

Intro to Multithreading

CS110L

February 14, 2022

Logistics

- Project 1 due on Sunday
 - Post questions in #proj1-discussion
 - Please let me know if I can help!!!
 - Collaborate with each other — it'll make things easier. :)

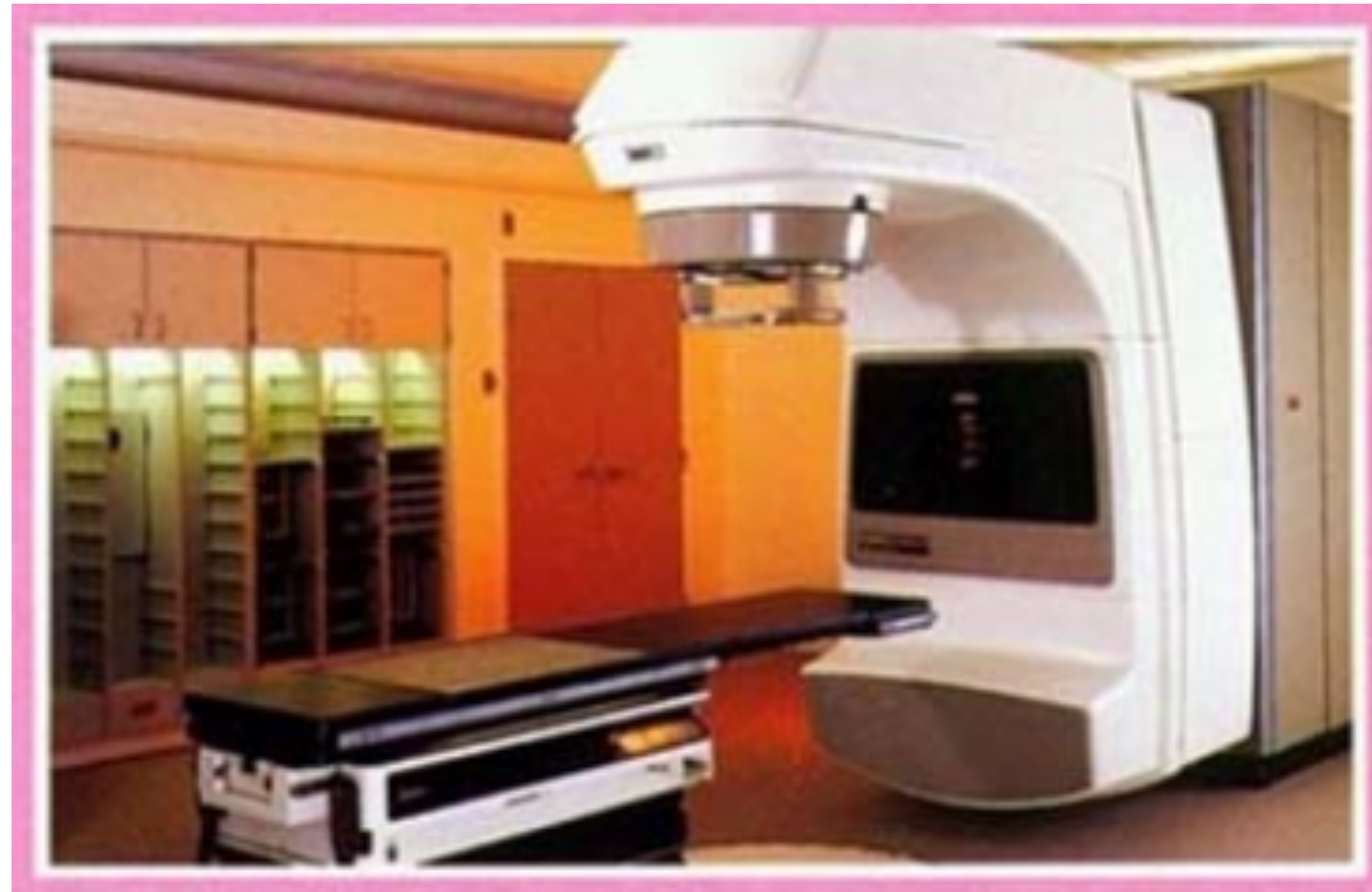
Why is multithreading nice?

- Parallelism and concurrency!
- Particularly helpful if...
 - Program is *I/O bound*: e.g., time spent waiting for a response from the network or data from file system. While one thread is idle, other threads can use the CPU. Allow the scheduler to *interleave* for you.
 - There are *multiple cores*: do multiple CPU-intensive tasks in *parallel*.

Perils of concurrency

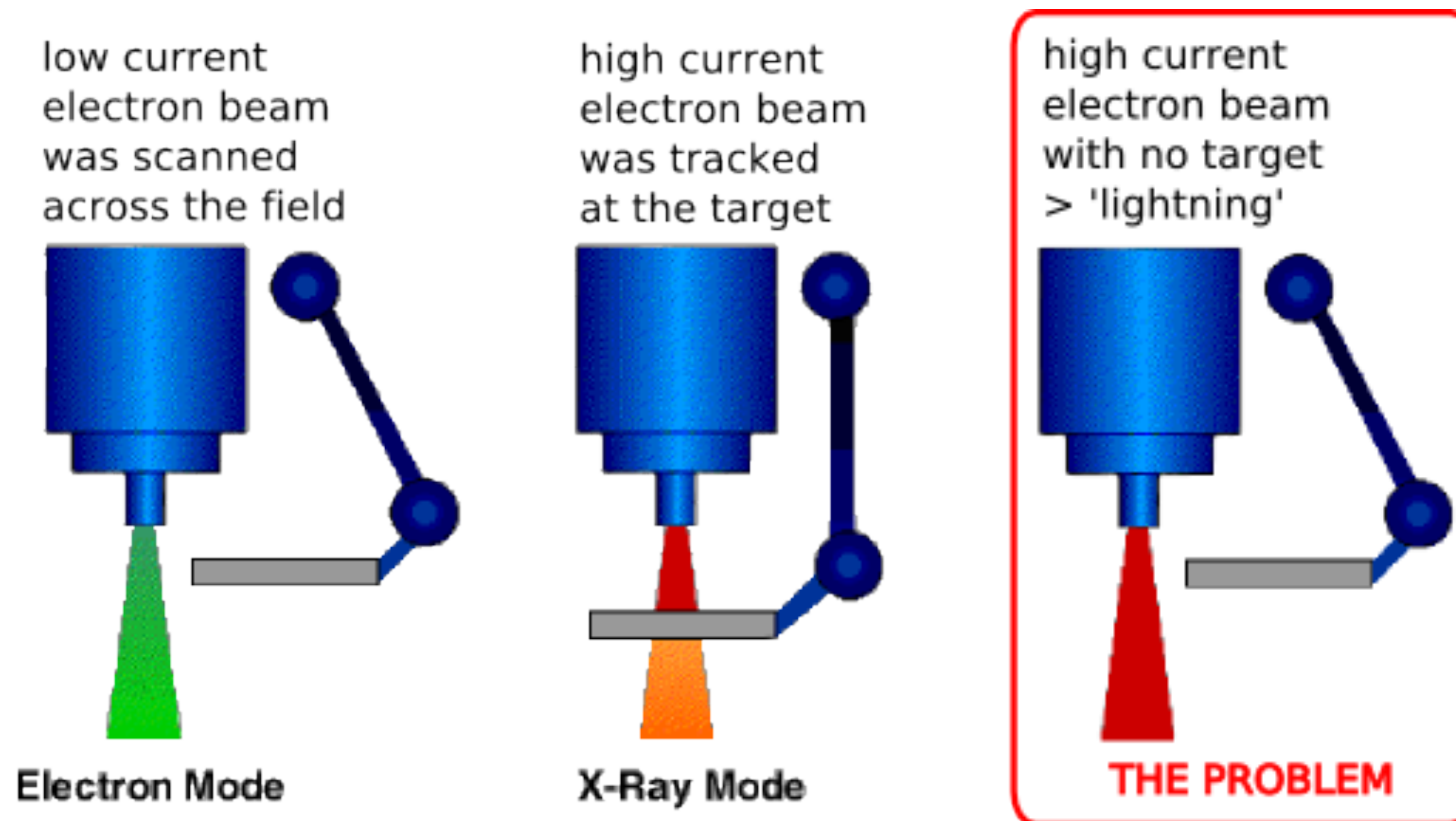
- Why is multithreading dangerous?
 - Race conditions
 - *Cause the program to not work... **but only sometimes!** They're easy to miss in testing and extremely hard to debug.*
 - *If you've taken CS110, I probably don't have to tell you that race conditions are frustrating.*
 - Deadlock (more next week)

Intense race condition example: Therac-25



<https://hci.cs.siu.edu/NSF/Files/Semester/Week13-2/PPT-Text/Slide13.html>
<https://hackaday.com/2015/10/26/killed-by-a-machine-the-therac-25/>

Intense race condition example: Therac-25



Intense race condition example: Therac-25

*After each overdose the creators of Therac-25 were contacted. After the first incident the AECL responses was simple: “**After careful consideration, we are of the opinion that this damage could not have been produced by any malfunction of the Therac-25 or by any operator error** (Leveson, 1993).”*

*After the 2nd incident the AECL sent **a service technician** to the Therac-25 machine, he **was unable to recreate the malfunction and therefore conclude nothing was wrong with the software.** Some minor adjustments to the hardware were changed but the main problems still remained.*

*It was not until the fifth incident that any formal action was taken by the AECL. However it was **a physicist at the hospital** where the 4th and 5th incident took place in Tyler, Texas who actually **was able to reproduce the mysterious "malfunction 54"**. The AECL **finally took action** and made a variety of changes in the software of the Therac-25 radiation treatment system.*

Intense race condition example: Therac-25

- Investigation results:
- ***The failure occurred only when a particular nonstandard sequence of keystrokes was entered on the VT-100 terminal which controlled the PDP-11 computer: an "X" to (erroneously) select 25 MeV photon mode followed by "cursor up", "E" to (correctly) select 25 MeV Electron mode, then "Enter", all within eight seconds.***
- *The equipment control task did not properly synchronize with the operator interface task, so that race conditions occurred if the operator changed the setup too quickly. **This was missed during testing, since it took some practice before operators were able to work quickly enough to trigger this failure mode.***
- <https://en.wikipedia.org/wiki/Therac-25> and <http://sunnyday.mit.edu/papers/therac.pdf>

Intense race condition example: Therac-25

- Key point: the race condition was **not caught during testing**, because it was not triggered during testing.
- The company concluded: “we’ve tested it, and it didn’t happen, so it can’t happen. Y’all must be wrong. Everything’s fine.”

Race conditions are everywhere!

- Starbucks: [possible to get unlimited coffee](#)
- GitHub: [possible to get logged in as a different user](#)
- [Unlimited bitcoin, voting multiple times, using Instacart coupons multiple times](#) (from Jack Cable)
- [Kernel race condition in CPlayground](#) (via Ryan Eberhardt - great blog post!)
- *Usually arise from an **unpredictable** input/behavior that is **unknown at compile time** and uncaught in testing.*
 - *In 110, we talk about the OS **scheduler***
 - *Other examples: user input (e.g., pressing two keys in quick succession), an I/O operation, ...*

Small probabilities are deceiving

- “Given the scale that Twitter is at, a one-in-a-million chance happens 500 times a day.” ([Del Harvey, 2014](#))

Compounding effects

- “I’m just working on my hot new social media app... Who cares if it breaks 0.01% of the time?”
- Let’s say that downloading/displaying a post involves 20 steps
 - Selecting the post to display, serializing, transmitting over the network, receiving, rendering, etc...
- You weren’t very careful, and 5 of those steps have race conditions that each manifest 0.01% of the time. Displaying a post will crash 0.05% of the time
- Let’s say the average user quickly scrolls through 300 posts/day. A user now has a ~15% chance of crashing the app every day
- Next, you add a messaging feature. Sending/receiving a message also fails 0.05% of the time
- A typical user sends/receives 100 messages a day. Now your app has a ~20% chance of crashing for a user on any given day
 - Who would want to use an app like this? (Not me!)

Recap:

- Some factors are fundamentally unpredictable
- Race conditions happen only sometimes.
 - No hard guarantees that they will be caught by static/dynamic analysis
- Scale matters in systems; small probabilities compound.
 - What happens when we scale this up?
 - What happens in a complicated codebase (e.g., millions of lines of code)
- When working with concurrency, you must be meticulous and disciplined
- Even the very best programmers make mistakes! **We need more tools to help us prevent and identify problems.**

Preventing data races

What are race conditions?

- Race condition:

A race condition or race hazard is the condition of an electronics, software, or other system where the system's substantive behavior is dependent on the sequence or timing of other uncontrollable events. ([Wikipedia](#))

- Data race:

Multiple threads access a value, where at least one of them is writing

- This should sound familiar!

Rust's design pays off

- Rust's design goals:
 - How do you do safe systems programming?
 - How do you make concurrency painless?
 - How do you make it fast?
- *“Initially these [first two] problems seemed orthogonal, but to our amazement, the solution turned out to be identical: the same tools that make Rust safe also help you tackle concurrency head-on.”* ([Rust blog](#))
- Compiler enforces rules for safe concurrency. *“Thread safety isn't just documentation; it's law.”*
- There's very little in the core language specific to threading! (Only two traits!) Almost all thread safety comes from the ownership model you already know

Hello world!

```
use std::{thread, time};
use rand::Rng;
```

```
const NUM_THREADS: u32 = 20;
```

```
fn main() {
    let mut threads = Vec::new();
    println!("Spawning {} threads...", NUM_THREADS);
    for _ in 0..NUM_THREADS {
        threads.push(thread::spawn(|| {
            let mut rng = rand::thread_rng();
            thread::sleep(time::Duration::from_millis(rng.gen_range(0, 5000)));
            println!("Thread finished running!");
        }));
    }
    // wait for all the threads to finish
    for handle in threads {
        handle.join().expect("Panic happened inside of a thread!");
    }
    println!("All threads finished!");
}
```

Parameters for closure function (none, in this case)

Closure/lambda function borrows any referenced variables

A panic in a thread will not crash the entire program
Can check if the thread panicked (and deal with it)

[Playground](#)

[Helpful explainer of closures in Rust](#)

Extroverts demo (CS 110)

```
static const char *kExtroverts[] = {
    "Frank", "Jon", "Lauren", "Marco", "Julie", "Patty",
    "Tagalong Introvert Jerry"
};
static const size_t kNumExtroverts = sizeof(kExtroverts)/sizeof(kExtroverts[0]) - 1;

int main() {
    vector<thread> threads;
    for (size_t i = 0; i < kNumExtroverts; i++) {
        threads.push_back(thread([&i]() {
            cout << "Hello from extrovert " << kExtroverts[i] << "!" << endl;
        }));
    }
    // wait for all the threads to finish
    for (thread& handle : threads) {
        handle.join();
    }
    return 0;
}
```

Passes a reference/pointer to i, but then the main thread changes i on the next iteration of the for loop. By the time the new thread runs, i is 7

Can we do the same in Rust?

Can we do the same in Rust?

```
use std::thread;

const NAMES: [&str; 7] = ["Frank", "Jon", "Lauren", "Marco", "Julie", "Patty",
    "Tagalong Introvert Jerry"];

fn main() {
    let mut threads = Vec::new();
    for i in 0..6 {
        threads.push(thread::spawn(|| {
            println!("Hello from extrovert {}", NAMES[i]);
        }));
    }
    // wait for all the threads to finish
    for handle in threads {
        handle.join().expect("Panic occurred in thread!");
    }
}
```

Closure/lambda function *borrow*s referenced variables by default (whenever possible)



`NAMES[i]`

Can we do the same in Rust?

error[E0373]: closure may outlive the current function, but it borrows `i`, which is owned by the current function

[--> src/main.rs:9:36](#)

```
9 |         threads.push(thread::spawn(|| {
    |                                   ^^ may outlive borrowed value `i`
10 |             println!("Hello from extrovert {}", NAMES[i]);
    |                                                     - `i` is borrowed here
```

note: function requires argument type to outlive `static`

[--> src/main.rs:9:22](#)

```
9 |         threads.push(thread::spawn(|| {
    |                                   ^
10 |             println!("Hello from extrovert {}", NAMES[i]);
11 |         }));
    |         ^
```

help: to force the closure to take ownership of `i` (and any other referenced variables), use the `move` keyword

```
9 |         threads.push(thread::spawn(move || {
    |                                   ^^^^^^^
```


Can we do the same in Rust?

```
use std::thread;

const NAMES: [&str; 7] = ["Frank", "Jon", "Lauren", "Marco", "Julie", "Patty",
    "Tagalong Introvert Jerry"];

fn main() {
    let mut threads = Vec::new();
    for i in 0..6 {
        threads.push(thread::spawn(move || {
            println!("Hello from extrovert {}!", NAMES[i]);
        }));
    }
    // wait for all the threads to finish
    for handle in threads {
        handle.join().expect("Panic occurred in thread!");
    }
}
```

i is moved into the closure function;
closure now has ownership

Ticket agents demo (CS 110)

```
static void ticketAgent(size_t id, size_t& remainingTickets) {
    while (remainingTickets > 0) {
        handleCall(); // sleep for a small amount of time to emulate conversation time.
        remainingTickets--;
        cout << oslock << "Agent #" << id << " sold a ticket! (" << remainingTickets
            << " more to be sold)." << endl << osunlock;
        if (shouldTakeBreak()) // flip a biased coin
            takeBreak(); // if comes up heads, sleep for a random time to take a break
    }
    cout << oslock << "Agent #" << id << " notices all tickets are sold, and goes home!"
        << endl << osunlock;
}
```

Multiple threads get mutable reference to remainingTickets

Value decremented simultaneously

```
int main(int argc, const char *argv[]) {
    thread agents[10];
    size_t remainingTickets = 250;
    for (size_t i = 0; i < 10; i++)
        agents[i] = thread(ticketAgent, 101 + i, ref(remainingTickets));
    for (thread& agent: agents) agent.join();
    cout << "End of Business Day!" << endl;
    return 0;
}
```


Ticket agents demo (CS 110)

- Race condition issue 1:

```
while (remainingTickets > 0) {  
    handleCall(); // sleep for a small amount of time to emulate conversation time.  
    remainingTickets--;
```

- Thread A reads that `remainingTickets` is 1, then goes to handle a call.
- Thread B reads that `remainingTickets` is 1, then goes to handle a call.
- Thread A decrements `remainingTickets` to 0.
- Thread B decrements `remainingTickets` to... `UINT_MAX`

Ticket agents demo (CS 110)

- Race condition issue 2:

```
remainingTickets--;
```

- Operations involved (roughly): read `remainingTickets` from memory, do some arithmetic in registers, write result of this arithmetic back to memory
- Say that remainingTickets is 5
 - Thread A starts this operation: reads `5` from memory
 - Thread B starts this operation: reads `5` from memory
 - Thread B decrements; writes back its result — `4` — to memory
 - Thread A decrements; writes back its result — `4` — to memory
- Value is out of sync! Should have decremented to `3`. The result of thread A's calculation *overwrote* the result of thread B's calculation.

Let's rewrite it in Rust!



Rewrite it in Rust, they said. It will be fun, they said

Attempt 1

```
fn main() {
    let mut remaining_tickets = 250;

    let mut threads = Vec::new();
    for i in 0..10 {
        threads.push(thread::spawn(move || {
            ticket_agent(i, &mut remaining_tickets)
        }));
    }
    // wait for all the threads to finish
    for handle in threads {
        handle.join().expect("Panic occurred in thread!");
    }
    println!("End of business day!");
}
```

This code only compiles because `i32` is Copy. Every thread is getting its own copy of the number! Not at all what we want!

If `remaining_tickets` were a non-Copy type, we would get an error when trying to give ownership to multiple threads

[Rust playground](#)

Attempt 2: Shared ownership

- We want to have *one* `remaining_tickets` counter that is shared between all threads
- ...but shared in a “safe” way (without data races)
- Rust allows shared ownership using *reference counting*
 - Take the thing you want to share and allocate it on the heap, along with a reference count
 - Whenever you share the object with another owner, increment the reference count



1

Attempt 2: Shared ownership

- We want to have *one* `remaining_tickets` counter that is shared between all threads
- Rust allows shared ownership using *reference counting*
 - Take the thing you want to share and allocate it on the heap, along with a reference count
 - Whenever you share the object with another owner, increment the reference count
 - Whenever an owner drops the object, decrement the reference count



2



Attempt 2: Shared ownership

- We want to have *one* `remaining_tickets` counter that is shared between all threads
- Rust allows shared ownership using *reference counting*
 - Take the thing you want to share and allocate it on the heap, along with a reference count
 - Whenever you share the object with another owner, increment the reference count
 - Whenever an owner drops the object, decrement the reference count
 - When the reference count hits 0, free the memory



1



Attempt 2: Shared ownership

- We want to have *one* `remaining_tickets` counter that is shared between all threads
- Rust allows shared ownership using *reference counting*
 - Take the thing you want to share and allocate it on the heap, along with a reference count
 - Whenever you share the object with another owner, increment the reference count
 - Whenever an owner drops the object, decrement the reference count
 - When the reference count hits 0, free the memory

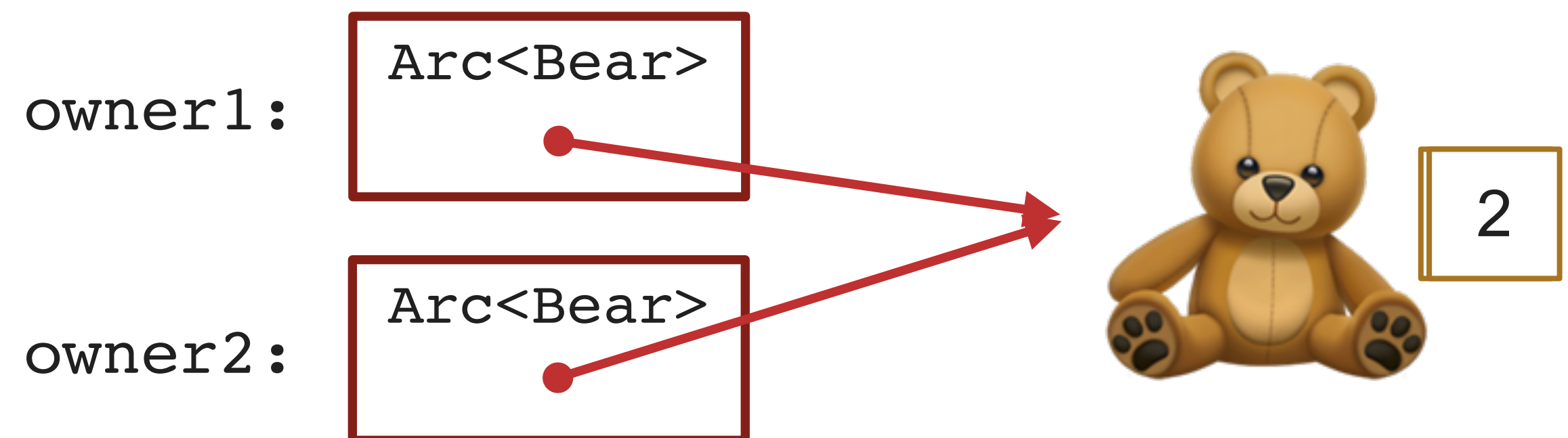


Note that this is **NOT A REFERENCE**. Entirely different implementation and functionality! References cannot outlive their owners, but with shared ownership, owners don't need to worry about each others' lifetimes

Attempt 2: Shared ownership

- We want to have *one* `remaining_tickets` counter that is shared between all threads
- Rust allows shared ownership using *reference counting*
 - Take the thing you want to share and allocate it on the heap, along with a reference count
 - Whenever you share the object with another owner, increment the reference count
 - Whenever an owner drops the object, decrement the reference count
 - When the reference count hits 0, free the memory
- [Arc type](#): Atomically Reference Counted
 - Atomic: safe for multithreaded use
 - You may see the [Rc](#) type used in non-multithreaded settings (not used in this class)

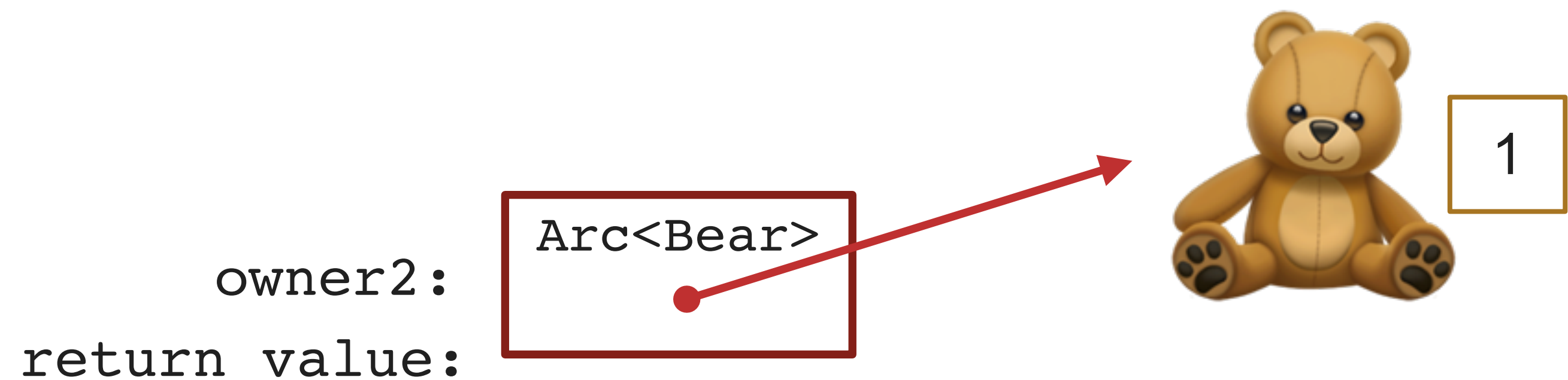
```
fn make_bear() -> Arc<Bear> {  
    let owner1 = Arc::new(Bear {});  
    let owner2 = owner1.clone();  
    return owner2;  
}
```



Attempt 2: Shared ownership

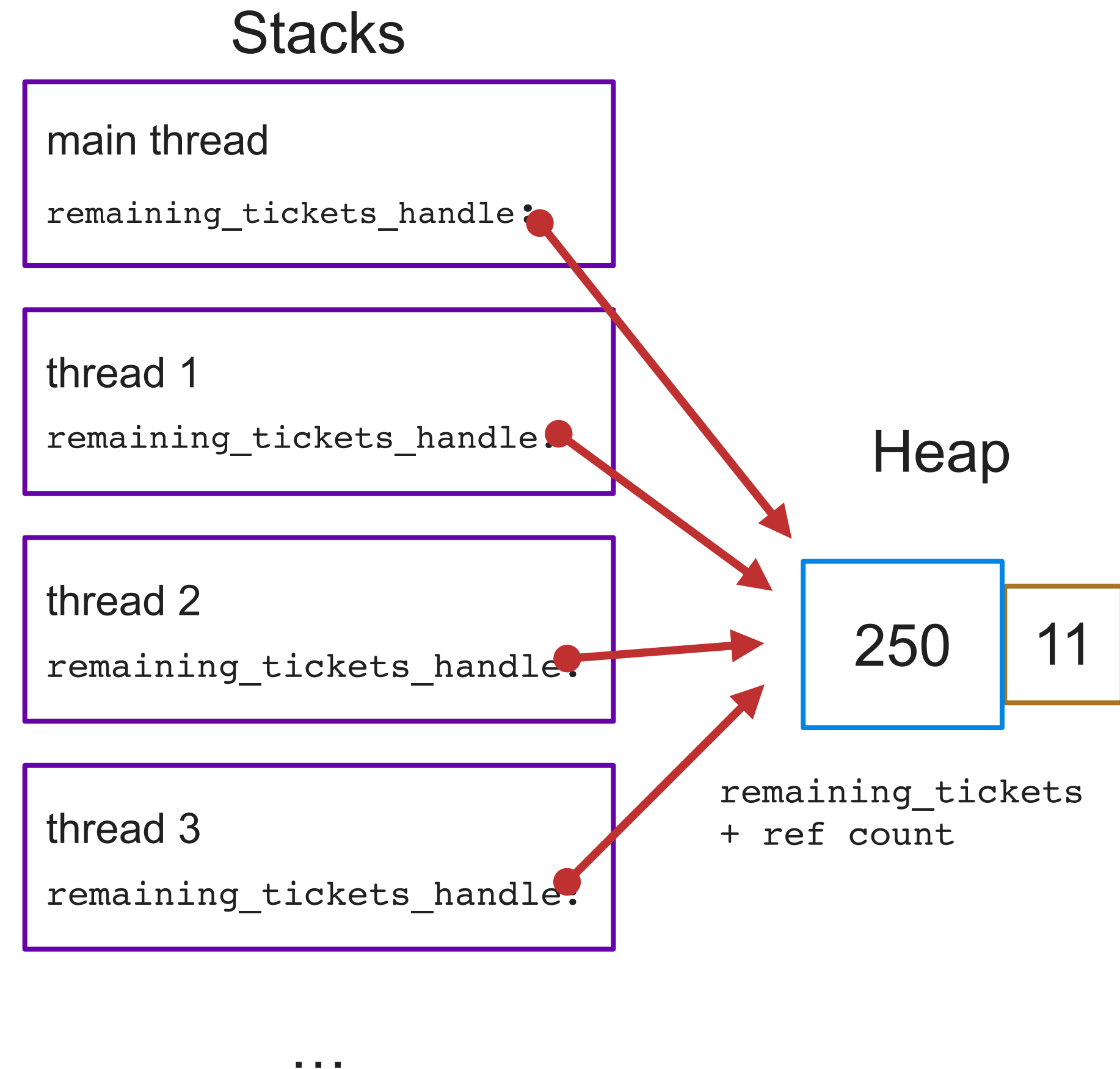
- We want to have *one* `remaining_tickets` counter that is shared between all threads
- Rust allows shared ownership using *reference counting*
 - Take the thing you want to share and allocate it on the heap, along with a reference count
 - Whenever you share the object with another owner, increment the reference count
 - Whenever an owner drops the object, decrement the reference count
 - When the reference count hits 0, free the memory
- [Arc type](#): Atomically Reference Counted
 - Atomic: safe for multithreaded use
 - You may see the [Rc](#) type used in non-multithreaded settings (not used in this class)

```
fn make_bear() -> Arc<Bear> {  
    let owner1 = Arc::new(Bear {});  
    let owner2 = owner1.clone();  
    return owner2;  
}
```



Attempt 2: Shared ownership

```
fn main() {  
    let remaining_tickets = Arc::new(250);  
  
    let mut threads = Vec::new();  
    for i in 0..10 {  
        let remaining_tickets_handle = remaining_tickets.clone();  
        threads.push(thread::spawn(move || {  
            ticket_agent(i, remaining_tickets_handle)  
        }));  
    }  
    // wait for all the threads to finish  
    for handle in threads {  
        handle.join().expect("Panic occurred in thread!");  
    }  
    println!("End of business day!");  
}
```



[Rust playground](#)

Problem: We can't modify data in an Arc!

```
error[E0594]: cannot assign to data in an `Arc`  
--> src/main.rs:24:9  
24 |         *remaining_tickets -= 1;  
   |         ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ cannot assign  
= help: trait `DerefMut` is required to modify through a dereference, but it is not implemented for `Arc<usize>`
```

- Arc allows us to have multiple owners, but multiple ownership is only safe if the data is immutable
 - Otherwise, we could have someone altering the bear while someone else is painting it
- We need a way to safely coordinate access so that if someone wants to modify the bear, we ensure no one else is currently using it

Attempt 3: Coordinated access with mutexes

- In Rust, the data goes *inside* the mutex
- The [Mutex](#) acts like a bathroom lock, where only one owner can pass at a time
- Unlike in C/C++, it is *impossible* to forget to lock a mutex! You can't access the data without going inside and locking the lock



Attempt 3: Coordinated access with mutexes

```
fn main() {  
    let remaining_tickets: Arc<Mutex<usize>>  
        = Arc::new(Mutex::new(250));  
  
    let mut threads = Vec::new();  
    for i in 0..10 {  
        let remaining_tickets_handle = remaining_tickets.clone();  
        threads.push(thread::spawn(move || {  
            ticket_agent(i, remaining_tickets_handle);  
        }));  
    }  
    // wait for all the threads to finish  
    for handle in threads {  
        handle.join().expect("in thread!");  
    }  
}
```

Takes care of
reference counting
(Drop when no longer
in use)

Takes care of
mutability — only
one thread can
write at the same
time.

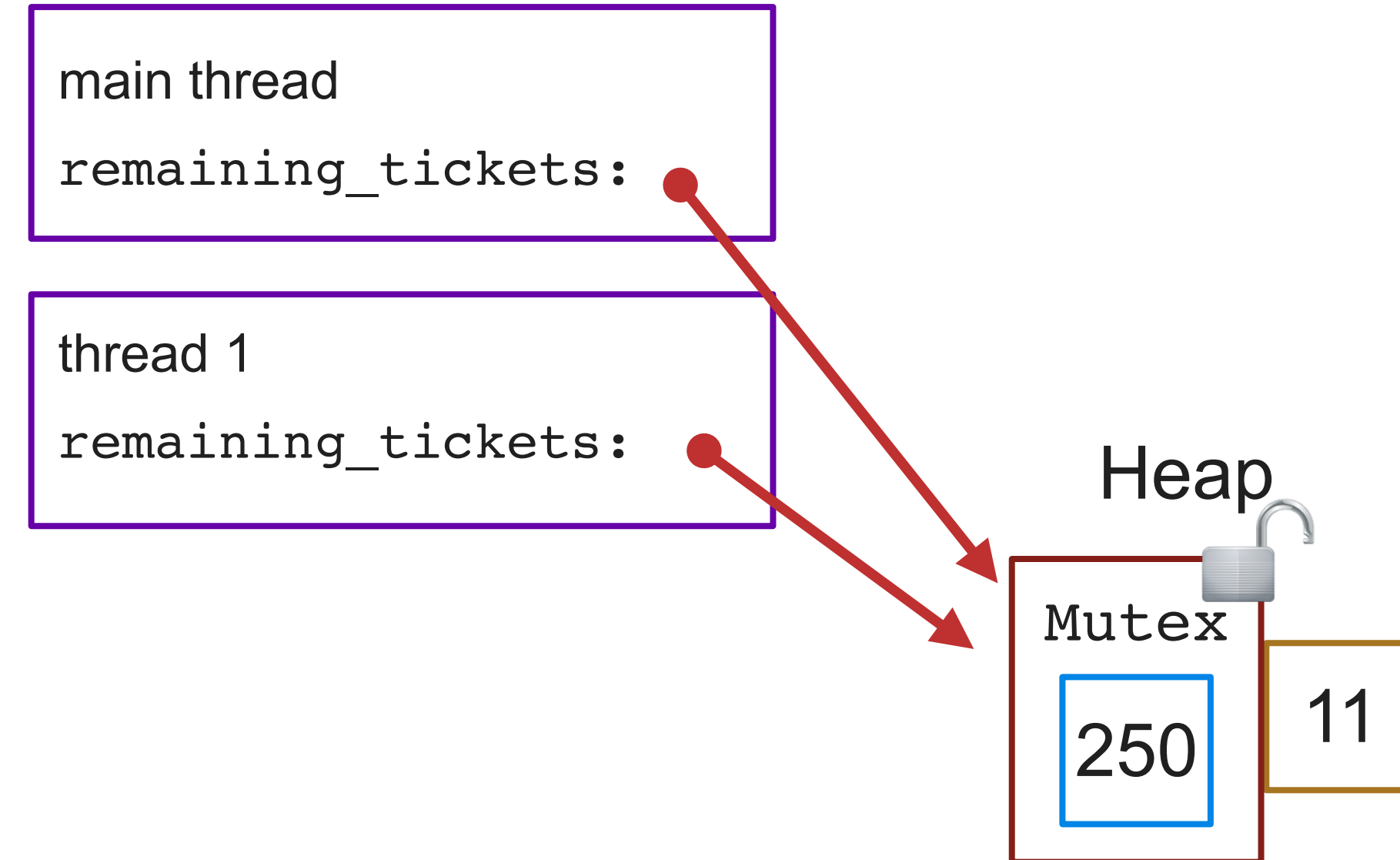
[Rust playground](#)



Attempt 3: Coordinated access with mutexes

```
fn ticket_agent(id: usize, remaining_tickets: Arc<Mutex<usize>>) {  
    loop {  
        let mut remaining_tickets_ref =  
            remaining_tickets.lock().unwrap();  
        if *remaining_tickets_ref == 0 {  
            break;  
        }  
        handle_call();  
        *remaining_tickets_ref -= 1;  
        println!("Agent #{} sold a ticket! ({} more to be sold)",  
            id, *remaining_tickets_ref);  
        if should_take_break() {  
            take_break();  
        }  
    }  
  
    println!("Agent #{} notices all tickets are sold, and goes home!", id);  
}
```

Stacks



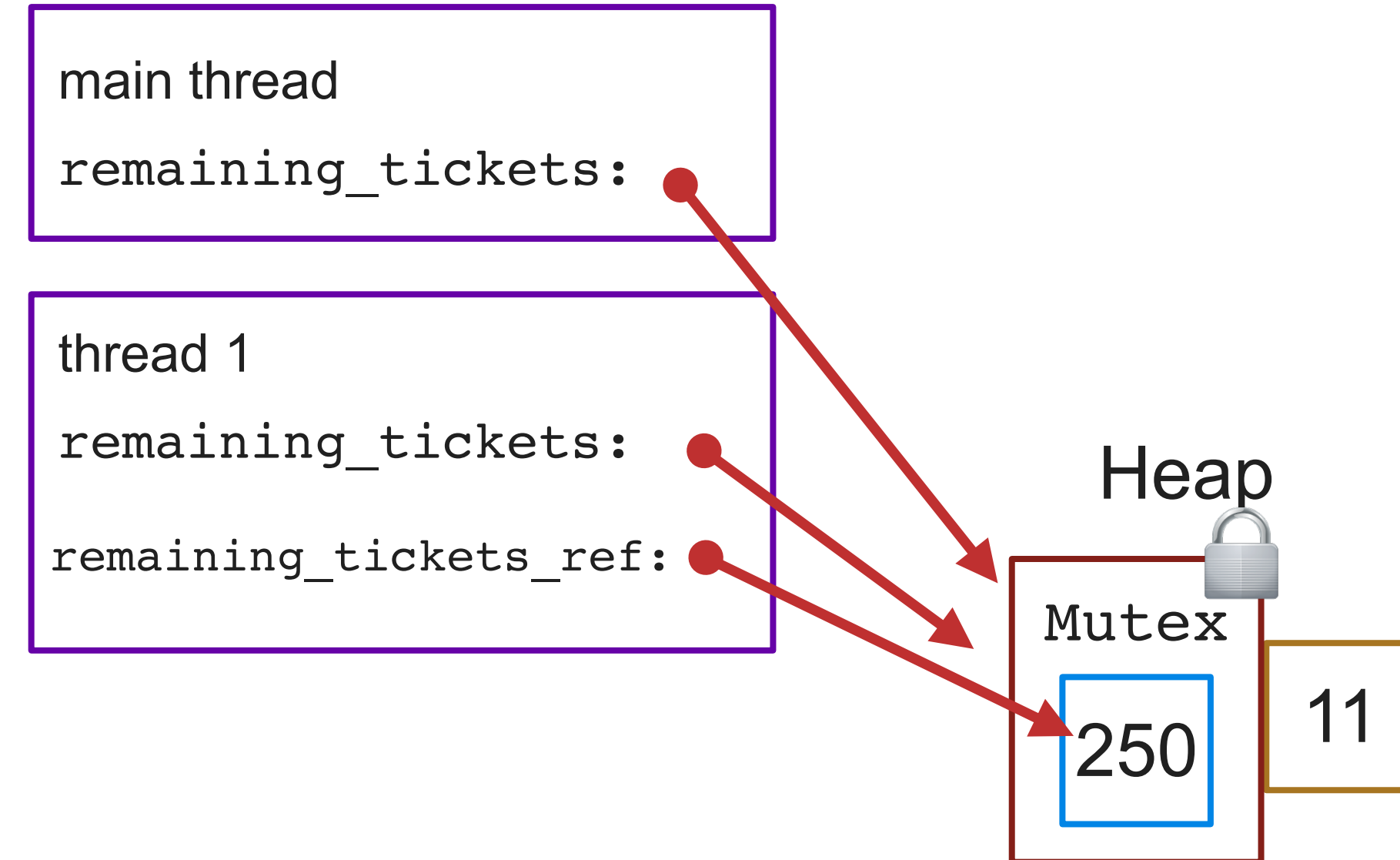
[Rust playground](#)

[Mutex documentation](#)

Attempt 3: Coordinated access with mutexes

```
fn ticket_agent(id: usize, remaining_tickets: Arc<Mutex<usize>>) {  
    loop {  
        let mut remaining_tickets_ref =  
            remaining_tickets.lock().unwrap();  
        if *remaining_tickets_ref == 0 {  
            break;  
        }  
        handle_call();  
        *remaining_tickets_ref -= 1;  
        println!("Agent #{} sold a ticket! ({} more to be sold)",  
            id, *remaining_tickets_ref);  
        if should_take_break() {  
            take_break();  
        }  
    }  
    remaining_tickets_ref dropped at  
    end of scope, lock is unlocked  
    println!("Agent #{} notices all tickets are sold, and goes home!", id);  
}
```

Stacks

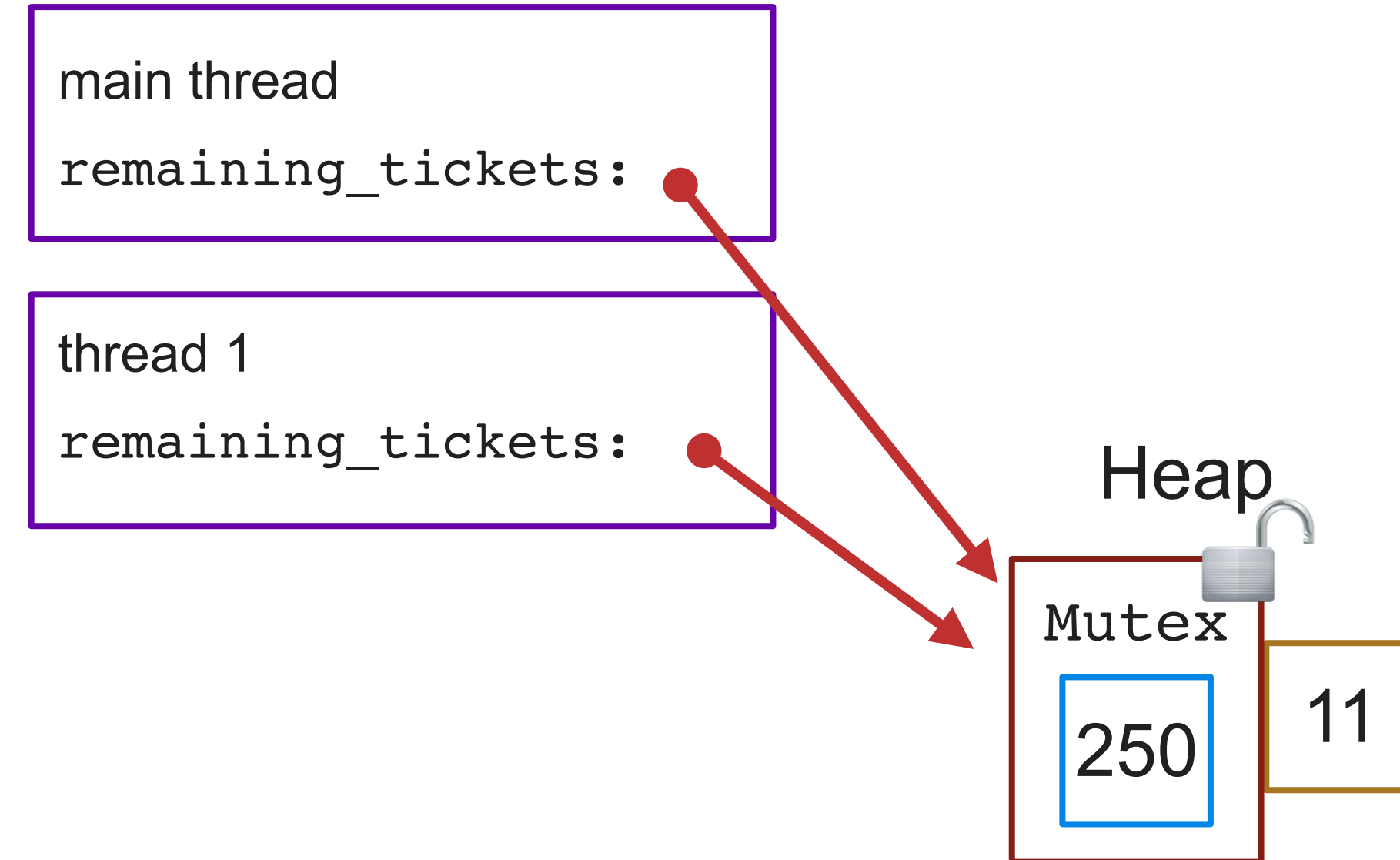


[Rust playground](#)

Attempt 3: Coordinated access with mutexes

```
fn ticket_agent(id: usize, remaining_tickets: Arc<Mutex<usize>>) {  
    loop {  
        let mut remaining_tickets_ref =  
            remaining_tickets.lock().unwrap();  
        if *remaining_tickets_ref == 0 {  
            break;  
        }  
        handle_call();  
        *remaining_tickets_ref -= 1;  
        println!("Agent #{} sold a ticket! ({} more to be sold)",  
            id, *remaining_tickets_ref);  
        if should_take_break() {  
            take_break();  
        }  
    }  
    remaining_tickets_ref dropped at  
    end of scope, lock is unlocked  
  
    println!("Agent #{} notices all tickets are sold, and goes home!", id);  
}
```

Stacks



Can't forget to unlock the lock 👍
But this code is completely serialized!!

[Rust playground](#)

Attempt 4: Releasing lock early

```
fn ticket_agent(id: usize, remaining_tickets: Arc<Mutex<usize>>) {  
    loop {  
        let mut remaining_tickets_ref =  
            remaining_tickets.lock().unwrap();  
        if *remaining_tickets_ref == 0 {  
            break;  
        }  
        handle_call();  
        *remaining_tickets_ref -= 1;  
        println!("Agent #{} sold a ticket! ({} more to be sold)",  
            id, *remaining_tickets_ref);  
        drop(remaining_tickets_ref);  
        if should_take_break() {  
            take_break();  
        }  
    }  
  
    println!("Agent #{} notices all tickets are sold, and goes home!", id);  
}
```

[Rust playground](#)

Note: alternate syntax - “create” a scope to induce `drop`

```
fn ticket_agent(id: usize, remaining_tickets: Arc<Mutex<usize>>) {  
    loop {  
        {  
            let mut remaining_tickets_ref =  
                remaining_tickets.lock().unwrap();  
            if *remaining_tickets_ref == 0 {  
                break;  
            }  
            handle_call();  
            *remaining_tickets_ref -= 1;  
            println!("Agent #{} sold a ticket! ({} more to be sold)",  
                id, *remaining_tickets_ref);  
        }  
        if should_take_break() {  
            take_break();  
        }  
    }  
  
    println!("Agent #{} notices all tickets are sold, and goes home!", id);  
}
```

Alternative: decomposition! E.g., create new scope by creating a new function. Example implementation [here](#). This is what we recommend in the week 7 exercises.

remaining_tickets_ref dropped at end of scope, lock is unlocked

Note: documentation for Mutex

- [Lock](#): “Acquires a mutex, blocking the current thread until it is able to do so.”
- Returns [LockResult](#) (a Result type)
- Ok variant will contain a [MutexGuard](#)
 - A “scoped lock” on a mutex: when this structure is dropped, the lock will be unlocked
 - Data protected by the mutex can be accessed by dereferencing.
- Error variant indicates the mutex is “poisoned”:
 - Poisoning = another thread panicked while holding the mutex.
 - Data in the mutex may be in a corrupted state.
 - In the previous example, we unwrapped (crash/exit if mutex is poisoned), but you could handle this error case differently.
- See also - [lock_guard](#) in C++!

Still need to put `drop` in the right place

- Race condition issue 2: `remaining_tickets--` concurrently
 - Can no longer happen in this code. Must lock mutex in order to modify the value. (Like `atomic` wrapper in C++.)
- Race condition issue 1: context switch between check (`remaining_tickets == 0`) and decrement.
 - Can still happen if you release the lock early — even if both of these operations, separately, are protected.

Still need to put `drop` in the right place

```
fn ticket_agent(id: usize, remaining_tickets: Arc<Mutex<usize>>) {
    loop {
        let remaining_tickets_ref =
            remaining_tickets.lock().unwrap();
        if *remaining_tickets_ref == 0 {
            break;
        }
        drop(remaining_tickets_ref);
        handle_call();
        let mut remaining_tickets_ref = remaining_tickets.lock().unwrap();
        *remaining_tickets_ref -= 1;
        println!("Agent #{} sold a ticket! ({} more to be sold)",
            id, *remaining_tickets_ref);
        drop(remaining_tickets_ref);
        if should_take_break() {
            take_break();
        }
    }
}

println!("Agent #{} notices all tickets are sold, and goes home!", id);
}
```

*This sometimes works...
and sometimes crashes
with an "underflow error".*

Summary

- Rust does not prevent all race conditions, but it does prevent *data races*
 - *Multiple threads access a value, where at least one of them is writing*
 - *Most common type of race condition in systems programming — big win!*
 - *This is also a huge advantage over other memory-safe languages. Garbage collection provides memory safety but not thread safety*
- You still need to be sure to drop in the right place
- You still must be careful to avoid inadvertently serializing your code
- Deadlock can still be a problem

Week 7 exercises

- Very short! Get to know the syntax. Due next Wednesday.
- [Here's](#) some commented code for the ticket agents problem. Use it as an example for:
 - Spawning and joining threads
 - Establishing and moving data — should be *shareable* and *mutable* and *protected* (from data races) across multiple threads.
 - Accessing this data within threads by “locking” the mutex — without serializing the code and without introducing race conditions.
- Next week: more multithreading practice!