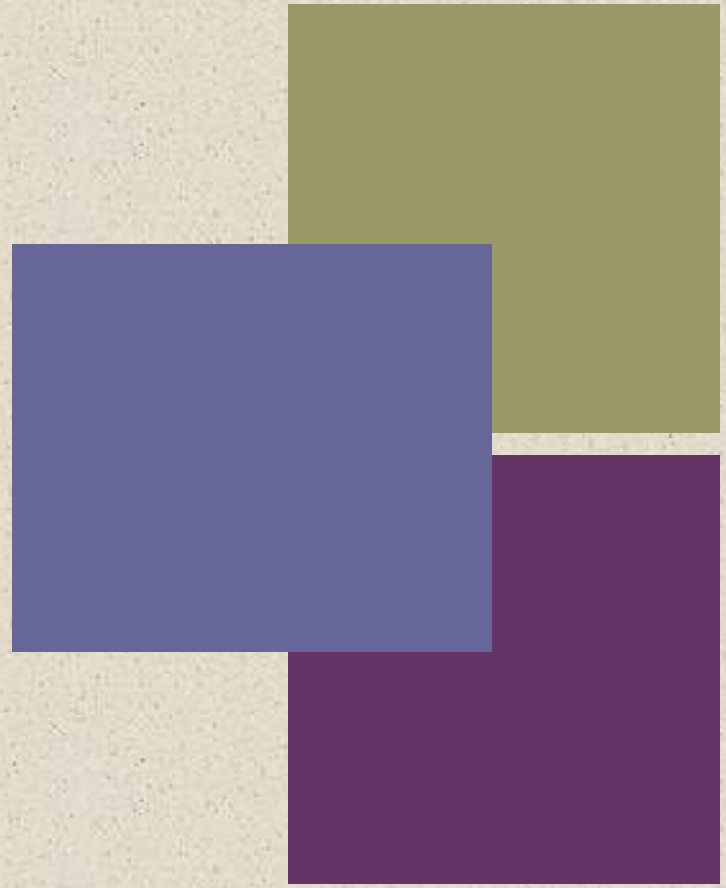


Chapter 4

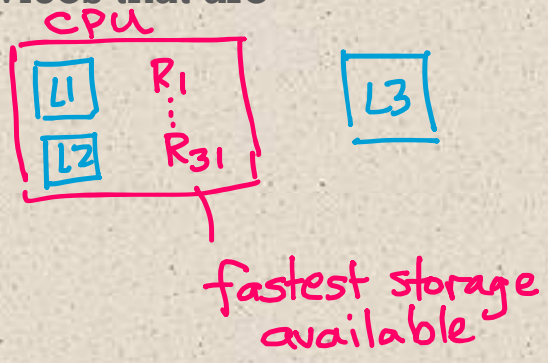
Cache Memory

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

Table 4.1
Key Characteristics of Computer Memory Systems

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



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

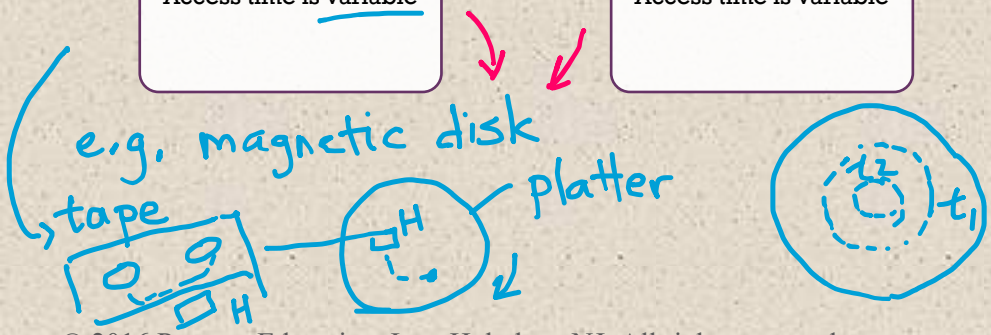
- Involves a shared read-write mechanism
- Individual blocks or records have a unique address based on physical location
- Access time is variable

Random access

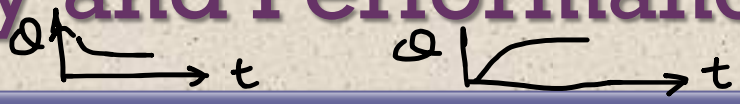
- Each addressable location in memory has a unique, physically wired-in addressing mechanism
- The time to access a given location is independent of the sequence of prior accesses and is constant
- Any location can be selected at random and directly addressed and accessed
- e.g. DRAM
  Main memory and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

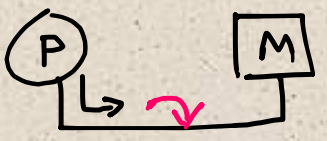
Associative

- 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
- e.g.
  Cache memories may employ associative access



Capacity and Performance:





The two most important characteristics of memory

Three performance parameters are used: (metrics)

1

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 e.g. head above the data

2

Memory cycle time (T)

- Access time plus any additional time required before ~~second~~ next 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

3

Transfer rate (throughput)

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

$1/T$

Memory

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

Memory Hierarchy

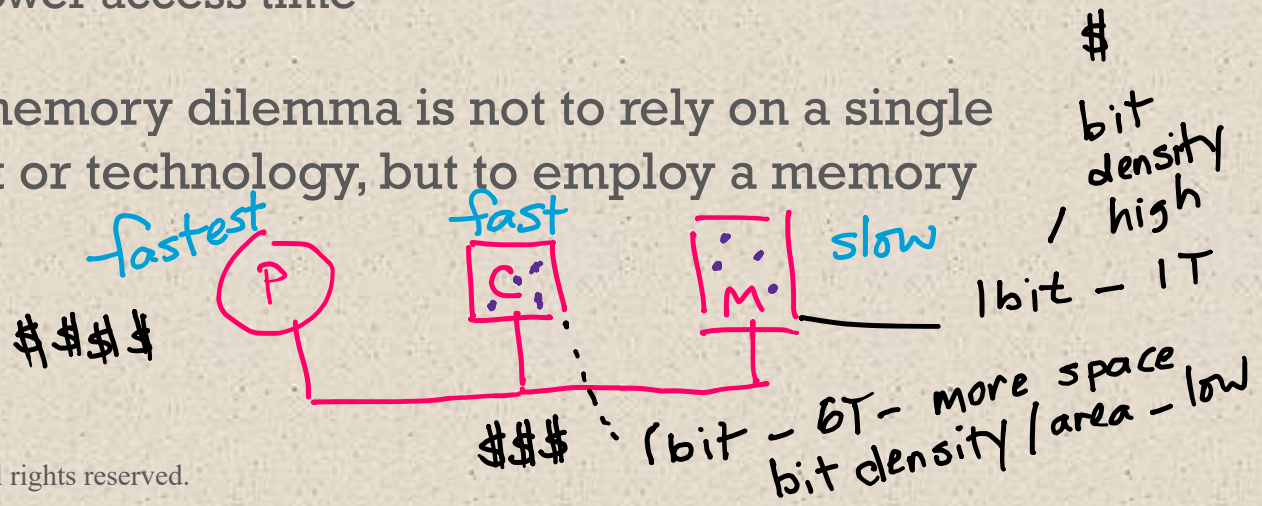
- overall system perf ↑

- 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

① real estate
② $

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



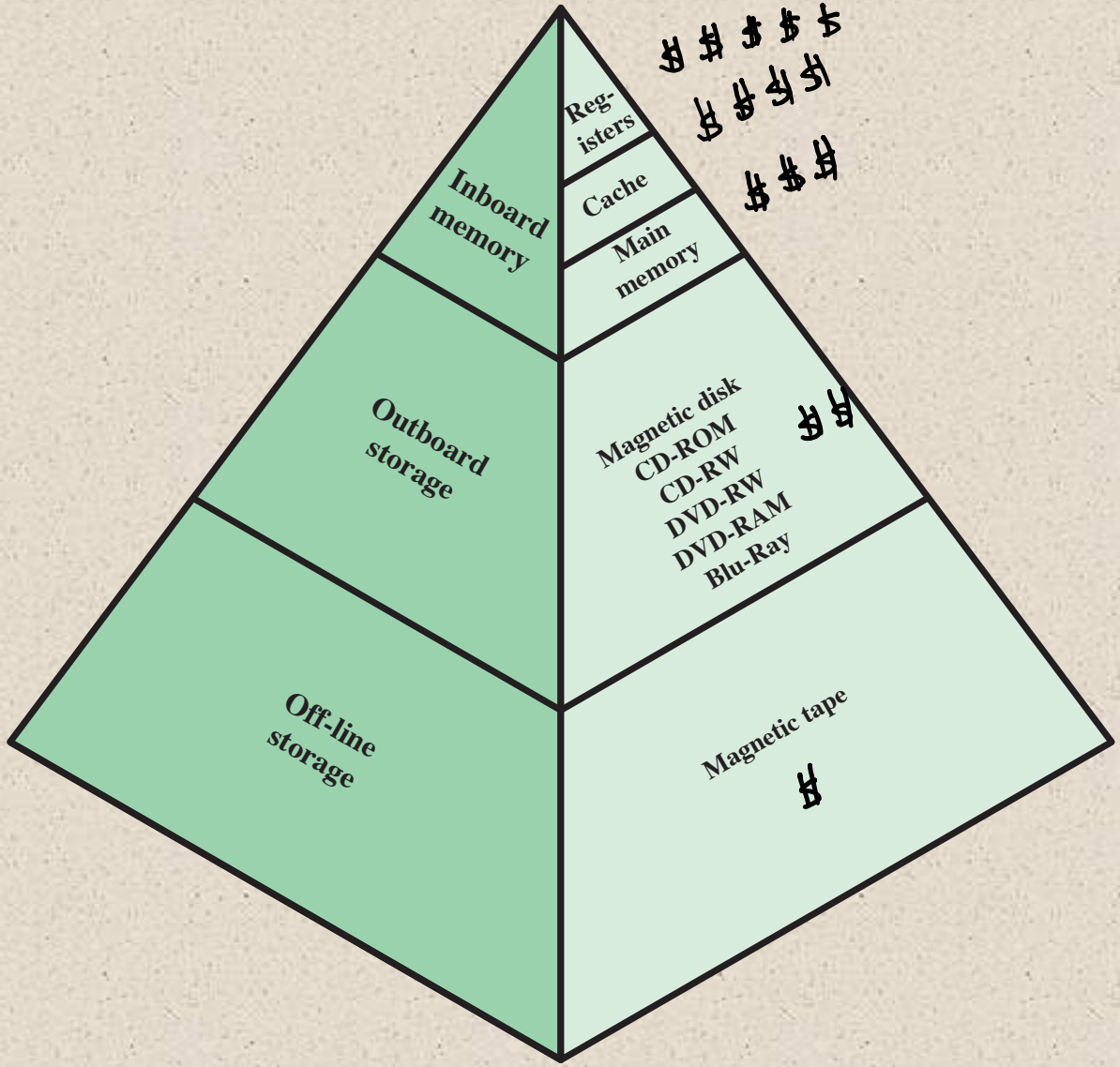


Figure 4.1 The Memory Hierarchy

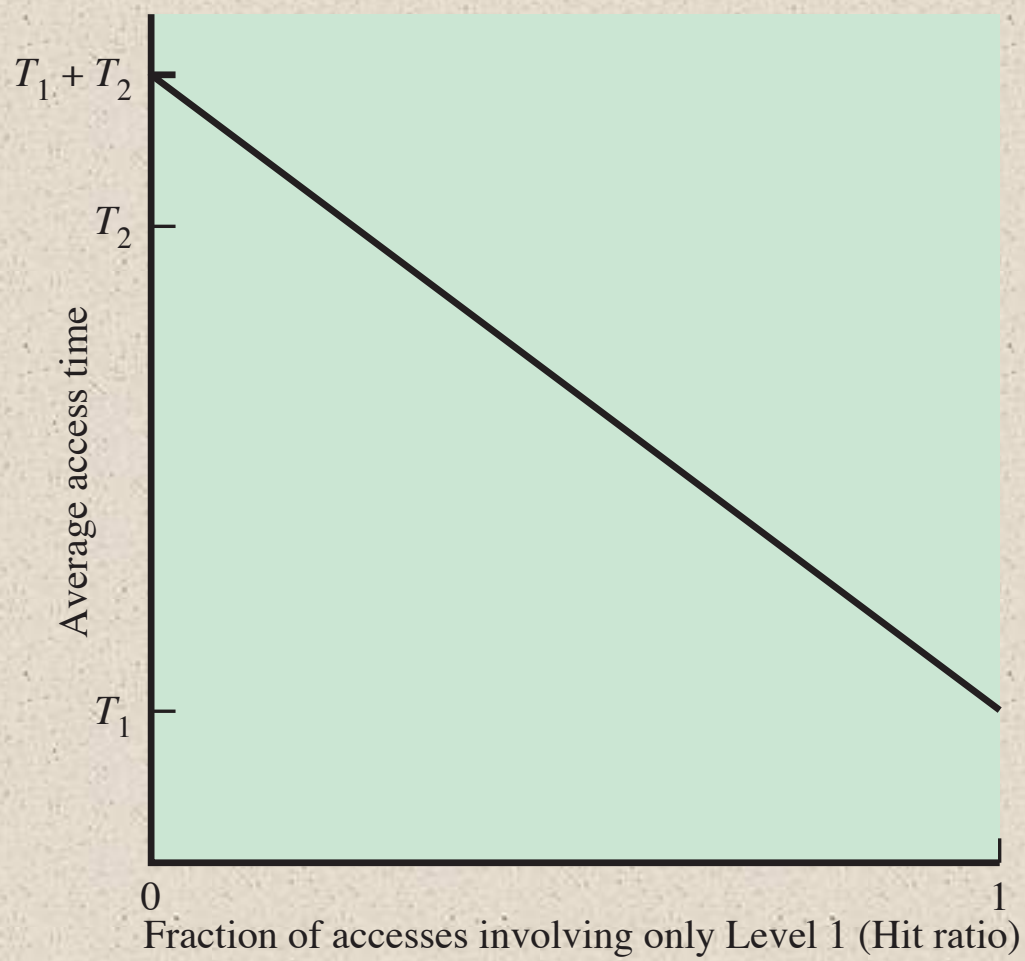


Figure 4.2 Performance of a Simple Two-Level Memory

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

Unsigned numbers

ALU – 2's complement number system

Memory adresses – unsigned " "

e.g. 3-bits

binary	decimal
000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	7

using position / Weighted number system

$1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$

$= 4 + 0 + 1 = 5$

$\cdots 2^3 - 1$

$2^2\ 2^1\ 2^0$

8 unique binary patterns

2-ways

0/1 0/1 0/1

$2 \cdot 2 \cdot 2$

total # of ways of filling 3-bits

$= 2^3$

$(0 .. 2^3 - 1)$

n-bits: $(0 .. 2^n - 1)$

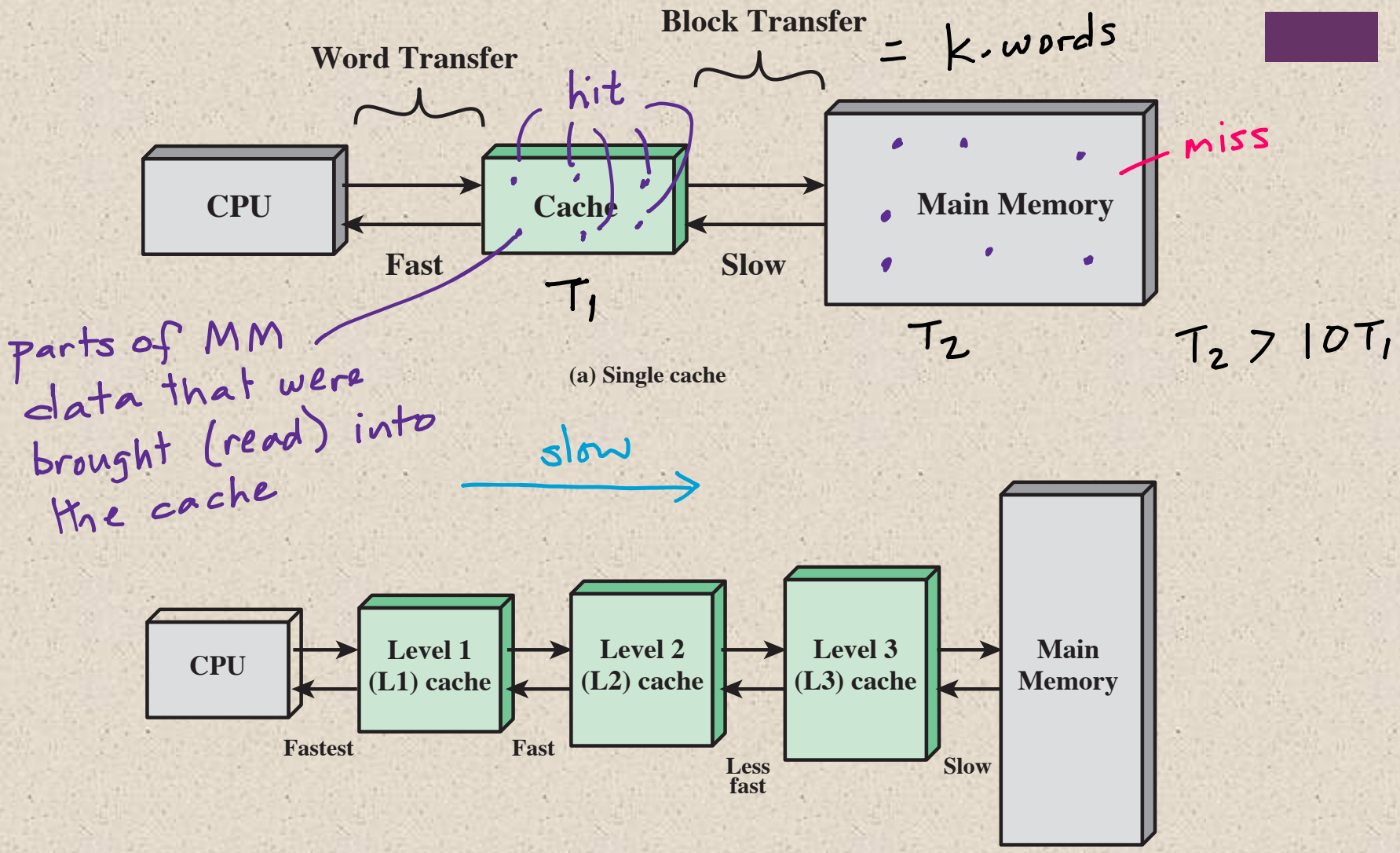


Figure 4.3 Cache and Main Memory

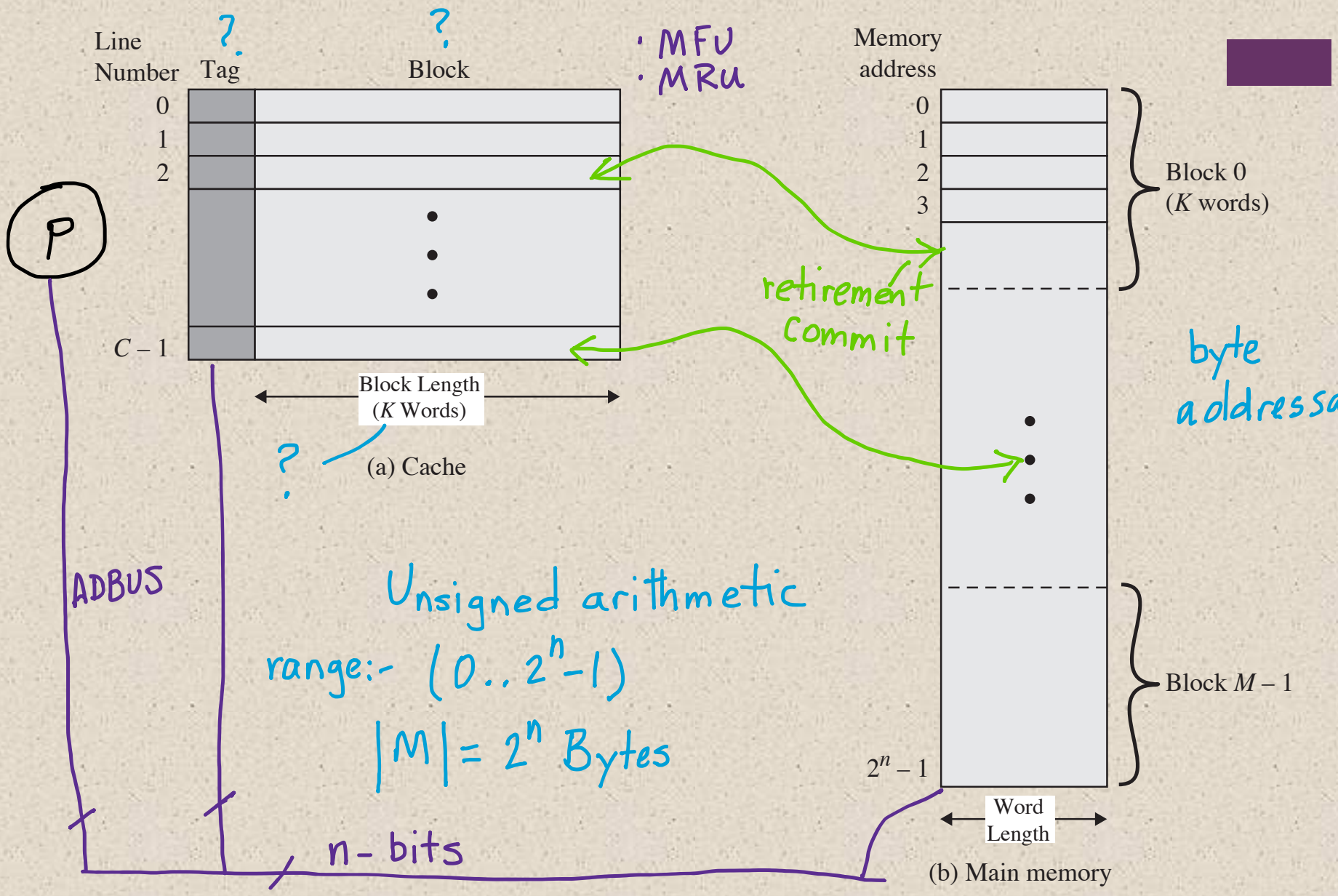
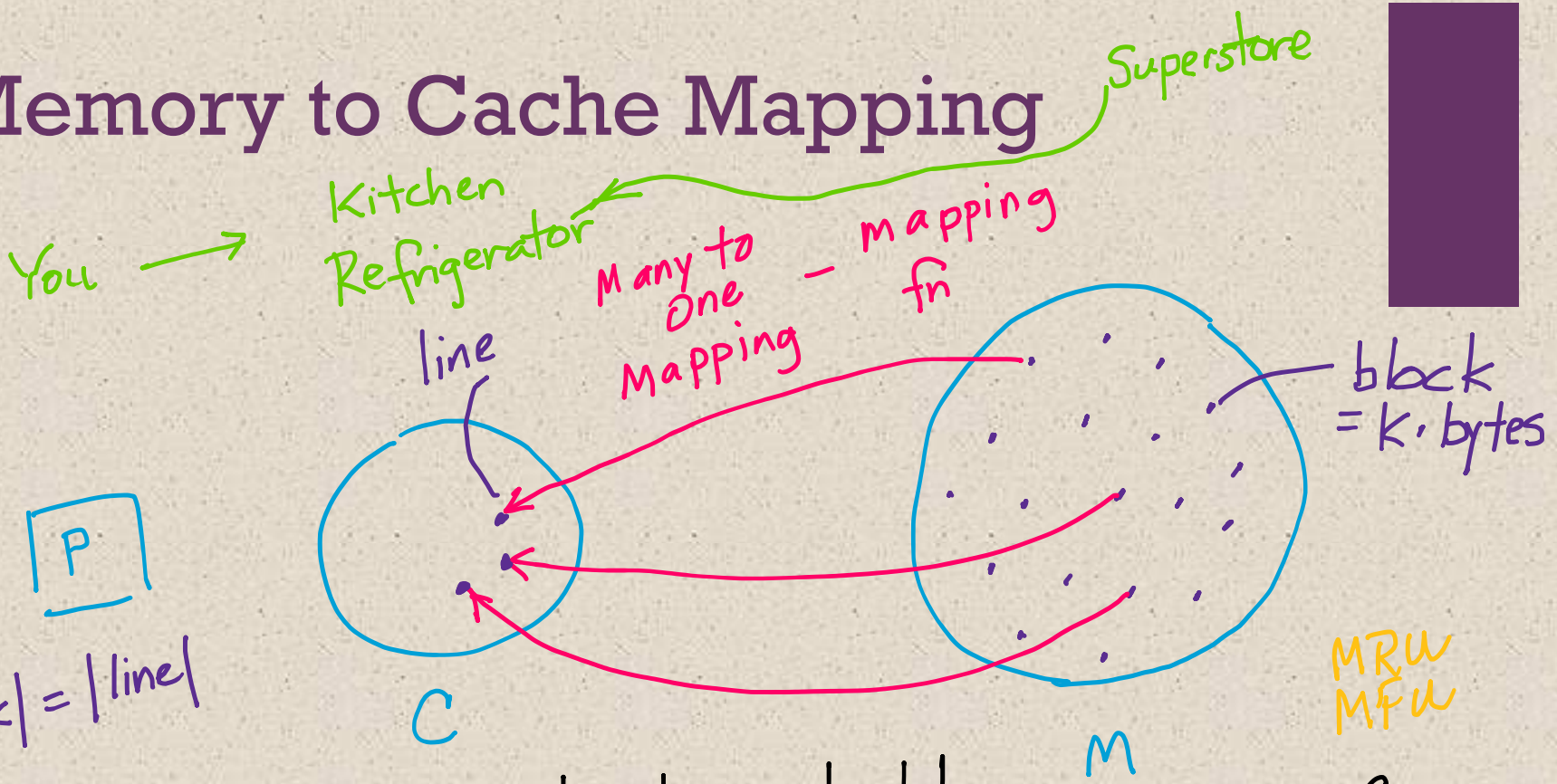


Figure 4.4 Cache/Main-Memory Structure

Memory to Cache Mapping



Q1. which data items should be mapped from MM to C? $g: M \rightarrow C$

Q2. how can 'g' be such that it is evaluated quickly at run-time, upon each access?

Cache Addresses
- Logical
- Physical

Cache Size

Mapping Function
- Direct
- Associative
- Set Associative

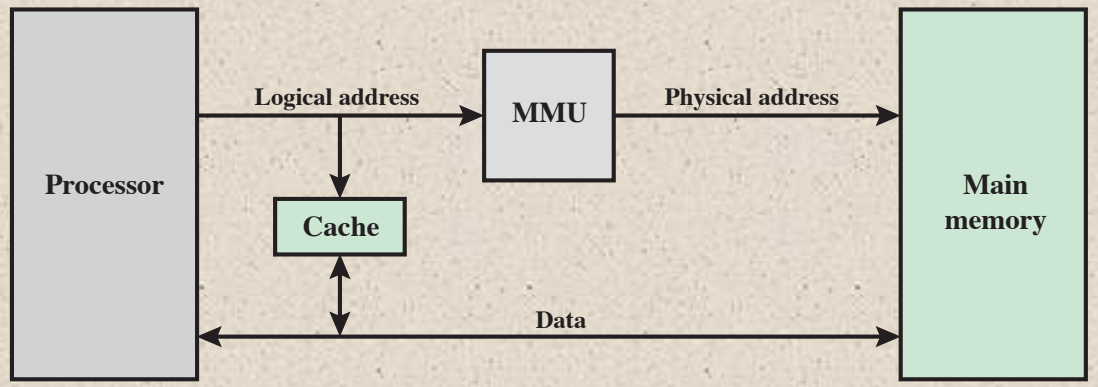
Replacement Algorithm
- Least recently used (LRU)
- First in first out (FIFO)
- Least frequently used (LFU)
- Random

Write Policy
- Write through
- Write back

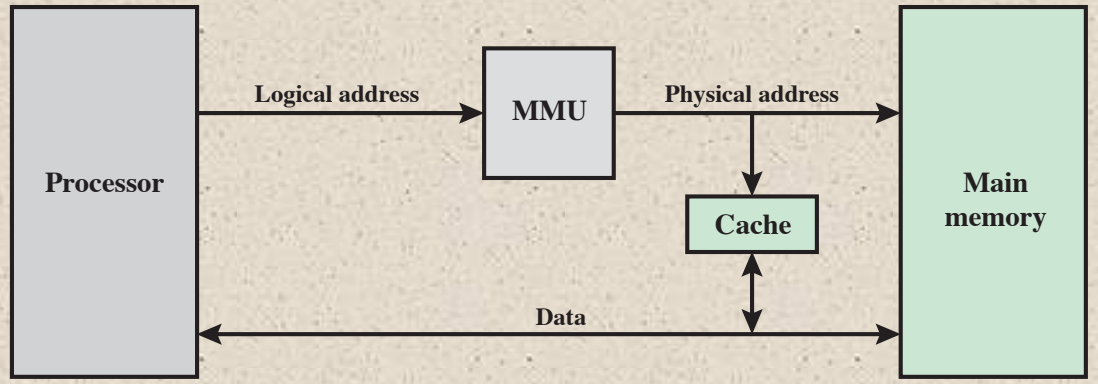
Line Size

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

Table 4.2
Elements of Cache Design



(a) Logical Cache



(b) Physical Cache

Figure 4.7 Logical and Physical Caches

Processor	Type	Year of Introduction	L1 Cache[a]	L2 cache	L3 Cache
IBM 360/85	Mainframe	1968	16 to 32 kB	—	—
PDP-11/70	Minicomputer	1975	1 kB	—	—
VAX 11/780	Minicomputer	1978	16 kB	—	—
IBM 3033	Mainframe	1978	64 kB	—	—
IBM 3090	Mainframe	1985	128 to 256 kB	—	—
Intel 80486	PC	1989	8 kB	—	—
Pentium	PC	1993	8 kB/8 kB	256 to 512 KB	—
PowerPC 601	PC	1993	32 kB	—	—
PowerPC 620	PC	1996	32 kB/32 kB	—	—
PowerPC G4	PC/server	1999	32 kB/32 kB	256 KB to 1 MB	2 MB
IBM S/390 G6	Mainframe	1999	256 kB	8 MB	—
Pentium 4	PC/server	2000	8 kB/8 kB	256 KB	—
IBM SP	High-end server/ supercomputer	2000	64 kB/32 kB	8 MB	—
CRAY MTAb	Supercomputer	2000	8 kB	2 MB	—
Itanium	PC/server	2001	16 kB/16 kB	96 KB	4 MB
Itanium 2	PC/server	2002	32 kB	256 KB	6 MB
IBM POWER5	High-end server	2003	64 kB	1.9 MB	36 MB
CRAY XD-1	Supercomputer	2004	64 kB/64 kB	1MB	—
IBM POWER6	PC/server	2007	64 kB/64 kB	4 MB	32 MB
IBM z10	Mainframe	2008	64 kB/128 kB	3 MB	24-48 MB
Intel Core i7 EE 990	Workstaton/ server	2011	6 × 32 kB/32 kB	1.5 MB	12 MB
IBM zEnterprise 196	Mainframe/ Server	2011	24 × 64 kB/ 128 kB	24 × 1.5 MB	24 MB L3 192 MB L4

Table 4.3

Cache Sizes of Some Processors

[a] Two values separated by a slash refer to instruction and data caches.

[b] Both caches are instruction only; no data caches.

(Table can be found on page 134 in the textbook.)

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	Associative	Set Associative
• The simplest technique • Maps each block of main memory into only one possible cache line	• 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	• A compromise that exhibits the strengths of both the direct and associative approaches while reducing their disadvantages

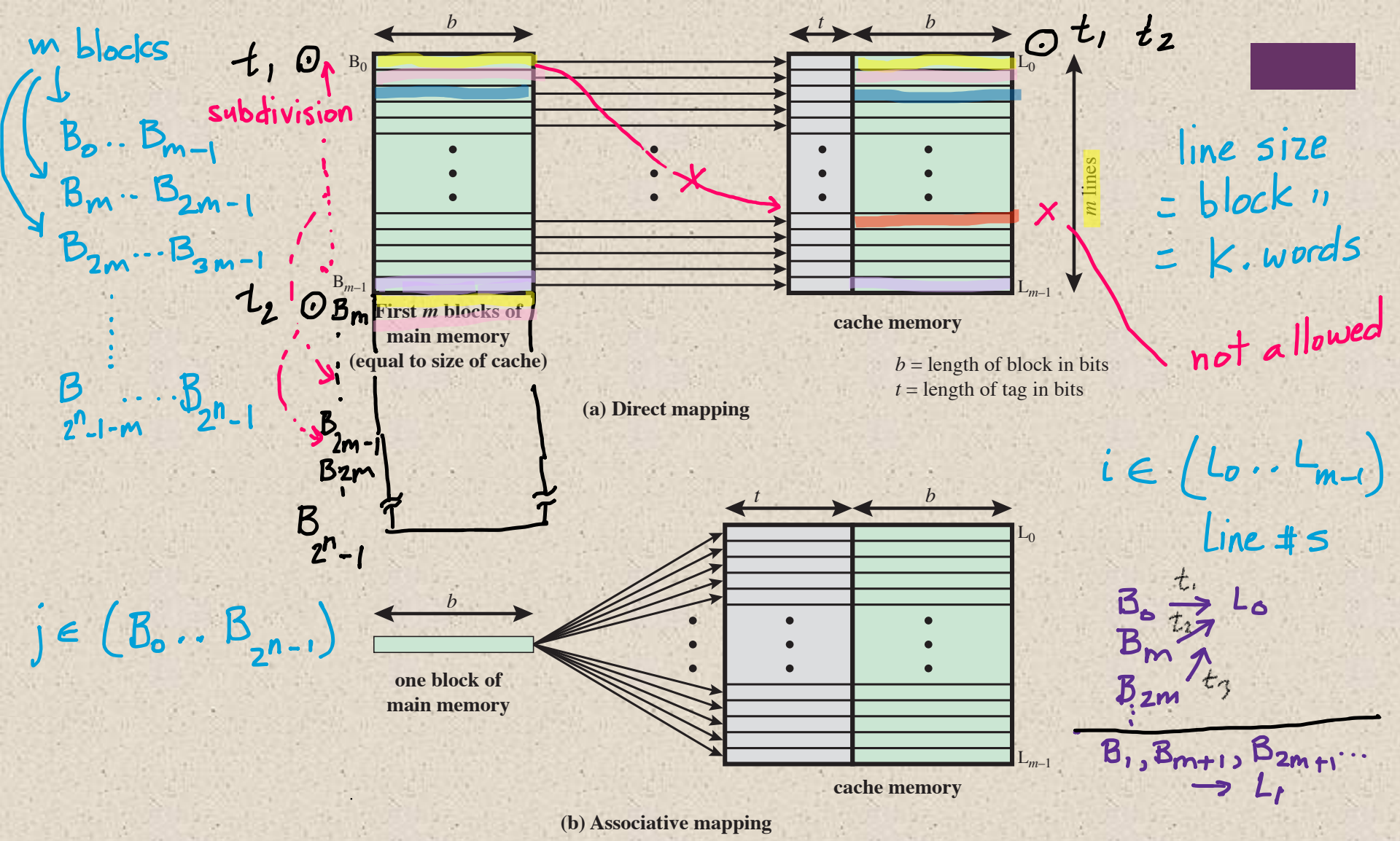


Figure 4.8 Mapping From Main Memory to Cache: Direct and Associative

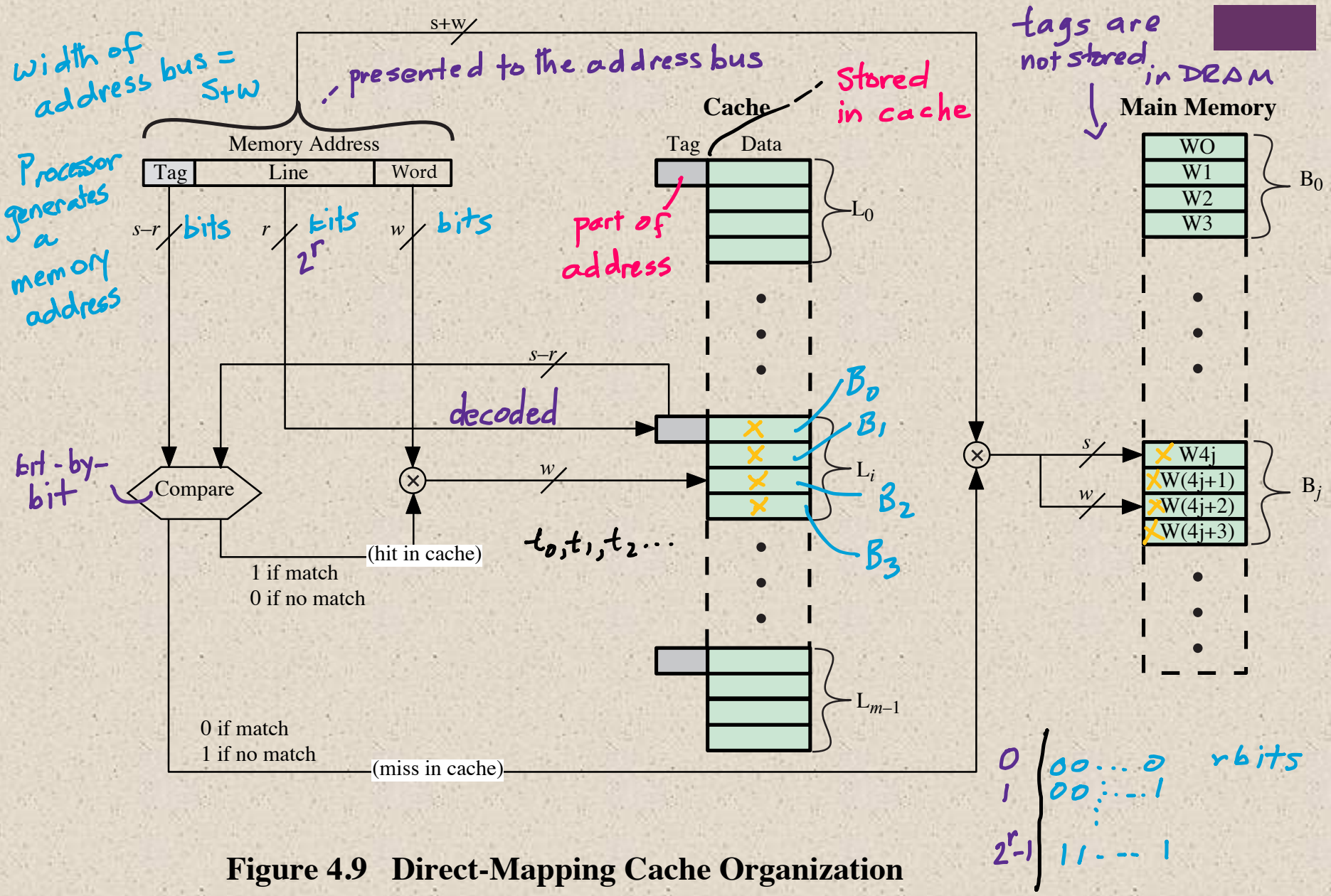


Figure 4.9 Direct-Mapping Cache Organization

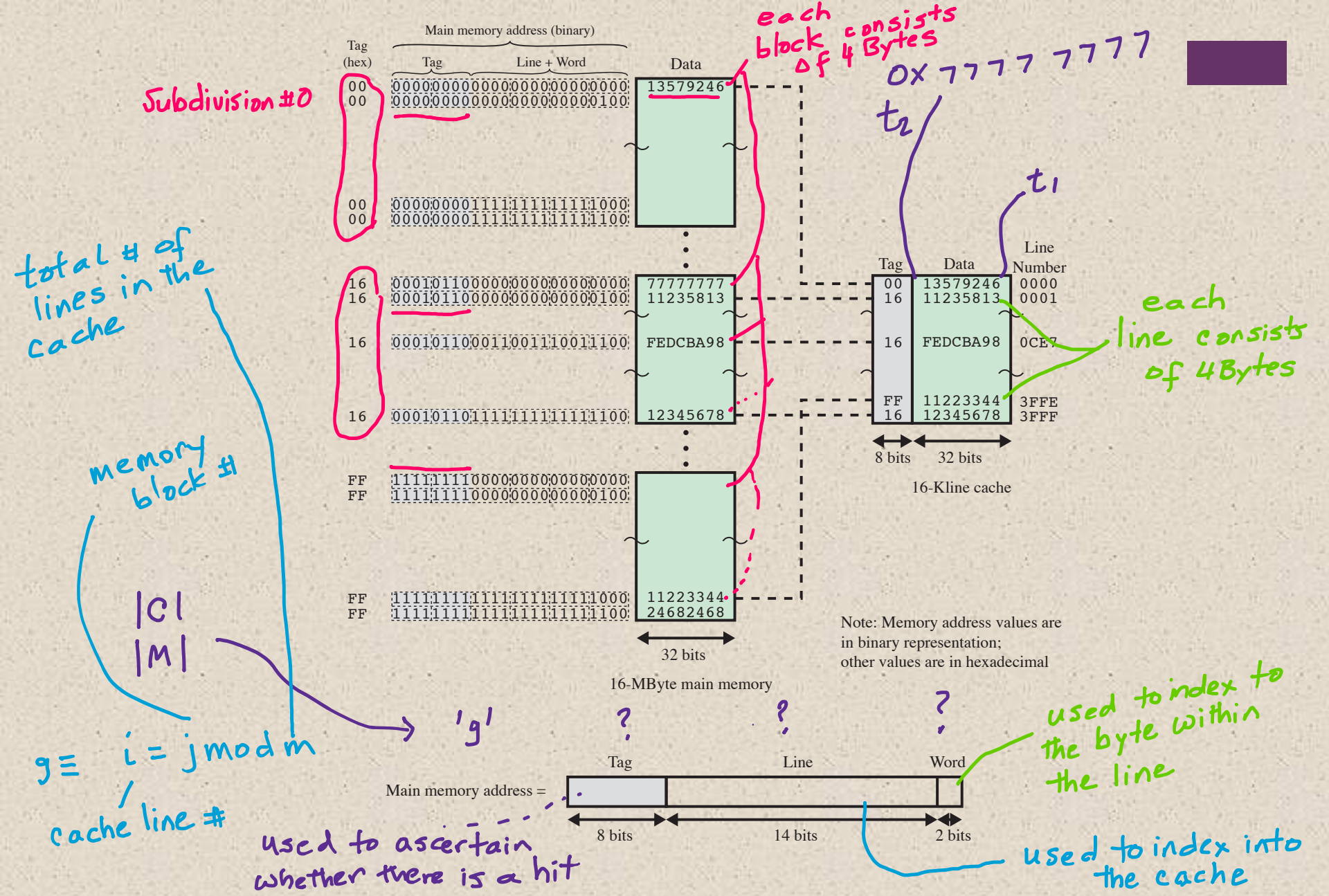
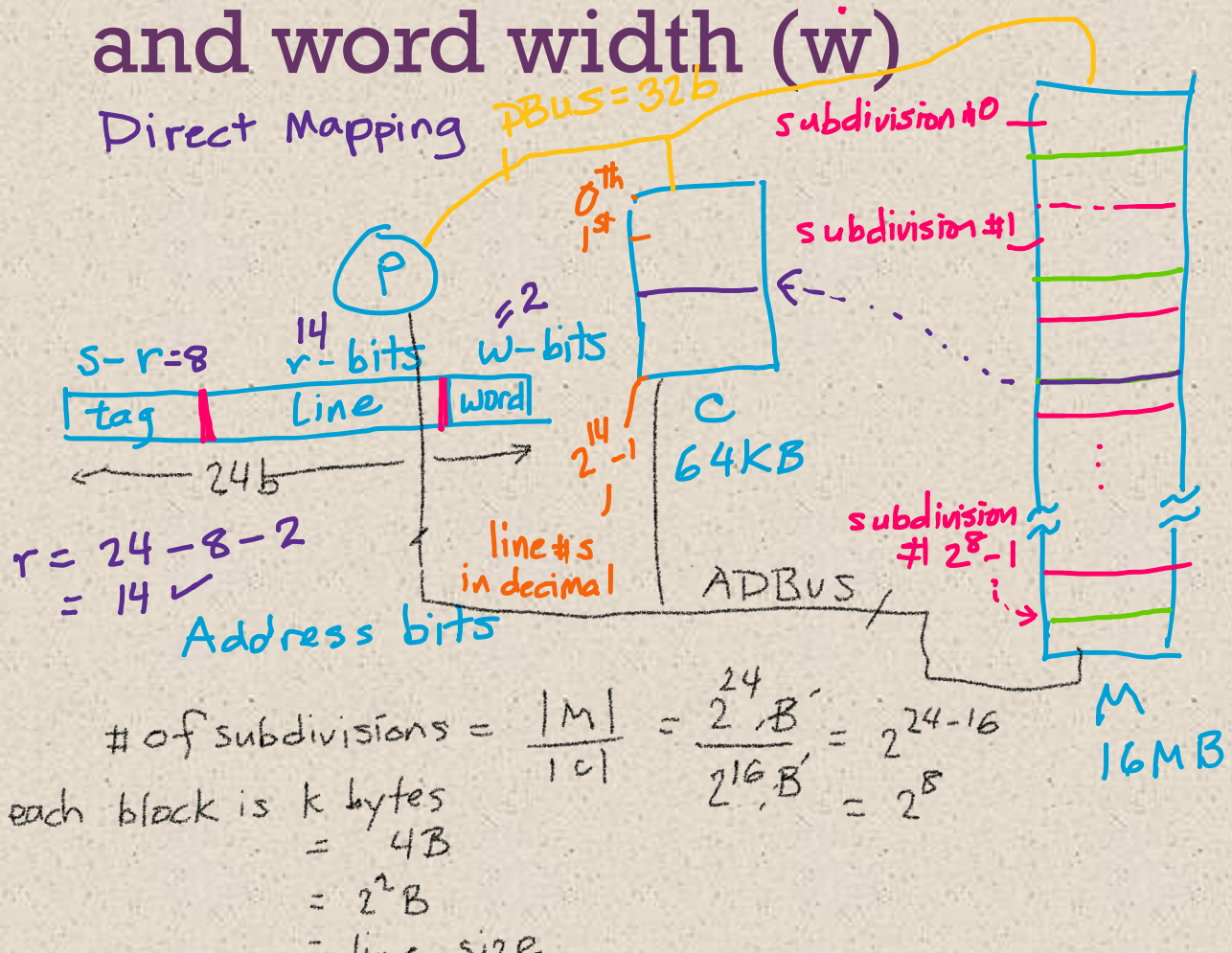


Figure 4.10 Direct Mapping Example

Find line width(r), tag width (s-r) and word width (w)



r = 24 - 8 - 2
= 14 ✓
Address bits

# of subdivisions = $\frac{|M|}{|C|} = \frac{2^{24} B}{2^{16} B} = 2^{24-16} = 2^8$

each block is k bytes
= 4B
= 2^2 B
= line size

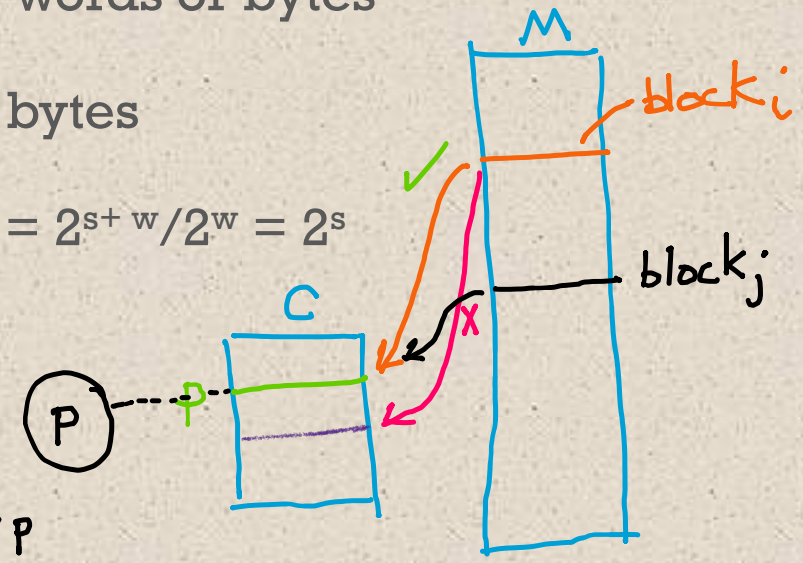
$|M| = 16MB$
$= 2^4 \cdot 2^{20} B$
$= 2^{24} B$
$0 .. 2^{24} - 1$ addresses

$|C| = 64KB$
$= 16.4 KB$
$= 2^4 \cdot 2^2 KB$
$= 2^6 \cdot 2^{10} B$
$= 2^{16} B$
$= \frac{2^{16} B}{2^2 B}$ lines
$= 2^{14}$ lines ✓

Direct Mapping Summary

observation { A block can go to only one unique position in the cache. – restrictive

- 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



$v_1, v_2, v_3, \ldots$ → mod m

slow computer clown

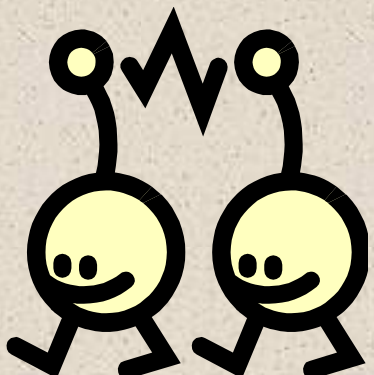
block$_i$ → t_1 → line p

block$_j$ → t_2

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



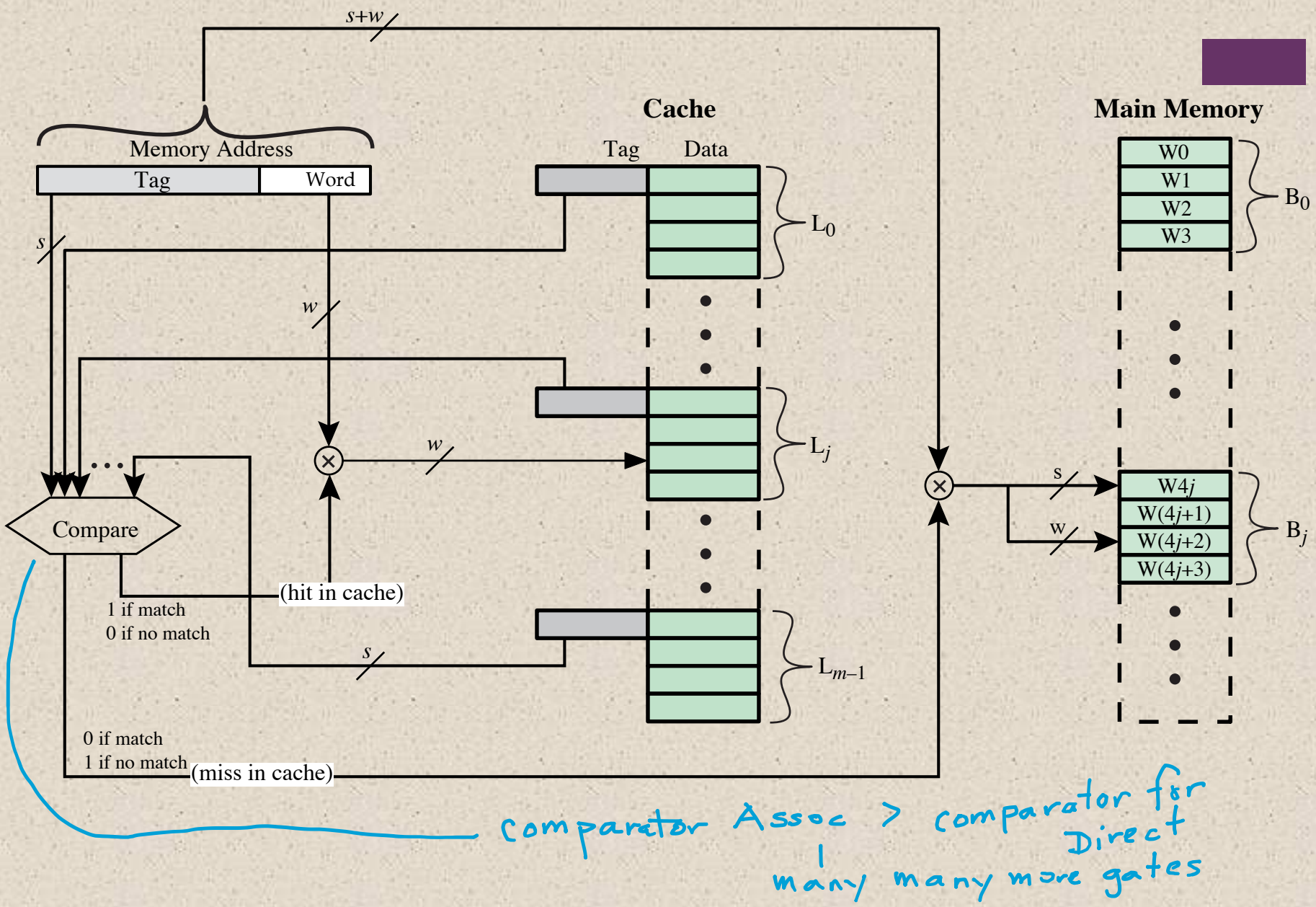


Figure 4.11 Fully Associative Cache Organization

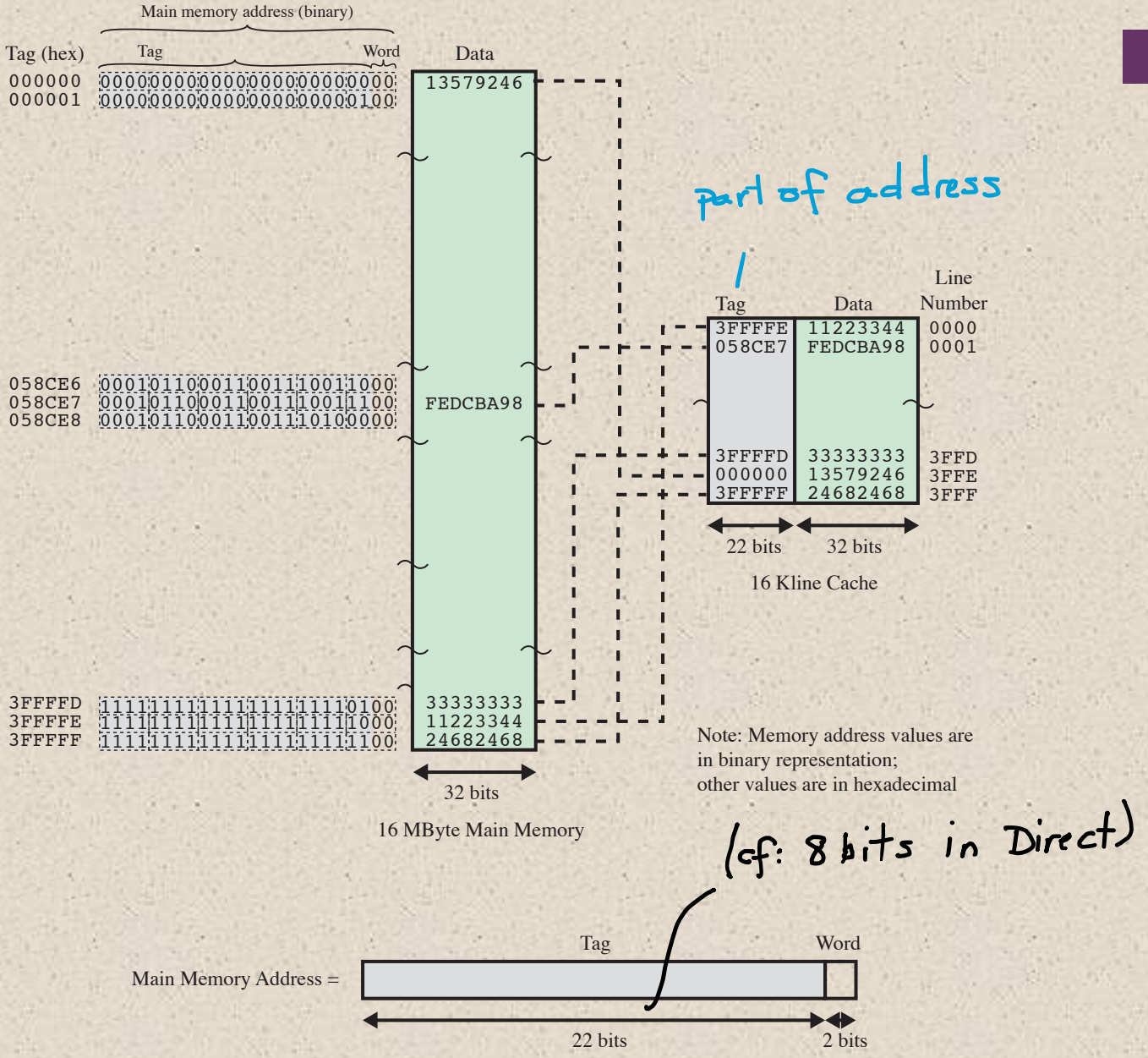


Figure 4.12 Associative Mapping Example

Associative Mapping Summary

Observation:- too many gates for tag comparison

- 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 = undetermined
- Size of tag = s bits

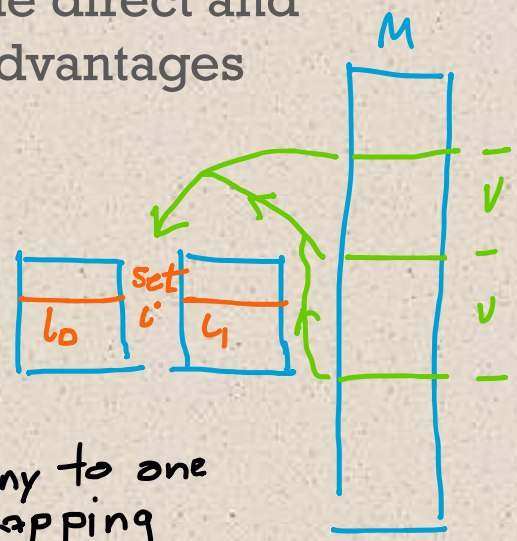
Set Associative Mapping

$i = j \bmod v$

i: set #, j: block #, v: # of sets

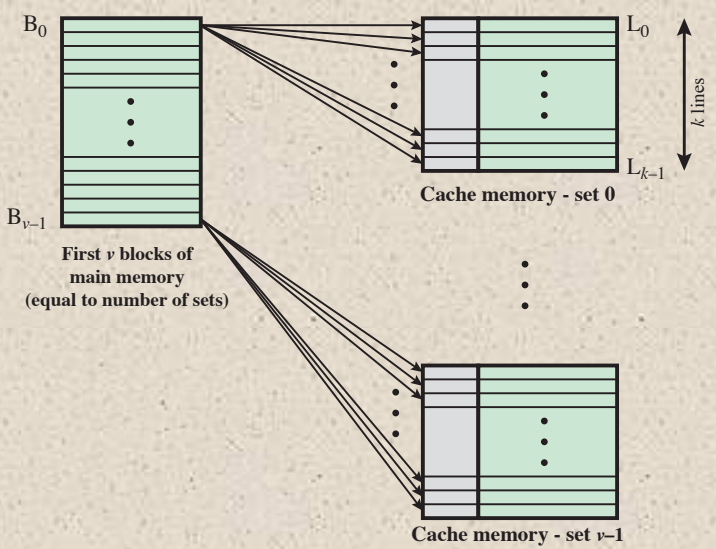
k-way set assoc

- Compromise that exhibits the strengths of both the direct and associative approaches while reducing their disadvantages
- Cache consists of a number of sets
- Each set contains a number of lines, k,
- A given block maps to any line in a given set
- e.g. 2 lines per set
  - 2 way associative mapping
  - A given block can be in one of 2 lines in only one set

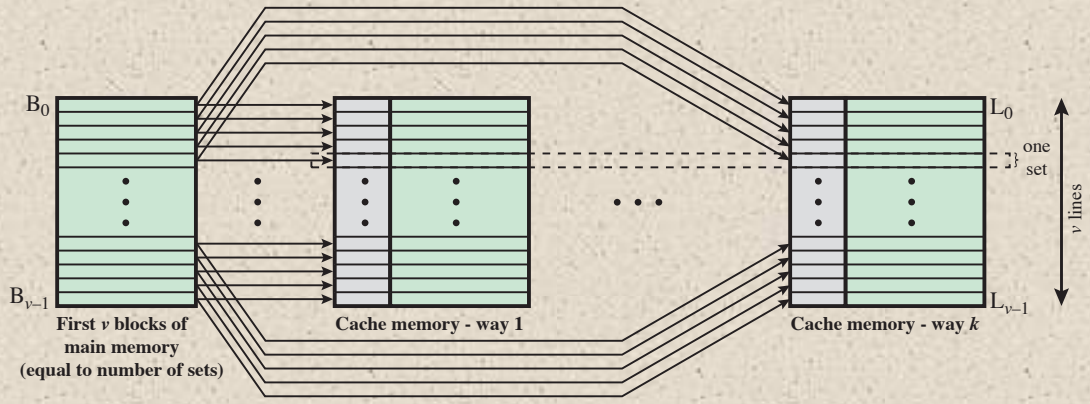


Many to one mapping

Many blocks map to one set



(a) *v* associative-mapped caches



(b) *k* direct-mapped caches

Figure 4.13 Mapping From Main Memory to Cache: *k*-way Set Associative

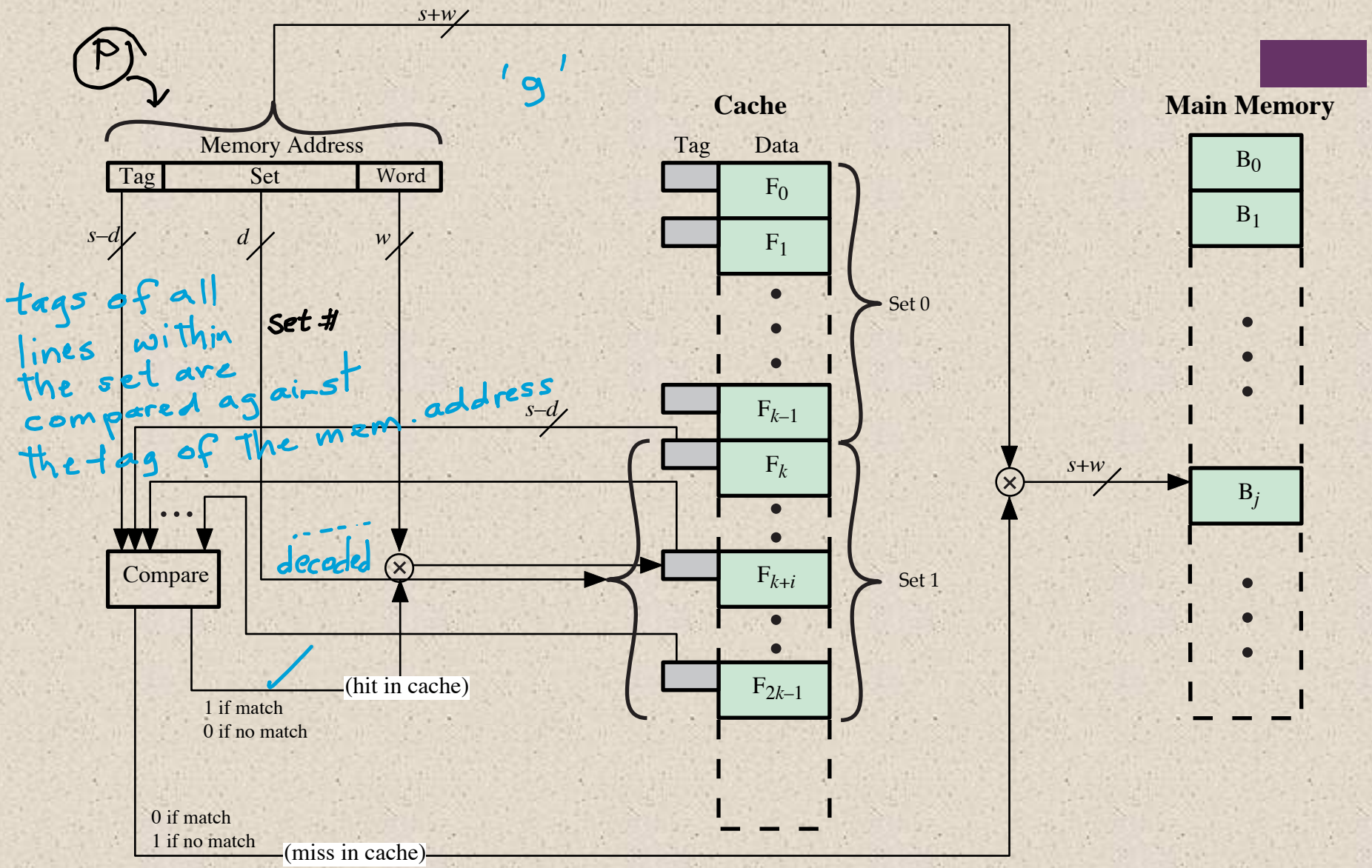


Figure 4.14 ***k*-Way Set Associative Cache Organization**

Set Associative 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 set = k
- Number of sets = v = 2^{d}
- Number of lines in cache = m=kv = k * 2^{d}
- Size of cache = $k * 2^{d+w}$ words or bytes
- Size of tag = (s – d) bits

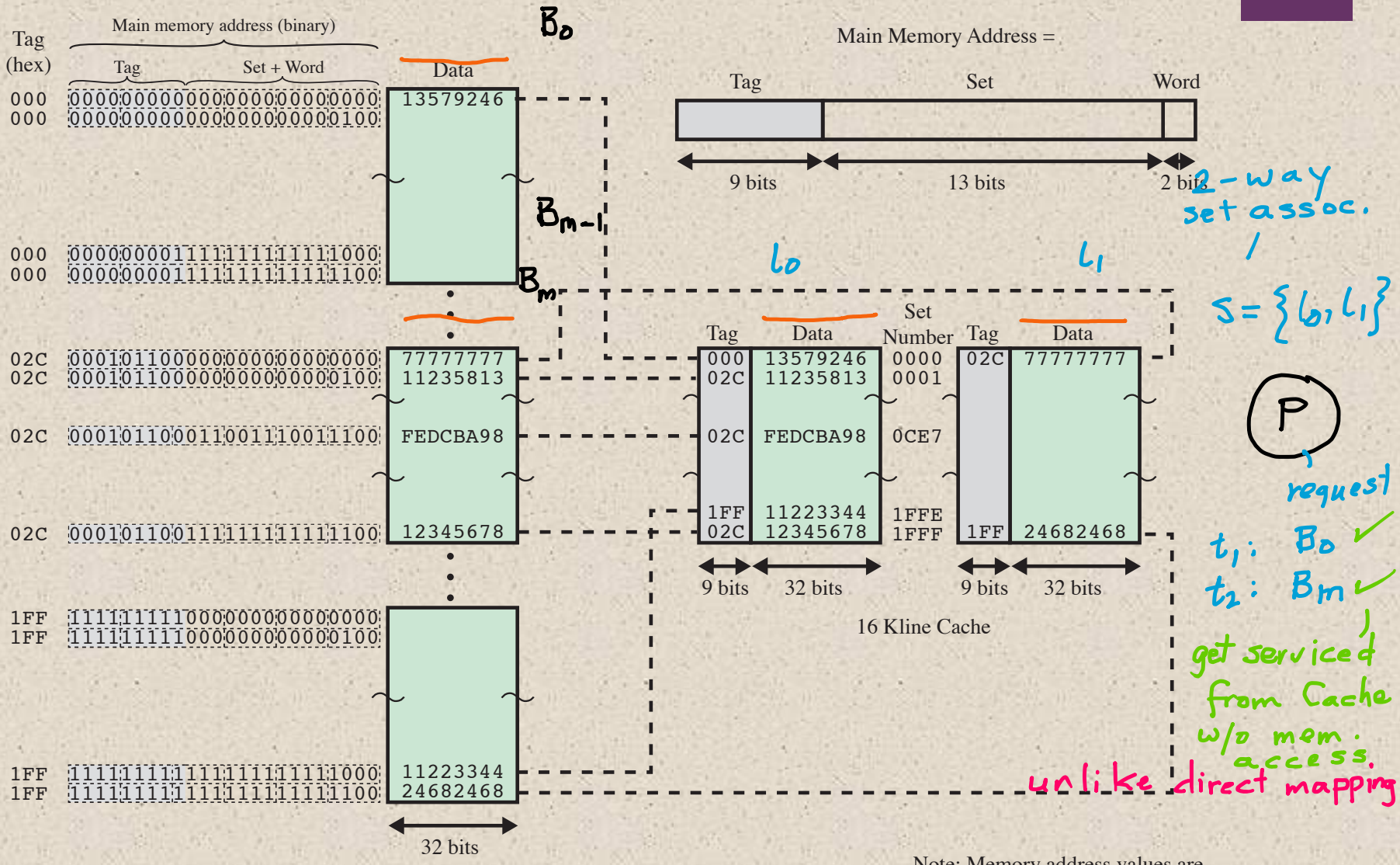


Figure 4.15 Two-Way Set Associative Mapping Example

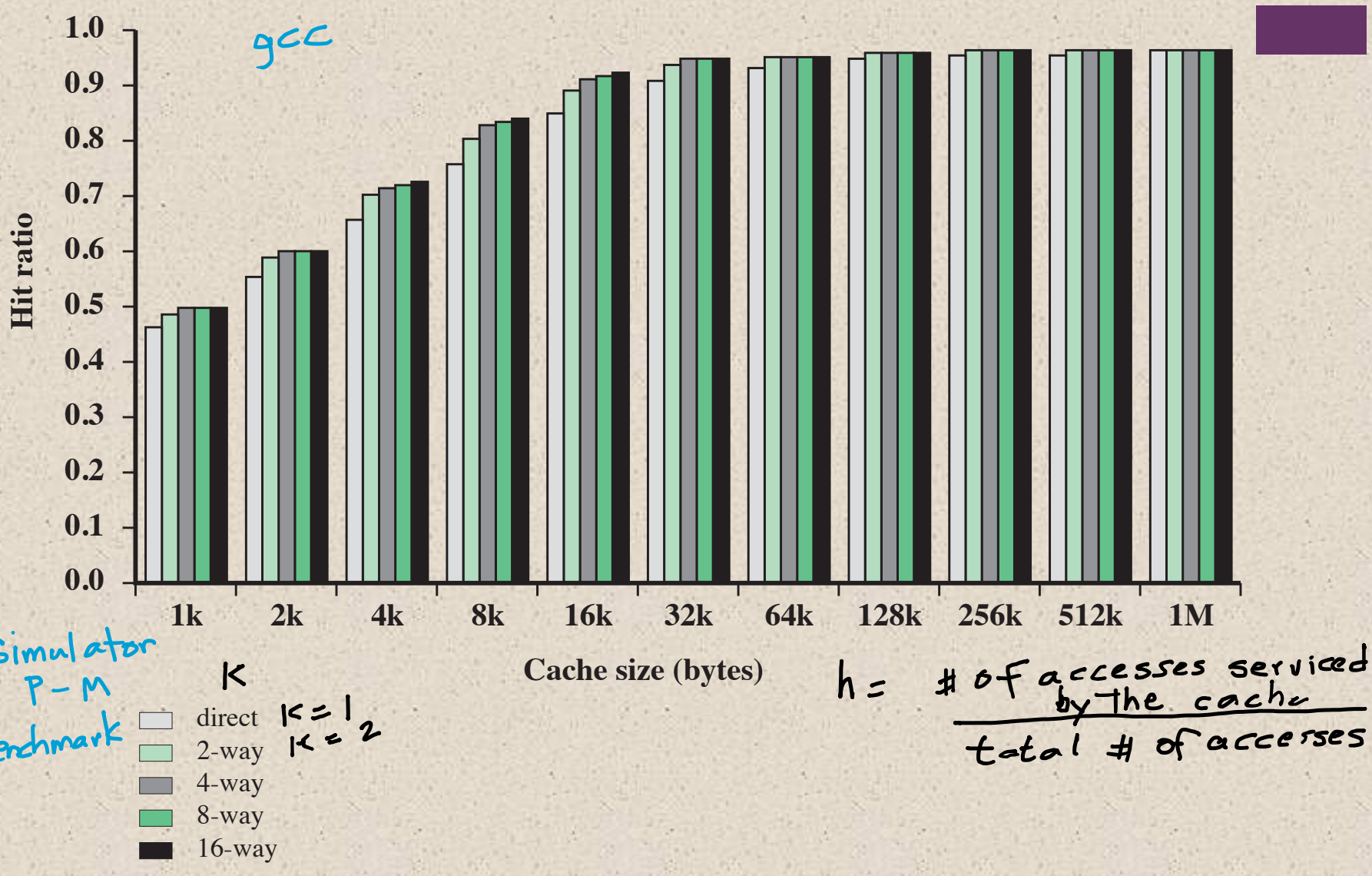
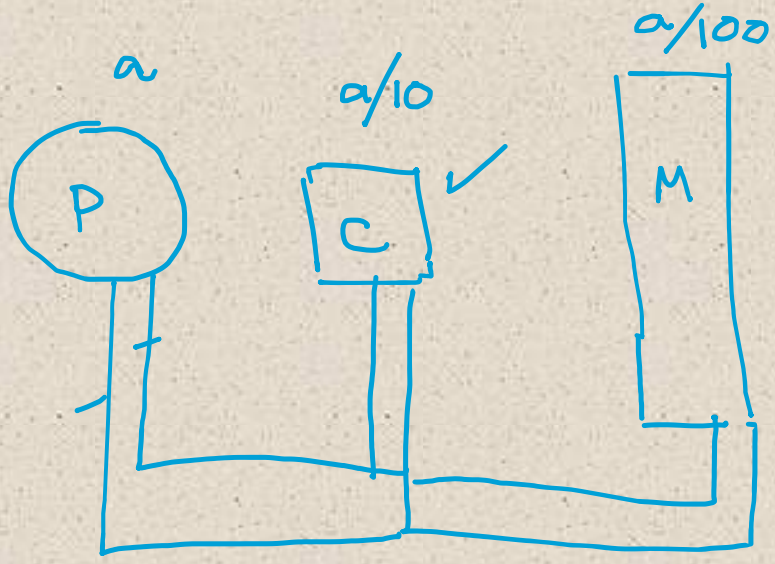
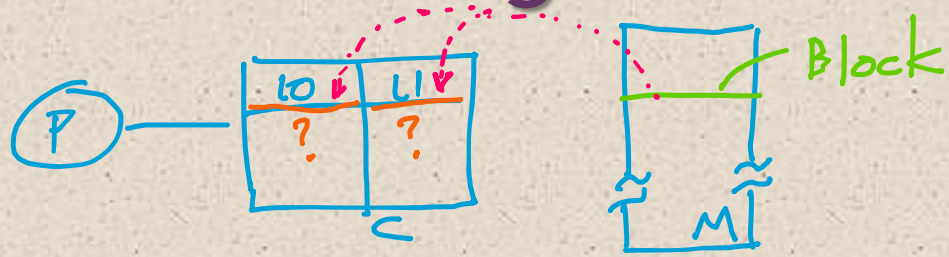


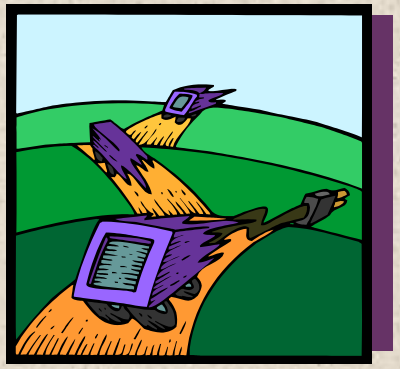
Figure 4.16 Varying Associativity over Cache Size

Processor memory speeds



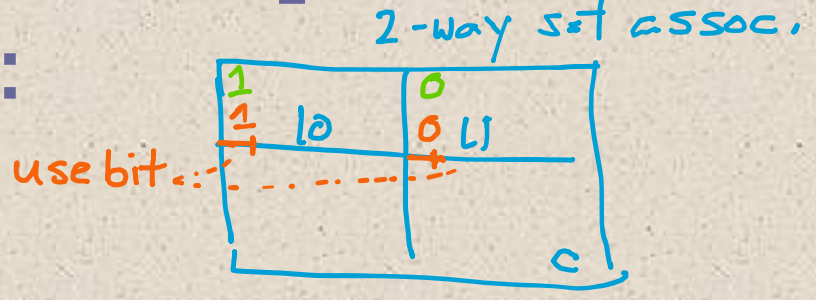
Replacement Algorithms



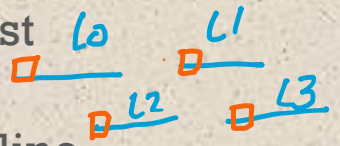


- Once the cache has been filled, when a new block is brought into the cache, one of the existing blocks must be replaced
- For direct mapping there is only one possible line for any particular block and no choice is possible
- For the associative and set-associative techniques a replacement algorithm is needed
  LRU, LFU, FIFO
  high hit rate
- To achieve high speed, an algorithm must be implemented in hardware

The most common replacement algorithms are:



- Least recently used (LRU)
  - Most effective
  - Replace that block in the set that has been in the cache longest with no reference to it
  - Because of its simplicity of implementation, LRU is the most popular replacement algorithm
- First-in-first-out (FIFO)
  - Replace that block in the set that has been in the cache longest
  - Easily implemented as a round-robin or circular buffer technique
- Least frequently used (LFU)
  - Replace that block in the set that has experienced the fewest references
  - Could be implemented by associating a counter with each line



Replacement algorithms

Q. Which line to replace when bringing a new block from MM to Cache?

A. LRU, FIFO, LFU

|

hardware

Write Policy

When a block that is resident in the cache is to be replaced there are two cases to consider:

If the old block in the cache has not been altered then it may be overwritten with a new block without first writing out the old block

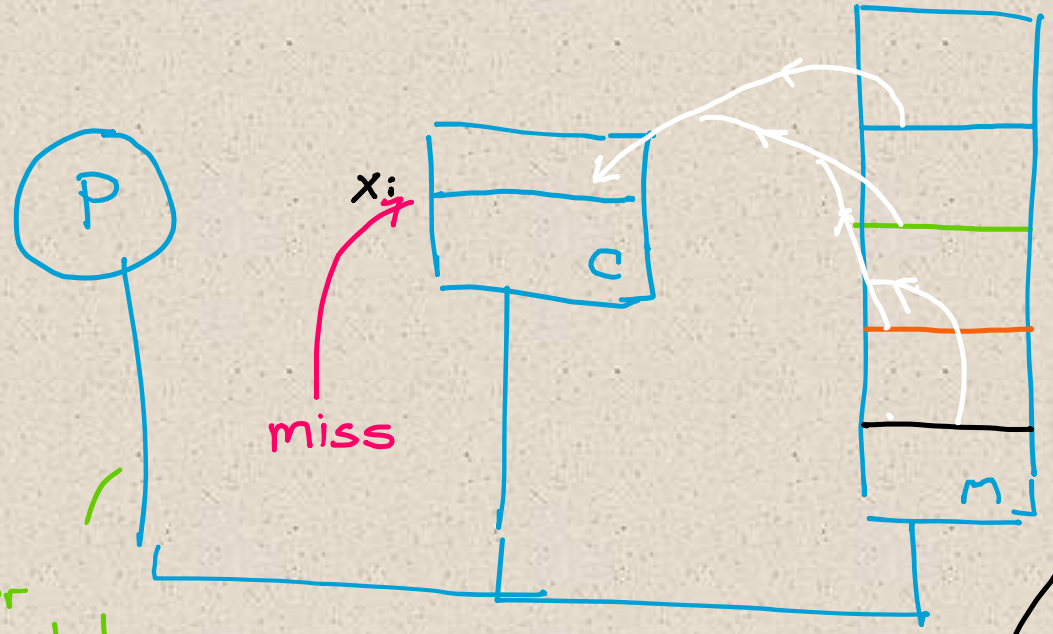
If at least one write operation has been performed on a word in that line of the cache then main memory must be updated by writing the line of cache out to the block of memory before bringing in the new block

There are two problems to contend with:

More than one device may have access to main memory

A more complex problem occurs when multiple processors are attached to the same bus and each processor has its own local cache - if a word is altered in one cache it could conceivably invalidate a word in other caches

Write Policy

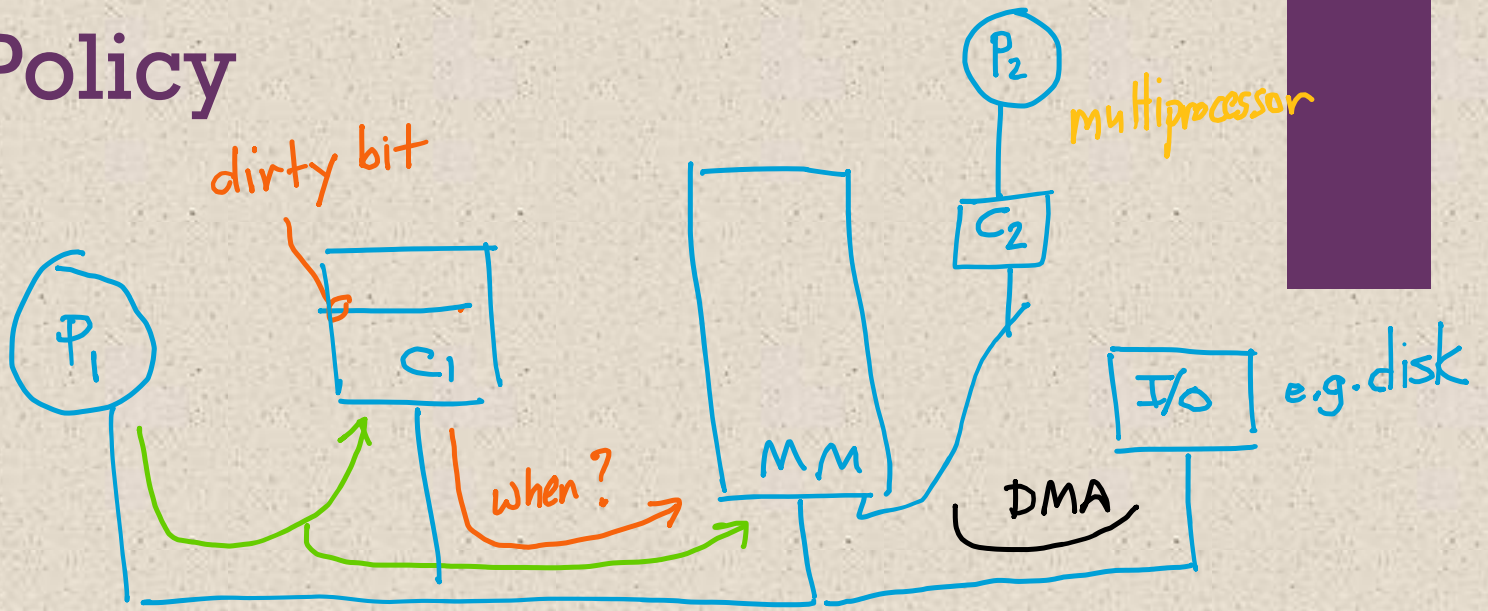


processor requests data in the green block

if the line was altered it must be written back to MM for consistency

blue line must be evicted from the cache and the green block from MM must be brought into cache at line # x.

Write Policy



- Write through - C & MM are always synch'ed; more bus traffic
- Write back - write to cache
  set the dirty bit

e.g. loop - overwriting the same variable with each iteration of the loop

+ bus traffic is lowered
- e.g. DMA stale dump to disk; complex \ckt to handle
- cache coherency problem

Write Through and Write Back

Q: When to write to MM?

- Write through
  - Simplest technique
  - All write operations are made to main memory as well as to the cache
  - The main disadvantage of this technique is that it generates substantial memory traffic and may create a bottleneck
- Write back
  - Minimizes memory writes
  - Updates are made only in the cache
  - Portions of main memory are invalid and hence accesses by I/O modules can be allowed only through the cache
  - This makes for complex circuitry and a potential bottleneck

Cache space overheads

m lines in the cache

- dirty bit; m lines → m/8 Bytes
- tag ; 8 bits of tag per line
  1 B/line of overhead; m bytes
- use bits; $\frac{m}{8} \cdot 3$ Bytes
  ↳ for 8-way set assoc.

$$\left(\frac{m}{8} \cdot 3 + \frac{m}{8} + m\right) B$$

usable → 1 GB −

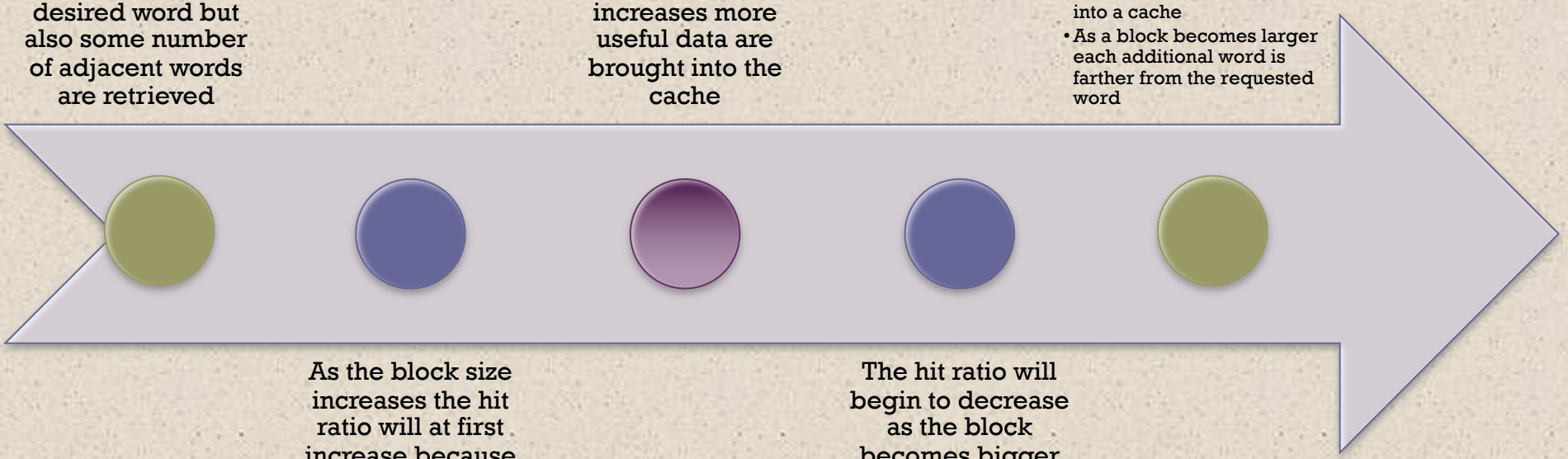
Line Size / Block size

When a block of data is retrieved and placed in the cache not only the desired word but also some number of adjacent words are retrieved

As the block size increases more useful data are brought into the cache

Two specific effects come into play:

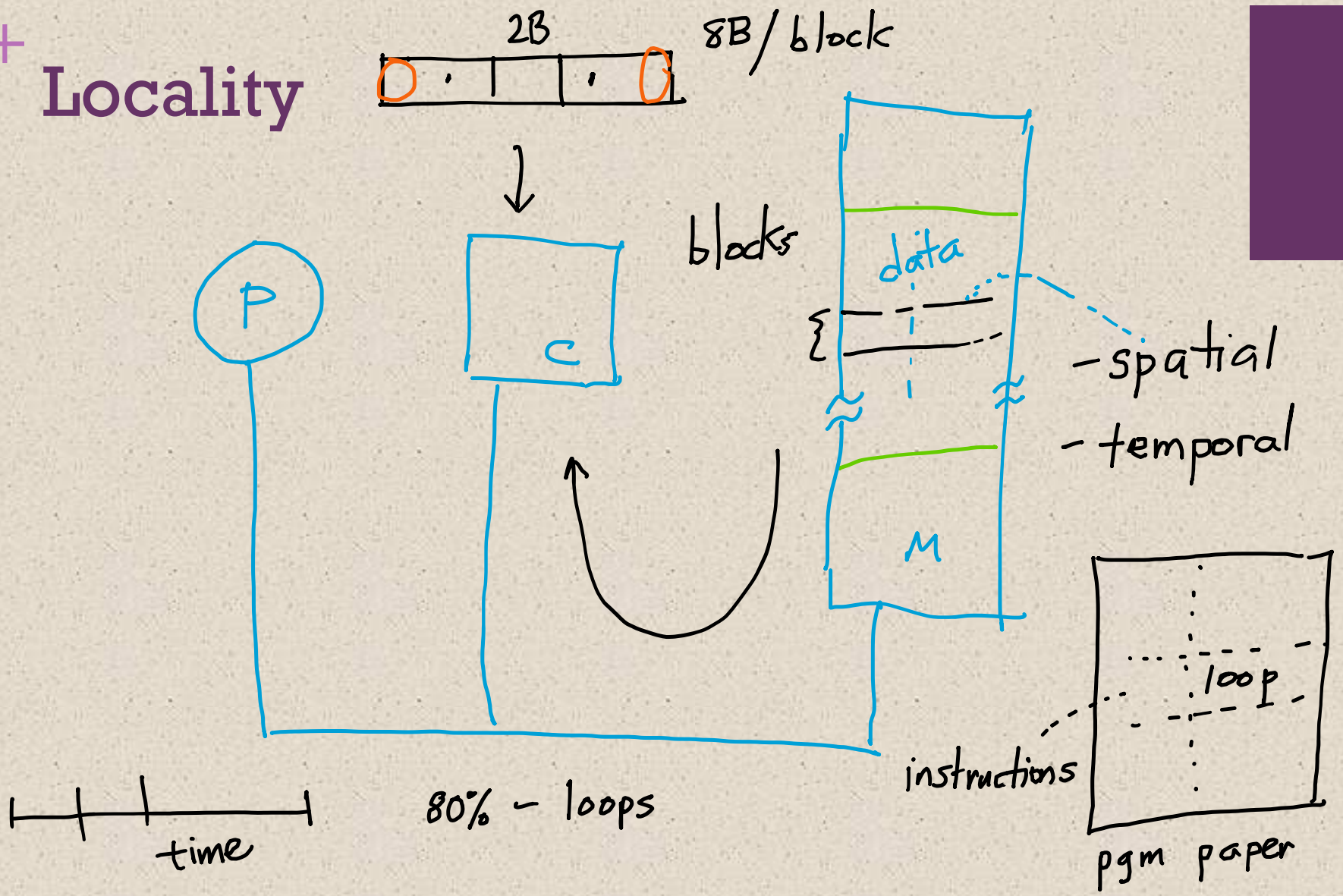
- Larger blocks reduce the number of blocks that fit into a cache
- As a block becomes larger each additional word is farther from the requested word



As the block size increases the hit ratio will at first increase because of the principle of locality

The hit ratio will begin to decrease as the block becomes bigger and the probability of using the newly fetched information becomes less than the probability of reusing the information that has to be replaced

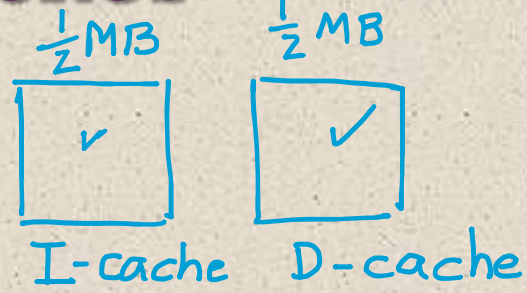
Locality



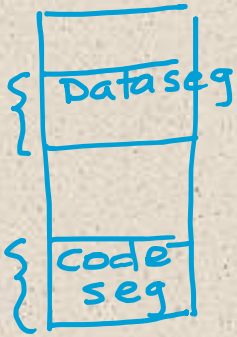
Multilevel Caches

- As logic density has increased it has become possible to have a cache on the same chip as the processor
- The on-chip cache reduces the processor's external bus activity and speeds up execution time and increases overall system performance
  - When the requested instruction or data is found in the on-chip cache, the bus access is eliminated
  - On-chip cache accesses will complete appreciably faster than would even zero-wait state bus cycles
  - During this period the bus is free to support other transfers
- Two-level cache:
  - Internal cache designated as level 1 (L1)
  - External cache designated as level 2 (L2)
- Potential savings due to the use of an L2 cache depends on the hit rates in both the L1 and L2 caches
- The use of multilevel caches complicates all of the design issues related to caches, including size, replacement algorithm, and write policy

Unified Versus Split Caches



- Has become common to split cache:
  - One dedicated to instructions
  - One dedicated to data
  - Both exist at the same level, typically as two L1 caches



- Advantages of unified cache:
  - Higher hit rate
    - Balances load of instruction and data fetches automatically
    - Only one cache needs to be designed and implemented
- Trend is toward split caches at the L1 and unified caches for higher levels
- Advantages of split cache:
  - Eliminates cache contention between instruction fetch/decode unit and execution unit
    - Important in pipelining

prefetching

FI DI CO FO EI WO

contention to access unified

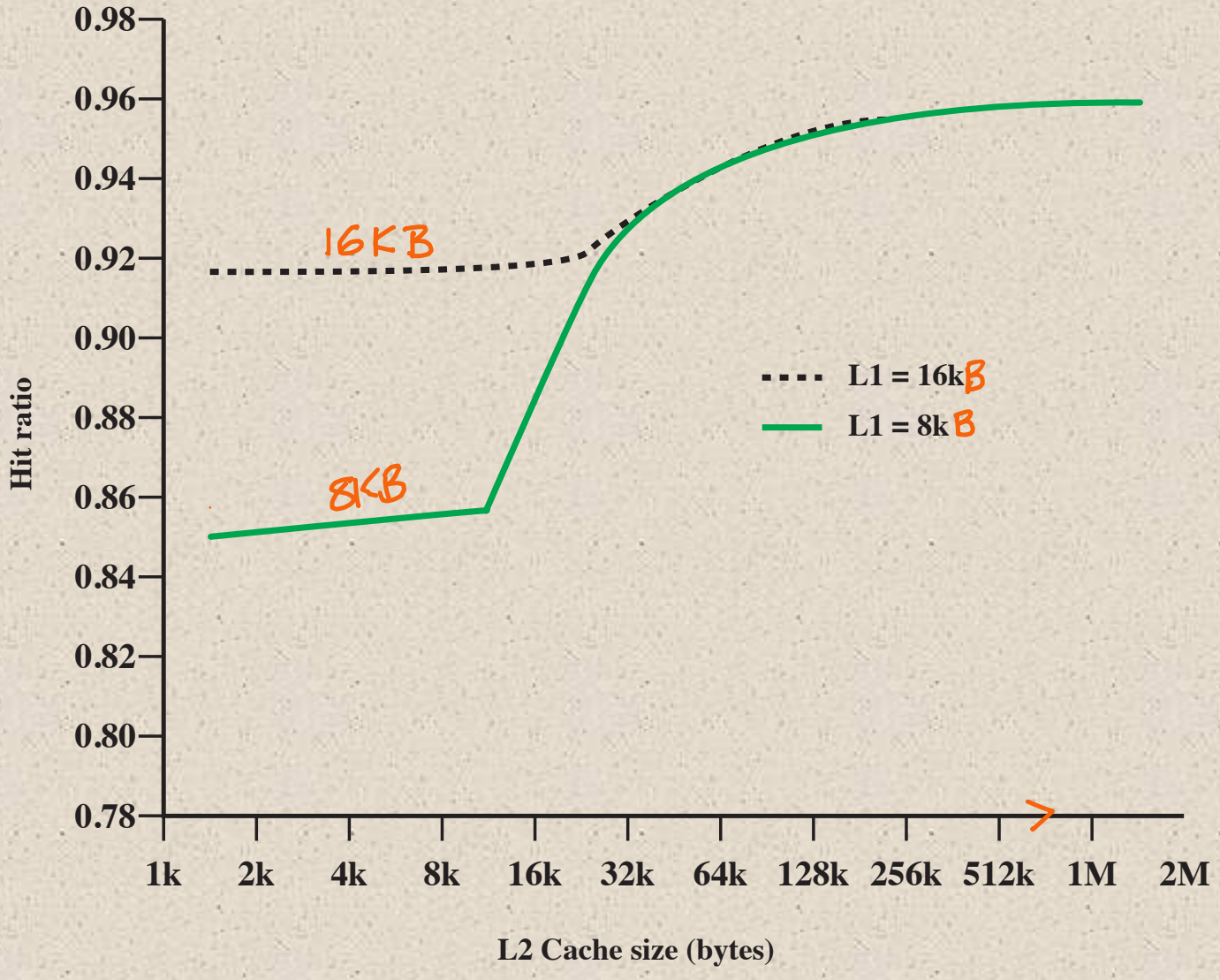
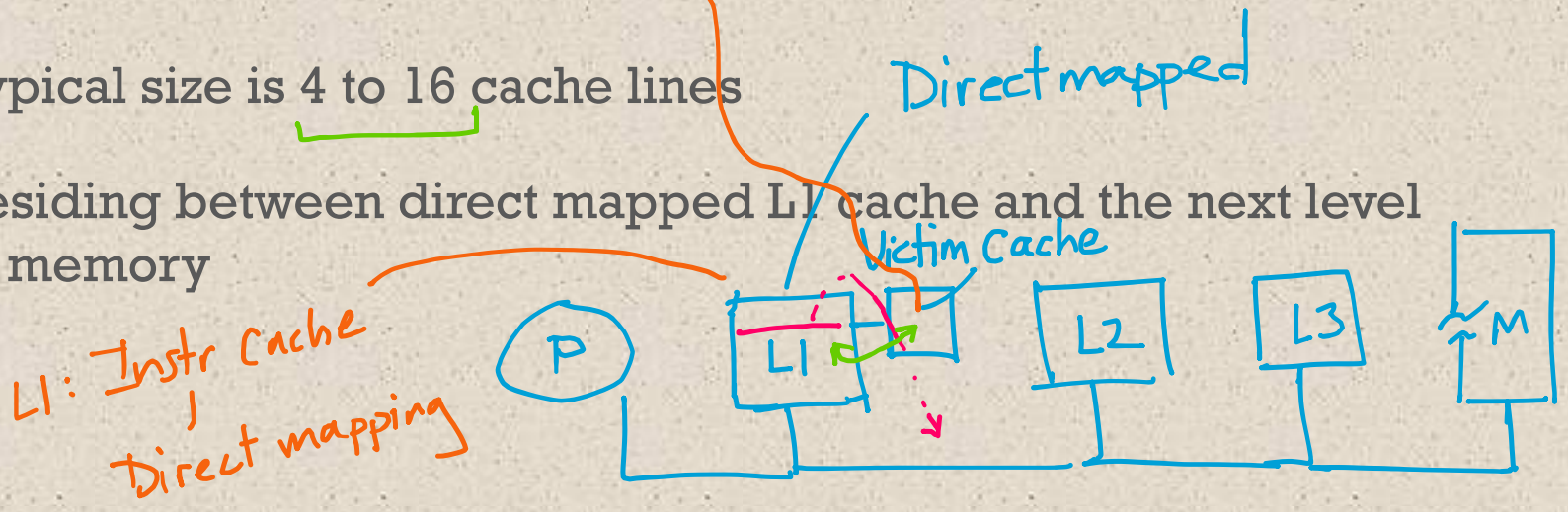


Figure 4.17 Total Hit Ratio (L1 and L2) for 8 Kbyte and 16 Kbyte L1

Victim Cache

Norman Jouppi

- Originally proposed as an approach to reduce the conflict misses of direct mapped caches without affecting its fast access time
- Fully associative cache
- Typical size is 4 to 16 cache lines
- Residing between direct mapped L1 cache and the next level of memory



Types of cache misses

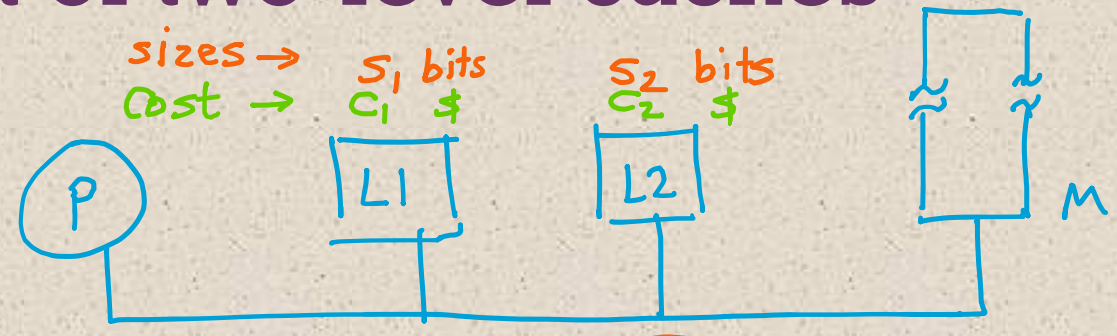
1. Compulsory miss – cold start
2. Conflict miss – evicted ←
3. Capacity miss – size of the cache on large data sets

- hit
- miss

= total # of misses − (# of compulsory misses + # of conflict misses)

✓ Direct – fastest, logic gates; victim cache
Assoc – slowest
Set Assoc – faster

Cost of two-level caches



$$C = \frac{C_1 S_1 + C_2 S_2}{S_1 + S_2}$$

— decides the total cost

Effective Cost per bit

$

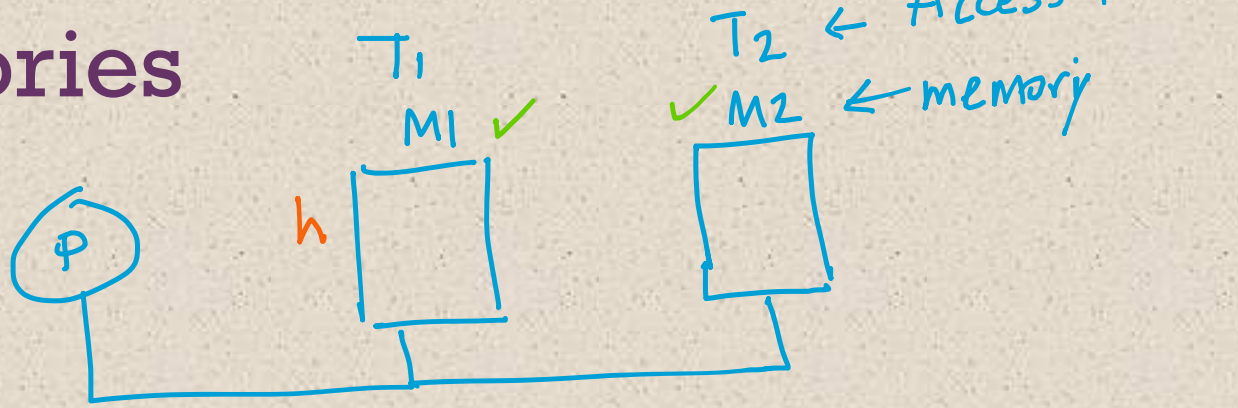
C_1/bit > C_2/bit

$S_2 > S_1$

Goal- minimize $

maximize perf.

Access time of multi-level memories



$$h = \frac{\text{\# of times the access succeeded}}{\text{total \# of accesses}}$$

$$T_e = h \cdot T_1 + (1-h)(T_1 + T_2)$$

Problem	Solution	Processor on which Feature First Appears
External memory slower than the system bus.	Add external cache using faster memory technology.	386 mid 1980s
Increased processor speed results in external bus becoming a bottleneck for cache access.	Move external cache on-chip, operating at the same speed as the processor.	486 late 1980s
Internal cache is rather small, due to limited space on chip	Add external L2 cache using faster technology than main memory	486
Contention occurs when both the Instruction Prefetcher and the Execution Unit simultaneously require access to the cache. In that case, the Prefetcher is stalled while the Execution Unit's data access takes place.	Create separate data and instruction caches.	Pentium
Increased processor speed results in external bus becoming a bottleneck for L2 cache access.	Create separate back-side bus that runs at higher speed than the main (front-side) external bus. The BSB is dedicated to the L2 cache.	Pentium Pro
	Move L2 cache on to the processor chip.	Pentium II
Some applications deal with massive databases and must have rapid access to large amounts of data. The on-chip caches are too small.	Add external L3 cache.	Pentium III
	Move L3 cache on-chip.	Pentium 4 2005 ..

Table 4.4

Intel Cache Evolution

(Table is on page 150 in the textbook.)

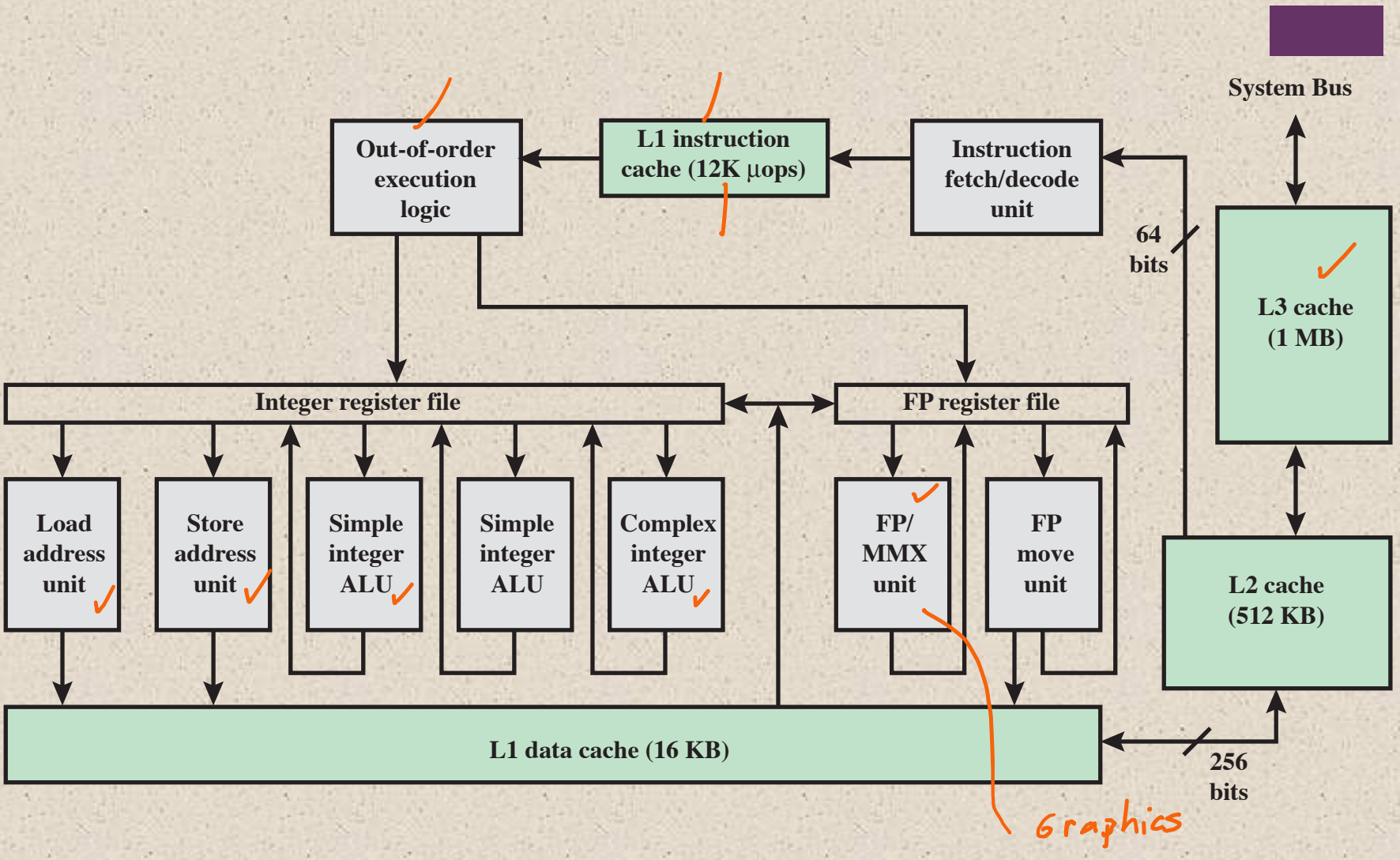


Figure 4.18 Pentium 4 Block Diagram

Table 4.5 Pentium 4 Cache Operating Modes

Control Bits		Operating Mode		
CD	**NW**	**Cache Fills**	**Write Throughs**	**Invalidates**
0	0	Enabled	Enabled	Enabled
1	0	Disabled	Enabled	Enabled
1	1	Disabled	Disabled	Disabled

Note: CD = 0; NW = 1 is an invalid combination.

Summary

Chapter 4

Cache Memory

- Computer memory system overview
  - Characteristics of Memory Systems
  - Memory Hierarchy
- Cache memory principles
- Pentium 4 cache organization

- Elements of cache design
  - Cache addresses
  - Cache size
  - Mapping function
  - Replacement algorithms
  - Write policy
  - Line size
  - Number of caches