CS 350: Operating Systems

Charles Shen

Fall 2016, University of Waterloo

Notes written from Gregor Richards's lectures.

Feel free to email feedback to me at ccshen902@gmail.com.

Contents

1 Introduction

There are three views of an operating system:

1. **Application View** (Section 1.1): what service does it provide?
2. **System View** (Section 1.2): what problems does it solve?
3. **Implementation View** (Section 1.3): how is it built?

An operating system is part cop, part facilitator.

kernel: The operating system kernel is the part of the operating system that responds to system calls, interrupts and exception.

operating system (OS): The operating system as a whole includes the kernel, and may include other related programs that provide services for application such as utility programs, command interpreters, and programming libraries.

1.1 Application View of an Operating System

The OS provides an execution environment for running programs.

- The execution environment provides a program with the processor time and memory space that it needs to run.
- The execution environment provides interfaces through which a program can use networks, storage, I/O devices, and other system hardware components.
  Interfaces provide a simplified, abstract view of hardware to application programs.
- The execution environment isolates running programs from one another and prevents undesirable interactions among them.

1.2 System View of an Operating System

The OS manages the hardware resources of a computer system.

- Resources include processors, memory, disks and other storage devices, network interfaces, I/O devices such as keyboards, mice and monitors, and so on.
- The operating system allocates resources among running programs.
  It controls the sharing of resources among programs.
- The OS itself also uses resources, which it must share with application programs.

1.3 Implementation View of an Operating System

The OS is a concurrent, real-time program.

- Concurrency arises naturally in an OS when it supports concurrent applications, and because it must interact directly with the hardware.
- Hardware interactions also impose timing constraints.

1.4 Operating System Abstractions

The execution environment provided by the OS includes a variety of abstract entities that can be manipulated by a running program.
Examples:

- **files and file systems**: abstract view of secondary storage
- **address spaces**: abstract view of primary memory
- **processes, threads**: abstract view of program execution
- **sockets, pipes**: abstract view of network or other message channels

2 Threads and Concurrency

Threads provide a way for programmers to express *concurrency* in a program.
A normal *sequential program* consists of a single thread of execution.
In threaded concurrent programs, there are multiple threads of executions that are all occurring at the same time.

2.1 OS/161's Thread Interface

Create a new thread:

```
int thread_fork(
  const char *name,             // name of new thread
  struct proc *proc,            // thread's process
  void (*func)                  // new thread's function
    (void *, unsigned long),
  void *datat1,                 // function's first param
  unsigned long data2           // function's second param
);
```

Terminating the calling thread:

```
void thread_exit(void);
```

Voluntarily yield execution:

```
void thread_yield(void);
```

2.2 Why Threads?

Reason 1: parallelism exposed by threads enables parallel execution if the underlying hardware supports it. Programs can run faster!

Reason 2: parallelism exposed by threads enable better processor utilization.
If one thread has to *block*, another may be able to run.

Concurrent Program Execution (Two Threads)
Conceptually, each thread executes sequentially using its private register contents and stack.

2.3 Some Reviews

The Fetch/Execute Cycle

1. fetch instruction PC points to
2. decode and execute instruction
3. advance PC

Table 1: MIPS Registers

num	name	use	num	name	use
0	z0	always zero	24-25	t8-t9	temps (caller-save)
1	at	assembler reserved	26-27	k0-k1	kernel temps
2	v0	return val/syscall #	28	gp	global pointer
3	v1	return value	29	sp	stack pointer
4-7	a0-a3	subroutine args	30	s8/fp	frame ptr (callee-save)
8-15	t0-t7	temps (caller-save)	31	ra	return addr (for jal)
16-23	s0-s7	saved (calle-save)			

2.4 Implementing Concurrent Threads

Option 1: multiple processors, multiple cores, hardware multithreading per core

- P processors, C cores per processor, M multithreading degree per core $\Rightarrow P \cdot C \cdot M$ threads can execute simultaneously
- separate register set for each running thread, to hold its execution context

Option 2: *timesharing*

- multiple threads take turns on the same hardware
- rapidly switch from thread to thread so that all make progress

In practice, both techniques can be combined!

2.5 Timesharing and Context Switches

When timesharing, the switch from one thread to another is called a *context switch*.

What happens during a context switch:

1. decide which thread will run next (scheduling)
2. save register contents of current thread
3. load register contents of next thread

Thread context must be saved/restored carefully, since thread execution continuously changes the context!

Context Switch on the MIPS

```
switchframe_switch:
  /* a0: address of switchframe pointer of old thread. */
  /* a1: address of switchframe pointer of new thread. */
  /* Allocate stack space for saving 10 registers. 10*4 = 40 */
  addi sp, sp, -40
```

```
sw   ra, 36(sp)  /* Save the registers */
sw   gp, 32(sp)
sw   s8, 28(sp)
sw   s6, 24(sp)
sw   s5, 20(sp)
sw   s4, 16(sp)
sw   s3, 12(sp)
sw   s2, 8(sp)
sw   s1, 4(sp)
sw   s0, 0(sp)

/* Store the old stack pointer in the old thread */
sw   sp, 0(a0)

/* Get the new stack pointer from the new thread */
lw   sp, 0(a1)
nop             /* delay slot for load */

lw   s0, 0(sp)  /* Now, restore the registers */
lw   s1, 4(sp)
lw   s2, 8(sp)
lw   s3, 12(sp)
lw   s4, 16(sp)
lw   s5, 20(sp)
lw   s6, 24(sp)
lw   s8, 28(sp)
lw   gp, 32(sp)
lw   ra, 36(sp)
nop             /* delay slot for load */

/* and return. */
j ra
addi sp, sp, 40 /* in delay slot */
.end switchframe_switch
```

2.6 What Causes a Context Switch?

The running thread calls `thread_yield`, running thread *voluntarily* allows other threads to run.
So we have the following stack (in growth order) after voluntary context switch:

- thread_yield() stack frame
- thread_switch() stack frame

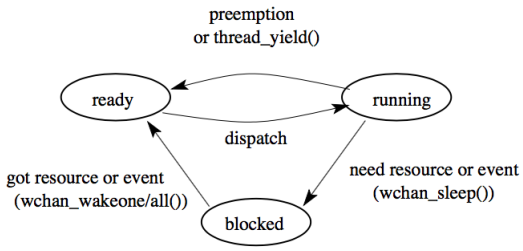
- saved thread context (switchframe)

The running thread calls `thread_exit`, running thread is terminated.

The running thread *blocks*, via a call to `wchan_sleep`.

The running thread is *preempted*, running thread *involuntarily* stops running.
So we have the following stack (in growth order) after preemption:

- trap frame
- interrupt handling stack frame(s)
- thread_yield() stack frame
- thread_switch() stack frame
- saved thread context (switchframe)



running: currently executing

ready: ready to execute

blocked: waiting for something, so not ready to execute.

Figure 1: Thread States

2.7 Review on Interrupts

An interrupt is an event that occurs during execution of a program.
Interrupts are caused by system devices (hardware), e.g., a timer, a disk controller, a network interface.
When an interrupt occurs, the hardware automatically transfer control to a fixed location in memory. At that memory location, the thread library places a procedure called an *interrupt handler*.
The interrupt handler normally:

1. create a *trap frame* to record thread context at the time of the interrupt

2. determines which device caused the interrupt and performs device-specific processing

3. restores the saved thread context from the trap frame and resumes executions of the thread

2.8 Preemption

Without preemption, a running thread could potentially run forever, without yielding, blocking, or exiting.
Preemption means forcing a running thread to stop running, so that another thread can have a chance.
To implement preemption, the thread library must have a means of "getting control" (causing thread library code to be executed) even though the running thread has not called a thread library function.
This is normally accomplished using *interrupts*.

2.8.1 Preemptive Scheduling

A preemptive scheduler imposes a limit, called the *scheduling quantum* on how long a thread can run before being preempted.
The quantum is an *upper bound* on the amount of time that a thread can run. It may block or yield before its quantum has expired.
Periodic timer interrupts allow running time to be tracked.
If a thread has run too long, the timer interrupt handler preempts the thread by calling `thread_yield`.
The preempted thread changes state from running to ready, and it is placed on the *ready queue*.

OS/161 threads use *preemptive round-robin scheduling*.

2.8.2 Two-Thread Example

Thread 1 is running, thread two had previously yielded voluntarily.
Thread 1: program stack frame(s).
Thread 2: program stack frame(s), thread_yield, thread_switch, switch frame.

A time interrupt occurs. Interrupt handler runs.
Thread 1: program stack frame(s), trap frame, interrupt handler.
Thread 2: program stack frame(s), thread_yield, thread_switch, switch frame.

Interrupt handler decides THread 1 quantum has expired.
Thread 1: program stack frame(s), trap frame, interrupt handler, thread_yield.
Thread 2: program stack frame(s), thread_yield, thread_switch, switch frame.

Scheduler chooses Thread 2 to run. Context switch.
Thread 1: program stack frame(s), trap frame, interrupt handler, thread_yield, thread_switch,

switch frame.
Thread 2: program stack frame(s), thread_yield, thread_switch, switch frame.

Thread 2 context is restored.
Thread 1: program stack frame(s), trap frame, interrupt handler, thread_yield, thread_switch, switch frame.
Thread 2: program stack frame(s), thread_yield.

`thread_yield` finishes, Thread 2 program resumes.
Thread 1: program stack frame(s), trap frame, interrupt handler, thread_yield, thread_switch, switch frame.
Thread 2: program stack frame(s).

Later, Thread 2 yields again. Scheduler chooses Thread 1.
Thread 1: program stack frame(s), trap frame, interrupt handler, thread_yield, thread_switch, switch frame.
Thread 2: program stack frame(s), thread_yield, thread_switch, switch frame.

Thread 1 context is restored, interrupt handler resumes.
Thread 1: program stack frame(s), trap frame, interrupt handler.
Thread 2: program stack frame(s), thread_yield, thread_switch, switch frame.

Interrupt handler restores state from trap frame and returns.
Thread 1: program stack frame(s).
Thread 2: program stack frame(s), thread_yield, thread_switch, switch frame.

3 Synchronization

3.1 Thread Synchronization

All threads in a concurrent program *share access* to the program's global variables and the heap.
The part of a concurrent program in which a shared object is accessed is called a *critical section*.

`volatile` keyword: Without it, the compiler could optimize the code on the variable.
`volatile` forces the compiler to load and store the value on every use. Otherwise, we may have a variable that is loaded before a loop and stored after the loop terminates which may not be what we want/expect.

3.2 Mutual Exclusion

To prevent race conditions, we can enforce *mutual exclusion* on critical sections in the code.

A *race condition* is an undesirable situation that occurs when a device or system attempts to perform two or more operations at the same time, but because of the nature of the device or system, the operations must be done in the proper sequence to be done correctly.

3.2.1 Enforcing Mutual Exclusions With Locks

Acquire/Release must ensure that only one thread at a time can hold the lock, even if both attempt to Acquire at the same time.
If a thread cannot Acquire the lock immediately, it must wait until the lock is available.

3.3 Hardware-Specific Synchronization Instructions

Used to implement synchronization primitives like locks.
Provide a way to *test and set* a lock in a single *atomic* (indivisible) operation.

Example. x86 `xchg` instruction:

```
xchg  src, addr
```

where `src` is typically a register, and `addr` is a memory address.
Value in register `src` is written to memory at address `addr`, and the old value at `addr` is placed into `src`.

Logical behaviour of `xchg` can be thought of as an *atomic* function that behaves like this:

```
Xchg(value,addr) {
  old = *addr;
  *addr = value;
  return(old);
}
```

3.3.1 Lock Acquire and Release with Xchg

```
Acquire(bool *lock) {
  while (Xchg(true,lock) == true) ;
}
Release(book *lock) {
  *lock = false; /* give up the lock */
}
```

If `Xchg` returns `true`, the lock was already set, and we must continue to loop.
If `Xchg` returns `false`, then the lock was free, and we have now acquired it.

This construct is known as a *spin lock*, since a thread busy-waits (loops) in Acquire until the lock is free!

3.3.2 Other Synchronization Instructions

SPARC `cas` instruction

```
cas   addr, R1, R2
```

if value at `addr` matches value in `R1` then swap contents of `addr` and `R2`.

Compare-And-Swap

```
CompareAndSwap(addr, expectedval, newval)
  old = *addr;            // get old value at addr
  if (old == expectedval)
    *addr = newval;
  return old;
```

MIPS load-linked and store-conditional
Load-linked returns the current value of a memory location, while a subsequent store-conditional to the same memory location will store a new value only if no updates have occurred to that location since the load-linked.

3.4 Spinlocks in OS/1616

```
struct spinlock {
  volatile spinlock_data_t lk_lock;
  struct cpu *lk_holder;
};
void spinlock_init(struct spinlock *lk);
void spinlock_acquire(struct spinlock *lk);
void spinlock_release(struct spinlock *lk);
```

`spinlock_acquire` calls `spinlock_data_testandset` in a loop until the lock is acquired.
Using Load-Linked/Store-Conditional:

```
/* return value 0 indicates lock was acquired */
spinlock_data_testandset(volatile spinlock_data_t *sd) {
  spinlock_data_t x,y;
  y = 1;
  __asm volatile(
    ".set push;"      /* save assembler mode */
    ".set mips32;"    /* allow MIPS32 instructions */
    ".set volatile;"  /* avoid unwanted optimization */
    "ll %0, 0(%2);"   /* x = *sd */
    "sc %1, 0(%2);"   /* *sd = y; y = success? */
    ".set pop"        /* restore assembler mode */
    : "=r" (x), "+r" (y) : "r" (sd));
  if (y == 0) { return 1; }
  return x;
}
```

3.5 OS/161 Locks

In addition to spinlocks, OS/161 also has *locks*.
Like spinlocks, locks are used to enforce mutual exclusion.

```
struct lock *mylock = lock_create("LockName");

lock_acquire(mylock);
  critical section
lock_release(mylock);
```

Spinlock *spins*, locks *blocks*:

- a thread that calls `spinlock_acquire` spins until the lock can be acquired
- a thread that called `lock_acquire` *blocks* until the lock can be acquired

3.6 Thread Blocking

Sometimes a thread will need to wait for something, e.g.:

- wait for a lock to be released by another thread
- wait for data from a (relatively) slow device
- wait for input from a keyboard
- wait for busy device to become idle

When a thread blocks, it stops running:

i. the scheduler chooses a new thread to run

ii. a context switch from the blocking thread to the new thread occurs

iii. the blocking thread is queued in a *wait queue* (not on the ready list)

Eventually, a blocked thread is signalled and awakened by another thread.

3.7 Wait Channels in OS/161

Wait channels are used to implement thread blocking in OS/161.

- `void wchan_sleep (struct wchan *wc);`
  blocks calling thread on wait channel `wc`
  causes a context switch, like `thread_yield`
- `void wchan_wakeall (struct wchan *wc);`
  unblocks all threads sleeping on waiting channel `wc`
- `void wchan_wakeone (struct wchan *wc);`
  unblocks one thread sleeping on waiting channel `wc`
- `void wchan_lock (struct wchan *wc);`
  prevent operations on wait channel `wc`

There can be many different wait channels, holding threads that are blocked for different reasons.

Ready threads are queued on the ready queue, blocked threads are queued on wait channels (refer to Figure 1).

3.8 Semaphores

A semaphore is a synchronization primitive that can be used to enforce mutual exclusion requirements. It can also be used to solve other kinds of synchronization problems.

A semaphore is an object that has an integer value, and that supports two operations (*atomic* by definition):

P: if the semaphore value is greater than 0, decrement the value.
Otherwise, wait until the value is greater than 0 and then decrement it.

V: increment the value of the semaphore

3.9 Condition Variables in OS/161

Each cv is purposed to work together along with a lock: cvs are used *from within the critical section protected by the lock.*
Supported operations:

wait: causes calling thread to block, and it releases the lock associated within the cv. Once the thread is unblocked, it reacquires the lock.

signal: one of the threads blocked on the signalled cv is unblocked

broadcast: all threads blocked on the cv are unblocked

3.9.1 Using Condition Variables

Cvs allow threads to wait for arbitrary conditions to become true inside of a critical section. By convention, each cv corresponds to a particular condition of interest to an application. When a condition is not true, a thread can `wait` on the corresponding cv until it becomes true.
When a thread detects that a condition is true, it uses `signal` or `broadcast` to notify any threads that may be waiting.

Note: signalling (or broadcasting to) a cv that has no waiting threads has *no effect*. Signals *do not* accumulate.

3.9.2 Waiting on Condition Variables

The OS/161 condition variables follows the Mesa-style condition variables.
When a blocked thread is unblocked, it reacquires the lock before returning from the `wait` call.
A thread is in the critical section when it calls `wait`, and it will be in the critical section when `wait` returns. In between the call and the return, while the caller is blocked, the caller is *out* of the critical section, and other threads may enter!
The thread that calls `signal` (or `broadcast`) to wake up the waiting thread will itself be in the critical section when it signals. The waiting threads need to wait until the signaller releases the lock before it unblocks and return from the `wait` call.

3.9.3 Example

```
int volatile count = 0;
struct lock *mutex;
struct cv *notfull, *notempty;
/* Initialization Note: the lock and cv's must be created
 * using lock_create() and cv_create() before Produce()
 * and Consume() are called
 */
Produce(itemType item) {
  lock acquire(mutex);
  while (count == N) {
    cv_wait(notfull, mutex);
  }
  add item to buffer
```

```
itemType Consume() {
  lock_acquire(mutex);
  while (count == 0)
    cv_wait(nonempty, mutex);
  remove item from buffer
  count = count - 1;
```

```
    count = count + 1;
    cv_signal(notempty, mutex);
    lock release(mutex);
  }
```

```
    cv_signal(notfull, mutex);
    lock_release(mutex);
    return(item);
  }
```

3.10 Deadlocks

Threads are *deadlocked* when none of the threads can make progress (i.e. waiting on a lock indefinitely).
Waiting does not resolve the deadlock.
Those threads are permanently stuck.

3.10.1 Two Techniques for Deadlock Prevention

No Hold and Wait: prevent a thread from requesting resources if it currently has resources allocated to it.
A thread may hold several resources, but to do so it must make a single request for all of them.

Resource Ordering: Order the resource types, and require that each thread acquire resources in increasing resource type order.
That is, a thread may make no request for resources of type less than or equal to i if it is holding resources of type i.

4 Processes and System Calls

4.1 What is a Process?

A *process* is an environment in which an application program runs.
A process includes virtualized *resources* that its program can use:

- one (or more) threads
- virtual memory, used for the program's code and data
- other resources

Processes are created and managed by the *kernel*.
Each program's process *isolates* it from other programs in other processes.

4.2 System Calls

System calls (syscalls) are the interface between processes and the kernel.
A process uses syscalls to request operating system services.

Table 2: Syscall examples

Services	OS/161 Examples
create, destroy, manage processes	`fork, execv, waitpid, getpid`
create, destroy, read, write files	`open, close, remove, read, write`
manage file system and directories	`mkdir, rmdir, link, sync`
interprocess communication	`pipe, read, write`
manage virtual memory	`sbrk`
query, manage system	`reboot, _time`

4.3 Kernel Privilege

Kernel code runs at a higher level of *execution privilege* than application code.
Privilege levels are implemented by the CPU.

The kernel's higher privilege level allows it to do things that the CPU prevents less-privileged (application) programs from doing, like

- application programs cannot modify the page tables that the kernel uses to implement process virtual memories
- application programs cannot halt the CPU

These restrictions allow the kernel to keep processes isolated from one another — and from the kernel.

Note: application programs cannot directly call kernel functions or access kernel data structures.

4.4 How System Calls Work

There are only *two* things that make kernel code run!

On the MIPS, the same mechanism handles exceptions and interrupts, and there is a single handler for both in the kernel. The handler uses these codes to determine what triggered it to run.

4.4.1 Interrupts

Interrupts are generated by devices.
An interrupt means a device (hardware) needs attention.

Recall that an interrupt causes the hardware to transfer control to a fixed location in memory, where an *interrupt handler* is located.

Interrupt handlers are part of the kernel.
If an interrupt occurs while an application program is running, control will jump from the application to the kernel's interrupt handler.

When an interrupt occurs, the processor switches to privileged execution mode when it transfers control to the interrupt handler. This is how the kernel gets its execution privilege.

4.4.2 Exceptions

Exceptions are caused by instruction execution.
An exception means that a running program needs attention.

Exceptions are conditions that occur during the execution of a program instruction.
Examples: arithmetic overflows, illegal instructions, or page faults.

Exceptions are detected by the CPU during instruction execution.

The CPU handles exceptions like it handles interrupts:

- control is transferred to a fixed location, where an *exception handler* is located
- the processor is switched to privileged execution mode

The exception handler is part of the kernel

MIPS Exception Types:

```
EX_IRQ    0     /* Interrupt */
EX_MOD    1     /* TLB Modify (write to read-only page) */
EX_TLBL   2     /* TLB miss on load */
EX_TLBS   3     /* TLB miss on store */
```

```
EX_ADEL   4     /* Address error on load */
EX_ADES   5     /* Address error on store */
EX_IBE    6     /* Bus error on instruction fetch */
EX_DBE    7     /* Bus error on data load *or* store */
EX_SYS    8     /* Syscall */
EX_BP     9     /* Breakpoint */
EX_RI     10    /* Reserved (illegal) instruction */
EX_CPU    11    /* Coprocessor unusable */
EX_OVF    12    /* Arithmetic overflow */
```

4.4.3 Performing a Syscall

To perform a syscall, the application program needs to cause an exception to make the kernel execute!
The kernel's exception handler checks the exception code (set by the CPU when the exception is generated) to distinguish syscall exceptions from other types of exceptions.

On the MIPS, `EX_SYS` is the syscall exception.
To cause this exception on the MIPS, the application executes a special purpose instruction: `syscall`. Other processor instruction sets include similar instructions, e.g., `syscall` on x86.

4.4.4 System Call Timeline

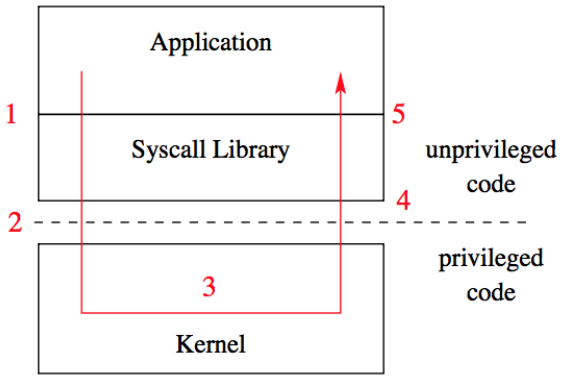


Figure 2: System call software stack

1. application calls library wrapper function for desired system call
2. library function performs `syscall` instruction
3. kernel exception handler runs
   - creates a trap frame to save application program state
   - determines that this is a syscall exception
   - determines which syscall is being requested

- does the work for the requested system call
- restores the application program state from the trap frame
- returns from the exception

4. library wrapper function finishes and returns from its call
5. application continues execution

4.4.5 Which Syscall?

The kernel uses system call codes to determine which syscall the application is requesting,

- the kernel defines a code for each syscall it understands
- the kernel expects the application to place a code in a specified location before executing the `syscall` instruction.
  For OS/161 on the MIPS, the code goes in register `v2`
- the kernel's exception handler checks this code to determine which system call has been requested
- the codes and and code location are part of the *kernel ABI* (Application Binary Interface)

4.4.6 Syscall Parameters

The application places parameter values in kernel-specified locations before the `syscall`, and looks for return values in kernel-specified locations after the exception handler returns,

- The locations are part of the kernel ABI
- Parameter and return value placement is handled by the application system call library functions
- On the MIPS
  - parameters go in registers `a0, a1, a2, a3`
  - result success/fail code is in `a3` on return
  - return value or error code is in `v0` on return

4.5 User and Kernel Stacks

Every OS/161 process thread has two stacks, although it only uses one at a time.

User (Application) Stack: used while application code is executing,

- this stack is located in the application's virtual memory

- it holds activation records for application functions
- the kernel creates this stack when it sets up the virtual address memory for the process

Kernel Stack: used while the thread is executing kernel code, after an exception or interrupt,

- this stack is a kernel structure
- in OS/161, the `t_stack` field of the `thread` structure points to this stack
- this stack holds activation records for kernel functions
- this stack also holds *trap frames* and *switch frames* (because the kernel creates trap frames and switch frames)

4.6 Exception Handling in OS/161

First to run is careful assembly code that

- saves the application stack pointer
- switches the stack pointer to point to the thread's kernel stack
- carefully saves application state and the address of the instruction that was interrupted in a trap frame on the thread's kernel stack
- calls `mips_trap`, passing a pointer to the trap frame as a parameter

After `mips_trap` is finished, the handler will

- restore application state (including the application stack pointer) from the trap frame on the thread's kernel stack
- jump back to the application instruction that was interrupted, and switch back to unprivileged execution mode

4.6.1 mips_trap

`mips_trap` determines what type of exception this is by looking at the exception code. There is a separate handler in the kernel for each type of exception:

- interrupt? call `mainbus_interrupt`
- address translation exception? call `vm_fault`
- system call? call `syscall` (kernel function), passing it the trap frame pointer

4.7 Multiprocessing

Multiprocessing (or multitasking) means having multiple processes existing at the same time.

All processes share the available hardware resources, with the sharing coordinated by the OS:

- Each process' virtual memory is implemented using some of the available physical memory.
  The OS decides how much memory each process gets.
- Each process' threads are scheduled onto the available CPUs (or CPU cores) by the OS.
- Processes share access to other resources (e.g., disks, network devices, I/O devices) by making syscalls.
  The OS controls this sharing.

The OS ensures that processes are isolated from one another.
Interprocess communication should be possible, but only at the explicit request of the processes involved.

Table 3: Syscalls for Process Management

	Linux	OS/161
Creation	`fork, execv`	`fork, execv`
Destruction	`_exit, kill`	`_exit`
Synchronization	`wait, waitpid, pause, ...`	`waitpid`
Attribute Mgmt	`getpid, getuid, nice, getrusage, ...`	`getpid`

4.8 `fork`, `_exit`, and `waitpid`

`fork` creates a new process (the *child*) that is a clone of the original (the *parent*),

- after `fork`, both parent and child are executing copies of the same program
- virtual memories of parent and child are identical at the time of the fork, but may diverge afterwards
- `fork` is called by the parent, but returns in *both* the parent and the child!
  Parent and child see different return values from `fork`, parent gets child's pid and child gets 0.

`_exit` terminates the process that calls it,

- process can supply an exit status code when it exits
- kernel records the exit status code in case another process asks for it (via `waitpid`)

`waitpid` lets a process wait for another to terminate, and retrieve its exit status code.

4.9 The `execv` system call

`execv` changes the program that a process is running.
The calling process's current virtual memory is destroyed.
The process gets a new virtual memory, initialized with the code and data of the new program to run.
After `execv`, the new program starts executing.

5 Assignment 2A Review

5.1 `fork`

i. Create process structure for child process.
Use `proc_create_runprogram` to create the process structure, sets up the VFS and console.

ii. Create and copy address space (and data) from parent to child.
Child process must be identical to the parent process.
`as_copy` creates a new address space, and copies the pages from the old address space to the new one.
Address space is not associated with the new process yet.

iii. Attach the newly created address space to the child process structure.

iv. Assign PID to child process and create the parent/child relationship.
PIDs should be unique (no two processes should have the same PID).
PIDs should be reusable.

v. Create thread for child process (need a safe way to pass the trapframe to the child thread).
Use `thread_fork` to create a new thread.
Need to pass trapframe to the child thread.

vi. Child thread needs to put the trapframe onto the stack and modify it so that it returns the current value (and executes the next instruction)

vii. Call `mips_usermode` in the child to go back to userspace

Note: need to ensure that the new thread does not go to user mode until its address space and trap frame have been set up (requires synchronization)! In addition, the parent process must not return to user mode until its address space and trap frame have been copied.

5.2 `waitpid`

Only the parent can call `waitpid` on its children.
If `waitpid` is called before the child process exits, then the parent must wait/block.
If `waitpid` is called after the child process has exited, then the parent should immediately get the exit status and exit code.
PID cleanup should not rely on `waitpid`. Parent process is not guaranteed to call `waitpid` when it exits.

5.3 `getpid`

Returns the PID of the current process.
Need to perform process assignment even without/before any fork calls. The first user process might call `getpid` before creating any children. `getpid` needs to return a valid PID for this process.

5.4 `_exit`

Causes the current process to exit.
Exit code is passed to the parent process.

6 Assignment 2B Review

```
int execv (const char* program, char** args);
```

Replaces currently executing program with a newly loaded program image.
Process id remains unchanged.
Path of the program is passed in as `program`.
Arguments to the program (`args`) is an array of `NULL` terminated strings.
The array is terminated by a `NULL` pointer.
In the new user program, `argv[argc]` should `== NULL`.

6.1 `runprogram`

`execv` is very similar to `runprogram`
`runprogram` is used to load and execute the first program from the `menu`,

1. Opens the program file using `vfs_open(progname, ...)`
2. Creates a new address space (`as_create`), switches the process to that address space (`curproc_setas`) and then activate it (`as_activate`)
3. Using the opened program file, load the program image using `load_elf`
4. Define the user stack using `as_define_stack`
5. Call `enter_new_process` with no parameters, the stack pointer (determined by `as_define_stack`) *and* entry point for the executable (determined by `load_elf`)

6.2 `execv`

- Count the number of arguments and copy them into the kernel
- Copy the program path into the kernel
- Open the program file using `vfs_open(progname, ...)`
- Create new address space, set process to the new address space, and activate it
- Using the opened program file, load the program image using `load_elf`
- Need to copy the arguments into the new address space.
  Consider copying the arguments (both the array and the strings) onto the user stack as part of `as_define_stack`
- Delete old address space
- Call `enter_new_process` with address to the arguments on the stack, the stack pointer (from `as_define_stack`), and the program entry point (from `vfs_open`)

6.3 Argument Passing

When copying from/to userspace:

- Use `copyin`/`copyout` for fixed size variables (integers, arrays, etc.)
- Use `copyinstr`/`copyoutstr` when copying `NULL` terminated strings

Useful defines/macros:

- `USERSTACK` (base address of the stack)
- `ROUNDUP` (useful for memory alignment)

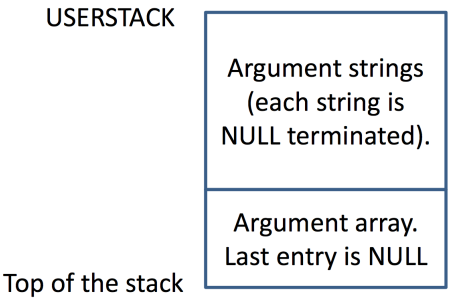
Common mistakes:

- Remember that `strlen` does not count the `NULL` terminator.
  Make sure to include space for the `NULL` terminator
- User pointers should be of the type `userptr_t`
- Make sure to pass a pointer to the top of the stack to `enter_new_process`

6.4 Alignment

When storing items on the stack, pad each item such that they are 8-byte aligned.
e.g., `args_size = ROUNDUP(args_size, 8);`

Strings don't have to be 4 or 8-byte aligned. However, pointers to strings need to be 4-byte aligned.



7 Virtual Memory

7.1 Physical Memory and Addresses

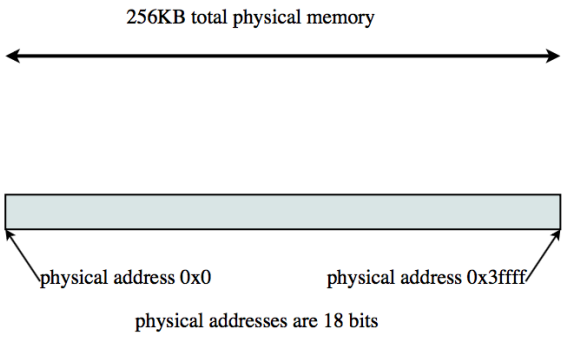


Figure 3: An example physical memory, $P = 18$

If physical addresses have P *bits*, the maximum amount of addressable physical memory is 2^P *bytes* (assuming a byte-addressable machine).

- Sys/161 MIPS processor uses 32 bit physical addresses ($P = 32$) $\Rightarrow$ maximum physical memory size of 2^{32} bytes, or 4GB
- Larger values of P are common on modern processors, e.g., $P = 48$, which allows 256TB of physical memory to be addressed

The actual amount of physical memory on a machine may be less than the maximum amount that can be addressed.

7.2 Virtual Memory and Addresses

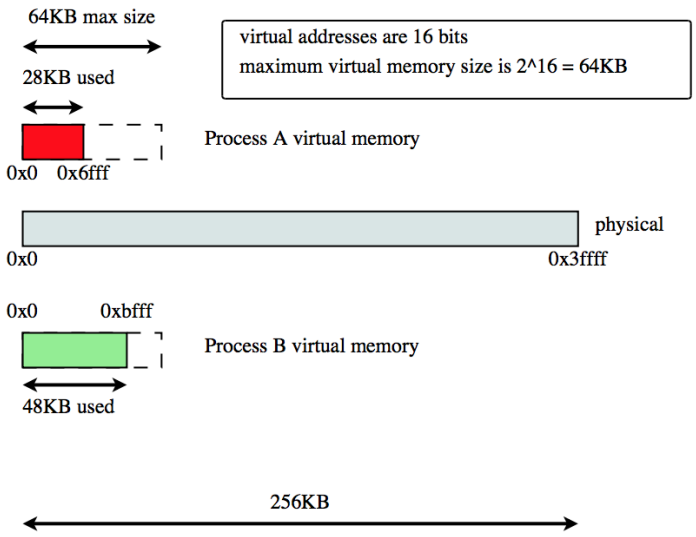


Figure 4: An example of virtual memory, $V = 16$

The kernel provides a separate, private *virtual* memory for each process.
The virtual memory of a process holds the code, data, and stack for the program that is running in that process.
If virtual addresses are V bits, the *maximum* size of a virtual memory is 2^V bytes.

- For the MIPS, $V = 32$

Running applications see only virtual addresses, e.g.,

- program counter and stack pointer hold *virtual addresses* of the next instruction and the stack
- pointers to variables are *virtual addresses*
- jumps/branches refer to *virtual addresses*

Each process is isolated in its virtual memory, and cannot access other process' virtual memories.

7.3 Address Translation

Each virtual memory is mapped to a different part of physical memory.
Since virtual memory is not real, when an process tries to access (load or store) a virtual address, the virtual address is *translated* (mapped) to its corresponding physical address, and the load or store is performed in physical memory.
Address translation is performed in hardware, using information provided by the kernel.

7.4 Address Translation for Dynamic Relocation

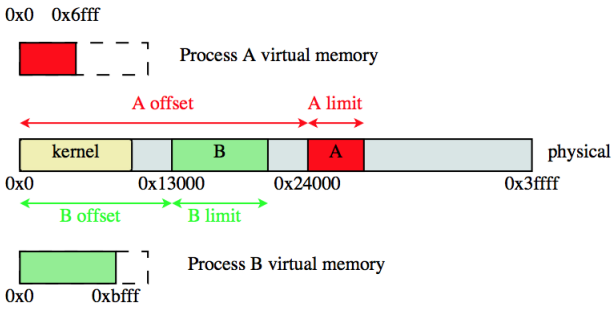


Figure 5: Example of dynamic relocation

CPU includes a *memory management unit* (MMU), with a *relocation register* and a *limit register*.

- relocation register holds the physical offset (R) for the running process' virtual memory
- limit register holds the size L of the running process' virtual memory

To translate a virtual address v to a physical address p:

```
if v ≥ L then generate exception
else
    p ← v + R
```

Translation is done in hardware by the MMU.
The kernel maintains a separate R and L for each process, and changes the values in the MMU registers when there is a context switch between processes.

Example. $v =$ `0x102c` $p =$ `0x102c` $+$ `0x24000` $=$ `0x2502c`
$v =$ `0x8800` $p =$ `exception` since `0x8800` $\geq$ `0x7000`

7.5 Properties of Dynamic Relocation

Each virtual address space corresponds to a *contiguous range of physical addresses*.
The kernel is responsible for deciding *where* each virtual address space should map in physical memory.

- the OS must track which part of physical memory are in use, and which parts are free
- since different address spaces may have different sizes, the OS must allocate/deallocate variable-sized chunks of physical memory
- hence creates potential for *fragmentation* of physical memory

7.6 Paging

7.6.1 Paging: Physical Memory

Physical memory is divided into fixed-size chunks called *frames* or *physical pages*.

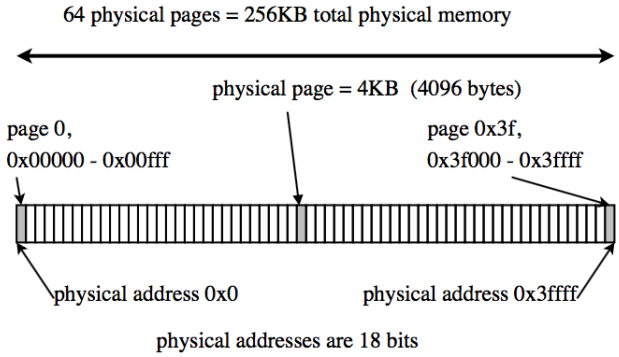


Figure 6: The frame size if 2^{12} bytes (4KB)

7.6.2 Paging: Virtual Memory

Virtual memories are divided into fixed-size chunks called *pages*.
Page size is equal to frame size.

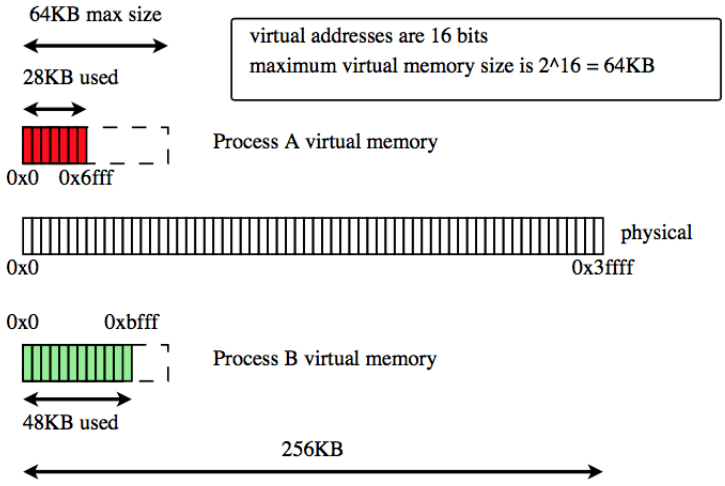
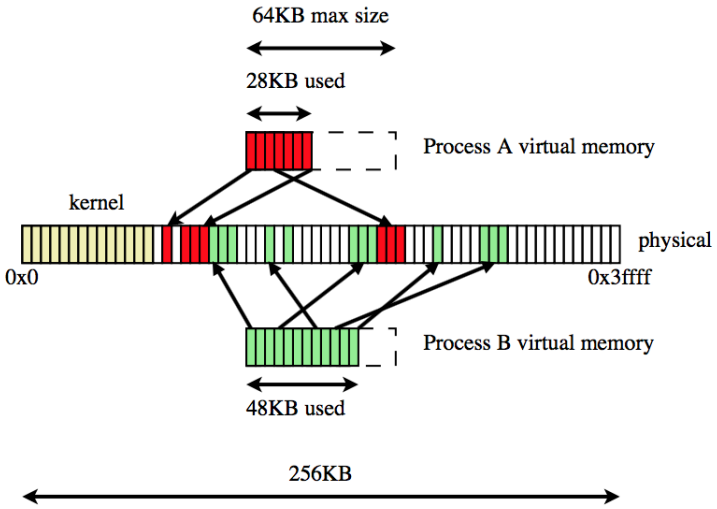


Figure 7: Page size is 4KB here

7.6.3 Paging: Address Translation



Each page maps to a different frame. Any page can map to any frame.

7.6.4 Paging: Address Translation

The MMU includes a *page table base register* which points to the page table for the current process.

How the MMU translate a virtual address:

1. determines the *page number* and *offset* of the virtual address
   - page number is the virtual address divided by the page size
   - offset is the virtual address modulo the page size
2. looks up the page's entry (PTE) in the current process page table, using the page number
3. if the PTE is not valid, raise an exception
4. otherwise, combine page's frame number from the PTE with the offset to determine the physical address; physical address is (frame number $\cdot$ frame size) + offset

7.6.5 Other Information Found in PTEs

PTEs may contain other fields, in addition to the frame number and valid bit.

Example 1: write protection bit

- can be set by the kernel to indicate that a page is read-only

- if a write operation (e.g., MIPS `lw`) uses a virtual address on a read-only page, the MMU will raise an exception when it translate the virtual address

Example 2: bits to track page usage

- reference (use) bit: has the process used this page recently?
- dirty bit: have contents of this page been changed?
- these bits are set by the MMU, and read by the kernel

7.6.6 Page Tables: How Big?

A page table has one PTE for each page in the virtual memory

- page table size = number of pages $\cdot$ size of PTE
- number of pages = $\frac{\text{virtual memory size}}{\text{page size}}$

The page table of a 64KB virtual memory, with 4KB pages, is 64 bytes, assuming 32 *bits* for each PTE.
Page tables for larger virtual memories are larger.

7.6.7 Page Tables: Where?

Page tables are kernel data structures, i.e. page tables for all processes are in the kernel's stack.

7.7 Summary: Roles of the Kernel and the MMU

Kernel:

- Manage MMU (memory management unit) registers on address space switches (context switch from thread in one process to thread in a different process)
- Create and manage page tables
- Manage (allocate/deallocate) physical memory
- Handle exceptions raised by the MMU

Memory Management Unit, MMU, (hardware):

- Translate virtual addresses to physical addresses
- Check for and raise exceptions when necessary

7.8 TLBs

Execution of each machine instruction may involve one, two, or more memory operations

- one to fetch instruction
- one or more for instruction operands

Address translation through a page table adds one extra memory operation (for page table entry lookup) for each memory operation performed during instruction execution.
This can be slow!

Solution: include a *Translation Lookaside Buffer* (TLB) in the MMU,

- TLB is a small, fast, dedicated cache for address translation in the MMU
- Each TLB entry stores a (page number ⇒ frame number) mapping

7.8.1 TLB Use

What the MMU does to translate a virtual address on page p:

```
if there is an entry (p, f) in the TLB then
  return f  /* TLB hit! */
else
  find p's frame number (f) from the page table
  add (p, f) to the TLB, evicting another entry if full
  return f  /* TLB miss */
```

If the MMU cannot distinguish TLB entries from different address spaces, then the kernel must *clear* or *invalidate* the TLB on *each* context switch from one process to another!

This is a *hardware-managed TLB*,

- the MMU handles TLB misses, including page table lookup and replacement of TLB entries
- MMU must understand the kernel's page table format

7.8.2 Software-Managed TLBs

The MIPS has a *software-managed TLB*, which translates a virtual address on page p like this:

```
if there is an entry (p,f) in the TLB then
  return f        /* TLB hit! */
else
  raise exception /* TLB miss */
```

In case of a TLB miss, the kernel must

1. determine the frame number of p
2. add (p, f) to the TLB, evicting another entry if necessary

After the miss is handled, the instruction that caused the exception is re-tried.

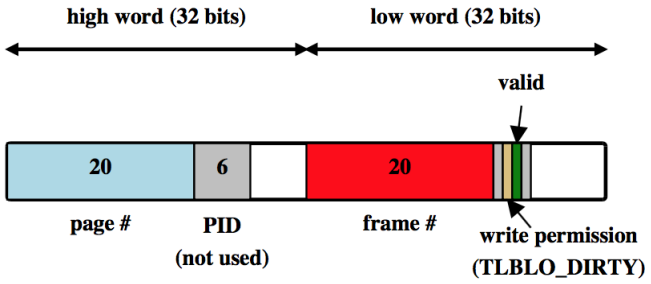


Figure 8: The MIPS R3000 TLB

The MIPS TLB has room for 64 entries. Each entry is 64 bits (8 bytes) long.

7.9 Large, Sparse Virtual Memories

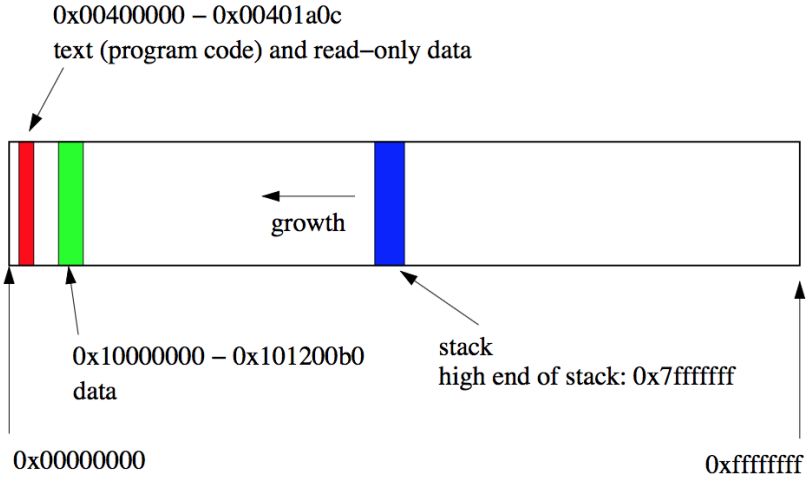


Figure 9: A more realistic virtual memory; this is a layout of the virtual address space for `user/testbin/sort` in OS/161

Virtual memory may be large,

MIPS: $V = 32$, max virtual memory is 2^{32} bytes (4 GB)

x86-64: $V = 48$, max virtual memory is 2^{48} bytes (256 TB)

Much of the virtual memory may be unused (see Figure 9)!

Application may use *discontinuous segments* of the virtual memory.
One reason is that we have to give room for the stack to grow in the virtual memory!

7.10 Limitations of Simple Address Translation Approaches

A kernel that used simple dynamic relocation would have to allocate 2 GB of contiguous physical memory for `testbin/sort`'s virtual memory, even though `sort` only uses about 1.2 MB.

A kernel that used simple paging would require a page table with 2^{20} PTEs (assuming page size is 4 KB) to map `testbin/sort`'s address space,

- this page table is actually larger than the virtual memory that `sort` needs to use
- most of the PTEs are marked as invalid
- this page table has to be contiguous in kernel memory

7.11 Segmentation

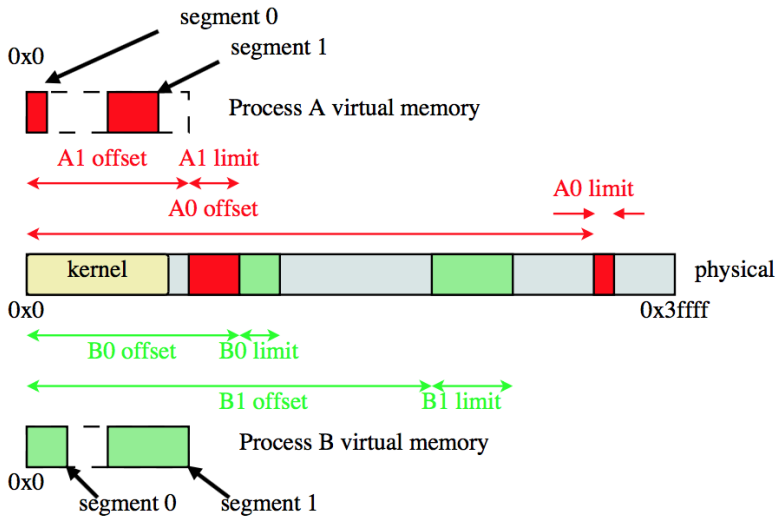


Figure 10: Example of segment address space

Instead of mapping the entire virtual memory to physical, we can provide a separate mapping for each segment of the virtual memory that the application actually uses.

Instead of a single offset and limit for the entire address space, the kernel maintains an offset and limit for each segment,

- The MMU has multiple offset and limit registers, one pair for each segment

With segmentation, a virtual address can be thought of as having two parts: segment ID and offset within segment.

With K bits for the segment ID, we can have up to:

- 2^K segments
- 2^{V-K} bytes per segment

The kernel decides where each segment is placed in physical memory. Fragmentation of physical memory is possible!

7.12 Translating Segmented Virtual Addresses

The MMU needs a relocation register and a limit register for each segment,

- let R_i be the relocation offset for the ith segment
- let L_i be the limit of the ith segment

To translate virtual address v to a physical address p:

```
split p into segment number (s) and address within segment (a)
if a ≥ L_s then generate exception
else
  p ← a + R_s
```

As for dynamic relocation, the kernel maintains a separate set of relocation offsets and limits for each process, and changes the values in the MMU's registers when there is a context switch between processes.

7.13 Two-Level Paging

Instead of having a single page table to map an entire virtual memory, we can split the page table into smaller page tables, and add page table directory.

- Instead of one larger, contiguous table, we have multiple smaller tables
- If all PTEs in a small table are invalid, we can avoid creating that table entirely

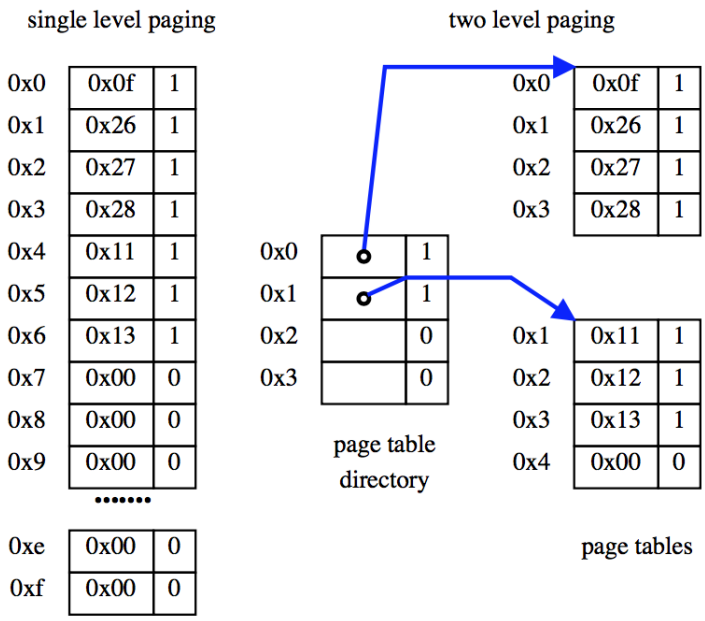


Figure 11: Single vs. Two Level Paging

Each virtual address has three parts:

1. Level one page number: used to index the directory
2. Level two page number: used to index a page table
3. Offset within the page

7.14 Address Translation with Two-Level Paging

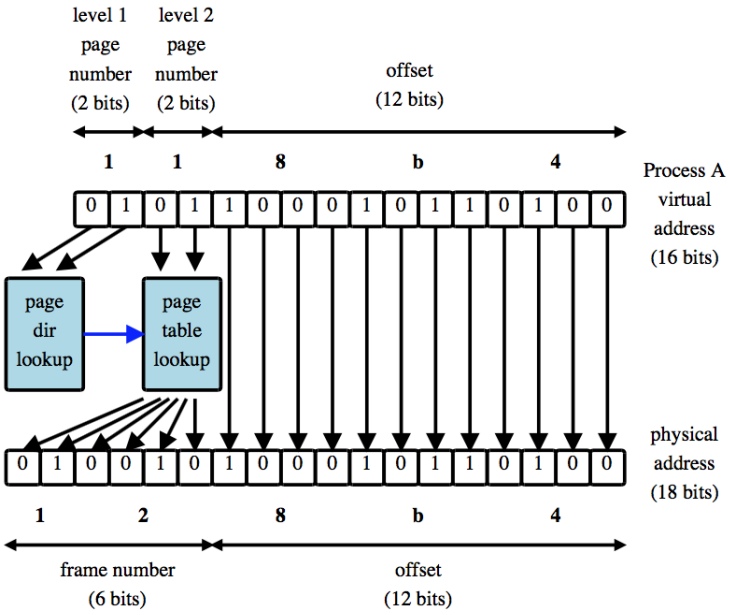


Figure 12: Example of two-level address translation

The MMU's *page table base register* points to the page table directory for the current process.
Each virtual address v has three parts: (p_1, p_2, o)
How the MMU translate a virtual address:

1. Index into the page table directory using p_1 to get a pointer to a 2nd level page table
2. If the directory entry is not valid, raise an exception
3. Index into the 2nd level page table using p_2 to find the PTE for the page being accessed
4. If the PTE is not valid, raise an exception
5. Otherwise, combine the frame number from the PTE with o to determine the physical address (as for single-level paging)

7.15 Limits of Two-Level Paging

A goal of two-level paging was to keep individual page tables small.
Suppose we have 40 bit virtual addresses ($V = 40$) and that

- the size of a PTE is 4 bytes
- page size is 4KB (2^{12} bytes)
- we'd like to limit each page table's size to 4KB

Problem: for large address spaces, we may need a large page table directory!

- There can be up to 2^{28} pages in a virtual memory
- A single page table can hold 2^{10} PTEs
- We may need up to 2^{18} page tables
- Our page table directory will have to have 2^{18} entries
- If a directory entry is 4 bytes, the directory will occupy 1MB

This is the problem we were trying to avoid by introducing a second level!

7.16 Multi-Level Paging

We can solve the large directory problem by introducing additional levels of directories.
Example: 4-level paging in x86-64 architecture.
Properties of Multi-Level Paging:

- can map large virtual memories by adding more levels
- individual page table/directories can remain small
- can avoid allocating page tables and directories that are not needed for programs that use a small amount of virtual memory
- TLB misses become *more* expensive as the number of levels goes up, since more directories must be accessed to find the correct PTE

7.17 Virtual Memory in OS/161 on MIPS: `dumbvm`

The MIPS uses 32-bit paged virtual and physical addresses.
The MIPS has a software-managed TLB,

- TLB raises an exception on every TLB miss
- kernel is free to record page-to-frame mappings however it wants to

TLB exceptions are handled by a kernel function called `vm_fault`
`vm_fault` uses information from an `addrspace` structure to determine a page-to-frame mapping to load into the TLB,

- there is a separate `addrspace` structure for each process
- each `addrspace` structure describes where its process's pages are stored in physical memory
- an `addrspace` structure does the same job as a page table, but the `addrspace` structure is simpler because OS/161 places all pages of each segment *contiguously* in physical memory

7.17.1 The `addrspace` Structure

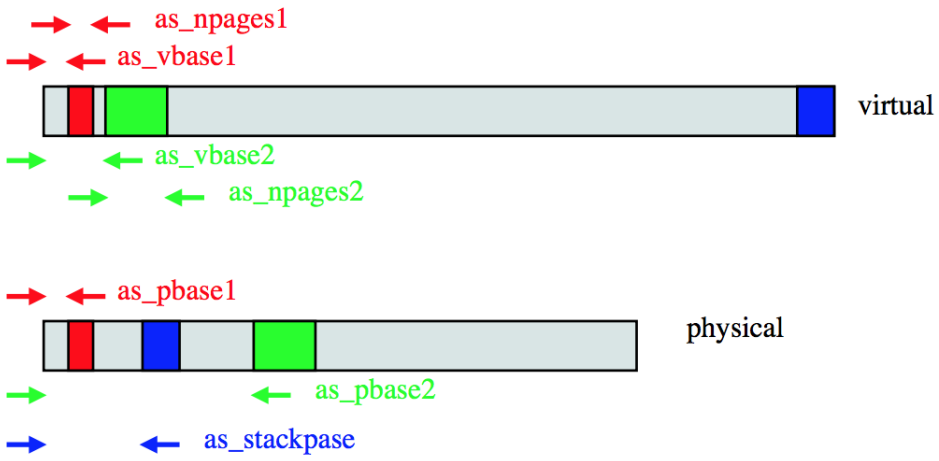


Figure 13: `addrspace` diagram

```
struct addrspace {
  vaddr_t as_vbase1;      /* base virtual address of code segment */
  paddr_t as_pbase1;      /* base physical address of code segment */
  size_t as_npages1;      /* size (in pages) of code segment */
  vaddr_t as_vbase2;      /* base virtual address of data segment */
  paddr_t as_pbase2;      /* base physical address of data segment */
  size_t as_npages2;      /* size (in pages) of data segment */
  paddr_t as_stackpbase;  /* base physical address of stack */
};
```

7.17.2 Address Translation: OS/161 dumbvm Example

Note: in OS/161, the stack is 12 pages and the page size is 4 KB = 0x1000.

Variable/Field	Process 1	Process 2
as_vbase1	0x0040 0000	0x0040 0000
as_pbase1	0x0020 0000	0x0050 0000
as_bpages1	0x0000 0008	0x0000 0002
as_vbase2	0x1000 0000	0x1000 0000
as_pbase2	0x0080 0000	0x00A0 0000
as_npages	0x0000 0010	0x0000 0008
as_stackpbase	0x0010 0000	0x00B0 0000

7.18 Initializing an Address Space

When the kernel creates a process to run a particular program, it must create an address space for the process, and load the program's code and data into that address space.

OS/161 *pre-loads* the address space before the program runs. Many other OS load pages on *demand*.

A program's code and data is described in an *executable file*, which is created when the program is compiled and linked.

OS/161 (and some other operating systems) expect executable files to be in ELF (**E**xecutable and **L**inking **F**ormat) format.

The OS/161 execv system call re-initializes the address space of a process

```
int execv (const char* program, char** args)
```

The `program` parameter of the `execv` system call should be the name of the ELF executable file for the program that is to be loaded into the address space.

7.19 ELF Files

ELF files contain address space segment descriptions, which are useful to the kernel when it is loading a new address space.

The ELF file identifies the (virtual) address of the program's first instruction.

The ELF file also contains lots of other information (e.g., section descriptors, symbol tables) that is useful to compilers, linkers, debuggers, loaders, and other tools used to build programs.

7.19.1 Address Space Segments in ELF Files

The ELF file contains a header describing the segments and segment *images*.

Each ELF segment describes a contiguous region of the virtual address space.

The header includes an entry for each segment which describes:

- the virtual address of the start of the segment
- length of the segment in the virtual address space
- location of the start of the segment image in the ELF file (if present)
- the length of the segment image in the LEF file (if present)

The image is an exact copy of the binary data that should be loaded into specified portion of the virtual address space.

The image may be smaller than the address space segment, in which case the rest of the address space segment is expected to be zero-filled.

In OS/161, the kernel copies segment images from the ELF file to the specified portions of the virtual address space.

7.19.2 ELF Files and OS/161

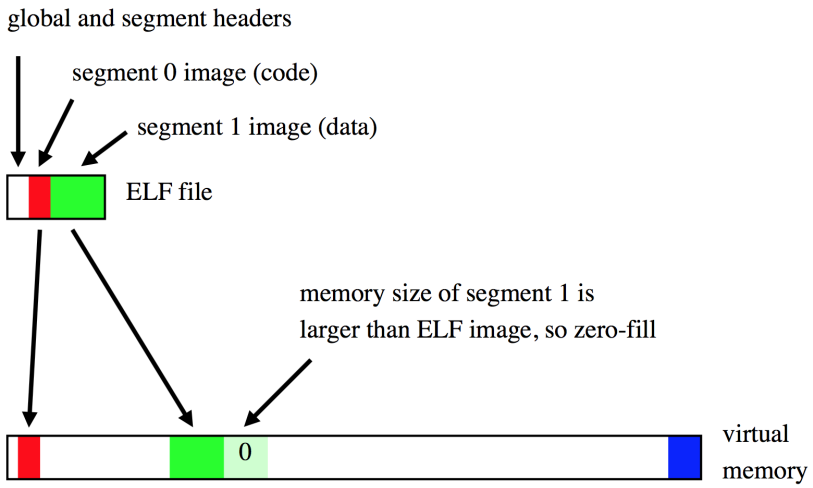


Figure 14: ELF File Diagram

OS/161's `dumbvm` implementation assumes that an LEF file contains two segments:

- a *text segment*, containing the program code and any read-only data
- a *data segment*, containing any other global program data

The ELF file does *not* describe the stack.

`dumbvm` creates a *stack segment* for each process. It is 12 pages long, ending at virtual address `0x7fffffff`.

7.20 Virtual Memory for the Kernel

We would like the kernel to live in virtual memory, but there are some challenges:

Bootstrapping: since the kernel helps to implement virtual memory, how can the kernel run in virtual memory when it is just starting?

Sharing: sometimes data need to be copied between the kernel and application programs? How can this happen if they are in different virtual address spaces?

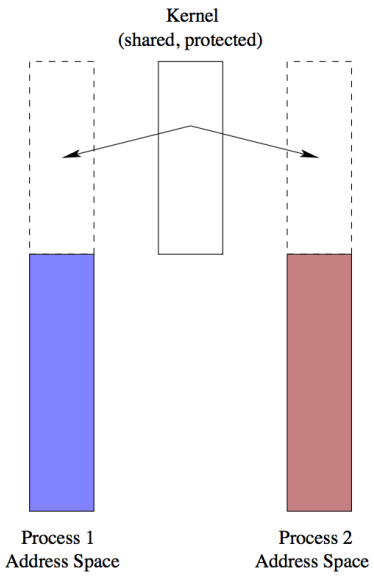


Figure 15: The Kernel in Process' Address Spaces

The sharing problem can be addressed by making the kernel's virtual memory *overlap* with process' virtual memories.
Attempts to access kernel code/data is user mode results in memory protection exceptions, not invalid address exception.

Solutions to the bootstrapping problem are architecture-specific.

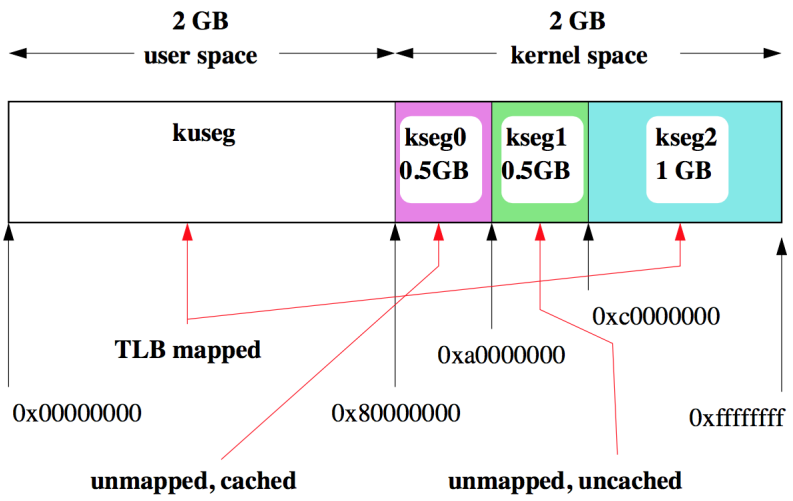


Figure 16: User Space and Kernel Space on the MIPS R3000

In OS/161, user programs live in kuseg, kernel code and data structures live in kseg0, devices are accessed through kseg1, and kseg2 is not used.

7.21 Exploiting Secondary Storage

Goals:

- Allow virtual address spaces that are larger than the physical address space
- Allow greater multiprogramming levels by using less of the available (primary) memory for each process

Method:

- Allow pages from virtual memories to be stored in secondary storage, i.e., on disks or SSDs
- Swap pages (or segments) between secondary storage and primary storage so that they are in primary memory when needed

7.22 Resident Sets and Present Bits

When swapping is used, some pages of each virtual memory will be in memory, and others will not be in memory.

- The set of virtual pages present in physical memory is called the *resident set* of a process.
- A process's resident set will change over time as pages are swapped in and out of physical memory

To track which pages are in physical memory, each PTE needs to contain an extra bit, called the present bit:

- valid = 1, present = 1: page is valid and in memory
- valid = 1, present = 0: page is valid, but not in memory
- value = 0, present - x: invalid page

7.23 Page Faults

When a process tries to access a page that is not in memory, the problem is detected because the page's *present* bit is zero:

- on a machine with a hardware-managed TLB, the MMU detects this when it checks the page's PTE, and generates an exception, which the kernel must handle
- on a machine with a software-managed TLB, the kernel detects the problem when it checks the page's PTE after a TLB miss

This event (attempting to access a non-resident page) is called a *page fault*.

When a page fault happens, it is the kernel's job to:

1. swap the page into memory from secondary storage, evicting another page from memory if necessary
2. update the PTE (set the *present* bit)
3. return from the exception so that the application can retry the virtual memory access that caused the page fault

7.24 Secondary Storage is Slow

Access times for disks are measured in *milli*seconds, SSD read latencies are 10âĂŹs-100âĂŹs of *micro*seconds.

Both of these are much higher than memory access times (100âĂŹs of *nano*seconds)

Suppose that secondary storage access is 1000 times slower than memory access. Then:

- if there is one page fault every 10 memory accesses (on average), the average memory access time with swapping will be about 100 times larger than it would be without swapping
- if there is one page fault every 100 memory accesses (on average), the average memory access time with swapping will be about 10 times larger than it would be without swapping
- if there is one page fault every 1000 memory accesses (on average), the average memory access time with swapping will be about 2 times larger than it would be without swapping.

7.25 Performance with Swapping

To provide good performance for virtual memory accesses, the kernel should try to ensure that page faults are rare.

Some techniques the kernel can use to improve performance:

- limit the number of processes, so that there is enough physical memory per process
- try to be smart about *which* pages are kept in physical memory, and which are evicted
- hide latencies, e.g., by *prefetching* pages before a process needs them

7.25.1 A Simple Replacement Policy: FIFO

Replace the page that has been in memory the longest.

Table 4: Three-frame example

Num	1	2	3	4	5	6	7	8	9	10	11	12
Refs	a	b	c	d	a	b	e	a	b	c	d	e
Frame 1	a	a	a	d	d	d	e	e	e	e	e	e
Frame 2		b	b	b	a	a	a	a	a	c	c	c
Frame 3			c	c	c	b	b	b	b	b	d	d
Fault ?	x	x	x	x	x	x	x			x	x	

7.25.2 Optimal Page Replacement

MIN: replace the page that will not be referenced for the longest time.
MIN requires knowledge of the future.

Table 5: Three-frame example

Num	1	2	3	4	5	6	7	8	9	10	11	12
Refs	a	b	c	d	a	b	e	a	b	c	d	e
Frame 1	a	a	a	a	a	a	a	a	a	c	c	c
Frame 2		b	b	b	b	b	b	b	b	b	d	d
Frame 3			c	d	d	d	e	e	e	e	e	e
Fault ?	x	x	x	x			x			x	x	

7.25.3 Least Recently Used (LRU) Page Replacement

Table 6: Three-frame example

Num	1	2	3	4	5	6	7	8	9	10	11	12
Refs	a	b	c	d	a	b	e	a	b	c	d	e
Frame 1	a	a	a	d	d	d	e	e	e	c	c	c
Frame 2		b	b	b	a	a	a	a	a	a	d	d
Frame 3			c	c	c	b	b	b	b	b	b	e
Fault ?	x	x	x	x	x	x	x			x	x	x

7.26 Locality

Real programs do not access their virtual memories randomly.
Instead, they exhibit *locality*:

- **temporal locality**: programs are more likely to access pages that they have accessed recently than pages that they have not accessed recently
- **spatial locality**: programs are likely to access parts of memory that are close to parts of memory they have accessed recently

Locality helps the kernel keep page fault rates low.

7.27 Measuring Memory Accesses

The kernel is not aware which pages a program is using unless there is an exception.
This makes it difficult for the kernel to exploit locality by implementing a replace policy like LRU.

The MMU can help solve this program by tracking page accesses in hardware.
Simple scheme: add a *use bit* (or *reference bit*) to each PTE. This bit:

- is set by the MMU each time the page is used, i.e., each time the MMU translates a virtual address on that page
- can be read and cleared by the kernel

The use bit provides a small amount of memory usage information that can be exploited by the kernel.

7.28 The Clock Replacement Algorithm

The clock algorithm (also known as "second chance") is one of the simplest algorithms that exploits the use bit.

Clock is identical to FIFO, except that a page is "skipped" if its use bit is set.

The clock algorithm can be visualized as a victim pointer that cycles through the page frames.
The pointer moves whenever a replacement is necessary:

```
while use bit of victim is set
  clear use bit of victim
  victim = (victim + 1) % num_frames
choose victim for replacement
victim = (victim + 1) % num_frames
```

8 Scheduling

8.1 Simple Scheduling Model

We are given a set of *jobs* to schedule.
Only one job can run at a time.
For each job, we are given

- job arrival time (a_i)
- job run time (r_i)

For each job, we define

- *response time*: time between the job's arrival and when the job starts to run
- *turnaround time*: time between the job's arrival and when the job finishes running

We must decide when each job should run, to achieve some goal, e.g., minimize average turnaround time, or minimize average response time.

8.2 Basic Non-Preemptive Schedulers: FCFS and SJF

FCFS: run jobs in arrival time order.

- simple, avoids starvation
- pre-emptive variant: round-robin

SFJ: shortest job first — run jobs in increasing order of r_i

- minimizes average *turnaround* time
- long jobs may starve
- pre-emptive variant: SRTF (shortest remaining time first)

8.3 CPU Scheduling

In CPU scheduling, the "jobs" to be scheduled are the threads.
CPU scheduling typically differs from the simple scheduling model:

- the run times of threads are normally not known
- threads are sometimes not runnable: when they are blocked
- threads may have different priorities

The objective of the scheduler is normally to achieve a balance between

- responsiveness (ensure that threads get to run regularly)
- fairness
- efficiency

8.4 Multi-level Feedback Queues

Objective: good responsiveness for *interactive* threads, non-interactive threads make as much progress as possible.
Key idea: interactive threads are frequently blocked.
Approach: given higher priority to interactive threads, so that they run whenever they are ready.
Problem: how to determine which threads are interactive and which are not?

Algorithm
Scheduler maintains n round-robin ready queues $(Q_1, \ldots, Q_n)$.
Scheduler always chooses a thread from Q_n, unless it is empty

- if Q_n is empty, choose a thread from Q_{n-1}, unless it is empty too
- and so on, choosing a thread from Q_1 only if all other queues are empty.

Threads in queue Q_i uses quantum q_i, typically larger quanta for lower-priority threads $(q_i \geq q_{i+1})$.
If the running thread from Q_i uses its entire quantum and gets preempted, demote it to queue Q_{i-1}.
If a thread blocks, put it into Q_n when it wakes up.
To prevent starvation, periodically move all threads to Q_n.

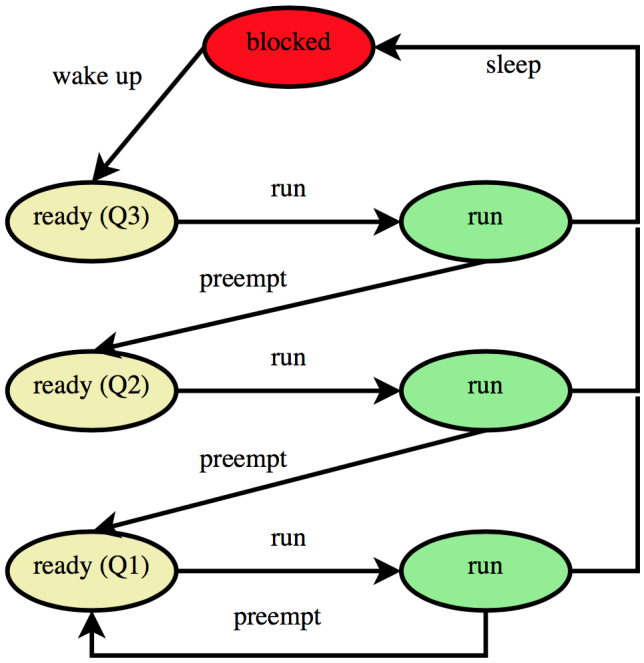


Figure 17: 3 Level Feedback Queue State Diagram

8.5 Linux Complexity Fair Scheduler (CFS) - Main Ideas

Each thread can be assigned a *weight*.
The goal of the scheduler is to ensure that each thread gets a "share" of the processor in proportion to its weight.

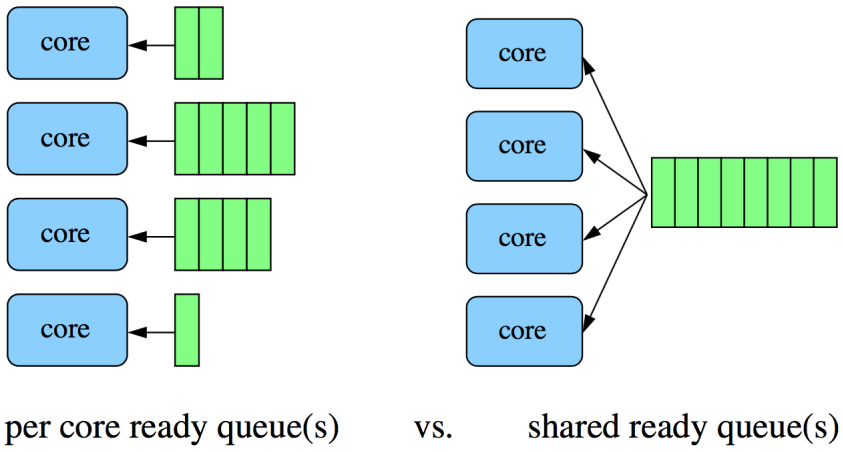
Basic operation:

- track the "virtual" run time of each run-able thread
- always run the thread with the lowest virtual run time

Virtual run time is actual run time adjusted by the thread weights

- suppose w_i is the weight of the ith thread
- actual run time of ith thread is multiplied by $\frac{\sum_j w_j}{w_i}$
- virtual run time advances slowly for threads with high weights, quickly for threads with low weights

8.6 Scheduling on Multi-Core Processors



8.6.1 Scalability and Cache Affinity

Contention and Scalability

- access to shared ready queue is a critical section, mutual exclusion needed
- as number of cores grows, contention for ready queue becomes a problem
- per core design *scales* to a larger number of cores

CPU cache affinity

- as thread runs, data it accesses is loaded into CPU cache(s)
- moving the thread to another core means data must be reloaded into that core's caches
- as thread runs, it acquires an *affinity* for one core because of the cached data
- per core design benefits from affinity by keeping threads on the same core
- shared queue design does not

8.6.2 Load Balancing

In per-core design, queues may have different lengths.
This results in *load imbalance* across the cores,

- cores may be idle while others are busy
- threads on lightly loaded cores get more CPU time than threads on heavily loaded cores

Not an issue in shared queue design.

Per-core designs typically need some mechanism for *thread migration* to address load imbalances,

- migration means moving threads from heavily loaded cores to lightly loaded cores

9 Assignment 3 Review

`dumbvm` is a very limited virtual memory system,

- A full TLB leads to a kernel panic
- Text segment is not read-only
- Uses fixed segmentation (external fragmentation)
- Never reuses physical memory (required restarting the OS after each test)

9.1 TLB Replacement

VM related exceptions are handled by `vm_fault`

`vm_fault` performs address translation and loads the virtual address to physical address mapping into the TLB.

- Iterates through the TLB to find an unused/invalid entry
- Overwrites the unused entry with the virtual to physical address mapping required by the instruction that generated the TLB exception

If the TLB is full, call `tlb_random` to write the entry into a random TLB slot,

- that's it for TLB replacement!
- make sure that virtual page fields in the TLB are unique.

9.2 Read-Only Text Segment

Currently, TLB entries are loaded with `TLBLO_DIRTY` on.

- Pages are therefore read and writable.

Text segment should be read-only.

- Load TLB entries for the text segment with `TLBLO_DIRTY` off.
- `elo &= ~TLBLO_DIRTY;`

Determine the segment of the fault address by looking at the `vbase` and `vtop` addresses.

Unfortunately, this will cause `load_elf` to throw a `VM_FAULT_READONLY` exception.

- It is trying to write to a memory location that is read-only.

We must instead load TLB entries with `TLBLO_DIRTY` on until `load_elf` has completed.

- Consider adding a flag to `struct addrspace` to indicate whether or not `load_elf` has completed.
- When `load_elf` completes, flush the TLB, and ensure that all future TLB entries for the text segment has `TLBLO_DIRTY` off.

Writing to read-only memory address will lead to a `VM_FAULT_READONLY` exception.
This will currently cause a kernel panic.

Instead of panicking, your VM system should kill the process.
Modify `kill_curthread` to kill the current thread,

- very similar to `sys__exit`. However, the exit code/status will be different.
- Consider which signal number this will trigger (look at the beginning of `kill_curthread`).

9.3 Managing Memory

During bootstrap, the kernel allocates memory by calling `getppages`, which in turn calls `ram_stealmem(pages)`.

`ram_stealmem` just allocates pages without providing any mechanism to release pages (see `free_kpages`).

We want to manage our own memory after bootstrap.

In `vm_bootstrap`, call `ram_getsize` to get the remaining physical memory in the system.
Do *not* call `ram_stealmem` again!

Logically partition the memory into fixed size frames.
Each frame is `PAGE_SIZE` bytes.

Keep track of the status of each frame (core-map data structure).

Where should we store the core-map data structure?
Store it in the start of the memory return by `ram_getsize`.
The frames should start after the core-map data structure.
The core-map should track which frames are used and available.
It should also keep track of contiguous memory allocations, because frames need to be contiguous.

9.4 Alloc and Free

`alloc_kpages(int npages)`:

- Allocate frames for both `kmalloc` and address spaces

- Frames need to be contiguous

`kfree_kpages(vaddr_t addr)`:

- Free pages allocated with `alloc_kpages`
- We don't specify how many pages we need to free. It should free the same number of pages that was allocated.
- Update core-map to make frames available after `free_kpages`

Consider adding some information in the core-map to help determine the number of pages that need to be free.
e.g. If 4 contiguous frames were allocated using `alloc_kpages`, then store 4 in the core-map entry for the start for the start of the four frames.

9.5 Page Tables

In order to avoid external fragmentation, we need to introduce paging.
New VM system combines segmentation with page.
Three segments:

- Text (read-only)
- Data
- Stack

Create a page table for each segment,

- each page table entry should include the frame number

In `dumbmvm`, `struct addrspace` has the following fields:

- `vaddr_t as_vbase1`
- `paddr_t as_pbase1`
- `size_t as_npages1`
- `vaddr_t as_vbase2`
- `paddr_t as_pbase2`
- `size_t as_npages2`
- `paddr_t as_stackpbase`

With segmentation and paging, replace `pbase` with page table.

`as_create`:

- Initialize address space data structures

`as_define_region`:

- A region is essentially a segment
- Allocate (`kmalloc`) and initialize the page table for the specified segment
- Size of the segment is a parameter of `as_define_region`
  - Because we perform preloading, segment size will never grow
  - Size of the page table is based on the segment size
- Setup the read/write permissions for this segment
- Optionally have permissions per page

`as_prepare_load`:

- Pre-allocate frames for each page in the segment
- Frames do not need to be contiguous
- Allocate each frame one at a time

`as_define_stack`:

- Always allocate `NUM_STACK_PAGES` for the stack
- Need to create a page table for the stack
- Need to allocate frames for the stack
  - `as_prepare_load` only allocates frames for segments that were defined by `load_elf`
  - Stack segment is not defined by `load_elf`

`as_copy`:

- Call `as_create` to create the address space
- Create segments based on old address space
- Allocate frames for the segments
- `memcpy` frames from the old address space to the frames in the new address space

`as_destroy`:

- Call `free_kpages` on the frames for each segment
- `kfree` the page tables

9.6 User Address/Kernel Virtual Address/Physical Address

Remember that you are always working with virtual addresses,

- only use physical addresses when loading entries in the TLB.
- virtual addresses are converted either by the TLB or by the MMU directly.

Addresses below 0x8000 0000 are user space addresses that are TLB mapped.

Addresses between 0x8000 0000 and 0xa000 0000 are kernel virtual addresses that are converted by the MMU directly, kernel virtual address − 0x8000 0000 = physical address.

`kmalloc` always return a kernel virtual address.
Do *not* use `kmalloc` to allocate frames.

10 Devices and I/O

10.1 Sys/161 Device Examples

- timer/clock — current time, timer, beep
- disk drive — persistent storage
- serial console — character input/output
- text screen — character-oriented graphics
- network interface — packet input/output

Table 7: Sys/161 timer/clock

Offset	Size	Type	Description
0	4	status	current time (seconds)
4	4	status	current time (nanoseconds)
8	4	command	restart-on-expiry
12	4	status and command	interrupt (reading clears)
16	4	status and command	countdown time (microseconds)
20	4	command	speaker (cause beeps)

Table 8: Serial Console

Offset	Size	Type	Description
0	4	command and data	character buffer
4	4	status	writeIRQ
8	4	status	readIRQ

10.2 Device Drivers

A device driver is a part of the kernel that interacts with a device.

Example. Write character to serial output device.

```
// only one writer at a time
P(output device write semaphore)
// trigger the write operation
write character to device data register
repeat {
  read writeIRQ register
} until status is "complete"
// make the device ready again
write writeIRQ register to acknowledge completion
V(output device write semaphore)
```

This illustrates *polling*: the kernel driver repeatedly checks device status

10.2.1 Using Interrupts to Avoid Polling

Device Driver Write Handler:

```
// only one writer at a time
P(output device write semaphore)
// trigger write operation
write character to device data register
```

Interrupt Handler for Serial Device:

```
// make the device ready again
write writeIRQ register to acknowledge completion
V(output device write semaphore)
```

10.2.2 Accessing Devices

Option 1: special I/O instructions

- such as `in` and `out` instructions on x86
- device registers are assigned "port" numbers
- instructions transfer data between a specified port and a CPU register

Option 2: memory-mapped I/O

- each device register has a physical memory address
- device drivers can read from or write to device registers using normal load and store instructions, as though accessing memory

10.3 MIPS/OS161 Physical Address Space

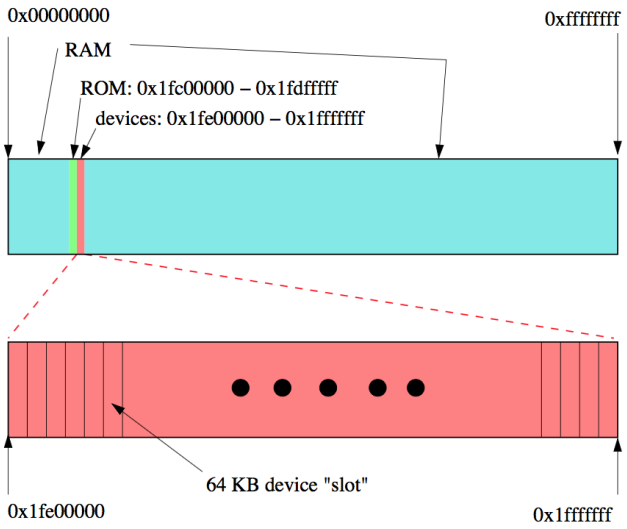


Figure 18: Each device is assigned to one of 32 64KB device "slots". A device's registers and data buffers are memory-mapped into its assigned slot.

10.4 Logical View of a Disk Drive

Disk is an array of of numbered blocks (or sectors).
Each block is the same size (e.g. 512 bytes).

Blocks are the unit of transfer between the disk and memory.
Typically, one or more contiguous blocks can be transferred in a single operation.

Storage is *non-volatile*, i.e., data persists even when the device is without power.

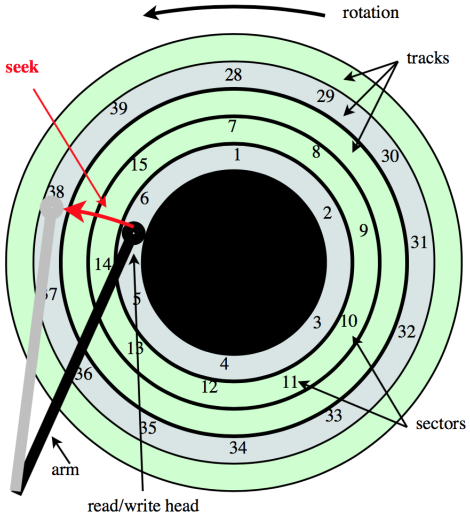


Figure 19: A disk platter's surface

10.4.1 Cost Model for Disk I/O

Moving data to/from a disk involves:

- **seek time**: move the read/write heads to the appropriate cylinder. Depends on the distance (in tracks) between previous request and current request — called the *seek distance*.
- **rotational latency**: wait until the desired sectors spin to the read/write heads. Depends on the rotational speed of the disk.
- **transfer time**: wait while the desired sectors spin past the read/write heads. Depends on the rotational speed of the disk and the amount of data being read/written.

Request service time is the *sum* of seek time, rotational latency, and transfer time.

10.4.2 Seek, Rotation, and Transfer

Seek Time:

- If the next request is for data on the same track as the previous request, seek distance and seek time will be zero.
- In the worst case, e.g., seek from outermost track to innermost track, seek time may be 10 milliseconds or more.

Rotational Latency:

- Consider a disk that spins at $12,000$ RPM.
- One complete rotation takes 5 milliseconds.

- Rotational latency ranges from 0ms to 5ms.

Transfer Time:

- Once positioned, the 12,000 RPM disk can read/write all data on a track in one rotation (5ms).
- If only x% of the track's sectors are being read/written, transfer time will be x% of the complete rotation time (5ms).

10.4.3 Performance Implications of Disk Characteristics

Larger transfers to/from a disk device are *more efficient* than smaller ones.
That is, the cost (time) per byte is smaller for larger transfers.

Sequential I/O is faster than non-sequential I/O.
Sequential I/O operations eliminate the need for (most) seeks.

10.5 Disk Head Scheduling

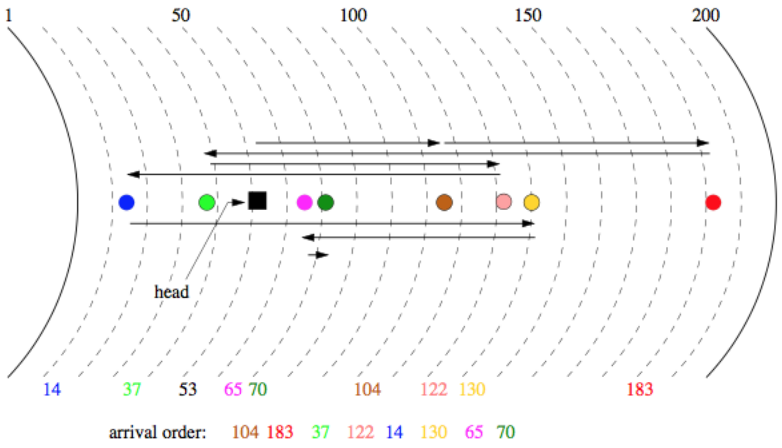
Goal: reduce seek times by controlling the order in which requests are serviced.

Disk head scheduling may be performed by the device, by the operating system, or both.

For disk head scheduling to be effective, there must be a queue of outstanding disk requests (otherwise there is nothing to reorder).

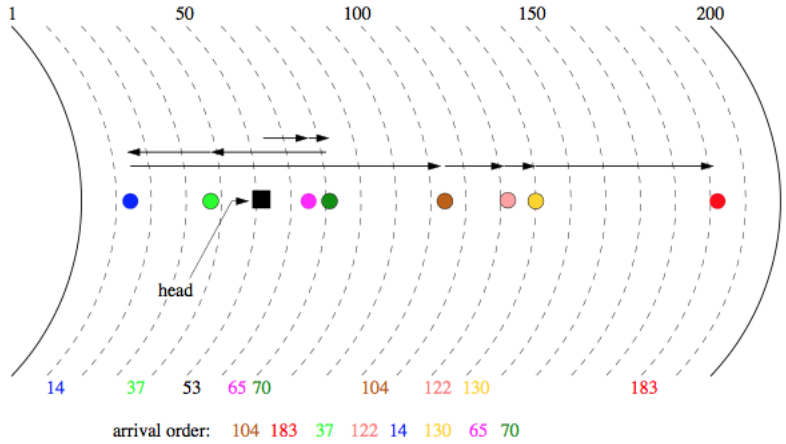
An on-line approach is required: new I/O requests may arrive at any time.

10.5.1 FCFS Disk Head Scheduling



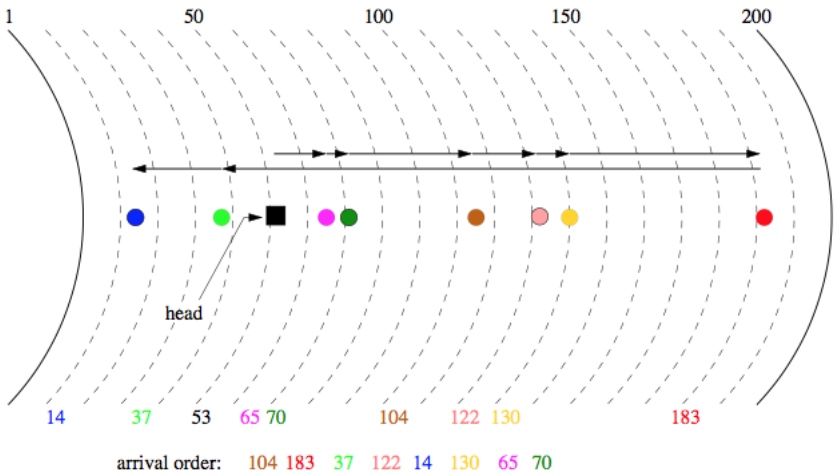
Handle requests in the order in which they arrive.
Fair and simple, but no optimization of seek times.

10.5.2 Shortest Seek Time First (SSTF)



Choose closest request (a greedy approach).
Seek times are reduced, but requests may starve.

10.5.3 Elevator Algorithms (SCAN)



Under SCAN, aka the elevator algorithm, the disk head moves in one direction until there are no more requests in front of it, then reverses direction.
There are many variations on this idea.

SCAN reduces seek times (relative to FCFS), while avoiding starvation.

10.6 Data Transfer To/From Devices

Option 1: *program-controlled I/O*
The device driver moves the data between memory and a buffer on the device.

- Simple, but the CPU is *busy* while the data is being transferred.

Option 2: *direct memory access* (DMA)

- The device itself is responsible for moving data to/from memory. CPU is *not busy* during this data transfer, and is free to do something else.

Sys/161 disks do program-controlled I/O.

Table 9: Sys/161 Disk Controller

Offset	Size	Type	Description
0	4	status	number of sectors
4	4	status and command	status
8	4	command	sector number
12	4	status	rotational speed (RPM)
32768	512	data	transfer buffer

10.6.1 Writing to a Sys/161 Disk

Device Driver Write Handler

```
// only one disk request at a time
P(disk semaphore)
copy data from memory to device transfer buffer
write target sector number to disk sector number register
write "write" command to disk status register
// wait for request to complete
P(disk completion semaphore)
V(disk semaphore)
```

Interrupt Handler for Disk Device

```
// make the device ready again
write disk status register to acknowledge completion
V(disk completion semaphore)
```

10.6.2 Reading from a Sys/161 Disk

Device Driver Read Handler

```
// only one disk request at a time
P(disk semaphore)
write target sector number to disk sector number register
write "read" command to disk status register
// wait for request to complete
P(disk completion semaphore)
copy data from device transfer buffer to memory
V(disk semaphore)
```

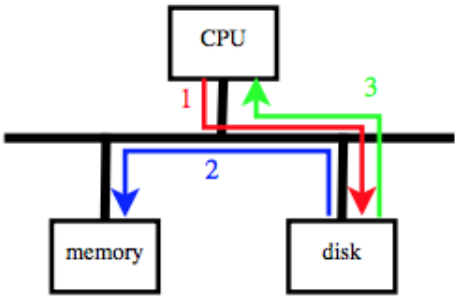
Interrupt Handler for Disk Device

```
// make the device ready again
write disk status register to acknowledge completion
V(disk completion semaphore)
```

10.7 Direct Memory Access (DMA)

DMA is used for block data transfers between devices (e.g., a disk) and memory.

Under DMA, the CPU initiates the data transfer and is notified when the transfer is finished. However, the device (not the CPU) controls the transfer itself.



1. CPU issues DMA request to device
2. Device directs data transfer
3. Device interrupts CPU on completion

11 File Systems

11.1 Files and File Systems

Files: persistent, named data objects

- data consists of a sequence of numbered bytes
- file may change size over time
- file has associated meta-data. Ex., owner, access controls, file type, creation and access timestamps

11.1.1 File Interface: Basics

open

- open returns a file identifier (or handle or descriptor), which is used in subsequent operations to identify the file
- other operations (e.g., read, write) require file descriptor as a parameter

close

- kernel tracks while file descriptors are currently valid for each process
- close invalidates a valid file descriptor

read, write, seek

- read copies data from a file into a virtual address space. `read(fileID, vaddr, length)`
- write copies data from a virtual address space into a file
- seek enables non-sequential reading/writing

get/set file meta-data

11.1.2 File Position and Seeks

Each file descriptor (open file) has an associated file position.

Read and write operations,

- start from the current file position
- update the current file position

This makes sequential file I/O easy for an application to request.

Seeks (`lseek`) are used for achieve for non-sequential file I/O,

- `lseek` changes the file position associated with a descriptor
- next read ore write from that descriptor will use the new position

11.1.3 Directories and File Names

A directory maps *files names* (strings) to *i-numbers*

- an i-number is a unique (within a file system) identifier for a file or directory
- given an i-number, the file system can find the data and meta-data for the file

Directories provide a way for applications to group related files.

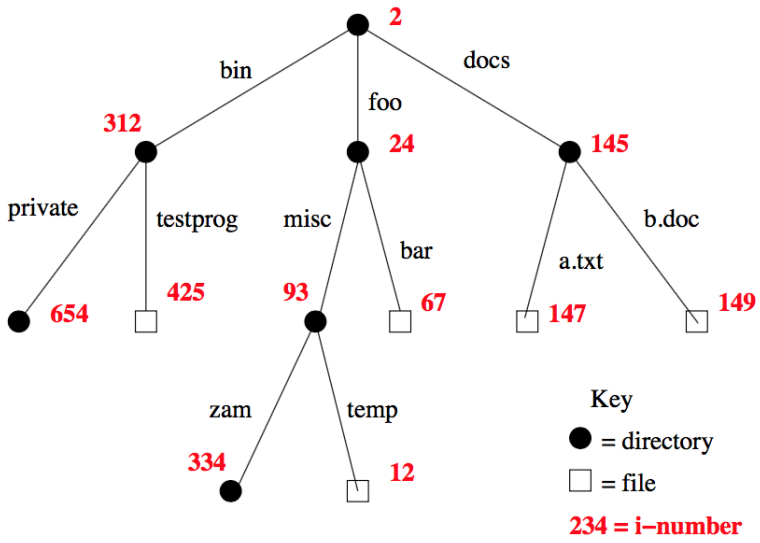


Figure 20: Hierarchical Namespace Example

Since directories can be nested, a file system's directories can be viewed as a tree, with a single *root* directory. In a directory tree, files are leaves.
Files may be identified by *path names*, which describe a path through the directory tree from the root directory to the file. Directories also have path names.
Applications refer to files using path names, not i-numbers.

11.2 Links

A *hard link* is an association between a name (string) and an i-number.
Each entry in a directory is a hard link.

When a file is created, so is a hard link to that file; ex.,

- `open(/foo/misc/biz, O_CREAT|O_TRUNC)`
- this creates a new file if a file called `/foo/misc/biz` does not already exist
- it also creates a hard link to the file in the directory `/foo/misc`

Once a file is created, *additional* hard links can be made to it.
Ex., `link(/docs/a.txt, /foo/myA)` creates a new hard link `myA` in directory `/foo`. The link refers to the i-number of file `/docs/a.txt`, which must exist.

Linking to an existing file creates a new path name for that file.
Each file has a *unique* i-number, but may have *multiple* path names.

Not possible to `link` to a directory (to avoid cycles)!

11.2.1 Unlinking

Hard links can be removed,

- `unlink(/docs/b.doc)`
- this removes the link `b.doc` from the directory `/docs`

When the last hard link to a file is removed, the file is also removed!
Since there are no links to the file, it has no path name, and can no longer to opened.

11.3 Multiple File Systems

It is not uncommon for a system to have multiple file systems.

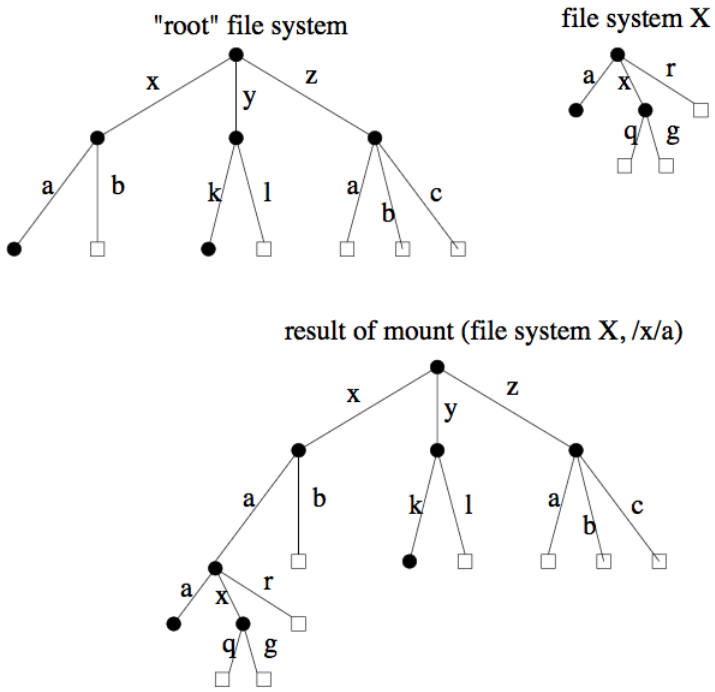
Some kind of global file namespace is required. Examples,

- **DOS/Windows**: use two-part file names: file system name, path name within file system. Example: `C:`
  `user`
  `cs350`
  `schedule.txt`
- **Unix**: create single hierarchical namespace that combines that namespaces of two file systems. Unix `mount` system call does this.

Mounting does *not* two file systems into one file system,

- it merely creates a single, hierarchical namespace that combines the namespaces of two file systems
- the new namespace is temporary — it exists only until the file system in unmounted

11.3.1 Unix mount Example



11.4 File System Implementation

What needs to be stored persistently?

- file data
- file meta-data
- directories and links
- file system meta-data

Non-persistent information,

- open files per process
- file position for each open file
- *cached* copies of persistent data

11.4.1 File System Example

Use an extremely small disk as an example:

- 256 KB disk

- Most disks have a sector size of 512 bytes. Memory is usually *byte addressable*. Disk is usually *sector addressable*.
- 512 total sectors on this disk.

Group every 8 consecutive sectors into a block.

- Better spatial locality (fewer seeks).
- Reduces the number of block pointers.
- 4 KB block is a convenient size for demand paging.
- 64 total blocks on this disk.

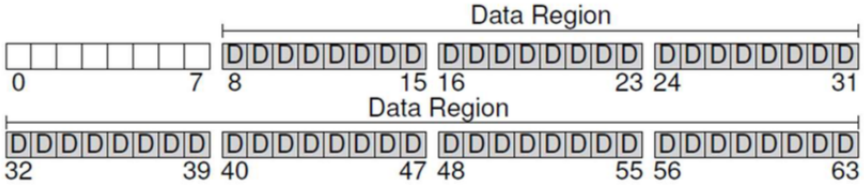
11.5 i-nodes

An i-node is a *fixed-size* index structure that holds both file meta-data and a small number of pointers to data blocks.
i-node fields may include:

- file type
- file permission
- file length
- number of file blocks
- time of last file access
- time of last i-node update, last file update
- number of hard links to this file
- direct data block pointers
- single, double, and triple indirect data block pointers

11.6 VSFS: Very Simple File System

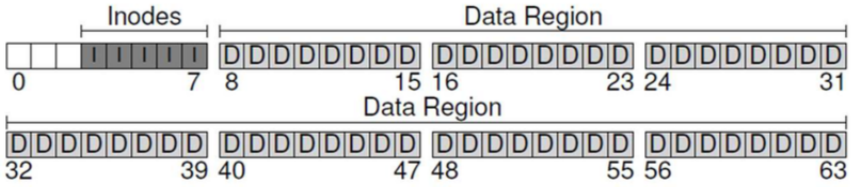
Most of the blocks should be for storing user data (last 56 blocks).



Need some way to map files to data blocks.

Create an array of i-nodes, where each i-node contains the meta-data for a file.
The index into the array is the file's index number (i-number).

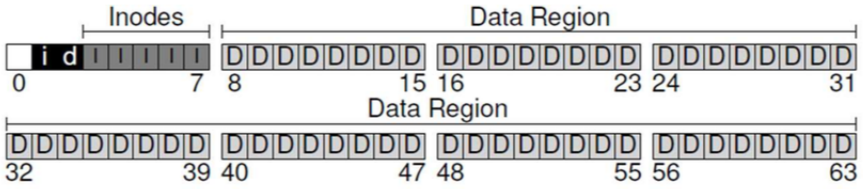
Assume each i-node is 256 bytes, and we dedicate 5 blocks for i-nodes. This allows for 80 total i-nodes/files.



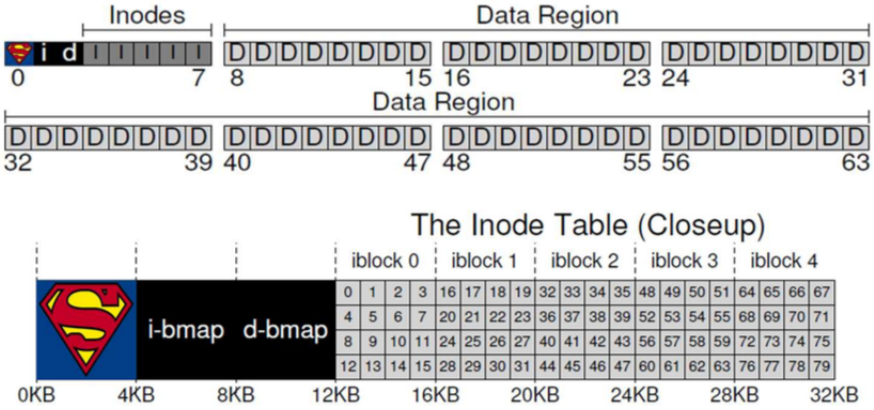
We also need to know which i-nodes and blocks are unused.
Many ways of doing this:

- in VSFS, we use a bitmap for each.
- can also use a free list instead of a bitmap.

A block size of 4 KB means we can track 32K i-nodes and 32K blocks. This is much more than actually required.



Reserve the first block as the **superblock**.
A superblock contains meta-information about the entire file system. e.g., how many i-nodes and blocks are in the system, where the i-node table begins, etc.



11.6.1 VSFS: i-node

Assume disk blocks can be referenced based on a 4 byte address.
2^{32} blocks, 4 KB blocks. Maximum disk size is 16TB.

In VSFS, an i-node is 256 bytes.
Assume there is enough room for 12 direct pointers to blocks.
Each pointer points to a different block for storing user data.
Pointers are ordered: first pointer points to the first block in the file, etc.

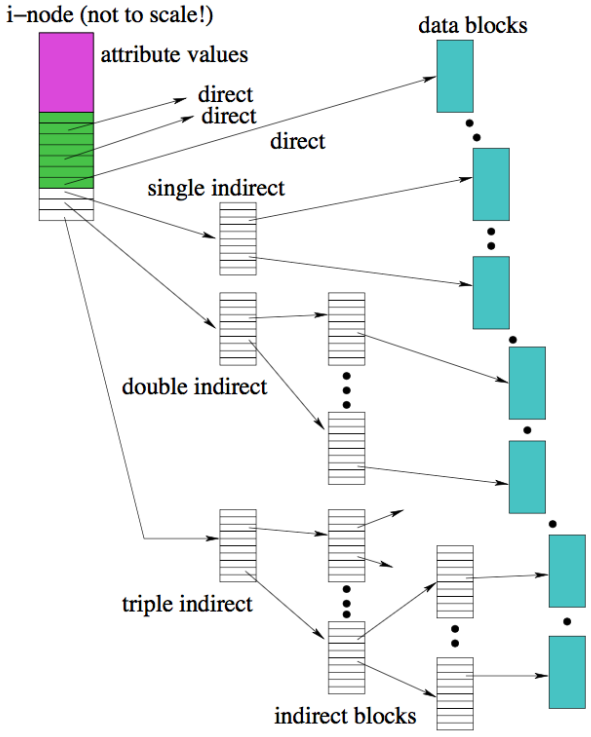
What is the maximum file size if we only have direct pointers? 12×4 KB $= 48$ KB.
Great for small files (which are common).
Not great if we want to store larger files.

11.6.2 VSFS: Indirect Blocks

In addition to 12 direct pointers, we can also introduce an **indirect pointer**.
An indirect pointer points to a block full of direct pointers.

4 KB block of direct pointers = 1024 pointers.
Maximum file size is: $(12 + 1024) \times 4$ KB $= 4144$ KB

What if the disk were larger?
Add a **double indirect pointer**, it points to a 4 KB block of indirect pointers.
Maximum file size is: $(12 + 1024 + 1024 \times 1024) \times 4$ KB (just over 4 GB).



11.7 File System Design

File system parameters:

- *How many i-nodes should a file system have?*
- *How many direct and indirect blocks should an i-node have?*
- *What is the "right" block size?*

For a general purpose file system, design it to be efficient for the common case,

Most files are small	Roughly 2K is the most common size
Average file size is growing	Almost 200K is the average
Most bytes are stored in large files	A few big files use most of the space
File systems contains lots of files	Almost 100K on average
File systems are roughly half full	Even as disks grow, file system remains ~50% full
Directories are typically small	Many have few entries; most have 20 or fewer

11.8 DIrectories

Implemented as a special type of file.
Directory file contains directory entries, each consisting of a file name (component of a path name) and the corresponding i-number.

name	i-number
.	5
..	2
foo	12
bar	13
foobar	24

Directory files can be read by application programs.
Directory files are only updated by the kernel, in response to file system operations.
Application programs cannot write directly to directory files.

11.9 In-Memory (Non-Persistent) Structures

- per process
  - descriptor table
    - which file descriptors does this process have open?
    - to which file does each open descriptor refer?
    - what is the current file position for each descriptor?
- system wide

- open file table
  - which files are currently open (by any process)?
- i-node cache
  - in-memory copies of recently-used i-nodes
- block cache
  - in-memory copies of data blocks and indirect blocks

11.10 Reading from a File (/foo/bar)

First, read the root i-node.

- At "well known" position (i-node 2)
- i-node 1 is usually for tracking bad blocks

Read the directory information from root.

- Find the i-number for foo
- Read the foo i-node

Read the directory information from foo.

- Find the i-number for bar
- Read the bar i-node

Permission check. (is the user allowed to read this file?)
Allocate a file descriptor in the per-process descriptor table.
Increment the counter for this i-number in the global open file table.
Find the block using a direct/indirect pointer and read the data.
Update the i-node with a new access time.
Update the file position in the per-process descriptor table.
Closing a file deallocates the file descriptor and decrements the counter for this i-number in the global open file table.

11.11 Problems Caused by Failures

A single logical file system operation may require several disk I/O operations.

Example. Deleting a file,

- remove entry from directory
- remove file index (i-node) form i-node table
- mark file's data blocks free in free space index

What if, due to a failure, some but not all of these changes are reflected on the disk?

- System failure will destroy in-memory file system structures.
- Persistent structures should be *crash consistent*, i.e., should be consistent when system restarts after a failure.

11.11.1 Fault Tolerance

Special-purpose consistency checkers (e.g., Unix `fsck` in Kerkeley FFS, Linux ext2),

- runs after a crash, before normal operations resume
- find and attempt to repair inconsistent file system data structures, e.g.:
  - file with no directory entry item free space that is not marked as free

Journaling (e.g., Veritas, NTFS, Linux ext3),

- record file system meta-data changes in a journal (log), so that sequences of changes can be written to disk in a single operation
- *after* changes have been journaled, update the disk data structures (*write-ahead logging*)
- after a failure, redo journaled updates in case they were not done before the failure

12 Inter-process Communications and Networking

Note: bonus material. See slides.