MEMORY SYSTEM

Mahdi Nazm Bojnordi

Assistant Professor

School of Computing

University of Utah

Overview

- This lecture
  - Cache terminologies
  - Cache performance
  - Cache addressing

Cache Terminology

- Block (cache line): unit of data access

Cache Terminology

- Block (cache line): unit of data access
- Hit: accessed data found at current level
  - *hit rate: fraction of accesses that finds the data*
  - *hit time: time to access data on a hit*

Cache Terminology

- Block (cache line): unit of data access
- Hit: accessed data found at current level
  - *hit rate: fraction of accesses that finds the data*
  - *hit time: time to access data on a hit*
- Miss: accessed data NOT found at current level
  - miss rate: 1 – hit rate
  - miss penalty: time to get block from lower level

Cache Terminology

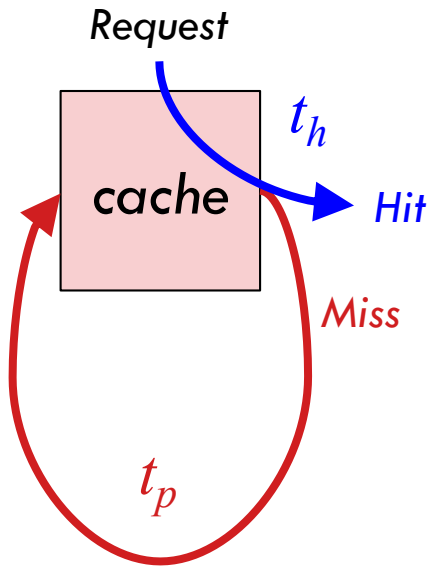
- Block (cache line): unit of data access
- Hit: accessed data found at current level
  - *hit rate: fraction of accesses that finds the data*
  - *hit time: time to access data on a hit*
- Miss: accessed data NOT found at current level
  - miss rate: 1 – hit rate
  - miss penalty: time to get block from lower level

hit time << miss penalty

Cache Performance

- Average Memory Access Time (AMAT)

Outcome	Rate	Access Time
Hit	r_h	t_h
Miss	r_m	$t_h + t_p$



$$AMAT = r_h t_h + r_m (t_h + t_p)$$

$$r_h = 1 - r_m$$

$$AMAT = t_h + r_m t_p$$

Example

- *Assume that hit rate is 90%; hit time is 2 cycles; and accessing the lower level takes 200 cycles; find the average memory access time?*

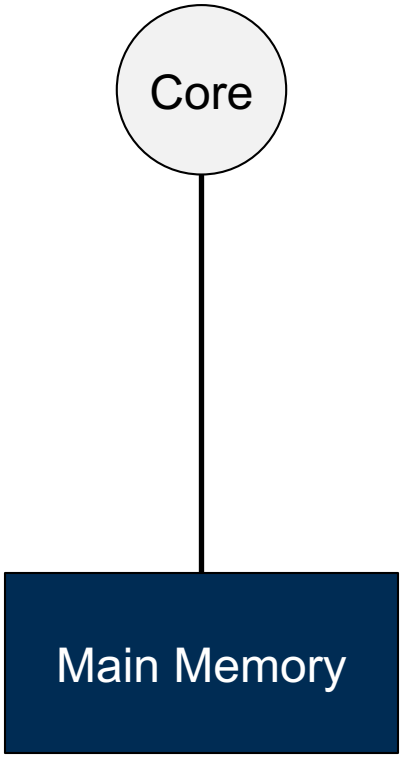
Example

- *Assume that hit rate is 90%; hit time is 2 cycles; and accessing the lower level takes 200 cycles; find the average memory access time?*

AMAT = 2 + 0.1x200 = 22 cycles

Summary: Cache Performance

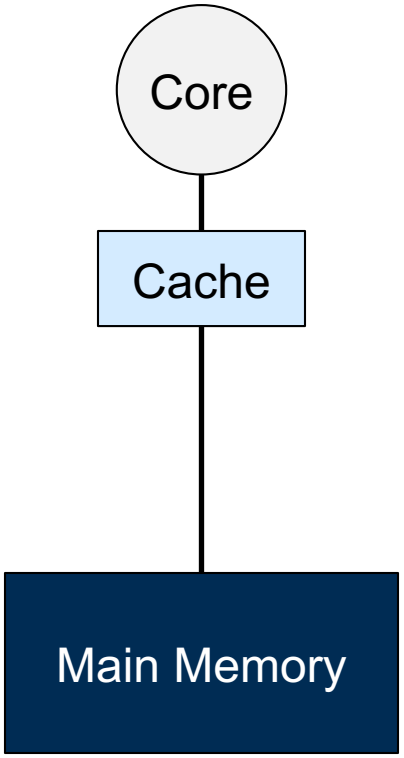
- Bridging the processor-memory performance gap



Main memory access time: 300 cycles

Summary: Cache Performance

- Bridging the processor-memory performance gap



Main memory access time: 300 cycles

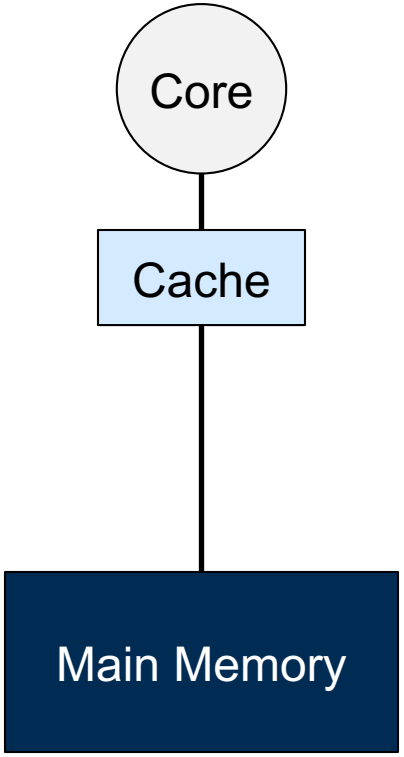
Cache

- L1: 2 cycles hit time; 60% hit rate

What is the average mem access time?

Summary: Cache Performance

- Bridging the processor-memory performance gap



Main memory access time: 300 cycles

Cache

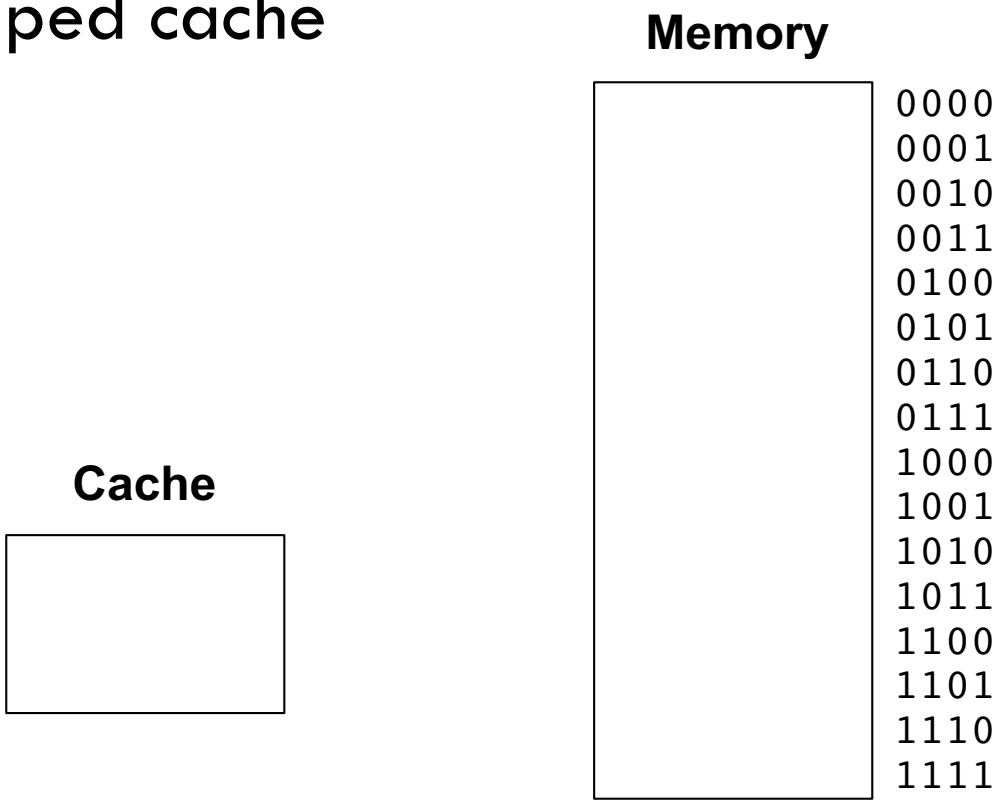
- L1: 2 cycles hit time; 60% hit rate

What is the average mem access time?

$$AMAT = t_h + r_m\, t_p$$
$$= 2 + 0.4 \times 300$$
$$= 122$$

Cache Addressing

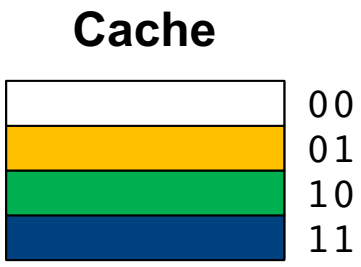
- Instead of specifying cache address we specify main memory address
- Simplest: direct-mapped cache

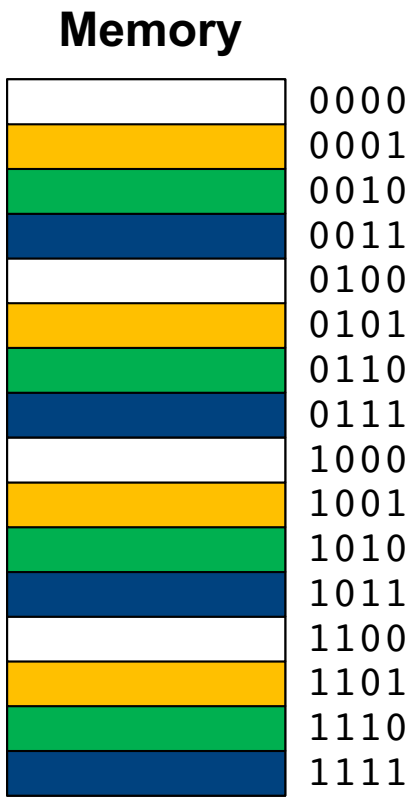


Cache Addressing

- Instead of specifying cache address we specify main memory address
- Simplest: direct-mapped cache

Note: each memory address maps to a single cache location determined by modulo hashing



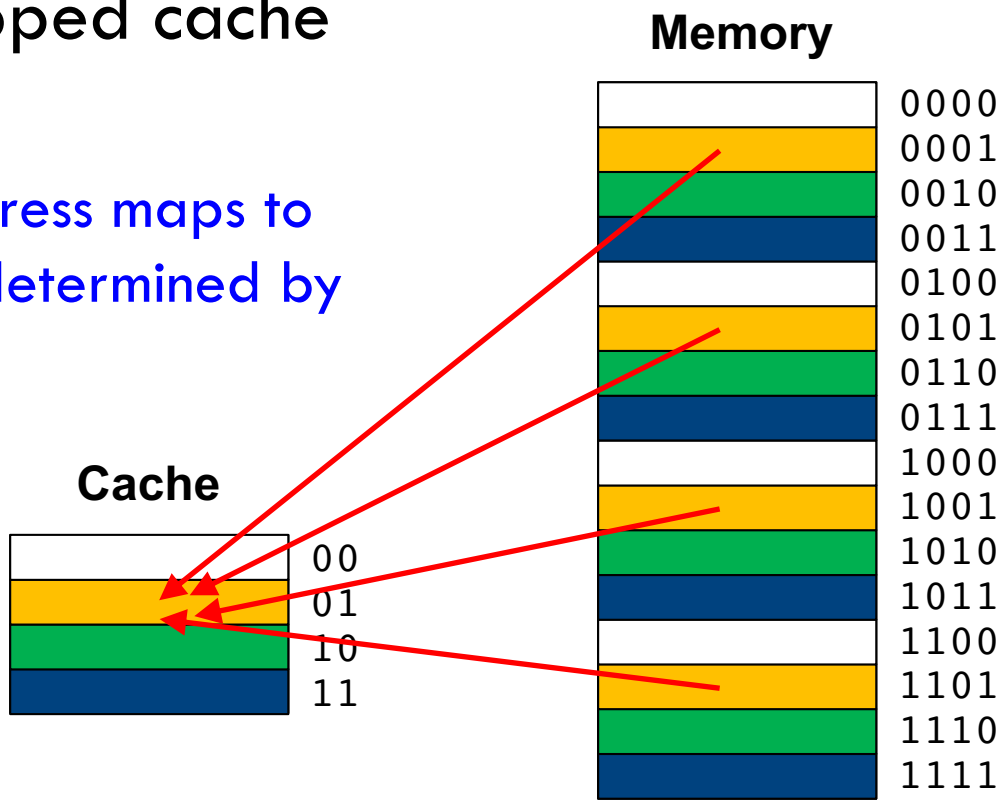


Cache Addressing

- Instead of specifying cache address we specify main memory address
- Simplest: direct-mapped cache

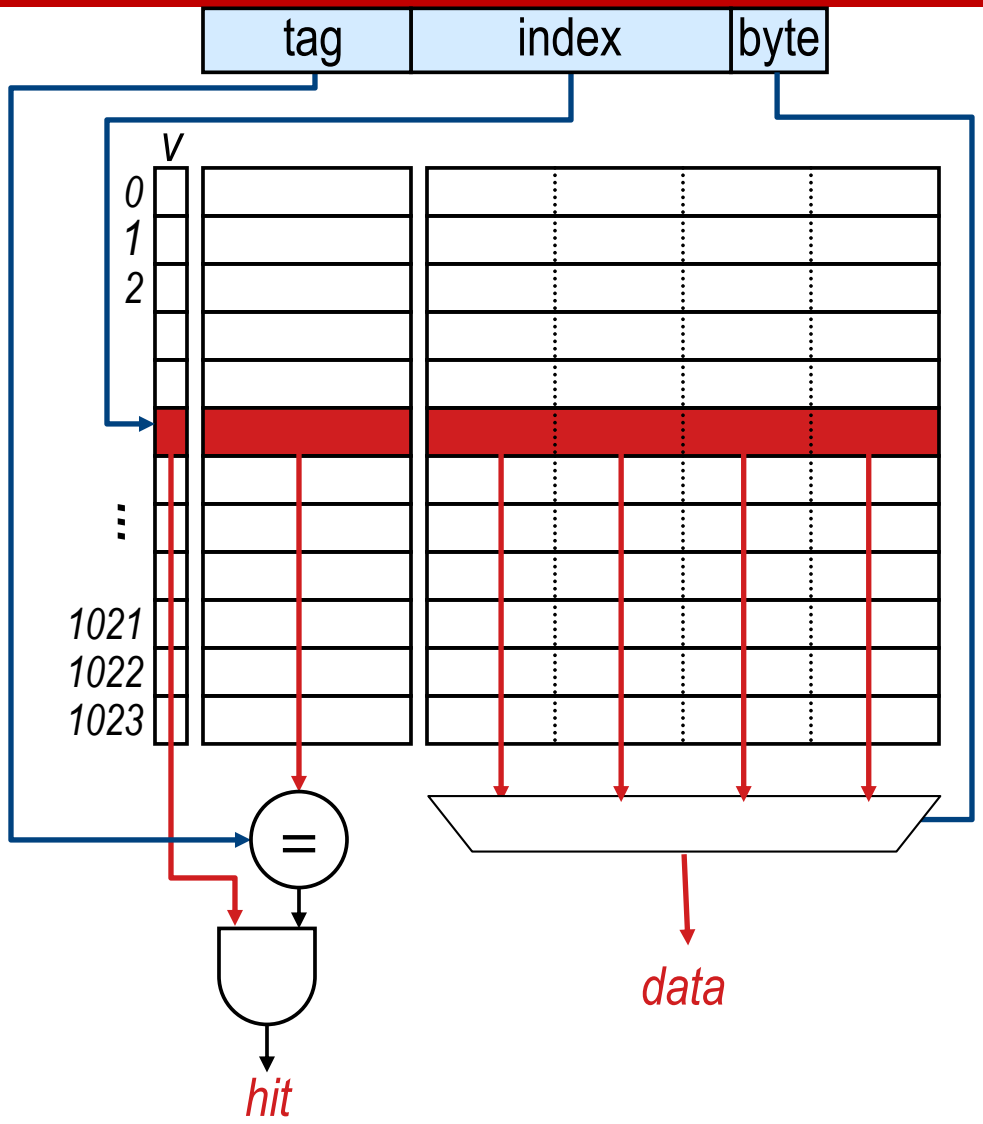
Note: each memory address maps to a single cache location determined by modulo hashing

How to exactly specify which blocks are in the cache?



Direct-Mapped Lookup

- Byte offset: to select the requested byte
- Tag: to maintain the address
- Valid flag (v): whether content is meaningful
- Data and tag are always accessed



Example Problem

- Find the size of tag, index, and offset bits for an 8MB, direct-mapped L3 cache with 64B cache blocks. Assume that the processor can address up to 4GB of main memory.

Example Problem

- Find the size of tag, index, and offset bits for an 8MB, direct-mapped L3 cache with 64B cache blocks. Assume that the processor can address up to 4GB of main memory.
- $4GB = 2^{32}$ B → address bits = 32
- $64B = 2^{6}$ B → byte offset bits = 6
- $8MB/64B = 2^{17}$ → index bits = 17
- tag bits = 32 – 6 – 17 = 9