

Lecture 13

Cache Memories

CS213 – Intro to Computer Systems
Branden Ghen a – Spring 2021

Slides adapted from:

St-Amour, Hardavellas, Bustamente (Northwestern), Bryant, O'Hallaron (CMU), Garcia, Weaver (UC Berkeley)

Announcements

- Homework 3 due today
- Attack Lab due next week
 - Get started ASAP!
- Homework 4 releasing tomorrow
 - Will be the last homework
 - Focusing on cache parts including Thursday material

Today's Goals

- Discuss organization of various cache designs
 - Direct-mapped caches
 - N-way set-associative caches
 - Fully-associative caches
- Understand how cache memories are used to reduce the average time to access memory

Outline

- **Cache Organization**
- Associativity
- Cache Performance

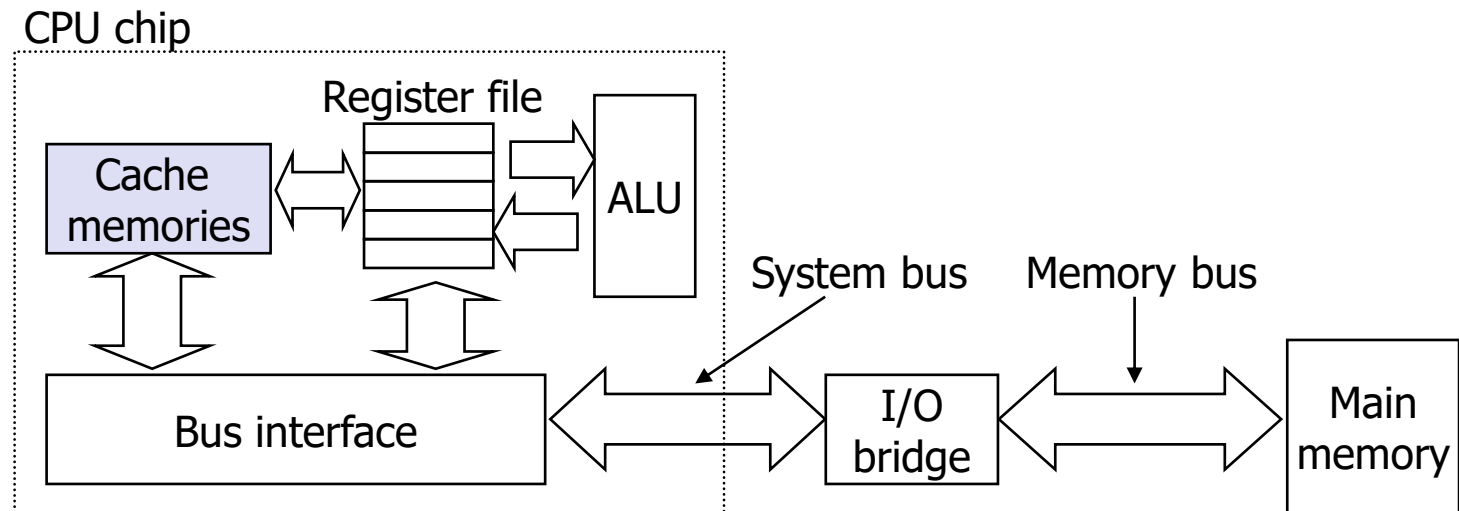
Caching speeds up code

- Cache: smaller, faster storage device that keeps copies of a subset of the data in a larger, slower device
 - If the data we access is already in the cache, we win!
 - Can get access time of faster memory, with overall capacity of larger
- Locality helps predict which data code is likely to access
 - So want to design caches to take advantage of it!
 - **Spatial locality**: if you need a byte, you're likely to need its neighbors
 - Caches should load whole blocks, not single bytes!
 - **Temporal locality**: if you need a byte, you're likely to need it again
 - Caches should try to keep recently cached data in the cache!
 - Most code has good locality
 - Well-written code has great locality!

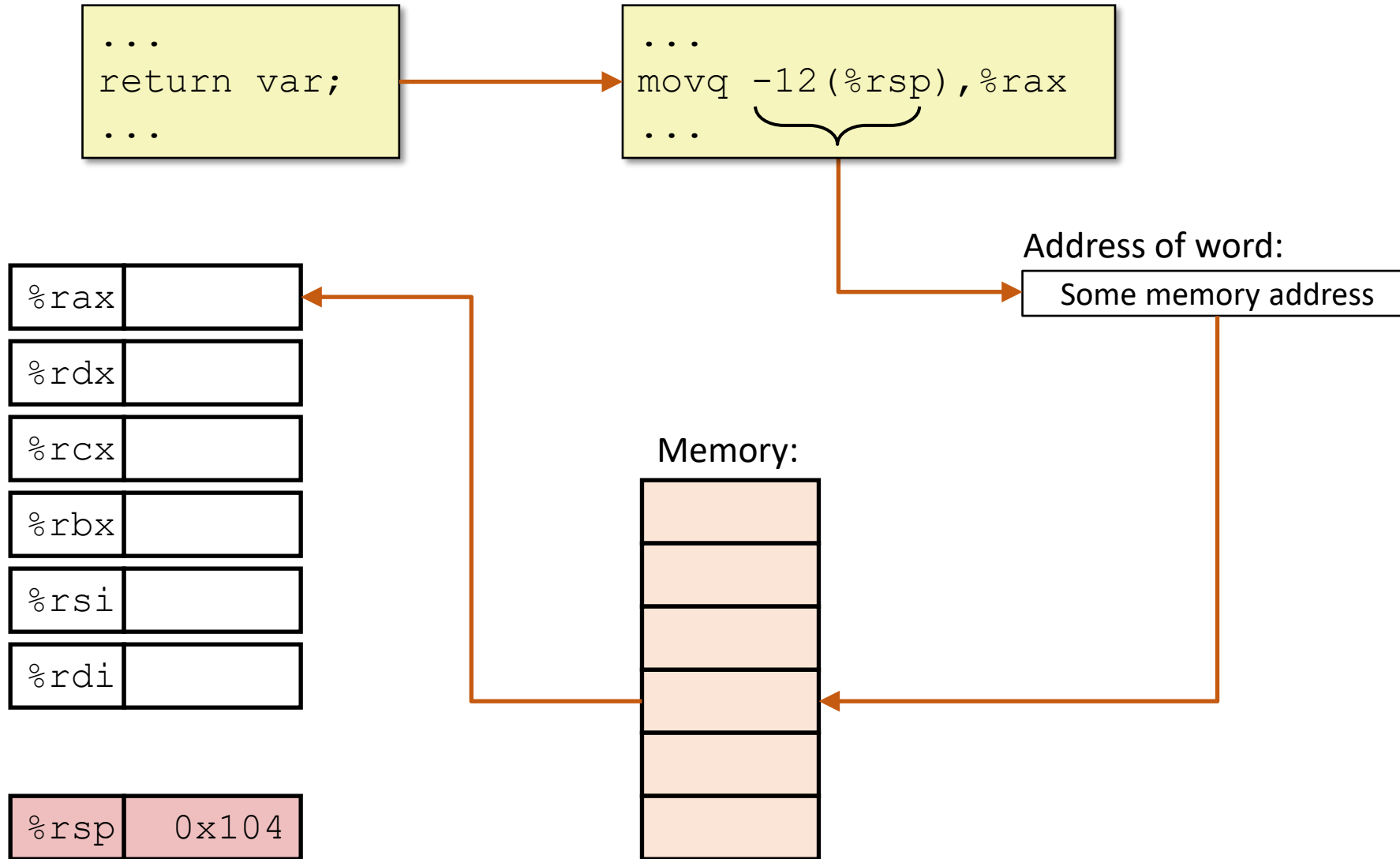
Cache memories

- A specific instance of the general principle of caching
 - Small, fast SRAM-based memories between CPU and main memory
 - Can include multiple levels
 - L1 = small, but really fast, L2 = larger, slower, L3, etc.
- CPU looks for data in caches first
 - e.g., L1, then L2, then L3, then finally in main memory as a last resort
- Mechanisms we'll see today are implemented in *hardware*

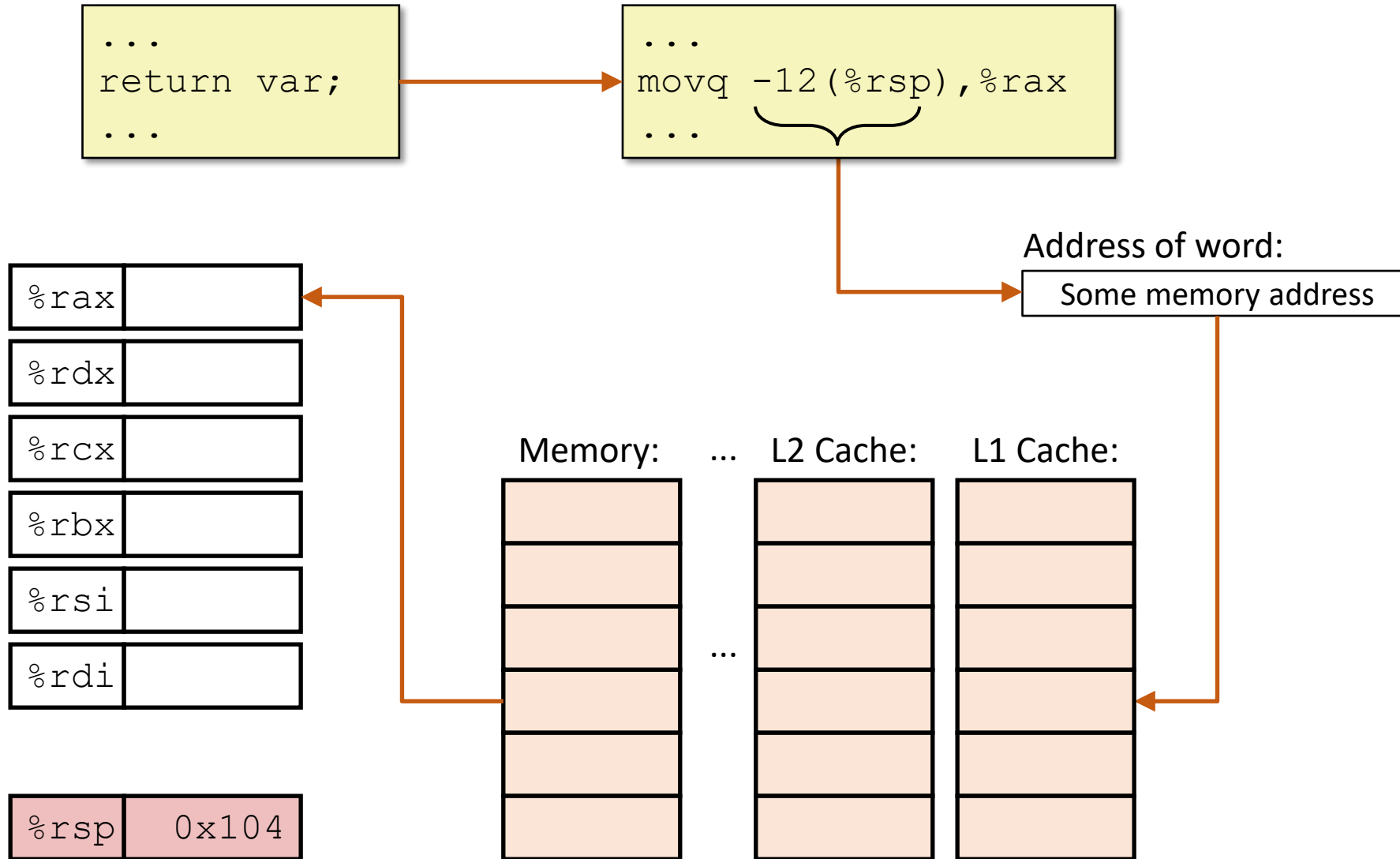
- Typical system structure:



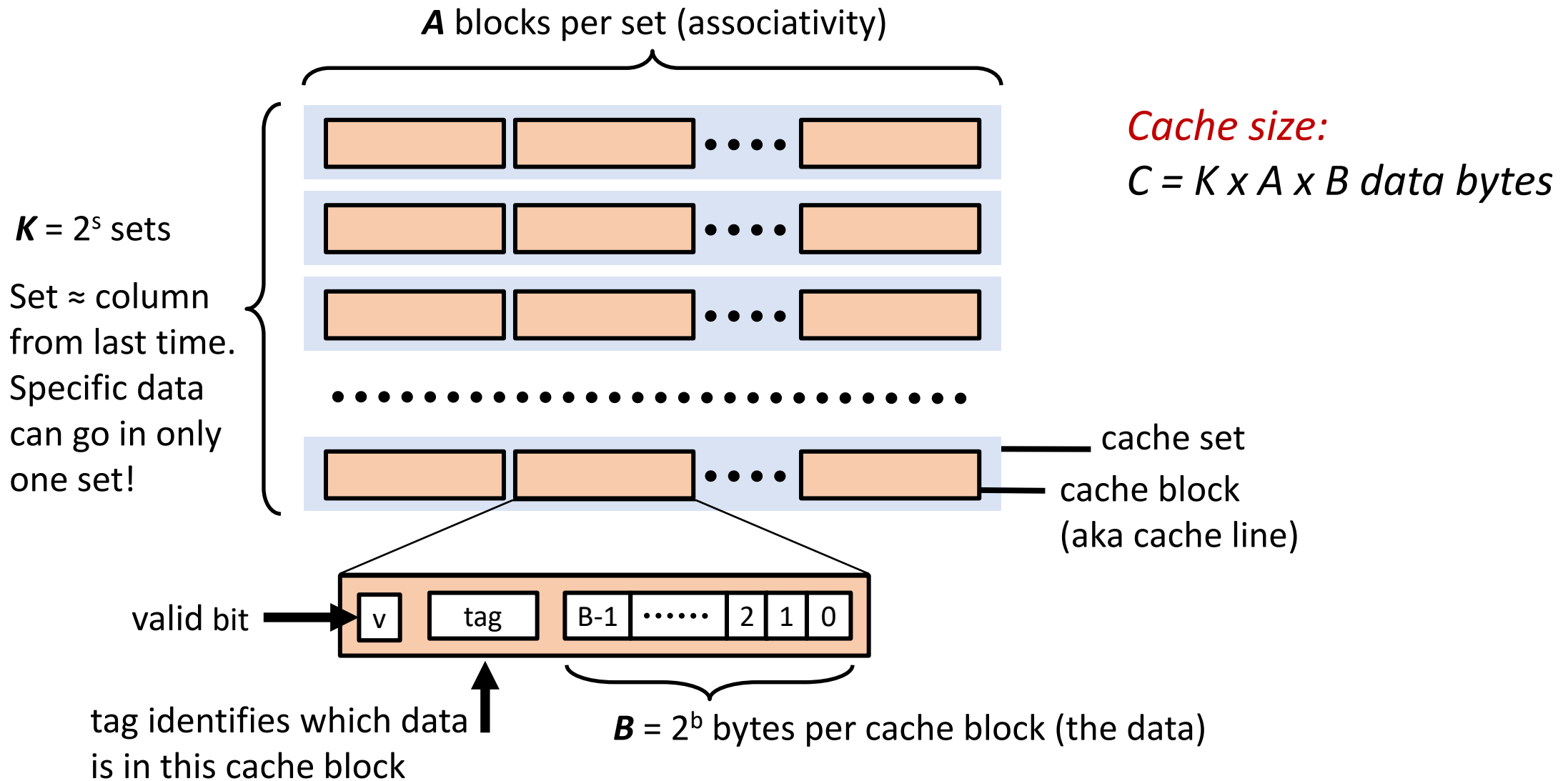
How You Thought a Memory Access Worked



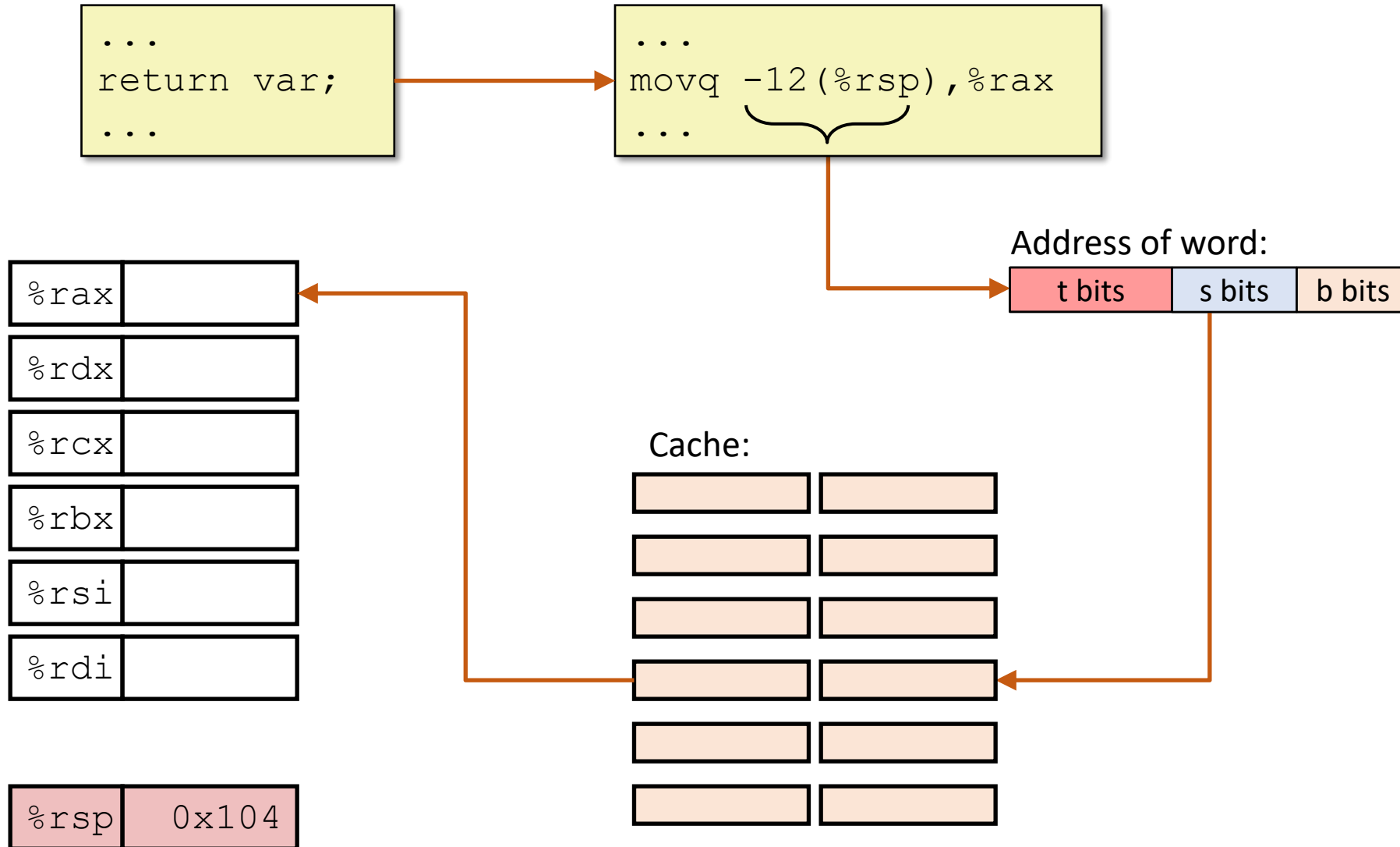
How a Memory Access Actually Works



General Cache Organization (S, A, B)

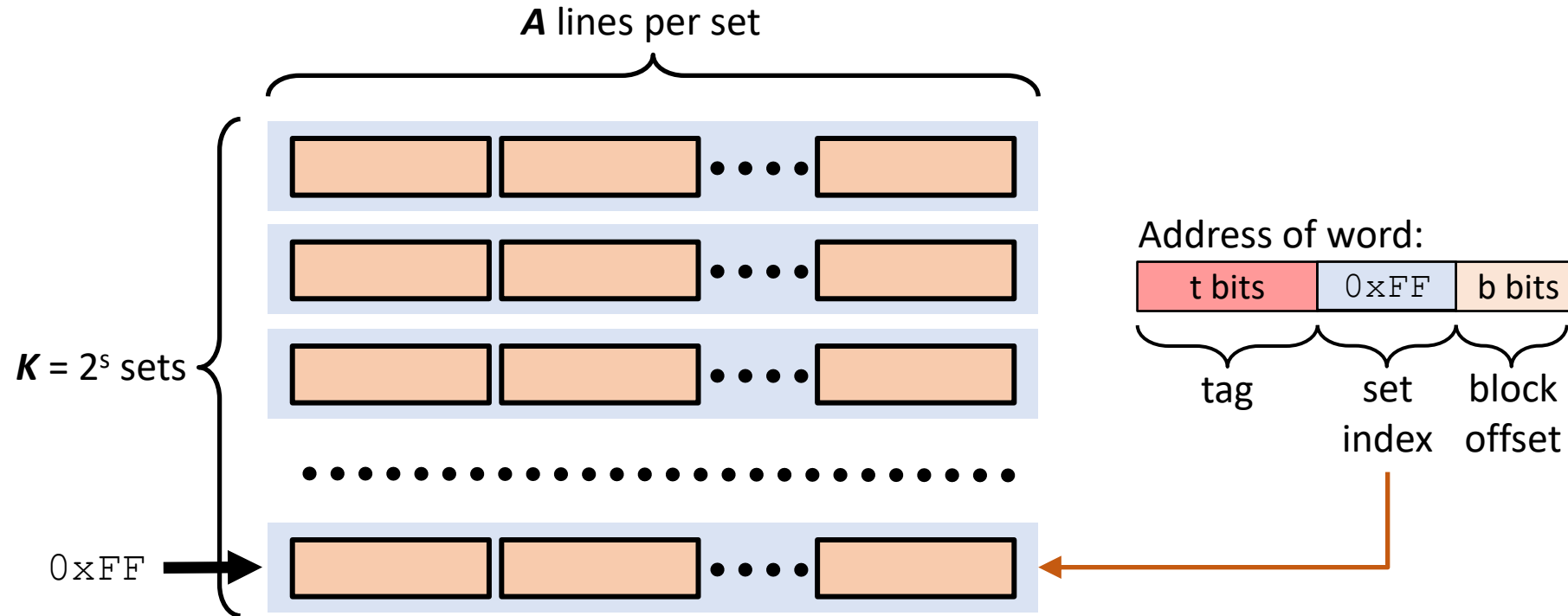


Cache Access



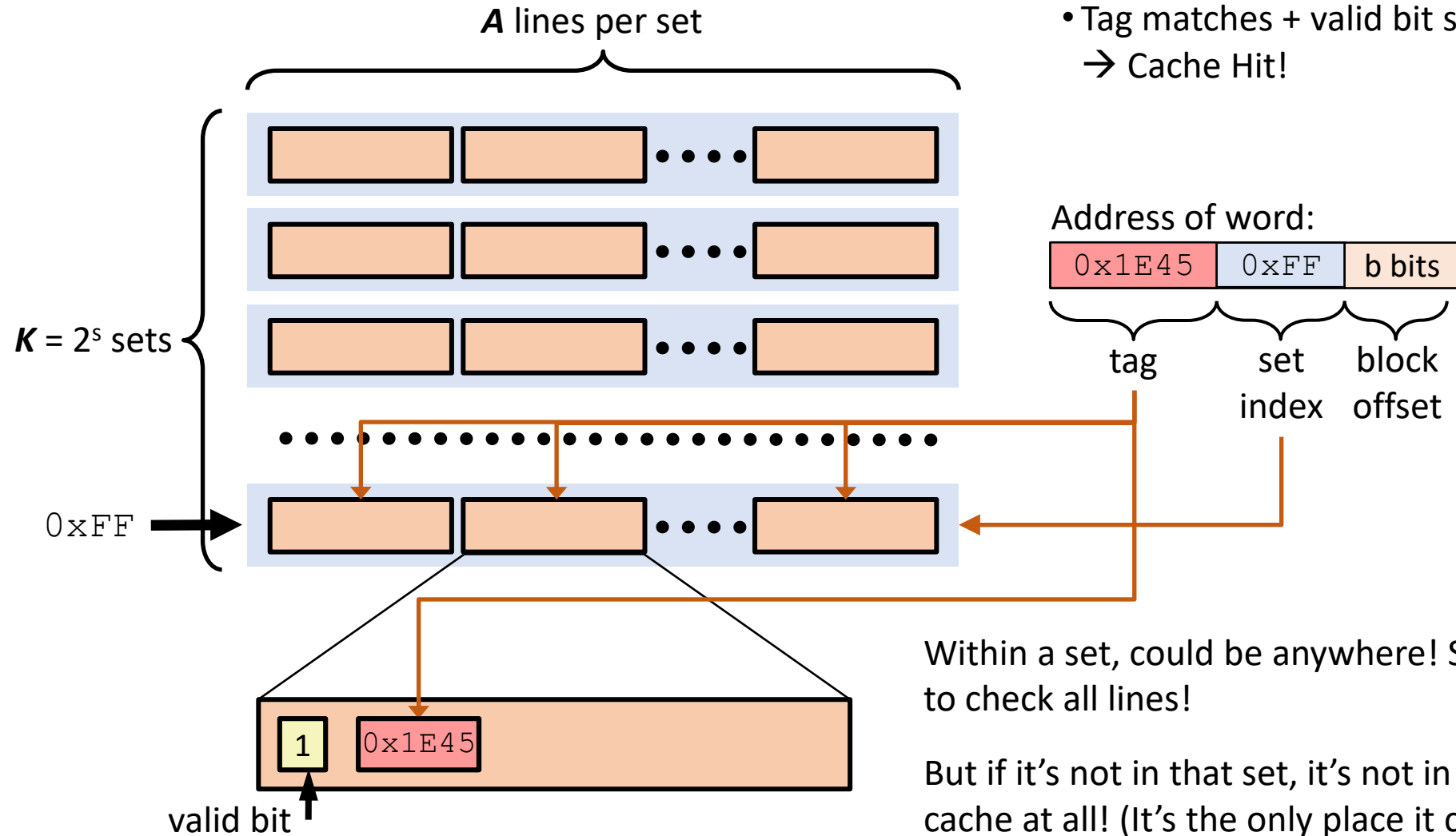
Cache Read (1): Locate Set

- Locate set



Each address maps to a particular set!
Data has to be stored at that particular set!
Even if that set is full and there “is space” elsewhere!
(That’s where conflict misses come from.)

Cache Read (2): Tag Match + Valid

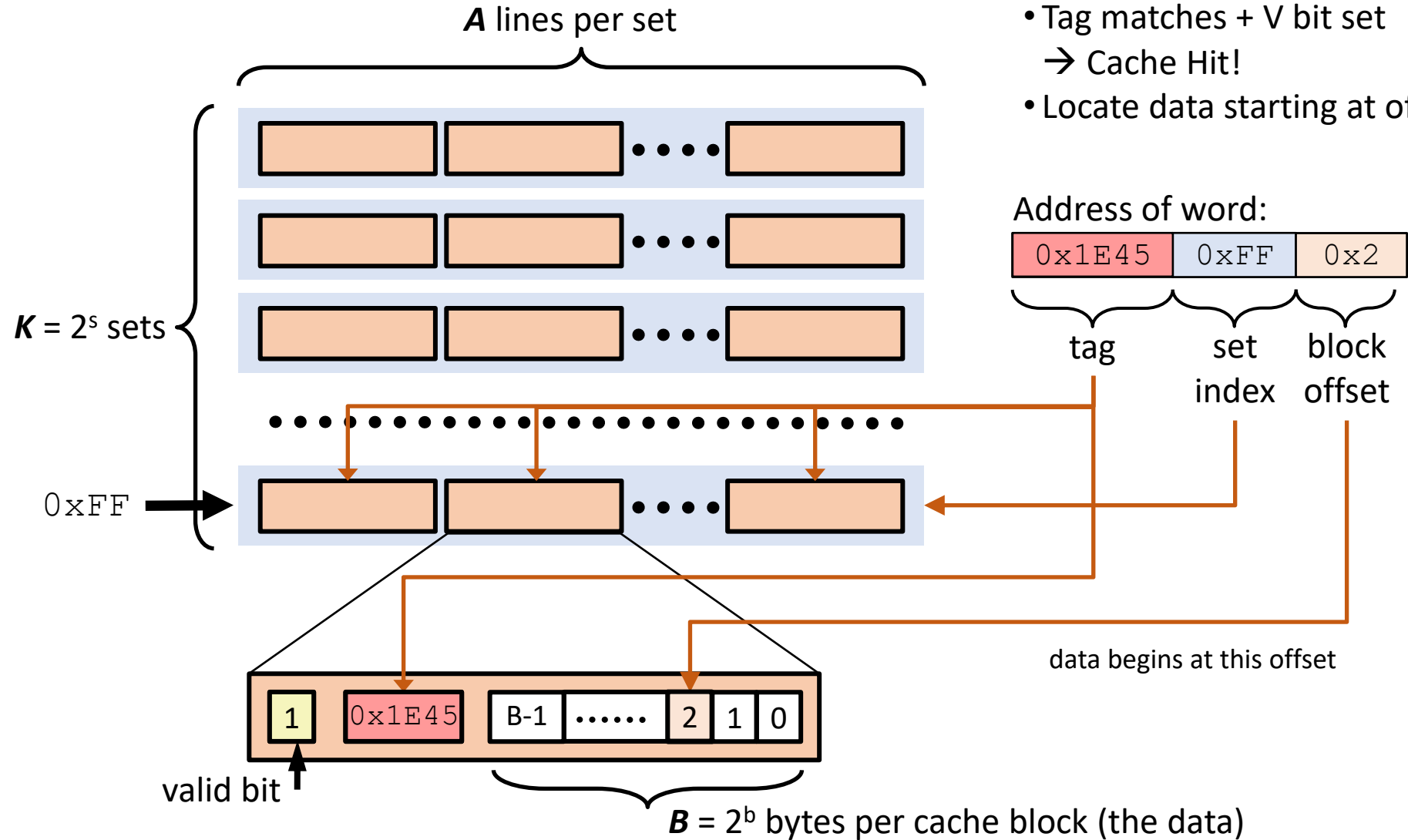


- Locate set
- Locate block in set
- Tag matches + valid bit set
→ Cache Hit!

Within a set, could be anywhere! So, need to check all lines!

But if it's not in that set, it's not in the cache at all! (It's the only place it could be.)

Cache Read (3): Block Offset

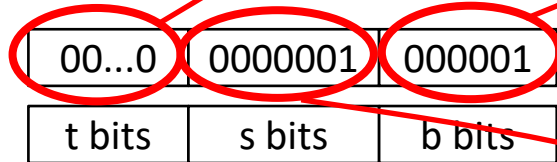


- Locate set
- Locate line in set
- Tag matches + V bit set
→ Cache Hit!
- Locate data starting at offset

Example: 128 sets, 64 bytes per block

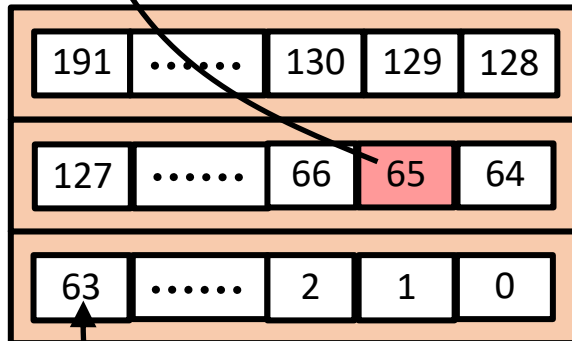
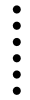
Goal: Get byte M[65] from cache

$$65_{10} = 100\ 0001_2$$

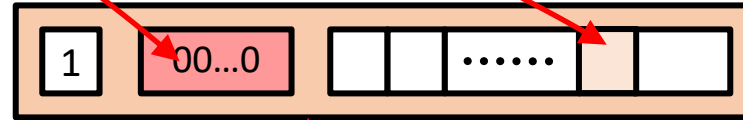


64 bytes per block \rightarrow $b = 6$ bits
 128 sets \rightarrow $s = 7$ bits
 remaining address bits \rightarrow t bits

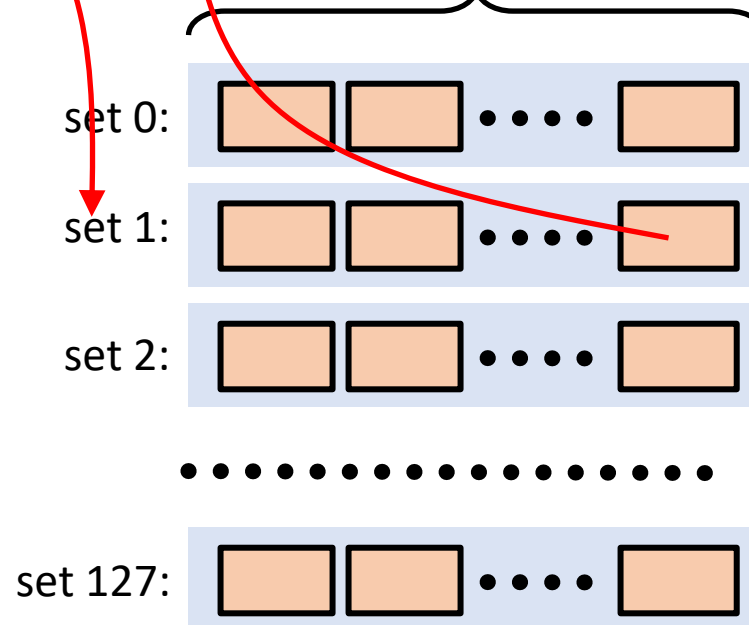
Memory:



address of a byte in memory



A lines per set

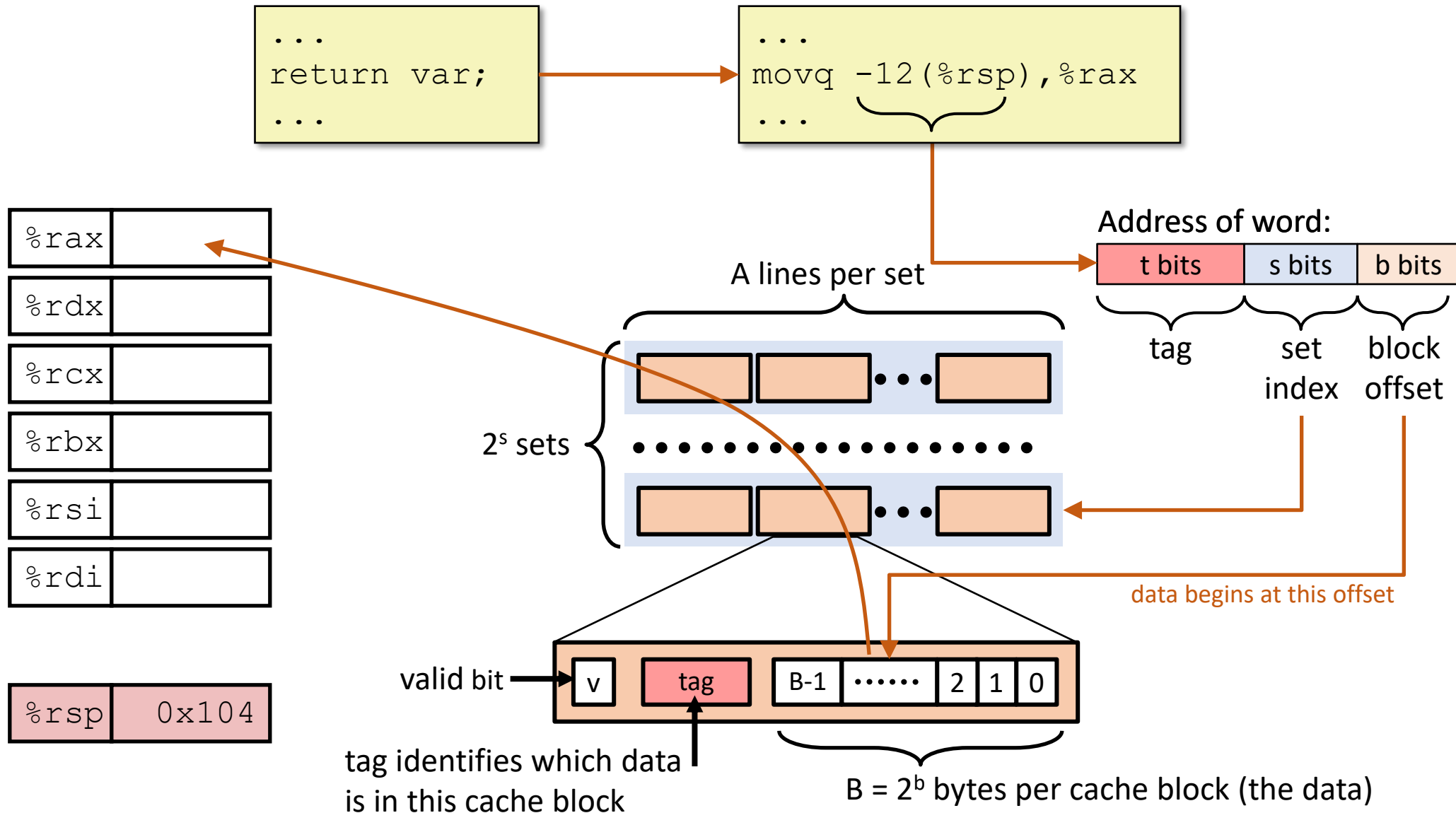


QUIZ #1: which set should we look in?

QUIZ #2: which tag are we looking for?

QUIZ #3: which byte within the block is the one that we want?

Cache access overview



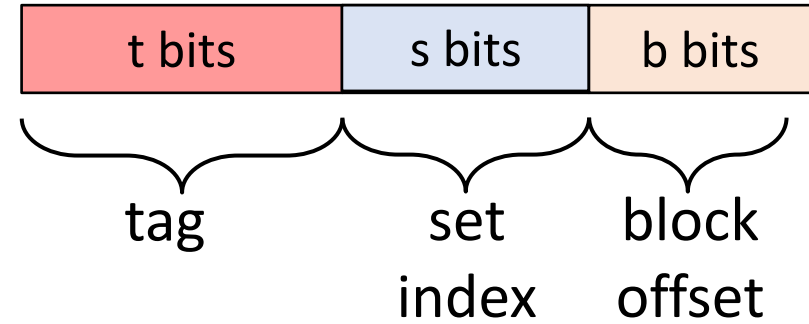
What about writes?

- Multiple copies of data exist:
 - L1, L2, Main Memory, Disk
 - Don't want them to get (or at least not to stay) out of sync!
 - Otherwise, who do you believe?
- 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)
 - We had an example of that last time
- 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 immediately to memory, doesn't bring into cache)
- Typical
 - **Write-back + Write-allocate** ← **by far the most common**
 - Write-through + No-write-allocate

Break + Question

- 64-bit, byte-addressed system
- 32 kB cache
 - 512 sets and 64-byte blocks
- How many bits for Tag?
 - A: 6 bits
 - B: 9 bits
 - C: 17 bits
 - D: 49 bits

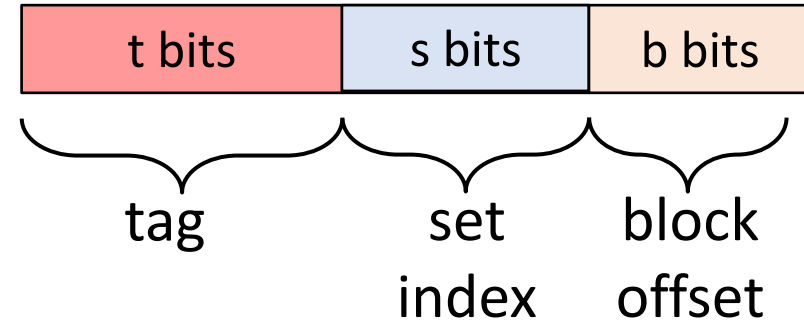
Address of word:



Break + Question

- 64-bit, byte-addressed system
- 32 kB cache
 - 512 sets and 64-byte blocks

Address of word:



- How many bits for Tag? (6 bits for block, 9 bits for set)
 - A: 6 bits
 - B: 9 bits
 - C: 17 bits
 - **D: 49 bits** (Tag is remaining bits. $64 - 6 - 9 = 49$)

Outline

- Cache Organization
- **Associativity**
- Cache Performance

Cache memory associativity

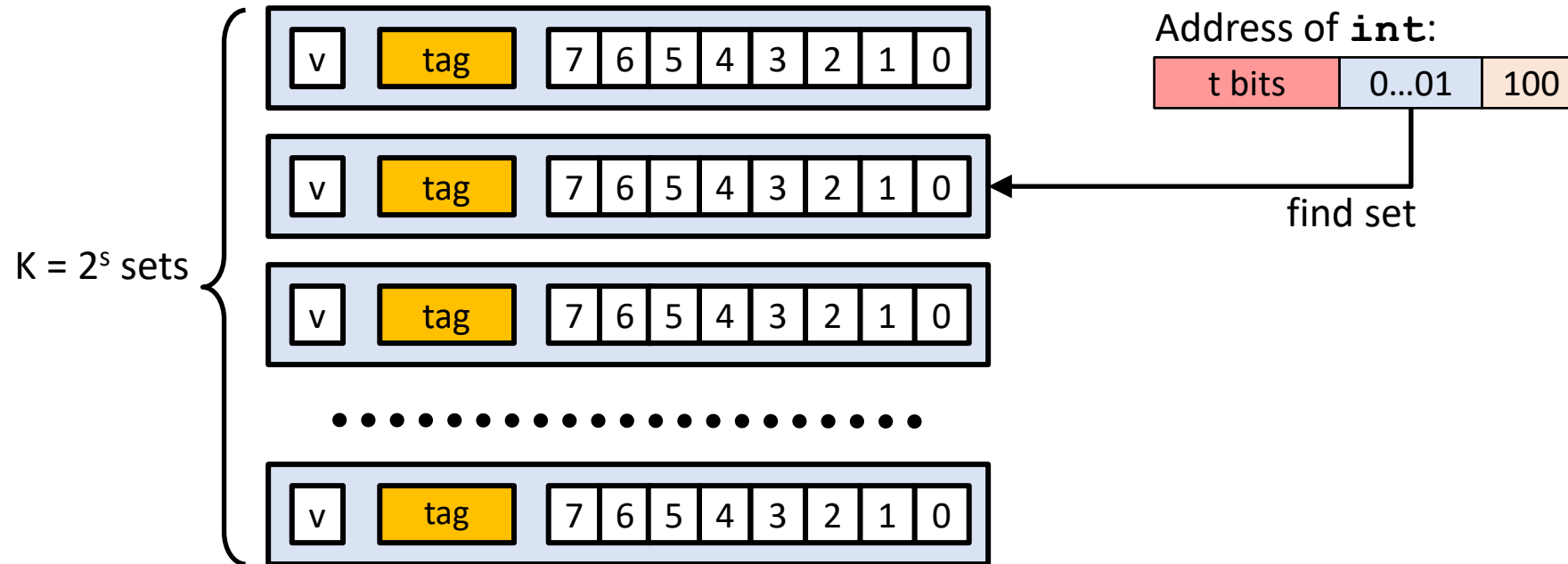
- When designing a cache, a number of parameters to choose
 - Total size (C), cache block size (B), number of sets (K), ...
- The most interesting one: associativity (A)
 - i.e., how many cache blocks per set
 - Has a significant impact on effectiveness (and complexity!)

Associativity choices

- Associativity 1 → **direct-mapped caches**
 - One cache block per set, blocks can only go in that one block
 - Whenever we place data in a set, must evict whatever is there
- Associativity >1 → **set-associative caches**
 - Can keep multiple cache blocks that would map to the same set
- Single set → **fully-associative caches**
 - Any cache block can go anywhere, 1 big set, tag is all that matters
 - Very rare for cache memories due to expensive hardware

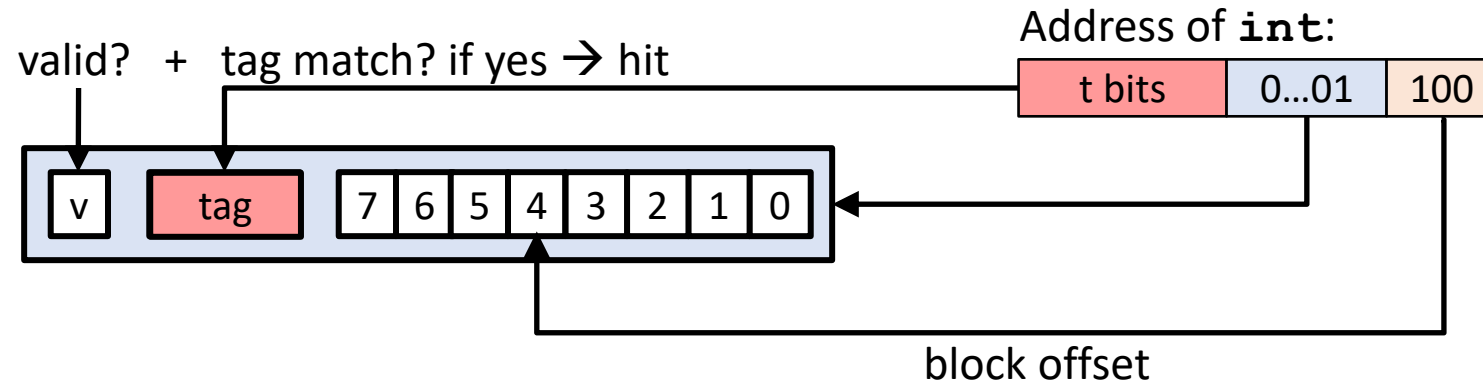
Direct-mapped cache (associativity = 1)

Direct mapped: One line per set
Assume: cache block size 8 bytes



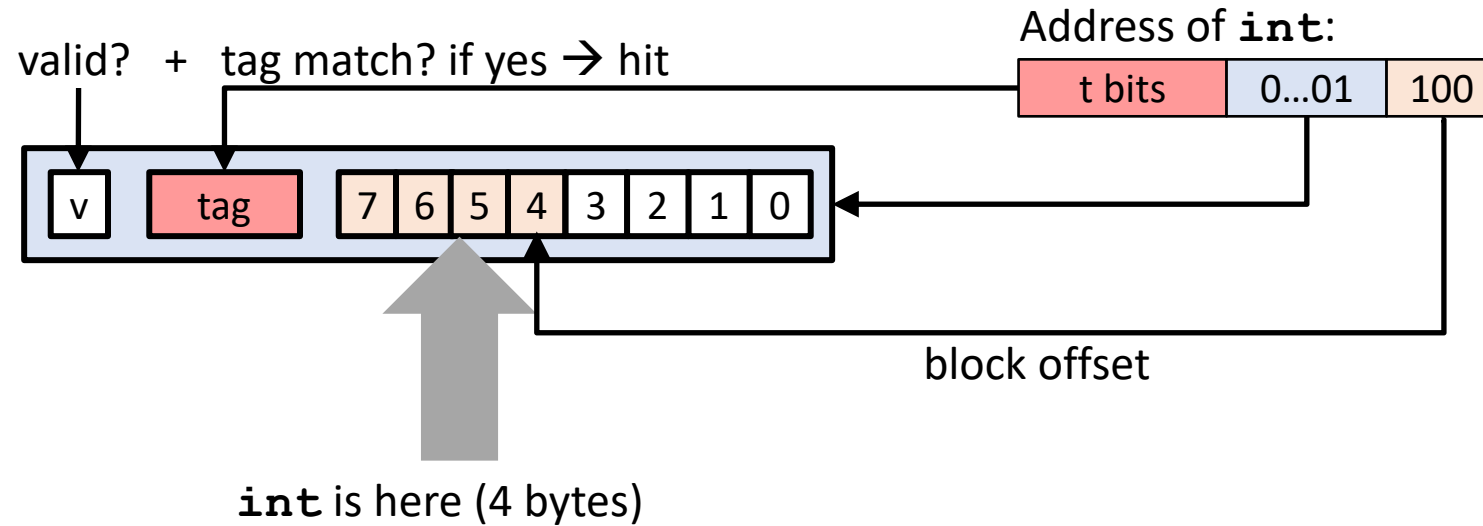
Direct-mapped cache (associativity = 1)

Direct mapped: One line per set
Assume: cache block size 8 bytes



Direct-mapped cache (associativity = 1)

Direct mapped: One line per set
Assume: cache block size 8 bytes



Direct-mapped cache simulation

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

Address trace

(reads, one byte per read):

0 [0 00 0₂] miss

1 [0 00 1₂] ~~hit~~

7 [0 11 1₂] miss

8 [1 00 0₂] miss

0 [0 00 0₂] miss

	v	tag	block	
set 00 ₂	0	0	m[9]	m[8]
set 01 ₂	0			
set 10 ₂	0			
set 11 ₂	0	0	m[7]	m[6]

M=16 addresses,
 byte-addressable
 B=2 bytes/block
 K=4 sets
 A=1 blocks/set

Conflict misses:

There is “room” in the cache,
 but two blocks map to the same set;
 one evicts the other!

Pause for questions on direct-mapped caches

Associativity choices

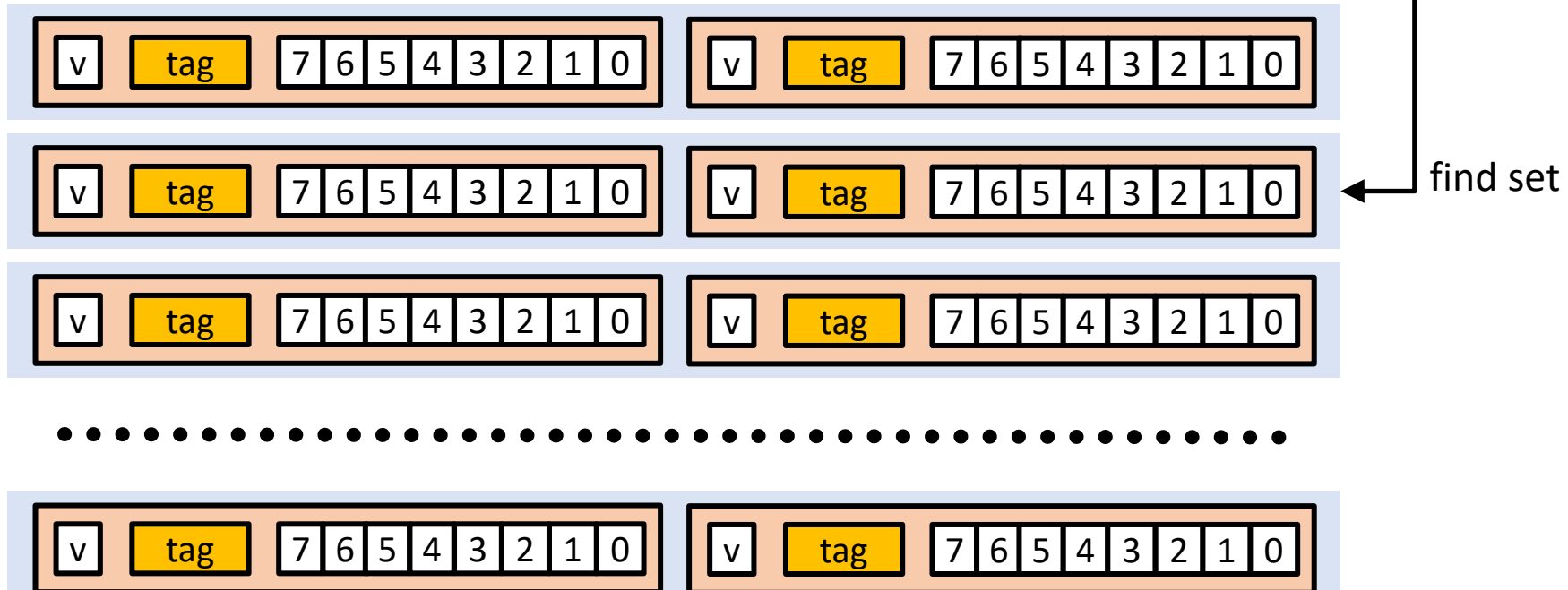
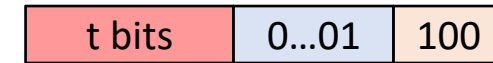
- Associativity 1 → **direct-mapped caches**
 - One cache block per set, blocks can only go in that one block
 - Whenever we place data in a set, must evict whatever is there
- Associativity >1 → **set-associative caches**
 - Can keep multiple cache blocks that would map to the same set
- Single set → **fully-associative caches**
 - Any cache block can go anywhere, 1 big set, tag is all that matters
 - Very rare for cache memories due to expensive hardware

2-way set-associative cache (associativity = 2)

A = 2: Two lines per set

Assume: cache block size 8 bytes

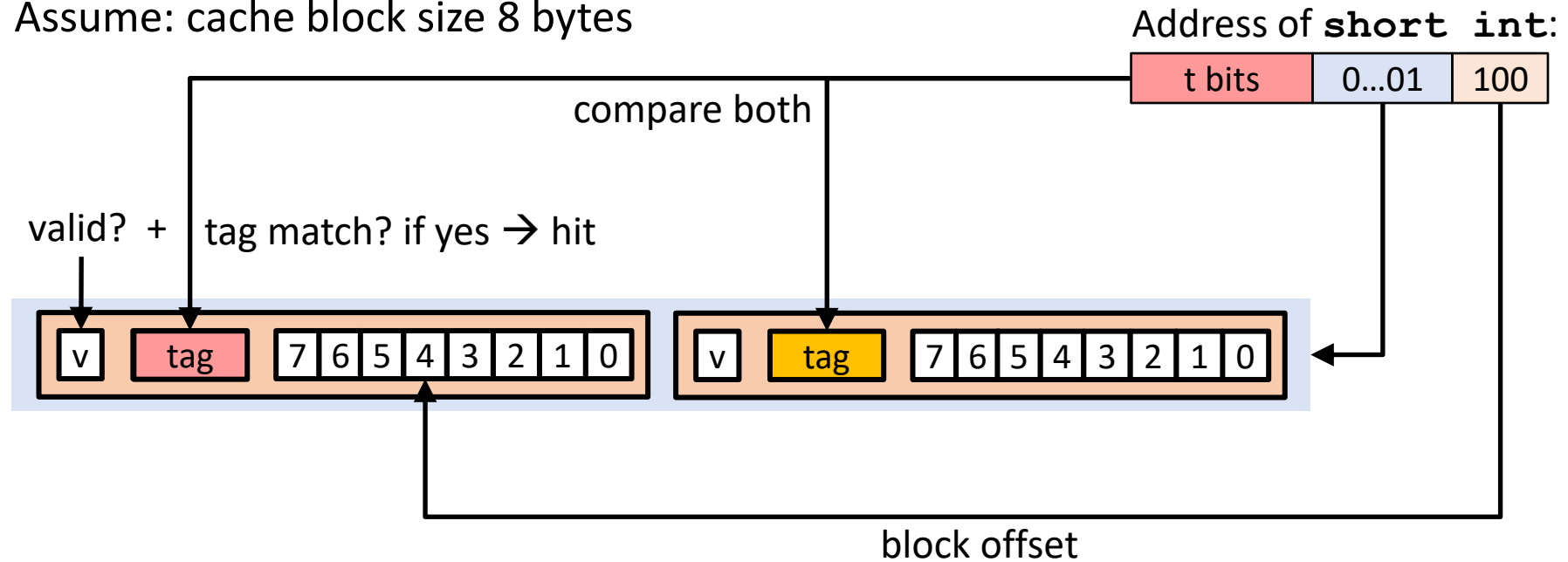
Address of **short int**:



2-way set-associative cache (associativity = 2)

A = 2: Two lines per set

Assume: cache block size 8 bytes

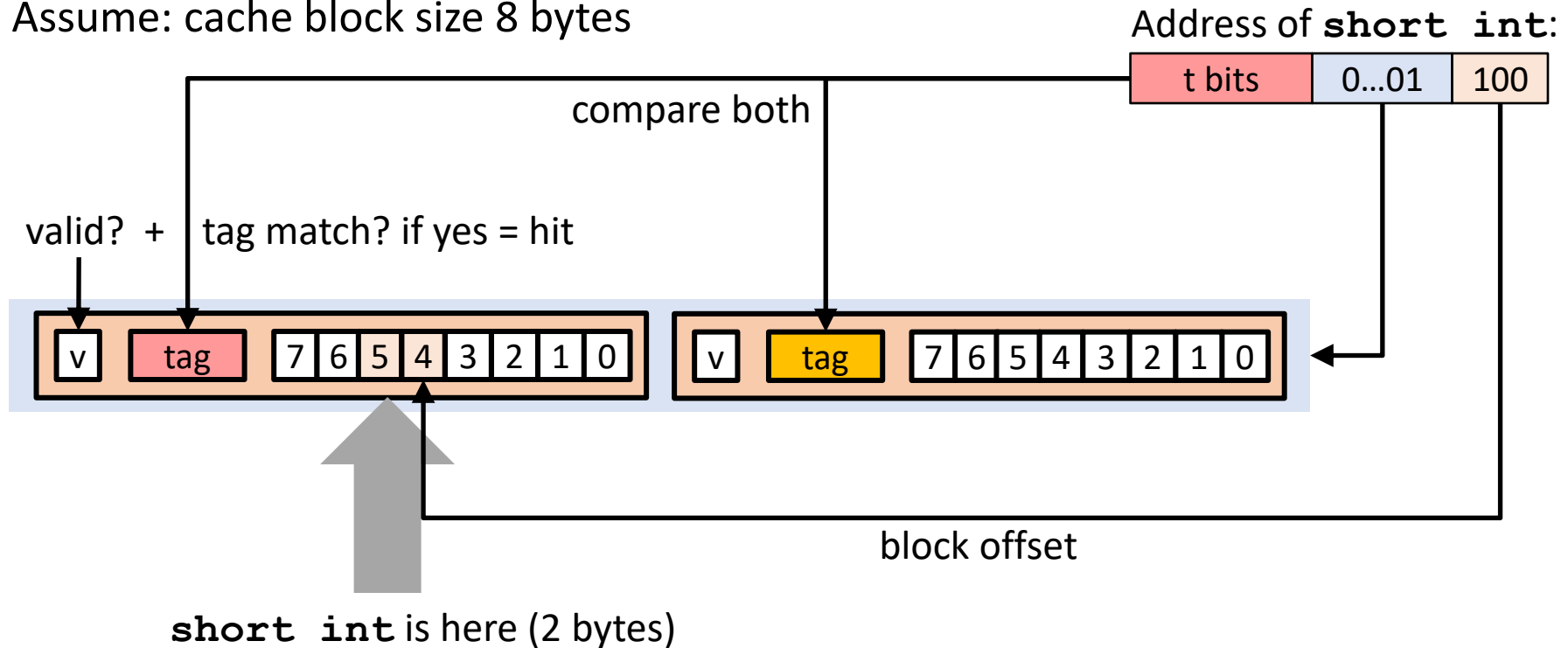


The data we want is either on the left, or on the right, or not in the cache at all. It can't be anywhere else! Addresses map to a single set!

2-way set-associative cache (associativity = 2)

A = 2: Two lines per set

Assume: cache block size 8 bytes



If no match:

- One line in set is selected for eviction and replacement
- Replacement policies: random, least recently used (LRU), ...
 - More clever → lower miss rate, but harder to implement in hardware

2-way set-associative cache simulation

M=16 addresses, byte-addressable,
B=2 bytes/block, K=2 sets, A=2 blocks/set

Same total size and block size as before.
Associativity (and thus # of sets) changed.

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

Address trace (reads, one byte per read):

0	[00 0 0 ₂]	miss
1	[00 0 1 ₂]	hit
7	[01 1 1 ₂]	miss
8	[10 0 0 ₂]	miss
0	[00 0 0 ₂]	hit

The same address sequence in the direct mapped cache resulted in:

miss
hit
miss
miss
miss

Higher associativity =
Less likely to have to evict!

Temporal locality: want data
in cache to *stay* in cache!

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

	v	Tag	Block
Set 1	1	01	M[7-6]
	0		

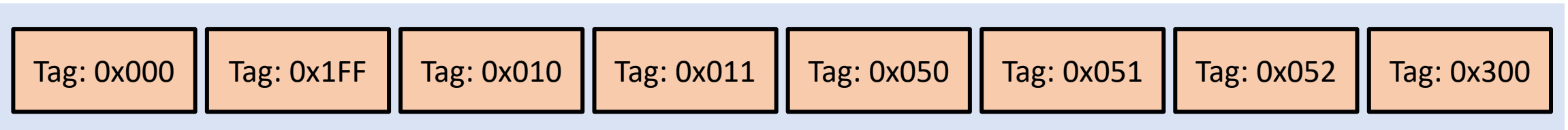
Pause for questions on set-associative caches

Fully-associative caches

- What changes with fully-associative caches?
 - Anything can go anywhere
 - Only one set ($s = 0$ bits)
- Otherwise, same steps as for a set-associative cache
 - Compare tag against all blocks in the set

Break + Question

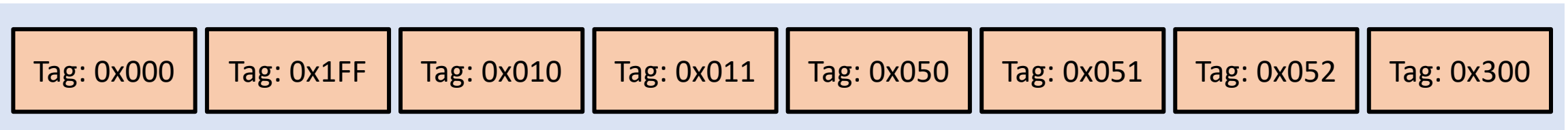
- Fully-associative cache on a 16-bit system
 - One set (fully associative!)
 - Eight, 64-byte blocks



- Are the following addresses in the cache?
 - 0x0400
 - 0x0410
 - 0xC002
 - 0xC048

Break + Question

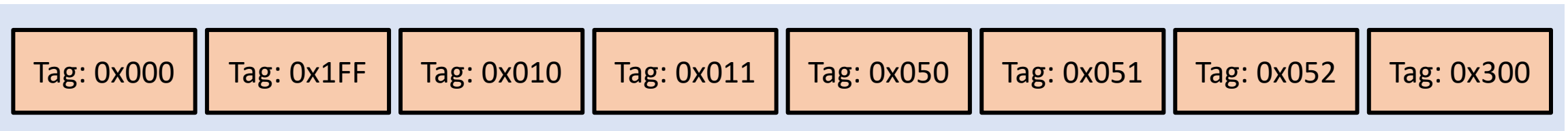
- Fully-associative cache on a 16-bit system
 - One set (fully associative!)
 - Eight, 64-byte blocks



- Are the following addresses in the cache?
 - 0x0400 \Rightarrow 0b0000 0100 0000 0000
 - 0x0410 \Rightarrow 0b0000 0100 0001 0000
 - 0xC002 \Rightarrow 0b1100 0000 0000 0010
 - 0xC048 \Rightarrow 0b1100 0000 0100 1000

Break + Question

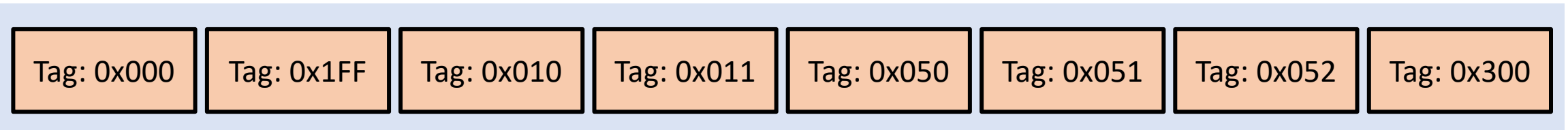
- Fully-associative cache on a 16-bit system
 - One set (fully associative!)
 - Eight, 64-byte blocks



- Are the following addresses in the cache?
 - $0x0400 \Rightarrow 0b\underline{0000\ 0100\ 0000\ 0000}$ → Tag 0x010 **HIT**
 - $0x0410 \Rightarrow 0b\underline{0000\ 0100\ 0001\ 0000}$ → Tag 0x010 (same block!) **HIT**
 - $0xC002 \Rightarrow 0b\underline{1100\ 0000\ 0000\ 0010}$
 - $0xC048 \Rightarrow 0b\underline{1100\ 0000\ 0100\ 1000}$

Break + Question

- Fully-associative cache on a 16-bit system
 - One set (fully associative!)
 - Eight, 64-byte blocks



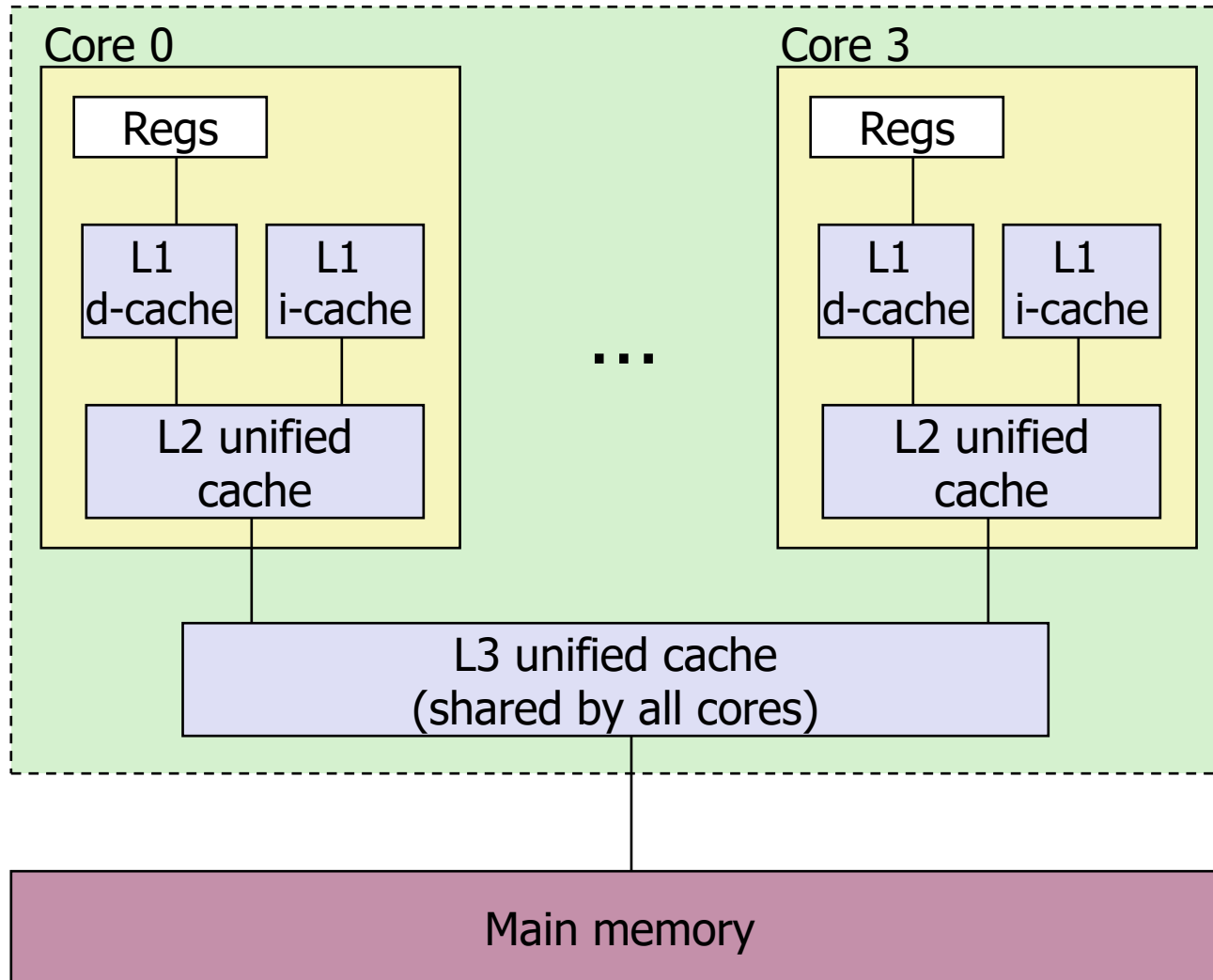
- Are the following addresses in the cache?
 - $0x0400 \Rightarrow 0b\underline{0000\ 0100\ 0000\ 0000}$ → Tag 0x010 **HIT**
 - $0x0410 \Rightarrow 0b\underline{0000\ 0100\ 0001\ 0000}$ → Tag 0x010 (same block!) **HIT**
 - $0xC002 \Rightarrow 0b\underline{1100\ 0000\ 0000\ 0010}$ → Tag 0x300 **HIT**
 - $0xC048 \Rightarrow 0b\underline{1100\ 0000\ 0100\ 1000}$ → Tag 0x301 (different block!) **MISS**

Associativity Pros and Cons

- Direct-mapped
 - Simplest to implement: look-up compares tag with 1 cache line
→ requires fewer transistors, which can be used elsewhere on the chip
 - Conflicts can easily lead to *thrashing*
 - Two cache lines map to the same set, program needs both, and they keep kicking each other out of the cache. Lots of misses. Bad times.
- Set-associative
 - More complex implementation: requires more (HW) tag comparators
 - Lower miss rate than direct-mapped caches (fewer conflict misses)
 - 2-way is a significant improvement over direct-mapped
 - 4-way is a more modest improvement over 2-way, and so on
- Fully-associative
 - One comparator per cache line in the cache. Ouch.
 - Often a deal-breaker for hardware
 - Very low miss rate!

Intel Core i7 Cache Hierarchy

Processor package



L1 i-cache and d-cache:

32 KB, 8-way,

Access: 4 cycles

Keep separate caches for instructions and data. Don't want them to step on each other's toes!

L2 unified cache:

256 KB, 8-way,

Access: 11 cycles

L3 unified cache:

8 MB, 16-way,

Access: 30-40 cycles

Last resort before going to main memory (slow!) So want this large and highly-associative, to have very few misses.

Block size: 64 bytes for all caches.

Outline

- Cache Organization
- Associativity
- **Cache Performance**

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 dataset size, etc.
 - However, many applications have $>30\%$ miss rate in L2 cache
- 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:
 - 1-2 clock cycles for L1
 - 5-20 clock cycles for L2
- Miss Penalty
 - Additional time required because of a miss
 - Typically 50-200 cycles for main memory

Let's think about those numbers

- Huge difference between a hit and a miss
 - Could be 100x, if comparing L1 and main memory
- Would you believe a 99% hit rate is twice as good as 97%?
 - Consider:
cache hit time of 1 cycle
miss penalty of 100 cycles
 - Average access time:
97% hits: 100 instructions: 100 cycles (1 per instruction) + 3*100 (misses)
on average: 1 cycle/instr. + 0.03 * 100 cycles/instr. = **4 cycles/instr.**
99% hits: on average: 1 cycle/instr. + 0.01 * 100 cycles/instr. = **2 cycles/instr.**
- This is why "miss rate" is used instead of "hit rate"
 - In our example, 1% miss rate vs. 3% miss rate
 - Brings the radical performance difference across better
- "Computation is what happens between cache misses."

Average Memory Access Time (AMAT)

- $AMAT = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty}$
 - Generalization of previous formula
- Can extend for multiple layers of caching
 - $AMAT = \text{Hit Time L1} + \text{Miss Rate L1} \times \text{Miss Penalty L1}$
 - $\text{Miss Penalty L1} = \text{Hit Time L2} + \text{Miss Rate L2} \times \text{Miss Penalty L2}$
 - $\text{Miss Penalty L2} = \text{Hit Time Main Memory}$
- Multi-level caching helps minimize AMAT

Outline

- Cache Organization
- Associativity
- Cache Performance