UESTC4019: Real-Time Computer Systems and Architecture

Lecture 8

Memory Technologies (Part-1)

Key Characteristics of Computer Memory Systems (1 of 3)

- **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

Key Characteristics of Computer Memory Systems (2 of 3)

- **Access Method**
  - Sequential
  - Direct
  - Random
  - Associative
- **Performance**
  - Access time
  - Cycle time
  - Transfer rate

Key Characteristics of Computer Memory Systems (3 of 3)

- **Physical Type**
  - Semiconductor
  - Magnetic
  - Optical
  - Magneto-optical
- **Physical Characteristics**
  - Volatile/nonvolatile
  - Erasable/nonerasable
- **Organization**
  - Memory modules

Characteristics of Memory Systems (1 of 2)

- 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

Characteristics of Memory Systems (2 of 2)

- 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

Method of Accessing Units of Data (1 of 3)

- 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

Method of Accessing Units of Data (2 of 3)

- 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

Method of Accessing Units of Data (3 of 3)

- 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

Capacity and Performance (1 of 2)

The two most important characteristics of memory are Capacity and Performance.

- 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

Capacity and Performance (2 of 2)

- 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)

Memory (1 of 3)

- The most common forms are:
  - Semiconductor memory
  - Magnetic surface memory
  - Optical
  - Magneto-optical

Memory (2 of 3)

- 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

Memory (3 of 3)

- Nonerasable memory
  - Cannot be altered, except by destroying the storage unit
  - Semiconductor memory of this type is known as read-only memory (ROM)

Memory Hierarchy

- Design constraints on a computer’s memory can be summed up by three questions:
  - How much, how fast, how expensive

- There is a trade-off among capacity, access time, and cost
  - Faster access time, greater cost per bit
  - Greater capacity, smaller cost per bit
  - Greater capacity, slower access time

- The way out of the memory dilemma is not to rely on a single memory component or technology, but to employ a memory hierarchy

The Memory Hierarchy

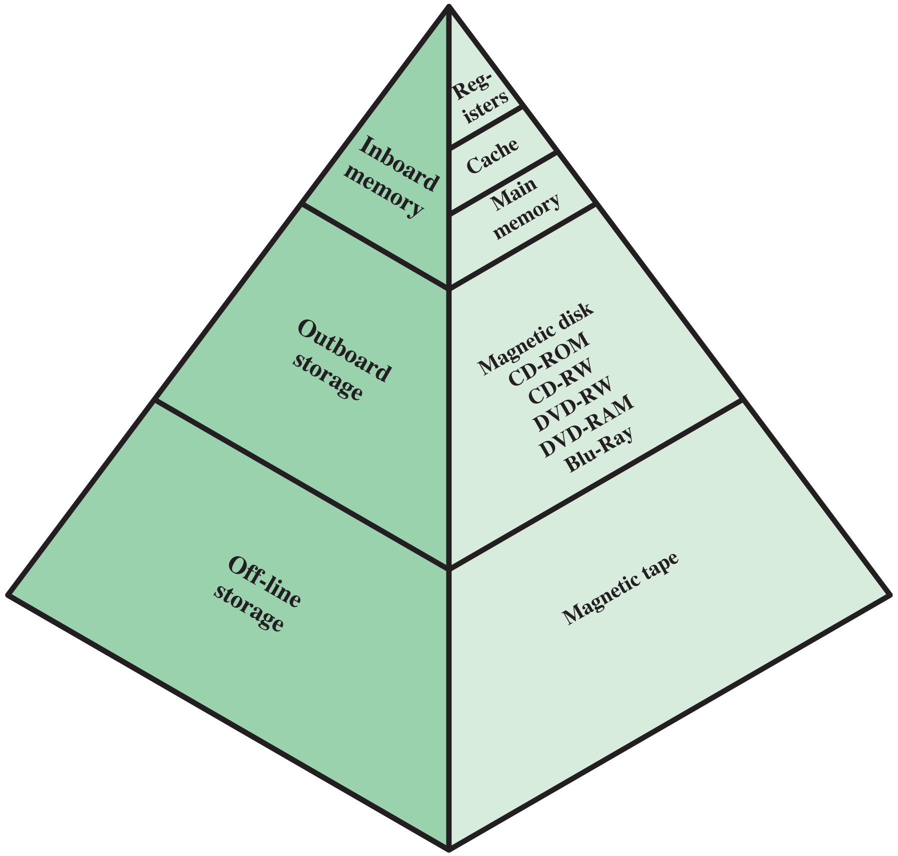
- A typical hierarchy is illustrated in Figure 4.1. As one goes down the hierarchy, the following occur:

  a. Decreasing cost per bit

  b. Increasing capacity

  c. Increasing access time

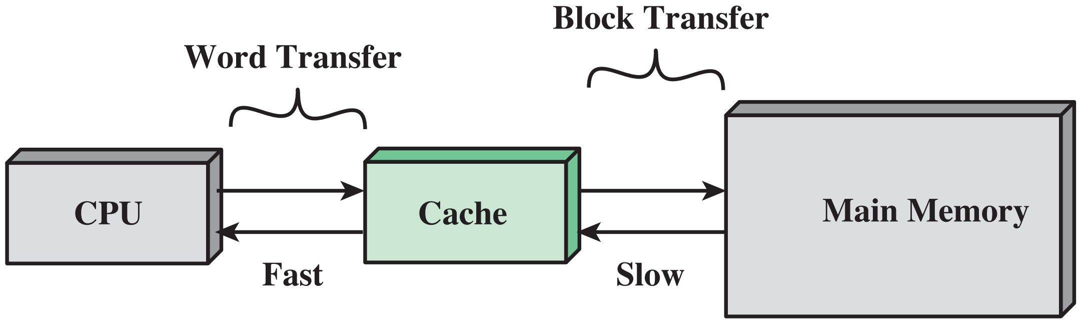
  d. Decreasing frequency of access of the memory by the processor



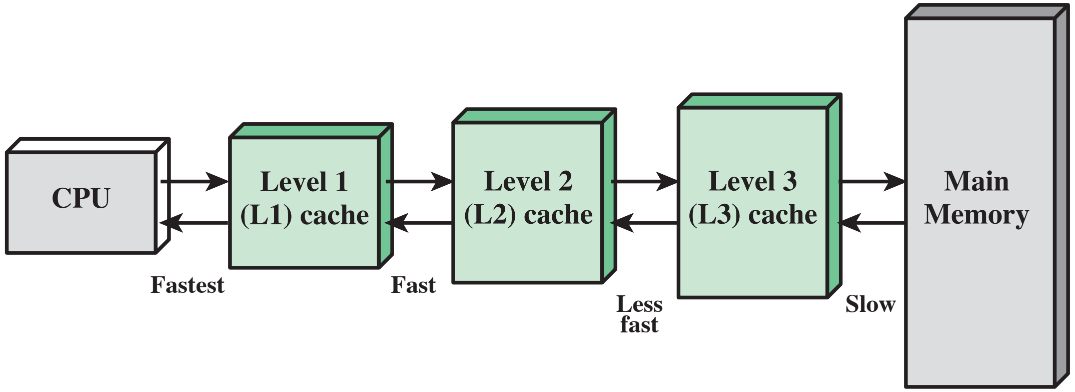
Memory

- The use of three levels exploits the fact that semiconductor memory comes in a variety of types which differ in speed and cost
- Data are stored more permanently on external mass storage devices
- External, nonvolatile memory is also referred to as secondary memory or auxiliary memory
- Disk cache
  - A portion of main memory can be used as a buffer to hold data temporarily that is to be read out to disk
  - A few large transfers of data can be used instead of many small transfers of data
  - Data can be retrieved rapidly from the software cache rather than slowly from the disk

Cache and Main Memory

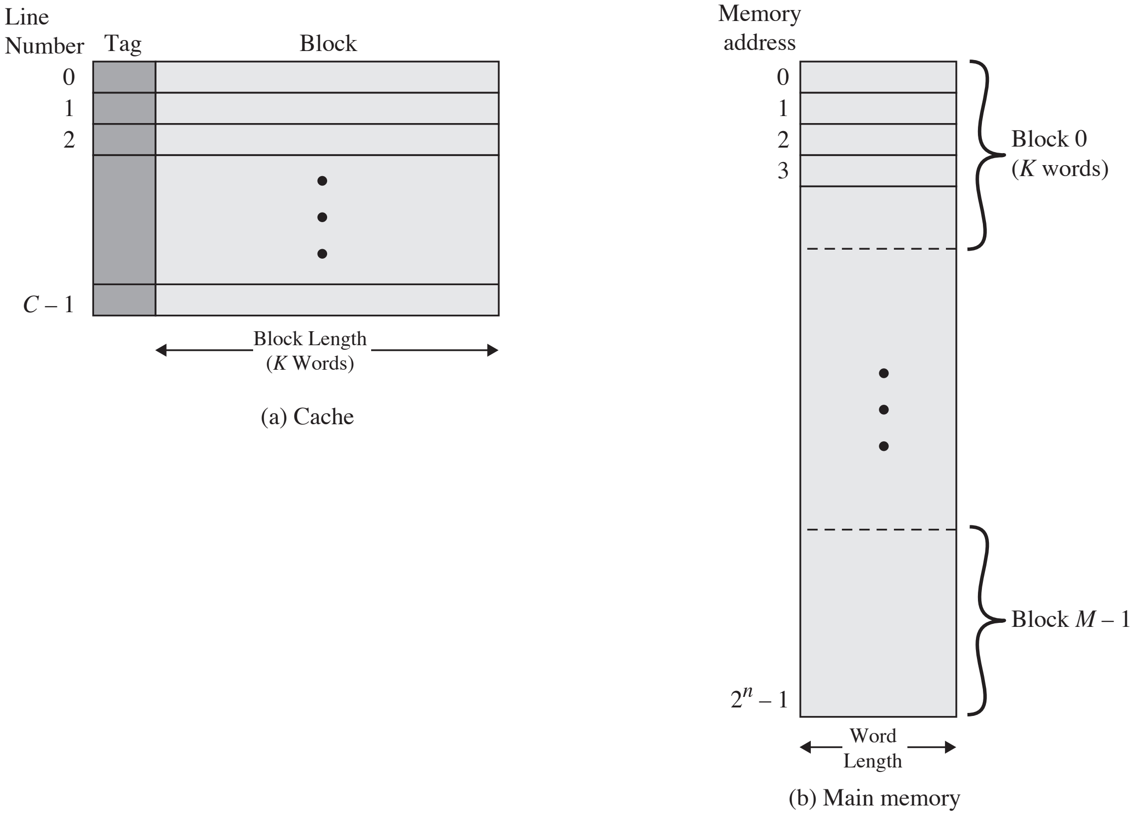


(a) Single cache



(b) Three-level cache organization

Cache/Main-Memory Structure

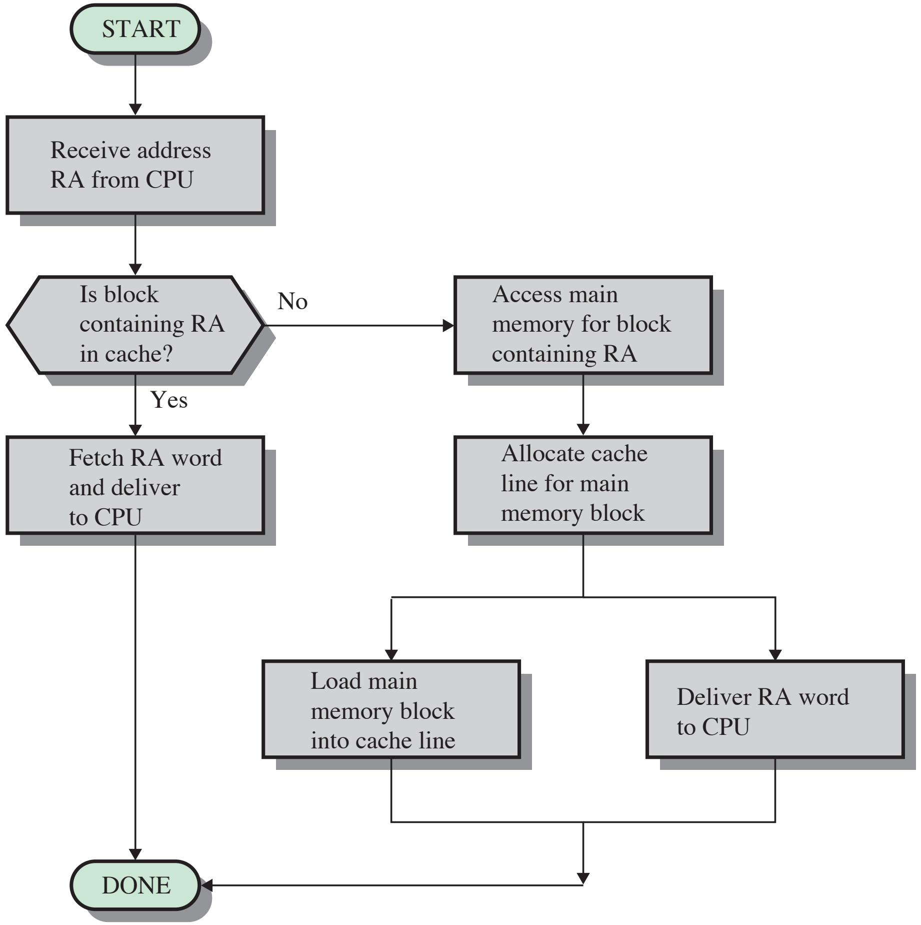


(a) Cache

(b) Main memory

Cache Read Operation

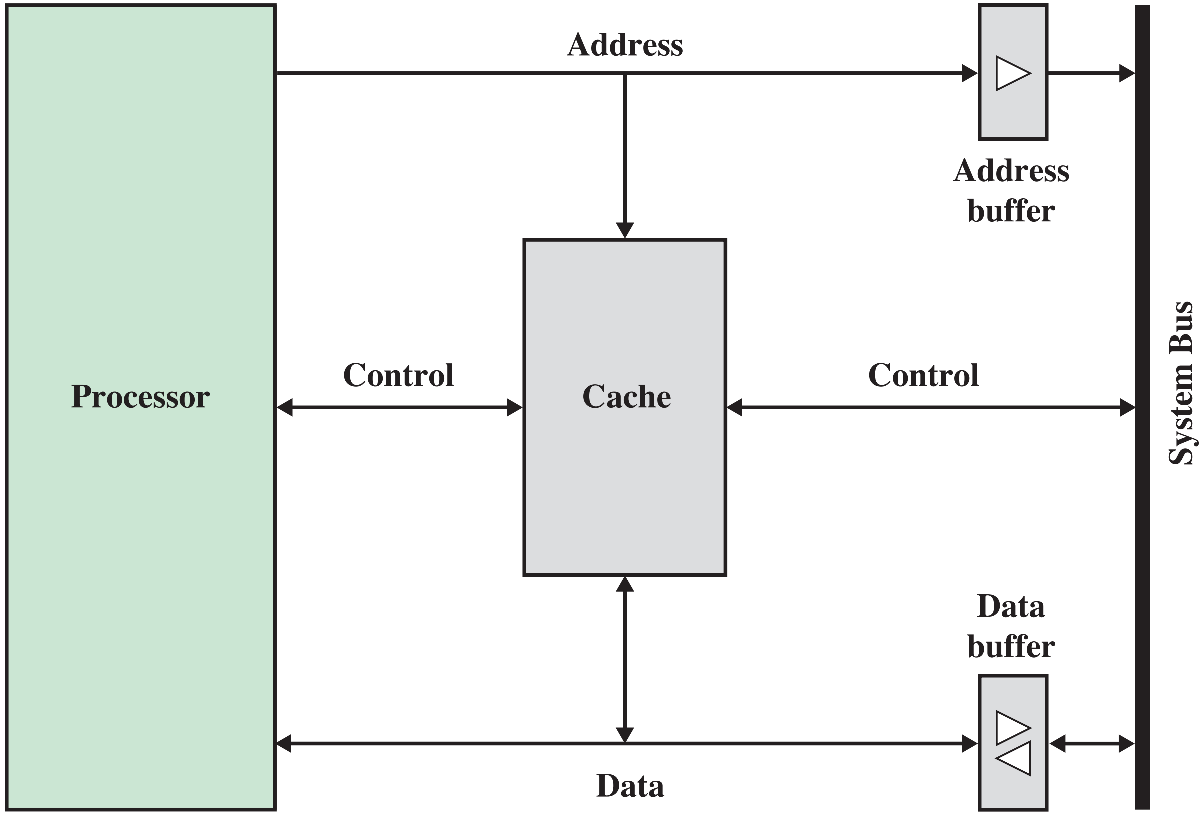
- The processor generates the read address (RA) of a word to be read
- If the word is contained in the cache, it is delivered to the processor
- Otherwise, the block containing that word is loaded into the cache, and the word is delivered to the processor


START
Receive address RA from CPU
Is block containing RA in cache?
No
Yes
Access main memory for block containing RA
Fetch RA word and deliver to CPU
Allocate cache line for main memory block
Load main memory block into cache line
Deliver RA word to CPU
DONE

Typical Cache Organization

- When a cache hit occurs, the data and address buffers are disabled and communication is only between processor and cache, with no system bus traffic

- When a cache miss occurs, the desired address is loaded onto the system bus and the data are returned through the data buffer to both the cache and the processor



Address
Address buffer
Processor
Control
Cache
Control
System Bus
Data buffer
Data