Memory Systems

Sandhya Dwarkadas

Memory Hierarchies

Problem: Want unlimited fast memory

Solutions:

- Caches - level of memory hierarchy between CPU and main memory
- Virtual Memory - level between disk and main memory that creates the illusion of a large address space

Principle of locality

- Temporal locality - address referenced in the past will be tend to be referenced again soon
- Spatial locality - if an address is referenced, addresses close by will tend to be referenced soon

Cache Organization

Placement - where is a block placed

Location - how do you locate a block

Re-placement - which block do you replace

Write policy - what happens on a write

Cache Placement Policy

Direct-mapped

Set-associative

Fully-associative

Cache Re-placement Policy

Least Recently Used

FIFO - First-in, First-Out

Random

Write Policy

Write through

Write back

Handling Cache Misses

Instruction cache miss -

- Send the original PC value to memory
- Instruct main memory to perform a read and wait for access to complete
- Write the cache entry - data, tag, and valid bit
- Restart the instruction at the fetch stage

Memory Organization

Memory interleaving -

- memory organized in banks
- full latency incurred only once
- addresses allocated in a round-robin fashion using low-order address bits

Wide memory

Page mode (e.g., EDO (Extended Data Out RAMS)) or synchronous DRAMs (SDRAMs)

Cache Performance

Qualitative Categorization of Misses:
Compulsory, Capacity, Conflict

Design Change	Effect on Miss Rate	Negative Performance Effect
Increase size	Decrease capacity misses	May increase access time
Increase associativity	Decrease conflict misses	May increase access time
Increase block size	May decrease compulsory misses	May increase miss penalty May increase capacity misses

Improving Cache Performance

at the Application/Compiler Level

Merging arrays into structures if accessed in a similar manner to improve spatial locality and conflict misses

Loop interchange - to improve spatial locality and reduce capacity misses

Loop fusion - to improve temporal locality and capacity misses

Blocking - to improve temporal locality and capacity and conflict misses