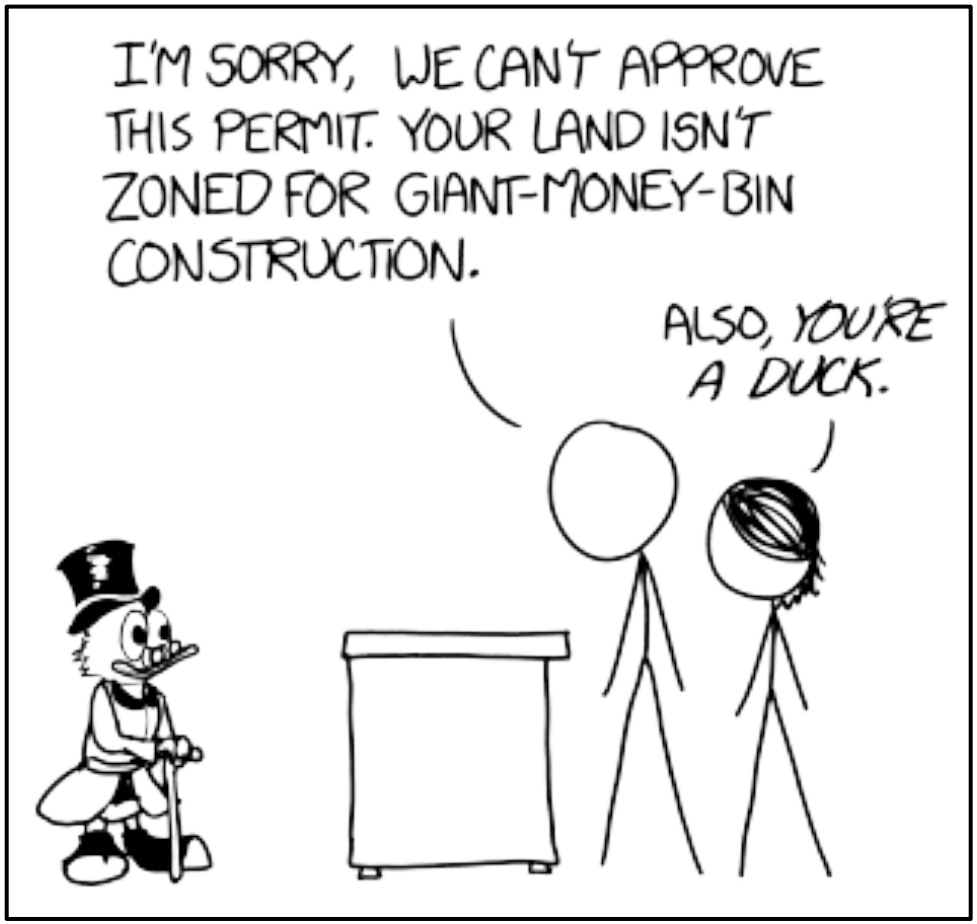
Caches III



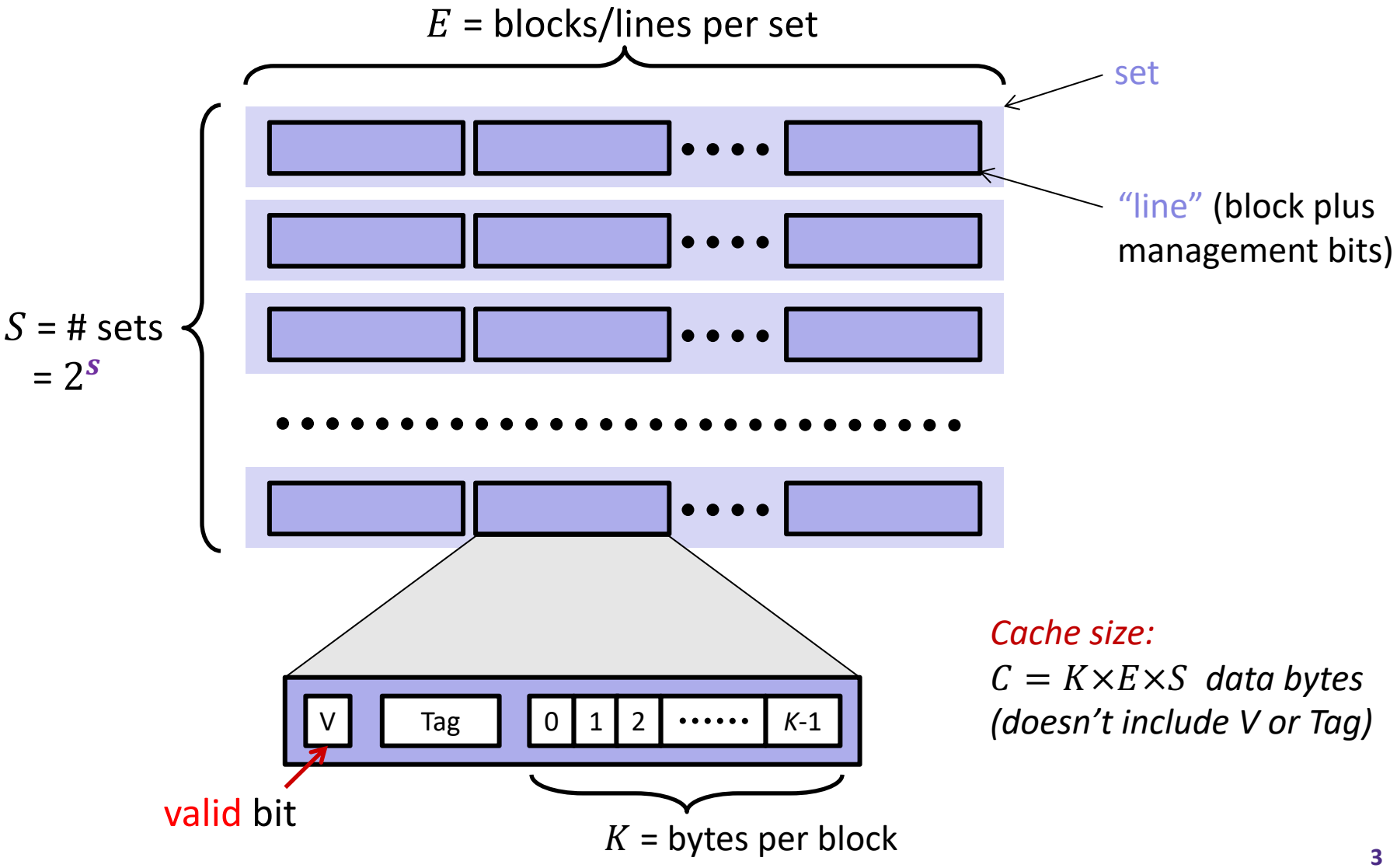


https://what-if.xkcd.com/111/

Making memory accesses fast!

- Cache basics
- Principle of locality
- Memory hierarchies
- Cache organization
  - Direct-mapped (*sets*; index + tag)
  - **Associativity (*ways*)**
  - **Replacement policy**
  - **Handling writes**
- Program optimizations that consider caches

General Cache Organization (S, E, K)



Cache size:
$C = K \times E \times S$ *data bytes*
(doesn't include V or Tag)

Notation Review

- We just introduced a lot of new variable names!
  - Please be mindful of block size notation when you look at past exam questions or are watching videos

Variable	This Quarter	Formulas
Block size	K (B in book)	$M = 2^m \leftrightarrow m = \log_2 M$ $S = 2^s \leftrightarrow s = \log_2 S$ $K = 2^k \leftrightarrow k = \log_2 K$ $C = K \times E \times S$ $s = \log_2(C/K/E)$ $m = t + s + k$
Cache size	C	
Associativity	E	
Number of Sets	S	
Address space	M	
Address width	m	
Tag field width	t	
Index field width	s	
Offset field width	k (b in book)	

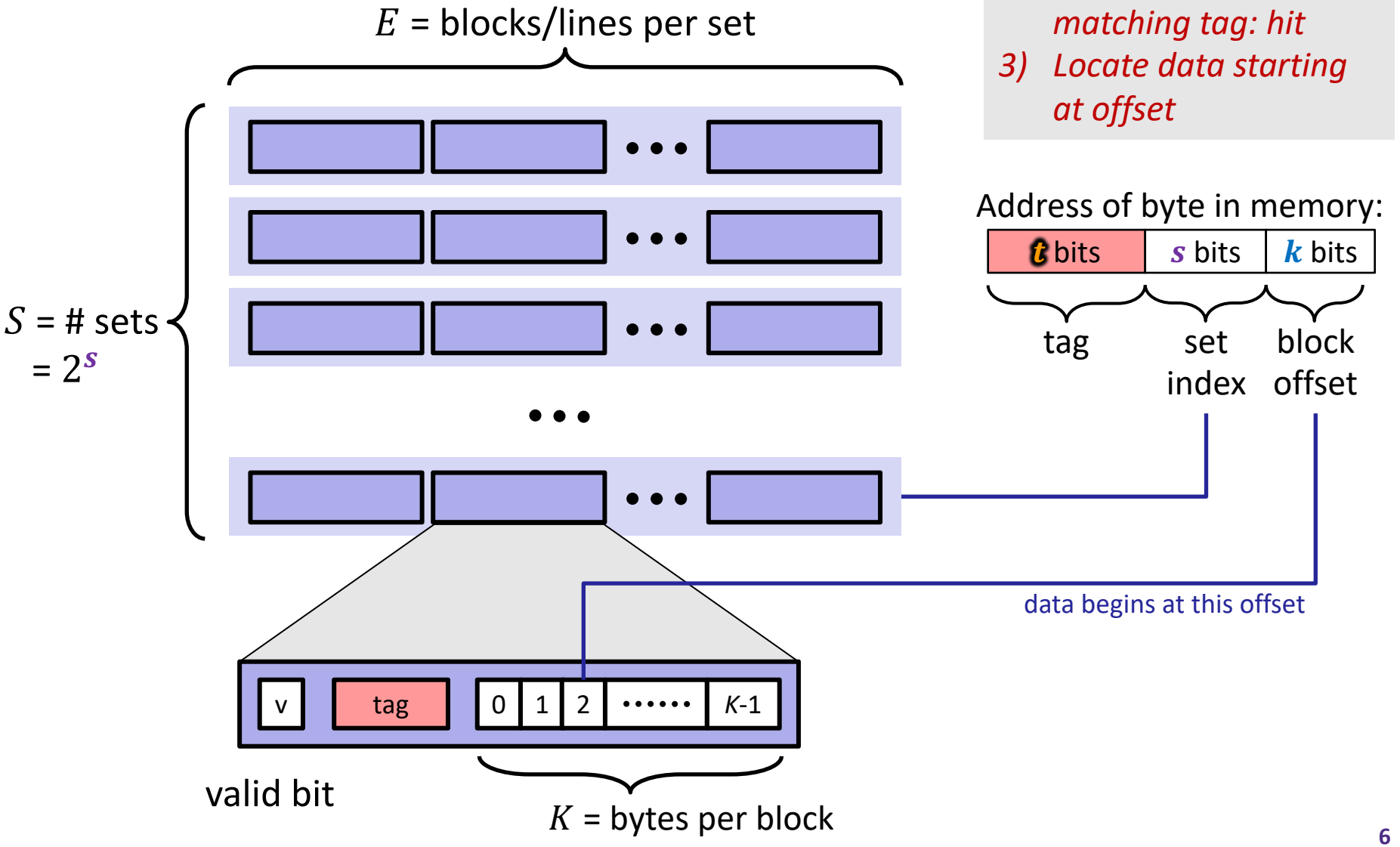
Example Cache Parameters Problem

- 4 KiB address space, 125 cycles to go to memory. Fill in the following table:

Cache Size	256 B
Block Size	32 B
Associativity	2-way
Hit Time	3 cycles
Miss Rate	20%
Tag Bits	
Index Bits	
Offset Bits	
AMAT	

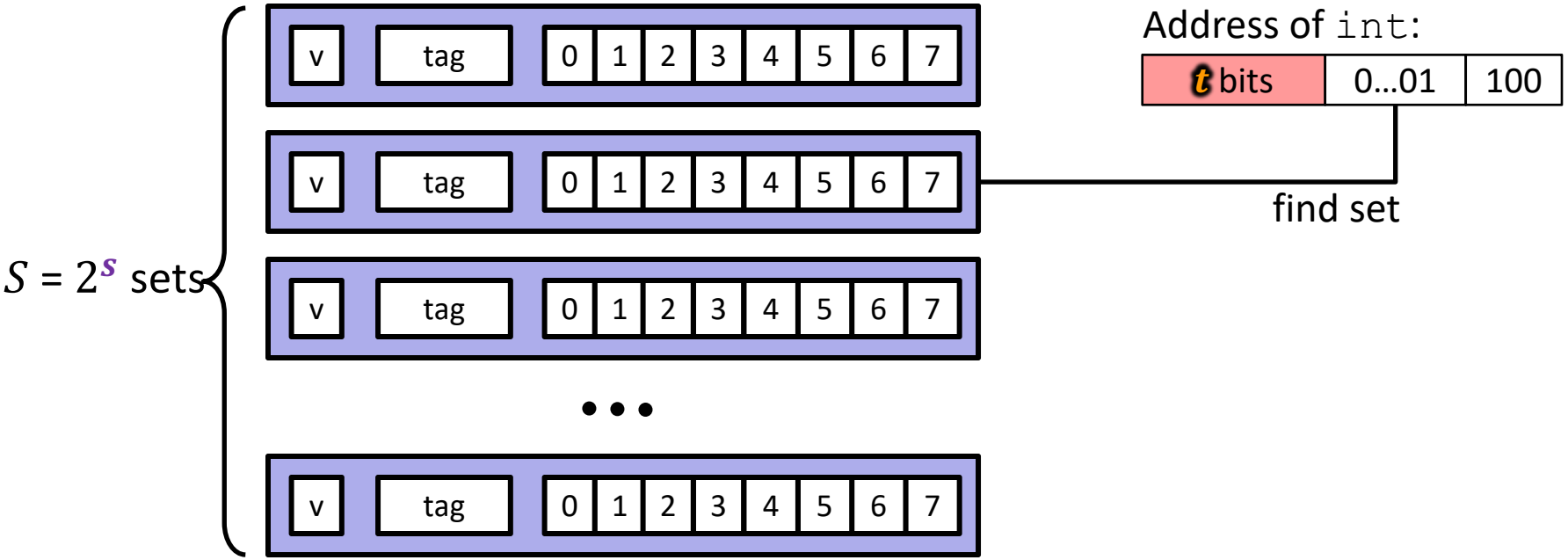
Cache Read

1) *Locate set*
2) *Check if any line in set is valid and has matching tag: hit*
3) *Locate data starting at offset*



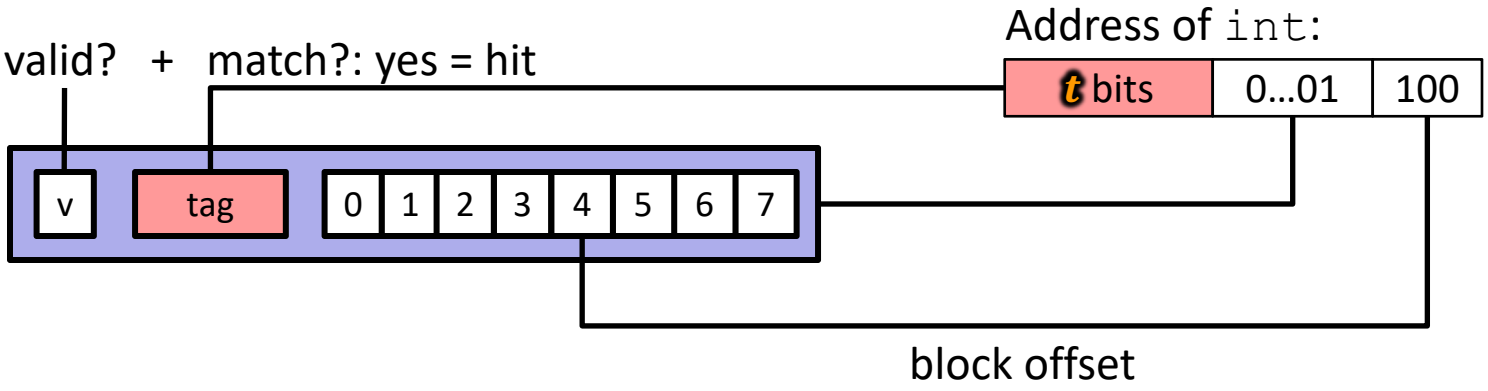
Example: Direct-Mapped Cache (E = 1)

Direct-mapped: One line per set
Block Size K = 8 B



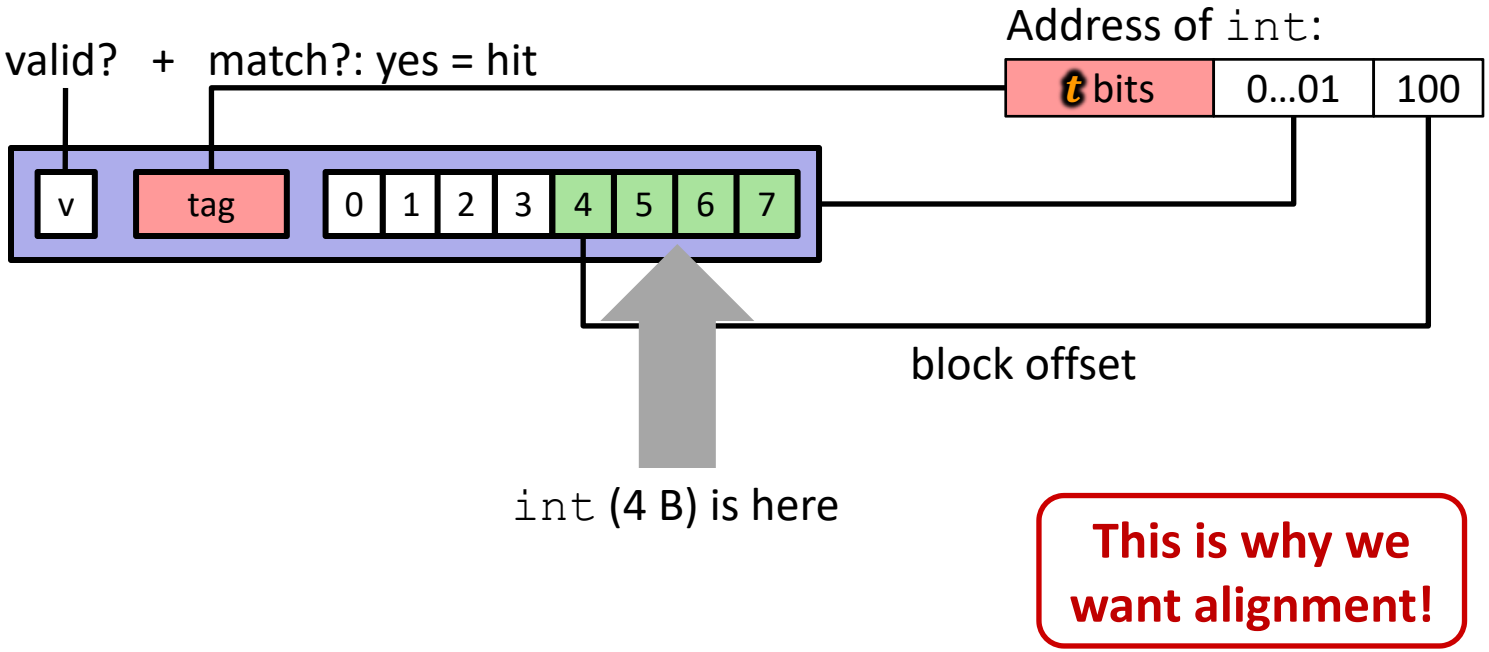
Example: Direct-Mapped Cache (E = 1)

Direct-mapped: One line per set
Block Size K = 8 B



Example: Direct-Mapped Cache (E = 1)

Direct-mapped: One line per set
Block Size K = 8 B

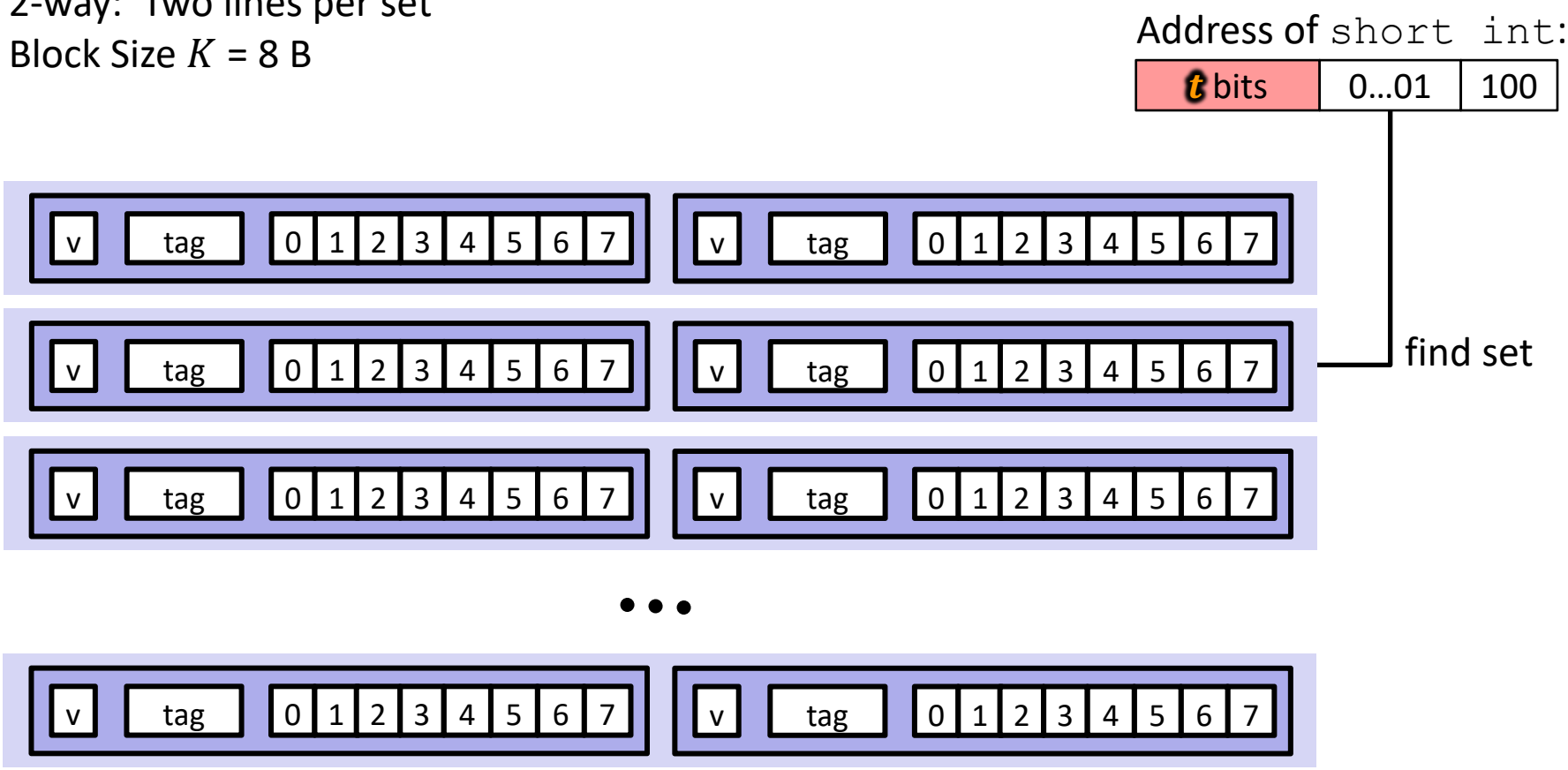


This is why we want alignment!

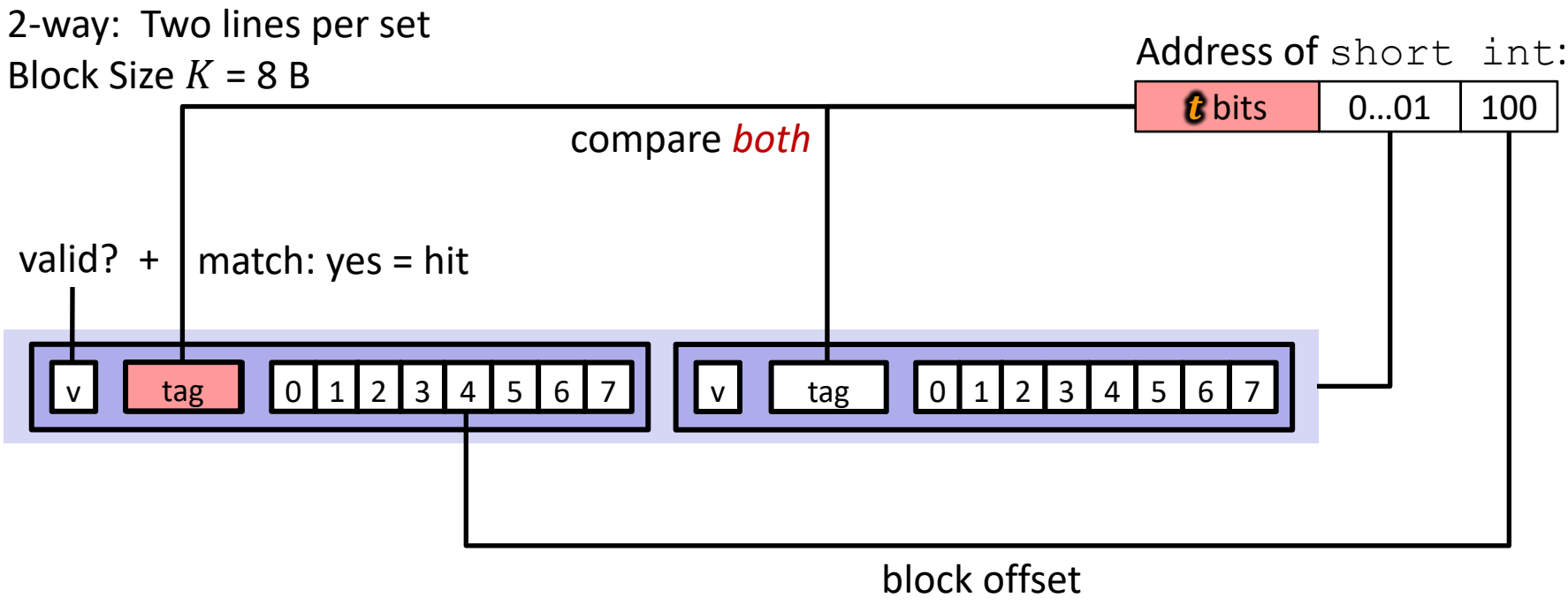
No match? Then old line gets evicted and replaced

Example: Set-Associative Cache (E = 2)

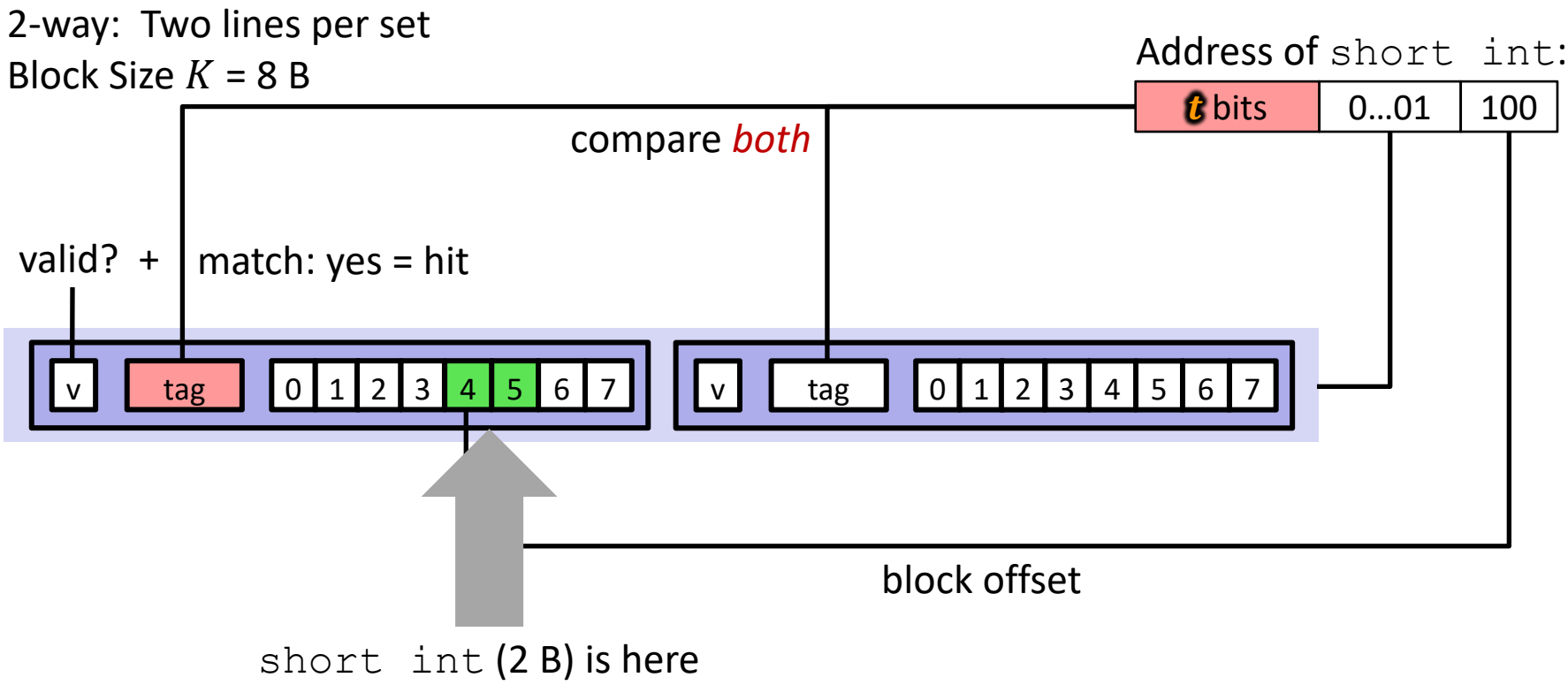
2-way: Two lines per set
Block Size K = 8 B



Example: Set-Associative Cache (E = 2)



Example: Set-Associative Cache (E = 2)



No match?

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

Example Code Analysis Problem

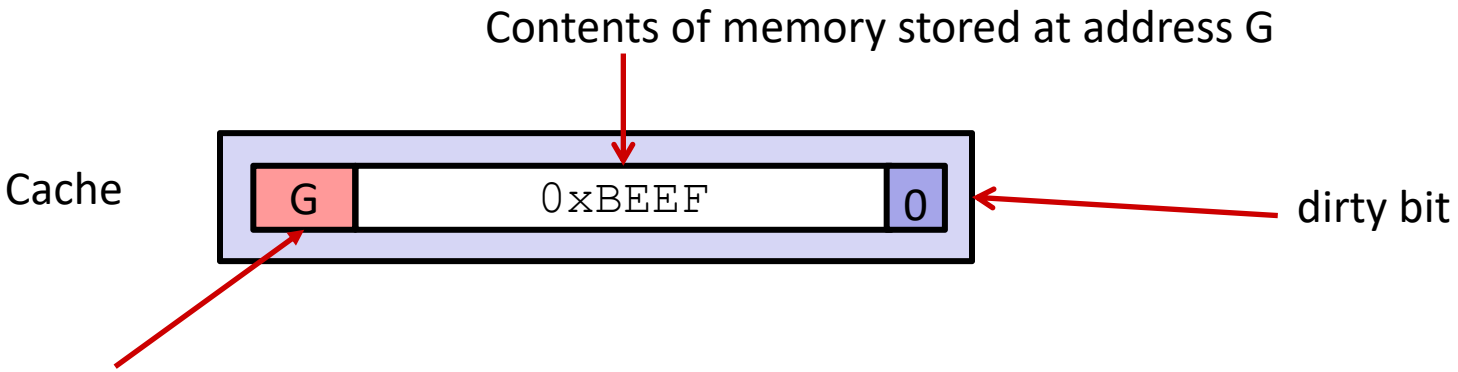
- Assuming the cache starts cold (all blocks invalid) and `sum` is stored in a register, calculate the **miss rate**:
  - m = 12 bits, C = 256 B, K = 32 B, E = 2

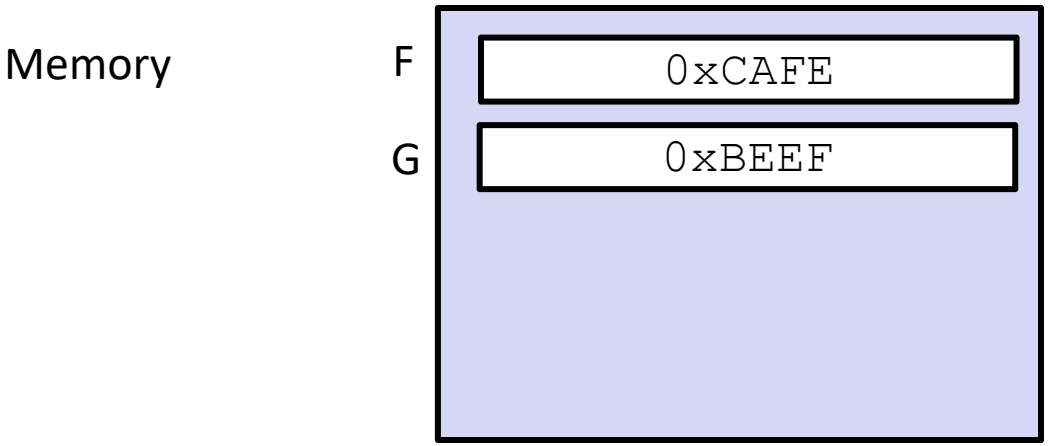
```
#define SIZE 8
long ar[SIZE][SIZE], sum = 0;  // &ar=0x800
for (int i = 0; i < SIZE; i++)
   for (int j = 0; j < SIZE; j++)
      sum += ar[i][j];
```

What about writes?

- Multiple copies of data exist:
  - L1, L2, possibly L3, main memory
- What to do on a write-hit?
  - Write-through: write immediately to next level
  - Write-back: defer write to next level until line is evicted (replaced)
    - Must track which cache lines have been modified ("*dirty bit*")
- What to do on a write-miss?
  - Write-allocate: ("fetch on write") load into cache, update line in cache
    - Good if more writes or reads to the location follow
  - No-write-allocate: ("write around") just write immediately to memory
- Typical caches:
  - Write-back + Write-allocate, usually
  - Write-through + No-write-allocate, occasionally

Write-back, write-allocate example



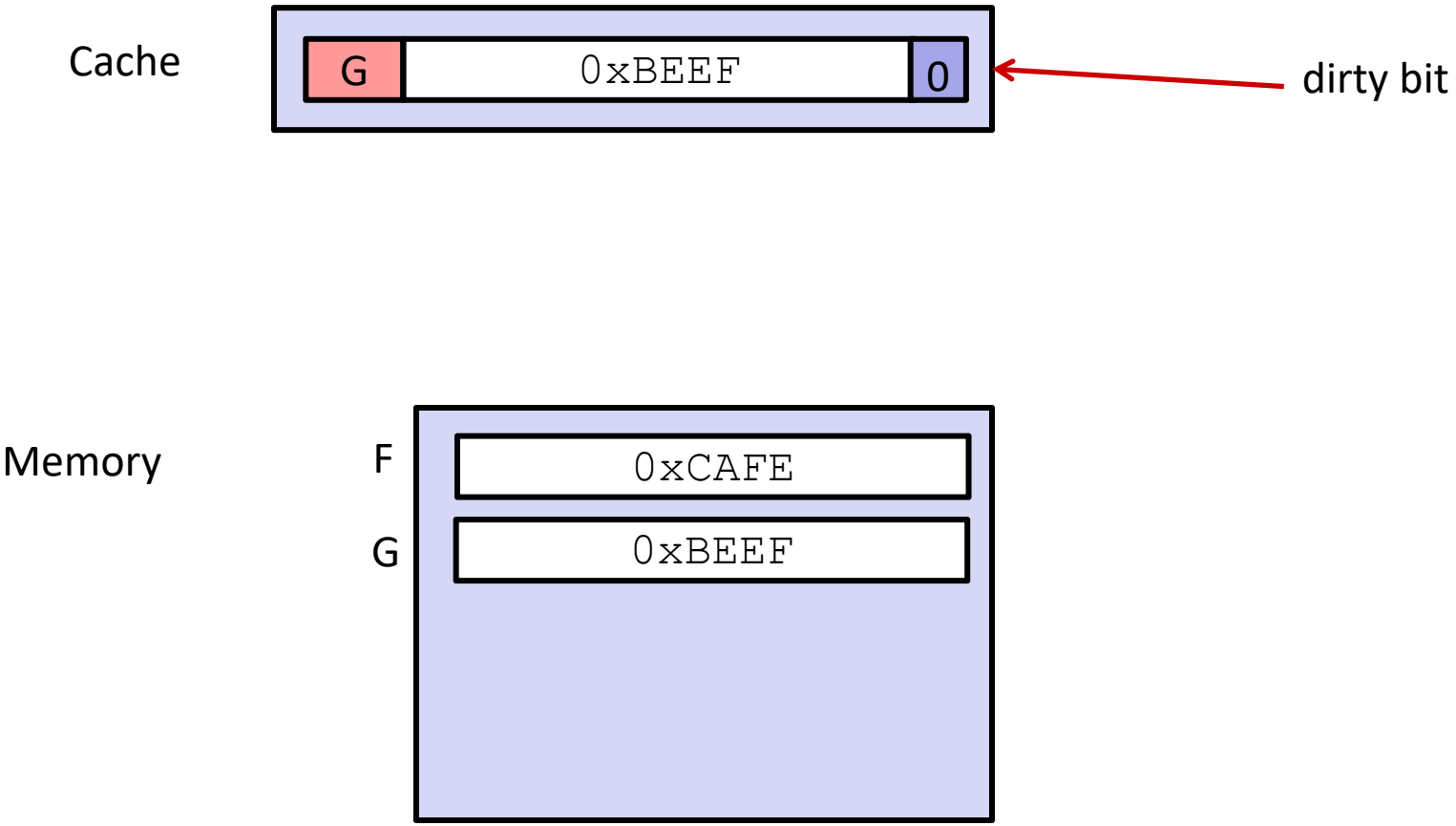


In this example we are sort of ignoring block offsets. Here a block holds 2 bytes (16 bits, 4 hex digits).

Normally a block would be much bigger and thus there would be multiple items per block. While only one item in that block would be written at a time, the entire line would be brought into cache.

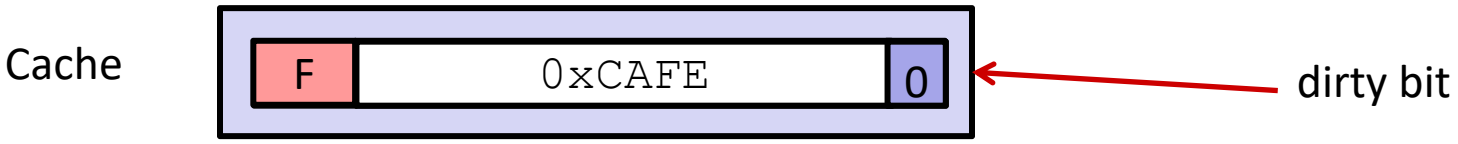
Write-back, write-allocate example

```
STORE 0xFACE, F
```

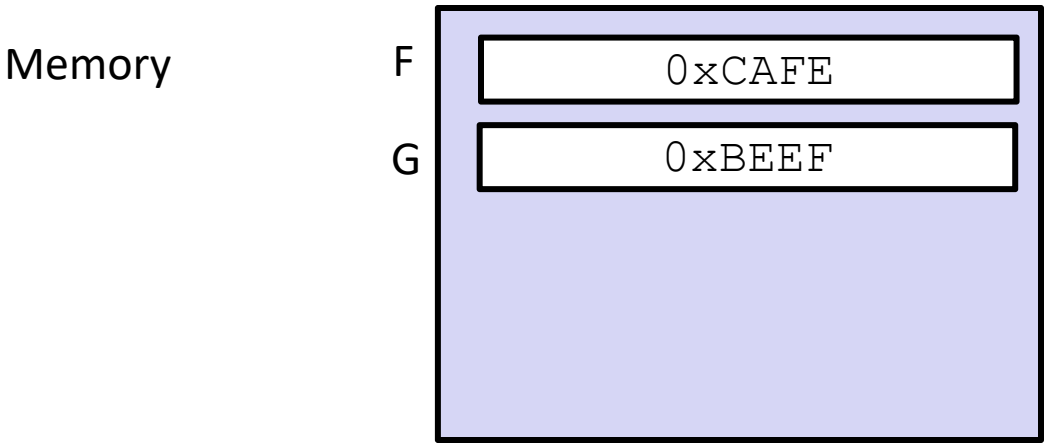


Write-back, write-allocate example

```
STORE 0xFACE, F
```

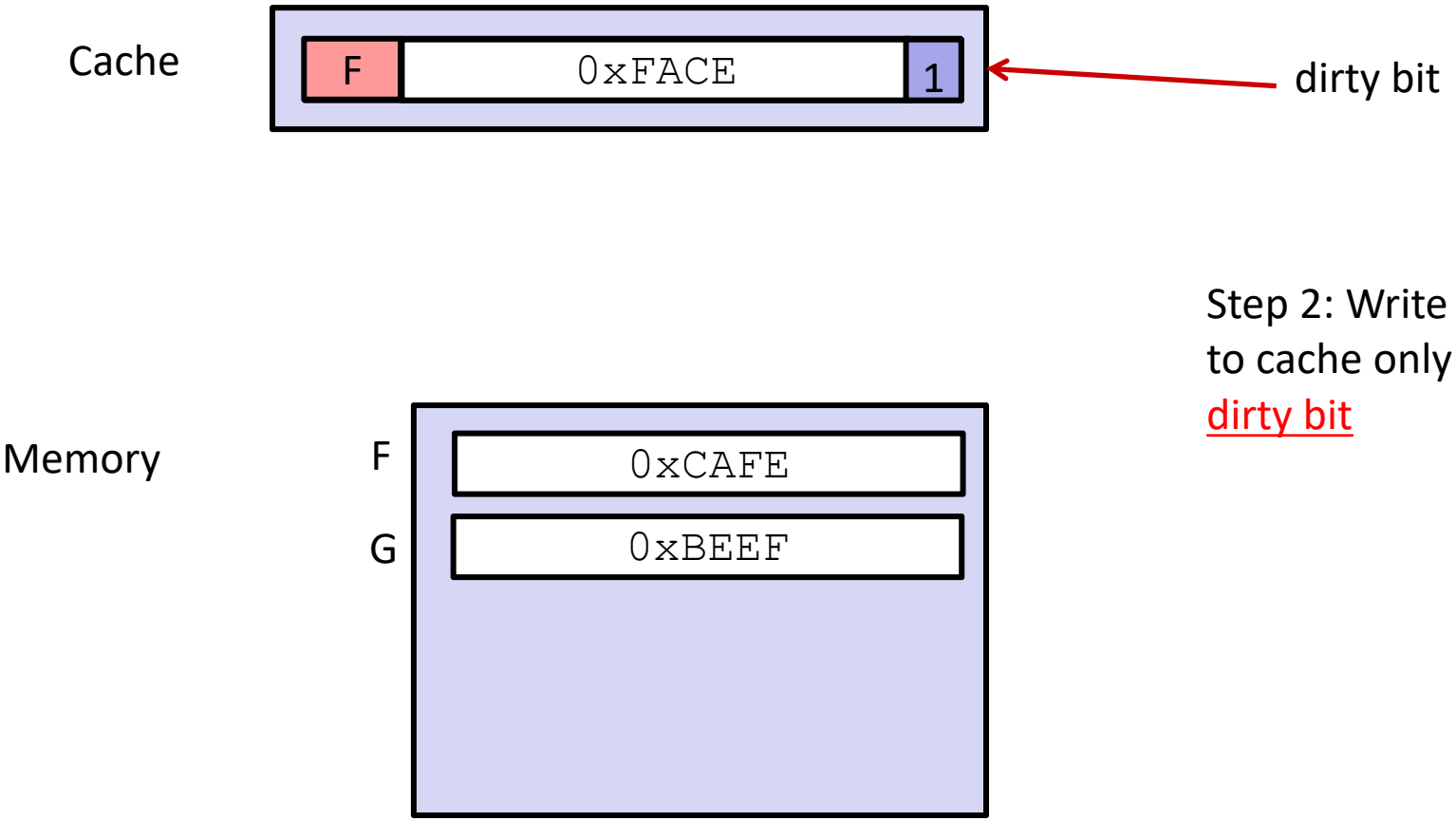


Step 1: Bring **F** into cache



Write-back, write-allocate example

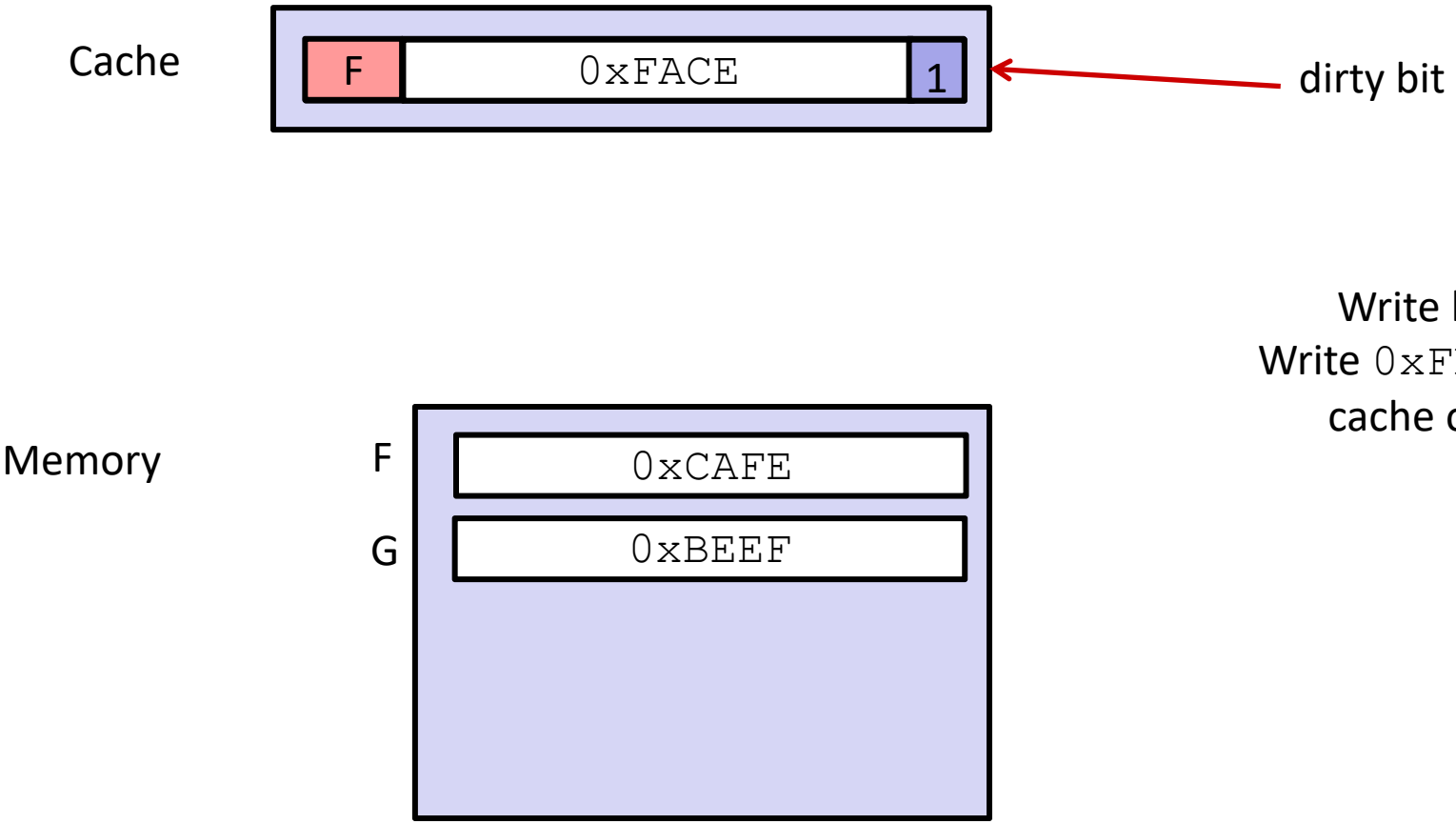
```
STORE 0xFACE, F
```



Step 2: Write `0xFACE` to cache only and set dirty bit

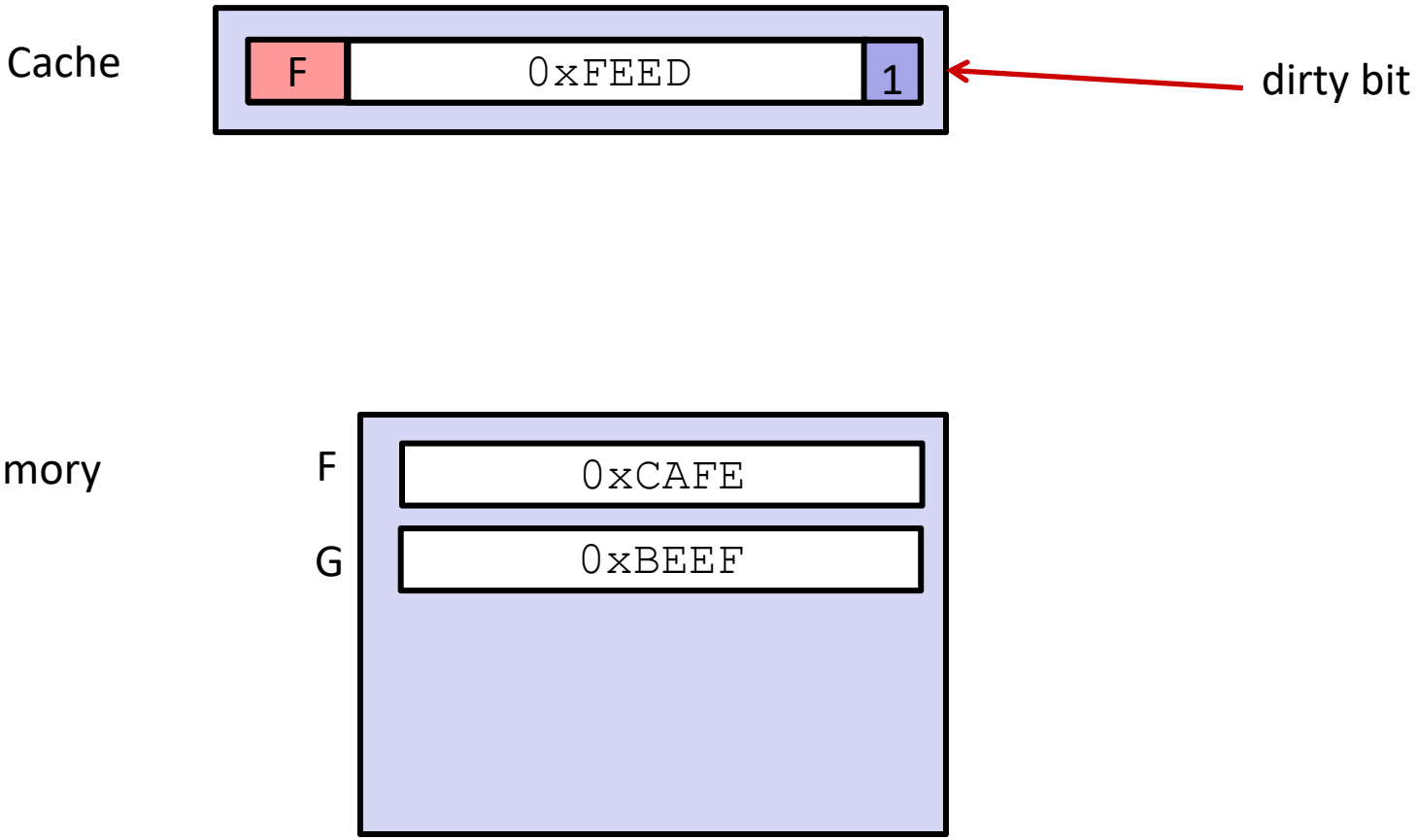
Write-back, write-allocate example

`STORE 0xFACE, F` `mov 0xFEED, F`



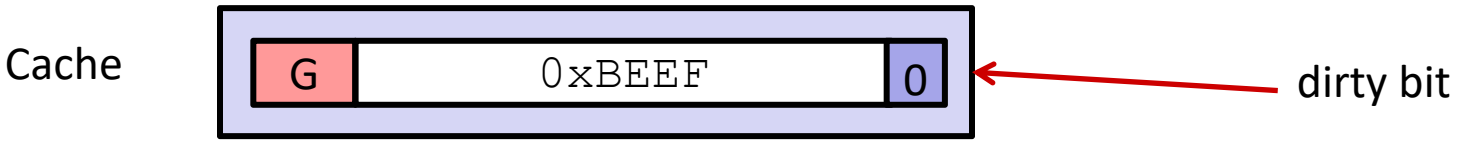
Write-back, write-allocate example

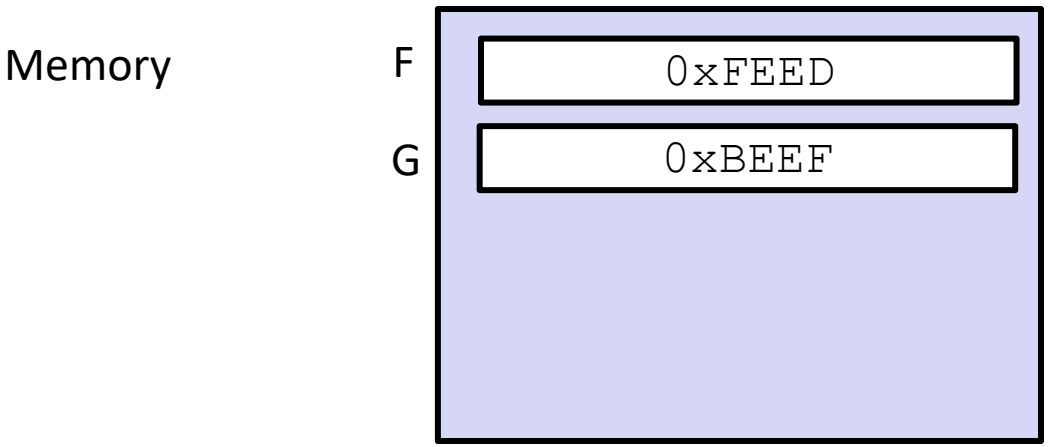
```
STORE 0xFACE, F   STORE 0xFEED, F        LOAD G, %t1
```



Write-back, write-allocate example

```
STORE 0xFACE, F   STORE 0xFEED, F        LOAD G, %t1
```





1. Write **F** back to memory since it is dirty
2. Bring **G** into the cache so we can copy it into `%t1`