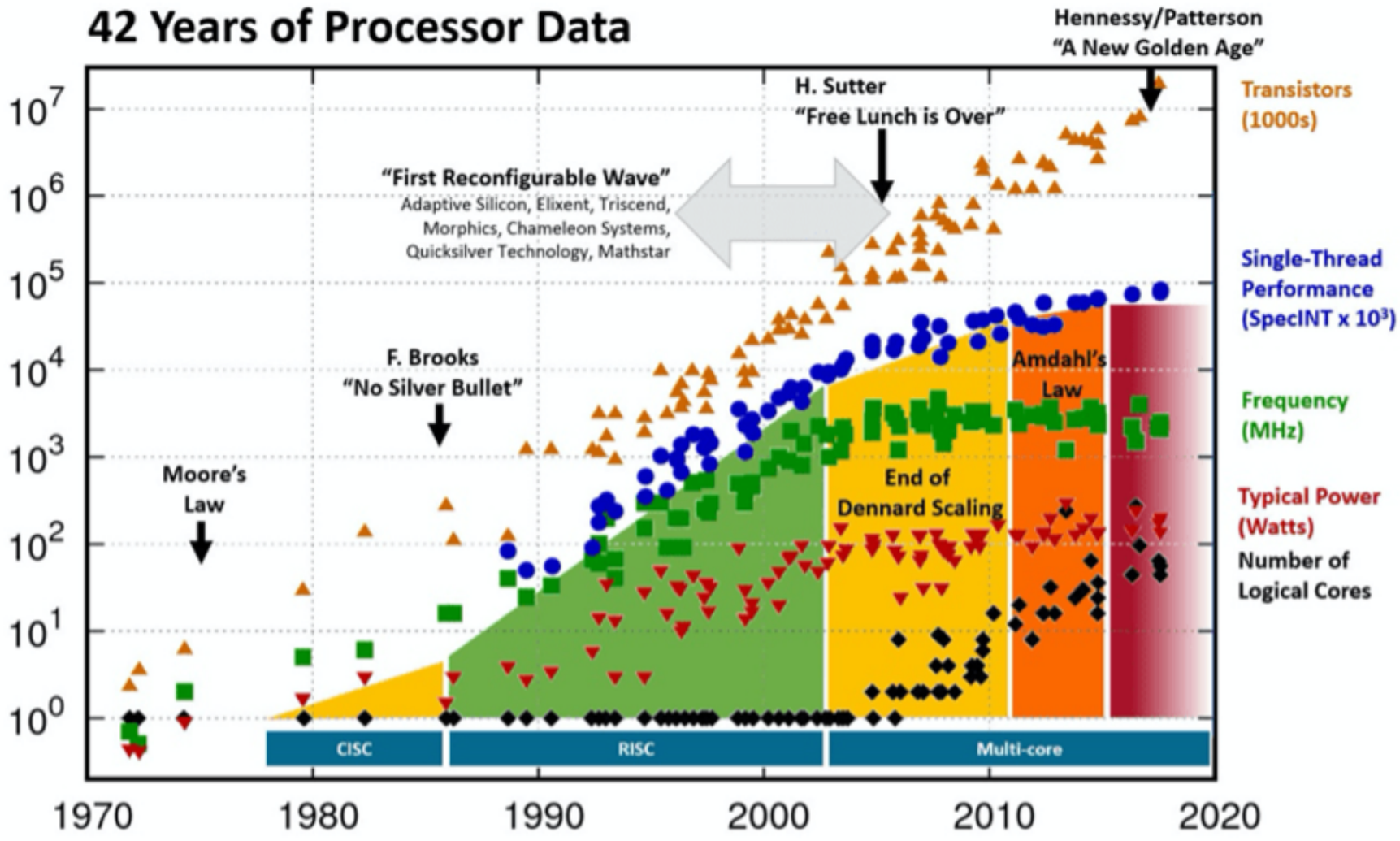
互斥和同步

Section 4: Part I

2026

多线程编程

"美好的" 单处理器时代 (Moore's Law) 已经过去了

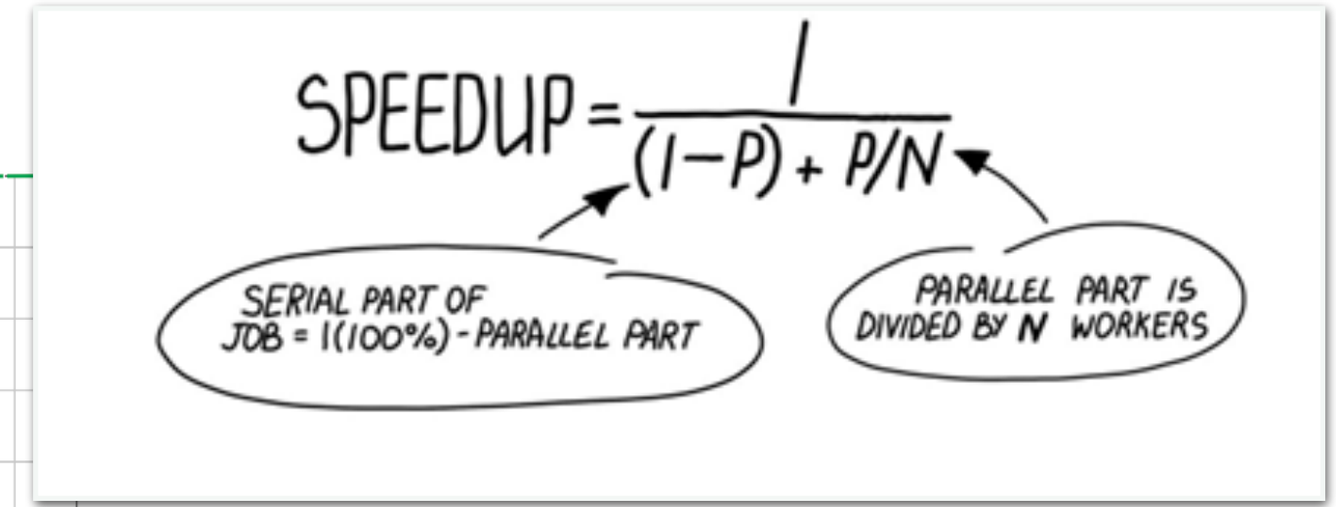


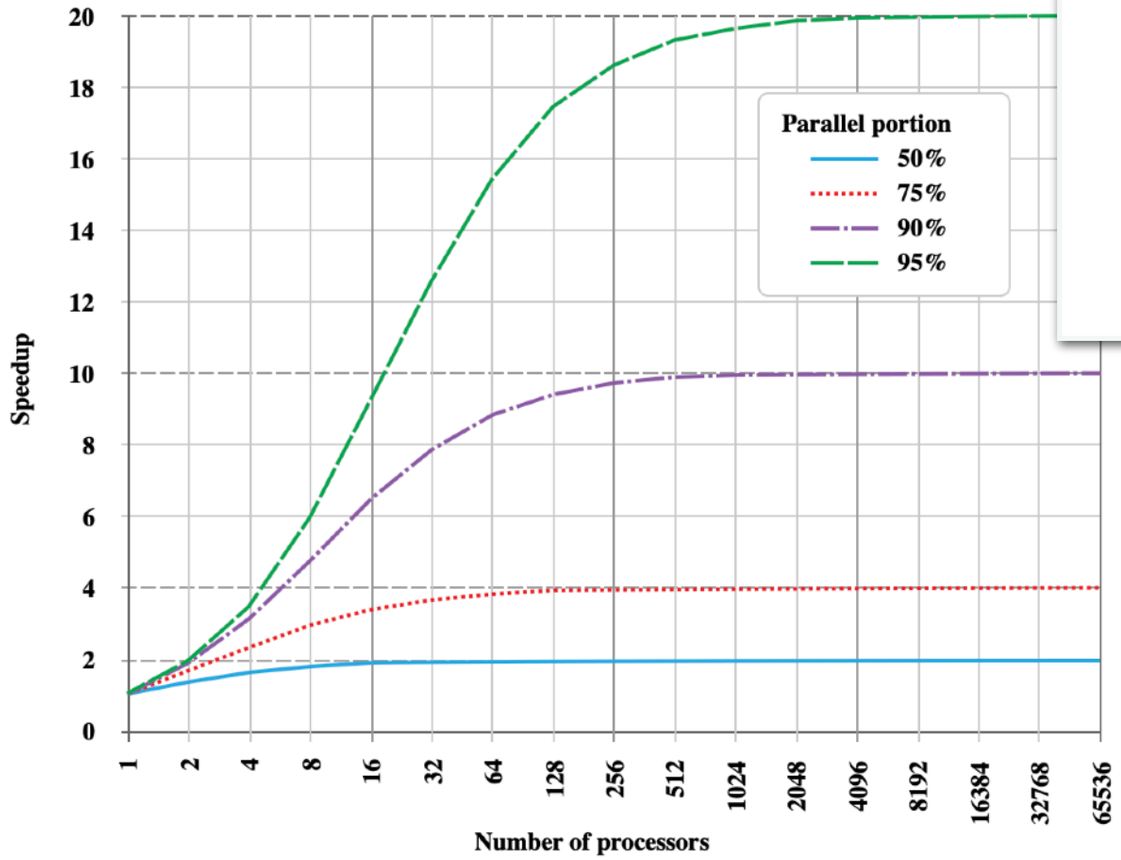
Hennessy and Patterson, Turing Lecture 2018, overlaid over "42 Years of Processors Data"
https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/;"First Wave" added by Les Wilson, Frank Schirrmeister
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

多线程编程

如果单核性能无法大幅提高，只能依赖多核来并行计算

- 进程 (Process) 和线程 (Thread) 已提供了实现 Multitasking 的能力
- 此时，一个计算任务的加速极限受限于任务中无法并行化的部分 (Amdahl's Law)





多线程编程

但人类并不是擅长处理 Multitasking 的生物 😵‍💫

Human multitasking

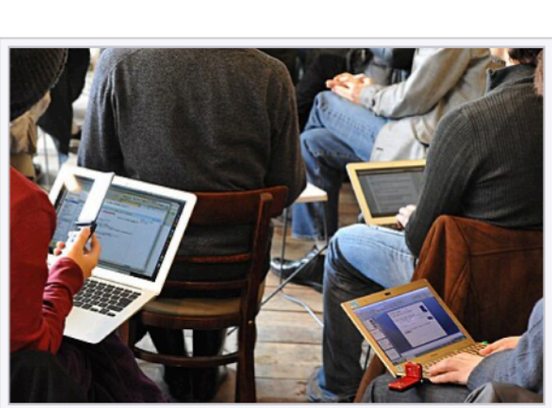
18 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Multitasking (disambiguation).

Human multitasking is the concept that one can split their attention on more than one task or activity at the same time, such as speaking on the phone while driving a car.



Laptop and mobile phone

Multitasking can result in time wasted due to human context switching (e.g., determining which step is next in the task just switched to) and becoming prone to errors due to insufficient attention. Some people may be proficient at the tasks in question and also be able to rapidly shift attention between the tasks, and therefore perform the tasks well; however, self-perception of being good at multitasking or getting more done while multitasking is frequently inaccurate.[1][2]

Multitasking is mentally and physically stressful for everyone,[3] to the point that multitasking is used in laboratory experiments to study stressful environments.[4] Research suggests that people who are multitasking in a learning environment are worse at learning new information compared to those who do not have their attention divided among different tasks.[5][6][7]

多线程编程

如下程序 (创建了两个线程) 可能的输出结果是什么?

```
void* T1(void *arg) {
  while(1) { printf("A\n"); }
}

void* T2(void *arg) {
  while(1) { printf("B\n"); }
}

int main() {
  pthread_t t1, t2;
  pthread_create(&t1, NULL, T1, NULL);
  pthread_create(&t2, NULL, T2, NULL);

  pthread_join(t1, NULL)
  pthread_join(t2, NULL)
  return 0;
}
```

多线程编程

我们不能假设多线程程序的执行遵循与单线程程序类似的行为 (即按某种结构化方式顺序执行)

- 在单核 CPU 上，不同线程的指令会交叠 (interleaved) 在一起执行
  - 任何时刻都有可能产生 Context Switch
  - 调度器的行为是不确定的
  - 即使调度器实现了某种 “确定” 的算法，但影响因素太多 (e.g., unexpected interrupts, different CPU frequencies, cache hit rate, …)
- 对于多核 CPU，多个线程的指令根本就是并行执行的
- 此时，即使我们已经很小心地编程，但还是会出很多问题

多线程编程: 从入门到放弃

無 Atomicity

原子性 (Atomicity): 一个操作要么根本就没有执行、要么按照预期执行完毕，不会处于任何中间状态 (All or nothing)

- 当多个线程并发对一个共享数据进行操作时，会存在竞争同一个数据的情况 (Data Race)
- 最终结果取决于操作的具体执行顺序
- 如果 Context Switch 发生的不巧，则会产生预料之外的结果

多线程编程: 从入门到放弃

無 Atomicity

如果多个线程同时执行 `pay(100)`?

```
unsigned long balance = 100;

void pay(int amount) {
  if (balance >= amount) {
    balance -= amount;
  }
}
```

Thread 1 在执行完 `if` 但还没更新 `balance` 时产生 *context switch* 切换到 *Thread 2* 执行 ...

多线程编程: 从入门到放弃

無 Atomicity

如果多个线程同时对全局变量 `count` 进行更新?

```
#define NUM 1000000
int count = 0;

void* func(void *arg) {
  for (int i = 0; i < NUM; i++) {
    count++;
  }
  return NULL;
}
```

从高级程序语言看，只有一行代码；

但实际上，是三条指令 (不同线程的指令可能会被交叠在一起执行)

```
mov $count, %eax
add 1, %eax
mov %eax, $count
```

多线程编程: 从入门到放弃

無 Atomicity

如果多个线程同时对全局变量 `count` 进行更新?

Thread 1

```
mov $count, %eax
add 1, %eax
mov %eax, $count
```

Thread 2

```
mov $count, %eax
add 1, %eax
mov %eax, $count
```

`Thread 1` 和 `Thread 2` 都分别做了一次 `count++`

此时 `count += 2`

Thread 1

```
mov $count, %eax
add 1, %eax


mov %eax, $count
```

Thread 2

```
mov $count, %eax
add 1, %eax
mov %eax, $count
```

此时 `count += 1`

多线程编程: 从入门到放弃

無 Atomicity

如果多个线程同时对全局变量 `count` 进行更新?

```
#define NUM 1000000
int count = 0;

void* func(void *arg) {
  for (int i = 0; i < NUM; i++) {
    asm volatile("incq %0" : "+m" (count));
  }
  return NULL;
}
```

把 `count++` 变成一条指令就正确了吗?

多线程编程: 从入门到放弃

無 Sequential Order

顺序性 (In-order): 程序按照代码语句编写的顺序执行

- 只要不影响语义，指令是否按照顺序执行并不重要
  - 编译器会通过改写指令来提高程序执行速度
  - 这些优化在单线程下往往没有问题 (safe)，但在多线程下就未必

多线程编程: 从入门到放弃

無 Sequential Order

如果使用 -O1 和 -O2 选项来编译执行下列代码?

```
#define NUM 1000000
int count = 0;

void* func(void *arg) {
  for (int i = 0; i < NUM; i++) {
    count++;
   }
  return NULL;
}
```

-O1 ➡️ 1000000

count 被移出循环，最后又被加回来

-O2 ➡️ 2000000 看起来结果对了，但是真的吗?

直接优化为 count += 1000000

多线程编程: 从入门到放弃

無 Sequential Order

如果我们想让线程 `T1` 等待线程 `T2` (实现一种互相等待)?

```
int flag = 0;
int x = 0;

void *T1(void* arg) {
  while (flag == 0)
    ;
  printf("%d\n", x); // use x
  return NULL;
}

void *T2(void* arg) {
  x = 10;              // prepare x
  flag = 1;
  return NULL;
}
```

在 `-O2` 下会被优化为

```
if (flag == 0)
  while(1) ;
```

多线程编程: 从入门到放弃 *

無 Consistency *

内存一致性模型 (Memory Consistency Model) 定义了不同处理器对于共享内存操作需要遵循的顺序 (how multiple threads see the world)

- 现代处理器往往允许指令乱序执行 (同样是为了优化性能)
  - 例如，对于高时延的访存指令 (e.g., cache miss)，处理器可以选择调度后续其他指令执行，从而缓解访存操作的时延
  - 这种乱序会导致多个处理器看到不一致的访存顺序
- 内存一致性模型就是一种硬件和软件之间的约定

多线程编程: 从入门到放弃 *

🈚 Consistency *

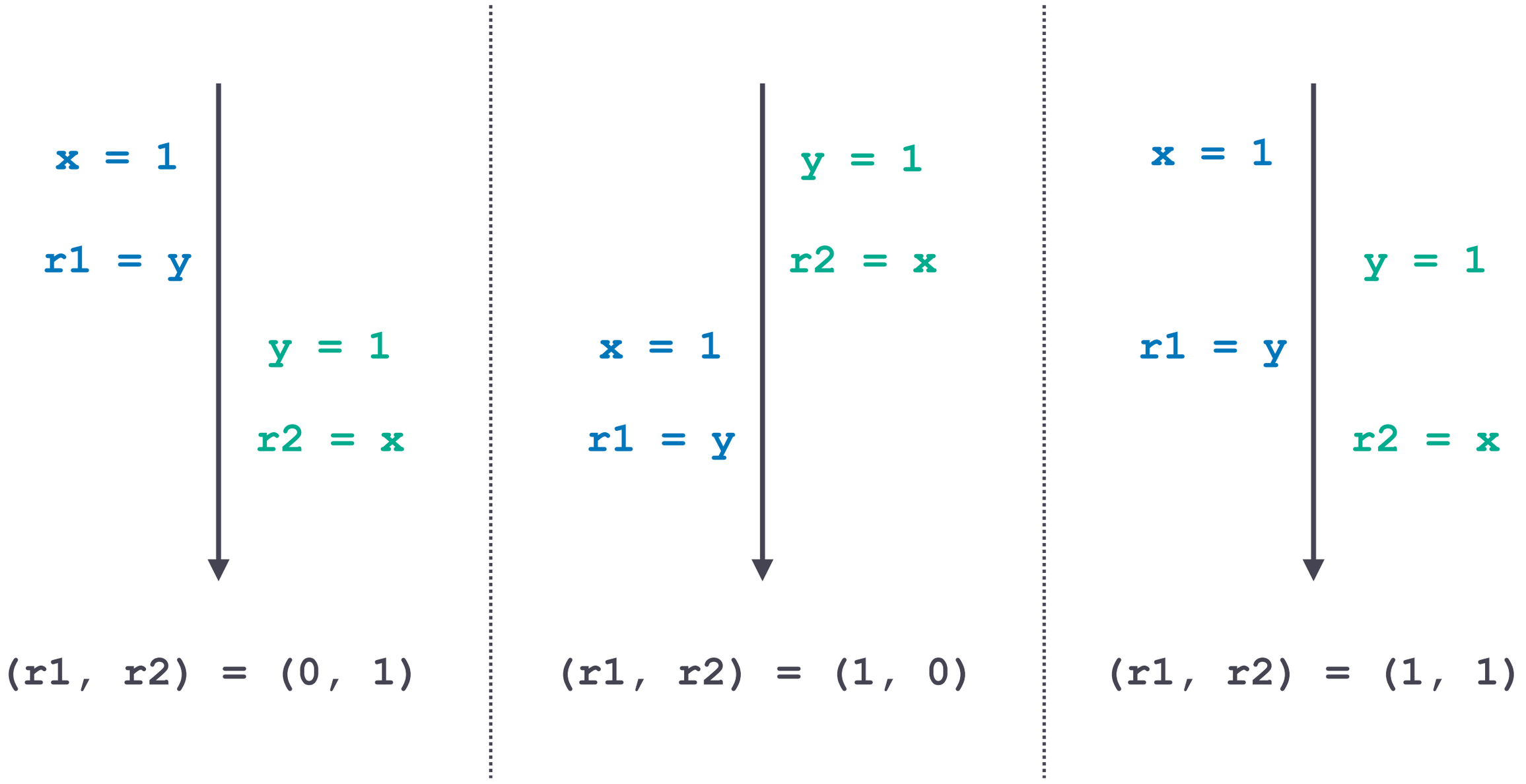
线程 `T1` 和 `T2` 运行结束后 `r1` 和 `r2` 的值分别可能是多少?

```
volatile int x = 0, y = 0;
volatile int r1, r2;

void* T1(void *arg) {
  x = 1;   // store(x)
  r1 = y;  // load(y)
}

void* T2(void *arg) {
  y = 1;   // store(y)
  r2 = x;  // load(x)
}
```

多线程编程: 从入门到放弃 *

無 Consistency *

线程 `T1` 和 `T2` 运行结束后 `r1` 和 `r2` 的值分别可能是多少?

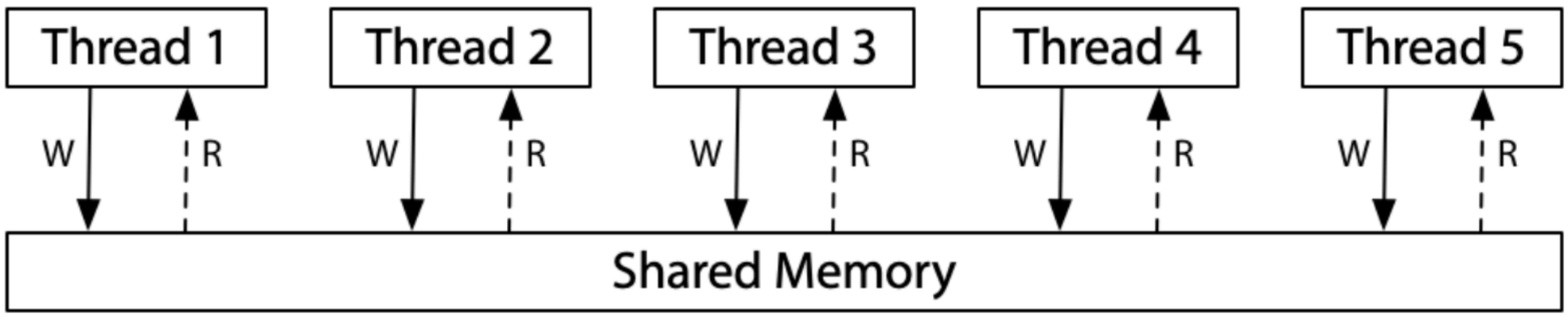


多线程编程: 从入门到放弃 *

無 Consistency *

Sequential Consistency (an intuitive model of parallelism)

- The result of any execution is the same as if the operations of all the processors were executed in some sequential order, and the operations of each individual processor appear in this sequence in the order specified by its program (everything must happen in order)



Every time a processor needs to read from or write to memory, that request goes to the shared memory

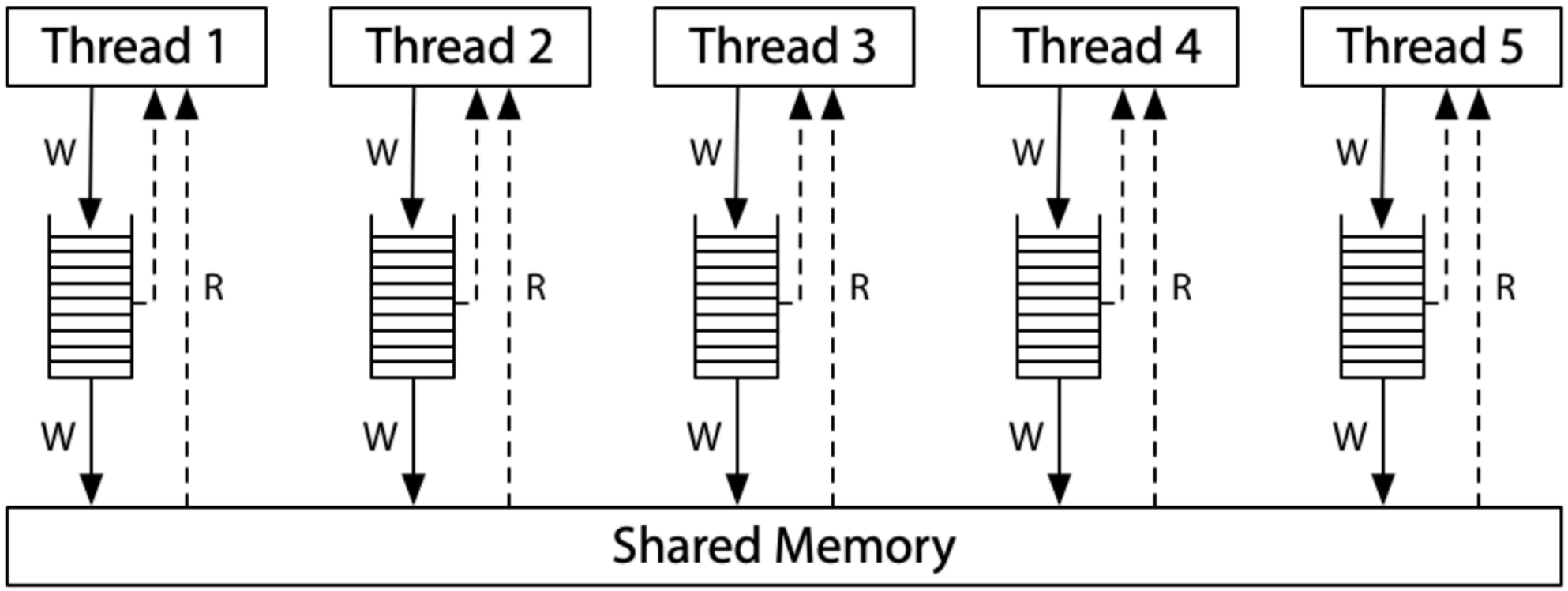
https://research.swtch.com/hwmm

多线程编程: 从入门到放弃 *

無 Consistency *

Total Store Order (x86)

- A popular memory model that allows store buffering



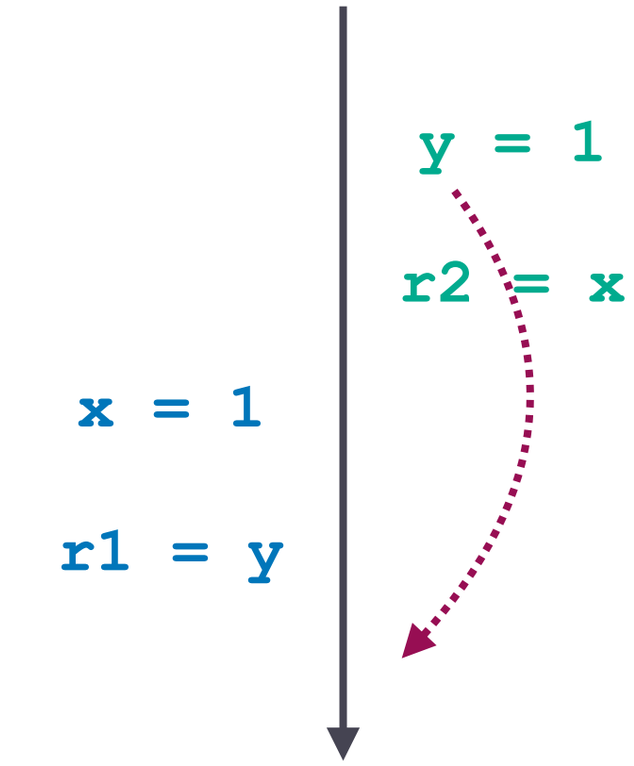
Each processor queues writes to that memory in a local write queue

https://research.swtch.com/hwmm

多线程编程: 从入门到放弃 *

無 Consistency *

线程 `T1` 和 `T2` 运行结束后 `r1` 和 `r2` 的值分别可能是多少?



Both threads could queue their writes and then read from memory before either write makes it to memory.

`(r1, r2) = (0, 0)`

多线程编程: 从入门到放弃

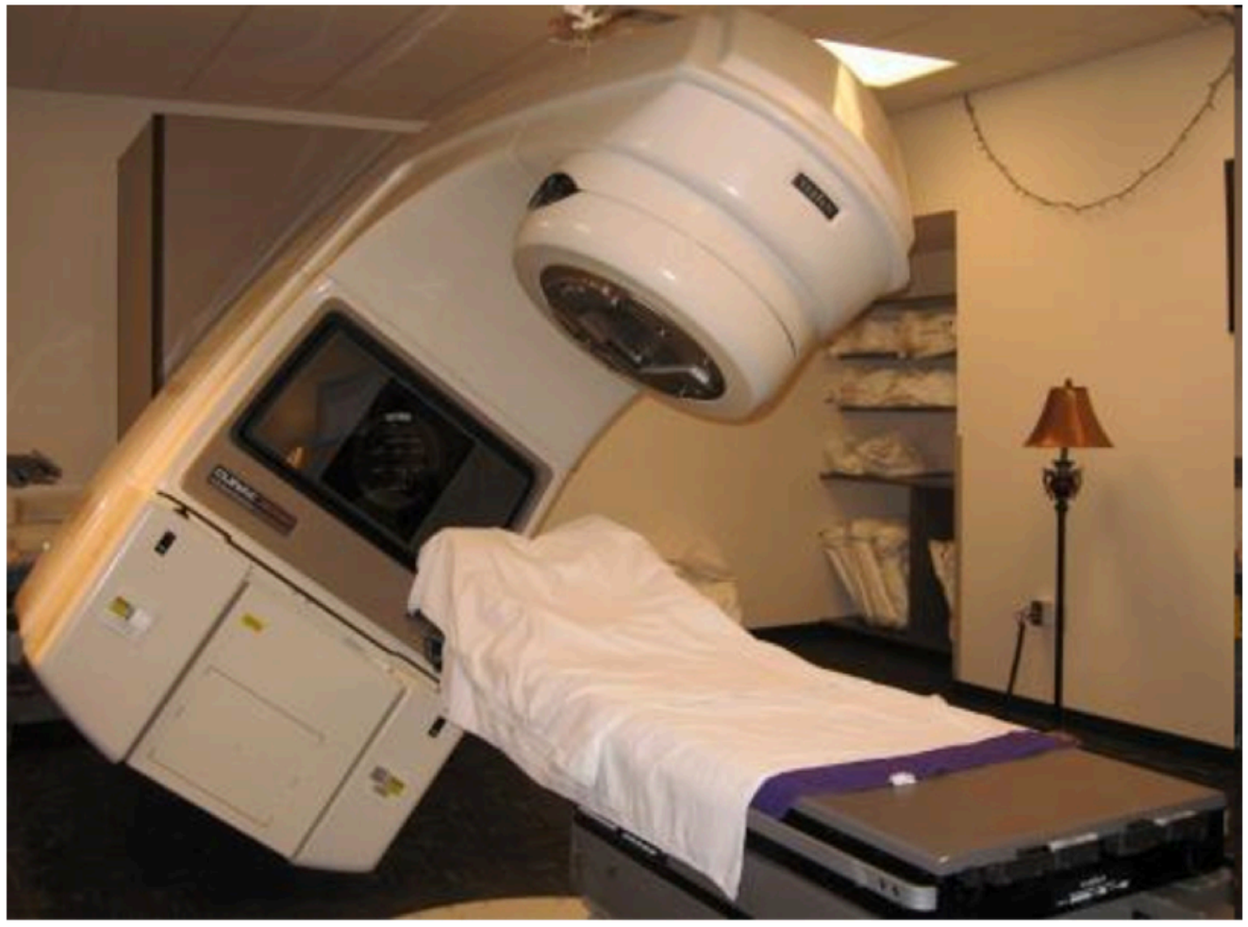
多线程并发程序很难写对，而且也很难检测其中的并发 Bugs
(the failure might be improbable to trigger, but possible)

- 多线程指令能以很多不同的方式交叠执行
- 其中只有部分交叠执行顺序会触发 Bugs (程序的非确定性行为)
  - 同样的输入在不同次执行下会产生不同的结果
  - 即使运行很久没有出问题也不代表并发程序就一定是对的
- 还需要了解编译器、以及底层硬件的很多细节，才能尝试去理解并发程序的真实表现行为

经典的并发 Bugs

Therac-25 放射线治疗仪 (1980s)

- 由于系统运行过程中的一个 race condition，导致至少六名患者因收到过量辐射而死亡或严重受伤
- Therac-25 之前的机型通过一个硬件锁来避免此情况，但在 Therac-25 中改由软件来进行检查和处理



https://en.wikipedia.org/wiki/Therac-25

经典的并发 Bugs

北美大停电 (2023)

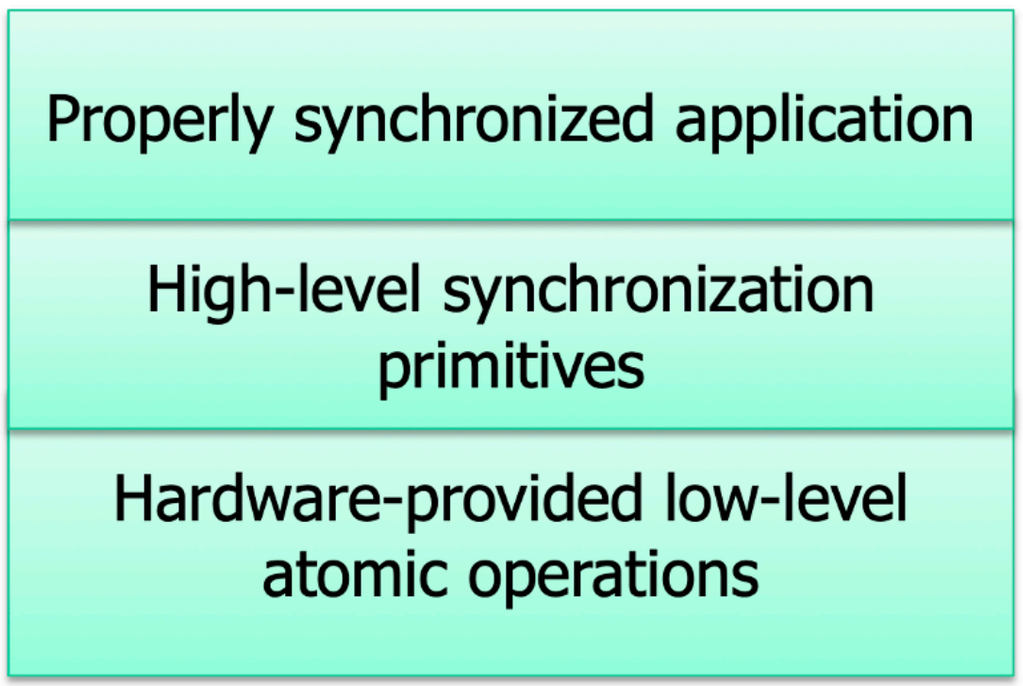
- 约有 5500 万人受到影响，经济损失估计 250～300 亿美元
- 监控系统中的一个 race conditon 部分导致警报系统失效
- 系统代码自 1990 年开始运行，在持续运行超过 300 万小时中从未出现任何 bug

https://en.wikipedia.org/wiki/Northeast_blackout_of_2003

互斥和同步

我们需要特定的机制来重新获得多线程执行的某种确定性 (determinism)

- 协调多个线程以达成某种一致 (某种特定的指令执行序列)
- 一种分层的实现方式
  - 由硬件提供一些必要的原子性指令
  - 在此基础上，构建一些 synchronization primitives



互斥和同步

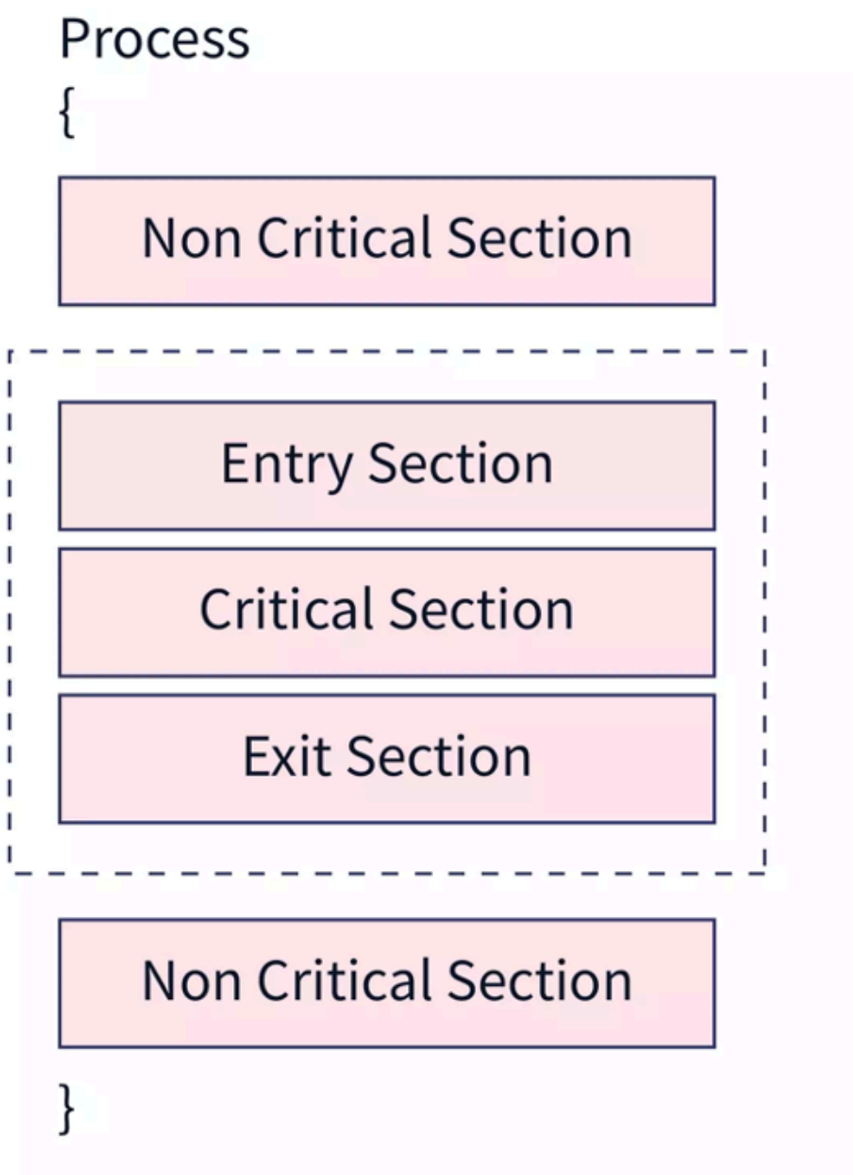
为什么我们要在操作系统课程里学习这个主题？

- 因为所有操纵系统教材里都有这一章 😵
- 操作系统就是一个重要的并发程序
  - 内核的很多内部数据结构 (例如 PID、进程列表、页表、文件系统结构等) 都存在数据竞争
- 解决并发的很多技术都源自于操作系统的设计需求及其相应的解决方案

Locks

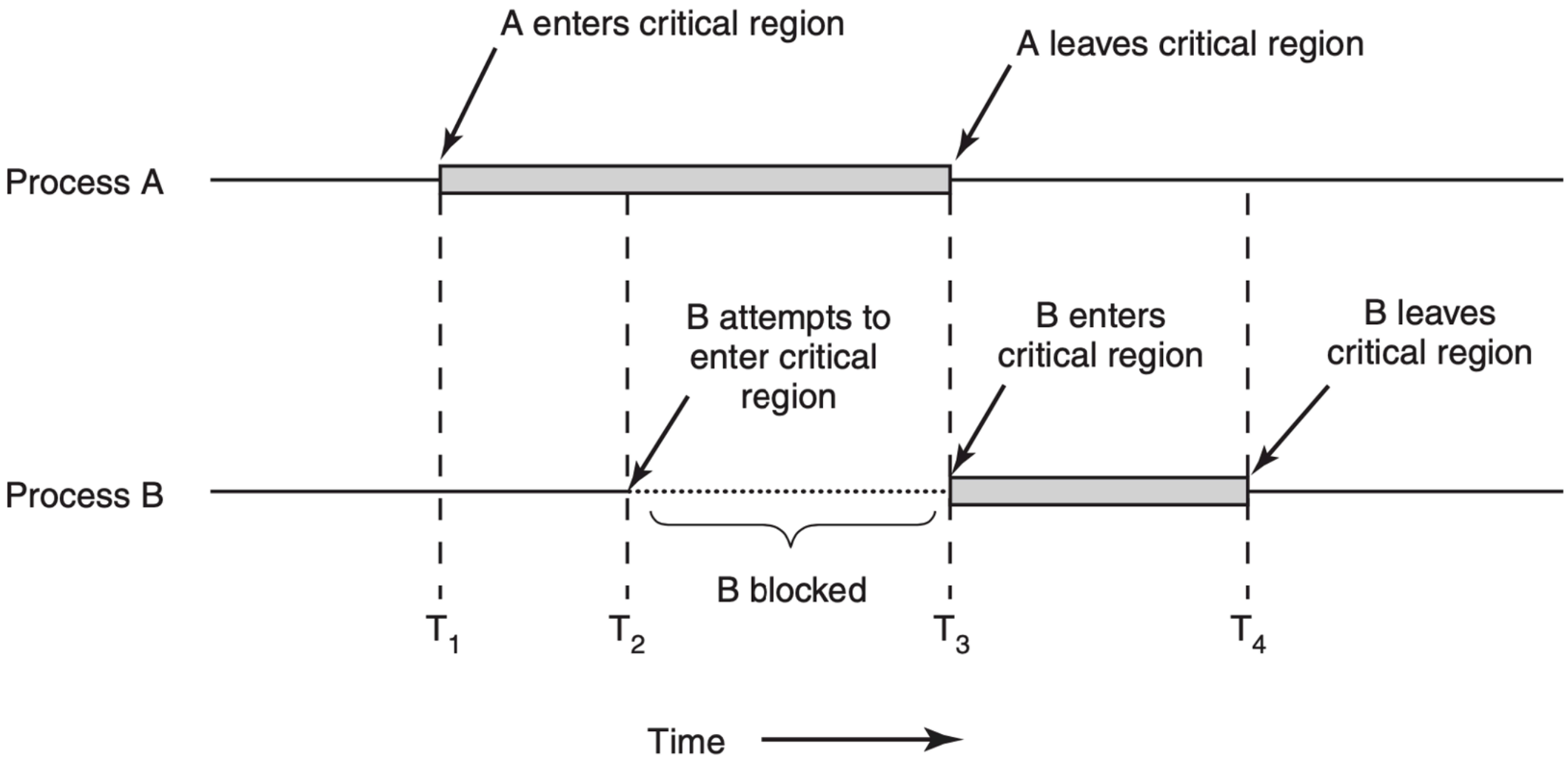
Critical Section

使用临界区 (Critical Section) 来刻画访问共享资源的一段代码



Critical Section

互斥 (Mutual Exclusion): 如果一个线程在临界区内执行，则其他线程将被阻止进入临界区 (确保临界区中的代码以某种方式顺序执行)



Locks

一个锁 (lock) 就是一个变量，其记录了锁当前的状态 (available 或者 acquired)

🔒 **`lock()`**

- 如果没有其它线程持有这个锁 (lock is available)，则执行该操作的线程获得锁，并进入临界区
- 如果其它线程持有这个锁 (lock is acquired)，则调用不会返回

🔓 **`unlock()`**

- 当持有锁的线程执行该操作时，锁变为 available 状态
- 如果此时有正在等待的线程，则其中一个会感知到此时锁状态的变化，从而获得锁并进入临界区

Locks

有了锁之后，程序员就能实现互斥了 (获得对并发程序最基本的控制)

- 通过给代码片段加锁，确保该段代码执行的原子性
- 但并发程序的正确性仍取决于 "有没有正确使用锁"

```
#define NUM 1000000
int count = 0;

void* func(void *arg) {
  for (int i = 0; i < NUM; i++) {
    count++;
   }
  return NULL;
}
```

在哪里加锁? 使用几个锁? 可不可以不使用锁?

How to build a lock

实现临界区需要满足的条件

- 临界区内最多只能有一个线程执行 (Mutual Exclusion)
- 如果一个线程在临界区外，则该线程不能阻止其它线程进入临界区 (Progress / Liveness)
- 如果一个线程正在等待进入临界区，则该线程最终一定有机会进入 (Bounded Waiting / Fairness)
- 此外，进入临界区和退出临界区这两个操作对程序运行造成的额外开销应尽可能小 (Performance)

How to build a lock

Disabling Interrupts

尝试 1️⃣ : 在进入临界区的时候关中断

- 临界区内线程的执行不会被打断
- 相应地，共享资源就不会被其它线程所访问

```
void lock() {
  DisableInterrupts();
}

void unlock() {
  EnableInterrupts();
}
```

How to build a lock

Disabling Interrupts

尝试 1️⃣ : 在进入临界区的时候关中断

- 如果允许用户程序执行关中断指令？
  - 关中断是一个特权指令 (privileged instruction)
  - 但在操作系统内核中，使用关中断是一个常见的操作
- 如果临界区代码死循环 (操作系统无法重新获得控制权)？
- 中断关闭时间过长会导致外部的重要事件丢失？
- 如果本身就是多处理器系统？
  - 中断是每个处理器内部状态 (每个处理器有独立的寄存器组)

How to build a lock

Just Use Loads & Stores

尝试 2️⃣ : 使用一个 `flag` 变量来记录当前锁的状态
(假设 load 和 store 操作是原子的)

```
// 0 -> lock is available, 1 -> held
flag = 0

void lock() {
  while(flag == 1) // TEST the flag
    ;
  flag = 1;        // now SET it
}

void unlock() {
  flag = 0;
}
```

What could possibly go wrong?
(pretend you are a malicious scheduler)

How to build a lock

Just Use Loads & Stores

尝试 2️⃣ : 使用一个 `flag` 变量来记录当前锁的状态
(假设 load 和 store 操作是原子的)

```
        // initially, flag = 0

┆       Thread 1                    Thread 2

        // lock()                     // lock()
        while(flag == 1);
                                      while(flag == 1);
                                      flag = 1;
        flag = 1;
        // critical section           // critical section
```

How to build a lock

Just Use Loads & Stores

尝试 2️⃣：使用一个 `flag` 变量来记录当前锁的状态
(假设 load 和 store 操作是原子的)

- 存在两个线程同时将 `flag` 置为 `1` 的情况，即同时进入临界区
  (不满足互斥需求)

- 关键问题在于对 `flag` 的 *TEST & SET* 不是原子操作
  - *TEST:* 看一下当前是什么状态，但不知道后面会变成什么样
  - *SET:* 对状态做更新，但不知道后面会不会又被修改

How to build a lock

Just Use Loads & Stores

尝试 3️⃣：使用一个 `flag` 变量来标记当前轮到谁进入临界区

```
// assume two threads: 0 and 1
flag = 0;

void lock() {
  // wait for my turn
  while(flag == 1 - self)
    ;
}

void unlock() {
  // I am done, your turn
  flag = 1 - self;
}
```

How to build a lock

Just Use Loads & Stores

尝试 3️⃣：使用一个 `flag` 变量来标记当前轮到谁进入临界区

- 一种通过严格轮转 (strict alternation) 来实现互斥的方法
- 但是，一个线程能不能进入临界区取决于另一个线程是否进入过临界区 (不满足 Progress 性质)

- `T0` enters its critical section and exits; sets `turn = 1`; then executes a long time non-critical procedure
- `T1` enters its critical section and exits; sets `turn = 0`; then tries to enter its critical section again
- Now, `T0` is in its non-critical section, and `T1` is waiting for `turn` to become `0`

Peterson's Algorithm

在 1960s 人们尝试了很多纯软件方法来实现互斥，但都是错的

- 直到 Dijkstra 的一个数学家朋友 Dekker 给出了第一个正确的算法
- Peterson 在 1989 年对该算法进行了改进和简化 (结合 flag variable 和 strict alternation 两种思想)

The original solution due to Dekker is discussed at length by Dijkstra in [1]. Of the many reformulations given since, perhaps the best appears in [3]. (Unfortunately the authors believe their correct solution is incorrect.) The solutions of Doran and Thomas are slight improvements which eliminate the 'loop inside a loop' structure of the previously published solutions. The solution presented here has an extremely simple structure and, as shown later, is easy to prove correct.

Gary L. Peterson. Myths about the mutual exclusion problem. Information Processing Letters, 1981

Peterson's Algorithm

```
// indicate the intend to hold the lock
bool flag[2] = {false, false};
// whose turn is it (thread 0 or 1)
int turn;

void lock() {
  flag[self] = true;  // I would like to enter
  turn = 1 - self;    // but make it other's turn
  while ((flag[1-self] == true) && (turn == 1 - self))
    ;
}

void unlock() {
  // undo the intent
  flag[self] = false;
}
```

Peterson's Algorithm

Thread 0

```
while(1) {
  flag[0] = true;
  turn = 1;
  while(flag[1] && turn == 1)
    ;
  // critical section
  flag[0] = false;
  // reminder section
}
```

Thread 1

```
while(1) {
  flag[1] = true;
  turn = 0;
  while(flag[0] && turn == 0)
    ;
  // critical section
  flag[1] = false;
  // reminder section
}
```

如果两个线程都想进入临界区，则该算法让谁先进?

Peterson's Algorithm

✅ 满足 Mutual Exclusion

T1 此时 while 条件是 True

```
T0:while()
T1:while()
turn=0
flag[0]=true
flag[1]=true
```

```
T0:turn = 1
T1:critical
turn=?
flag[0]=true
flag[1]=true
```

T0 会将 turn 的值设置为 1

T0 一定是从该状态进入临界区 (只有其 while 条件能被不满足)

```
T0:while()
T1:critical
turn=0
flag[0]=true
flag[1]=true
```

不失一般性，假设这个状态发生了 (T0 和 T1 都进入了临界区)

```
T0:critical
T1:critical
turn=0
flag[0]=true
flag[1]=true
```

所以不会执行到这个状态

Peterson's Algorithm

✅ **满足 Mutual Exclusion**

✅ **满足 Progress**

- 如果 `T0` 不在临界区，则一定有 `flag[0] = false`
- 此时 `T1` 一定能进入临界区

✅ **满足 Bounded Waiting**

- 如果 `T0` 正在等待进入临界区，则 `flag[1] = true` 且 `turn = 1`
- 此时如果 `T1` 想要再次进入临界区，则其一定会设置 `turn = 0`
- 此时，`T0` 将进入临界区

Peterson's Algorithm

通过手工证明一个程序是否满足特定性质是很繁琐且易错的事情

- 如果对 Peterson 算法稍作修改，其性质是否还能满足：
  - 交换 `flag` 和 `turn` 的赋值顺序？
  - 交换 `while` 语句两个条件的顺序？
  - …
- 可以借助模型检查 (Model Checking) 来实现自动化证明
  - 但需要有效地解决状态空间爆炸的问题

Peterson's Algorithm

虽然上述 Peterson 算法 "理论上" 是正确的，但在目前硬件架构上并不能保证实现互斥

- Load 和 Store 操作不一定是原子的 (e.g., write a 64-bit integer on 32-bit CPU requires two store instructions)
- CPU 指令乱序执行引起的内存一致性问题 (store-load in total store order of x86)
  - 可以使用内存屏障 (memory barrier) 来强制操作执行的顺序
- 此外，Peterson 算法的原始版本只支持两个线程
  - 可以扩展到 *N* 个线程，但需要提前知道 *N* 的值

Build Working Locks

不忘初心: 只要让 *TEST & SET* 是一个原子操作就可以了

```
// 0 -> lock is available, 1 -> held
flag = 0

void lock() {
  while(flag == 1) // TEST the flag
    ;
  flag = 1;          // now SET it
}

void unlock() {
  flag = 0;
}
```

Build Working Locks

Atomic Exchange (`xchg`)

返回 `ptr` 指向的值 (test the old value)，同时原子性地将该值更新为 `new` 对应的值 (set the memory location to a new value)

```
int xchg(int *ptr, int new) {
   int old = *ptr;  // fetch old value at ptr
   *ptr = new;      // store new into ptr
   return old;
}
```

Build Working Locks

Atomic Exchange (`xchg`)

```
typedef struct __lock_t {
   int flag;
} lock_t;

void init(lock_t *lock) {
   // 0 -> available, 1 -> held
   lock->flag = 0;
}

void lock(lock_t *lock) {
   while (xchg(&lock->flag, 1) == 1)
      ;
}

void unlock(lock_t *lock) {
   lock->flag = 0;
}
```

Build Working Locks

Compare-and-Swap (`cmpxchg`)

测试 `ptr` 指向的值是否等于 `expected`

- 如果相等，使用 `new` 的值进行更新
- 如果不相等，不做任何操作

```
int cmpxchg(int *ptr, int expected, int new) {
   int old = *ptr;
   if (old == expected)
      *ptr = new;
   return old;
}
```

Build Working Locks

Compare-and-Swap (`cmpxchg`)

```
typedef struct __lock_t {
   int flag;
} lock_t;

void init(lock_t *lock) {
   // 0 -> available, 1 -> held
   lock->flag = 0;
}

void lock(lock_t *lock) {
   while (cmpxchg(&lock->flag, 0, 1) == 1)
      ;
}

void unlock(lock_t *lock) {
   lock->flag = 0;
}
```

Spin Lock

上述基于 Test-And-Set 原子指令实现的锁也被称为自旋锁 (Spin Lock)

- 线程获取不到锁时不断在一个 `while` 循环中空转 (Busy Waiting)，直到锁的状态再次变为 available 为止

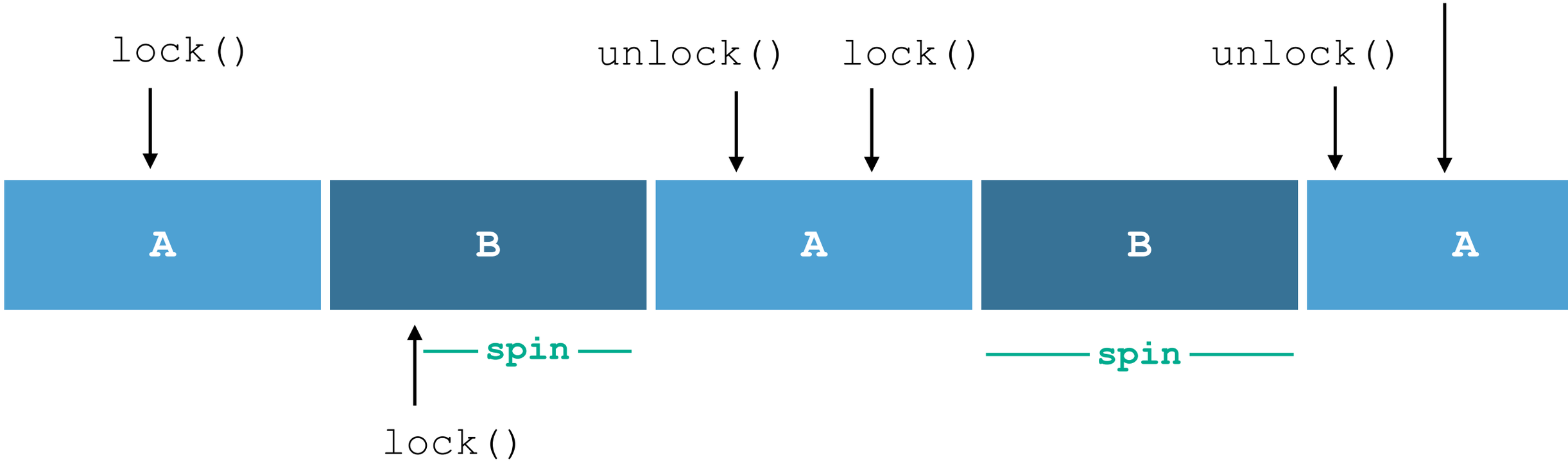
Spin Lock

自旋锁的公平性 (Fairness) 如何？

- 一个想进入临界区的线程一定能进入临界区吗？
- 持有锁的线程释放锁时，哪个线程会获得锁？

Spin Lock

一个在自旋等待锁的线程可能会永远自旋下去 (starvation)



Ticket Lock

Fetch-and-Add (`xadd`)

原子性地增加 `ptr` 指向的值 (同时返回该地址的旧值)

```
int xadd(int *ptr) {
   int old = *ptr;
   *ptr = old + 1;
   return old;
}
```

Ticket Lock

Fetch-and-Add (`xadd`)

```
typedef struct __lock_t {
   int ticket;
   int turn;
} lock_t;

void init(lock_t *lock) {
   lock->ticket = 0;
   lock->turn = 0;
}

void lock(lock_t *lock) {
   int my_turn = xadd(&lock->ticket);
   while (lock->turn != my_turn)
      ;
}

void unlock(lock_t *lock) {
   lock->turn = lock->turn + 1;
}
```

线程每次想要进入临界区就拿一个“号” (ticket)，并等待“叫号”

Ticket Lock

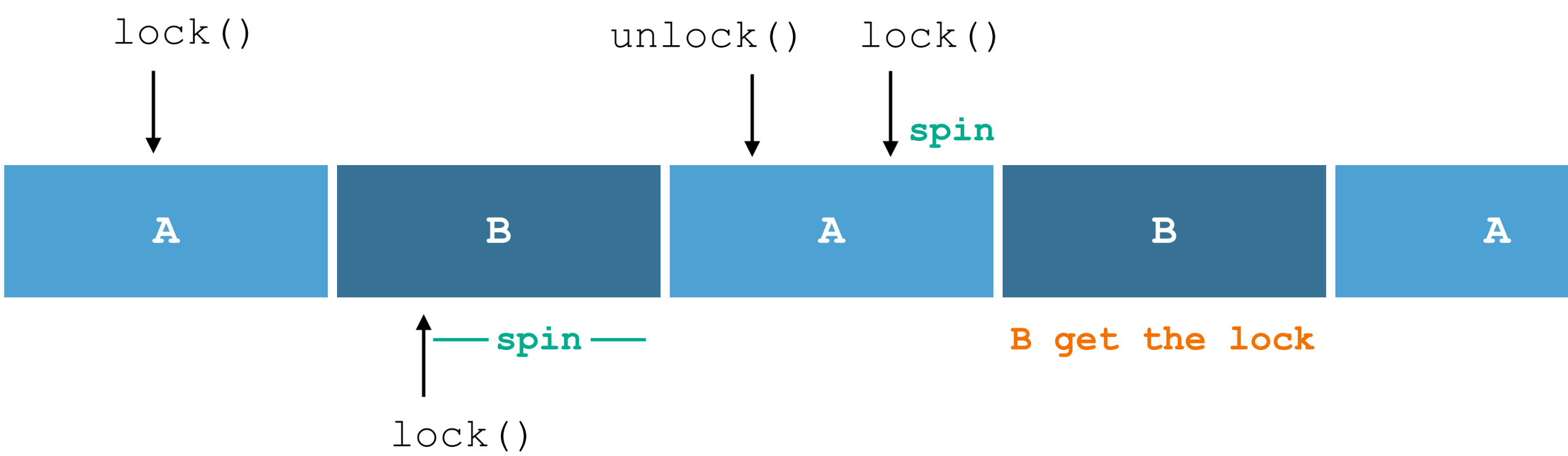
Fetch-and-Add (`xadd`)

```
init ticket = 0, turn = 0
```

```
get ticket 0,
spin until turn == 0
```

```
turn = 1
```

```
get ticket 2,
spin until turn == 2
```




```
get ticket 1,
spin until turn == 1
```

Spin Lock

自旋锁的性能 (Performance) 如何？

- 除了当前持有锁的线程外，其它线程都在空转
  - 对于比较小的临界区，自旋一会儿是一种合理的选择 (但在单核下仍旧会导致性能下降)
  - 对于比较大的临界区 (锁被长时间持有)，较大的计算资源浪费
- 更糟糕的是，如果当前持有锁的线程被切换出去了 (当并发执行的线程个数超过处理器核个数)
  - 其它所有线程的时间片都用来自旋，不断检查一个注定不会发生改变的状态 (100% 的资源浪费)

Locks for User Apps

Just Yield

当线程无法获得锁时，利用 `yield` 系统调用主动让出 CPU (线程状态由 Running 变为 Ready)

```
void lock() {
  while (xchg(&flag, 1) == 1)
    yield();
}

void unlock() {
  flag = 0;
}
```

Locks for User Apps

Just Yield

当线程无法获得锁时，利用 `yield` 系统调用主动让出 CPU (线程状态由 Running 变为 Ready)

- 只是暂时让出 CPU，线程处于 Ready 状态可随时被再次调度
  - 在获得锁之前，反复的 “scheduled → yield” 会带来大量的 context switch 开销 (设想 100 个想获取锁的线程通过 RR 方式调度)
- 同时也存在公平性 (Fairness) 的问题

Locks for User Apps

Sleep and Wakeup

当线程无法获得锁时，通过 `sleep` 将其置于 Blocked 状态，并在锁释放时通过 `wakeup` 唤醒正在等待的一个线程 (block when waiting)

```
void lock(lock_t *lock) {
  // if cannot acquire the lock -> blocked
  while(xchg(&lock->flag, 1) == 1) {
    sleep(lock->queue);
  }
}

void unlock(lock_t *lock) {
  // if there are threads waiting -> ready
  lock->flag = 0;
  if(!is_empty(lock->queue))
    wakeup(lock->queue);
}
```

What could possibly go wrong?

Locks for User Apps

Sleep and Wakeup

Thread 0

```
// assume flag = 1
// lock()
while(xchg(&lock->flag, 1) == 1)


  sleep(lock->queue);
```

Thread 1

```


// unlock()
flag = 0;
if(!is_empty(lock->queue))
  wakeup(lock->queue);
```

Lost wakeup

Locks for User Apps

Sleep and Wakeup

Thread 0

```
// assume flag = 1
// lock()
while()
  sleep();
```

Thread 1

```
// unlock()
flag = 0;

if(!is_empty())
  wakeup();
```

Thread 2

```
// lock()
while()
// critical section
```

Wrong thread gets the lock

```
void lock(lock_t *m) {
  while (xchg(&m->guard, 1) == 1)
    ; // acquire guard lock
  if (m->flag == 0)
    m->flag = 1; // lock is acquired
    m->guard = 0;
  else
    queue_add(m->q, gettid());
    setpark();  // are going to sleep
    m->guard = 0;
    park();      // sleep
}
```

```
typedef struct {
  int flag = 0;
  int guard = 0;
  queue *q;
} lock_t
```

```
void unlock(lock_t *m) {
  while (xchg(&m->guard, 1) == 1)
    ; // acquire guard lock
  if (queue_empty(m->q))
    m->flag = 0; // no one wants the lock
  else
    unpark(queue_remove(m->q));  // wakeup
  m->guard = 0;
}
```

Example of Solaris (see textbook for more details)

Locks for User Apps

Linux 系统中提供 **`futex` (fast user space mutex)** 系统调用

`futex_wait(int *address, int expected)`

- 原子性地判断 `address` 地址上的值是否和 `expected` 相等，并在相等时阻塞调用该操作的线程
- 如果不相等，则直接返回

`futex_wake(int *address)`

- 唤醒一个正阻塞在 `address` 上的线程

Spin or Block

选择 Spin 还是 Block 实现取决于临界区的长短、以及对锁的争用情况

- Spin 实现
  - 成功获得锁时直接进入临界区 (low cost on success)
  - 但在失败时会浪费 CPU 时间自旋 (high cost on failed)
- Block 实现
  - 避免了无法获得锁时的自旋开销 (reduce cost on failed)
  - 但每次申请和释放锁时 (即使当前无人争抢锁) 都要陷入内核 (certain cost even on success)

Two-Phase Locking

Combine the best of spin and block

结合 Spin 和 Block 的优点实现一种两阶段的锁 (a hybrid approach)

- 现实中解决性能优化问题的一种常见思路
  - **Fast path:** 在获取不到锁时先自旋一会儿，期望锁能尽快被释放
  - **Slow path:** 如果还是无法获得锁，则阻塞自己

- `pthread_mutex_lock()`和 `pthread_mutex_unlock()` 提供了一个高可扩展的实现

```
#define UNLOCK    0
#define ONE_HOLD  1
#define WAITERS   2
```

```
void unlock(mutex_lock_t* lk){
  // state can only be ONE_HOLD or WAITERS
  if (atomic_dec(lk) != ONE_HOLD) {
    // has more than one waiters
    lk = UNLOCK;
    futex_wake(lk);
  }
}
```

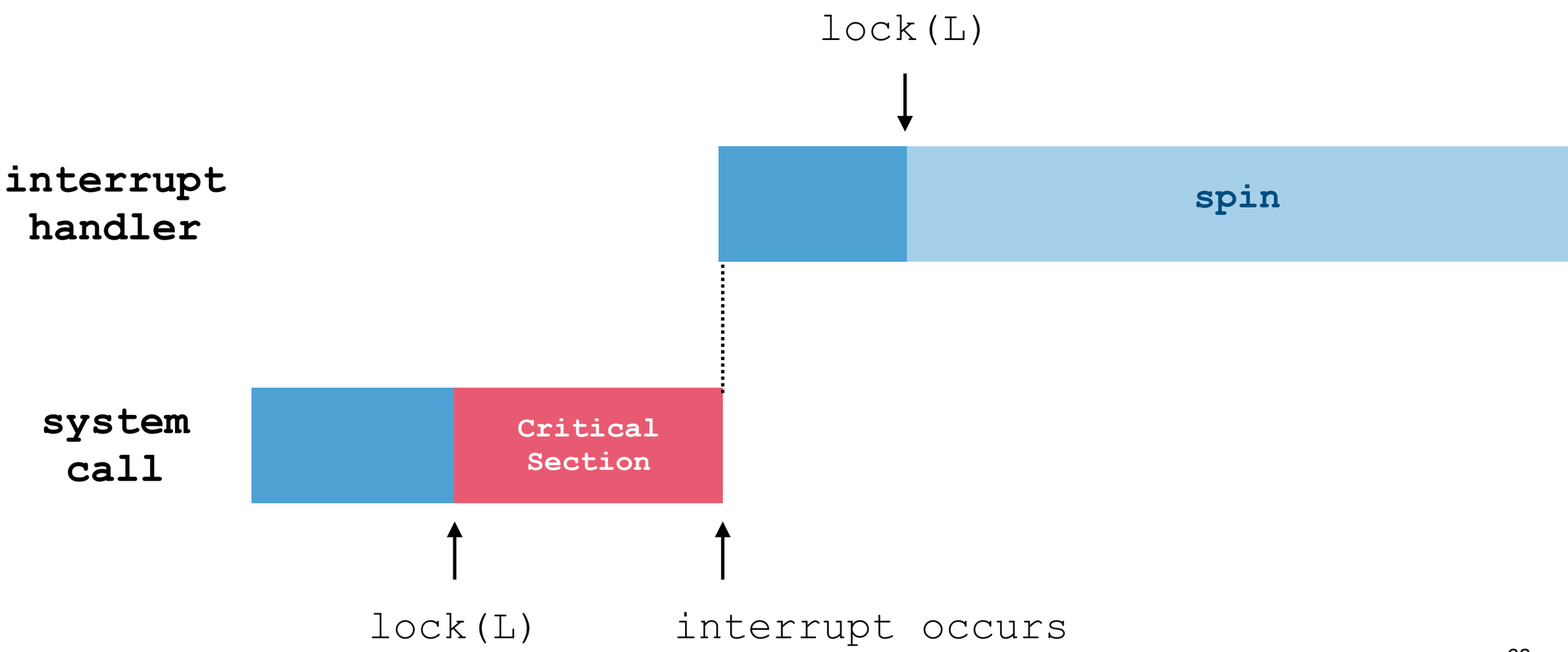
```
void lock(mutex_lock_t* lk) {
  int c = cmpxchg(lk, UNLOCK, ONE_HOLD);
  // if the lock is previously UNLOCKED, there is nothing else to do
  // otherwise, we will probably have to wait
  if (c != UNLOCK){
    do {
      // if the lock is ONE_HOLD, now there are waiters (cmpxchg)
      if (c == WAITERS || cmpxchg(lk, ONE_HOLD, WAITERS) != UNLOCK)
        futex_wait(lk, WAITERS);
      // once futex_wait returns, or we did not make the call
      // another attempt to take the lock
     } while ((c = cmpxchg(&lk, UNLOCK, WAITERS)) != UNLOCK);
    }
  }
}
```

See Futexes are tricky by Ulrich Drepper for more details

More Problems: Interrupt

在中断处理代码中使用 Spin Lock 会带来新的问题

- 假设线程在临界区内发生中断，并且 Interrupt Handler 试图获取相同的 Spin Lock；由于 Interrupt Handler 通常具有更高的优先级，因此线程没有机会离开临界区 …



More Problems: Interrupt

一种在内核中实现互斥的方法是在自旋之前关中断
(spin lock + disable interrupt)

```
void lock() {
  disable_interrupt();
  while(xchg(&lock->flag, 1) == 1)
    ;
}

void unlock() {
  lock->flag = 0;
  enable_interrupt();  // can we directly enable
                       // interrupt here?
}
```

假设按如下的方式使用锁？

lock(A) -> lock(B) -> unlock(B) -> unlock(A)

More Problems: Interrupt

一种在内核中实现互斥的方法是在自旋之前关中断

- 需要记录自旋之前的中断状态
- **xv6** 里分别使用 `push_off()` 和 `pop_off()` 记录中断关闭和打开的次数，只有当打开次数等于关闭次数时才能真正打开中断

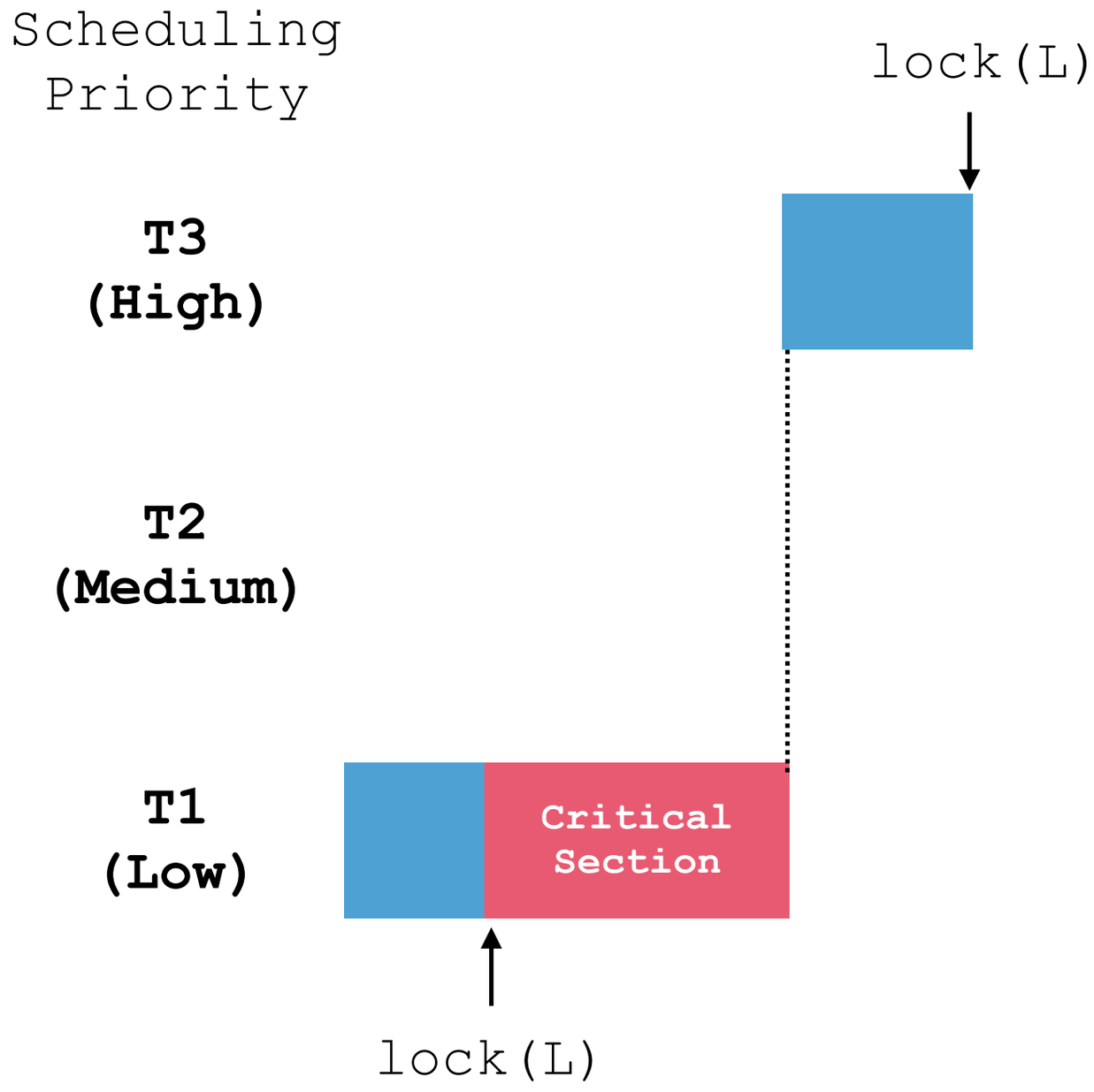
```
void
push_off(void)
{
  int old = intr_get();

  intr_off();
  if(mycpu()->noff == 0)
    mycpu()->intena = old;
  mycpu()->noff += 1;
}
```

```
pop_off(void)
{
  struct cpu *c = mycpu();
  if(intr_get())
    panic("pop_off - interruptible");
  if(c->noff < 1)
    panic("pop_off");
  c->noff -= 1;
  if(c->noff == 0 && c->intena)
    intr_on();
}
```

https://github.com/mit-pdos/xv6-riscv/blob/riscv/kernel/spinlock.c

More Problems: Priority Inversion

优先级调度和互斥这两个需求一起出现也会带来新的问题 (优先级反转)

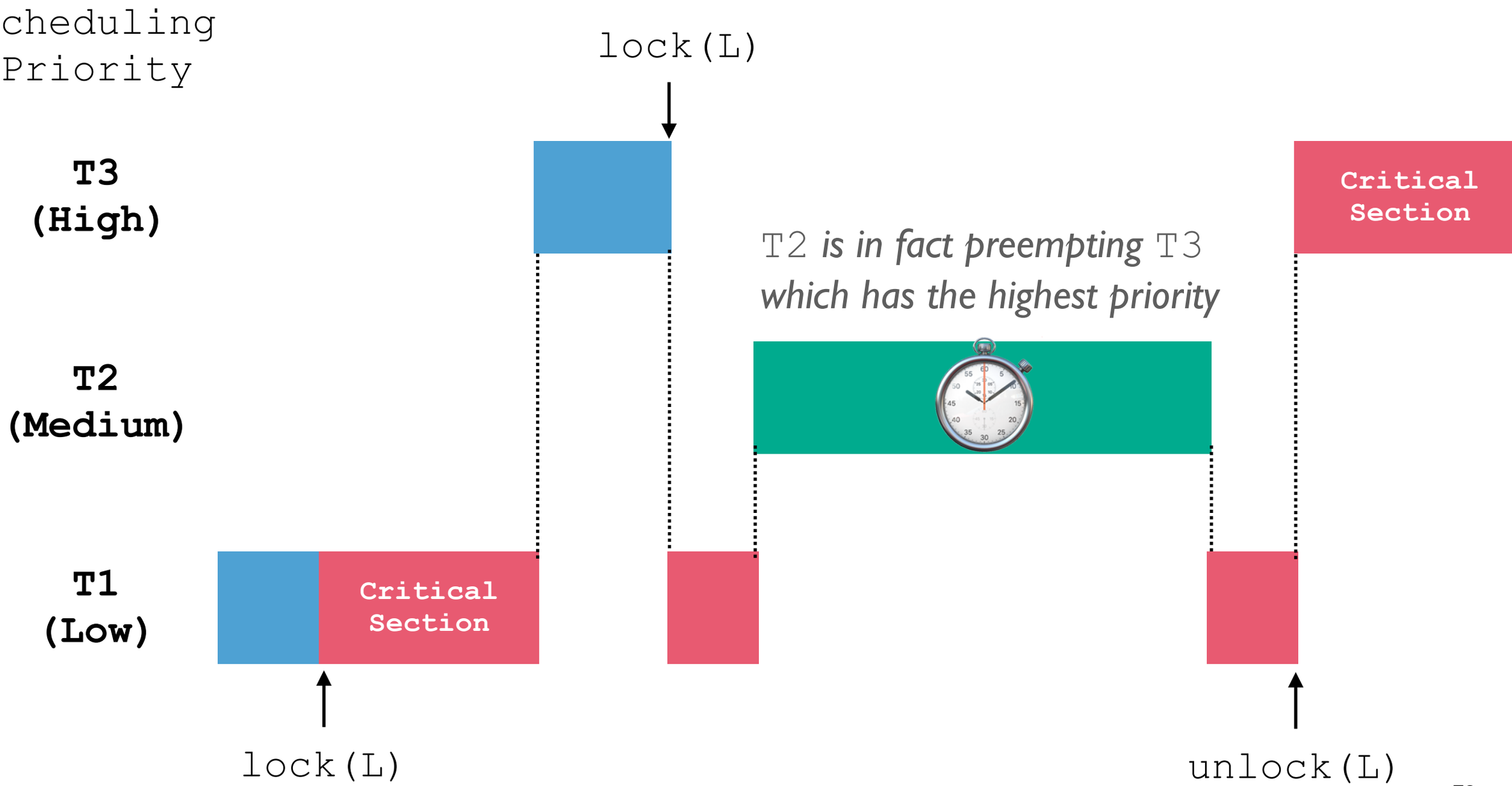


如果是 *Spin* 实现的 *Lock* ?

如果是 *Block* 实现的 *Lock* ?

More Problems: Priority Inversion

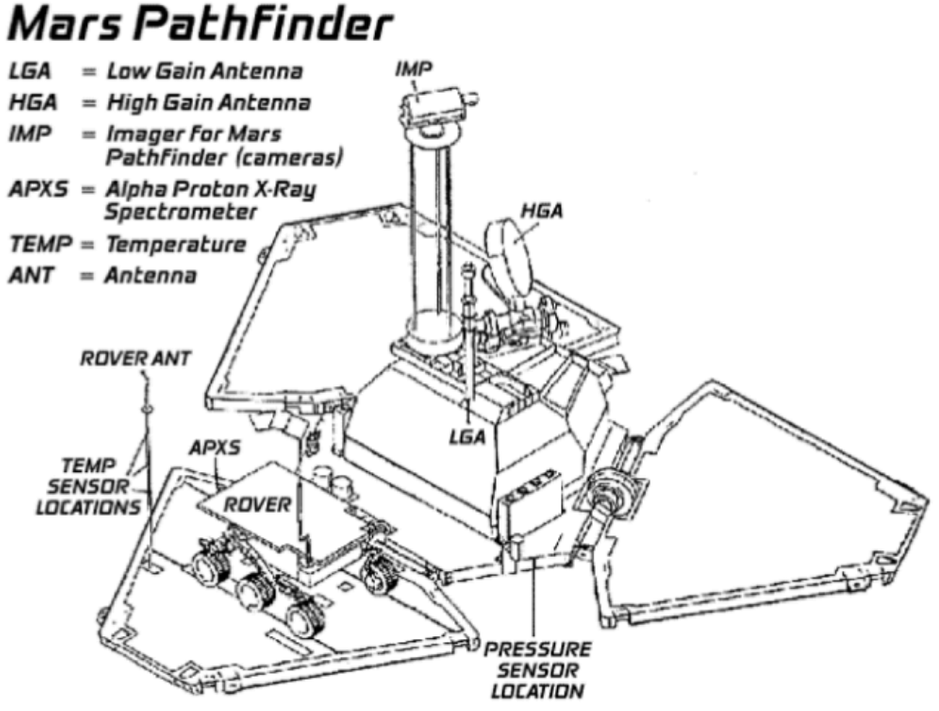
优先级调度和互斥这两个需求一起出现也会带来新的问题 (优先级反转)



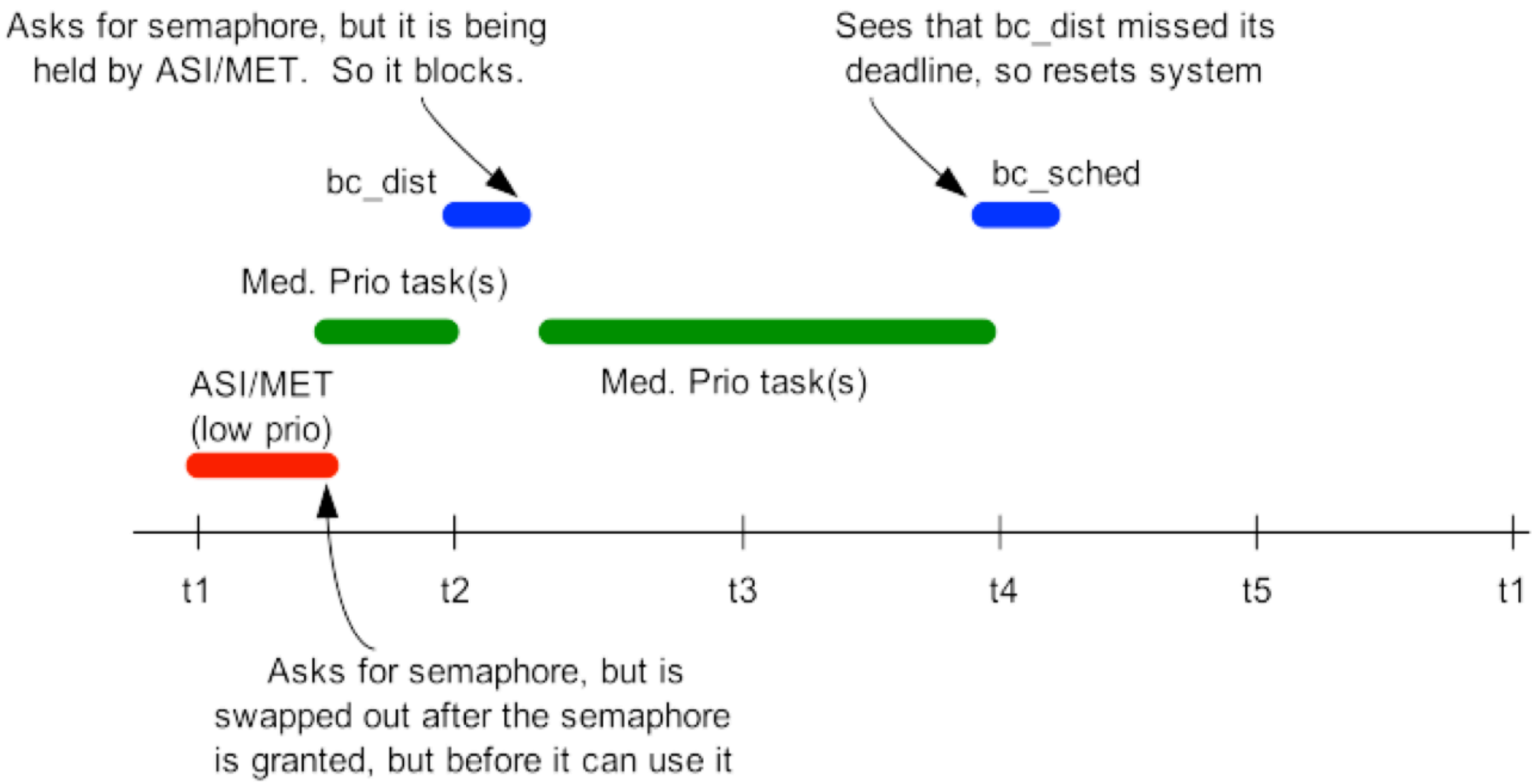
The First Bug on Mars

Mars Pathfinder (1997)

- 优先级反转问题几乎终结了火星探路者任务
- 在部署火星车着陆器后，整个系统每隔几天就因为间歇性的优先级反转问题而被 watchdog timer 重启一次
- 工程师最终发现了这个 Bug，并向着陆器发送了修复补丁



The First Bug on Mars

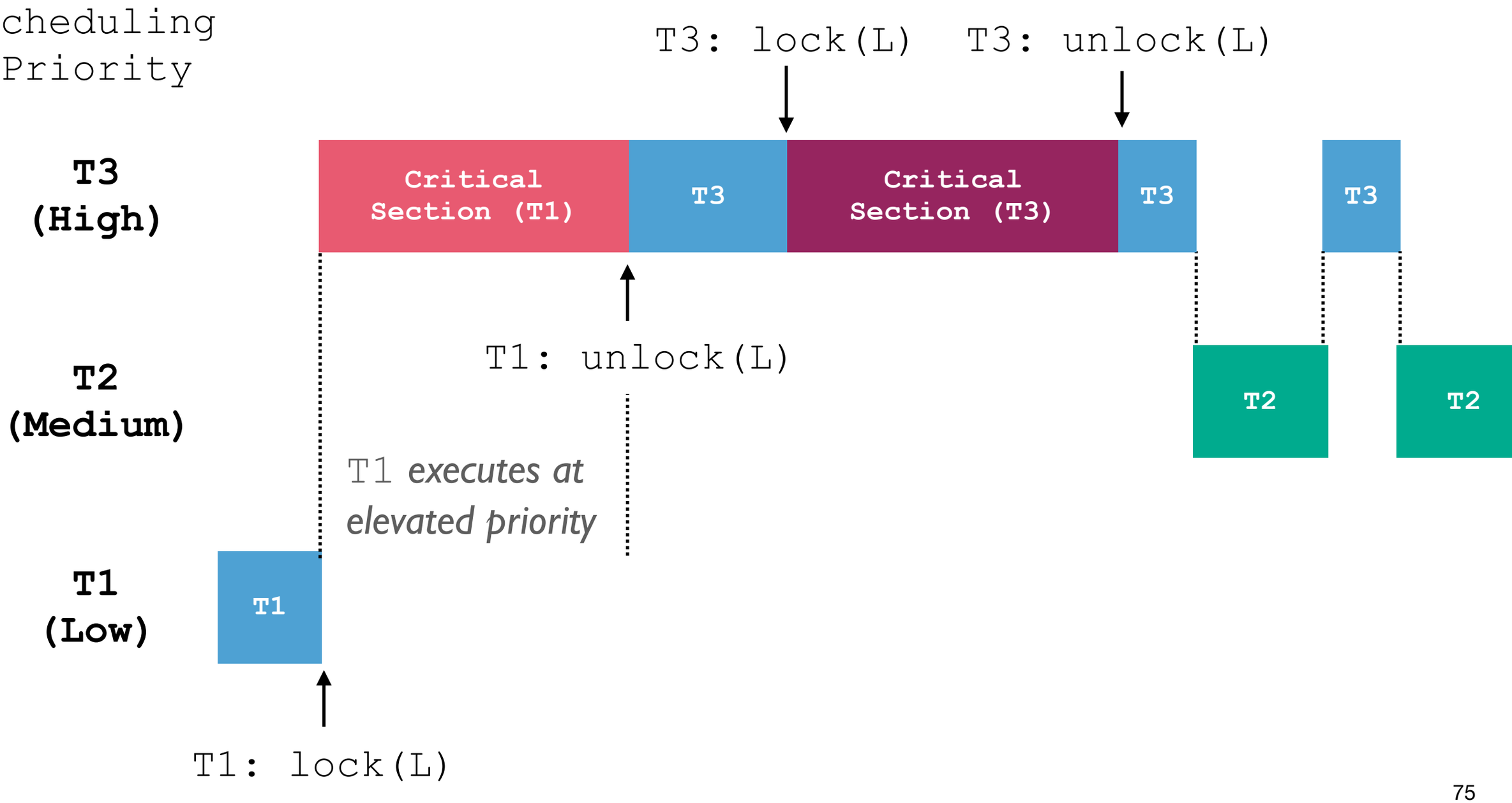


Even when you think you've tested everything that you can possibly imagine, you're wrong.

— Glenn E. Reeves
(Pathfinder's Software Team Leader)

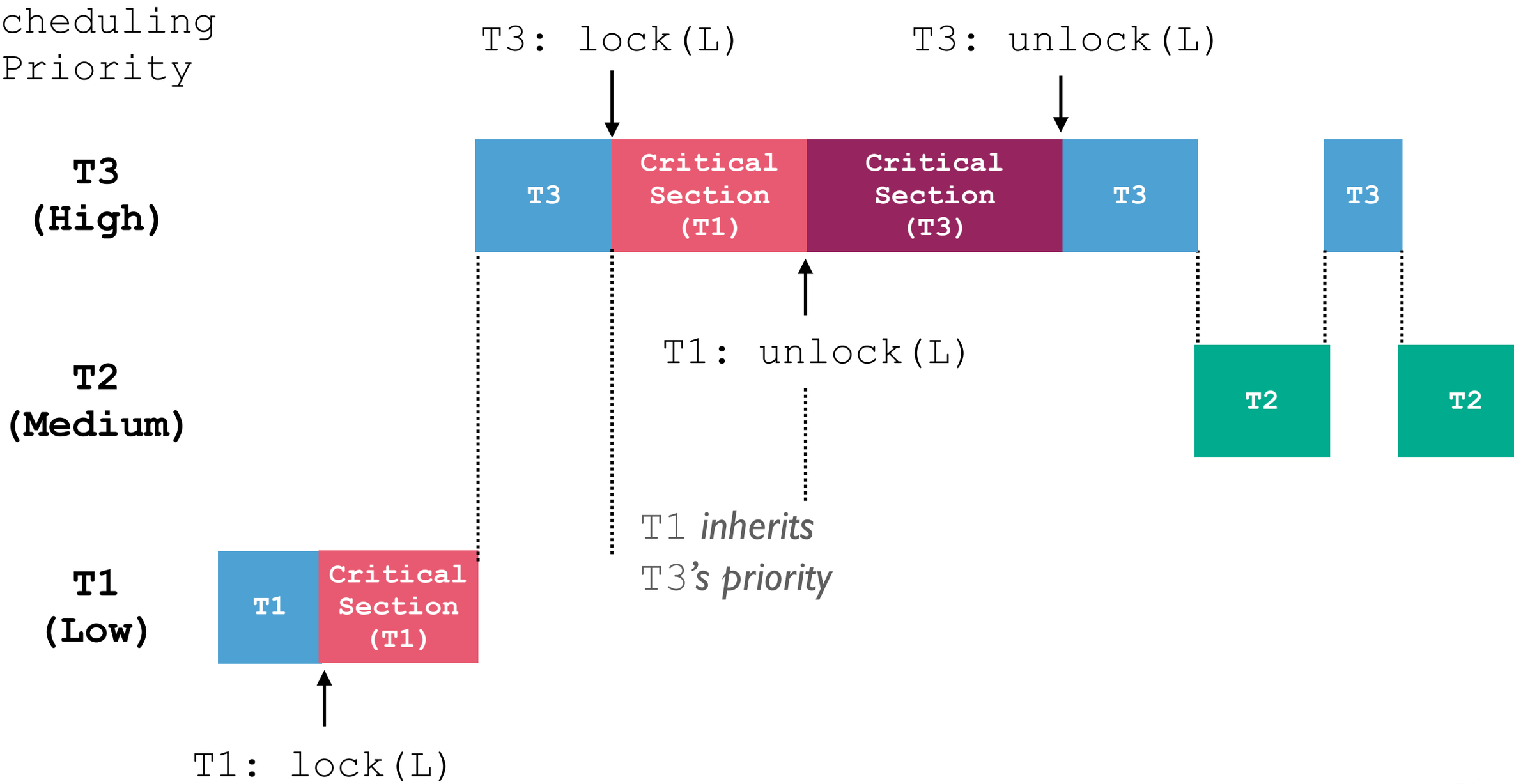
Priority Ceiling

当某个任务获得一个锁时，将该任务的优先级提升到与该锁关联的所有任务的优先级上限



Priority Inheritance

当某个高优先级任务想要获取一个低优先级任务持有的锁时，将该持有锁的任务的优先级提高到当前任务的优先级



Condition Variables

Waiting For Another

除了支持对共享资源的互斥访问外，还需要确保一个线程只有在满足特定条件时才能继续执行 (多个线程之间的协作)

```
void student() {
  while (1) {
    // wait until some condition is satisfied
    wait_until(lab_is_ready);
    write_code();
    submit_code();
  }
}

void TA() {
  while (1) {
    debug_lab();
    // the action to change the system state
    release_lab();
  }
}
```

Producer/Consumer Problem

假设有一个或多个生产者和消费者线程，其共享一个有界大小的缓冲区用于交换数据 (also known as the bounded buffer problem, first posed by Dijkstra)

A bounded buffer

(a shared resource, and of course must require synchronised access to it)

Producer/Consumer Problem

```
int buffer[MAX];
int in = 0;
int out = 0;
int count = 0;
```

```
void producer() {
  while(1) {
    produce_value();
    // put(value)
    buffer[in] = value;
    in = (in + 1) % MAX;
    count++;
  }
}
```

```
void consumer() {
  while(1) {
    // get(value)
    value = buffer[out];
    out = (out + 1) % MAX;
    count--;
    consume_value();
  }
}
```

Producer/Consumer Problem

生产者和消费者线程对 Buffer 的操作需要满足特定的同步条件

- 同一时刻只能有一个线程 (生产者或消费者) 对 Buffer 中的数据进行修改 (mutual exclusion)
- 当 Buffer 已满时 (`count = MAX`)，生产者应停止向其中写入数据，避免数据被覆盖 (wait until there is a space in the buffer)
- 当 Buffer 为空时 (`count = 0`)，消费者应停止从其中取出数据，避免获得无效数据 (wait until there is something in the buffer)

Producer/Consumer Problem

```
void producer() {
  while(1) {
    produce_value();
    while(count == MAX)
      ;
    // put(value)
    buffer[in] = value;
    in = (in + 1) % MAX;
    count++;
  }
}
```

```
void consumer() {
  while(1) {
    while(count == 0)
      ;
    // get(value);
    value = buffer[out];
    out = (out + 1) % MAX;
    count—;
    consume_value();
  }
}
```

共享变量的 Race Condition

Producer/Consumer Problem

```
mutex_t mutex;
```

```
void producer() {
  while(1) {
    produce_value();
retry:
    lock(mutex);
    if(count == MAX) {
      unlock(mutex);
      goto retry;
    }
    put(value);
    unlock(mutex);
  }
}
```

```
void consumer() {
  while(1) {
retry:
    lock(mutex);
    if(count == 0) {
      unlock(mutex);
      goto retry;
    }
    get(value);
    unlock(mutex);
    consume_value();
  }
}
```

Busy Waiting 浪费计算资源

Producer/Consumer Problem

```
void producer() {
  while(1) {
    produce_value();
    if(count == MAX)
      sleep();
    put(value);
    wakeup();
  }
}
```

```
void consumer() {
  while(1) {
    if(count == 0)
      sleep();
    get(value);
    wakeup();
    consume_value();
  }
}
```

Producer/Consumer Problem

```
mutex_t mutex;
```

```
void producer() {
  while(1) {
    produce_value();
    lock(mutex);
    if(count == MAX)
      sleep();
    put(value);
    wakeup();
    unlock(mutex);
  }
}
```

```
void consumer() {
  while(1) {
    lock(mutex);
    if(count == 0)
      sleep();
    get(value);
    wakeup();
    unlock(mutex);
    consume_value();
  }
}
```

一个持有锁的线程执行了 `sleep()`？

Producer/Consumer Problem

```
mutex_t mutex;
```

```
void producer() {
  while(1) {
    produce_value();
    lock(mutex);
    if(count == MAX)
      unlock(mutex);
      sleep();
      lock(mutex);
    put(value);
    wakeup();
    unlock(mutex);
  }
}
```

```
void consumer() {
  while(1) {
    lock(mutex);
    if(count == 0)
      unlock(mutex);
      sleep();
      lock(mutex);
    get(value);
    wakeup();
    unlock(mutex);
    consume_value();
  }
}
```

在一个线程执行了 unlock()但还没 sleep()之前，
另一个线程执行了 wakeup()？

Producer/Consumer Problem

```
mutex_t mutex;
```

```
void producer() {
  while(1) {
    produce_value();
    lock(mutex);
    if(count == MAX)
      unlock(mutex);
      futex_wait();
      lock(mutex);
    put(value);
    futex_wake();
    unlock(mutex);
  }
}
```

```
void consumer() {
  while(1) {
    lock(mutex);
    if(count == 0)
      unlock(mutex);
      futex_wait();
      lock(mutex);
    get(value);
    futex_wake();
    unlock(mutex);
    consume_value();
  }
}
```

We have implemented condition variables

Condition Variables

条件变量 (condition variables) 提供了一种让线程只有在满足特定条件才能继续执行的方法

💤 **`wait(cond cv, mutex m)`**

- 调用该操作时需持有互斥锁 (用于保护对条件的判断和修改)
- 原子性地释放互斥锁，并阻塞调用该操作的线程
- 在阻塞的线程被唤醒之后，需要在返回之前再次获得互斥锁

⏰ **`signal(cond cv)`**

- 唤醒一个正阻塞在该条件变量上的线程
- 如果当前没有线程等待，则不产生任何效果 (the signal is lost)

Condition Variables

一个利用 futex 的简化版实现

```
typedef struct __cond_t {
  unsigned int value;
} cond_t;

void wait(cond_t *cv, mutex_t *lk) {
  int val = atomic_load(&cv->value);
  mutex_unlock(&lk);
  futex_wait(&cv->value, val);  // if value = val, sleep
  mutex_lock(&lk);
}

void signal(cond_t *cv) {
  atomic_fetch_add(&cv->value, 1);
  futex_wake(&cv->value);
}
```

Producer/Consumer Problem

```
cond_t cv;  // condition variable
mutex_t mutex;
```

```
void producer() {
  while(1) {
    produce_value();
    lock(mutex);
    if(count == MAX)
      wait(cv, mutex);
    put(value);
    signal(cv);
    unlock(mutex);
  }
}
```

```
void consumer() {
  while(1) {
    lock(mutex);
    if(count == 0)
      wait(cv, mutex);
    get(value);
    signal(cv);
    unlock(mutex);
    consume_value();
  }
}
```

从 `wait()` 返回时系统状态一定符合预期吗?

Producer/Consumer Problem

Producer P1

```
// full buffer
lock(mutex);
if(count == MAX)
  wait(cv, mutex);




put(value)?
```

Consumer C1

```
lock(mutex);
get(value);
signal(cv);
unlock(mutex);
```

Producer P2

```
// P2 sneaks in
lock(mutex);
if(count == MAX)
  wait(cv, mutex)
put(value);
// now, count = MAX
```

在唤醒 P1 但 P1 还没执行之前，另一个生产者 P2 可能被调度执行，
并进而改变了当前系统的状态 (count 的值)

Producer/Consumer Problem

```
cond_t cv;  // condition variable
mutex_t mutex;
```

```
void producer() {
  while(1) {
    produce_value();
    lock(mutex);
    while(count == MAX)
      wait(cv, mutex);
    put(value);
    signal(cv);
    unlock(mutex);
  }
}
```

```
void consumer() {
  while(1) {
    lock(mutex);
    while(count == 0)
      wait(cv, mutex);
    get(value);
    signal(cv);
    unlock(mutex);
    consume_value();
  }
}
```

Interpretation of Signal

条件变量 `singal()` 操作有不同的语义

- **Mesa Semantics** (always use a while loop)
  - 唤醒一个线程时只将其设置为 Ready 状态 (易于实现)
  - 提示当前系统状态发生了变化，但不保证被唤醒线程调度执行时系统仍处于预期状态 (仅保证线程从 `wait()` 返回时一定持有锁)
- **Hoare Semantics**
  - 当一个线程被唤醒时，其得到立即调度执行
    - 互斥锁转移到被唤醒的线程，同时阻塞执行 `signal()` 的线程
  - 提供了更强的保证 (尤其有助于证明程序的某种性质)，但在真实系统中难以实现

Producer/Consumer Problem

Producer P1

```
// full buffer
// MAX = 1
while(count == MAX)
  wait(cv, mutex)




put(value)
// now, count = MAX
signal(cv)
```

Producer P2

```
while(count == MAX)
  wait(cv, mutex)


// if wakeup P2
while(count == MAX)
  wait(cv, mutex)
```

Consumer C1

```
get(value)
signal(cv) // -> P1
while(count == 0)
  wait(cv, mutex)


// should wakeup C1
get(value)
```

signal()应该是有向的(生产者和消费者只应该互相唤醒)

Producer/Consumer Problem

```
cond_t empty, fill;
mutex_t mutex;
```

```
void producer() {
  while(1) {
    produce_value();
    lock(mutex);
    while(count == MAX)
      wait(empty, mutex);
    put(value);
    signal(fill);
    unlock(mutex);
  }
}
```

```
void consumer() {
  while(1) {
    lock(mutex);
    while(count == 0)
      wait(fill, mutex);
    get(value);
    signal(empty);
    unlock(mutex);
    consume_value();
  }
}
```

Solution 1: 使用两个条件变量

(生产者和消费者 wait 在不同的条件变量上)

Producer/Consumer Problem

```
cond_t cv;
mutex_t mutex;
```

```
void producer() {
  while(1) {
    produce_value();
    lock(mutex);
    while(count == MAX)
      wait(cv, mutex);
    put(value);
    broadcast(cv);
    unlock(mutex);
  }
}
```

```
void consumer() {
  while(1) {
    lock(mutex);
    while(count == 0)
      wait(cv, mutex);
    get(value);
    broadcast(cv);
    unlock(mutex);
    consume_value();
  }
}
```

Solution 2: 使用 `broadcast()` 唤醒所有线程

(也称为 covering conditions: cover 所有可能需要被唤醒的情况)

Rules of Thumb

- 使用共享变量来记录系统状态 (确定线程执行的条件是否满足)
- 使用互斥锁来对共享变量进行保护
- 总是在持有锁的时候执行 `wait()` 和 `signal()/broadcast()`
- 当线程被唤醒时，使用 `while` 循环来再次检查系统状态
- 为不同的条件定义不同的条件变量，或者总是使用 `broadcast()` 来唤醒所有线程