Lecture 24: Multiprocessors

• Today’s topics:
  ▪ Directory-based cache coherence protocol
  ▪ Synchronization
  ▪ Consistency
  ▪ Writing parallel programs

• Reminder:
  ▪ Assignment 9 will be posted later today, due in a week
  ▪ Next Tuesday: recap lecture
  ▪ Next Thursday: guest lecture by Al Davis on future technologies
Snooping-Based Protocols

- Three states for a block: invalid, shared, modified
- A write is placed on the bus and sharers invalidate themselves
- The protocols are referred to as MSI, MESI, etc.
Example

- P1 reads X: not found in cache-1, request sent on bus, memory responds, X is placed in cache-1 in shared state
- P2 reads X: not found in cache-2, request sent on bus, everyone snoops this request, cache-1 does nothing because this is just a read request, memory responds, X is placed in cache-2 in shared state

• P1 writes X: cache-1 has data in shared state (shared only provides read perms), request sent on bus, cache-2 snoops and then invalidates its copy of X, cache-1 moves its state to modified
• P2 reads X: cache-2 has data in invalid state, request sent on bus, cache-1 snoops and realizes it has the only valid copy, so it downgrades itself to shared state and responds with data, X is placed in cache-2 in shared state
Cache Coherence Protocols

• Directory-based: A single location (directory) keeps track of the sharing status of a block of memory

• Snooping: Every cache block is accompanied by the sharing status of that block – all cache controllers monitor the shared bus so they can update the sharing status of the block, if necessary

➤ Write-invalidate: a processor gains exclusive access of a block before writing by invalidating all other copies
➤ Write-update: when a processor writes, it updates other shared copies of that block
Coherence in Distributed Memory Multiprocs

• Distributed memory systems are typically larger → bus-based snooping may not work well

• Option 1: software-based mechanisms – message-passing systems or software-controlled cache coherence

• Option 2: hardware-based mechanisms – directory-based cache coherence
Distributed Memory Multiprocessors
Directory-Based Cache Coherence

• The physical memory is distributed among all processors

• The directory is also distributed along with the corresponding memory

• The physical address is enough to determine the location of memory

• The (many) processing nodes are connected with a scalable interconnect (not a bus) – hence, messages are no longer broadcast, but routed from sender to receiver – since the processing nodes can no longer snoop, the directory keeps track of sharing state
Cache Block States

• What are the different states a block of memory can have within the directory?

• Note that we need information for each cache so that invalidate messages can be sent

• The directory now serves as the arbitrator: if multiple write attempts happen simultaneously, the directory determines the ordering
Directory-Based Example

A: Rd X
B: Rd X
C: Rd X
A: Wr X
B: Rd X
A: Wr X
B: Wr X
B: Wr Y
Directory Actions

• If block is in uncached state:
  ➢ Read miss: send data, make block shared
  ➢ Write miss: send data, make block exclusive

• If block is in shared state:
  ➢ Read miss: send data, add node to sharers list
  ➢ Write miss: send data, invalidate sharers, make excl

• If block is in exclusive state:
  ➢ Read miss: ask owner for data, write to memory, send data, make shared, add node to sharers list
  ➢ Data write back: write to memory, make uncached
  ➢ Write miss: ask owner for data, write to memory, send data, update identity of new owner, remain exclusive
Constructing Locks

- Applications have phases (consisting of many instructions) that must be executed atomically, without other parallel processes modifying the data.

- A lock surrounding the data/code ensures that only one program can be in a critical section at a time.

- The hardware must provide some basic primitives that allow us to construct locks with different properties.

```
<table>
<thead>
<tr>
<th>Bank balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1000</td>
</tr>
</tbody>
</table>

Parallel (unlocked) banking transactions

- Rd $1000
- Add $100
- Wr $1100
- Rd $1000
- Add $200
- Wr $1200
```
Synchronization

- The simplest hardware primitive that greatly facilitates synchronization implementations (locks, barriers, etc.) is an atomic read-modify-write

- Atomic exchange: swap contents of register and memory

- Special case of atomic exchange: test & set: transfer memory location into register and write 1 into memory (if memory has 0, lock is free)

- lock:  t&s  register, location
      bnz  register, lock
      CS  location, #0

      When multiple parallel threads execute this code, only one will be able to enter CS
Coherence Vs. Consistency

- Recall that coherence guarantees (i) write propagation (a write will eventually be seen by other processors), and (ii) write serialization (all processors see writes to the same location in the same order).

- The consistency model defines the ordering of writes and reads to different memory locations – the hardware guarantees a certain consistency model and the programmer attempts to write correct programs with those assumptions.
Consistency Example

• Consider a multiprocessor with bus-based snooping cache coherence and a write buffer between CPU and cache

<table>
<thead>
<tr>
<th>Initially A = B = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
</tr>
<tr>
<td>A ← 1</td>
</tr>
</tbody>
</table>
|                   ...
| if (B == 0)        |
| Crit.Section       |

| P2                  |
| B ← 1              |
|                   ...
| if (A == 0)        |
| Crit.Section       |

The programmer expected the above code to implement a lock – because of write buffering, both processors can enter the critical section

The consistency model lets the programmer know what assumptions they can make about the hardware’s reordering capabilities
Sequential Consistency

• A multiprocessor is sequentially consistent if the result of the execution is achieveable by maintaining program order within a processor and interleaving accesses by different processors in an arbitrary fashion.

• The multiprocessor in the previous example is not sequentially consistent.

• Can implement sequential consistency by requiring the following: program order, write serialization, everyone has seen an update before a value is read – very intuitive for the programmer, but extremely slow.
Shared-Memory Vs. Message-Passing

Shared-memory:
- Well-understood programming model
- Communication is implicit and hardware handles protection
- Hardware-controlled caching

Message-passing:
- No cache coherence → simpler hardware
- Explicit communication → easier for the programmer to restructure code
- Software-controlled caching
- Sender can initiate data transfer
Ocean Kernel

Procedure Solve(A)
begin
  diff = done = 0;
  while (!done) do
    diff = 0;
    for i ← 1 to n do
      for j ← 1 to n do
        temp = A[i,j];
        A[i,j] ← 0.2 * (A[i,j] + neighbors);
        diff += abs(A[i,j] − temp);
      end for
    end for
    if (diff < TOL) then done = 1;
  end while
end procedure
Shared Address Space Model

int n, nprocs;
float **A, diff;
LOCKDEC(diff_lock);
BARDEC(bar1);

main()
begin
  read(n); read(nprocs);
  A $\leftarrow$ G_MALLOC();
  initialize (A);
  CREATE (nprocs,Solve,A);
  WAIT_FOR_END (nprocs);
end main

procedure Solve(A)

  int i, j, pid, done=0;
  float temp, mydiff=0;
  int mymin = 1 + (pid * n/procs);
  int mymax = mymin + n/nprocs -1;
  while (!done) do
    mydiff = diff = 0;
    BARRIER(bar1,nprocs);
    for i $\leftarrow$ mymin to mymax
      for j $\leftarrow$ 1 to n do
        ...
      endfor
    endfor
    LOCK(diff_lock);
    diff += mydiff;
    UNLOCK(diff_lock);
    BARRIER (bar1, nprocs);
    if (diff < TOL) then done = 1;
    BARRIER (bar1, nprocs);
  endwhile
Message Passing Model

main()
read(n); read(nprocs);
CREATE (nprocs-1, Solve);
Solve();
WAIT_FOR_END (nprocs-1);

procedure Solve()
int i, j, pid, nn = n/nprocs, done=0;
float temp, tempdiff, mydiff = 0;
myA ← malloc(…)
initialize(myA);
while (!done) do
  mydiff = 0;
  if (pid != 0)
    SEND(&myA[1,0], n, pid-1, ROW);
  if (pid != nprocs-1)
    SEND(&myA[nn,0], n, pid+1, ROW);
  if (pid != 0)
    RECEIVE(&myA[0,0], n, pid-1, ROW);
  if (pid != nprocs-1)
    RECEIVE(&myA[nn+1,0], n, pid+1, ROW);

  for i ← 1 to nn do
    for j ← 1 to n do
      …
    endfor
  endfor

  if (pid != 0)
    SEND(mydiff, 1, 0, DIFF);
  RECEIVE(done, 1, 0, DONE);
  else
    for i ← 1 to nprocs-1 do
      RECEIVE(tempdiff, 1, *, DIFF);
      mydiff += tempdiff;
    endfor
    if (mydiff < TOL) done = 1;
    for i ← 1 to nprocs-1 do
      SEND(done, 1, i, DONE);
    endfor
  endif
endwhile
Title

• Bullet