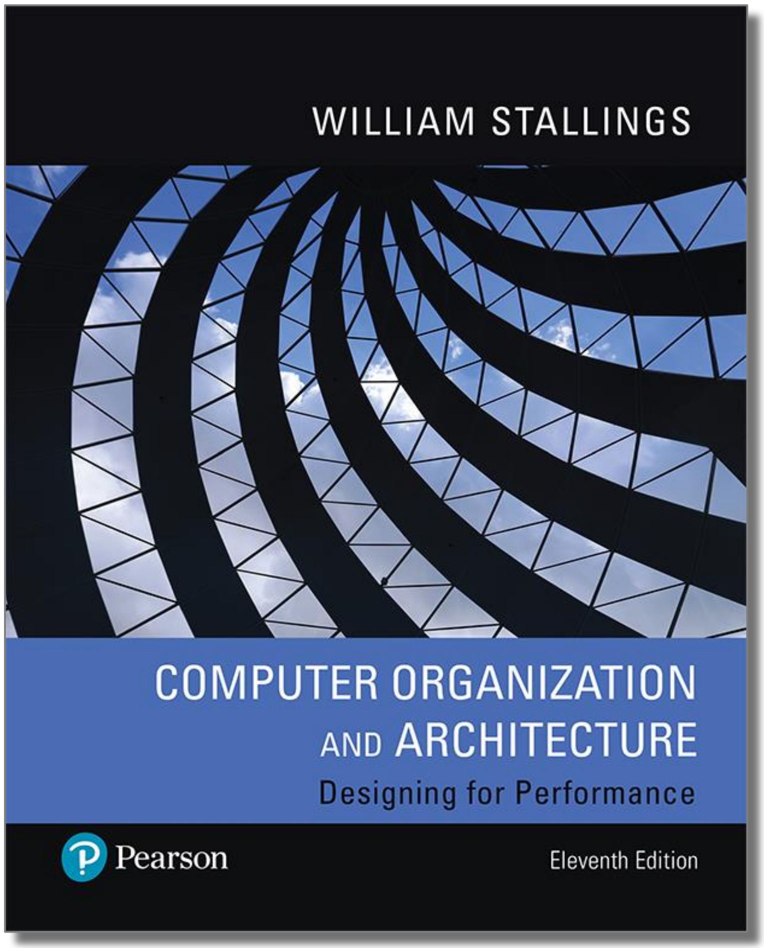
Computer Organization and Architecture

Designing for Performance

11th Edition



Chapter 4

The Memory Hierarchy: Locality and Performance

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Principle of Locality (1 of 2)

- Also referred to as the *locality of reference*
- Reflects the observation that during the course of execution of a program, memory references by the processor tend to **cluster**
- Think about memory access when:
  - Searching in an array – elements are sequentially stored in memory
  - Using loops – repeated small set of instructions
  - Executing a subroutine

Pearson

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Principle of Locality (1 of 2)

- Locality is based on three assertions:
  - During any interval of time, a program references memory location **non-uniformly**, meaning some units of memory are **more likely to be accessed** than others.
  - As a function of time, the probability that a given unit of memory is referenced tends to change slowly. The **probability distribution** of memory references across the entire memory space tends to **change slowly over time**.
  - The **correlation** between **immediate past and immediate future memory reference patterns is high** and is weakened as the time interval increases

Pearson

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Principle of Locality (2 of 2)

- Two forms of locality
  - Temporal locality
    - Refers to the tendency of a program to reference in the near future those units of memory referenced in the recent past
    - Constants, temporary variables, and working stacks are also constructs that lead to this principle
- Spatial locality
  - Spatial locality
    - Refers to the tendency of a program to reference units of memory whose addresses are near one another
    - Also reflects the tendency of a program to access data locations sequentially, such as when processing a table of data

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Figure 4.1

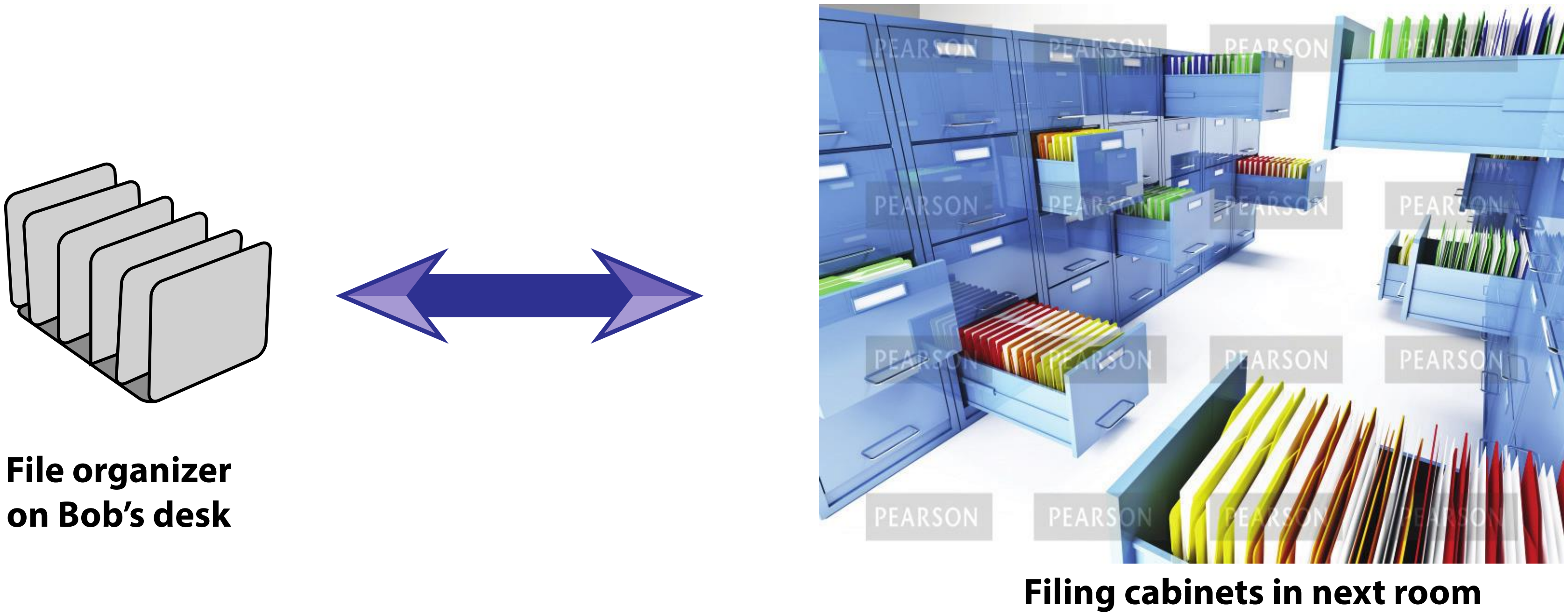


Figure 4.1 Moving File Folders Between Smaller, Faster-Access Storage and Larger, Slower-Access Storage

Pearson
Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Analogy

- Bob spends most of his time dealing with file folders.
- The folder are stored in the cabinets next room.
- For convenience, Bob has a file organizer on his desk which has slots that can hold few folders at one time.

Pearson

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Analogy – Bob Exploits Temporal Locality

- When Bob is working on a file and temporarily is finished, it may be likely that he will need to read or write one of the documents in that file in the near future, so he keeps it in his desk organizer.
- This is an example of temporal locality

Pearson Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Analogy – Bob Exploits Spatial Locality

- Bob also observes that when he retrieves a folder from the filing cabinets, it is likely that in the near future he will need access to some of the nearby folders as well, so he retrieves the folder he needs plus a few folders on either side at the same time.
- This is an example of spatial locality.

Pearson Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Analogy – Returning Folders

- Soon Bob’s desk organizer is going to be filled up.
- Next time he goes to retrieve a folder from the cabinets, he must return some/all of the folders in his organizer.
- What policy should Bob follow in such case?
  - Replace the folder that has been in his desk organizer the longest or the least folder that has been used recently – Temporal Locality
  - Replace all the folders in his desk organizer with a new batch of contiguous folders including the one he needs plus nearby folders – Spatial Locality
- Neither policy is optimal. How about using both to a certain degree? E.g., batches of 10-20% of his desk organizer cap.

Pearson

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Figure 4.2

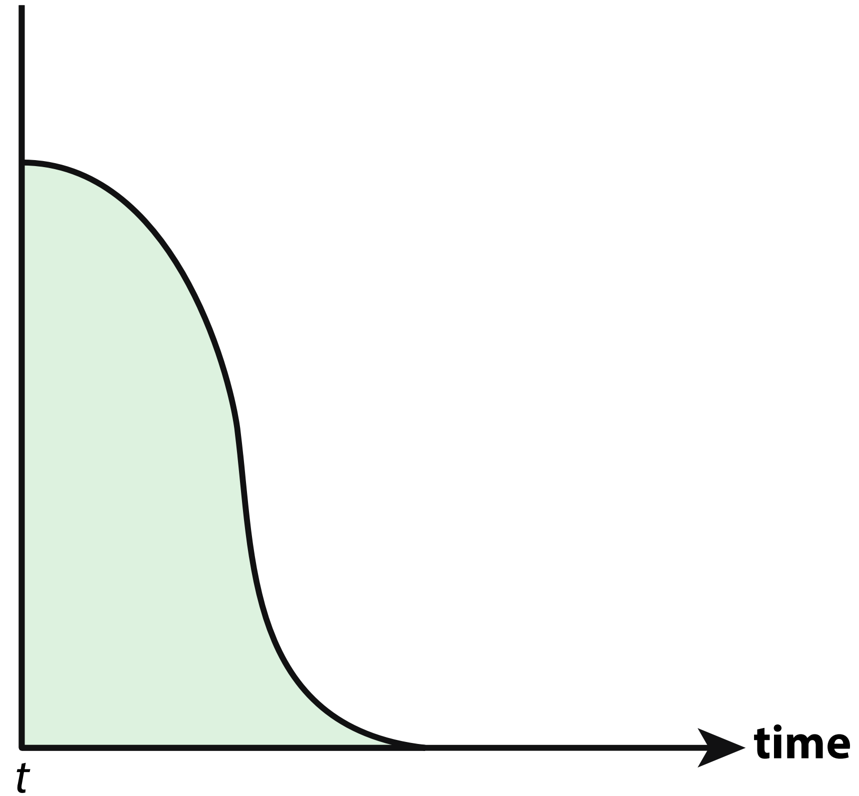


Figure 4.2 Idealized Temporal Locality Behavior: Probability Distribution for Time of Next Memory Access to Memory Unit Accessed at Time *t*

Pearson
Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Figure 4.3

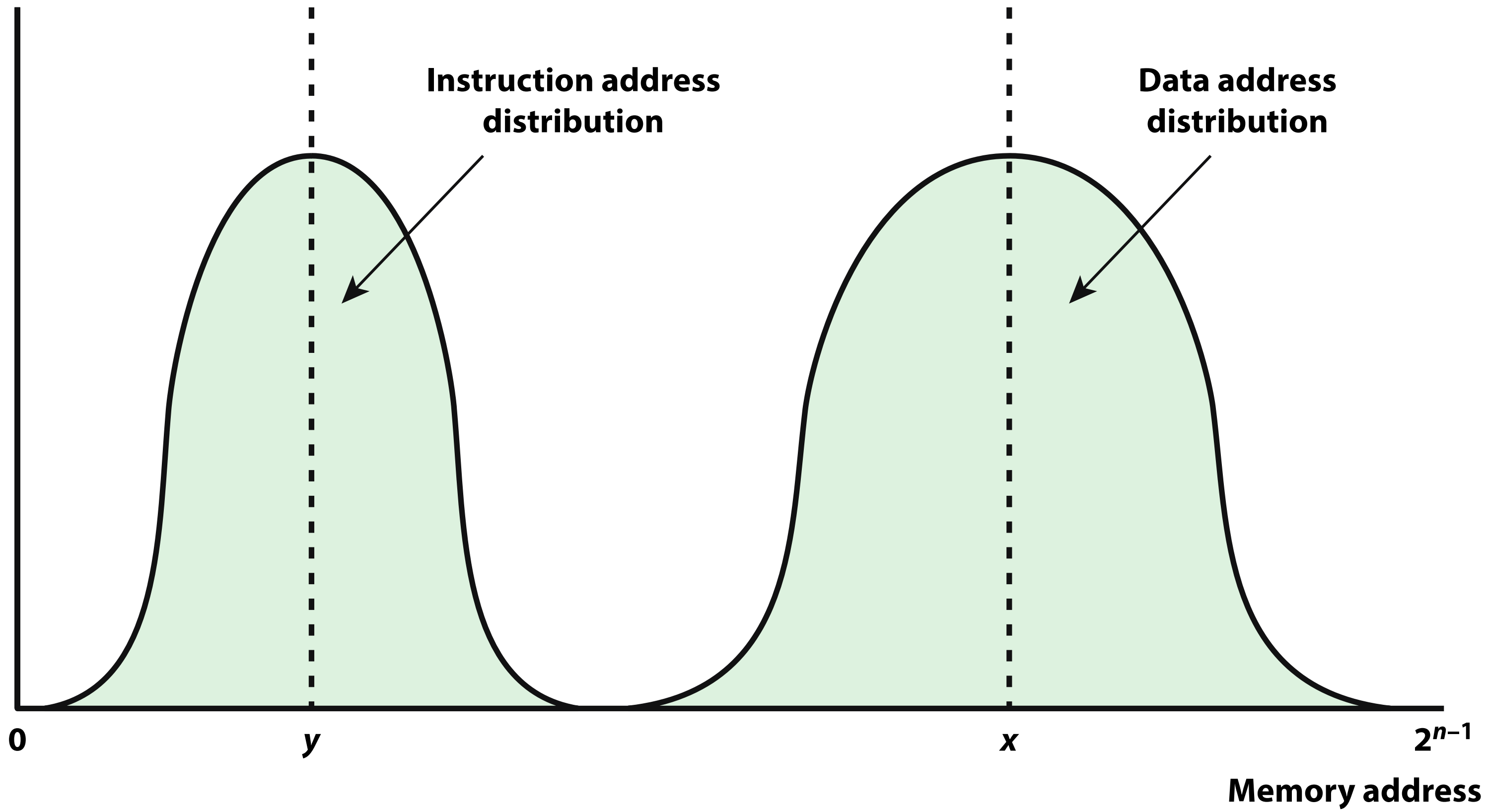


Figure 4.3 Idealized Spatial Locality Behavior: Probability Distribution for Next Memory Access (most recent data memory access at location *x*; most recent instruction fetch at location *y*)

Pearson

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Table 4.1
Key Characteristics of Computer Memory Systems

Location Internal (e.g., processor registers, cache, main memory) External (e.g., optical disks, magnetic disks, tapes) **Capacity** Number of words Number of bytes **Unit of Transfer** Word Block **Access Method** Sequential Direct Random Associative	**Performance** Access time Cycle time Transfer rate **Physical Type** Semiconductor Magnetic Optical Magneto-optical **Physical Characteristics** Volatile/nonvolatile Erasable/nonerasable **Organization** Memory modules

Pearson

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Characteristics of Memory Systems

- Location
  - Refers to whether memory is internal and external to the computer
  - Internal memory is often equated with main memory
  - Processor requires its own local memory, in the form of registers
  - Cache is another form of internal memory
  - External memory consists of peripheral storage devices that are accessible to the processor via I/O controllers
- Capacity
  - Memory is typically expressed in terms of bytes
- Unit of transfer
  - For internal memory the unit of transfer is equal to the number of electrical lines into and out of the memory module

Pearson Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Method of Accessing Units of Data

Sequential access

- Memory is organized into units of data called records
- Access must be made in a specific linear sequence
- Access time is variable

Direct access

- Involves a shared read-write mechanism
- Individual blocks or records have a unique address based on physical location
- Access time is variable

Random access

- Each addressable location in memory has a unique, physically wired-in addressing mechanism
- The time to access a given location is independent of the sequence of prior accesses and is constant
- Any location can be selected at random and directly addressed and accessed
- Main memory and some cache systems are random access

Associative

- A word is retrieved based on a portion of its contents rather than its address
- Each location has its own addressing mechanism and retrieval time is constant independent of location or prior access patterns
- Cache memories may employ associative access

Pearson

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Capacity and Performance:

The two most important characteristics of memory

Three performance parameters are used:

Access time (latency)
- For random-access memory it is the time it takes to perform a read or write operation
- For non-random-access memory it is the time it takes to position the read-write mechanism at the desired location

Memory cycle time
- Access time plus any additional time required before second access can commence
- Additional time may be required for transients to die out on signal lines or to regenerate data if they are read destructively
- Concerned with the system bus, not the processor

Transfer rate
- The rate at which data can be transferred into or out of a memory unit
- For random-access memory it is equal to 1/(cycle time)

Pearson Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Memory

- The most common forms are:
  - Semiconductor memory
  - Magnetic surface memory
  - Optical
- Several physical characteristics of data storage are important:
  - Volatile memory
    - Information decays naturally or is lost when electrical power is switched off
  - Nonvolatile memory
    - Once recorded, information remains without deterioration until deliberately changed
    - No electrical power is needed to retain information
  - Magnetic-surface memories
    - Are nonvolatile
  - Semiconductor memory
    - May be either volatile or nonvolatile
  - Nonerasable memory
    - Cannot be altered, except by destroying the storage unit
    - Semiconductor memory of this type is known as read-only memory (ROM)
- For random-access memory the organization is a key design issue
  - Organization refers to the physical arrangement of bits to form words

Pearson Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved

Copyright

This work is protected by United States copyright laws and is provided solely for the use of instructions in teaching their courses and assessing student learning. dissemination or sale of any part of this work (including on the World Wide Web) will destroy the integrity of the work and is not permitted. The work and materials from it should never be made available to students except by instructors using the accompanying text in their classes. All recipients of this work are expected to abide by these restrictions and to honor the intended pedagogical purposes and the needs of other instructors who rely on these materials.

Copyright © 2019, 2016, 2013 Pearson Education, Inc. All Rights Reserved