18-640 Foundations of Computer Architecture

Lecture 14: "Multicore Cache Coherence"

John Paul Shen
October 23, 2014

- Required Reading Assignment:
  - "Parallel Computer Organization and Design," by Michel Dubois, Murali Annavaram, Per Stenstrom, Chapters 5 and 7, 2012.

18-640 Foundations of Computer Architecture

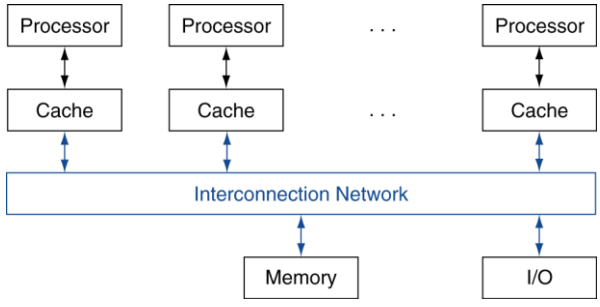
Lecture 14: "Multicore Cache Coherence"

A. Cache Coherence Problem
B. Cache Coherence Protocols
   - Write Update
   - Write Invalidate
C. Bus-Based Snoopy Protocols
   - VI & MI Protocols
   - MSI, MESI, MOESI Protocols
D. Directory-Based Protocols

Shared Memory Multiprocessors

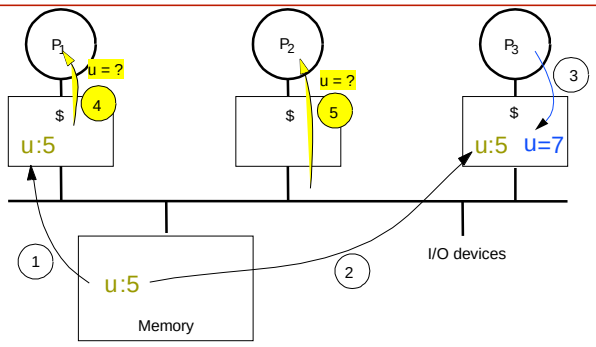
- All processor cores have access to unified physical memory
  - They can (implicitly) communicate using loads and stores
- Advantages
  - Looks like a better multithreaded processor (multitasking)
  - Requires evolutionary changes to the OS
  - Threads within an app communicate implicitly without using OS
    - Simpler to code for and lower overhead
  - App development: first focus on correctness, then on performance
- Disadvantages
  - Implicit communication is hard to optimize
  - Synchronization can get tricky
  - Higher hardware complexity for cache management

Shared Memory Multiprocessors



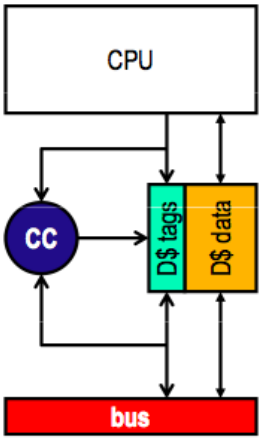
- Caches are (equally) helpful with multiprocessors
  - Reduce access latency, reduce bandwidth requirements
  - For both private and shared data across processors
- But caches introduce the problems of coherence & consistency

Cache Coherence Problem: Example



- Processors may see different values for u after event 3
- With write back caches, value written back to memory depends on which cache flushes or writes back value when doing write back
  - Threads or processes accessing main memory may see very stale value
- Unacceptable for programming, and it's frequent!

Hardware Cache Coherence Using Snooping



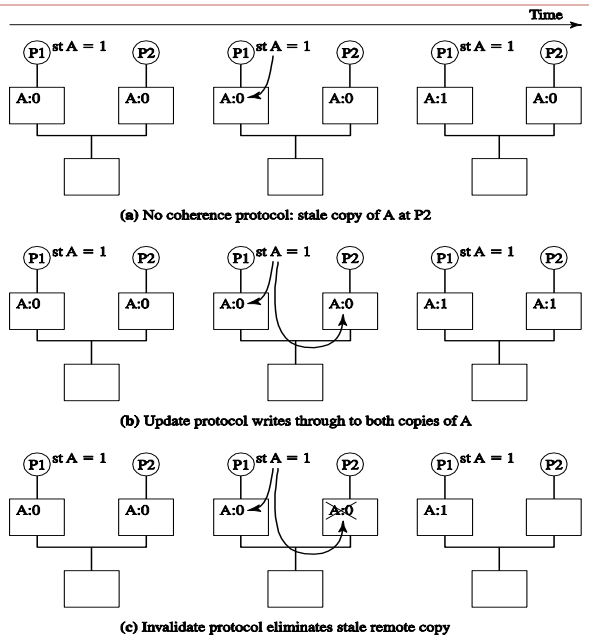
- Hardware guarantees that loads from all cores will return the value of the latest store
- Coherence mechanisms
  - Metadata to track state for cached data
  - Controller that snoops bus (or interconnect) activity and reacts if needed to adjust the state of the cache data
- There needs to be a serialization point
  - Bus, shared L2/L3, memory controller, …

Cache Coherence

- Informally, with coherent caches: accesses to a memory location *appear* to occur simultaneously in all copies of the memory location
  - "copies" ⇒ caches
- Cache coherence suggests an absolute time scale -- this is not necessary
  - What is required is the "appearance" of coherence... not absolute coherence
  - E.g. temporary incoherence between memory and a write-back cache may be OK.

Write Update vs. Write Invalidate

- Coherent Shared Memory
  - All processors see the effects of others' writes
- How/when writes are propagated
  - Determine by coherence protocol

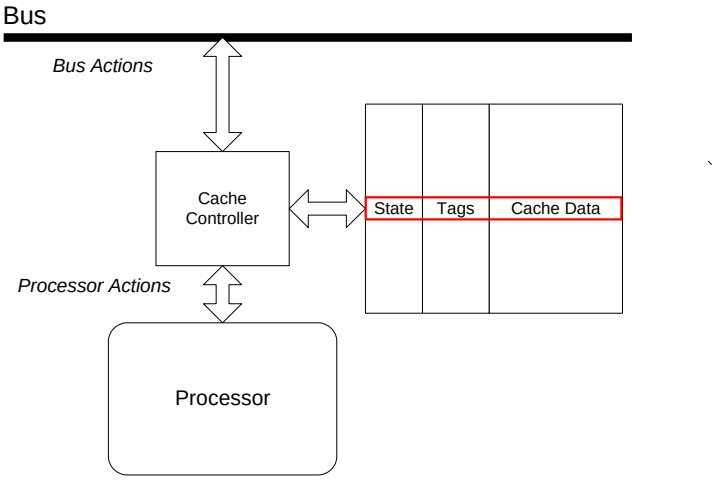


Bus-Based Snoopy Cache Coherence

- All requests broadcast on bus
- All processors and memory snoop and respond
- Cache blocks writeable at one processor or read-only at several
  - Single-writer protocol
- Snoops that hit dirty lines?
  - Flush modified data out of cache
  - Either write back to memory, then satisfy remote miss from memory, or
  - Provide dirty data directly to requestor
  - Big problem in MP systems
    - Dirty/coherence/sharing misses

Bus-Based Protocols

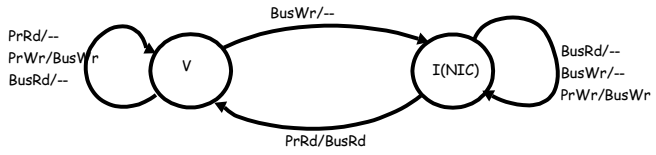
- Protocol consists of states and actions (state transitions)
- Actions can be invoked from processor or bus to the cache controller
- Coherence based on per cache line (block)



A Simple Protocol for Write-Through Caches

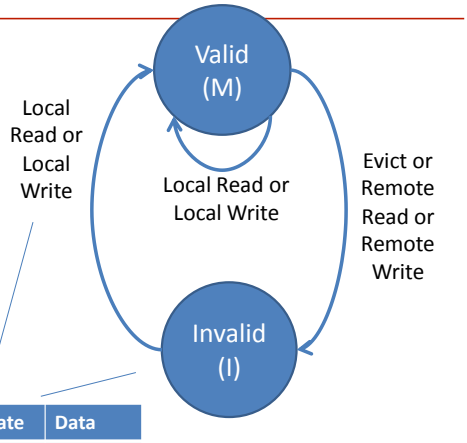
- TO SIMPLIFY ASSUME NO ALLOCATE ON STORE MISSES
  - ALL STOREs AND LOAD MISSES PROPAGATE ON THE BUS
- STOREs MAY UPDATE OR INVALIDATE CACHES

- STATE DIAGRAM
  - EACH CACHE IS REPRESENTED BY A FINITE STATE MACHINE
  - IMAGINE P IDENTICAL FSMs WORKING TOGETHER, ONE PER CACHE
  - FSM REPRESENTS THE BEHAVIOR OF A CACHE W.R.T. A MEMORY BLOCK (CACHE LINE)



Minimal Coherence Protocol for Write-Back Caches

- Blocks are always private or exclusive
- State transitions:
  - Local read: I->M, fetch, invalidate other copies
  - Local write: I->M, fetch, invalidate other copies
  - Evict: M->I, write back data
  - Remote read: M->I, write back data
  - Remote write: M->I, write back data



Cache

Tag	State	Data
A	M	…
B	I	…

Invalidate Protocol Optimization

- Observation: data often read shared by multiple CPUs
  - Add S (shared) state to protocol: MSI
- State transitions:
  - Local read: I->S, fetch shared
  - Local write: I->M, fetch modified; S->M, invalidate other copies
  - Remote read: M->I, write back data
  - Remote write: M->I, write back data

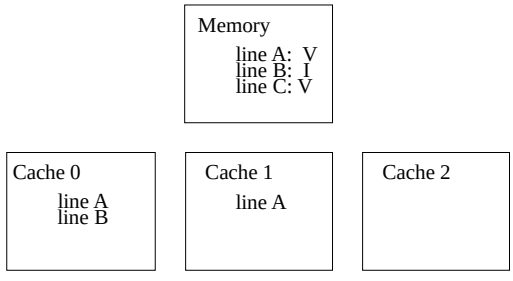
Global Coherence States

- A memory line can be present (valid) in any of the caches and/or memory
- Represent global state with an N+1 element vector
  - First N components => cache states (valid/invalid)
  - N+1st component => memory state (valid/invalid)
- Example:

  Line A: <1,1,0,1>

  Line B: <1,0,0,0>

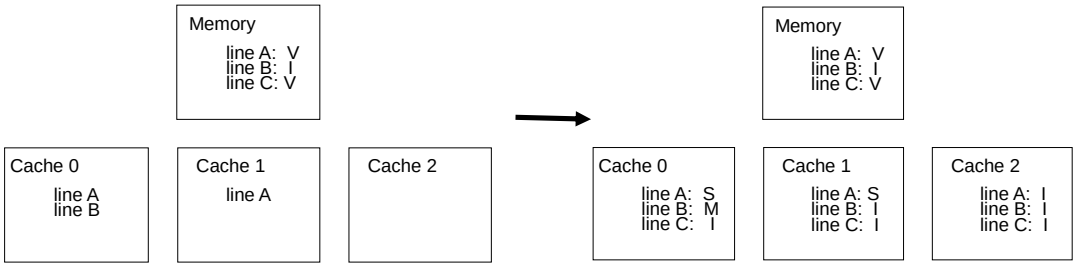
  Line C: <0,0,0,1>



Local Coherence States

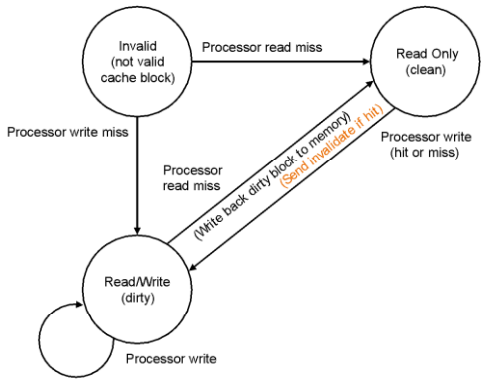
- Individual caches can maintain a summary of the state of memory lines, from a "local" perspective
  - Reduces storage for maintaining state
  - May have only partial information
- Invalid (I): <0,X,X,X....X> -- local cache does not have a valid copy; (cache miss)
  - Don't confuse invalid state with empty frame
- Shared (S): <1,X,X,X,…,1> -- local cache has a valid copy, main memory has a valid copy, other caches ??
- Modified(M): <1,0,0,..0,…0> -- local cache has only valid copy.
- Exclusive(E): <1,0,0,..0,…1> -- local cache has a valid copy, no other caches do, main memory has a valid copy.
- Owned(O): <1,X,X,X,….X> -- local cache has a valid copy, all other caches and memory may have a valid copy.
  - Only one cache can be in O state
  - <1,X,1,X,… 0> is included in O, but not included in any of the others.

Example

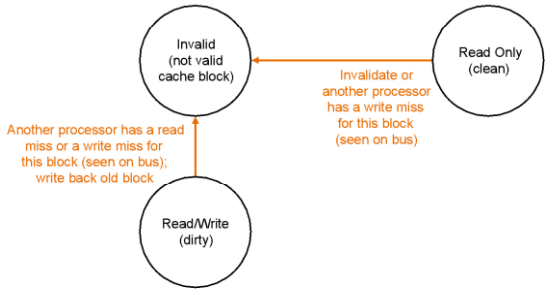


Simple Coherence Protocol FSM

[Source: Patterson/Hennessy, Comp. Org. & Design]



a. Cache state transitions using signals from the processor



b. Cache state transitions using signals from the bus

MSI Protocol

	Action and Next State						
Current State	***Processor Read***	***Processor Write***	***Eviction***		***Cache Read***	***Cache Read&M***	***Cache Upgrade***
I	***Cache Read*** **Acquire Copy → S**	***Cache Read&M*** **Acquire Copy → M**			**No Action → I**	**No Action → I**	**No Action → I**
S	**No Action → S**	***Cache Upgrade*** **→ M**	**No Action → I**		**No Action → S**	***Invalidate Frame*** **→ I**	***Invalidate Frame*** **→ I**
M	**No Action → M**	**No Action → M**	***Cache Write back*** **→ I**		**Memory inhibit; Supply data; → S**	***Invalidate Frame;*** **Memory inhibit; Supply data; → I**	

MSI Example

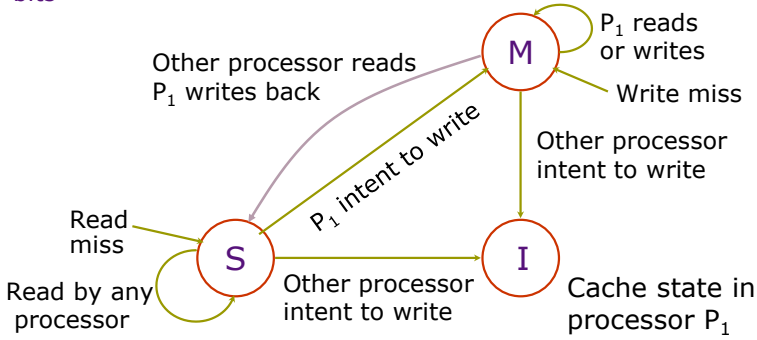
Thread Event	Bus Action	Data From	Global State	Local States: C0	C1	C2
0. Initially:			<0,0,0,1>	I	I	I
1. T0 read→	CR	Memory	<1,0,0,1>	S	I	I
2. T0 write→	CU		<1,0,0,0>	M	I	I
3. T2 read→	CR	C0	<1,0,1,1>	S	I	S
4. T1 write→	CRM	Memory	<0,1,0,0>	I	M	I

- If line is in no cache
  - Read, modify, Write requires 2 bus transactions
  - Optimization: add **E**xclusive state

MSI: A Coherence Protocol for Write Back Caches

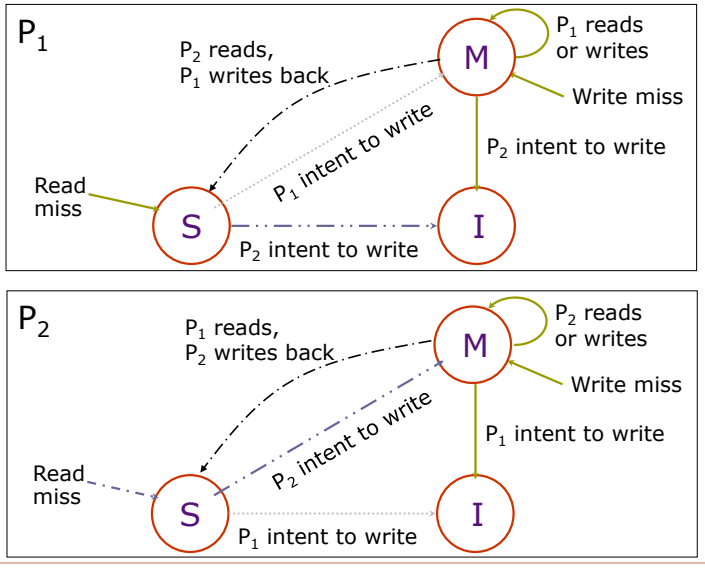
Each cache line has a tag

M: Modified
S: Shared
I: Invalid



MSI Coherence Protocol Example with 2 Cores

P_1 reads
P_1 writes
P_2 reads
P_2 writes
P_1 reads
P_1 writes
P_2 writes
P_1 writes



Invalidate Protocol Optimizations

- Observation: data can be write-private (e.g. stack frame)
  - Avoid invalidate messages in that case
  - Add E (exclusive) state to protocol: MESI
- State transitions:
  - Local read: I->E if only copy, I->S if other copies exist
  - Local write: E->M silently, S->M, invalidate other copies

MESI Protocol

- Variation used in many Intel processors
- 4-State Protocol
  - **M**odified: <1,0,0…0>
  - **E**xclusive: <1,0,0,…,1>
  - **S**hared: <1,X,X,…,1>
  - **I**nvalid: <0,X,X,…X>
- Bus/Processor Actions
  - Same as MSI
- Adds *shared* signal to indicate if other caches have a copy

MESI Protocol

	Action and Next State						
Current State	***Processor Read***	***Processor Write***	***Eviction***		***Cache Read***	***Cache Read&M***	***Cache Upgrade***
I	*Cache Read* If no sharers: → E If sharers: → S	*Cache Read&M* → M			No Action → I	No Action → I	No Action → I
S	No Action → S	*Cache Upgrade* → M	No Action → I		Respond Shared: → S	No Action → I	No Action → I
E	No Action → E	No Action → M	No Action → I		Respond Shared; → S	No Action → I	
M	No Action → M	No Action → M	*Cache Write-back* → I		Respond dirty; Write back data; → S	Respond dirty; Write back data; → I	

MESI Example

Thread Event	Bus Action	Data From	Global State	Local States: C0	C1	C2
0. Initially:			<0,0,0,1>	I	I	I
1. T0 read→	CR	Memory	<1,0,0,1>	E	I	I
2. T0 write→	none		<1,0,0,0>	M	I	I

Cache-to-Cache Transfers

- Common in many workloads:
  - T0 writes to a block: <1,0,…,0> (block in M state in T0)
  - T1 reads from block: T0 must write back, then T1 reads from memory
- In shared-bus system
  - T1 can *snarf* data from the bus during the writeback
  - Called *cache-to-cache transfer* or *dirty miss* or *intervention*
- Without shared bus
  - Must explicitly send data to requestor and to memory (for writeback)
- Known as the 4th C (cold, capacity, conflict, communication)

MESI Example 2

Thread Event	Bus Action	Data From	Global State	Local States: C0	C1	C2
0. Initially:			<0,0,0,1>	I	I	I
1. T0 read→	CR	Memory	<1,0,0,1>	E	I	I
2. T0 write→	none		<1,0,0,0>	M	I	I
3. T1 read→	CR	C0	<1,1,0,1>	S	S	I
4. T2 read→	CR	Memory	<1,1,1,1>	S	S	S

MOESI Optimization

- Observation: shared ownership prevents cache-to-cache transfer, causes unnecessary memory read
  - Add O (owner) state to protocol: MOSI/MOESI
  - Last requestor becomes the owner
  - Avoid writeback (to memory) of dirty data
  - Also called *shared-dirty* state, since memory is stale

MOESI Protocol

- Used in AMD Opteron
- 5-State Protocol
  - **M**odified: <1,0,0…0>
  - **E**xclusive: <1,0,0,…,1>
  - **S**hared: <1,X,X,…,1>
  - **I**nvalid: <0,X,X,…X>
  - **O**wned: <1,X,X,X,0> ; only one owner, memory not up to date
- Owner can supply data, so memory does not have to
  - Avoids lengthy memory access

MOESI Protocol

	Action and Next State						
Current State	*Processor Read*	*Processor Write*	*Eviction*		*Cache Read*	*Cache Read&M*	*Cache Upgrade*
I	*Cache Read* If no sharers: → E If sharers: → S	*Cache Read&M* → M			No Action → I	No Action → I	No Action → I
S	No Action → S	*Cache Upgrade* → M	No Action → I		Respond shared; → S	No Action → I	No Action → I
E	No Action → E	No Action → M	No Action → I		Respond shared; Supply data; → S	Respond shared; Supply data; → I	
O	No Action → O	*Cache Upgrade* → M	*Cache Write-back* → I		Respond shared; Supply data; → O	Respond shared; Supply data; → I	
M	No Action → M	No Action → M	*Cache Write-back* → I		Respond shared; Supply data; → O	Respond shared; Supply data; → I	

MOESI Example

Thread Event	Bus Action	Data From	Global State	local states C0	C1	C2
0. Initially:			<0,0,0,1>	I	I	I
1. T0 read→	CR	Memory	<1,0,0,1>	E	I	I
2. T0 write→	none		<1,0,0,0>	M	I	I
3. T2 read→	CR	C0	<1,0,1,0>	O	I	S
4. T1 write→	CRM	C0	<0,1,0,0>	I	M	I

MOESI Coherence Protocol

- A protocol that tracks validity, ownership, and exclusiveness
  - Modified: dirty and private
  - Owned: dirty but shared
    - Avoid writeback to memory on M->S transitions
  - Exclusive: clean but private
    - Avoid upgrade misses on private data
  - Shared
  - Invalid
- There are also some variations (MOSI and MESI)
- What happens when 2 cores read/write different words in a cache line?

Snooping with Multi-level Caches

- Private L2 caches
  - If inclusive, snooping traffic checked at the L2 level first
  - Only accesses that refer to data cached in L1 need to be forwarded
  - Saves bandwidth at the L1 cache
- Shared L2 or L3 caches
  - Can act as serialization points even if there is no bus
  - Track state of cache line and list of sharers (bit mask)
  - Essentially the shared cache acts like a coherence directory

Scaling Coherence Protocols

- The problem
  - Too much broadcast traffic for snooping (probing)
- Solution: probe filters
  - Maintain info of which address ranges that are definitely not shared or definitely shared
  - Allows filtering of snoop traffic
- Solution: directory based coherence
  - A directory stores all coherence info (e.g., sharers)
  - Consult directory before sending coherence messages
  - Caching/filtering schemes to avoid latency of 3-hops

Implementing Cache Coherence

- Snooping implementation
  - Origins in shared-memory-bus systems
  - All CPUs could observe all other CPUs requests on the bus; hence "snooping"
    - Bus Read, Bus Write, Bus Upgrade
  - React appropriately to snooped commands
    - Invalidate shared copies
    - Provide up-to-date copies of dirty lines
      - Flush (writeback) to memory, or
      - Direct intervention (*modified intervention* or *dirty miss*)
- Snooping suffers from:
  - Scalability: shared busses not practical
  - Ordering of requests without a shared bus
  - Lots of recent and on-going work on scaling snoop-based systems

Snooping Cache Coherence

- Basic idea: broadcast snoop to all caches to find owner
- Not scalable?
  - Address traffic roughly proportional to square of number of processors
  - Current implementations scale to 64/128-way (Sun/IBM) with multiple address-interleaved broadcast networks
- Inbound snoop bandwidth: big problem

$$OutboundSnoopRate = s_o = \langle CacheMissRate \rangle + \langle BusUpgradeRate \rangle$$

$$InboundSnoopRate = s_i = n \times s_o$$

Snoop Bandwidth

- Snoop filtering of various kinds is possible
- Filter snoops at sink: Jetty filter [Moshovos et al., HPCA 2001]
  - Check small "filter cache" that summarizes contents of local cache
  - Avoid power-hungry lookups in each tag array
- Filter snoops at source: Multicast snooping [Bilir et al., ISCA 1999]
  - Predict likely sharing set, snoop only predicted sharers
  - Double-check at directory to make sure
- Filter snoops at source: Region coherence
  - Concurrent work: [Cantin/Smith/Lipasti, ISCA 2005; Moshovos, ISCA 2005]
  - Check larger region of memory on every snoop; remember when no sharers
  - Snoop only on first reference to region, or when region is shared
  - Eliminate 60%+ of all snoops

Snoop Latency

- Snoop latency:
  - Must reach all nodes, return and combine responses
  - Topology matters: ring, mesh, torus, hypercube
  - No obvious solutions
- Parallelism: fundamental advantage of snooping
  - Broadcast exposes parallelism, enables speculative latency reduction

Scaleable Cache Coherence

- No physical bus but still snoop
  - Point-to-point tree structure (indirect) or ring
  - Root of tree or ring provide ordering point
  - Use some scalable network for data (ordering less important)
- Or, use level of indirection through directory
  - Directory at memory remembers:
    - Which processor is "single writer"
    - Which processors are "shared readers"
  - Level of indirection has a price
    - Dirty misses require 3 hops instead of two
      - Snoop: Requestor->Owner->Requestor
      - Directory: Requestor->Directory->Owner->Requestor

Implementing Cache Coherence

- Directory implementation
  - Extra bits stored in memory (directory) record state of line
  - Memory controller maintains coherence based on the current state
  - Other CPUs' commands are not snooped, instead:
    - Directory forwards relevant commands
  - Powerful filtering effect: only observe commands that you need to observe
  - Meanwhile, bandwidth at directory scales by adding memory controllers as you increase size of the system
    - Leads to very scalable designs (100s to 1000s of CPUs)
- Directory shortcomings
  - Indirection through directory has latency penalty
  - If shared line is dirty in other CPU's cache, directory must forward request, adding latency
  - This can severely impact performance of applications with heavy sharing (e.g. relational databases)

Directory Protocol Implementation

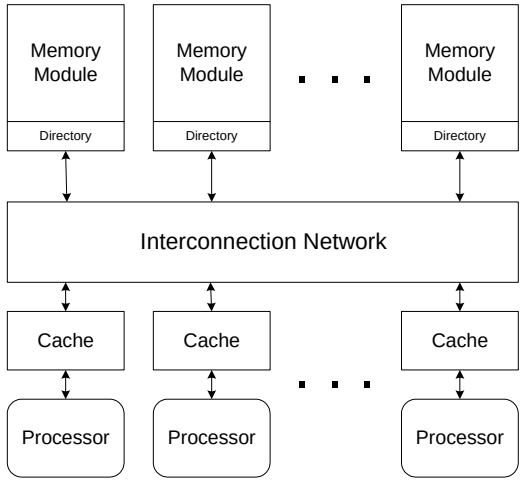
- Basic idea: Centralized directory keeps track of data location(s)
- Scalable
  - Address traffic roughly proportional to number of processors
  - Directory & traffic can be distributed with memory banks (interleaved)
  - Directory cost (SRAM) or latency (DRAM) can be prohibitive
- Presence bits track sharers
  - Full map (N processors, N bits): cost/scalability
  - Limited map (limits number of sharers)
  - Coarse map (identifies board/node/cluster; must use broadcast)
- Vectors track sharers
  - Point to shared copies
  - Fixed number, linked lists (SCI), caches chained together
  - Latency vs. cost vs. scalability

Directory Protocol Latency

- Access to non-shared data
  - Overlap directory read with data read
  - Best possible latency given distributed memory
- Access to shared data
  - Dirty miss, modified intervention
  - Shared intervention?
    - If DRAM directory, no gain
    - If directory cache, possible gain (use F state)
  - No inherent parallelism
  - Indirection adds latency
  - Minimum 3 hops, often 4 hops

Directory-Based Cache Coherence

- An alternative for large, scalable MPs
- Can be based on any of the protocols discussed thus far
  - We will use MSI
- Memory Controller becomes an active participant
- Sharing info held in memory directory
  - Directory may be distributed
- Use point-to-point messages
- Network is not totally ordered

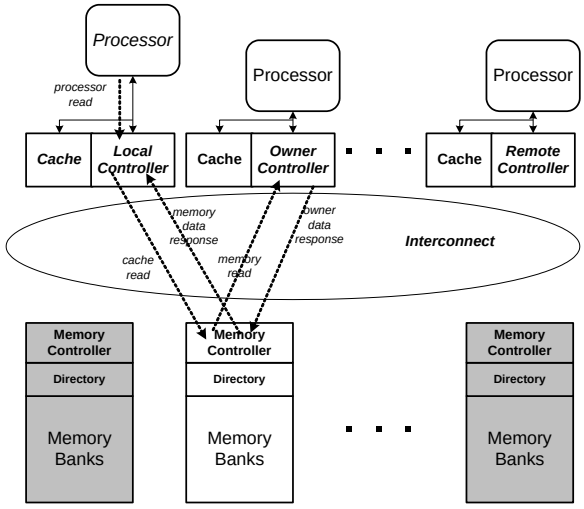


Example: Simple Directory Protocol

- Local cache controller states
  - M, S, I as before
- Local directory states
  - **S**hared: <1,X,X,…1>; one or more proc. has copy; memory is up-to-date
  - **M**odified: <0,1,0,….,0> one processor has copy; memory does not have a valid copy
  - **U**ncached: <0,0,…0,1> none of the processors has a valid copy
- Directory also keeps track of sharers
  - Can keep global state vector in full
  - e.g. via a bit vector

Example

- *Local cache* suffers load miss
- Line in *remote cache* in M state
  - It is the *owner*
- Four messages send over network
  - Cache read from local controller to home memory controller
  - Memory read to remote cache controller
  - Owner data back to memory controller; change state to S
  - Memory data back to local cache; change state to S



Cache Controller State Table

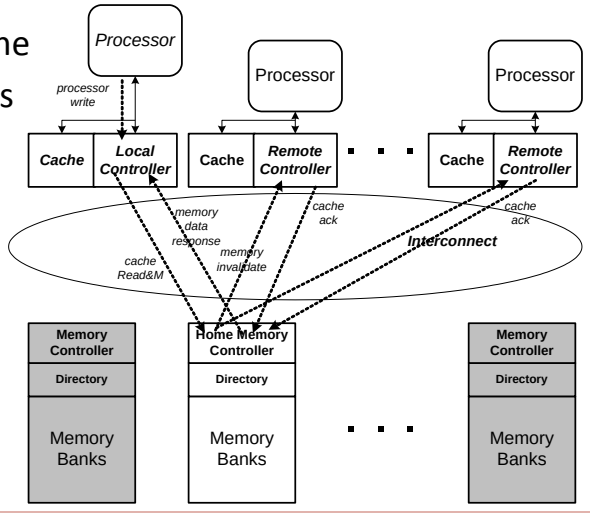
	Cache Controller Actions and Next States								
	from Processor Side				from Memory Side				
Current State	*Processor Read*	*Processor Write*	*Eviction*		*Memory Read*	*Memory Read&M*	*Memory Invalidate*	*Memory Upgrade*	*Memory Data*
I	*Cache Read* → I'	*Cache Read&M* → I''					No Action → I		
S	No Action → S	*Cache Upgrade* → S'	No Action* → I				Invalidate Frame; *Cache ACK*; → I		
M	No Action → M	No Action → M	*Cache Write-back* → I		*Owner Data;* → S	*Owner Data;* → I	Invalidate Frame; *Cache ACK*; → I		
I'									Fill Cache → S
I''									Fill Cache → M
S'								No Action → M	

Memory Controller State Table

	Memory Controller Actions and Next States					
	command from Local Cache Controller				response from Remote Cache Controller	
Current Directory State	*Cache Read*	*Cache Read&M*	*Cache Upgrade*		*Data Write-back*	*Cache ACK*	*Owner Data*
U	*Memory Data*; Add Requestor to Sharers; → S	*Memory Data*; Add Requestor to Sharers; → M					
S	*Memory Data*; Add Requestor to Sharers; → S	*Memory Invalidate* All Sharers; → M'	*Memory Upgrade* All Sharers; → M''		No Action → I		
M	*Memory Read* from Owner; → S'	*Memory Read&M*; to Owner → M'			Make Sharers Empty; → U		
S'							*Memory Data* to Requestor; Write memory; Add Requestor to Sharers; → S
M'						When all ACKS *Memory Data*; → M	*Memory Data* to Requestor; → M
M''						When all ACKS then → M	

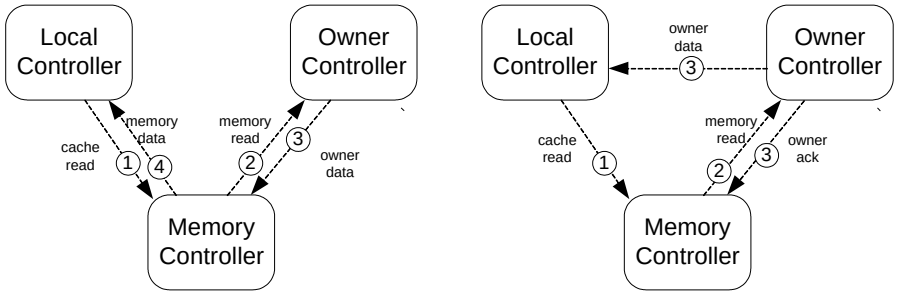
Another Example

- Local write (miss) to shared line
- Requires invalidations and acks



Variation: Three Hop Protocol

- Have owner send data directly to local controller
- Owner Acks to Memory Controller in parallel



Example Sequence

- Similar to earlier sequences

Thread Event	Controller Actions	Data From	global state	local states: C0	C1	C2
0. Initially:			<0,0,0,1>	I	I	I
1. T0 read→	CR,MD	Memory	<1,0,0,1>	S	I	I
2. T0 write→	CU, MU*,MD		<1,0,0,0>	M	I	I
3. T2 read→	CR,MR,MD	C0	<1,0,1,1>	S	I	S
4. T1 write→	CRM,MI,CA,MD	Memory	<0,1,0,0>	I	M	I

Directory Protocol Optimizations

- Remove dead blocks from cache:
  - Eliminate 3- or 4-hop latency
  - Dynamic Self-Invalidation [Lebeck/Wood, ISCA 1995]
  - Last touch prediction [Lai/Falsafi, ISCA 2000]
  - Dead block prediction [Lai/Fide/Falsafi, ISCA 2001]
- Predict sharers
  - Prediction in coherence protocols [Mukherjee/Hill, ISCA 1998]
  - Instruction-based prediction [Kaxiras/Goodman, ISCA 1999]
  - Sharing prediction [Lai/Falsafi, ISCA 1999]
- Hybrid snooping/directory protocols
  - Improve latency by snooping, conserve bandwidth with directory
  - Multicast snooping [Bilir et al., ISCA 1999; Martin et al., ISCA 2003]
  - Bandwidth-adaptive hybrid [Martin et al., HPCA 2002]
  - Token Coherence [Martin et al., ISCA 2003]
  - Virtual Tree Coherence [Enright Jerger MICRO 2008]

Update Protocols

- Basic idea:
  - All writes (updates) are made visible to all caches:
    - (address,value) tuples sent "everywhere"
    - Similar to write-through protocol for uniprocessor caches
  - Obviously not scalable beyond a few processors
  - No one actually builds machines this way
- Simple optimization
  - Send updates to memory/directory
  - Directory propagates updates to all known copies: less bandwidth
- Further optimizations: combine & delay
  - Write-combining of adjacent updates (if consistency model allows)
  - Send write-combined data
  - Delay sending write-combined data until requested
- Logical end result
  - Writes are combined into larger units, updates are delayed until needed
  - Effectively the same as invalidate protocol
- Of historical interest only (Firefly and Dragon protocols)

Update vs Invalidate

- [Weber & Gupta, ASPLOS3]
  - Consider sharing patterns
- No Sharing
  - Independent threads
  - Coherence due to thread migration
  - Update protocol performs many wasteful updates
- Read-Only
  - No significant coherence issues; most protocols work well
- Migratory Objects
  - Manipulated by one processor at a time
  - Often protected by a lock
  - Usually a write causes only a single invalidation
  - E state useful for Read-modify-Write patterns
  - Update protocol could proliferate copies

Update vs Invalidate, contd.

- Synchronization Objects
  - Locks
  - Update could reduce spin traffic invalidations
  - Test & Test&Set w/ invalidate protocol would work well
- Many Readers, One Writer
  - Update protocol may work well, but writes are relatively rare
- Many Writers/Readers
  - Invalidate probably works better
  - Update will proliferate copies
- What is used today?
  - Invalidate is dominant
  - CMP *may* change this assessment
    - more on-chip bandwidth