synchronization 3: rwlocks / deadlock

Changelog

Changes not seen in first lecture:

- 20 Feb 2020: moving two files graphs: make directory names consistent with code
- 20 Feb 2020: is deadlock exercise: correct option lettering
- 20 Feb 2020: livelock example: don't use trylock to acquire first lock

last time

counting semaphores

- up: increment counter
- down: decrement counter, but wait first if count is zero
- intuition: track available quantity of resource

binary semaphores

semaphores and monitors accomplish the same things

- can implement one with the other

reader/writer locks

- implementing with monitors
- problem: priority

reader/writer-priority

policy question: writers first or readers first?

- writers-first: no readers go when writer waiting
- readers-first: no writers go when reader waiting

previous implementation: whatever randomly happens

- writers signalled first, maybe gets lock first?
- ...but non-determinstic in pthreads

can make explicit decision

reader/writer-priority

policy question: writers first or readers first?

- writers-first: no readers go when writer waiting
- readers-first: no writers go when reader waiting

previous implementation: whatever randomly happens

- writers signalled first, maybe gets lock first?
- ...but non-determinstic in pthreads

can make explicit decision

key method: track number of waiting readers/writers

writer-priority (1)

```
mutex_t lock; cond_t ok_to_read_cv; cond_t ok_to_write_cv;
int readers = 0, writers = 0;
int waiting_writers = 0;
ReadLock() {
  mutex_lock(&lock);
  while (writers != 0
         || waiting_writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  ++readers;
  mutex_unlock(&lock);
}

ReadUnlock() {
  mutex_lock(&lock);
  --readers;
  if (readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
  mutex_unlock(&lock);
}
```

```
WriteLock() {
  mutex_lock(&lock);
  ++waiting_writers;
  while (readers + writers != 0) {
    cond_wait(&ok_to_write_cv, &lock);
  }
  --waiting_writers;
  ++writers;
  mutex_unlock(&lock);
}

WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  if (waiting_writers != 0) {
    cond_signal(&ok_to_write_cv);
  } else {
    cond_broadcast(&ok_to_read_cv);
  }
  mutex_unlock(&lock);
}
```

writer-priority (1)

```
mutex_t lock; cond_t ok_to_read_cv; cond_t ok_to_write_cv;
int readers = 0, writers = 0;
int waiting_writers = 0;
ReadLock() {
  mutex_lock(&lock);
  while (writers != 0
         || waiting_writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  ++readers;
  mutex_unlock(&lock);
}

ReadUnlock() {
  mutex_lock(&lock);
  --readers;
  if (readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
  mutex_unlock(&lock);
}
```

```
WriteLock() {
  mutex_lock(&lock);
  ++waiting_writers;
  while (readers + writers != 0) {
    cond_wait(&ok_to_write_cv, &lock);
  }
  --waiting_writers;
  ++writers;
  mutex_unlock(&lock);
}

WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  if (waiting_writers != 0) {
    cond_signal(&ok_to_write_cv);
  } else {
    cond_broadcast(&ok_to_read_cv);
  }
  mutex_unlock(&lock);
}
```

writer-priority (1)

```
mutex_t lock; cond_t ok_to_read_cv; cond_t ok_to_write_cv;
int readers = 0, writers = 0;
int waiting_writers = 0;
ReadLock() {
  mutex_lock(&lock);
  while (writers != 0
         || waiting_writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  ++readers;
  mutex_unlock(&lock);
}

ReadUnlock() {
  mutex_lock(&lock);
  --readers;
  if (readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
  mutex_unlock(&lock);
}
```

```
WriteLock() {
  mutex_lock(&lock);
  ++waiting_writers;
  while (readers + writers != 0) {
    cond_wait(&ok_to_write_cv, &lock);
  }
  --waiting_writers;
  ++writers;
  mutex_unlock(&lock);
}

WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  if (waiting_writers != 0) {
    cond_signal(&ok_to_write_cv);
  } else {
    cond_broadcast(&ok_to_read_cv);
  }
  mutex_unlock(&lock);
}
```

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0

```
mutex_lock(&lock);
while (writers != 0 || waiting_writers != 0) {
  cond_wait(&ok_to_read_cv, &lock);
}
++readers;
mutex_unlock(&lock);
```

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
`ReadLock`				0	1	0
`(reading)`	`ReadLock`			0	2	0

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0
(reading)	ReadLock			0	2	0
(reading)	(reading)	WriteLock wait		0	2	1

```
mutex_lock(&lock);
++waiting_writers;
while (readers + writers != 0) {
  cond_wait(&ok_to_write_cv, &lock);
}
```

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0
(reading)	ReadLock			0	2	0
(reading)	(reading)	WriteLock wait		0	2	1
(reading)	(reading)	WriteLock wait	ReadLock wait	0	2	1

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0
(reading)	ReadLock			0	2	0
(reading)	(reading)	WriteLock wait		0	2	1
(reading)	(read	wait	ReadLock wait	0	2	1
ReadUnlock	(r	wait	ReadLock wait	0	1	1

```
mutex_lock(&lock);
--readers;
if (readers == 0)
  ...
```

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	**reader 2**	**writer 1**	**reader 3**	**W**	**R**	**WW**
				0	0	0
ReadLock				0	1	0
(reading)	ReadLock			0	2	0
(reading)	(reading)	WriteLock wait		0	2	1
(reading)	(reading)	WriteLock wait	ReadLock wait	0	2	1
ReadUnlock	(reading)	Write			1	1
	ReadUnlock				0	1

```
mutex_lock(&lock);
--readers;
if (readers == 0)
  cond_signal(&ok_to_write_cv)
mutex_unlock(&lock);
```

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0
(reading)	Read			0	2	0
(reading)	(rea			0	2	1
(reading)	(rea		it	0	2	1
ReadUnlock	(reading)	WriteLock wait	ReadLock wait	0	1	1
	ReadUnlock	WriteLock wait	ReadLock wait	0	0	1
		WriteLock	ReadLock wait	1	0	0

```
while (readers + writers != 0) {
  cond_wait(&ok_to_write_cv, &lock);
}
--waiting_writers; ++writers;
mutex_unlock(&lock);
```

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0
(reading)	ReadLock			0	2	0
(reading)	(reading)	WriteLock wait		0	2	1
(reading)	(reading)	WriteLock wait	ReadLock wait	0	2	1
ReadUnlock	(reading)	WriteLock wait	ReadLock wait	0	1	1
	ReadUnlock	WriteLock wait	ReadLock wait	0	0	1
		WriteLock	ReadLock wait	1	0	0
		(read+writing)	ReadLock wait	1	0	0

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0
(reading)	ReadLo			0	2	0
(reading)	(readi			0	2	1
(reading)	(readi		wait	0	2	1
ReadUnlock	(readi		wait	0	1	1
	ReadUr		wait	0	0	1
		WriteLock	ReadLock wait	1	0	0
		(read+writing)	ReadLock wait	1	0	0
		WriteUnlock	ReadLock wait	0	0	0

```
mutex_lock(&lock);
if (waiting_writers != 0) {
  cond_signal(&ok_to_write_cv);
} else {
  cond_broadcast(&ok_to_read_cv);
}
```

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0
(reading)	ReadLock			0	2	0
(reading)	(reading)					
(reading)	(reading)					
ReadUnlock	(reading)					
	ReadUnlock					
		WriteLock	ReadLock wait	1	0	0
		(read+writing)	ReadLock wait	1	0	0
		WriteUnlock	ReadLock wait	0	0	0
			ReadLock	0	1	0

```
while (writers != 0 && waiting_writers != 0) {
  cond_wait(&ok_to_read_cv, &lock);
}
++readers;
mutex_unlock(&lock);
```

simulation of reader/write lock

writer-priority version

W = writers, R = readers, WW = waiting_writers

reader 1	reader 2	writer 1	reader 3	W	R	WW
				0	0	0
ReadLock				0	1	0
(reading)	ReadLock			0	2	0
(reading)	(reading)	WriteLock wait		0	2	1
(reading)	(reading)	WriteLock wait	ReadLock wait	0	2	1
ReadUnlock	(reading)	WriteLock wait	ReadLock wait	0	1	1
	ReadUnlock	WriteLock wait	ReadLock wait	0	0	1
		WriteLock	ReadLock wait	1	0	0
		(read+writing)	ReadLock wait	1	0	0
		WriteUnlock	ReadLock wait	0	0	0
			ReadLock	0	1	0

reader-priority (1)

```
...
int waiting_readers = 0;
ReadLock() {
  mutex_lock(&lock);
  ++waiting_readers;
  while (writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  --waiting_readers;
  ++readers;
  mutex_unlock(&lock);
}

ReadUnlock() {
  ...
  if (waiting_readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
}
```

```
WriteLock() {
  mutex_lock(&lock);
  while (waiting_readers +
         readers + writers != 0) {
    cond_wait(&ok_to_write_cv);
  }
  ++writers;
  mutex_unlock(&lock);
}
WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  if (readers == 0 && waiting_readers == 0) {
    cond_signal(&ok_to_write_cv);
  } else {
    cond_broadcast(&ok_to_read_cv);
  }
  mutex_unlock(&lock);
}
```

reader-priority (1)

```
...
int waiting_readers = 0;
ReadLock() {
  mutex_lock(&lock);
  ++waiting_readers;
  while (writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  --waiting_readers;
  ++readers;
  mutex_unlock(&lock);
}

ReadUnlock() {
  ...
  if (waiting_readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
}
```

```
WriteLock() {
  mutex_lock(&lock);
  while (waiting_readers +
         readers + writers != 0) {
    cond_wait(&ok_to_write_cv);
  }
  ++writers;
  mutex_unlock(&lock);
}
WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  if (readers == 0 && waiting_readers == 0) {
    cond_signal(&ok_to_write_cv);
  } else {
    cond_broadcast(&ok_to_read_cv);
  }
  mutex_unlock(&lock);
}
```

rwlock exercise

suppose we want something in-between reader and writer priority:

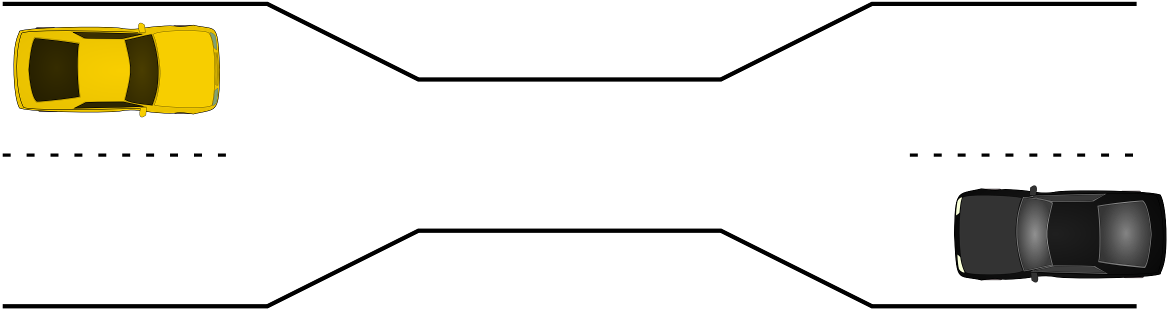
reader-priority except if writers wait more than 1 second

exercise: what do we change?

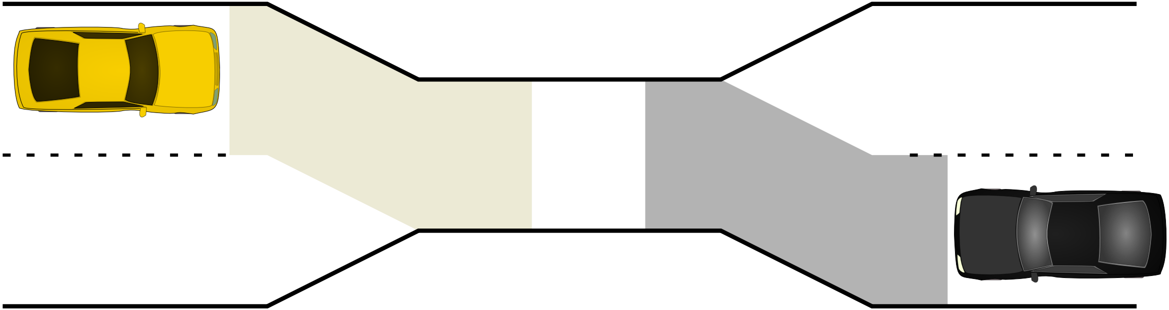
```
...
int waiting_readers = 0;
ReadLock() {
  mutex_lock(&lock);
  ++waiting_readers;
  while (writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  --waiting_readers;
  ++readers;
  mutex_unlock(&lock);
}

ReadUnlock() {
  mutex_lock(&lock);
  --readers;
  if (waiting_readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
  mutex_unlock(&lock);
```

```
WriteLock() {
  mutex_lock(&lock);
  while (waiting_readers + readers + writers != 0) {
    cond_wait(&ok_to_write_cv);
  }
  ++writers;
  mutex_unlock(&lock);
}
WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  if (waiting_readers == 0) {
    cond_signal(&ok_to_write_cv);
  } else {
    cond_broadcast(&ok_to_read_cv);
  }
  mutex_unlock(&lock);
}
```

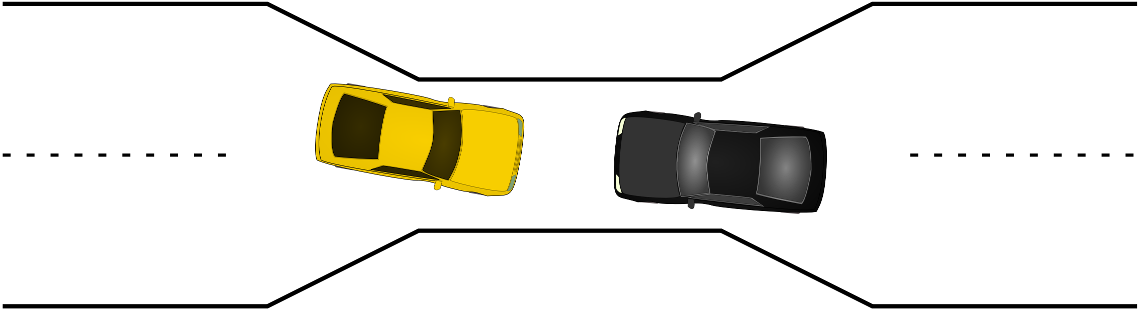
the one-way bridge



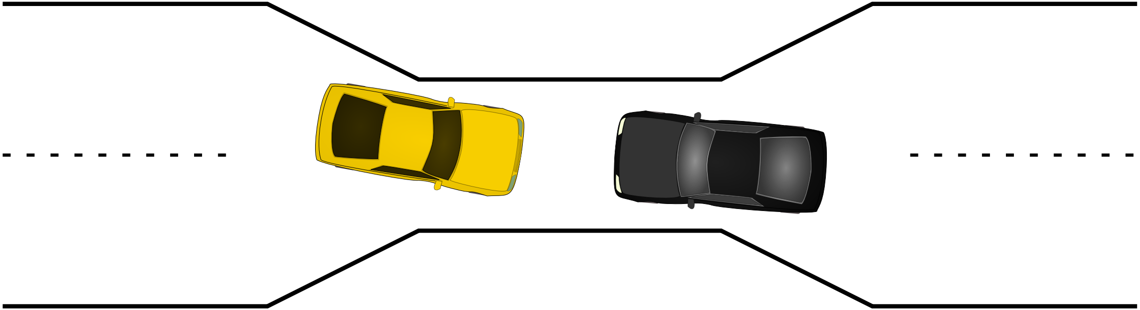
the one-way bridge



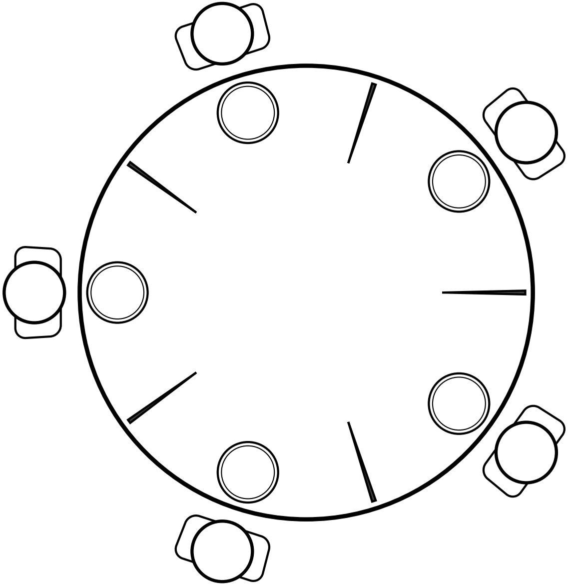
the one-way bridge



the one-way bridge

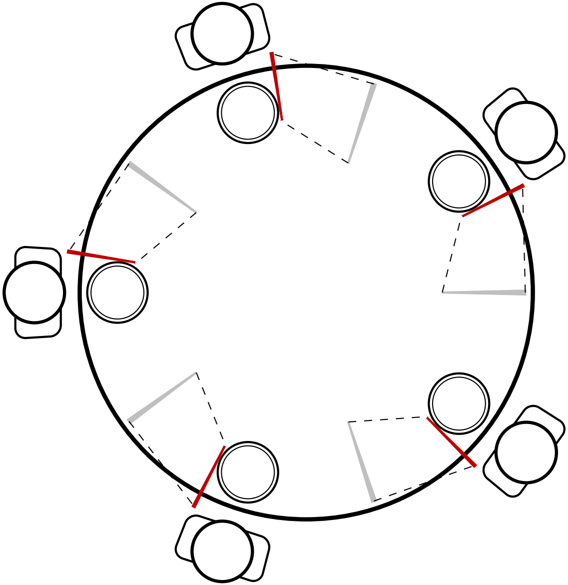


dining philosophers



five philosophers either think or eat
to eat, grab chopsticks on either side

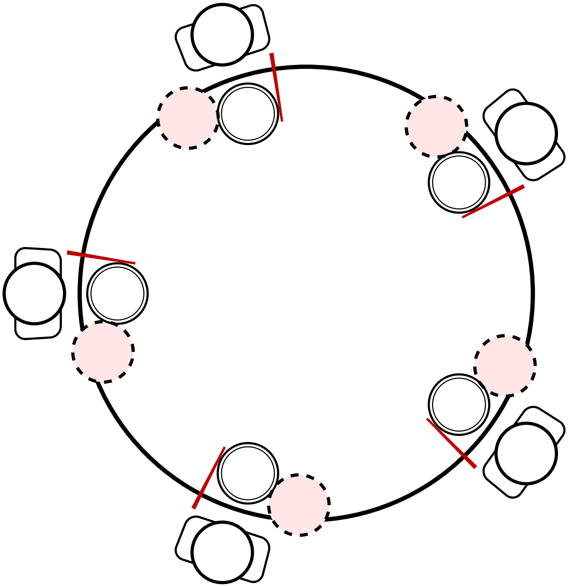
dining philosophers



everyone eats at the same time?

grab left chopstick, then...

dining philosophers



everyone eats at the same time?
grab left chopstick, then
try to grab right chopstick, …
we're at an impasse

pipe() deadlock

BROKEN example:

```
int child_to_parent_pipe[2], parent_to_child_pipe[2];
pipe(child_to_parent_pipe); pipe(parent_to_child_pipe);
if (fork() == 0) {
    /* child */
    write(child_to_parent_pipe[1], buffer, HUGE_SIZE);
    read(parent_to_child_pipe[0], buffer, HUGE_SIZE);
    exit(0);
} else {
    /* parent */
    write(parent_to_child_pipe[1], buffer, HUGE_SIZE);
    read(child_to_parent[0], buffer, HUGE_SIZE);
}
```

This will hang forever (if HUGE_SIZE is big enough).

deadlock waiting

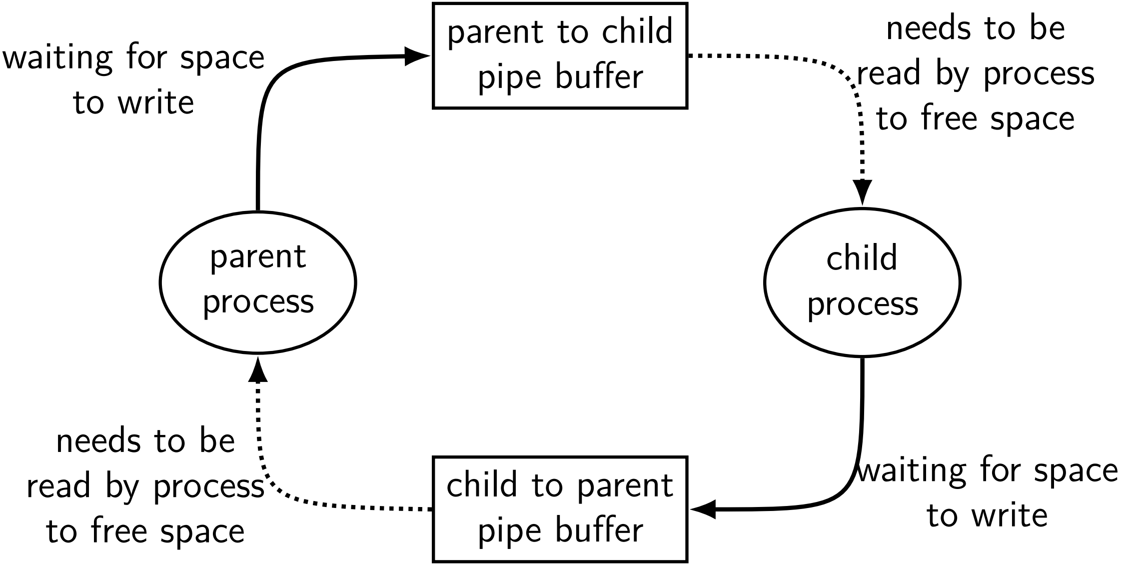
child writing to pipe waiting for free buffer space

…which will not be available until parent reads

parent writing to pipe waiting for free buffer space

…which will not be available until child reads

circular dependency



moving two files

```
struct Dir {
  mutex_t lock; map<string, DirEntry> entries;
};
void MoveFile(Dir *from_dir, Dir *to_dir, string filename) {
  mutex_lock(&from_dir->lock);
  mutex_lock(&to_dir->lock);

  to_dir->entries[filename] = from_dir->entries[filename];
  from_dir->entries.erase(filename);

  mutex_unlock(&to_dir->lock);
  mutex_unlock(&from_dir->lock);
}
```

Thread 1: `MoveFile(A, B, "foo")`

Thread 2: `MoveFile(B, A, "bar")`

moving two files: lucky timeline (1)

Thread 1	**Thread 2**
`MoveFile(A, B, "foo")`	`MoveFile(B, A, "bar")`
`lock(&A->lock);`	
`lock(&B->lock);`	
(do move)	
`unlock(&B->lock);`	
`unlock(&A->lock);`	
	`lock(&B->lock);`
	`lock(&A->lock);`
	(do move)
	`unlock(&B->lock);`
	`unlock(&A->lock);`

moving two files: lucky timeline (2)

Thread 1	Thread 2
`MoveFile(A, B, "foo")`	`MoveFile(B, A, "bar")`
`lock(&A->lock);`	
`lock(&B->lock);`	
	`lock(&B->lock…`
(do move)	(waiting for B lock)
`unlock(&B->lock);`	
	`lock(&B->lock);`
	`lock(&A->lock…`
`unlock(&A->lock);`	
	`lock(&A->lock);`
	(do move)
	`unlock(&A->lock);`
	`unlock(&B->lock);`

moving two files: unlucky timeline

Thread 1	Thread 2
`MoveFile(A, B, "foo")`	`MoveFile(B, A, "bar")`
`lock(&A->lock);`	
	`lock(&B->lock);`

moving two files: unlucky timeline

Thread 1	**Thread 2**
`MoveFile(A, B, "foo")`	`MoveFile(B, A, "bar")`
`lock(&A->lock);`	
	`lock(&B->lock);`
`lock(&B->lock`… stalled	
(waiting for lock on B)	`lock(&A->lock`… stalled
(waiting for lock on B)	(waiting for lock on A)

moving two files: unlucky timeline

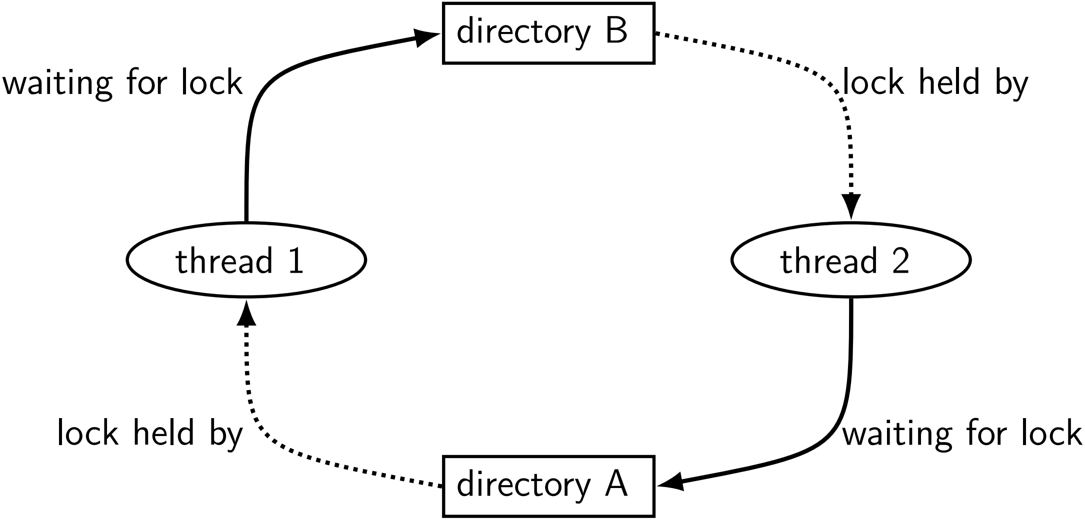
Thread 1 `MoveFile(A, B, "foo")`	Thread 2 `MoveFile(B, A, "bar")`
`lock(&A->lock);`	
	`lock(&B->lock);`
`lock(&B->lock`… stalled	
(waiting for lock on B)	`lock(&A->lock`… stalled
(waiting for lock on B)	(waiting for lock on A)
~~(do move)~~ unreachable	~~(do move)~~ unreachable
~~`unlock(&B->lock);`~~ unreachable	~~`unlock(&A->lock);`~~ unreachable
~~`unlock(&A->lock);`~~ unreachable	~~`unlock(&B->lock);`~~ unreachable

moving two files: unlucky timeline

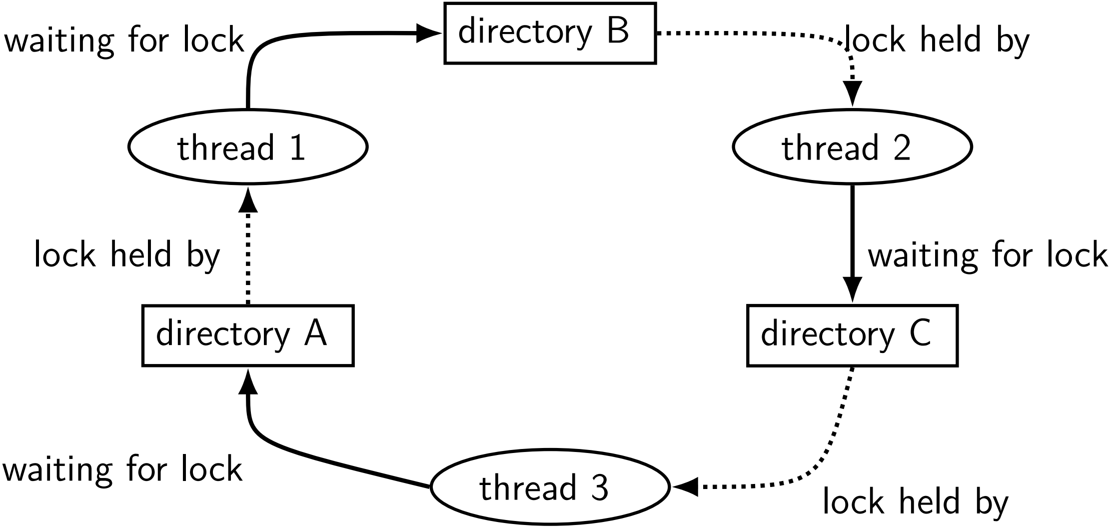
Thread 1 `MoveFile(A, B, "foo")`	Thread 2 `MoveFile(B, A, "bar")`
`lock(&A->lock);`	
	`lock(&B->lock);`
`lock(&B->lock…` stalled	
(waiting for lock on B)	`lock(&A->lock…` stalled
(waiting for lock on B)	(waiting for lock on A)
~~(do move)~~ unreachable	~~(do move)~~ unreachable
~~`unlock(&B->lock);`~~ unreachable	~~`unlock(&A->lock);`~~ unreachable
~~`unlock(&A->lock);`~~ unreachable	~~`unlock(&B->lock);`~~ unreachable

Thread 1 holds A lock, waiting for Thread 2 to release B lock
Thread 2 holds B lock, waiting for Thread 1 to release A lock

moving two files: dependencies



moving three files: dependencies



moving three files: unlucky timeline

Thread 1	Thread 2	Thread 3
`MoveFile(A, B, "foo")`	`MoveFile(B, C, "bar")`	`MoveFile(C, A, "quux")`
`lock(&A->lock);`		
	`lock(&B->lock);`	
		`lock(&C->lock);`
`lock(&B->lock…` stalled		
	`lock(&C->lock…` stalled	
		`lock(&A->lock…` stalled

deadlock with free space

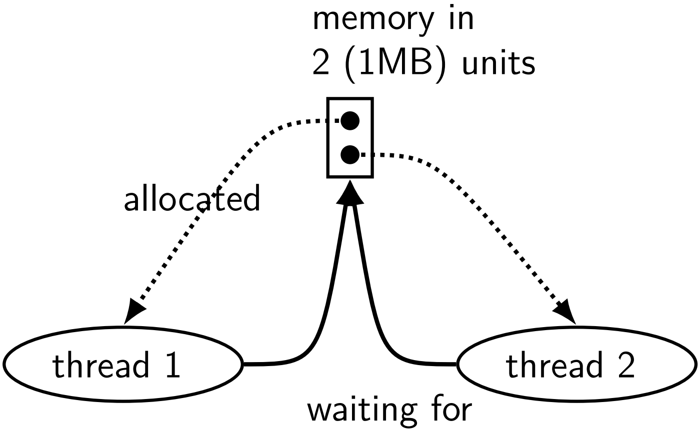
Thread 1	Thread 2
`AllocateOrWaitFor(1 MB)`	`AllocateOrWaitFor(1 MB)`
`AllocateOrWaitFor(1 MB)`	`AllocateOrWaitFor(1 MB)`
`(do calculation)`	`(do calculation)`
`Free(1 MB)`	`Free(1 MB)`
`Free(1 MB)`	`Free(1 MB)`

2 MB of space — deadlock possible with unlucky order

deadlock with free space (unlucky case)

Thread 1	Thread 2
`AllocateOrWaitFor(1 MB)`	
	`AllocateOrWaitFor(1 MB)`
`AllocateOrWaitFor(1 MB...` stalled	
	`AllocateOrWaitFor(1 MB...` stalled

free space: dependency graph



deadlock with free space (lucky case)

Thread 1	Thread 2
`AllocateOrWaitFor(1 MB)`	
`AllocateOrWaitFor(1 MB)`	
`(do calculation)`	
`Free(1 MB);`	
`Free(1 MB);`	
	`AllocateOrWaitFor(1 MB)`
	`AllocateOrWaitFor(1 MB)`
	`(do calculation)`
	`Free(1 MB);`
	`Free(1 MB);`

deadlock

deadlock — circular waiting for resources

resource = something needed by a thread to do work

- locks
- CPU time
- disk space
- memory
- ...

often non-deterministic in practice

most common example: when acquiring multiple locks

deadlock

deadlock — circular waiting for resources

resource = something needed by a thread to do work

- locks
- CPU time
- disk space
- memory
- ...

often non-deterministic in practice

most common example: when acquiring multiple locks

deadlock versus starvation

starvation: one+ unlucky (no progress), one+ lucky (yes progress)

- example: low priority threads versus high-priority threads

deadlock: no one involved in deadlock makes progress

deadlock versus starvation

starvation: one+ unlucky (no progress), one+ lucky (yes progress)

- example: low priority threads versus high-priority threads

deadlock: no one involved in deadlock makes progress

starvation: once starvation happens, taking turns will resolve

- low priority thread just needed a chance...

deadlock: once it happens, taking turns won't fix

deadlock requirements

mutual exclusion

one thread at a time can use a resource

hold and wait

thread holding a resources waits to acquire *another* resource

no preemption of resources

resources are only released voluntarily

thread trying to acquire resources can't 'steal'

circular wait

there exists a set $\{T_1, \ldots, T_n\}$ of waiting threads such that

- T_1 is waiting for a resource held by T_2
- T_2 is waiting for a resource held by T_3
- ...
- T_n is waiting for a resource held by T_1

how is deadlock possible?

Given list: A, B, C, D, E

```
RemoveNode(LinkedListNode *node) {
    pthread_mutex_lock(&node->lock);
    pthread_mutex_lock(&node->prev->lock);
    pthread_mutex_lock(&node->next->lock);
    node->next->prev = node->prev;
    node->prev->next = node->next;
    pthread_mutex_unlock(&node->next->lock);
    pthread_mutex_unlock(&node->prev->lock);
    pthread_mutex_unlock(&node->lock);
}
```

Which of these (all run in parallel) can deadlock?

A. RemoveNode(B) and RemoveNode(D)
B. RemoveNode(B) and RemoveNode(C)
C. RemoveNode(B) and RemoveNode(C) and RemoveNode(D)
D. A and C
E. B and C
F. all of the above
G. none of the above

how is deadlock — solution

Remove B	Remove C
lock B	lock C
lock A (prev)	wait to lock B (prev)
wait to lock C (next)	

With B and D — only overlap in in node C — no circular wait possible

deadlock prevention techniques

infinite resources or at least enough that never run out	no *mutual exclusion*
no shared resources	no *mutual exclusion*
no waiting "busy signal" — abort and retry revoke/preempt resources	no *hold and wait*/*preemption*
acquire resources in **consistent order**	no *circular wait*
request **all resources at once**	no *hold and wait*

deadlock prevention techniques

infinite resources
or at least enough that never run out

no *mutual exclusion*

no shared resources

no *mutual exclusion*

no waiting
"busy signal" — abort and retry
revoke/preempt resources

no *hold and wait*/*preemption*

acquire resources in **consistent order**

no *circular wait*

request **all resources at once**

no *hold and wait*

deadlock prevention techniques

infinite resources or at least enough that never run out	no *mutual exclusion*
no shared resources	no *mutual exclusion*
no waiting “busy signal” — abort and retry revoke/preempt resources	no *hold and wait*/ *preemption*
acquire resources in **consistent order**	no *circular wait*
request **all resources at once**	no *hold and wait*

deadlock prevention techniques

infinite resources or at least enough that never run out	no *mutual exclusion*
no shared resources	no *mutual exclusion*
no waiting "busy signal" — abort and retry revoke/preempt resources	no *hold and wait/ preemption*
acquire resources in **consistent order**	no *circular wait*
request **all resources at once**	no *hold and wait*

AllocateOrFail

Thread 1	Thread 2
`AllocateOrFail(1 MB)`	
	`AllocateOrFail(1 MB)`
`AllocateOrFail(1 MB)` fails!	
	`AllocateOrFail(1 MB)` fails!
`Free(1 MB)` (cleanup after failure)	
	`Free(1 MB)` (cleanup after failure)

okay, now what?

- give up?
- both try again? — maybe this will keep happening? (called livelock)
- try one-at-a-time? — gaurenteed to work, but tricky to implement

AllocateOrSteal

Thread 1	Thread 2
`AllocateOrSteal(1 MB)`	
	`AllocateOrSteal(1 MB)`
`AllocateOrSteal(1 MB)`	`Thread killed to free 1MB`
(do work)	

problem: can one actually implement this?

problem: can one kill thread and keep system in consistent state?

fail/steal with locks

pthreads provides pthread_mutex_trylock — “lock or fail”

some databases implement *revocable locks*

- do equivalent of throwing exception in thread to ‘steal’ lock
- need to carefully arrange for operation to be cleaned up

deadlock prevention techniques

infinite resources or at least enough that never run out	no *mutual exclusion*
no shared resources	no *mutual exclusion*
no waiting "busy signal" — abort and retry revoke/preempt resources	no *hold and wait*/ *preemption*
acquire resources in **consistent order**	no *circular wait*
request **all resources at once**	no *hold and wait*

abort and retry limits?

abort-and-retry

how many times will you retry?

moving two files: abort-and-retry

```
struct Dir {
  mutex_t lock; map<string, DirEntry> entries;
};
void MoveFile(Dir *from_dir, Dir *to_dir, string filename) {
  while (true) {
    mutex_lock(&from_dir->lock);
    if (mutex_trylock(&to_dir->lock) == LOCKED) break;
    mutex_unlock(&from_dir->lock);
  }

  to_dir->entries[filename] = from_dir->entries[filename];
  from_dir->entries.erase(filename);

  mutex_unlock(&to_dir->lock);
  mutex_unlock(&from_dir->lock);
}
```

Thread 1: `MoveFile(A, B, "foo")`

Thread 2: `MoveFile(B, A, "bar")`

moving two files: lots of bad luck?

Thread 1	Thread 2
`MoveFile(A, B, "foo")`	`MoveFile(B, A, "bar")`
`lock(&A->lock)` → `LOCKED`	
	`lock(&B->lock)` → `LOCKED`
`trylock(&B->lock)` → `FAILED`	
	`trylock(&A->lock)` → `FAILED`
`unlock(&A->lock)`	
	`unlock(&B->lock)`
`lock(&A->lock)` → `LOCKED`	
	`lock(&B->lock)` → `LOCKED`
`trylock(&B->lock)` → `FAILED`	
	`trylock(&A->lock)` → `FAILED`
`unlock(&A->lock)`	
	`unlock(&B->lock)`

livelock

livelock: keep aborting and retrying without end

like deadlock — no one's making progress

potentially forever

unlike deadlock — threads are not waiting

preventing livelock

make schedule random — e.g. random waiting after abort

make threads run one-at-a-time if lots of aborting

other ideas?

deadlock prevention techniques

infinite resources or at least enough that never run out	no *mutual exclusion*
no shared resources	no *mutual exclusion*
no waiting "busy signal" — abort and retry revoke/preempt resources	no *hold and wait*/ *preemption*
acquire resources in **consistent order**	no *circular wait*
request **all resources at once**	no *hold and wait*

stealing locks???

how do we make stealing locks possible

unclean: just kill the thread

- problem: inconsistent state?

clean: have code to undo partial oepration

- some databases do this

won't go into detail in this class

revokable locks?

```
try {
    AcquireLock();
    use shared data
} catch (LockRevokedException le) {
    undo operation hopefully?
} finally {
    ReleaseLock();
}
```

deadlock prevention techniques

infinite resources or at least enough that never run out	no *mutual exclusion*
no shared resources	no *mutual exclusion*
no waiting "busy signal" — abort and retry revoke/preempt resources	no *hold and wait*/ *preemption*
acquire resources in **consistent order**	no *circular wait*
request **all resources at once**	no *hold and wait*

acquiring locks in consistent order (1)

```
MoveFile(Dir* from_dir, Dir* to_dir, string filename) {
  if (from_dir->path < to_dir->path) {
    lock(&from_dir->lock);
    lock(&to_dir->lock);
  } else {
    lock(&to_dir->lock);
    lock(&from_dir->lock);
  }
  ...
}
```

acquiring locks in consistent order (1)

```
MoveFile(Dir* from_dir, Dir* to_dir, string filename) {
  if (from_dir->path < to_dir->path) {
    lock(&from_dir->lock);
    lock(&to_dir->lock);
  } else {
    lock(&to_dir->lock);
    lock(&from_dir->lock);
  }
  ...
}
```

any ordering will do
e.g. compare pointers

acquiring locks in consistent order (2)

often by convention, e.g. Linux kernel comments:

```
/*
 * ...
 * Lock order:
 *      contex.ldt_usr_sem
 *        mmap_sem
 *          context.lock
 */
```

```
/*
 * ...
 * Lock order:
 *   1. slab_mutex (Global Mutex)
 *   2. node->list_lock
 *   3. slab_lock(page) (Only on some arches and for debugging)
 * ...
 */
```

deadlock prevention techniques

infinite resources or at least enough that never run out	no *mutual exclusion*
no shared resources	no *mutual exclusion*
no waiting "busy signal" — abort and retry revoke/preempt resources	no *hold and wait*/ *preemption*
acquire resources in **consistent order**	no *circular wait*
request **all resources at once**	no *hold and wait*

allocating all at once?

for resources like disk space, memory

figure out maximum allocation when starting thread

"only" need conservative estimate

only start thread if those resources are available

okay solution for embedded systems?

deadlock detection

idea: search for cyclic dependencies

detecting deadlocks on locks

let's say I want to detect deadlocks that only involve mutexes

goal: help programmers debug deadlocks

…by modifying my threading library:

```
struct Thread {
    ... /* stuff for implementing thread */
    /* what extra fields go here? */


};

struct Mutex {
    ... /* stuff for implementing mutex */
    /* what extra fields go here? */


};
```

deadlock detection

idea: search for cyclic dependencies

need:

- list of all contended resources
- what thread is waiting for what?
- what thread 'owns' what?

aside: divisible resources

deadlock is possible with divislbe resources like memory,...

example: suppose 6MB of RAM for threads total:

- thread 1 has 2MB allocated, waiting for 2MB
- thread 2 has 2MB allocated, waiting for 2MB
- thread 3 has 1MB allocated, waiting for keypress

cycle: thread 1 waiting on memory owned by thread 2?

not a deadlock — thread 3 can still finish

- and after it does, thread 1 or 2 can finish

aside: divisible resources

deadlock is possible with divislbe resources like memory,...

example: suppose 6MB of RAM for threads total:

- thread 1 has 2MB allocated, waiting for 2MB
- thread 2 has 2MB allocated, waiting for 2MB
- thread 3 has 1MB allocated, waiting for keypress

cycle: thread 1 waiting on memory owned by thread 2?

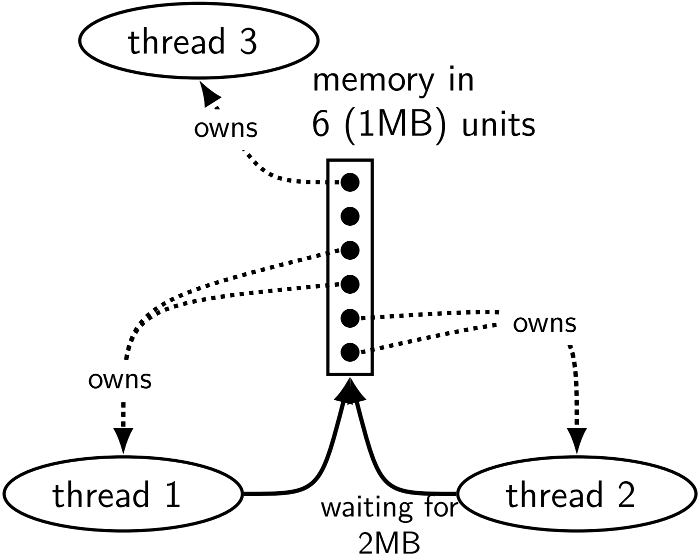
not a deadlock — thread 3 can still finish

- and after it does, thread 1 or 2 can finish

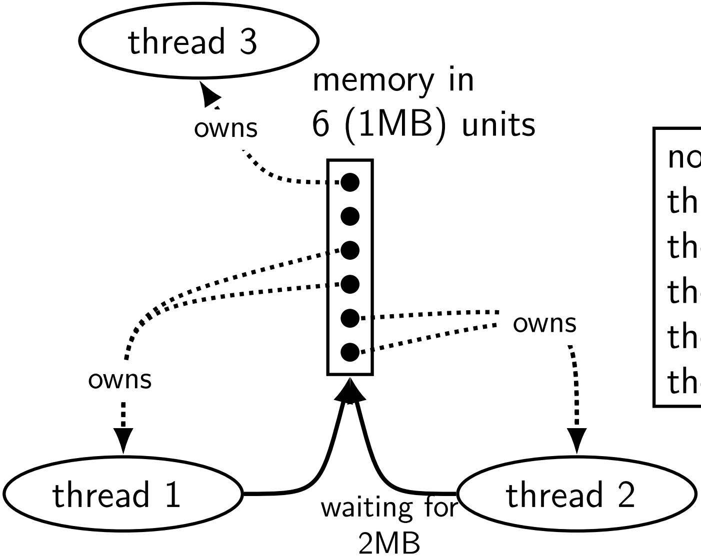
...but would be deadlock

- ...if thread 3 waiting lock held by thread 1
- ...with 5MB of RAM

divisible resources: not deadlock

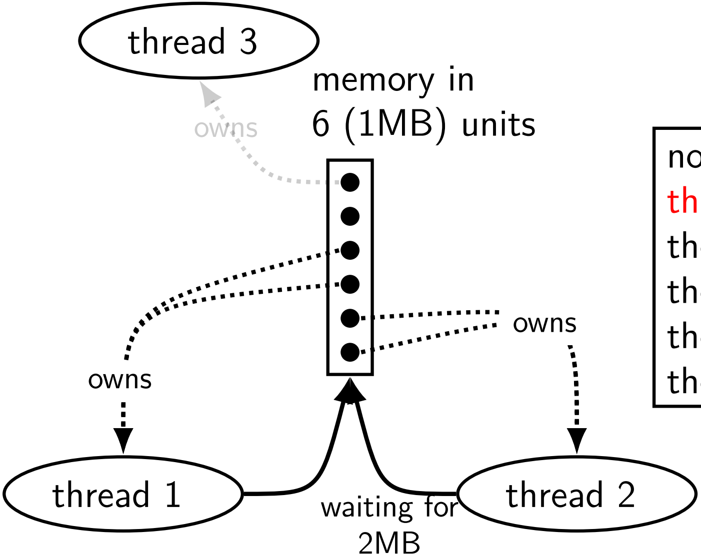


divisible resources: not deadlock



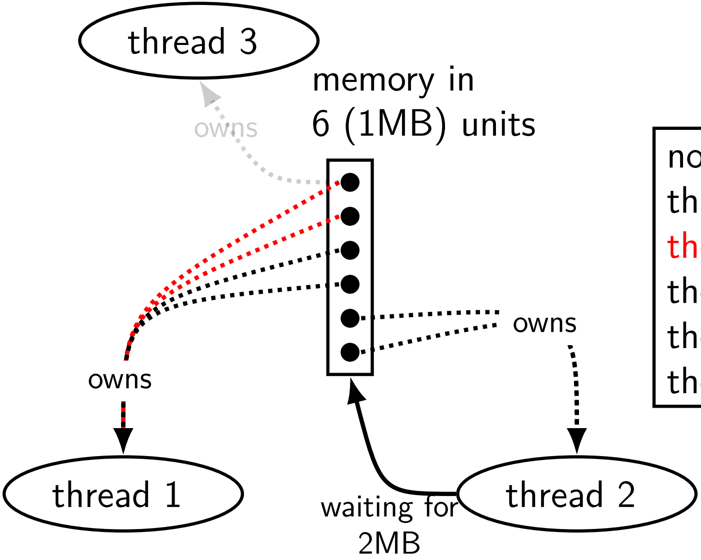
not deadlock:
thread 3 finishes
then thread 1 can get memory
then thread 1 finishes
then thread 2 can get resources
then thread 2 can finish

divisible resources: not deadlock



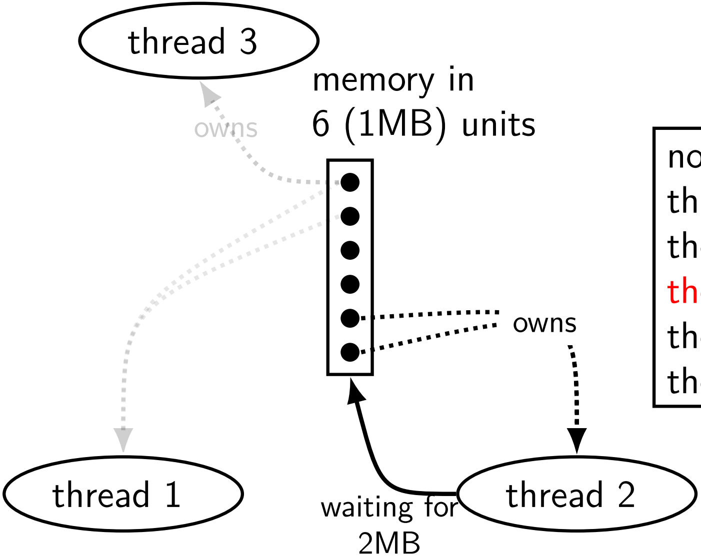
not deadlock:
thread 3 finishes
then thread 1 can get memory
then thread 1 finishes
then thread 2 can get resources
then thread 2 can finish

divisible resources: not deadlock



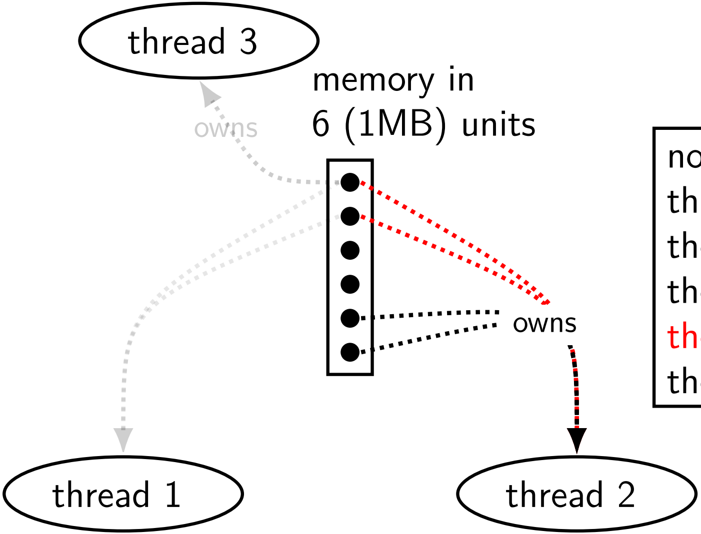
not deadlock:
thread 3 finishes
then thread 1 can get memory
then thread 1 finishes
then thread 2 can get resources
then thread 2 can finish

divisible resources: not deadlock



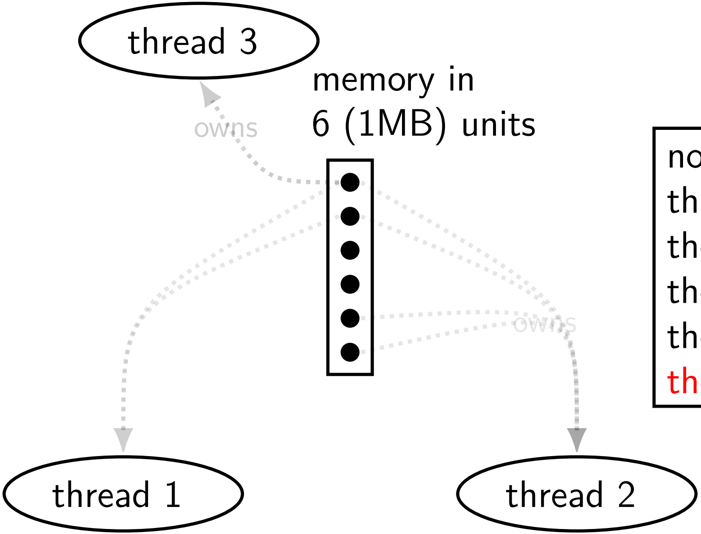
not deadlock:
thread 3 finishes
then thread 1 can get memory
then thread 1 finishes
then thread 2 can get resources
then thread 2 can finish

divisible resources: not deadlock



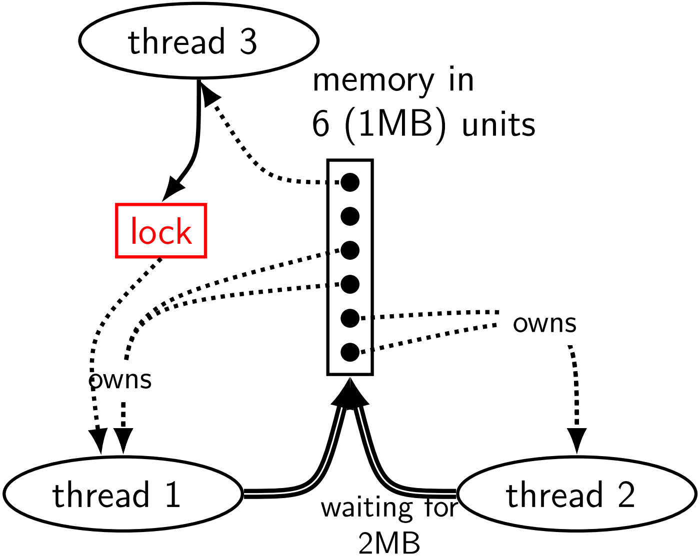
not deadlock:
thread 3 finishes
then thread 1 can get memory
then thread 1 finishes
then thread 2 can get resources
then thread 2 can finish

divisible resources: not deadlock

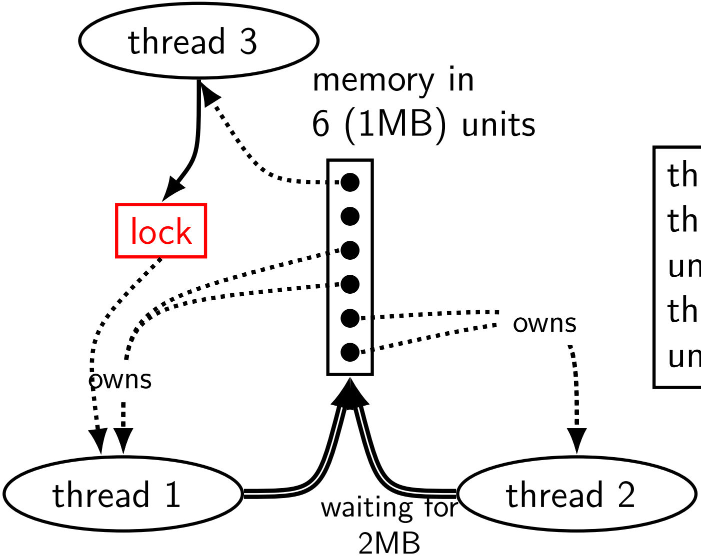


not deadlock:
thread 3 finishes
then thread 1 can get memory
then thread 1 finishes
then thread 2 can get resources
then thread 2 can finish

divisible resources: is deadlock

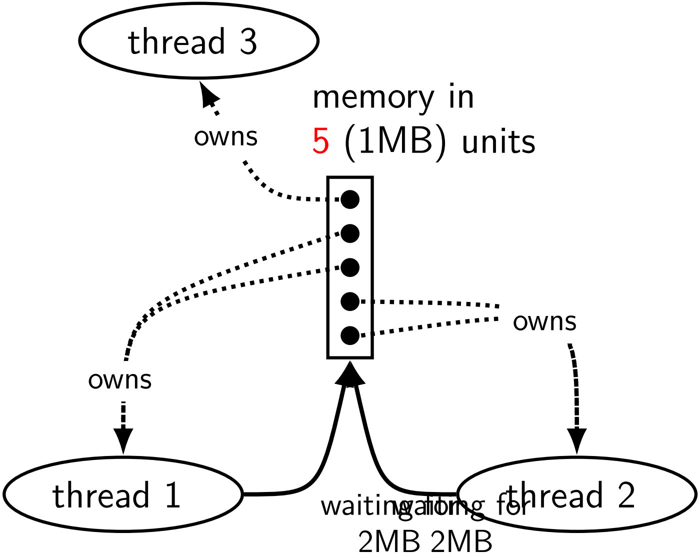


divisible resources: is deadlock

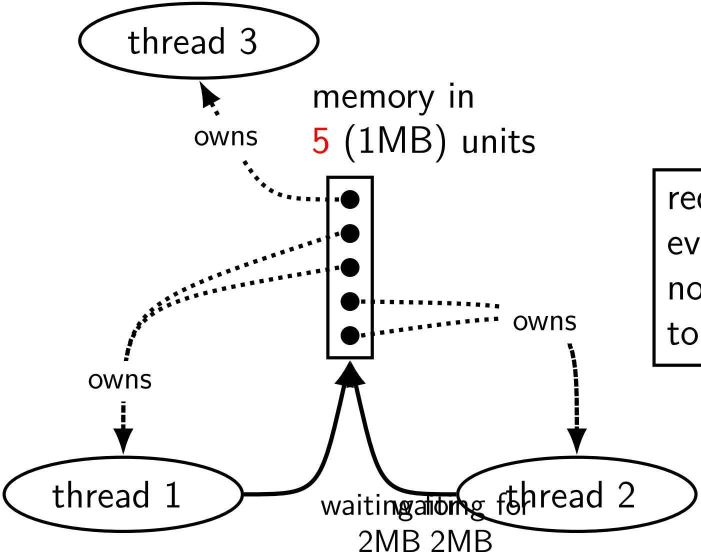


this is deadlock:
thread 3 can't finish
until thread 1 releases lock, but
thread 1 can't finish
until thread 3 releases memory

divisible resources: is deadlock

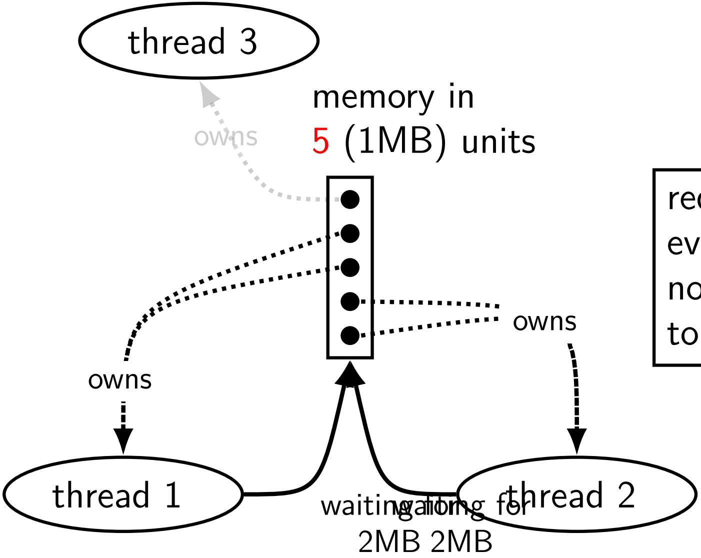


divisible resources: is deadlock



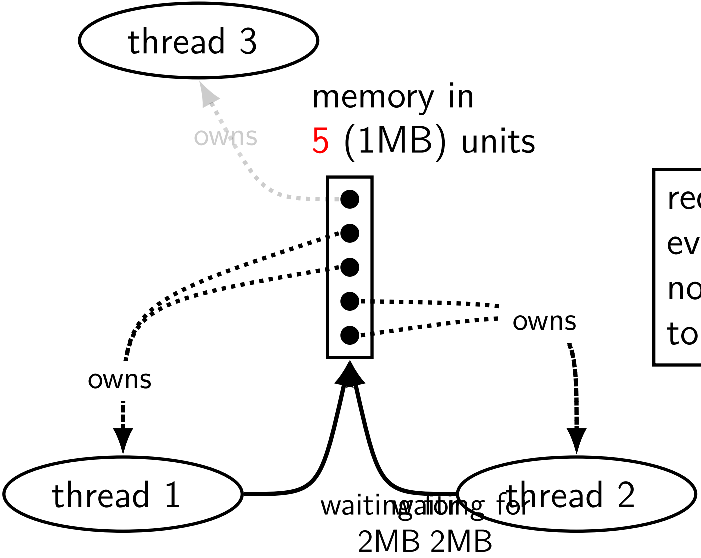
reducing memory: this is deadloc
even after thread 3 finishes
no way for thread 1+2
to get what they want

divisible resources: is deadlock



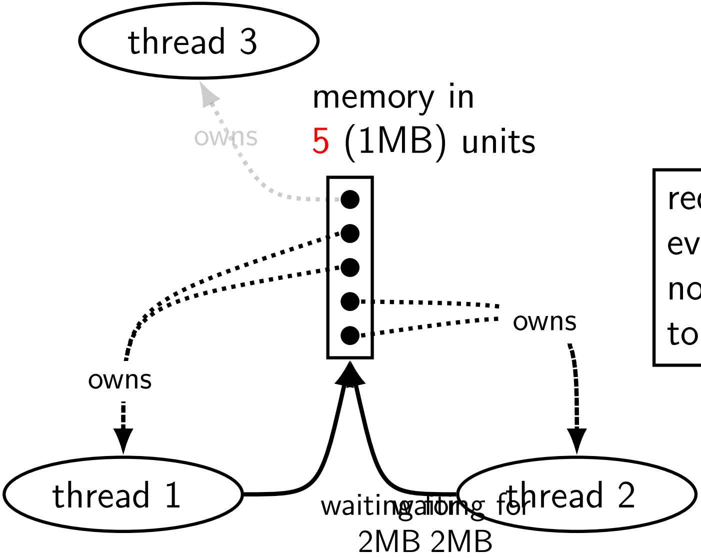
reducing memory: this is deadloc
even after thread 3 finishes
no way for thread 1+2
to get what they want

divisible resources: is deadlock



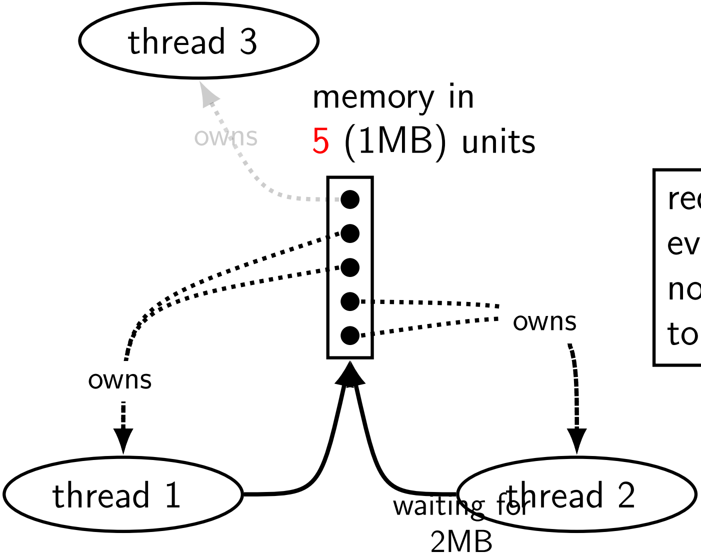
reducing memory: this is deadloc
even after thread 3 finishes
no way for thread 1+2
to get what they want

divisible resources: is deadlock



reducing memory: this is deadloc
even after thread 3 finishes
no way for thread 1+2
to get what they want

divisible resources: is deadlock



reducing memory: this is deadloc
even after thread 3 finishes
no way for thread 1+2
to get what they want

deadlock detection with divisibe resources

can't rely on cycles in graphs in this case

alternate algorithm exists

- similar technique to how we showed no deadlock

high-level intuition: simulate what could happen

- find threads that could finish based on resources available now

full details: look up Baker's algorithm

backup slides

rwlocks with monitors (attempt 1)

```
mutex_t lock;
```

lock to protect shared state

rwlocks with monitors (attempt 1)

```
mutex_t lock;
unsigned int readers, writers;
```

state: number of active readers, writers

rwlocks with monitors (attempt 1)

```
mutex_t lock;
unsigned int readers, writers;
/* condition, signal when writers becomes 0 */
cond_t ok_to_read_cv;
/* condition, signal when readers + writers becomes 0 */
cond_t ok_to_write_cv;
```

conditions to wait for (no readers or writers, no writers)

rwlocks with monitors (attempt 1)

```
mutex_t lock;
unsigned int readers, writers;
/* condition, signal when writers becomes 0 */
cond_t ok_to_read_cv;
/* condition, signal when readers + writers becomes 0 */
cond_t ok_to_write_cv;
```

```
ReadLock() {
  mutex_lock(&lock);
  while (writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  ++readers;
  mutex_unlock(&lock);
}
ReadUnlock() {
  mutex_lock(&lock);
  --readers;
  if (readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
  mutex_unlock(&lock);
}
```

```
WriteLock() {
  mutex_lock(&lock);
  while (readers + writers != 0) {
    cond_wait(&ok_to_write_cv);
  }
  ++writers;
  mutex_unlock(&lock);
}
WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  cond_signal(&ok_to_write_cv);
  cond_broadcast(&ok_to_read_cv);
  mutex_unlock(&lock);
}
```

broadcast — wakeup all readers when no writers

rwlocks with monitors (attempt 1)

```
mutex_t lock;
unsigned int readers, writers;
/* condition, signal when writers becomes 0 */
cond_t ok_to_read_cv;
/* condition, signal when readers + writers becomes 0 */
cond_t ok_to_write_cv;
```

```
ReadLock() {
  mutex_lock(&lock);
  while (writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  ++readers;
  mutex_unlock(&lock);
}
ReadUnlock() {
  mutex_lock(&lock);
  --readers;
  if (readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
  mutex_unlock(&lock);
}
```

```
WriteLock() {
  mutex_lock(&lock);
  while (readers + writers != 0) {
    cond_wait(&ok_to_write_cv);
  }
  ++writers;
  mutex_unlock(&lock);
}
WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  cond_signal(&ok_to_write_cv);
  cond_broadcast(&ok_to_read_cv);
  mutex_unlock(&lock);
}
```

wakeup a single writer when no readers or writers

rwlocks with monitors (attempt 1)

```
mutex_t lock;
unsigned int readers, writers;
/* condition, signal when writers becomes 0 */
cond_t ok_to_read_cv;
/* condition, signal when readers + writers becomes 0 */
cond_t ok_to_write_cv;
```

```
ReadLock() {
  mutex_lock(&lock);
  while (writers != 0) {
    cond_wait(&ok_to_read_cv, &lock);
  }
  ++readers;
  mutex_unlock(&lock);
}
ReadUnlock() {
  mutex_lock(&lock);
  --readers;
  if (readers == 0) {
    cond_signal(&ok_to_write_cv);
  }
  mutex_unlock(&lock);
}
```

```
WriteLock() {
  mutex_lock(&lock);
  while (readers + writers != 0) {
    cond_wait(&ok_to_write_cv);
  }
  ++writers;
  mutex_unlock(&lock);
}
WriteUnlock() {
  mutex_lock(&lock);
  --writers;
  cond_signal(&ok_to_write_cv);
  cond_broadcast(&ok_to_read_cv);
  mutex_unlock(&lock);
}
```

problem: wakeup readers first or writer first?

this solution: wake them all up and they fight! inefficient!

resource allocation graphs

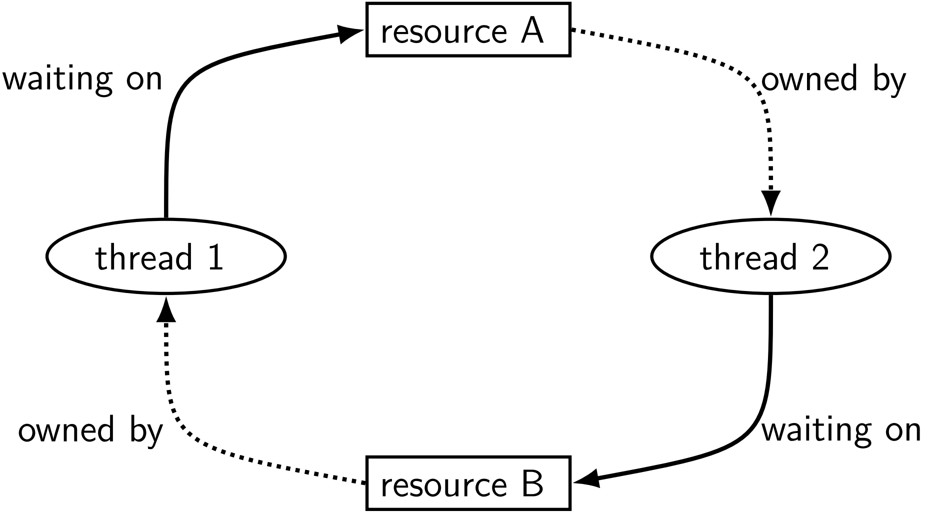
nodes: resources or threads

edge thread$\rightarrow$resource: thread waiting for resource

edge resource$\rightarrow$thread: resource is "owned" by thread

- holds lock on
- will be deallocated by
- ...

resource allocate graphs



searching for cycles

cycle $\rightarrow$ deadlock happened!

finding cycles: recall 2150 topological sort (maybe???)

resource allocation graphs and quantity

so far: assuming resource is fully taken or not at all taken

what about resources like memory?

- two processes can take parts of resource
- …but deadlock still possible

there's a version of resource allocation graphs for this case

using deadlock detection for prevention

suppose you know the *maximum resources* a process could request

make decision when starting process ("*admission control*")

using deadlock detection for prevention

suppose you know the *maximum resources* a process could request

make decision when starting process ("*admission control*")

ask "what if every process was waiting for maximum resources"

including the one we're starting

would it cause deadlock? then don't let it start

called Baker's algorithm

building semaphore with monitors (version B)

```
pthread_mutex_t lock;
unsigned int count;
/* condition, broadcast when becomes count > 0 */
pthread_cond_t count_is_positive_cv;
void down() {
    pthread_mutex_lock(&lock);
    while (!(count > 0)) {
        pthread_cond_wait(
            &count_is_positive_cv,
            &lock);
    }
    count -= 1;
    pthread_mutex_unlock(&lock);
}
```

```
void up() {
    pthread_mutex_lock(&lock);
    count += 1;
    /* condition *just* became true */
    if (count == 1) {
        pthread_cond_broadcast(
            &count_is_positive_cv
        );
    }
    pthread_mutex_unlock(&lock);
}
```

before: signal every time

can check if condition just became true instead?

building semaphore with monitors (version B)

```
pthread_mutex_t lock;
unsigned int count;
/* condition, broadcast when becomes count > 0 */
pthread_cond_t count_is_positive_cv;
void down() {
    pthread_mutex_lock(&lock);
    while (!(count > 0)) {
        pthread_cond_wait(
            &count_is_positive_cv,
            &lock);
    }
    count -= 1;
    pthread_mutex_unlock(&lock);
}
```

```
void up() {
    pthread_mutex_lock(&lock);
    count += 1;
    /* condition *just* became true */
    if (count == 1) {
        pthread_cond_broadcast(
            &count_is_positive_cv
        );
    }
    pthread_mutex_unlock(&lock);
}
```

before: signal every time

can check if condition just became true instead?

but do we really need to broadcast?

exercise: why broadcast?

```
pthread_mutex_t lock;
unsigned int count;
/* condition, broadcast when becomes count > 0 */
pthread_cond_t count_is_positive_cv;
void down() {
    pthread_mutex_lock(&lock);
    while (!(count > 0)) {
        pthread_cond_wait(
            &count_is_positive_cv,
            &lock);
    }
    count -= 1;
    pthread_mutex_unlock(&lock);
}
```

```
void up() {
    pthread_mutex_lock(&lock);
    count += 1;
    if (count == 1) { /* became > 0 */
        pthread_cond_broadcast(
            &count_is_positive_cv
        );
    }
    pthread_mutex_unlock(&lock);
}
```

exercise: why can't this be `pthread_cond_signal`?

hint: think of two threads calling down + two calling up?

brute force: only so many orders they can get the lock in

broadcast problem

Thread 1	Thread 2	Thread 3	Thread 4
Down()			
lock			
count == 0? yes			
unlock/wait			
	Down()		
	lock		
	count == 0? yes		
	unlock/wait		
		Up()	
		lock	
		count += 1 (now 1)	Up()
stop waiting on CV		signal	wait for lock
wait for lock		unlock	wait for lock
wait for lock			lock
wait for lock			count += 1 (now 2)
wait for lock			count != 1: don't signal
lock			unlock
count == 0? no			
count -= 1 (becomes 1)			
unlock			
	still waiting???		

broadcast problem

Thread 1	Thread 2	Thread 3	Thread 4
Down()			
lock			
count == 0? yes			
unlock/wait			
	Down()		
	lock		
	count == 0? yes		
	unlock/wait		
		Up()	
		lock	
		count += 1 (now 1)	Up()
stop waiting on CV		signal	wait for lock
wait for lock		unlock	wait for lock
wait for lock			lock
wait for lock			count += 1 (now 2)
wait for lock			count != 1: don't signal
lock			unlock
count == 0? no			
count -= 1 (becomes 1)			
unlock			
	still waiting???		

broadcast problem

Thread 1	Thread 2	Thread 3	Thread 4
Down()			
lock			
count == 0? yes			
unlock/wait			
	Down()		
	lock		
	count == 0? yes		
	unlock/wait		
		Up()	
		lock	
		count += 1 (now 1)	Up()
stop waiting on CV		signal	wait for lock
wait for lock		unlock	wait for lock
wait for lock			lock
wait for lock			count += 1 (now 2)
wait for lock			count != 1: don't signal
lock			unlock
count == 0? no			
count -= 1 (becomes 1)			
unlock			
	still waiting???		

Mesa-style monitors signalling doesn't “hand off” lock

semaphores with monitors: no condition

```
pthread_mutex_t lock;
unsigned int count;
/* condition, broadcast when becomes count > 0 */
pthread_cond_t count_is_positive_cv;
void down() {
    pthread_mutex_lock(&lock);
    while (!(count > 0)) {
        pthread_cond_wait(
            &count_is_positive_cv,
            &lock);
    }
    count -= 1;
    pthread_mutex_unlock(&lock);
}
```

```
void up() {
    pthread_mutex_lock(&lock);
    count += 1;
    pthread_cond_signal(
        &count_is_positive_cv
    );
    pthread_mutex_unlock(&lock);
}
```

same as where we started...

semaphores with monitors: alt w/ signal

```
pthread_mutex_t lock;
unsigned int count;
/* condition, broadcast when becomes count > 0 */
pthread_cond_t count_is_positive_cv;
void down() {
    pthread_mutex_lock(&lock);
    while (!(count > 0)) {
        pthread_cond_wait(
            &count_is_positive_cv,
            &lock);
    }
    count -= 1;
    if (count > 0) {
        pthread_cond_signal(
            &count_is_positive_cv
        );
    }
    pthread_mutex_unlock(&lock);
}

void up() {
    pthread_mutex_lock(&lock);
    count += 1;
    if (count == 1) {
        pthread_cond_signal(
            &count_is_positive_cv
        );
    }
    pthread_mutex_unlock(&lock);
}
```

on signal/broadcast generally

whenever using signal need to ask
what if more than one thread is waiting?

need to explain why those threads will be signalled eventually

…even if next thread signalled doesn't run right away

another problem that would be avoided with Hoare scheduling