Cache Coherence

CSE 471 Spring 2015
Mark Wyse
May 7, 2015

Parallelism

- ILP – Instruction
  - Instructions with no dependency can be executed in parallel
- DLP – Data
  - Same operation on multiple pieces of data
- TLP – Thread
  - Multiple instruction streams simultaneously (with or without communication)
- RLP – Request
  - Warehouse Scale Computers, Data Centers

Multiprocessors

- Single core performance limited, diminishing ILP, growing energy/power concerns
- Billions of transistors, what to do?
- Symmetric, Shared-Memory Multiprocessors (SMP)
  - Small number of cores, UMA
- Distributed, Shared-Memory Multiprocessors (DSM)
  - Large number of processors (each possibly SMP), NUMA
- Shared Address Space
  - Cores communicate through memory operations

Cache Coherence

- Private and Shared Data
  - Cache holds copy of data, used privately by a core or shared my many readers
- Cache Coherence Problem: if core X writes address A, then core Y reads address A, what value does it read?
  - Assuming A was cached locally by both X and Y

A Definition of Coherence

1. A read by P to location X following write by P to X, with no writes to X by another processor between always returns the value written to X by P

2. A read by P to location X following write by another processor to X returns value written by other processor given sufficient time and no other writes to X between

3. Writes to same location are *serialized*, that is two writes to the same location are seen in the same order by all processors.

A Definition of Coherence

1. Program Order
2. Coherent View of Memory
3. Write Serialization

Coherence Protocols

- Hardware (usually) protocol to maintain coherent view of memory between processors
  - Software Coherence: IBM Cell (Sony PS3)
- Track state at Cache Block granularity
- Snooping (Bus) vs. Directory (Multi-path interconnect)
- Write Invalidate
  - On write, all other copies are invalidated
  - Coherence Miss
- 3, 4, 5 state protocols

MSI Protocol

- Modified (Exclusive), Shared, Invalid Protocol
- Modified/Exclusive is a <u>dirty</u> state
- Shared is <u>clean</u>

MESI Protocol

- Modified, Exclusive, Shared, Invalid
- Exclusive State
  - New from MSI
  - <u>Clean</u>, data in one cache
- What transitions does this eliminate?
- What transitions does this add/duplicate?