02 - Design Principles

CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

Defining Requirements

- Before we can implement an OS, we must identify why we want to and what requirements must be satisfied
  - Not every computer requires and operating system
    - Example: Arduino Uno can do many complex things without an OS

Defining Requirements (cont.)

- What is required of our OS?
  - What hardware should it run on?
    - There are many types of processors
    - Should we support them all or focus on one?
  - What type of system should it be?
    - Does our system need to be RTOS, multiprocessor, multitasking, etc?
  - What are the goals of our users?
    - What do our users want from our OS?
  - What are the goals of the system?
    - What do developers want from our OS?

User’s Goals

- Think about the average user, they probably want an OS to be:
  - Convenient to use
  - Easy to learn and use
  - Reliable
  - Safe
  - Fast

System Goals

- Think about YOU having to write code, you probably want to work with a system that is:
  - Easy to design, implement, maintain, and operate
  - Flexible
  - Reliable
  - Error free
  - Efficient
- Imagine writing a lot of really complex code without meeting any of these requirements, nobody would want to expand upon it

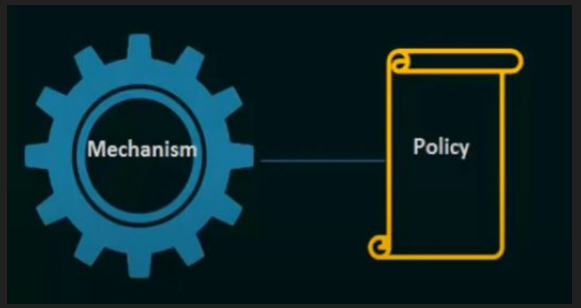
The Separation of Mechanisms and Policies

- Mechanisms determine how to do something
  - A car drives on a road by using a motor, controlled by a pedal, to spin the wheels, which moves the car
- Policies determine what will be done
  - A car driving on the road is forced to obey laws like speed limits, staying within road lines, and maintain distance from other cars

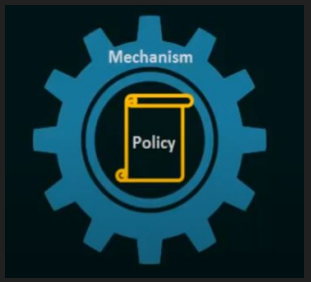
The Separation of Mechanisms and Policies (cont.)

- Mechanisms and policies must be separated
  - A car should not be allowed to change the speed limit that the other cars abide by
  - The speed limit should not be allowed to change how the car drives
- Keeping these separate is critical in working with a complex OS

Good



Bad



Separation of Mechanisms and Policies Example

- Scheduling CPU access
  - Mechanism: the CPU scheduler itself that switches processes in and out of the CPU
  - Policy: the scheduling algorithm which determines which process runs next
- The CPU scheduler (mechanism) should be able to operate with any scheduling algorithm (policy)
- The two should not be dependent or coupled with each other

Hardware Requirements

- In the early days of computing, OSs were written in assembly
  - Assembly language is unique to the CPU's architecture
  - If you write assembly language to run on x86 it will NEVER run on ARM
  - MS-DOS was written in 8088 assembly, and is only available on Intel family of CPUs
- Now most OSs are written in C or C++ with a little bit of assembly
  - Linux is mostly written in C and is available on x86, ARM, RISC-V, MIPS, etc.

Using a High Level Language

- Writing in a high level language grants you:
  - Writing code faster
  - Code is compact
  - Code is easier to understand and debug
  - Easier to port to numerous architectures

Requirements of the OS

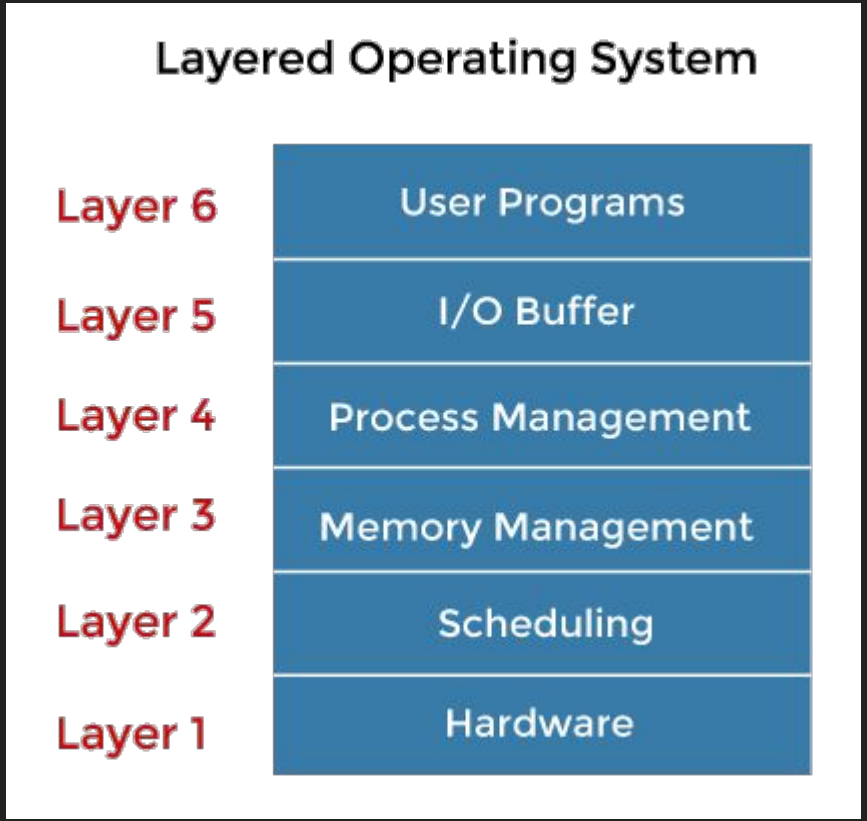
- The OS is a resource manager
  - Allocates time for programs to run on the processor
  - Allocates memory for the programs to use
  - Allocates usage of devices for programs to use, hard drive, printer, keyboard, etc.
- Do we need ensure tasks meet timing deadlines?
  - RTOS
- Do we need to ensure we can run multiple tasks?
  - Multitasking

Kernels

- The kernel is the protected part of the OS that runs in kernel mode
  - This is opposed to programs the user runs which run in user mode
- Protects the critical OS data structures and device registers from user programs

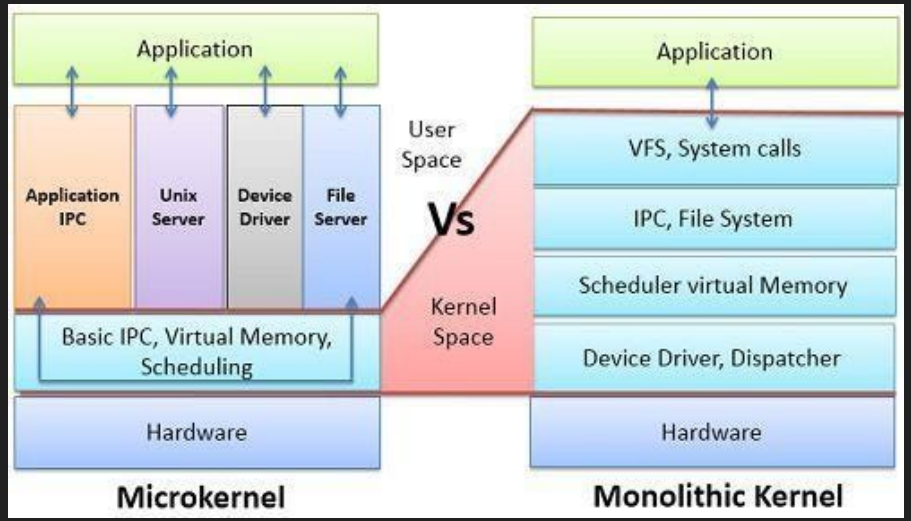
Structuring an OS

- Layered OS Structure
  - All layers exist separately to perform certain functionality
  - Modification in one layer does not affect other layers
  - Lower layers have higher privileges for accessing hardware compared to higher layers
  - Communication overhead between layers



Structuring an OS (cont.)

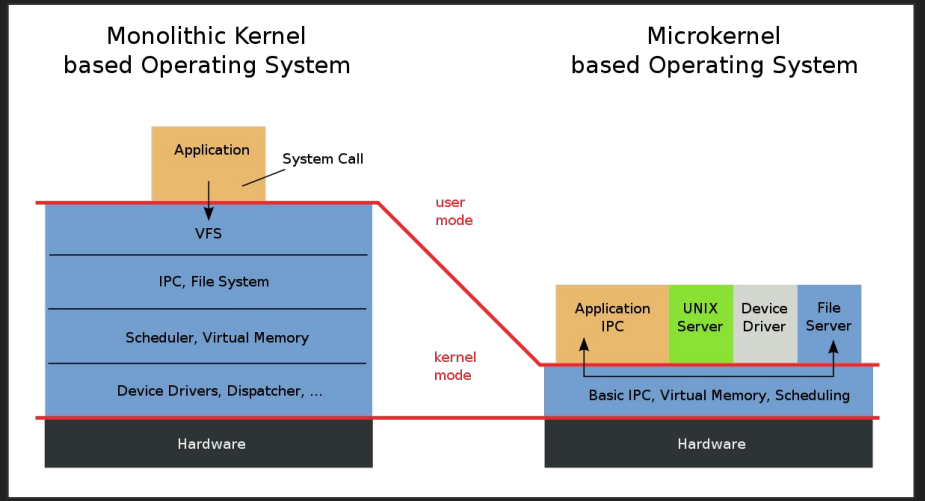
- Monolithic OS Structure
  - Entire OS is working in kernel space
  - Can be hard to write or update
  - Modules stacked on top of modules on top of modules



“Unix is structured like an onion. If you look closely at its insides, you will cry.”
- [Oscar Vivo on Twitter]

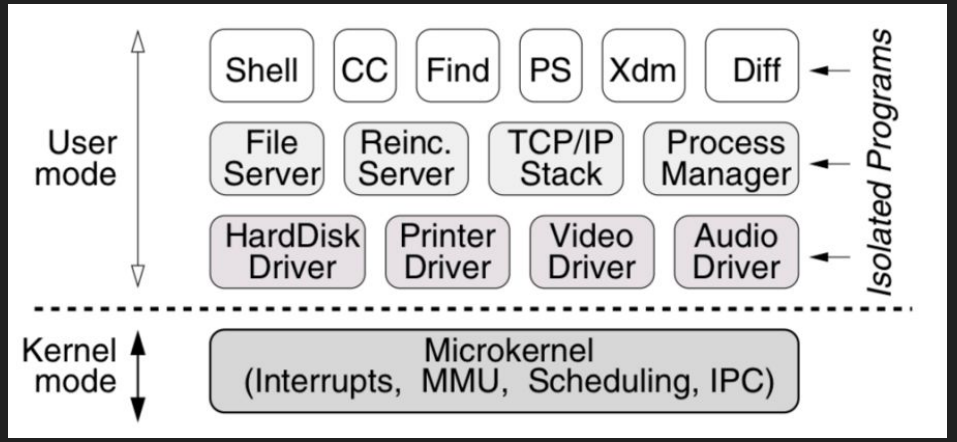
Structuring an OS (cont.)

- Microkernel OS Structure
  - Only basic functionality is provided in the core of the software system
  - Contains the required minimum amount of functions, data, and features to implement an operating system
  - Not as efficient as monolithic OS but is fast to implement



Microkernel

- A microkernel only implements the most basic tools required of the OS
  - Hardware interfaces
  - Task scheduling
- The file system, device drivers, and user programs are all ran in user mode

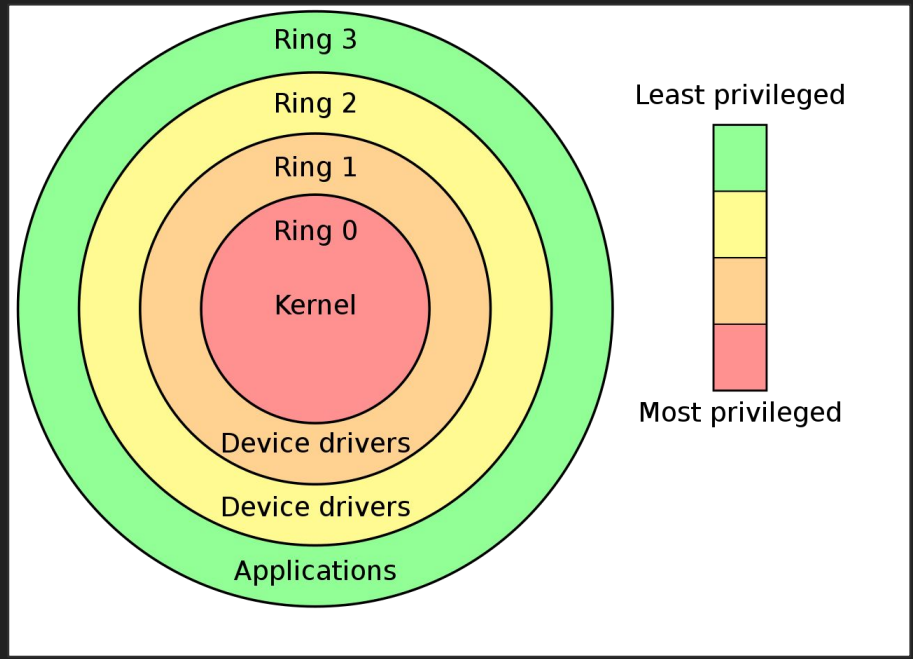


Protections

- An OS must offer certain protections
  - Protect memory
  - Protect I/O
  - Protect CPU
  - Protect user programs from each other
  - Protect the OS from user programs

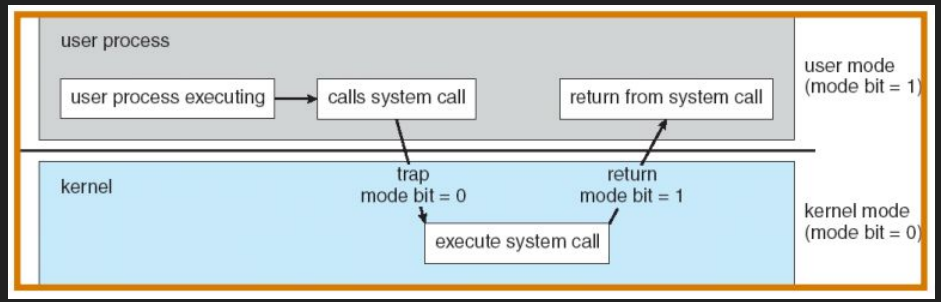
Kernel Mode vs User Mode

- Kernel Mode
  - Has most privileges for accessing memory, I/O, interrupts, special instructions, and halting the CPU
  - This gives the operating system complete control over the hardware
- User Mode
  - Has least privileges to prevent rogue programs from tampering with memory, I/O, interrupts, and halting
- Some OS implement multiple rings of protection for other purposes
- The hardware itself must have a way of enabling and disabling the protection layers usually through bits in a register



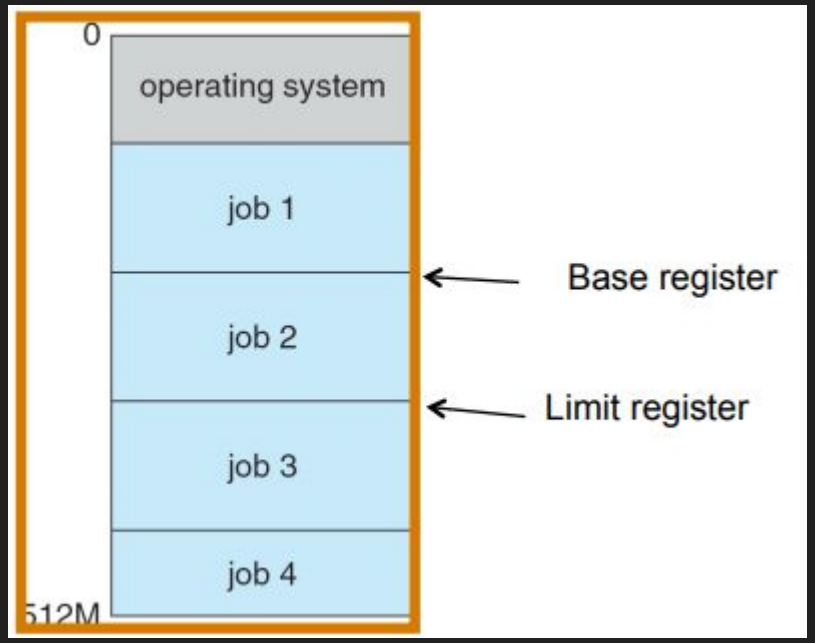
Kernel Mode vs User Mode (cont.)

- Switching between modes should be through system calls
- A system call is a mechanism used by programs to request services from the OS
  - System calls exist so the user programs can “ask” the OS if they can use sensitive resources
  - The OS determines how the system call is executed, so the OS is still in control



Memory Protection

- The OS must protect user programs from tampering with each other
- The OS must protect itself from user programs tampering with it
- The hardware must allow for checking if an instruction or data address is within a specific range (base and limit registers)
    - Example:
      base register points to:
      x0000 F000
      limit register points to:
      x0000 FFFF
      The hardware should not allow memory accesses outside of this range

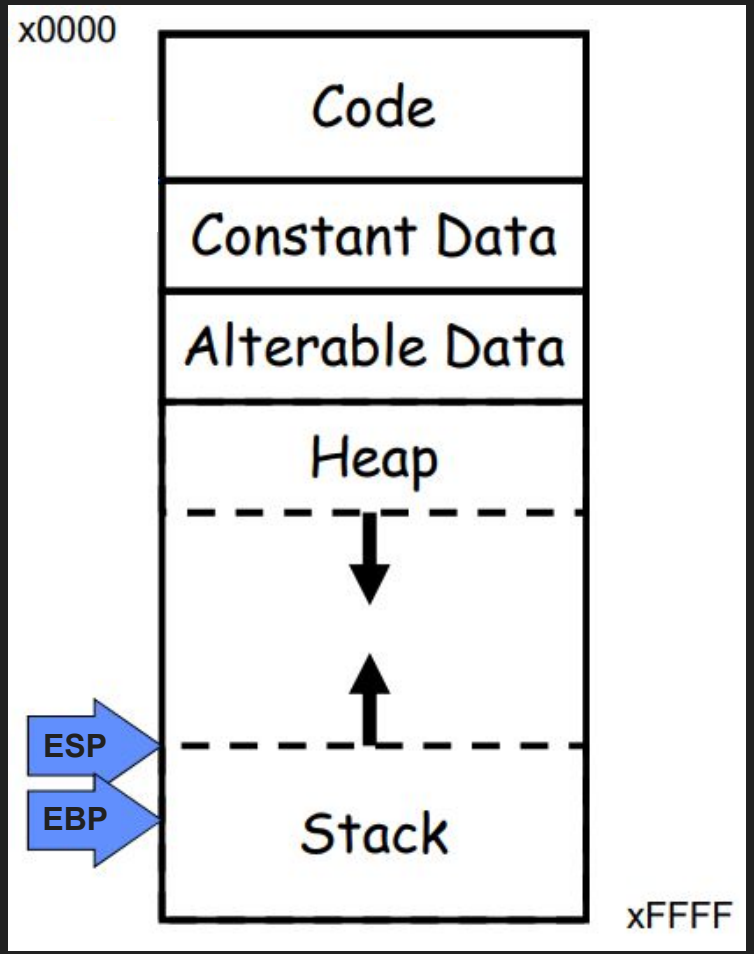


I/O Access

- Computers are not very useful without inputs and outputs (I/O)
- Interrupt based I/O
  - I/O sends an interrupt request when it needs CPU to handle it
  - I/O can be handled immediately
- Memory Mapped I/O
  - Uses a portion of the regular address space (memory addresses) for the I/O
  - I/O can be accessed by reading and writing to memory
- Port Mapped I/O
  - Uses a separate address space (not memory addresses) to identify the port address and special instructions to read from or write to the port
  - The distinction between the two addressing spaces is made in the control bus

Runtime Stack

- To manage data for multiple programs or even basic programs a runtime stack is required
- The runtime stack keeps track of all the variables and data in our program
- The stack pointer
  - (ESP or SP for 32 bit or 16 bit)
  - Used to keep track of the top of the the runtime stack
- The frame pointer
  - (EBP or BP for 32 bit or 16 bit)
  - Use to keep track of the start of variables in the current frame
  - The current frame is the function we're inside



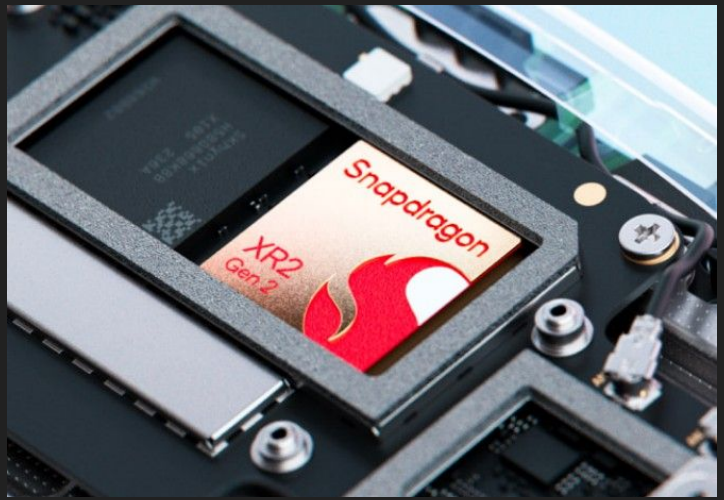
03 - Hardware and The OS

CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

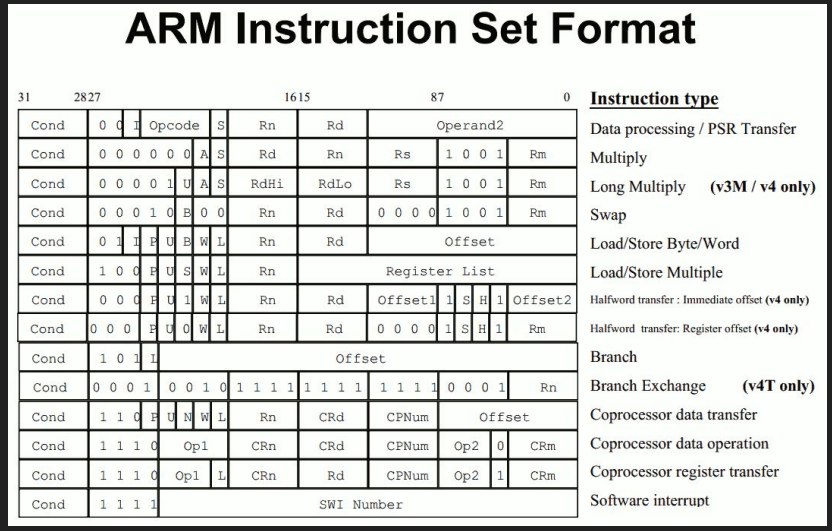
The CPU

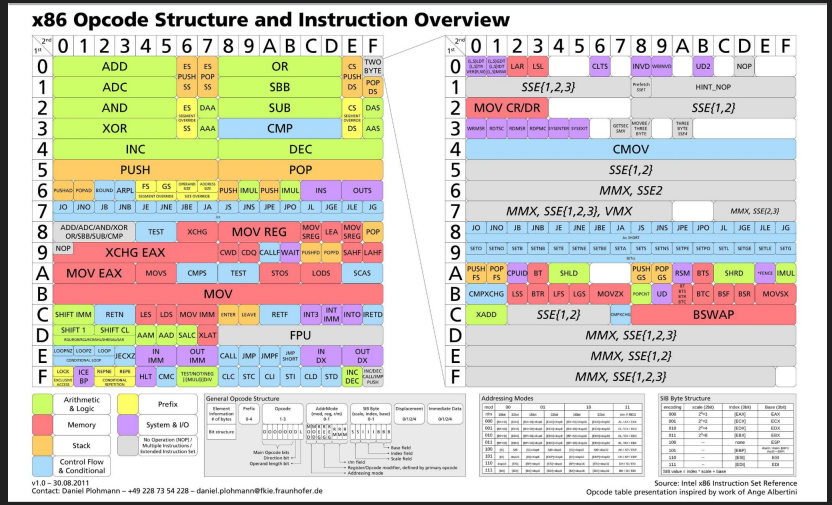
- The Central Processing Unit (CPU) executes the instructions we provide to it
- Without the CPU, the computer cannot run programs or do anything useful
- The CPU has a specific instruction set architecture (ISA) that defines what instructions it runs and how it runs them



The ISA

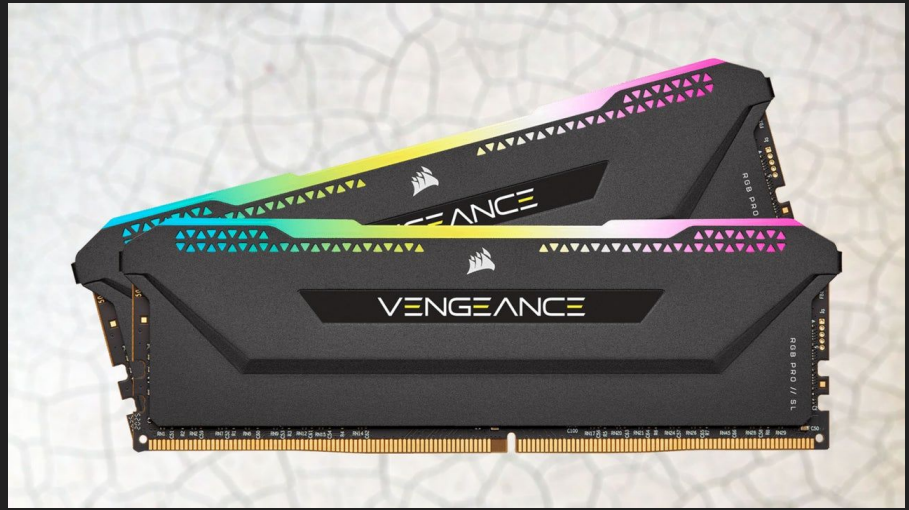
- The ISA is a “blueprint” for the tools the CPU has at its disposal
- The ISA defines what registers, instructions, and systems are available for the CPU to use
- x86, ARM, RISC-V, MIPS are examples
  - If you write code for one ISA it cannot run on another ISA without recompiling or rewriting code
  - Assembly code written for x86 will never run on ARM without a complete rewrite from scratch
  - A C program can be compiled for multiple ISAs





The RAM

- The Random Access Memory (RAM) stores our programs and data
- Without RAM, the CPU wouldn’t have instructions to execute
- Any program you want to run has to loaded into RAM



The Storage Device

- The storage device holds our files and programs
  - Hard Disk Drive (HDD)
  - Solid State Drive (SSD)
  - Floppy Disk
- Our storage device is an I/O device
- When you want to execute a program or read a file, it is copied from storage and put into RAM
- If you hear the word “disk” we are talking about these

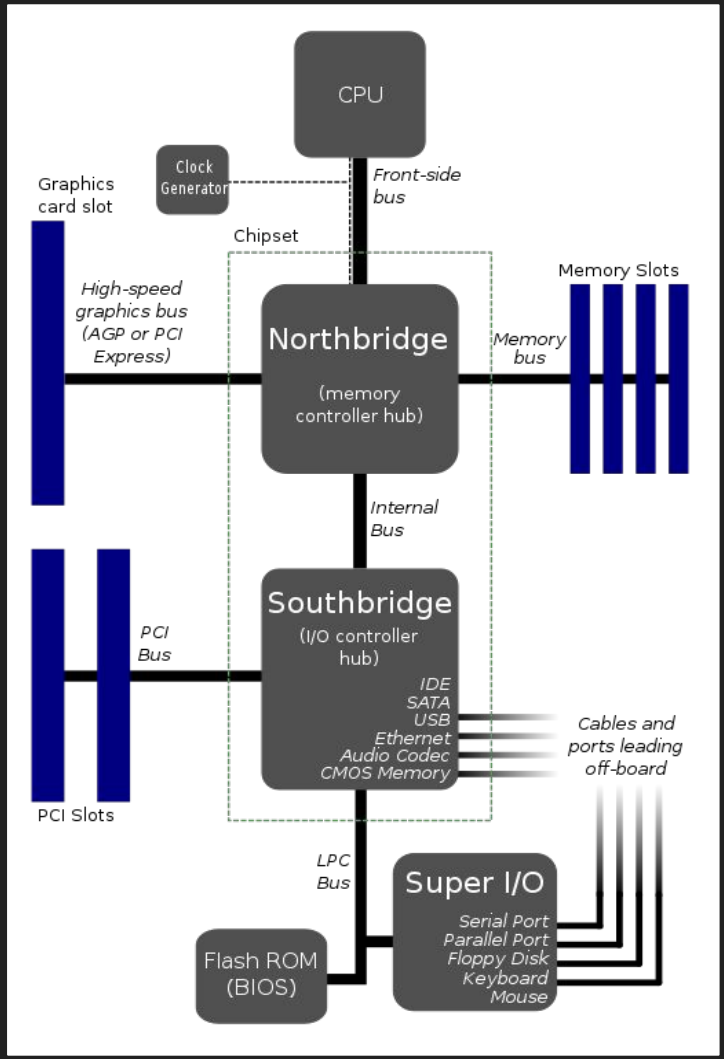
SSD

HDD

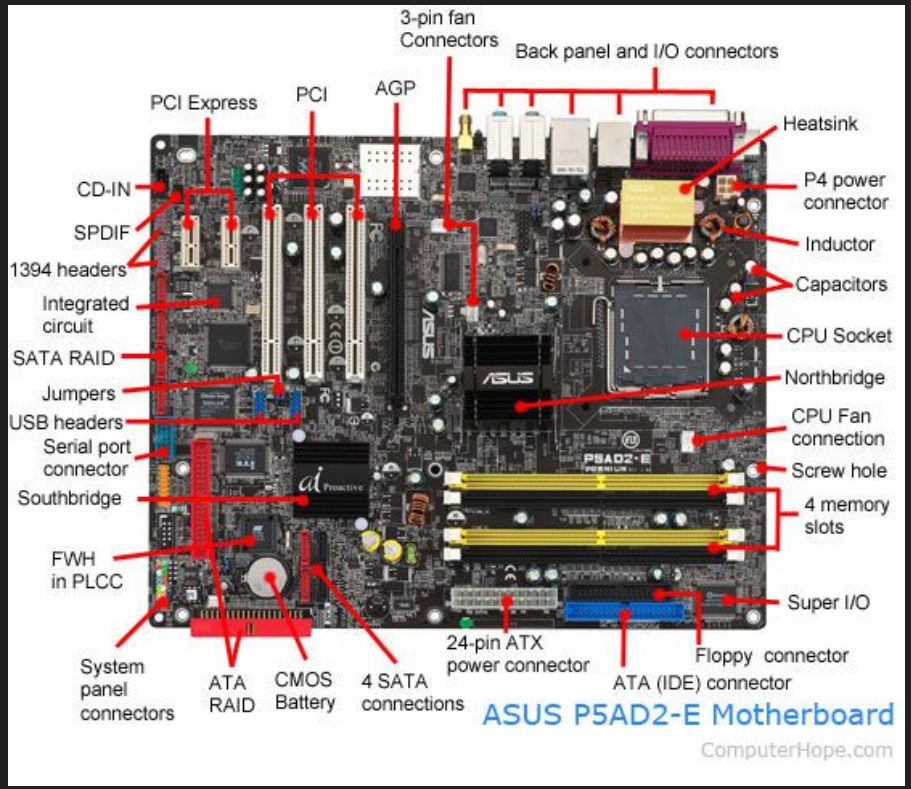
Floppy Disk

The Motherboard

- The motherboard is a physical component of a computer that connects CPU, RAM, and I/O devices together
- The CPU, RAM, and I/O all plug into the motherboard as individual modules
- There are other components on the motherboard that assist in the transfer of data between these components
  - Northbridge interconnects CPU, RAM, and Southbridge
    - Fast stuff
  - Southbridge interconnects I/O
    - Slower stuff
  - Busses are the paths for data to travel

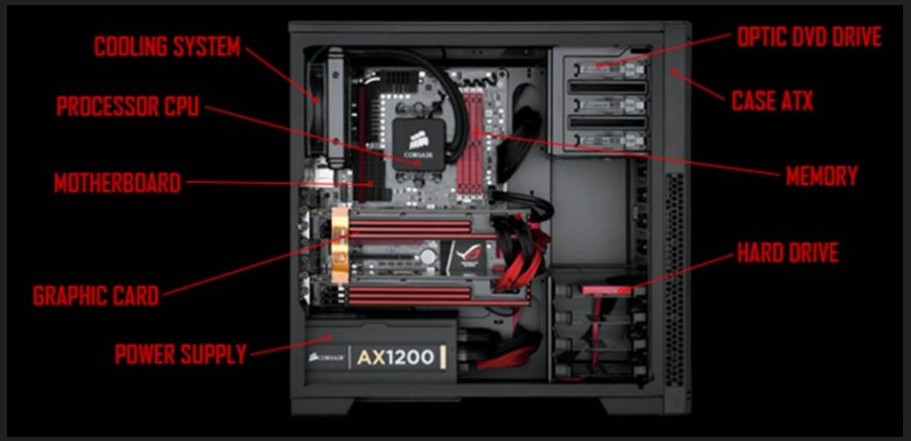


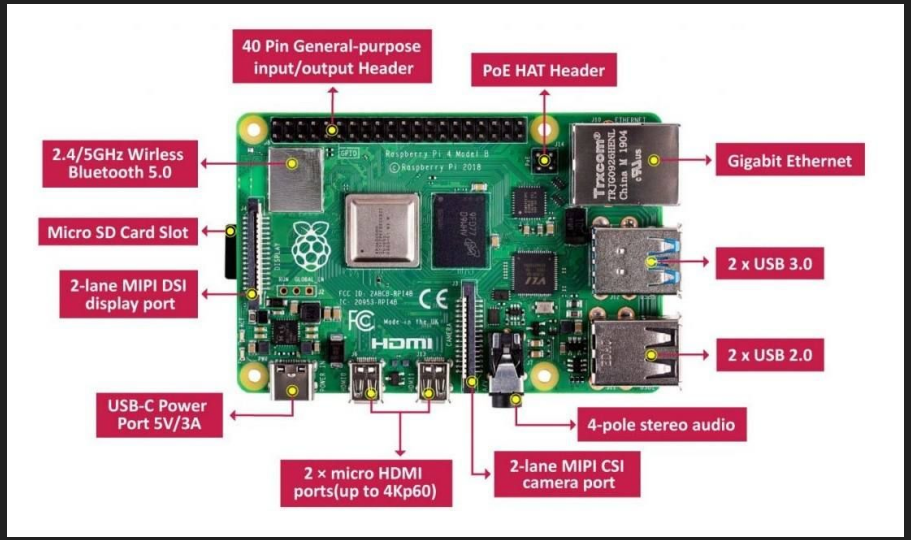
The Motherboard (cont.)



Some Extras

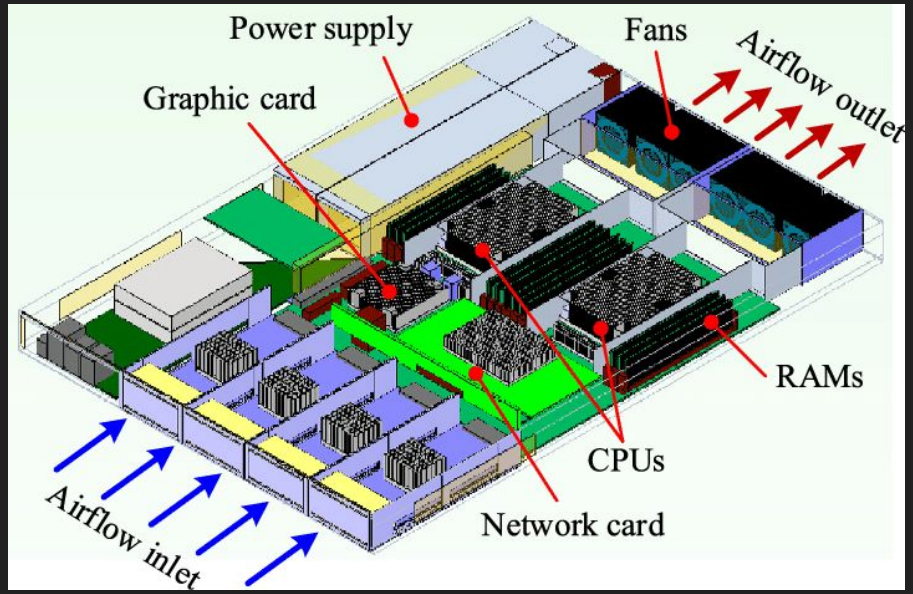
- The CPU, RAM, and Storage are our main components
- We need something to power our computer
  - Power Supply Unit (PSU)
  - Battery (embedded system)
- We *might* need some other things
  - Mouse, Keyboard, Graphics Card
  - Case
  - CPU cooler and fans
  - Network connectivity





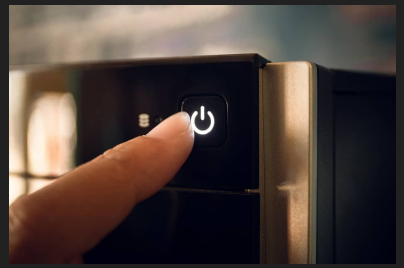
There are Many Ways To Design a Computer

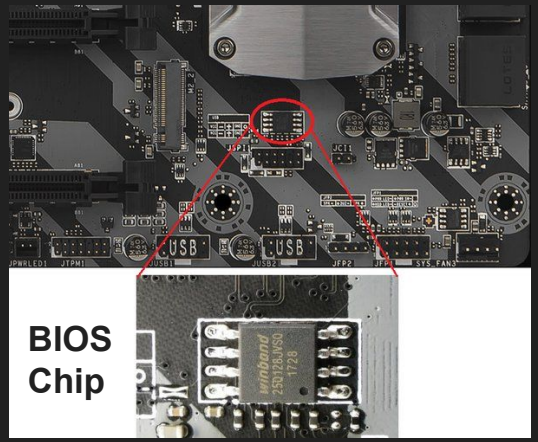
- Computers are ubiquitous and complex
- Applications for computers vary but general purpose computers can do a lot
  - Desktops/laptops
- Before trying to solve a problem, ensure you're working with the correct hardware
  - Locally running a LLM AI on a Raspberry Pi is technically possible but produces lackluster results



The Boot Process

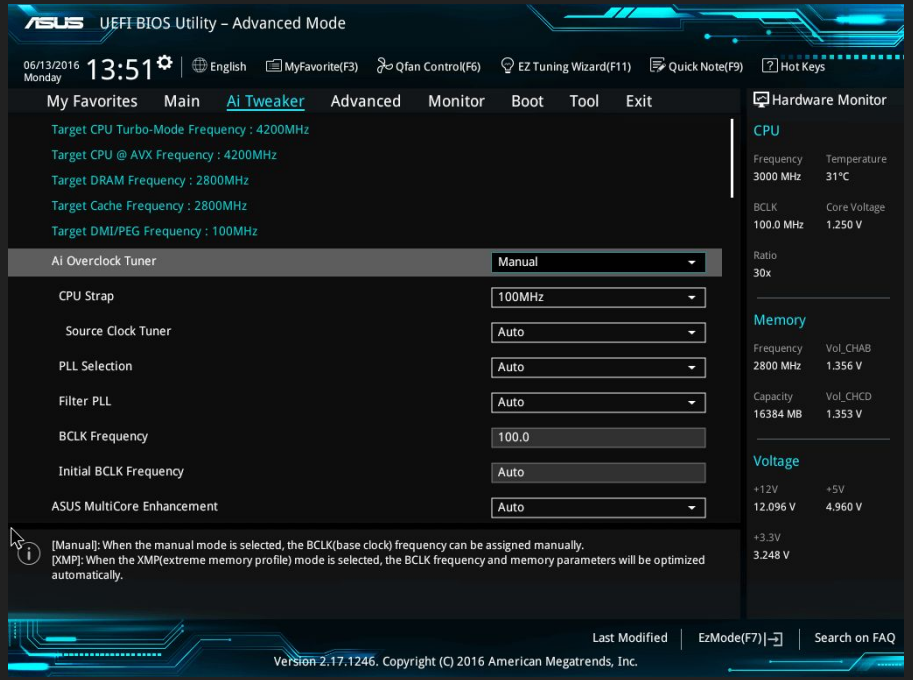
1. The power button is pressed
    a. The PSU begins sending power to the system
2. The Basic Input and Output System (BIOS) prepares the hardware and loads the bootloader from the Master Boot Record (MBR) into memory
    a. The BIOS is a program stored on a chip on the motherboard usually 16MB max size
    b. The MBR is the first 512 bytes on the storage device
    c. Power-On Self Test (POST) initializes RAM, search for storage, USB devices, performs quick tests (i.e. keyboard check), initializes video card
    d. If a bootable disk is found, start its bootloader





Unified Extensible Firmware Interface

- More modern machines use Unified Extensible Firmware Interface (UEFI) instead of BIOS
- UEFI does everything BIOS does and more
  - Provides nice GUI interface
  - Mouse support
  - Secure boot
  - Faster boot times
  - More options for configuring boot
    - Up to 128 physical partitions (BIOS has 4)



The Boot Process (cont.)

3. The bootloader stored in the MBR sets up hardware for the OS and loads the kernel into memory
   a. The bootloader is a very small (512 bytes maximum) program
   b. The bootloader must run in “Real Mode” and will switch to “Protected Mode” on x86 ISA to execute the kernel
   c. The MBR also specifies which partitions are available on the storage
   d. Sometimes, there is little bootloader code in the MBR and it simply jumps the computer’s execution to the active partition’s code
   e. The small bootloader in the MBR is the First-Stage Bootloader, the code in the main partition is the Second-Stage Bootloader
   f. Multiple partitions can be used to boot different OSs from the same storage drive

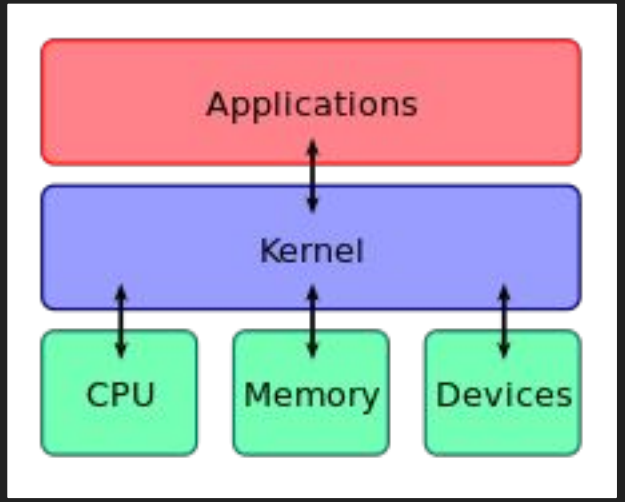
Structure of a classical generic MBR

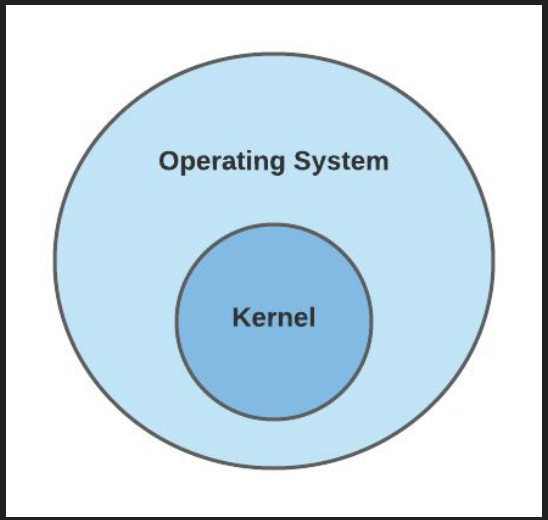
Address		Description		Size (bytes)
Hex	Dec			
000_{hex}	+0	Bootstrap code area		446
$1BE_{hex}$	+446	Partition entry №1	*Partition table* (for primary partitions)	16
$1CE_{hex}$	+462	Partition entry №2		16
$1DE_{hex}$	+478	Partition entry №3		16
$1EE_{hex}$	+494	Partition entry №4		16
$1FE_{hex}$	+510	55_{hex}	*Boot signature*[a]	2
$1FF_{hex}$	+511	AA_{hex}		
			Total size: 446 + 4×16 + 2	**512**

Element (offset)	Size	Description
0	byte	Boot indicator bit flag: 0 = no, 0x80 = bootable (or "active")
1	byte	Starting head
2	6 bits	Starting sector (Bits 6-7 are the upper two bits for the Starting Cylinder field.)
3	10 bits	Starting Cylinder
4	byte	System ID
5	byte	Ending Head
6	6 bits	Ending Sector (Bits 6-7 are the upper two bits for the ending cylinder field)
7	10 bits	Ending Cylinder
8	uint32_t	Relative Sector (to start of partition -- also equals the partition's starting LBA value)
12	uint32_t	Total Sectors in partition

The Boot Process (cont.)

4. The kernel is now running and begins to initialize and execute the core functions of the operating system
    a. The kernel connects and manages resources for the computer's operating system
    b. Task management, memory management, and I/O interfaces
    c. The kernel sits between the hardware and the user’s programs
5. The operating system is now running and takes inputs from the user and user’s programs to do complex and useful tasks
    a. Running Netflix, playing video games, crypto mining, etc.





The Shutdown Process

- When the computer is instructed to shutdown, the OS may do some things first
  - It is not a good idea for the OS to “pull the plug”
- If there is unsaved file data in RAM, it needs to be saved to storage (or it is lost)
- If there are processes that are running, the OS should request them to terminate
  - This takes a long time because the processes may be in virtual memory (waiting in slow storage) and have to finish up before being forced to terminate
- Once processes have stopped and the file systems are unloaded, the kernel sends a signal to the BIOS that turns the PSU off

The Kernel

- The kernel is the lowest level of the operating system
  - There are many components and often involves low level programming (assembly)
  - Low level programming might be necessary to interface the hardware directly
  - Kernels are usually a combination of both C and assembly
  - Even though C is a high level programming language, it gets us close enough to the hardware most of the time

Execution Modes

- On x86 based processors we have two main CPU modes:
  - Real Mode
    - 16 bit instructions, 16 bit registers, 16 bit addresses
    - CAN call BIOS interrupts
    - No safety nets or protections
    - CPU boots into this mode
    - Kept for legacy purposes
  - Protected Mode
    - 32 bit instructions, 32 bit registers, 32 bit addresses
    - CANNOT call BIOS interrupts
    - Extra protections, multitasking, virtual memory
    - The CPU must be forced into this mode from running in real mode
    - All major OSs run in protected mode
- Protected mode is enabled by setting bit 0 to “1” in register CR0
  - It’s not that simple though
  - Switching back to real mode is not easy
- There are more than these two modes but these are what we will focus on for now

Hardware Abstraction Layer

- The OS must implement a Hardware Abstraction Layer (HAL)
- The HAL is a software component that acts as an interface between the hardware and the operating system
- At its most basic, we must have a HAL that allows us to:
  - Write to display
  - Write to storage
  - Read from keyboard
  - Read from storage

VGA Text Mode

- Using VGA text mode, writing to the display is the simplest hardware interface to implement
  - VGA text mode makes things easy, otherwise we need more advanced techniques, i.e. writing drivers for graphics card 😨
  - VGA text mode is limited, we cannot display nice graphics
- To set VGA text mode the computer must be in “Real Mode”
  - Once you are still in real mode, set AH = x00 and AL = x##
  - AL should be set to your desired BIOS Video Mode found here: https://www.minuszerodegrees.net/video/bios_video_modes.htm
  - Once these values are set call the INT 0x10 interrupt
- More info on INT 0x10 interrupt (including how to disable the blinking cursor) can be found here:
  - https://en.wikipedia.org/wiki/INT_10H

Write to Display

- You can write to the display by writing to memory locations starting at 0xB8000
- Each character on the display is comprised of two bytes, first the ASCII character and second the color
  - For example, if you want to display the character “a” in monochrome green at the first character of the display, you must set memory to:
    - 0xB8000 <- 0x61
    - 0xB8001 <- 0x2A

Write to Display (cont.)

- In the previous example:
  - 0xB8000 <- 0x61
  - 0xB8001 <- 0x2A
- 0x61 is an ASCII character ‘a’
  - Available characters: https://www.asciitable.com/
- 0x2A is the color code for green text on a light green background
  - Available colors: https://wiki.osdev.org/Printing_To_Screen
    - Hint: look for “Color Number” in the “Color Table”
  - The most significant 3 bits specifies the background color
  - The least significant 4 bits specifies the text color
  - 0x2A = 0010 1010
- For light grey text on black background use 0x07
  - 0x07 = 0000 0111

Write to Display (cont.)

- In C it is very easy to write to video memory:
  - char *vidmem = (char *) 0xB8000;
    - Creates a pointer pointing to address 0xB8000
  - vidmem[0] = 0x61;
    - Sets the first location the pointer points to
    - 0xB8000 to 0x61
  - vidmem[1] = 0x2A;
    - Sets the second location the pointer points to
    - 0xB8001 to 0x2A

Reading From Keyboard

- Reading from the keyboard is not as easy as reading from a memory location
- To read from the keyboard port 0x60 and 0x64 must be read from
  - Port 0x60 is NOT a memory address
  - Port 0x60 is the Keyboard Data I/O port
  - Port 0x60 is the keyboard data port (provides a scancode of what was typed)
  - Port 0x64 is the keyboard status port (indicates if the keyboard is ready to send data)
    - The least significant bit (bit 0) will indicate keyboard readiness (1 for ready, 0 for waiting)
- There are many different types of keyboards
  - US English, UK (British) English, Chinese, Spanish, etc.
  - To ensure support for these keyboards, scancodes are used (not ASCII)
  - There is no ASCII equivalent for special keys like SHIFT or INSERT so scancodes must be used to register all the keys
- Keyboard scancodes can be found here: https://wiki.osdev.org/PS/2_Keyboard

Reading From Keyboard (cont.)

- Reading from the port must be done in assembly:

```
in al, 0x60             ; put byte from port 80 into al
```

  - The above code reads port 0x60 and puts the result into the AL register
  - If using inline assembly, the `inb` (input byte) and `inw` (input word) can be used for reading from ports
- The scancode must be converted to an ASCII character, which can be implemented with an array of ASCII characters mapped to scancode indices
- This assembly code can be called from a C program, or written in a C program using inline assembly, to create your own barebones scanf function
  - This is the preferred method to avoid having to swap between C and assembly in different .c and .asm files

Accessing Ports using Inline Assembly in C

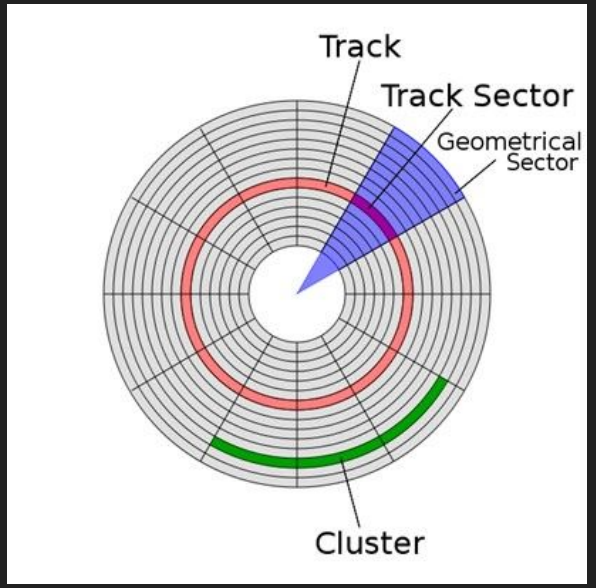
```
typedef unsigned char  uint8;
typedef unsigned short uint16;

void outb(uint16 port, uint8 value)
{
    asm volatile ("outb %1, %0" : : "dN" (port), "a" (value));
}

void outw(uint16 port, uint16 value)
{
    asm volatile ("outw %1, %0" : : "dN" (port), "a" (value));
}

uint8 inb(uint16 port)
{
   uint8 ret;
   asm volatile("inb %1, %0" : "=a" (ret) : "dN" (port));
   return ret;
}

uint16 inw(uint16 port)
{
   uint16 ret;
   asm volatile ("inw %1, %0" : "=a" (ret) : "dN" (port));
   return ret;
}
```

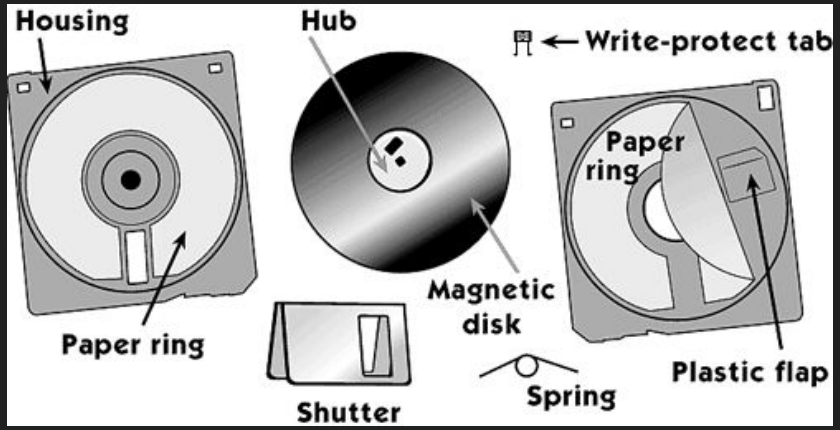
Reading from and Writing to Storage

- We will only focus on floppy disks for our storage
- Floppy disks are the easiest to interface with and are still supported to this day
  - Even though first introduced in 1967
- HDD or SSD drives are a more difficult storage to interface with
- Even though floppies are easier than HDD/SSD, it is still very complex
- For those interested in just how complicated this can get:
  - https://wiki.osdev.org/Floppy_Disk_Controller

Anatomy of a Floppy Disk

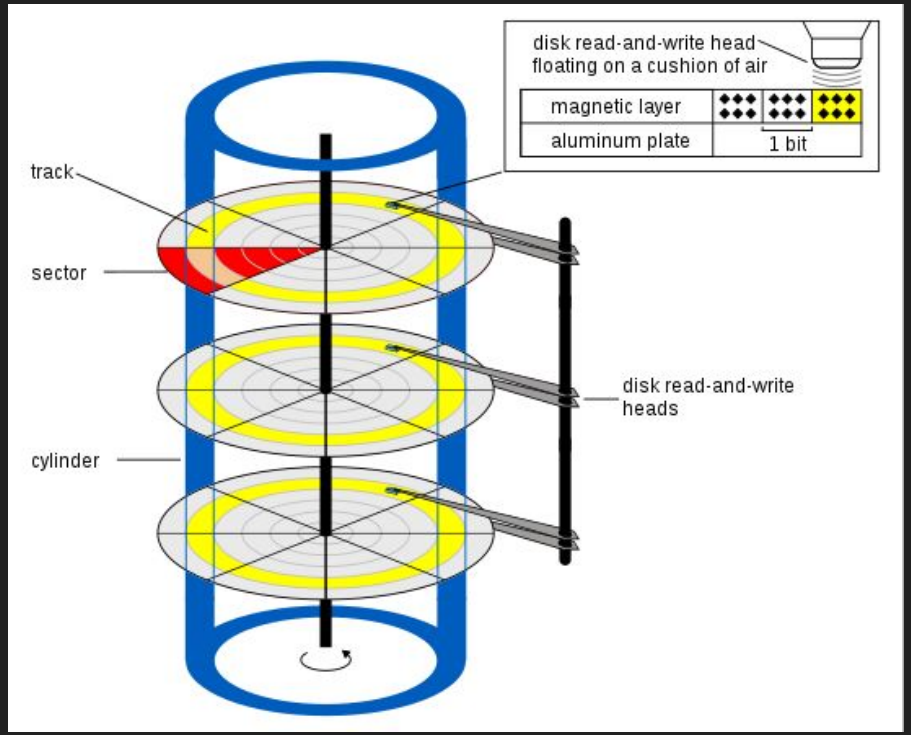
- 512 bytes per sector
- 18 sectors per track
- 80 tracks per side
- 2 heads
- Total 1,474,560 bytes per disk
- 3.5 inch 1.44 MB disks are the most common





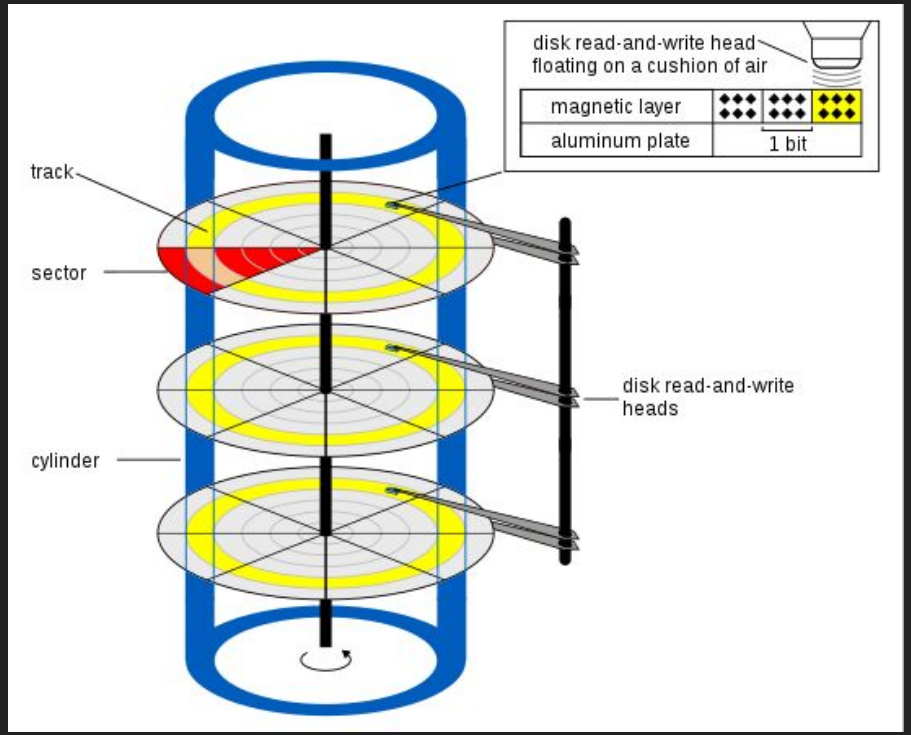
Cylinder-Head-Sector Addressing

- The cylinder-head-sector (CHS) addressing scheme is a way of accessing a specific memory location in storage
- This is an old school way of addressing and is required by INT 13,2 BIOS interrupt



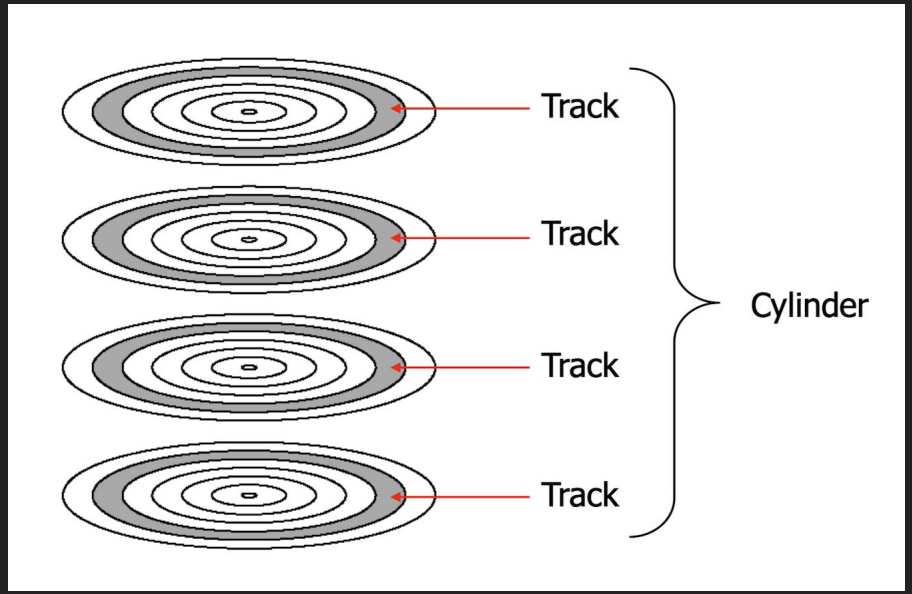
Cylinder-Head-Sector Addressing (cont.)

- To read from cylinder 0, head 0, sector 2:
  - CHS = 0,0,2
- Note: Sectors start at 1, there is NO sector 0!
  - Cylinder and head start from 0



Cylinder-Head-Sector Addressing (cont.)

- The cylinder is the aggregate of all tracks on all platters (disks)
- Cylinder count is not much of a concern today since most OSs use logical block addressing
  - Often the OS will report inaccurate numbers for the physical CHS that exists



Logical Block Addressing

- It is more intuitive to use logical block addressing (LBA)
  - We don’t care where our data is physically on the drive, we just want to access it specifically
- LBA provides a linear address space for dealing with storage
- Instead of specifying the cylinder, head, and sector, the next sector is +1 value
- To calculate the CHS value for a given LBA value use the following equations
  - Hint: We know a floppy drive has 18 sectors per track and 2 heads

$$C = (\mathrm{LBA} \,/\, \text{sectors per track}) \,/\, \text{number of heads}$$
$$H = (\mathrm{LBA} \,/\, \text{sectors per track}) \,\%\, \text{number of heads}$$
$$S = (\mathrm{LBA} \,\%\, \text{sectors per track}) + 1$$

LBA and CHS equivalence with 16 heads per cylinder

LBA value	CHS tuple
0	0, 0, 1
1	0, 0, 2
2	0, 0, 3
62	0, 0, 63
63	0, 1, 1
945	0, 15, 1
1007	0, 15, 63
1008	1, 0, 1
1070	1, 0, 63
1071	1, 1, 1
1133	1, 1, 63
1134	1, 2, 1
2015	1, 15, 63
2016	2, 0, 1
16,127	15, 15, 63
16,128	16, 0, 1
32,255	31, 15, 63
32,256	32, 0, 1
16,450,559	16319, 15, 63
16,514,063	16382, 15, 63

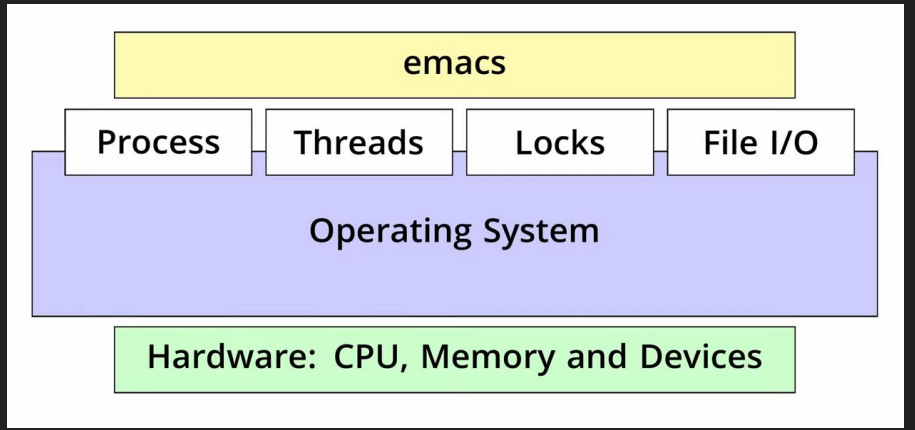
04 - Processes and Threads

CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

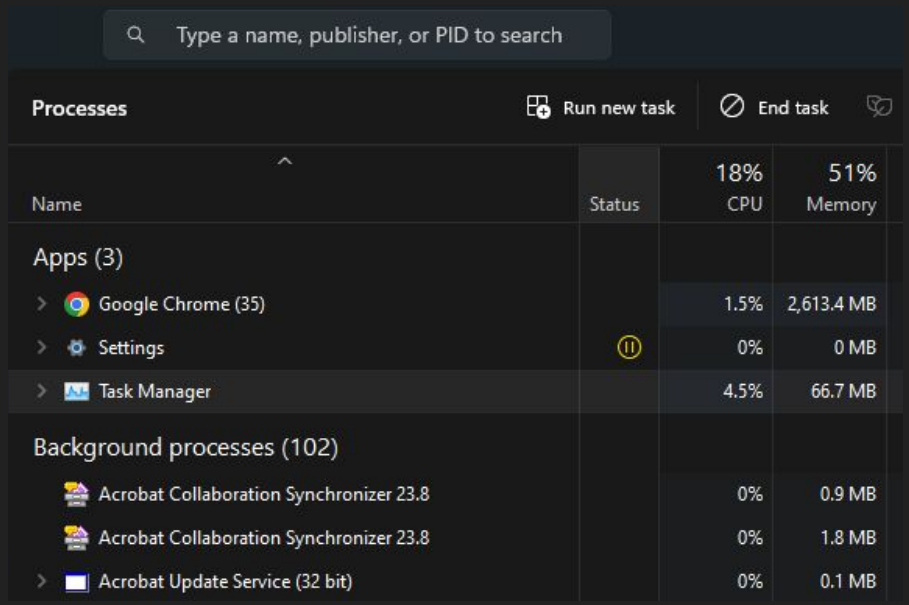
The Process Abstraction

- The OS provides abstractions for our programs
  - Abstraction is breaking down complex problems into smaller, manageable parts
  - An abstraction is basically a simplified interface
- The OS provides a process abstraction for our programs to use to get access to CPU time, memory, or other resources
- Emacs is a text editor available on Linux



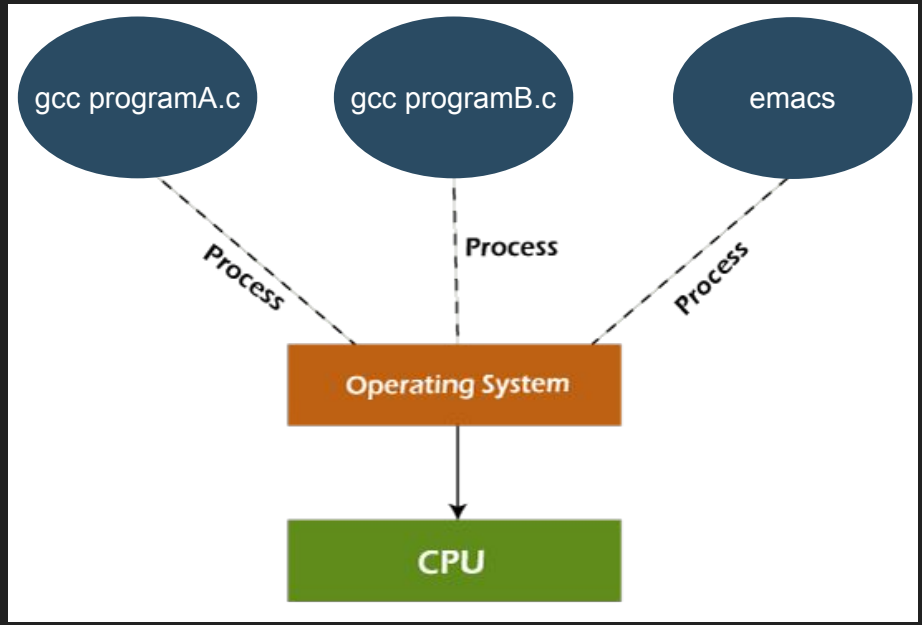
The Process Abstraction (cont.)

- A process (sometimes called a job or task) is an instance of a program running: Chrome, Fortnite, Notepad++, etc.
- Typically, multiple open windows of a program are still one process
  - Exception: Chrome creates a process for each site visited or tab opened
- Most OS can run multiple processes simultaneously
  - Notice the 35 Chrome processes ->



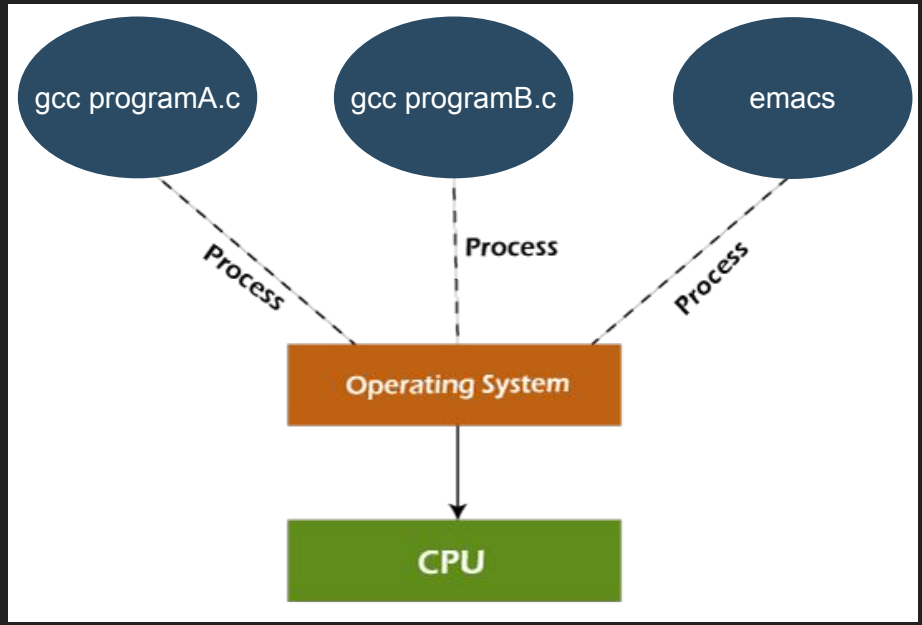
The Process Abstraction (cont.)

- The process abstraction ensures that data inside each process is unique to the process itself
  - Example: Starting an instance of gcc to compile programA.c and simultaneously starting another instance of gcc to compile programB.c
  - The two processes won't accidentally combine programA and programB together or get their data mixed up while compiling
  - The processes are kept separate



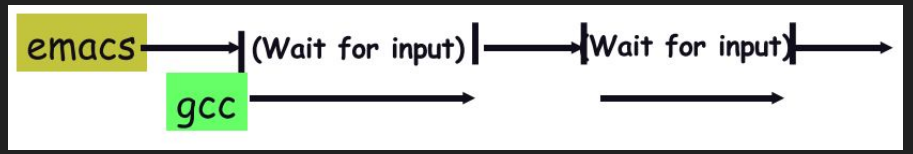
The Process Abstraction (cont.)

- This separation of processes states and data keeps our jobs as programmers “simple”
  - When gcc is compiling programA it doesn’t have to worry about any other instance of gcc that might be currently running
  - When new processes are ran the states and data are basically reset
- Running multiple processes allow us to better utilize our CPU’s computing power
  - Imagine only being able to run 1 process that spends 1 clock cycles running and 1,000,000 clock cycles waiting!



Multiple Processes Can Improve Performance

- Emacs (a text editor) spends most of its time waiting for you to type
  - So why not give up that CPU time spent waiting to a program that needs it (a compiler)

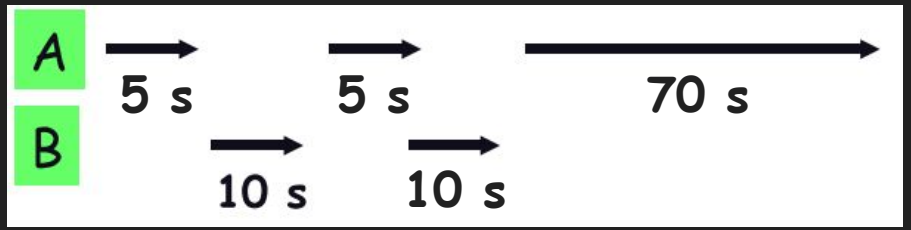


Multiple Processes Can Improve Performance (cont.)

- Old school computers could only run one process at a time
  - One right after the other
- Modern OSs allow us to switch between processes
  - From the user's perspective, the computer is running much faster!
- Note: Context switching (switching between processes) takes slightly more time than if ran one right after the other

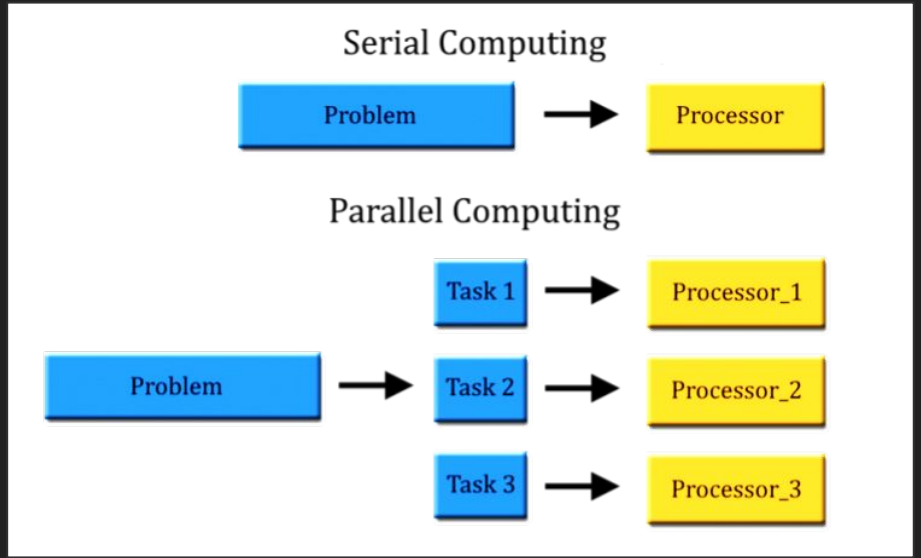
Old way:

Modern way:



Parallelism

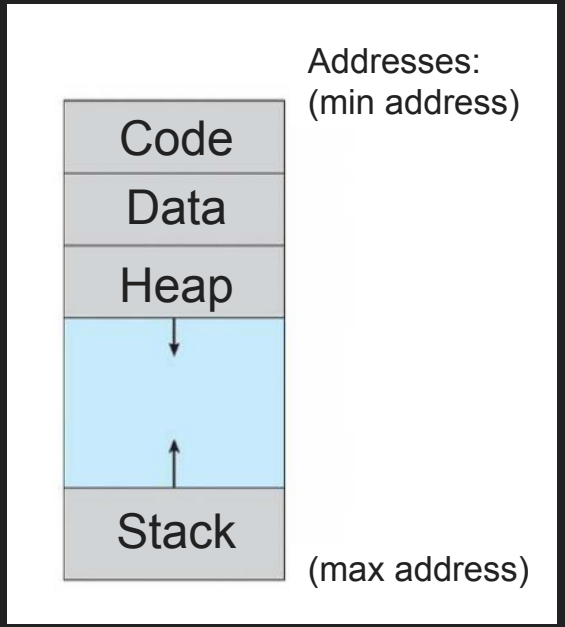
- Parallelism is the simultaneous execution of multiple tasks or processes in order to increase efficiency and speed
  - Imagine it takes a factory worker 1 day to make a product
  - If you want to make 100 products it will take this worker 100 days
  - Hire 100 workers it will take 1 day to make 100 products if they work perfectly in parallel
- Multi-core CPUs are how real parallelism is possible
- A 4 core computer can run 4 processes in parallel which yields 4x throughput



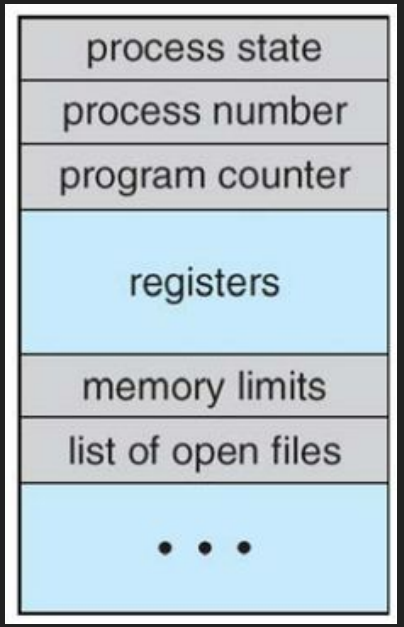
From the Process's Perspective

- Each process has its own unique view of the machine it's running on
  - Address space for code and data
  - Opened files (in memory)
  - "Virtual" CPU
    - The OS can take away CPU access whenever it wants even though the process does not know this
- One process of gcc is isolated from another process of gcc in memory
- To keep track of multiple processes information the OS uses a Process Control Block (PCB)
  - We'll take about this later

Process Memory (Process's View)

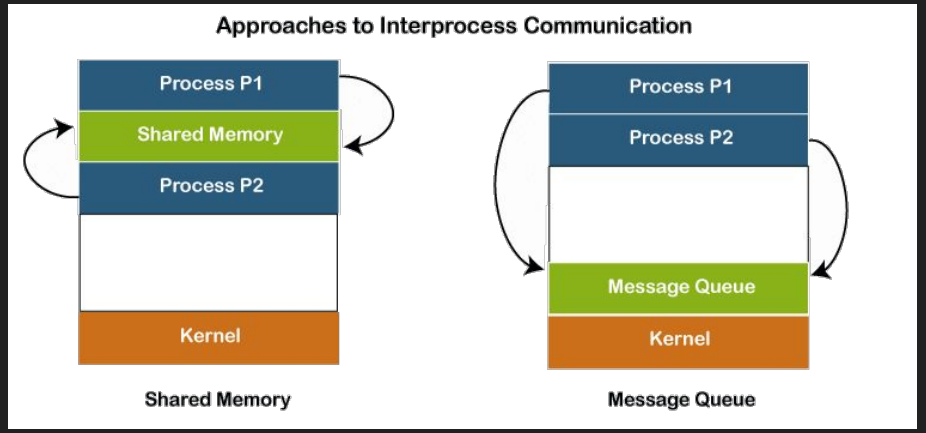


PCB (OS's View)



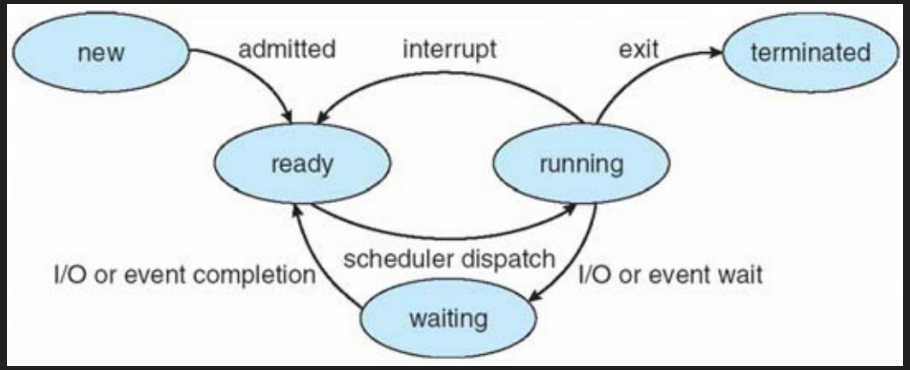
Processes Communicating

- Sometimes, we want processes to communicate with each other
- This is called Inter-Process Communication (IPC)
  - Shared files
    - i.e. edit a file with a emacs, save it, then compile that file with gcc
  - Shared memory
  - Message passing



Processes States

- New
  - A new process wants to run
  - The OS *admits* the process to the pool of other processes that want to run
- Ready
  - The process is now ready to run but must wait for the OS to dispatch it
- Running
  - The process is now executing on the CPU but it can be interrupted or forced to wait for I/O (i.e. keyboard or disk)
- Waiting
  - The process cannot run because it is waiting for something to happen
- Terminated
  - The process has finished or was forcefully terminated (i.e. closing out a window)



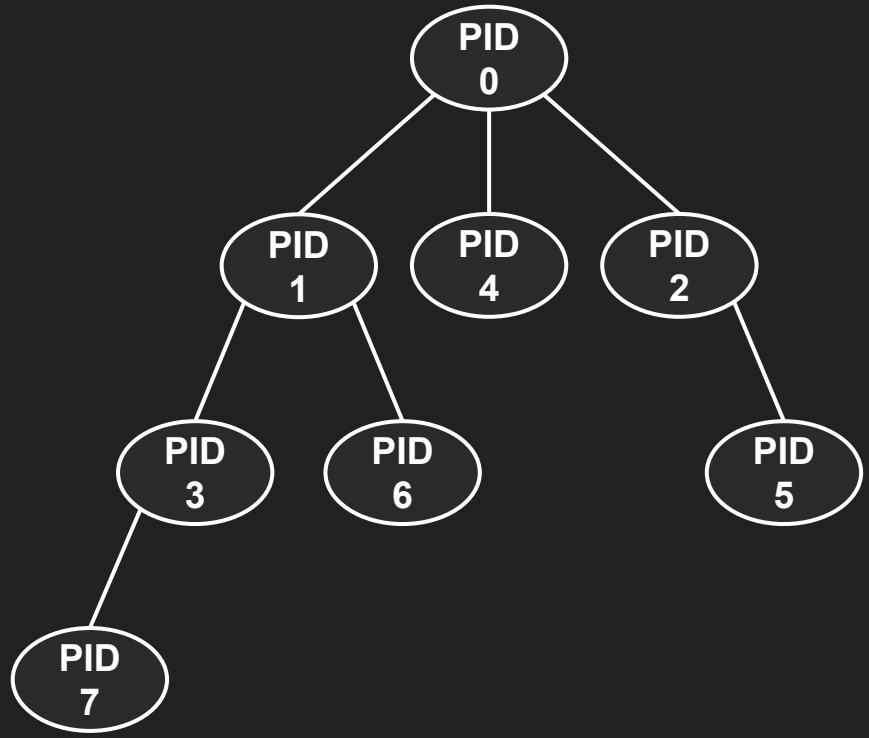
Creating a New Process in Linux

- Linux system calls make it easy to create new processes of the current program in C using the `fork()` function
- `int fork(void);`
  - Creates a process that is an exact copy of the current one that starts immediately after the fork() call
  - Returns process ID of new process to “parent”
  - Returns 0 to “child”
  - Returns negative if error

```
#include <stdio.h>
#include <sys/types.h>
#include <unistd.h>
int main()
{
    fork();
    fork();
    fork();
    printf("Hello World!\n");
    return 0;
}

// "Hello World!" will get printed
8 times, why?
```

Creating a New Process in Linux (cont.)

```
#include <stdio.h>
#include <sys/types.h>
#include <unistd.h>
int main()
{
    fork();
    fork();
    fork();
    printf("Hello World!\n");
    return 0;
}

// "Hello World!" will get
printed 8 times, why?
```



Creating a New Process in Linux (cont.)

- Your program can also wait for your child processes to finish before continuing using the `waitpid()` function
- `int waitpid(int pid, int *stat, int opt)`
  - `pid` - the process id to wait for, or -1 for any
  - `stat` - contains exit value of finished process
  - `opt` - custom options
  - Returns process ID or -1 on error

Creating a New Process in Linux (cont.)

- Sometimes you might want to execute a different program instead of creating a child process
- For this you can use `execve()`, `execvp()`, or `execlp()`
- `int execve(char *prog, char **argv, char **envp)`
  - `prog` - full path of program to run
  - `argv` - arguments the program should run with
  - `envp` - environment variables such as `PATH`, and `HOME`
  - Returns -1 if an error has occurred calling the program, i.e. program could not be found
  - The current process is now overtaken by the program we have specified
- `int execvp(char *prog, char **argv)`
  - Search `PATH` for `prog` using the current environment
- `int execlp(char *prog, char *argv, ...)`
  - List arguments one at a time, finishing with `NULL`

Terminating a Process in Linux

- You can forcefully terminate the current process (or another process) using `exit(status)` or `kill(pid, SIGTERM)`
- `void exit (int status)`
  - `status` - status code returned to `waitpid`
  - Terminates the current process
  - By convention, `status` of 0 is success, non-zero is error
- `int kill(int pid, int sig)`
  - `pid` - the process id you want to terminate/kill
  - `sig` - either `SIGTERM` (15) or `SIGKILL` (9)
  - `SIGTERM` - safely terminate the process but allow the process to do some clean up before it is forced to terminate
  - `SIGKILL` - immediately terminate the process and give no warning or time for process to clean up

Creating a New Process in Windows

- Creating a new process in Windows is not so straightforward and requires many arguments using the Windows API, `WINAPI`
- There are even multiple functions that can be called to create a process, which makes it more confusing:
  - `CreateProcess()`
  - `CreateProcessAsUser()`
  - `CreateProcessWithLogonW()`
  - `CreateProcessWithTokenW()`
  - `...`

```
BOOL WINAPI CreateProcess(
  _In_opt_     LPCTSTR lpApplicationName,
  _Inout_opt_  LPTSTR lpCommandLine,
  _In_opt_     LPSECURITY_ATTRIBUTES lpProcessAttributes,
  _In_opt_     LPSECURITY_ATTRIBUTES lpThreadAttributes,
  _In_         BOOL bInheritHandles,
  _In_         DWORD dwCreationFlags,
  _In_opt_     LPVOID lpEnvironment,
  _In_opt_     LPCTSTR lpCurrentDirectory,
  _In_         LPSTARTUPINFO lpStartupInfo,
  _Out_        LPPROCESS_INFORMATION lpProcessInformation
);
```

From the OS's Perspective

- The OS maintains a data structure for each process called a Process Control Block (PCB)
  - Sometimes called Task Control Block (TCB)
- Tracks state of process
  - Running, ready, waiting, etc.
- Includes necessary information for running
  - Current registers being used, virtual memory mappings (what's in memory?), etc.
  - Open files
- Includes other details
  - User credentials, priority, etc.

PCB
(OS's View)

Scheduling Processes

- If more than 1 process needs to run, the OS must schedule them to run individually
  - If the machine has multiple cores and is capable of parallelism, multiple processes can run simultaneously
  - Without parallelism, your computer just appears to be doing many things at the same time because it's just switching between them very quickly
- The OS will look at its list of PCBs, find all that are "Ready", and decide which one gets to run
- How should the OS decide which one gets to run if there are multiple ready processes?
  - FIFO - First process that was ready is the first to run
  - Round Robin - Arrival time, burst time, and quantum
  - Priority - Highest priority runs first

Preemptive Multitasking

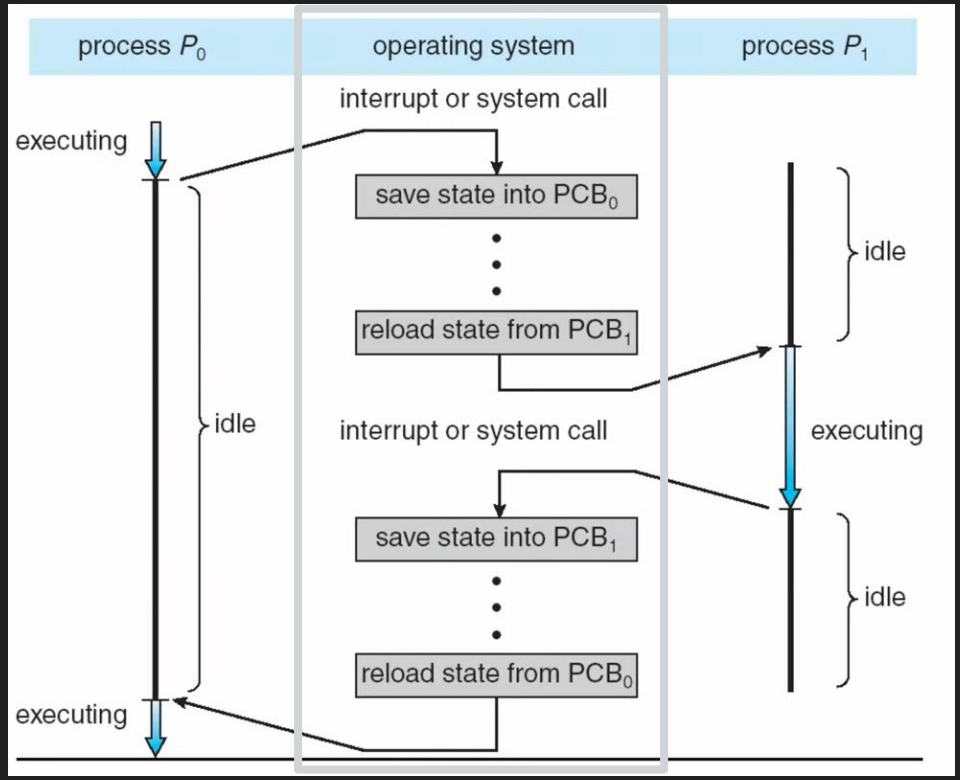
- A process can be preempted by the OS
  - Preempted means, pause for a short time, and allow something else to run, then resume later
  - Without preemption, 1 process could hog the CPU forever and your computer would appear frozen!
- The process may also preempt itself:
  - Makes a system call, waits to read disk, makes another process runnable (i.e. `fork()`), etc.
- Periodic timers interrupt the process
  - If the OS gives a fixed maximum amount of time the process can run, and the process exceeds this time, preempt the process so another process gets a chance to run
- Device interrupts
  - If process A is stuck waiting for our keyboard while process B is running, and we finally press a key, preempt process B and let the process A get a chance to capture the keyboard input
- The changeover between running processes is called *context switching*

Cooperative Multitasking

- A process must willingly yield or give up control for another task to run
  - Cooperative means every process must cooperate
  - At some point a process must give other tasks/processes a chance to run
  - The downside is if you have a process that does not play fair and hogs the CPU!
- This type of multitasking is much easier to implement because it does not involve any interrupts (timers)

Context Switch

- Switching between running processes is called context switching
- Process P_0 is currently executing but we want to run P_1:
  1. Save P_0's PCB data
  2. Reload P_1's PCB data
  3. Run P_1
- When the OS wants to switch back to P_0 the steps are the same
- If a process is not executing, it is not doing anything useful and is just waiting for the OS to give it a chance to run



Context Switch (cont.)

- Context switching is not free, it costs CPU time and memory (to store and access PCB data)
- Context switching is very hardware dependent
  - Which registers should get saved in the PCB and restored when we run the process?
    - What about floating point or special registers?
  - Are there flags that must be maintained?
  - Save/restore memory translations

Implementing Basic Multiprocessing in C

- To allow your OS to run multiple processes, we need to implement multiprocessing
  - Also known as multitasking
- Our OS will switch between multiple processes and execute each of them individually
- How does our system know when it needs to switch between processes?
  - Preemptive - hard to implement but better solution
  - Cooperative - easy to implement but worse solution

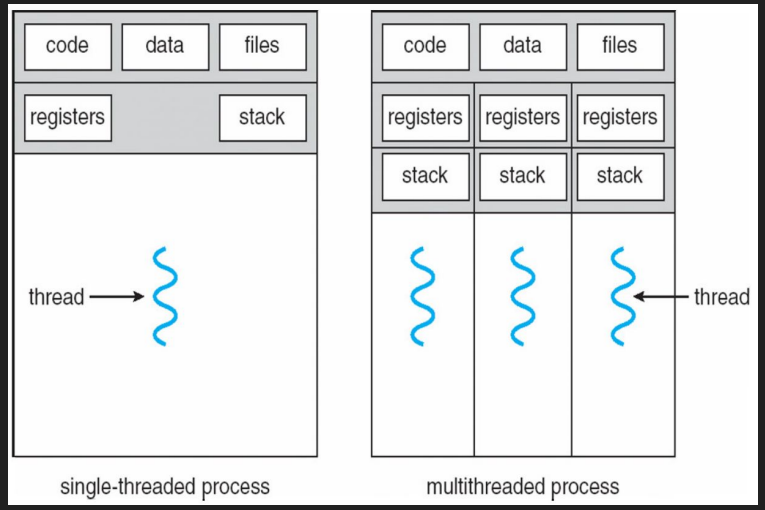
Implementing Basic Multiprocessing in C (cont.)

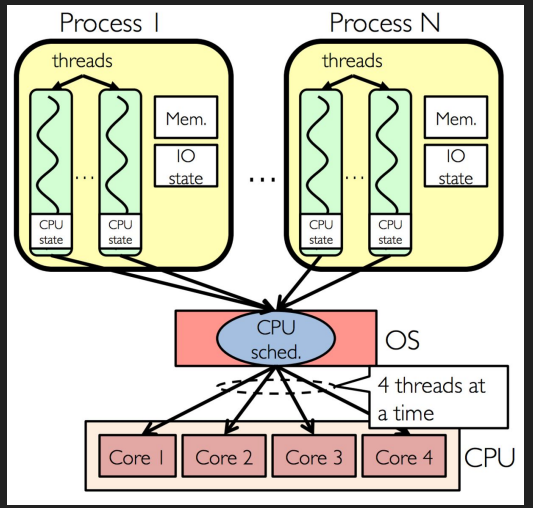
- To implement cooperative multitasking we must:
  1. Write a set of tasks (functions) that we want to execute
  2. Write a scheduler that can schedule tasks from a maintained list of PCBs
  3. Pass the tasks to a scheduler that will maintain PCBs for each task passed to it and construct a sequence of these tasks to execute them in order
  4. Write a `yield()` function that will save the state of the current task in the PCB in the scheduler and resume the state or start the next task in the scheduler
     a. Make sure to put the `yield()` function in each task (function) so it can give time to another task at some point
     b. Make sure to write an `exit()` function in each task so it can communicate to the scheduler when it should be removed from the task list
  5. The scheduler should loop through the list of PCBs until they have all exited

The remaining slides focus on ***threads*** not ***processes***

Threads

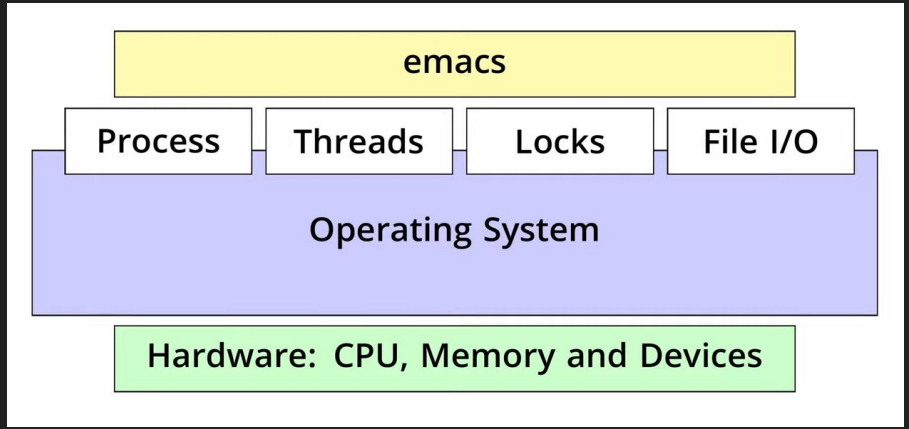
- A thread is a schedulable execution context
  - An execution context is all the information a CPU needs to execute a stream of instructions
- A process may have 1 or more threads
- Multithreaded processes share the same address space
  - All the threads share code, data, and any open files
  - Each thread has its own registers and stack
- Threads can be executed simultaneously by using CPU *cores*
  - A core is the CPU hardware
  - A thread is the instructions/data provided to the core





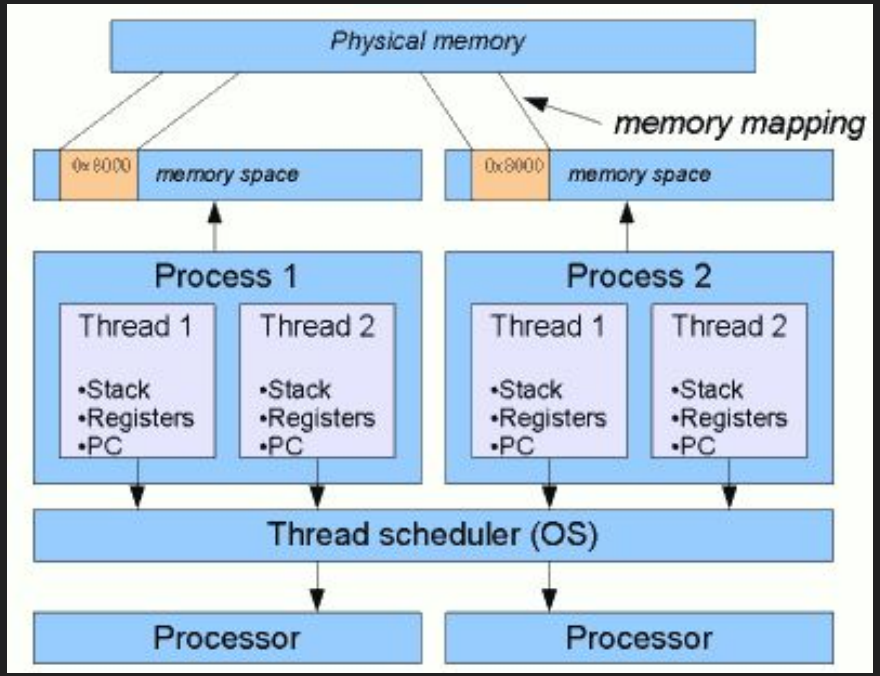
The Thread Abstraction

- Why separate the thread abstraction from the process abstraction?
  - What if you have 4 processes that want 4 threads?
  - What if you have 1 process that wants 4 threads?
- Keeping the threads as a separate system from processes allows for more flexibility
- The kernel usually has its own thread internally for every user mode thread or process
  - This internal thread keeps an eye on the user's processes/threads
  - Also has threads for every user currently logged into the system
- Just like processes, threads must be scheduled by the OS *or* by the process



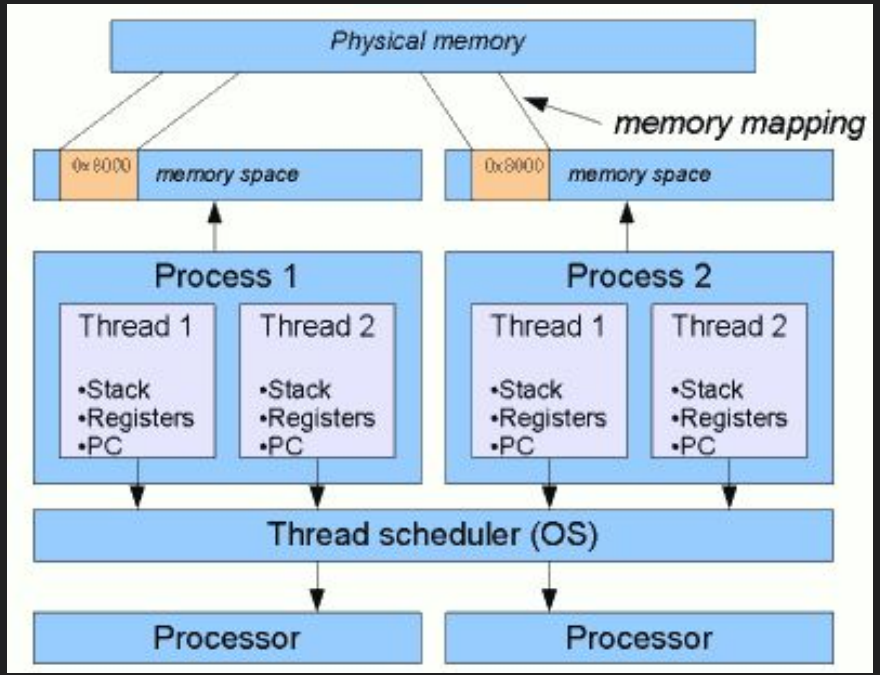
The Thread Abstraction (cont.)

- Threads are great for concurrency
  - Concurrency is executing multiple computations at the same time
  - This requires multiple physical CPU cores
- Threads are lighter-weight than processes
  - Threads share memory, files, etc.
  - Threads are easier to assign rather than spinning up a new process for each execution context
- Threads allow a process to do many things simultaneously
  - Execute code WHILE reading from a file, keyboard, etc.
  - I/O is really slow so this is very beneficial
  - Execute multiple pieces of code for faster results (physics simulations, games, etc.)
- The OS or user processes can create as much threads as they want and are not bound by how many physical cores exist



Multiprocess ≠ Multithreaded

- Threads and processes are two different things!
  - An OS can be multiprocess without being multithreaded and vice versa
- A multiprocess OS can run or switch between multiple processes
- A multithreaded OS can allow processes to use multiple threads
  - As long as there are CPU cores to use



Portable Operating System Interface

- The Portable Operating System Interface (POSIX) is a set of standards used to maintain compatibility between OSs
- POSIX standards allow for many OSs to use standardized methods for threads, I/O, file operations, and other OS functions
- The goal of POSIX is to define both the kernel and user-level application programming interfaces (APIs), along with command line shells and utility interfaces
    - This allows for software to be ported easier to other variants of Unix and other operating systems

Threads in POSIX

- `int pthread_create(pthread_t *thr, pthread_attr_t *attr, void *(*fn)(void *), void *arg);`
  - Create a new thread `thr`, with attributes `attr`, to run a function `fn`, using arguments `arg`
  - If the syntax for `fn` looks confusing, just know it is a pointer to a function, that accepts a pointer as its arguments, and a pointer as its return value, each are void to allow the developer flexibility to choose any type they want to return or pass as arguments
- `void pthread_exit(void *return_value);`
  - Exit or terminate the current thread and return a value `return_value`
- `int pthread_join(pthread_t thread, void **return_value);`
  - Wait for thread `thread`, to exit, and capture its return value `return_value`
- `void pthread_yield();`
  - Tell the OS to allow other threads to execute and pause this thread
  - Helpful when this thread needs to wait for something (like I/O)
- There are more APIs that allow for thread synchronization and other benefits but this is all we care about for now

Implementing POSIX Compliant Threads

- The kernel can implement thread creation using a system call and maintains user threads with kernel threads
  - Every process that wants a thread must use this system call
  - The OS has the final say on whether or not a process gets a thread and controls access to threads
- Implement `pthread_create`, and other thread APIs as a system call
- Create the process abstraction in kernel
- Allow processes to utilize `pthread_create`
- When a process calls `pthread_create`, create a new thread that uses the same address space, file table, code, as the process
- Assigns one kernel thread to this user thread using one-to-one thread model

Kernel Threads vs User Threads

- There are two levels of threads that we use in an OS
- Kernel Threads
  - Managed and scheduled within the kernel
  - Has direct access to the CPU cores
- User Threads
  - Managed and scheduled within the user program
  - Allows multiple parts of the user program to execute simultaneously
- In order for a process to execute, there must exist a relationship between user threads and kernel threads

Contention Scope

- Contention scope is the level at which contention for resources occurs between threads
  - This can be in user space and/or kernel space
- There are two methods for scheduling threads:
  - Process-Contention Scope (PCS)
  - System-Contention Scope (SCS)

Process-Contention Scope

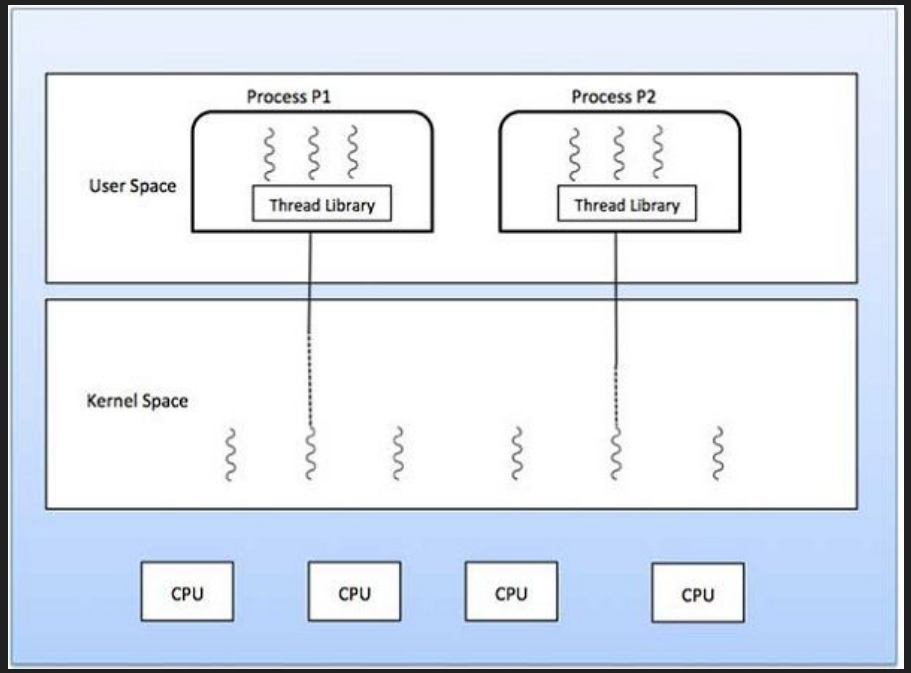
- In PCS, threads within the process compete with each other
- The scheduling mechanism for the thread is local to the process
  - The thread's library has full control over which thread will be scheduled
  - This is typically handled by the programmer assigning priorities to the threads when writing the multithreaded program

System-Contention Scope

- In SCS, threads within the kernel compete with each other
- The scheduling mechanism for the thread is within the OS
  - The OS determines, out of all kernel threads, which get to execute on the CPU(s)
- Many modern OSs only allow for SCS scheduling and the one-to-one thread model

Many-To-One Thread Model

- Each user-level thread is mapped to a single kernel-level thread per process
  - The operating system handles multiple threads as a single task
- This is implemented using a user level library rather than a system call
- If one user thread blocks, the entire kernel thread will become blocked
  - This prevents the other threads from getting a chance to execute
- Lacks true parallelism, the single kernel thread can only execute the user threads on 1 CPU



Implementing User Level Threads with Many-To-One

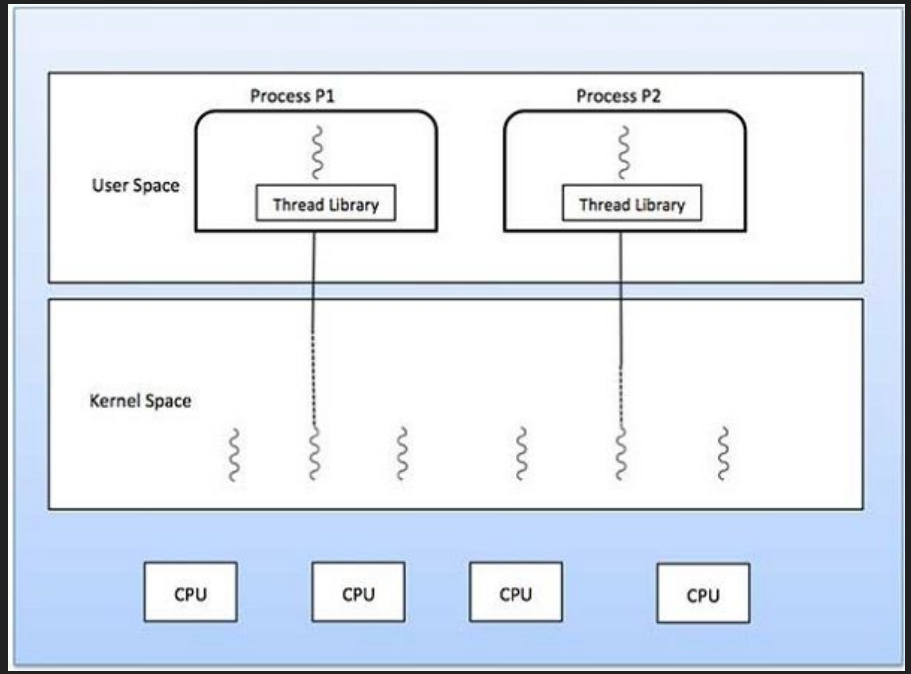
- In Many-To-One the user level threads are required to do a few things:
  - Allocate a new stack for each thread
  - Maintain a queue of *runnable* threads
    - Threads not waiting for I/O or other system calls
- A thread scheduler (running in user space) selects the next runnable thread from the queue and gives it to the kernel thread to execute
- When we want to execute a function using these threads we need to:
  - Write non-blocking versions of any blocking system calls so the thread can yield temporarily, a different thread can run, then resume this thread when done waiting (for I/O, network, storage, etc.)

Problems With User Level Threads in Many-to-One

- Our process only has access to 1 CPU core through the kernel thread
  - The only way to use multiple cores is to have multiple processes running
- Often, it is impossible to write a non-blocking disk read function (OSs don't often give you the tools to do so)
  - This blocks the thread, waiting for the disk
  - Once the thread is blocked it can't yield to other threads so the whole process is waiting for that 1 thread!
  - This gets even worse when working with virtual memory
- If one thread blocks another thread, the whole process may get stuck in deadlock

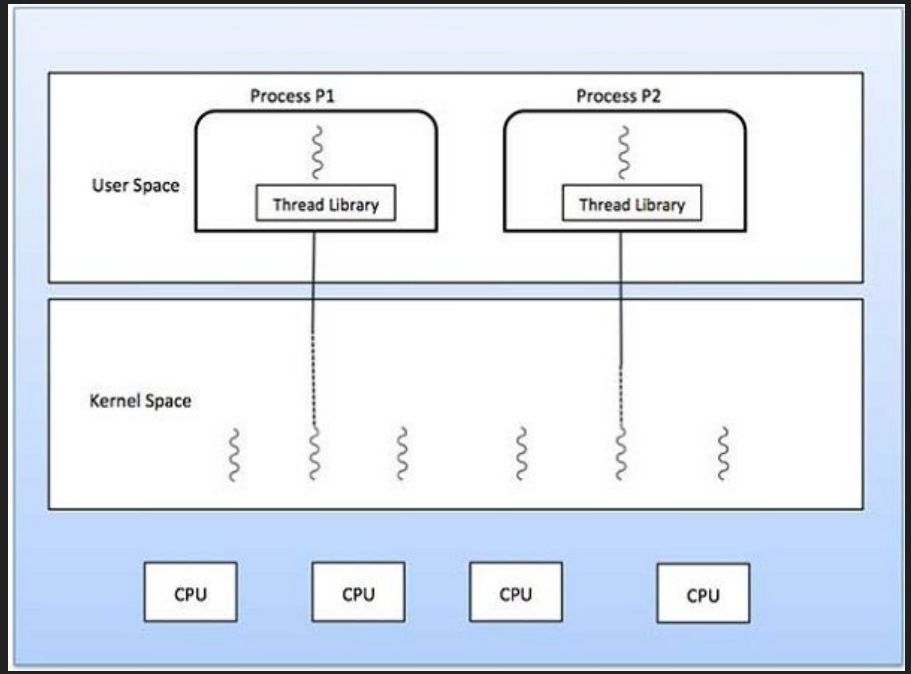
One-To-One Thread Model

- Each user-level thread corresponds to exactly one kernel-level thread
- Creates a direct association with a kernel-level thread managed by the operating system
- Threads can execute simultaneously on multiple processors
- When creating a user thread a kernel thread must also be created
    - This typically results in a lot of memory and processing overhead, creating kernel threads is very slow



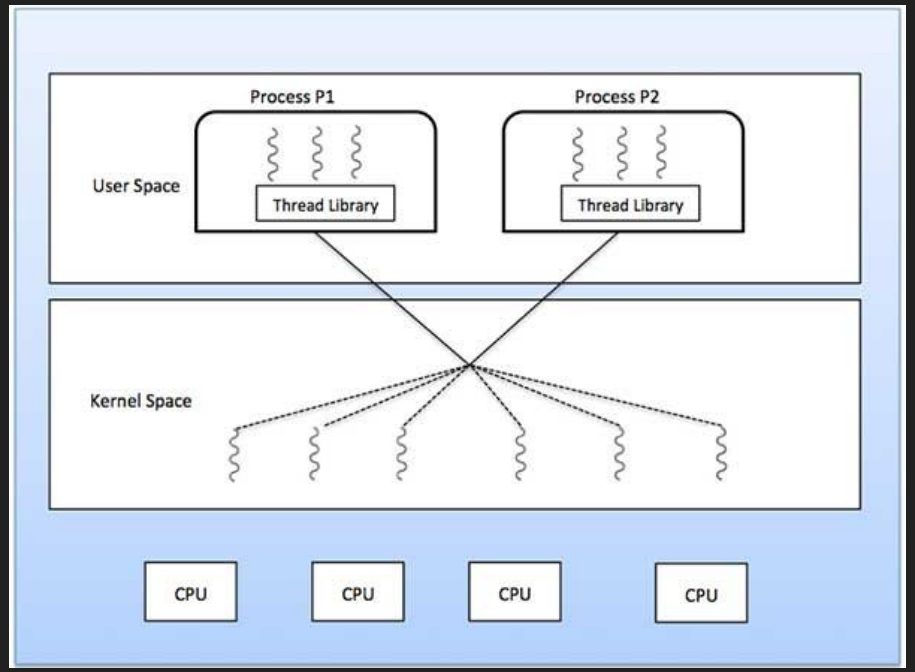
One-To-One Thread Model (cont.)

- What if a process wants 8 user threads on a 4 CPU system?
  - Only 4 can run in parallel
- What if 2 processes want 4 user threads each?
  - Only 1 process can execute its threads in parallel
- The application is required to limit its maximum user threads to the number of cores on the system



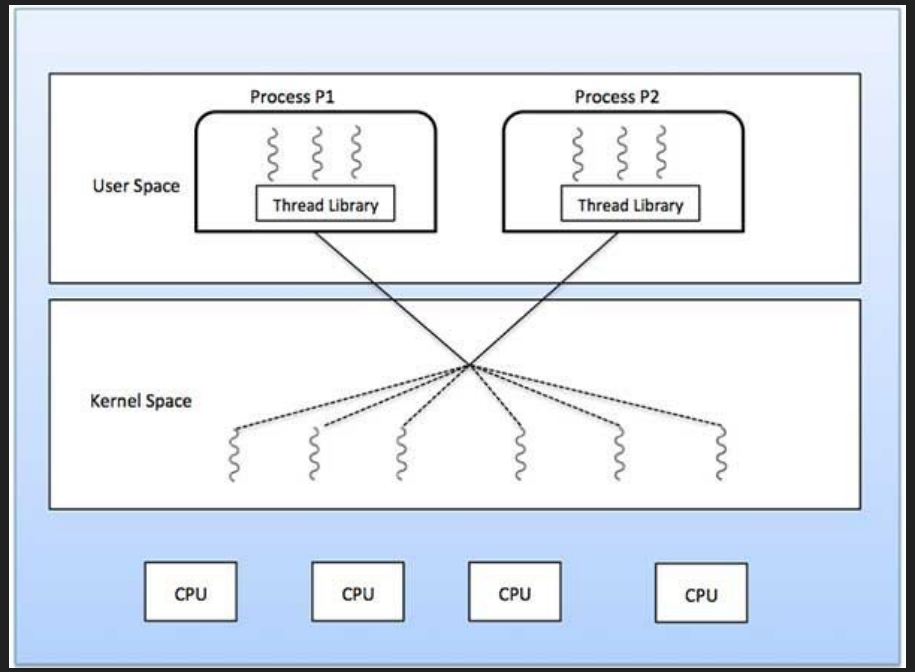
Many-To-Many Thread Model (aka n:m Threading)

- (n) number user-level threads are mapped to an equal or smaller number (m) of kernel-level threads
- Provides a balance between parallelism and efficiency, achieves better performance and flexibility
- User-level threads can run in parallel on multiple processors, and the operating system can manage and schedule a smaller number of kernel-level threads
    - Typically the kernel threads are between 1 thread to the number of cores on the CPU



Many-To-Many Thread Model (aka n:m Threading) (cont.)

- The OS determines how many kernel threads to assign
  - This may be specific to the application or to the hardware
- If a user thread blocks, the other user threads can continue to execute on separate kernel threads
- Many-to-many is not as commonly used due to its complexity



Problems With n:m Threading

- Our processes don’t really know what is happening with the actual CPU cores
  - Only the kernel knows how many CPU cores are available
  - The user space thread scheduler might constantly schedule a user thread that has a blocked *kernel* thread
- The kernel doesn’t know if one user thread is more important than another
  - If one user thread holds an important resource the kernel thread will eventually preempt and lock that thread for some time

Process Scheduling vs Thread Scheduling

- When using a multithreaded OS, threads are typically the main schedulable entity
- Each thread can, generally, be thought of as its own process
- On Linux and POSIX:
  - Threads of a multithreaded process are scheduled independently, rather than the whole process itself
  - Some threads may be allowed to be grouped, or be assigned priority, which may allow all the threads of one process to run simultaneously
  - For single threaded processes, only the single thread is scheduled amongst all the other threads in the system

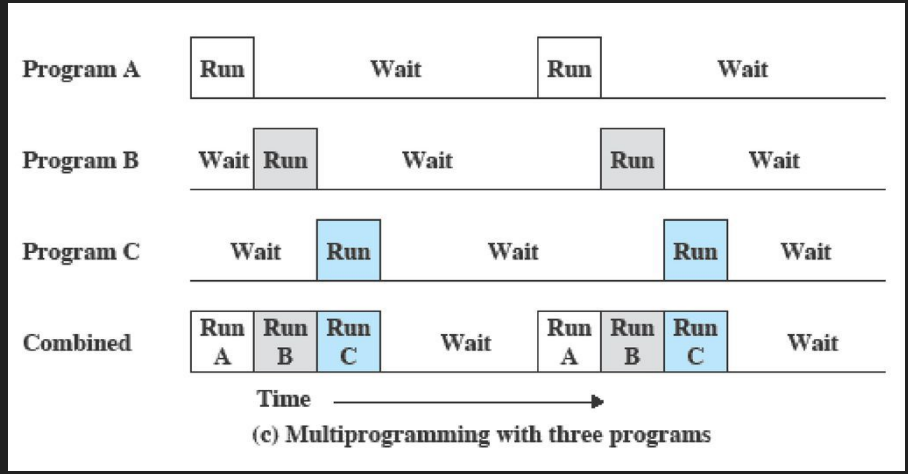
05 - CPU Scheduling

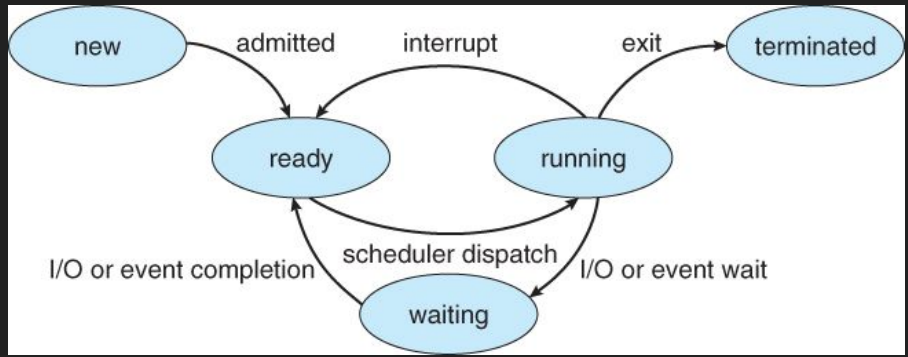
CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

Scheduling Processes

- Multiprocessing allows us to better utilize the CPU and execute multiple processes “at once”
  - In reality, only one process is running on the CPU at a time
  - The OS switches between multiple processes very quickly
  - If our hardware has multiple CPU cores we can execute multiple processes simultaneously
- If a process needs to wait for I/O it can give another process a chance to run
- All of our processes reside in a state queue
  - i.e. all ready processes live inside a “ready” queue that the OS will pull from when choosing to execute a new process





Scheduling Processes (cont.)

- Long Term Scheduling
  - How many processes will the OS allow to exist
  - We are limited by memory so we can’t have an infinite amount
- Short Term Scheduling
  - How should the OS select a “ready” process from the ready queue

Short Term Scheduling

- The process scheduler resides in the kernel and can make a decision when:
  - A process switches from running to waiting
    - i.e. if the process wants to access I/O
  - An interrupt occurs
    - i.e. if a process wants the keyboard data and the keyboard sends and interrupt
  - A processes is created or terminated
- Non-preemptive scheduling method must wait for one of the above events to occur before switching processes
- Preemptive scheduling method allows the scheduler to interrupt a process

What Makes a Good Scheduling Algorithm?

- CPU Utilization
  - How much is the CPU being utilized to its fullest potential?
  - Ideal: as close to 100% as possible
- I/O Utilization
  - How much of our I/O or storage is being used to its fullest potential?
  - Ideal: as close to 100% as possible
- Throughput
  - How fast are the processes completing?
  - Ideal: very fast
- Turnaround Time
  - How much time are processes taking to complete from “new” to “terminated”?
  - Ideal: very short
- Waiting Time
  - How much time do processes stay in the “ready” queue?
  - Ideal: very short
- Response Time
  - How much time between a process being “ready” and its next I/O request?
  - Ideal: very short

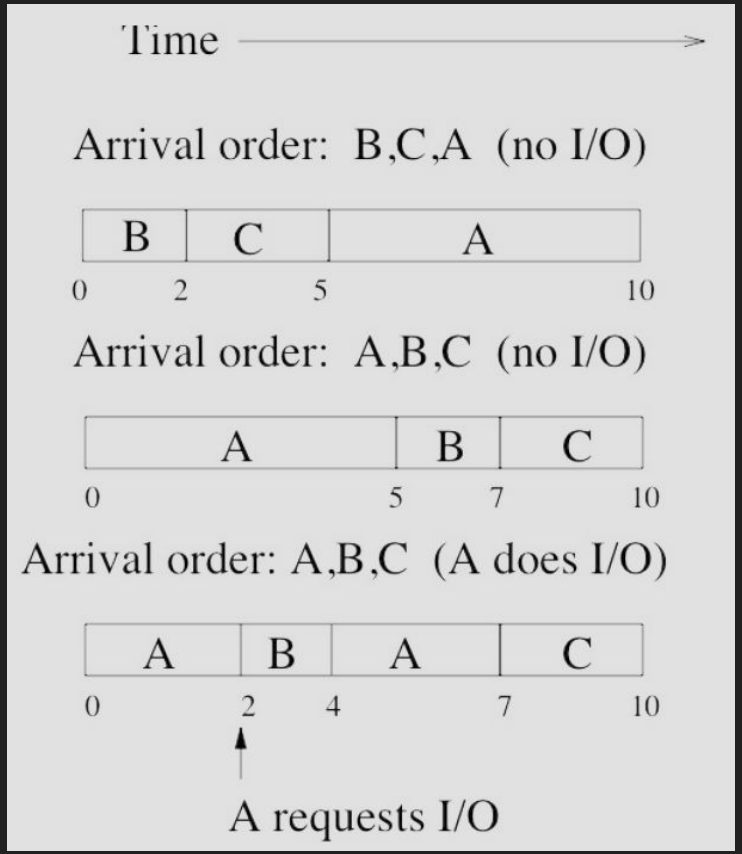
Scheduling Policies

- Ideally, you want a process scheduler to optimize all criteria but this is not realistic
- Instead, choose a scheduler that optimizes your most important metric(s):
  - Shortest response time
    - Good for when the user has to interact with the system (i.e. moving the mouse, typing on keyboard, loading screens, etc.)
  - Lowest response time *variance*
    - Having a consistent response time might make the system less frustrating to work with
  - Maximize throughput
    - Minimize OS overhead, context switching overhead, and efficiently use I/O and resources
  - Minimize waiting time
    - Give each process the same amount of CPU time but this may increase response time

First In First Out (FIFO)

- The scheduler will execute processes in the order in which they are created or arrive
  - Not necessarily the order that they are added to the ready queue
- Assume processes only release the CPU when they are waiting for I/O
- Notice, A still gets to finish before C
  - Even though A had to wait for I/O, it still arrived before C
- This method is cooperative (requires process to release the CPU)
- Advantage: simple to implement
- Disadvantage: process's time spent waiting is highly variable
- Disadvantage: I/O bound processes are not prioritized and forced to wait for CPU bound processes

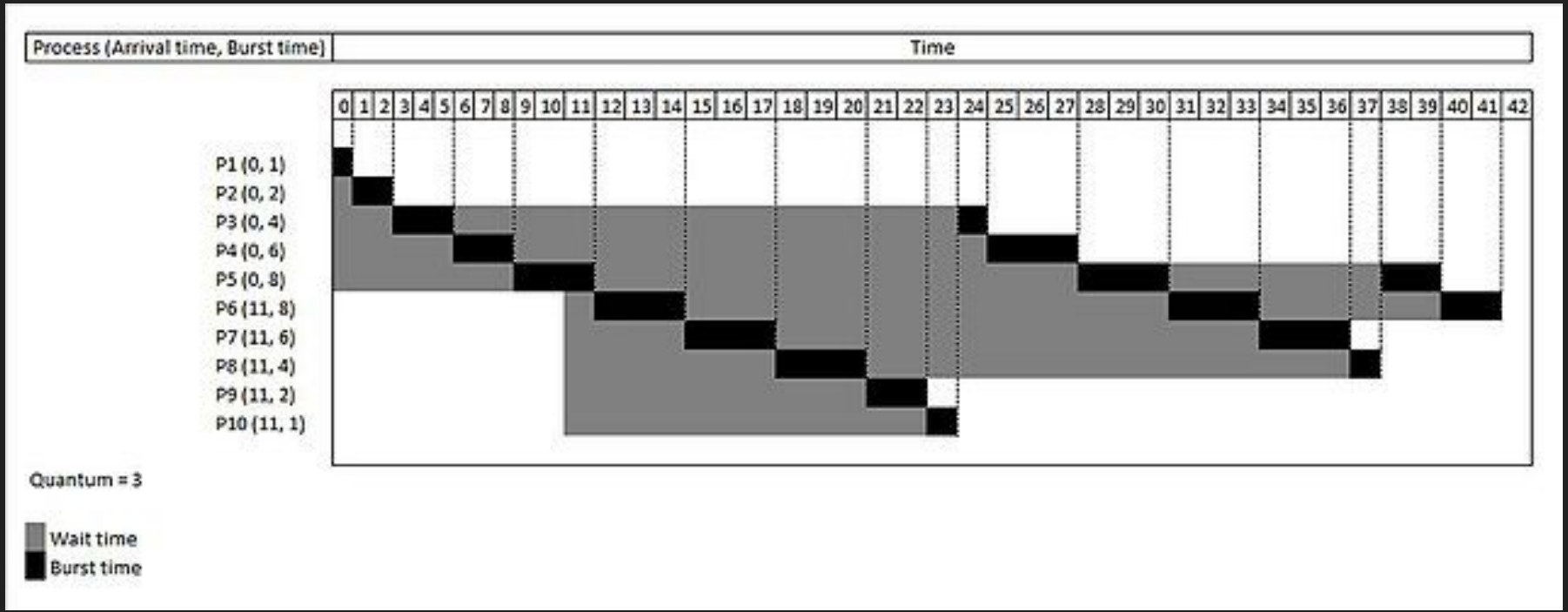
First In First Out (FIFO) (cont.)



Round Robin (RR)

- The scheduler schedules processes evenly or fairly giving each process a *quanta* of time to run on the CPU
  - The quantum is a predefined slice of time (e.g. 100 clock cycles)
  - The quanta is how many of these slices a process may use at a maximum (e.g. 3 quanta max. = 300 clock cycles max. CPU time)
- This method is preemptive (uses a clock to force processes to stop)
- Quanta too large - processes spend way too much time waiting
- Quanta too small - most of your time is spent context switching
  - Try to find a quanta where context switching is 1% of the total quanta
- Advantage: Consistent response time, all processes have equal time to access CPU
- Disadvantage: Process's average time spent waiting can be long (system may feel slow with 100's of processes running)

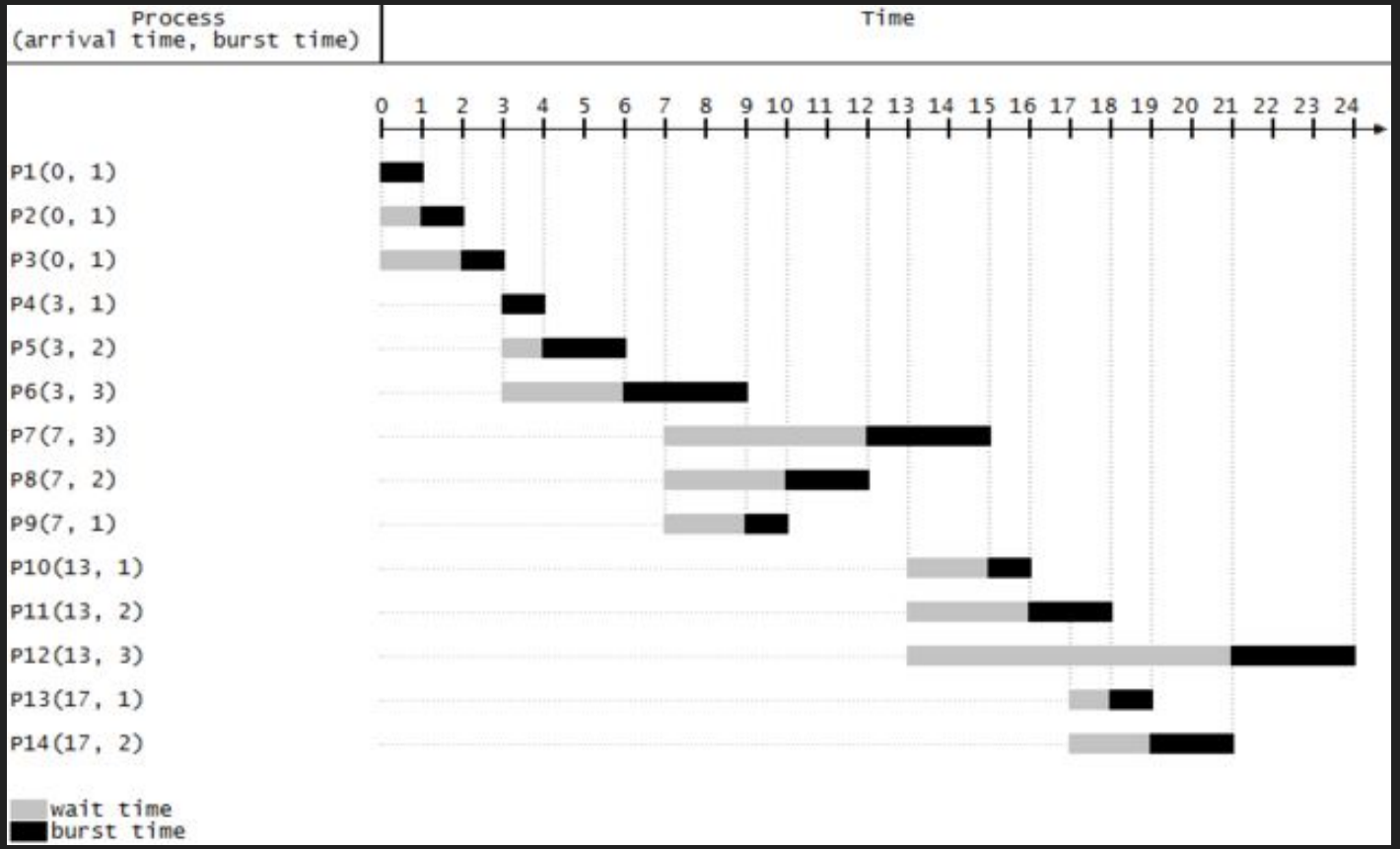
Round Robin (RR) (cont.)



Shortest Process Next (SPN)

- The scheduler schedules the process with the least (expected) amount of work (CPU time) to run until the process has an I/O request or terminates
- Can be preemptive or cooperative
  - Preemptive version uses shortest remaining time first
  - Preemptive version prioritizes I/O bound jobs over CPU bound jobs
- Advantage: Optimally minimizes (on average) process's waiting time
  - This is an optimal solution for this metric
- Disadvantage: impossible to predict how long a process needs to run
- Disadvantage: long processes may never get a chance to run

Shortest Process First (cont.)

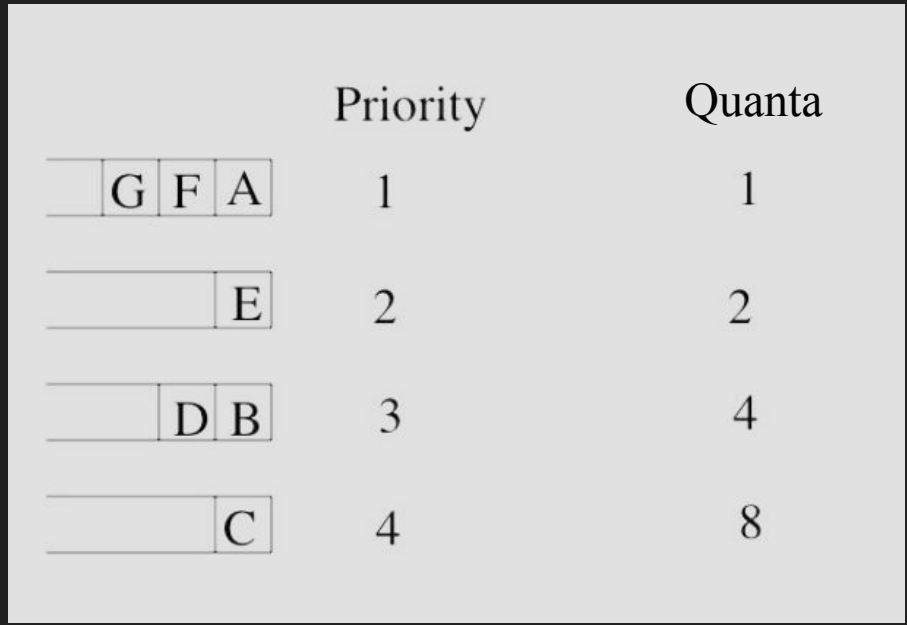


Multilevel Feedback Queue (MLFQ)

- The scheduler schedules based on past behavior of the process to attempt to predict the future and assign process priorities
- Used in most modern UNIX like systems
  - Overcomes the limitations of preemptive SPN
- If a process was I/O bound in the past, it is likely to be I/O bound in the future
- The scheduler can favor jobs that have used the least amount of CPU time (I/O bound processes), thus approximating SPN
- This policy is adaptive because it adapts to how the processes run and changes it scheduling behavior based on this past history
- The kernel keeps track of how often a process waits for I/O and prioritizes the processes that often access I/O

Multilevel Feedback Queue (MLFQ) (cont.)

- Multiple queues with priority based on predicted run time
- Use RR scheduling for each priority queue
  - Once finished, run the next priority level queue with RR
    - This can lead to starvation!
- Increase RR quanta exponentially for each priority level
  - This gives CPU bound processes more time to get stuff done



Multilevel Feedback Queue (MLFQ) (cont.)

- New processes start in the highest priority queue
- If the process wants to exceed its quanta, decrease the process's priority level by 1
  - The process is using more CPU time than expected
- If the process does not exceed its quanta, increase the priority level by 1 up to the highest priority level
  - The process is using less CPU time than expected
  - This can happen if the process begins waiting for I/O very quickly
- I/O bound jobs become higher priority
- CPU bound processes become lower priority

Improving Fairness

- Since SPN is optimal, but unfair and can starve long processes, increasing fairness must increase the waiting time
- Possible solutions:
  - Give each queue a fraction of CPU time
    - This is only fair if priority level queues have the same number of processes
  - Adjust the priority of processes if they are not getting ran
    - Unix originally took this approach
    - Avoids starvation but waiting time suffers when system is overloaded
      - This is because every process becomes high priority!

Lottery Scheduling

- Give each process a set of tickets
  - Assign more tickets to short running processes
  - Assign fewer tickets to long running processes
- Each quantum, randomly pick a winning ticket
- On average, CPU time is proportional to the number of tickets given to a process
- This approximates SPN while avoiding starvation since every process has a ticket that can be picked to run
- As the CPU load changes, adding or removing processes affects all other processes proportionately
  - Regardless of how many tickets each process has

Lottery Scheduling Performance

- In the example to the right, assume:
  - Short jobs get 10 tickets
  - Long jobs get 1 ticket
- How do we know how long the processes run?
  - Look at past history and estimate
- How do we determine how many tickets to give out?
  - Let the user decide (basically allows user to choose priority)
  - Let the OS determine based on the time

# of short processes / # of long processes	% of CPU each short process gets	% of CPU each long process gets
1/1	91% (10/11)	9% (1/11)
0/2	N/A	50% (1/2)
2/0	50% (10/20)	N/A
10/1	~10% (10/101)	< 1% (1/101)
1/10	50% (10/20)	5% (1/20)

Scheduling Algorithm Summary

- FIFO
  - Not fair, and average waiting time is poor
- Round Robin
  - Fair, response time variance minimized, but average waiting time is poor
- SPN
  - Not fair, but average waiting time is minimized assuming we can accurately predict the length of the next CPU burst
  - Starvation is possible
- Multilevel Queuing
  - An implementation (approximation) of SJF.
- Lottery Scheduling
  - Fairer with a low average waiting time, but less predictable.

06 - Main Memory

CEG 4350/5350 Operating Systems Internals and Design

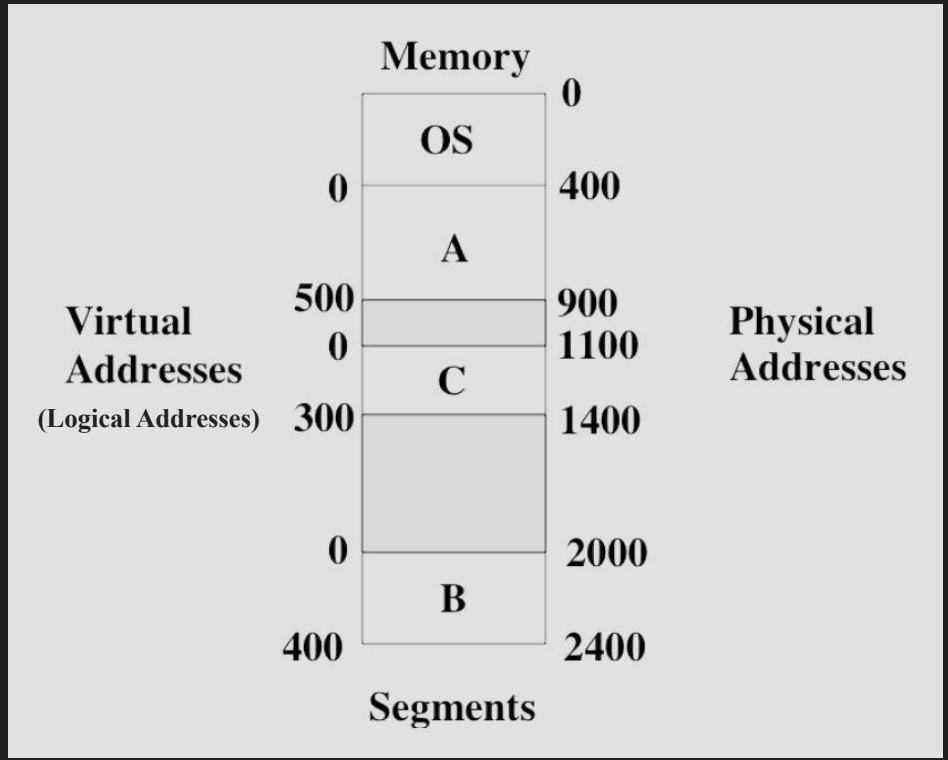
Max Gilson

Background

- All of our programs exist on our disk
- Before we can execute them they have to be brought into main memory (RAM) by the OS
- The CPU will fetch instructions from RAM to execute our program
- Your OS project does this already for your kernel
    - Before your kernel can be executed, it must be loaded off of the floppy disk

Memory Terminology

- Segment - a chunk of memory assigned to a process
- Physical address - a real address in memory that corresponds to an actual physical memory location
  - e.g. x000B8000 or xFFFF1234
- Virtual address - an address relative to the start of the process's address space
  - e.g. if a process exists within the range x1234 to x1FFF it will have virtual addresses from x000 to xDCB
  - This gives our programs more flexibility when accessing memory
  - Also called "logical address"
- Contiguous memory - memory that is contained within one region, one address after the other

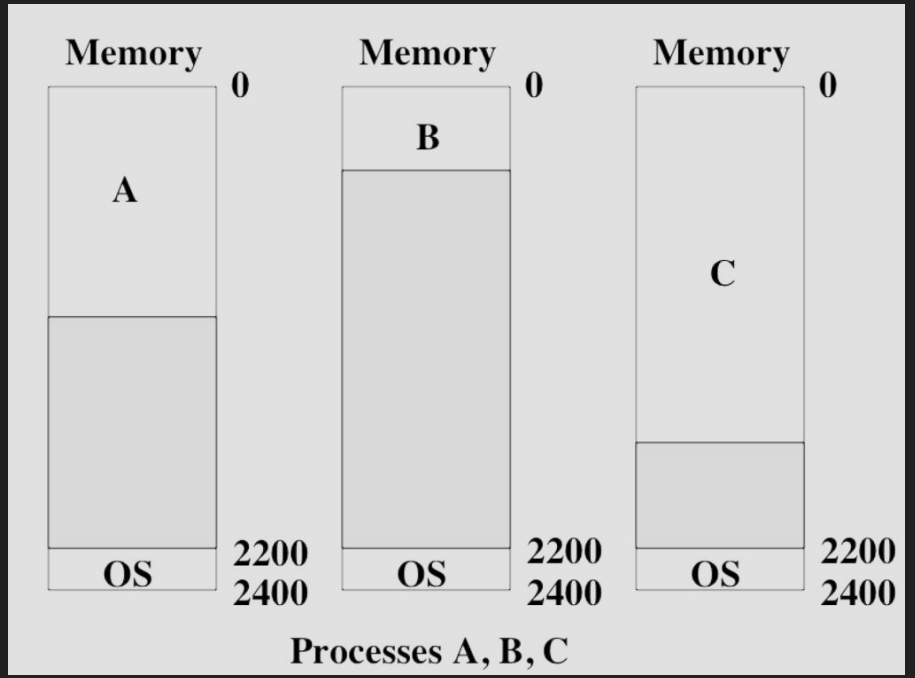


Where Do Addresses Come From?

- How do programs generate addresses for instructions and data?
- 3 different methods:
  - Compile time
    - The compiler decides where the program starts in memory at a fixed physical memory address
    - The OS does nothing
  - Load time
    - The compiler decides a starting address
    - The OS determines where this starting address gets placed in physical memory
    - Once loaded, the process does not move in memory
  - Execution time
    - The compiler decides a starting address
    - The OS determines where this starting address gets placed in physical memory
    - When the process is running, the OS will translate the virtual addresses to physical addresses

Uniprogramming

- Uniprogramming - running only 1 process + OS
- Assume the OS gets a fixed part of memory up to the highest address (DOS-like)
- Process is loaded at physical address x00000000
  - Process executes in a contiguous section of memory
- Maximum address = Memory Size - OS Size
- Simple but does not allow for overlap of I/O and CPU usage

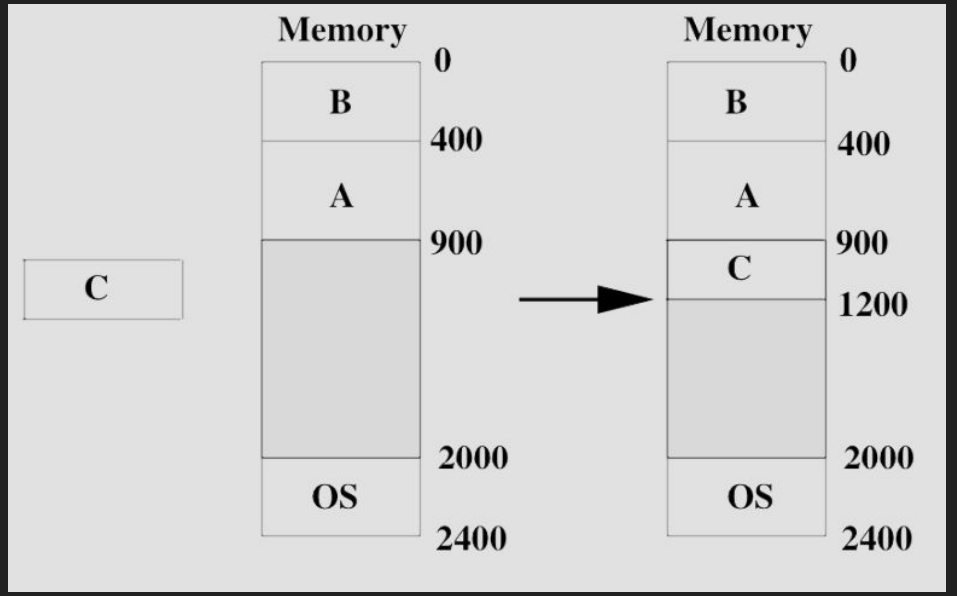


Requirements of Multiprogramming

- Transparency
  - We want multiple processes to coexist in physical memory
  - No process should be aware memory is shared
  - Processes should not care where they are in memory
- Safety
  - Processes must not corrupt each other or the OS
- Efficiency
  - Performance of CPU and memory should not be degraded

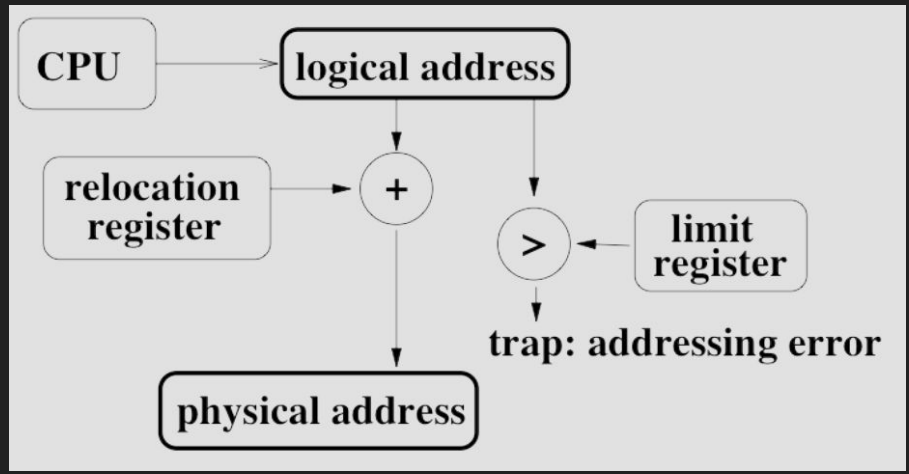
Relocation

- Relocation is moving a process to any location in memory
- Assume the OS gets a fixed part of memory up to the highest address (DOS-like)
- Assume at compile time the process starts at address 0 with: Maximum address = Memory Size - OS Size
- Base address
  - The first physical address of the process
- Limit address
  - The last physical address of the process



Types of Relocation

- Static Relocation
  - When the process is loaded, the OS offsets all addresses to reflect the process's new location in memory
  - Once the process is assigned to memory it cannot be moved
    - Moving would require re-offsetting all the addresses for every instruction in the program!
- Dynamic Relocation
  - Hardware has base register that gets added to virtual address, the result is the physical address
  - Hardware compares address with limit address
    - If the address exceeds the limit address, execute a trap service routine to handle addressing error and ignore physical address
  - Assume all logical addresses are positive



Benefits of Dynamic Relocation

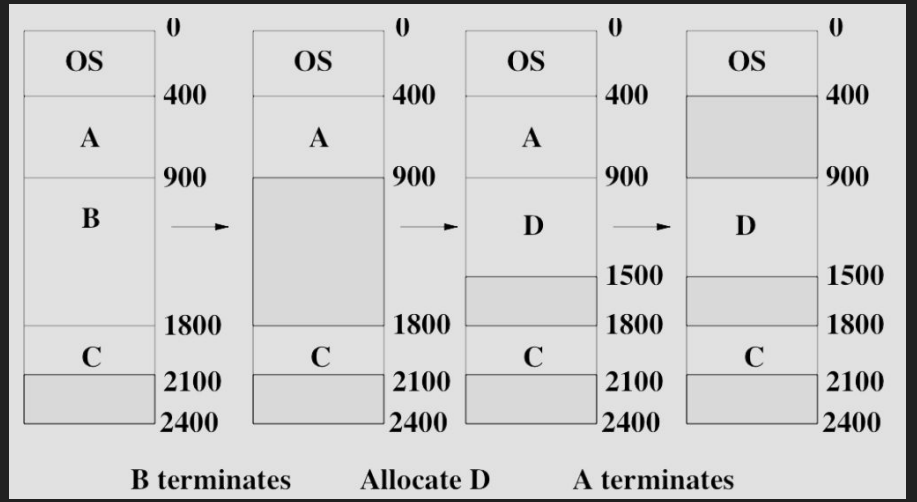
- Advantage:
  - OS can move a process in memory much easier
  - OS can allow a process to grow over time
  - Hardware requirements are simple
    - 2 extra registers, addition, and comparison
- Disadvantage:
  - Extra hardware may increase time to access memory
  - Sharing memory between processes is impossible
    - Each process is restricted to its segment
  - All running processes must coexist in physical memory
  - We are still using contiguous memory

Benefits of Dynamic Relocation (cont.)

- Transparency
  - Processes are unaware of other processes in memory
- Safety
  - Each memory access is checked by hardware
- Efficiency
  - Slightly slower but still very fast
  - Moving a process to a new location in memory is very slow
    - This may be necessary if a process grows

Memory Allocation

- As processes start, grow, and terminate, the OS tracks all used and unused memory
- When a new process starts, the OS must decide where the process goes into memory
  - A *hole* is a location in memory that is not filled
  - The OS tries to fill these holes with new processes
- What if we tried to allocate B again in the last diagram?
  - Hint: it won't fit

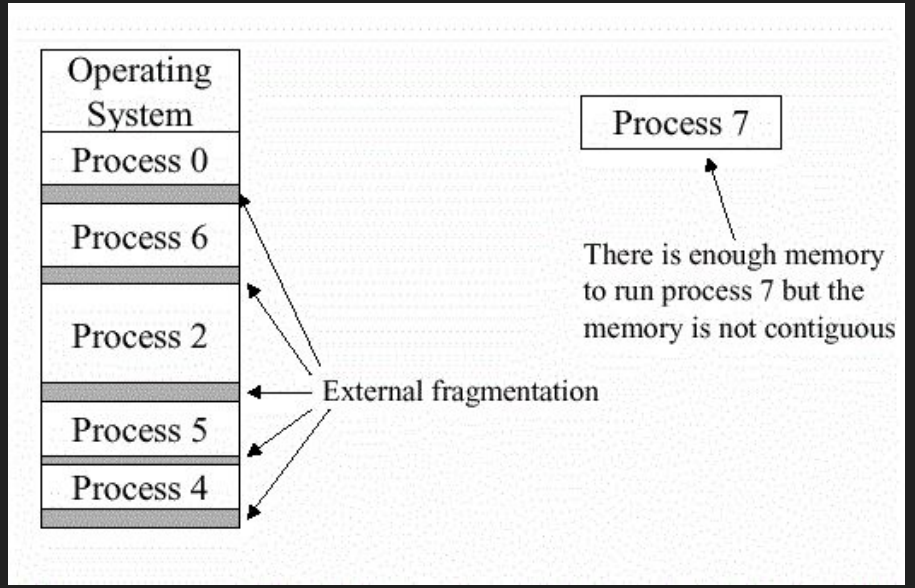


Memory Allocation Policies

- First Fit
  - Allocate the process to the first hole it fits in
  - Generally faster than Best Fit
- Best Fit
  - Allocate the process to the smallest hole that the process still fits in
  - Generally better storage utilization than Worst Fit
- Worst Fit
  - Allocate the process to the largest hole in memory
  - The remaining memory after the process might be large enough to fit another process or allow the process itself to grow
- For Best Fit and Worst Fit, the OS must search the entire list of holes to find the desired location to put the process
  - This can be slow if there are 1,000’s of holes!

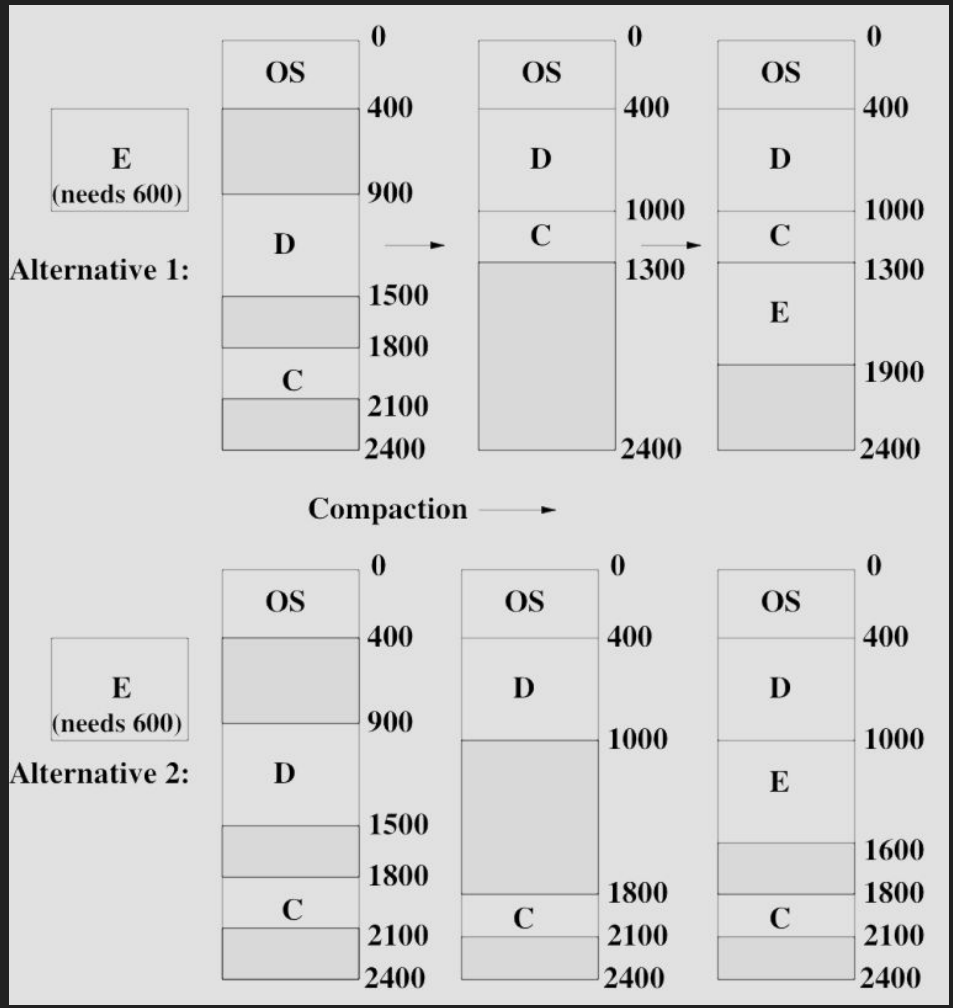
Fragmentation

- External Fragmentation
  - The memory between processes that is too small to be used
- Internal Fragmentation
  - Occurs if memory split into fixed sized chunks
  - If a process occupies one fixed sized chunk, but doesn't use it all



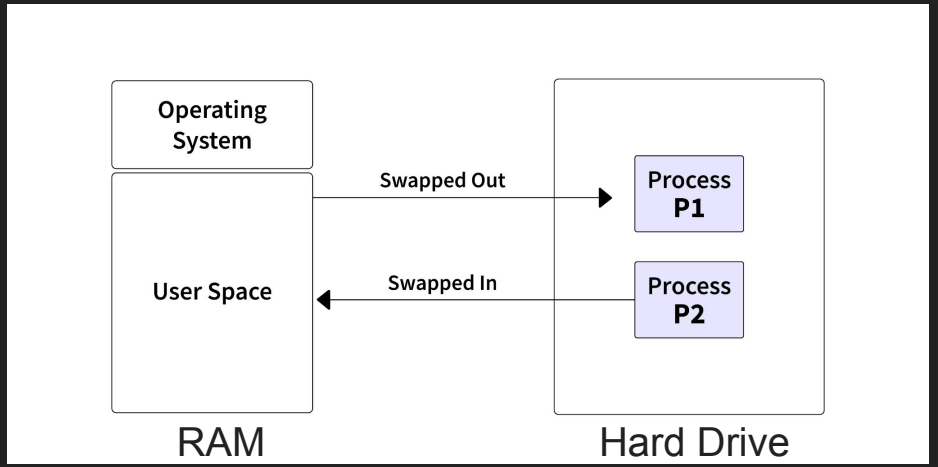
Compaction

- Compaction involves moving processes inside of memory to remove external fragmentation
- Advantage:
  - We can now load another process into memory
- Disadvantage:
  - Expensive operation, takes a lot of time
- Question:
  - We only need processes in memory if we're executing them. So, why do all 3 processes have to be in memory?



Swapping

- When we aren't executing a process, save it to the disk so we can resume it later
  - This frees up memory for other processes to use
- If the process becomes active again, reload it into memory
  - If using static relocation, process must go into same memory location
  - If using dynamic relocation, process can go anywhere in memory and OS must update relocation/limit registers
    - Compaction can be performed at this time
- Swapping processes takes a long time
  - If our scheduler wants to execute a process that is on disk, it should load that process while scheduling other processes first

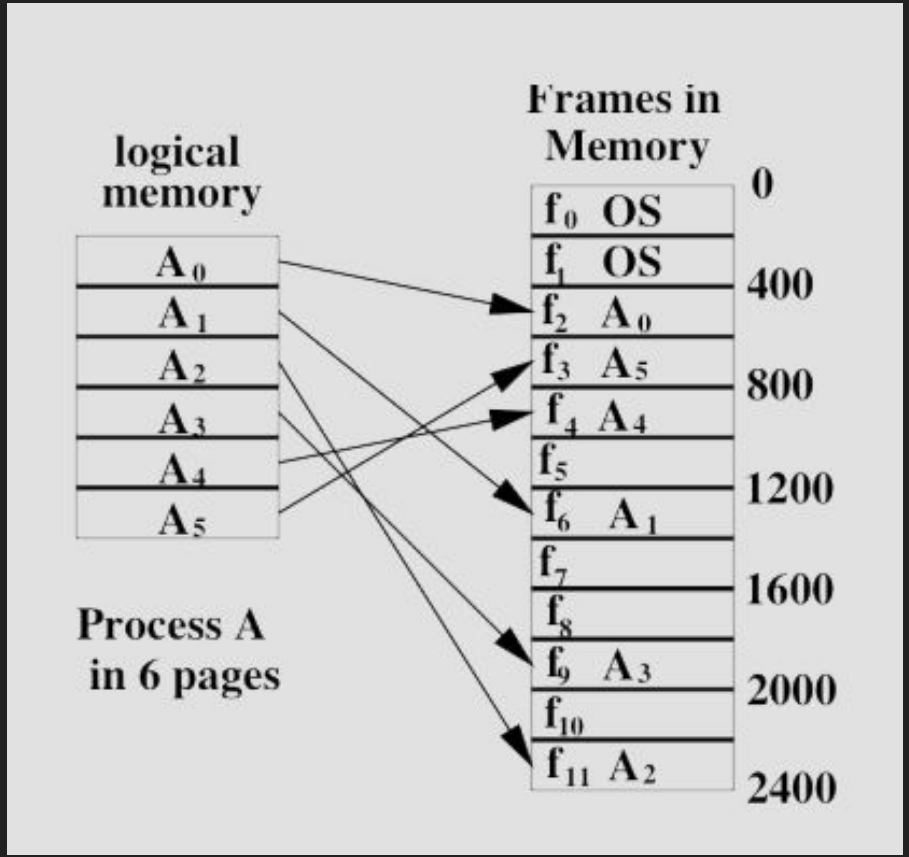


Paging

- Let’s assume that processes spend 90% of their time accessing 10% of their memory
  - Let’s only keep that 10% of their memory *in* memory unless the process needs more
- The logical memory of a process is contiguous
- The physical memory of a process is NOT contiguous
- Memory is divided into fixed size pages
  - This eliminates external fragmentation
  - Internal fragmentation still exists, about ½ a page is wasted per process
- Each process has its own table that translates pages to frames in memory
  - This table is managed by the OS

Paging (cont.)

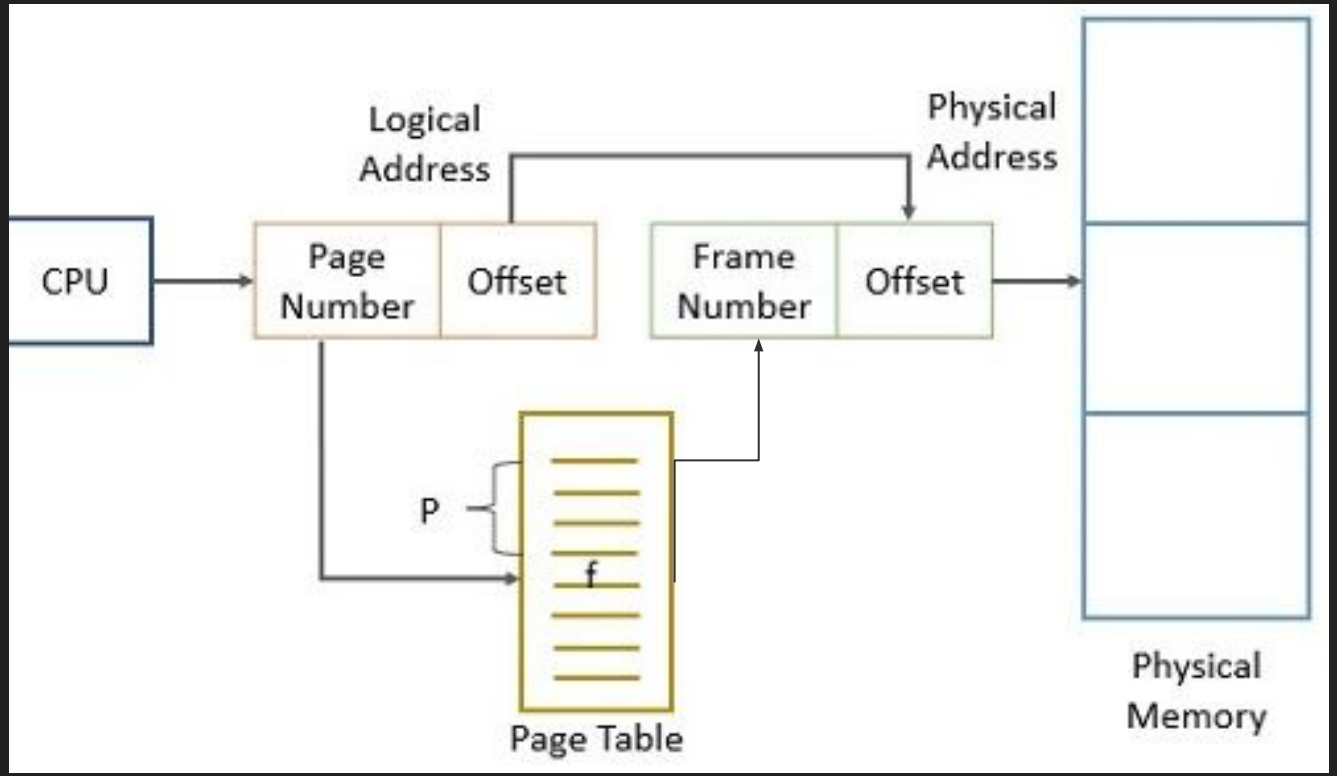
- The entirety of our process is divided up into fixed size pages
- These pages get mapped to frames in physical memory
- Pages can be moved, removed, added to frames in memory
- Our process should not care where its mapped to in physical memory
    - Our OS must do extra work keeping track of and maintaining processes page mapping
    - Our OS must translate logical addresses to physical addresses at runtime



Creating and Mapping Pages

- Processes refer to their memory locations with a virtual address
- Process generates contiguous virtual addresses from 0 to the size of the process
- The OS separates the process into pages and keeps track of their mapping to frames in the page table
- The paging hardware translates a virtual address (page number and page offset) into a physical address (frame number and frame offset)

Paging Hardware



Finding a Physical Address

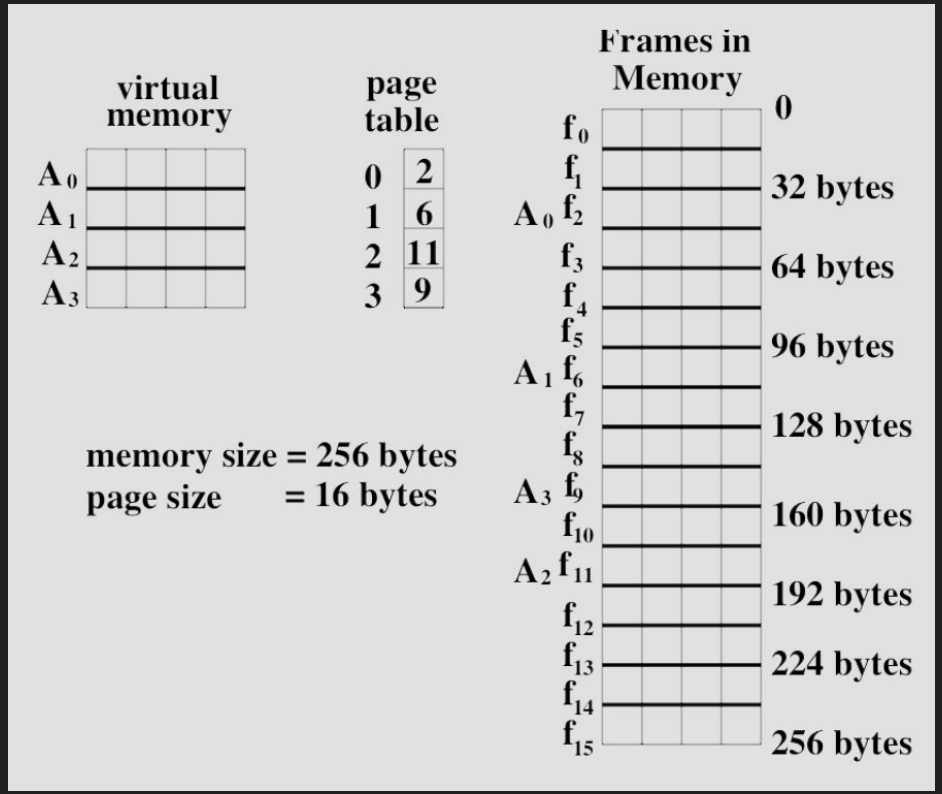
- Pages are typically 4096 bytes
- Assume 32 bit addresses with byte addressable memory
- Assume that the logical address requested by the CPU is:
  - x01A2C003
- Assume that there is an entry in the page table:
  - x01A2C : x20002
- What is the physical address that will be accessed?
  - x20002003

Paging Hardware

- Paging is a form of dynamic relocation, each virtual address is bound by the paging hardware to a physical address
- The page table acts as a set of relocation registers, one for each frame
- Mapping is invisible to the process, the OS does the mapping and the hardware does the translation
- Protections are provided with the same mechanisms used in dynamic relocation

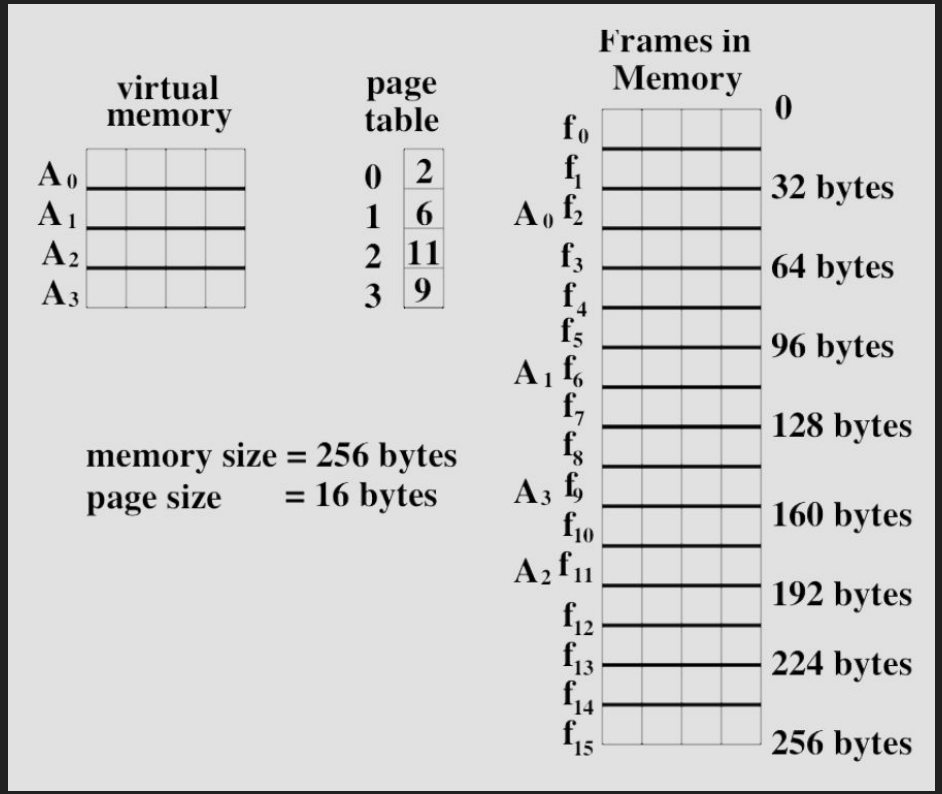
Address Translation Example 1

- Assume 256 bytes of total memory
- Assume byte addressable
- Assume page size of 16 bytes
- How big is the page table?
  - 16 entries = 256 total bytes / 16 bytes per page
- How many bits for a physical address?
  - 4 bits for frame (16 frames)
  - 4 bits for offset (16 bytes per frame)
  - 8 bits = 4 + 4 (256 addresses)
- How many bits for a virtual address?
  - 4 bits for 16 entries (page)
  - 4 bits for 16 bytes per page (offset)
  - 8 bits = 4 + 4
- Given a virtual address 0001 1000 in process A, what is the frame and offset of the physical address?
  - Page = 1, offset = 8
  - Frame = 6 (from page table), offset = 8
  - Physical address = 0110 1000



Address Translation Example 2

- Assume 256 bytes of total memory
- Assume word addressable with 4 byte word size
- Assume page size of 16 bytes
- How big is the page table?
  - 16 entries = 256 total bytes / 16 bytes per page
- How many bits for a physical address?
  - 4 bits for frame (16 frames)
  - 2 bits for offset (2 words per frame)
  - 6 bits = 4 + 2 (64 addresses) (64 = 256 / 4)
- How many bits for a virtual address?
  - 4 bits for 16 entries (page)
  - 2 bits for 4 words per page (offset)
  - 6 bits = 4 + 2
- Given a virtual address 0011 01 in process A, what is the frame and offset of the physical address?
  - Page = 3, offset = 1
  - Frame = 9 (from page table), offset = 1
  - Physical address = 1001 01



Where is The Page Table?

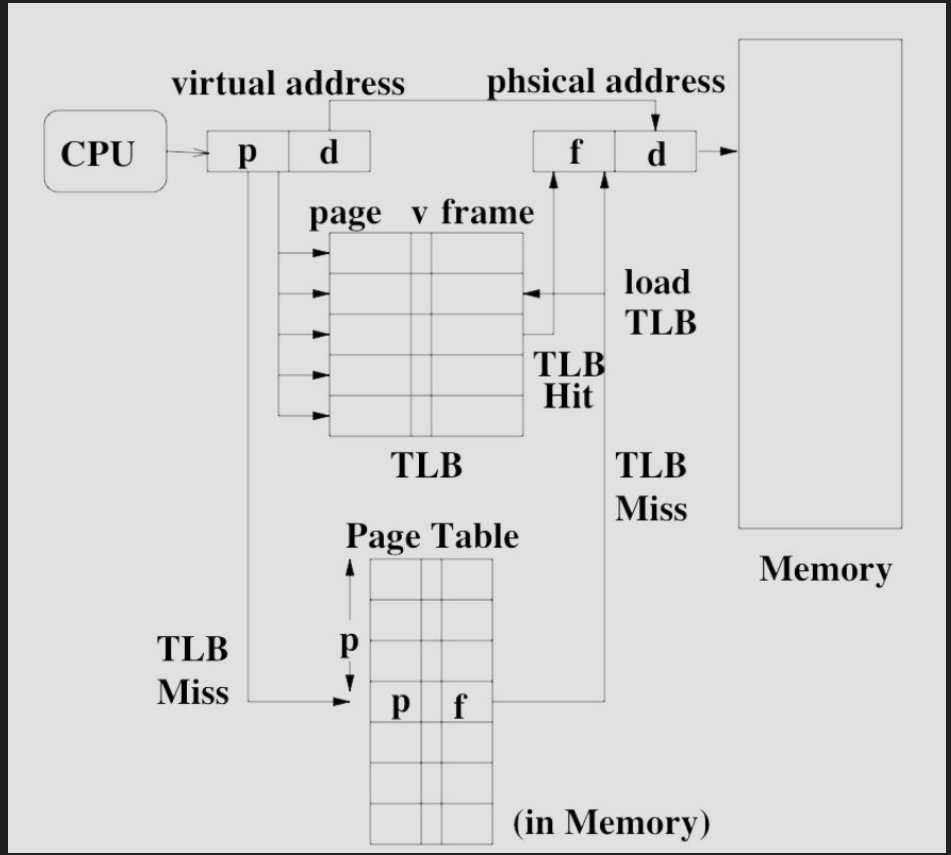
- Registers?
  - Very fast but we only have a few fixed number to work with
- Memory?
  - Slow but we probably have plenty of space
- Translation Look-Aside Buffer (TLB)
  - Works like a cache
  - Much larger than registers
  - Much faster than memory
  - Holds only a portion of our page table, the rest can exist in memory

Where is The Page Table?

- Registers?
  - Very fast but we only have a few fixed number to work with
- Memory?
  - Slow but we probably have plenty of space
- Translation Lookaside Buffer (TLB)
  - Works like a cache
  - Much larger than registers
  - Much faster than memory
  - Holds only a portion of our page table, the rest can exist in memory

TLB

- Have a small cache (TLB) that's going to store most of our page table
- TLB Hit:
  - If the page we requested exists in the TLB, translate the virtual address to a physical address
- TLB Miss:
  - If the page we requested does not exist in the TLB, go to memory to get it and put the missing page into the TLB
- Valid bit indicates if the TLB entry matches the page table entry
- A replacement algorithm must be implemented to replace entries in the TLB
  - Least recently used page can be replaced



Costs of Using the TLB

- How much time does it take to access memory if the page table is inside of memory?
  - T = total time, ma = memory access time
  - T = 2 * ma
  - We have to access the page table, then the physical memory, so 2 memory accesses are required
- How much additional time does the TLB add?
  - tlb = TLB access time, p = hit rate
  - T = (ma + tlb) * p + (2 * ma + tlb) * (1 - p)
- A larger TLB size decreases average memory access time

Initializing Memory With a New Process

1. Process arrives and needs k number of pages
2. If k frames are free, allocate the frames to pages, otherwise, free up frames that are no longer needed
3. OS puts each page inside each frame and adds the frame number to the page table
4. OS marks all TLB entries as invalid (flushes TLB)
5. OS starts executing process
6. OS loads TLB entries as each page is accessed, replacing existing entries if needed

The Process Control Block

- The PCB must be extended to contain
  - The entire process's page table
  - A copy of the TLB maybe
- During a context switch
  - Copy the page table base register value into the PCB for the process we are switching out
    - Page table base register (PTBR) points to the page table in memory for this process
  - Copy the TLB into the PCB (optional)
  - Flush the TLB
  - Restore the page table base register for process we're switching to
  - Restore the TLB if it was saved in the new process's PCB

Sharing Code Between Processes

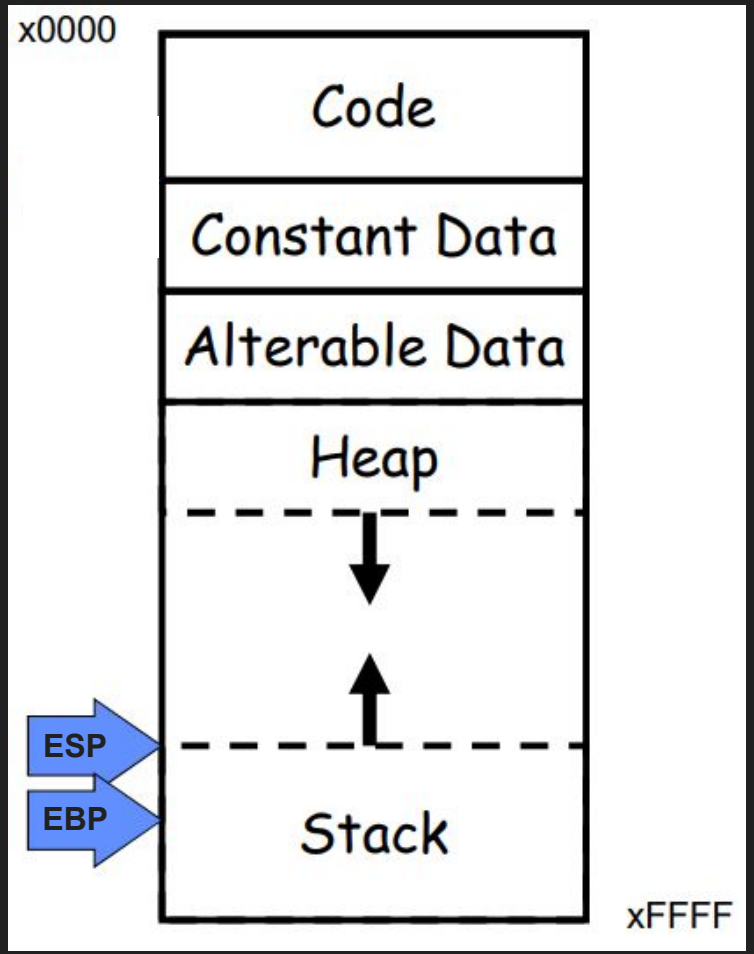
- Since a process's physical memory does not need to be contiguous, multiple processes could be allowed to access the same frame
- Any shared code must be reentrant, meaning processes should not be allowed to change it
  - Shared libraries (like standard C libraries) can easily be shared amongst processes and only one instance of the library is needed to be in memory

Paging Summary

- Advantages to paging:
  - Eliminates external fragmentation and compaction
  - Allows for sharing of code amongst processes
  - Processes can be partially loaded into memory
- Disadvantages to paging:
  - Translating a virtual address to a physical address takes more time
  - Requires hardware support with TLB
  - Requires more complex OS to maintain page table

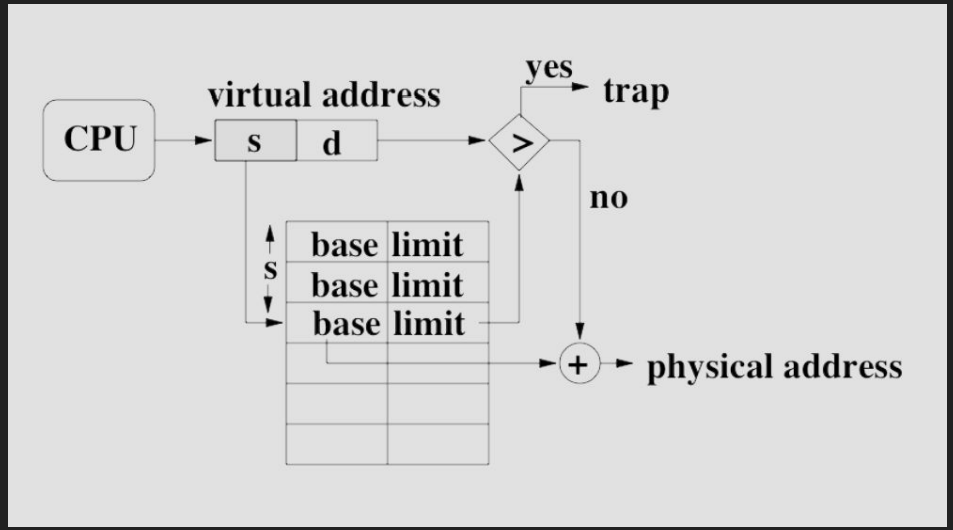
Segmentation

- Segmentation is another form of memory management
  - OSs may use segmentation instead of paging or they may use a combination of both
  - Windows and Linux use a combination of paging and segmentation
- Processes are made of different segments:
  - Code (instructions), variables (data), stack, heap
- Segmentation refers to these segments directly, instead of treating a process's memory like a linear array of bytes
  - The virtual address is a combination of a base register that identifies the segment and an offset inside that segment



Segmentation Hardware

- A segment table contains the base and limit registers of the segment
  - Additional information may be included to specify if a segment can be shared, read, written, etc.
- Typically, each process only has a few segments
  - x86 architecture only allows for 6 segments with the following registers: CS, SS, DS, ES, FS, GS
  - If a system uses many segments a TLB-like system may need to be implemented
- Segmentation does not eliminate external fragmentation
  - Segments must be contiguous and are of arbitrary size



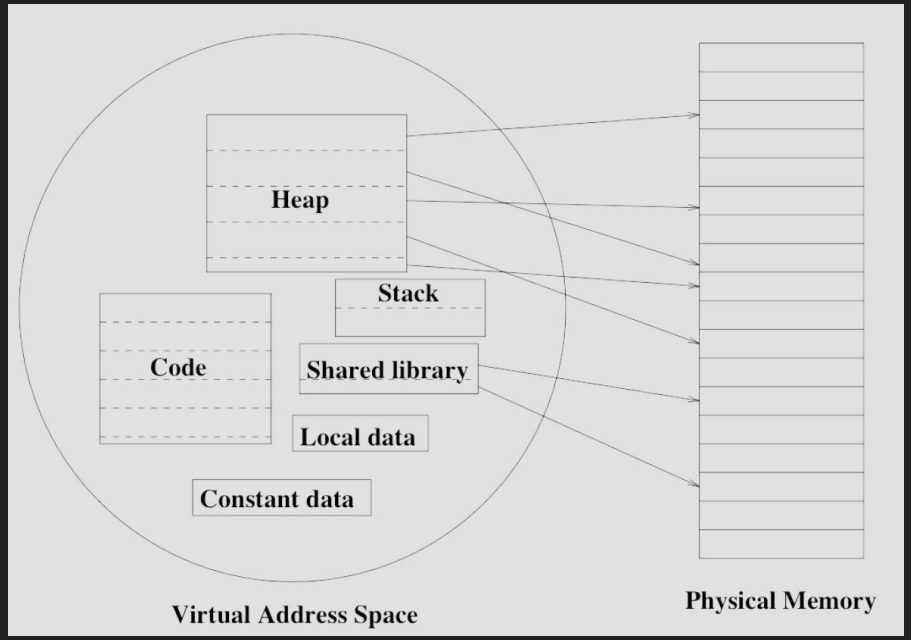
CS	Code Segment
DS	Data Segment
SS	Stack Segment
ES	Extra Segment
FS	General Purpose Segments
GS	

Segmentation Implementation

- At compile time, virtual addresses are generated where the most significant bit(s) are the segment number and the least significant bits are the offset into that segment
- Segmentation could be combined with dynamic or static relocation
  - Each segment is contiguous and could be assigned to any region of physical memory

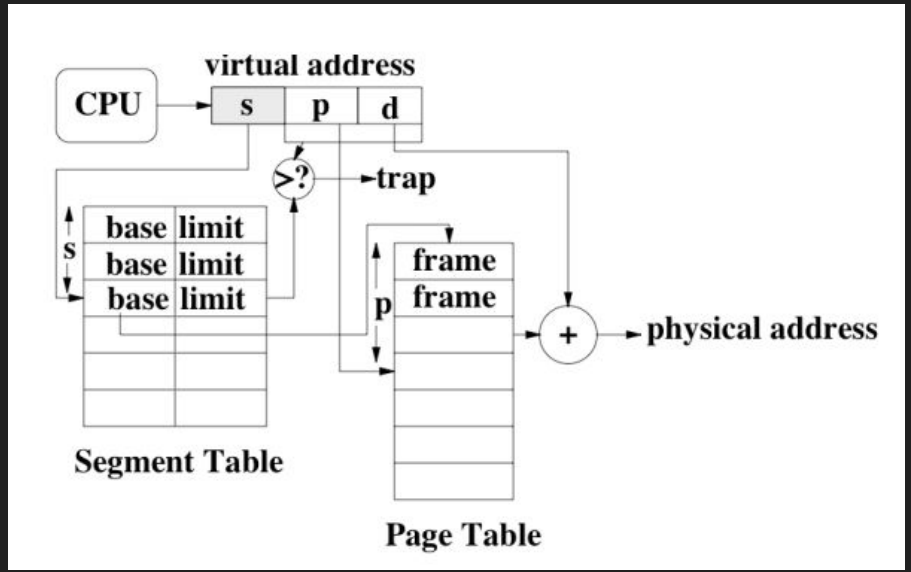
Segments and Pages

- Treat virtual address space as a collection of segments
- Treat physical memory as a sequence of fixed size frames
- Split the virtual segments into pages and map onto multiple frames



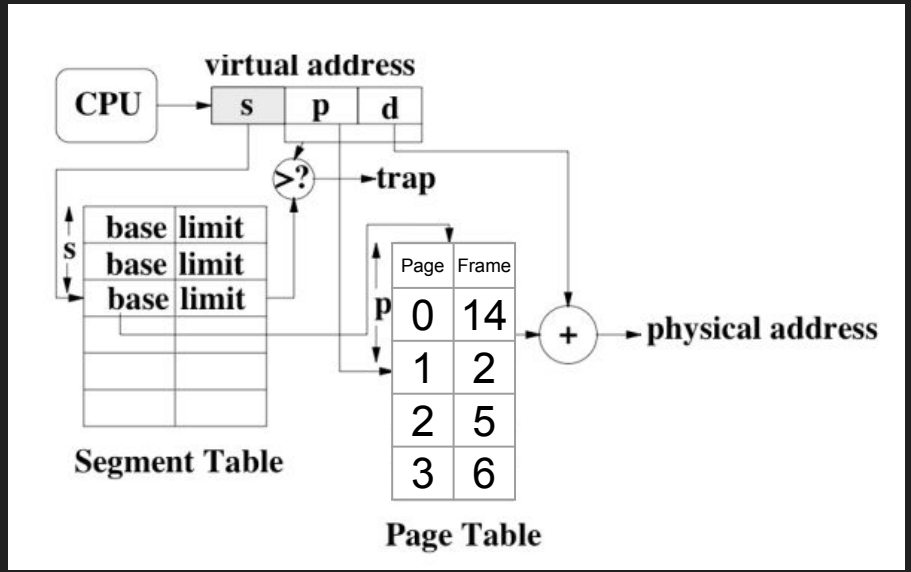
Translating Addresses in Segmented Paging

- The CPU uses a virtual address that contains a segment number, page number, and offset
- Both a segment table and page table are required
- Segment table can exist inside of registers or memory
    - Limits the maximum number of segments
- Page table can exist inside of memory and utilize a TLB
- Each base segment register points to a different page table that corresponds to the given segment



Segmented Paging Example 1

- Assume 256 bytes of total memory
- Assume byte addressable memory
- Assume page size of 32 bytes
- Assume maximum segments of 8
- How big is the page table?
  - 8 entries = 256 total bytes / 32 bytes per page
- How many bits for a physical address?
  - 3 bits for frame (8 frames)
  - 5 bits for offset (32 bytes per frame)
  - 8 bits (256 bytes)
- How many bits for a virtual address?
  - 3 bits (8 segments)
  - 3 bits (8 entries) (page)
  - 5 bits (32 bytes per page) (offset)
  - 11 bits = 3 + 3 + 5
- Given a virtual address 011 001 10101 in process A, what is the segment, frame, and offset of the physical address? Assume the segment points to the page table listed in the diagram.
  - Segment = 3
  - Page = 1, offset = 21
  - Frame = 2 (from page table), offset = 21
  - Physical address = 010 10101



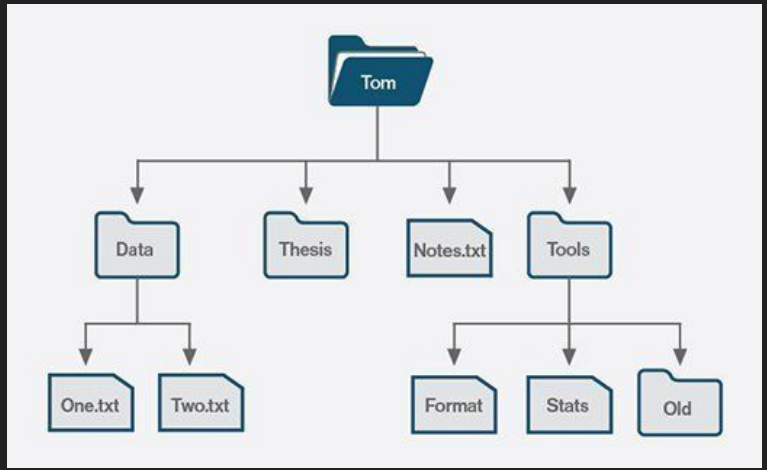
08 - File Systems

CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

File Systems

- File systems are traditionally the hardest part of an operating system to implement
- A file system keeps our files in storage organized and accessible
- Usually, file systems are organized with directories (folders) which contain files or more directories
- File systems are OS independent, but an OS must be written to comply with one or more file systems



File System Types

- There are many ways file systems may be implemented, a few are:
  - NTFS - New Technology File System
    - Modern Windows (90’s onward)
  - FAT - File Allocation Table
    - Old Windows (Windows 95) and flash drives
  - APFS - Apple File System
    - Modern Apple devices, MacOS, iOS, etc.
  - HFS/HFS+ - Hierarchical File System
    - Old Apple devices
  - Ext - Extended File System
    - Linux devices
- Comprehensive list here: https://en.wikipedia.org/wiki/List_of_file_systems

Goals of a File System

- A file system should have the following features:
  - Persistence
    - Keep files forever (unless we delete them)
  - Ease of use
    - Give human readable names to files and directories
  - Speed and efficiency
    - Quickly access any file
  - Protection
    - Protect files from other files and prevent unauthorized access to files from users

Storage is Different Than Memory

- Memory (RAM) stores data volitally
  - Once the computer is shut off RAM is gone forever
- Storage stores data non-volitally
  - The data on a HDD/SDD/Floppy disk persists even when power is removed
- The CPU and RAM are much faster than storage
- Storage holds much more data (100x to 1000x) than RAM

Storage is Different Than Memory (cont.)

	Disk	MLC NAND Flash	DRAM
Smallest write	sector	sector	byte
Atomic write	sector	sector	byte/word
Random read	8 ms	75 μs	50 ns
Random write	8 ms	300 μs*	50 ns
Sequential read	100 MB/s	250 MB/s	$>$ 1 GB/s
Sequential write	100 MB/s	170 MB/s*	$>$ 1 GB/s
Cost	$0.04/GB	$0.65/GB	$10/GiB
Persistence	Non-volatile	Non-volatile	Volatile

*Flash write performance degrades over time

Writing to Storage

- The smallest write/read for storage is a sector
- A sector typically consists of 512 bytes
- This means if you want to write 1 byte of data to storage, you have to rewrite all 511 other bytes that haven't changed
  - Writing and reading is time expensive
- To write to storage we "Read-modify-write"
  - Read the entire sector from storage into memory
  - Modify the byte(s) that we want to change in memory
  - Write the entire sector back from memory to storage

Writing to Storage (cont.)

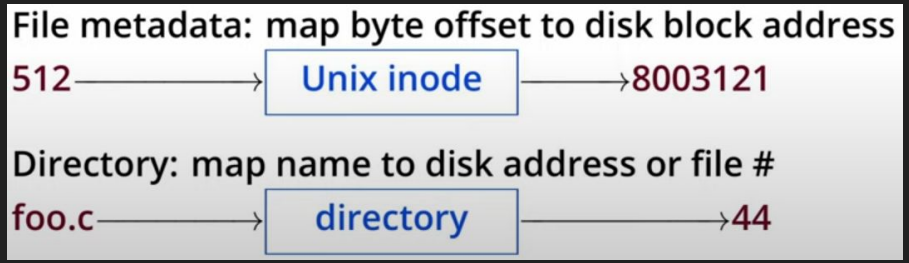
- A sector is the *storage's* unit of atomicity
  - A unit of atomicity is the smallest type/size of data that a device can work with
  - *Memory's* unit of atomicity is a single byte
    - Single bytes can be written/read from memory
- Generally, a physically spinning disk storage has enough physical momentum to complete an entire write of a sector even when the computer experiences a crash
  - This is not always the case though
- Any larger atomic units must be defined by the OS (files)

Files Are Bytes on Storage

- Files are an abstraction of the underlying storage hardware for ease of use for the user
  - When you want to access a file on your computer, you don’t care which cylinder, head, sector it is in or how many sectors it takes up, you just want your file
- File operations for the user:
  - Create a file / Delete a file
  - Read from file / Write to file
- User’s perspective
  - a file called foo.c containing some C code
- File system’s perspective
  - unique identifier (name and offset) that is translated to location(s) on storage
- Storage’s perspective
  - many sectors spread out not necessarily related to each other

A File System Groups Blocks

- File system metadata are data structures that construct mappings to map files to locations on storage
  - Metadata - data about other data
- A file system must be capable of efficiently mapping all the sectors of files or the starting sector of a file and have it persist between shutdowns, crashes, and changes to the file system
- A block is the smallest size of data we can allocate to a file defined by the file system, this may be 1 sector or multiple sectors



File System Realities

- A file system’s performance for accessing files is dominated by number of storage accesses
  - Assume each storage access takes 10ms
  - Access the storage 100 extra times = 1 second of wasted time
  - The CPU can process 3+ billion instructions in that same time

File System Realities (cont.)

- Access time is dominated by physical movement (for hard disk drives):

$$\text{access time} = \text{seek time} + \text{rotational delay} + \#\,\text{bytes} \,/\, \text{disk throughput}$$

- Seek time: the time required by the read/write head to move from one track to another
  - For HDD assume 5ms
  - For SSD assume 0.16ms (SSD has no moving head but still has to find the sector)
- Rotational delay: time required by the read/write head to rotate to the requested sector from the current position
  - For HDD assume 4ms
  - For SSD assume 0ms (SSD has no rotating disk)
- Disk throughput: the speed the storage can read/write in bytes/bits per second
  - For HDD assume 100 MBps (Megabytes per second)
  - For SSD assume 250 MBps
- Calculate the access time for 1 sector (512 bytes) on HDD/SSD and 50 sectors on HDD/SSD
  - 50x the data (on HDD) with only 3% overhead!

File System Realities (cont.)

- All blocks in a file tend to be accessed together and sequentially
  - Nobody opens half of a file unless using a database or virtual memory
- All files in a directory tend to be used together
  - When you use a file in a folder it typically relies on the other files in that folder (games, programs, etc.)
- For these two observations to be useful, the file system should organize files using *contiguous* sectors on storage

Addressing Patterns

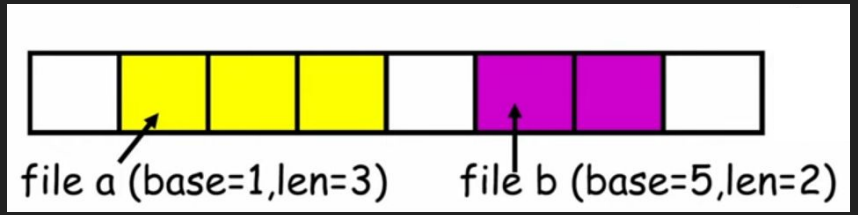
- Sequential
  - Access all blocks of a file one right after the other
  - When you open a file (foo.c) you can read the entire file front to back
- Random
  - Access a specific block of file
  - Uncommon, used for databases and virtual memory
- Keyed access
  - Search for a block with a particular value
  - Uncommon, used for databases

How Files are Tracked

- Storage management:
  - Need to keep track of where file contents exist on storage
  - Need to map files to blocks on storage
  - Index node (inode) is a structure that tracks a file's sectors
  - Inodes must also be stored on storage otherwise they'd be lost forever!
- Intuitions when designing a file structure:
  - Assume *most* files are small
  - Assume *most* of the storage is taken up by large files
  - Assume *most* I/O operations are made to large files

Extent-Based Allocation

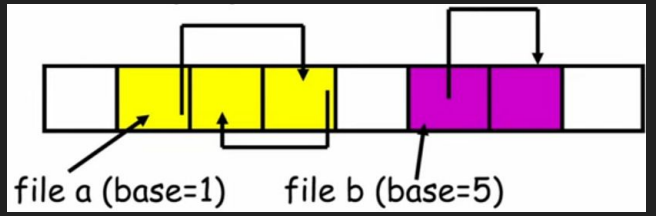
- Extent-based allocation: the *extents* or total size of the file are defined ahead of time and this space is allocated to the file (even if not all of it is used)
  - Inode contains the location in storage and the size of file
- Pro: Provides simple and fast access to files
- Con: Can lead to fragmentation and requires moving files around (time intensive) to store more files



- File A takes up 3 blocks and exists at location 1
- File B takes up 2 blocks and exists at location 5
- 3 blocks are unused and fragmented
- File C requires 2 blocks but cannot be placed anywhere!
  - The only way we could store File C is to move either File A or B to make enough room

Linked Allocation

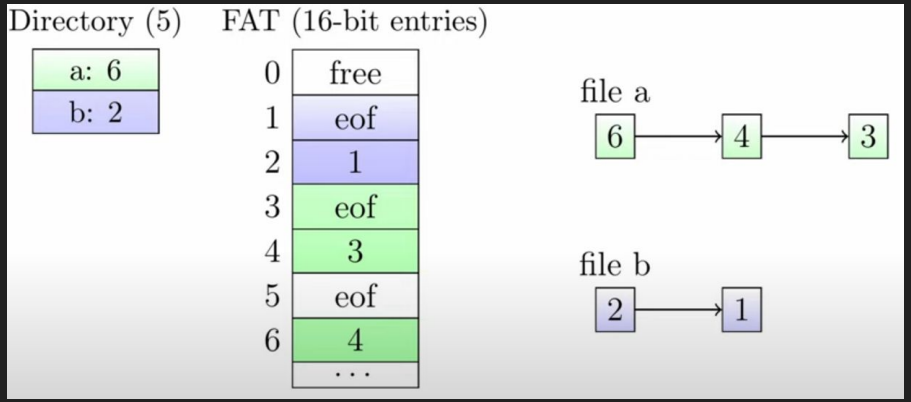
- Linked allocation: each block maintains a link to the next block of allocated storage for the file
  - Inode contains the location of the first location on storage
- Pro: Files can grow or shrink dynamically without fragmentation
- Con: Files are not necessarily sequential (time intensive to access) and reading the next block of the file requires reading the pointer to the next block



- File A takes up 3 blocks and exists at location 1
- File B takes up 2 blocks and exists at location 5
- 3 blocks are unused and fragmented
- File C requires 2 blocks and could now exist at locations 0 and 4!

FAT File System Simplified

- Uses linked files where links reside in the File Allocation Table (FAT) and directories maintain the starting links to files
  - Directory (5) has links to files A and B, at FAT entries 6 and 2, respectively
  - Entry 6 points to 4, entry 4 points to 3
- In FAT16 each entry is 16-bits so the entire FAT is easily stored in RAM for quick access
  - The entire FAT is usually kept in memory to allow quick access of the file structure
  - Remember, the FAT only keeps track of links to files NOT the contents of the files themselves



FAT File System Simplified (cont.)

- FAT12, FAT16, FAT32 use 12, 16, 32 bit entries, respectively
- Max file system sizes:
  - FAT12: 32 MB (256 MB for 64 KB clusters)
  - FAT16: 2 GB (4 GB for 64 KB clusters)
  - FAT32: 2 TB (16 TB for 4 KB sectors)

FAT File System Simplified (cont.)

- Alternatively, we could increase the block size (say 1024 bytes per block), but if files are routinely smaller than our block size, we lose some performance
- For reliability, the FAT is usually duplicated, so in case of corruption there is always a backup to prevent losing links to files
  - Have you ever unplugged a flash drive and had files get corrupted? The links were not all able to be saved!

FAT File System Simplified (cont.)

- A FAT file system contains 4 regions:

  1. Reserved sectors - Contains boot sector which has information about the file system and stores the operating system’s bootloader

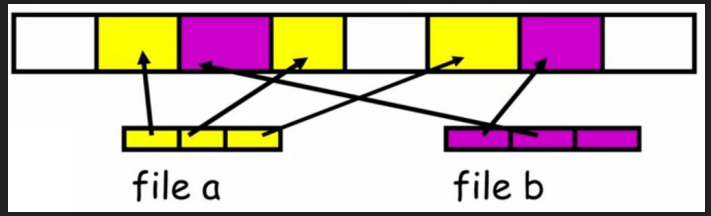
  2. FAT Region - Contains the actual File Allocation Table and any duplicates

  3. Root Directory - Contains information about the root directory and files/directories in the root directory (FAT12 and FAT16 only)

  4. Data Region - Contains all file and directory data (FAT32 uses this to store the root directory)

Indexed Allocation

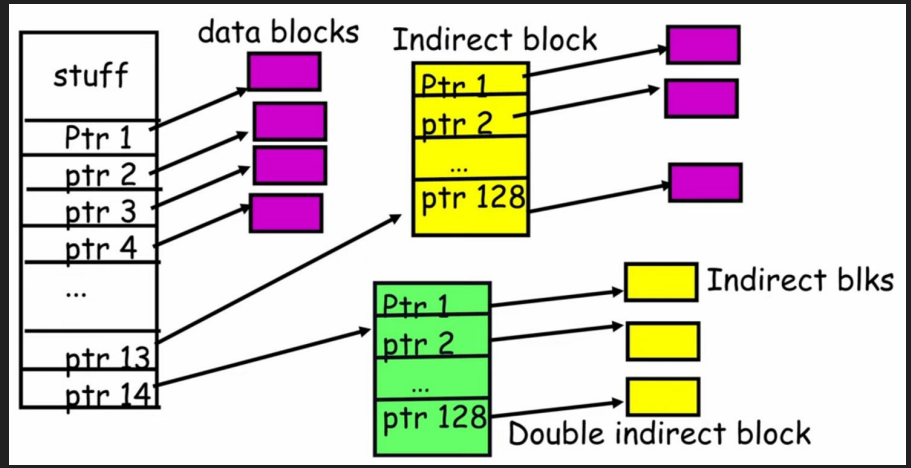
- Indexed allocation: an array of pointers points to all the blocks allocated for a particular file
  - This array is essentially the inode
- Pro: Sequential and random access is done easily (just follow wherever the pointer points to for a given block)
- Con: The table storing the arrays require large amounts of contiguous memory so fragmentation is still a problem



- File A takes up 3 blocks and a reference to each block is stored in an array
- File B takes up 2 blocks and a reference to each block is stored in an array
- File C requires 2 blocks and could now exist at locations 0 and 4 and have an array pointing to those locations

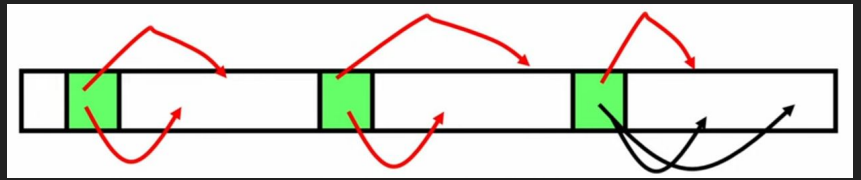
Multi-Level Indexed Allocation

- The inode is comprised of 14 block pointers and “stuff” (metadata)
- The first few pointers point directly to data blocks
  - Assume 4 pointers
  - Small files (up to 4 blocks or 2048 bytes) can now be easily accessed
- The remaining pointers point to more pointers to allow for larger files to exist
  - A pointer that points to a pointer is an indirect block
  - A pointer that points to a pointer, that points to a pointer, is a double indirect block
- The last few pointers (13 and 14) are used for indirect and double indirect blocks



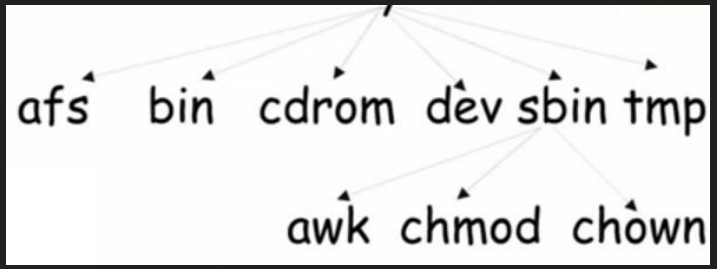
Inode Placement and Storage Structure

- Consider the two options on the right:
  - Fixed size inode array at the start of storage
    - Everytime you want to read a file you have to seek to the front of the disk, then seek to the file (slow)
  - Inodes collocated with the blocks they reference
    - Seek times reduced significantly, best case scenario is 0ms seek time
- Inodes can also be referenced by number which translates to their location in the array
- A superblock appears at the front of the storage which gives system information



Hierarchical Unix

- A single namespace describes the root directory
  - Any additional file systems (like cdrom) can be accessed like any other file
- Directories are essentially just special files
  - Directories cannot be directly modified, only created, destroyed, or linked to other files
    - Enforced by the OS
  - Reading a directory retrieves the list of files or directories within it
    - Retrieves a name and inode number for each file or directory



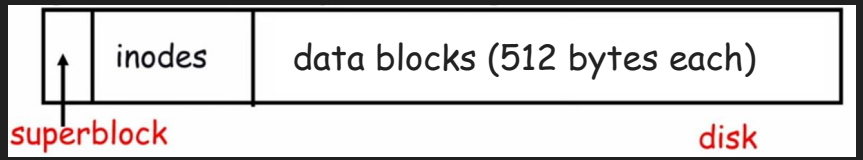
<name,inode#>
<afs,1021>
<tmp,1020>
<bin,1022>
<cdrom,4123>
<dev,1001>
<sbin,1011>
⋮

Naming Conventions

- Where is the root directory?
  - Inode #2 always points to the root directory
  - Inodes #0 and #1 are historically preserved
- Special names:
  - Root directory: “/”
  - Current directory: “.”
  - Parent directory: “..”
- The shell or OS may implement other special naming:
  - User’s home directory: “~”
  - Searching for “foo.*” returns all files starting with “foo.”
- To navigate the entire file system we only need:
  - cd *name*: move into directory *name*
  - ls: display all file and directory names in current directory

Original Unix Format

- Data blocks (files)
- Inodes linked to files (hard link)
- Superblock
  - Specifies # of blocks, max # of files, and pointer to head of free list
- Con: This is very slow to access (down to 2% of maximum storage throughput)



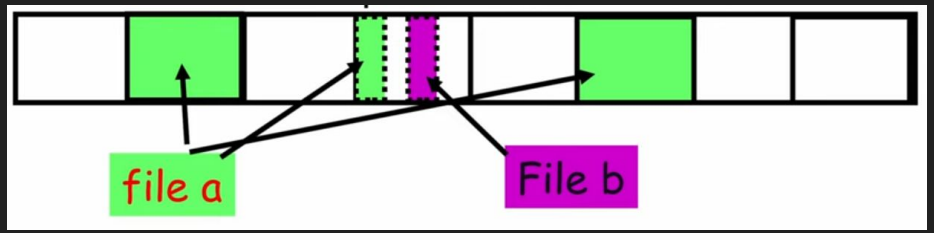
Issues With The Original Unix Format

- Blocks too small
  - Most files leave fragments of blocks unless they are exactly 512 byte multiples
  - To read a file, you retrieve one block at a time (slow)
- Poor clustering
  - File blocks are randomly thrown around storage
  - Inodes are far away from file blocks
  - Inodes for directories may be far away from each other in the inode region
  - Using “ls” or “grep foo *.c” (search) takes forever going back and forth between inode array and randomly placed file blocks

Solving Block Size Problem

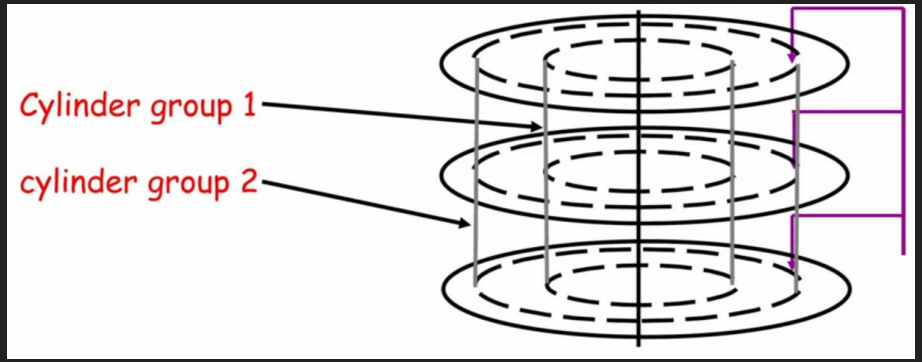
- Block size is too small
- Make it bigger!
  - Doesn’t completely solve the problem
- Put fragments into dedicated block using 4096 byte blocks
  - Now small files, and tail end of large files, can be stored in the fragment blocks
  - These fragments are 1024 or 2048 bytes

Block size	space wasted	file bandwidth
512	6.9%	2.6%
1024	11.8%	3.3%
2048	22.4%	6.4%
4096	45.6%	12.0%
1MB	99.0%	97.2%



Solving Seek Time Problems

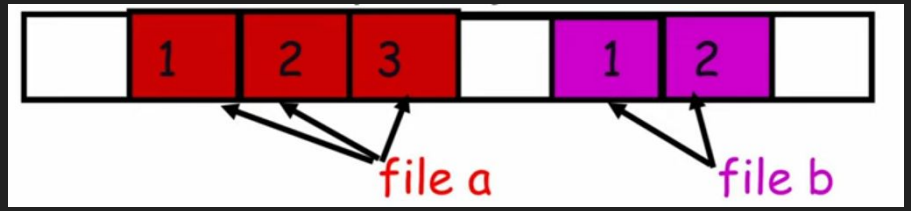
- Put related data into the same cylinder group
- A cylinder group is the same vertical slice of storage locations across all disks in a platter
- Now the head does not have to seek at all to read multiple blocks of a files data



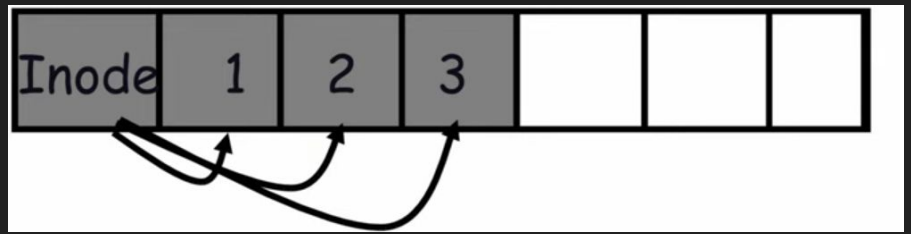
Solving Clustering Problems

- Try to keep sequential blocks in adjacent sectors
- Try to keep inode in same cylinder as file data

Sectors:

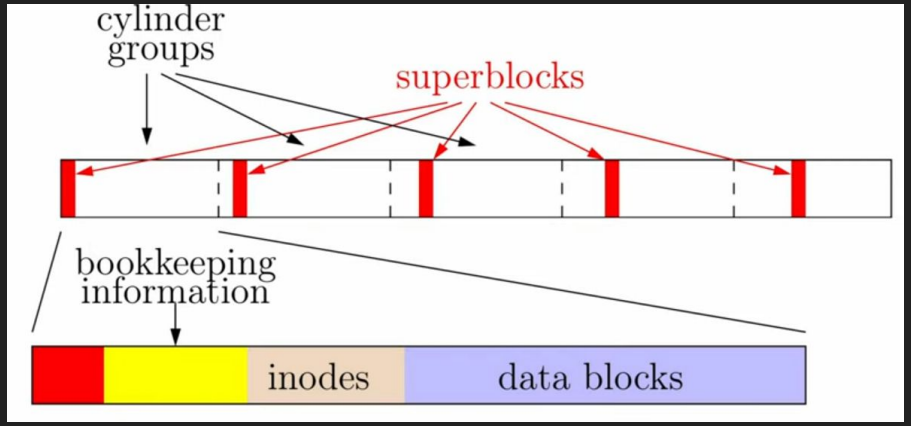


Cylinders:



Big Picture Layout

- Each cylinder group acts as a mini file system
  - Superblock
  - Bookkeeping
  - Inodes
  - Data blocks
- Files in a directory are collocated in the same cylinder group
- Growing files grow within the same cylinder group
- Files over 1MB get remainder sent to different cylinder group
- New directories can have a completely new cylinder group created just for them!
- This system is much faster and more space efficient



09 - Input and Output Systems

CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

Input and Output

- An input is some device that provides data to the computer
  - Keyboard, mouse, microphone, webcam
- An output is some device that receives data from the computer
  - Monitor, headphones, printer, graphics card
- Together, these are called I/O for short
- The term “peripheral” or “device” refers to any I/O device
- Your hard drive, floppy drive, and SSD are I/O devices too!

I/O Overview

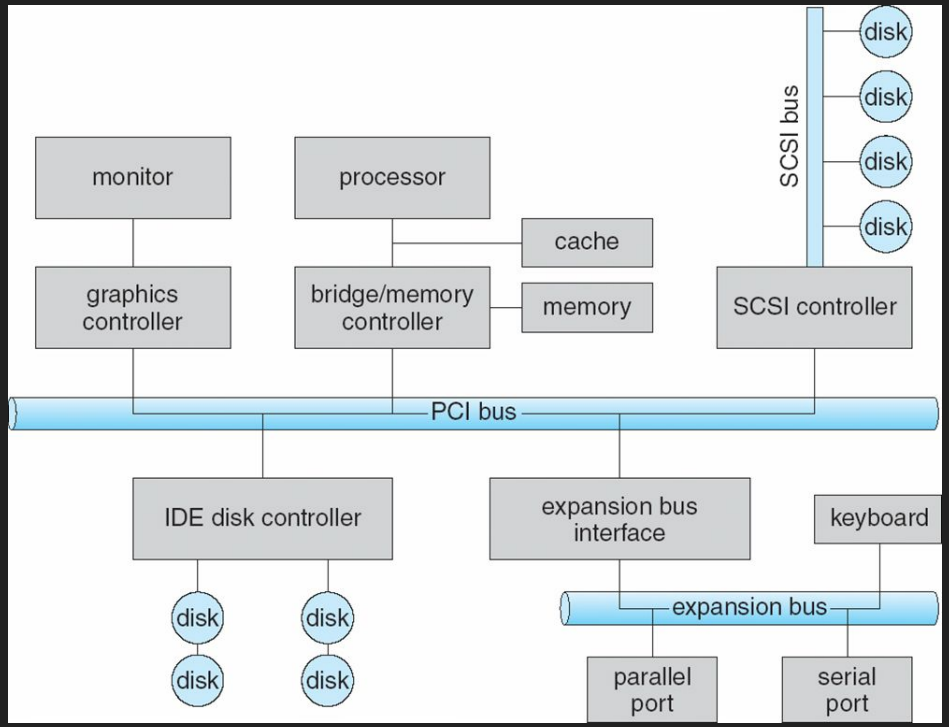
- I/O management is a major component of operating system design and operation
  - Important aspect of computer operation
  - I/O devices vary greatly
  - Various methods to control them
  - Performance management
  - New types of devices frequent
- Ports, busses, device controllers connect your computer to various devices
- Device drivers encapsulate device details
  - Present uniform device-access interface to I/O subsystem

I/O System Architecture

- System Bus: links multiple devices directly to the CPU for communication
- Device Port: the interface which a CPU accesses devices
  - There are typically 4 types of device ports or registers:
    - Status port: indicates if a device is busy, ready, or has an error
    - Control port: allows the device to receive some command from the CPU (i.e. read a sector vs write a sector for a hard drive)
    - Data-in port: data sent from the device to the CPU
    - Data-out port: data sent from the CPU to the device
    - These ports are typically accessed using 1 byte or 1 word using a single machine language instruction
- Controller: a piece of hardware that receives data over the system bus and translates them into actions performed by the device
- The device itself

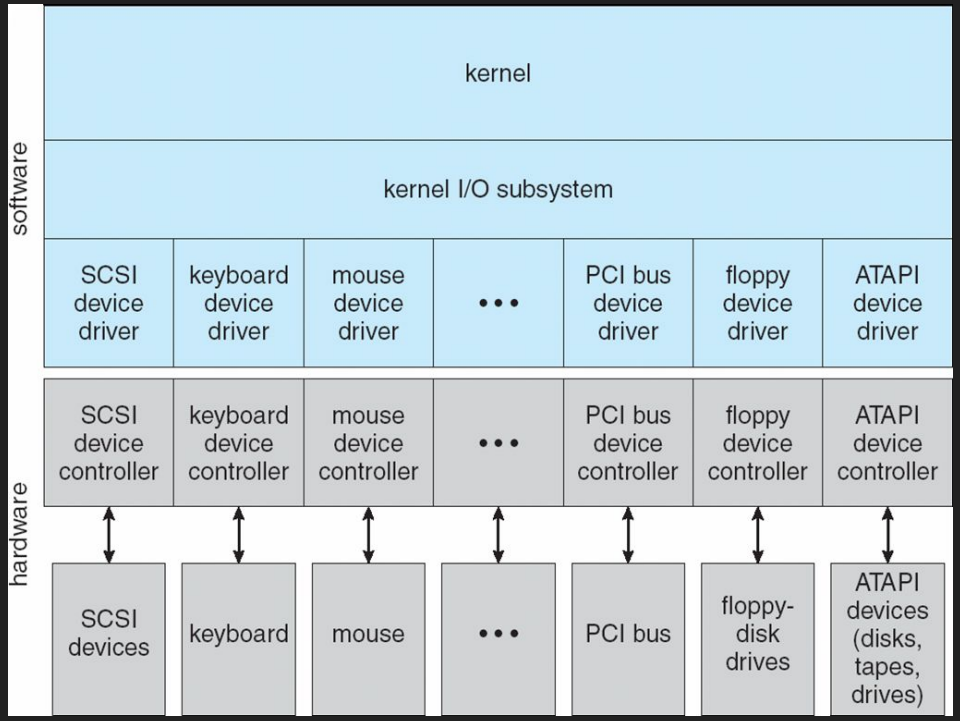
PCI Bus Structure

- Many controllers are all connected to a single bus
- These controllers may control multiple devices
- Through the bus and their controllers, the devices can communicate with the CPU



Kernel I/O Structure

- Your kernel interfaces with devices through drivers
- A driver is a software component that lets the operating system and a device communicate with each other
- The drivers provide a standard interface for our devices
- Drivers must know the specific commands a device controller understands



I/O Port Locations

- Port locations are sometimes hard coded by the ISA
  - The diagram shows common port addresses for x86
- Device controllers are listening to specific address ranges

I/O address range (hexadecimal)	device
000–00F	DMA controller
020–021	interrupt controller
040–043	timer
200–20F	game controller
2F8–2FF	serial port (secondary)
320–32F	hard-disk controller
378–37F	parallel port
3D0–3DF	graphics controller
3F0–3F7	diskette-drive controller
3F8–3FF	serial port (primary)

Accessing Devices with the OS

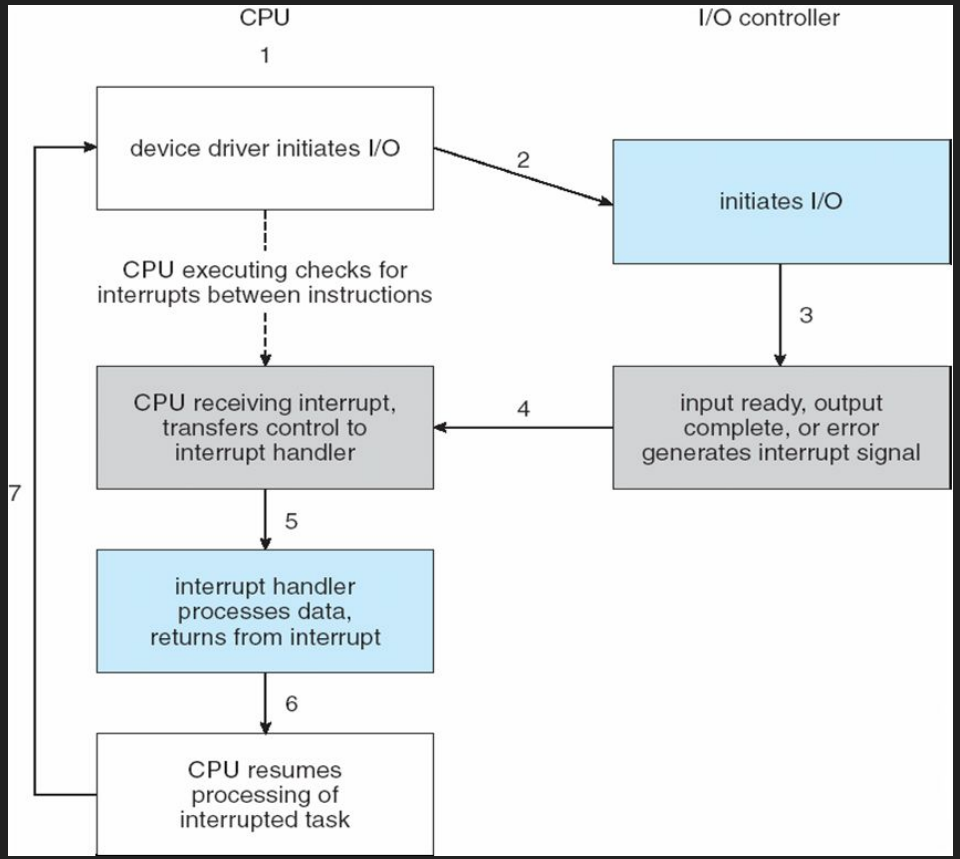
- Unless you are writing drivers, accessing ports directly is not necessary
- Most OSs provide a more simplified interface for accessing devices
  - Ease of use: On Unix, devices appear as files in the /dev directory
  - Access control: The OS must limit how and when devices are accesses to prevent errors or malicious programs
  - Buffering and caching: Allows for device data to be efficiently accessible
  - Scheduling: If many programs want access to the same device, scheduling access may be necessary (think about the hard drive!)
  - Error handling and failure recovery

Polling

- CPU waits until a device’s controller status is ready
- CPU sets the command register and data registers on the device
- Controller indicates it is busy and performs the command
- If the operation succeeds, the controller changes its status to ready
- CPU reads the data port from the controller if the CPU sent an input command
- Restart back at step 1 if CPU needs to send another command
- If the device is very slow the CPU is going to wait for a long time doing nothing by using polling!

Interrupts

- Rather than wasting CPU time by waiting, we can use interrupts
- An interrupt will *interrupt* the CPU when it completes the I/O operation
- The CPU is forced to handle the I/O operation immediately
    - 1. Determine which device caused the interrupt
    - 2. If the last command was an input operation, retrieve the data from the device register
    - 3. Start the next operation for the device
    - 4. Resume executing code previous to the interrupt



Interrupt Vector Table

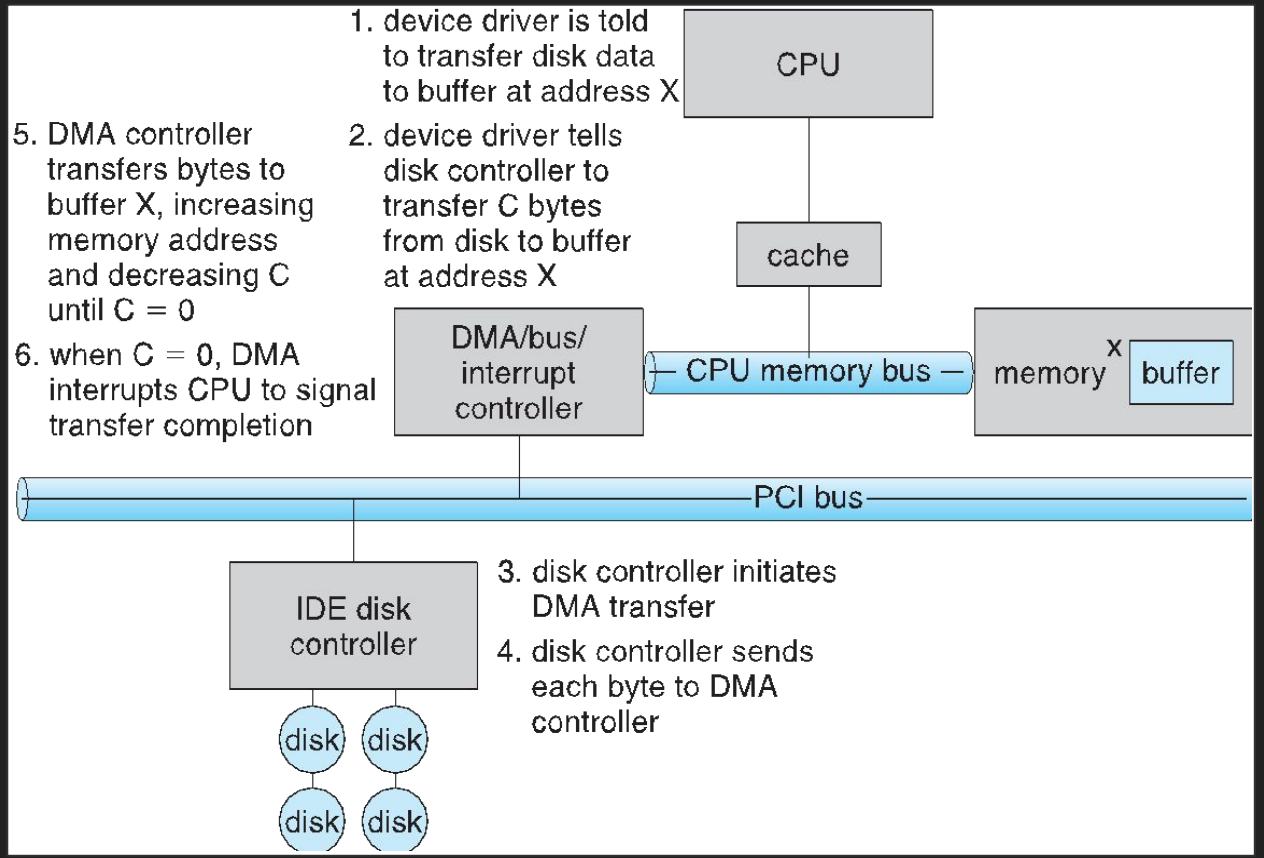
- The interrupt vector table provides addresses for routines that get executed when interrupts occur
  - Software interrupts (TRAPs and exceptions)
  - Hardware interrupts

vector number	description
0	divide error
1	debug exception
2	null interrupt
3	breakpoint
4	INTO-detected overflow
5	bound range exception
6	invalid opcode
7	device not available
8	double fault
9	coprocessor segment overrun (reserved)
10	invalid task state segment
11	segment not present
12	stack fault
13	general protection
14	page fault
15	(Intel reserved, do not use)
16	floating-point error
17	alignment check
18	machine check
19–31	(Intel reserved, do not use)
32–255	maskable interrupts

Direct Memory Access

- Many devices want to transfer large amounts of data at a time
  - Receiving data using single bytes from ports is too slow!
  - Imagine a high quality webcam only being able to transfer 1 byte at a time
- Direct Memory Access (DMA) allows devices to read/write directly from/to memory without involving the CPU
- DMA requires a DMA controller that will map inputs/outputs from devices to specific areas of memory
- The CPU tells the DMA controller the source (specific device) and destination (memory address) for I/O accesses in memory
- Once the DMA controller has transferred all device data to memory, it interrupts the CPU so the CPU can go fetch that data
- The DMA controller and CPU are now competing for memory bus time but is still more efficient than requiring the CPU to read individual bytes from the device

Direct Memory Access (cont.)



Application Programmer’s View of I/O Devices

- An OS provides a high level interface for working with devices
- Device characteristics:
  - Transfer units: character or block
    - A block device reads/writes many bytes (hard drive)
    - A character device reads/writes single bytes (keyboard)
  - Access methods: sequential or random
    - Is data accessed in a specific order or not
  - Timing: synchronous or asynchronous
    - Most devices are asynchronous
    - Some embedded systems are specifically designed to work with one I/O device, this access may be synchronous
  - Access protection: shared or dedicated
    - Hard drive is shared amongst many processes or users
  - Speed: fast or slow
    - Keyboard is very slow, graphics card is very fast
  - Operations: input, output, or both

Block and Character Devices

- Block devices include disk drives
  - Commands include read, write, seek
  - Raw I/O, direct I/O, or file-system access
  - Memory-mapped file access possible
    - File mapped to virtual memory and clusters brought via demand paging
  - DMA
- Character devices include keyboards, mice, serial ports
  - Commands include get(), put()
  - Libraries layered on top allow line editing

I/O Buffering

- I/O devices typically contain a small on-board memory where they can store data temporarily before transferring to/from the CPU.
- A disk buffer stores a block when it is read from the disk.
- It is transferred over the bus by the DMA controller into a buffer in physical memory
- The DMA controller interrupts the CPU when the transfer is done.

Why Use I/O Buffering Inside the OS?

- To cope with speed mismatches between device and CPU.
  - Example: Receive file over network (slow) and store to disk (faster)
- To cope with devices that have different data transfer sizes.
  - Example: ftp brings the file over the network one packet at a time. Stores to disk happen one block at a time.
- To minimize the time a user process is blocked on a write.
  - Writes => copy data to a kernel buffer and return control to the user program. The write from the kernel buffer to the disk is done later.

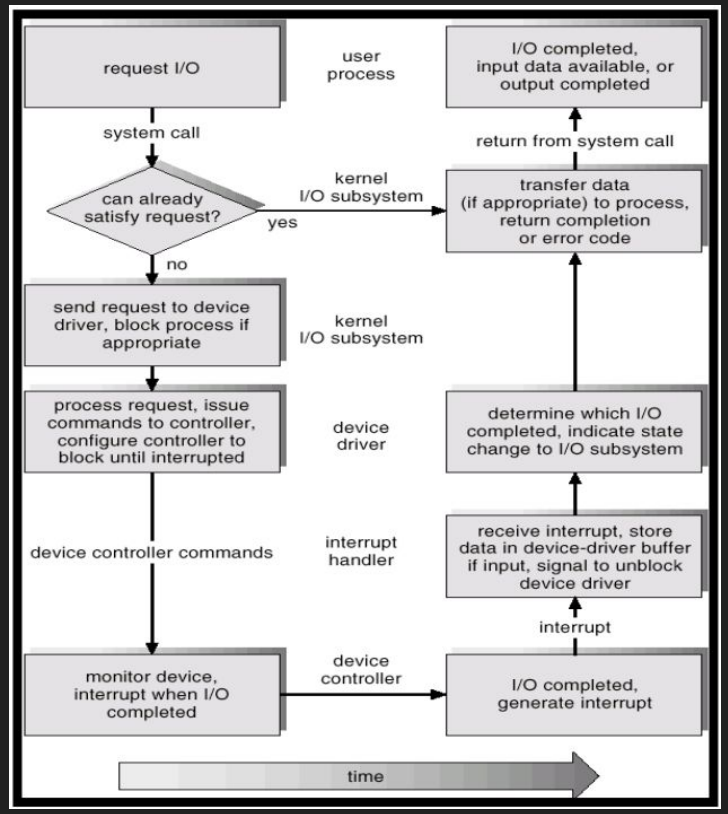
Caching

- Improve disk performance by reducing the number of disk accesses.
- Idea: keep recently used disk blocks in main memory after the I/O call that brought them into memory completes.
  - Example: Read (diskAddress)
    If (block in memory) return value from memory
    Else ReadSector(diskAddress)
  - Example: Write (diskAddress)
    If (block in memory) update value in memory
    Else Allocate space in memory, read block from disk, and update value in memory
- What should happen when we write to a cache?
  - Write-through policy (write changes to memory and disk). High reliability. More common for older systems
  - Write-back policy (write changes to memory and write to disk some time later). Faster. More common for newer systems

OS Processing a Read Call

- User process requests to read the disk
- The OS checks if the data exists inside a cache
- If it is not in the cache:
  - OS tells the device driver to perform input
  - Device driver tells the DMA controller what to do and blocks itself
  - DMA controller transfers the data to the kernel buffer
  - DMA controller interrupts the CPU when transfer is complete
- OS transfers the data to the user process and places the process in the ready queue
- The the process gets CPU access, it resumes executing after the read call

OS Processing a Read Call (cont.)



IO Performance

- IO accesses are computationally expensive
  - Most IO devices are slow
  - Contention from multiple processes
  - IO is typically handled through system calls and interrupt handling which adds extra time
- How to improve performance:
  - Caching
  - Reduce interrupt frequency by using larger data transfers
  - DMA controllers
  - Use multiple IO devices to reduce contention
  - Increase RAM to reduce paging

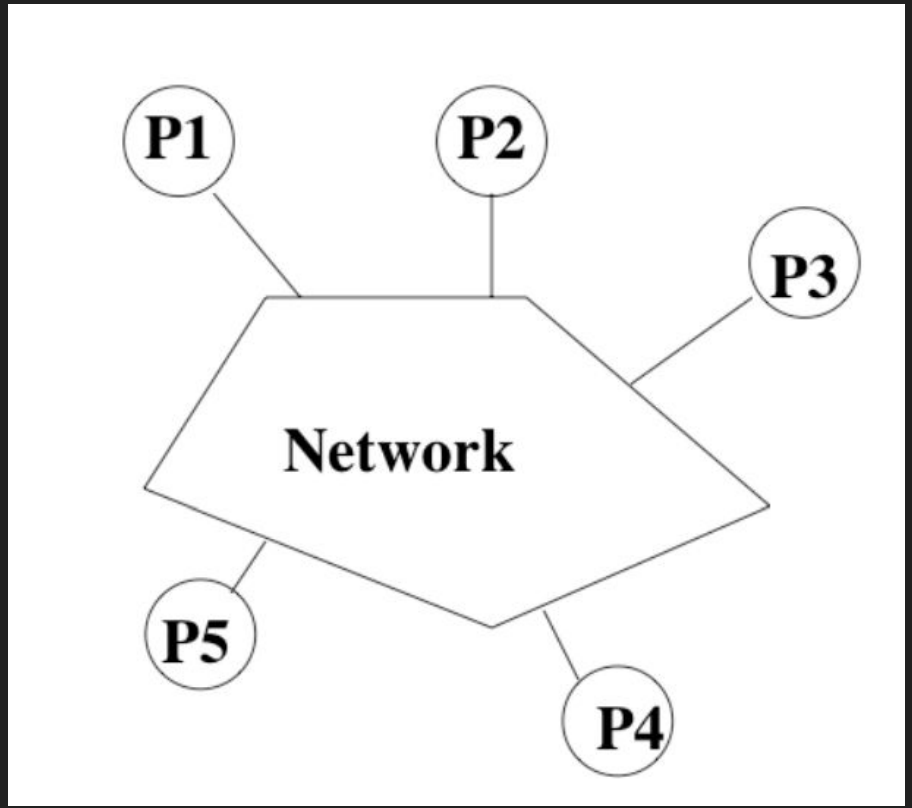
10 - Networks and Distributed Systems

CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

Distributed Systems

- A distributed system is a set of physically separate processors connected by one or more communication links
- Nearly all systems today are distributed in some way
  - Email, file servers, network printers remote backups, web sites, online gaming



Parallel Systems

- Parallel Systems
  - Tightly coupled
  - Processors share clock, memory, and are all running the same OS
  - If your computer has multiple cores, it is a parallel system
  - Frequent communication over fast system bus
- Distributed Systems
  - Each processor has its own clock, memory, and OS
  - Communication is less frequent and over a network
    - Networks are even slower than disks!

Advantages of Distributed Systems

- Resource sharing
  - Resources don’t need to be replicated for each processor
    - Shared files on a network can be accessed by multiple computers but reside on a single server
  - Expensive or scarce resources can be shared
    - Printers accessed by multiple computers over a network
  - Each processor can present the same environment to the user
    - A user could login to their system from any device on the network

Advantages of Distributed Systems (cont.)

- Computation speedup
  - *n* number of processors can potentially give you *n* times the computational power
    - Tasks must be able to be broken down into subtasks
    - Processors must be synchronized and communicate with each other efficiently

Advantages of Distributed Systems (cont.)

- Reliability
  - Replication of resources builds fault tolerance
    - If you only have 1 server and it goes down, there is no backup
    - If you have multiple servers with the same resources and data, the loss of 1 server could be recovered
      - Google has thousands of systems to avoid down time
    - If there is a central or critical component of the system, this could introduce a single point of failure
      - A shared file system
      - Version control, git, svn

Advantages of Distributed Systems (cont.)

- Communication
  - Users and processes can communicate with each other
  - One processor can be on one side of the world and still communicate with processors on the other side
  - Email/messaging, banking systems, online gaming, web sites, message boards

Distributed Systems Considerations

- Communication systems and networks
  - Internet? Local network?
- Transparency
  - Does the system hide its distributed nature from its users?
  - Google is very transparent, to the end user it “feels” like one system
- Security
  - How does the system protect other users or the system itself?
- Reliability
  - “How many 9’s?” - 5 nines reliability says a system is required to be fully operational 99.999% of the time
    - 31,536,000 seconds in a year with 5 nines reliability means the system should only go down less than 315 seconds every year
- Performance and scalability
  - Applications must be designed to break big tasks into subtasks
  - Should the application and system be designed to service 10 simultaneous users? 100? 1,000? 100,000? 1,000,000+?
- Programming models
  - What techniques should developers use to properly write applications for these systems to ensure all of the above?

Distributed System Design

- What are the challenges when moving from a standalone system to a distributed system?
  - Resource sharing
    - Resources are limited and many processors might all want access
  - Timing and synchronization
    - In the real world clocks aren't exactly synchronized
      - Make compiles your programs based on last modified date
      - If your development machine's date is out of sync with a build machine's date, the build machine will not recognize your changes
  - Critical sections
    - How do we reconcile two processors attempting to modify the same thing at the same time?
  - Failure recovery
    - Can your application withstand a failure of some of the machines on the network? How is this managed? Performance downgrade or a complete halt?

Networks

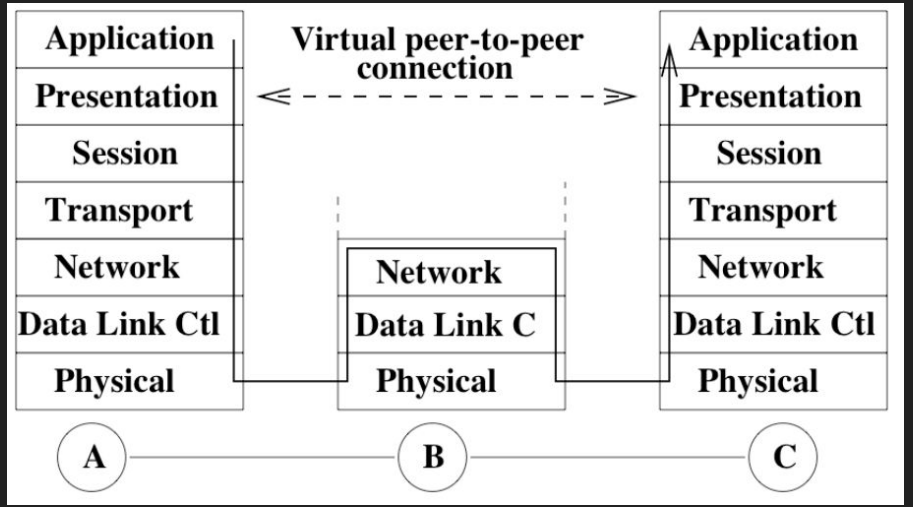
- Networks provide efficient, correct, and robust message passing between multiple machines
- Local Area Network (LAN) connects nodes in a single building and needs to be fast and reliable (Ethernet)
  - Uses twisted pair (ethernet cable), coaxial, fiber optics
  - Throughput: 10 Mb/s to 100,000 Mb/s
- Wide Area Network (WAN) connects nodes across the state, country, or planet
  - WANs (e.g. internet) are slower and less reliable than LAN
  - Private WANs may be used by global companies or governments
  - Uses satellite channels, cellular networks (4G, 5G, LTE), telephone lines
  - Throughput: 1 Mb/s to 20,000 Mb/s
- Context: 25 Mb/s is recommended for watching a 4K video stream

Communication Over Networks

- Data chopped up and is sent in “packets”, the network’s basic transmission unit
  - A single file sent over a network is made of many packets
- There is a source node that sends packets and a destination node that receives packets
  - The packet may pass through multiple machines along the way (e.g. routers)
  - When you connect to any website, there are often many routers that pass your packets to and from the final destination
- The road analogy
  - Cars = Packets
  - Highways = Network
  - Road Signs = Routers
    - Routers ensure traffic is flowing even when there are large amounts of data coming through
    - Shared resources can lead to contention (traffic jam)

Communication Protocols

- A protocol is a set of rules for communication that all parties agree to
  - As long as two devices speak the same protocol they can have different OSs and hardware but still communicate
  - IP, http, https, ftp, etc.
- A protocol stack is how networking software and hardware is structured into layers
  - Layers application to session are provided by the user application
  - Layers transport and lower are provided by the OS
  - Physical layer is provided by the physical electronic components that transmit the raw bits in the form of electrical signals

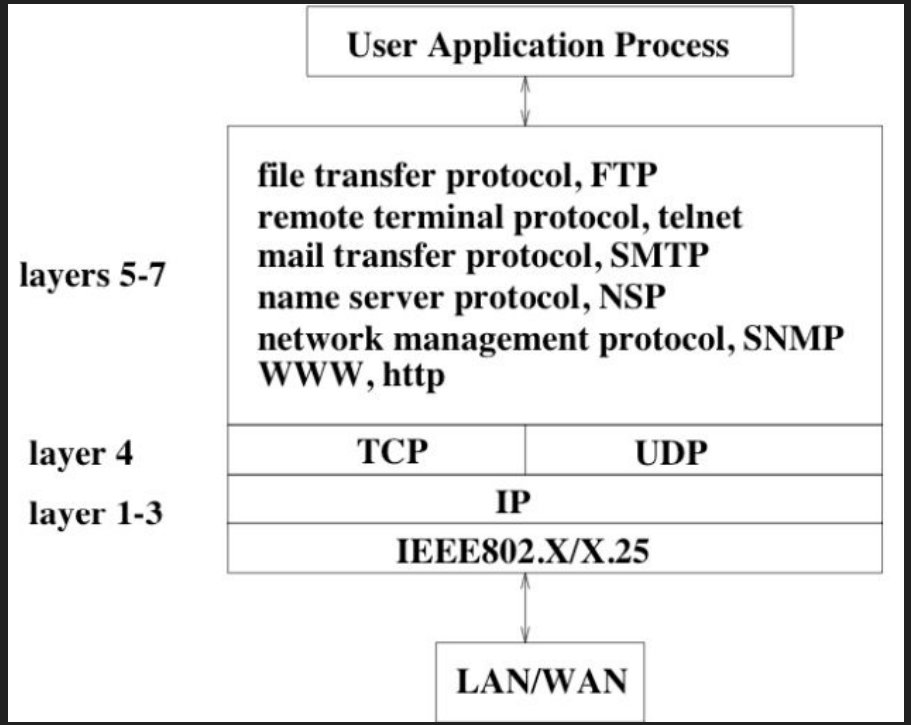


ISO Network Protocol Stack

- Application layer
  - applications that use the net, e.g., Outlook, Chrome, Xservices, ftp, telnet, these provide a UI
- Presentation layer
  - data format conversion, e.g., big/little endian integer format)
- Session layer
  - implements the communication strategy, such as RPC. Provided by libraries.
- Transport layer
  - reliable end-to-end communication between any set of nodes. Provided by OS.
- Network layer
  - routing and congestion control. Usually implemented in OS.
- Data Link Control layer
  - reliable point-to-point communication of packets over an unreliable channel.
  - Sometimes implemented in hardware, sometimes in software (PPP).
- Physical layer
  - electrical/optical signaling across a “wire”. Deals with timing issues. Implemented in hardware.

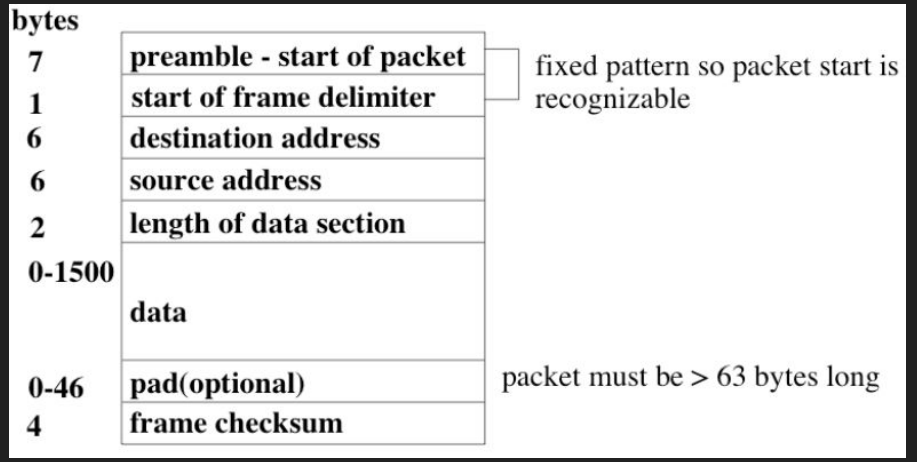
TCP/IP Protocol Stack

- Most websites use Transmission Control Protocol (TCP)/Internet Protocol (IP)
  - These protocols are used in the transport layer
  - Fewer layers than ISO for efficiency
  - Consists of a suite of protocols UDP, TCP, IP, etc.
  - TCP is a reliable protocol, packets are received in the order they are sent
  - UDP is an unreliable protocol, packets are not guaranteed to be delivered in any order or even at all!
- IEEE802 is at the physical layer



Packets

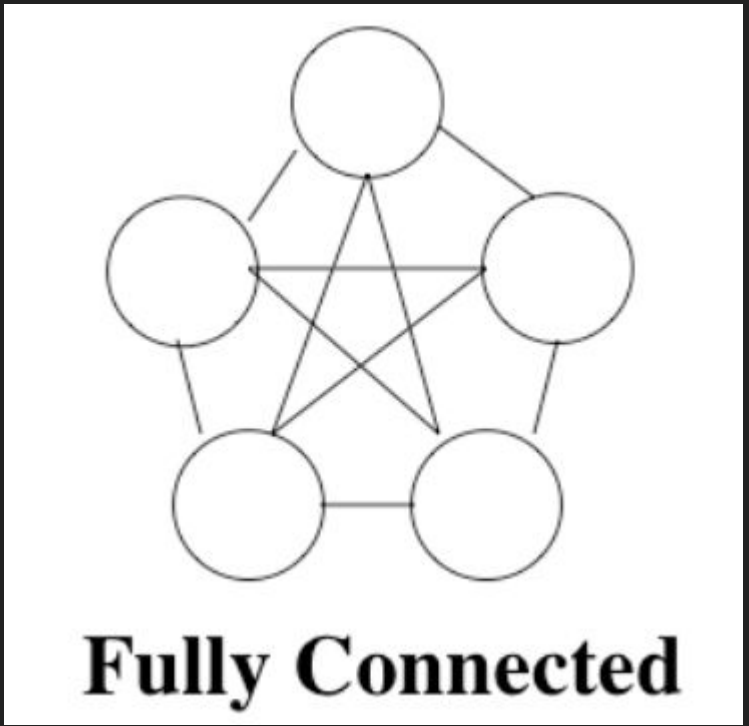
- Each message sent over the network is chopped into packets
  - Each packet contains enough information to recreate the original message
    - Ex. if packets arrive out of order, the destination node should be able to reorder the packets
  - The data section contains headers for higher protocol layers and actual application data



Ethernet Packet

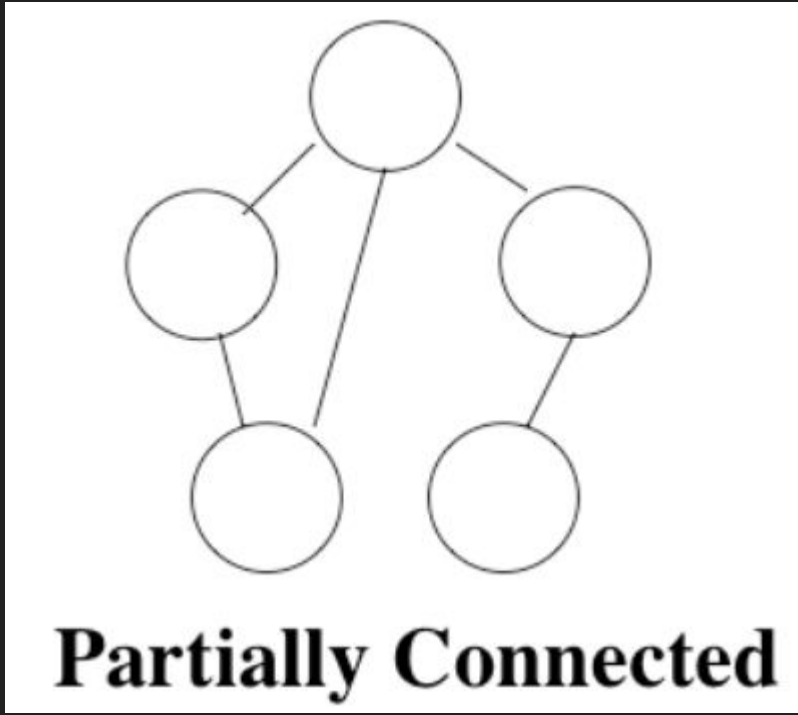
Point-to-Point Network Topologies

- Fully connected topology
  - All nodes connected to all other nodes
  - Each message goes directly from source to destination
- If one node fails, the other nodes can still communicate with other nodes
- This topology is extremely expensive



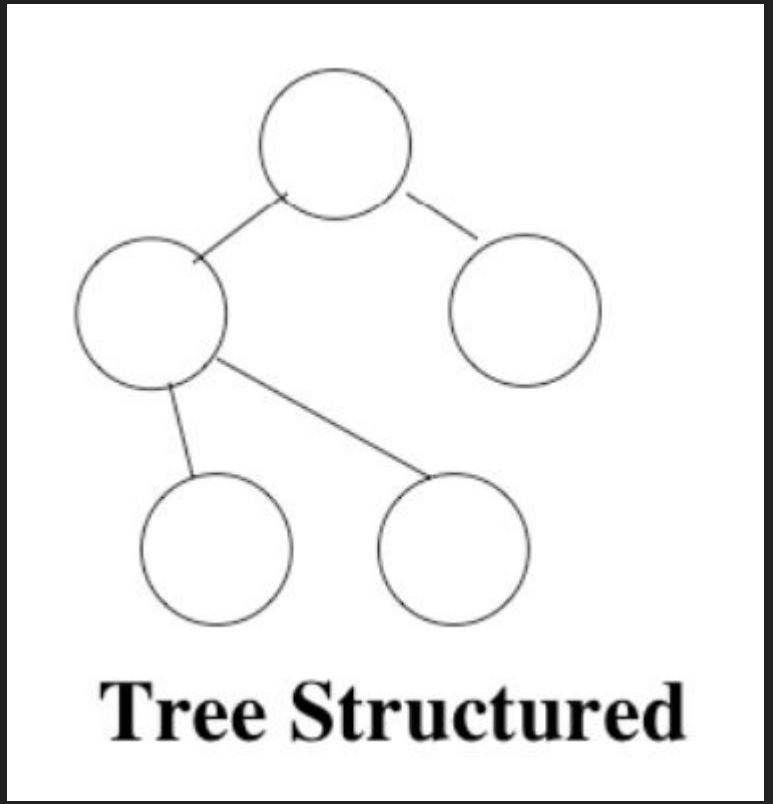
Point-to-Point Network Topologies (cont.)

- Partially connected topology
  - Links between some, but not all
- If one node fails, the network may become partitioned or separated
- Sending a message may require a routing algorithm to select which nodes the message passes through



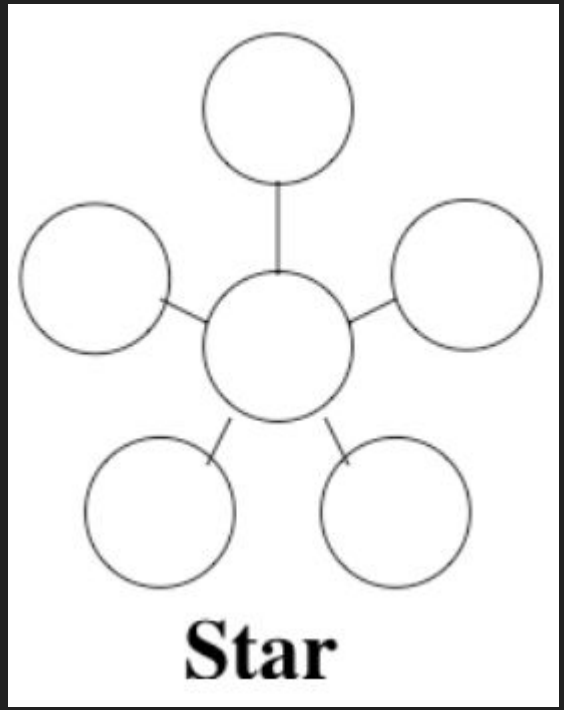
Point-to-Point Network Topologies (cont.)

- Tree structure topology
  - Links exist in a hierarchy
- All messages between parent and child are fast, but “cousins” must communicate through a common ancestor
- Not tolerant of failures, one node can break the communication chain for many



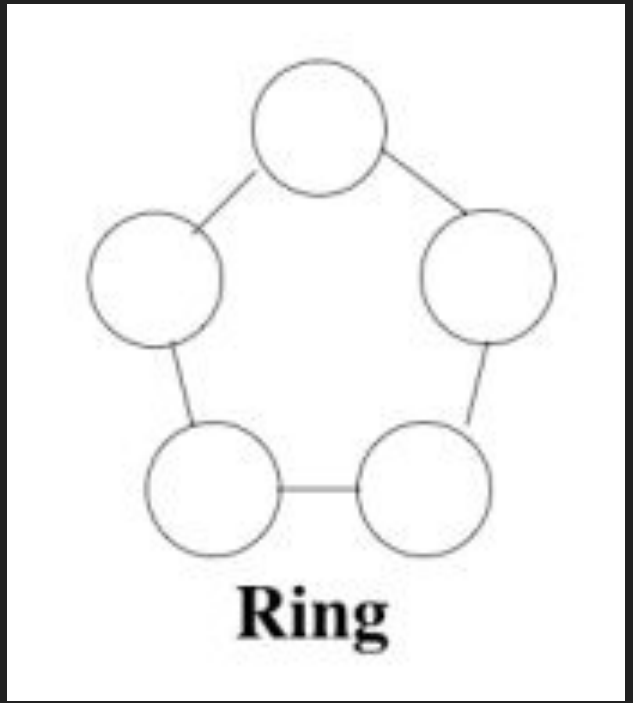
Point-to-Point Network Topologies (cont.)

- Star topology
  - All nodes connected to a centralized node
- The central node is generally dedicated to network traffic
- Each message only takes two hops
- If the central node fails, the entire network is lost
- Inexpensive, often used for LAN



Point-to-Point Network Topologies (cont.)

- Ring topology
  - Used by old LAN networks
  - Inexpensive, but failure prone
- One directional ring
  - Nodes can only send data in one direction
  - Maximum *n-1* hops given *n* number of nodes
  - One failure partitions the network
- Bidirectional ring
  - Nodes can send data in both directions
  - Maximum *n/2* hops given n number of nodes
  - Two failures can partition the network

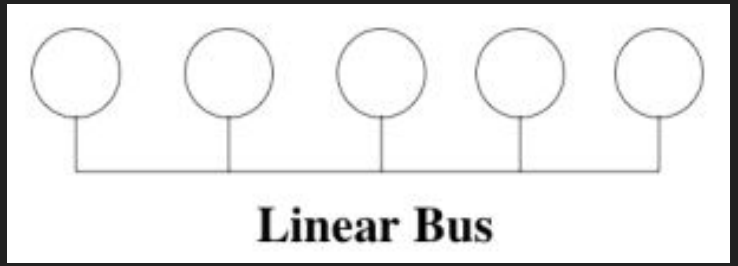


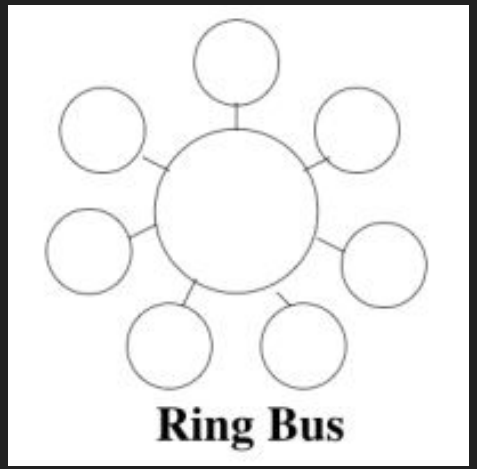
Point-to-Point Network Topologies (cont.)

- Doubly Linked Ring topology
  - Connections between neighbors and second neighbor
  - Expensive, but more tolerant of failures
  - Maximum *n/4* hops given *n* number of nodes

Point-to-Point Network Topologies (cont.)

- Bus nodes connect to a common network
  - Two devices sending packets that the same time is a constant problem
  - This is a bus collision
- Linear bus
  - Single shared link
  - Nodes directly connected to each other
  - Inexpensive and tolerant of node failures
  - Ethernet uses this
- Ring bus
  - Single shared circular link
  - Differs from star topology in that only one packet can be sent at a time





Resource Sharing for Distributed Systems

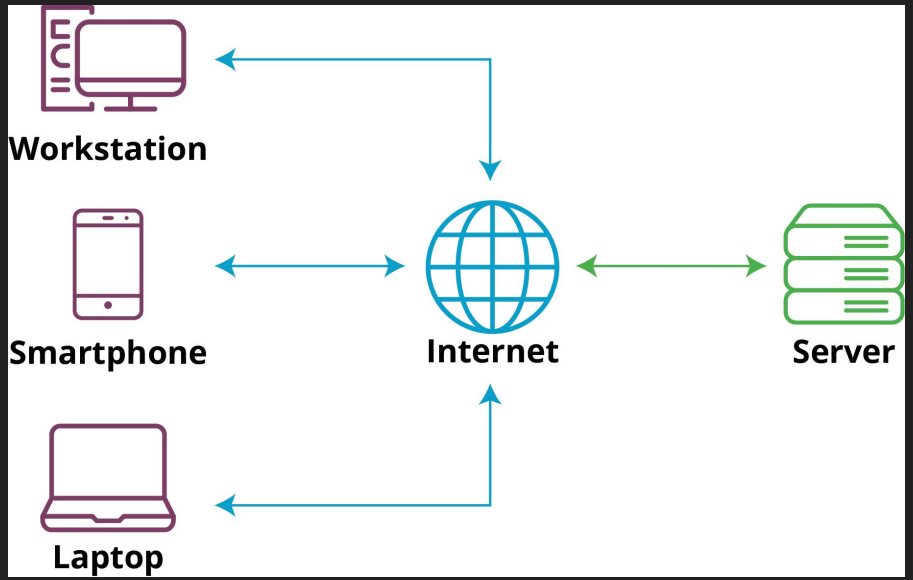
- Mechanisms of sharing resources (hardware, software, data):
  - Data Migration
    - Moving data between systems
    - Email, messages, file transfers
  - Computation Migration
    - Move the data to the a computation
    - Search engines take your query, compute, and provide results
    - ChatGPT requires powerful hardware to execute which is hosted by a server that you send messages to
  - Process Migration
    - Move the process (computation and data) or part of a process
    - SETI@home uses your processor for compute, then sends results back to a server

Resource Sharing for Distributed Systems (cont.)

- The tradeoff in resource sharing is to complete instructions as fast and as cheaply as possible
  - When performing a Google search, it would take forever to compute the search on your device
  - Instead, the query is computed on a server with large amounts of CPU power, memory, and storage to ensure your query gets fulfilled as fast as possible
- If communication is really cheap: use all your resources in your distributed system
- If communication is really expensive: compute locally and send small amounts of data
- Distributed systems live somewhere between these two extremes

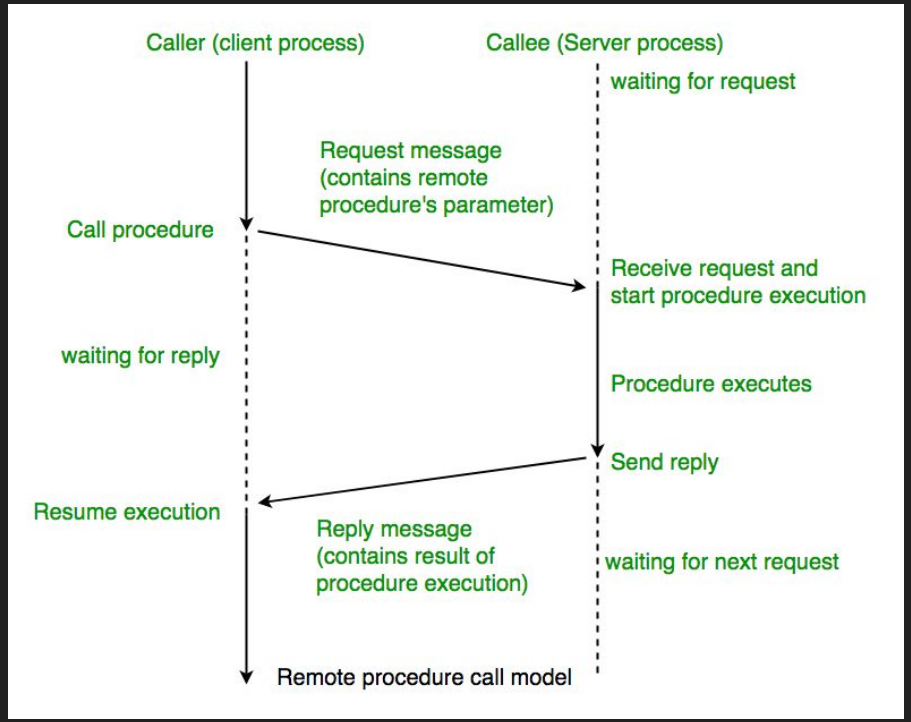
Client/Server Model

- Client/server is the most common distributed system computation paradigms
- A server is a process that provides a service
  - The server may exist on one or more nodes
  - The server sends responses to clients requests
- A client is a process that uses the service
  - The client binds to the server by locating it in the network and establishing a connection
  - The clients to send requests to the server and receive responses
- Remote procedure call (RPC) can be used to implement this structure
- This differs from the Peer-to-Peer model



Remote Procedure Call

- A remote procedure call is a way of requesting a server to execute a procedure or function on the client's behalf
- Servers advertise their procedures
- Clients call the server's procedures and receive the result
- RPCs are very common in many types of client/server systems
    - In online gaming, this can be used to prevent hackers

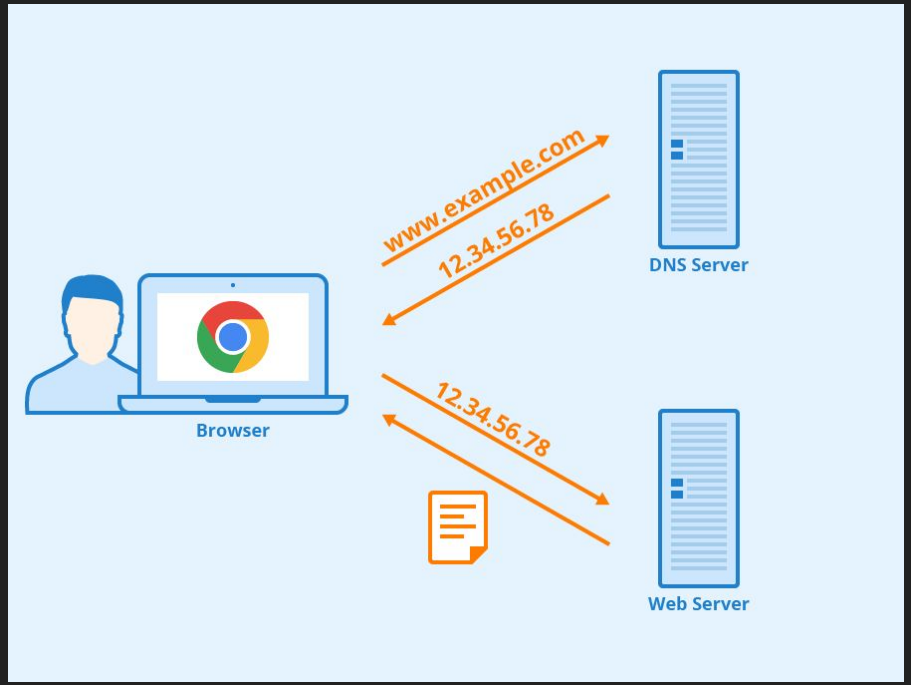


Remote Procedure Call (cont.)

- To implement an RPC function we need a procedure signature
  - The signature includes the number and types of arguments and return values
- “Stubs” are used to send and receive the data required for the RPC
  - The client stub bundles up the RPC arguments into a message that get sent to the server
  - The server stub unpacks the message and makes a function call with the provided arguments

Remote Procedure Call (cont.)

- How does the client know the location of the server?
  - Binding can be static - fixed at compile time
  - Binding can be dynamic - fixed at runtime
- In most RPC systems, dynamic binding is used using a name server (NS)
  - In DNS, a NS translates domain names into IP addresses
  - The servers that have RPCs attach themselves to NS and identify the RPCs they implement
  - The client asks the NS for a connection to a server that the NS knows and then establishes a connection
  - Try: nslookup google.com



RPC in Python

```
# Server
from xmlrpc.server import SimpleXMLRPCServer

def add_numbers(x, y):
    return x + y

server = SimpleXMLRPCServer(("localhost", 8000))
print("Listening on port 8000...")

server.register_function(add_numbers, "add")

server.serve_forever()
```

```
# Client
import xmlrpc.client

proxy =
xmlrpc.client.ServerProxy("http://localhost:8000/")

while True:
    x = int(input("Type first number to add: "))
    y = int(input("Type second number to add: "))
    print("Waiting for RPC result from server...")
    result = proxy.add(x, y)
    print(f"{x} + {y} = {result}")
```

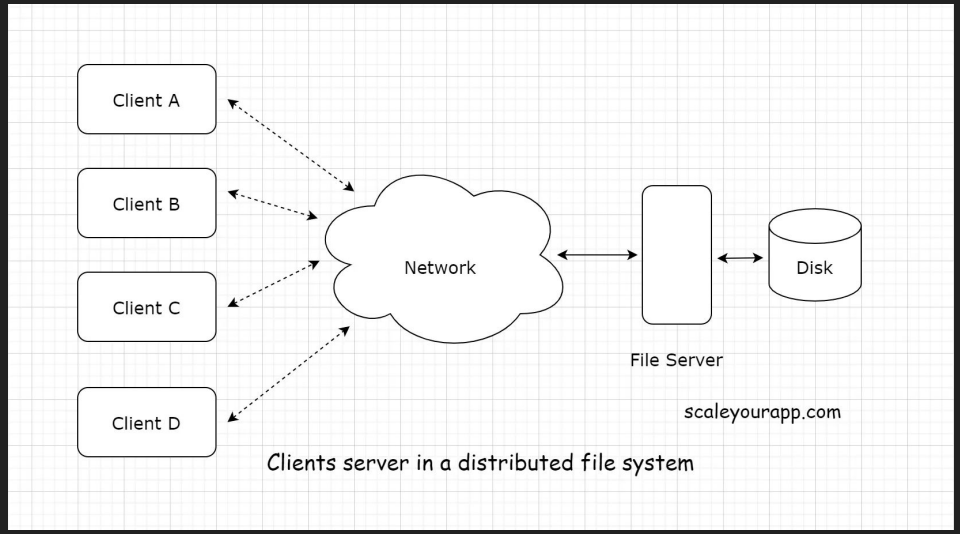
11 - Distributed File Systems

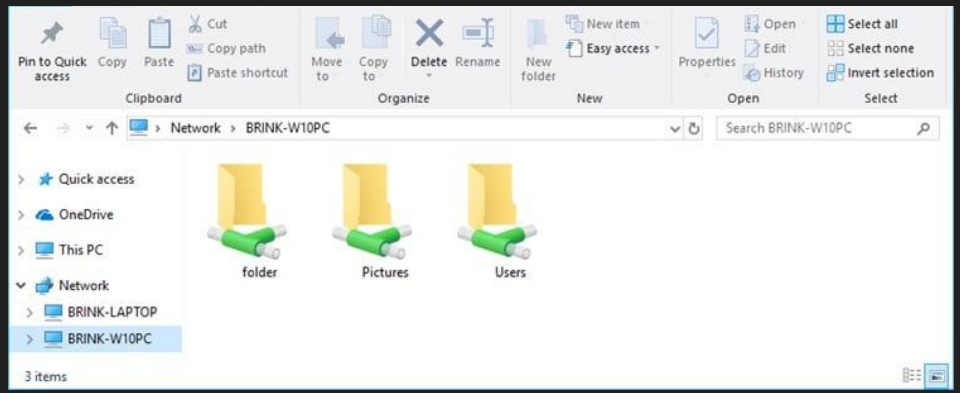
CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

Distributed File Systems

- Distributed file systems (DFS) is one of the most common uses of distributed systems and for RPCs
- A DFS is capable of sharing disks between nodes on the network as if the disks were attached to every node
- Example: video editing for large companies
  - If there are many video editors at a company that work with large amounts of video footage, having a central server to hold footage can be very beneficial for easily sharing between the video editors





Distributed File System Issues

- Naming and transparency
  - How should files be named so they can be easily found?
  - What does a file's name reveal about which node it is actually stored on? (Files are stored on many different machines)
- Remote access
  - How can we read or write to storage that is not directly attached to our machine?
- Caching
  - How can we store a cache of file data if there are many machines on the network who can change the file at any time?
- Stateless or stateful
  - Do we force the user to provide all the details about a file they want to access on our machine or do we keep track client requests overtime?

Naming and Transparency

- Issues with naming
  - How are files named?
  - Do file names reveal their location?
  - Do file names change if the file moves?
  - Do file names change if the user moves?
- Location transparency: The name of the file does not reveal the physical storage location
  - Example: Amazon S3
    - Users interact with S3 through a simple API or a web interface without needing to know the physical location of their stored objects (http://s3.amazonaws.com/bucket/)
    - S3 handles the data distribution and retrieval transparently, ensuring that users can access their data reliably regardless of its actual storage location
- Location independence: The name of the file is not linked to the physical storage location
  - Example: Google Drive
    - Users can access their files from any device with an internet connection without needing to know where the files are stored physically
    - Google Drive abstracts the storage infrastructure, providing users with seamless access to their files regardless of the underlying storage servers.

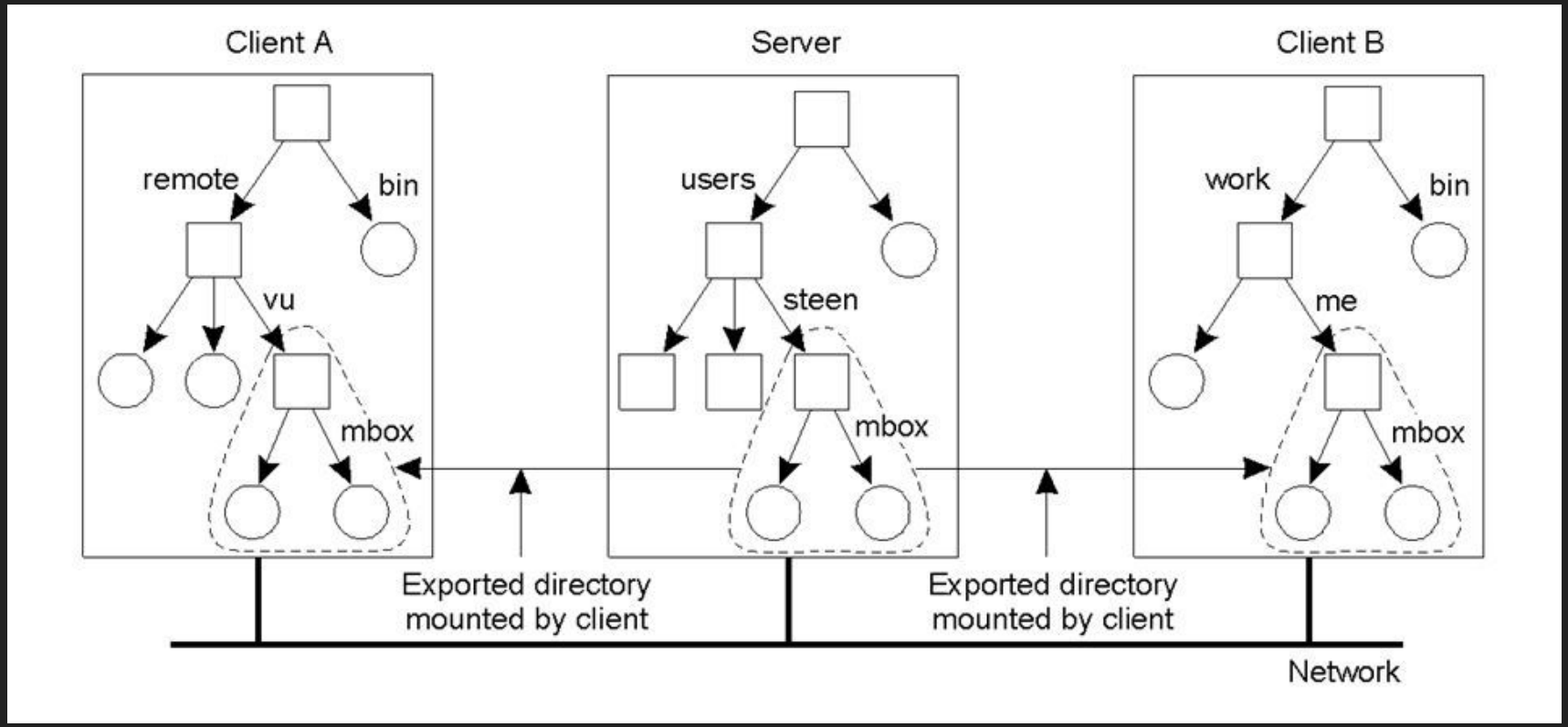
Absolute Naming Scheme

- The absolute naming scheme specifies a server name and a path to that file on that server
  - <server name: path name>
- Advantages:
  - Easy to find where files are at if you know the server
  - Scales easily (no naming conflicts between multiple machines)
- Disadvantages:
  - User has to know the server specifically
  - File is not location independent
  - Not fault tolerant (if the server goes down you cannot access the file!)

Mount Point Naming Scheme

- Special directories on your machine refers to directories in remote file system
- Mappings provided by a configuration file
  - `<local path name>` is bound to `<remote path name @ machine name>`
  - These directories are mapped at boot time and the user accesses them as if they were local
- Advantages:
  - Location transparency (remote path can change across reboots)
- Disadvantages:
  - Same file can have multiple names (remote path can be mounted to different local paths for multiple clients)

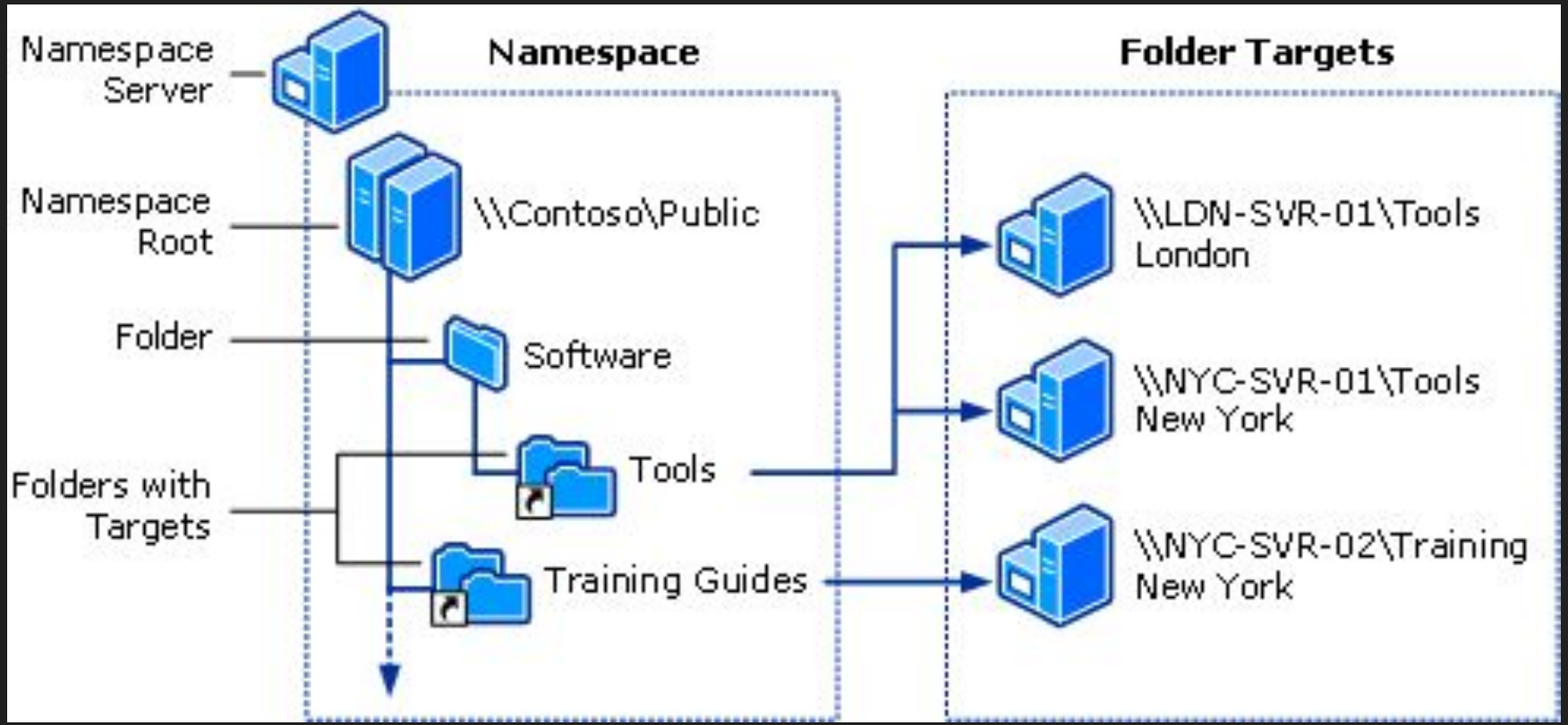
Mount Point Naming Scheme Example (NFS)



Global Namespace Naming Scheme

- A global namespace naming scheme presents a single, unified namespace to users, encompassing all files and directories stored across the DFS
- This namespace spans multiple servers and storage devices, providing users with a seamless view of the entire DFS
- When the client connects, it gets the file name structure form the name server
- Advantages:
  - Naming is consistent for a single file across multiple clients
  - Transparent (users can access files as if they were stored locally, regardless of their actual physical location)
- Disadvantages:
  - More performance overhead with large file systems (requires a name server to keep track of all the files)

Global Namespace Naming Scheme Example



Remote File Access and Caching

- If a user wants to access a remote file we have two options:
  - Remote Access Model
    - Client requests reads/writes parts of a file using RPC and server sends results using RPC
    - Server handles all the reads and writes (computationally expensive)
    - Files between multiple clients are always consistent
  - Caching Model
    - Client requests entire file which gets copied from the server and stored locally for reads/writes
    - When and where are file blocks cached?
    - When are modifications propagated back to the remote file?
    - What happens if multiple clients cache the same file?

Where to Store the Copy with Caching?

- If we are caching a file, where can the copy of the file be stored?
- Local Memory
  - Advantages:
    - Really fast access time
    - Works well if you have a diskless workstation
  - Disadvantages:
    - Memory is much smaller than disk
    - If the client goes down, all progress is lost

Where to Store the Copy with Caching?

- Local Disk
  - Advantages:
    - Safer if client fails
  - Disadvantages:
    - Disk is much slower than memory
    - Client requires a disk

How to Update the Cache?

- Write Through: every time the cache is written to on the local machine, update the remote copy also
  - This method is equivalent to the remote access model for writes, only exploits caching for reads
- Write Back: the cache is written back to the remote copy when the file is finally closed
  - Reduces network traffic if the same block of a file is written to multiple times
  - If the client crashes, changes can be lost

Cache Consistency

- Client-initiated consistency: Client contacts the server and ask if its copy is consistent with the server’s copy
  - Client can ask for every access, at some given interval, or only when opening the file after its been downloaded
  - This approach is much simpler for the server (it doesn’t care if clients have inconsistent caches)
- Server-initiated consistency: Server detects potential conflicts and invalidates caches
  - Server must know which clients have cached with parts of which files and which clients are readers or writers

Stateless and Stateful Servers

- Stateless: each request from a client is treated independently
  - Communication requires more information to be sent
  - More resilient to failures because they don't rely on maintaining state across multiple servers
  - Less complex than stateful and has less synchronization issues with replicated server instances
- Stateful: the server maintains information about the state of each client's connection or session
  - Communication requires less information to be sent
  - Performance benefits by caching frequently accessed data or optimizing requests based on previous interactions
  - More complexity especially with replicated server instances
    - Storing backups

Sun’s Network File System (NFS)

- NFS is the standard for distributed UNIX file access
  - Originally designed at a time when clients were diskless (their OS would be booted over the network)
  - Stateless protocl
  - Uses mount point protocol for file names
- NFSv3
  - Still commonly used today
  - Better performance and support for 64-bit file sizes and offsets compared to NFSv2
  - Lacks strong authentication and encryption
- NFSv4
  - Improved security, better support for large files, and better performance for some workloads
  - Has authentication and encryption
  - Supports file locking (protecting a file from being accessed by multiple users simultaneously)

Sun’s Network File System (NFS) (cont.)

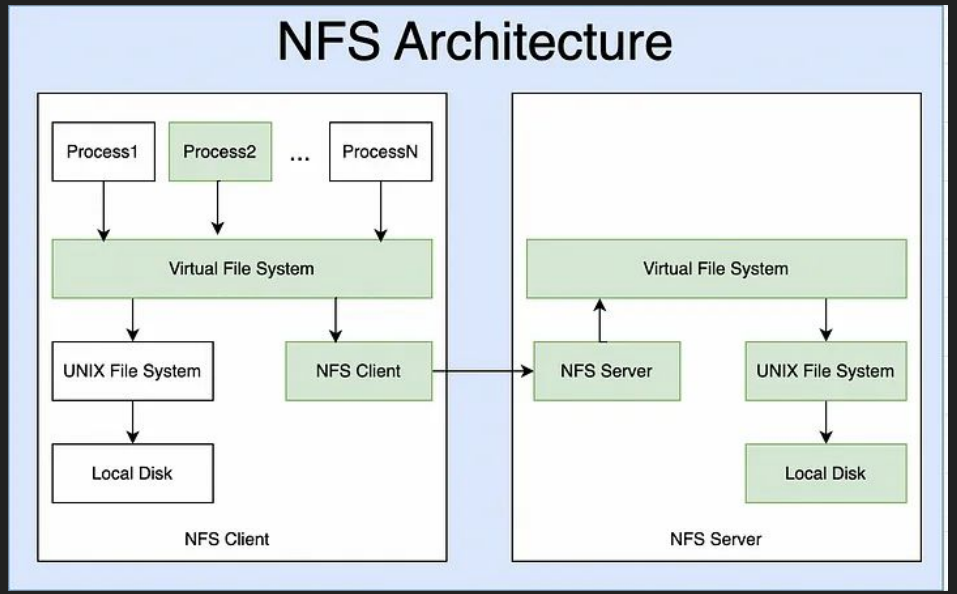
- Mount point protocol uses two files:
  - /etc/fstab (used by clients):
    - NFS clients use this file to define how remote file systems should be mounted on the local system at boot time
  - /etc/exports (used by servers):
    - NFS servers use this file to define which directories or filesystems on the server should be shared with remote clients

Sun’s Network File System (NFS) (cont.)

- NFS defines a set of RPC operations:
  - Directory search and reading directory entries
  - Manipulating links and directories
  - Accessing file attributes (file size, creation date, etc.)
  - Reading/writing files
- The underlying local file systems for the clients and servers does not need to be the same
  - Both client and server must only support NFS and RPC
  - Example: a Mac system with APFS (Apple File System) and a Linux system with Ext4 (Fourth Extended File System) can both communicate with the NFS standard
- There are no open/close file RPC operations
  - NFS is stateless

Sun’s Network File System (NFS) (cont.)

- In UNIX-like OSs, we have a virtual file system
  - Supports both local file system and remote file system
  - The VFS provides a standard interface for file access (regardless if remote or local)
- NFS module sends and receives RPCs
- If a read/write request is received on the server, this request is then passed to the local file system module (UFS, ext4, APFS, etc.)



Sun’s Network File System (NFS) (cont.)

- In NFS, all the consistency work is offloaded to the client
  - The client *could* read every block of a file, and store that in some cache
  - Then, if it makes any updates (writes) send those to the server every 30 seconds (cache flush) or when the file is closed
  - The server uses a write through policy, committing all changes immediately when client performs a write RPC
- In NSFv4, clients can lock a file (stateful)
  - Allows clients to work on cached version of file without having to cache flush or to check if their file has changed remotely

12 - Protection

CEG 4350/5350 Operating Systems Internals and Design

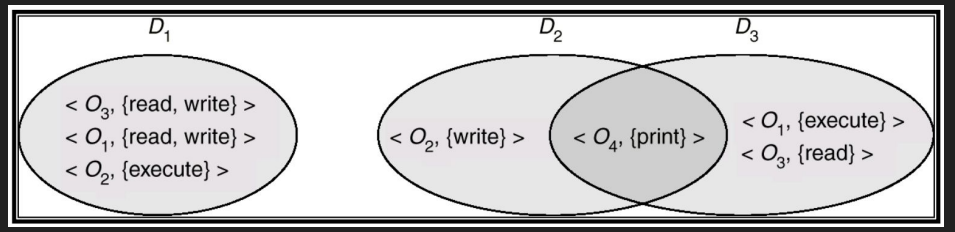
Max Gilson

Protection

- An OS is just a collection of modules or objects either hardware or software
  - CPUs, memory locations, files, devices, drivers, etc.
- Each object has a unique name and is accessed through well defined set of operations
  - Example: the storage device must be accessed by a driver that the OS should be protective over
- An OS must ensure any object is accessed correctly and only by the processes that are allowed to access them
- Principle of least privilege: we want to give every process or the least amount of privileges required, while still allowing everything to run as intended
  - If we gave every user and every process unlimited privilege, we could have severe security issues

Domain Structure

- Access-right = <object-name, rights-set>
  - Where rights-set is a subset of all valid operations that can be performed on the object
  - Example: A user may have the access-right to read or execute a file object, but not write the object
- Domain = set of access-rights
  - Associated with users, user groups, and processes
  - Example: An admin usually has the most access-rights, so its domain could include all valid operations
    - A guest account's domain should be smaller, or have fewer access-rights

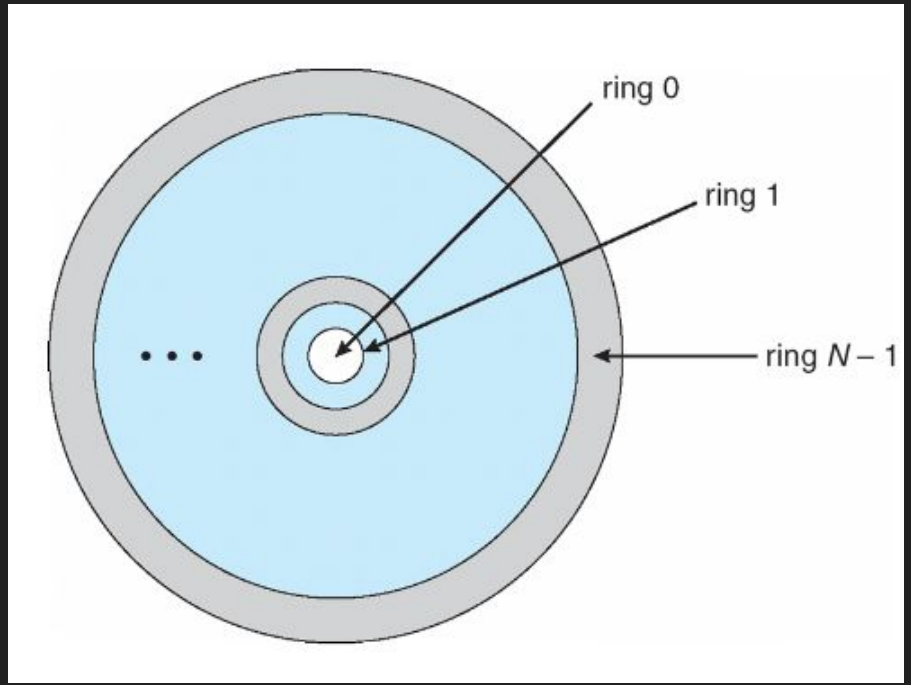


Domain Implementation in UNIX

- There are two types of domains:
  - User - only a select few privileges
  - Supervisor (aka root) - all possible privileges
- UNIX
  - Domains are governed by a user ID
  - Processes inherit the domain of a user when executed
  - Files may have a setuid bit in their metadata, which allows the process to execute within the domain of the user that *created* the file instead
    - Can be useful for processes that require more privileges that a regular user might not have (directly interacting with hardware)

Domain Implementation in Multics

- Domain rings are a hierarchical form of domains
- The innermost ring has the most amount of privileges
- The outermost ring has the least amount of privileges
- If D_i and D_j are two domains in the domain ring, and $j < i$, then D_i is a subset of D_j
  - Simply put, the innermost ring has all of the privileges surrounding it



Access Matrix

- Let’s view protection as a matrix called an access matrix
  - Rows represent domains
  - Columns represent objects
- Access(i, j) is the set of operations that a process executing in $Domain_i$ can invoke on $Object_j$
- Problems can arise if the access matrix is huge
  - Lots of files and/or lots of domains and most of the matrix entries will be empty

domain \ object	F_1	F_2	F_3	printer
D_1	read		read	
D_2				print
D_3		read	execute	
D_4	read write		read write	

Using an Access Matrix

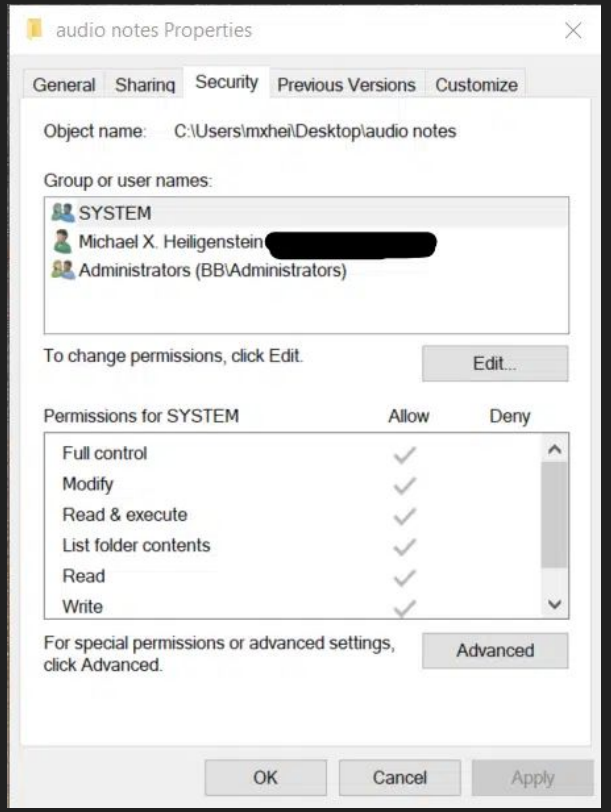
- If a process in Domain D_i tries to perform an operation on object O_j then check if the operation is in the access matrix at location (i, j)
- Can be expanded to dynamic protection
  - Have the creator of an object be the “owner”
  - Give the owner the ability to modify the access-rights for any domain for objects they own

Separation of Mechanism and Policy

- An access matrix separates mechanism from policy
  - Mechanism: How we allow
    - OS provides access matrix + rules
    - OS ensures access matrix is only manipulated by authorized users and ensures rules are enforced
  - Policy: What we want to allow
    - User dictates the policies of who can access what

Implementing and Access Matrix

- The access matrix is typically implemented either with only rows or only columns
- Access-control list:
  - Each object has a column that defines what domain can perform what operation for that object
    - Used in Windows and Linux
    - Easy to change permissions for a specific file
- Capability List:
  - Each domain has a row that defines what operations are allowed for that domain
    - Hard to change permissions for a specific file (need to search all user's capabilities)



Language-Based Protection

- Some programming languages provide protection independent of the OS
- This is particularly apparent with Java
  - Java runs inside of the Java Virtual Machine (JVM)
  - The JVM has its own protection domain that defines what your code is and isn’t allowed to do
  - Example: Write a program that loads code from a network and executes it with a less permissions in the same JVM
  - Example: JVM does not allow pointer arithmetic or low level stack/memory operations

13 - Security

CEG 4350/5350 Operating Systems Internals and Design

Max Gilson

Security

- A system is secure if resources are always used and accessed as intended under all circumstances (unachievable)
- *Intruders* (hackers) attempt to breach security
- A *threat* or *exploit* is potential security violation
- An *attack* is attempt to breach security
  - Attacks can be accidental or malicious
- *Malware* is any software intentionally designed to cause harm to a computer system, network, or user

Types of Security Violations

- Breach of confidentiality
  - Unauthorized reading of data
  - Identity theft (2017 Equifax data breach)
- Breach of integrity
  - Unauthorized modification of data
  - Nuclear centrifuge control system modification (Stuxnet Worm)
- Breach of availability
  - Unauthorized removal of access
  - Ransomware (“Pay me or I will lock your computer forever”)
- Theft of service
  - Unauthorized use of resources
  - Bypassing your ISP or cable provider
- Denial of service (DOS)
  - Prevention of legitimate use
  - Overwhelm a system with traffic

Methods of Security Violations

- Masquerading (breach authentication)
  - Pretending to be an authorized user
  - A hacker logging into your email after stealing your password
- Replay attack
  - A hacker capturing your data and “replaying” it to appear as you
- Man-in-the-middle attack
  - A hacker intercepts communication between two systems
  - RFID skimming
- Session hijacking
  - A hacker uses an already-established session to bypass authentication
  - LTT YouTube Channel Hack (2023)

Levels of Security Measures

- Perfect security is impossible but we must make it difficult to hackers to get in!
- The four levels of security:
  - Physical
    - Physical access to data centers, servers, connected terminals
  - Human
    - Social engineering, email phishing, dumpster diving, scam calls
  - Operating System
    - Protection mechanisms, debugging
  - Network
    - Intercepted communications, interruption, DOS

Program Threats

- Trojan Horses
  - Code segment that misuses its environment
  - Spyware, pop-up browser windows, covert channels
  - Zeus Trojan read all of your keystrokes, captured screenshots, and stole banking credentials
- Backdoor
  - Specific user identifier or password that circumvents normal security procedures
  - Sony BMG Rootkit installed music copy protection software on your computer that created a backdoor on your computer

Program Threats (cont.)

- Logic Bomb
  - Program that initiates a security incident under certain circumstances
  - Your coworker hides secret code in your company's program to force it to crash and erase all data if he's ever fired
- Stack and Buffer Overflow
  - Overflows the stack or a buffer (like a string) to modify the program that is currently running
  - Tony Hawk's Pro Skater Strcpy hack for Xbox, PS2, Gamecube, and Xbox 360 consoles
    - This even works over a network so you can hack other people!

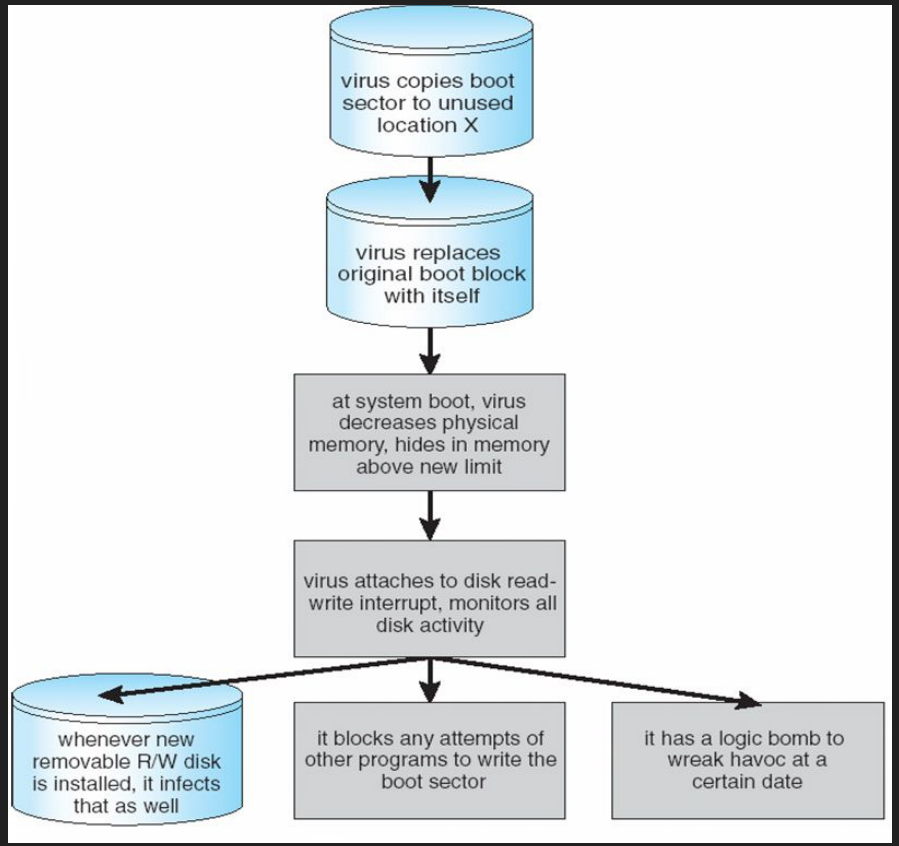
Program Threats (cont.)

- Viruses
  - Code fragment embedded in legitimate program
  - Usually very specific to CPU architecture, operating system, applications
  - Melissa Virus (1999) hid inside of a Word Doc and used a Visual Basic Macro to send itself to the first 50 people in your Outlook address book
- Virus dropper
  - A sneaky program that drops malware onto systems but usually does not attack the system itself
  - When opening a fake PDF, it installs other malware

Buffer Overflow Condition

```
#include <stdio.h>
#define BUFFER SIZE 256
int main(int argc, char *argv[])
{
    char buffer[BUFFER SIZE];
    if (argc < 2)
        return -1;
    else {
        strcpy(buffer,argv[1]);
        return 0;
    }
}
```

Boot Sector Virus

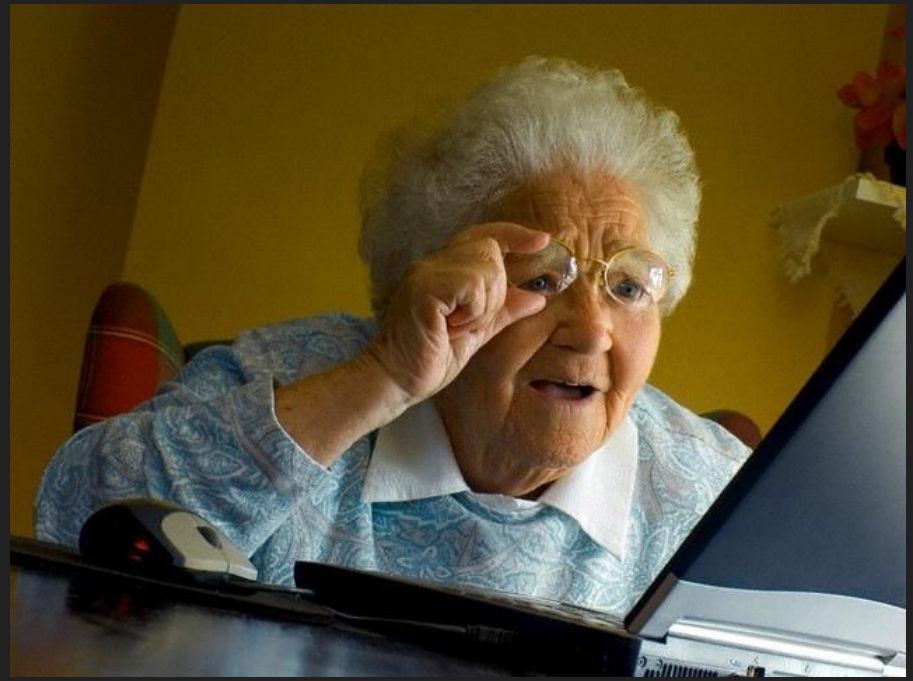


Programming Skills Required

- To determine the bug in someone else’s security and to write the exploit code requires high level of programming skills
- After this is done, if a script is written to exploit the bug it can be easy for anyone to hack!

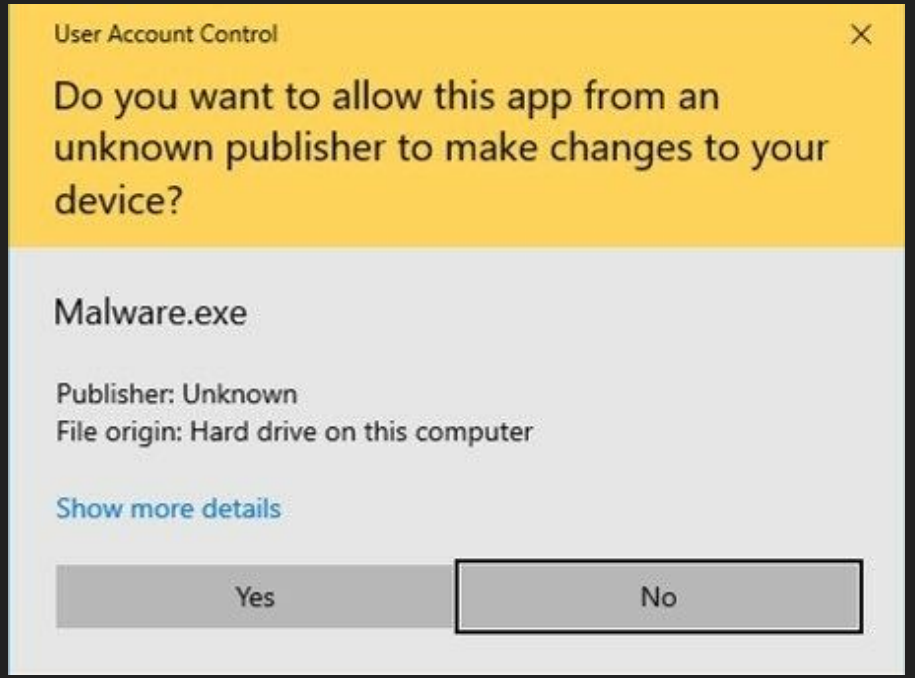
Attacks are Common

- Attacks on systems are very common even today
- Attacks used to be fun science experiments, now they are tools for organized crime, military, and espionage
- Windows OS the most common target
  - Most widely used by non-technical users (grandma’s favorite OS)
  - Users are administrators by default
  - Many companies run outdated versions of Windows



Systems and Network Threats

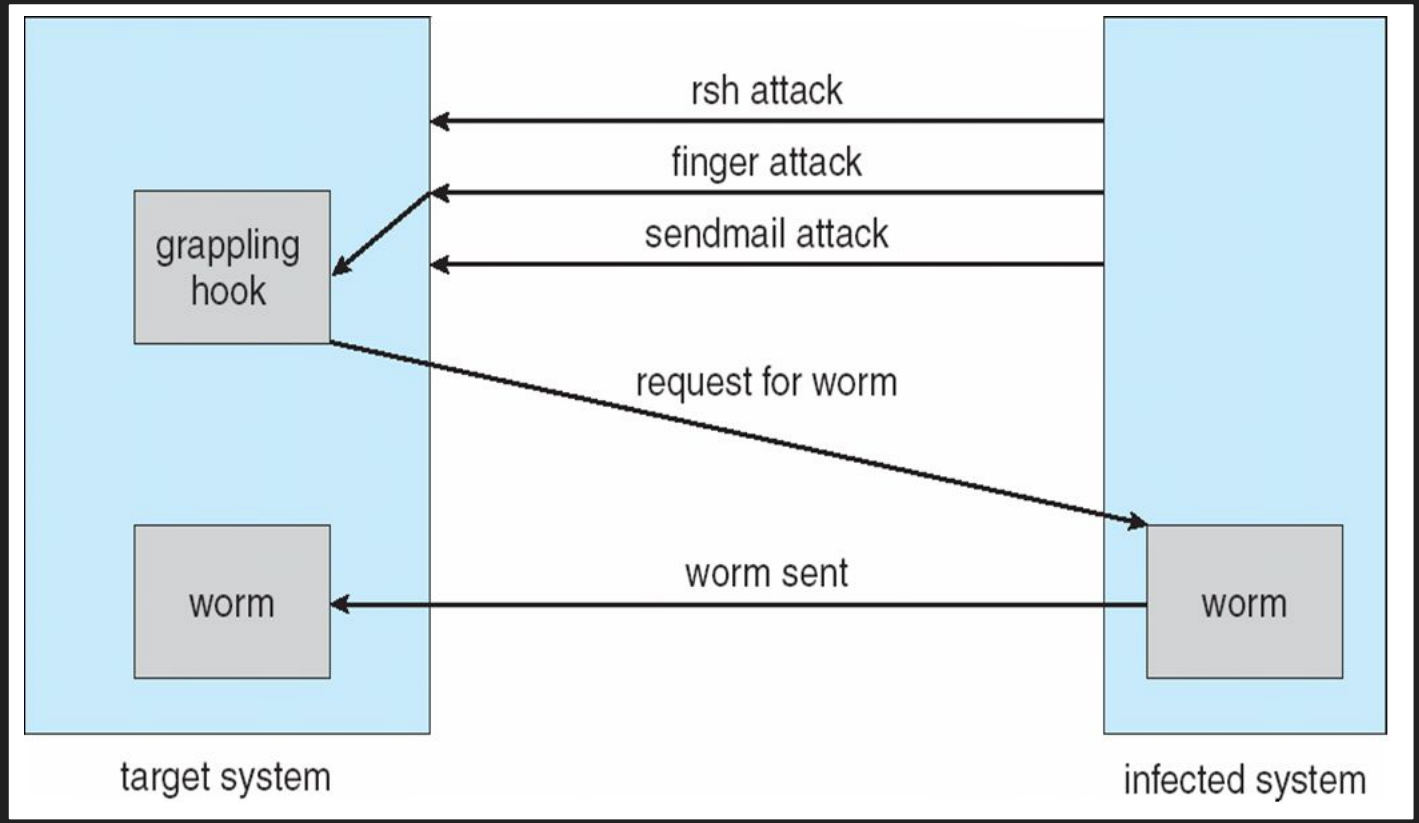
- Should a system be secure by default?
  - Could make it harder to use
- Should a system be “open” by default?
  - Could make it easier for hackers to get in



Systems and Network Threats (cont.)

- Network threats harder to detect, prevent
  - More difficult to have a shared secret on which to base access (what if someone eavesdropped our secret code?)
  - If attached to the internet, there is no limit
  - Determining location of an attacking system difficult
- Morris Internet Worm
  - Grappling hook (virus loader) uploaded the worm program
  - Exploited remote shell access to replicate itself to other machines on a network
  - “Accidentally” overloaded the computer it infected by reinfecting multiple times, making it unusable
  - Broke the entire internet

Systems and Network Threats (cont.)



Systems and Network Threats (cont.)

- Port scanning
  - First, automate connections to a range of ports on one or a range of IP addresses
  - Next, detect service protocol and OS on the responding port
  - Finally, use this port for malicious purposes
  - Frequently launched from zombie systems
    - A zombie system is a compromised computer controlled by an attacker
    - Used to decrease trace-ability

Systems and Network Threats (cont.)

- Denial of Service
  - Overload the targeted computer preventing it from doing any useful work
  - Distributed denial-of-service (DDOS) come from multiple sites at once
  - Can happen with a SYN flood
    - Attacker sends a SYN (start of connection request)
    - Attacker receives the server’s acknowledgement
    - Attacker never sends an acknowledgement back to the server
    - Server is stuck waiting for acknowledgement forever
  - Consider traffic to a web site
    - How can you tell the difference between being a target and being really popular?

Cryptography

- Cryptography is the science of protecting information by transforming it into a form that only intended recipients can understand
  - Based on secret keys
  - Internal to a given computer, source and destination of messages can be known and protected
  - Source and destination of messages on network cannot be trusted without cryptography
    - People can’t read your messages unless they’re supposed to
    - Messages can’t be changed without it being obvious
    - The sender of a message is who they say they are