Distributed Systems Synchronization

Marcus Völp

2008

Purpose of this Lecture

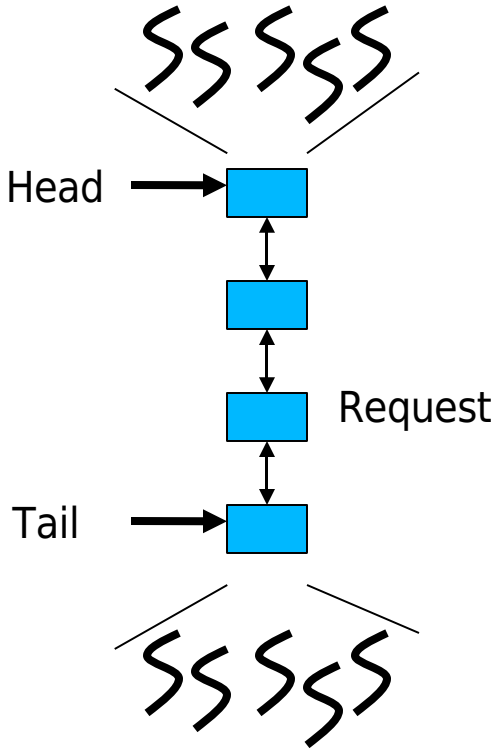
- Synchronization
- Locking
- Analysis / Comparison

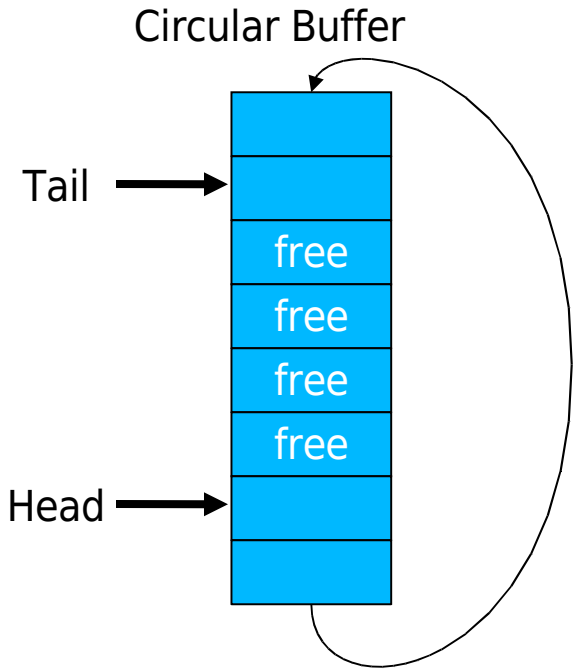
Overview

- Introduction
- Hardware Primitives
- Locking
  - Spin Lock (Test & Set Lock)
  - Test & Test & Set Lock
  - Ticket Locks
  - MCS Locks
- Lock-free Synchronization
- Special Issues
  - Timeouts
  - Reader Writer Locks
  - Lockholder Preemption
  - Monitor, Mwait
- Performance

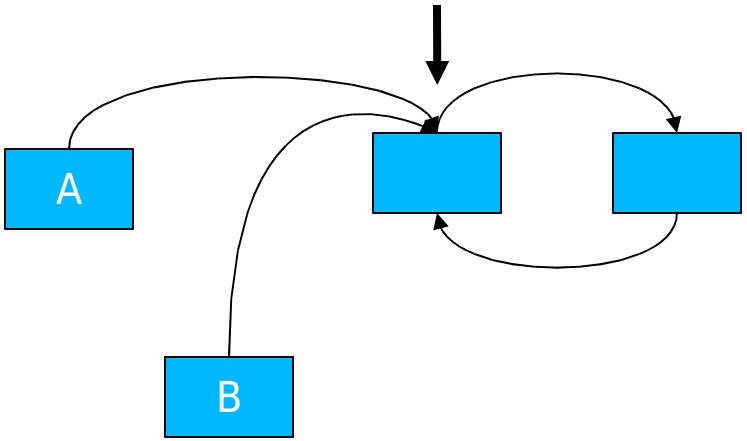
Introduction

- An example: Request Queue



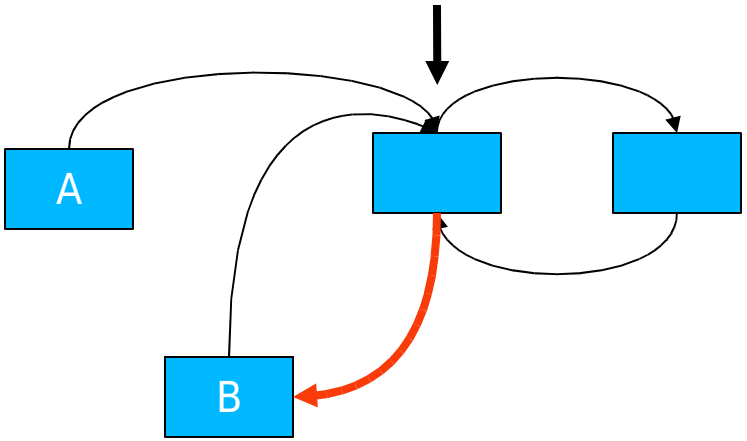


Introduction



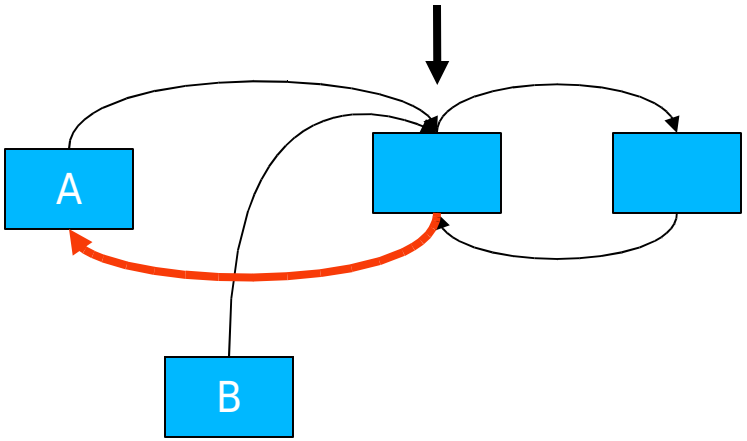
1) A,B create list elements
2) A,B set next pointer to head

Introduction



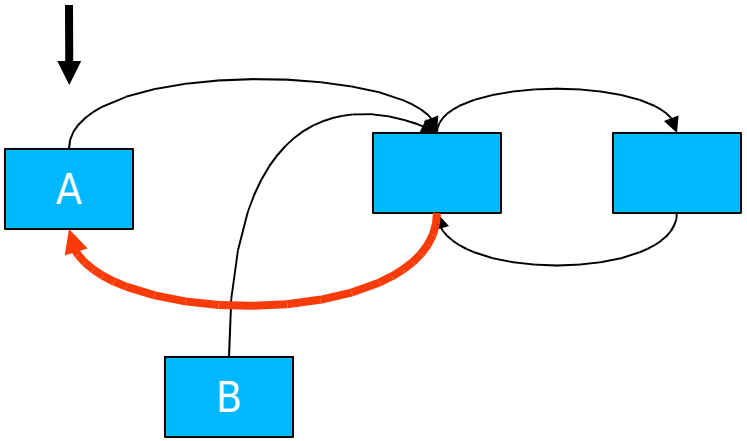
1) A,B create list elements
2) A,B set next pointer to head
3) B set prev pointer

Introduction



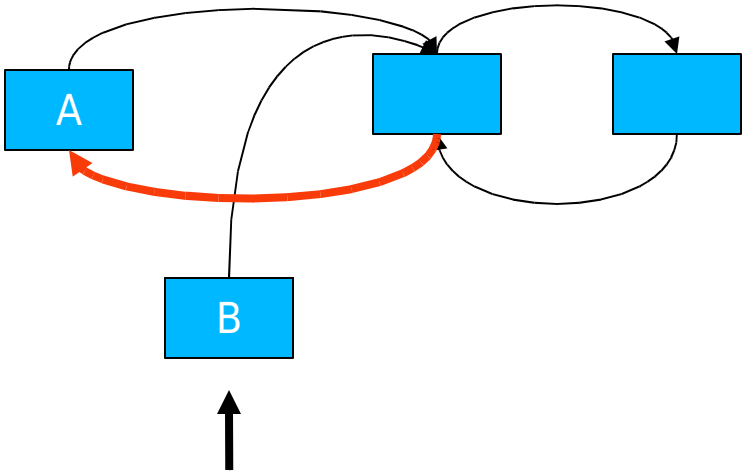
1) A,B create list elements
2) A,B set next pointer to head
3) B set prev pointer
4) A set prev pointer

Introduction



1) A,B create list elements
2) A,B set next pointer to head
3) B set prev pointer
4) A set prev pointer
5) A update head pointer

Introduction



1) A,B create list elements
2) A,B set next pointer to head
3) B set prev pointer
4) A set prev pointer
5) A update head pointer
6) B update head pointer

Introduction

- First Solution
  - Locks
    - List Lock
    - List Element Lock

lock_list;
insert_element;
unlock list;

Hardware Primitives

- How to make instructions atomic
  - Bus lock
    - Lock memory bus for duration of single instruction (e.g., lock add)
  - Observe Cache (ARM v6, Alpha, x86: monitor, mwait)
    - Load Linked: Load value and watch location
    - Store Conditional: Store value if no other store has accessed location
- Atomic operations
  - loads, stores
  - swap (XCHG) !! x86 implementation requires no bus lock !!
  - bit test and set (BTS)
    - **if** bit clear **then** set bit; return true **else** return false
  - compare and swap (CAS m, old, new)
    - **if** m == old **then** m := new ; return true **else** new := m, return false

Locking Algorithms

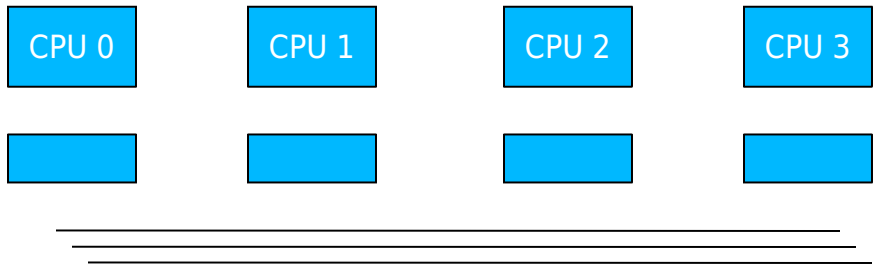
- Peterson's Algorithm
  - **Works only for 2 Threads**
  - atomic stores, atomic loads
  - sequential consistency (memory fences)

```
bool interested[2];
int blocked;

void entersection(int thread) {
        int other;                      /* number of other thread */
        other = 1 - thread;             /* the other thread: 1 for thread 0, 0 for thread 1 */
        interested[thread] = true;      /* show that you are interested */
        blocked=thread;
        while (blocked == thread && interested[other] == true){};/*wait*/
}

void leavesection(int thread) {
        interested[thread] = false;
}
```

Locking Algorithms

- Spin Lock (Test and Set Lock)
  - atomic swap

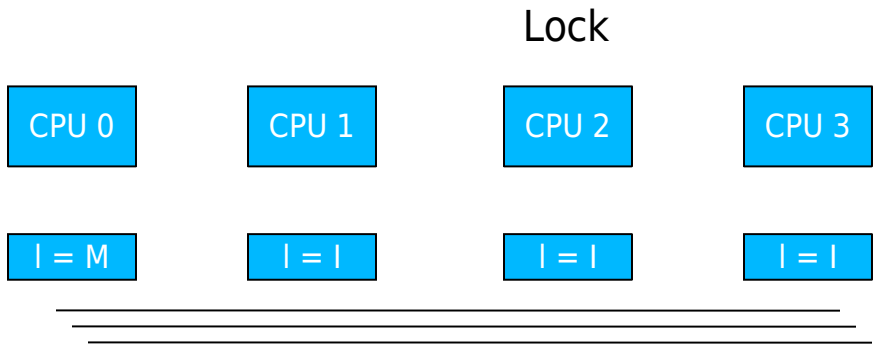
```
lock (lock_var l) {
  do {
     reg = 1;
     swap (l, reg)
  } while (reg == 1);
}

unlock (lock_var l) {
  l = 0;
}
```



Locking Algorithms

- Spin Lock (Test and Set Lock)
  - atomic swap

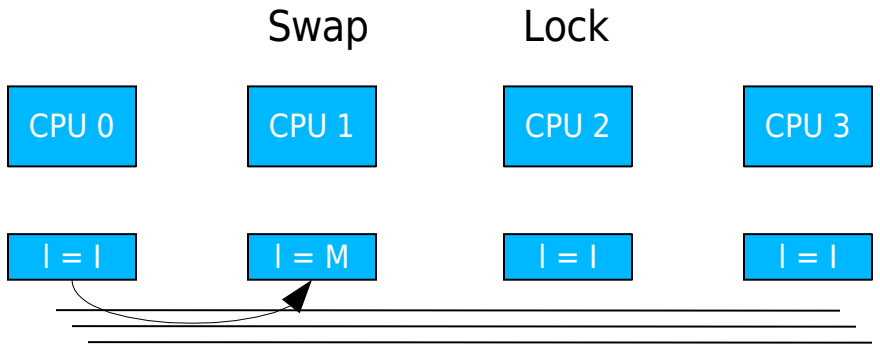
```
lock (lock_var l) {
  do {
     reg = 1;
     swap (l, reg)
  } while (reg == 1);
}

unlock (lock_var l) {
  l = 0;
}
```



Locking Algorithms

- Spin Lock (Test and Set Lock)
  - atomic swap

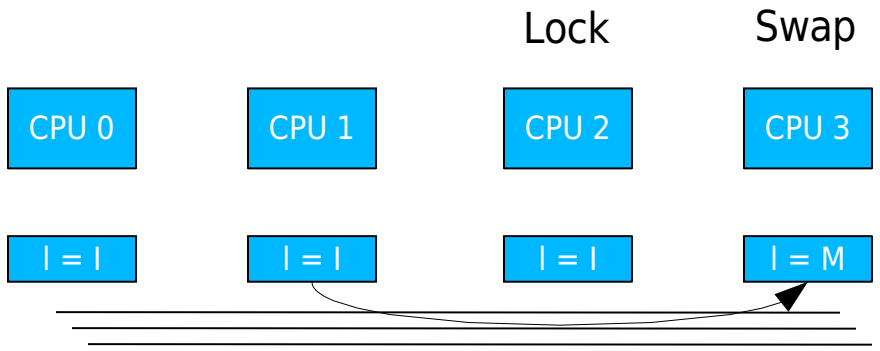
```
lock (lock_var l) {
  do {
     reg = 1;
     swap (l, reg)
  } while (reg == 1);
}

unlock (lock_var l) {
  l = 0;
}
```



Locking Algorithms

- Spin Lock (Test and Set Lock)
  - atomic swap

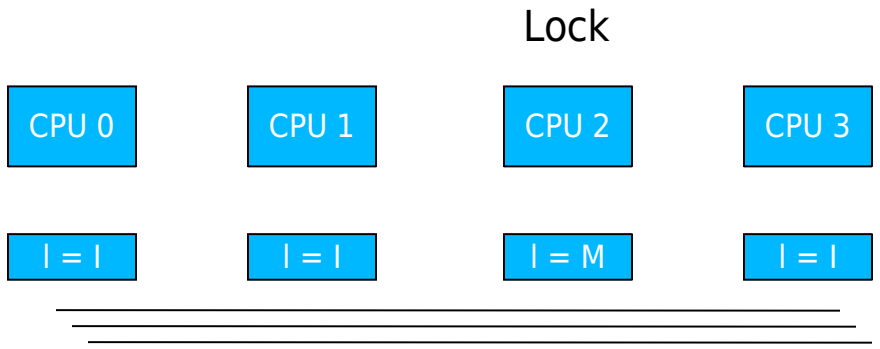
```
lock (lock_var l) {
  do {
     reg = 1;
     swap (l, reg)
  } while (reg == 1);
}

unlock (lock_var l) {
  l = 0;
}
```



Locking Algorithms

- Spin Lock (Test and Test and Set Lock)
  - atomic swap

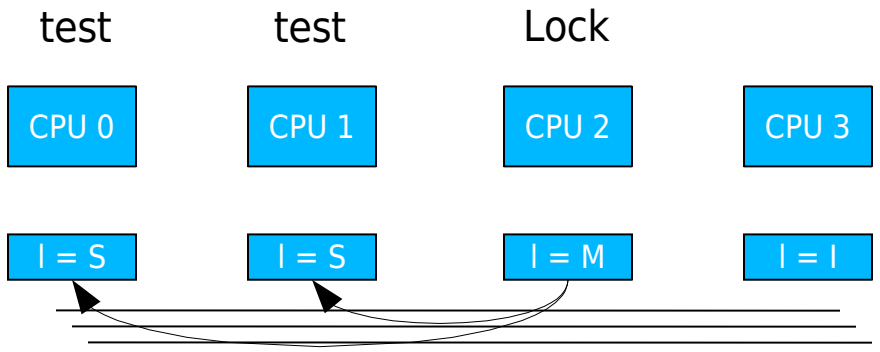
```
lock (lock_var l) {
  do {
     reg = 1;
     do { } while (l == 1);
     swap (l, reg)
  } while (reg == 1);
}

unlock (lock_var l) {
  l = 0;
}
```



Locking Algorithms

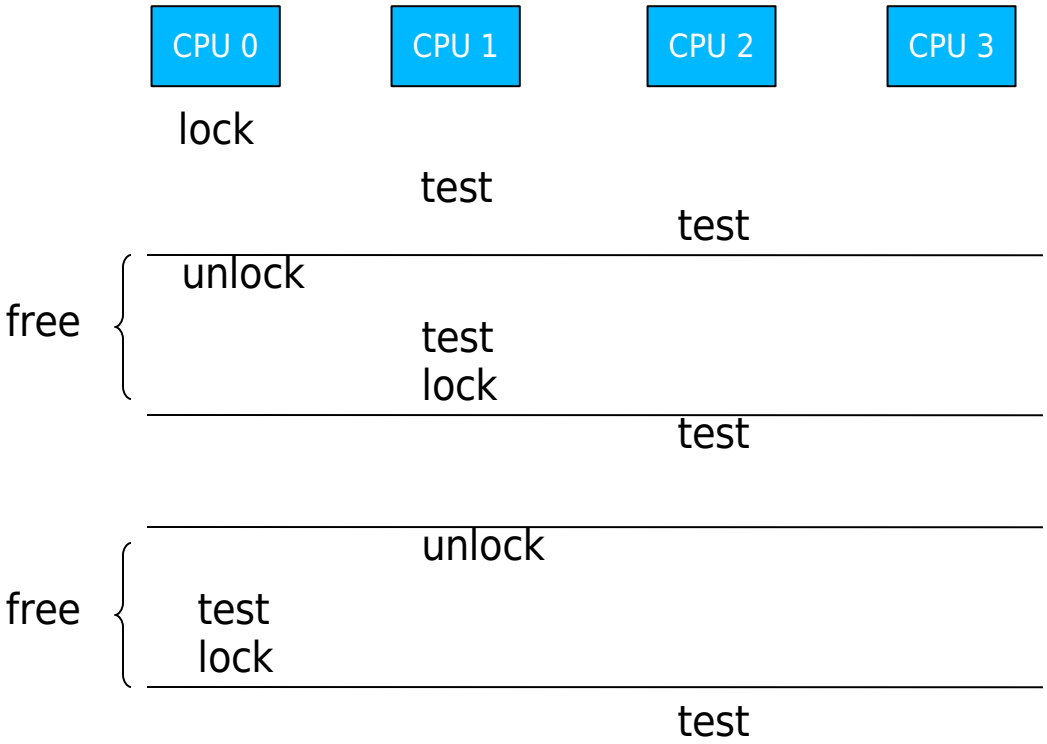
- Spin Lock (Test and Test and Set Lock)
  - atomic swap

```
lock (lock_var l) {
  do {
     reg = 1;
     do { } while (l == 1);
     swap (l, reg)
  } while (reg == 1);
}

unlock (lock_var l) {
  l = 0;
}
```



Locking Algorithms

- Fairness



Locking Algorithms

- Fairness: Ticket Lock
  - fetch and add (xadd)

```
lock_struct {
  next_ticket,
  current_ticket
}

ticket_lock (lock_struct l) {
  my_ticket = xadd (l.next_ticket, 1)
  do { } while (l.current_ticket != my_ticket);
}

unlock (lock_var l) {
  current_ticket ++;
}
```

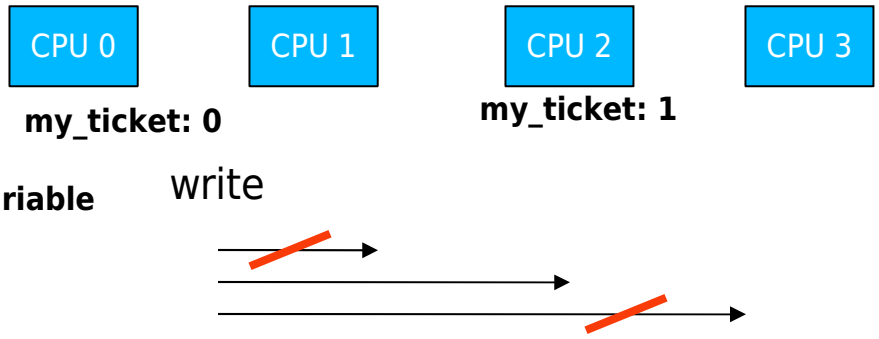
	current	next	
	0	0	
L.CPU0 [0]:	**0**		**1 => Lockholder = CPU0**
L.CPU1 [1]:	**0**		**2**
L.CPU2 [2]:	**0**		**3**
U.CPU0 [0]:	**1**		**3 => Lockholder = CPU1**
L.CPU3 [3]:	**1**		**4**
L.CPU0 [4]:	**1**		**5**
U.CPU1 [1]:	**2**		**5 => Lockholder = CPU 2**

Locking Algorithms

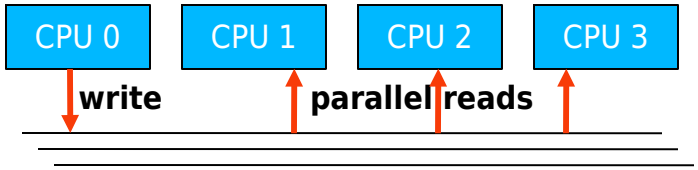
- Fairness: Ticket Lock
  - fetch and add (xadd)

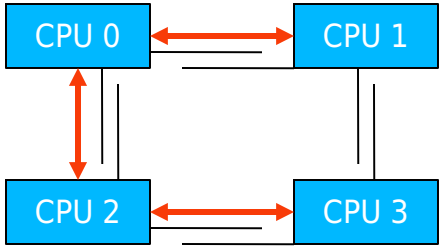
```
lock_struct {
  next_ticket,
  current_ticket
}

ticket_lock (lock_struct l) {
  my_ticket = xadd (l.next_ticket, 1)
  do { } while (l.current_ticket != my_ticket);
}

unlock (lock_var l) {
  current_ticket ++;
}
```

Spin on global variable

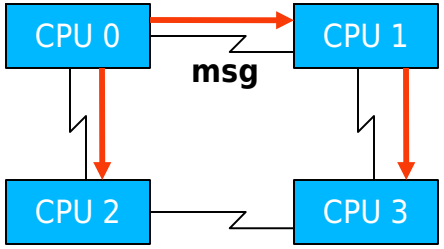


Local Spinning





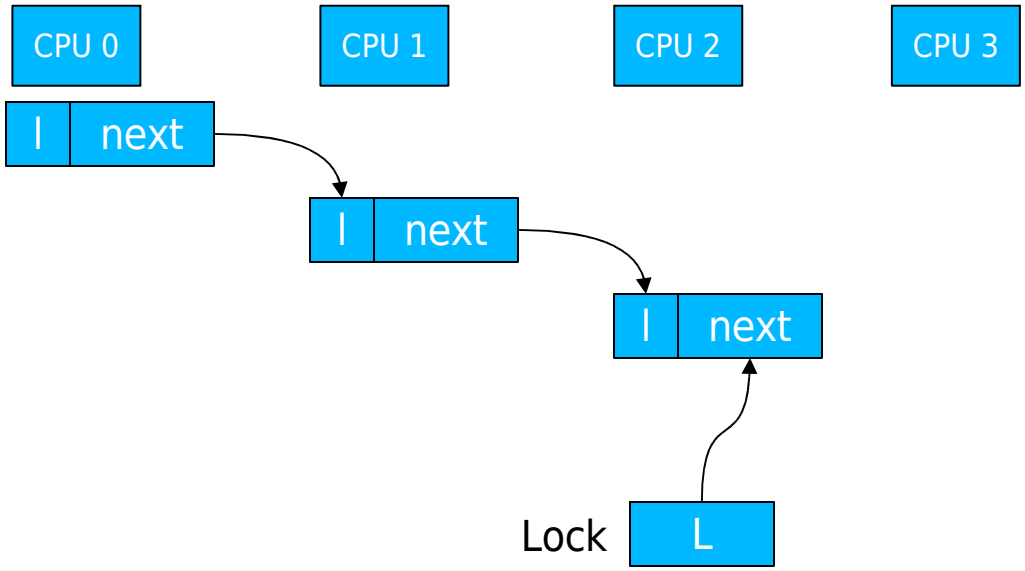
Need to propagate write on Bus 2-3 (or 1 – 3)



3 Network Messages

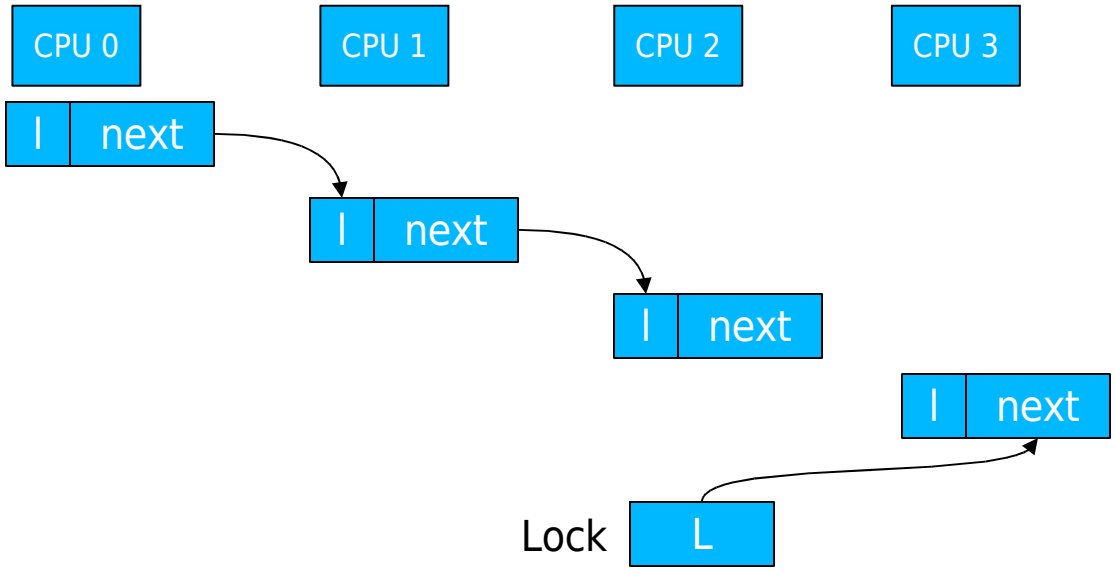
MCS-Lock (Mellor Crummey Scott)

- Fair Lock with Local Spinning



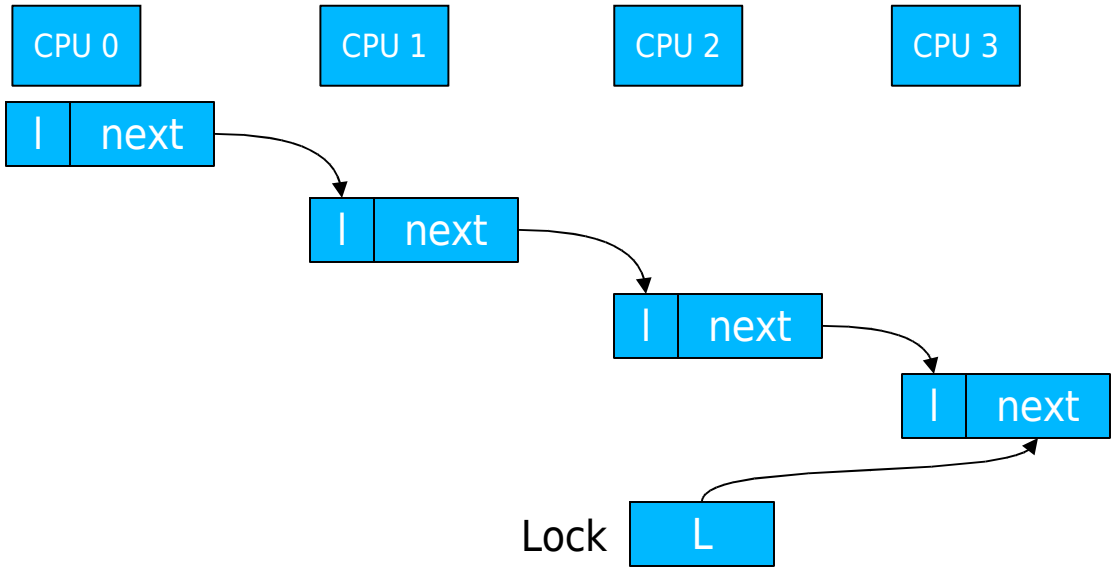
MCS-Lock (Mellor Crummey Scott)

- Fair Lock with Local Spinning



MCS-Lock (Mellor Crummey Scott)

- Fair Lock with Local Spinning

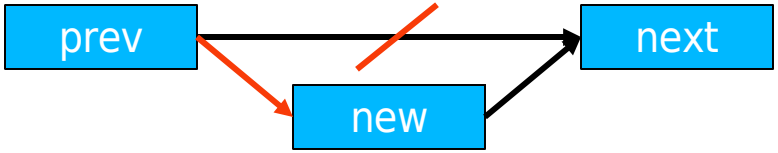


MCS Locks

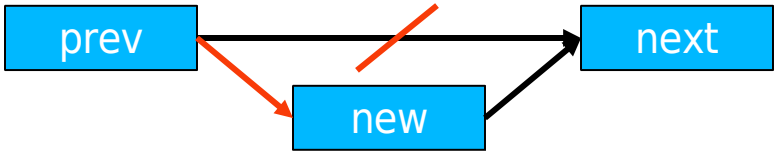
- Fair, local spinning
  - atomic compare exchange: cmpxchg (L == Old, New)

```
lock (L, lock_element I) {
  I.next = null; I.lock = false;
  prev = xchg (L, I);
  if (prev != null) {
    prev.next = I;
    do { } while (I.lock == false);
  }
}

unlock (L, I) {
 if (I.next == Null) {
   if (cmpxchg (L == I, Null)) return; // no waiting cpu
   do { } while (I.next == Null); // spin until the following process updates the next pointer
 }
 I.next.lock = true;
}
```

Lock-free synchronization




```
insert(new, prev) {
  retry:
    new.next = next;
    if (not CAS(prev.next == next, new)) goto retry;
}
```




```
insert(new, prev) {
  retry:
    new.next = next;
    new.prev = prev;
    if (not DCAS(prev.next == next && next.prev ==prev, prev.next = new, next.prev = new))
      goto retry;
}
```

Lock-free Synchronization

- Load Linked, Store Conditional

```
insert (prev, new) {
  retry:
   ll (prev.next);
   new.next = prev.next;
   if (not stc (prev.next, new)) goto retry;
}
```

Overview

- Introduction
- Hardware Primitives
- Locking
  - Spin Lock (Test & Set Lock)
  - Test & Test & Set Lock
  - Ticket Locks
  - MCS Locks
- Lock-free Synchronization
- Special Issues
  - Timeouts
  - Reader Writer Locks
  - Lockholder Preemption
  - Monitor, Mwait
- Performance

Special Issues

- Timeouts
  - No longer apply for lock after timeout
  - Dequeue from MCS queue
- Reader Writer Locks
  - Lock differentiates two types of lockers: reader, writer
    - Multiple readers may hold lock at same time
    - Writers hold lock exclusively
  - Fairness
  - Improve reader latency by allowing readers to overtake writers (unfair lock)

Special Issues

- Fair Ticket Reader-Writer Lock
  - combine read, write ticket in single word

```
lock read (next, current) {
  my_ticket = xadd (next, 1);
  do {} while (current.write != my_ticket.write);
}

lock write (next, current) {
   my_ticket = xadd (next.write, 1);
   do {} while (current != my_ticket);
}


unlock_read () {
   xadd (current.read, 1);
}

unlock write () {
   current.write ++;
}
```

current	next	R0	R1	W2	R3
0 0	0 0	0 0			
	0 1		0 1		
	0 2			0 2	
	1 2				1 2

Special Issues

- Lockholder preemption
  - Spin-time of other CPUs increases by preemption time of lockholder
  - E.g., no packets can be sent when OS network code is preempted while holding xmit queue lock

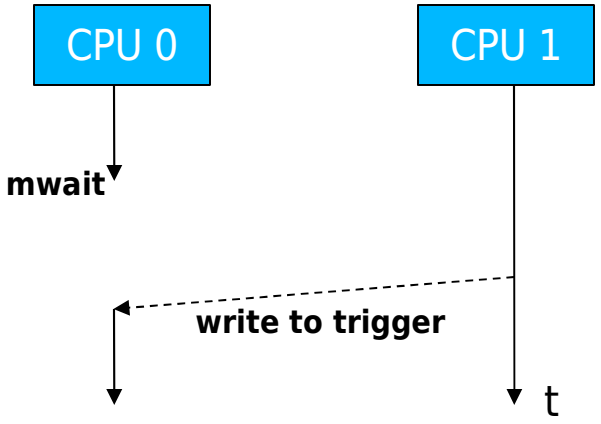
```
spin_lock(lock_var) {
  pushf; // store whether interrupts were already closed
  do {
     popf;
     reg = 1;
     do {} while (lock_var == 1);
     pushf;
     cli;
     swap(lock_var, reg);
  } while (reg == 1);
}
```

```
spin_unlock(lock_var) {
        lock_var = 0;
  popf;
}
```

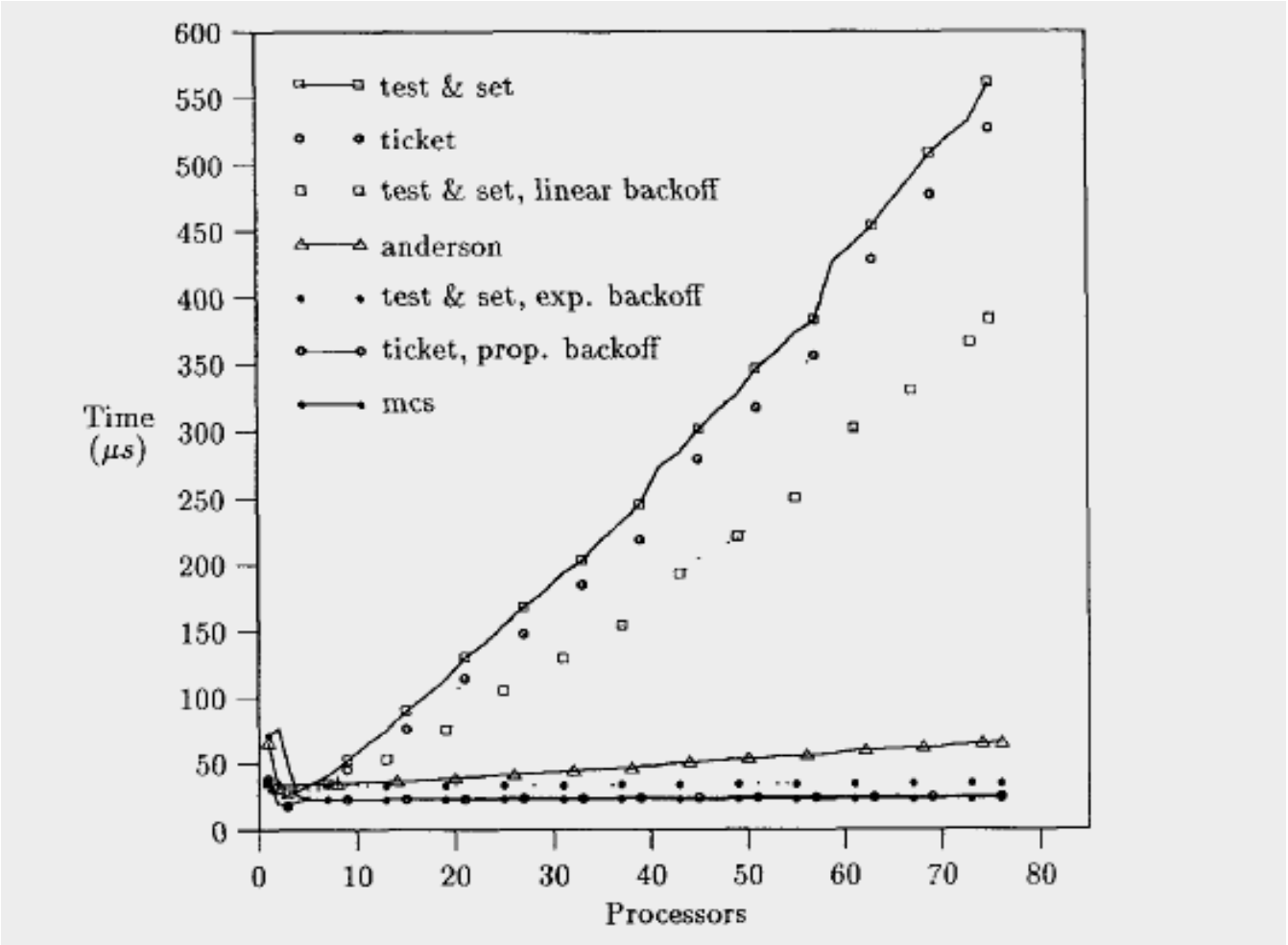
Special Issues

- Monitor, Mwait
  - Stop CPU / HT while waiting for lock (signal)
    - Saves power
    - Frees up processor resources (HT)
  - Monitor: watch cacheline
  - Mwait: stop CPU / HT until:
    - cacheline has been written, or
    - interrupt occurs

```
while (trigger[0] != value) {
  monitor (&trigger[0])
  if (trigger[0] != value) {
    mwait
  }
}
```



Performance



Source: Mellor Crummey, Scott : Algorithms for Scalable Synchronization on Shared Memory Multiprocessors

References

- *Scheduler-Conscious Synchronization*
  LEONIDAS I. KONTOTHANASSIS, ROBERT W. WISNIEWSKI, MICHAEL L. SCOTT
- *Scalable Reader- Writer Synchronization for Shared-Memory Multiprocessors*
  John M. Mellor-Crummey, Michael L. Scottt
- *Algorithms for Scalable Synchronization on Shared-Memory Multiprocessors*
  JOHN M. MELLOR-CRUMMEY, MICHAEL L.
- *Concurrent Update on Multiprogrammed Shared Memory Multiprocessors*
  Maged M. Michael, Michael L. Scott
- *Scalable Queue-Based Spin Locks with Timeout*
  Michael L. Scott and William N. Scherer III
- Reactive Synchronization Algorithms for Multiprocessors
  B. Lim, A. Agarwal
- Lock Free Data Structures
  John D. Valois (PhD Thesis)