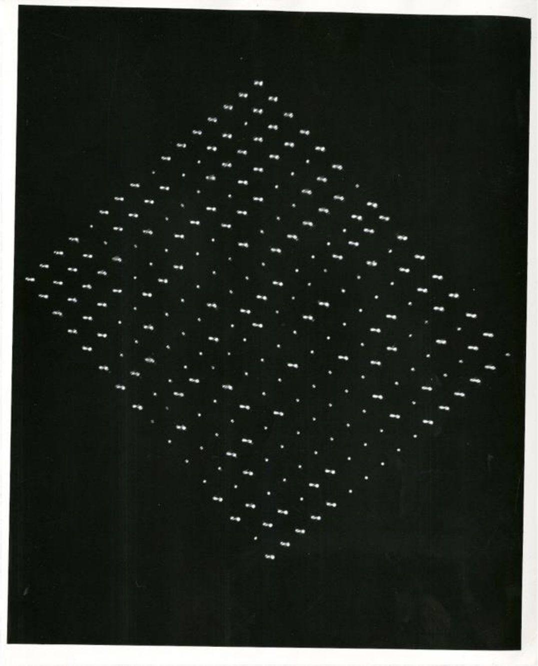
CSCI 210: Computer Architecture
Lecture 34: Caches III

Stephen Checkoway

Slides from Cynthia Taylor

CS History: The Williams Tube

ArnoldReinhold, CC BY-SA 3.0



National Institute of Standards and Technology, Public domain, via Wikimedia Commons

- First random-access storage device
- Developed in 1946
- Displays a grid of dots over a cathode ray tube (using an electron beam to strike phosphor)
- Each dot represents a bit
- Each dot creates a small static electricity charge
- Charge at each location is read by a metal sheet in front of the display
- Needs to be periodically refreshed as charge fades over time

Three types of cache misses

- Compulsory (or cold-start) misses
  - first access to the data.
- Capacity misses
  - we missed only because the cache isn't big enough.
- Conflict misses
  - we missed because the data maps to the same index as other data that forced it out of the cache.

block address of misses

4
8
12
4
8
20
4
8
20
24
12
8
4

tag	data

DM cache

Cache miss example (from StackOverflow)

32 kB direct-mapped cache

1. You repeatedly iterate over a 128 kB array
   – All misses but the first access to each block are capacity misses because the array does not fit in cache; the first are compulsory misses
2. You iterate over two 8 kB arrays that map to the same cache indices
   – These are conflict misses because if you changed the locations of the arrays to be consecutive, then both would fit in the cache

https://stackoverflow.com/a/33336918

Cache Miss Type

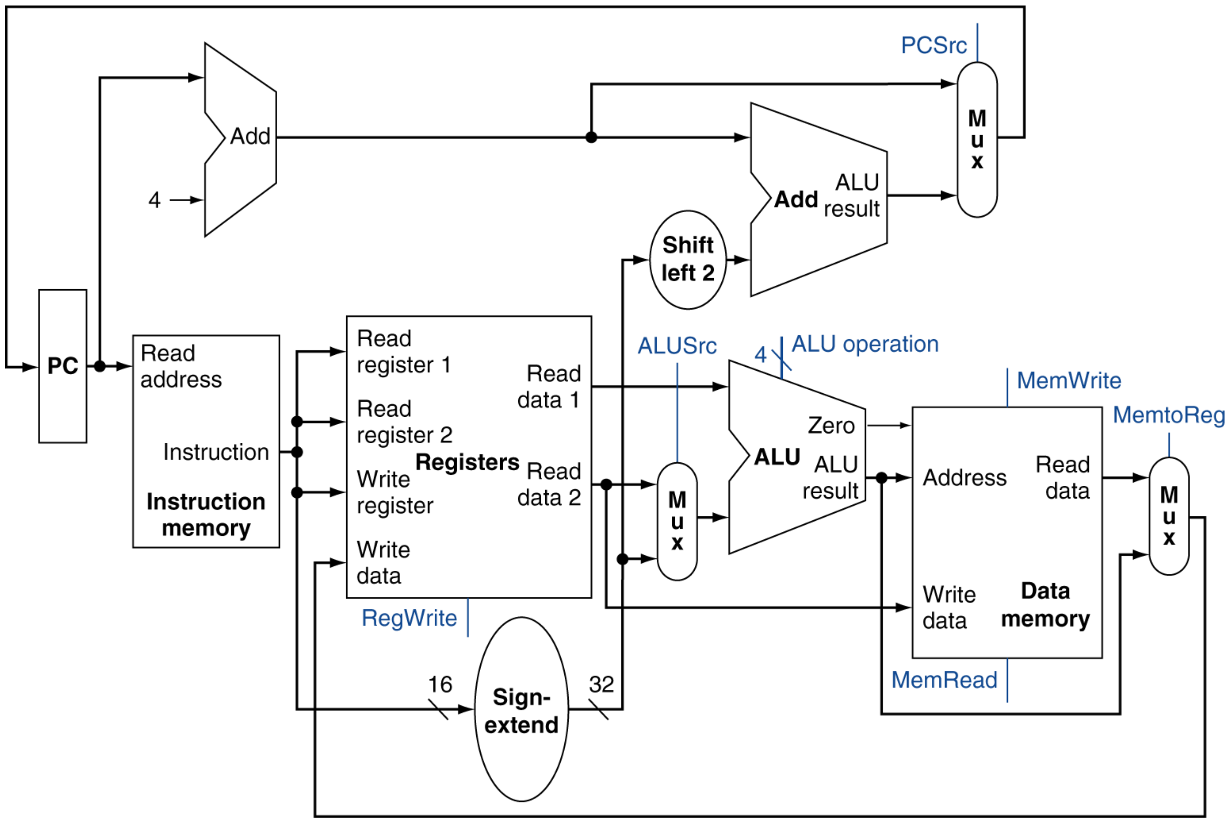
Suppose you experience a cache miss on a block (let's call it block A). You have accessed block A in the past. There have been precisely 1027 different blocks accessed between your last access to block A and your current miss. Your block size is 32-bytes and you have a 64 kB cache (recall a kB = 1024 bytes). What kind of miss was this?

Selection	Cache Miss
A	Compulsory
B	Capacity
C	Conflict
D	Both Capacity and Conflict
E	None of the above

Questions on associativity, replacement?

CACHE PERFORMANCE

I-cache vs D-cache



- Separate caches for instruction memory and data memory
- I-cache: instruction cache
- D-cache: data cache

Measuring Cache Performance

- Components of CPU time
  - Program execution cycles
    - Includes cache hit time
  - Memory stall cycles
    - Mainly from cache misses
- With simplifying assumptions:

$$\text{Memory stall cycles}$$

$$= \frac{\text{Memory accesses}}{\text{Program}} \times \text{Miss rate} \times \text{Miss penalty}$$

$$= \frac{\text{Instructio ns}}{\text{Program}} \times \frac{\text{Misses}}{\text{Instructio n}} \times \text{Miss penalty}$$

Miss Cycles Per Instruction

Given

- I-cache miss rate = 2%
- D-cache miss rate = 4%
- Miss penalty = 100 cycles
- Base CPI (ideal cache) = 2
- Load & stores are 36% of instructions

	I-cache	D-cache
A	.02 * 100	.04 * 100
B	.02	.04
C	.02 * .36 * 100	.04 * .36 * 100
D	.02 * 100	.04 * .36 * 100

Cache Performance Example

- Given
  - I-cache miss rate = 2%
  - D-cache miss rate = 4%
  - Miss penalty = 100 cycles
  - Base CPI (ideal cache) = 2
  - Load & stores are 36% of instructions
- Miss cycles per instruction
  - I-cache: 0.02 × 100 = 2
  - D-cache: 0.36 × 0.04 × 100 = 1.44
- Actual CPI = 2 + 2 + 1.44 = 5.44

Average Access Time

- Hit time is also important for performance
- Average memory access time (AMAT)
  - AMAT = Hit time + Miss rate × Miss penalty
- Example
  - hit time = 1 cycle, miss penalty = 20 cycles, I-cache miss rate = 5%
  - AMAT =

Cache Speed Factors

- Memory lookup time

- Hit rate

- Size

- Frequency of collisions

Reading

- Next lecture: More Caches!