Common Multiprocessor Spin Lock

```c
void mp_spinlock (volatile lock t *l) {
    cli(); // prevent preemption
    while (test and set(l)) ; // lock
}

void mp_unlock (volatile lock t *l) {
    *l = 0;
    sti();
}
```

Only good for short critical sections
Does not scale for large number of processors
Relies on bus-arbitrator for fairness
Not appropriate for user-level
Used in practice in small SMP systems
Need a more systematic analysis

Compares Simple Spinlocks

Test and Set

```c
void lock (volatile lock_t *l) {
  while (test_and_set(l)) ;
}
```

Test and Test and Set

```c
void lock (volatile lock_t *l) {
  while (*l == BUSY || test_and_set(l)) ;
}
```
test_and_test_and_set LOCK

Avoid bus traffic contention caused by test_and_set until it is likely to succeed
Normal read spins in cache
Can starve in pathological cases
Benchmark

for i = 1 .. 1,000,000 {
    lock(l)
    crit_section()
    unlock()
    compute()
}

Compute chosen from uniform random distribution of mean 5 times critical section

Measure elapsed time on Sequent Symmetry (20 CPU 30386, coherent write-back invalidate caches)
Results

Test and set performs poorly once there is enough CPUs to cause contention for lock
  • Expected

Test and Test and Set performs better
  • Performance less than expected
  • Still significant contention on lock when CPUs notice release and all attempt acquisition

Critical section performance degenerates
  • Critical section requires bus traffic to modify shared structure
  • Lock holder competes with CPU that missed as they test and set
    – lock holder is slower
  • Slower lock holder results in more contention
Idea

Can inserting delays reduce bus traffic and improve performance

Explore 2 dimensions

• Location of delay
  – Insert a delay after release prior to attempting acquire
  – Insert a delay after each memory reference

• Delay is static or dynamic
  – Static – assign delay “slots” to processors
    » Issue: delay tuned for expected contention level
  – Dynamic – use a back-off scheme to estimate contention
    » Similar to ethernet
    » Degrades to static case in worst case.
Examining Inserting Delays

### TABLE III
**DELAY AFTER SPINNER NOTICES RELEASED LOCK**

| Lock | while (lock = BUSY or TestAndSet (Lock) = BUSY) begin while (lock = BUSY) ; Delay () ; end ; |

### TABLE IV
**DELAY BETWEEN EACH REFERENCE**

| Lock | while (lock = BUSY or TestAndSet (lock) = BUSY) Delay () ; |
Queue Based Locking

Each processor inserts itself into a waiting queue

- It waits for the lock to free by spinning on its own separate cache line
- Lock holder frees the lock by “freeing” the next processors cache line.
Results
Results

Static backoff has higher overhead when backoff is inappropriate

Dynamic backoff has higher overheads when static delay is appropriate

- as collisions are still required to tune the backoff time

Queue is better when contention occurs, but has higher overhead when it does not.

- Issue: Preemption of queued CPU blocks rest of queue (worse than simple spin locks)
MCS Locks

Each CPU enqueues its own private lock variable into a queue and spins on it

- No contention

On lock release, the releaser unlocks the next lock in the queue

- Only have bus contention on actual unlock
- No livelock (order of lock acquisitions defined by the list)
MCS Lock

Requires

• compare_and_swap()
• exchange()
  – Also called fetch_and_store()
type qnode = record
    next : \qnode
    locked : Boolean
end

type lock = ^qnode

// parameter I, below, points to a qnode record allocated
// (in an enclosing scope) in shared memory locally-accessible
// to the invoking processor

procedure acquire_lock (L : ^lock, I : ^qnode)
    I->next := nil
    predecessor : ^qnode := fetch_and_store (L, I)
    if predecessor != nil       // queue was non-empty
        I->locked := true
    predecessor->next := I
    repeat while I->locked       // spin

procedure release_lock (L : ^lock, I: ^qnode)
    if I->next = nil            // no known successor
        if compare_and_swap (L, I, nil)
            return
        // compare_and_swap returns true iff it swapped
        repeat while I->next = nil       // spin
    I->next->locked := false
Sample MCS code for ARM MPCore

```c
void mcs_acquire(mcs_lock *L, mcs_qnode_ptr I)
{
    I->next = NULL;
    MEM_BARRIER;
    mcs_qnode_ptr pred = (mcs_qnode*) SWAP_PTR( L, (void *)I);
    if (pred == NULL)
    {
        /* lock was free */

        MEM_BARRIER;
        return;
    }
    I->waiting = 1; // word on which to spin
    MEM_BARRIER;
    pred->next = I; // make pred point to me
}
```
Selected Benchmark

Compared
• test and test and set
• Anderson’s array based queue
• test and set with exponential back-off
• MCS
Fig. 17. Performance of spin locks on the Symmetry (empty critical section).
Confirmed Trade-off

Queue locks scale well but have higher overhead
Spin Locks have low overhead but don’t scale well
What do we use?
Idea

Can we dynamically switch locking methods to suit the current contention level???
Issues

How do we determine which protocol to use?
• Must not add significant cost

How do we correctly and efficiently switch protocols?

How do we determine when to switch protocols?
Protocol Selection

Keep a “hint”

Ensure both TTS and MCS lock a never free at the same time

- Only correct selection will get the lock
- Choosing the wrong lock with result in retry which can get it right next time
- Assumption: Lock mode changes infrequently
  - hint cached read-only
  - infrequent protocol mismatch retries
Changing Protocol

Only lock holder can switch to avoid race conditions

• It chooses which lock to free, TTS or MCS.
When to change protocol

Use threshold scheme

- Repeated acquisition failures will switch mode to queue
- Repeated immediate acquisition will switch mode to TTS
Results
The multicore evolution and operating systems

Frans Kaashoek

Joint work with: Silas Boyd-Wickizer, Austin T. Clements, Yandong Mao, Aleksey Pesterev, Robert Morris, and Nickolai Zeldovich

MIT
Non-scalable locks are dangerous.
How well does Linux scale?

- **Experiment:**
  - Linux 2.6.35-rc5 (relatively old, but problems are representative of issues in recent kernels too)
  - Select a few inherent parallel system applications
  - Measure throughput on different # of cores
  - Use tmpfs to avoid disk bottlenecks

- **Insight 1:** Short critical sections can lead to sharp performance collapse
Off-the-shelf 48-core server (AMD)

- Cache-coherent and non-uniform access
- An approximation of a future 48-core chip
Poor scaling on stock Linux kernel

Y-axis: (throughput with 48 cores) / (throughput with one core)
Exim on stock Linux: collapse

The graph shows the throughput (in messages/second) of Exim with increasing cores. As the number of cores increases, the throughput also increases until it reaches a peak around 40 cores, after which it decreases sharply, indicating a collapse or bottleneck in performance.
Exim on stock Linux: collapse
Exim on stock Linux: collapse

![Graph showing throughput and kernel time against cores](image)
Oprofile shows an obvious problem

<table>
<thead>
<tr>
<th>Samples</th>
<th>%</th>
<th>App Name</th>
<th>Symbol Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2616</td>
<td>7.3522</td>
<td>vmlinux</td>
<td>radix_tree_lookup_slot</td>
</tr>
<tr>
<td>2329</td>
<td>6.5456</td>
<td>vmlinux</td>
<td>unmap_vmas</td>
</tr>
<tr>
<td>2197</td>
<td>6.1746</td>
<td>vmlinux</td>
<td>filemap_fault</td>
</tr>
<tr>
<td>1488</td>
<td>4.1820</td>
<td>vmlinux</td>
<td>__do_fault</td>
</tr>
<tr>
<td>1348</td>
<td>3.7885</td>
<td>vmlinux</td>
<td>copy_page_c</td>
</tr>
<tr>
<td>1182</td>
<td>3.3220</td>
<td>vmlinux</td>
<td>unlock_page</td>
</tr>
<tr>
<td>966</td>
<td>2.7149</td>
<td>vmlinux</td>
<td>page_fault</td>
</tr>
</tbody>
</table>

40 cores: 10000 msg/sec

<table>
<thead>
<tr>
<th>Samples</th>
<th>%</th>
<th>App Name</th>
<th>Symbol Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>13515</td>
<td>34.8657</td>
<td>vmlinux</td>
<td>lookup_mnt</td>
</tr>
<tr>
<td>2002</td>
<td>5.1647</td>
<td>vmlinux</td>
<td>radix_tree_lookup_slot</td>
</tr>
<tr>
<td>1661</td>
<td>4.2850</td>
<td>vmlinux</td>
<td>filemapFault</td>
</tr>
<tr>
<td>1497</td>
<td>3.8619</td>
<td>vmlinux</td>
<td>unmap_vmas</td>
</tr>
<tr>
<td>1026</td>
<td>2.6469</td>
<td>vmlinux</td>
<td>__do_fault</td>
</tr>
<tr>
<td>914</td>
<td>2.3579</td>
<td>vmlinux</td>
<td>atomic_dec</td>
</tr>
<tr>
<td>896</td>
<td>2.3115</td>
<td>vmlinux</td>
<td>unlock_page</td>
</tr>
</tbody>
</table>

48 cores: 4000 msg/sec
Oprofile shows an obvious problem

<table>
<thead>
<tr>
<th>samples</th>
<th>%</th>
<th>app name</th>
<th>symbol name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2616</td>
<td>7.3522</td>
<td>vmlinux</td>
<td>radix_tree_lookup_slot</td>
</tr>
<tr>
<td>2329</td>
<td>6.5456</td>
<td>vmlinux</td>
<td>unmap_vmas</td>
</tr>
<tr>
<td>2197</td>
<td>6.1746</td>
<td>vmlinux</td>
<td>filemap_fault</td>
</tr>
<tr>
<td>1488</td>
<td>4.1820</td>
<td>vmlinux</td>
<td>__do_fault</td>
</tr>
<tr>
<td>1348</td>
<td>3.7885</td>
<td>vmlinux</td>
<td>copy_page_c</td>
</tr>
<tr>
<td>1182</td>
<td>3.3220</td>
<td>vmlinux</td>
<td>unlock_page</td>
</tr>
<tr>
<td>966</td>
<td>2.7149</td>
<td>vmlinux</td>
<td>page_fault</td>
</tr>
</tbody>
</table>

40 cores: 10000 msg/sec

<table>
<thead>
<tr>
<th>samples</th>
<th>%</th>
<th>app name</th>
<th>symbol name</th>
</tr>
</thead>
<tbody>
<tr>
<td>13515</td>
<td>34.8657</td>
<td>vmlinux</td>
<td>lookup_mnt</td>
</tr>
<tr>
<td>2002</td>
<td>5.1647</td>
<td>vmlinux</td>
<td>radix_tree_lookup_slot</td>
</tr>
<tr>
<td>1661</td>
<td>4.2850</td>
<td>vmlinux</td>
<td>filemap_fault</td>
</tr>
<tr>
<td>1497</td>
<td>3.8619</td>
<td>vmlinux</td>
<td>unmap_vmas</td>
</tr>
<tr>
<td>1026</td>
<td>2.6469</td>
<td>vmlinux</td>
<td>__do_fault</td>
</tr>
<tr>
<td>914</td>
<td>2.3579</td>
<td>vmlinux</td>
<td>atomic_dec</td>
</tr>
<tr>
<td>896</td>
<td>2.3115</td>
<td>vmlinux</td>
<td>unlock_page</td>
</tr>
</tbody>
</table>

48 cores: 4000 msg/sec
Oprofile shows an obvious problem

<table>
<thead>
<tr>
<th>samples</th>
<th>%</th>
<th>app name</th>
<th>symbol name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2616</td>
<td>7.3522</td>
<td>vmlinux</td>
<td>radix_tree_lookup_slot</td>
</tr>
<tr>
<td>2329</td>
<td>6.5456</td>
<td>vmlinux</td>
<td>unmap_vmas</td>
</tr>
<tr>
<td>2197</td>
<td>6.1746</td>
<td>vmlinux</td>
<td>filemap_fault</td>
</tr>
<tr>
<td>1488</td>
<td>4.1820</td>
<td>vmlinux</td>
<td>__do_fault</td>
</tr>
<tr>
<td>1348</td>
<td>3.7885</td>
<td>vmlinux</td>
<td>copy_page_c</td>
</tr>
<tr>
<td>1182</td>
<td>3.3220</td>
<td>vmlinux</td>
<td>unlock_page</td>
</tr>
<tr>
<td>966</td>
<td>2.7149</td>
<td>vmlinux</td>
<td>page_fault</td>
</tr>
</tbody>
</table>

40 cores: 10000 msg/sec

<table>
<thead>
<tr>
<th>samples</th>
<th>%</th>
<th>app name</th>
<th>symbol name</th>
</tr>
</thead>
<tbody>
<tr>
<td>13515</td>
<td>34.8657</td>
<td>vmlinux</td>
<td>lookup_mnt</td>
</tr>
<tr>
<td>2002</td>
<td>5.1647</td>
<td>vmlinux</td>
<td>radix_tree_lookup_slot</td>
</tr>
<tr>
<td>1661</td>
<td>4.2850</td>
<td>vmlinux</td>
<td>filemap_fault</td>
</tr>
<tr>
<td>1497</td>
<td>3.8619</td>
<td>vmlinux</td>
<td>unmap_vmas</td>
</tr>
<tr>
<td>1026</td>
<td>2.6469</td>
<td>vmlinux</td>
<td>__do_fault</td>
</tr>
<tr>
<td>914</td>
<td>2.3579</td>
<td>vmlinux</td>
<td>atomic_dec</td>
</tr>
<tr>
<td>896</td>
<td>2.3115</td>
<td>vmlinux</td>
<td>unlock_page</td>
</tr>
</tbody>
</table>

48 cores: 4000 msg/sec
Bottleneck: reading mount table

- Delivering an email calls `sys_open`
- `sys_open` calls

```c
struct vfsmount *lookup_mnt(struct path *path)
{
    struct vfsmount *mnt;
    spin_lock(&vfsmount_lock);
    mnt = hash_get(mnts, path);
    spin_unlock(&vfsmount_lock);
    return mnt;
}
```
Bottleneck: reading mount table

- sys_open calls:

```c
struct vfsmount *lookup_mnt(struct path *path)
{
    struct vfsmount *mnt;
    spin_lock(&vfsmount_lock);
    mnt = hash_get(mnts, path);
    spin_unlock(&vfsmount_lock);
    return mnt;
}
```
Bottleneck: reading mount table

- **sys_open calls:**

```c
struct vfsmount *lookup_mnt(struct path *path)
{
    struct vfsmount *mnt;
    spin_lock(&vfsmount_lock);
    mnt = hash_get(mnts, path);
    spin_unlock(&vfsmount_lock);
    return mnt;
}
```

Serial section is short. Why does it cause a scalability bottleneck?
What causes the sharp performance collapse?

- Linux uses ticket spin locks, which are non-scalable
  - So we should expect collapse [Anderson 90]

- But why so sudden, and so sharp, for a short section?
  - Is spin lock/unlock implemented incorrectly?
  - Is hardware cache-coherence protocol at fault?
Scalability collapse caused by non-scalable locks [Anderson 90]

```c
void spin_lock(spinlock_t *lock)
{
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket) ; /* Spin */
}

void spin_unlock(spinlock_t *lock)
{
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
```
Scalability collapse caused by non-scalable locks [Anderson 90]

```c
void spin_lock(spinlock_t *lock) {
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket) ;  /* Spin */
}

void spin_unlock(spinlock_t *lock) {
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
```
Scalability collapse caused by non-scalable locks [Anderson 90]

```c
void spin_lock(spinlock_t *lock) {
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket) ; /* Spin */
}

void spin_unlock(spinlock_t *lock) {
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
```
Scalability collapse caused by non-scalable locks [Anderson 90]

```
void spin_lock(spinlock_t *lock)
{
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket)
        ;  /* Spin */
}

void spin_unlock(spinlock_t *lock)
{
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
```
Scalability collapse caused by non-scalable locks [Anderson 90]

```c
void spin_lock(spinlock_t *lock)
{
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket)
        ; /* Spin */
}

void spin_unlock(spinlock_t *lock)
{
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
```
Scalability collapse caused by non-scalable locks [Anderson 90]

```c
void spin_lock(spinlock_t *lock) {
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket) ; /* Spin */
}

void spin_unlock(spinlock_t *lock) {
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
```

500 cycles
Scalability collapse caused by non-scalable locks [Anderson 90]

void spin_lock(spinlock_t *lock)
{
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket)
        ; /* Spin */
}

void spin_unlock(spinlock_t *lock)
{
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
Scalability collapse caused by non-scalable locks [Anderson 90]

```c
void spin_lock(spinlock_t *lock) {
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket) ;  /* Spin */
}

void spin_unlock(spinlock_t *lock) {
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
```
Scalability collapse caused by non-scalable locks [Anderson 90]

```c
void spin_lock(spinlock_t *lock) {
    t = atomic_inc(lock->next_ticket);
    while (t != lock->current_ticket) ; /* Spin */
}

void spin_unlock(spinlock_t *lock) {
    lock->current_ticket++;
}

struct spinlock_t {
    int current_ticket;
    int next_ticket;
}
```

Previous lock holder notifies next lock holder after sending out N/2 replies
Why collapse with short sections?

- Arrival rate is proportional to # non-waiting cores
- Service time is proportional to # cores waiting \((k)\)
  - As \(k\) increases, waiting time goes up
  - As waiting time goes up, \(k\) increases
- System gets stuck in states with many waiting cores
Short sections result in collapse

- Experiment: 2% of time spent in critical section
- Critical sections become “longer” with more cores
- Lesson: non-scalable locks fine for long sections
Avoiding lock collapse

- Unscalable locks are fine for long sections
- Unscalable locks collapse for short sections
  - Sudden sharp collapse due to “snowball” effect
- Scalable locks avoid collapse altogether
  - But requires interface change
Scalable lock scalability

- It doesn't matter much which one
- But all slower in terms of latency
Avoiding lock collapse is not enough to scale

- “Scalable” locks don't make the kernel scalable
  - Main benefit is avoiding collapse: total throughput will not be lower with more cores
  - But, usually want throughput to keep increasing with more cores