Caches II

CSE 351 Spring 2018

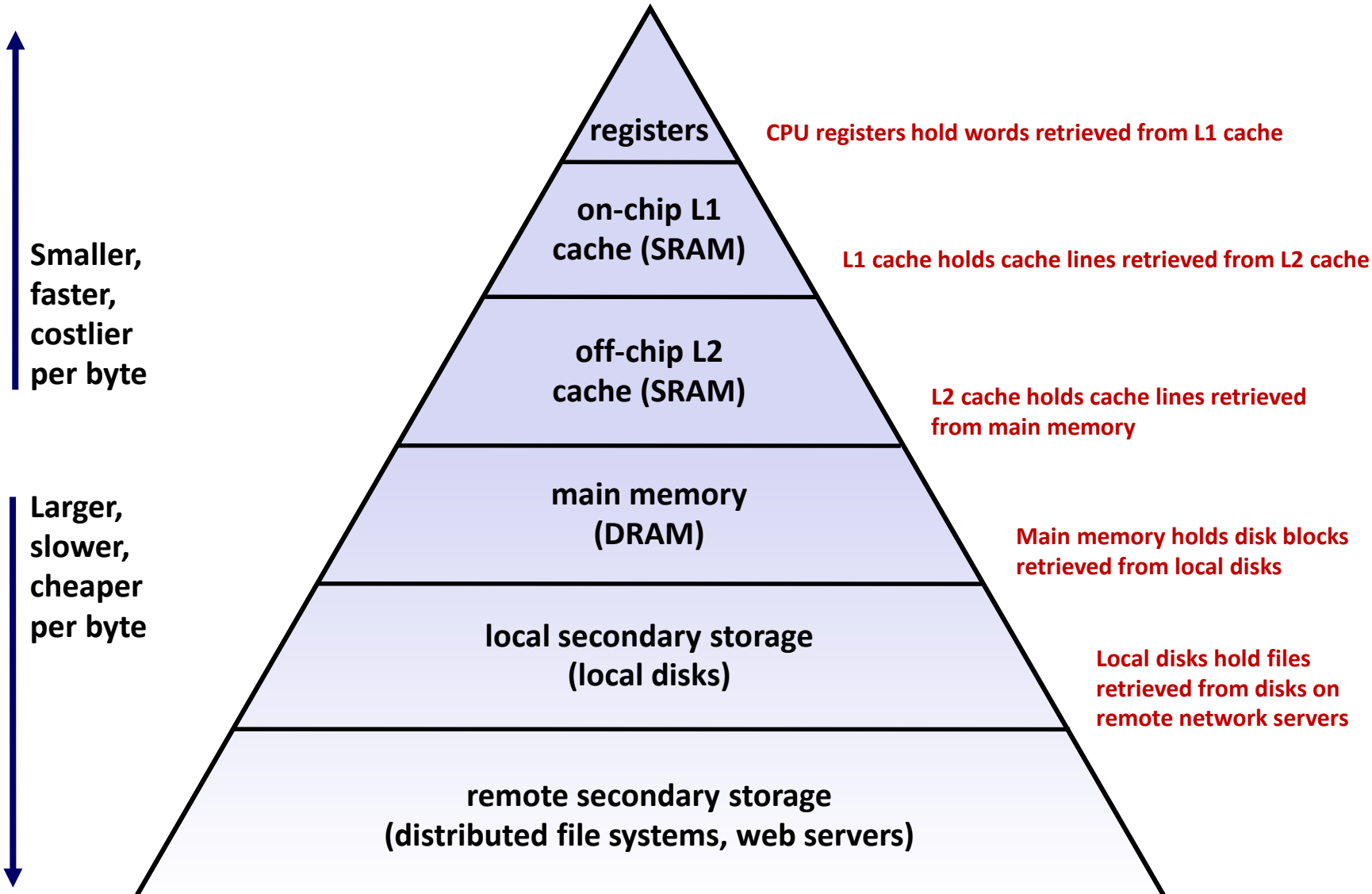
```
#include <yoda.h>
int is_try(int do_flag) {
  return !(do_flag
           || !do_flag);
}
```

Memory Hierarchies

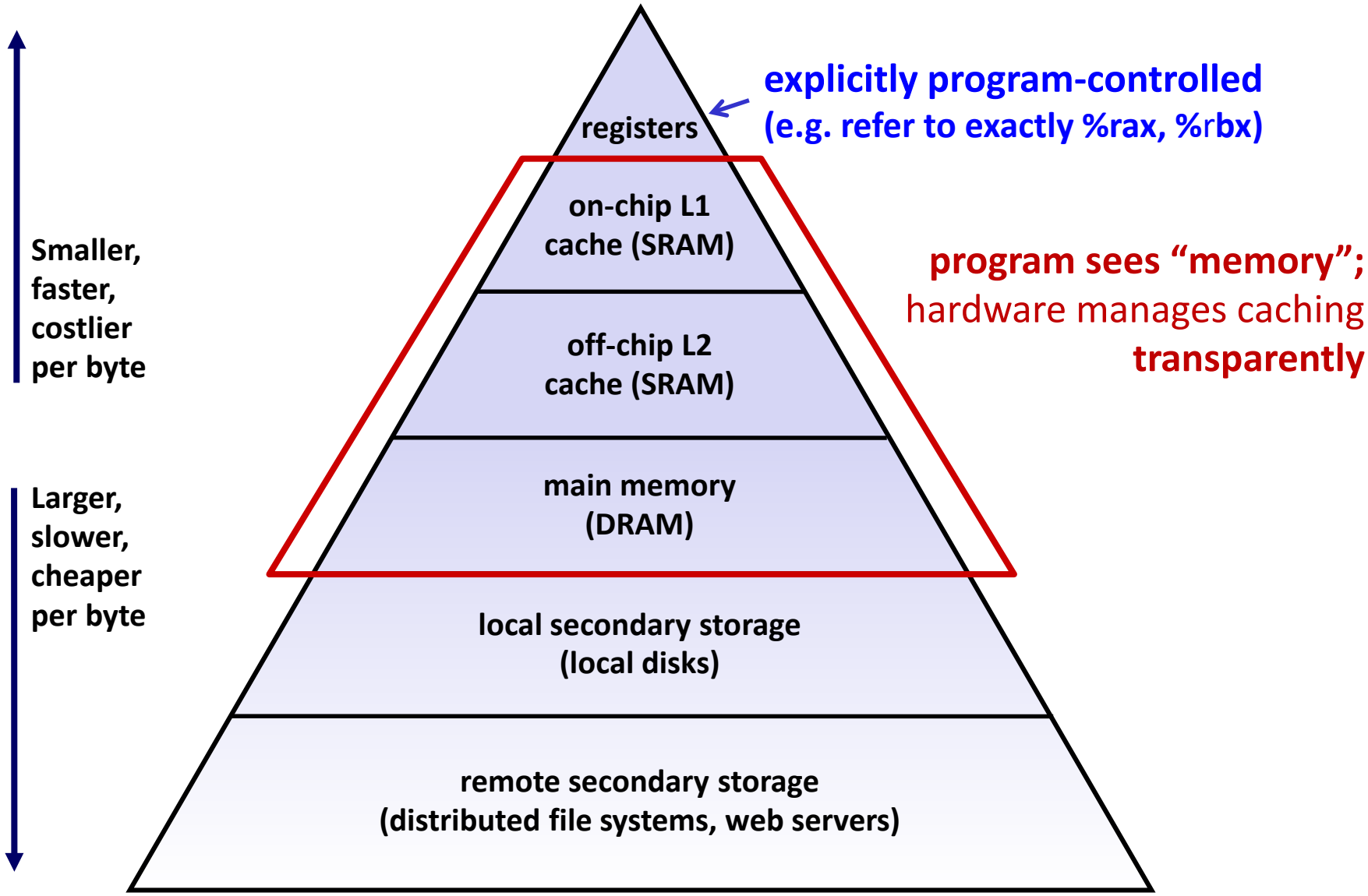
- Some fundamental and enduring properties of hardware and software systems:
  - Faster storage technologies almost always cost more per byte and have lower capacity
  - The gaps between memory technology speeds are widening
    - True for: registers ↔ cache, cache ↔ DRAM, DRAM ↔ disk, etc.
  - Well-written programs tend to exhibit good locality

- These properties complement each other beautifully
  - They suggest an approach for organizing memory and storage systems known as a memory hierarchy

An Example Memory Hierarchy



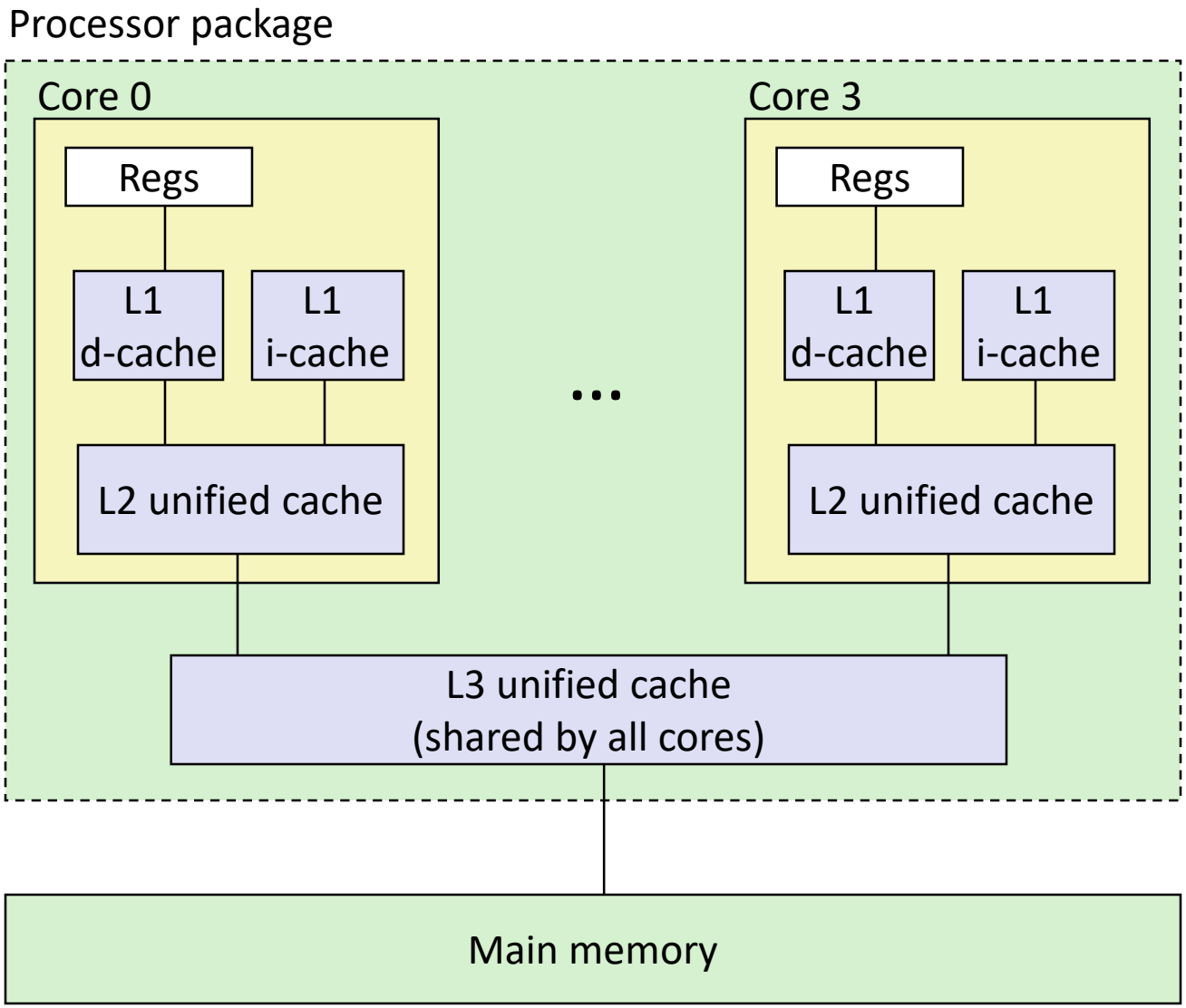
An Example Memory Hierarchy



Memory Hierarchies

- Fundamental idea of a memory hierarchy:
  - For each level k, the faster, smaller device at level k serves as a cache for the larger, slower device at level k+1
- Why do memory hierarchies work?
  - Because of locality, programs tend to access the data at level k more often than they access the data at level k+1
  - Thus, the storage at level k+1 can be slower, and thus larger and cheaper per bit
- *Big Idea:* The memory hierarchy creates a large pool of storage that costs as much as the cheap storage near the bottom, but that serves data to programs at the rate of the fast storage near the top

Intel Core i7 Cache Hierarchy



Block size:
64 bytes for all caches

L1 i-cache and d-cache:
32 KiB, 8-way,
Access: 4 cycles

L2 unified cache:
256 KiB, 8-way,
Access: 11 cycles

L3 unified cache:
8 MiB, 16-way,
Access: 30-40 cycles

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

Cache Organization (1)

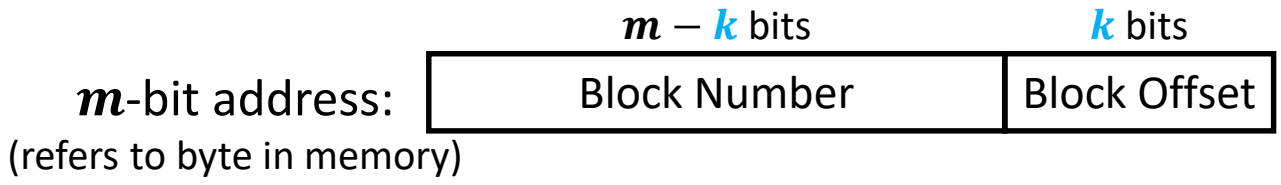
Note: The textbook uses "B" for block size

- Block Size (K): unit of transfer between $ and Mem
  - Given in bytes and always a power of 2 (*e.g.* 64 B)
  - Blocks consist of adjacent bytes (differ in address by 1)
    - Spatial locality!

Cache Organization (1)

Note: The textbook uses "b" for offset bits

- Block Size (K): unit of transfer between $ and Mem
  - Given in bytes and always a power of 2 (*e.g.* 64 B)
  - Blocks consist of adjacent bytes (differ in address by 1)
    - Spatial locality!
- Offset field
  - Low-order $\log_2(K) = \boldsymbol{k}$ bits of address tell you which byte within a block
    - (address) mod 2^n = n lowest bits of address
  - (address) modulo (# of bytes in a block)



Cache Organization (2)

- Cache Size (C): amount of *data* the $ can store
  - Cache can only hold so much data (subset of next level)
  - Given in bytes (C) or number of blocks (C/K)
  - Example: C = 32 KiB = 512 blocks if using 64-B blocks

- Where should data go in the cache?
  - We need a mapping from memory addresses to specific locations in the cache to make checking the cache for an address **fast**

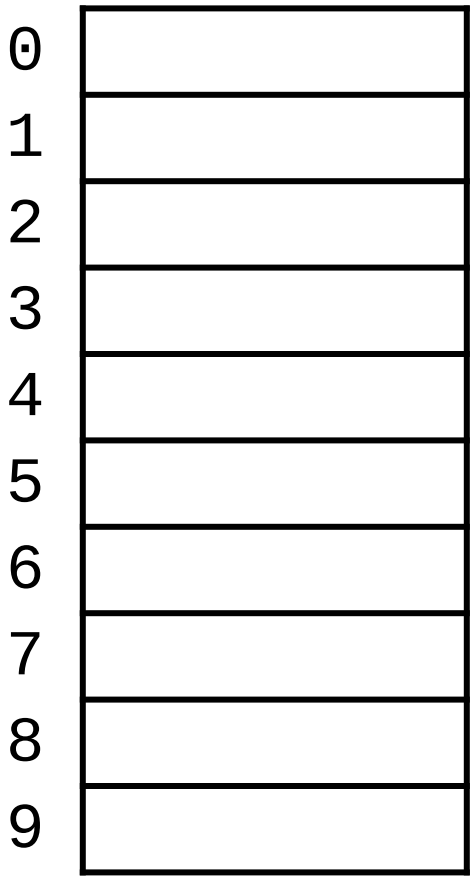
- What is a data structure that provides fast lookup?
  - Hash table!

Review: Hash Tables for Fast Lookup

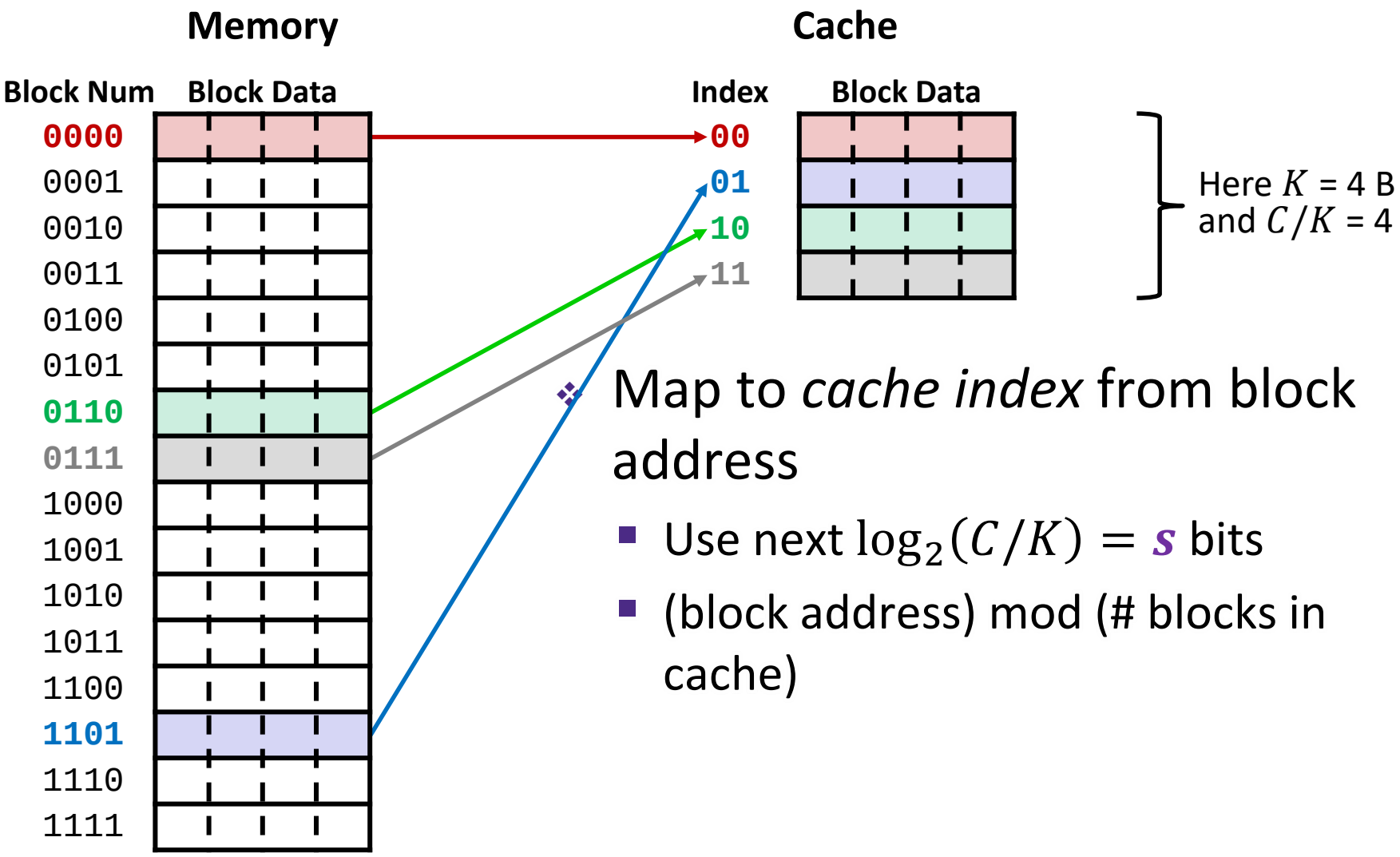
Insert:

5
27
34
102
119

Apply hash function to map data to "buckets"



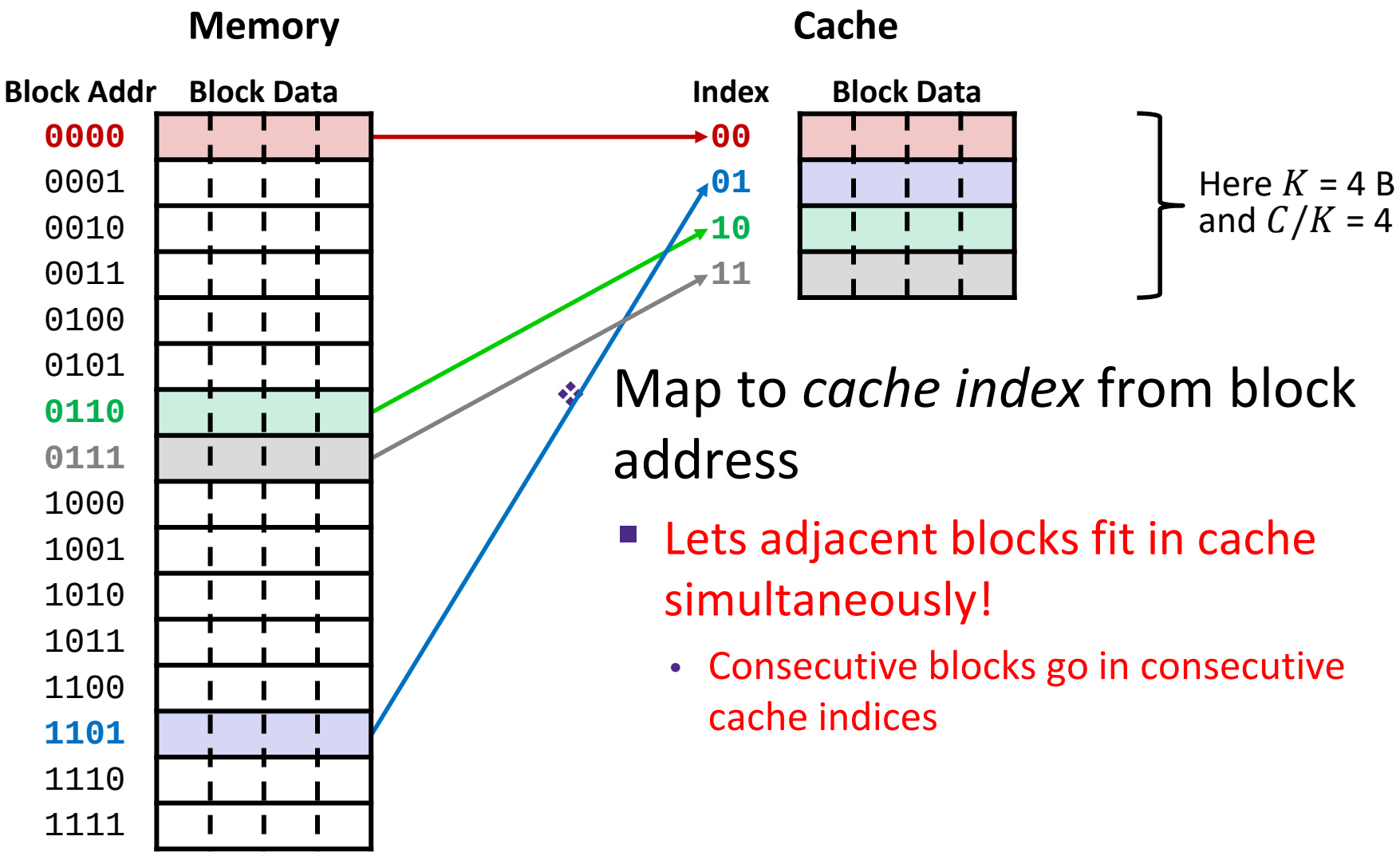
Place Data in Cache by Hashing Address



Map to *cache index* from block address

- Use next $\log_2(C/K) = \boldsymbol{s}$ bits
- (block address) mod (# blocks in cache)

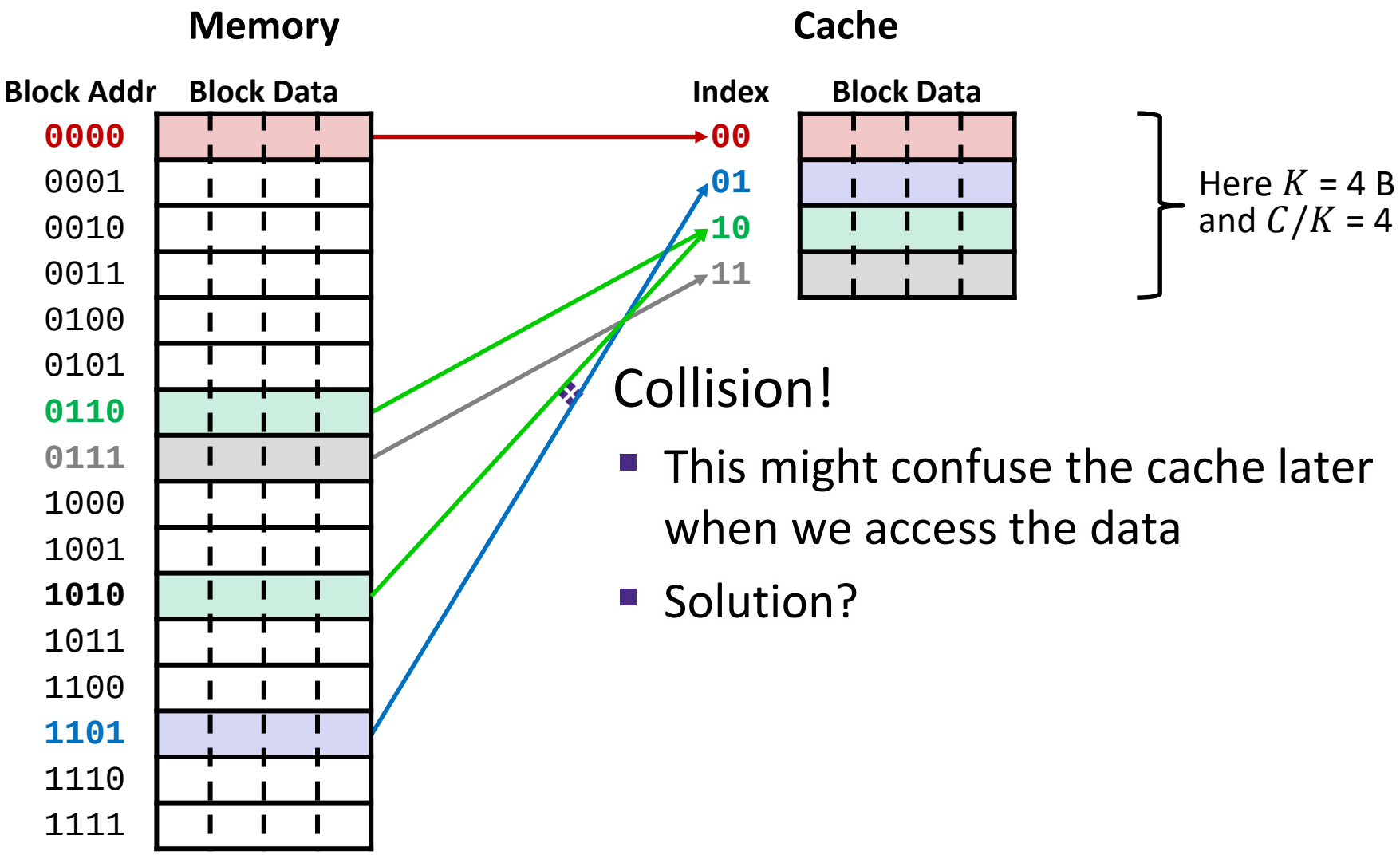
Place Data in Cache by Hashing Address



Map to *cache index* from block address

- Lets adjacent blocks fit in cache simultaneously!
  - Consecutive blocks go in consecutive cache indices

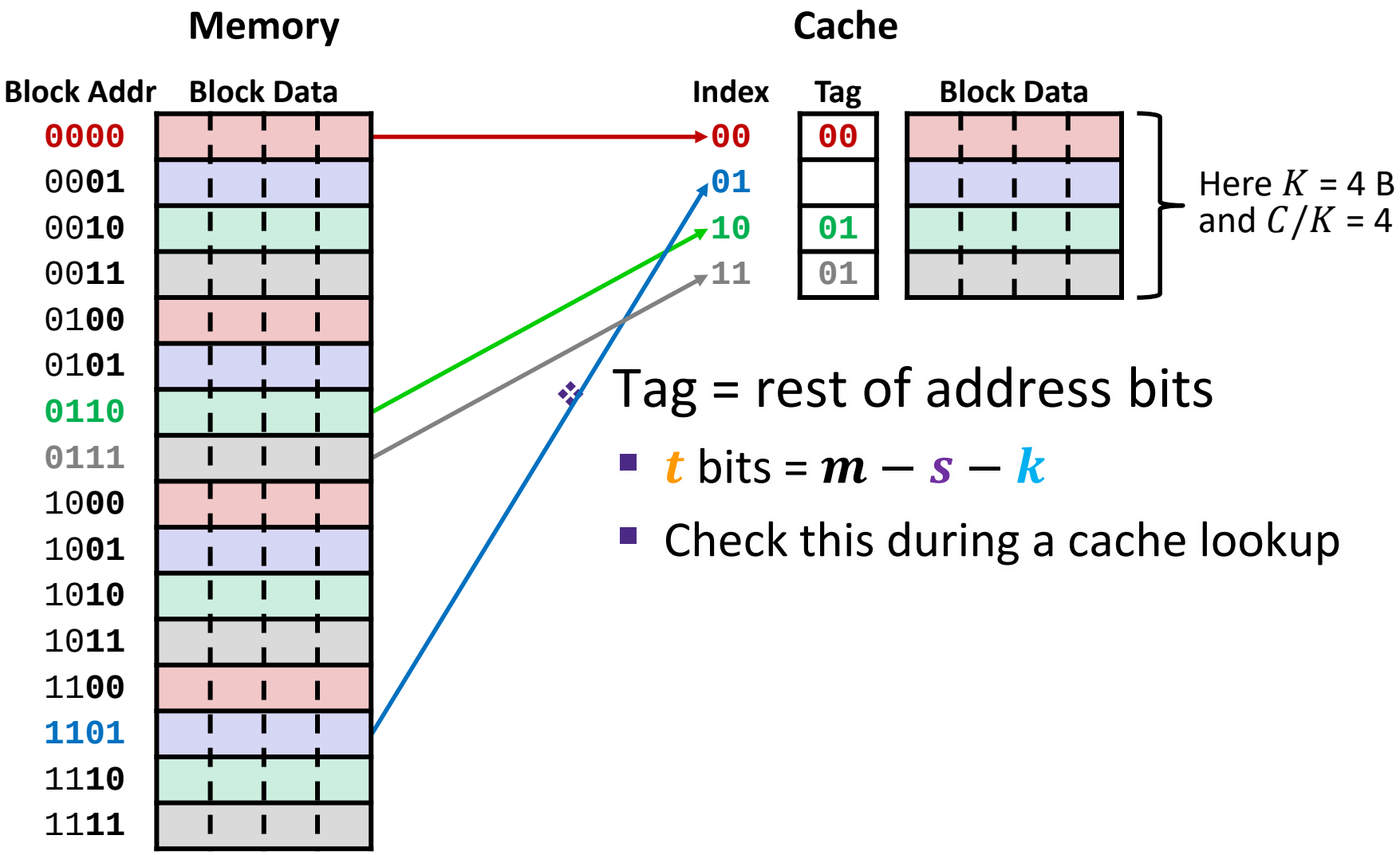
Place Data in Cache by Hashing Address



Collision!

- This might confuse the cache later when we access the data
- Solution?

Tags Differentiate Blocks in Same Index



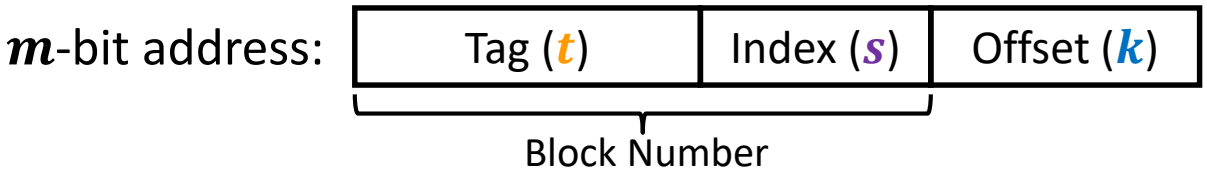
Tag = rest of address bits

- t bits = $m - s - k$
- Check this during a cache lookup

Checking for a Requested Address

- CPU sends address request for chunk of data
  - Address and requested data are not the same thing!
    - Analogy: your friend ≠ his or her phone number

- TIO address breakdown:



  - **Index** field tells you where to look in cache
  - **Tag** field lets you check that data is the block you want
  - **Offset** field selects specified start byte within block

  - [So far] All a function of (a) block size K, (b) cache size C

Cache Puzzle #1

- Based on the following behavior, which of the following block sizes is NOT possible for our cache?
  - Cache starts *empty*, also known as a *cold cache*
  - Access (addr: hit/miss) stream:
    - (14: miss), (15: hit), (16: miss)

A. 4 bytes

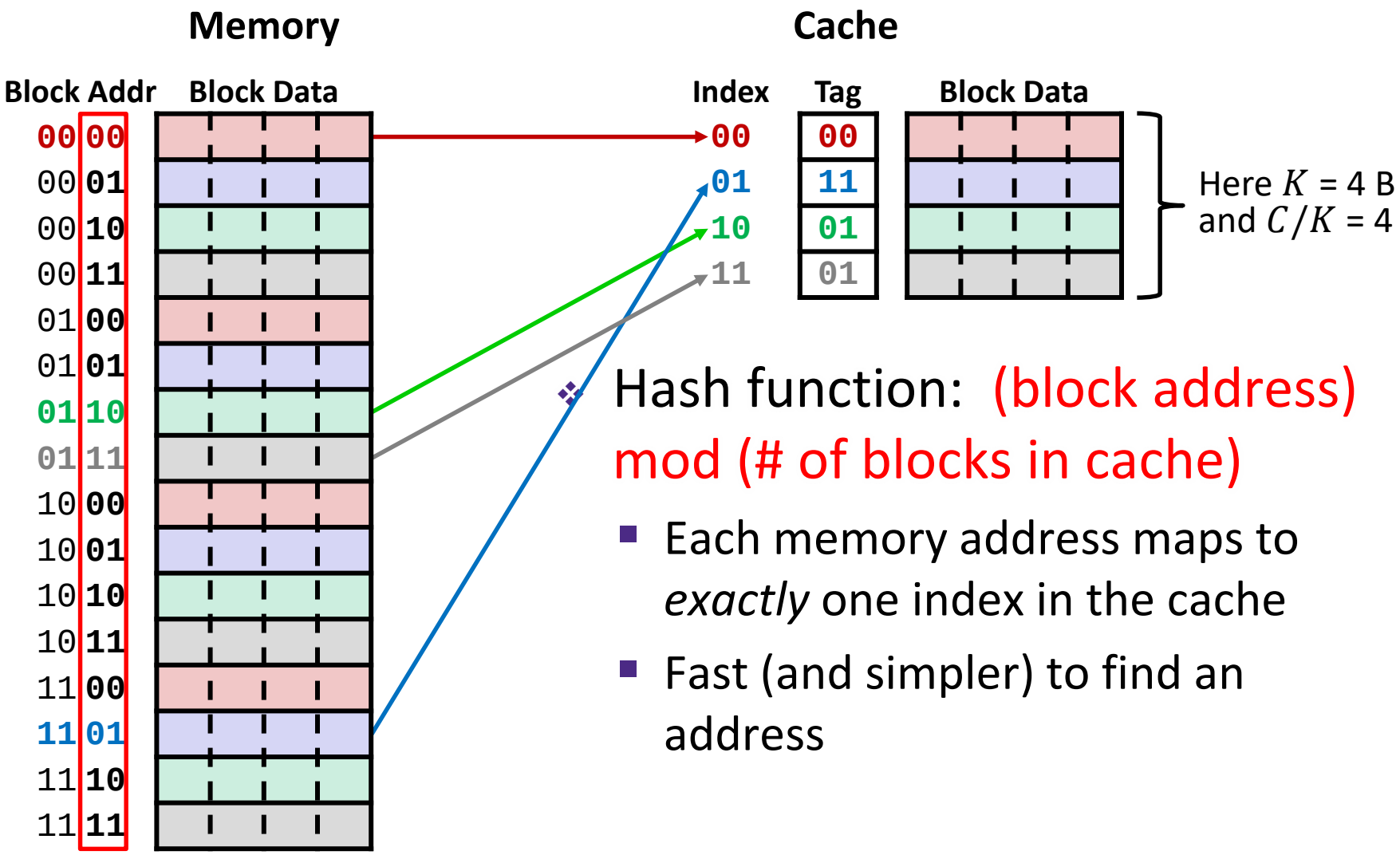
B. 8 bytes

C. 16 bytes

D. 32 bytes

E. We're lost…

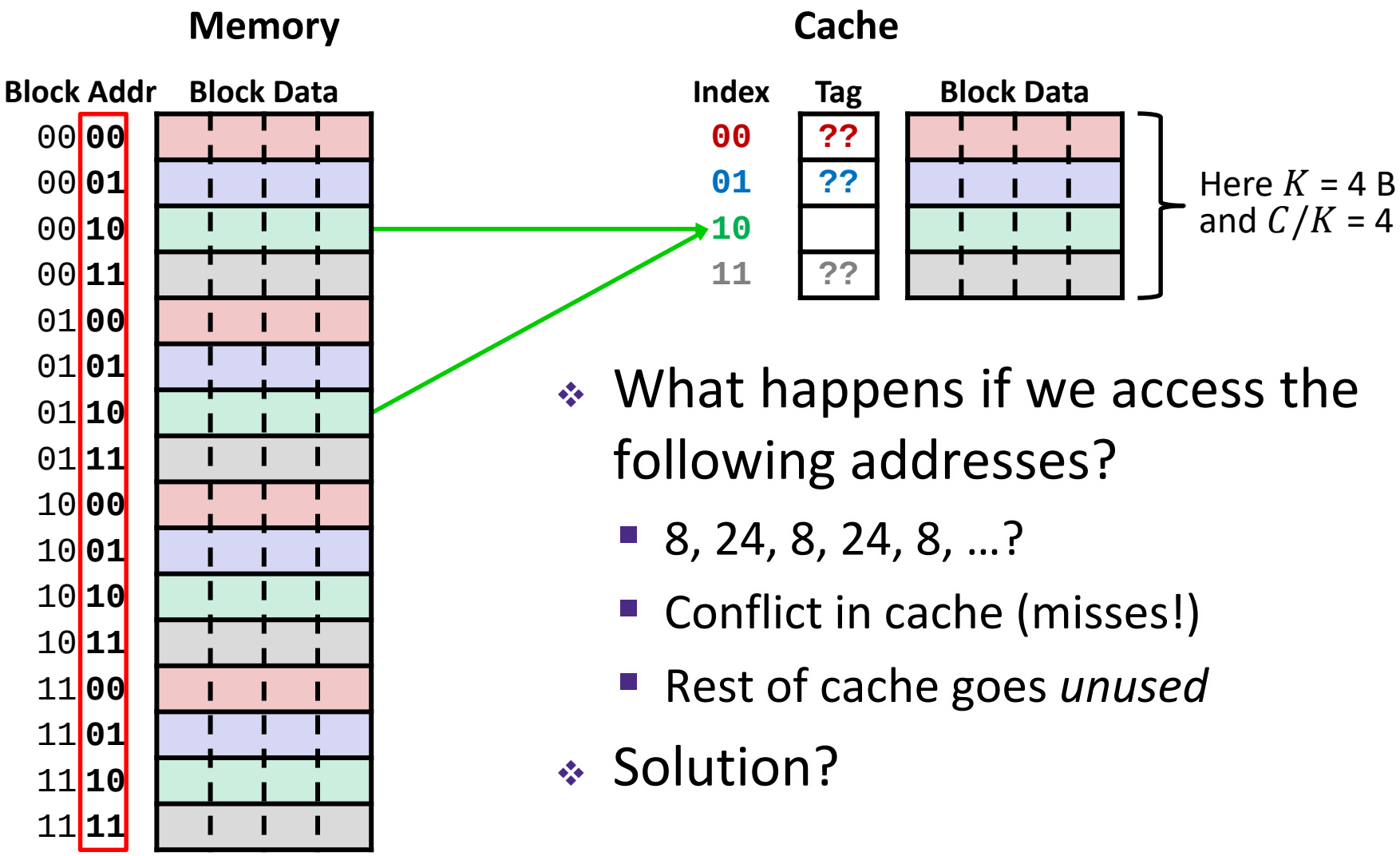
Direct-Mapped Cache



Hash function: (block address) mod (# of blocks in cache)

- Each memory address maps to *exactly* one index in the cache
- Fast (and simpler) to find an address

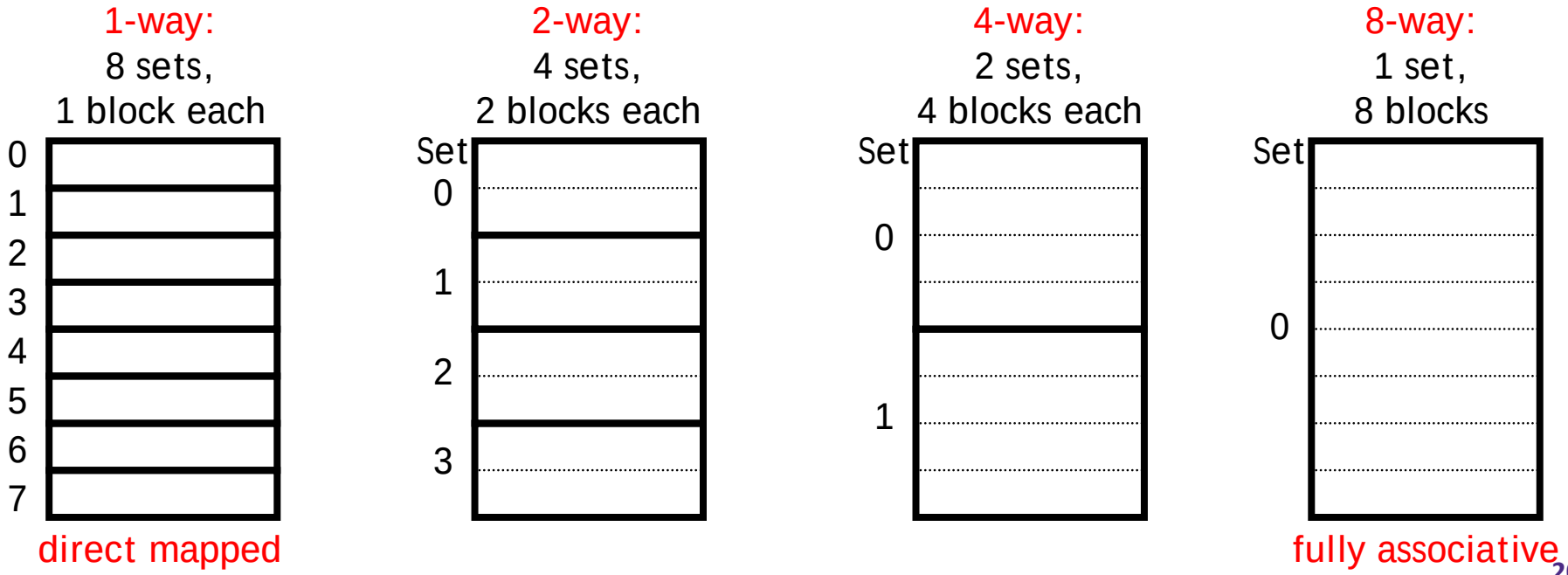
Direct-Mapped Cache Problem



- What happens if we access the following addresses?
  - 8, 24, 8, 24, 8, …?
  - Conflict in cache (misses!)
  - Rest of cache goes *unused*
- Solution?

Associativity

- What if we could store data in any place in the cache?
  - More complicated hardware = more power consumed, slower
- So we *combine* the two ideas:
  - Each address maps to exactly one **set**
  - Each set can store block in more than one **way**



A step back: Block size

- Cache mechanics are a lot of details
  - Perfect for hardware instead of humans ☺
- And/but there are high-level, beautiful engineering trade-offs throughout
- Advantages of bigger block size, K
  - Fewer cache loads when spatial locality expands to more bytes (e.g., all of large array)
  - Can grow block size [within reason] without much more *latency*
    - *Already going to the store, so put a few more avocados in your bag*
- Advantages of smaller block size, K
  - Less energy/effort moving around data that may not be used
  - More blocks for same C, so less *eviction* of data you might be using due to other data you're using

The real world

- So hardware designers pick a K that seems to do well performance-wise for Programs People Care About
  - Rigorously studied both empirically (experiments) and analytically (math)

- Then [new] software that Cares About Performance tries can take the cache parameters (C, K, …) into account
  - Data layout, loop structure, …