Universiti Malaysia PAHANG
Engineering • Technology • Creativity

BCN1043

COMPUTER ARCHITECTURE & ORGANIZATION

By
Dr. Mritha Ramalingam

Faculty of Computer Systems & Software Engineering
mritha@ump.edu.my

AUTHORS

- **Dr. Mohd Nizam Mohmad Kahar** (mnizam@ump.edu.my)
- **Jamaludin Sallim** (jamal@ump.edu.my)
- **Dr. Syafiq Fauzi Kamarulzaman** (syafiq29@ump.edu.my)
- **Dr. Mritha Ramalingam** (mritha@ump.edu.my)

Faculty of Computer Systems & Software Engineering

BCN1043

COMPUTER ARCHITECTURE & ORGANIZATION

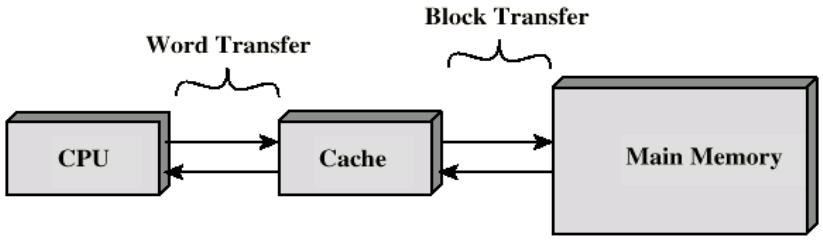
Chapter 5 continues…

COMPUTER MEMORY

1. Storage System & Technology
2. Memory Hierarchy
3. Memory Organization and Operations
4. Cache Memory

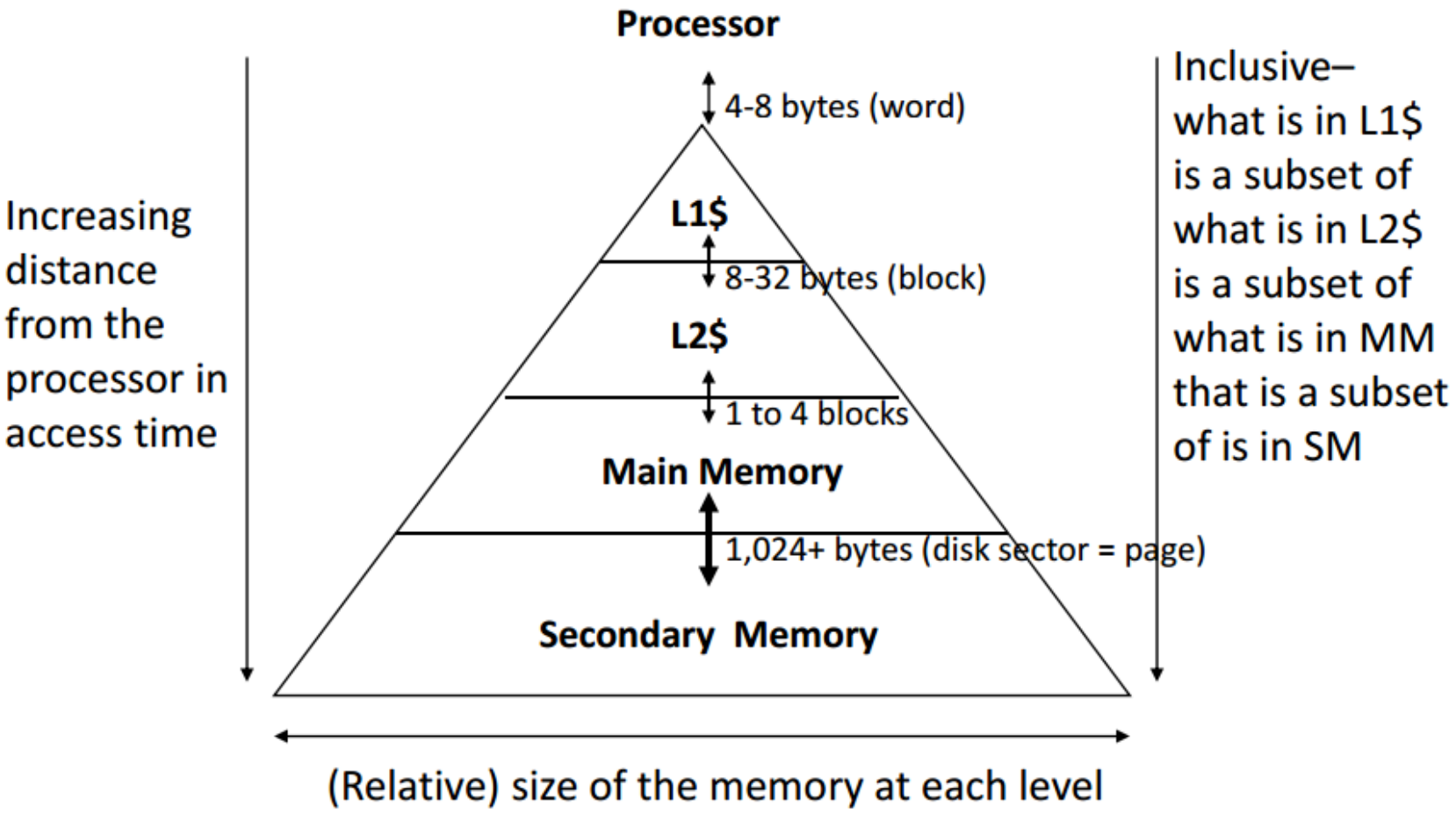
Cache Memory

- Cache memory is
  - relatively small memory
  - Faster
  - Very expensive
  - Cache sits between main memory and CPU
  - Will hold copies of main memory section
  - When the CPU reads a memory word, firstly, the word is checked in cache
  - If present, the word is transferred to CPU
  - otherwise, a main memory block is read into cache, then the word is transferred to CPU.



Source: http://docplayer.net

Characteristics of the Memory Hierarchy



Factors in Cache Design

- Size
- Mapping Function
  - Direct
  - Associative
  - Set Associative
- Line Replacement Algorithm
  - Least recently used (LRU)
  - First in first out (FIFO)
  - Least Frequently used (LFU)
  - Random
- Write Policy
  - Write through
  - Write back
- Block/Line Size
- Number and type of Caches
  - Single or two level
  - Unified or split

- mapping techniques
  - Direct mapping
  - Associative mapping
  - Set associative mapping

Cache of direct mapping includes

- ➔**Tag identifier**
- ➔ **Line number identifier**
- ➔ **Word identifier (offset)**

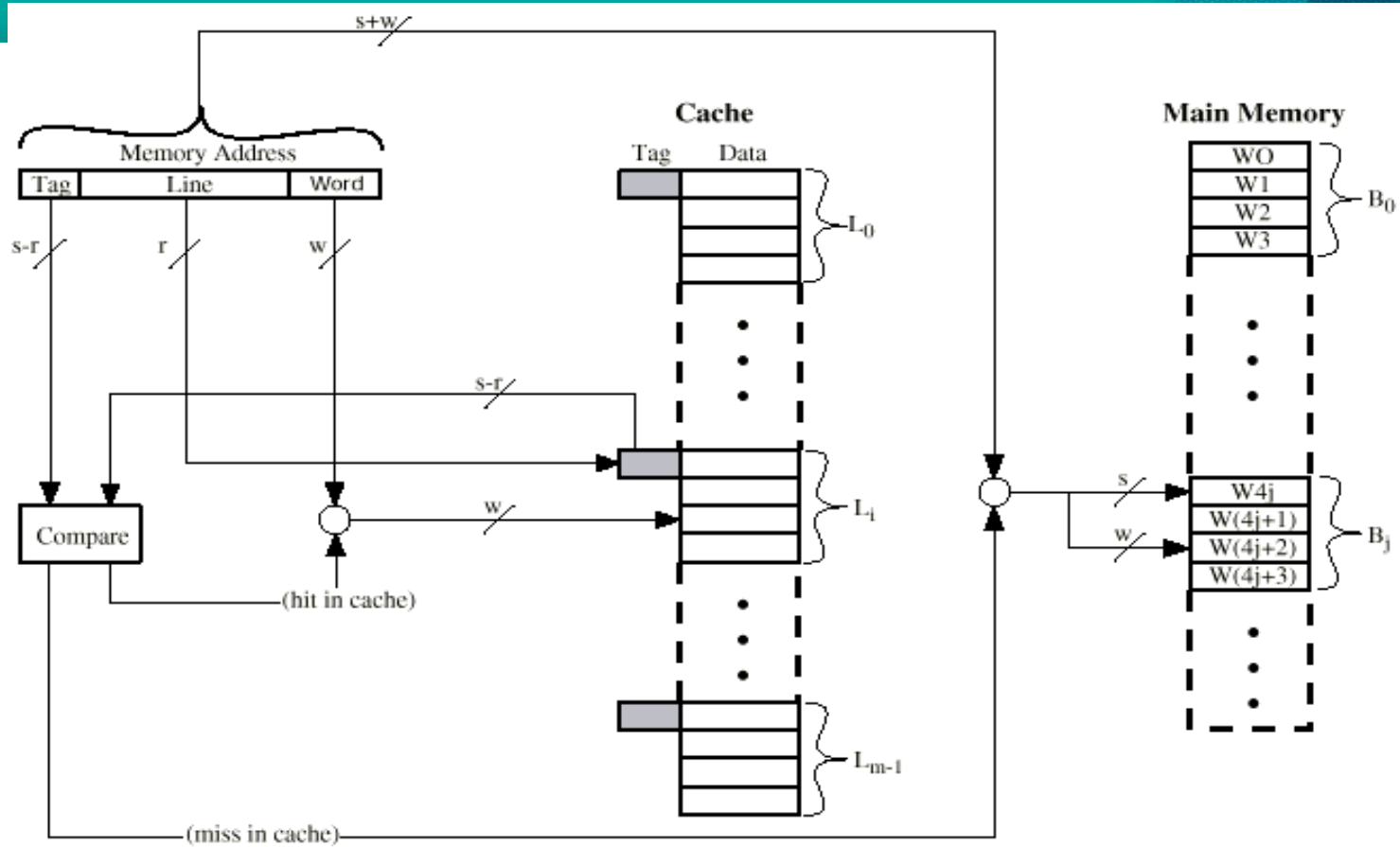
Of main memory address

tag - stored with data

Line identifier - physical line that holds address in cache

Word identifier - specific word in a cache to be read

Direct Mapping Cache Organization



Source: William Stallings, Computer Organization and Architecture,10th Edn

Advantages and disadvantages of Direct Mapping

- Advantages
  - Easy implementation
  - Relatively less implementation cost
  - Easy to determine in cache
- Disadvantages
  - Each main memory block is mapped to a specific cache line
  - Can refer the blocks that map to the same line number

Associative mapping

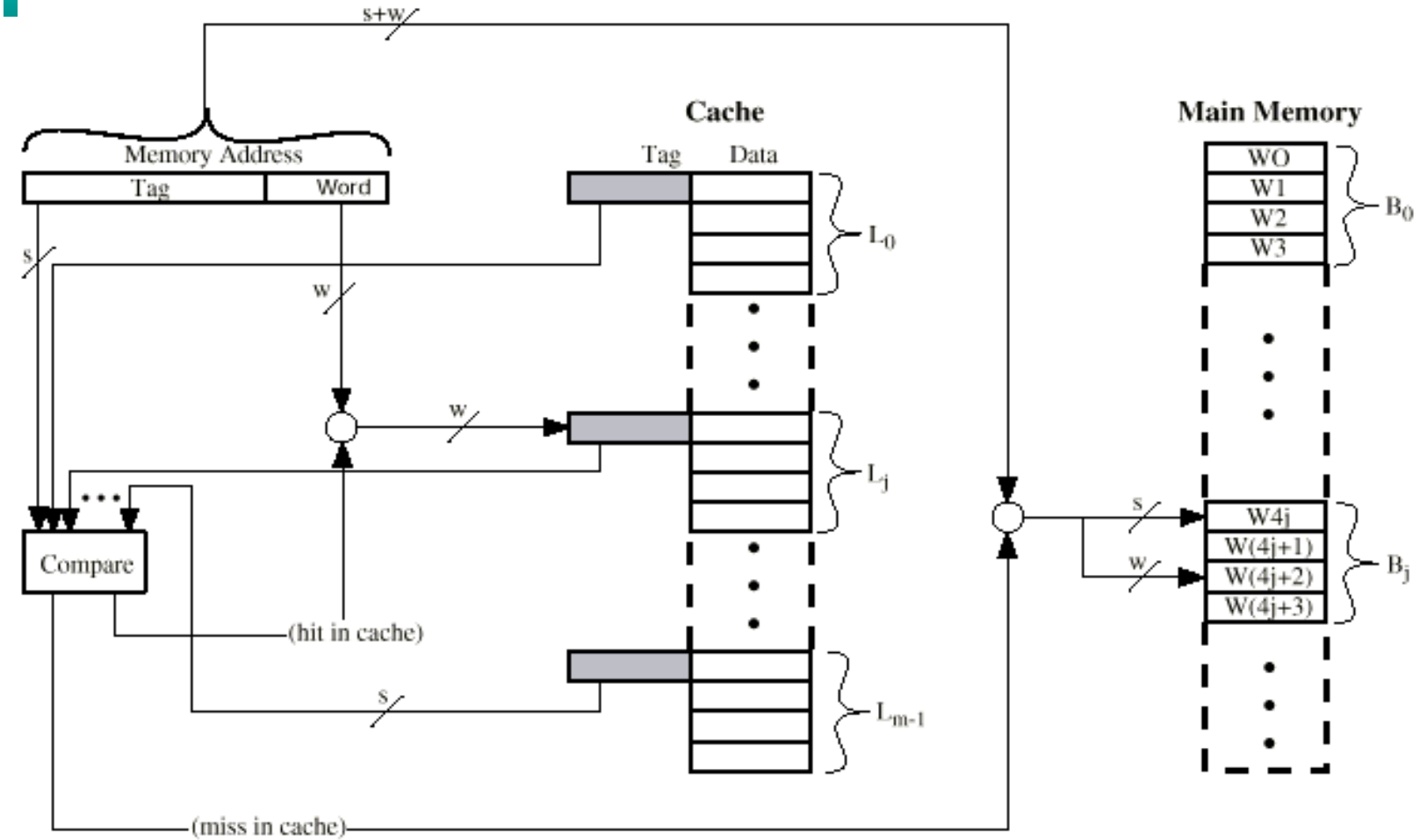
- Advantages

Fast & Flexible

- Disadvantages

Implementation cost

Fully Associative Cache Organization



Source: William Stallings, Computer Organization and Architecture,10th Edn

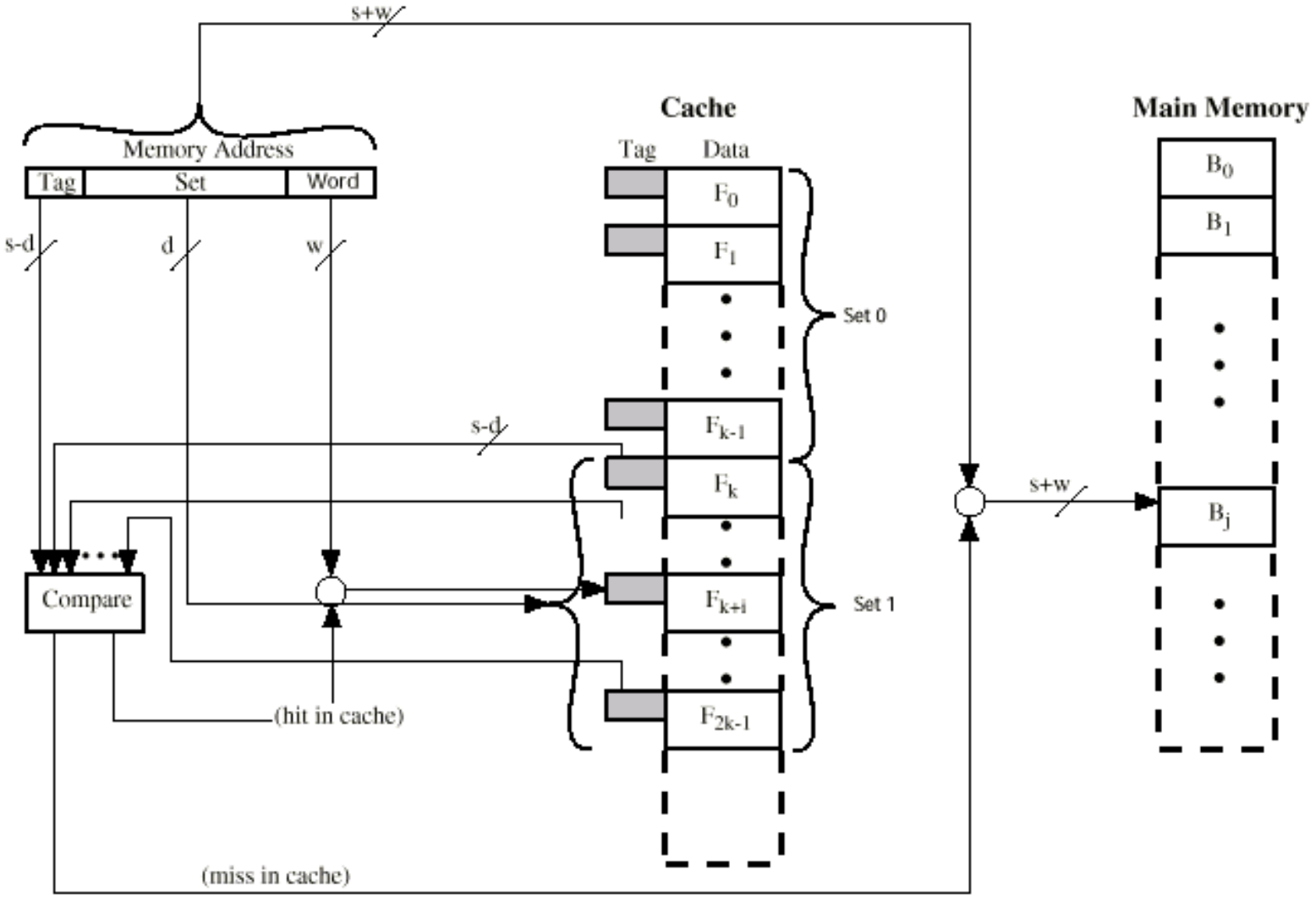
Set Associative Mapping

- Compromise between direct and fully associative mappings that builds on the strengths of both
- Divide cache into a number of sets (v), each set holding a number of lines (k)
- A main memory block can be stored in any one of the k lines in a set such that

  set number = j modulo v

- If a set can hold X lines, the cache is referred to as an *X-way set associative cache*
  - Most cache systems today that use set associative mapping are 2- or 4-way set associative
  - 2 lines per set (2 way set associative mapping)

Two Way Set Associative Cache Organization



Source: William Stallings, Computer Organization and Architecture,10th Edn

Line Replacement Algorithms

- When an associative cache or a set associative cache set is full, which line to replace with the new line that is to be read from memory?
  - Least Recently used (LRU)
    - e.g. in 2 way set associative; Which of the 2 block is LRU?
  - First in first out (FIFO)
    - Replace block that has been in cache longest
  - Least frequently used
    - Replace block which has had fewest hits
  - Random

Write Policy

- While replacing a line, the original copy of the line in main memory must be updated
- E.g
  - Write through
  - Write back

Write through

Anytime a word in cache is changed, it is also changed in main memory

Both copies always agree

Write back

During a write, only change the contents of the cache

Update main memory only when the cache line is to be replaced

Block / line sizes

- The amount of data to be transferred to cache from main memory
- Maintains Complex relationship between block size and hit ratio

Number of caches

- L1
  - Modern CPU chips have on-board cache (L1)
  - L1 provides best performance
  - L2 provides high speed access to main memory
  - L2 is usually 512KB or less – above that is not cost-effective

Cache Types

- Unified cache stores data and instructions in 1 cache
  - Only 1 cache to design and operate
  - Cache is flexible and can balance "allocation" of space to instructions or data to best fit the execution of the program -- higher hit ratio
- Split cache uses 2 caches -- 1 for instructions and 1 for data
  - Must build and manage 2 caches
  - Static allocation of cache sizes
  - Can out perform unified cache in systems that support parallel execution and pipelining (reduces cache contention)

Chapter 5 Review

1. Storage System & Technology
2. Memory Hierarchy
3. Memory Organization and Operations
4. Cache Memories

Chapter 5 ends!