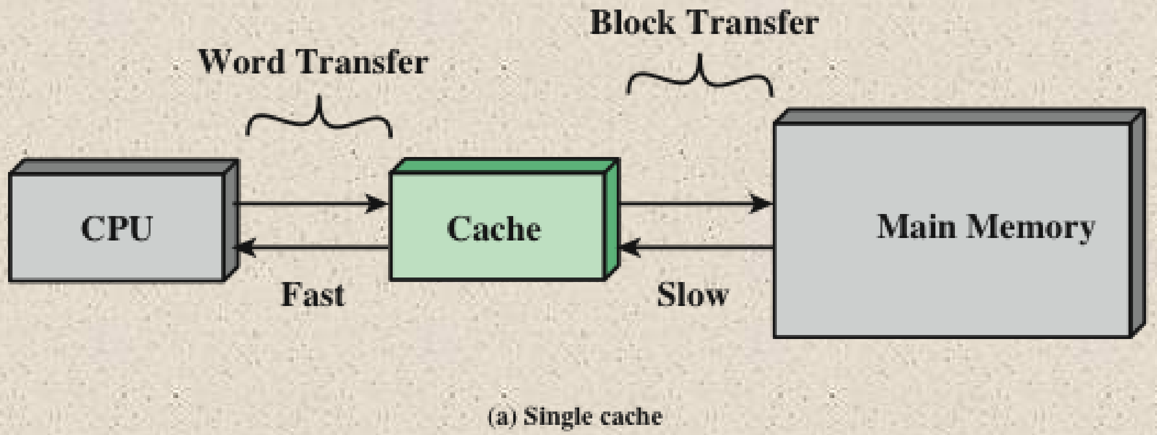
William Stallings

Computer Organization and Architecture
9th Edition

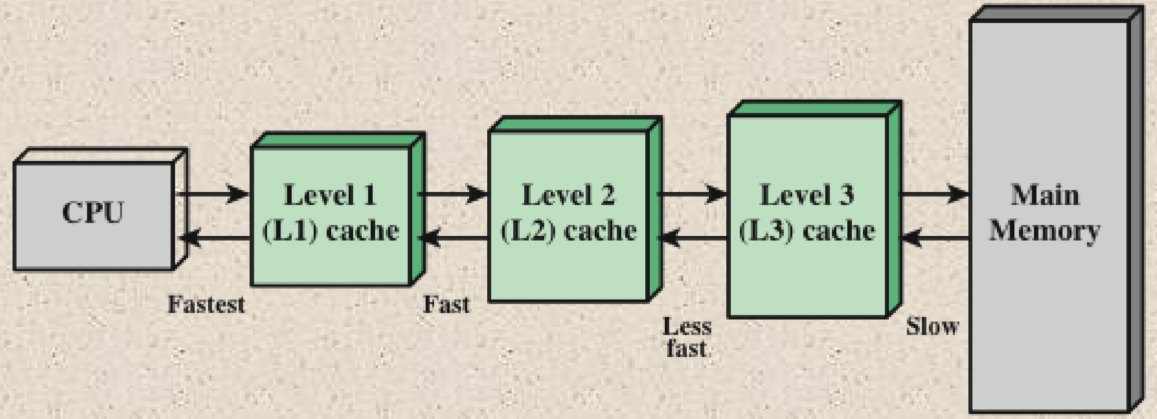
Chapter 4

Cache Memory

Cache and Main Memory



(a) Single cache



(b) Three-level cache organization

Figure 4.3 Cache and Main Memory

Cache/Main Memory Structure

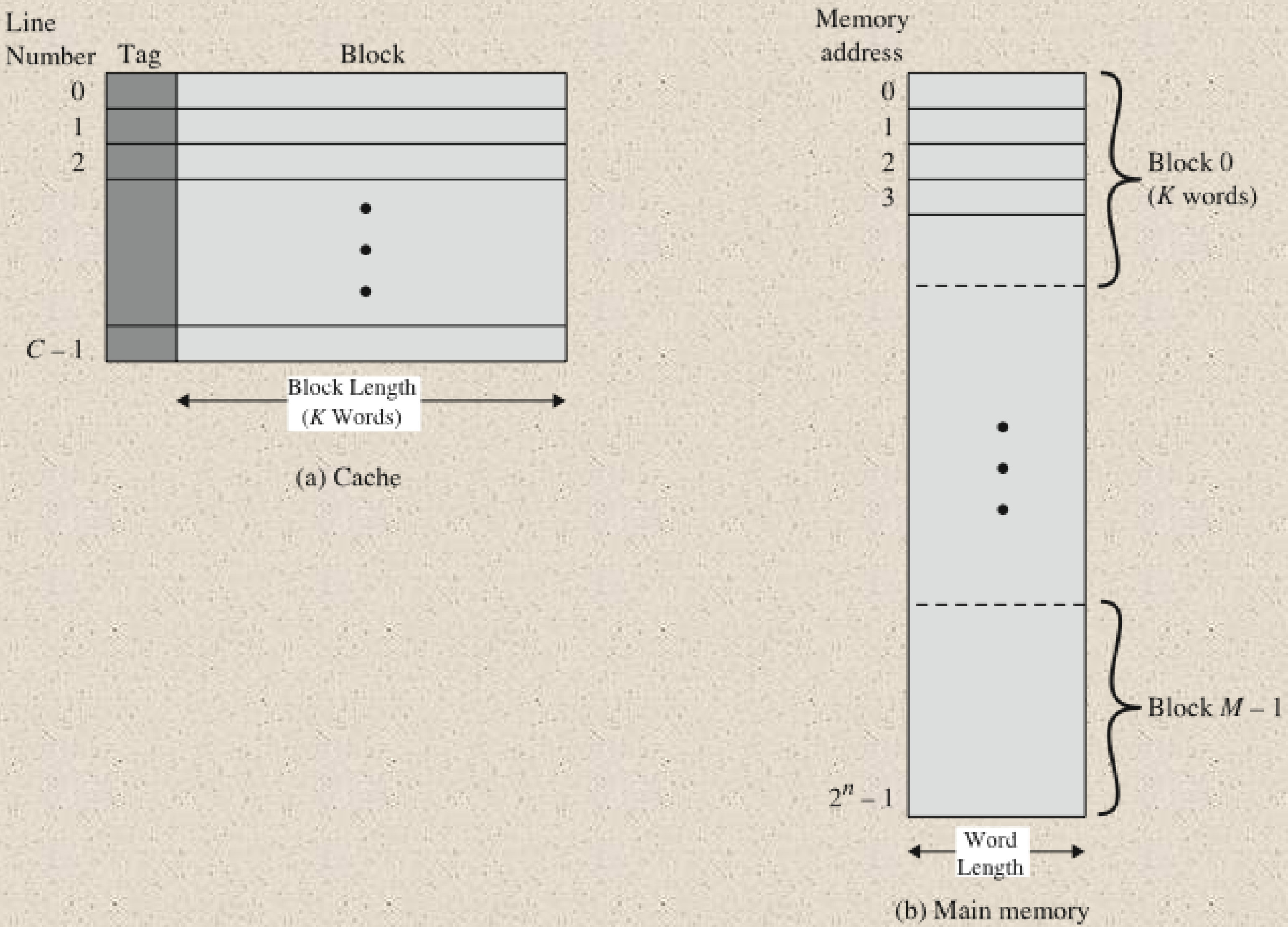


Figure 4.4 Cache/Main-Memory Structure

Cache Read Operation

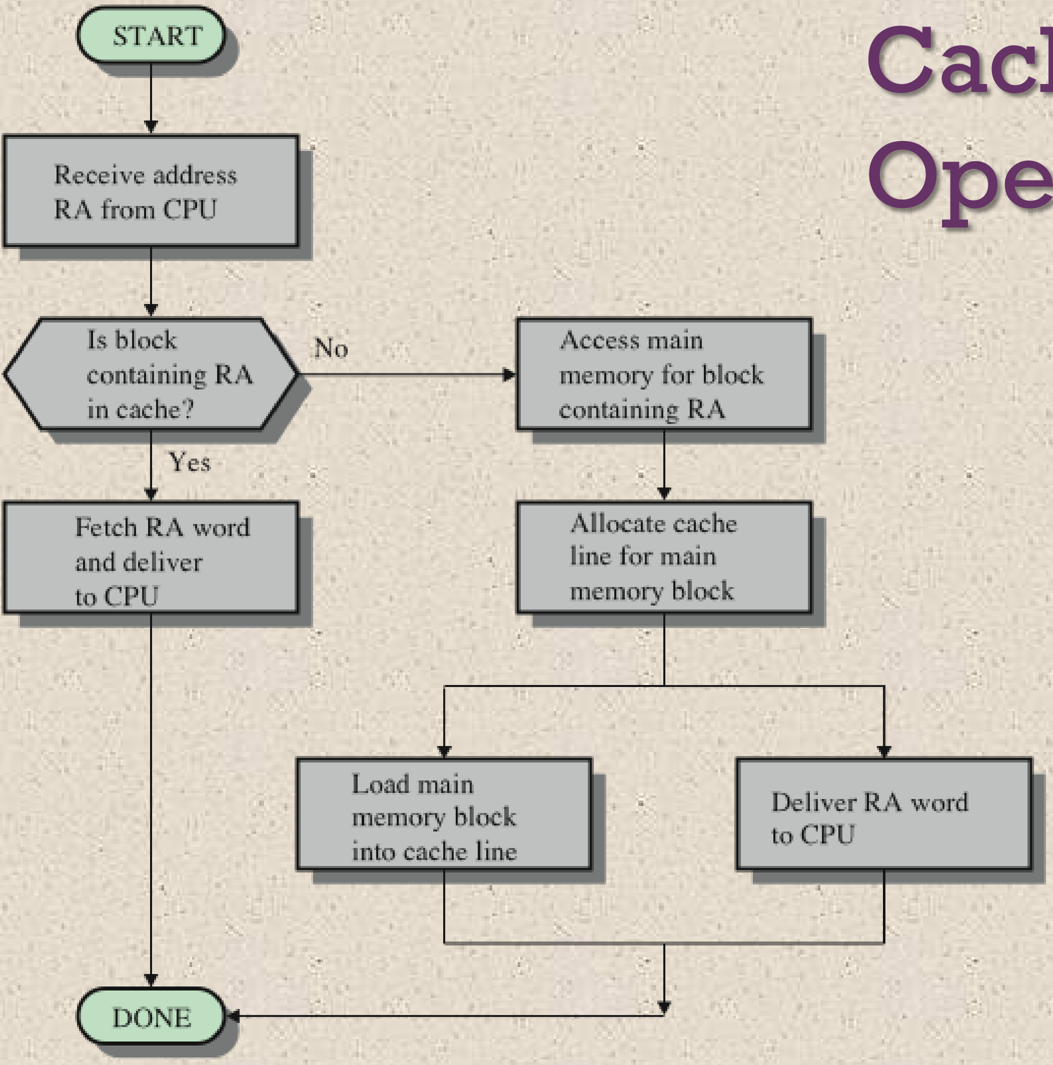

START
Receive address RA from CPU
Is block containing RA in cache?
No
Access main memory for block containing RA
Yes
Fetch RA word and deliver to CPU
Allocate cache line for main memory block
Load main memory block into cache line
Deliver RA word to CPU
DONE

Figure 4.5 Cache Read Operation

Typical Cache Organization

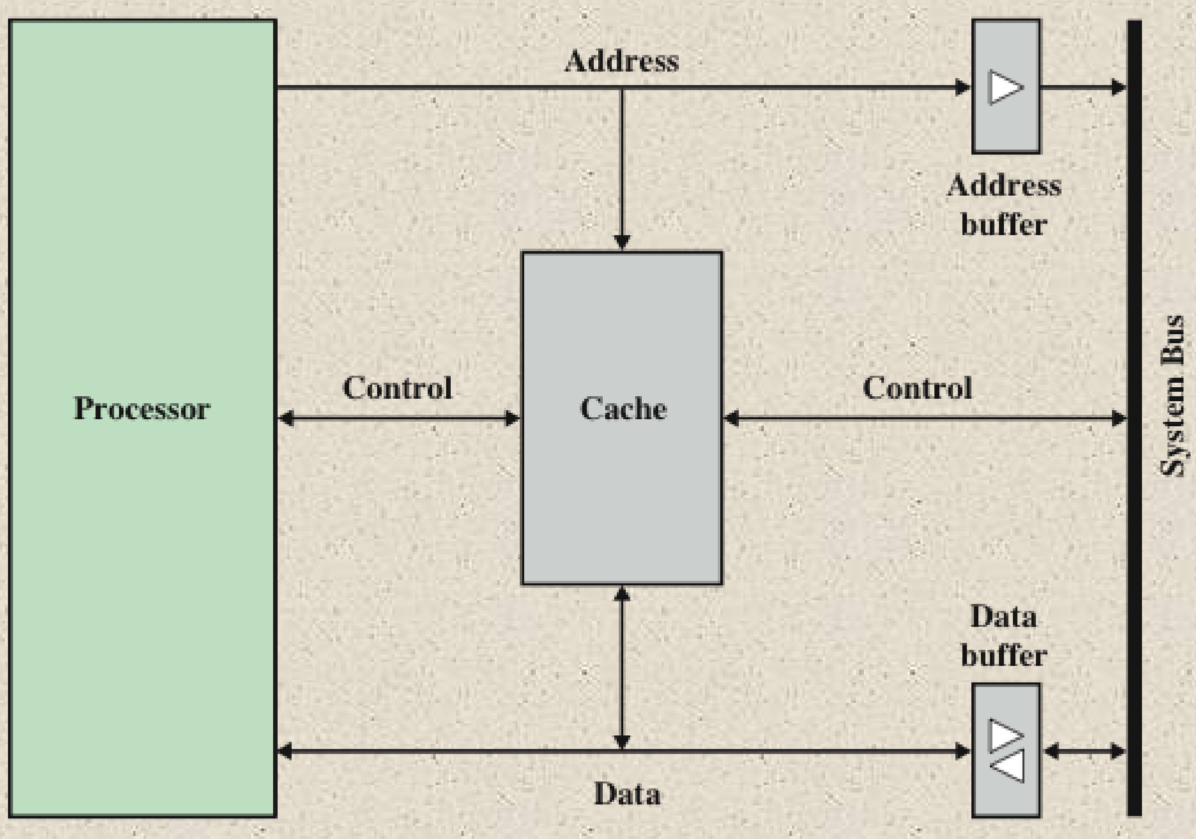
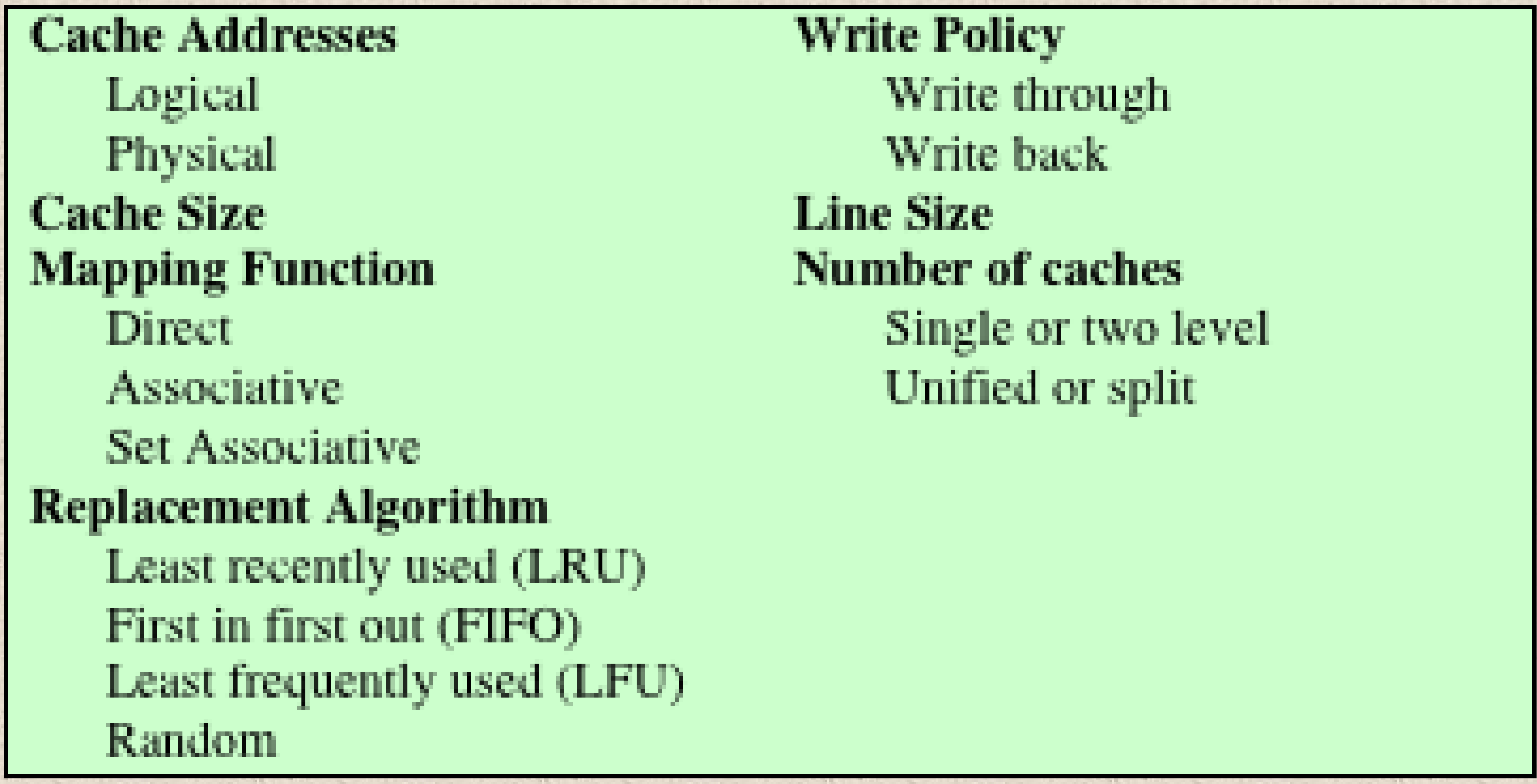


Figure 4.6 Typical Cache Organization

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

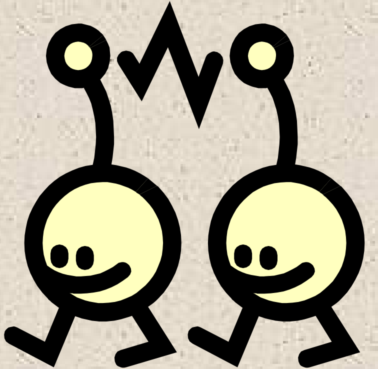
Table 4.2 Elements of Cache Design

+

Cache Addresses

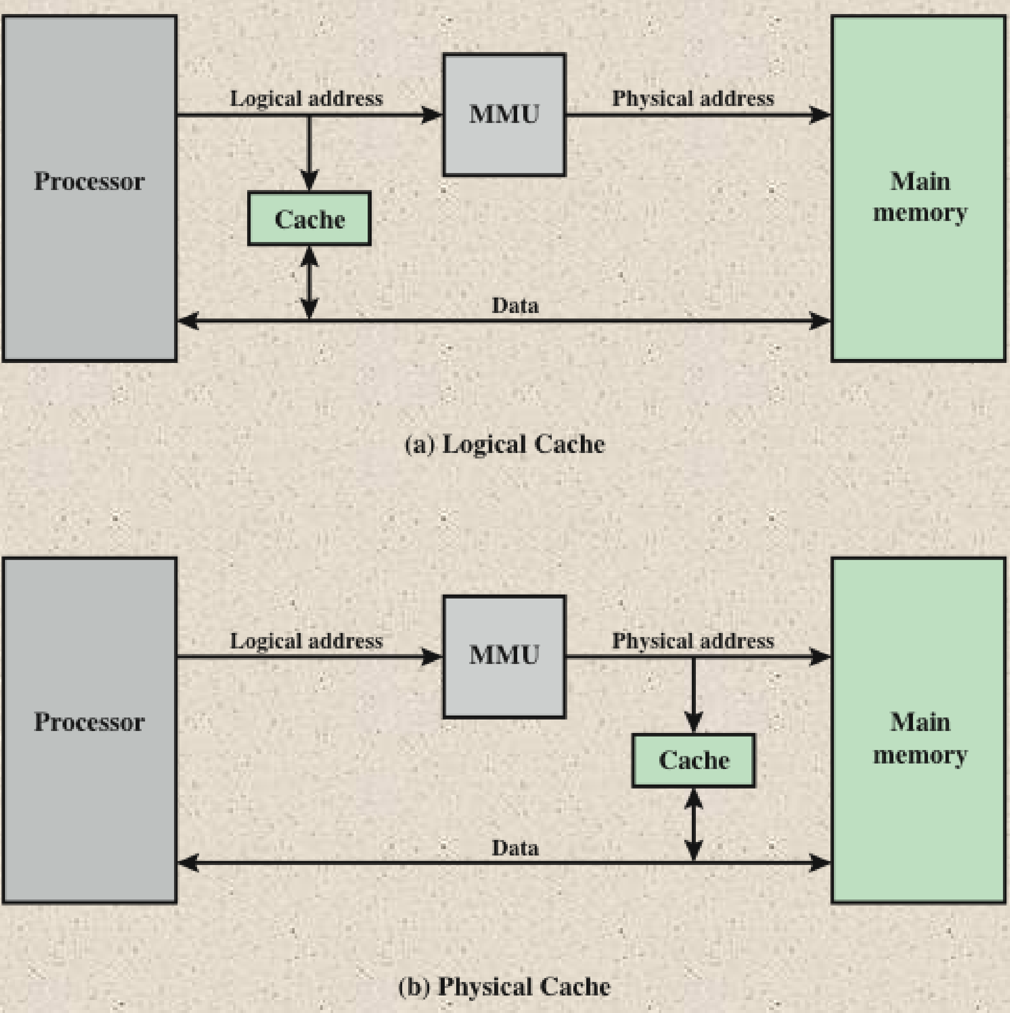
Virtual Memory

- Virtual memory
  - Facility that allows programs to address memory from a logical point of view, without regard to the amount of main memory physically available
  - When used, the address fields of machine instructions contain virtual addresses
  - For reads to and writes from main memory, a hardware memory management unit (MMU) translates each virtual address into a physical address in main memory



+

Logical and Physical Caches



(a) Logical Cache

(b) Physical Cache

Figure 4.7 Logical and Physical Caches

Mapping Function

- Because there are fewer cache lines than main memory blocks, an algorithm is needed for mapping main memory blocks into cache lines

- Three techniques can be used:

Direct

- The simplest technique
- Maps each block of main memory into only one possible cache line

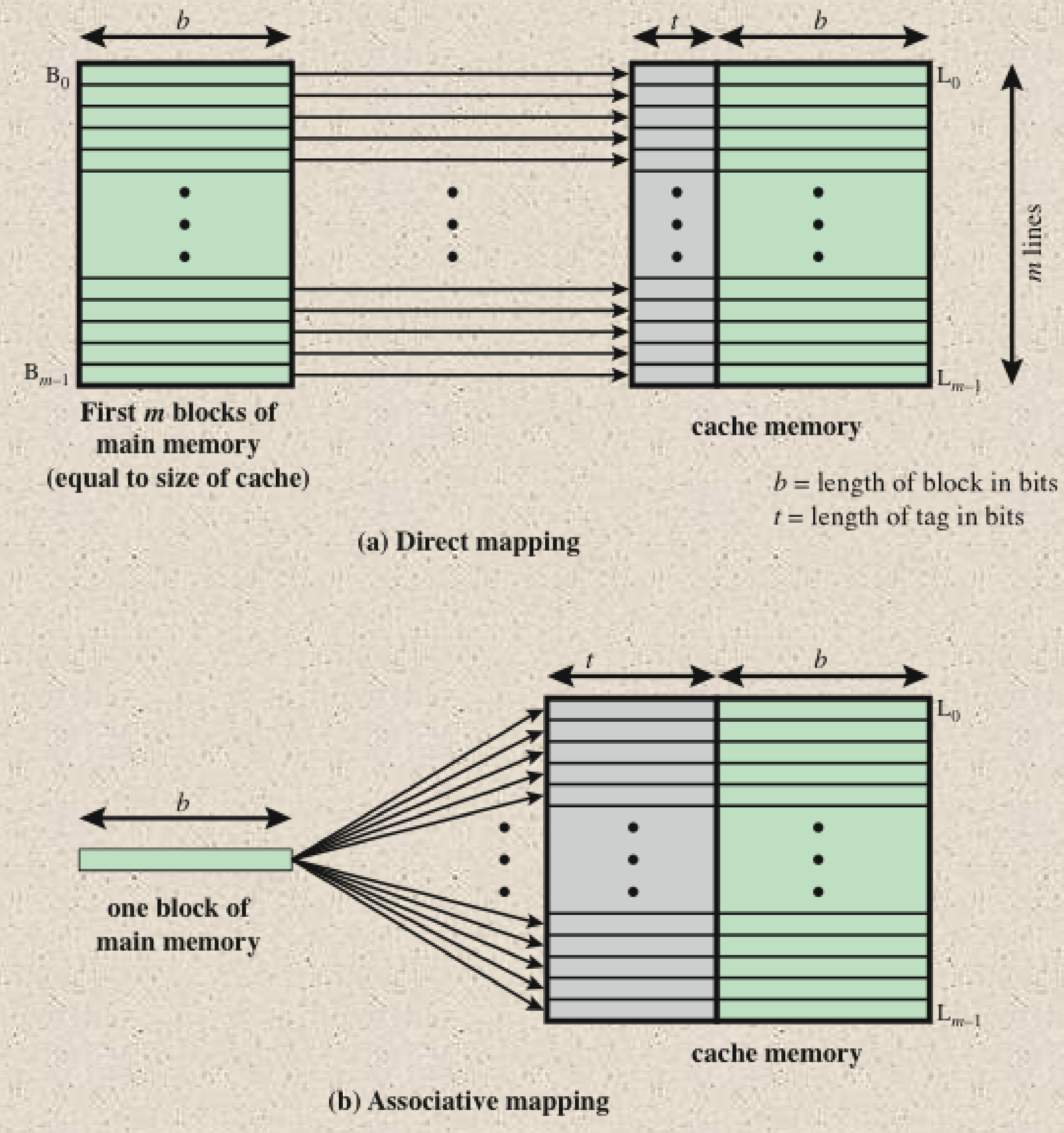
Associative

- Permits each main memory block to be loaded into any line of the cache
- The cache control logic interprets a memory address simply as a Tag and a Word field
- To determine whether a block is in the cache, the cache control logic must simultaneously examine every line's Tag for a match

Set Associative

- A compromise that exhibits the strengths of both the direct and associative approaches while reducing their disadvantages

Direct Mapping


b
B$_0$
B$_{m-1}$
First *m* blocks of main memory (equal to size of cache)
t
b
L$_0$
L$_{m-1}$
m lines
cache memory
b = length of block in bits
t = length of tag in bits
t
b
L$_0$
L$_{m-1}$
b
one block of main memory
cache memory

(a) Direct mapping

(b) Associative mapping

Figure 4.8 Mapping From Main Memory to Cache: Direct and Associative

Direct Mapping Cache Organization

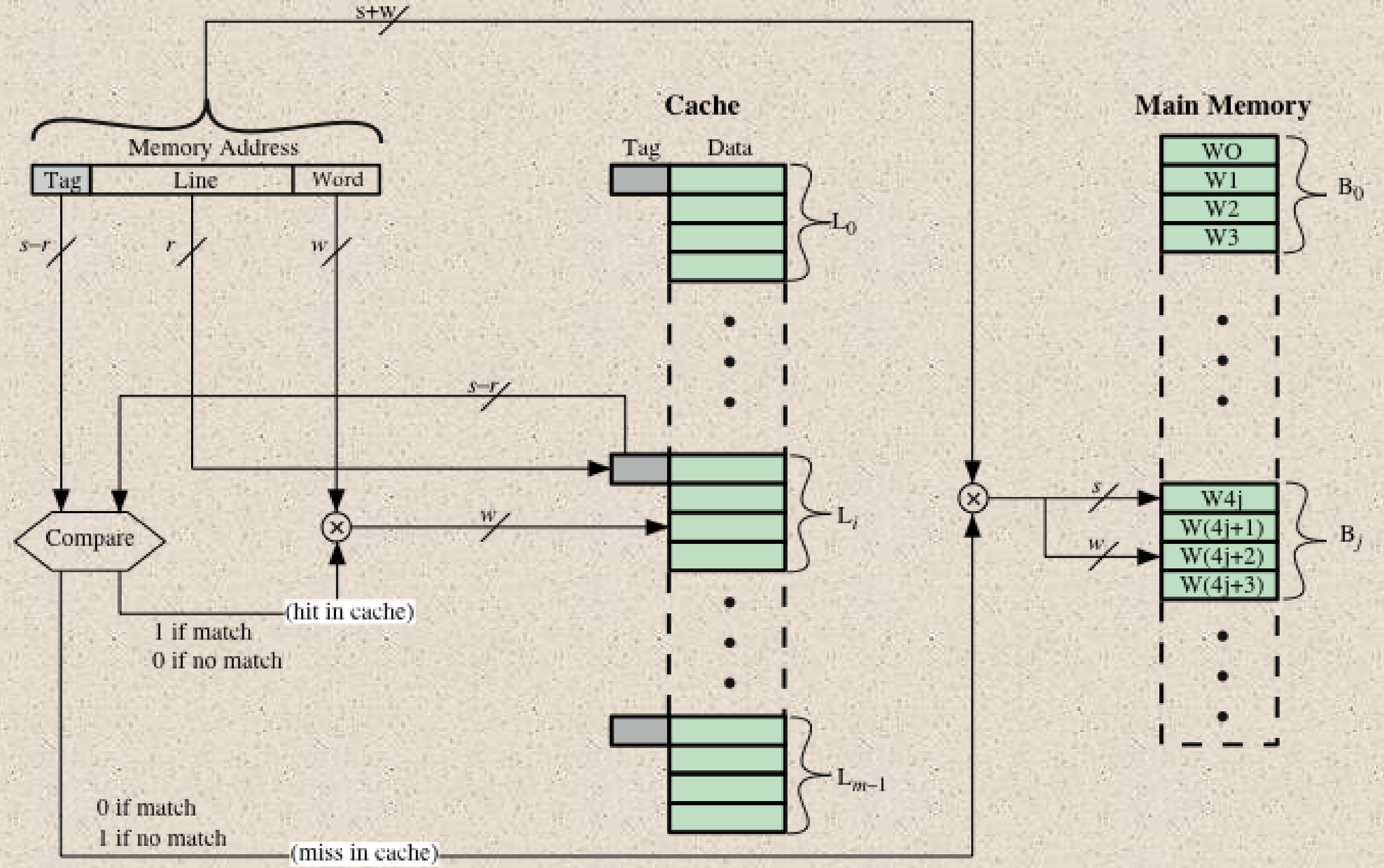


Figure 4.9 Direct-Mapping Cache Organization

+

Direct Mapping Example

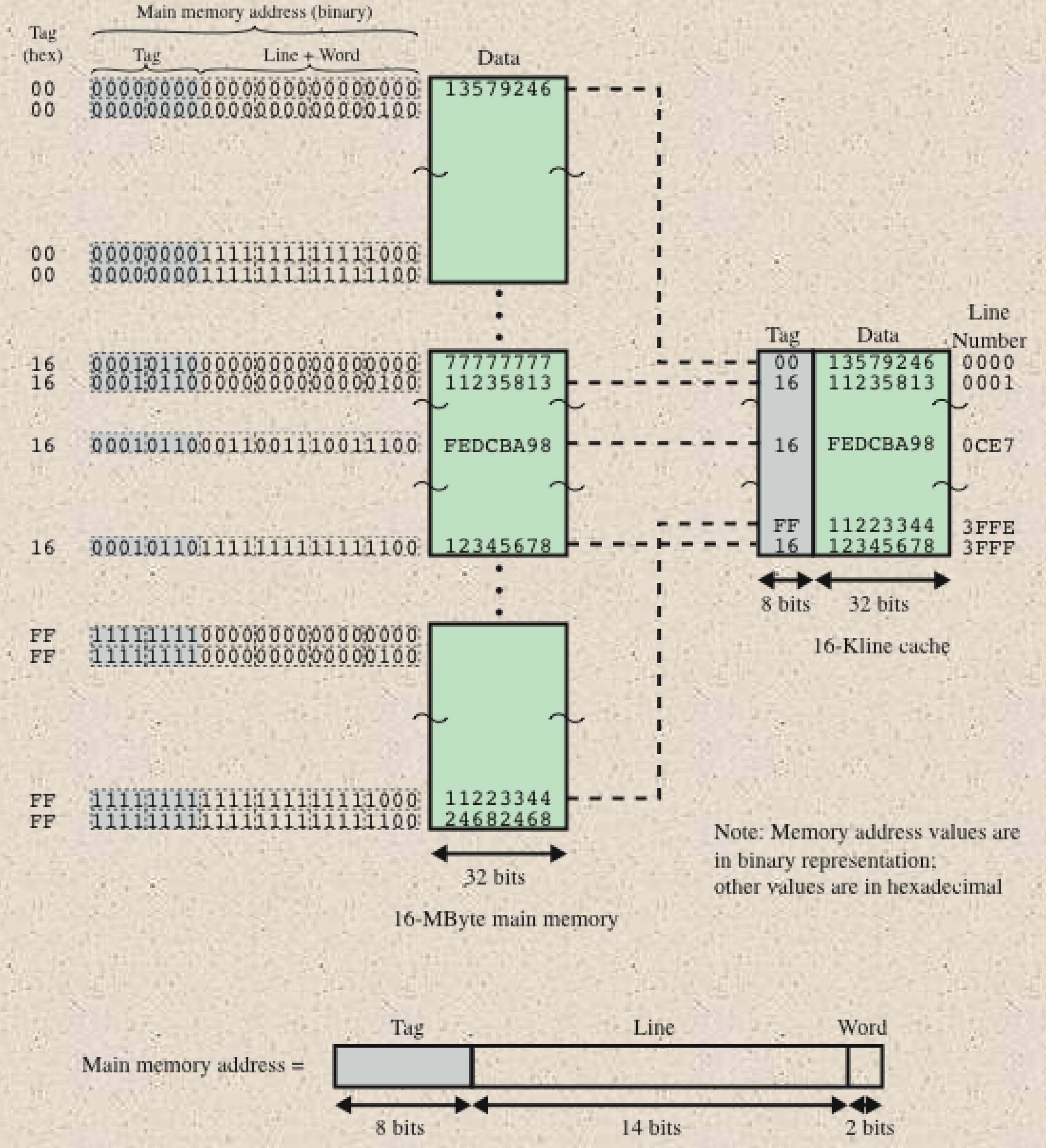


Figure 4.10 Direct Mapping Example

Direct Mapping Summary

- Address length = $(s + w)$ bits
- Number of addressable units = 2^{s+w} words or bytes
- Block size = line size = 2^w words or bytes
- Number of blocks in main memory = $2^{s+w}/2^w = 2^s$
- Number of lines in cache = $m = 2^r$
- Size of tag = $(s – r)$ bits