



# Cache Memories

CS:APP 6.4

These slides adapted from materials provided by the textbook authors.

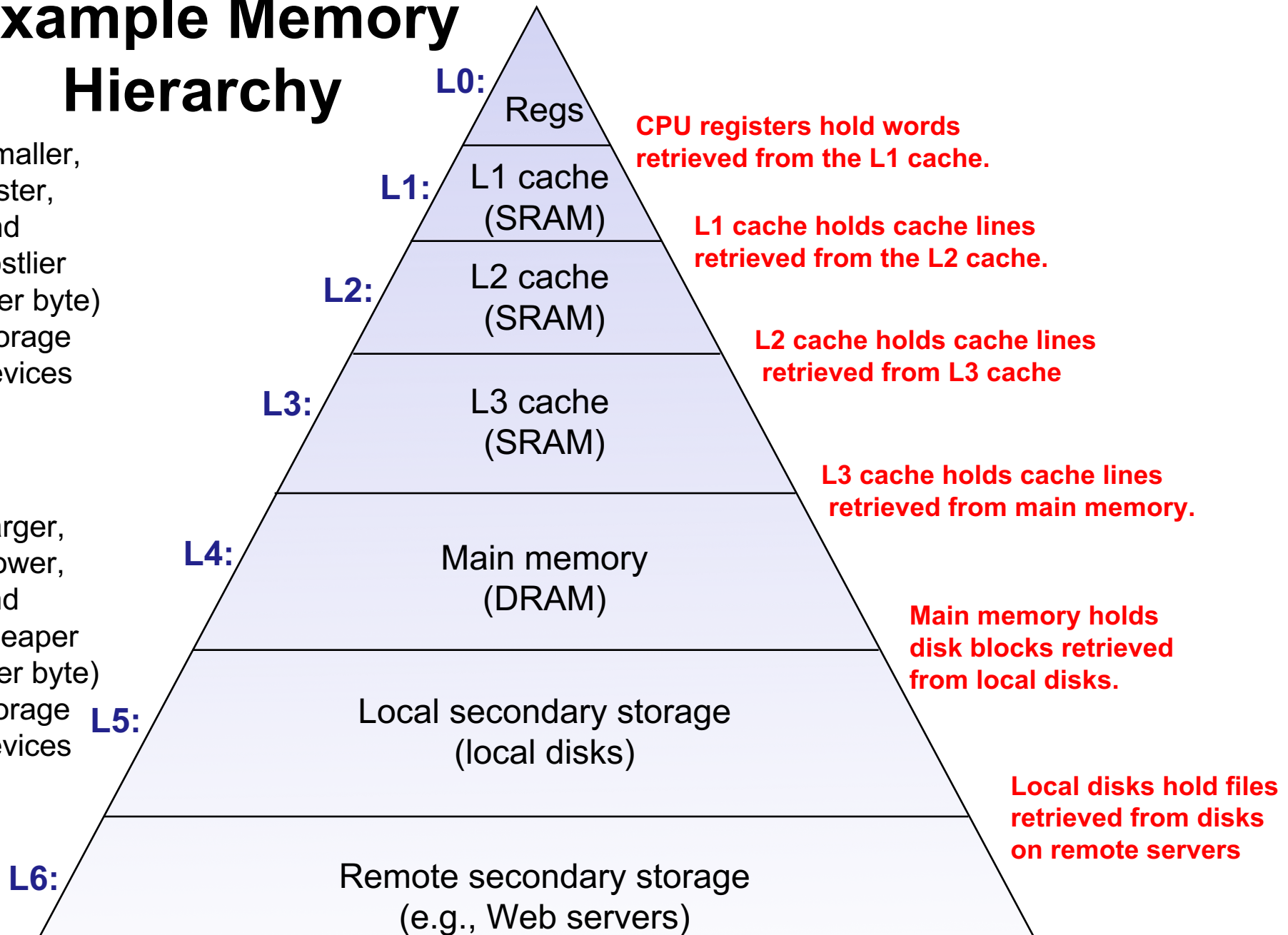
# Cache Memories

- **Cache memory organization and operation**
- **Performance impact of caches**
  - The memory mountain

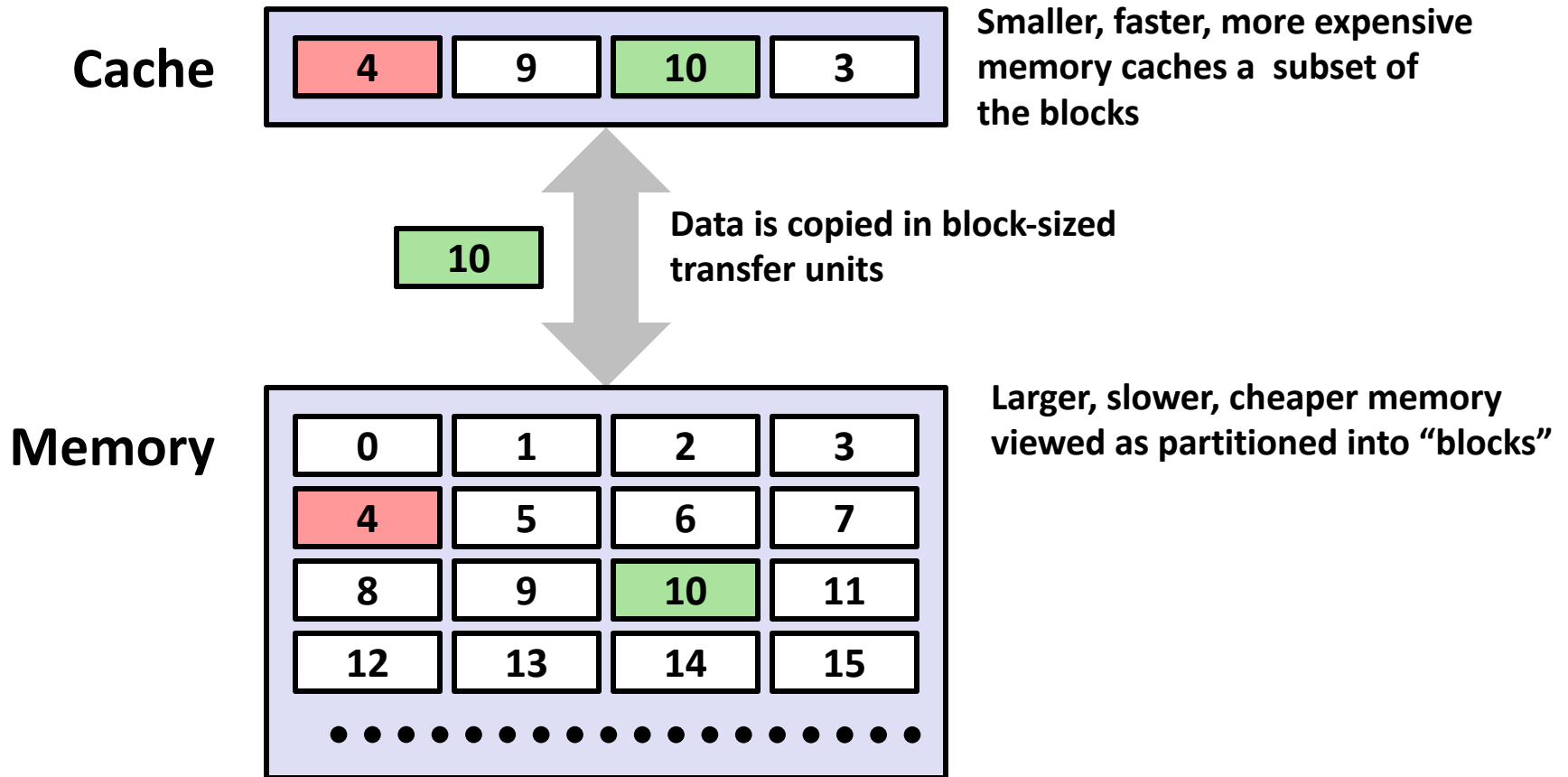
# Example Memory Hierarchy

↑  
Smaller,  
faster,  
and  
costlier  
(per byte)  
storage  
devices

↓  
Larger,  
slower,  
and  
cheaper  
(per byte)  
storage  
devices

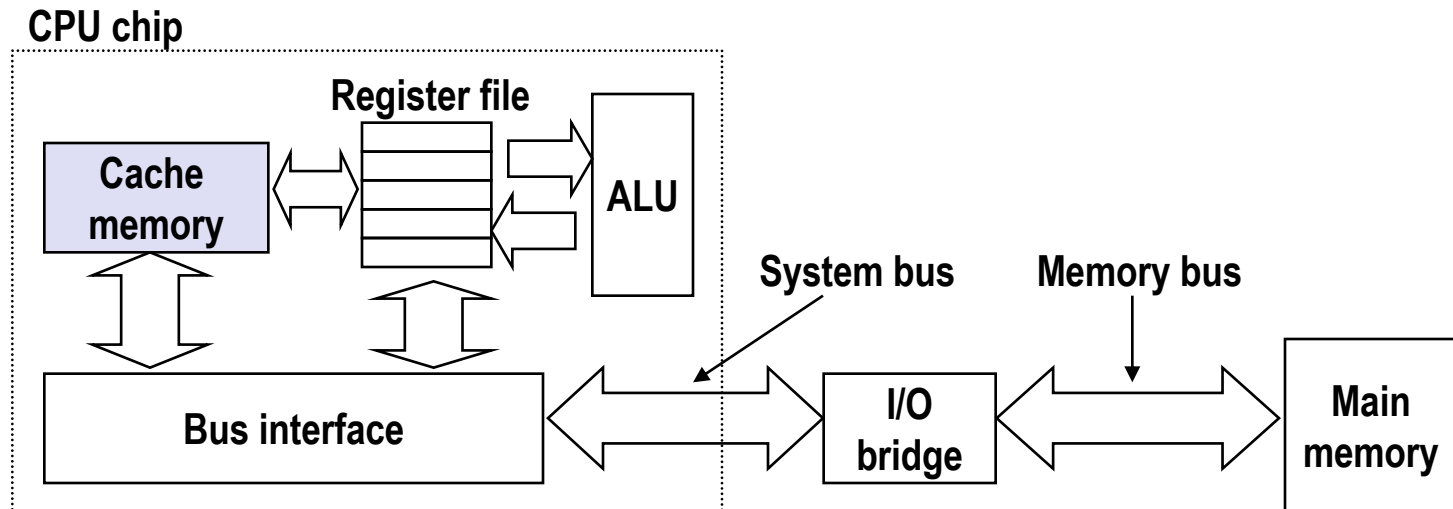


# General Cache Concept



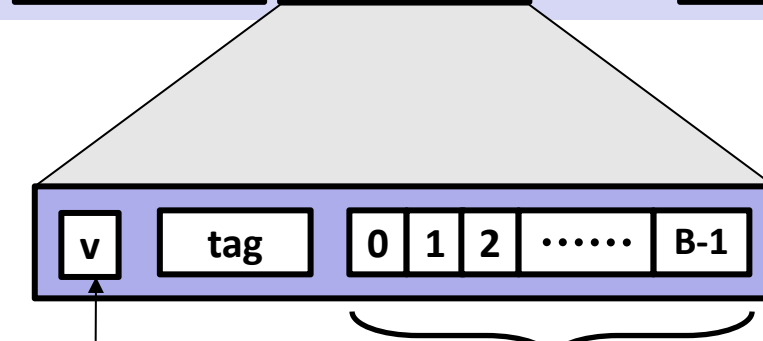
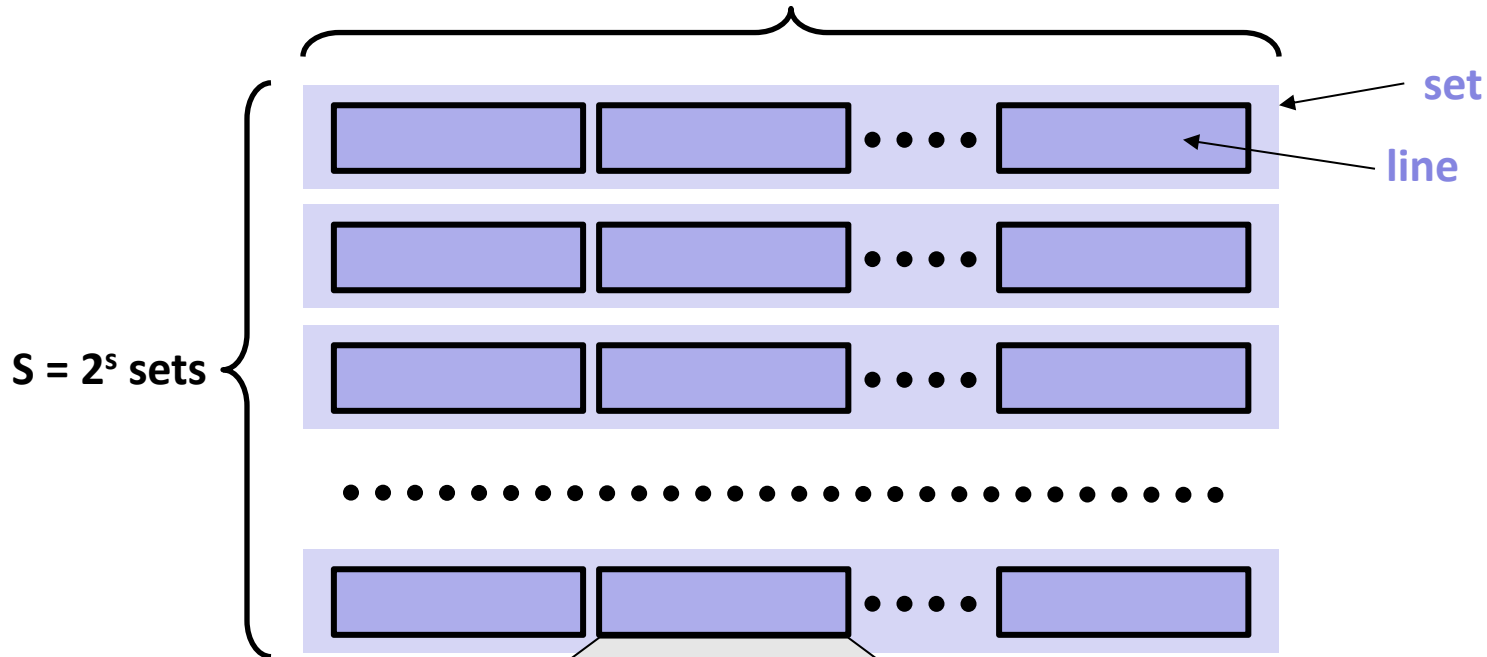
# Cache Memories

- **Cache memories** are small, fast SRAM-based memories managed automatically in hardware
  - Hold frequently accessed blocks of main memory
- CPU looks first for data in cache
- Typical system structure:



# General Cache Organization (S, E, B)

$E = 2^e$  lines per set



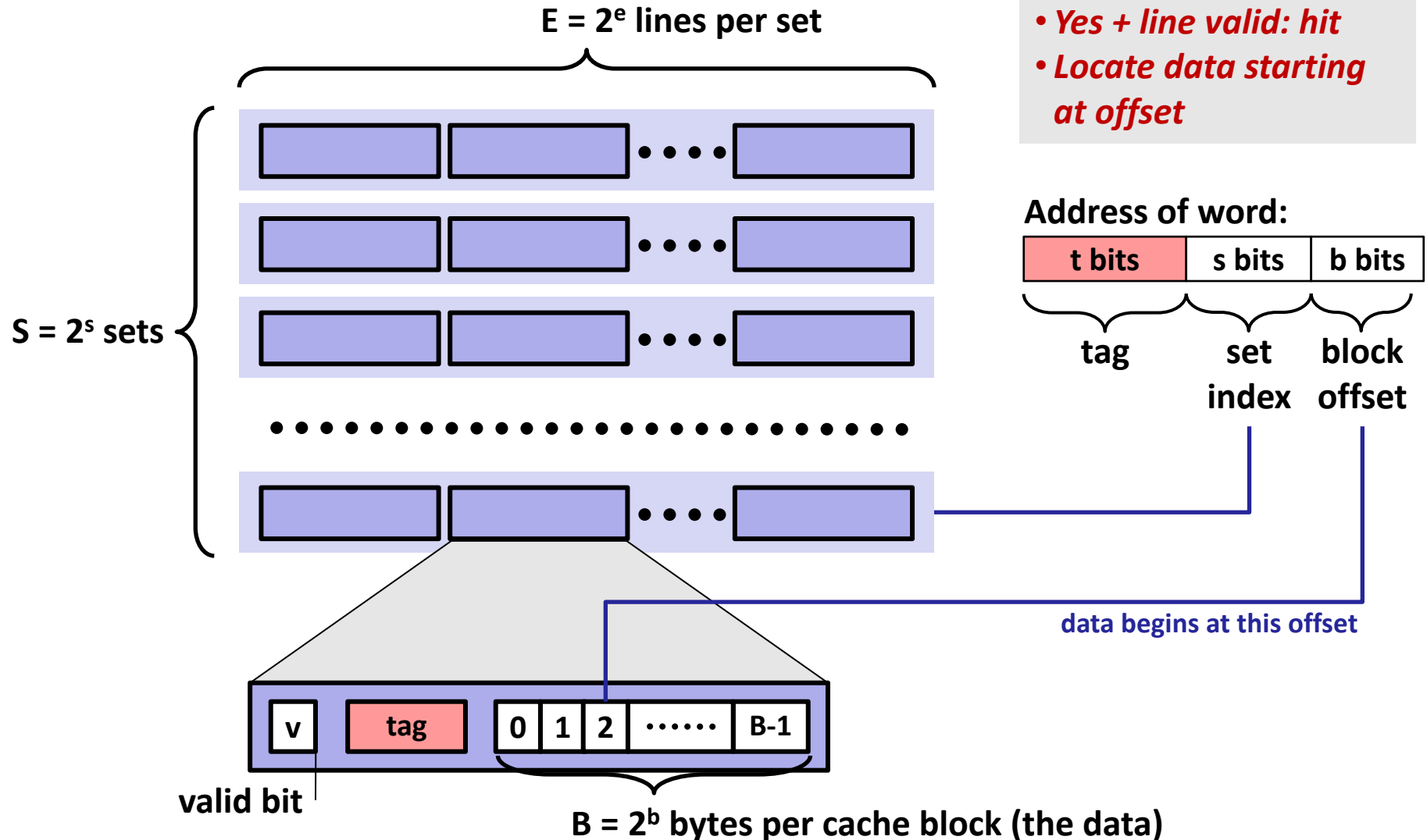
valid bit

$B = 2^b$  bytes per cache block (the data)

**Cache size:**

$C = S \times E \times B$  data bytes

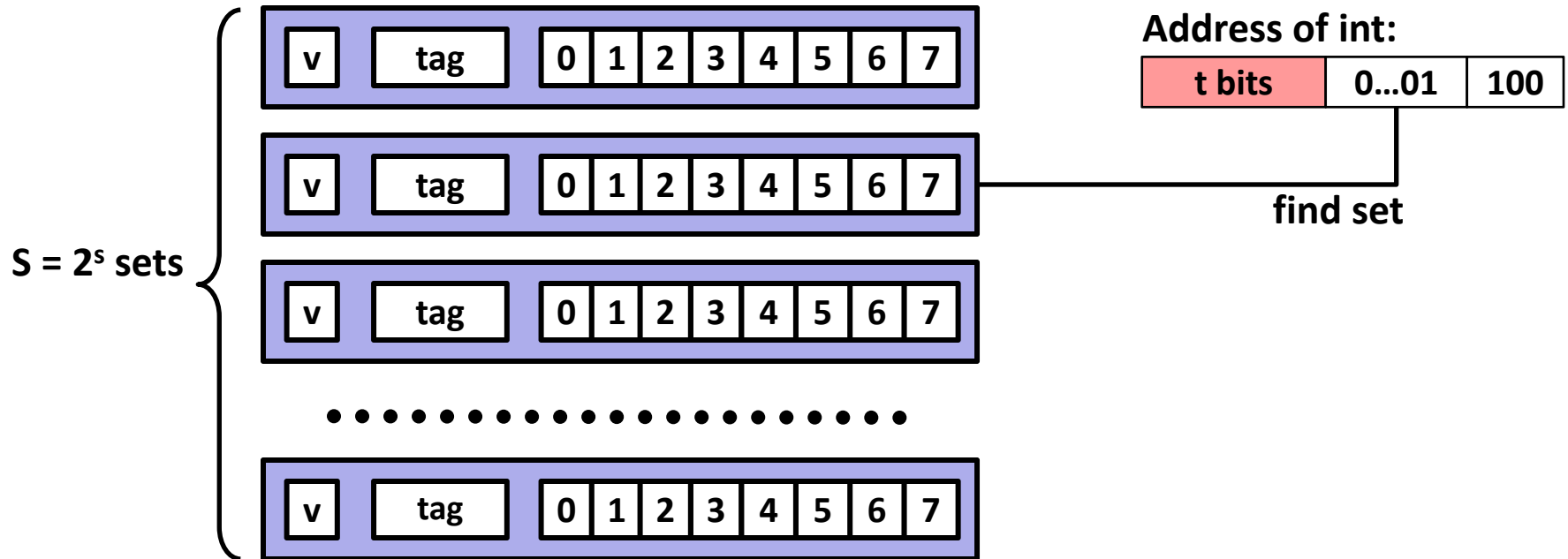
# Cache Read



# Example: Direct Mapped Cache (E = 1)

Direct mapped: One line per set

Assume: cache block size 8 bytes

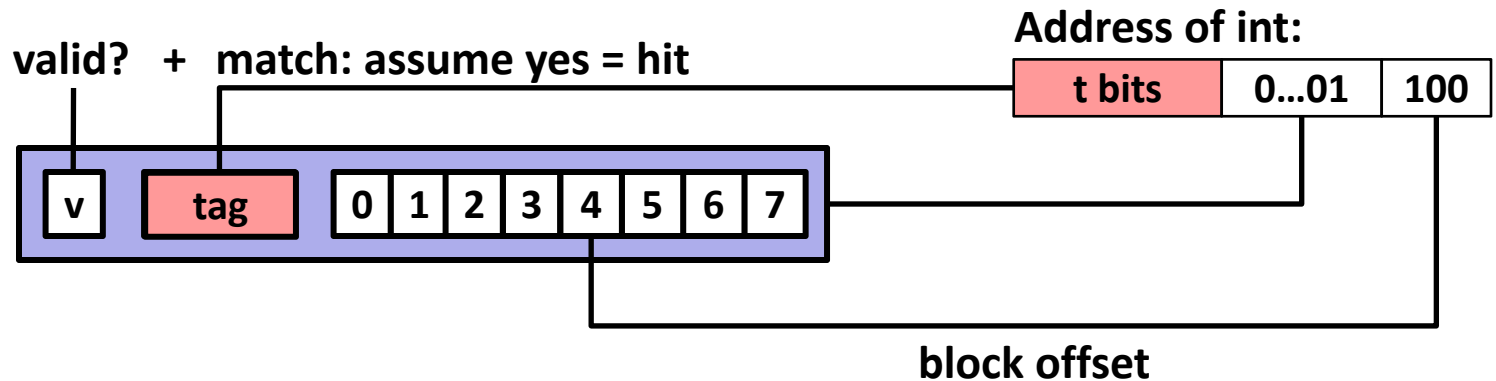




# Example: Direct Mapped Cache (E = 1)

Direct mapped: One line per set

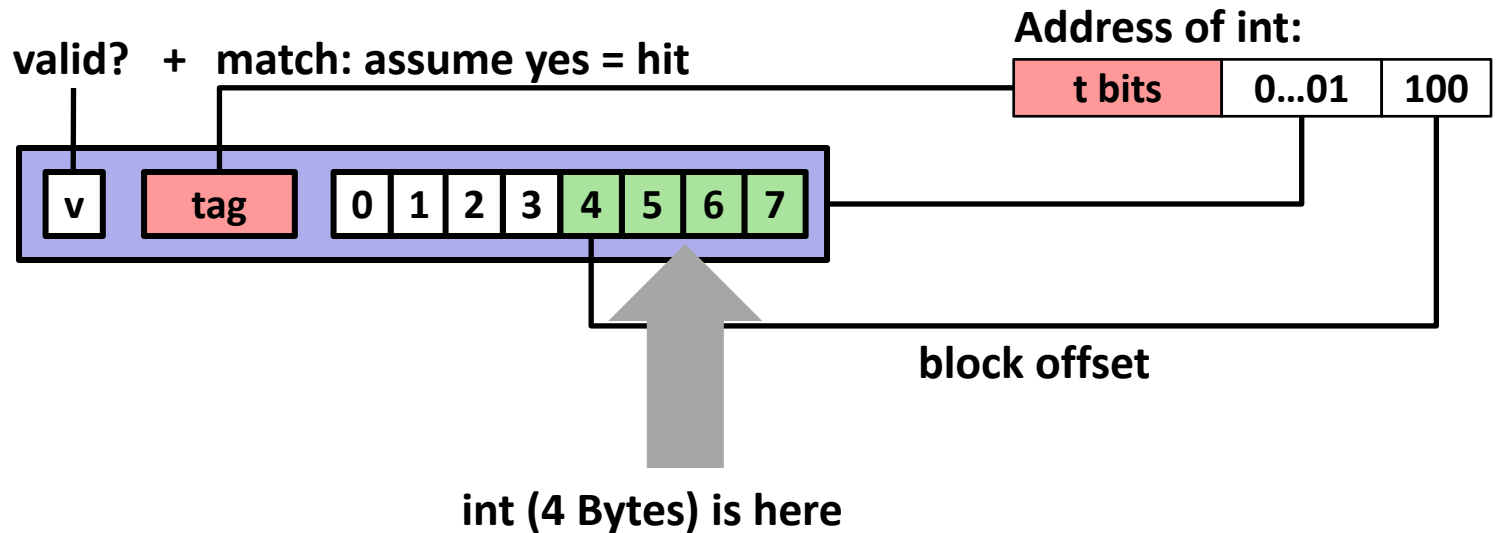
Assume: cache block size 8 bytes



# Example: Direct Mapped Cache (E = 1)

Direct mapped: One line per set

Assume: cache block size 8 bytes



**If tag doesn't match:** old line is evicted and replaced

# Direct-Mapped Cache Simulation

t=1	s=2	b=1
x	xx	x

M=16 bytes (4-bit addresses), B=2 bytes/block,  
S=4 sets, E=1 Blocks/set

Address trace (reads, one byte per read):

0	[ <u>0000</u> <sub>2</sub> ],	miss
1	[ <u>0001</u> <sub>2</sub> ],	hit
7	[ <u>0111</u> <sub>2</sub> ],	miss
8	[ <u>1000</u> <sub>2</sub> ],	miss
0	[ <u>0000</u> <sub>2</sub> ]	miss

	v	Tag	Block
Set 0	1	0	M[0-1]
Set 1			
Set 2			
Set 3	1	0	M[6-7]

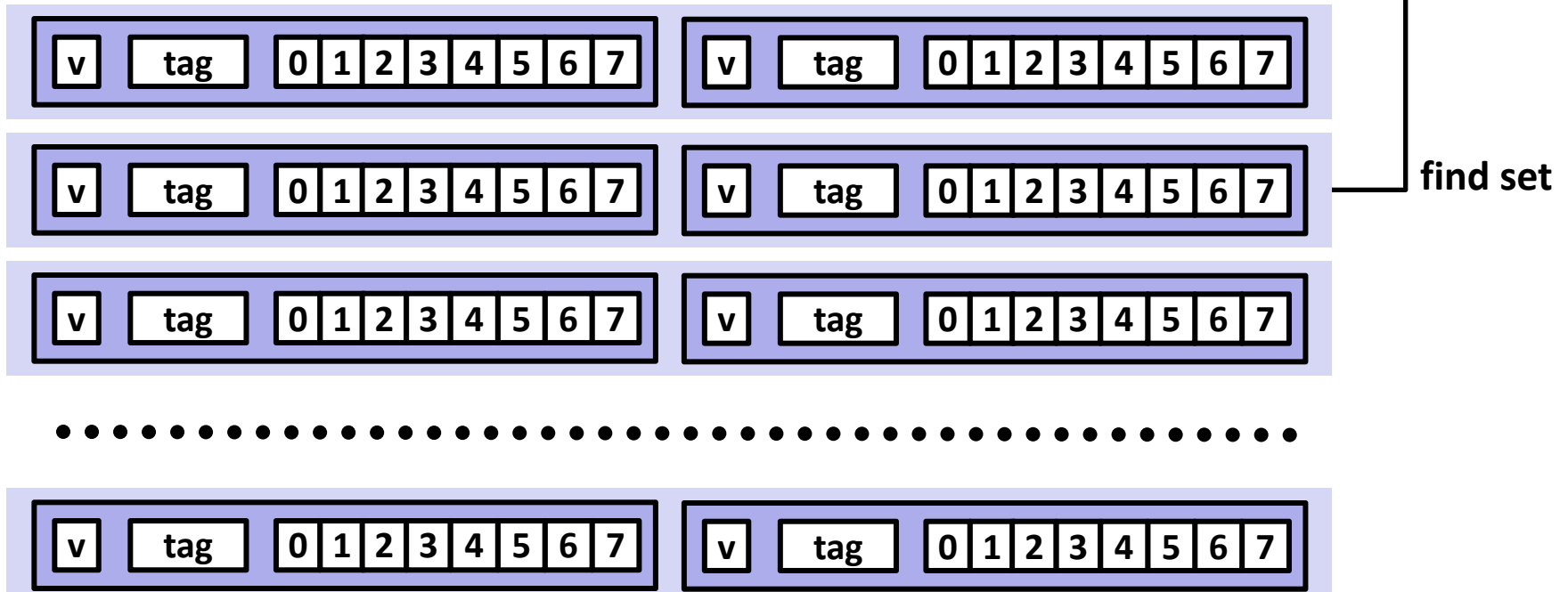
# E-way Set Associative Cache (Here: E = 2)

E = 2: Two lines per set

Assume: cache block size 8 bytes

Address of short int:

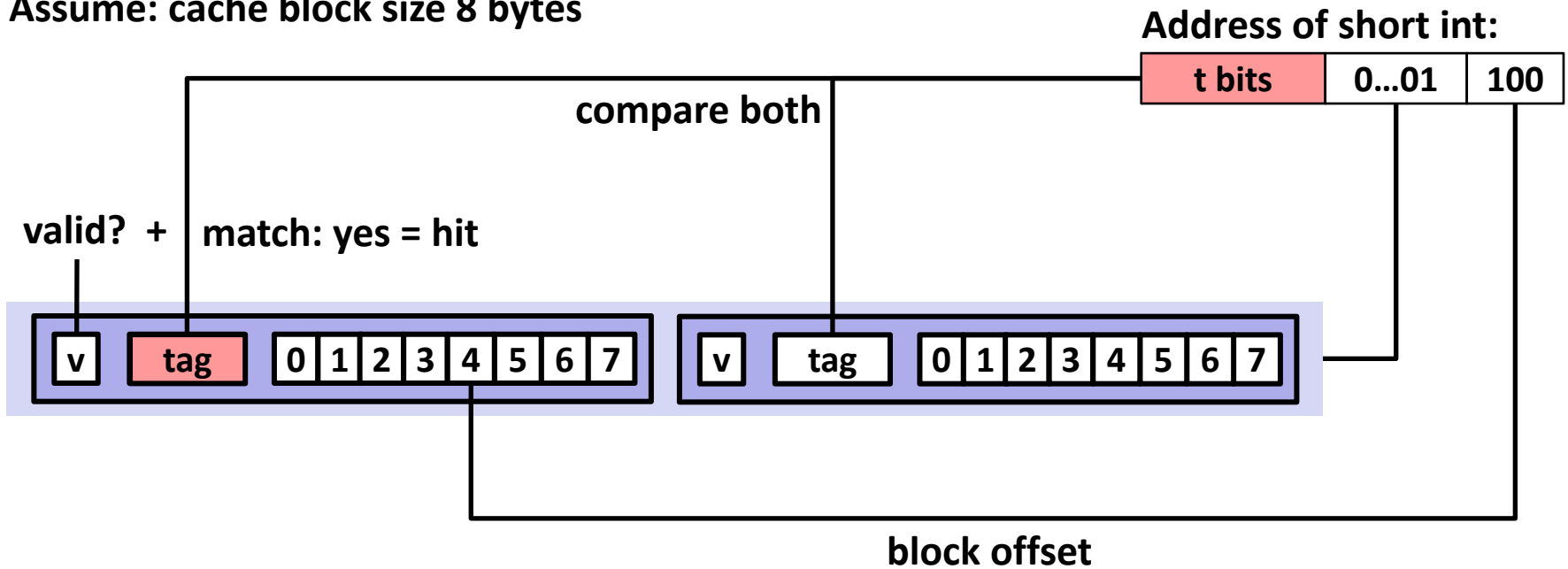
t bits	0...01	100
--------	--------	-----



# E-way Set Associative Cache (Here: E = 2)

E = 2: Two lines per set

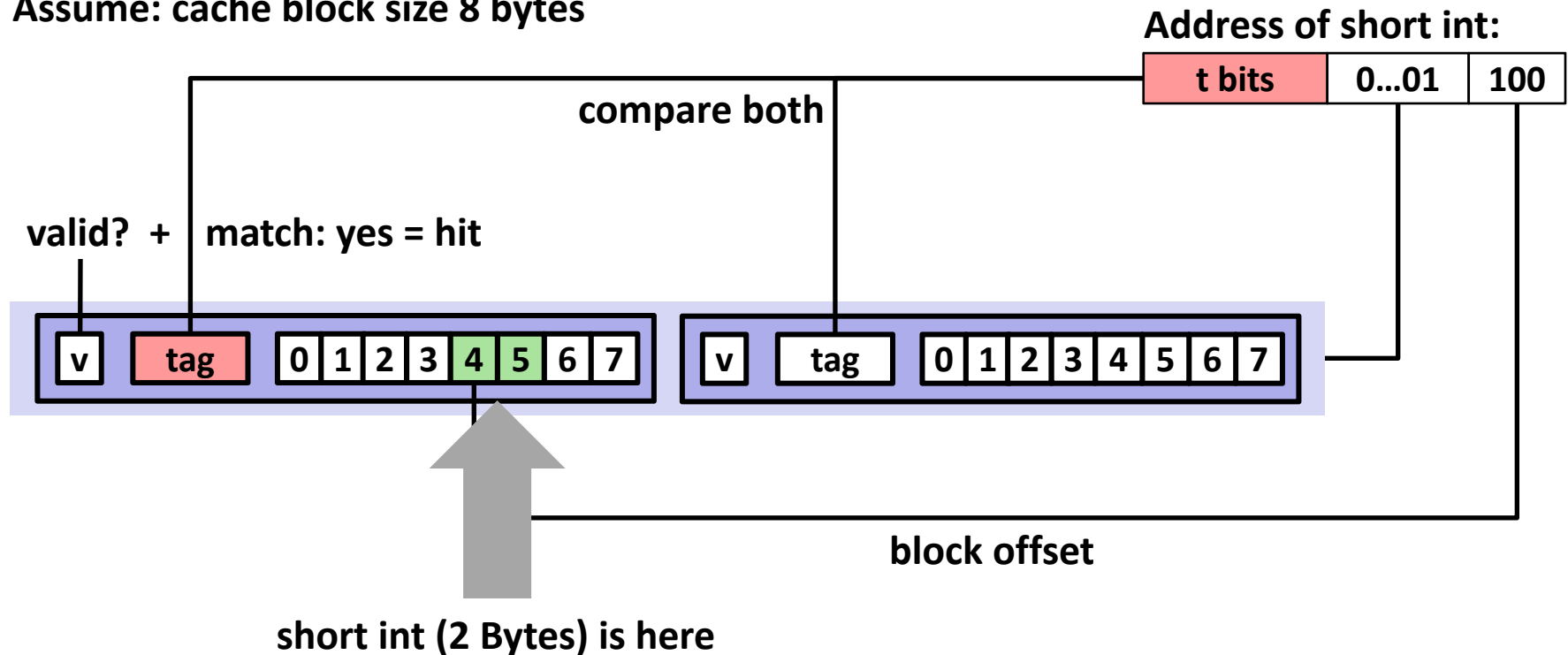
Assume: cache block size 8 bytes



# E-way Set Associative Cache (Here: E = 2)

E = 2: Two lines per set

Assume: cache block size 8 bytes



## No match:

- One line in set is selected for eviction and replacement
- Replacement policies: random, least recently used (LRU), ...

# 2-Way Set Associative Cache Simulation

t=2	s=1	b=1
xx	x	x

M=16 byte addresses, B=2 bytes/block,  
S=2 sets, E=2 blocks/set

Address trace (reads, one byte per read):

0	[0000 <sub>2</sub> ],	miss
1	[0001 <sub>2</sub> ],	hit
7	[0111 <sub>2</sub> ],	miss
8	[1000 <sub>2</sub> ],	miss
0	[0000 <sub>2</sub> ]	hit

	v	Tag	Block
Set 0	1	00	M[0-1]
	1	10	M[8-9]
Set 1	1	01	M[6-7]
	0		

# What about writes?

## ■ Multiple copies of data exist:

- L1, L2, L3, Main Memory, Disk

## ■ What to do on a write-hit?

- **Write-through** (write immediately to memory)
- **Write-back** (defer write to memory until replacement of line)
  - Need a dirty bit (line different from memory or not)

## ■ What to do on a write-miss?

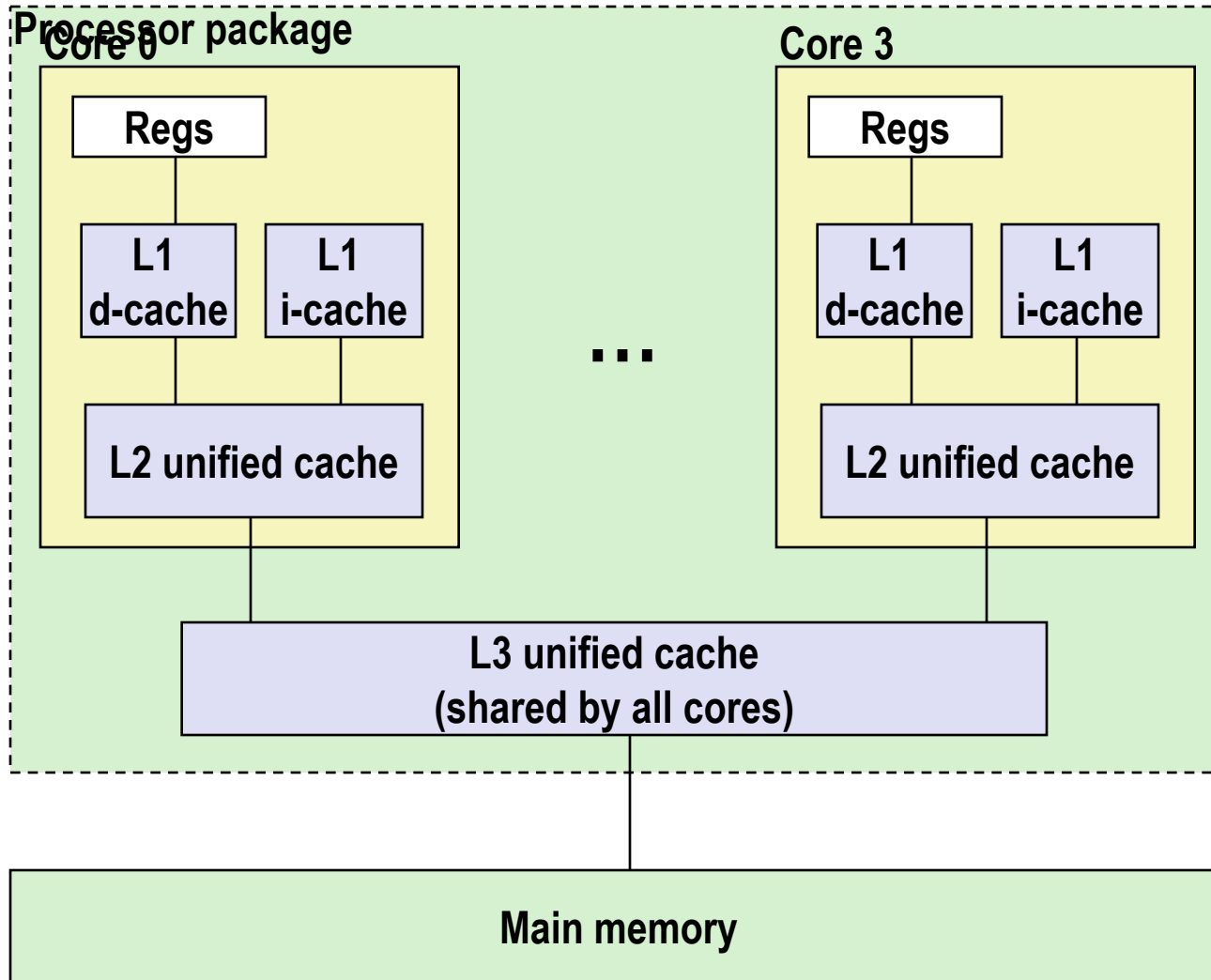
- **Write-allocate** (load into cache, update line in cache)
  - Good if more writes to the location follow
- **No-write-allocate** (writes straight to memory, does not load into cache)

## ■ Typical

- Write-through + No-write-allocate
- **Write-back + Write-allocate**



# Intel Core i7 Cache Hierarchy



## L1 i-cache and d-cache:

32 KB, 8-way,  
Access: 4 cycles

## L2 unified cache:

256 KB, 8-way,  
Access: 10 cycles

## L3 unified cache:

8 MB, 16-way,  
Access: 40-75 cycles

**Block size:** 64 bytes for  
all caches.

# Cache Performance Metrics

## ■ Miss Rate

- Fraction of memory references not found in cache (misses / accesses)  
=  $1 - \text{hit rate}$
- Typical numbers (in percentages):
  - 3-10% for L1
  - can be quite small (e.g.,  $< 1\%$ ) for L2, depending on size, etc.

## ■ Hit Time

- Time to deliver a line in the cache to the processor
  - includes time to determine whether the line is in the cache
- Typical numbers:
  - 4 clock cycle for L1
  - 10 clock cycles for L2

## ■ Miss Penalty

- Additional time required because of a miss
  - typically 50-200 cycles for main memory (Trend: increasing!)

# Let's think about those numbers

- **Huge difference between a hit and a miss**
  - Could be 100x, if just L1 and main memory
- **Would you believe 99% hits is twice as good as 97%?**
  - Consider:  
cache hit time of 1 cycle  
miss penalty of 100 cycles
  - Average access time:  
97% hits:  $1 \text{ cycle} + 0.03 * 100 \text{ cycles} = 4 \text{ cycles}$   
99% hits:  $1 \text{ cycle} + 0.01 * 100 \text{ cycles} = 2 \text{ cycles}$
- **This is why “miss rate” is used instead of “hit rate”**

# Writing Cache Friendly Code

- **Make the common case go fast**
  - Focus on the inner loops of the core functions
- **Minimize the misses in the inner loops**
  - Repeated references to variables are good (**temporal locality**)
  - Stride-1 reference patterns are good (**spatial locality**)

**Key idea: Our qualitative notion of locality is quantified through our understanding of cache memories**