CSE113: Parallel Programming

March 8, 2023

- **Topics**:
  - Memory consistency models:
    - Examples

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

Announcements

- Striving for HW 2 grades to be out by end of today.
- Work on Homework 4 (Due March 10, you have until March 14)
- We are working on HW 5 to be released by end of the week or Monday
- Last day on Module 4, we will move to Module 5 on Friday (Javascript and GPU)

Memory Consistency

- We have been very strict about using atomic types in this class
  - and the methods (.load and .store)
  - why?

  - Architectures do very strange things with memory loads and stores
  - Compilers do too (but we won’t talk too much about them today)

  - C++ gives us sequential consistency if we use atomic types and operations
  - What do we remember sequential consistency from?

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

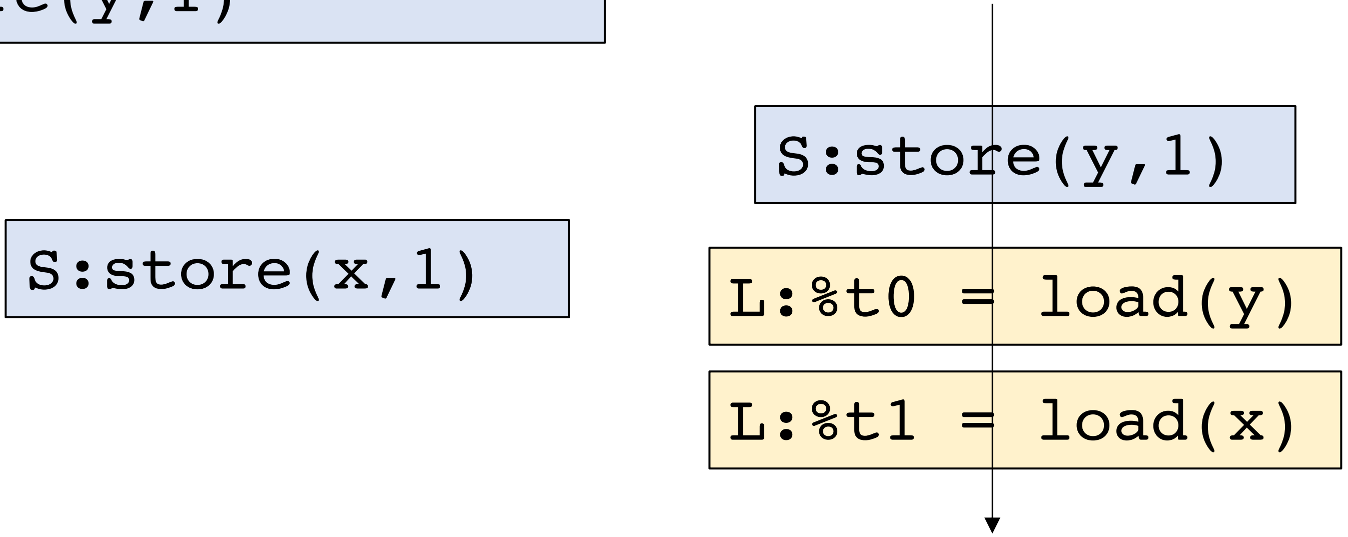
Question: can `t0 == 1 and t1 == 0?`

Thread 0:

```
S:store(x,1)
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
S:%t1 = load(x)
```



```
S:store(y,1)
S:store(x,1)
L:%t0 = load(y)
L:%t1 = load(x)
```

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

Question: can `t0 == 1 and t1 == 0`?

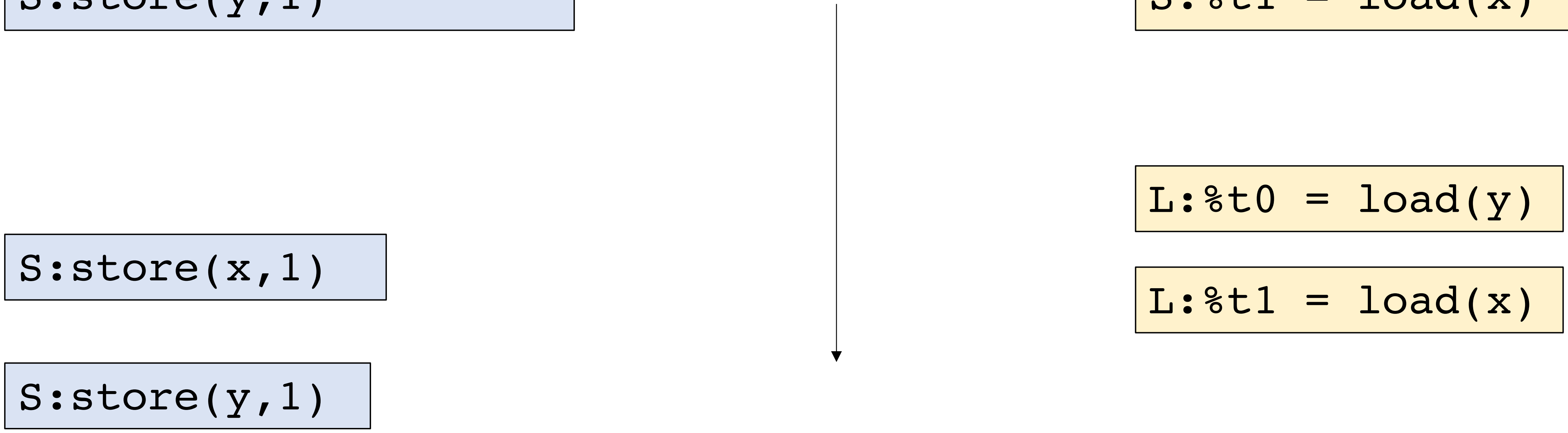
start off thinking about sequential consistency

Thread 0:

```
S:store(x,1)
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
S:%t1 = load(x)
```



```
S:store(x,1)
S:store(y,1)
```

```
L:%t0 = load(y)
L:%t1 = load(x)
```

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

Question: can `t0 == 1 and t1 == 0`?

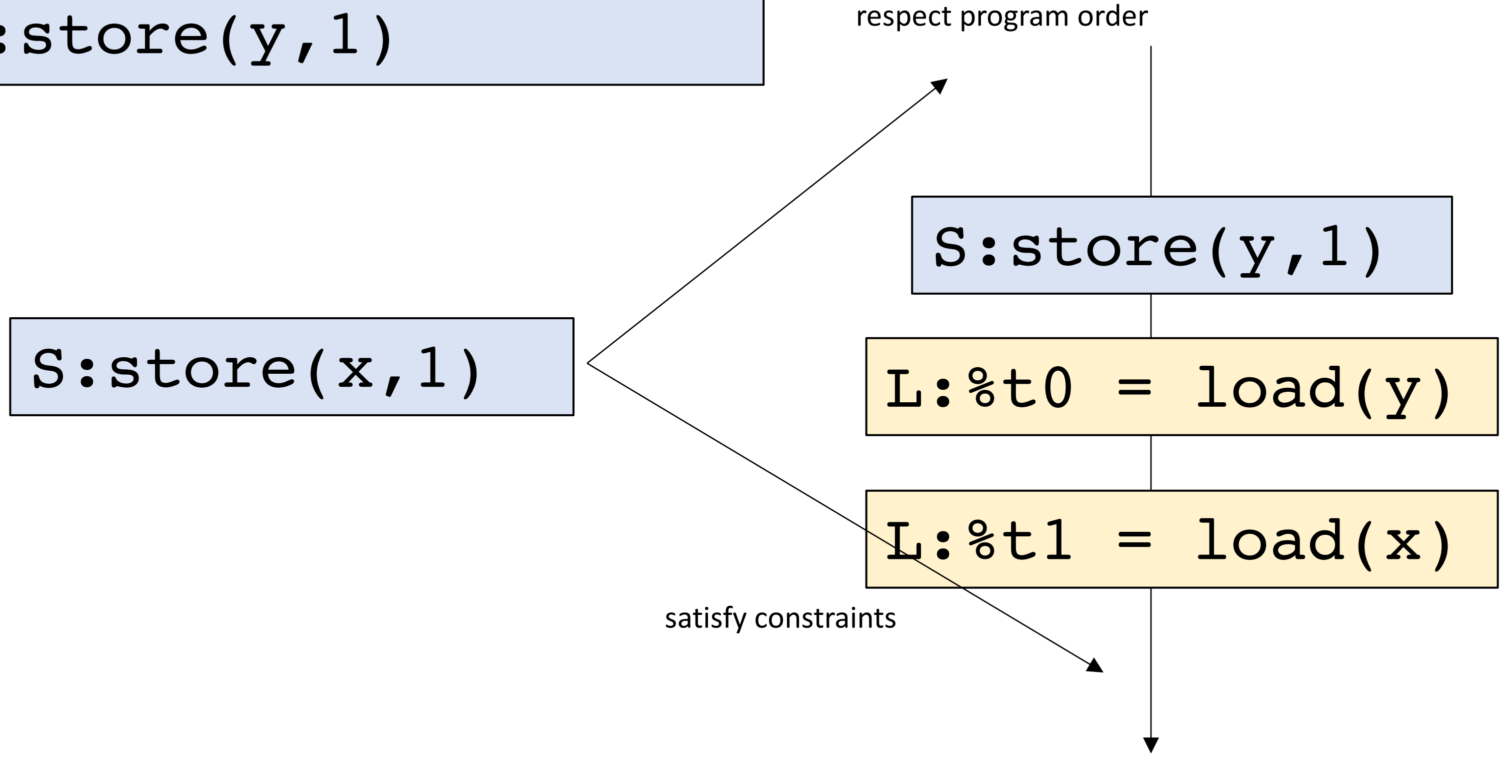
start off thinking about sequential consistency

Thread 0:

```
S:store(x,1)
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
S:%t1 = load(x)
```



Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

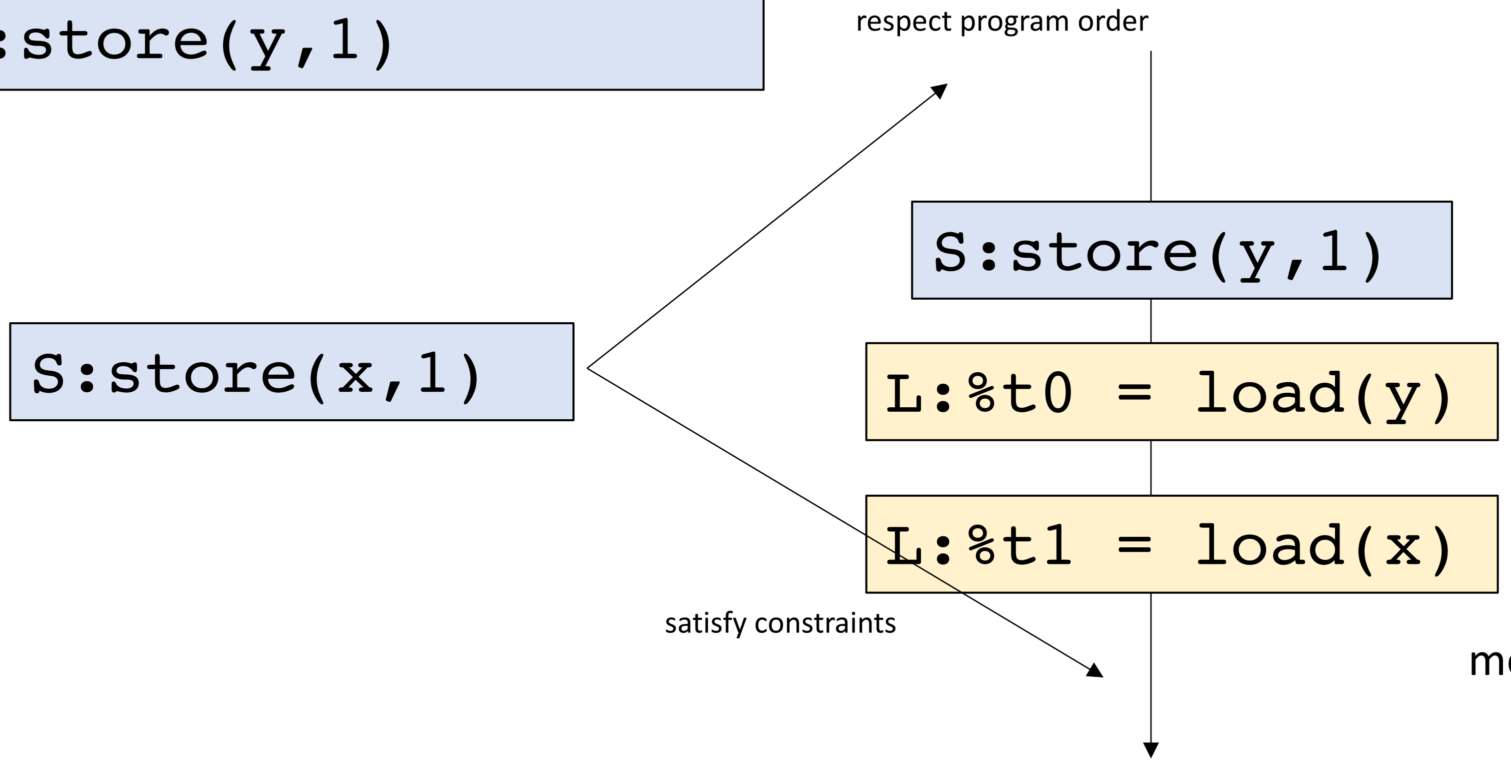
Question: can `t0 == 1 and t1 == 0`?

Thread 0:

```
S:store(x,1)
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
L:%t1 = load(x)
```



```
S:store(x,1)
```

```
S:store(y,1)
L:%t0 = load(y)
L:%t1 = load(x)
```

memory access 0

memory access 1

	L	S
L	NO	Different address
S	NO	NO

What about TSO?

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

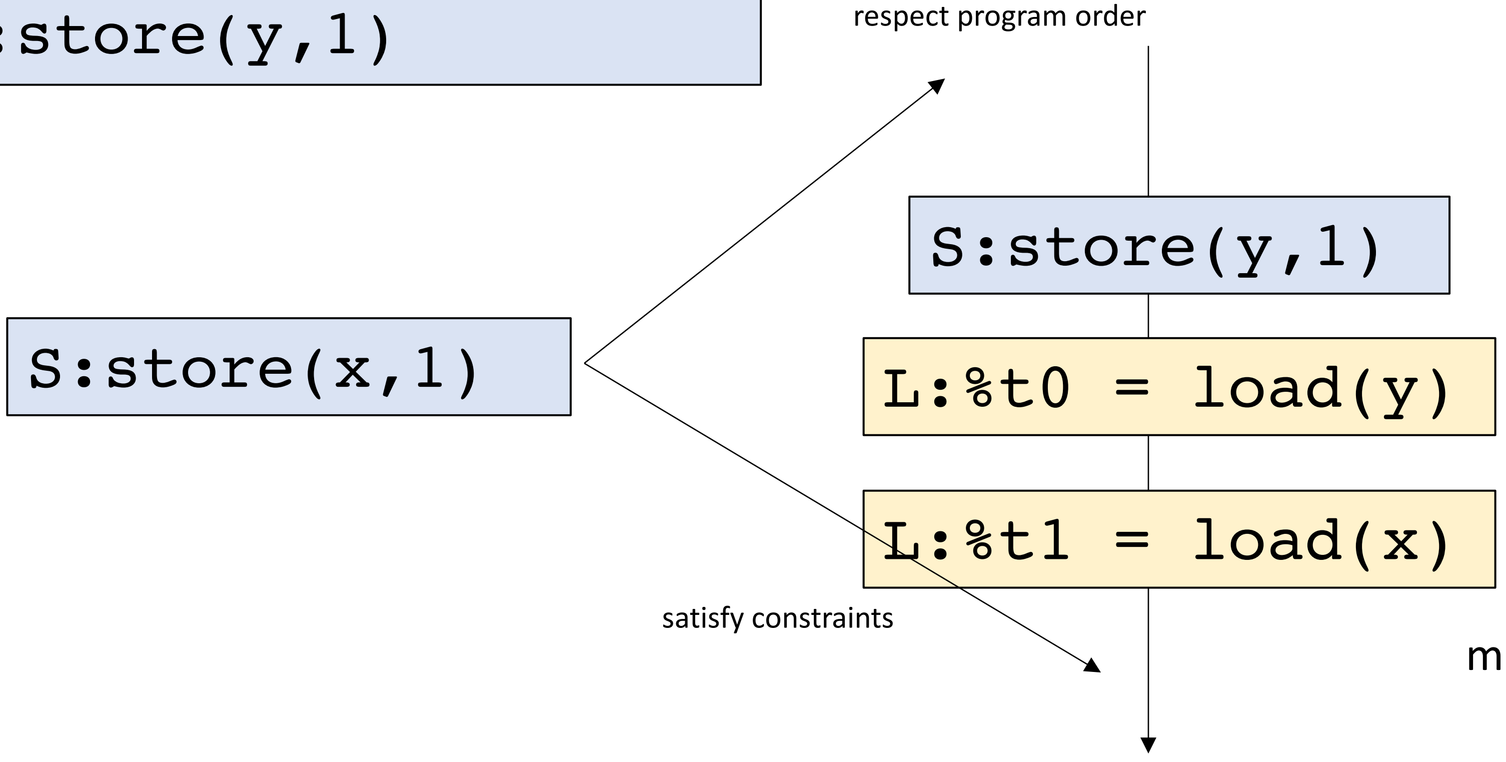
Question: can `t0 == 1 and t1 == 0`?

Thread 0:

```
S:store(x,1)
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
L:%t1 = load(x)
```



memory access 0

		L	S
memory access 1	L	NO	Different address
	S	NO	NO

What about TSO? NO

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

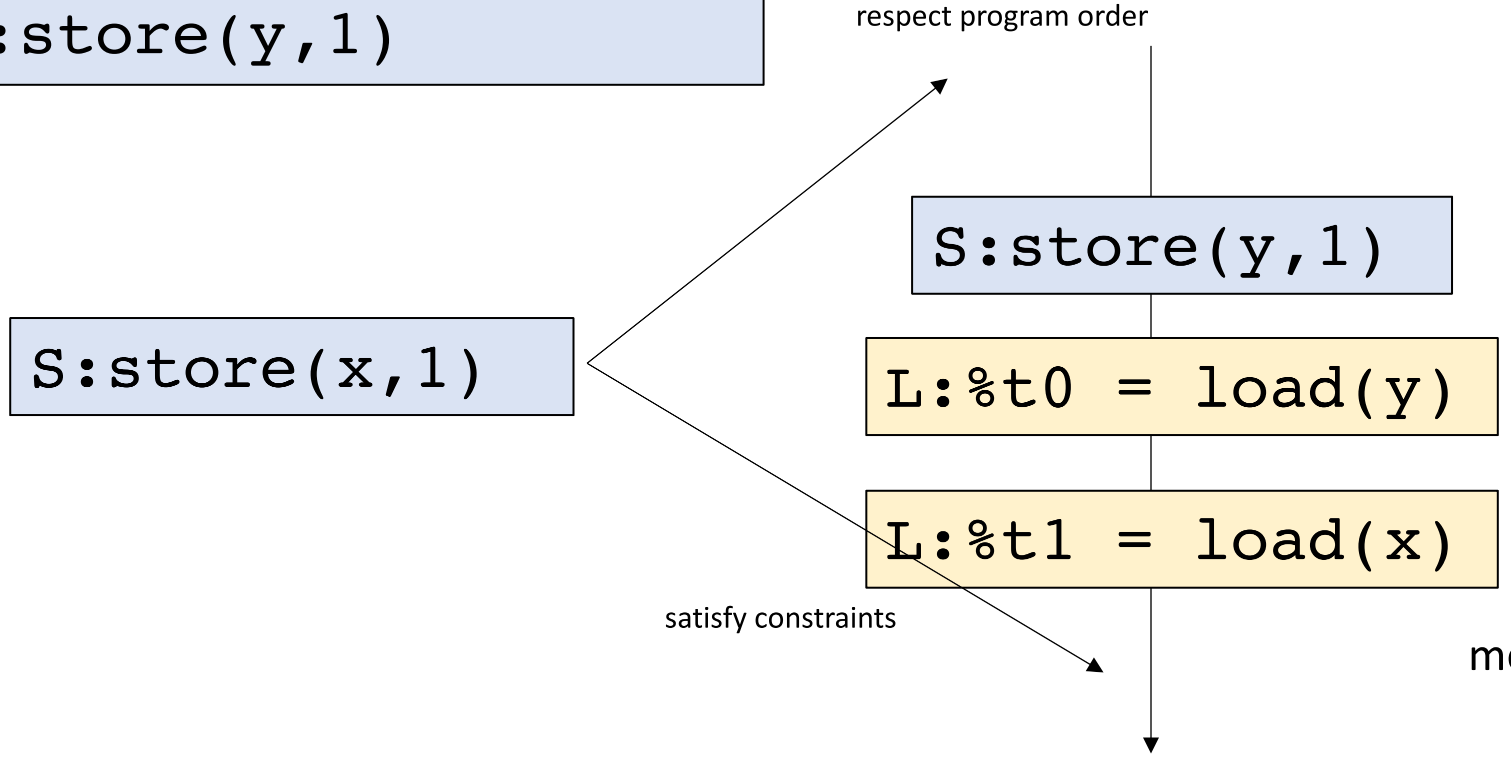
Question: can `t0 == 1 and t1 == 0`?

Thread 0:

```
S:store(x,1)
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
L:%t1 = load(x)
```



memory access 1 \ memory access 0	L	S
L	NO	Different address
S	NO	Different address

What about PSO?

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

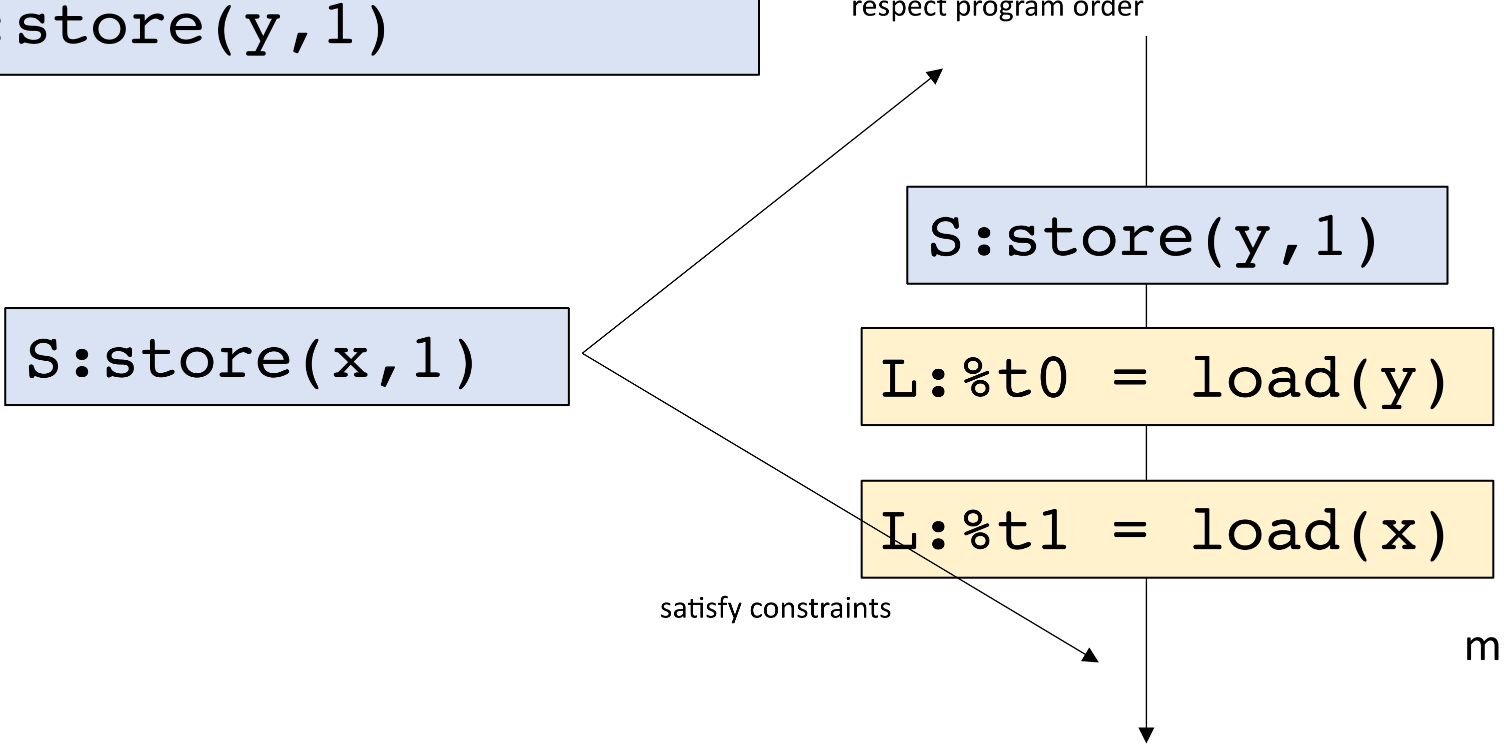
Question: can `t0 == 1 and t1 == 0?`

Thread 0:

```
S:store(x,1)
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
S:%t1 = load(x)
```



```
S:store(y,1)
L:%t0 = load(y)
L:%t1 = load(x)
```

```
S:store(x,1)
```

memory access 1 \ memory access 0	L	S
L	NO	Different address
S	NO	🟩 Different address

What about PSO?

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

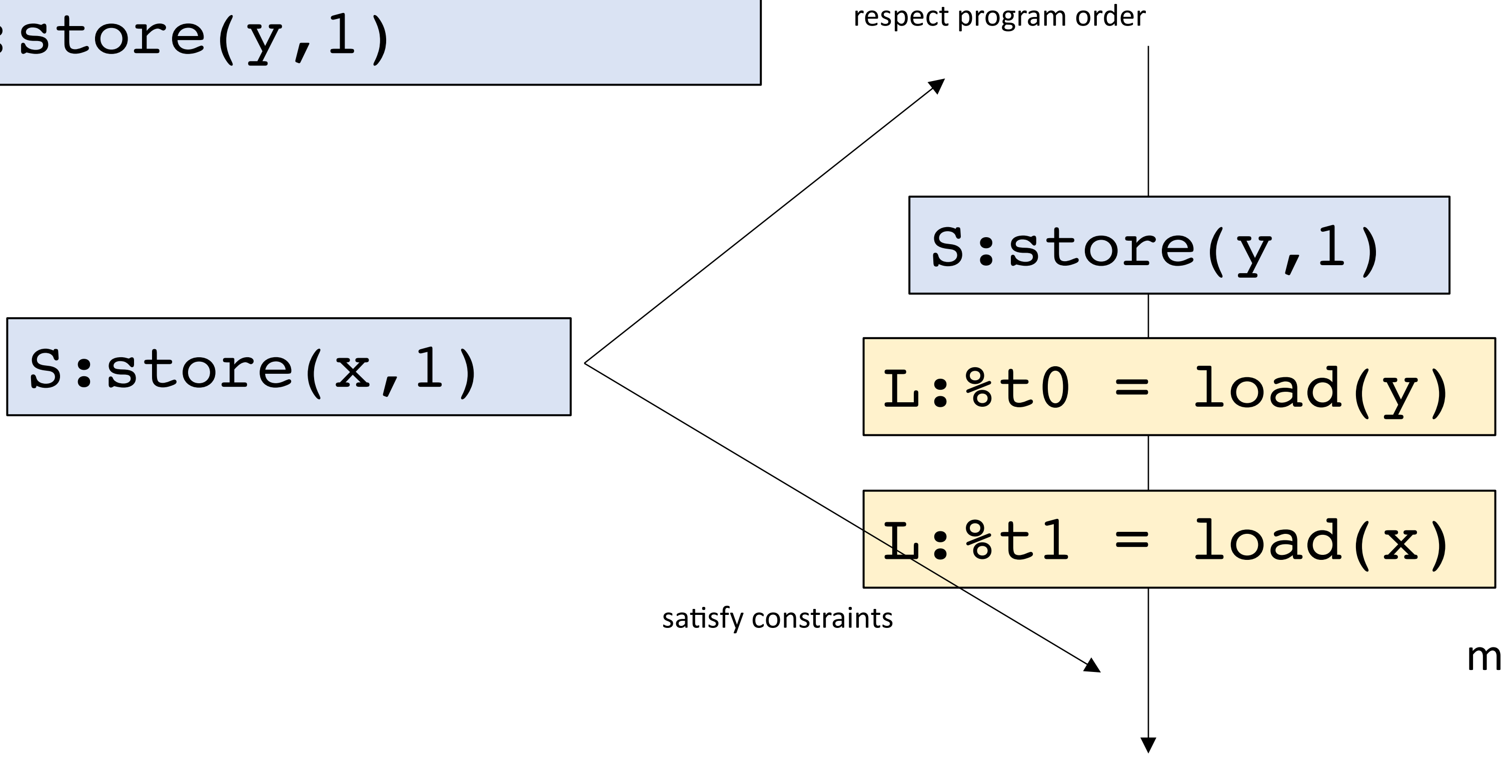
Question: can `t0 == 1 and t1 == 0`?

Thread 0:

```
S:store(x,1)
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
S:%t1 = load(x)
```



		memory access 0	
		L	S
memory access 1	L	NO	Different address
	S	NO	Different address

What about PSO? YES

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

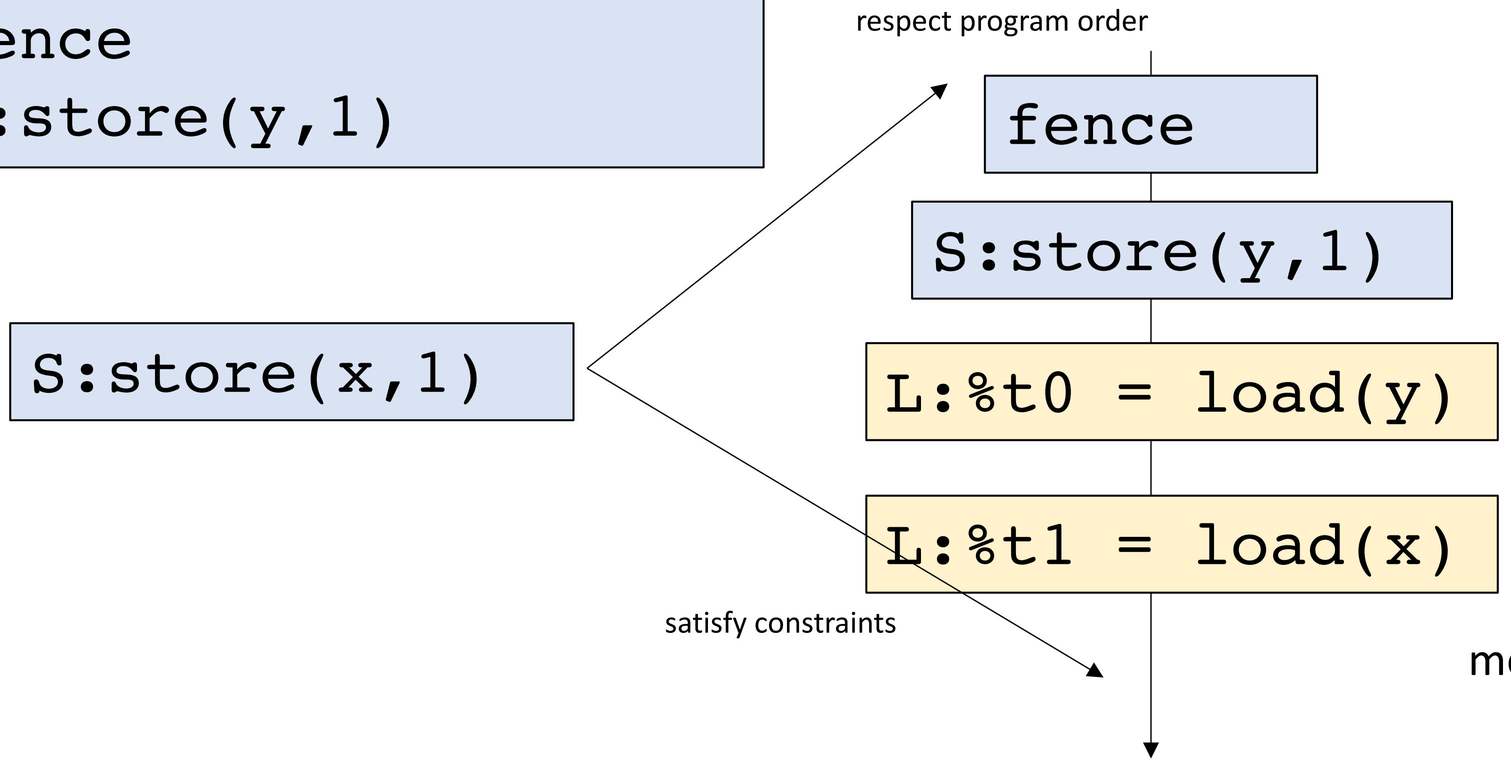
Question: can `t0 == 1 and t1 == 0`?

Thread 0:

```
S:store(x,1)
fence
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
L:%t1 = load(x)
```



		memory access 0	
		L	S
memory access 1	L	NO	Different address
	S	NO	Different address

Now it is disallowed in PSO

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

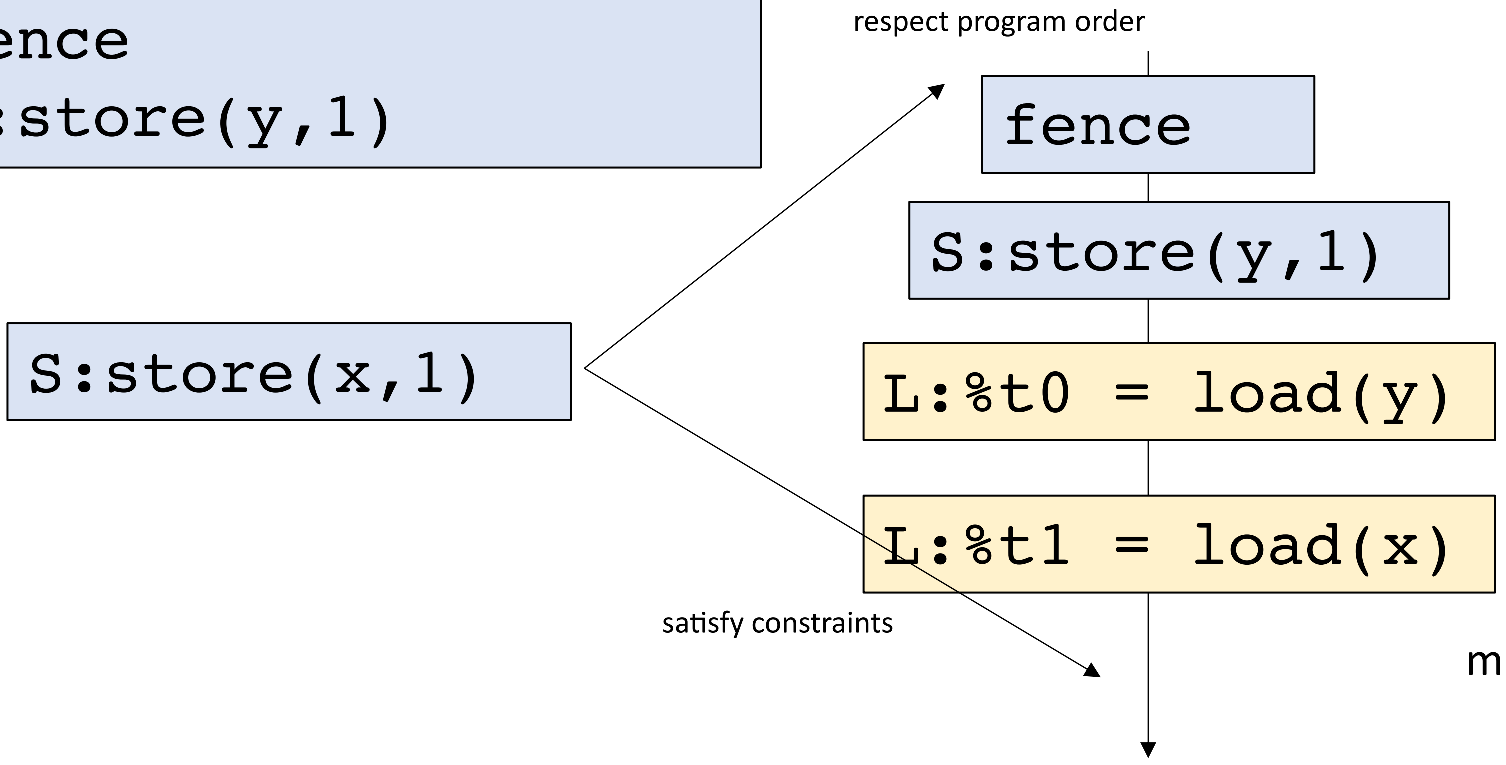
Question: can `t0 == 1 and t1 == 0?`

Thread 0:

```
S:store(x,1)
fence
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
S:%t1 = load(x)
```



memory access 0 (columns), memory access 1 (rows)

	L	S
L	YES	Different address
S	Different address	Different address

What about RMO?

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

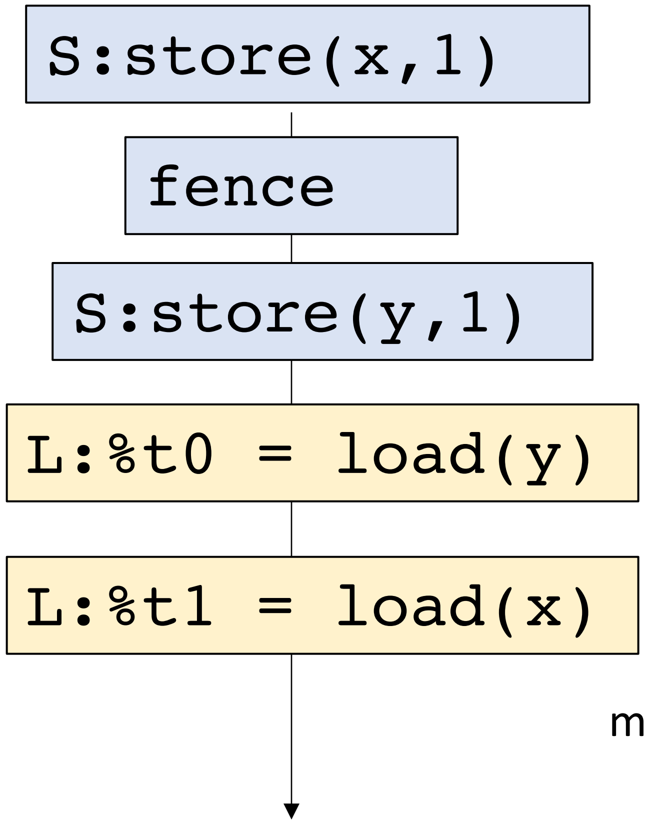
Question: can `t0 == 1 and t1 == 0?`

Thread 0:

```
S:store(x,1)
fence
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
S:%t1 = load(x)
```



```
S:store(x,1)
fence
S:store(y,1)
L:%t0 = load(y)
L:%t1 = load(x)
```

		memory access 0	
		L	S
memory access 1	L	YES	Different address
	S	Different address	Different address

What about RMO?

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

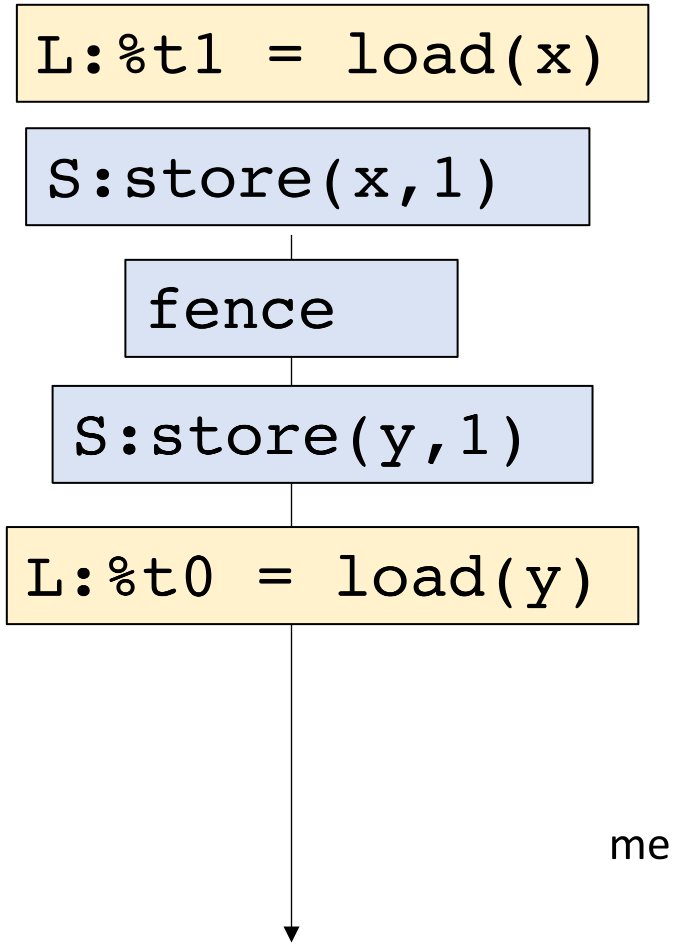
Question: can `t0 == 1 and t1 == 0`?

Thread 0:

```
S:store(x,1)
fence
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
S:%t1 = load(x)
```



```
L:%t1 = load(x)
S:store(x,1)
fence
S:store(y,1)
L:%t0 = load(y)
```

		memory access 0	
		L	S
memory access 1	L	YES	Different address
	S	Different address	Different address

What about RMO? The loads can be reordered also!

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

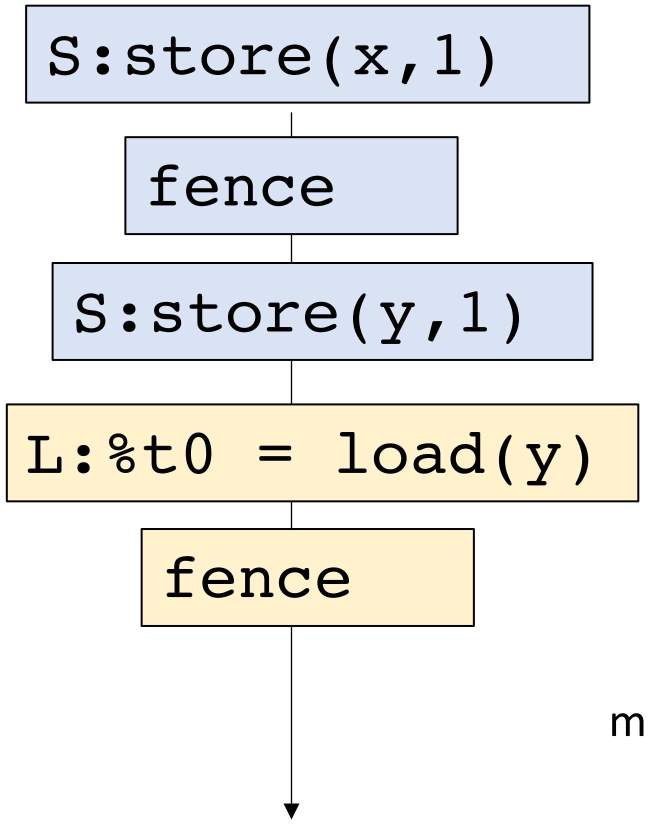
Question: can `t0 == 1 and t1 == 0`?

Thread 0:

```
S:store(x,1)
fence
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
fence
S:%t1 = load(x)
```



```
S:store(x,1)
fence
S:store(y,1)
L:%t0 = load(y)
fence
```

```
L:%t1 = load(x)
```

memory access 0

		L	S
memory access 1	L	🟩 YES	Different address
	S	Different address	🟩 Different address

What about RMO? add a fence

Global variable:

```
int x[1] = {0};
int y[1] = {0};
```

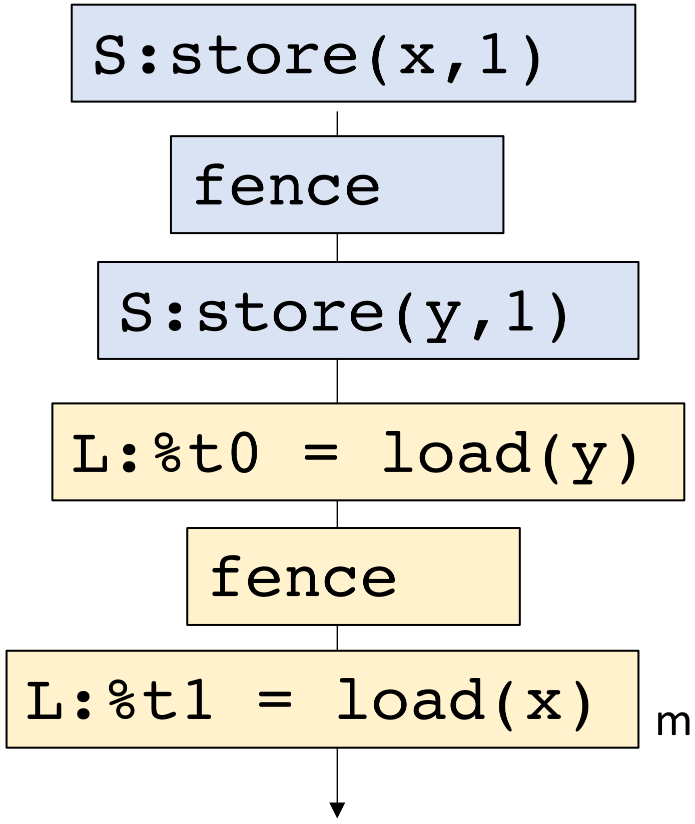
Question: can `t0 == 1 and t1 == 0?`

Thread 0:

```
S:store(x,1)
fence
S:store(y,1)
```

Thread 1:

```
L:%t0 = load(y)
fence
S:%t1 = load(x)
```



memory access 1

memory access 0

	L	S
L	🟩 YES	Different address
S	Different address	🟩 Different address

Now the relaxed behavior is disallowed

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

target machine

	L	S
L	?	?
S	?	?

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

find mismatch

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

find mismatch

Two options:

make sure stores
are not reordered
with later loads

make sure loads
are not reordered
with earlier stores

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

C++	ISA
`x.store(1);`	`store(x,1);` `fence;`

$$\begin{tikzcd} \boxed{\texttt{x.store(1);}} \arrow[r] & \boxed{\begin{array}{l}\texttt{store(x,1);}\\ \texttt{fence;}\end{array}} \end{tikzcd}$$

or

target machine
TSO (x86)

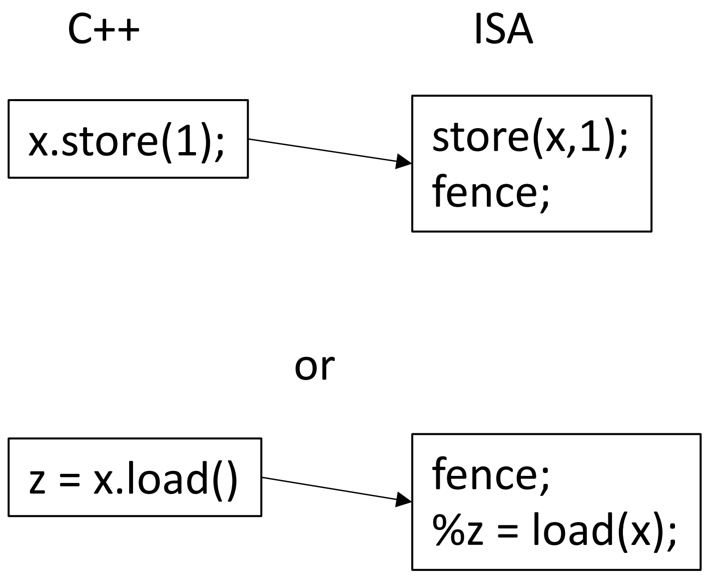
	L	S
L	NO	different address
S	NO	No

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO



```
x.store(1);
```

```
store(x,1);
fence;
```

or

```
z = x.load()
```

```
fence;
%z = load(x);
```

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

C++	ISA

x.store(1);	store(x,1);
	fence;

or

C++	ISA

z = x.load()	fence;
	%z = load(x);

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

This should help you see why you want to reduce the number of atomic load/stores in your program

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

How about this one?

target machine
PSO

	L	S
L	NO	different address
S	NO	different address

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

target machine
PSO

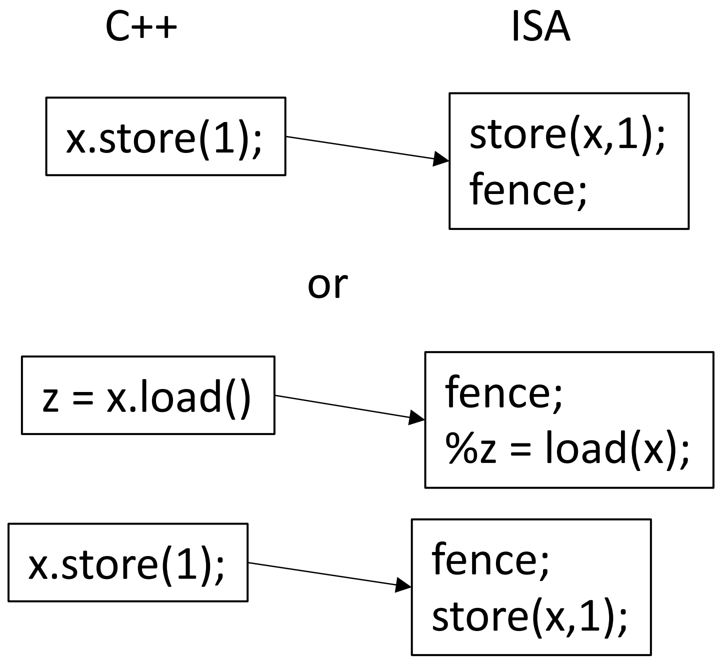
	L	S
L	NO	different address
S	NO	different address

C++11 atomic operation compilation

start with both both of the grids for the two different memory models

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO



```
x.store(1);    →    store(x,1);
                    fence;
```

or

```
z = x.load()   →    fence;
                    %z = load(x);

x.store(1);    →    fence;
                    store(x,1);
```

target machine
PSO

	L	S
L	NO	different address
S	NO	different address

Memory orders

- Atomic operations take an additional “memory order” argument
  - `memory_order_seq_cst` - default
  - `memory_order_relaxed` - weakest

Where have we seen `memory_order_relaxed`?

Relaxed memory order

language
C++11 (sequential consistency)

	L	S
L	NO	NO
S	NO	NO

language
C++11 (memory_order_relaxed)

	L	S
L	different address	different address
S	different address	different address

basically no orderings except for accesses to the same address

Compiling memory order relaxed

language
C++11 (memory_order_relaxed)

	L	S
L	different address	different address
S	different address	different address

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

Compiling memory order relaxed

language
C++11 (memory_order_relaxed)

	L	S
L	different address	different address
S	different address	different address

lots of mismatches!

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

Compiling memory order relaxed

language
C++11 (memory_order_relaxed)

	L	S
L	different address	different address
S	different address	different address

lots of mismatches!

But language is more relaxed than machine

so no fences are needed

target machine
TSO (x86)

	L	S
L	NO	different address
S	NO	No

Compiling memory order relaxed

Do any of the ISA memory models need any fences for relaxed memory order?

language
C++11 (memory_order_relaxed)

	L	S
L	different address	different address
S	different address	different address

TSO

	L	S
L	NO	Different address
S	NO	NO

PSO

	L	S
L	NO	Different address
S	NO	Different address

RMO

	L	S
L	YES	Different address
S	Different address	Different address

Memory order relaxed

- Very few use-cases! Be very careful when using it
  - Peeking at values (later accessed using a heavier memory order)
  - Counting (e.g. number of finished threads in work stealing)
  - ***DO NOT USE FOR QUEUE INDEXES***

More memory orders: we will not discuss in class

- Atomic operations take an additional “memory order” argument
  - `memory_order_seq_cst` - default
  - `memory_order_relaxed` - weakest
- More memory orders (useful for mutex implementations):
  - `memory_order_acquire`
  - `memory_order_release`
- EVEN MORE memory orders (complicated: in most research it is omitted)
  - `memory_order_consume`

A cautionary tale

Consider the following example: a graphics program where each thread wants to display a triangle; the display is a queue (not thread safe)

Thread 0:
```
m.lock();
display.enq(triangle0);
m.unlock();
```

Thread 1:
```
m.lock();
display.enq(triangle1);
m.unlock();
```

Consider the following example: a graphics program where each thread wants to display a triangle; the display is a queue (not thread safe)

Thread 0:

```
m.lock();
display.enq(triangle0);
m.unlock();
```

Thread 1:

```
m.lock();
display.enq(triangle1);
m.unlock();
```

We know how lock and unlock are implemented

Consider the following example: a graphics program where each thread wants to display a triangle; the display is a queue (not thread safe)

__Thread 0:__

```
SPIN:CAS(mutex,0,1);
display.enq(triangle0);
store(mutex,0);
```

__Thread 1:__

```
SPIN:CAS(mutex,0,1);
display.enq(triangle1);
store(mutex,0);
```

We know how lock and unlock are implemented
We also know how a queue is implemented

Consider the following example: a graphics program where each thread wants to display a triangle; the display is a queue (not thread safe)

<u>Thread 0:</u>

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

<u>Thread 1:</u>

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```

We know how lock and unlock are implemented
We also know how a queue is implemented

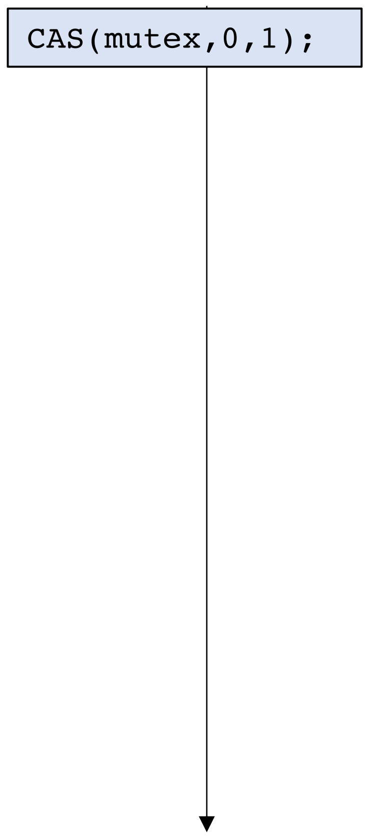
What is an execution?

Thread 0:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```

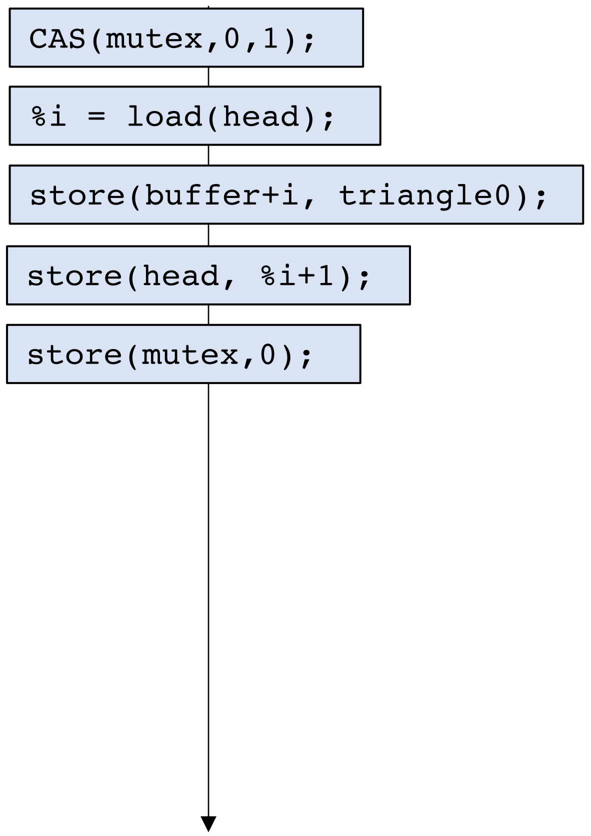


```
CAS(mutex,0,1);
```

if blue goes first
it gets to complete
its critical section
while thread 1 is spinning

Thread 0:
```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 1:
```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```

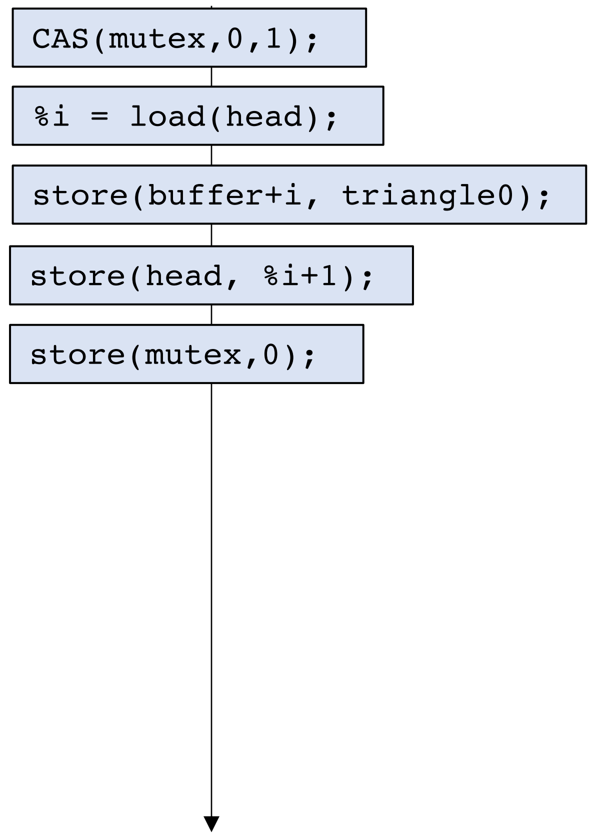

```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 0:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```



```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

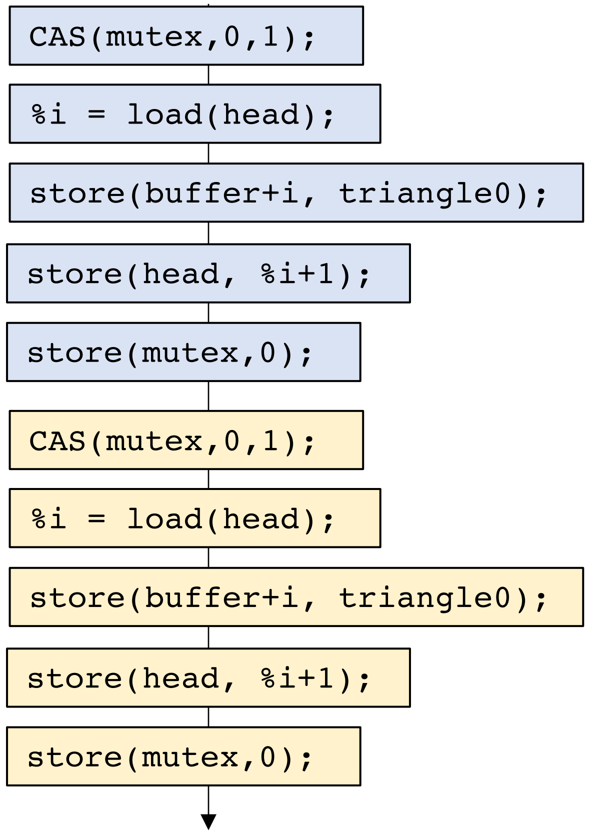
now yellow gets a change to go

Thread 0:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```



```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

now yellow gets a change to go

Thread 0:

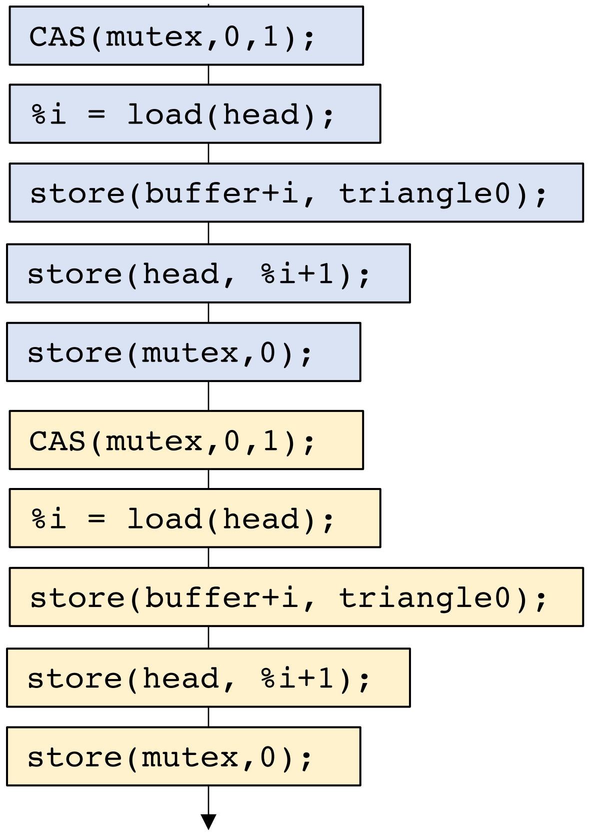
```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```

what can happen in a PSO memory model?

	L	S
L	NO	Different address
S	NO	Different address



```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 0:

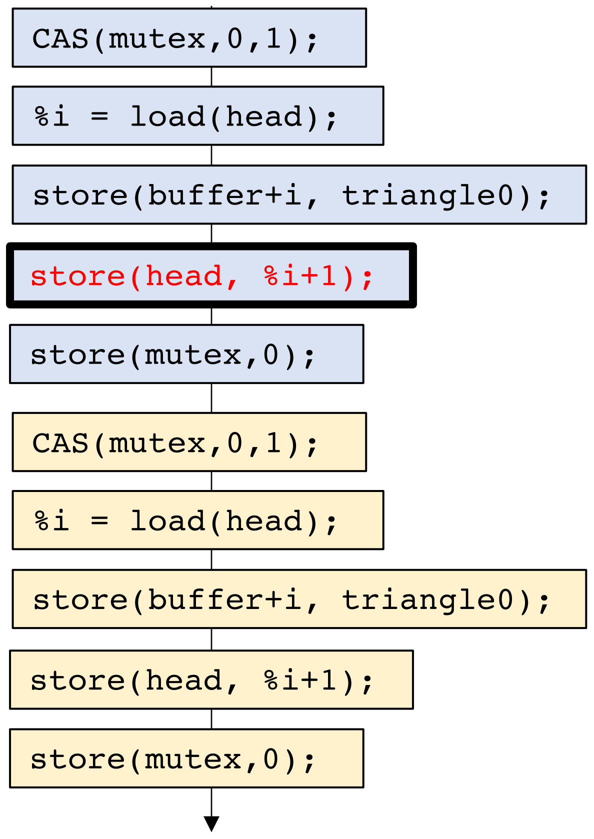
```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```

what can happen in a PSO memory model?

	L	S
L	NO	Different address
S	NO	Different address



```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 0:

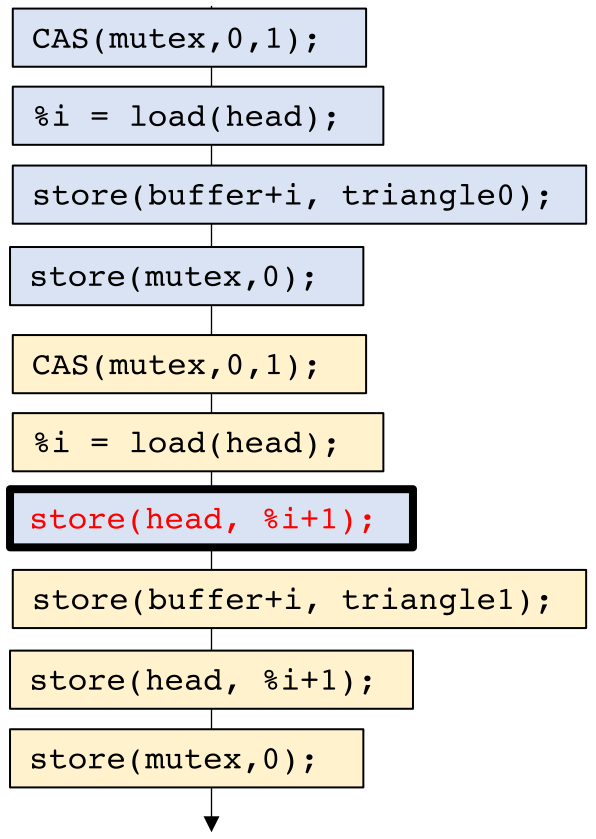
```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```

what can happen in a PSO memory model?

	L	S
L	NO	Different address
S	NO	Different address



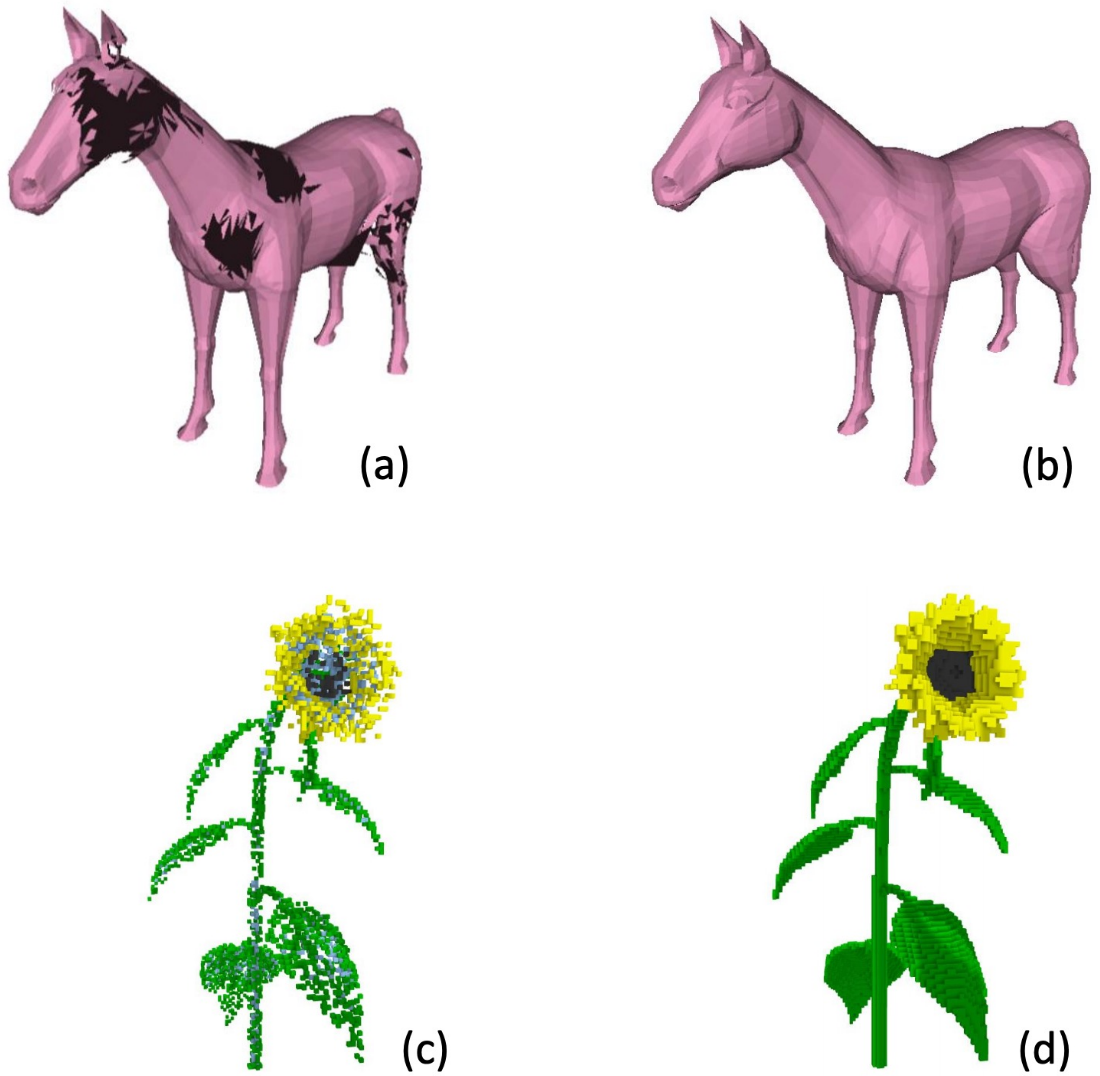
```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(mutex,0);
CAS(mutex,0,1);
%i = load(head);
store(head, %i+1);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```

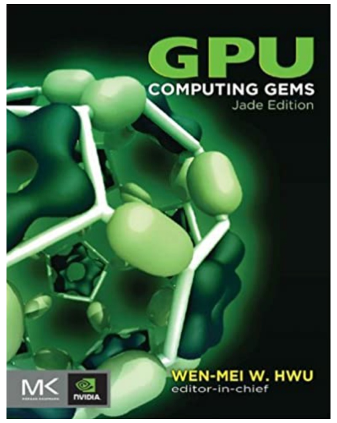
What just happened if this store moves?

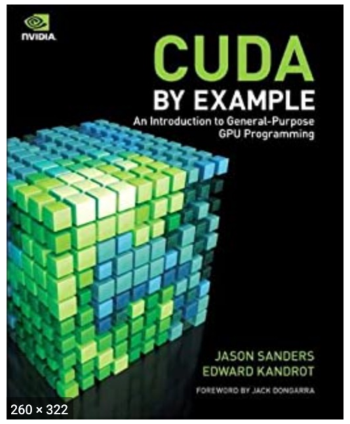
Nvidia in 2015

- Nvidia architects implemented a weak memory model
- Nvidia programmers expected a strong memory model
- Mutexes implemented without fences!

Nvidia in 2015







bug found in two Nvidia textbooks

We implemented a side-channel attack that made the bugs appear more frequently

These days Nvidia has a very well-specified memory model!

Thread 0:

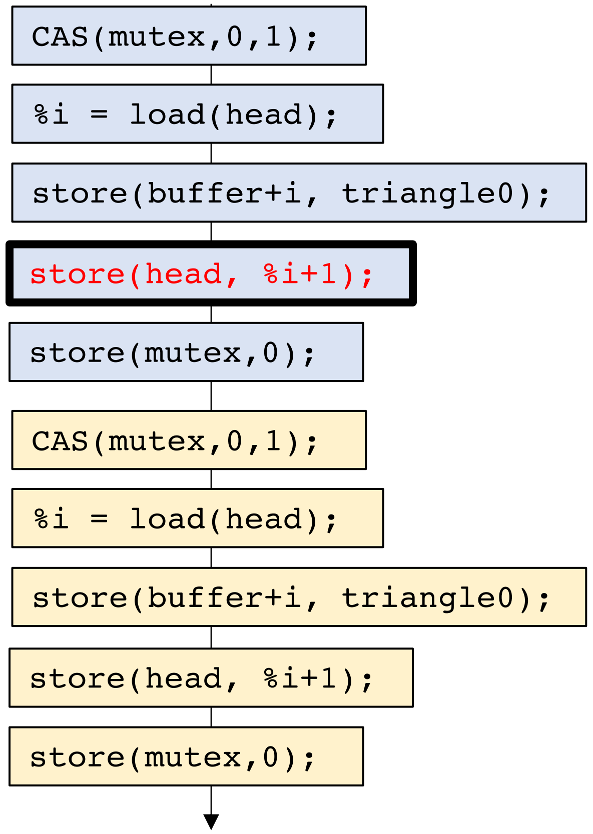
```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
store(mutex,0);
```

what can happen in a PSO memory model?

	L	S
L	NO	Different address
S	NO	🟩 Different address



```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

How to fix the issue?

Thread 0:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
fence;
store(mutex,0);
```

unlock contains fence before store!

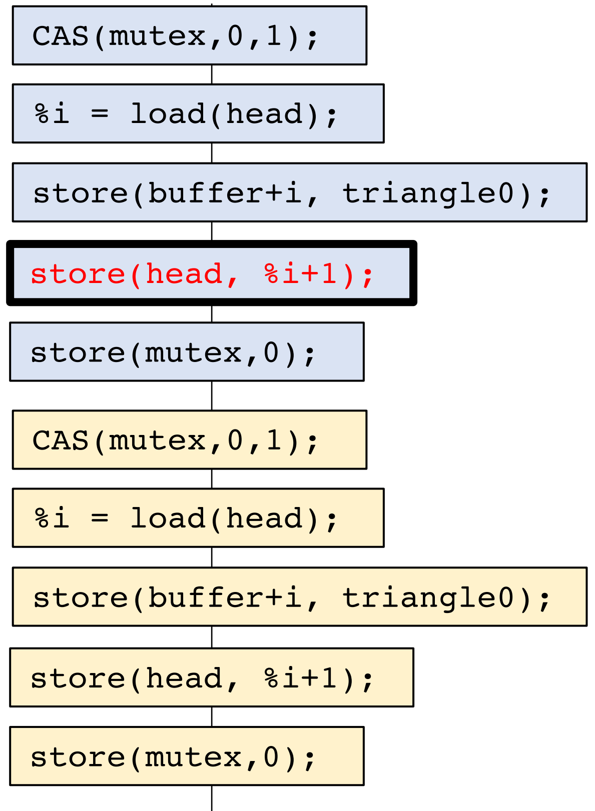
Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
fence;
store(mutex,0);
```

unlock contains fence before store!

what can happen in a PSO memory model?

	L	S
L	NO	Different address
S	NO	Different address



```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
store(mutex,0);
```

How to fix the issue?

your unlock function should contain a fence!

Thread 0:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
fence;
store(mutex,0);
```

unlock contains fence before store!

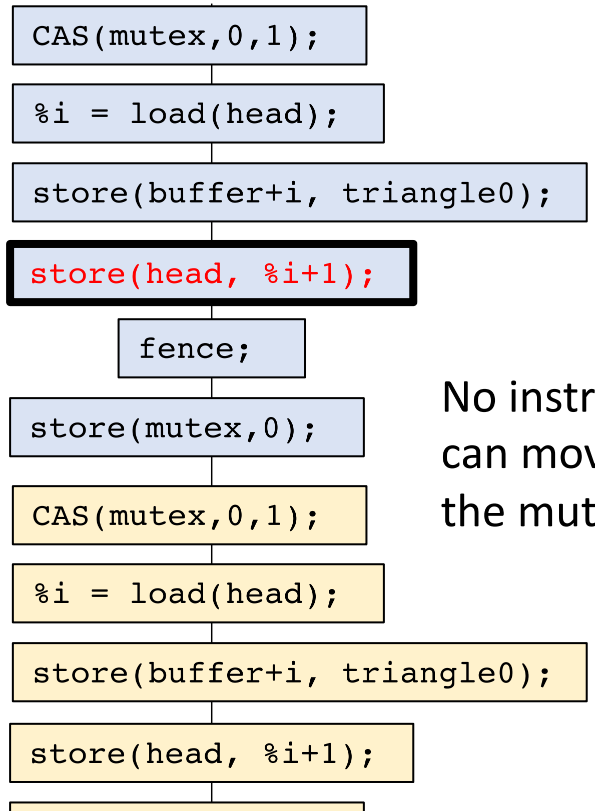
Thread 1:

```
SPIN:CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle1);
store(head, %i+1);
fence;
store(mutex,0);
```

unlock contains fence before store!

what can happen in a PSO memory model?

	L	S
L	NO	Different address
S	NO	🟩 Different address



```
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
fence;
store(mutex,0);
CAS(mutex,0,1);
%i = load(head);
store(buffer+i, triangle0);
store(head, %i+1);
```

No instructions can move after the mutex store!

How to fix the issue?

your unlock function should contain a fence!

Memory Model Strength

- If one memory model M0 allows more relaxed behaviors than another memory model M1, then M0 is more *relaxed* (or *weaker*) than M1.

- It is safe to run a program written for M0 on M1. But not vice versa

	L	S
L	NO	Different address
S	NO	NO

TSO

	L	S
L	NO	Different address
S	NO	Different address

PSO

	L	S
L	YES	Different address
S	Different address	Different address

RMO

Memory Model Strength

- Many times specifications are weaker than implementations:
  - A chip might document PSO, but implement TSO:
    - Why?

	L	S
L	NO	Different address
S	NO	NO

TSO

	L	S
L	NO	Different address
S	NO	Different address

PSO

	L	S
L	YES	Different address
S	Different address	Different address

RMO

Schedules and Liveness

Safety property

- ***Something bad will never happen***
    - i.e. the program will not exit with the mutex taken
    - Two threads will never be in the critical section at the same time
    - can be specified with assert statements in the program

However…

- *Safety is only half of the picture*

- Self driving car example:
    - Design a car that never crashes (safety property)

However…

- *Safety is only half of the picture*

- Self driving car example:
  - Design a car that never crashes (safety property)
  - **Easy**! Just design a car that can’t move!

  - We need include something else in the specification:

Liveness property

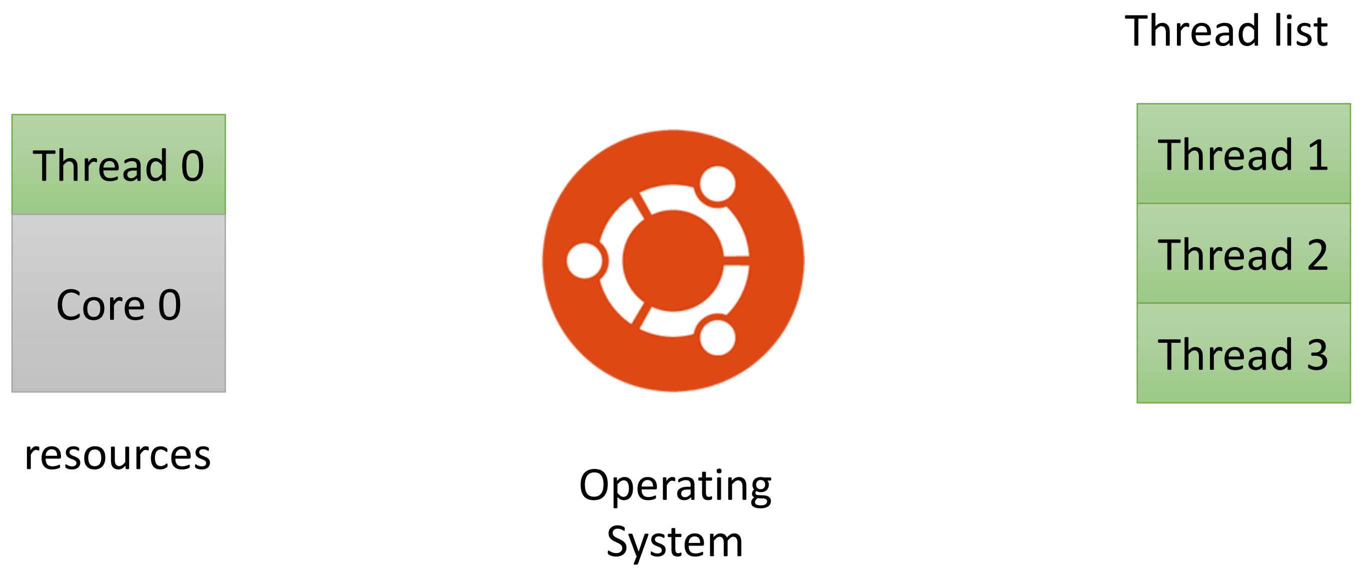
- Something good will eventually happen

- Examples:
  - The mutex program *will eventually terminate*
  - The self driving car *will eventually reach its destination*

- More difficult to reason about that safety properties

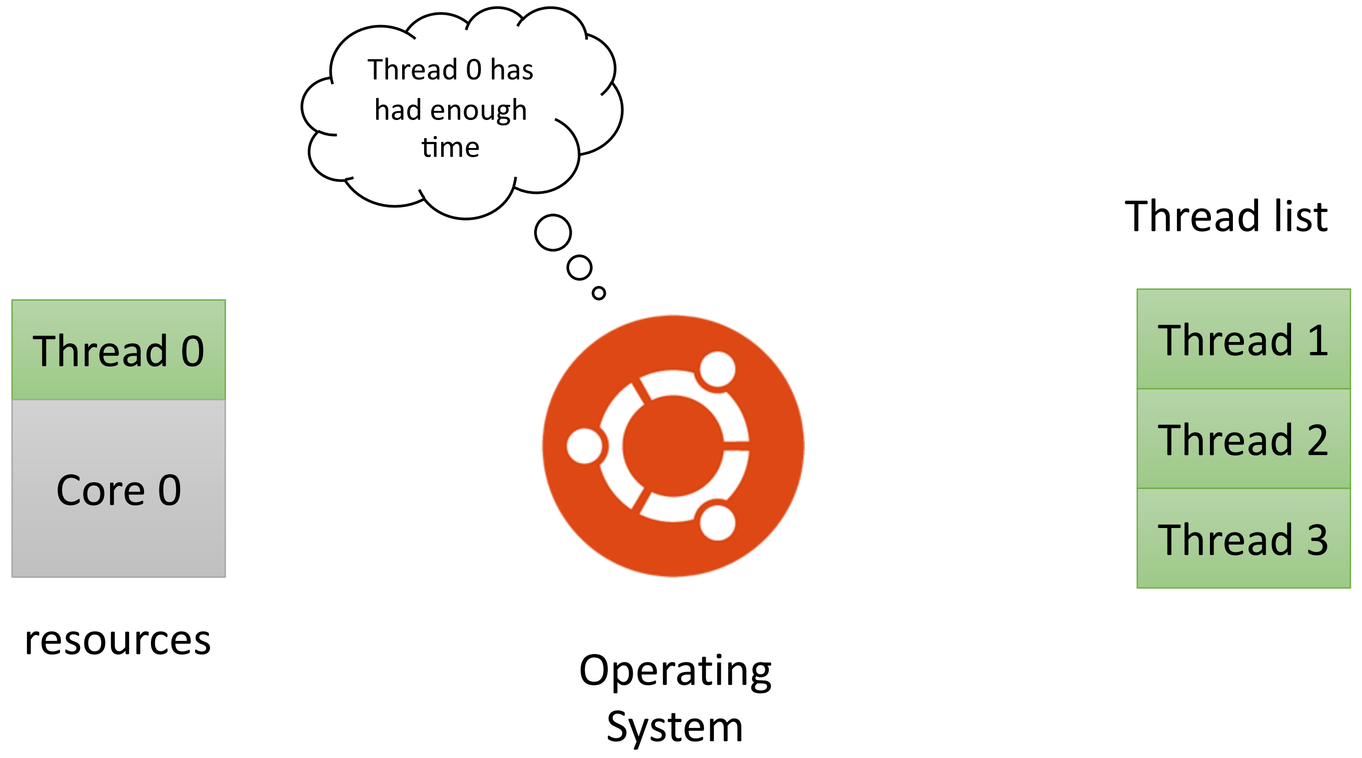
Schedulers

- A fair scheduler typically requires preemption



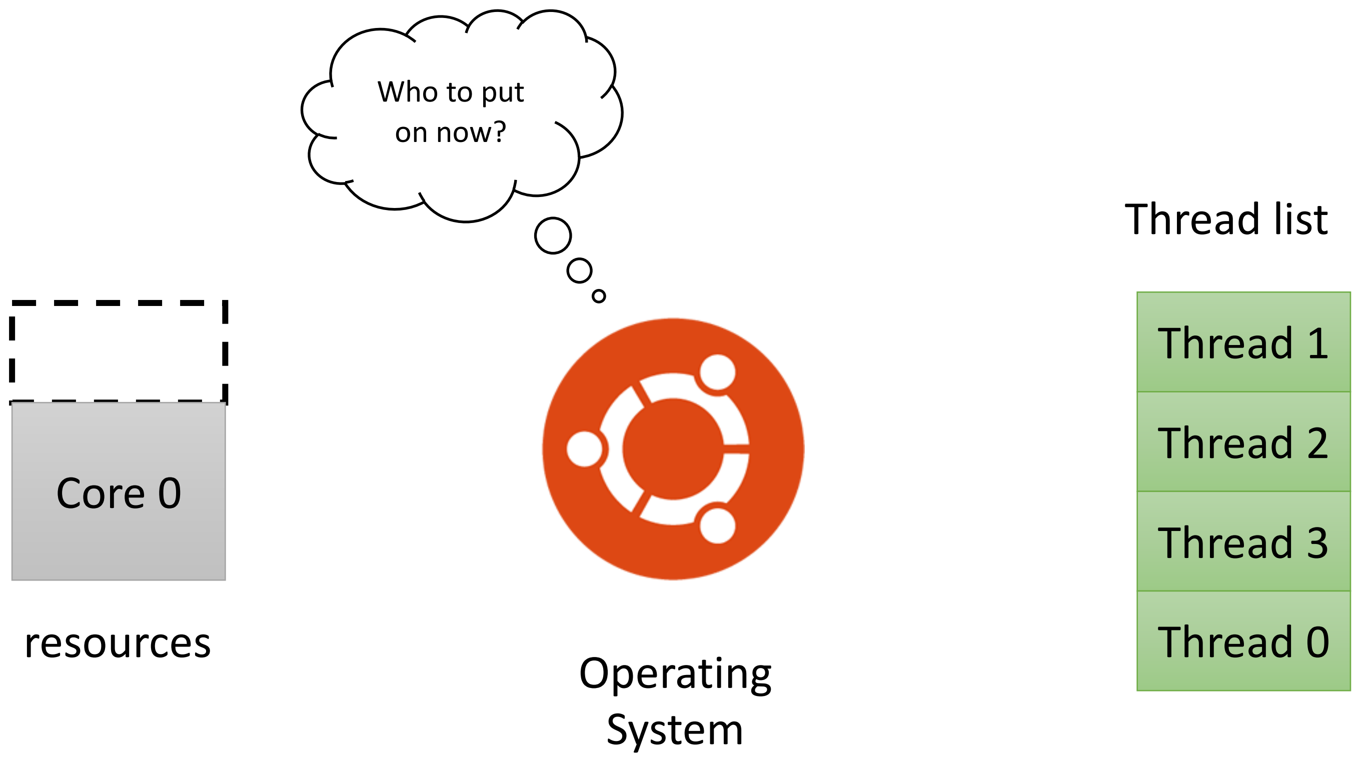
Schedulers

- A fair scheduler typically requires preemption



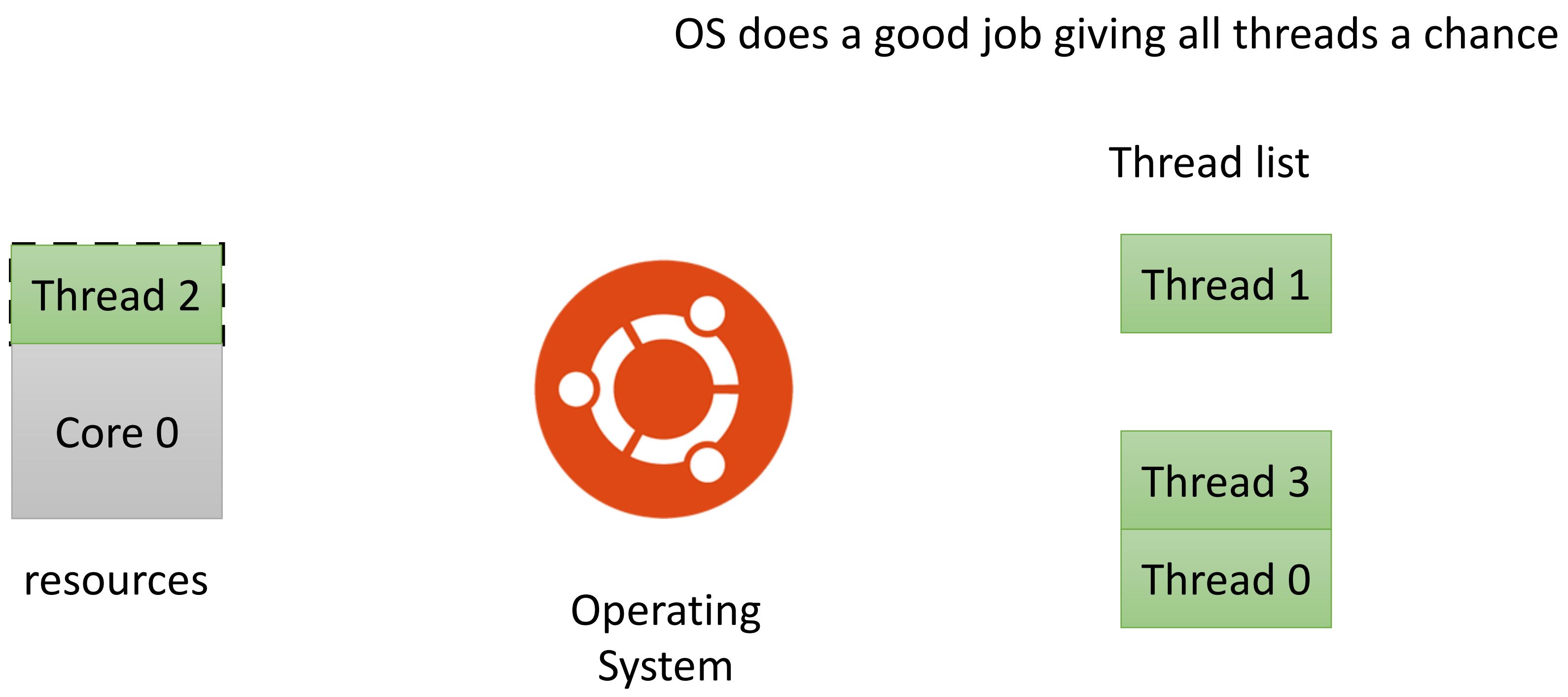
Schedulers

- A fair scheduler typically requires preemption



Schedulers

- A fair scheduler typically requires preemption

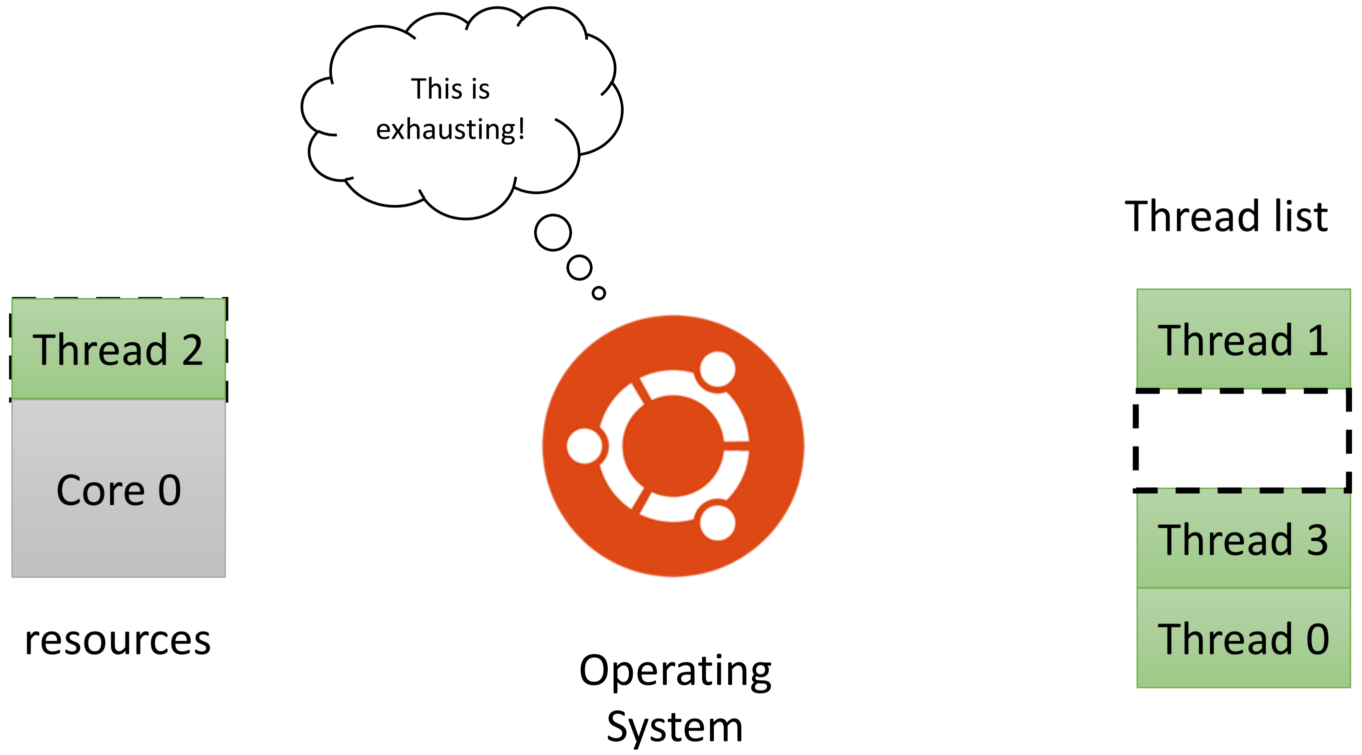


The fair scheduler

- every thread that has not terminated will “eventually” get a chance to execute.
  - “concurrent forward progress”: defined by C++
    not guaranteed, but encouraged (and likely what you will observe)
  - “weakly fair scheduler”: defined by classic concurrency textbooks

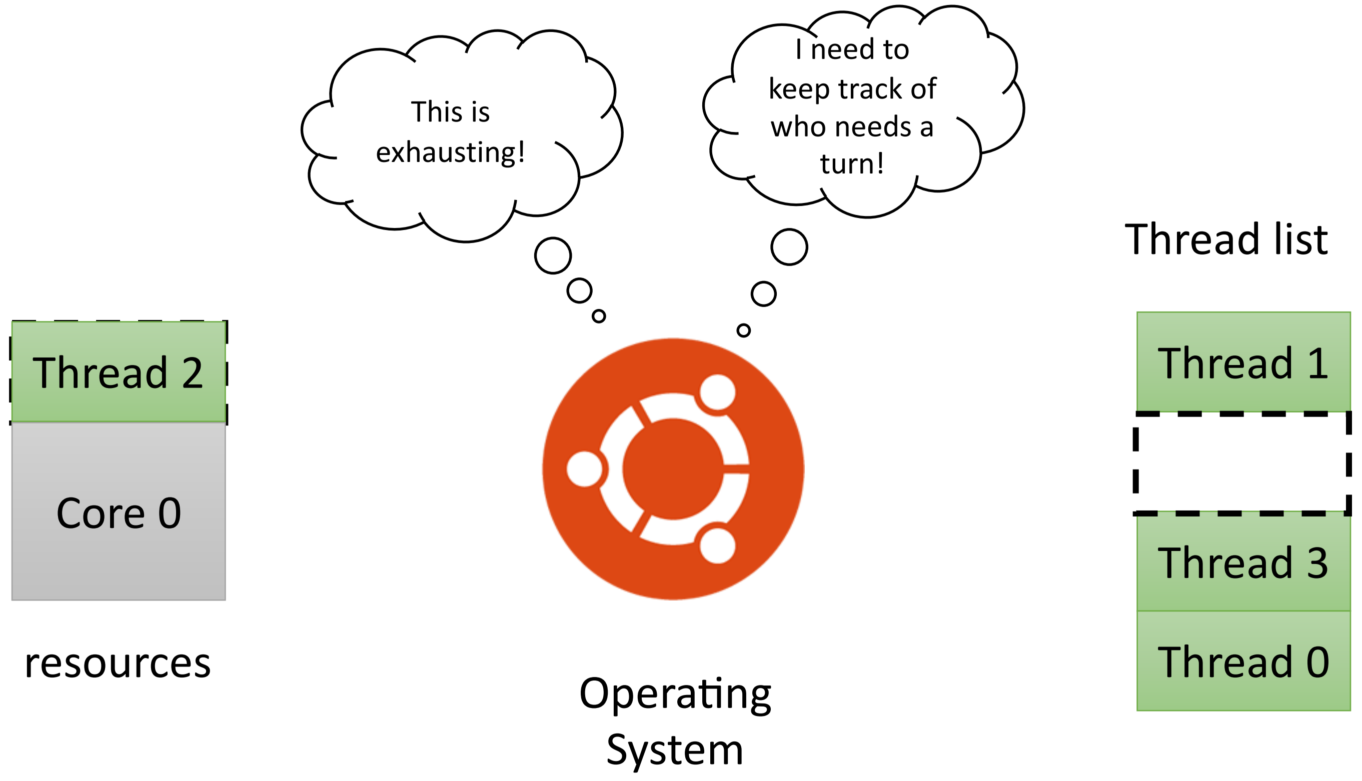
Schedulers

- A fair scheduler typically requires preemption



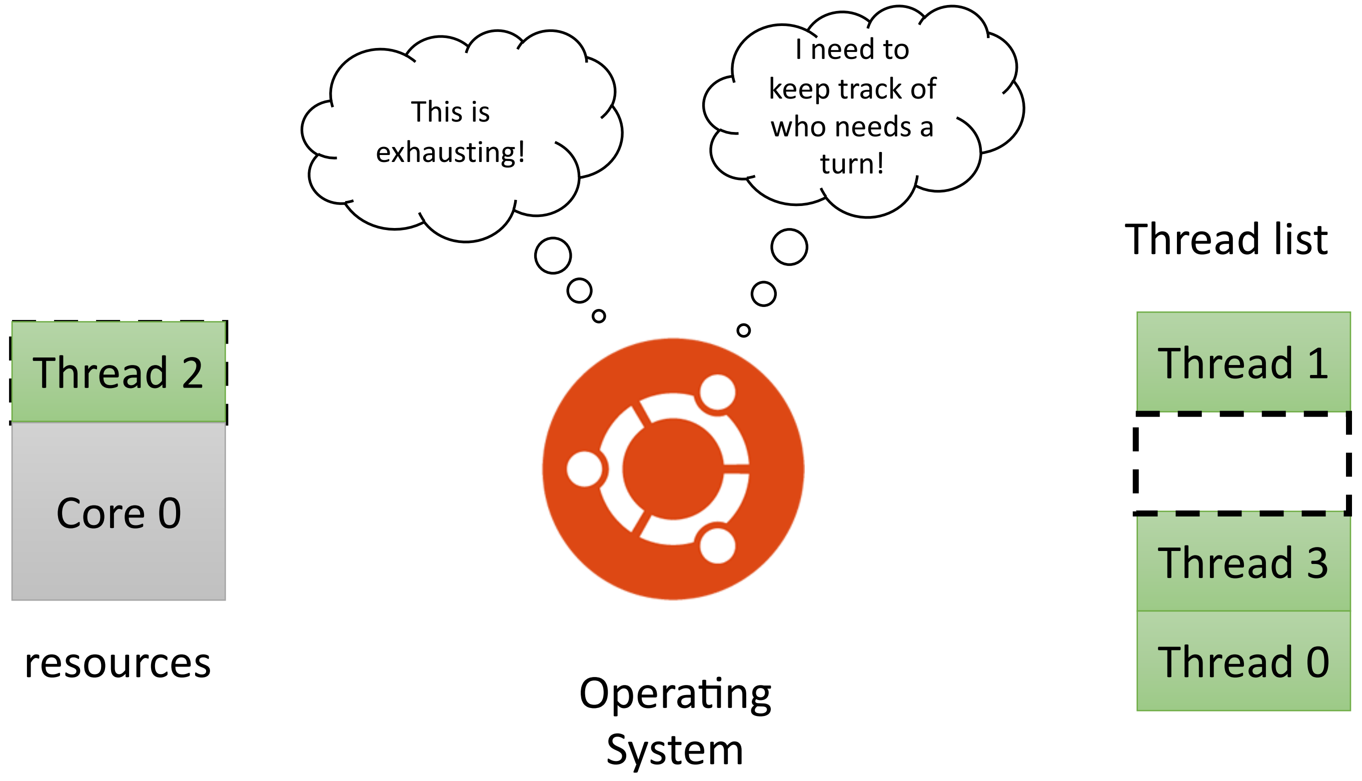
Schedulers

- A fair scheduler typically requires preemption



Schedulers

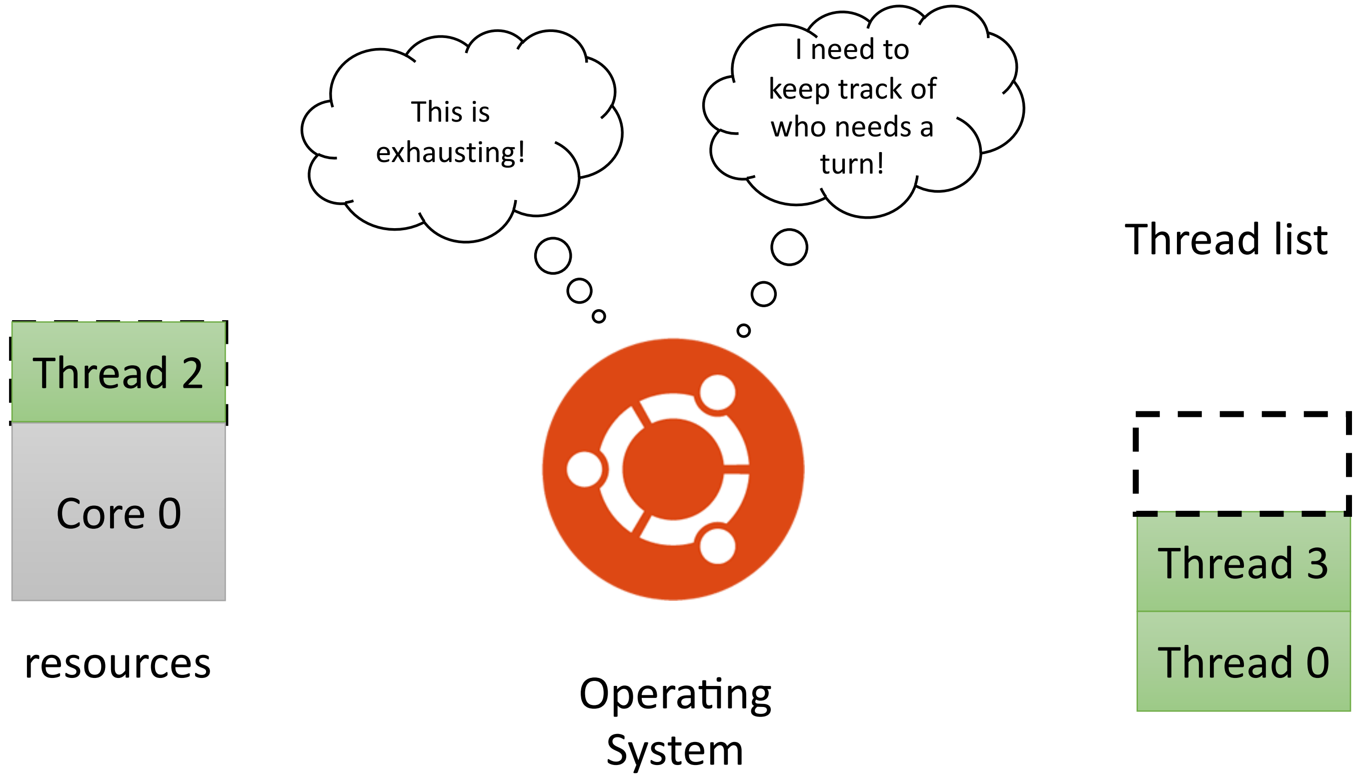
- A fair scheduler typically requires preemption



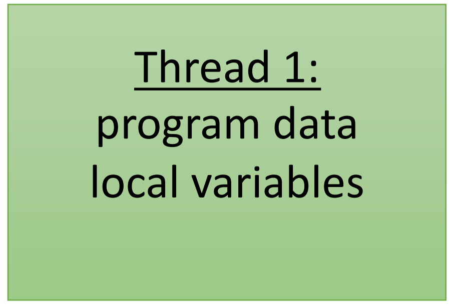
peak into a thread object:

Schedulers

- A fair scheduler typically requires preemption



peak into a thread object:

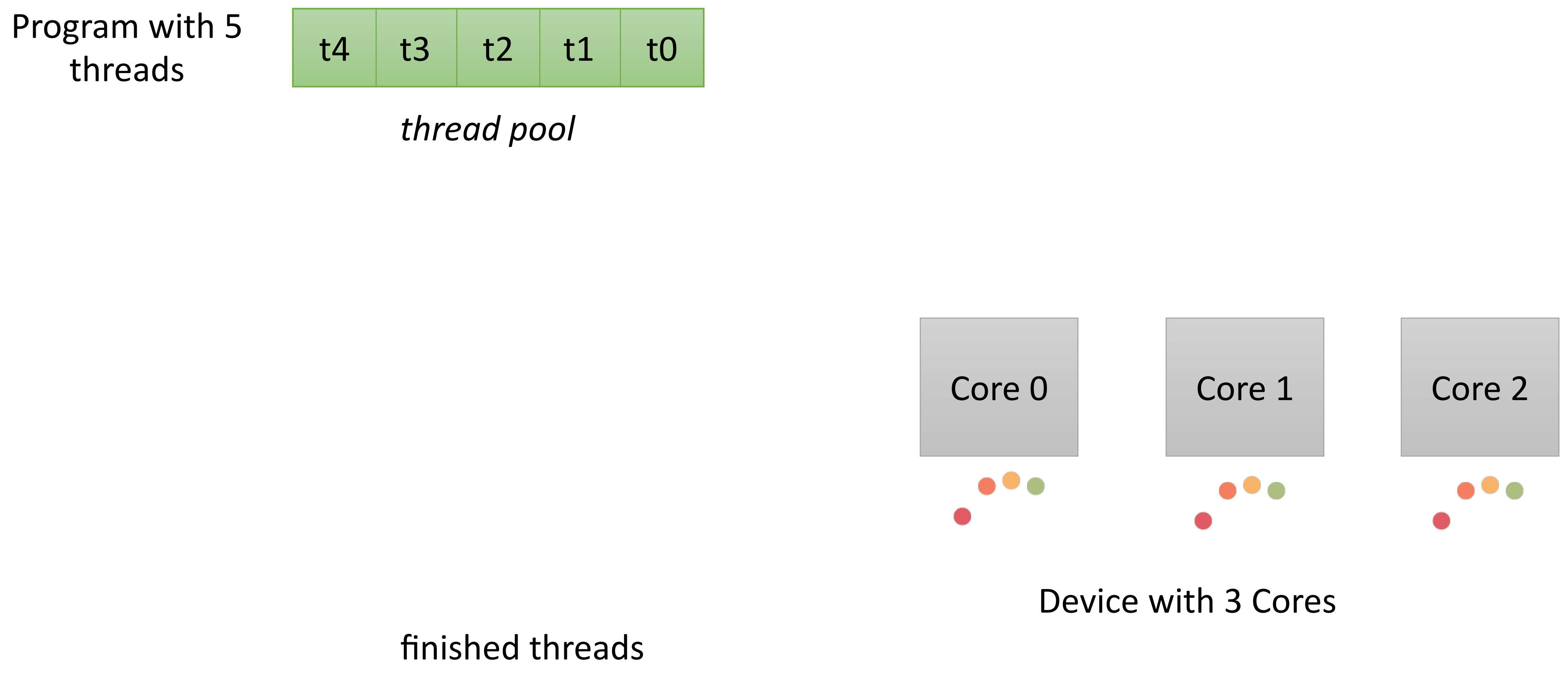


Estimated to be ~30K cycles to context switch between threads

Schedulers

- Systems might not support preemption: e.g. GPUs
- Frameworks might not implement preemption (e.g. OpenCL on CPUs)

simplified execution model



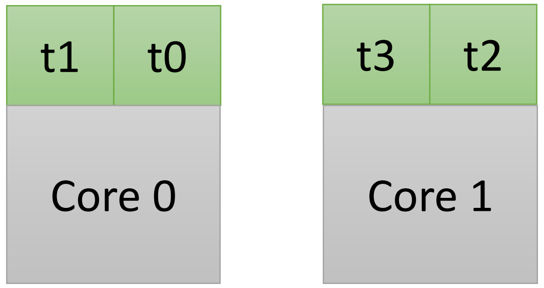
Solutions?

- I have N cores, only run N threads?

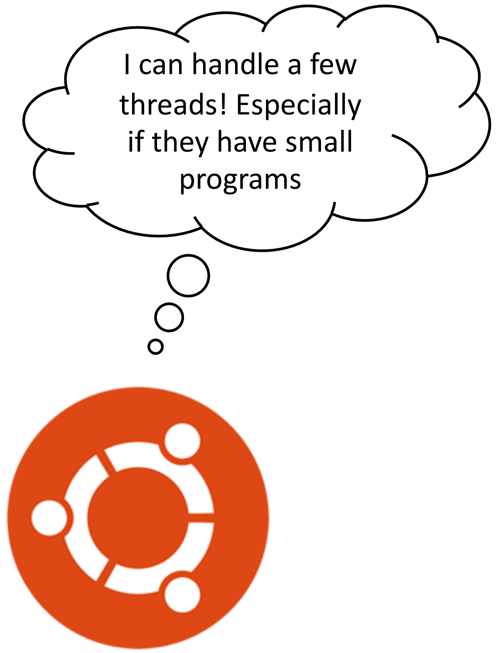
Solutions?

- I have N cores, only run N threads?

sometimes concurrency can help hide latency! Don't want to completely disallow it!



Device with 2 cores


I can handle a few threads! Especially if they have small programs

Solutions?

- I have N cores, only run N threads?
- GPU examples:
  - Depending on program size Nvidia GPUs support
    - 32 threads per core for small programs
    - 2 threads per core for big programs
- We need a better specification

Parallel Forward Progress

- “Any thread that has executed at least 1 instruction, is guaranteed to continue to be fairly executed”

- Also called:
  - “Parallel Forward Progress”: by C++
  - “Persistent Thread Model”: by GPU programmers
  - ”Occupancy Bound Execution Model”: in some of my papers

example

- Producer - consumer
  - Thread 0 waits for Thread 1 to write a flag

Thread 0:
```
0.0: while(flag.load() == 0);
```

Thread 1:
```
1.0: flag.store(1);
```

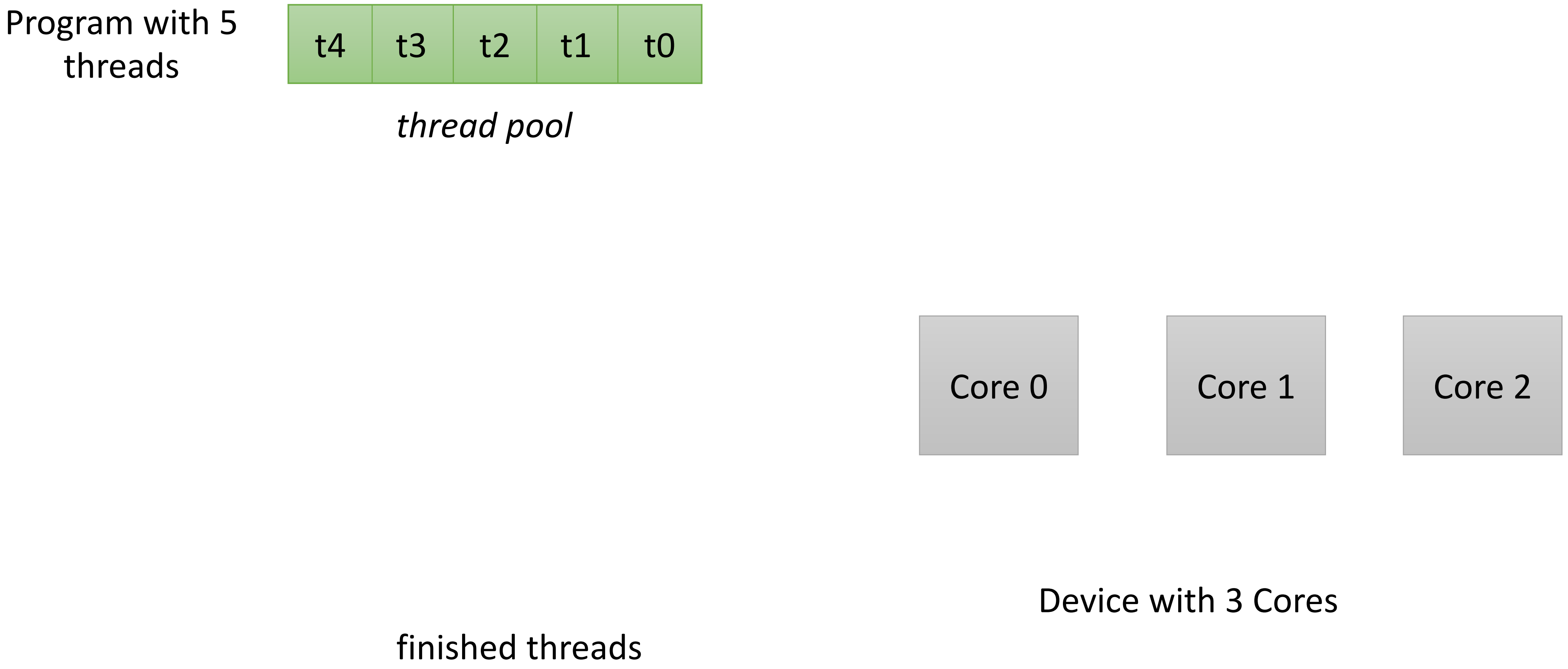
Is this program guaranteed to terminate under the fair scheduler?

Is this program guaranteed to terminate under the parallel scheduler?

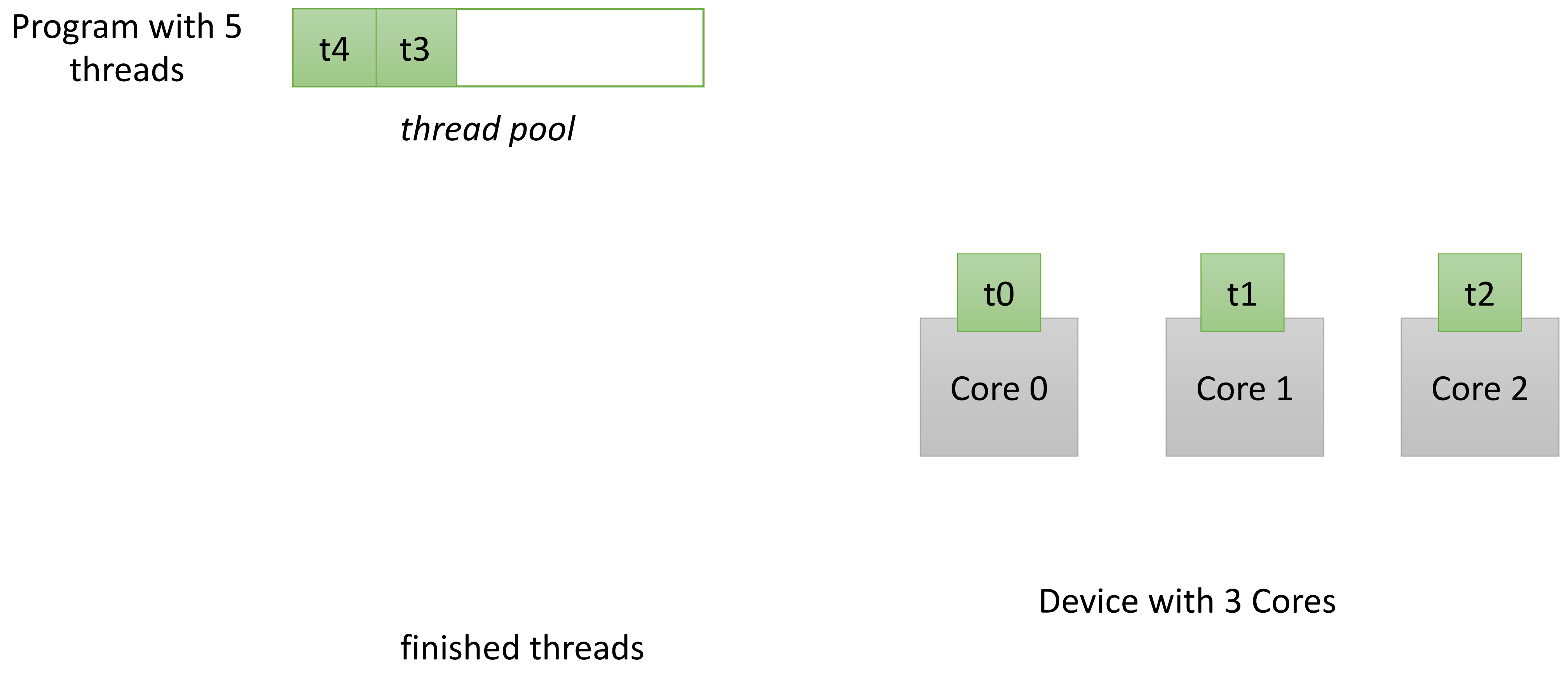
Schedulers

- In some cases the Parallel scheduler might be too strong
- For example dynamic power management on mobile devices

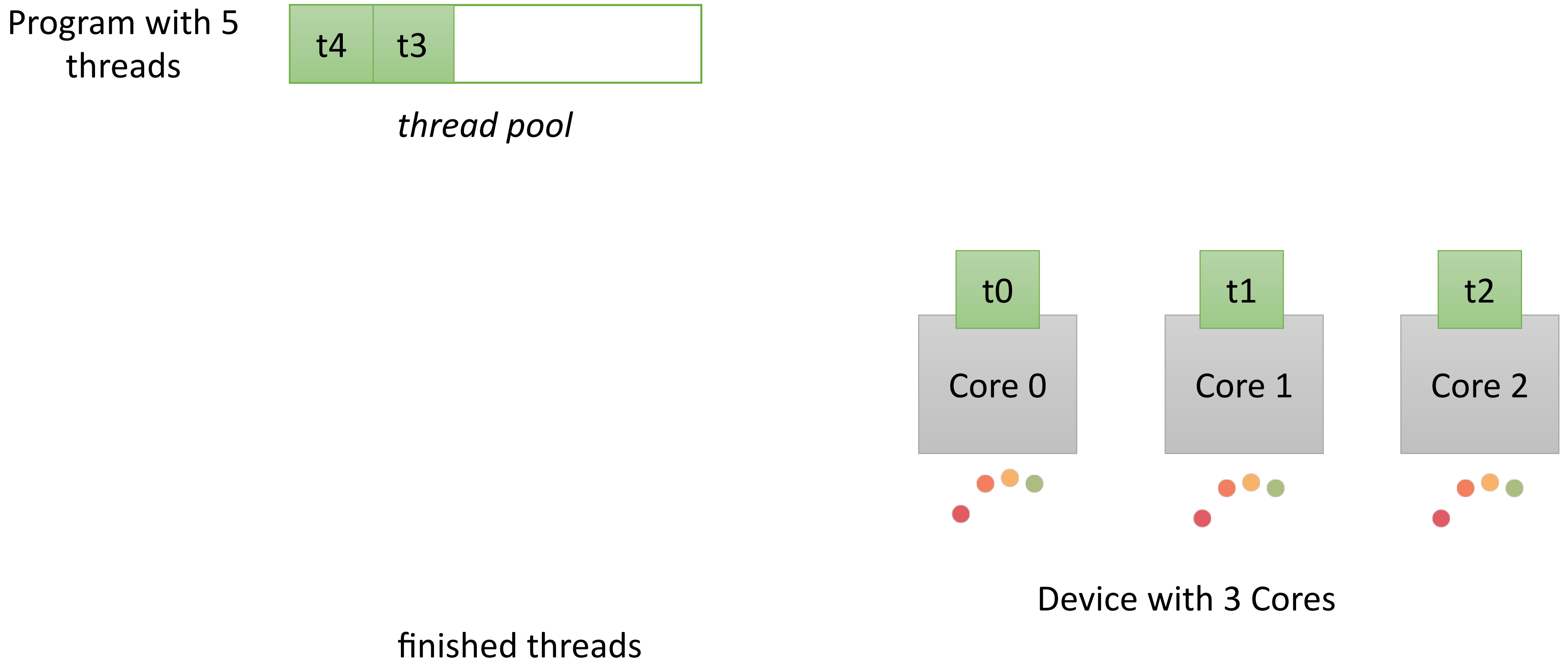
A power-saving scheduler



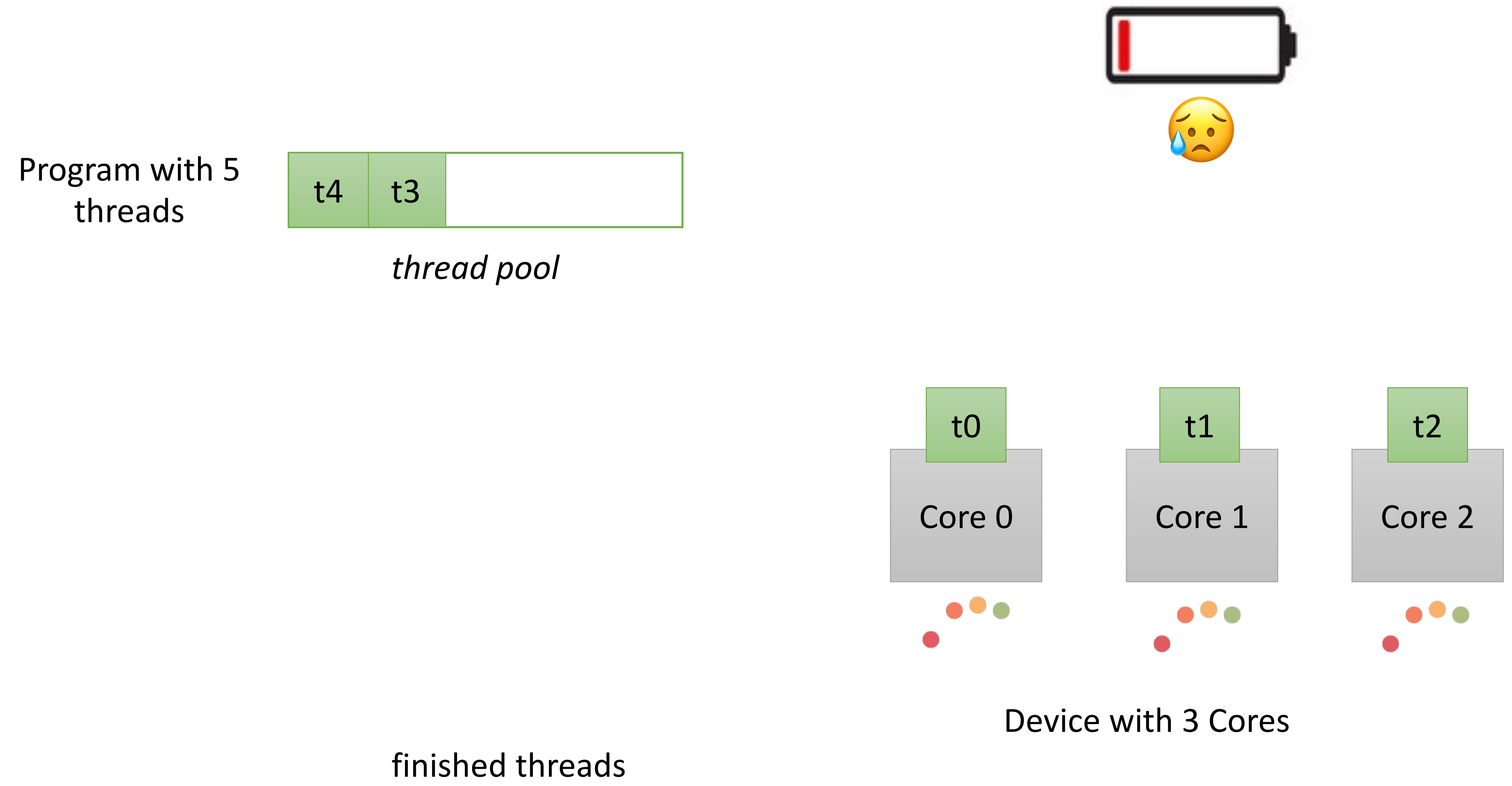
A power-saving scheduler



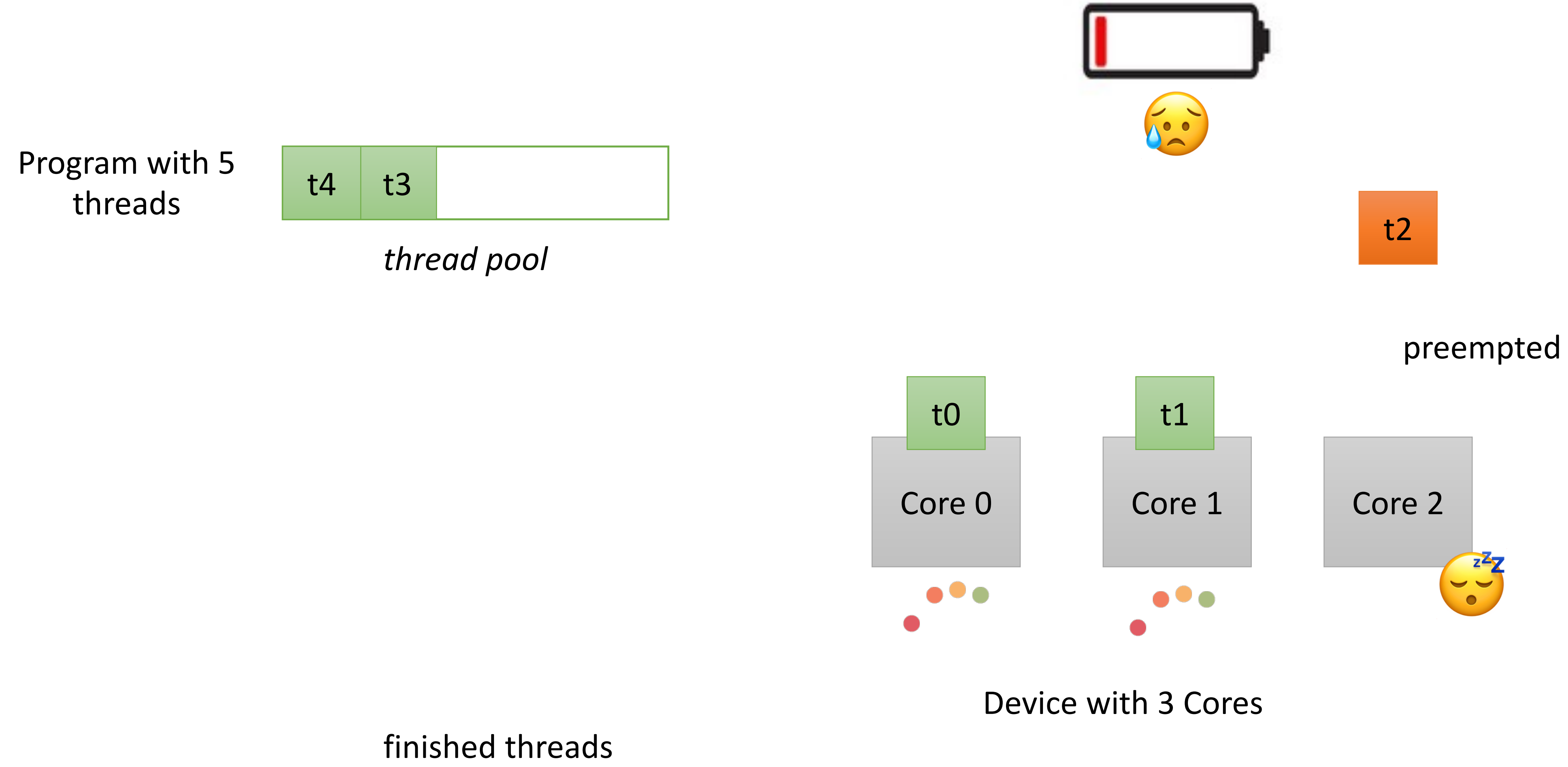
A power-saving scheduler



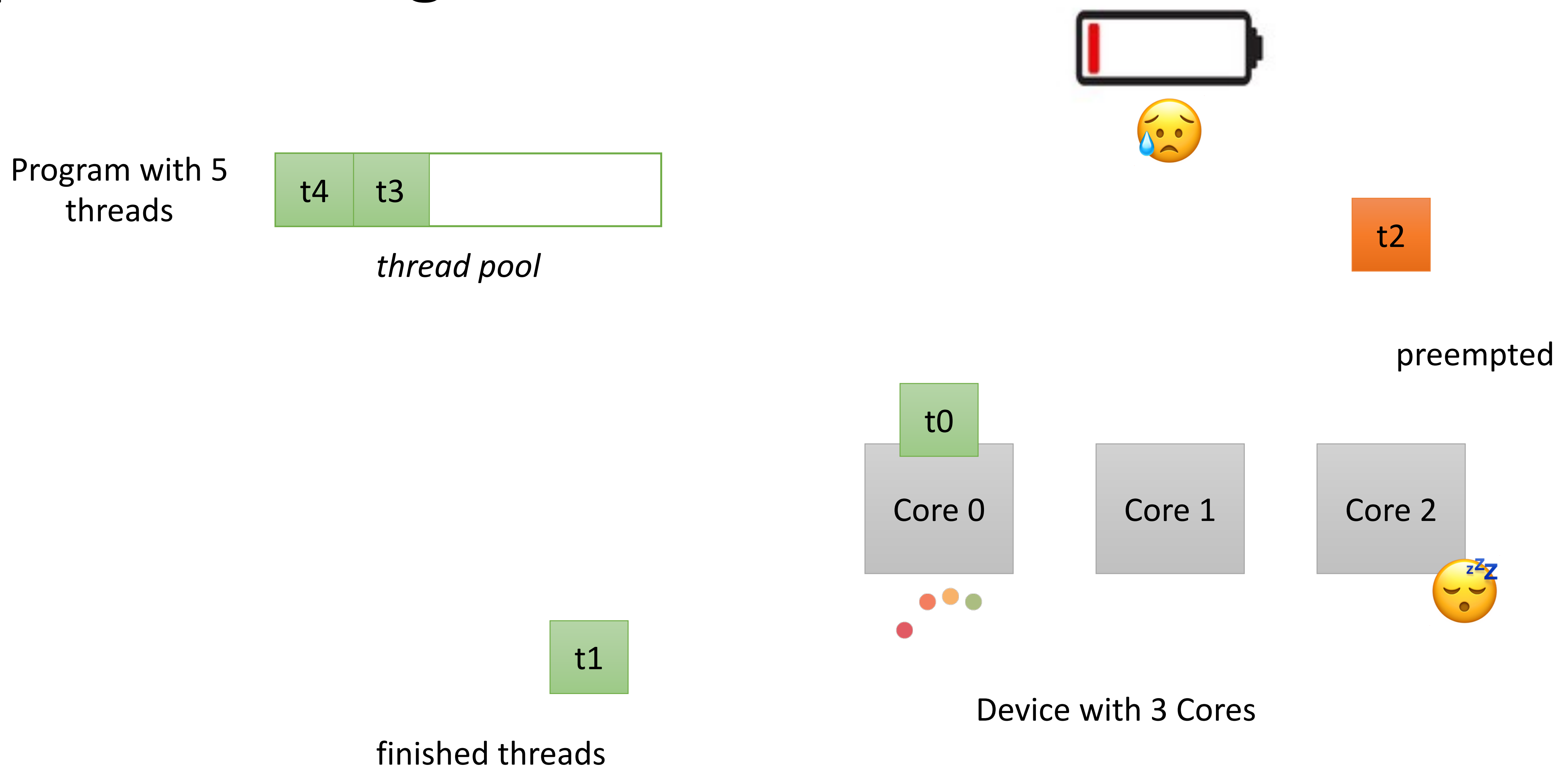
A power-saving scheduler



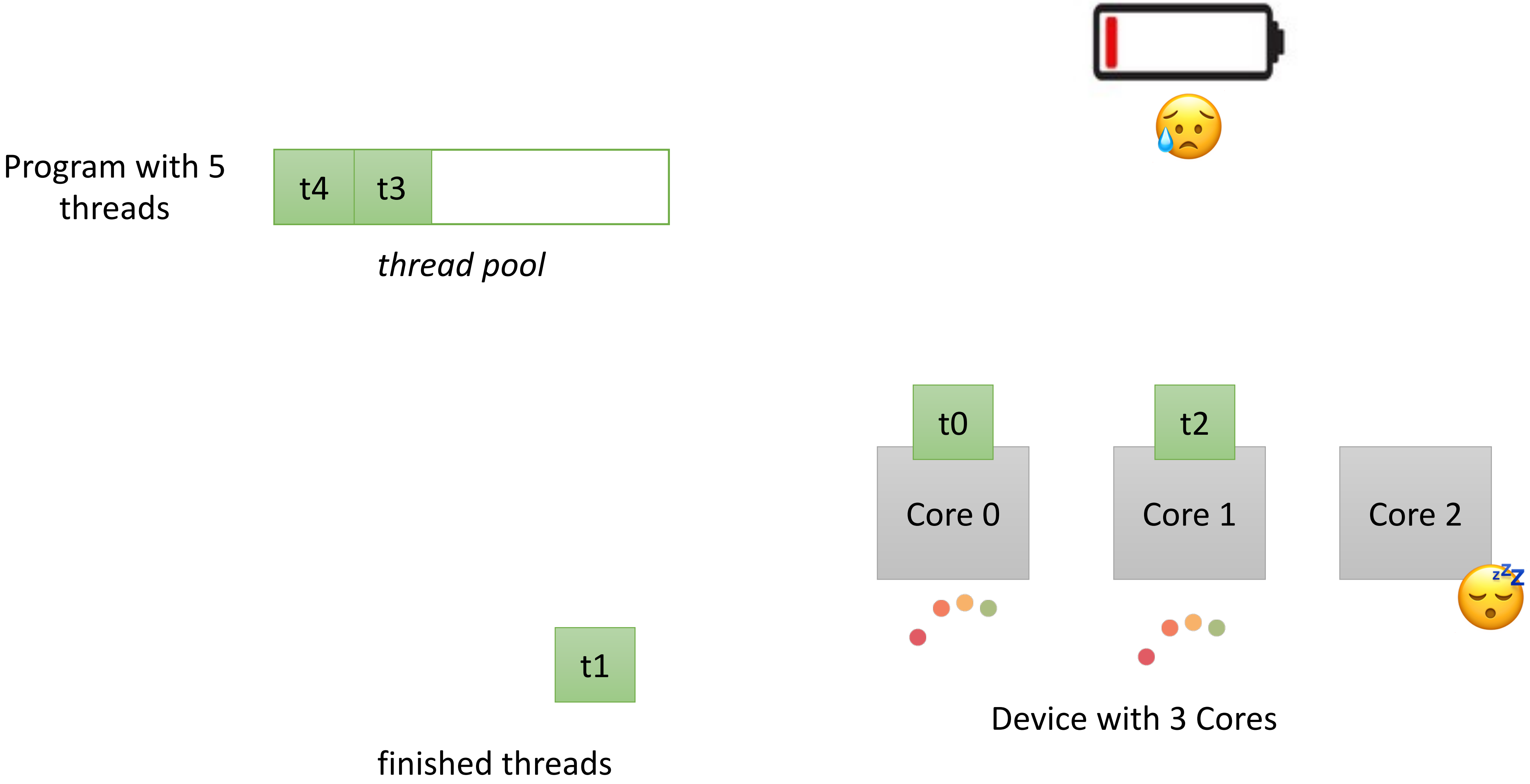
A power-saving scheduler



A power-saving scheduler



A power-saving scheduler



Schedulers

- This power-saving optimization messes up the Parallel Scheduler guarantees

- Can we do anything interesting with a scheduler like this?

Schedulers

- This power-saving optimization messes up the Parallel Scheduler guarantees

- Can we do anything interesting with a scheduler like this?

- The OS can give guarantees about the threads that it preempts for energy savings.

Schedulers

- This power-saving optimization messes up the Parallel Scheduler guarantees
- Can we do anything interesting with a scheduler like this?
- The OS can give guarantees about the threads that it preempts for energy savings.
- The OS could target threads with higher ids and give priority with threads with the lower id.

The HSA scheduler

- The thread with the lowest ID that hasn’t terminated is guaranteed to eventually be executed.

- Called:
  - “HSA” - Heterogeneous System Architecture, programming language proposed by AMD for new systems.

  - The HSA language appears to be defunct now, but the scheduler is a good fit for mobile devices (esp. mobile GPUs).

Thread 0:
```
0.0: while(flag.load() == 0);
```

Thread 1:
```
1.0: flag.store(1);
```

Is this program guaranteed to terminate under the power saving scheduler?

What if we switched the threads?

Thread 0:

```
0.0: m.lock();
0.1: m.unlock();
```

Thread 1:

```
1.0: m.lock();
1.1: m.unlock();
```

What about a mutex? Which scheduler is it guaranteed to work with?

Liveness

- So where are we now?

- CPU Schedulers:
  - In practice, tend to provide weakly fair schedulers (usually assumed)
    - Pthreads
    - OpenMP
    - etc.
  - Some cases do not: e.g. OpenCL on CPUs
  - C++ is starting to provide a hierarchical specification
    - concurrent
    - parallel
    - weakly parallel - at least one thread will execute

Liveness

- So where are we now?

- GPU schedulers:
  - Nvidia provides Parallel Forward Progress
    - Allows mutexes, concurrent data structures, etc.

  - OpenCL, Vulkan, and Metal provide no documentation on scheduler behaviors.
    - In practice, many assume parallel forward progress
    - This is not portable (esp. to ARM and Apple)
    - Working with specification groups to try and provide these

See you on Friday

- keep working on HW 4
- Starting GPUs and Javascript on Friday