Review

- Caches contain all information on state of cached memory blocks
- Snooping cache over shared medium for smaller MP by invalidating other cached copies on write
- Sharing cached data ⇒ Coherence (values returned by a read), Consistency (when a written value will be returned by a read)

Outline

- Review
- Coherence traffic and Performance on MP
- Directory-based protocols and examples
- Administrivia
- Synchronization
- Relaxed Consistency Models
- Fallacies and Pitfalls
- Cautionary Tale
- Conclusion

Performance of Symmetric Shared-Memory Multiprocessors

- Cache performance is combination of
  1. Uniprocessor cache miss traffic
  2. Traffic caused by communication
     - Results in invalidations and subsequent cache misses
- 4th C: coherence miss
  - Joins Compulsory, Capacity, Conflict
Coherency Misses

1. True sharing misses arise from the communication of data through the cache coherence mechanism
   - Invalidates due to 1st write to shared block
   - Reads by another CPU of modified block in different cache
   - Miss would still occur if block size were 1 word

2. False sharing misses when a block is invalidated because some word in the block, other than the one being read, is written into
   - Invalidation does not cause a new value to be communicated, but only causes an extra cache miss
   - Block is shared, but no word in block is actually shared ⇒ miss would not occur if block size were 1 word

Example: True v. False Sharing v. Hit?

- Assume x1 and x2 in same cache block. P1 and P2 both read x1 and x2 before.

<table>
<thead>
<tr>
<th>Time</th>
<th>P1</th>
<th>P2</th>
<th>True, False, Hit? Why?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Write x1</td>
<td>True miss; invalidate x1 in P2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Read x2</td>
<td>False miss; x1 irrelevant to P2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Write x1</td>
<td>False miss; x1 irrelevant to P2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Write x2</td>
<td>False miss; x1 irrelevant to P2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Read x2</td>
<td>True miss; invalidate x2 in P1</td>
<td></td>
</tr>
</tbody>
</table>

MP Performance 4 Processor
Commercial Workload: OLTP, Decision Support (Database), Search Engine

- True sharing and false sharing unchanged going from 1 MB to 8 MB (L3 cache)
- Uniprocessor cache misses improve with cache size increase (Instruction, Capacity/Conflict, Compulsory)

MP Performance 2MB Cache
Commercial Workload: OLTP, Decision Support (Database), Search Engine

- True sharing, false sharing increase going from 1 to 8 CPUs
A Cache Coherent System Must:

- Provide set of states, state transition diagram, and actions
- Manage coherence protocol
  - (0) Determine when to invoke coherence protocol
  - (a) Find info about state of block in other caches to determine action
    - whether need to communicate with other cached copies
  - (b) Locate the other copies
  - (c) Communicate with those copies (invalidate/update)
- (0) is done the same way on all systems
  - state of the line is maintained in the cache
  - protocol is invoked if an “access fault” occurs on the line
- Different approaches distinguished by (a) to (c)

Bus-based Coherence

- All of (a), (b), (c) done through broadcast on bus
  - faulting processor sends out a “search”
  - others respond to the search probe and take necessary action
- Could do it in scalable network too
  - broadcast to all processors, and let them respond
- Conceptually simple, but broadcast doesn’t scale with p
  - on bus, bus bandwidth doesn’t scale
  - on scalable network, every fault leads to at least \( p \) network transactions
- Scalable coherence:
  - can have same cache states and state transition diagram
  - different mechanisms to manage protocol

Scalable Approach: Directories

- Every memory block has associated directory information
  - keeps track of copies of cached blocks and their states
  - on a miss, find directory entry, look it up, and communicate only with the nodes that have copies if necessary
  - in scalable networks, communication with directory and copies is through network transactions
- Many alternatives for organizing directory information

Basic Operation of Directory

- \( k \) processors.
- With each cache-block in memory: \( k \) presence-bits, 1 dirty-bit
- With each cache-block in cache: 1 valid bit, and 1 dirty (owner) bit

- Read from main memory by processor \( i \):
  - If dirty-bit OFF then \{ read from main memory; turn \( p[i] \) ON; \}
  - If dirty-bit ON then \{ recall line from dirty proc (cache state to shared); update memory; turn dirty-bit OFF; turn \( p[i] \) ON; supply recalled data to \( i \); \}

- Write to main memory by processor \( i \):
  - If dirty-bit OFF then \{ supply data to \( i \); send invalidations to all caches that have the block; turn dirty-bit ON; turn \( p[i] \) ON; … \}
### Directory Protocol

- **Similar to Snoopy Protocol: Three states**
  - **Shared**: ≥ 1 processors have data, memory up-to-date
  - **Uncached**: (no processor has it; not valid in any cache)
  - **Exclusive**: 1 processor (owner) has data; memory out-of-date

- In addition to cache state, must track which **processors** have data when in the shared state (usually bit vector, 1 if processor has copy)

- **Keep it simple(r):**
  - Writes to non-exclusive data ⇒ write miss
  - Processor blocks until access completes
  - Assume messages received and acted upon in order sent

### CS 252 Administrivia

- **Due Friday:** Problem Set and Comments on 2 papers
  - Problem Set Assignment done in pairs

- **Be sure to comment**
  - Amdahl: How long is paper? How much of it is Amdahl’s Law?
  - What other comments about parallelism besides Amdahl’s Law?
  - Hochstein: What programming styles investigated? What was methodology? How would you redesign the experiment they did? What other metrics would be important to capture? Assuming these results of programming productivity reflect the real world, what should architectures of the future do (or not do)?

- **Monday March 20 Quiz 5-8 PM 405 Soda**
- **Wed March 22 no class: project meetings in 635 Soda**

### Computers in the News

- **“Core” new microarchitecture; last Pentium 4 (2000)**
  - Wide Dynamic Execution: 4 issue + Combine 2 simple instructions into 1 powerful (“macrofusion”)
  - Advanced Digital Media Boost: All SSE instructions 1 clock cycle
  - Smart Memory Access: lets one core control the whole cache when the other core is idle, and governs how the same data can be shared by both cores
  - Intelligent Power Capability: shut down unneeded portions of chip

- **80% more performance, 40% less power**
- **4 core chips in 2007 (2 copies of dual core?)**
  - CTO: “Intel is taking a conservative approach that focuses on single-thread performance. You won't see mediocre thread performance just for the sake of getting multiple cores on a die.”

- **CTO urged software companies to support multicore designs with software that can efficiently divide tasks among multiple execution threads. "It's really time to get onboard the multithreaded train"**
**Directory Protocol Messages (Fig 4.22)**

<table>
<thead>
<tr>
<th>Message type</th>
<th>Source</th>
<th>Destination</th>
<th>Msg Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read miss</td>
<td>Local cache</td>
<td>Home directory</td>
<td>P, A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Write miss</td>
<td>Local cache</td>
<td>Home directory</td>
<td>P, A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invalidate</td>
<td>Home directory</td>
<td>Remote caches</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fetch</td>
<td>Home directory</td>
<td>Remote cache</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fetch/Invalidate</td>
<td>Home directory</td>
<td>Remote cache</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data value reply</td>
<td>Home directory</td>
<td>Local cache</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data write back</td>
<td>Remote cache</td>
<td>Home directory</td>
<td>A, Data</td>
</tr>
</tbody>
</table>

- Processor P reads data at address A; make P a read sharer and request data
- Processor P has a write miss at address A; make P the exclusive owner and request data
- Invalidate a shared copy at address A
- Fetch the block at address A and send it to its home directory; change the state of A in the remote cache to shared
- Fetch the block at address A and send it to its home directory; invalidate the block in the cache
- Return a data value from the home memory (read miss response)
- Write back a data value for address A (invalidate response)

---

**State Transition Diagram for One Cache Block in Directory Based System**

- States identical to snoopy case; transactions very similar
- Transitions caused by read misses, write misses, invalidates, data fetch requests
- Generates read miss & write miss message to home directory
- Write misses that were broadcast on the bus for snooping ⇒ explicit invalidate & data fetch requests
- Note: on a write, a cache block is bigger, so need to read the full cache block

---

**CPU -Cache State Machine**

- State machine for CPU requests for each memory block
- Invalid state if in memory

---

**State Transition Diagram for Directory**

- Same states & structure as the transition diagram for an individual cache
- 2 actions: update of directory state & send messages to satisfy requests
- Tracks all copies of memory block
- Also indicates an action that updates the sharing set, Sharers, as well as sending a message
## Example Directory Protocol

- **Message sent to directory causes two actions:**
  - Update the directory
  - More messages to satisfy request

- **Block is in Uncached state:** The copy in memory is the current value; only possible requests for that block are:
  - Read miss: requesting processor sent data from memory & requestor made only sharing node; state of block made Shared.
  - Write miss: requesting processor is sent the value & becomes the Sharing node. The block is made Exclusive to indicate that the only valid copy is cached. Sharers indicates the identity of the owner.

- **Block is Shared ⇒ the memory value is up-to-date:**
  - Read miss: requesting processor is sent back the data from memory & requesting processor is added to the sharing set.
  - Write miss: requesting processor is sent the value. All processors in the set Sharers are sent invalidate messages, & Sharers is set to identity of requesting processor. The state of the block is made Exclusive.

---

### Example

<table>
<thead>
<tr>
<th>Processor 1</th>
<th>Processor 2</th>
<th>Interconnect</th>
<th>Directory</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1: Write 10 to A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1: Read A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2: Read A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2: Write 20 to A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2: Write 40 to A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A1 and A2 map to the same cache block
A1 and A2 map to the same cache block
Example

Processor 1 Processor 2 Interconnect Directory Memory

<table>
<thead>
<tr>
<th>step</th>
<th>P1 State</th>
<th>P1 Value</th>
<th>P2 State</th>
<th>P2 Value</th>
<th>Action</th>
<th>Proc</th>
<th>Addr Value</th>
<th>Addr Value</th>
<th>State</th>
<th>Addr Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1: Write 10 to A1</td>
<td>Ext.</td>
<td>A1</td>
<td>10</td>
<td>A2Dk</td>
<td>P1</td>
<td>A1</td>
<td>10</td>
<td>Excl.</td>
<td>P1</td>
<td>A1</td>
</tr>
</tbody>
</table>

A1 and A2 map to the same cache block (but different memory block addresses A1 ≠ A2)

Implementing a Directory

- We assume operations atomic, but they are not; reality is much harder; must avoid deadlock when run out of buffers in network (see Appendix E)
- Optimizations:
  - read miss or write miss in Exclusive: send data directly to requestor from owner vs. 1st to memory and then from memory to requestor

Basic Directory Transactions

- Read request to directory
- Reply with owner identity
- Read req. to owner
- Data reply
- Dir. ctrl
- Read pA

Example Directory Protocol (1st Read)

- P1: pA
- R/req
- S
- P1
- Id va -> rd pA
- $ P2

Node with dirty copy
- Read miss to a block in dirty state
- Write miss to a block with sharers
Example Directory Protocol (Read Share)

Example Directory Protocol (Wr to shared)

Example Directory Protocol (Wr to Ex)

A Popular Middle Ground

- Two-level “hierarchy”
- Individual nodes are multiprocessors, connected non-hierarchically
  - e.g. mesh of SMPs
- Coherence across nodes is directory-based
  - directory keeps track of nodes, not individual processors
- Coherence within nodes is snooping or directory
  - orthogonal, but needs a good interface of functionality
- SMP on a chip directory + snoop?
Synchronization

• Why Synchronize? Need to know when it is safe for different processes to use shared data
• Issues for Synchronization:
  – Uninterruptable instruction to fetch and update memory (atomic operation);
  – User level synchronization operation using this primitive;
  – For large scale MPs, synchronization can be a bottleneck; techniques to reduce contention and latency of synchronization

Uninterruptable Instruction to Fetch and Update Memory

• Atomic exchange: interchange a value in a register for a value in memory
  0 ⇒ synchronization variable is free
  1 ⇒ synchronization variable is locked and unavailable
  – Set register to 1 & swap
  – New value in register determines success in getting lock
    0 if you succeeded in setting the lock (you were first)
    1 if other processor had already claimed access
  – Key is that exchange operation is indivisible
• Test-and-set: tests a value and sets it if the value passes the test
• Fetch-and-increment: it returns the value of a memory location and atomically increments it
  – 0 ⇒ synchronization variable is free

Uninterruptable Instruction to Fetch and Update Memory

• Hard to have read & write in 1 instruction: use 2 instead
• Load linked (or load locked) + store conditional
  – Load linked returns the initial value
  – Store conditional returns 1 if it succeeds (no other store to same memory location since preceding load) and 0 otherwise
• Example doing atomic swap with LL & SC:
  try: mov R3,R4; mov exchange value
  li R2,0(R1); load linked
  sc R3,0(R1); store conditional
  beqz R3,try; branch store fails (R3 = 0)
  mov R4,R2; put load value in R4
• Example doing fetch & increment with LL & SC:
  try: li R2,#1; load linked
  addi R2,2,#1; increment (OK if reg–reg)
  sc R2,0(R1); store conditional
  beqz R2,try; branch store fails (R2 = 0)

User Level Synchronization—Operation Using this Primitive

• Spin locks: processor continuously tries to acquire, spinning around a loop trying to get the lock
  li R2,#1
  exch R2,0(R1); atomic exchange
  bnez R2,lockit; already locked?
• What about MP with cache coherency?
  – Want to spin on cache copy to avoid full memory latency
  – Likely to get cache hits for such variables
• Problem: exchange includes a write, which invalidates all other copies; this generates considerable bus traffic
• Solution: start by simply repeatedly reading the variable; when it changes, then try exchange ("test and test&set"):
Another MP Issue: Memory Consistency Models

- What is consistency? When must a processor see the new value? e.g., seems that
  P1: A = 0; P2: B = 0;
  ....
  A = 1; B = 1;
  L1: if (B == 0) ...
  L2: if (A == 0) ...
- Impossible for both if statements L1 & L2 to be true?
  What if write invalidate is delayed & processor continues?

- Memory consistency models: what are the rules for such cases?
- Sequential consistency: result of any execution is the same as if the accesses of each processor were kept in order and the accesses among different processors were interleaved ⇒ assignments before ifs above
- SC: delay all memory accesses until all invalidates done

Memory Consistency Model

- Schemes faster execution to sequential consistency
- Not an issue for most programs; they are synchronized
  - A program is synchronized if all access to shared data are ordered by synchronization operations
  - write (x)
  - ... release (s) \{unlock\}
  - ... acquire (s) \{lock\}
  - ... read(x)
- Only those programs willing to be nondeterministic are not synchronized: “data race”: outcome f(proc. speed)
- Several Relaxed Models for Memory Consistency since most programs are synchronized; characterized by their attitude towards: RAR, WAR, RAW, WAW to different addresses

Relaxed Consistency Models: The Basics

- **Key idea**: allow reads and writes to complete out of order, but to use synchronization operations to enforce ordering, so that a synchronized program behaves as if the processor were sequentially consistent
  - By relaxing orderings, may obtain performance advantages
  - Also specifies range of legal compiler optimizations on shared data
  - Unless synchronization points are clearly defined and programs are synchronized, compiler could not interchange read and write of 2 shared data items because might affect the semantics of the program
- 3 major sets of relaxed orderings:
  1. W → R ordering (all writes completed before next read)
     - Because retains ordering among writes, many programs that operate under sequential consistency operate under this model, without additional synchronization. Called **processor consistency**
  2. W → W ordering (all writes completed before next write)
  3. R → W and R → R orderings, a variety of models depending on ordering restrictions and how synchronization operations enforce ordering
- Many complexities in relaxed consistency models: defining precisely what it means for a write to complete; deciding when processors can see values that it has written

Mark Hill observation

- Instead, use speculation to hide latency from strict consistency model
  - If processor receives invalidation for memory reference before it is committed, processor uses speculation recovery to back out computation and restart with invalidated memory reference
  1. Aggressive implementation of sequential consistency or processor consistency gains most of advantage of more relaxed models
  2. Implementation adds little to implementation cost of speculative processor
  3. Allows the programmer to reason using the simpler programming models
Cross Cutting Issues: Performance Measurement of Parallel Processors

- **Performance:** how well scale as increase Proc
- **Speedup fixed as well as scaleup of problem**
  - Assume benchmark of size $n$ on $p$ processors makes sense: how scale benchmark to run on $m \times p$ processors?
  - **Memory-constrained scaling:** keeping the amount of memory used per processor constant
  - **Time-constrained scaling:** keeping total execution time, assuming perfect speedup, constant
- **Example:** 1 hour on 10 P, time ~ $O(n^3)$, 100 P?
  - **Time-constrained scaling:** 1 hour $\Rightarrow 10^{10}n \Rightarrow 2.15n$ scale up
  - **Memory-constrained scaling:** 10n size $\Rightarrow 10^{3/10} \Rightarrow 100X$ or 100 hours! 10X processors for 100X longer?!
  - Need to know application well to scale: # iterations, error tolerance

Fallacy: Amdahl’s Law doesn’t apply to parallel computers

- **Since some part linear, can’t go 100X?**
- **1987 claim to break it, since 1000X speedup**
  - researchers scaled the benchmark to have a data set size that is 1000 times larger and compared the uniprocessor and parallel execution times of the scaled benchmark. For this particular algorithm the sequential portion of the program was constant independent of the size of the input, and the rest was fully parallel—hence, linear speedup with 1000 processors
- **Usually sequential scale with data too**

Fallacy: Linear speedups are needed to make multiprocessors cost-effective

- **Mark Hill & David Wood 1995 study**
- Compare costs SGI uniprocessor and MP
  - Uniprocessor = $38,400 + $100 * MB
  - MP = $81,600 + $20,000 * P + $100 * MB
  - 1 GB, uni = $138k v. mp = $181k + $20k * P
- What speedup for better MP cost performance?
  - 8 proc = $341k; $341k/138k $\Rightarrow 2.5X$
  - 16 proc $\Rightarrow$ need only 3.6X, or 25% linear speedup
- Even if need some more memory for MP, not linear

Fallacy: Scalability is almost free

- “build scalability into a multiprocessor and then simply offer the multiprocessor at any point on the scale from a small number of processors to a large number”
- **Cray T3E scales to 2048 CPUs vs. 4 CPU Alpha**
  - At 128 CPUs, it delivers a peak bisection BW of 38.4 GB/s, or 300 MB/s per CPU (uses Alpha microprocessor)
  - Compaq Alphaserver ES40 up to 4 CPUs and has 5.6 GB/s of interconnect BW, or 1400 MB/s per CPU
- Build apps that scale requires significantly more attention to load balance, locality, potential contention, and serial (or partly parallel) portions of program. 10X is very hard
Pitfall: Not developing SW to take advantage (or optimize for) multiprocessor architecture

- SGI OS protects the page table data structure with a single lock, assuming that page allocation is infrequent
- Suppose a program uses a large number of pages that are initialized at start-up
- Program parallelized so that multiple processes allocate the pages
- But page allocation requires lock of page table data structure, so even an OS kernel that allows multiple threads will be serialized at initialization (even if separate processes)

Answers to 1995 Questions about Parallelism

- In the 1995 edition of this text, we concluded the chapter with a discussion of two then current controversial issues.
- 1. What architecture would very large scale, microprocessor-based multiprocessors use?
- 2. What was the role for multiprocessing in the future of microprocessor architecture?
- Answer 1. Large scale multiprocessors did not become a major and growing market ⇒ clusters of single microprocessors or moderate SMPs
- Answer 2. Astonishingly clear. For at least for the next 5 years, future MPU performance comes from the exploitation of TLP through multicore processors vs. exploiting more ILP

Cautionary Tale

- Key to success of birth and development of ILP in 1980s and 1990s was software in the form of optimizing compilers that could exploit ILP
- Similarly, successful exploitation of TLP will depend as much on the development of suitable software systems as it will on the contributions of computer architects
- Given the slow progress on parallel software in the past 30+ years, it is likely that exploiting TLP broadly will remain challenging for years to come

And in Conclusion …

- Snooping and Directory Protocols similar; bus makes snooping easier because of broadcast (snooping ⇒ uniform memory access)
- Directory has extra data structure to keep track of state of all cache blocks
- Distributing directory ⇒ scalable shared address multiprocessor ⇒ Cache coherent, Non uniform memory access
- MPs are highly effective for multiprogrammed workloads
- MPs proved effective for intensive commercial workloads, such as OLTP (assuming enough I/O to be CPU-limited), DSS applications (where query optimization is critical), and large-scale, web searching applications