EE5250- COMPUTER ARCHITECTURE AND ORGANIZATION

Cache Memory

Miss. J M K A. Kumari
Dept. of Electrical & Information Engineering
University of Ruhuna

Last week Summery…?

- Interrupt service routine ?
- When will ISR be triggered in 3 stage instruction cycle ?

A program consists of 100 instructions running on a processor with a 5-stage pipeline. Suppose that each stage of the pipeline requires 10 clock cycles to complete.
i) Calculate the speedup of the processor due to pipelining compared to a processor without pipelining, assuming there are no branch instructions.
ii) Calculate the speedup of the processor when 20 of the instructions are branch instructions.
iii) Suppose that the branch prediction is occupied and the prediction accuracy is 90%. Calculate the speedup of the processor compared to a processor without pipelining

- A problem which is being addressed by introducing an **expansion bus.**
- A problem associated with this bus architecture ? → A solution to the problem ?
- Advantages of QPI over bus architecture ?

Characteristics of Memory Systems

- Sequential
  - Access must be made in a specific linear sequence
- Direct
  - Involves a shared read–write mechanism
- Random access
  - Each addressable location in memory has a unique, physically wired-in addressing mechanism
  - Main memory and some cache systems are random access

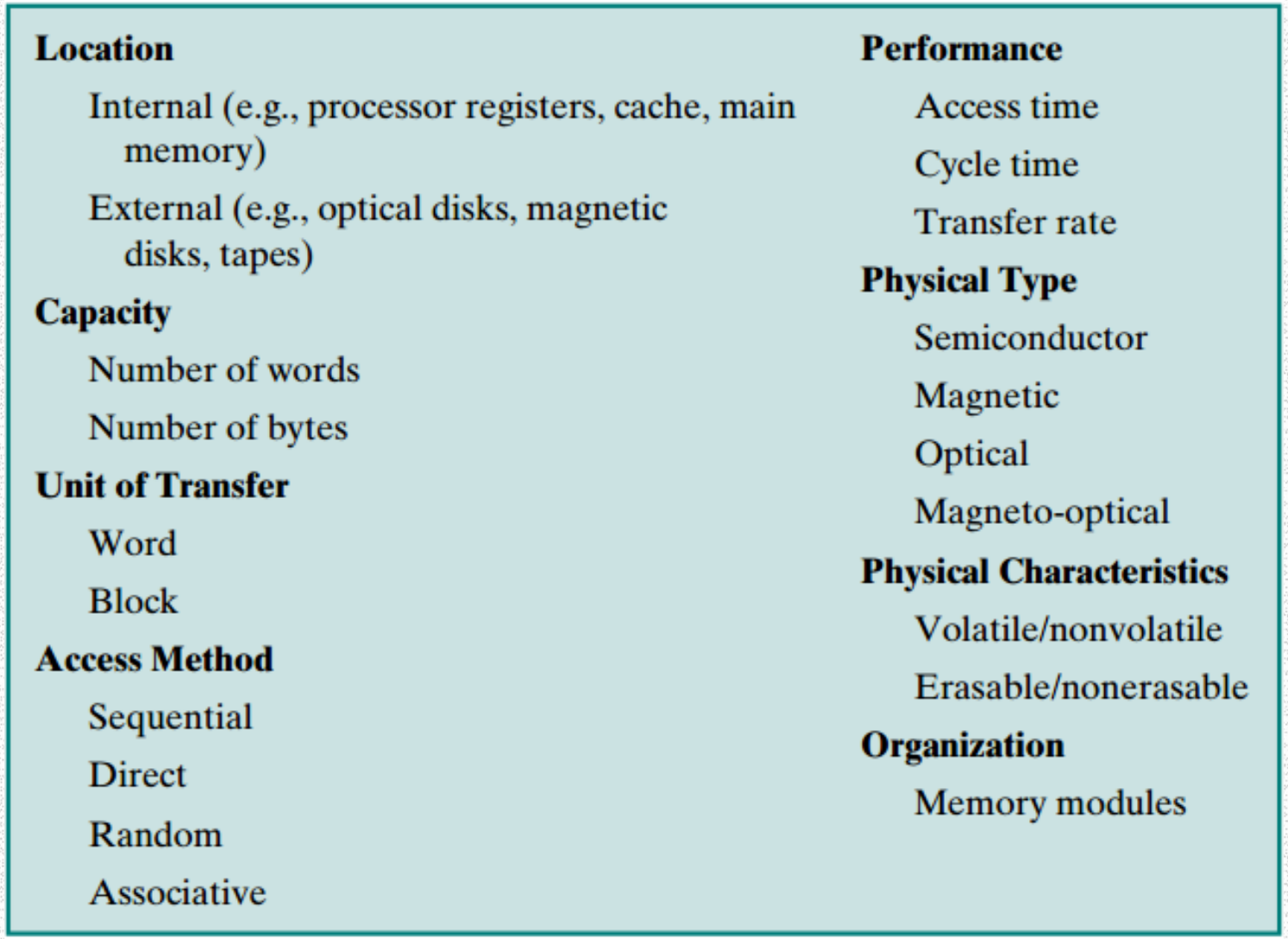


Image source: William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)

Characteristics of Memory Systems

- Performance parameters
  - Access time (latency)
    - Time it takes to perform a read or write operation
  - Memory cycle time
    - Consists of the access time plus any additional time required before a second access can commence
  - 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)

The Memory Hierarchy

- Design constraints on a computer's memory can be summed up by three questions
  - How much?
    - If the capacity is there, applications will likely be developed to use it.
  - How fast?
    - Memory must be able to keep up with the processor
  - How expensive?
    - The cost of memory must be reasonable in relationship to other components
- To meet performance requirements, the designer needs to use expensive, relatively lower-capacity memories with short access times

The Memory Hierarchy

- As one goes down the hierarchy
  - Decreasing cost per bit
  - Increasing capacity
  - Increasing access time
  - Decreasing frequency of access of the memory by the processor

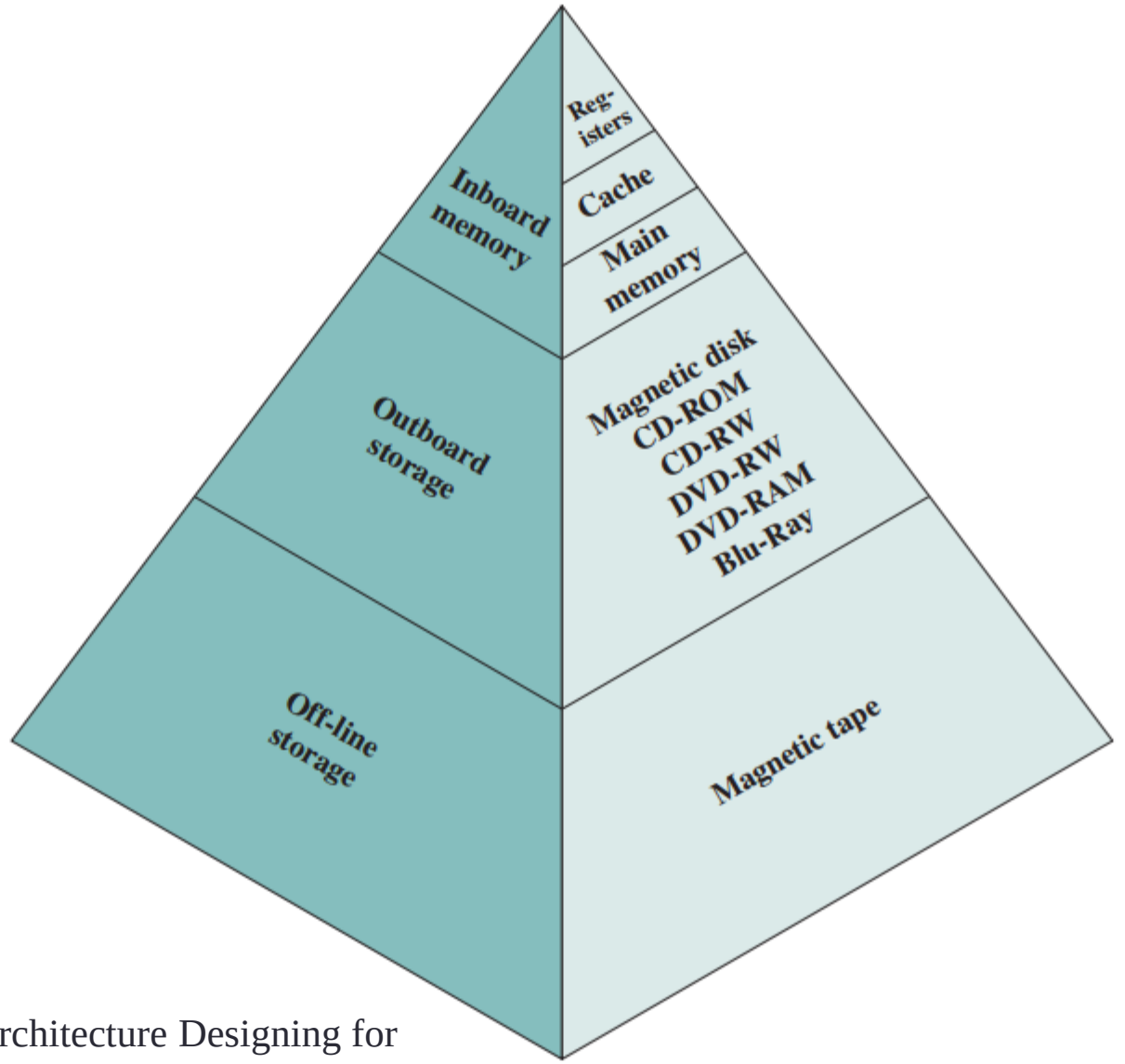


Image source: William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)

The Memory Hierarchy

Suppose that the processor has **access** to two levels of memory

Level 1 contains 1000 words and has an access time of 0.01 μs

Level 2 contains 100,000 words and has an access time of 0.1 μs

Assume that if a word to be accessed is in level 1, then the processor accesses it directly.

If it is in level 2, then the word is first transferred to level 1 and then accessed by the processor

We ignore the time required for the processor to determine whether the word is in level 1 or level 2

Suppose 95% of the memory accesses are found in level 1

- Average time to access a word

  (0.95)(0.01 μs) +(0.05)(0.01 μs +0.1 μs) =0.0095+0.0055=0.015 μs

Cache Memory Principles

- Cache memory enhances computer system performance by combining the speed of expensive, high-speed memory (like SRAM in CPU cache) with the capacity of larger, lower-speed memory (like DRAM). It stores frequently accessed data for quick retrieval, reducing the need to access slower main memory
- Check if the word is in the cache
- If the word is not in the cache, **a block of main memory is read into the cache and then the word is delivered to the processor**

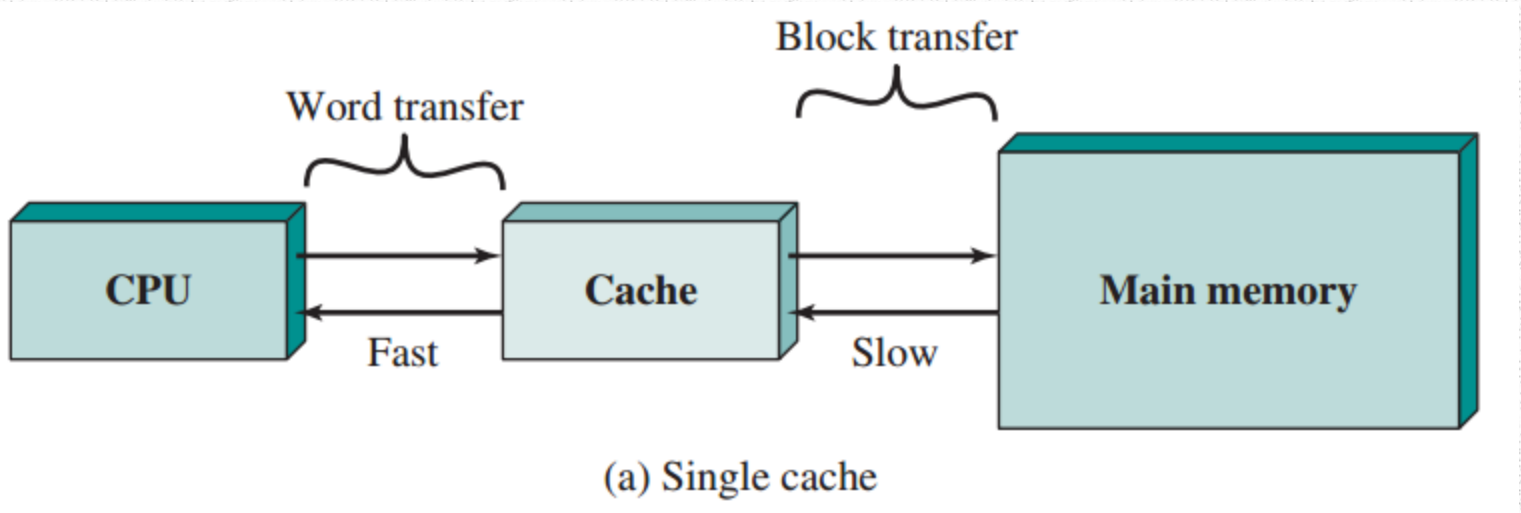
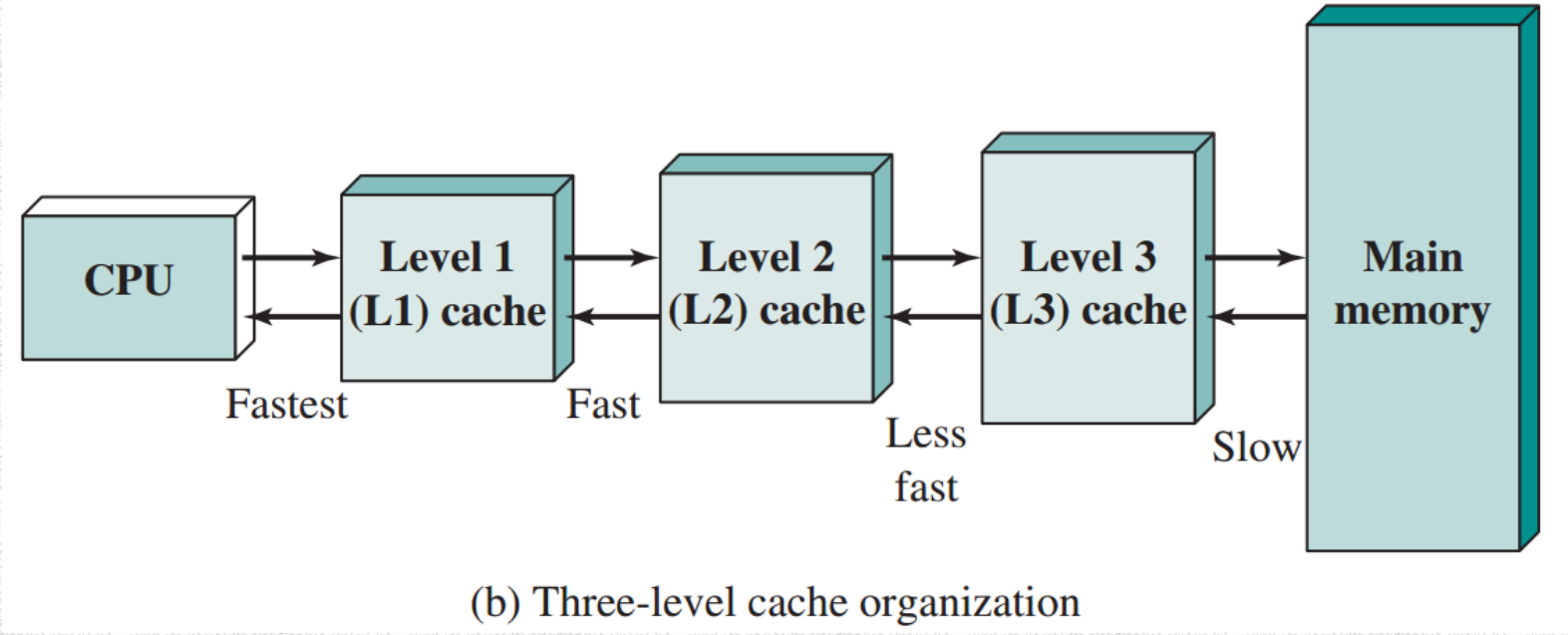


Image source: William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)

Cache Memory Principles

- Phenomenon of locality of reference
  - When a block of data is fetched into the cache to satisfy a single memory reference
  - It is likely that there will be future references to that same memory location or to other words in the block

- L2 cache is slower and typically larger than the L1 cache
- L3 cache is slower and typically larger than the L2 cache



(b) Three-level cache organization

Image source: William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)

Cache Memory Principles

- Main memory consists of up to 2^n addressable words
- Each word having a unique n-bit address
- Memory is consist of a number of fixed-length blocks of K words each
- $M=2^n/K$ blocks in main memory

- The cache consists of m blocks, called lines.
- Each line contains K words
- Each line includes a tag that identifies which particular block is currently being stored

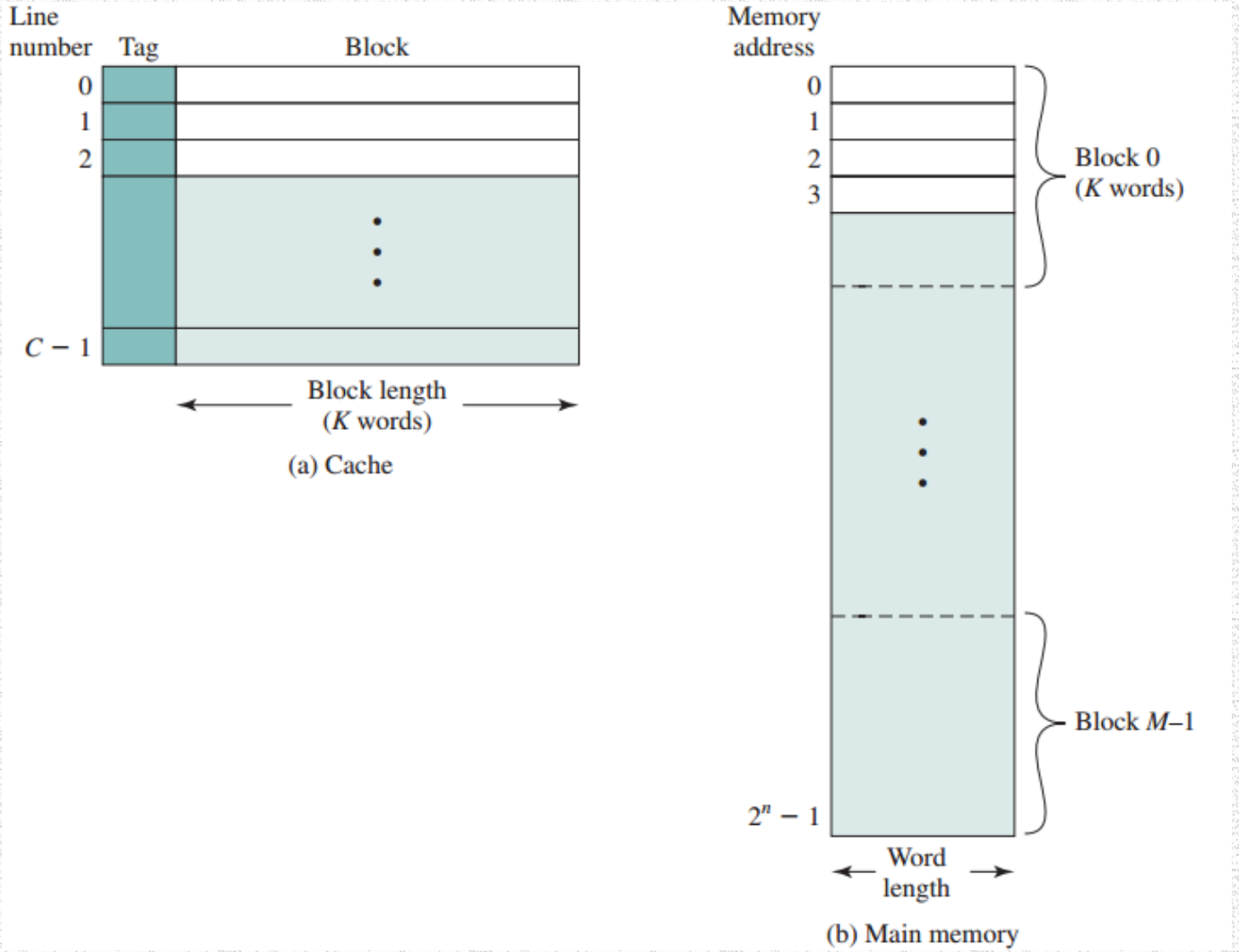


Image source: William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)

Cache Memory Principles

- Read operation
  - Processor generates the Read Address (RA)

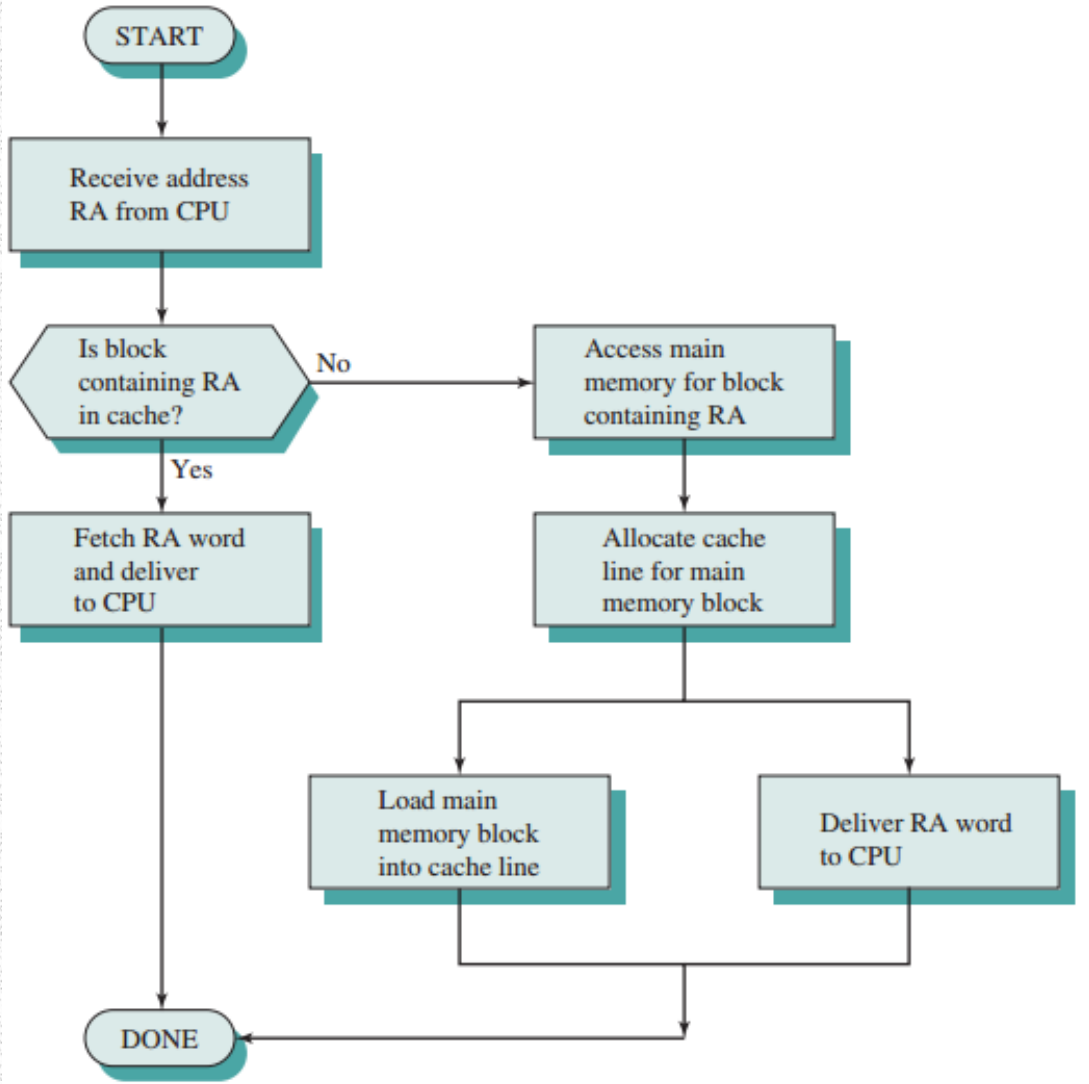


Image source: William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)

Cache Memory Principles

When a cache hit occurs

- The data and address buffers are disabled
- Communication is only between processor and cache

When a cache miss occurs

- The desired address is loaded onto the system bus
- The data are returned through the data buffer to both the cache and the processor
- Some organizations the word is first read into the cache and then transferred to processor

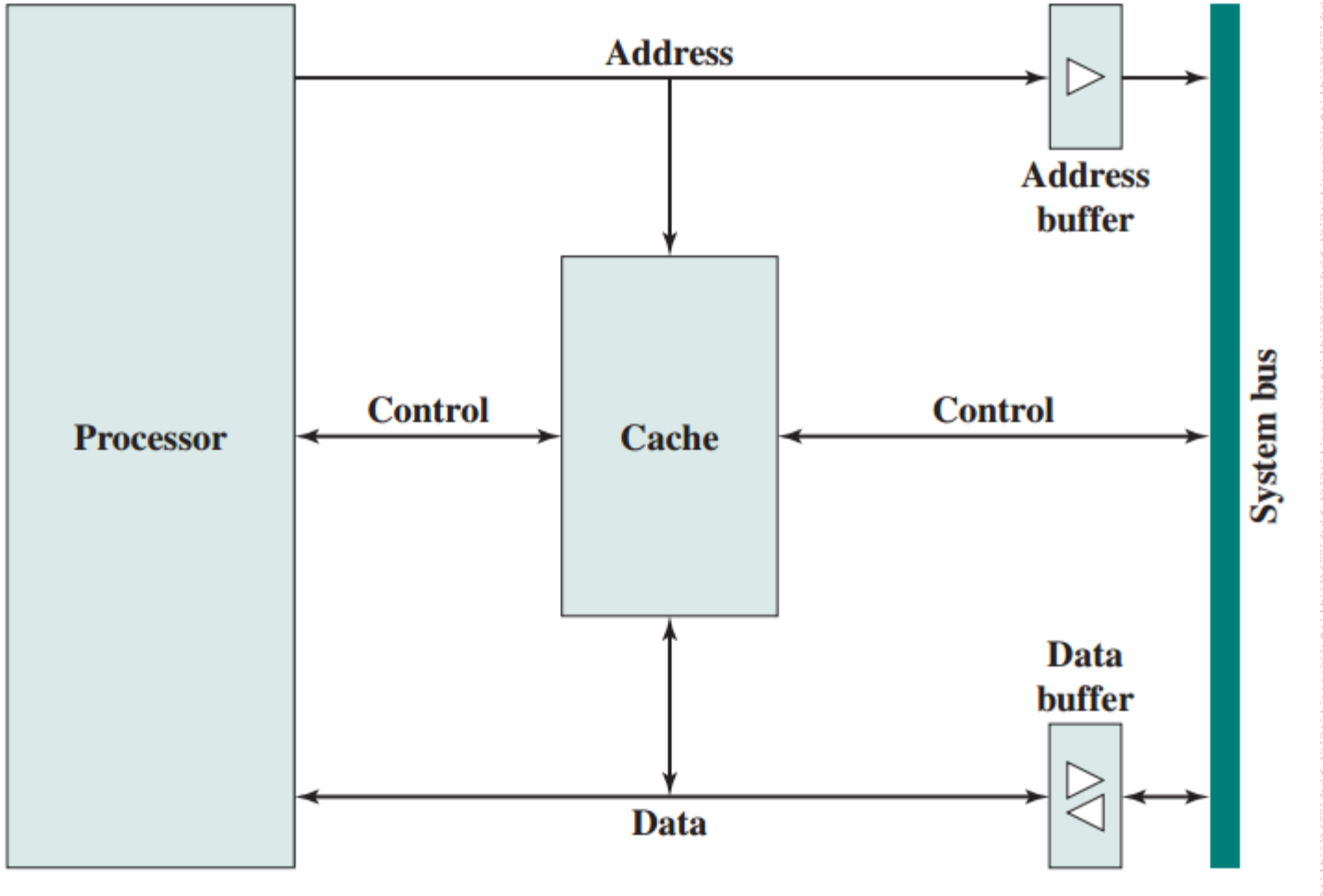


Image source: William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)

Elements of Cache Design

Cache Addresses
- Logical
- Physical

Cache Size

Mapping Function
- Direct
- Associative
- Set associative

Replacement Algorithm
- Least recently used (LRU)
- First in first out (FIFO)
- Least frequently used (LFU)
- Random

Write Policy
- Write through
- Write back

Line Size

Number of Caches
- Single or two level
- Unified or split

Image source: William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)

Exercise

Consider a computer system with a processor that has two levels of cache. The first-level cache (L1) has a hit ratio of 0.8 and an access time of 10ns, while the second-level cache (L2) has a hit ratio of 0.9 and an access time of 20ms. The main memory access time is 100ns.

Calculate the effective cache access time showing all your steps.

Exercise

Consider a computer system with a processor equipped with 8 Kbytes cache, connected to a main memory of 16 Gbytes. The offset is 16 bytes and uses a direct-mapped cache mapping technique.

i) Identify the number of cache lines in the cache memory. Clearly outline each step in your calculation.

ii) Given a memory address of 0xABCD1234 containing the value 0x56, identify the specific cache line this address would map. Show your calculation steps.

iii) Suppose that the memory address 0xDBAD9237 containing the value 0x31 is to be read from the memory. Briefly explain each step of the process along with the values stored at/removed from relevant memory locations, if the write back cache write policy is used

Reference

- William Stallings - Computer Organization and Architecture Designing for Performance (9th Edition)
  - 4.1 Computer Memory System Overview
  - 4.2 Cache Memory Principles
  - 4.3 Elements Of Cache Design
  - 4.4 Pentium 4 Cache Organization