

CSCI 211 Computer System Architecture

Lec 9 – Multiprocessor II

Xiuzhen Cheng
Department of Computer Sciences
The George Washington University

Adapted from the slides by Dr. David Patterson @ UC Berkeley

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 \Rightarrow Coherence (values returned by a read), Consistency (when a written value will be returned by a read)

3/27/2008

Csci 211 – Lec 9

2

Outline

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

3/27/2008

Csci 211 – Lec 9

3

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

3/27/2008

Csci 211 – Lec 9

4

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 \Rightarrow miss would not occur if block size were 1 word

3/27/2008

Csci 211 – Lec 9

5

Example: True v. False Sharing v. Hit?

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

Time	P1	P2	True, False, Hit? Why?
1	Write x1		True miss; invalidate x1 in P2
2		Read x2	False miss; x1 irrelevant to P2
3	Write x1		False miss; x1 irrelevant to P2
4		Write x2	False miss; x1 irrelevant to P2
5	Read x2		True miss; invalidate x2 in P1

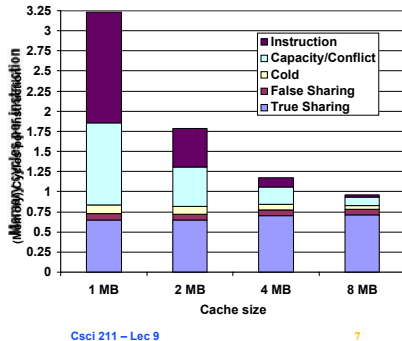
3/27/2008

Csci 211 – Lec 9

6

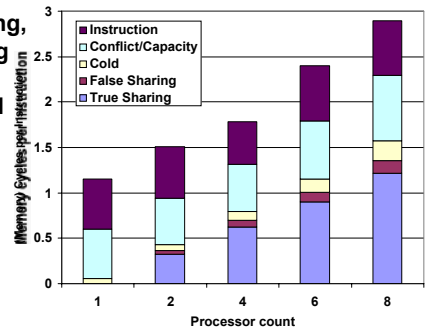
MP Performance 4 Processor Commercial Workload: OLTP, Decision Support (Database), Search Engine

- 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)

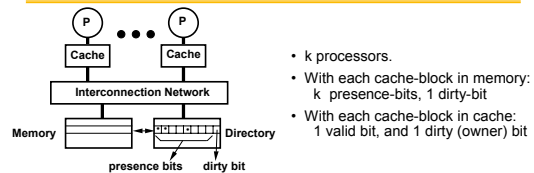
Bus-based Coherence

- 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 \Rightarrow write miss
 - Processor blocks until access completes
 - Assume messages received and acted upon in order sent

3/27/2008

Csci 211 – Lec 9

13

Directory Protocol

- No bus and don't want to broadcast:
 - interconnect no longer single arbitration point
 - all messages have explicit responses
- Terms: typically 3 processors involved
 - **Local node** where a request originates
 - **Home node** where the memory location of an address resides
 - **Remote node** has a copy of a cache block, whether exclusive or shared
- Example messages on next slide:
 - P = processor number, A = address

3/27/2008

Csci 211 – Lec 9

14

Directory Protocol Messages (Fig 4.20)

Message type	Source	Destination	Msg Content
Read miss	Local cache	Home directory	P, A
- Processor P reads data at address A; make P a read sharer and request data			
Write miss	Local cache	Home directory	P, A
- Processor P has a write miss at address A; make P the exclusive owner and request data			
Invalidate	Home directory	Remote caches	A
- Invalidate a shared copy at address A			
Fetch	Home directory	Remote cache	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/Invalidate	Home directory	Remote cache	A
- Fetch the block at address A and send it to its home directory; invalidate the block in the cache			
Data value reply	Home directory	Local cache	Data
- Return a data value from the home memory (read miss response)			
Data write back	Remote cache	Home directory	A, Data
- 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 \Rightarrow explicit invalidate & data fetch requests
- Note: on a write, a cache block is bigger, so need to read the full cache block

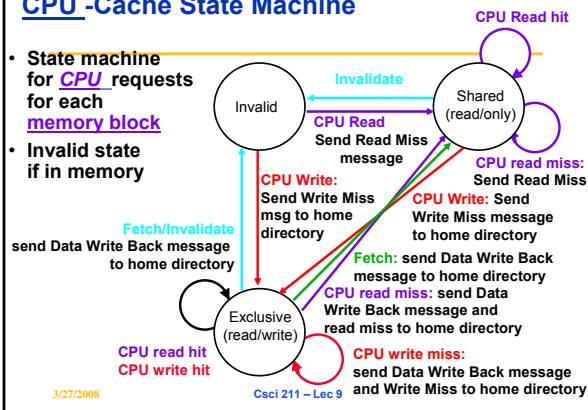
3/27/2008

Csci 211 – Lec 9

16

CPU -Cache State Machine

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



3/27/2008

Csci 211 – Lec 9

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

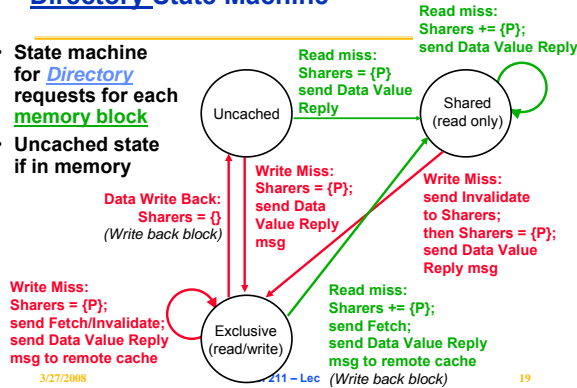
3/27/2008

Csci 211 – Lec 9

18

Directory State Machine

- State machine for **Directory** requests for each **memory block**
- Uncached state if in memory



3/27/2008

Csci 211 - Lec 9

19

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.

3/27/2008

Csci 211 - Lec 9

20

Example Directory Protocol

- Block is **Exclusive**: current value of the block is held in the cache of the processor identified by the set Sharers (the owner) ⇒ three possible directory requests:
 - Read miss:** owner processor is sent data fetch message, causing state of block in owner's cache to transition to Shared and causes owner to send data to directory, where it is written to memory & sent back to requesting processor. Identity of requesting processor is added to set Sharers, which still contains the identity of the processor that was the owner (since it still has a readable copy). State is shared.
 - Data write-back:** owner processor is replacing the block and hence must write it back, making memory copy up-to-date (the home directory essentially becomes the owner), the block is now Uncached, and the Sharer set is empty.
 - Write miss:** block has a new owner. A message is sent to old owner causing the cache to send the value of the block to the directory from which it is sent to the requesting processor, which becomes the new owner. Sharers is set to identity of new owner, and state of block is **made Exclusive**.

Csci 211 - Lec 9

21

Example

step	Processor 1		Processor 2		Interconnect		Directory		Memory					
	P1 State	Addr	P2 Value	State	Addr	Value	Action	Proc.	Addr	Value	Addr	State	(Procs)	Value
P1: Write 10 to A1	Excl	A1												
P1: Read A1														
P2: Read A1														
P2: Write 20 to A1														
P2: Write 40 to A2														

A1 and A2 map to the same cache block

3/27/2008

Csci 211 - Lec 9

22

Example

step	Processor 1		Processor 2		Interconnect		Directory		Memory					
	P1 State	Addr	P2 Value	State	Addr	Value	Action	Proc.	Addr	Value	Addr	State	(Procs)	Value
P1: Write 10 to A1	Excl	A1												
P1: Read A1														
P2: Read A1														
P2: Write 20 to A1														
P2: Write 40 to A2														

A1 and A2 map to the same cache block

3/27/2008

Csci 211 - Lec 9

23

Example

step	Processor 1		Processor 2		Interconnect		Directory		Memory					
	P1 State	Addr	P2 Value	State	Addr	Value	Action	Proc.	Addr	Value	Addr	State	(Procs)	Value
P1: Write 10 to A1	Excl	A1												
P1: Read A1														
P2: Read A1														
P2: Write 20 to A1														
P2: Write 40 to A2														

A1 and A2 map to the same cache block

3/27/2008

Csci 211 - Lec 9

24

Example

step	P1			P2			Interconnect			Directory			Memory	
	State	Addr	Value	State	Addr	Value	Action	Proc	Addr	Value	Addr	State	Proc	Value
P1: Write 10 to A1	Excl	A1	10				WtMs	P1	A1	0	A1	Ex	(P1)	
P1: Read A1	Excl	A1	10				DaRp	P1	A1	0				
P2: Read A1				Shar	A1	10	RdMs	P2	A1	10	A1	Ex	(P1)	10
P2: Write 20 to A1				Shar	A1	10	Ftch	P1	A1	10	A1	Shar	(P1,P2)	10
P2: Write 40 to A2							DaRp	P2	A1	10	A1	Shar	(P1,P2)	10

Write Back

A1 and A2 map to the same cache block

3/27/2008

Csci 211 – Lec 9

25

Example

step	P1			P2			Interconnect			Directory			Memory	
	State	Addr	Value	State	Addr	Value	Action	Proc	Addr	Value	Addr	State	Proc	Value
P1: Write 10 to A1	Excl	A1	10				WtMs	P1	A1	0	A1	Ex	(P1)	
P1: Read A1	Excl	A1	10				DaRp	P1	A1	0				
P2: Read A1				Shar	A1	10	RdMs	P2	A1	10	A1	Ex	(P1)	10
P2: Write 20 to A1				Shar	A1	10	Ftch	P1	A1	10	A1	Shar	(P1,P2)	10
P2: Write 40 to A2				Inv			DaRp	P2	A1	10	A1	Excl	(P2)	10

A1 and A2 map to the same cache block

3/27/2008

Csci 211 – Lec 9

26

Example

step	P1			P2			Interconnect			Directory			Memory	
	State	Addr	Value	State	Addr	Value	Action	Proc	Addr	Value	Addr	State	Proc	Value
P1: Write 10 to A1	Excl	A1	10				WtMs	P1	A1	0	A1	Ex	(P1)	
P1: Read A1	Excl	A1	10				DaRp	P1	A1	0				
P2: Read A1				Shar	A1	10	RdMs	P2	A1	10	A1	Ex	(P1)	10
P2: Write 20 to A1				Shar	A1	10	Ftch	P1	A1	10	A1	Shar	(P1,P2)	10
P2: Write 40 to A2				Inv			DaRp	P2	A1	10	A1	Excl	(P2)	10

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

3/27/2008

Csci 211 – Lec 9

27

Implementing a Directory

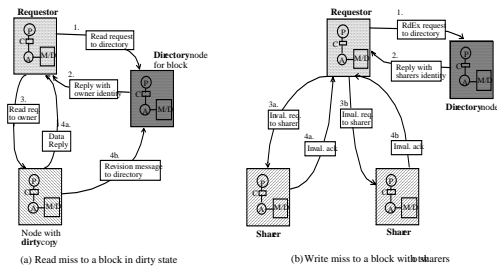
- We assume operations atomic, but they are not; reality is much harder; must avoid deadlock when run out of buffers in network (see Appendix E)
- Optimizations:
 - read miss or write miss in Exclusive: send data directly to requestor from owner vs. 1st to memory and then from memory to requestor

3/27/2008

Csci 211 – Lec 9

28

Basic Directory Transactions



3/27/2008

Csci 211 – Lec 9

29

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?

3/27/2008

Csci 211 – Lec 9

30

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

3/27/2008

Csci 211 – Lec 9

31

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

3/27/2008

Csci 211 – Lec 9

32

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
      ll     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:  ll     R2,0(R1)   ; load linked
      addi  R2,R2,#1   ; increment (OK if reg-reg)
      sc    R2,0(R1)   ; store conditional
      beqz  R2,try     ; branch store fails (R2 = 0)
```

3/27/2008

Csci 211 – Lec 9

33

User Level Synchronization—Operation Using this Primitive

- **Spin locks:** processor continuously tries to acquire, spinning around a loop trying to get the lock


```
lockit:  li     R2,#1      ;atomic exchange
          exch  R2,0(R1)   ;already locked?
          bnez  R2,lockit
```
- **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”):


```
try:    li     R2,#1      ;load var
lockit:  lw     R3,0(R1)   ;# 0 ⇒ not free ⇒ spin
          bnez  R3,lockit ;atomic exchange
          exch  R2,0(R1)   ;already locked?
          bnez  R2,try
```

3/27/2008

Csci 211 – Lec 9

34

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;          .....
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

3/27/2008

Csci 211 – Lec 9

35

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

3/27/2008

Csci 211 – Lec 9

36

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

3/27/2008

Csci 211 – Lec 9

37

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

3/27/2008

Csci 211 – Lec 9

38

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

3/27/2008

Csci 211 – Lec 9

39

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 ⇒ 2.5X
- 16 proc ⇒ need only 3.6X, or 25% linear speedup
- Even if need some more memory for MP, not linear

3/27/2008

Csci 211 – Lec 9

40

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 Alpha server 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

3/27/2008

Csci 211 – Lec 9

41

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)

3/27/2008

Csci 211 – Lec 9

42

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

3/27/2008

Csci 211 – Lec 9

43