System Calls, Kernel Mode, and Process Implementation

CS 256/456

Dept. of Computer Science, University of Rochester

Last Class …

- Processes
  - Process concept
  - Operations on processes
- Introduction to Signals
  - User-level events

Today

- Processes
  - A process's image in a computer
- System protection and kernel mode
- System calls and the interrupt interface
- More on signals
  - User-level events
- I/O and process groups
- Pipes
  - Inter-process communication

Processes

- Def: A *process* is an instance of a running program.
  - Not the same as "program" or "processor"
- Process provides each program with two key abstractions:
  - Logical control flow
    - Each program seems to have exclusive use of the CPU.
  - Private address space
    - Each program seems to have exclusive use of main memory.
- How are these Illusions maintained?
  - Process executions interleaved (multitasking)
  - Address spaces managed by virtual memory system

Process and Its Image

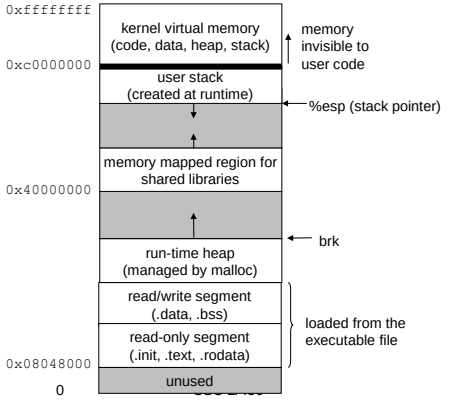
- An operating system executes a variety of programs:
  - A program that browses the Web,
  - A program that serves Web requests, …
- Process – a program in execution
- A process's state/image in a computer includes:
  - User-mode address space
  - Kernel data structures
  - Registers (including program counter and stack pointer)
- Address space and memory protection
  - Physical memory is divided into user memory and kernel memory
  - Kernel memory can only be accessed when in the kernel mode
  - Each process has its own exclusive address space in the user-mode memory space (sort-of)

Process Creation

- Actions/decisions when a process (parent) creates a new process (child)
  - Execution sequence
  - Address space sharing
  - Open files inheritance
  - … …
- UNIX examples
  - **fork** system call creates new process with a duplicated copy of everything.
  - **exec** system call used after a **fork** to replace the process' memory space with a new program.
  - child and parent compete for CPU like two normal processes.

Private Address Spaces

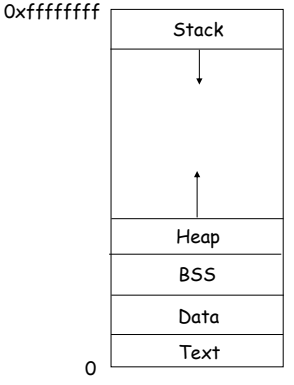
- Each process has its own private address space.



User-mode Address Space

User-mode address space for a process:

- Text: program code, instructions
- Data: initialized global and static variables (those data whose size is known before the execution)
- BSS (block started by symbol): uninitialized global and static variables
- Heap: dynamic memory (those being malloc-ed)
- Stack: local variables and other stuff for function invocations



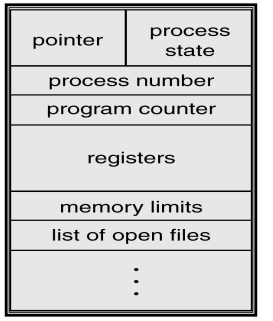
Process Management

- *A process* is a program in execution
  - Unit of work – A process needs certain resources, including CPU time, memory, files, and I/O devices, to accomplish its task
  - Protection domain
- OS responsibilities for process management:
  - Process creation and deletion
  - Resource allocation
  - Process scheduling, suspension, and resumption
  - Process synchronization, inter-process communication

Process Control Block (PCB)

OS data structure (in kernel memory) maintaining information associated with each process.

- Process state
- Program counter
- CPU registers
- CPU scheduling information
- Memory-management information
- Accounting information
- Information about open files
- Other



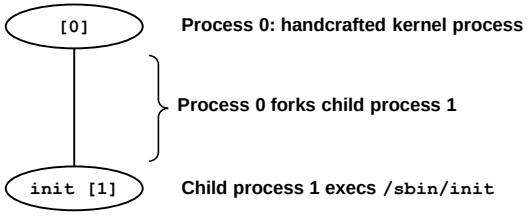
SYSTEM PROTECTION

System Protection

- User programs (programs not belonging to the OS) are generally not trusted
  - A user program may use an unfair amount of resource
  - A user program may maliciously cause other programs or the OS to fail
- Need protection against untrusted user programs; the system must differentiate between at least two modes of operations
  1. *User mode* – execution of user programs
     - untrusted
     - not allowed to have complete/direct access to hardware resources
  2. *Kernel mode* (also *system mode* or *monitor mode*) – execution of the operating system
     - trusted
     - allowed to have complete/direct access to hardware resources
- Hardware support is needed for such protection

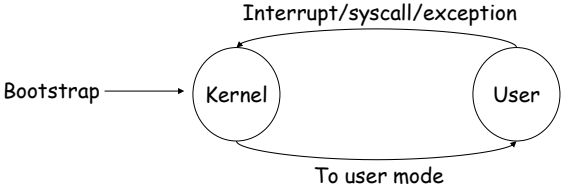
Unix Startup: Step 1

1. Pushing reset button loads the PC with the address of a small bootstrap program.
2. Bootstrap program loads the boot block (disk block 0).
3. Boot block program loads kernel binary (e.g., `/boot/vmlinux`)
4. Boot block program passes control to kernel.
5. Kernel handcrafts the data structures for process 0.



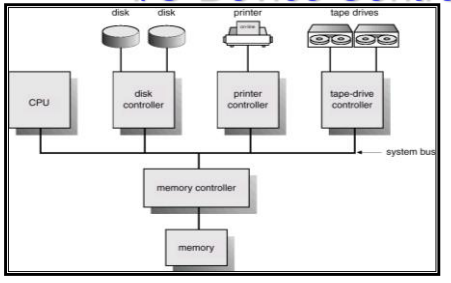
Transition between User/Kernel Mode

- When does the machine run in kernel mode?
  - after machine boot
  - interrupt handler
  - system call
  - exception



I/O Device Protection

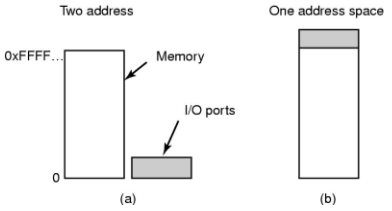
I/O Device Controllers



Device Controller State
Control registers
Status registers
Data buffers

- I/O devices have both mechanical component & electronic component
- The electronic component is the device controller
  - It contains control logic, command registers, status registers, and on-board buffer space

I/O Ports & Memory-Mapped I/O



I/O methods:
- Separate I/O and memory space; special I/O commands (IN/OUT)
- Memory-mapped I/O

Issues with them:
- Convenience/efficiency when using high-level language;
- Protection mechanisms;
- Data caching

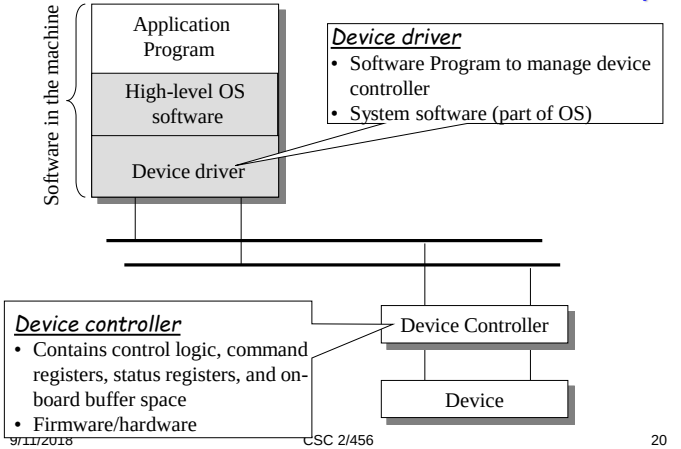
I/O Operations

- How is I/O done?
  - I/O devices are much slower than CPU
- Synchronous (polling)
  - After I/O starts, busy-wait while polling (or poll periodically) the device status register until it shows the operation completes
- Asynchronous (interrupt-driven)
  - After I/O starts, control returns to the user program without waiting for I/O completion
  - Device controller later informs CPU that it has finished its operation by causing an *interrupt*
  - When an interrupt occurs, current execution is put on hold; the CPU jumps to a service routine called an "interrupt handler"

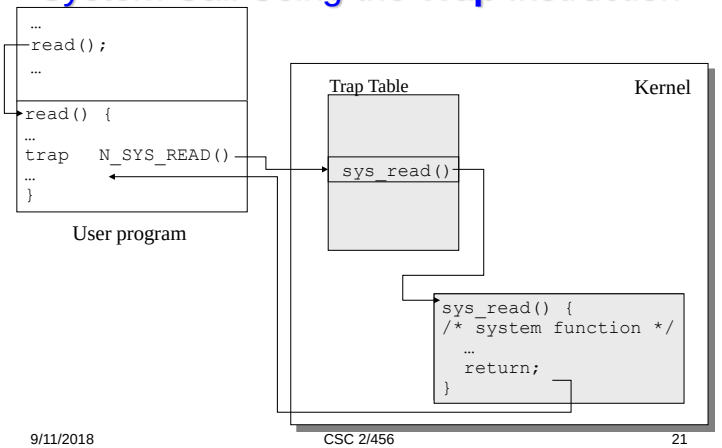
Protection of I/O Devices

- User programs are not allowed to directly access I/O devices
  - Special I/O instructions can only be used in kernel mode
  - Controller registers can only be accessed in kernel mode
- So device drivers, I/O interrupt handlers must run in kernel mode
- User programs perform I/O through requesting the OS (using system calls)

The Device-Controller-Software Relationship



System Call Using the **Trap** Instruction



Interrupt Handlers

1. Save registers of the old process
2. Set up context for interrupt service procedure (switch from the user space to kernel space: MMU, stack, …)
3. Run service procedure; when safe, re-enable interrupts
4. Run scheduler to choose the new process to run next
5. Set up context (MMU, registers) for process to run next
6. Start running the new process

How much cost is it? Is it a big deal?

For Gigabit Ethernet, each packet arrives once every 12us.

Interrupt Vectors

- Intel Pentium processor event-vector table
  - 0: divide by zero
  - 6: invalid opcode
  - 11: segment not present
  - 12: stack fault
  - 14: page fault
  - …31: non-maskable
  - 32-255: maskable interrupts

CPU Protection

CPU Protection

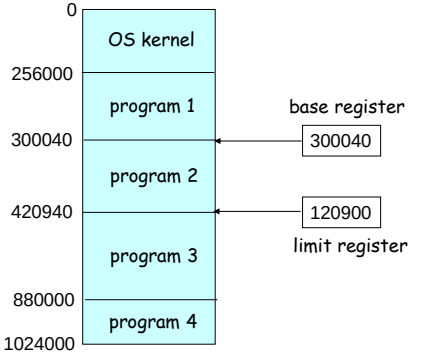
- Goal of CPU protection
  - A user program can't hold the CPU for ever
- *Timer* – interrupts computer after specified period to ensure the OS kernel maintains control
  - Timer is decremented every clock tick
  - When timer reaches the value 0, an interrupt occurs
  - CPU time sharing is implemented in the timer interrupt

Memory Protection

Memory Protection

- Goal of memory protection?
  - A user program can't use arbitrary amount of memory
  - A user program can't access data belonging to the operating system or other user programs
- How to achieve memory protection?
  - Add two registers that determine the range of legal addresses a program may access:
    - Base register – holds the smallest legal physical memory address
    - Limit register – contains the size of the range
    - Memory outside the defined range is protected

Hardware Address Protection



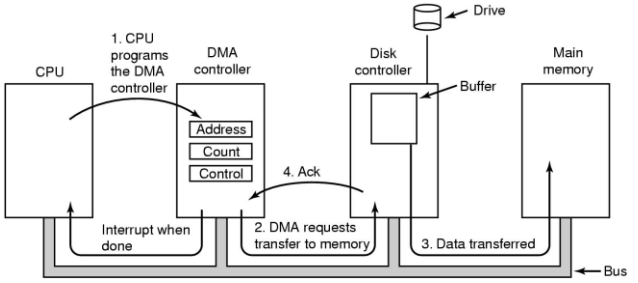
- Address of each memory address is checked against "base" and "base+limit"
- Trap to the OS kernel if it falls outside of the range (an exception)

In Practice Today: Virtual Memory

- Indirect memory access
  - Memory access with a virtual address which needs to be translated into physical address

29

Direct Memory Access (DMA)



- Are the addresses CPU sends to the DMA controller virtual or physical addresses?
- Can the disk controller directly read data into the main memory (bypassing the controller buffer)?

9/11/2018 CSC 2/456 30

Signals

- A *signal* is a small message that notifies a process that an event of some type has occurred in the system.
  - Kernel abstraction for exceptions and interrupts.
  - Sent from the kernel (sometimes at the request of another process) to a process.
  - Different signals are identified by small integer ID's
  - The only information in a signal is its ID and the fact that it arrived.

9/11/2018 CSC 2/456 31

Default Actions

- Each signal type has a predefined *default action*, which is one of:
  - The process terminates
  - The process terminates and dumps core.
  - The process stops until restarted by a SIGCONT signal.
  - The process ignores the signal.

9/11/2018 CSC 2/456 32

Some Common Signals and Their Defaults

ID	Name	Default Action	Corresponding Event
2	SIGINT	Terminate	Interrupt from keyboard (ctl-c)
9	SIGKILL	Terminate	Kill program (cannot override or ignore)
11	SIGSEGV	Terminate & Dump	Segmentation violation
14	SIGALRM	Terminate	Timer signal
17	SIGCHLD	Ignore	Child stopped or terminated

Installing Signal Handlers

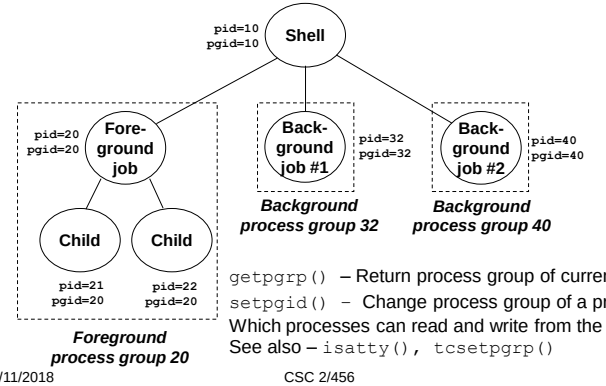
- The `signal` function modifies the default action associated with the receipt of signal `signum`:
  - `handler_t *signal(int signum, handler_t *handler)`
- Different values for `handler`:
  - SIG_IGN: ignore signals of type `signum`
  - SIG_DFL: revert to the default action on receipt of signals of type `signum`.
  - Otherwise, handler is the address of a *signal handler*
    - Called when process receives signal of type `signum`
    - Referred to as "*installing*" the handler.
    - Executing handler is called "*catching*" or "*handling*" the signal.
    - When the handler executes its return statement, control passes back to instruction in the control flow of the process that was interrupted by receipt of the signal.

Standard In, Out, and Error

- By convention, file descriptors 0, 1, and 2 are used for:
  - Standard Input
  - Standard Output
  - Standard Error

Process Groups

- Every process belongs to exactly one process group



`getpgrp()` – Return process group of current process
`setpgid()` – Change process group of a process
Which processes can read and write from the terminal?
See also – `isatty()`, `tcsetpgrp()`

Process Groups

- Every process belongs to exactly one process group
- Can send signal to an entire group
- getpgrp() – Return process group of current process
- setpgid() – Change process group of a process

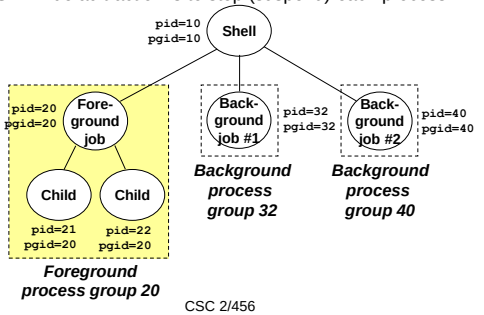
Sending Signals with `kill` Program

- `kill` program sends arbitrary signal to a process or process group
- Examples
  - `kill -9 24818`
    - Send SIGKILL to process 24818
  - `kill -9 -24817`
    - Send SIGKILL to every process in process group 24817.

```
linux> ./forks 16
linux> Child1: pid=24818 pgrp=24817
Child2: pid=24819 pgrp=24817

linux> ps
  PID TTY          TIME CMD
24788 pts/2    00:00:00 tcsh
24818 pts/2    00:00:02 forks
24819 pts/2    00:00:02 forks
24820 pts/2    00:00:00 ps
linux> kill -9 -24817
linux> ps
  PID TTY          TIME CMD
24788 pts/2    00:00:00 tcsh
24823 pts/2    00:00:00 ps
linux>
```


Sending Signals from the Keyboard

- Typing ctrl-c (ctrl-z) sends a SIGINT (SIGTSTP) to every job in the foreground process group.
  - SIGINT – default action is to terminate each process
  - SIGTSTP – default action is to stop (suspend) each process



Example of `ctrl-c` and `ctrl-z`

```
linux> ./forks 17
Child: pid=24868 pgrp=24867
Parent: pid=24867 pgrp=24867
 <typed ctrl-z>
Suspended
linux> ps a
  PID TTY      STAT   TIME COMMAND
24788 pts/2    S      0:00 -usr/local/bin/tcsh -i
24867 pts/2    T      0:01 ./forks 17
24868 pts/2    T      0:01 ./forks 17
24869 pts/2    R      0:00 ps a
bass> fg
./forks 17
<typed ctrl-c>
linux> ps a
  PID TTY      STAT   TIME COMMAND
24788 pts/2    S      0:00 -usr/local/bin/tcsh -i
24870 pts/2    R      0:00 ps a
```

Sending Signals with `kill` Function

```
void fork12()
{
    pid_t pid[N];
    int i, child_status;
    for (i = 0; i < N; i++)
        if ((pid[i] = fork()) == 0)
            while(1); /* Child infinite loop */

    /* Parent terminates the child processes */
    for (i = 0; i < N; i++) {
        printf("Killing process %d\n", pid[i]);
        kill(pid[i], SIGINT);
    }

    /* Parent reaps terminated children */
    for (i = 0; i < N; i++) {
        pid_t wpid = wait(&child_status);
        if (WIFEXITED(child_status))
            printf("Child %d terminated with exit status %d\n",
                   wpid, WEXITSTATUS(child_status));
        else
            printf("Child %d terminated abnormally\n", wpid);
    }
}
```

Receiving Signals

- Suppose kernel is returning from exception handler and is ready to pass control to process *p*.
- Kernel computes `pnb = pending & ~blocked`
  - The set of pending nonblocked signals for process *p*
- If (`pnb == 0`)
  - Pass control to next instruction in the logical flow for *p*.
- Else
  - Choose least nonzero bit *k* in `pnb` and force process *p* to receive signal *k*.
  - The receipt of the signal triggers some *action* by *p*
  - Repeat for all nonzero *k* in `pnb`.
  - Pass control to next instruction in logical flow for *p*.

Signal Handling Example

```
void int_handler(int sig)
{
    printf("Process %d received signal %d\n",
           getpid(), sig);
    exit(0);
}

void fork13()
{
    pid_t pid[N];
    int i, child_status;
    signal(SIGINT, int_handler);

    . . .
}
```

```
linux> ./forks 13
Killing process 24973
Killing process 24974
Killing process 24975
Killing process 24976
Killing process 24977
Process 24977 received signal 2
Child 24977 terminated with exit status 0
Process 24976 received signal 2
Child 24976 terminated with exit status 0
Process 24975 received signal 2
Child 24975 terminated with exit status 0
Process 24974 received signal 2
Child 24974 terminated with exit status 0
Process 24973 received signal 2
Child 24973 terminated with exit status 0
linux>
```

Non-queuing Nature of Signals

```
int ccount = 0;
void child_handler(int sig)
{
    int child_status;
    pid_t pid = wait(&child_status);
    ccount--;
    printf("Received signal %d from process %d\n",
           sig, pid);
}

void fork14()
{
    pid_t pid[N];
    int i, child_status;
    ccount = N;
    signal(SIGCHLD, child_handler);
    for (i = 0; i < N; i++)
        if ((pid[i] = fork()) == 0) {
            /* Child: Exit */
            exit(0);
        }
    while (ccount > 0)
        pause();/* Suspend until signal occurs */
}
```

- Pending signals are not queued
  - For each signal type, just have single bit indicating whether or not signal is pending
  - Even if multiple processes have sent this signal

Living With Nonqueuing Signals

- Must check for all terminated jobs
  - Typically loop with `wait(blocking)` or `waitpid` with appropriate parameter for non-blocking call

```
void child_handler2(int sig)
{
    int child_status;
    pid_t pid;
    while ((pid = wait(&child_status)) > 0) {
       ccount--;
       printf("Received signal %d from process %d\n",
sig, pid);
    }
}

void fork15()
{
    . . .
    signal(SIGCHLD, child_handler2);
    . . .
}
```

A Program That Reacts to Externally Generated Events (ctrl-c)

```
#include <stdlib.h>
#include <stdio.h>
#include <signal.h>

void handler(int sig) {
  printf("You think hitting ctrl-c will stop the bomb?\n");
  sleep(2);
  printf("Well...");
  fflush(stdout);
  sleep(1);
  printf("OK\n");
  exit(0);
}

main() {
  signal(SIGINT, handler); /* installs ctl-c handler */
  while(1) {
  }
}
```

A Program That Reacts to Internally Generated Events

```
#include <stdio.h>
#include <signal.h>

int beeps = 0;

/* SIGALRM handler */
void handler(int sig) {
  printf("BEEP\n");
  fflush(stdout);

  if (++beeps < 5)
    alarm(1);
  else {
    printf("BOOM!\n");
    exit(0);
  }
}
```

```
main() {
  signal(SIGALRM, handler);
  alarm(1); /* send SIGALRM in
               1 second */

  while (1) {
    /* handler returns here */
  }
}
```

```
linux> a.out
BEEP
BEEP
BEEP
BEEP
BEEP
BOOM!
bass>
```

Interprocess Communication: Pipes

- Conduit allowing two processes to communicate
  - Unidirectional or bidirectional
  - Full-duplex or half-duplex two-way communication
  - Is parent-child relationship required?
  - Is communication across a network allowed?

Ordinary Unix Pipes

- A unidirectional data channel that can be used for interprocess communication
- Treated as a special type of file, accessed using read() and write()
- Cannot be accessed from outside the process that created it unless inherited (by a child)
- Pipe ceases to exist once closed or when process terminates
- System calls
  - pipe (int fd[])
  - dup2

Example

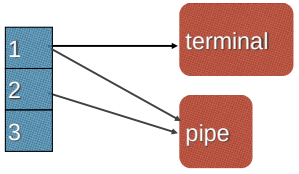
- pipe(int fd[])
  - fd[0] = read_end
  - fd[1] = write_end

fd[1] fd[0]

```
int fd[2];
pid_t pid;

pipe(fd);
pid = fork();
if (pid > 0) {
   /* Parent Process */
   close (fd[0]);

   /* Write a message to the child process */
   write (fd[1], write_msg, strlen(write_msg)+1);
   close (fd[1]);
} else {
   /* Child Process */
   close(fd[1]);

   /* Read a message from the parent process */
   read(fd[0], read_msg, BUFFER_SIZE);
   printf("read %s", read_msg);
   close(fd[0];
}
```


dup2() System Call

- Make one file descriptor point to the same file as another
- dup2 (old_fd, new_fd)
- Return value is -1 on error and new_fd on success
- dup2(1,2)



Standard In, Out, and Error

- By convention, file descriptors 0, 1, and 2 are used for:
  - Standard Input
  - Standard Output
  - Standard Error

Class so far …

- Processes
  - Process concept
  - Operations on processes
  - A process's image in a computer
- System protection and kernel mode
- System calls and the interrupt interface
- Signals
  - User-level events
- Pipes
  - Inter-process communication

Disclaimer

- Parts of the lecture slides contain original work from Gary Nutt, Andrew S. Tanenbaum, Abraham Silberschatz, Peter B. Galvin, Greg Gagne, Dave O'Hallaron, Randal Bryant, Kai Shen, and John Criswell. The slides are intended for the sole purpose of instruction of operating systems at the University of Rochester. All copyrighted materials belong to their original owner(s).