

Foundations: Concurrency Concerns Synchronization Basics

Chris Rossbach

CS378H

Multithreaded programming

Today

- Questions?
- Administrivia
 - You've started Lab 1 right?
- Foundations
 - Parallelism
 - Basic Synchronization
 - Threads/Processes/Fibers, Oh my!
 - Cache coherence (maybe)
- Acknowledgments: some materials in this lecture borrowed from
 - Emmett Witchel (who borrowed them from: Kathryn McKinley, Ron Rockhold, Tom Anderson, John Carter, Mike Dahlin, Jim Kurose, Hank Levy, Harrick Vin, Thomas Narten, and Emery Berger)
 - Mark Silberstein (who borrowed them from: Blaise Barney, Kunle Olukoton, Gupta)
 - Andy Tannenbaum
 - Don Porter
 - me...
 - Photo source: https://img.devrant.com/devrant/rant/r_10875_uRYQF.jpg



Faux Quiz (answer any 2, 5 min)

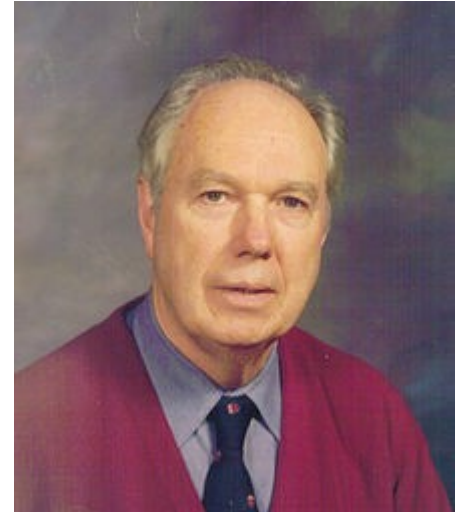
- Who was Flynn? Why is her/his taxonomy important?
- How does domain decomposition differ from functional decomposition? Give examples of each.
- Can a SIMD parallel program use functional decomposition? Why/why not?
- What is an RMW instruction? How can they be used to construct synchronization primitives? How can sync primitives be constructed without them?

Who is Flynn?

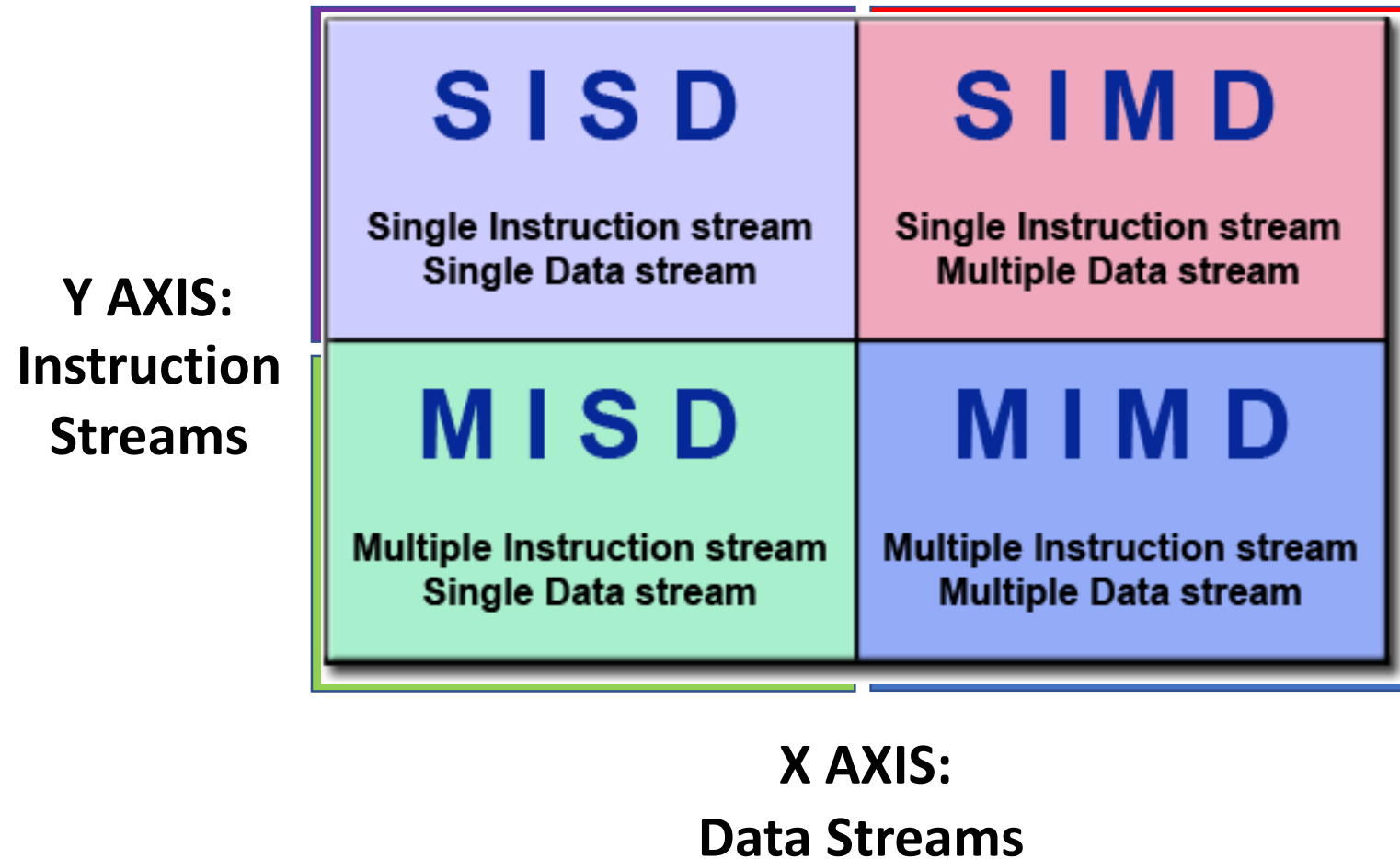
Michael J. Flynn

- Emeritus at Stanford
- Proposed taxonomy in 1966 (!!)
- 30 pages of publication titles
- Founding member of SIGARCH

- (Thanks Wikipedia)



Review: Flynn's Taxonomy



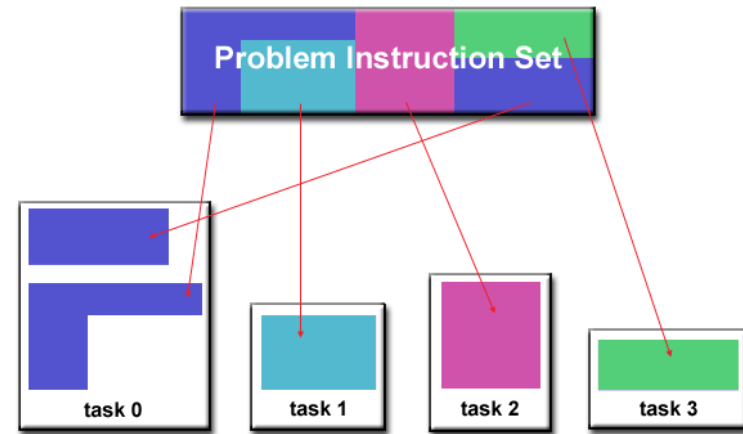
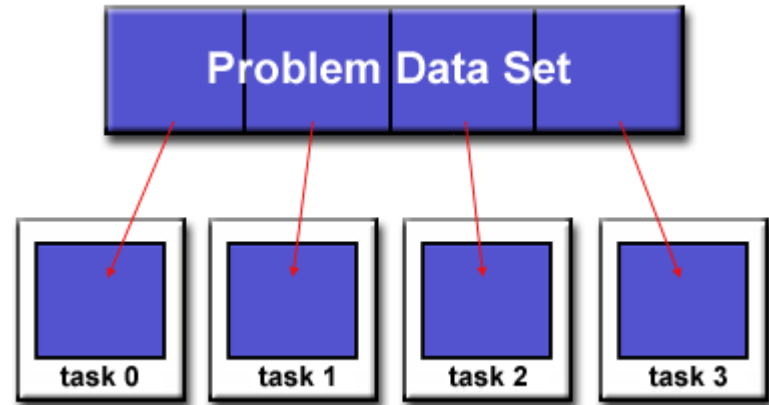
Review: Problem Partitioning

- Domain Decomposition

- SPMD
- Input domain
- Output Domain
- Both

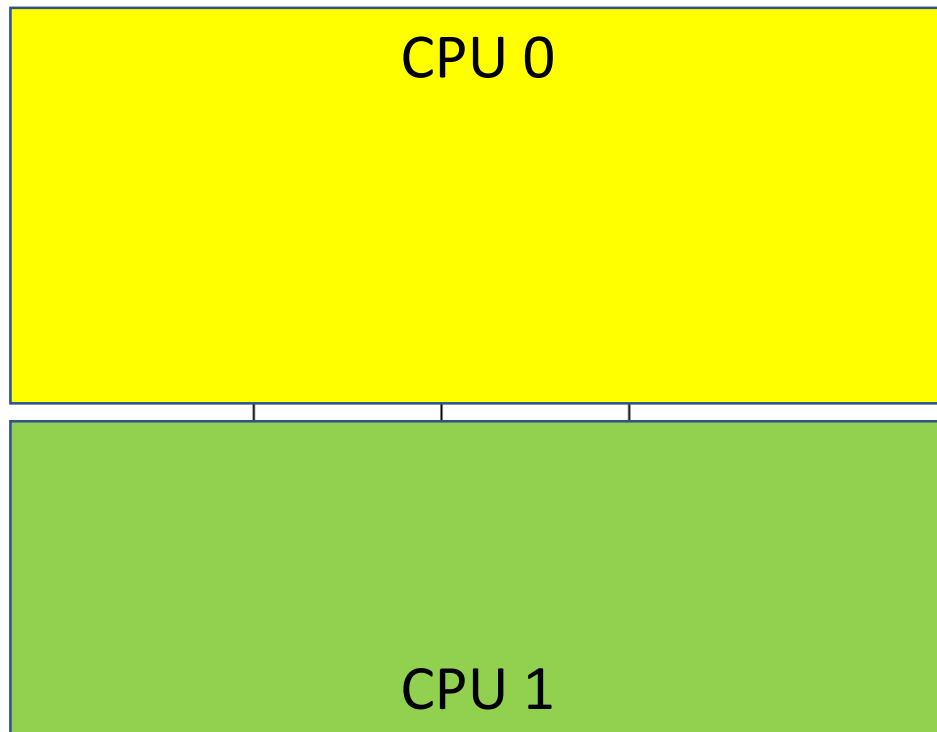
- Functional Decomposition

- MPMD
- Independent Tasks
- Pipelining



Domain decomposition

- Each CPU gets part of the input



Issues?

- Accessing Data
 - Can we access $v(i+1, j)$ from CPU 0
 - ...as in a “normal” serial program?
 - Shared memory? Distributed?
 - Time to access $v(i+1, j) ==$ Time to access $v(i-1, j)$?
 - *Scalability vs Latency*
- Control
 - Can we assign one vertex per CPU?
 - Can we assign one vertex per process/logical task?
 - *Task Management Overhead*
- *Load Balance*
- Correctness
 - order of reads and writes is non-deterministic
 - synchronization is required to enforce the order
 - *locks, semaphores, barriers, conditionals...*

Performance: Amdahl's law

- Speedup is bound by serial component

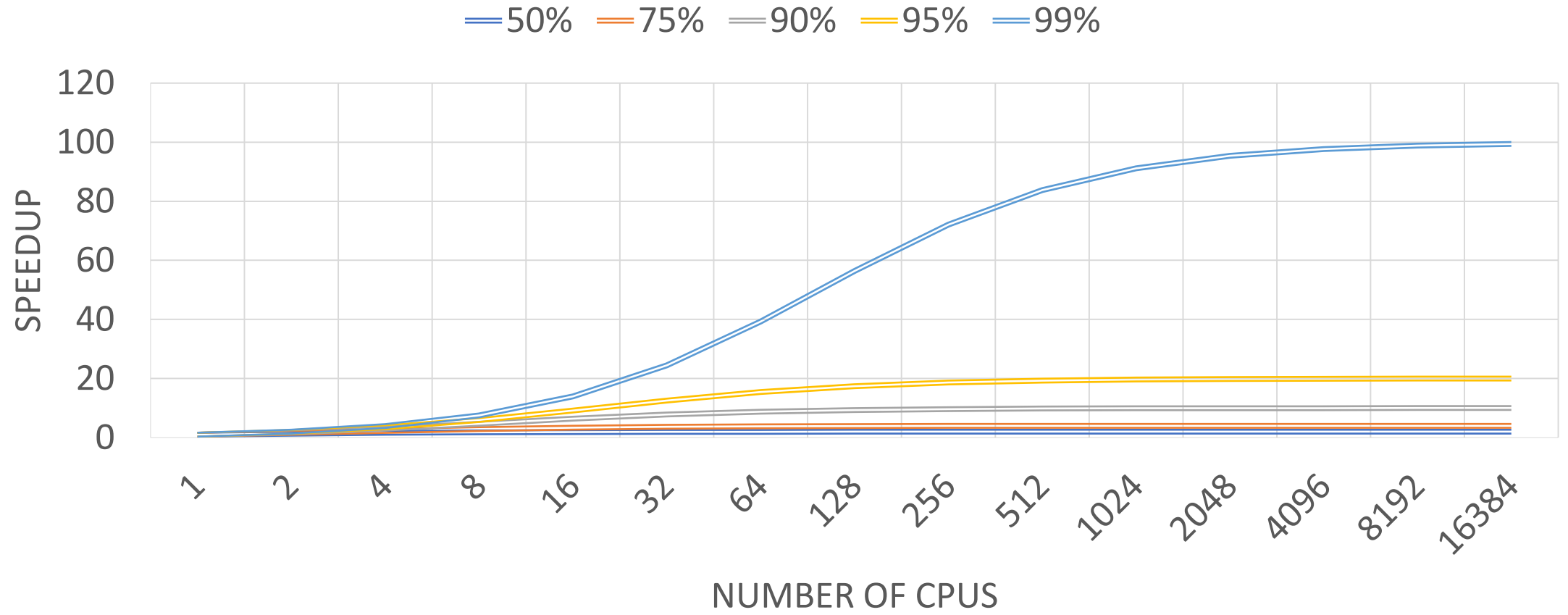
- Sp

$$Speedup = \frac{\text{serial run time}}{\text{parallel run time}}$$

parallel #CPUs

$$Speedup(\#CPUs) = \frac{T_{serial}}{T_{parallel}} = \frac{1}{\frac{A}{\#CPUs} + (1 - A)}$$

Amdahl Action Zone

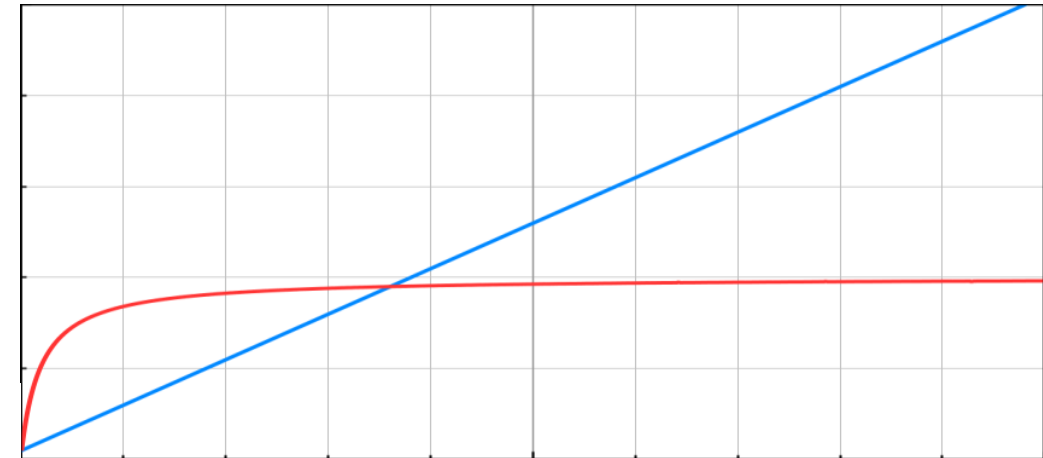


Strong Scaling vs Weak Scaling



Amdahl vs. Gustafson

- $N = \#CPUs$, $S = \text{serial portion} = 1 - A$
- Amdahl's law: $Speedup(N) = \frac{1}{\frac{A}{N} + S}$
 - **Strong scaling:** $Speedup(N)$ calculated given total amount of work is fixed
 - Solve same problems faster when problem size is fixed and #CPU grows
 - Assuming parallel portion is fixed, speedup soon ceases to increase
- Gustafson's law: $Speedup(N) = N + (N-1) \cdot S$
 - **Weak scaling:** $Speedup(N)$ calculated given amount of work per CPU is fixed
 - Keep the amount of work per CPU when adding more CPUs to keep the granularity fixed
 - Problem size grows: solve larger problems
 - **Consequence:** speedup upper bound much higher



When is Gustafson's law a better metric?
When is Amdahl's law a better metric?

Super-linear speedup

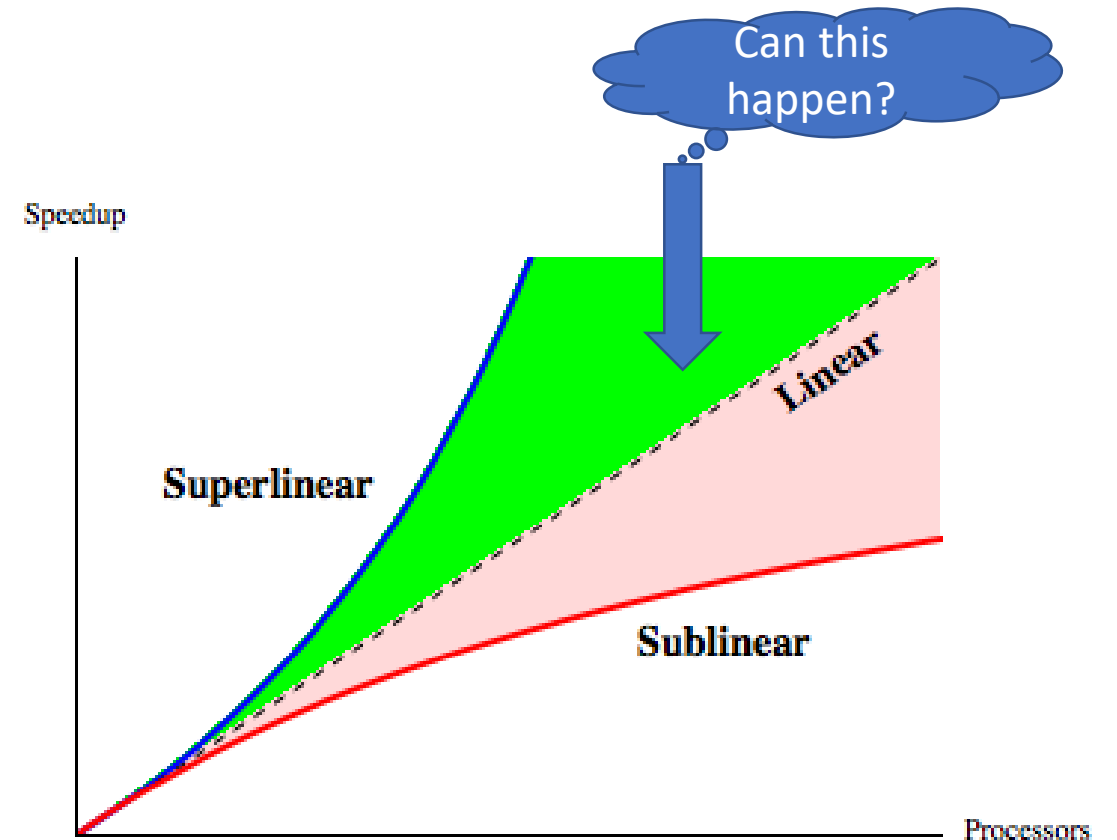
- Possible due to cache
- But usually just poor methodology
- Baseline: ***best*** serial algorithm
- Example:

Efficient **bubble sort**

- *Serial: 150s*
- *Parallel 40s*
- *Speedup: $\frac{150}{40} = 3.75$?*

NO NO NO!

- *Serial quicksort: 30s*
- *Speedup = $30/40 = 0.75X$*



Why insist on best serial algorithm as baseline?

Concurrency and Correctness

If two threads execute this program concurrently,
how many different final values of X are there?

Initially, X == 0.

Thread 1

```
void increment() {  
    int temp = X;  
    temp = temp + 1;  
    X = temp;  
}
```

Thread 2

```
void increment() {  
    int temp = X;  
    temp = temp + 1;  
    X = temp;  
}
```

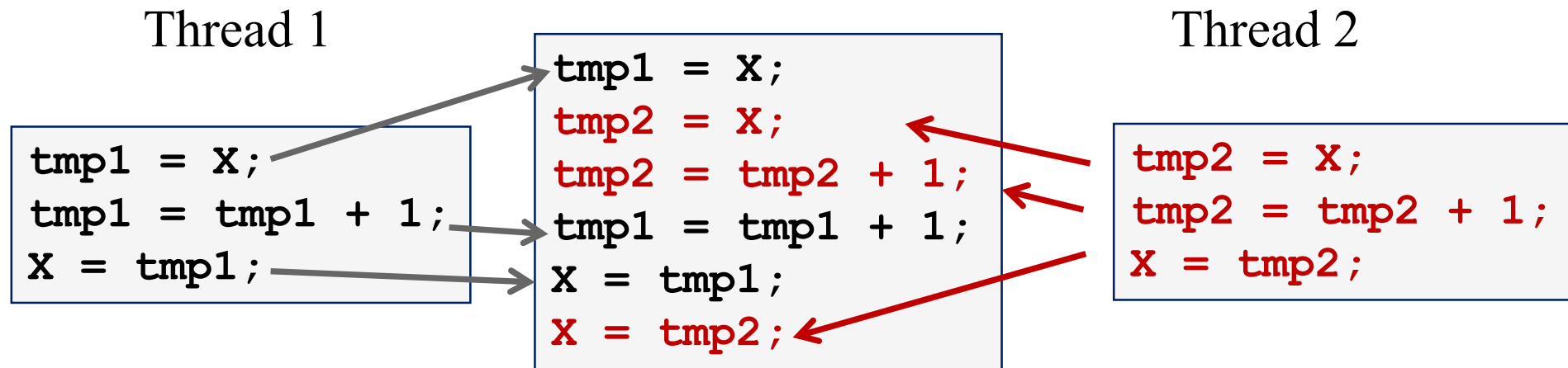
Answer:

- A. 0**
- B. 1**
- C. 2**
- D. More than 2**

Schedules/Interleavings

Model of concurrent execution

- Interleave statements from each thread into a single thread
- If **any** interleaving yields incorrect results, synchronization is needed



If X==0 initially, X == 1 at the end. WRONG result!

Locks fix this with Mutual Exclusion

```
void increment() {  
    lock.acquire();  
    int temp = X;  
    temp = temp + 1;  
    X = temp;  
    lock.release();  
}
```

Mutual exclusion ensures only safe interleavings

- *But it limits concurrency, and hence scalability/performance*

Is mutual exclusion a good abstraction?

Why are Locks “Hard?”

- Coarse-grain locks

- Simple to develop
- Easy to avoid deadlock
- Few data races
- Limited concurrency

```
// WITH FINE-GRAIN LOCKS
void move(T s, T d, Obj key) {
    LOCK(s);
    LOCK(d);
    tmp = s.remove(key);
    d.insert(key, tmp);
    UNLOCK(d);
    UNLOCK(s);
}
```

- Fine-grain locks

- Greater concurrency
- Greater code complexity
- Potential deadlocks
 - Not composable
- Potential data races
 - Which lock to lock?

Thread 0	Thread 1
move(a, b, key1);	
	move(b, a, key2);

DEADLOCK!

Review: correctness conditions

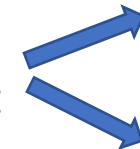
- Safety
 - Only one thread in the critical region
- Liveness
 - Some thread that enters the entry section eventually enters the critical region
 - Even if other thread takes forever in non-critical region
- Bounded waiting
 - ~~A thread that enters the entry section enters the critical section within some bounded number of operations.~~
 - *If a thread i is in entry section, then there is a bound on the number of times that other threads are allowed to enter the critical section before thread i 's request is granted*



Theorem: Every property is a combination of a safety property and a liveness property.

-Bowen Alpern & Fred Schneider

<https://www.cs.cornell.edu/fbs/publications/defliveness.pdf>

Mutex, spinlock, etc.
are ways to implement



```
while (1) {  
      
    Critical section  
      
    Non-critical section  
}
```

Did we get all the important conditions?
Why is correctness defined in terms of locks?

Implementing Locks

```
int lock_value = 0;  
int* lock = &lock_value;
```

```
Lock::Acquire() {  
    while (*lock == 1)  
        ; //spin  
    *lock = 1;  
}
```

```
Lock::Release() {  
    *lock = 0;  
}
```

Completely and utterly broken.
How can we fix it?

What are the problem(s) with this?

- A. CPU usage
- B. Memory usage
- C. Lock::Acquire() latency
- D. Memory bus usage
- E. Does not work

HW Support for Read-Modify-Write (RMW)

IDEA: hardware implements something like:

```
bool rmw(addr, value) {  
    atomic {  
        tmp = *addr;  
        newval = modify(tmp);  
        *addr = newval;  
    }  
}
```

Why is that hard?
How can we do it?

Preview of Techniques:

- Bus locking
- Single Instruction ISA extensions
 - Test&Set
 - CAS: Compare & swap
 - Exchange, locked increment, locked decrement (x86)
- Multi-instruction ISA extensions:
 - LLSC: (PowerPC, Alpha, MIPS)
 - Transactional Memory (x86, PowerPC)

More on this later...

Implementing Locks with Test&set

```
int lock_value = 0;  
int* lock = &lock_value;
```

```
Lock::Acquire() {  
    while (test&set(lock) == 1)  
        ; //spin  
}
```

```
Lock::Release() {  
    *lock = 0;  
}
```



(test & set ~ = CAS ~ = LLSC)

TST: *Test&set*

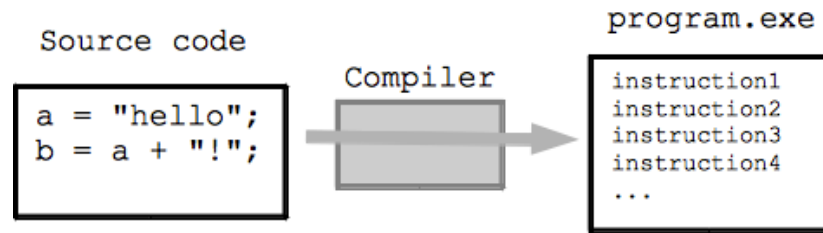
- Reads a value from memory
- Write "1" back to memory location

What are the problem(s) with this?

- A. CPU usage
- B. Memory usage
- C. Lock::Acquire() latency
- D. Memory bus usage
- E. Does not work

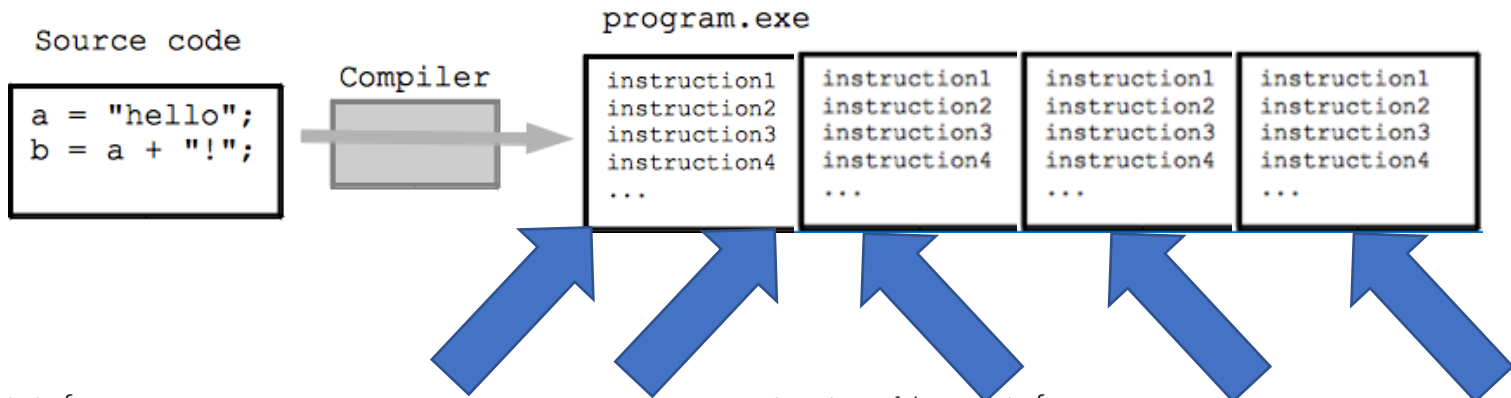
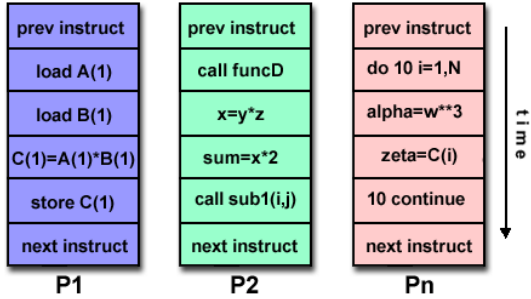
More on this later...

Programming and Machines: a mental model



```
struct machine_state{  
    uint64 pc;  
    uint64 Registers[16];  
    uint64 cr[6]; // control registers cr0-cr4 and EFER on AMD  
    ...  
} machine;  
while(1) {  
    fetch_instruction(machine.pc);  
    decode_instruction(machine.pc);  
    execute_instruction(machine.pc);  
}  
void execute_instruction(i) {  
    switch(opcode) {  
    case add_rr:  
        machine.Registers[i.dst] += machine.Registers[i.src];  
        break;  
    }  
}
```

Parallel Machines: a mental model



```

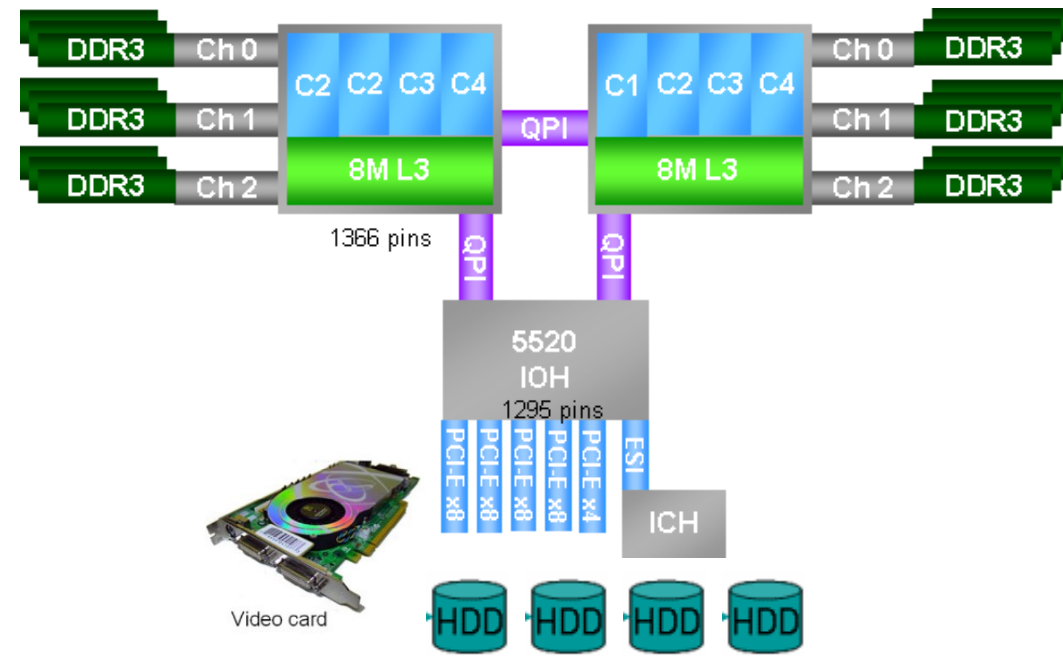
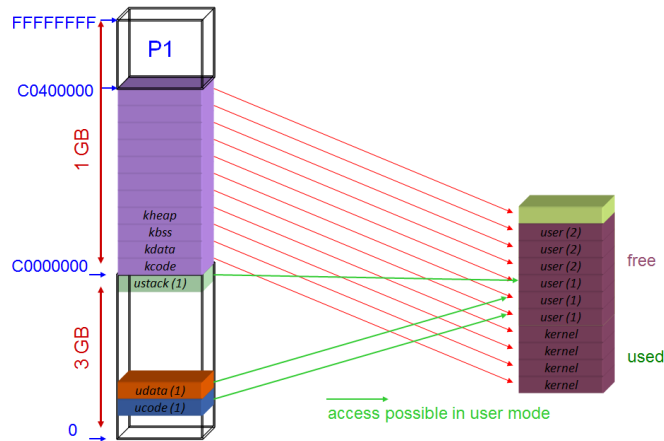
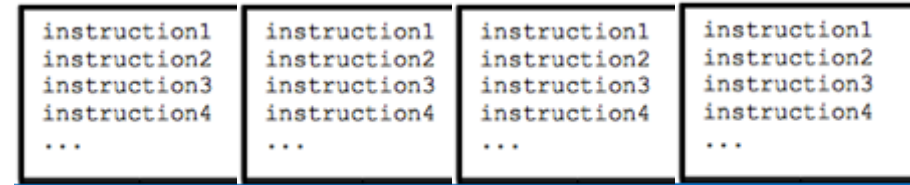
struct machine_state{
    uint64 pc;
    uint64 Registers[16];
    uint64 cr[6]; // control registers cr0-cr4 and EFER on AMD
    ...
} machine;
while(1) {
    fetch_instruction(machine.pc);
    decode_instruction(machine.pc);
    execute_instruction(machine.pc);
}
void execute_instruction(i) {
    switch(opcode) {
    case add_rr:
        machine.Registers[i.dst] += machine.Registers[i.src];
        break;
    }
}
    
```

```

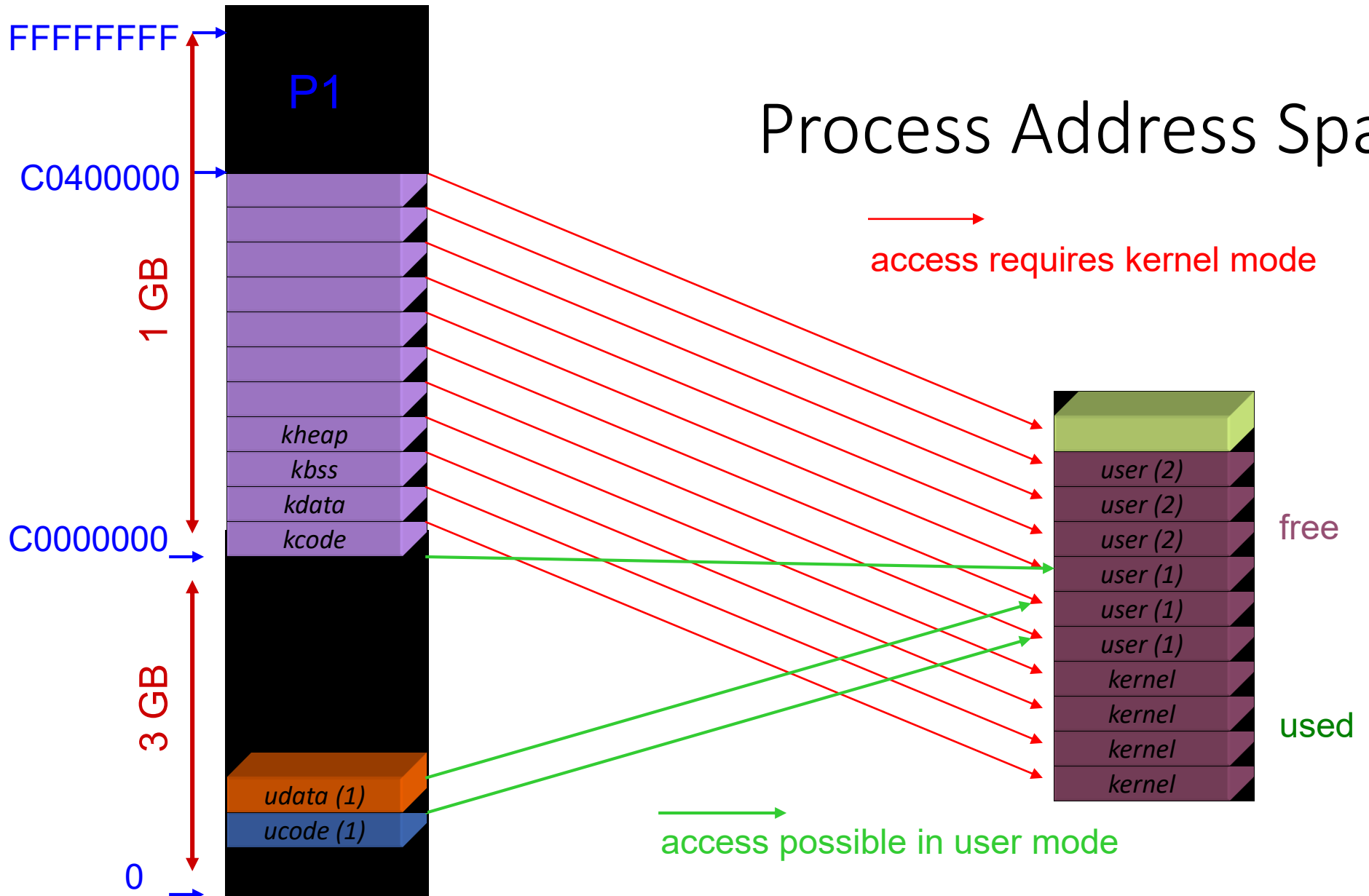
struct machine_state{
    uint64 pc;
    uint64 Registers[16];
    uint64 cr[6]; // control registers cr0-cr4 and EFER on AMD
    ...
} machine;
while(1) {
    fetch_instruction(machine.pc);
    decode_instruction(machine.pc);
    execute_instruction(machine.pc);
}
void execute_instruction(i) {
    switch(opcode) {
    case add_rr:
        machine.Registers[i.dst] += machine.Registers[i.src];
        break;
    }
}
    
```

Processes and Threads and Fibers...

- Abstractions
- Containers
- State
 - Where is shared state?
 - How is it accessed?
 - Is it mutable?



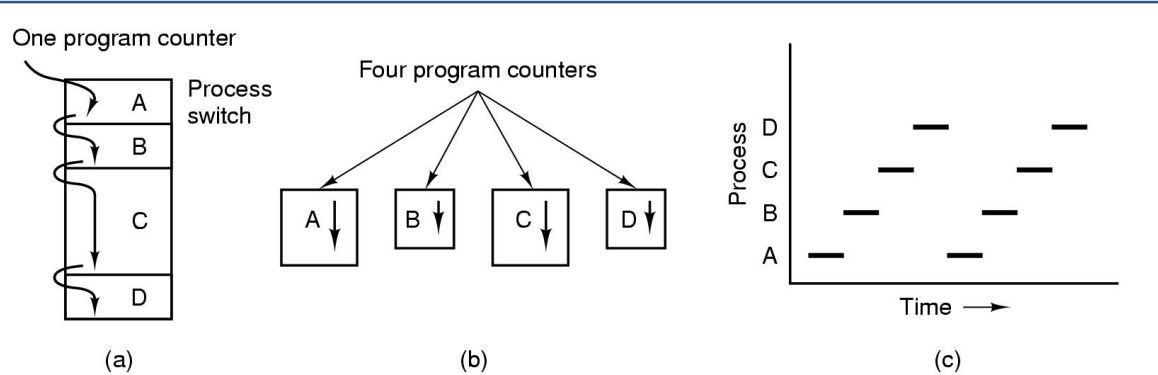
Process Address Space



Anyone see an issue?

Processes

Model

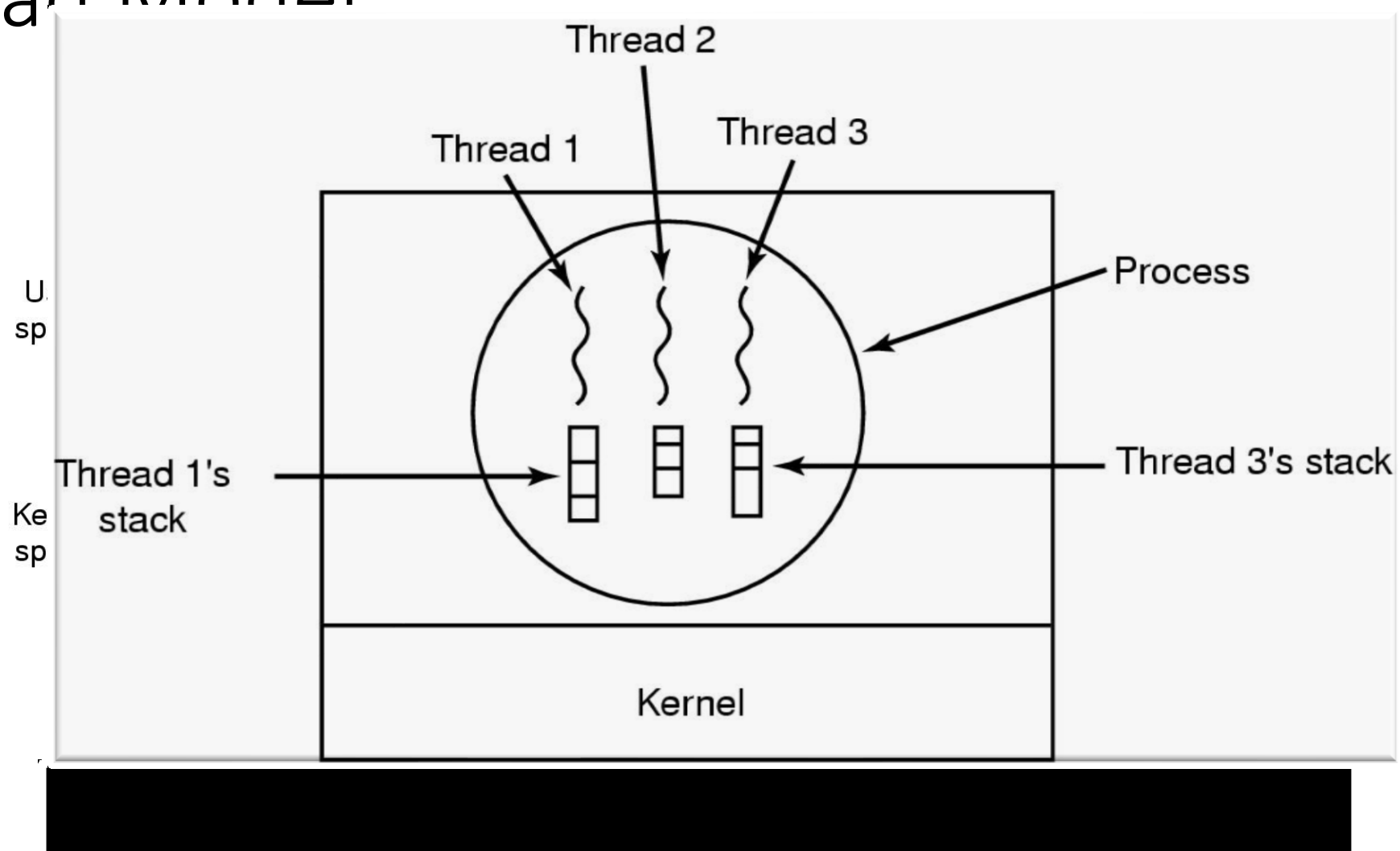


- Multiprogramming of four programs
- Conceptual model of 4 independent, sequential processes
- Only one program active at any instant

Implementation




Process management	Memory management	File management
Registers	Pointer to text segment	Root directory
Program counter	Pointer to data segment	Working directory
Program status word	Pointer to stack segment	File descriptors
Stack pointer		User ID
Process state		Group ID
Priority		
Scheduling parameters		
Process ID		
Parent process		
Process group		
Signals		
Time when process started		
CPU time used		
Children's CPU time		
Time of next alarm		

Thread Model



when might (a) be better than (b)? vice versa?

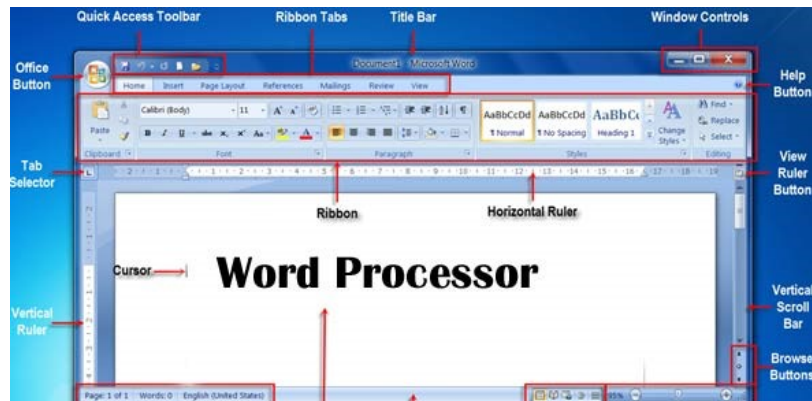
The Thread Model

Per process items Address space Global variables Open files Child processes Pending alarms Signals and signal handlers Accounting information	Per thread items Program counter Registers Stack State	Process management  Program status word  Process state  Process ID Parent process Process group Signals Time when process started CPU time used Children's CPU time Time of next alarm	Memory management Pointer to text segment Pointer to data segment Pointer to stack segment	File management Root directory Working directory File descriptors User ID Group ID
---	---	---	--	--

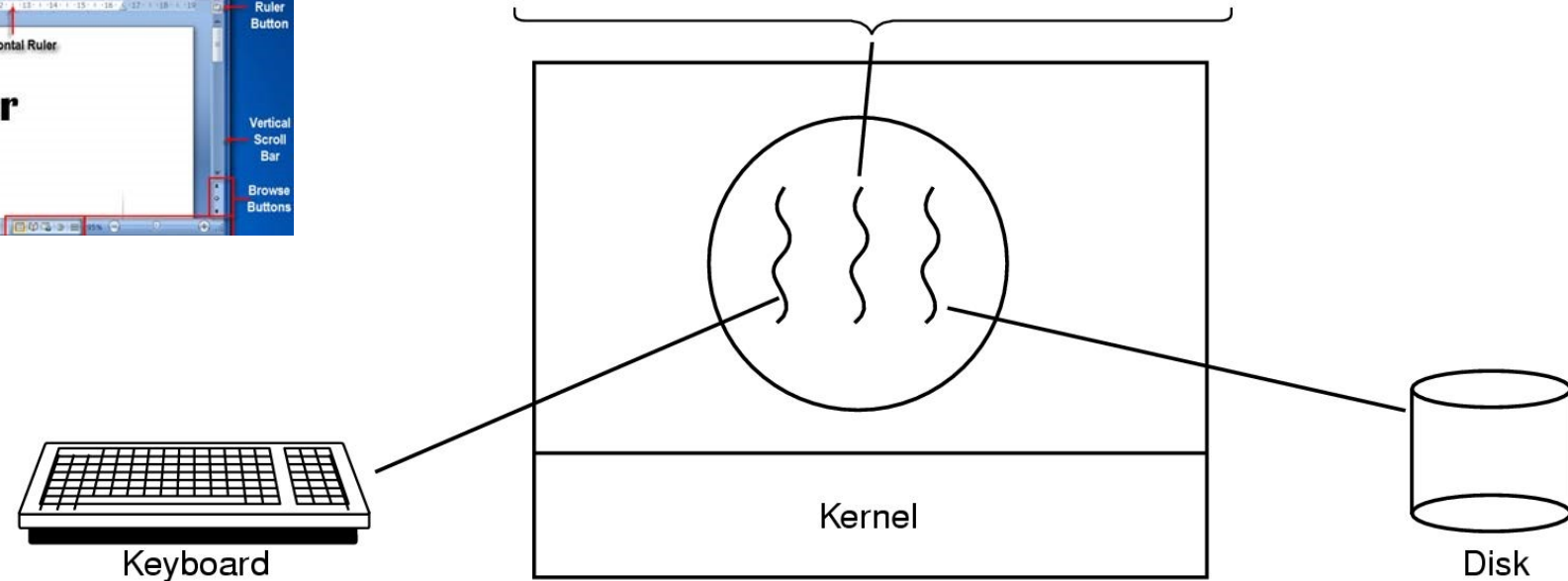
- Items shared by all threads in a process
- Items private to each thread

Using threads

Ex. How might we use threads in a word processor program?

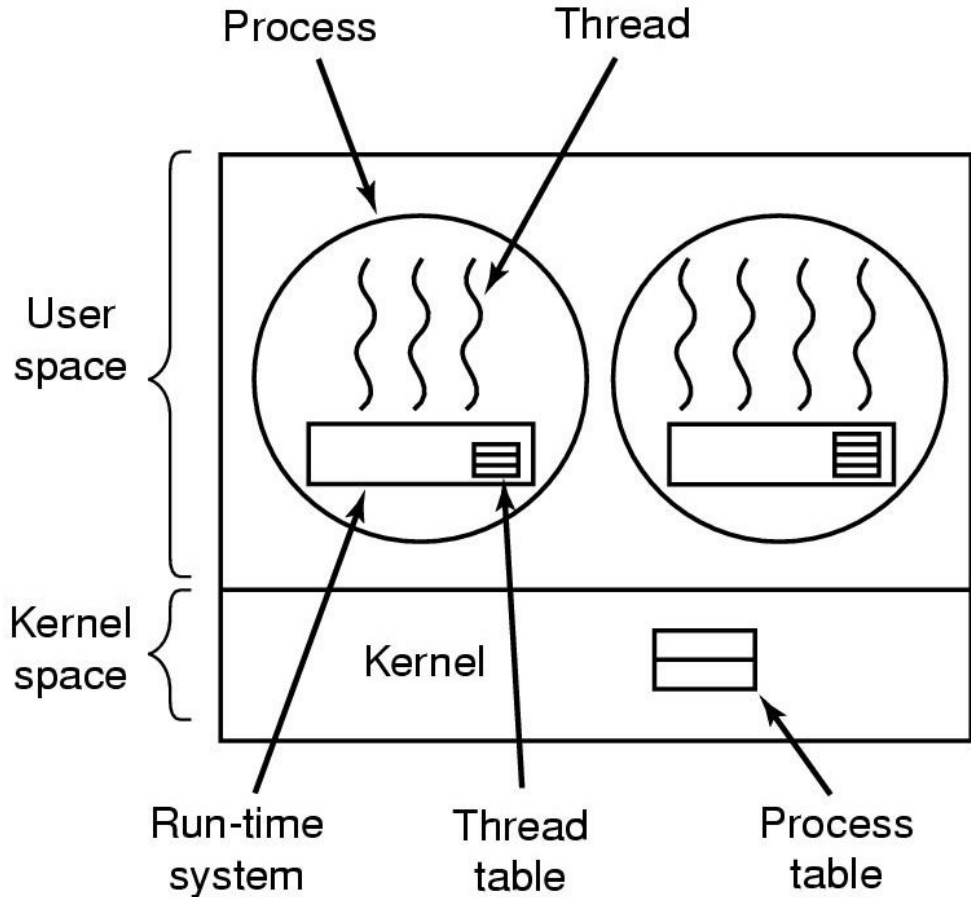


Four score and seven years ago, our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war testing whether that	nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting place for those who here gave their	lives that this nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate, we cannot consecrate we cannot hallow this ground. The brave men, living and dead,	who struggled here have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated	here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us, that from these honored dead we take increased devotion to that cause for which	they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain that this nation, under God, shall have a new birth of freedom and that government of the people, for the people,
---	--	---	---	---	--



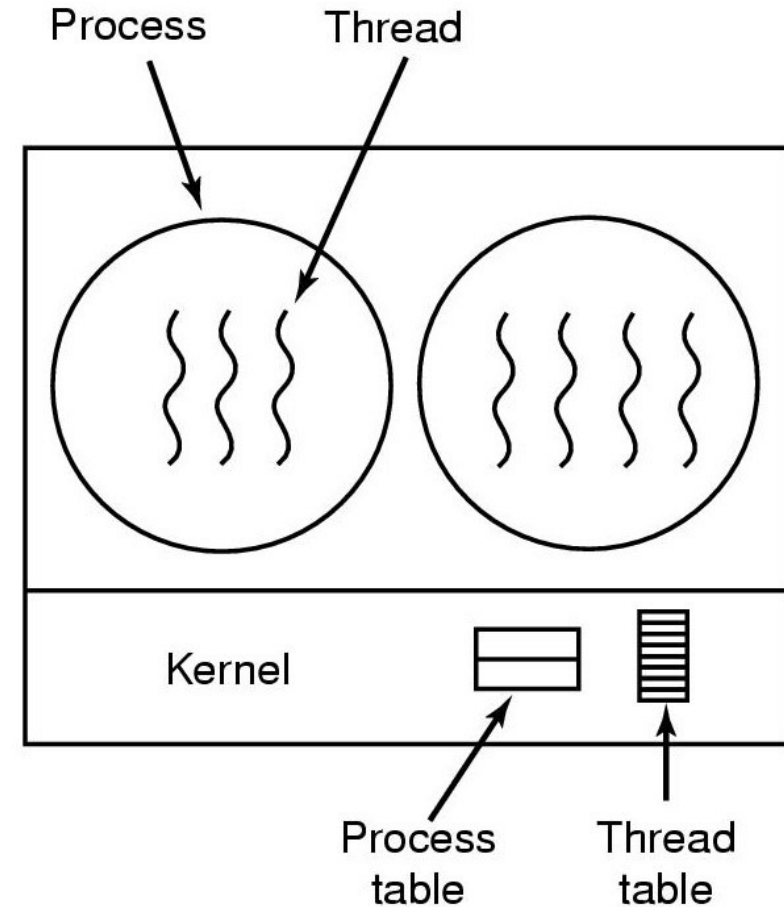
Where to Implement Threads:

User Space



A user-level threads package

Kernel Space



A threads package managed by the kernel

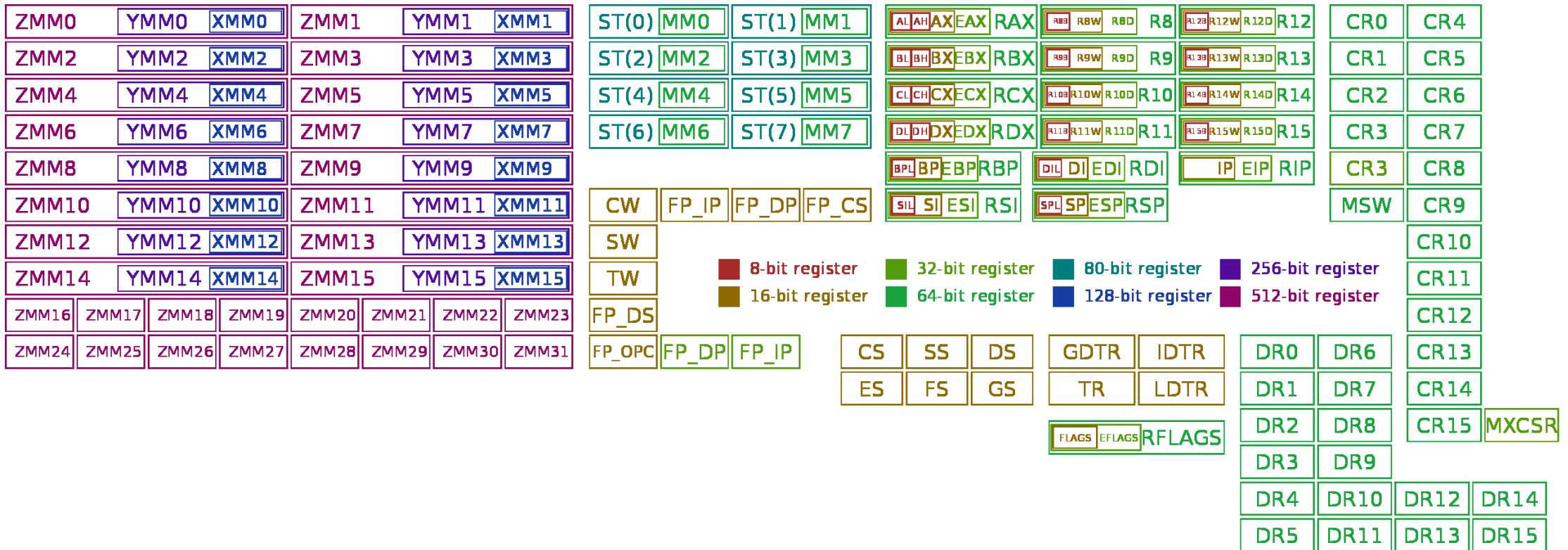
Threads vs Fibers

Blah blah **fibers**
blah **exception**
handling blah...



- Like threads, *just an abstraction* for flow of control
- *Lighter weight* than threads
 - In Windows, just a stack, subset of arch. registers, non-preemptive
 - *Not* just threads without exception support
 - stack management/impl has interplay with exceptions
 - Can be completely exception safe
- **Takeaway**: diversity of abstractions/containers for execution flows

x86_64 Architectural Registers



• Register map diagram courtesy of: By Immae - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=32745525>

```

/*
 * switch_to(x,y) should switch tasks from x to y.
 *
 * This could still be optimized:
 * - fold all the options into a flag word and test it with a single test.
 * - could test fs/gs bitsliced
 *
 * Kprobes not supported here. Set the probe on schedule instead.
 * Function graph tracer not supported too.
 */

```

Linux x86_64 context switch excerpt

Complete fiber context switch on Unix and Windows

```

__visible __notrace_funcgraph struct task_struct *
__switch_to(struct task_struct *prev_p, struct task_struct *next_p)
{
    struct thread_struct *prev = &prev_p->thread;
    struct thread_struct *next = &next_p->thread;
    struct fpu *prev_fpu = &prev->fpu;
    struct fpu *next_fpu = &next->fpu;
    int cpu = smp_processor_id();
    struct tss_struct *tss = &per_cpu(cpu_tss_rw, cpu);

    WARN_ON_ONCE(IS_ENABLED(CONFIG_DEBUG_ENTRY) &&
        this_cpu_read(irq_count) != -1);

    switch_fpu_prepare(prev_fpu, cpu);

    /* We must save %fs and %gs before load_TLS() because
     * %fs and %gs may be cleared by load_TLS().
     */
    /* (e.g. xen_load_tls())
     */
    save_fsgs(prev_p);

    /*
     * Load TLS before restoring any segments so that segment loads
     * reference the correct GDT entries.
     */
    load_TLS(next, cpu);

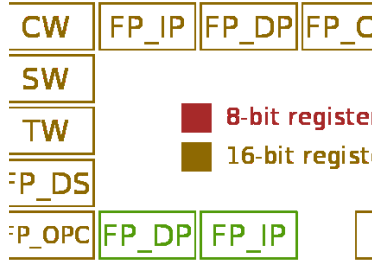
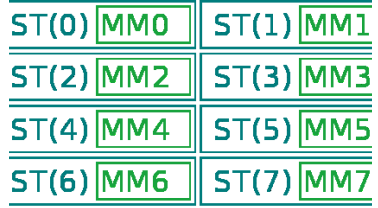
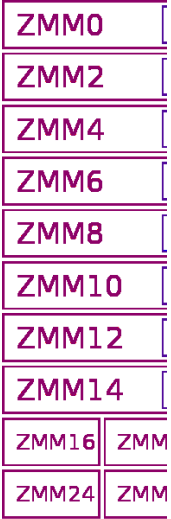
    /*
     * Leave lazy mode, flushing any hypercalls made here. This
     * must be done after loading TLS entries in the GDT but before
     * loading segments that might reference them, and and it must
     * be done before fpu_restore(), so the TS bit is up to
     * date.
     */
    arch_end_context_switch(next_p);

    /* Switch DS and ES.
     */
    /* Reading them only returns the selectors, but writing them (if
     * nonzero) loads the full descriptor from the GDT or LDT. The
     * LDT for next is loaded in switch_mm, and the GDT is loaded
     * above.
     */
    /* We therefore need to write new values to the segment
     * registers on every context switch unless both the new and old
     * values are zero.
     */
    /* Note that we don't need to do anything for CS and SS, as
     * those are saved and restored as part of pt_regs.
     */
    savesegment(es, prev->es);
    if (unlikely(next->es | prev->es))
        loadsegment(es, next->es);

    savesegment(ds, prev->ds);
    if (unlikely(next->ds | prev->ds))
        loadsegment(ds, next->ds);

    load_seg_legacy(prev->fsindex, prev->fsbase,
        next->fsindex, next->fsbase, FS);
    load_seg_legacy(prev->gsindex, prev->gsbase,
        next->gsindex, next->gsbase, GS);
}

```

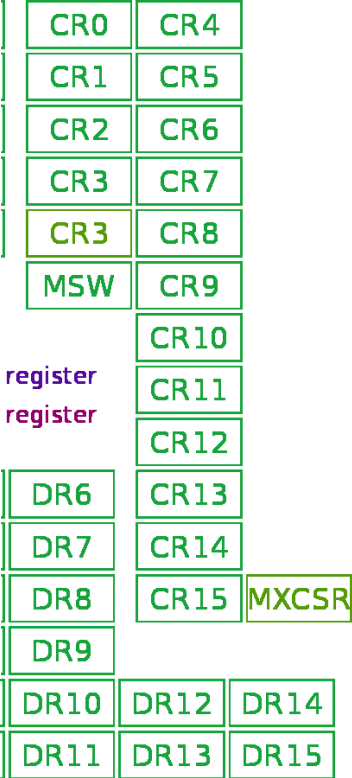


* The AMD64 architecture provides 16 general 64-bit registers together with 16 * 128-bit SSE registers, overlapping with 8 legacy 80-bit x87 floating point registers.

	Both	Unix only	Windows only
* rax	Result register		
* rbx	Must be preserved		
* rcx		Fourth argument	First argument
* rdx		Third argument	Second argument
* rsp	Stack pointer, must be preserved		
* rbp	Frame pointer, must be preserved		
* rsi		Second argument	Must be preserved
* rdi		First argument	Must be preserved
* r8		Fifth argument	Third argument
* r9		Sixth argument	Fourth argument
* r10-r11	Volatile		
* r12-r15	Must be preserved		
* xmm0-5	Volatile		
* xmm6-15		Volatile	Must be preserved
* fpcsr	Non volatile		
* mxcsr	Non volatile		

* Thus for the two architectures we get slightly different lists of registers * to preserve.

* Registers "owned" by caller:
 * Unix: rbx, rsp, rbp, r12-r15, mxcsr (control bits), x87 CW
 * Windows: rbx, rsp, rbp, rsi, rdi, r12-r15, xmm6-15



• Reg

x86_64 Registers and Threads

```
/*
 * switch_context() should switch tasks from x to y.
 *
 * This could still be optimized:
 * - If all the entries have a flag word and test it with a single test.
 * - could test flags offload.
 *
 * Kernels not supported here. Set the probe on schedule instead.
 * Function graph tracer not supported too.
 */
__attribute__((optimize("O3"))) void switch_context(struct task_struct *next,
                                                  struct task_struct *prev)
{
    struct thread_struct *tcpu = &prev->thread;
    struct thread_struct *ntcpu = &next->thread;
    struct fpu_state_fpu *fpu = &fpu;
    struct fpu_state_fpu *ntfpu = &ntfpu;
    struct task_struct *tcpu = &prev->cpu;
    struct task_struct *ntcpu = &next->cpu;

    WARN_ON_ONCE(!IS_ENABLED(CONFIG_DEBUG_ENTRY)) ||
        WARN_ON_ONCE(!IS_ENABLED(CONFIG_DEBUG_ENTRY));

    switch_fpu_prepare(prev, fpu, cpu);

    /* We must save R15 and R16 before load_TSS() because
     * R15 and R16 may be cleared by load_TSS().
     * (i.e. when load_TSS())
     */
    save_fpu(prev, fpu);

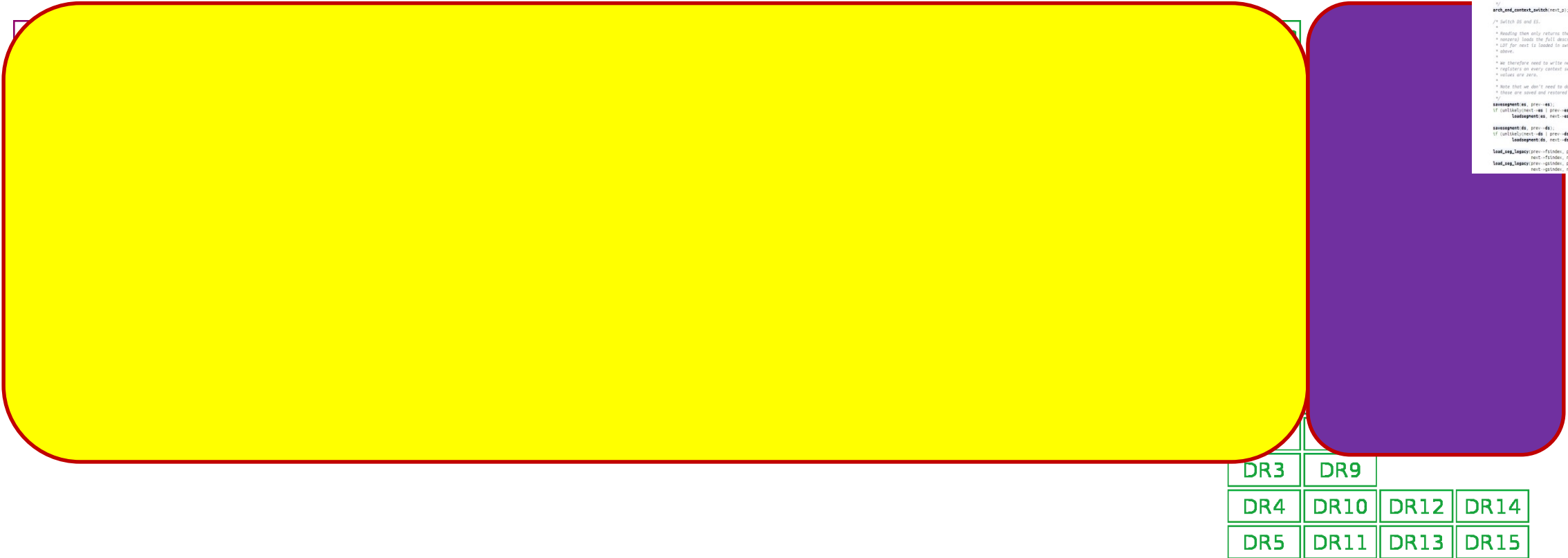
    /*
     * Load TSS before restoring any segments so that segment loads
     * reference the correct GDT entries.
     */
    load_TSS(next, ccpu);

    /*
     * Leave lazy mode, flushing any hypercalls made here. This
     * must be done after loading TSS entries in the GDT but before
     * loading segments that might reference them, and and it must
     * be done before fpu_restore(), so the TS bit is up to
     * 0.000.
     */
    arch_and_context_switch(next, fpu);

    /*
     * Switch DS and SS.
     *
     * Reading them only returns the selectors, but writing them (if
     * nonzero) loads the full descriptor from the GDT or LDT. The
     * LDT for next is loaded in switch_mm, and the GDT is loaded
     * above.
     *
     * We therefore need to write new values to the segment
     * registers on every context switch unless both the new and old
     * values are zero.
     *
     * Note that we don't need to do anything for CS and SS, as
     * those are saved and restored as part of pt_regs.
     */
    save_segment_ds_prev = ds;
    if (unlikely(next->ds != prev->ds))
        load_segment_ds_prev = ds;

    save_segment_ds_next = ds;
    if (unlikely(next->ds != prev->ds))
        load_segment_ds_next = ds;

    load_seg_legacy_prev = fpu->state;
    next = fpu->state;
    load_seg_legacy_prev = fpu->state;
    next = fpu->state;
}
```

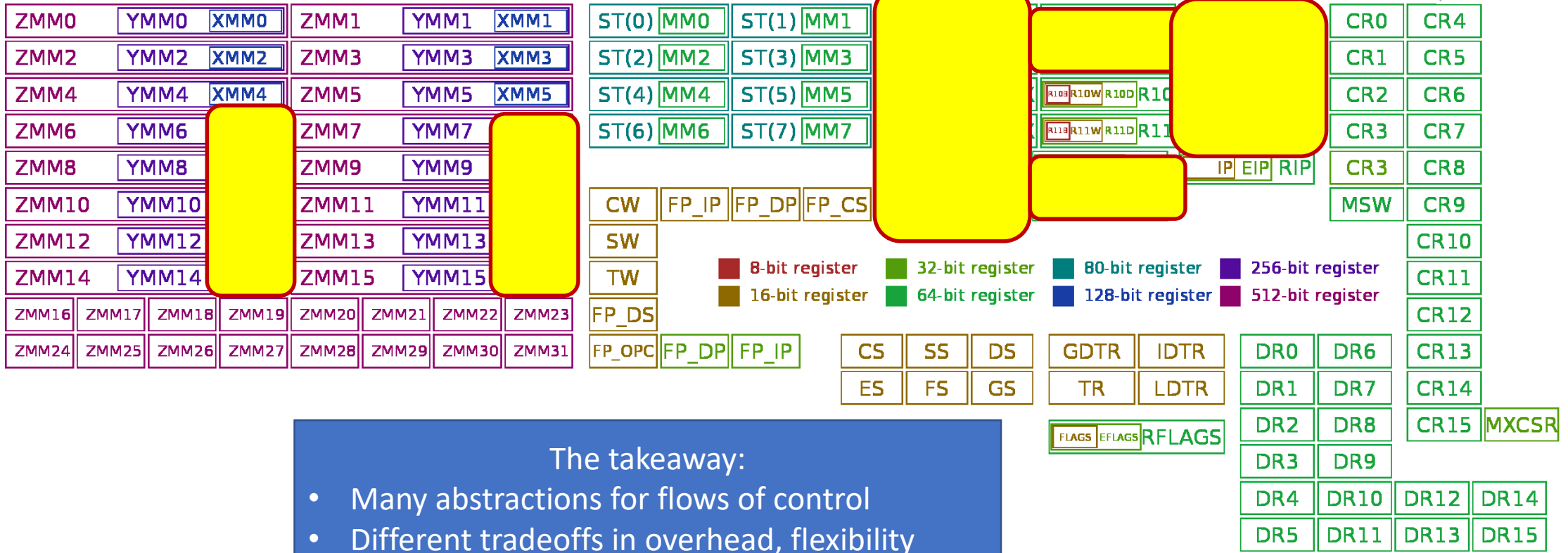


• Register map diagram courtesy of: By Immae - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=32745525>

x86_64 Registers and Fibers

```

* The AMD64 architecture provides 16 general 64-bit registers together with 16
* 128-bit SSE registers, overlapping with 8 legacy 80-bit x87 floating point
* registers.
*
* Both      Unix only      Windows only
* -----
* rax      Result register
* rbx      Must be preserved
* rcx      Fourth argument      First argument
* rdx      Third argument      Second argument
* rsp      Stack pointer, must be preserved
* rbp      Frame pointer, must be preserved
* rsi      Second argument      Must be preserved
* rdi      First argument      Must be preserved
* r8       Fifth argument      Third argument
* r9       Sixth argument      Fourth argument
* r10-r11  Volatile
* r12-r15  Must be preserved
* xmm0-5   Volatile
* xmm6-15  Volatile      Must be preserved
* fpcsr    Non volatile
* mxcsr    Non volatile
*
* Thus for the two architectures we get slightly different lists of registers
* to preserve.
*
* Registers "owned" by caller:
* Unix:   rbp, rsp, rbp, r12-r15, mxcsr (control bits), x87 Cw
* Windows: rbp, rsp, rbp, rsi, rdi, r12-r15, xmm0-15
    
```



The takeaway:

- Many abstractions for flows of control
- Different tradeoffs in overhead, flexibility
- Matters for concurrency: exercised heavily

• Register map diagram courtesy of: By Immae - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=32745525>

Pthreads

- POSIX standard thread model,
- Specifies the API and call semantics.
- Popular – most thread libraries are Pthreads-compatible

Preliminaries

- Include `pthread.h` in the main file
- Compile program with `-lpthread`
 - `gcc -o test test.c -lpthread`
 - may not report compilation errors otherwise but calls will fail
- Good idea to check return values on common functions

Thread creation

- Types: `pthread_t` – type of a thread
- Some calls:

```
int pthread_create(pthread_t *thread,  
                  const pthread_attr_t *attr,  
                  void * (*start_routine)(void *),  
                  void *arg);  
  
int pthread_join(pthread_t thread, void **status);  
int pthread_detach();  
void pthread_exit();
```

- No explicit parent/child model, except main thread holds process info
- Call `pthread_exit` in main, don't just fall through;
- When do you need `pthread_join` ?
 - `status` = exit value returned by joinable thread
- Detached threads are those which cannot be joined (can also set this at creation)

Creating multiple threads

```
#include <stdio.h>
#include <pthread.h>
#define NUM_THREADS 4

void *hello (void *arg) {
    printf("Hello Thread\n");
}

main() {
    pthread_t tid[NUM_THREADS];
    for (int i = 0; i < NUM_THREADS; i++)
        pthread_create(&tid[i], NULL, hello, NULL);

    for (int i = 0; i < NUM_THREADS; i++)
        pthread_join(tid[i], NULL);
}
```

Can you find the bug here?

What is printed for myNum?

```
void *threadFunc(void *pArg) {
    int* p = (int*)pArg;
    int myNum = *p;
    printf( "Thread number %d\n", myNum);
}

. . .
// from main():
for (int i = 0; i < numThreads; i++) {
    pthread_create(&tid[i], NULL, threadFunc, &i);
}
```

Pthread Mutexes

- Type: `pthread_mutex_t`

```
int pthread_mutex_init(pthread_mutex_t *mutex,  
                        const pthread_mutexattr_t *attr);  
int pthread_mutex_destroy(pthread_mutex_t *mutex);  
int pthread_mutex_lock(pthread_mutex_t *mutex);  
int pthread_mutex_unlock(pthread_mutex_t *mutex);  
int pthread_mutex_trylock(pthread_mutex_t *mutex);
```

- Attributes: for shared mutexes/condition vars among processes, for priority inheritance, etc.
 - use defaults
- Important: Mutex scope must be visible to all threads!

Pthread Spinlock

- Type: `pthread_spinlock_t`

```
int pthread_spinlock_init(pthread_spinlock_t *lock);  
int pthread_spinlock_destroy(pthread_spinlock_t *lock);  
int pthread_spin_lock(pthread_spinlock_t *lock);  
int pthread_spin_unlock(pthread_spinlock_t *lock);  
int pthread_spin_trylock(pthread_spinlock_t *lock);
```

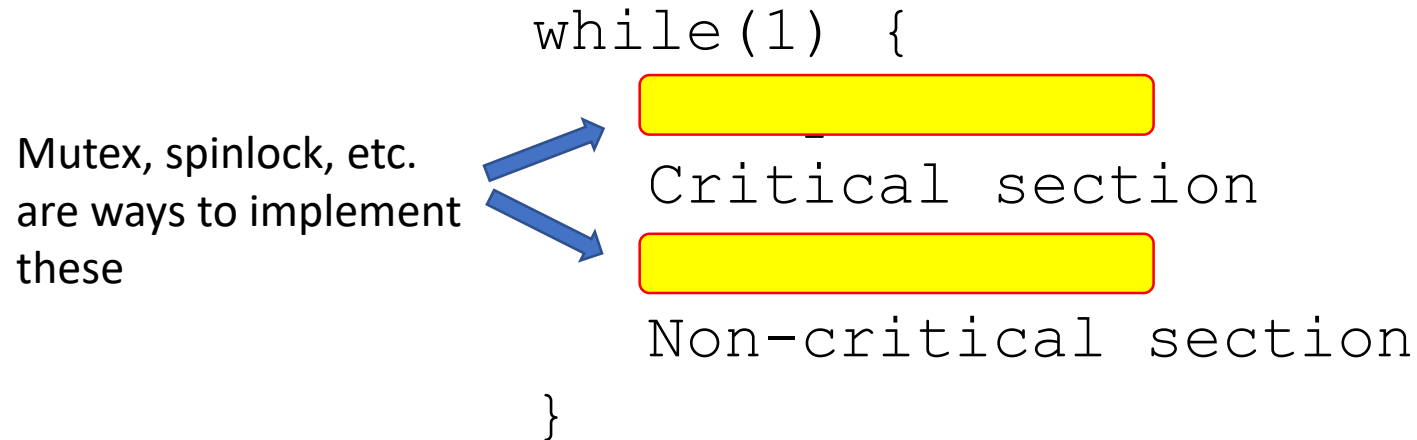
Wait...what's the
difference?



```
int pthread_mutex_init(pthread_mutex_t *mutex,...);  
int pthread_mutex_destroy(pthread_mutex_t *mutex);  
int pthread_mutex_lock(pthread_mutex_t *mutex);  
int pthread_mutex_unlock(pthread_mutex_t *mutex);  
int pthread_mutex_trylock(pthread_mutex_t *mutex);
```


Review: mutual exclusion model

- Safety
 - Only one thread in the critical region
- Liveness
 - Some thread that enters the entry section eventually enters the critical region
 - Even if other thread takes forever in non-critical region

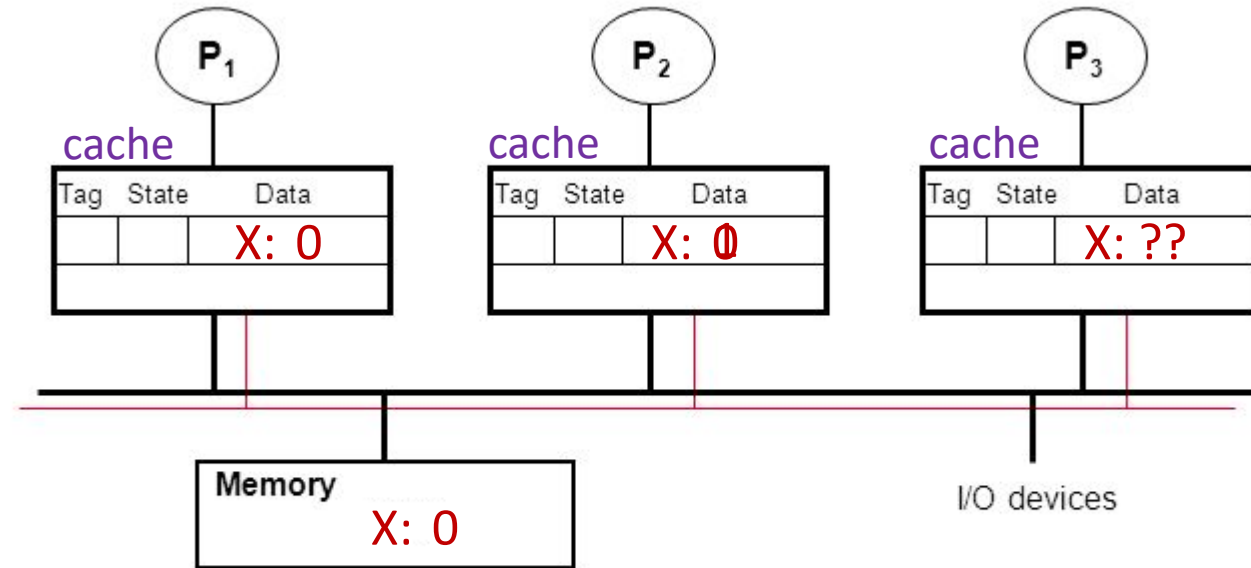


Multiprocessor Cache Coherence

Physics | Concurrency

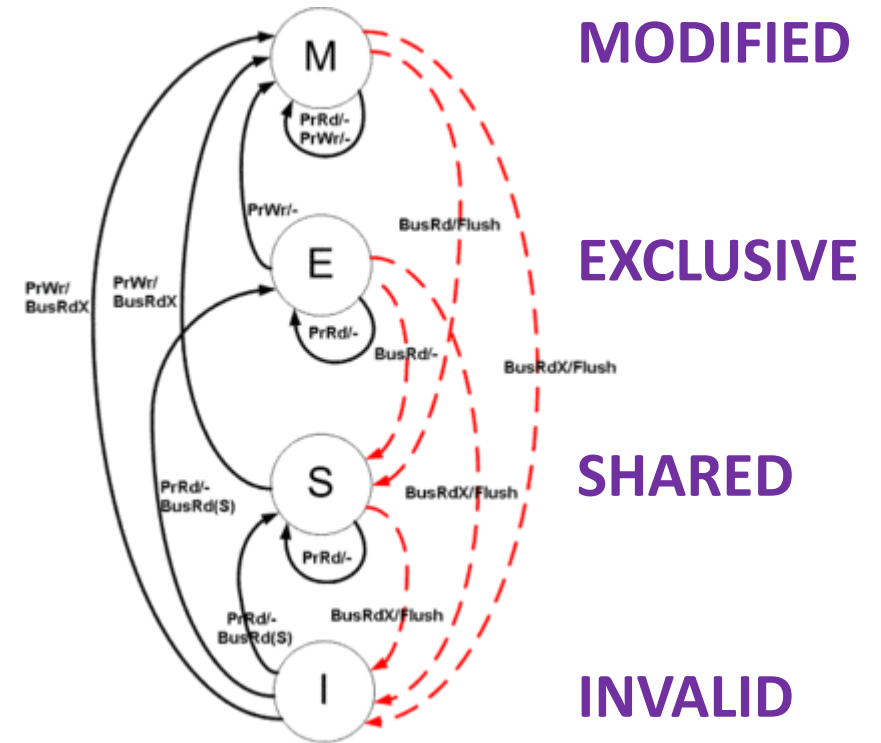
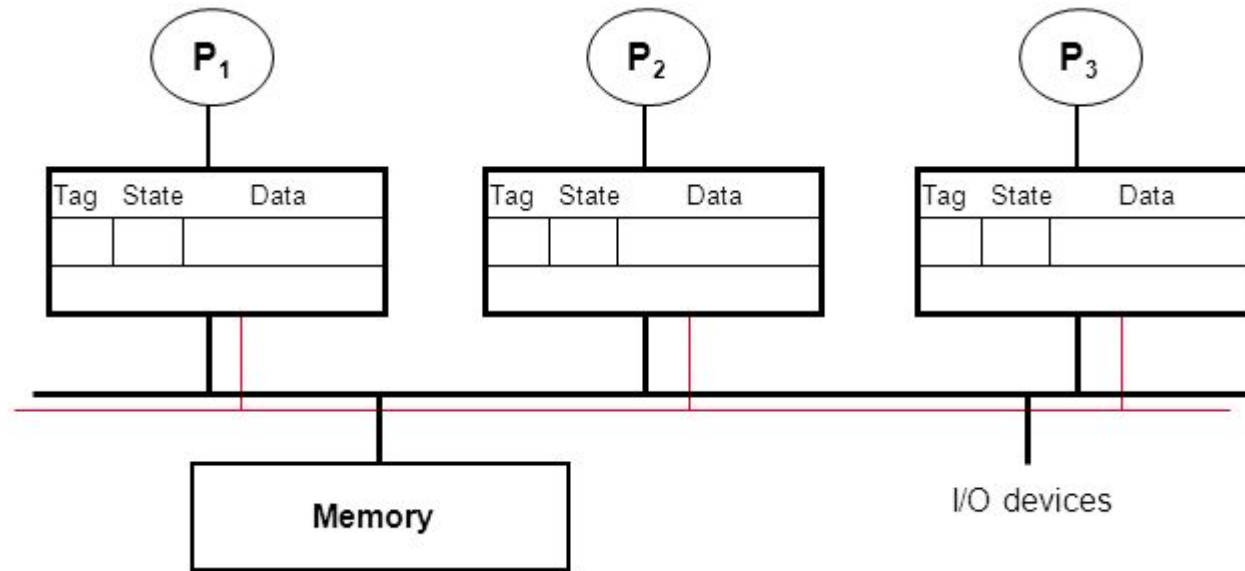
$F = ma \sim coherence$

Multiprocessor Cache Coherence



- P₁: read X
- P₂: read X
- P₂: X++
- P₃: read X

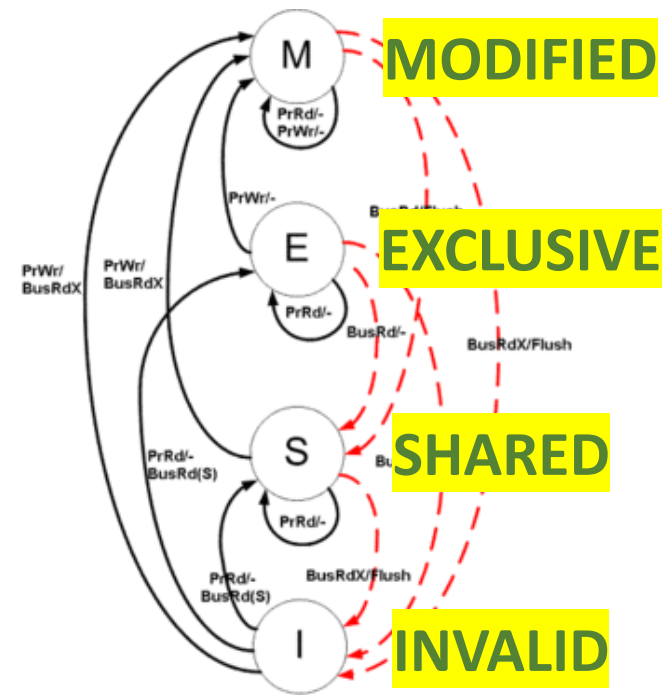
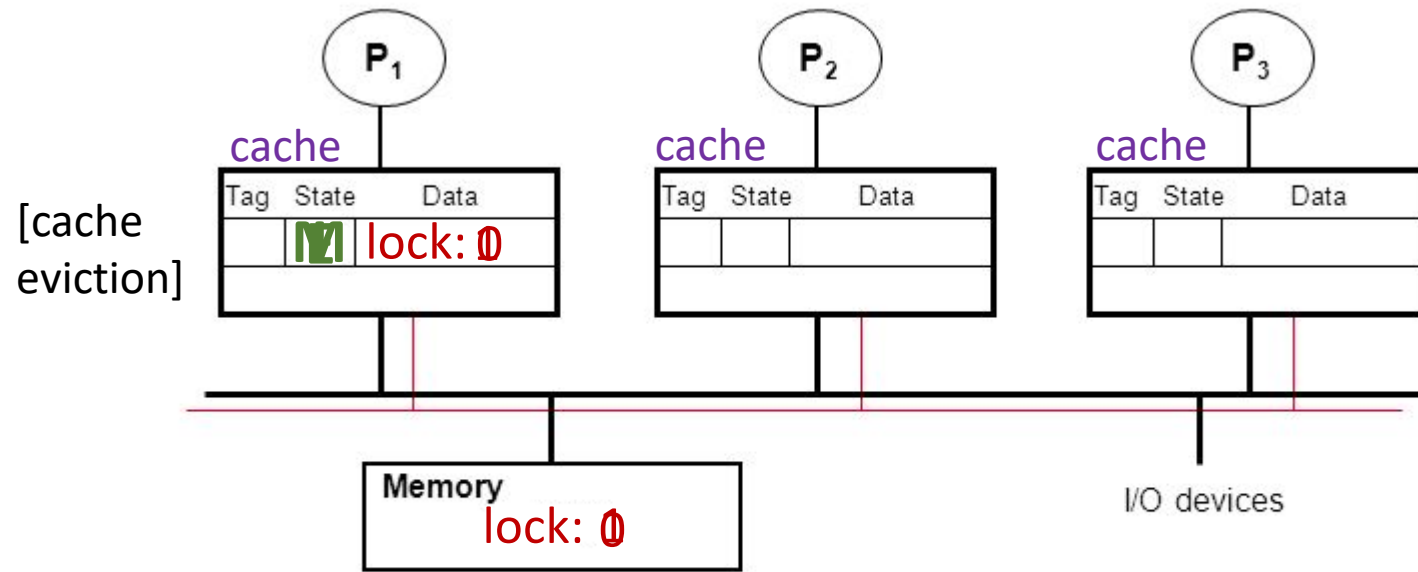
Multiprocessor Cache Coherence



Each cache line has a state (M, E, S, I)

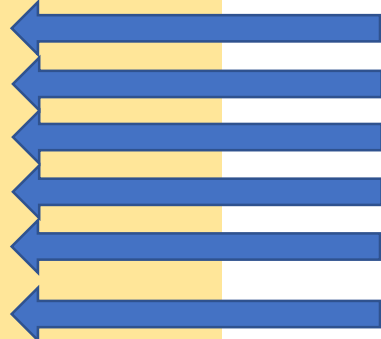
- Processors “snoop” bus to maintain states
- Initially → ‘I’ → Invalid
- Read one → ‘E’ → exclusive
- Reads → ‘S’ → multiple copies possible
- Write → ‘M’ → single copy → lots of cache coherence traffic

Cache Coherence: single-thread

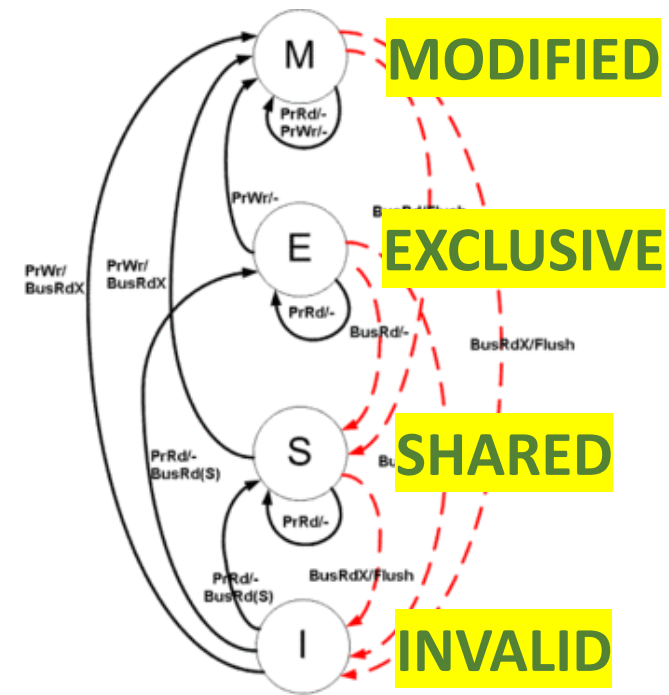
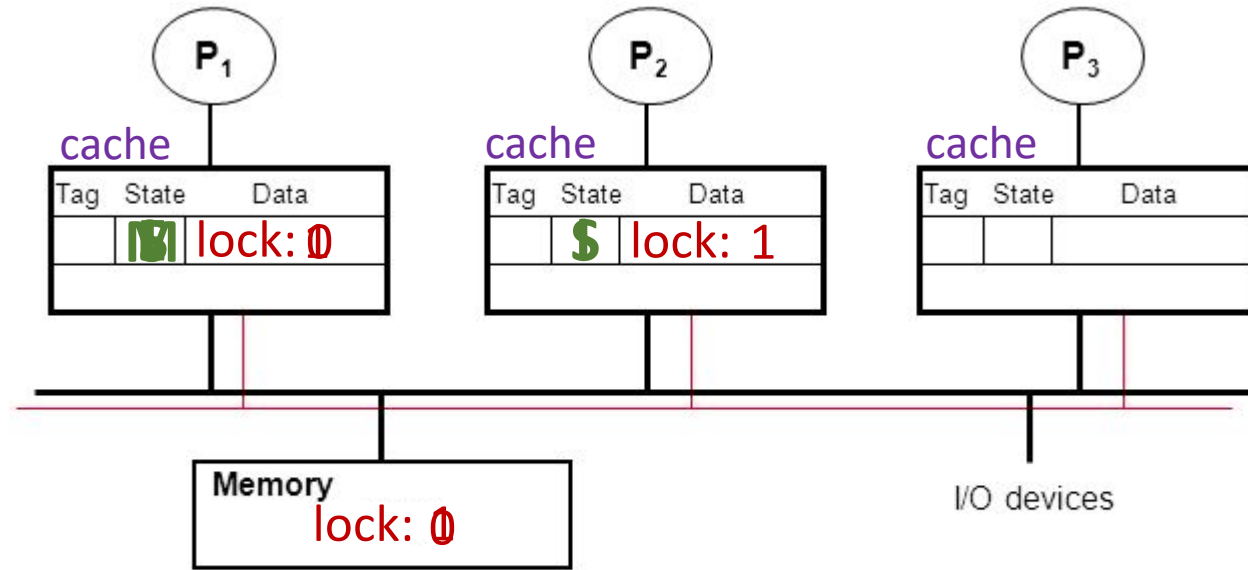


P1

```
// (straw-person lock impl)
// Initially, lock == 0 (unheld)
lock() {
try:  load lock, R0
      test R0
      bnz try
      store lock, 1
}
```



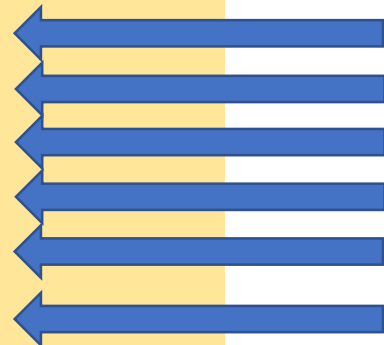
Cache Coherence Action Zone



P1

P2

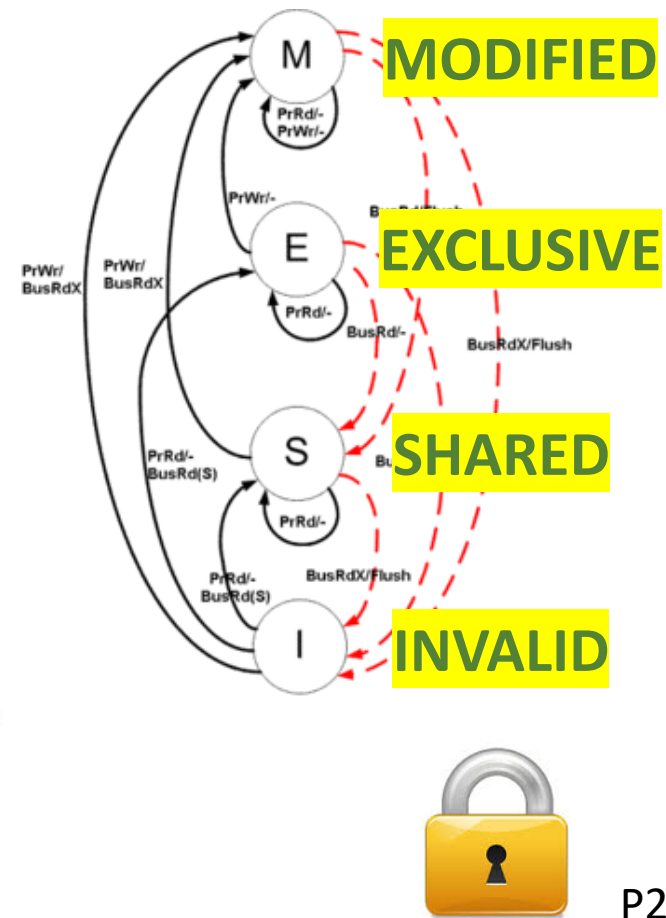
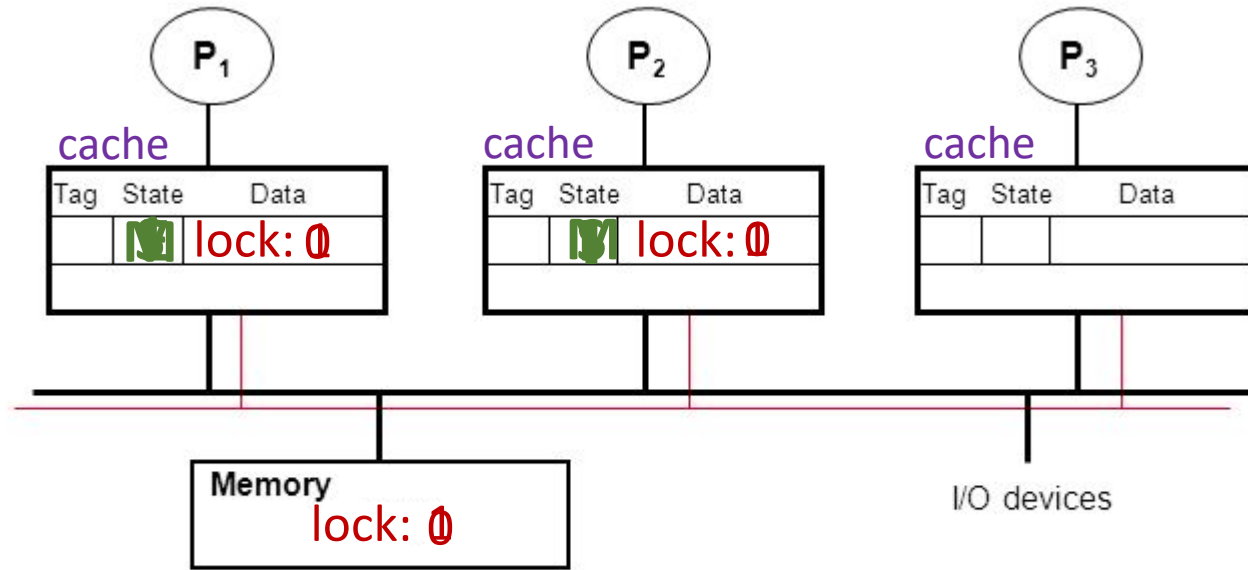
```
// (straw-person lock impl)
// Initially, lock == 0 (unheld)
lock() {
try:  load lock, R0
      test R0
      bnz try
      store lock, 1
}
```



SAFE!

```
// (straw-person lock impl)
// Initially, lock == 0 (unheld)
lock() {
try:  load lock, R0
      test R0
      bnz try
      store lock, 1
}
```

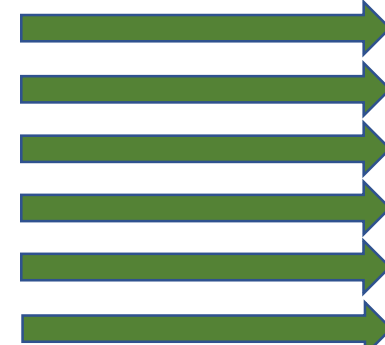
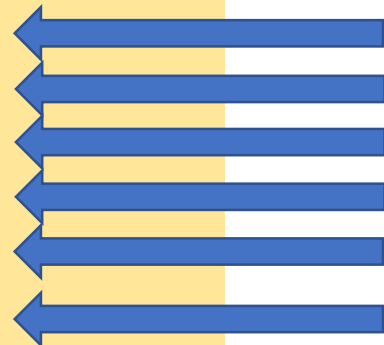
Cache Coherence Action Zone II



P1

P2

```
// (straw-person lock impl)
// Initially, lock == 0 (unheld)
lock() {
try:  load lock, R0
      test R0
      bnz try
      store lock, 1
}
```

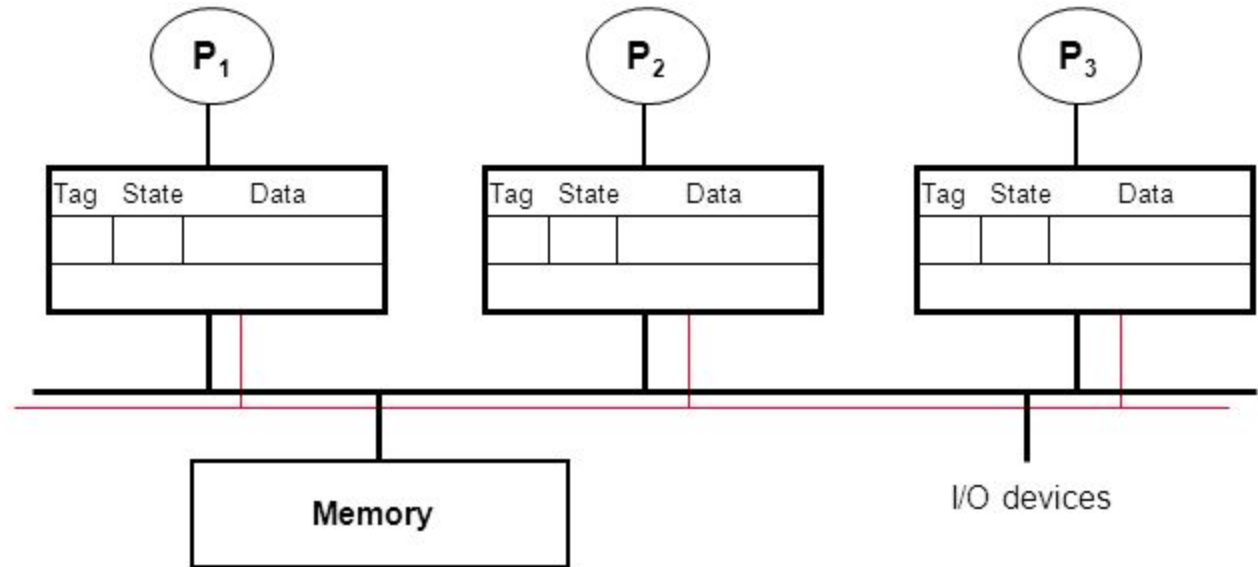


```
// (straw-person lock impl)
// Initially, lock == 0 (unheld)
lock() {
try:  load lock, R0
      test R0
      bnz try
      store lock, 1
}
```

Read-Modify-Write (RMW)

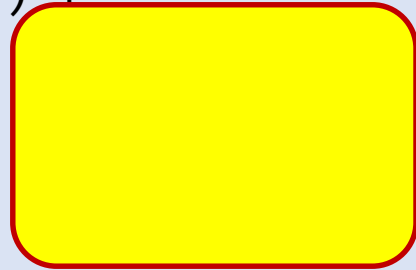
- ◆ Implementing locks requires read-modify-write operations
- ◆ Required effect is:
 - An atomic and isolated action
 1. read memory location **AND**
 2. write a new value to the location
 - RMW is *very tricky* in multi-processors
 - Cache coherence alone doesn't solve it

```
// (straw-person lock impl)
// Initially, lock == 0 (unheld)
lock() {
try:  load lock, R0
      test R0
      bnz try
      store lock, 1
}
```



Essence of HW-supported RMW

```
// (straw-person lock impl)
// Initially, lock == 0 (unheld)
lock() {
try:
}
}
```



Make this into a single
(atomic hardware instruction)

HW Support for Read-Modify-Write (RMW)

Test & Set	CAS	Exchange, locked increment/decrement,	LLSC: load-linked store-conditional
Most architectures	Many architectures	x86	PPC, Alpha, MIPS
<pre>int TST(addr) { atomic { ret = *addr; if(!*addr) *addr = 1; return ret; } }</pre>	<pre>bool cas(addr, old, new) { atomic { if(*addr == old) { *addr = new; return true; } return false; } }</pre>	<pre>int XCHG(addr, val) { atomic { ret = *addr; *addr = val; return ret; } }</pre>	<pre>bool LLSC(addr, val) { ret = *addr; atomic { if(*addr == ret) { *addr = val; return true; } } return false; }</pre>

```
void CAS_lock(lock) {
  while(CAS(&lock, 0, 1) != true);
}
```

HW Support for RMW: LL-SC

LLSC: load-linked store-conditional

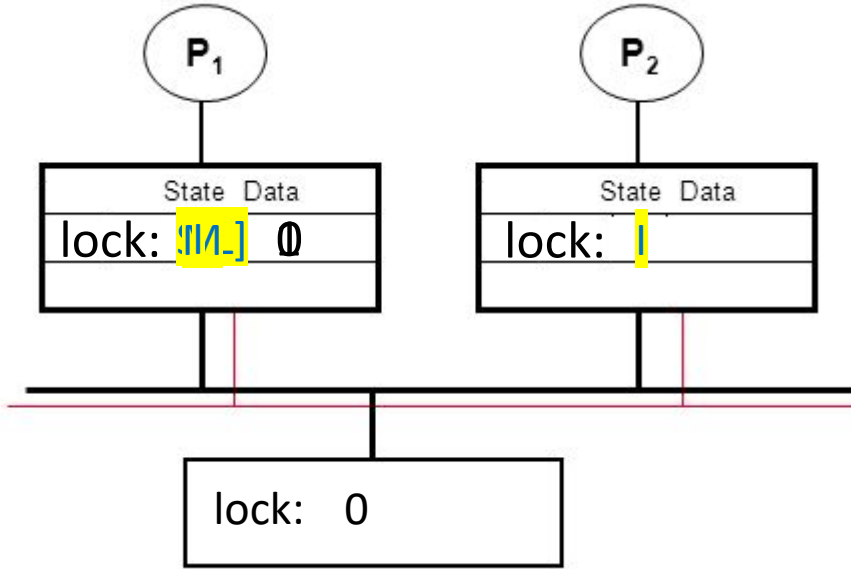
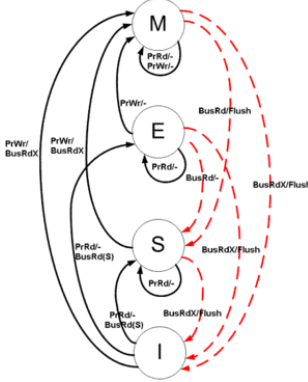
PPC, Alpha, MIPS

```
bool LLSC(addr, val) {
    ret = *addr;
    atomic {
        if(*addr == ret) {
            *addr = val;
            return true;
        }
        return false;
    }
}
```

```
void LLSC_lock(lock) {
    while(1) {
        old = load-linked(lock);
        if(old == 0 && store-cond(lock, 1))
            return;
    }
}
```

- load-linked is a load that is “linked” to a subsequent store-conditional
- Store-conditional only succeeds if value from linked-load is unchanged

LLSC Lock Action Zone



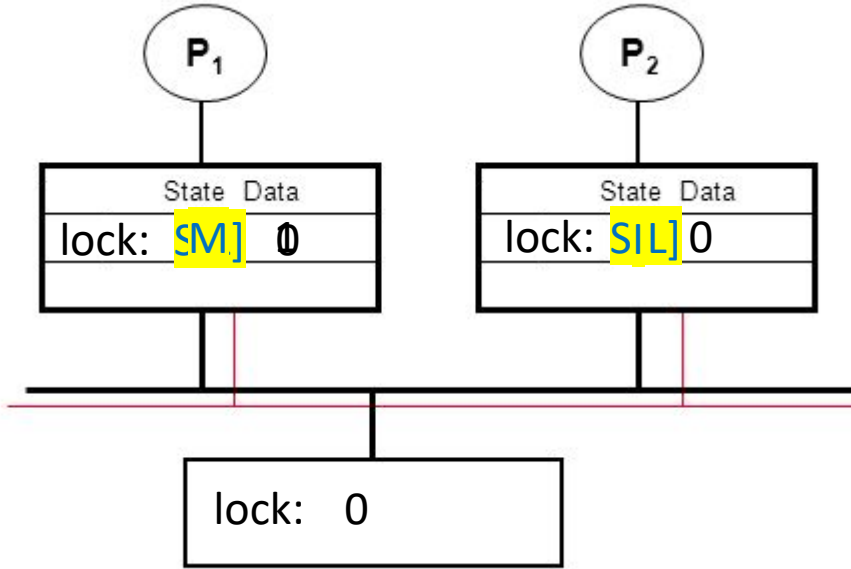
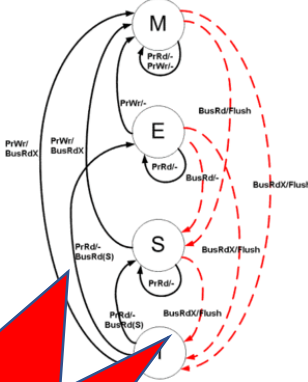
```

P1
lock(lock) {
  while(1) {
    old = ll(lock);
    if(old == 0)
      if(sc(lock, 1))
        return;
  }
}
    
```

```

P2
lock(lock) {
  while(1) {
    old = ll(lock);
    if(old == 0)
      if(sc(lock, 1))
        return;
  }
}
    
```

LLSC Lock Action Zone II



```

P1
lock(lock) {
  while(1) {
    old = ll(lock);
    if(old == 0)
      if(sc(lock, 1))
        return;
  }
}
    
```

```

P2
lock(lock) {
  while(1) {
    old = ll(lock);
    if(old == 0)
      if(sc(lock, 1))
        return;
  }
}
    
```

Implementing Locks with Test&set

```
int lock_value = 0;  
int* lock = &lock_value;
```

```
Lock::Acquire() {  
    while (test&set(lock) == 1)  
        ; //spin  
}
```

(test & set ~ CAS ~ LLSC)



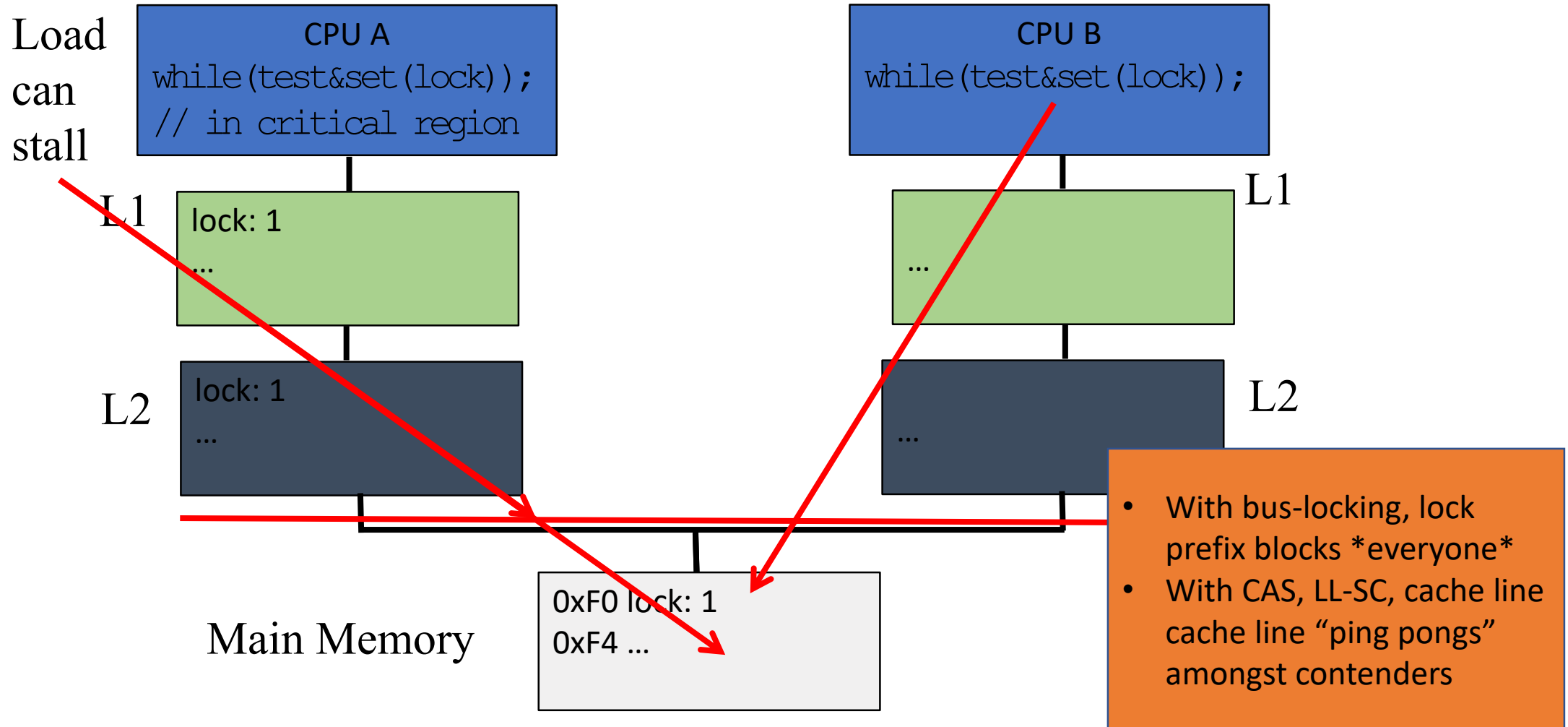
```
Lock::Release() {  
    *lock = 0;  
}
```

- ◆ What is the problem with this?
 - A. CPU usage B. Memory usage C. Lock::Acquire() latency
 - D. Memory bus usage E. Does not work

Test & Set with Memory Hierarchies

Initially, lock already held by some other CPU—A, B busy-waiting

What happens to lock variable's cache line when different cpu's contend?



TTS: Reducing busy wait contention

Test&Set

```
Lock::Acquire() {  
  while (test&set(lock) == 1);  
}
```

Busy-wait on in-memory copy

```
Lock::Release() {  
  *lock = 0;  
}
```

Test&Test&Set

```
Lock::Acquire() {  
  while(1) {  
    while (*lock == 1) ; // spin just reading  
    if (test&set(lock) == 0) break;  
  }  
}
```

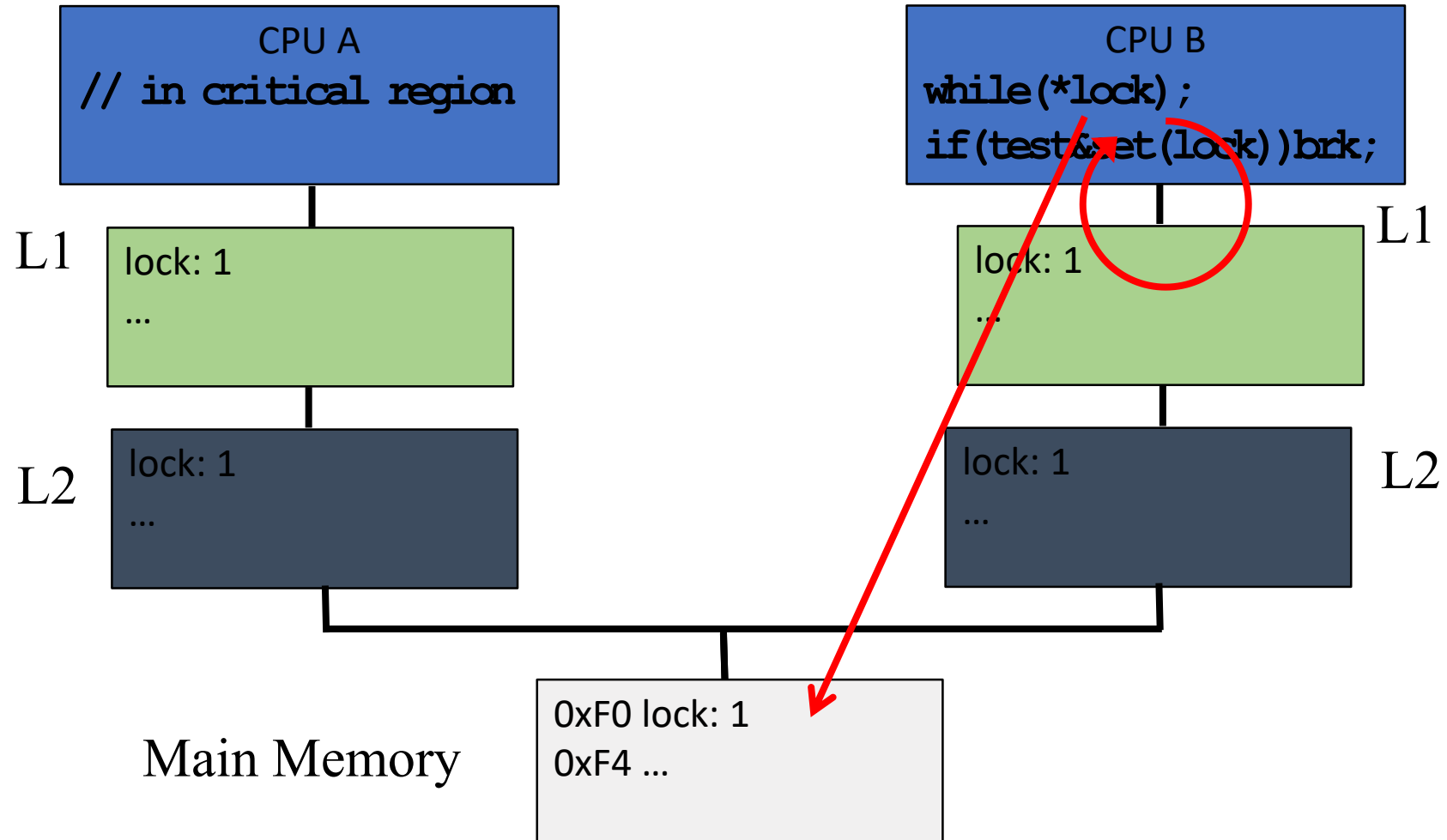
Busy-wait on cached copy

```
Lock::Release() {  
  *lock = 0;  
}
```

- What is the problem with this?
 - A. CPU usage B. Memory usage C. Lock::Acquire() latency
 - D. Memory bus usage E. Does not work

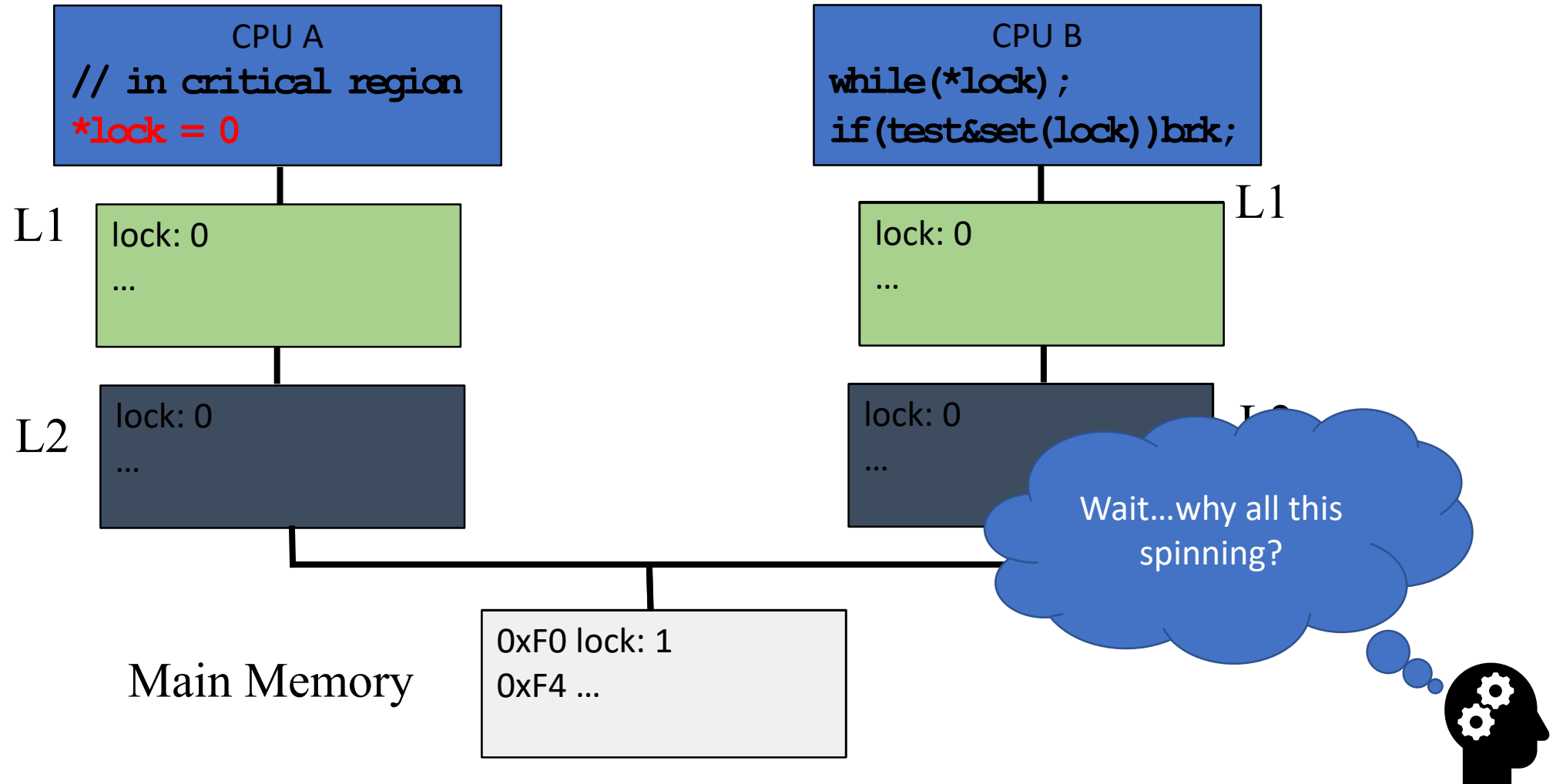
Test & Test & Set with Memory Hierarchies

What happens to lock variable's cache line when different cpu's contend for the same lock?



Test & Test & Set with Memory Hierarchies

What happens to lock variable's cache line when different cpu's contend for the same lock?



How can we improve over busy-wait?

```
Lock::Acquire() {  
  while(1) {  
    while (*lock == 1) ; // spin just reading  
    if (test&set(lock) == 0) break;  
  }  
}
```

Mutex

- Same abstraction as spinlock
- But is a “blocking” primitive
 - Lock available → same behavior
 - Lock held → yield/block
- Many ways to yield
- Simplest case of semaphore

```
void cm3_lock(u8_t* M) {
    u8_t LockedIn = 0;
    do {
        if (__LDREXB(Mutex) == 0) {
            // unlocked: try to obtain lock
            if (__STREXB(1, Mutex)) { // got lock
                __CLREXB(); // remove __LDREXB() lock
                LockedIn = 1;
            }
            else task_yield(); // give away cpu
        }
        else task_yield(); // give away cpu
    } while (!LockedIn);
}
```

- Is it better to use a spinlock or mutex on a uni-processor?
- Is it better to use a spinlock or mutex on a multi-processor?
- How do you choose between spinlock/mutex on a multi-processor?

Priority Inversion

A(prio-0) → enter(l);

B(prio-100) → enter(l); → must wait.

Solution?

Priority inheritance: A runs at B's priority

MARS pathfinder failure:

<http://wiki.csie.ncku.edu.tw/embedded/priority-inversion-on-Mars.pdf>

Other ideas?

Dekker's Algorithm

```

variables
  wants_to_enter : array of 2 booleans
  turn : integer

wants_to_enter[0] ← false
wants_to_enter[1] ← false
turn ← 0 // or 1
    
```

```

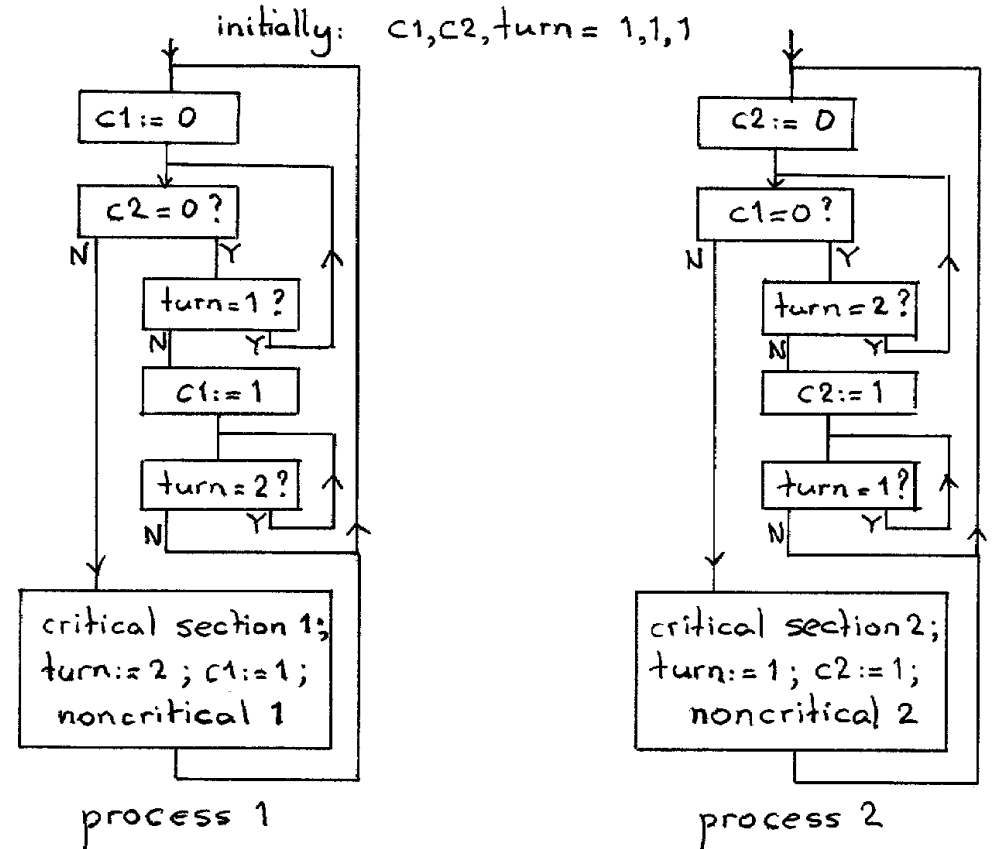
p0:
  wants_to_enter[0] ← true
  while wants_to_enter[1] {
    if turn ≠ 0 {
      wants_to_enter[0] ← false
      while turn ≠ 0 {
        // busy wait
      }
      wants_to_enter[0] ← true
    }
  }

  // critical section
  ...
  turn ← 1
  wants_to_enter[0] ← false
  // remainder section
    
```

```

p1:
  wants_to_enter[1] ← true
  while wants_to_enter[0] {
    if turn ≠ 1 {
      wants_to_enter[1] ← false
      while turn ≠ 1 {
        // busy wait
      }
      wants_to_enter[1] ← true
    }
  }

  // critical section
  ...
  turn ← 0
  wants_to_enter[1] ← false
  // remainder section
    
```



Th. J. Dekker's Solution

Lab #1

- Basic synchronization
- <http://www.cs.utexas.edu/~rossbach/cs378/lab/lab0.html>
- ***Start early!!!***

Questions?