case (e.g., (2,2,2))
  - isosceles case, considering all permutations of edge lengths (e.g., (1,1,2), (1,2,1), (2,1,1))
  - scalene case, considering all possible permutations of edge lengths (e.g., (3,4,5), (3,5,4), (4,3,5), (4,5,3), (5,3,4), (5,4,3))
  - incorrectly formatted data (e.g., “1 2 apple”)
  - insufficient data (i.e., less than three integers provided)
  - superfluous data (i.e., more than three integers provided)
  - lengths correspond to invalid triangle (e.g.: lengths not all strictly positive; or sum of any two lengths less than or equal to remaining length)
  - lengths chosen in attempt to induce overflow (i.e., very large lengths)
- if mission critical, we might perform more exhaustive testing

```
1  #include <iostream>
2  #include <string>
3  #include <utility>
4
5  std::string&& func1(std::string& x) {
6      return std::move(x);
7      // x? std::move(x)?
8  }
9
10 int main() {
11     const std::string hello("Hello");
12     std::string a;
13     std::string b;
14
15     a = hello + "!";
16     // hello? hello + "!"? a = hello + "!"?
17     std::cout << a << '\n';
18     // std::cout? std::cout << a?
19
20     a = std::string("");
21     // std::string("")? a = std::string("")?
22     ((a += hello) += "!");
23     // a += hello?
24     b = func1(a);
25     // func1(a)? b = func1(a)?
26     std::cout << b << '\n';
27 }
```



## [Q.11] What Is the Behavior of This Code?

```
1  class Widget {  
2  public:  
3      Widget() {}  
4      int i() const {return i_;}  
5  private:  
6      int i_;  
7  };  
8  
9  int main() {  
10     Widget w;  
11     return w.i();  
12 }
```

## [Q.12] What Is the Behavior of This Code?

```
1  class Widget {
2  public:
3      Widget() : i_() {}
4      int i() const {return i_;}
5  private:
6      int i_;
7  };
8
9  int main() {
10     Widget w;
11     return w.i();
12 }
```

## [Q.13] What Is the Behavior of This Code?

```
1  class Widget {  
2  public:  
3      Widget() {}  
4      int i() const {return i_;}  
5  private:  
6      int i_{};  
7  };  
8  
9  int main() {  
10     Widget w;  
11     return w.i();  
12 }
```

## [Q.14] What Is the Behavior of This Code?

```
1  #include <iostream>
2
3  // aggregate
4  struct Widget {
5      int i = 0;
6      int j = 0;
7  };
8
9  std::ostream& operator<<(std::ostream& out, const Widget& w) {
10     return out << w.i << ',' << w.j;
11 }
12
13 int main() {
14     Widget a;
15     Widget b{};
16     Widget c{1};
17     Widget d{1, 2};
18     std::cout << a << ' ' << b << ' ' << c << ' ' << d << '\n';
19 }
```

# [Q.MC1] Copy, Move, or Copy Elision?

```
1  #include <utility>
2
3  class Gadget {
4  public:
5      Gadget();
6      Gadget(const Gadget&);
7      Gadget(Gadget&&);
8      Gadget& operator=(const Gadget&);
9      Gadget& operator=(Gadget&&);
10     // ...
11 };
12
13 Gadget func_1() {return Gadget();}
14 Gadget func_2() {Gadget g; return g;}
15 void func_3(Gadget g) {/* ... */}
16
17 int main() {
18     Gadget s = Gadget();
19     Gadget t(std::move(s));
20     Gadget u(func_1());
21     Gadget v(func_2());
22     Gadget w(v);
23     s = std::move(t);
24     t = s;
25     s = func_1();
26     s = func_2();
27     func_3(s);
28     func_3(std::move(s));
29     func_3(func_1());
30 }
```

```
1  #include <utility>
2
3  class Gadget {
4  public:
5      Gadget();
6      Gadget(const Gadget&);
7      Gadget(Gadget&&);
8      Gadget& operator=(const Gadget&);
9      Gadget& operator=(Gadget&&);
10     // ...
11 };
12
13 Gadget func_1() {return Gadget();}
14 Gadget func_2() {Gadget g; return g;}
15 void func_3(Gadget g) {/* ... */}
16
17 int main() {
18     Gadget s = Gadget(); // default ctor, elided copy
19     Gadget t(std::move(s)); // move ctor
20     Gadget u(func_1()); // default ctor, elided copy
21     Gadget v(func_2()); // default ctor, maybe move ctor
22     Gadget w(v); // copy ctor
23     s = std::move(t); // move assign
24     t = s; // copy assign
25     s = func_1(); // default ctor, elided copy, move assign
26     s = func_2(); // default ctor, maybe move ctor, move assign
27     func_3(s); // copy ctor
28     func_3(std::move(s)); // move ctor
29     func_3(func_1()); // default ctor, elided copy
30 }
```

## [Q.MC2] Copy, Move, or Copy Elision?

```
1  #include <utility>
2
3  class Gadget {
4  public:
5      Gadget();
6      Gadget(const Gadget&);
7      Gadget(Gadget&&);
8      Gadget& operator=(const Gadget&);
9      Gadget& operator=(Gadget&&);
10     // ...
11 };
12
13 Gadget func_1() {return std::move(Gadget());} // BAD IDEA
14 Gadget func_2() {Gadget g; return std::move(g);} // BAD IDEA
15 const Gadget func_3() {return Gadget();} // BAD IDEA
16 const Gadget func_4() {Gadget g; return g;} // BAD IDEA
17
18 int main() {
19     Gadget s(func_1());
20     Gadget u(func_2());
21     Gadget t(func_3());
22     Gadget w(func_4());
23     s = func_1();
24     s = func_2();
25     s = func_3();
26     s = func_4();
27 }
```

```
1  #include <utility>
2
3  class Gadget {
4  public:
5      Gadget();
6      Gadget(const Gadget&);
7      Gadget(Gadget&&);
8      Gadget& operator=(const Gadget&);
9      Gadget& operator=(Gadget&&);
10     // ...
11 };
12
13 Gadget func_1() {return std::move(Gadget());} // BAD IDEA
14 Gadget func_2() {Gadget g; return std::move(g);} // BAD IDEA
15 const Gadget func_3() {return Gadget();} // BAD IDEA
16 const Gadget func_4() {Gadget g; return g;} // BAD IDEA
17
18 int main() {
19     Gadget s(func_1()); // default ctor, move ctor
20     Gadget u(func_2()); // default ctor, move ctor
21     Gadget t(func_3()); // default ctor, copy elided
22     Gadget w(func_4()); // default ctor, maybe move ctor
23     s = func_1(); // default ctor, move ctor, move assign
24     s = func_2(); // default ctor, move ctor, move assign
25     s = func_3(); // default ctor, copy elided, copy assign
26     s = func_4(); // default ctor, maybe move ctor, copy assign
27 }
```



## ■ What is wrong with the following code?

```
void func(const std::string&);  
std::string s("one");  
const char* p = "two";  
func(std::string(s) + std::string(", ") + std::string(p));  
func(std::string(p) + std::string(", ") + std::string(s));
```

## ■ *Unnecessary temporaries!*

### ■ Fix:

```
func(s + ", " + p);  
func(p + ", " + s);
```

## ■ What is wrong with the following code?

```
std::vector<std::string> v;  
std::string s("one");  
v.push_back(std::string(s));  
v.push_back(std::string(s + ", two"));  
v.push_back(std::string("three"));  
v.push_back(std::string());
```

## ■ Again, *unnecessary temporaries*.

## ■ Fix:

```
v.push_back(s);  
v.push_back(s + ", two");  
v.emplace_back("three");  
v.emplace_back();
```

# Classes Holding Multiple Resources

## ■ What is wrong with this code?

```
class TwoResources {  
public:  
    TwoResources() : x_(nullptr) : y_(nullptr) {  
        x_ = new X;  
        y_ = new Y;  
    }  
    ~TwoResources() {  
        delete x_;  
        delete y_;  
    }  
private:  
    X* x_;  
    Y* y_;  
};
```

- If an exception is thrown in a constructor, the object being constructed is deemed not to have started its lifetime and no destructor will ever be called for the object.

- So, for example, if **new** Y throws, x\_ will be leaked.

## ■ Fix:

```
class TwoResources {  
public:  
    TwoResources() : x_(make_unique<X>()),  
                    y_(make_unique<Y>()) {}  
private:  
    unique_ptr<X> x_;  
    unique_ptr<Y> y_;  
};
```

# Avoid Returning By Const Value

- What is wrong with the following code?

```
const std::string getMessage() {  
    return "Hello";  
}
```

- The const return value will *interact poorly with move semantics*, as the returned object cannot be used as the source for a move operation (since the source for a move operation must be modifiable).
- Fix:

```
std::string getMessage() {  
    return "Hello";  
}
```

- When specifying a literal, be careful to use a literal of the correct type, as the type can often be quite important.
- For example, what value will be printed by the following code and (more importantly) why:

```
std::vector<double> values;  
values.push_back(0.5);  
values.push_back(0.5);  
// Compute the sum of the elements in the vector values.  
double sum = std::accumulate(values.begin(),  
    values.end(), 0);  
std::cout << sum << '\n';
```

- Hint: The value printed for `sum` is not 1.
- In order to determine what values will be printed, look carefully at the definition of `std::accumulate`.
- Answer: The value printed for `sum` is 0.

This slide is a placeholder and has been intentionally left blank.

question

discussion

## Part 13

### **Additional Learning Resources**

- Know what you do not know.
- Ask questions when you are uncertain about something and be sure that the person whom you ask is knowledgeable enough to give a *correct* answer.
- Know what information resources can be *trusted*.
- Learn to use reference materials effectively (e.g., documentation on libraries, standards).



- Some good references on various topics related to the C++ programming language, C++ standard library, and other C++ libraries (such as Boost) are listed on the slides that follow.
- Any information on C++ (e.g., books, tutorials, videos, seminars) from the following individuals (who are held in very high regard by the C++ community) is highly recommended:
  - Bjarne Stroustrup (the creator of C++)
  - Scott Meyers
  - Herb Sutter (Convener of ISO C++ standards committee for over 10 years)
  - Andrei Alexandrescu
  - Stephan Lavavej

- 1 ISO/IEC 14882:2017 — information technology — programming languages — C++, Dec. 2017.

This is the definitive specification of the C++ language and standard library. This is an essential reference for any advanced programmer.

- 2 B. Stroustrup. *The C++ Programming Language*. Addison Wesley, 4th edition, 2013.

This is the classic book on the C++ programming language and standard library, written by the creator of the language. This is one of the best references for first learning C++.

- 3 Standard C++ Foundation web site. <http://www.isocpp.org>, 2014.

This is the web site of a non-profit organization whose purpose is to support the C++ software development community and promote the understanding and use of modern standard C++ on all compilers and platforms. This is an absolutely outstanding source of information on C++.

- 4 B. Stroustrup and H. Sutter (editors), C++ Core Guidelines, 2016, <http://isocpp.github.io/CppCoreGuidelines/CppCoreGuidelines>.  
This document provides a very detailed set of guidelines for writing good C++ code.
- 5 S. Meyers. *Effective Modern C++: 42 Specific Ways to Improve Your Use of C++11 and C++14*.  
O'Reilly Media, Cambridge, MA, USA, 2015.  
This book covers a list of 42 topics on how to better utilize the C++ language.
- 6 S. Meyers. *Effective C++: 50 Specific Ways to Improve Your Programs and Designs*.  
Addison Wesley, Menlo Park, California, 1992.  
This book covers a list of 50 topics on how to better utilize the C++ language.

- 7 S. Meyers. *More Effective C++: 35 New Ways to Improve Your Programs and Designs*.

Addison Wesley, Menlo Park, California, 1996.

This book covers a list of 35 topics on how to better utilize the C++ language. It builds on Meyers' earlier "Effective C++" book.

- 8 S. Meyers. *Effective STL: 50 Specific Ways to Improve Your Use of the Standard Template Library*.

Addison Wesley, 2001.

This book covers a list of 50 topics on how to better utilize the Standard Template Library (STL), an essential component of the C++ standard library.

- 9 N. M. Josuttis. *The C++ Standard Library: A Tutorial and Reference*.

Addison Wesley, Upper Saddle River, NJ, USA, 2nd edition, 2012.

This is a very comprehensive book on the C++ standard library. This is arguably the best reference on the standard library (other than the C++ standard).

- 10 D. Vandevoorde and N. M. Josuttis. *C++ Templates: The Complete Guide*. Addison Wesley, 2002.

This is a very comprehensive book on template programming in C++. It is arguably one of the best books on templates in C++.

- 11 A. Williams. *C++ Concurrency in Action*. Manning Publications, Shelter Island, NY, USA, 2012.

This is a fairly comprehensive book on concurrency and multithreaded programming in C++. It is arguably the best book available for those who want to learn how to write multithreaded code using C++.

- 12 H. Sutter. *Exceptional C++: 47 Engineering Puzzles, Programming Problems, and Solutions*. Addison Wesley, 1999.

This book covers topics including (but not limited to): proper resource management, exception safety, RAII, and good class design.

- 13 H. Sutter. *More Exceptional C++: 40 New Engineering Puzzles, Programming Problems, and Solutions*. Addison Wesley, 2001.

This book covers topics including (but not limited to): exception safety, effective object-oriented programming, and correct use of STL.

- 14 H. Sutter. *Exceptional C++ Style: 40 New Engineering Puzzles, Programming Problems, and Solutions*. Addison Wesley, 2004.

This book covers topics including (but not limited to): generic programming, optimization, resource management, and how to write modular code.

- 15 H. Sutter and A. Alexandrescu. *C++ Coding Standards: 101 Rules, Guidelines, and Best Practices*. Addison Wesley, 2004.

This book presents 101 best practices, idioms, and common pitfalls in C++ in order to allow the reader to become a more effective C++ programmer.

- 16 A. Langer and K. Kreft. *Standard C++ IOStreams and Locales*. Addison Wesley, 2000.

This book provides a very detailed look at C++ I/O streams and locales.

- 17 V. A. Punathambekar. *How to interpret complex C/C++ declarations*. <http://www.codeproject.com/Articles/7042/How-to-interpret-complex-C-C-declarations>, 2004.

This is a detailed tutorial on how to interpret complex C/C++ type declarations. This tutorial explains how type declarations are parsed in the language, which is essential for all programmers to understand clearly.

- 1 S. B. Lippman, J. Lajoie, and B. E. Moo. *C++ Primer*. Addison Wesley, Upper Saddle River, NJ, USA, 4th edition, 2005.
- 2 A. Koenig and B. E. Moo. *Accelerated C++: Practical Programming by Example*. Addison Wesley, Upper Saddle River, NJ, USA, 2000.
- 3 B. Eckel. *Thinking in C++—Volume 1: Introduction to Standard C++*. Prentice Hall, 2nd edition, 2000.
- 4 B. Eckel and C. Allison. *Thinking in C++—Volume 2: Practical Programming*. Prentice Hall, 1st edition, 2003.
- 5 B. Stroustrup. *Programming: Principles and Practice Using C++*. Addison Wesley, Upper Saddle River, NJ, USA, 2009.  
An introduction to programming using C++ by the creator of the language.



- 6 A. Alexandrescu. *Modern C++ Design*.  
Addison Wesley, Upper Saddle River, NJ, USA, 2001.
- 7 D. Abrahams and A. Gurtovoy. *C++ Template Metaprogramming: Concepts, Tools, and Techniques from Boost and Beyond*.  
Addison Wesley, Boston, MA, USA, 2004.
- 8 D. D. Gennaro. *Advanced C++ Metaprogramming*.  
CreateSpace Independent Publishing Platform, 2011.
- 9 Boost web site. <http://www.boost.org>, 2014.  
The web site for the Boost C++ libraries.
- 10 B. Karlsson. *Beyond the C++ Standard Library: An Introduction to Boost*.  
Addison Wesley, Upper Saddle River, NJ, USA, 2005.  
An introduction to (some parts of) the Boost library.

- 11 B. Schaling. *The Boost C++ Libraries*.

XML Press, 2nd edition, 2014.

An introduction to the Boost library. Online version at

<http://theboostcpplibraries.com>.

- 12 M. Kilpelainen. *Overload resolution — selecting the function*.

*Overload*, 66:22–25, Apr. 2005.

Available online at <http://accu.org/index.php/journals/268>.

# Yet More C++ References I

- 1 Herb Sutter's Web Site: <http://herbsutter.com>
- 2 Herb Sutter's Guru of the Week: <http://www.gotw.ca/gotw/>
- 3 Bjarne Stroustrup's Web Site: <http://www.stroustrup.com>
- 4 ISO C++ Working Group web site:  
<http://www.open-std.org/jtc1/sc22/wg21/>
- 5 ISO C++ Standards Committee GitHub site:  
<https://github.com/cplusplus>
- 6 C++ FAQ: <http://www.parashift.com/c++-faq/>
- 7 Newsgroup comp.lang.c++.moderated: <https://groups.google.com/forum/#!forum/comp.lang.c++.moderated>
- 8 <http://en.cppreference.com>
- 9 <http://www.cplusplus.com>
- 10 Stackoverflow: <http://stackoverflow.com>

- 11 Cpp Reddit (C++ discussions, articles, and news):  
<https://www.reddit.com/r/cpp>
- 12 Cplusplus Reddit (C++ questions, answers, and discussion):  
<https://www.reddit.com/r/cplusplus>
- 13 ACCU Overload Journal:  
<http://accu.org/index.php/journals/c78/>
- 14 The C++ Source: <http://www.artima.com/cppsource>

- 1 CppCon, <https://cppcon.org>
- 2 C++ Now, <http://cppnow.org>
- 3 Meeting C++, <https://meetingcpp.com>
- 4 code::dive, <https://codedive.pl>
- 5 ACCU Conference, <https://conference.accu.org>
- 6 Pacific++, <https://pacificplusplus.com>
- 7 NDC Conferences, <https://ndcconferences.com>
  - NDC TechTown, <https://ndctechtown.com>

1 Google Code Jam

<https://code.google.com/codejam/>

2 Topcoder

<https://www.topcoder.com/>

3 IEEEExtreme 24-Hour Programming Competition

<http://www.ieee.org/xtreme>

4 ACM International Collegiate Programming Contest (ICPC)

<http://icpcnews.com/>

5 CodeChef

<https://www.codechef.com/>

- Use as many information resources as you can to learn as much as you can about C++.
- Read books, articles, and other documents.
- Watch videos.
- Attend lectures and seminars.
- Participate in programming competitions.
- But most importantly:

**Write code!**

**Write lots and lots and lots of code!**

- The only way to truly learn a programming language well is to use it heavily (i.e., write lots of code using the language).