

Introduction

- Suppose some madman says “We shouldn’t use locks!”
- You know that this results (eventually!) in inconsistent data structures.
 - Loss of invariants within the data structure
 - Live pointers to dead memory
 - Live pointers to undead memory (Hey, my type changed! Stop poking there!)

Introduction

Locks Might Take A While

- Consider XCHG style locks which use `while(xchg(&locked, LOCKED) == LOCKED)` as their core operation.
- We could spend an unbounded amount of time here waiting...
- This implies we'll have very high latency *on contention*...
- Locks *by definition* reduce parallelism.

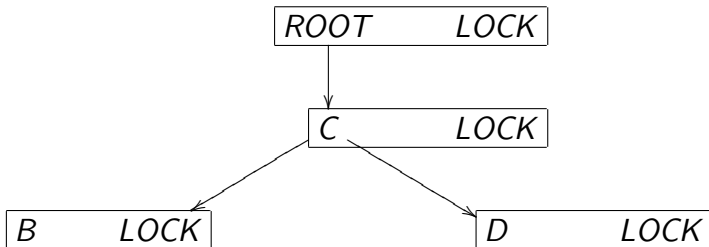
Introduction

Locks Might Take A While

- For a large data structure, we would *like* multiple *local* (independent) operations to be allowed concurrently.
 - e.g. “lookup” and “insert” in parallel threads
- Can somewhat get this with a data structure full of locks
- ...but order requirements mean that threads can still pile up while trying to get to their local site.

Locks Can Be...Not So Bad?

- Instead of a lock around a tree, we could have a tree with locks:

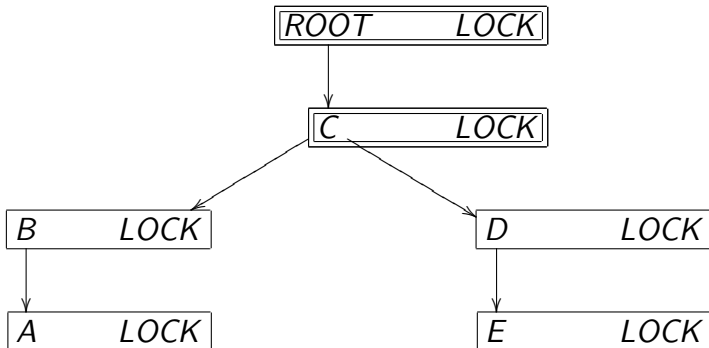


- The protocol is lock the root, then (lock child & unlock parent) as you go down.
 - This kind of *lock handoff* is a very common design.
- Here every time a thread decides to go down one branch, it gets out of roughly half of the others' ways.

Introduction

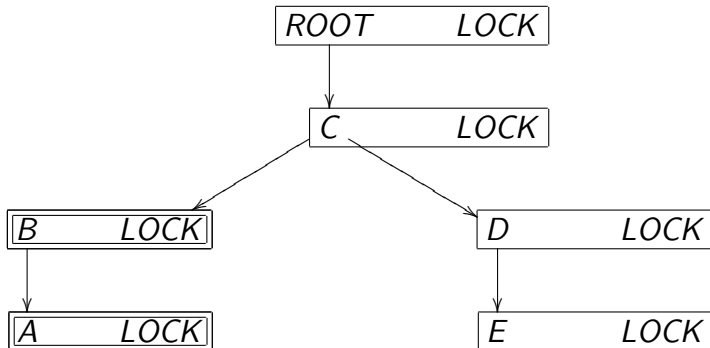
Locks Can Be...Not So Bad?

- Trying to find node A.
- Step 1: lock root pointer and top node



*Introduction**Locks Can Be... Not So Bad?*

- Trying to find node A.
- Step 3: lock left child and unlock parent



Lock-Free Linked List Node

- Node definition is simple:

label_t label

void* next

- When drawing, we'll use a shorthand:

label_t label = A

void* next = &B

 \Leftrightarrow

A	&B
---	----

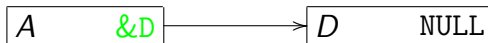
```
insertAfter(after, newlabel) {  
    //lockList();  
    new = newNode(newlabel);  
    prev = findLabel(after);  
    new->next = prev->next;  
    prev->next = new;  
    //unlockList();  
}
```

Insertion into a Linked List Without Locks
 “Good trace” in 410 notation

insertAfter(A,B)	insertAfter(A,C)
prev = &A	
B.next=A.next	
A.next=B	
	prev = &A
	C.next=A.next
	A.next=C

Insertion into a Linked List Without Locks

Precondition



- One list, two items on it: A and D .

Insertion into a Linked List Without Locks

First step

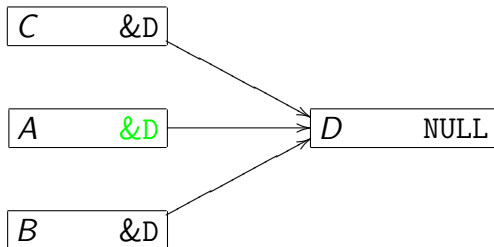


- Two threads get two nodes, B and C , and want to insert.

new = newNode(B);	new = newNode(C);
prev = &A	prev = &A

Insertion into a Linked List Without Locks

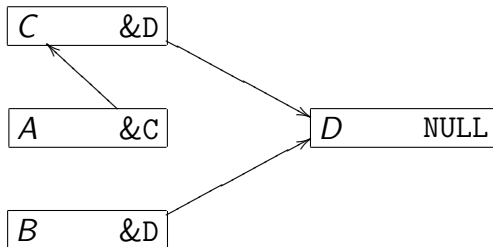
Second step



- Two threads point their respective nodes C and B into list at D

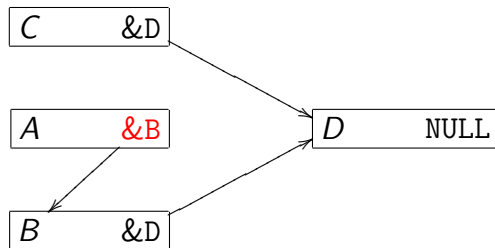
B.next=&D	C.next=&D
-----------	-----------

Insertion into a Linked List Without Locks
One thread goes



- Suppose the thread owning C completes its assignment first.





- And the other (owning B) completes second, overwriting

```
A.next=&B
```

- Node C is unreachable!

Insertion into a Linked List Without Locks

- Our assignments were really supposed to be

insertAfter(A,B)	insertAfter(A,C)
while(!done)	while(!done)
setup	setup
<i>ATOMICALLY</i> if (A->next == D) A->next = B done = 1	<i>ATOMICALLY</i> if (A->next == D) A->next = C done = 1

- If we do that, one critical section will *safely* fail out and tell us to try again.
- How do we do this **ATOMICALLY** without locking?

Review of Atomic Primitives

XCHG(ptr,new)	CAS(ptr, expect, new)
<i>ATOMICALLY</i>	<i>ATOMICALLY</i>
old = *ptr;	old = *ptr;
	if(old == expect)
*ptr = new;	*ptr = new;
return old;	return old;

Note that CAS is no harder:

- Still one read, one write under same lock.
- (logic time \ll memory time)

Insertion into a Lock-free Linked List

- Our assignments were really supposed to be

insertAfter(A,B)	insertAfter(A,C)
while(!done)	while(!done)
setup	setup
<i>ATOMICALLY</i> if (A->next == D) A->next = B done = 1	<i>ATOMICALLY</i> if (A->next == D) A->next = C done = 1

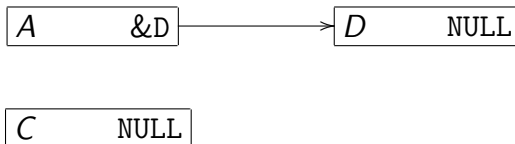
- This translates into

```
while(!done)
    prev = B->next = A->next;
    done = (CAS(&A->next,prev,B) == prev)
```

- CAS will assign if match, or bail otherwise.

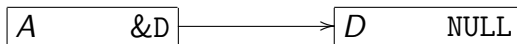
Insertion into a Lock-free Linked List

Simple case, setup

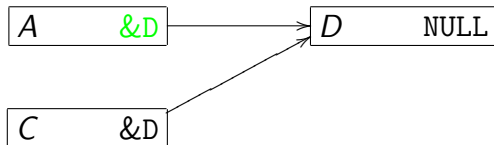


- Some thread constructs the bottom node C ; wishes to place it between the two above, A and D .
- `new = newNode(C);`
- `prev = findLabel(A); /* == &A */`

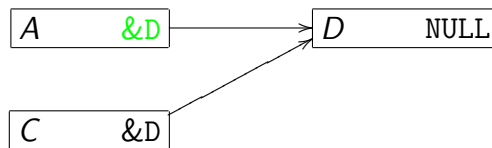
Insertion into a Lock-free Linked List
Simple case, first step



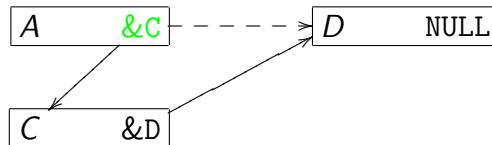
- Thread points *C* node's next into list at *D*.
- `C.next = A.next;`



Insertion into a Lock-free Linked List
Simple case, second step



- `CAS(&A.next, &D, &C);`



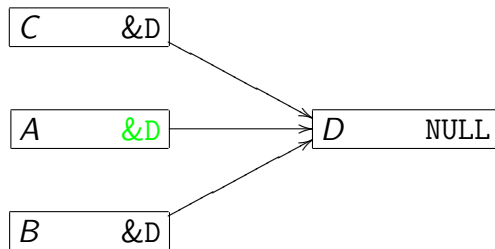
Insertion into a Lock-free Linked List
Race case, setup



- Two threads get their respective nodes B and C .

new = newNode(B);	new = newNode(C);
prev = &A	prev = &A

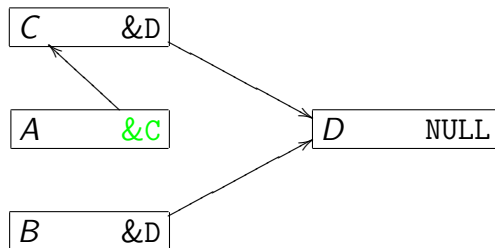
Insertion into a Lock-free Linked List
Race case, first step



- Both set their new node's next pointer.

B.next=&D	C.next=&D
-----------	-----------

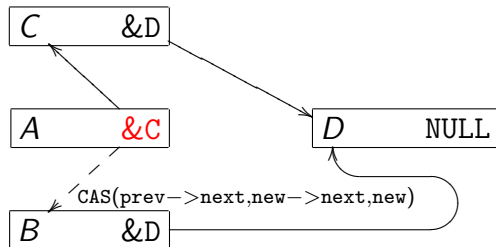
Insertion into a Lock-free Linked List
Race case, first thread



- Thread C goes first ...

	CAS(&A->next, D, C)
--	---------------------

Insertion into a Lock-free Linked List
Race case, second thread



- And the other (owning B)...

CAS(&A->next, D, B)

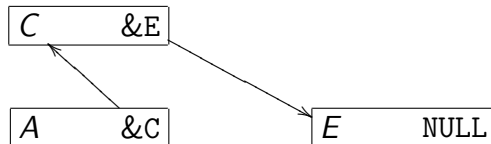
- ... fails since `A->next == C`, not `D`.
- So this thread tries again.

That's great!

- It works!
 - No locks!
 - Threads can simultaneously scan and scan the list...
 - Threads can simultaneously scan and *grow* the list!
 - Threads can simultaneously *grow* and grow the list!
- All those while loops... (retrying over and over?)
 - Remember, mutexes had while loops too...
 - maybe even around CAS()!
 - Here, whenever we retry we *know* somebody else got work done!
- Are we done?
 - Have we implemented all the standard operations?

Deletion is easy?

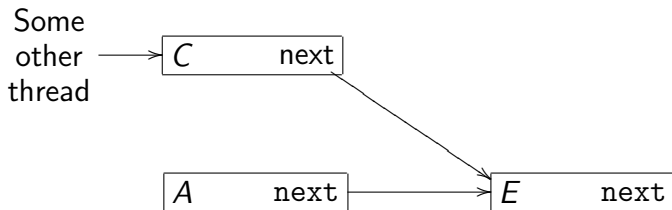
- Suppose we have



- And want to get rid of C.
- So CAS(&A.next, &C, &E)

Deletion is easy?
Continued

- But imagine there was another thread accessing C (say, scanning the list).



- We don't know when that thread is done with `C`!
- So we can never `free(C)`;

Deletion is easy?
What's to be done?

- We need *some* way to reclaim that memory for reuse..
- Some implementations cheat and assume a stop-the-world garbage collector.
 - (That's like a giant lock!)
- Doing deletion honestly is remarkably tricky!
 - We're not going to really have time to cover it.

Deletion is easy?

- Assume: once some memory is committed to being a LF list node that it's OK if it's *always* a LF list node.
- So we can have two lists: the “real” list and a “free” list.
 - This is not real free() but is hard enough.
- In particular, we run into the “ABA problem”.

ABA Problem

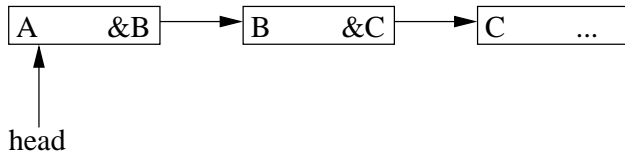
- A problem of confused identity

global = malloc(sizeof(Foo))		//0x1337
local ₁ = global	local ₂ = global	
global = NULL		
free(local ₁)		//0x1337
global = malloc(sizeof(Foo))		//0x1337
	/* Validity check */ if (global == local ₂) global->foo_baz = ...	

- Even though local_2 and global might point to the same address, they don't *really* mean the same thing.

ABA Problem Preliminaries

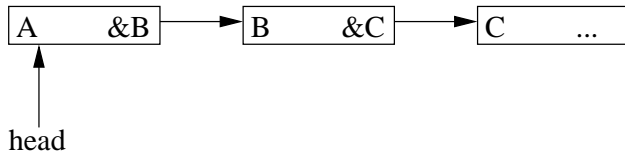
- We begin with an innocent linked list:



- Where head is a global pointer to the list.
- We're just going to do operations at the head – treating the list like a stack.

ABA Problem
Pop

- We begin with a linked list:



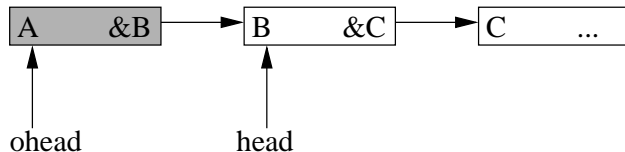
- Removing the head looks like

ohead = head	/* == &A */
onext = ohead->next	/* == &B */
CAS(head, ohead, onext);	

- If not, retry.

ABA Problem
Pop

- If successful,



- is the result of

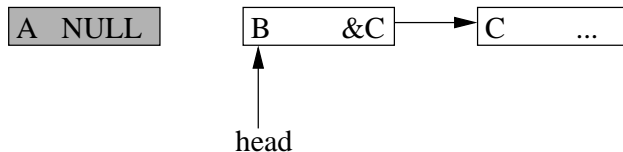
ohead = head	/* == &A */
onext = ohead->next	/* == &B */
CAS(head, ohead, onext);	

- If not, retry.

ABA Problem

Push

- We begin with a linked list and private item



- Inserting at the head looks like

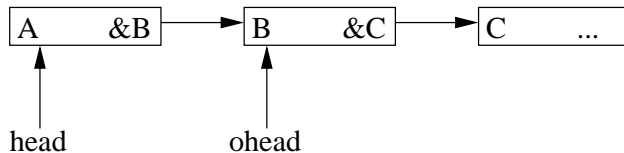
<code>ohead = head</code>	<code>/* == &B */</code>
<code>A.next = ohead</code>	<code>/* A points at B */</code>
<code>CAS(head, ohead, &A);</code>	

- If not, retry.

ABA Problem

Push

- If that works, we get



- from

ohead = head	/* == &B */
A.next = ohead	/* A points at B */
CAS(head, ohead, &A);	

- If not, retry.

ABA Problem

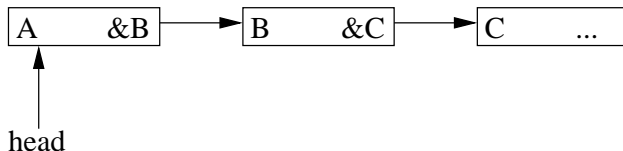
And now it breaks!

Three threads:

Thread 1 Pop an item.

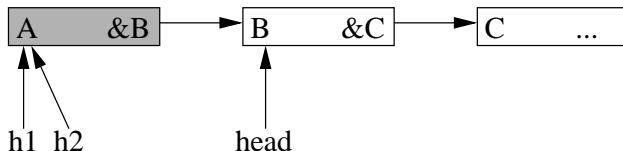
Thread 2 Pop an item, then push it right back.

Thread 3 Pop an item.

ABA Problem

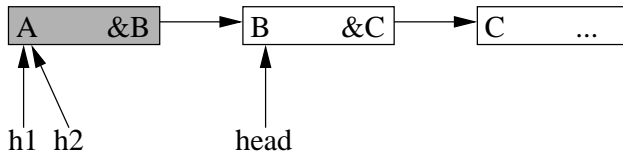
- The first thread gets one instruction into its pop, while
- The second thread completes its pop operation:

h1 = head	h2 = head	== &A
	n2 = h2->next	== &B
	CAS(head, h2, n2)	Success!

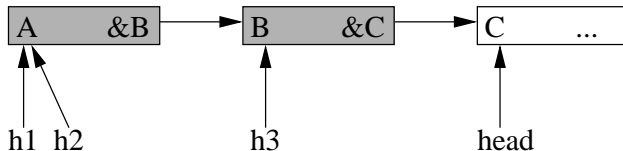
ABA Problem

- The first thread got one instruction into its pop, while
- The second thread completed its pop operation.

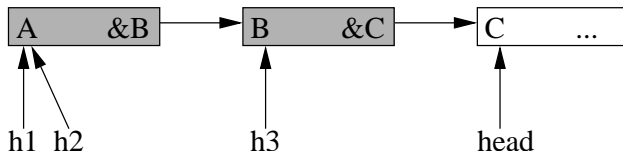
h1 = head	h2 = head	== &A
	n2 = h2->next	== &B
	CAS(head, h2, n2)	Success!

ABA Problem

- The third thread executes a pop operation.

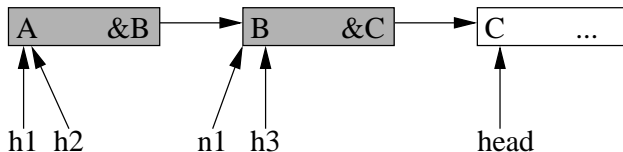
ABA Problem

- The third thread executed a pop operation.

ABA Problem

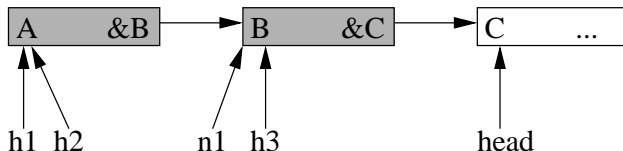
And the slower thread gets a few more instructions:

n1 = h1->next;		== &B
----------------	--	-------

ABA Problem

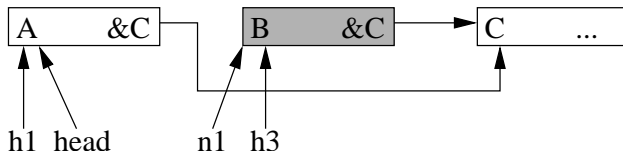
And the slower thread got a few more instructions:

n1 = h1->next;		== &B
----------------	--	-------

ABA Problem

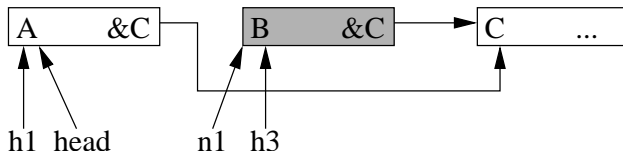
Now the second thread does its push operation...

	h2 = head;	== &C
	h2->next = h2;	A.next ← &C
	CAS(head, h2, &A)	Success!

ABA Problem

Now the second thread did its push operation...

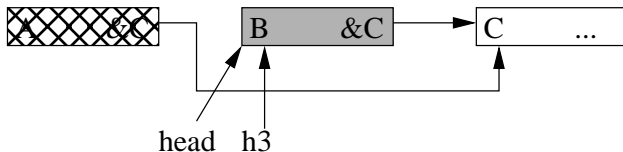
	<code>h2 = head;</code>	<code>== &C</code>
	<code>h2->next = h2;</code>	<code>A.next ← &C</code>
	<code>CAS(head, h2, &A)</code>	Success!

ABA Problem

And the slower thread finally completes its pop operation...

CAS(head, h1, n1)		Suc... hm!
-------------------	--	------------

ABA Problem



And the slower thread finally completed its pop operation...

CAS(head, h1, n1)

Suc... hm!

ABA Problem

- *B*, which was well and quite off the list, and not owned by Thread 1, is now at the head!
- Thread 1 missed its chance to be notified of having stale data.
 - All that matters is that *A* ended up back on the list head when Thread 1 was CAS-ing.
- There's relatively little that *thread 1* can do about this!
- In punishment, the data structure is now broken!
- For fun, try designing a different failure case.
 - Try getting a circular list.

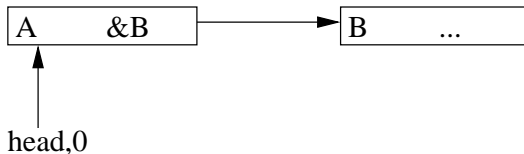
Fixing ABA

- Generation counters are a simple way to solve ABA
 - Let's replace all pointers with

```
struct versioned_ptr {  
    void * p; /* Pointer */  
    unsigned int v; /* Version */  
};
```
- This will allow a “reasonably large” number of pointer updates before we have to worry.

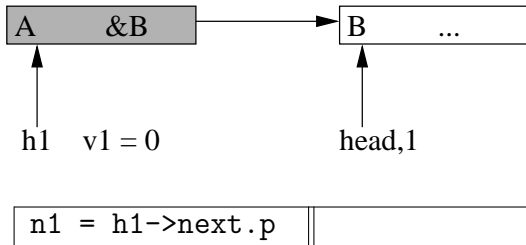
- Suppose we had a primitive which let us write things like

2nd thread pops...



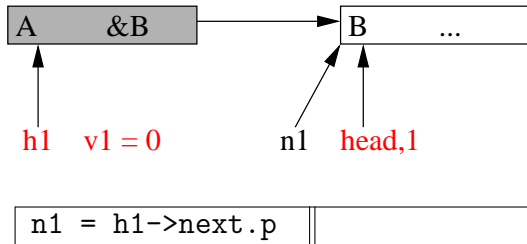
h1 = head.p v1 = head.v	h2 = head.p	== &A
	n2 = h2->next.p v2 = head.v	== &B == 0
	CAS2(head, {h2, v2}, {n2, v2+1})	Success!

1st thread reads n1



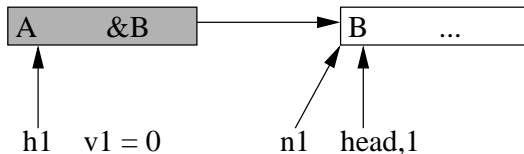
- n1 and v1 are just local variables in preparation for...
CAS2(head, {h1, v1}, {n1, v1+1})

1st thread read n1



- n1 and v1 are just local variables in preparation for...
CAS2(head, {h1, v1}, {n1, v1+1})
- So if that were to happen right now...

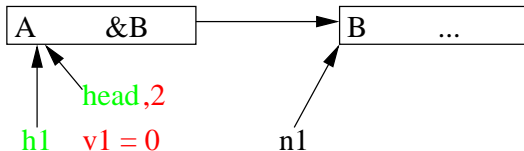
Fixing ABA
 2^{nd} thread pushes...



	h2 = head.p;
	v2 = head.v;
	A.next = h2;
	CAS2(head, {h2, v2}, {&A, v2+1})

Fixing ABA

2nd thread pushed...



	h2 = head.p; v2 = head.v;
	A.next.p = h2;
	CAS2(head, {h2, v2}, {&A, v2+1})

- CAS2(head, {h1, v1}, {n1, v1+1})
- head == h1 but v1 == 0 \neq 2. Hooray!

- Read-Copy-Update (RCU, [Wikc, McK03]; earlier papers) uses techniques from lock-free programming.
- Is used in several OSes, including Linux.
- It's a bit more complicated than the examples given here and not truly lock-free, but certainly interesting.

Read-Copy-Update Mutual Exclusion Preliminaries

- Many more readers than writers.
 - So we should make sure that the readers don't have to do much.
 - Kind of like a rwlock.
- Readers frequently can complete critical sections in bounded time (no `yield()` etc.).
 - Required property of RCU readers.
 - We'll see why this is important in a bit.
- Readers want to see a consistent data structure.
 - Not all consistency guarantees need to be kept, but, for example, we want to avoid use-after-free and the possibility of faulting.
 - But it might be the case that we let `node->next->prev != node` as readers only use these pointers to traverse.

Read-Copy-Update Mutual Exclusion Preliminaries

- Disclaimer: function names have been changed from, *e.g.*, the Linux implementation, to make the meanings more clear.
- Disclaimer 2: RCU comes in many flavors - the one here is a small toy model but works on real hardware (like Pebbles).

Read-Copy-Update Mutual Exclusion API

- Reader critical section functions.
 - `void rcu_read_lock(void);`
 - `void rcu_read_unlock(void);`
 - Note the absence of parameters (how odd!).
- Accessor functions:
 - `void * rcu_fetch(void *)`; is used to fetch a pointer from an RCU protected data structure.
 - `void * rcu_assign(void *, void *)`; is used to assign a new value to an RCU protected pointer.
- Synchronization points:
 - `void rcu_synchronize(void)`; is used once a writer is finished to signal that updates are complete.
 - Moves from “update” to “reclaim” phase.

- Suppose we have a global list, called `list`, that we want to read under RCU.
- The code for iteration looks like

```
rcu_read_lock();
list_head_t *llist = rcu_fetch(list);
list_node_t *node = rcu_fetch(llist->head);
while(node != NULL) {
    ... /* Do something reader-like */
    node = rcu_fetch(node->next);
}
rcu_read_unlock();
```


- This is kinda like a rwlock:
 - It allows an arbitrary number of readers to run against each other.
 - It prevents multiple writers from writing at once.
- It is absolutely unlike a rwlock because
 - readers and writers do not exclude each other!

Read-Copy-Update Mutual Exclusion

API: Wait, WHAT?

Readers can run alongside writers! There's no mechanism in the reader to serialize against the writer! See:

CPU 1 (reader)	CPU 2 (writer)
<code>rcu_read_lock();</code>	<code>mutex_lock(...);</code>
<code>llist = rcu_fetch(list);</code>	<code>...</code>
	<code>rcu_assign(list, new);</code>
	<code>rcu_synchronize();</code>
<code>rcu_fetch(llist->head);</code>	

Some Restrictions Apply™: Remember, only one writer, so `rcu_assign` doesn't use CAS.

Read-Copy-Update Mutual Exclusion Implementation: Key Ideas

- “All the magic is inside `rcu_synchronize()`” ...
- The deletion problem, and ABA, was a problem of not knowing when nobody had a stale reference.
- If
 - readers agree to drop *all* references in bounded time
 - AND writers can tell *when* readers have dropped references
- Then we know when it is safe to reclaim (*i.e.* `free()`) memory.
- Being safe for *reclaim* is exactly the same as being safe for *reuse*.

Read-Copy-Update Mutual Exclusion Implementation: Approximation

- Want:
 - readers agree to drop *all* references in bounded time
 - AND writers can tell when readers have dropped references
- You can imagine that there's an array of `reading[i]` values out there, with each thread having its own index...
- Each reader sets `reading[me] = 1`, reads, then sets `reading[me] = 0`.
- The writer then scans the array looking for all flags to be 0.
- When this happens, the writer knows that no readers have stale references, and is now OK to free deleted item(s).

Read-Copy-Update Mutual Exclusion
Pictures: Reader View

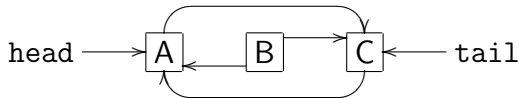
- Looking at that again, from the reader's side now. Originally



- The writer first sets it to



- And then



Read-Copy-Update Mutual Exclusion
Confessions of an Instructor






Real-world RCU once upon a time worked this way but more recent implementations are much fancier. For the really enthusiastic, see things like Linux's "Sleepable RCU" implementation [McK06].

Lessons

Write Your Own?

- It's *extremely hard* to roll your own lockfree algorithm.
- But moreover, it's *almost impossible* to debug one.
- Thus all the papers are long not because the algorithms are hard, ...
- ...but because they prove the correctness of the algorithm so they at least don't have to debug that.



-  Mikhail Fomitchev and Eric Ruppert, *Lock-free linked lists and skip lists*, PODC (2004), no. 1-58113-802-4/04/0007, 50–60,
<http://www.research.ibm.com/people/m/michael/podc-2002.pdf>.
-  Paul McKenney, *Kernel Korner - Using RCU in the Linux 2.5 Kernel*, <http://www.linuxjournal.com/article/6993>.
-  Paul McKenny, *Sleepable RCU*,
<http://lwn.net/Articles/202847/>.
-  Peter Memishian, *On locking*, July 2006,
http://blogs.sun.com/meem/entry/on_locking.
-  Maged M. Michael, *High performance dynamic lock-free hash tables and list-based sets*, SPAA (2002),



no. 1-58113-529-7/02/0008, 73–83,
[http://portal.acm.org/ft_gateway.cfm?id=564881&type=pdf
&coll=GUIDE&dl=ACM&CFID=73232202
&CFTOKEN=1170757](http://portal.acm.org/ft_gateway.cfm?id=564881&type=pdf&coll=GUIDE&dl=ACM&CFID=73232202&CFTOKEN=1170757).



———, *Safe memory reclamation for dynamic lock-free objects using atomic reads and writes*, PODC (2002), no. 1-58113-485-1/02/0007, 1–10,
<http://www.research.ibm.com/people/m/michael/podc-2002.pdf>.



———, *Hazard pointers: Safe memory reclamation for lock-free objects*, IEEECS (2004), no. TPDS-0058-0403, 1–10,
<http://www.research.ibm.com/people/m/michael/podc-2002.pdf>.



H. Sundell, *Wait-free reference counting and memory management*, International Parallel and Distributed Processing Symposium, no. 1530-2075/05, IEEE, April 2005,
<http://ieeexplore.ieee.org/iel5/9722/30685/01419843.pdf?tp=&arnumber=1419843&isnumber=30685>.



Wikipedia, *Lock-free and wait-free algorithms*,
http://en.wikipedia.org/wiki/Lock-free_and_wait-free_algorithms.



_____, *Non-blocking synchronization*,
http://en.wikipedia.org/wiki/Non-blocking_synchronization.



_____, *Read-copy-update*,
<http://en.wikipedia.org/wiki/Read-copy-update>.

Acknowledgements

- Dave Eckhardt (de0u) has seen this lecture about as often as I have, and has produced useful commentary on every release.
- Bruce Maggs (bmm) for moral support and big-picture guidance
- Jess Mink (jmink), Matt Brewer (mbrewer), and Mr. Wright (mrwright) for being victims of beta versions of this lecture.
- [Nobody on this list deserves any of the blame, but merely credit, for this lecture.]



Full fledged deletion & reclaim

- Even though we might be able to solve ABA, it still doesn't solve memory reclaim!
- Imagine that instead of being reclaimed by the list, the deleted node before had been reclaimed by something else...
 - A different list
 - A tree
 - For use as a thread control block



Full fledged deletion & reclaim

- What if we looked at ABA differently . . .
- It only matters if there is the possibility of confusion.
- In particular, might demonstrate strong interest in things that might confuse me
 - Hazard Pointers (“Safe Memory Reclamation” or just “SMR”) [Mic02b] and [Mic04]
 - Wait-free reference counters [Sun05]
- These are ways of asking “If I, Thread 189236, were to put something here, would anybody be confused?”
- This solves ABA, but really as a side effect: it lets us reclaim address space (and therefore memory) because we know nobody’s using it!



The SMR Algorithm

- Every thread comes pre-equipped with a *finite* list of “hazards”
- Memory reclaim involves scanning everybody’s hazards to see if there’s a collision
- Threads doing reclaim `yield()` (to the objecting thread) until the hazard is clear
- Difficulty
 - Show that hazards can only decrease when deletions are pending
 - Show that deletions eventually succeed (can’t deadlock on hazards)
 - Managing the list of threads’ hazards is difficult



Observation On Object Lifetime

Instance of a general problem [Mem06]:

*Things get tricky when the object must go away. [...]
Any thread looking up the object – by definition –
does not yet have the object and thus cannot hold
the object's lock during the lookup operation. [...]
Thus, whatever higher-level synchronization is used
to coordinate the threads looking up the object must
also be used as part of removing the object from
visibility.*



Miscellany

Locking vs. RCU

- Interestingly, this kind of RCU tends to decrease the number of (bus) atomic operations.
 - Uses scheduler to get per-CPU atomicity.
- RCU requires the ability to force a thread to run on every CPU or at least observe when every CPU has context switched.
 - Difficult to use RCU in userland!
- RCU, like lockfree, suffers a slowdown from cache line shuffling, but will make progress due to having at most one writer.