RISC Design:
Memory System Design

Virendra Singh
Associate Professor
Computer Architecture and Dependable Systems Lab
Department of Electrical Engineering
Indian Institute of Technology Bombay
http://www.ee.iitb.ac.in/~viren/
E-mail: viren@ee.iitb.ac.in

CP-226: Computer Architecture
Lecture 18 (05 April 2013)
Memory Hierarchy

Temporal Locality
- Keep recently referenced items at higher levels
- Future references satisfied quickly

Spatial Locality
- Bring neighbors of recently referenced to higher levels
- Future references satisfied quickly

CPU

I & D L1 Cache

Shared L2 Cache

Main Memory

Disk
Summary

- Memory technology
- Memory hierarchy
  - Temporal and spatial locality
- Caches
  - Placement
  - Identification
  - Replacement
  - Write Policy
- Pipeline integration of caches
Performance

CPU execution time = (CPU clock cycles + memory stall cycles) x Clock Cycle time

Memory Stall cycles = Number of misses x miss penalty

= IC x misses/Instruction x miss penalty
= IC x memory access/instruction x miss rate x miss penalty
Performance: Miss

• Miss rate
  ➢ Fraction of cache access that result in a miss

• Causes of misses
  ➢ Compulsory
    • First reference to a block
  ➢ Capacity
    • Blocks discarded and later retrieved
  ➢ Conflict
    • Program makes repeated references to multiple addresses from different blocks that map to the same location in the cache
Memory Optimization

\[
\frac{\text{Misses}}{\text{Instruction}} = \frac{\text{Miss rate} \times \text{Memory accesses}}{\text{Instruction count}} = \text{Miss rate} \times \frac{\text{Memory accesses}}{\text{Instruction}}
\]

Average memory access time = Hit time + Miss rate \times Miss penalty

- Reducing miss rate
  - Larger block size, larger cache size, higher associativity
- Reducing miss penalty
  - Multi-level caches, read priority over write
- Reducing time to hit in the cache
  - Avoid address translation when indexing caches
Memory Hierarchy Basics

- **Six** basic cache optimizations:
  - Larger block size
    - Reduces compulsory misses
    - Increases capacity and conflict misses, increases miss penalty
  - Larger total cache capacity to reduce miss rate
    - Increases hit time, increases power consumption
  - Higher associativity
    - Reduces conflict misses
    - Increases hit time, increases power consumption
Memory Hierarchy Basics

- Six basic cache optimizations:

  - Higher number of cache levels
    - Reduces overall memory access time

  - Giving priority to read misses over writes
    - Reduces miss penalty

  - Avoiding address translation in cache indexing
    - Reduces hit time
Summary

• Memory technology
• Memory hierarchy
  – Temporal and spatial locality
• Caches
  – Placement
  – Identification
  – Replacement
  – Write Policy
• Pipeline integration of caches
Thank You