CS162
Operating Systems and Systems Programming

Midterm Review

March 7, 2011
Ion Stoica
http://inst.eecs.berkeley.edu/~cs162

Synchronization, Critical section

Definitions

- Synchronization: using atomic operations to ensure cooperation between threads
- Mutual Exclusion: ensuring that only one thread does a particular thing at a time
  - One thread *excludes* the other while doing its task
- Critical Section: piece of code that only one thread can execute at once
  - Critical section is the result of mutual exclusion
  - Critical section and mutual exclusion are two ways of describing the same thing.

Locks: using interrupts

- Key idea: maintain a lock variable and impose mutual exclusion only during operations on that variable

```
int value = FREE;
```

```
Acquire() {
  disable interrupts;
  if (value == BUSY) {
    put thread on wait queue;
    Go to sleep();
    // Enable interrupts?
  } else {
    value = BUSY;
  }
  enable interrupts;
}
```

```
Release() {
  disable interrupts;
  if (anyone on wait queue) {
    take thread off wait queue
    Place on ready queue;
  } else {
    value = FREE;
  }
  enable interrupts;
}
```

Better locks: using test&set

```
• test&set (&address) {   /* most architectures */
      result = M[address];
      M[address] = 1;
      return result;
  }
```

```
int guard = 0;
int value = FREE;
```

```
Acquire() {
   // Short busy-wait time
   while (test&set(guard));
   if (value == BUSY) {
      put thread on wait queue;
      go to sleep() & guard = 0;
   } else {
      value = BUSY;
      guard = 0;
   }
}
```

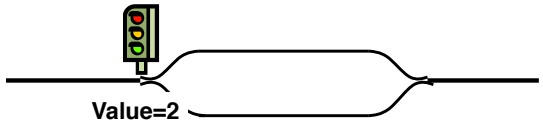
```
Release() {
   // Short busy-wait time
   while (test&set(guard));
   if anyone on wait queue {
      take thread off wait queue
      Place on ready queue;
   } else {
      value = FREE;
   }
   guard = 0;
```


Semaphores

- Semaphores are a kind of generalized lock
  - First defined by Dijkstra in late 60s
  - Main synchronization primitive used in original UNIX
- Definition: a Semaphore has a non-negative integer value and supports the following two operations:
  - P(): an atomic operation that waits for semaphore to become positive, then decrements it by 1
    - » Think of this as the wait() operation
  - V(): an atomic operation that increments the semaphore by 1, waking up a waiting P, if any
    - » This of this as the signal() operation
  - Note that P() stands for "*proberen*" (to test) and V() stands for "*verhogen*" (to increment) in Dutch

Semaphores Like Integers Except

- Semaphores are like integers, except
  - No negative values
  - Only operations allowed are P and V – can't read or write value, except to set it initially
  - Operations must be atomic
    - » Two P's together can't decrement value below zero
    - » Similarly, thread going to sleep in P won't miss wakeup from V – even if they both happen at same time
- Semaphore from railway analogy
  - Here is a semaphore initialized to 2 for resource control:



Condition Variables

- Condition Variable: a queue of threads waiting for something *inside* a critical section
  - Key idea: allow sleeping inside critical section by atomically releasing lock at time we go to sleep
  - Contrast to semaphores: Can't wait inside critical section
- Operations:
  - `Wait(&lock)`: Atomically release lock and go to sleep. Re-acquire lock later, before returning.
  - `Signal()`: Wake up one waiter, if any
  - `Broadcast()`: Wake up all waiters
- Rule: Must hold lock when doing condition variable ops!

Complete Monitor Example (with condition variable)

- Here is an (infinite) synchronized queue

```
Lock lock;
Condition dataready;
Queue queue;

AddToQueue(item) {
   lock.Acquire();           // Get Lock
   queue.enqueue(item);      // Add item
   dataready.signal();       // Signal any waiters
   lock.Release();           // Release Lock
}

RemoveFromQueue() {
   lock.Acquire();           // Get Lock
   while (queue.isEmpty()) {
      dataready.wait(&lock); // If nothing, sleep
   }
   item = queue.dequeue();   // Get next item
   lock.Release();           // Release Lock
   return(item);
}
```


Mesa vs. Hoare monitors

- Need to be careful about precise definition of signal and wait. Consider a piece of our dequeue code:

```
while (queue.isEmpty()) {
   dataready.wait(&lock); // If nothing, sleep
}
item = queue.dequeue();// Get next item
```

  - Why didn't we do this?

```
if (queue.isEmpty()) {
   dataready.wait(&lock); // If nothing, sleep
}
item = queue.dequeue();// Get next item
```

- Answer: depends on the type of scheduling
  - Hoare-style (most textbooks):
    - » Signaler gives lock, CPU to waiter; waiter runs immediately
    - » Waiter gives up lock, processor back to signaler when it exits critical section or if it waits again
  - Mesa-style (most real operating systems):
    - » Signaler keeps lock and processor
    - » Waiter placed on ready queue with no special priority
    - » Practically, need to check condition again after wait

Read/Writer Revisited

```
Reader() {
    // check into system
    lock.Acquire();
    while ((AW + WW) > 0) {
       WR++;
       okToRead.wait(&lock);
       WR--;
    }
    AR++;
    lock.release();

    // read-only
    AccessDbase

    // check out
    lock.Acquire
    AR--;
    if (AR == 0 && WW > 0)
       okToWrite.signal();
    lock.Release();
 }
```

```
Writer() {
    // check into system
    lock.Acquire();
    while ((AW + AR) > 0) {
       WW++;
       okToWrite.wait(&lock);
       WW--;
    }
    AW++;
    lock.release();

    // read/write access
    AccessDbase(ReadWrite);

    // check out of system
    lock.Acquire();
    AW--;
    if (WW > 0){
       okToWrite.signal();
    } else if (WR > 0) {
       okToRead.broadcast();
    }
    lock.Release();
 }
```

What if we remove this line?

Read/Writer Revisited

```
Reader() {
    // check into system
    lock.Acquire();
    while ((AW + WW) > 0) {
       WR++;
       okToRead.wait(&lock);
       WR--;
    }
    AR++;
    lock.release();

    // read-only
    AccessDbase

    // check out
    lock.Acquire
    AR--;
    if (AR == 0 && WW > 0)
       okToWrite.broadcast();
    lock.Release();
 }
```

```
Writer() {
    // check into system
    lock.Acquire();
    while ((AW + AR) > 0) {
       WW++;
       okToWrite.wait(&lock);
       WW--;
    }
    AW++;
    lock.release();

    // read/write access
    AccessDbase(ReadWrite);

    // check out of system
    lock.Acquire();
    AW--;
    if (WW > 0){
       okToWrite.signal();
    } else if (WR > 0) {
       okToRead.broadcast();
    }
    lock.Release();
 }
```

What if we turn signal to broadcast?

Read/Writer Revisited

```
Reader() {
    // check into system
    lock.Acquire();
    while ((AW + WW) > 0) {
       WR++;
       okContinue.wait(&lock);
       WR--;
    }
    AR++;
    lock.release();

    // read-only access
    AccessDbase(ReadOnly);

    // check out of system
    lock.Acquire();
    AR--;
    if (AR == 0 && WW > 0)
       okContinue.signal();
    lock.Release();
  }
```

```
Writer() {
    // check into system
    lock.Acquire();
    while ((AW + AR) > 0) {
       WW++;
       okContinue.wait(&lock);
       WW--;
    }
    AW++;
    lock.release();

    // read/write access
    AccessDbase(ReadWrite);

    // check out of system
    lock.Acquire();
    AW--;
    if (WW > 0){
       okToWrite.signal();
    } else if (WR > 0) {
       okContinue.broadcast();
    }
    lock.Release();
  }
```

What if we turn okToWrite and okToRead into okContinue?

Read/Writer Revisited

```
Reader() {
    // check into system
    lock.Acquire();
    while ((AW + WW) > 0) {
       WR++;
       okContinue.wait(&lock);
       WR--;
    }
    AR++;
    lock.release();

    // read-only access
    AccessDbase(ReadOnly);

    // check out of system
    lock.Acquire();
    AR--;
    if (AR == 0 && WW > 0)
       okContinue.signal();
    lock.Release();
  }
```

```
Writer() {
    // check into system
    lock.Acquire();
    while ((AW + AR) > 0) {
       WW++;
       okContinue.wait(&lock);
       WW--;
    }
    AW++;
    lock.release();

    // read/write access
    AccessDbase(ReadWrite);

    // check out of system
    lock.Acquire();
    AW--;
    if (WW > 0){
       okToWrite.signal();
    } else if (WR > 0) {
       okContinue.broadcast();
    }
    lock.Release();
  }
```

- R1 arrives
- W1, R2 arrive while R1 reads
- R1 signals R2

Read/Writer Revisited

```
Reader() {
    // check into system
    lock.Acquire();
    while ((AW + WW) > 0) {
       WR++;
       okContinue.wait(&lock);
       WR--;
    }
    AR++;
    lock.release();

    // read-only access
    AccessDbase(ReadOnly);

    // check out of system
    lock.Acquire();
    AR--;
    if (AR == 0 && WW > 0)
       okContinue.broadcast();
    lock.Release();
  }
```

```
Writer() {
    // check into system
    lock.Acquire();
    while ((AW + AR) > 0) {
       WW++;
       okContinue.wait(&lock);
       WW--;
    }
    AW++;
    lock.release();

    // read/write access
    AccessDbase(ReadWrite);

    // check out of system
    lock.Acquire();
    AW--;
    if (WW > 0){
       okToWrite.signal();
    } else if (WR > 0) {
       okContinue.broadcast();
    }
    lock.Release();
  }
```

Need to change to broadcast!
Why?

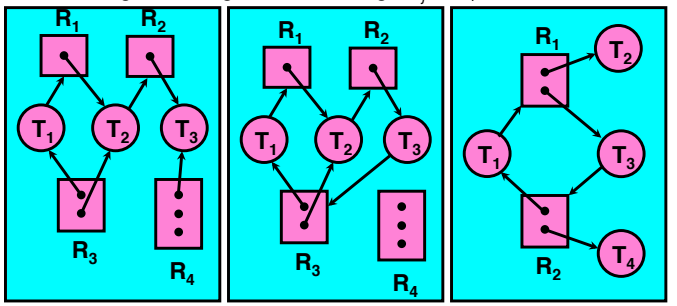
Deadlock

Four requirements for Deadlock

- Mutual exclusion
  - Only one thread at a time can use a resource.
- Hold and wait
  - Thread holding at least one resource is waiting to acquire additional resources held by other threads
- No preemption
  - Resources are released only voluntarily by the thread holding the resource, after thread is finished with it
- Circular wait
  - There exists a set $\{T_1, \ldots, T_n\}$ of waiting threads
    - » T_1 is waiting for a resource that is held by T_2
    - » T_2 is waiting for a resource that is held by T_3
    - » …
    - » T_n is waiting for a resource that is held by T_1

Resource Allocation Graph Examples

- Recall:
  - request edge – directed edge $T_1 \rightarrow R_j$
  - assignment edge – directed edge $R_j \rightarrow T_i$



Simple Resource Allocation Graph | **Allocation Graph With Deadlock** | **Allocation Graph With Cycle, but No Deadlock**

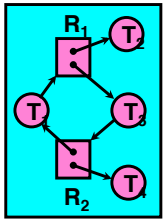
Deadlock Detection Algorithm

- Only one of each type of resource ⇒ look for loops
- More General Deadlock Detection Algorithm
  - Let [X] represent an m-ary vector of non-negative integers (quantities of resources of each type):

    [FreeResources]: Current free resources each type
    [Request$_X$]: Current requests from thread X
    [Alloc$_X$]: Current resources held by thread X

  - See if tasks can eventually terminate on their own

```
[Avail] = [FreeResources]
Add all nodes to UNFINISHED
do {
   done = true
   Foreach node in UNFINISHED {
      if ([Request_node] <= [Avail]) {
         remove node from UNFINISHED
         [Avail] = [Avail] + [Alloc_node]
         done = false
      }
   }
} until(done)
```



  - Nodes left in UNFINISHED ⇒ deadlocked

Banker's Algorithm for Preventing Deadlock

- Toward right idea:
  - State maximum resource needs in advance
  - Allow particular thread to proceed if:
    (available resources - #requested) ≥ max remaining that might be needed by any thread

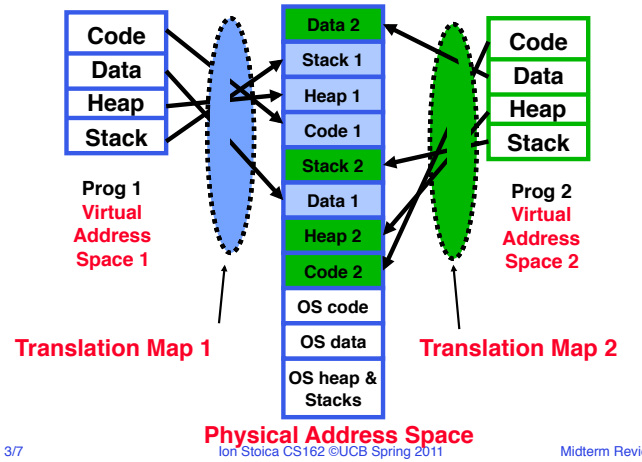
- Banker's algorithm (less conservative):
  - Allocate resources dynamically
    - » Evaluate each request and grant if some ordering of threads is still deadlock free afterward
    - » Technique: pretend each request is granted, then run deadlock detection algorithm, substituting ([Max$_{node}$]-[Alloc$_{node}$] ≤ [Avail]) for ([Request$_{node}$] ≤ [Avail]) Grant request if result is deadlock free (conservative!)
    - » Keeps system in a "SAFE" state, i.e. there exists a sequence $\{T_1, T_2, \ldots T_n\}$ with T_1 requesting all remaining resources, finishing, then T_2 requesting all remaining resources, etc..
  - Algorithm allows the sum of maximum resource needs of all current threads to be greater than total resources

Memory Multiplexing, Address Translation

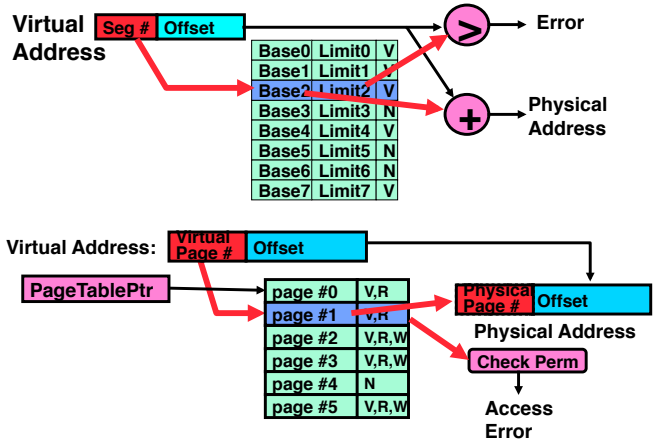
Important Aspects of Memory Multiplexing

- Controlled overlap:
  - Processes should not collide in physical memory
  - Conversely, would like the ability to share memory when desired (for communication)
- Protection:
  - Prevent access to private memory of other processes
    - » Different pages of memory can be given special behavior (Read Only, Invisible to user programs, etc).
    - » Kernel data protected from User programs
    - » Programs protected from themselves
- Translation:
  - Ability to translate accesses from one address space (virtual) to a different one (physical)
  - When translation exists, processor uses virtual addresses, physical memory uses physical addresses
  - Side effects:
    - » Can be used to avoid overlap
    - » Can be used to give uniform view of memory to programs

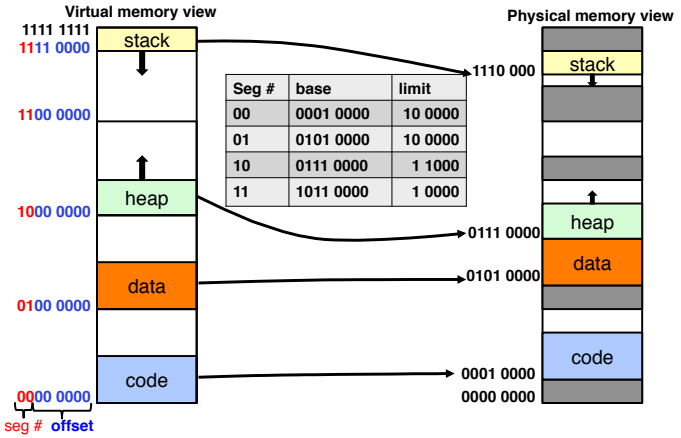
Why Address Translation?



Addr. Translation: Segmentation vs. Paging

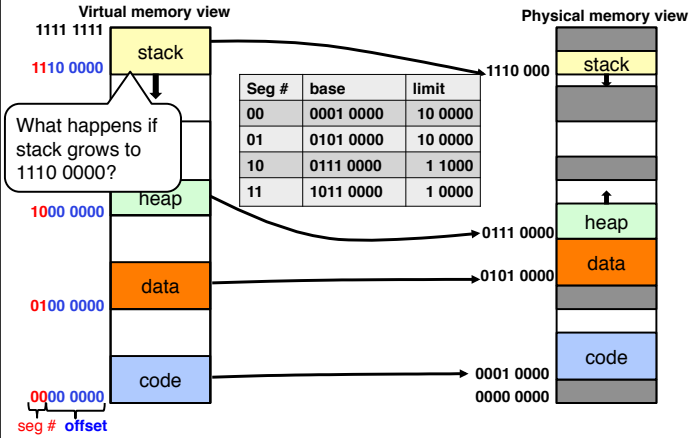


Review: Address Segmentation



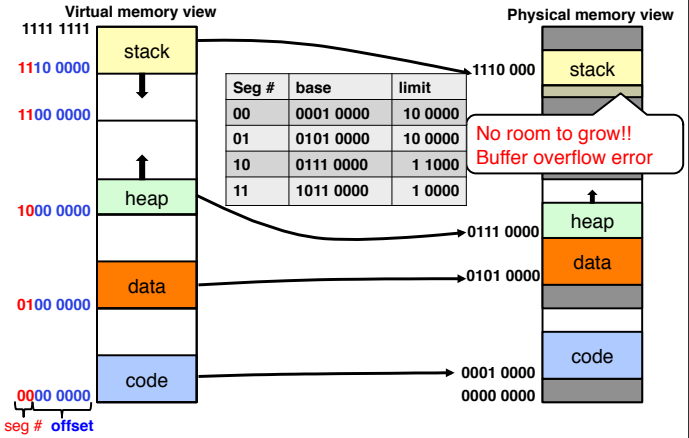
Seg #	base	limit
00	0001 0000	10 0000
01	0101 0000	10 0000
10	0111 0000	1 1000
11	1011 0000	1 0000

Review: Address Segmentation



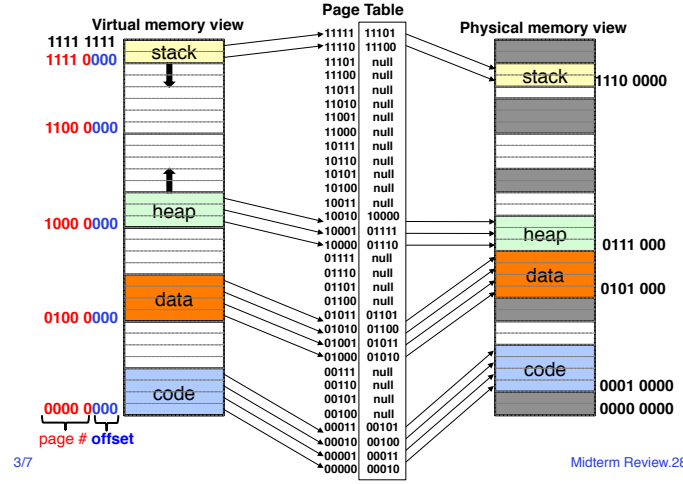
Seg #	base	limit
00	0001 0000	10 0000
01	0101 0000	10 0000
10	0111 0000	1 1000
11	1011 0000	1 0000

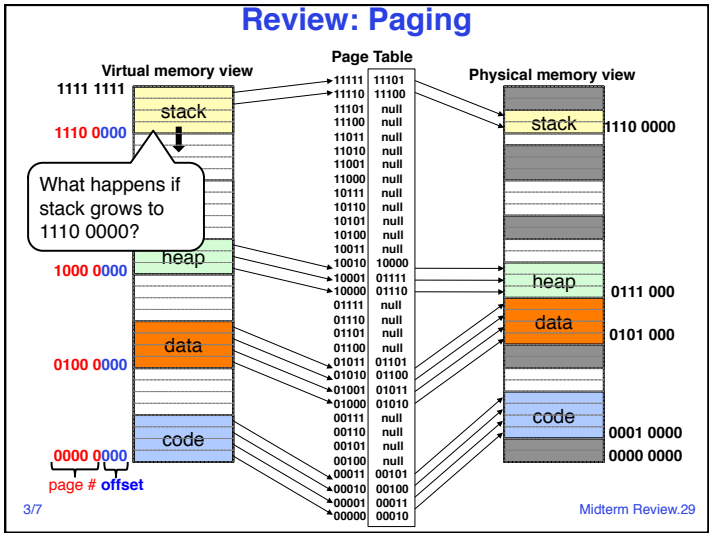
Review: Address Segmentation

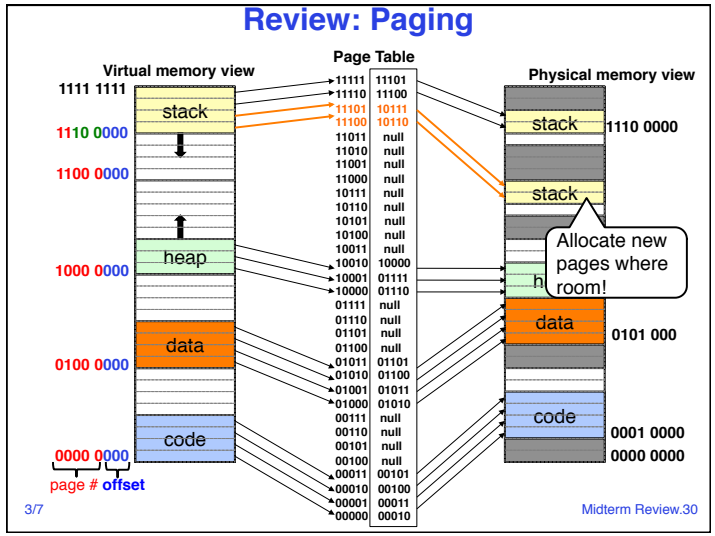


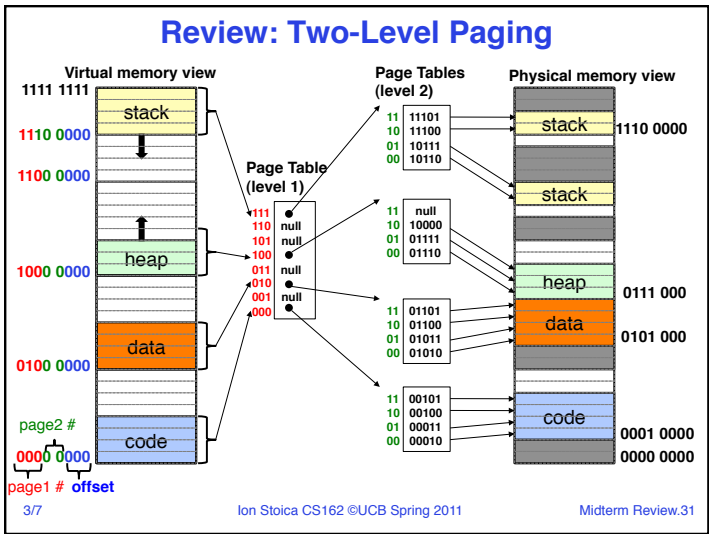
Seg #	base	limit
00	0001 0000	10 0000
01	0101 0000	10 0000
10	0111 0000	1 1000
11	1011 0000	1 0000

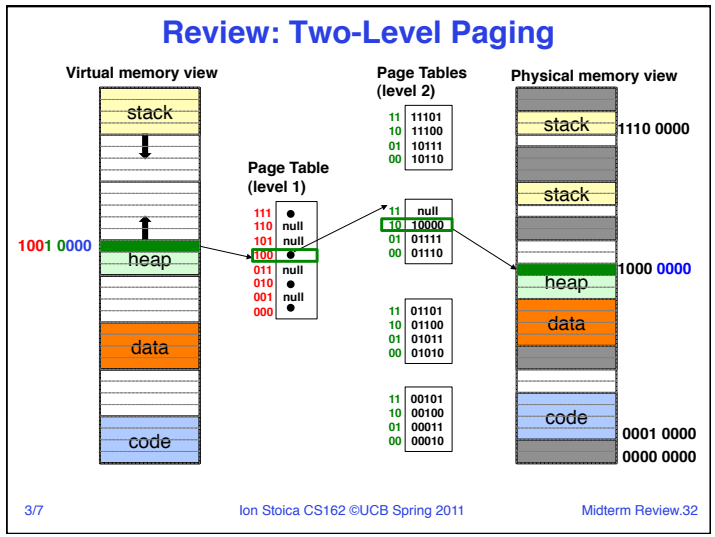
Review: Paging



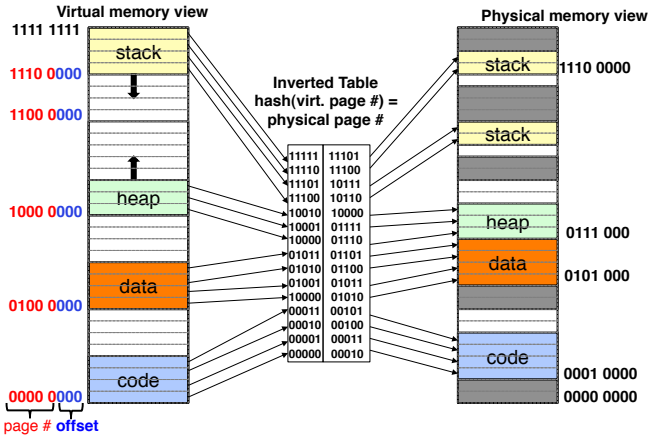

Review: Paging
Virtual memory view
Page Table
Physical memory view
1111 1111
1110 0000
1000 0000
0100 0000
0000 0000
stack
heap
data
code
What happens if stack grows to 1110 0000?
page # offset
1110 0000
0111 000
0101 000
0001 0000
0000 0000
3/7
Midterm Review.29


Review: Paging
Virtual memory view
Page Table
Physical memory view
1111 1111
1110 0000
1100 0000
1000 0000
0100 0000
0000 0000
stack
heap
data
code
Allocate new pages where room!
page # offset
1110 0000
0101 000
0001 0000
0000 0000
3/7
Midterm Review.30


Review: Two-Level Paging
Virtual memory view
Page Tables (level 2)
Physical memory view
Page Table (level 1)
1111 1111
1110 0000
1100 0000
1000 0000
0100 0000
stack
heap
data
code
page2 #
page1 # offset
1110 0000
0111 000
0101 000
0001 0000
0000 0000
3/7
Ion Stoica CS162 ©UCB Spring 2011
Midterm Review.31


Review: Two-Level Paging
Virtual memory view
Page Tables (level 2)
Physical memory view
Page Table (level 1)
1001 0000
stack
heap
data
code
1110 0000
1000 0000
0001 0000
0000 0000
3/7
Ion Stoica CS162 ©UCB Spring 2011
Midterm Review.32

Review: Inverted Table



Address Translation Comparison

	Advantages	Disadvantages
Segmentation	Fast context switching: Segment mapping maintained by CPU	External fragmentation
Paging (single-level page)	No external fragmentation	•Large size: Table size ~ virtual memory •Internal fragmentation
Paged segmentation	•No external fragmentation •Table size ~ memory used by program	•Multiple memory references per page access •Internal fragmentation
Two-level pages		
Inverted Table		Hash function more complex

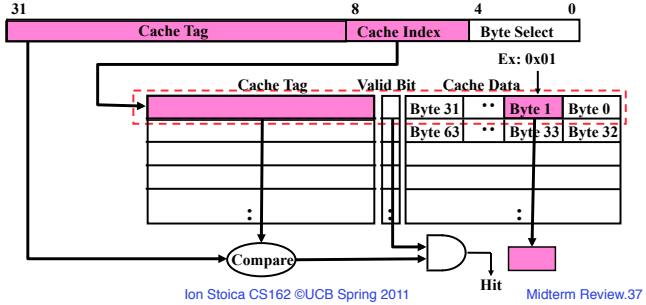
Caches, TLBs

Review: Sources of Cache Misses

- Compulsory (cold start): first reference to a block
  - "Cold" fact of life: not a whole lot you can do about it
  - Note: When running "billions" of instruction, Compulsory Misses are insignificant
- Capacity:
  - Cache cannot contain all blocks access by the program
  - Solution: increase cache size
- Conflict (collision):
  - Multiple memory locations mapped to same cache location
  - Solutions: increase cache size, or increase associativity
- Two others:
  - Coherence (Invalidation): other process (e.g., I/O) updates memory
  - Policy: Due to non-optimal replacement policy

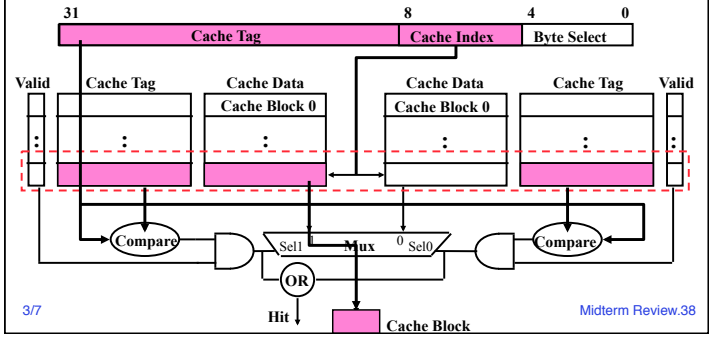
Direct Mapped Cache

- Cache index selects a cache block
- "Byte select" selects byte within cache block
  - Example: Block Size=32B blocks
- Cache tag fully identifies the cached data
- Data with same "cache tag" shares the same cache entry
  - Conflict misses



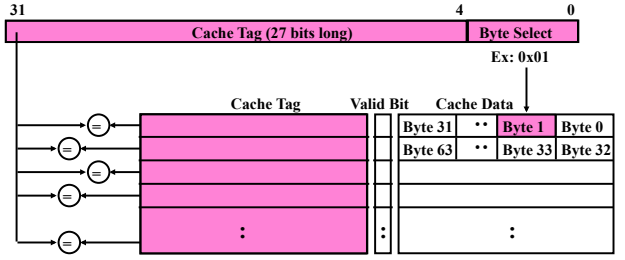
Set Associative Cache

- N-way set associative: N entries per Cache Index
  - N direct mapped caches operates in parallel
- Example: Two-way set associative cache
  - Two tags in the set are compared to input in parallel
  - Data is selected based on the tag result



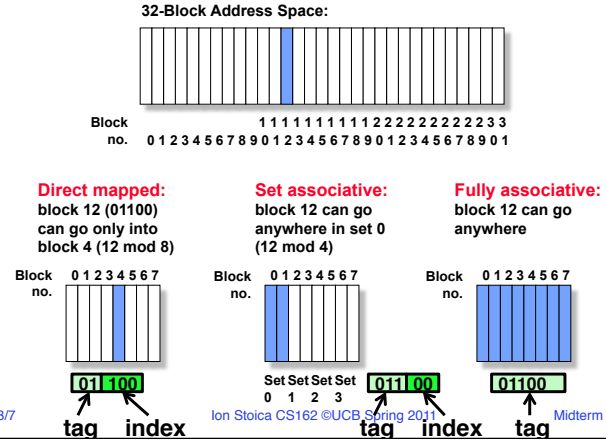
Fully Associative Cache

- Fully Associative: Every block can hold any line
  - Address does not include a cache index
  - Compare Cache Tags of all Cache Entries in Parallel
- Example: Block Size=32B blocks
  - We need N 27-bit comparators
  - Still have byte select to choose from within block



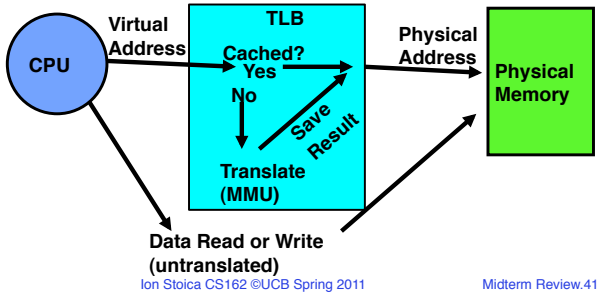
Where does a Block Get Placed in a Cache?

- Example: Block 12 placed in 8 block cache



Review: Caching Applied to Address Translation

- Problem: address translation expensive (especially multi-level)
- Solution: cache address translation (TLB)
  - Instruction accesses spend a lot of time on the same page (since accesses sequential)
  - Stack accesses have definite locality of reference
  - Data accesses have less page locality, but still some…

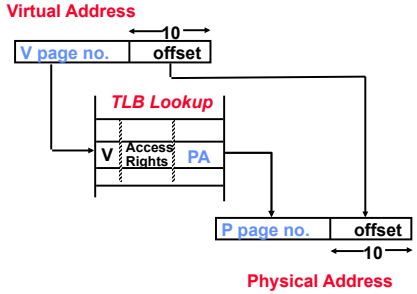


TLB organization

- How big does TLB actually have to be?
  - Usually small: 128-512 entries
  - Not very big, can support higher associativity
- TLB usually organized as fully-associative cache
  - Lookup is by Virtual Address
  - Returns Physical Address
- What happens when fully-associative is too slow?
  - Put a small (4-16 entry) direct-mapped cache in front
  - Called a "TLB Slice"
- When does TLB lookup occur?
  - Before cache lookup?
  - In parallel with cache lookup?

Reducing translation time further

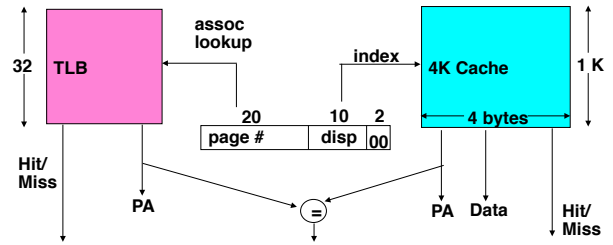
- As described, TLB lookup is in serial with cache lookup:



- Machines with TLBs go one step further: they overlap TLB lookup with cache access.
  - Works because offset available early

Overlapping TLB & Cache Access

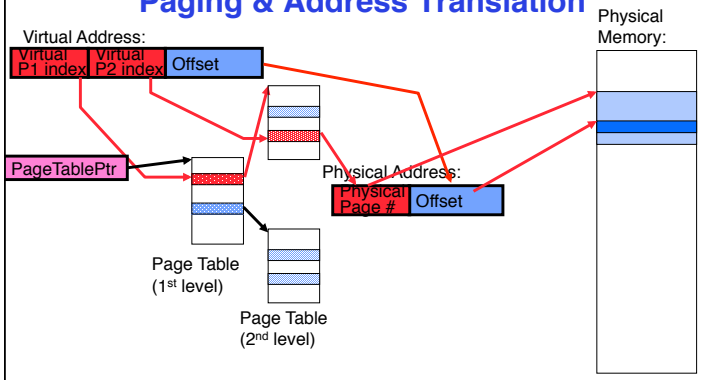
- Here is how this might work with a 4K cache:



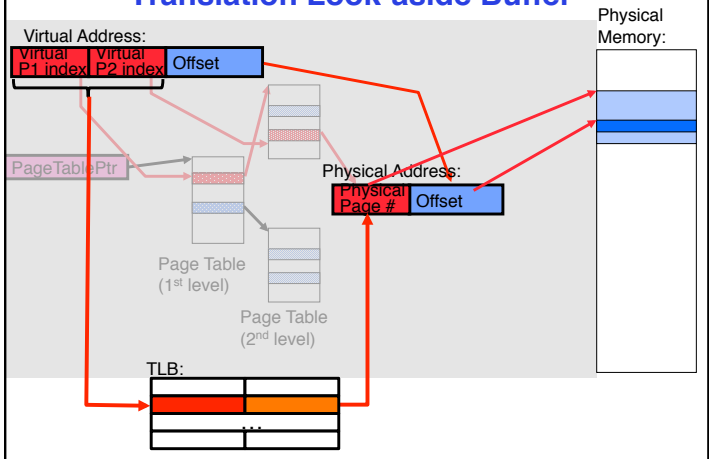
- What if cache size is increased to 8KB?
  - Overlap not complete
  - Need to do something else. See CS152/252
- Another option: Virtual Caches
  - Tags in cache are virtual addresses
  - Translation only happens on cache misses

Putting Everything Together

Paging & Address Translation



Translation Look-aside Buffer



Caching