

# **CSE 120**

# **Principles of Operating Systems**

**Fall 2024**

**Lecture 2: Architectural Support for  
Operating Systems**

Geoffrey M. Voelker

# Administrivia

---

- Project 0
  - ◆ Due 10/4 11:59pm, done **individually**
  - ◆ Troubleshoot login and permission problems
- Homework #1
  - ◆ Due 10/8
- Project groups
  - ◆ Google form on piazza to collect group member info
  - ◆ Just need one submission per group
  - ◆ **Fill out even if you are working alone**
- Lab hours
  - ◆ Posted as a Google calendar

# Why Start With Architecture?

---

- Operating system functionality fundamentally depends upon the architectural features of the computer
  - ◆ Particularly to enforce **protection** and **resource sharing**
  - ◆ Applications, though, are generally oblivious to HW details
- Architectural support can greatly simplify – or complicate – OS tasks
  - ◆ Early PC operating systems (DOS, MacOS) lacked virtual memory because the architecture did not support it
  - ◆ Hardware virtualization possible just with software (VMware), but easier and more efficient with hardware support

# Types of Arch Support

---

- **Manipulating privileged machine state**
  - ◆ Privileged instructions
  - ◆ Manipulate device registers
  - ◆ Manage memory protection (e.g., TLB entries)
- **Generating and handling “events”**
  - ◆ System calls, interrupts, exceptions, etc.
  - ◆ Respond to external events
  - ◆ CPU requires software intervention to handle fault or trap
- **Mechanisms for concurrency, synchronization**
  - ◆ Interrupts, atomic instructions

# Privileged Instructions

---

- A subset of instructions of every CPU is restricted to use only by the OS
  - ◆ Known as **privileged** (**protected**) instructions
- Only the operating system can ...
  - ◆ Directly access I/O devices (disks, printers, etc.)
    - » **Security, fairness (why?)**
  - ◆ Manipulate memory management state
    - » **Page table pointers, page protection, TLB management, etc.**
  - ◆ Manipulate protected control registers
    - » **Kernel mode, interrupt level**
  - ◆ Execute the halt instruction
    - » **Protection (why?)**

## HLT—Halt

| Opcode | Instruction | Op/En | 64-Bit Mode | Compat/Leg Mode | Description |
|--------|-------------|-------|-------------|-----------------|-------------|
| F4     | HLT         | Z0    | Valid       | Valid           | Halt        |

### Instruction Operand Encoding

| Op/En | Operand 1 | Operand 2 | Operand 3 | Operand 4 |
|-------|-----------|-----------|-----------|-----------|
| Z0    | NA        | NA        | NA        | NA        |

### Description

Stops instruction execution and places the processor in a HALT state. An enabled interrupt (including NMI and SMI), a debug exception, the BINIT# signal, the INIT# signal, or the RESET# signal will resume execution. If an interrupt (including NMI) is used to resume execution after a HLT instruction, the saved instruction pointer (CS:EIP) points to the instruction following the HLT instruction.

When a HLT instruction is executed on an Intel 64 or IA-32 processor supporting Intel Hyper-Threading Technology, only the logical processor that executes the instruction is halted. The other logical processors in the physical processor remain active, unless they are each individually halted by executing a HLT instruction.

The HLT instruction is a privileged instruction. When the processor is running in protected or virtual-8086 mode, the privilege level of a program or procedure must be 0 to execute the HLT instruction.

This instruction’s operation is the same in non-64-bit modes and 64-bit mode.

### Operation

Enter Halt state;

### Flags Affected

None

### Protected Mode Exceptions

#GP(0) If the current privilege level is not 0.

# OS Protection

---

- When can a privileged instruction execute?
  - ◆ Architecture must support (at least) two modes of operation: **kernel mode** and **user mode**
    - » VAX, x86 support four modes (earlier architectures even more)
    - » Why? Protect the OS from itself (software engineering)
  - ◆ Mode is indicated by a status bit in a protected control register
  - ◆ **User programs execute in user mode**
  - ◆ **OS executes in kernel, privileged mode (OS == “kernel”)**
- **Privileged instructions only execute in kernel mode**
  - ◆ CPU checks mode bit when privileged instruction executes
  - ◆ Setting mode bit must be a privileged instruction
  - ◆ Attempts to execute in user mode are detected and prevented

# **(Live Demo)**

---



# Memory Protection

---

- OS must be able to protect programs from each other
- OS must protect itself from user programs
- **May or may not protect user programs from OS**
  - ◆ Raises question of whether programs should trust the OS
  - ◆ Untrusted operating systems? (Intel SGX)
- Memory management hardware provides memory protection mechanisms
  - ◆ Page table pointers, page protection, segmentation, TLB
- Manipulating memory management hardware uses protected (privileged) operations

# Events

---

- An event is an **unnatural** change in control flow
  - ◆ Events immediately stop current execution
  - ◆ Changes mode, context (machine state), or both
- The OS defines a handler for each event type
  - ◆ Event handlers always execute in kernel mode
  - ◆ The specific types of events are defined by the machine
- Once the system is booted, all entry to the kernel occurs as the result of an event
  - ◆ In effect, the operating system is one big event handler
  - ◆ OS only executes in reaction to events (!)

# Types of Events

---

- Two kinds of events, **interrupts** and **exceptions**
- Interrupts are caused by an external event
  - ◆ Device finishes I/O, timer expires, etc.
  - ◆ Analogy: Receiving a phone call, text message
- Exceptions are caused by executing instructions
  - ◆ CPU requires software intervention to handle a fault or trap
- Events can be **unexpected** or **deliberate**
- Unexpected events are, well, unexpected
  - ◆ **What is an example?**
- Deliberate events are scheduled by OS or application
  - ◆ **Syscalls are most common deliberate event**

# Faults

---

- Hardware detects and reports **exceptional** conditions
  - ◆ Page fault, unaligned access, divide by zero
- Upon exception, hardware **faults** (verb)
  - ◆ **Must save state (PC, regs, mode, etc.) so that the faulting process can be restarted**
- Modern OSes use VM faults for many functions
  - ◆ Debugging, end-of-stack, garbage collection, copy-on-write
- Fault exceptions are a performance optimization
  - ◆ Could detect faults by inserting extra instructions into code (think array bounds checking), but at a significant performance penalty

# Handling Faults (Recovery)

---

- Some faults are handled by “fixing” the exceptional condition and returning to the faulting context
  - ◆ Page faults cause the OS to bring missing pages into memory
  - ◆ Fault handler resets PC of faulting context to re-execute instruction that caused the page fault
- Some faults are handled by notifying the process
  - ◆ Fault handler changes the saved context to transfer control to a user-mode handler on return from fault
  - ◆ Handler must be registered with OS
  - ◆ Unix **signals** or Win **user-mode Async Procedure Calls (APCs)**
    - » SIGALRM, SIGHUP, SIGTERM, SIGSEGV, etc.

# Handling Faults (Termination)

---

- The kernel may handle unrecoverable faults by killing the user process
  - ◆ Program fault with no registered handler
  - ◆ Halt process, write process state to file, destroy process
  - ◆ In Unix, the default action for many signals (e.g., SIGSEGV)
- What about faults in the kernel?
  - ◆ Dereference NULL, divide by zero, undefined instruction
  - ◆ These faults considered fatal, operating system crashes
  - ◆ Unix panic, Windows “Blue screen of death”
    - » Kernel is halted, state dumped to a core file, machine locked up

# **(Live Demo)**

---

# Language Support

---

- Structured exception handling is language support for handling exceptions

```
public class DivideByZero {
    public static void main (String args[]) {
        try {
            int x = 0;
            System.out.println (1/x);
        } catch (ArithmeticException e) {
            System.out.print (e.getMessage());
            System.out.println (": zoinks! divide by zero");
        }
    }
}
```



# System Calls

---

- For a user program to do something “privileged” (e.g., I/O) it must call an OS procedure
  - ◆ Known as **crossing the protection boundary**, or **protected procedure call**, or **protected control transfer**
- CPU ISA provides a **system call** instruction that:
  - ◆ Causes an exception, which vectors to a kernel handler
  - ◆ Passes a parameter determining the system routine to call
  - ◆ Saves caller state (PC, regs, mode) so it can be restored
  - ◆ Returning from system call restores this state
- Requires architectural support to:
  - ◆ **Restore saved state, reset mode, resume execution**

## INT *n*/INTO/INT3/INT1—Call to Interrupt Procedure

| Opcode       | Instruction     | Op/En | 64-Bit Mode | Compat/Leg Mode | Description  |
|--------------|-----------------|-------|-------------|-----------------|--|
| CC           | INT3            | Z0    | Valid       | Valid           | Generate breakpoint trap.  |
| CD <i>ib</i> | INT <i>imm8</i> | I     | Valid       | Valid           | Generate software interrupt with vector specified by immediate byte. |
| CE           | INTO            | Z0    | Invalid     | Valid           | Generate overflow trap if overflow flag is 1.                        |
| F1           | INT1            | Z0    | Valid       | Valid           | Generate debug trap.   |

### Instruction Operand Encoding

| Op/En | Operand 1 | Operand 2 | Operand 3 | Operand 4 |
|-------|-----------|-----------|-----------|-----------|
| Z0    | NA        | NA        | NA        | NA        |
| I     | imm8      | NA        | NA        | NA        |

### Description

The INT *n* instruction generates a call to the interrupt or exception handler specified with the destination operand (see the section titled “Interrupts and Exceptions” in Chapter 6 of the *Intel® 64 and IA-32 Architectures Software Developer’s Manual, Volume 1*). The destination operand specifies a vector from 0 to 255, encoded as an 8-bit unsigned intermediate value. Each vector provides an index to a gate descriptor in the IDT. The first 32 vectors are reserved by Intel for system use. Some of these vectors are used for internally generated exceptions.

The INT *n* instruction is the general mnemonic for executing a software-generated call to an interrupt handler. The INTO instruction is a special mnemonic for calling overflow exception (#OF), exception 4. The overflow interrupt checks the OF flag in the EFLAGS register and calls the overflow interrupt handler if the OF flag is set to 1. (The INTO instruction cannot be used in 64-bit mode.)

The INT3 instruction uses a one-byte opcode (CC) and is intended for calling the debug exception handler with a breakpoint exception (#BP). (This one-byte form is useful because it can replace the first byte of any instruction at which a breakpoint is desired, including other one-byte instructions, without overwriting other instructions.)

The INT1 instruction also uses a one-byte opcode (F1) and generates a debug exception (#DB) without setting any bits in DR6.<sup>1</sup> Hardware vendors may use the INT1 instruction for hardware debug. For that reason, Intel recom-

### A6.7.136 SVC (formerly SWI)

Generates a supervisor call. See *Exceptions in the ARM Architecture Reference Manual*.

Use it as a call to an operating system to provide a service.

**Encoding T1**      All versions of the Thumb ISA.

SVC<C> #<imm8>

|    |    |    |    |    |    |   |   |      |   |   |   |   |   |   |   |
|----|----|----|----|----|----|---|---|------|---|---|---|---|---|---|---|
| 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7    | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| 1  | 1  | 0  | 1  | 1  | 1  | 1 | 1 | imm8 |   |   |   |   |   |   |   |

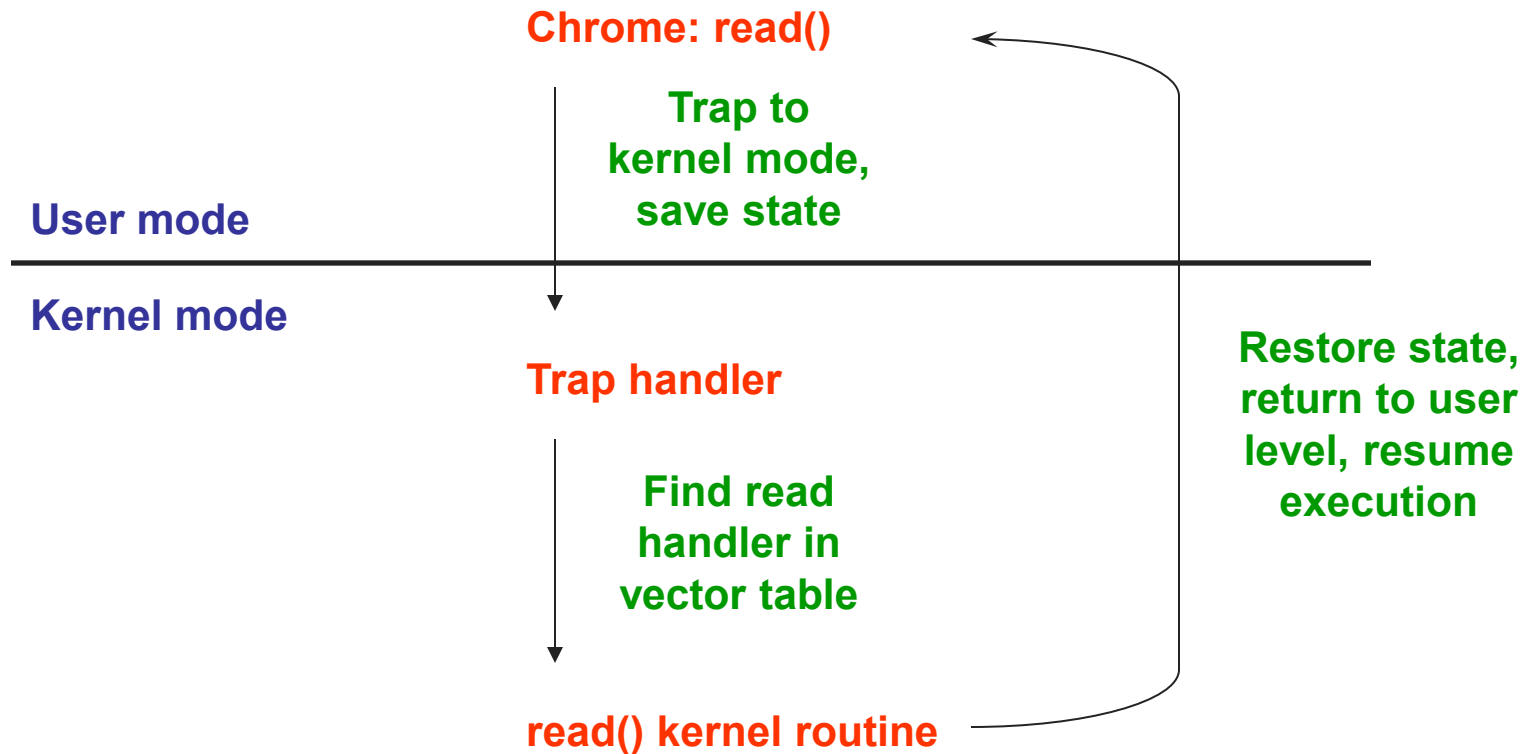
imm32 = ZeroExtend(imm8, 32);

// imm32 is for assembly/disassembly, and is ignored by hardware. SVC handlers in some  
// systems interpret imm8 in software, for example to determine the required service.

# Nachos (test/start.s)

```
/* -----  
 * System call stubs:  
 * Assembly language assist to make system calls to the Nachos kernel.  
 * There is one stub per system call, that places the code for the  
 * system call into register r2, and leaves the arguments to the  
 * system call alone (in other words, arg1 is in r4, arg2 is  
 * in r5, arg3 is in r6, arg4 is in r7)  
 *  
 * The return value is in r2. This follows the standard C calling  
 * convention on the MIPS.  
 * -----  
 */  
  
#define SYSCALLSTUB(name, number) \  
    .globl name          ; \  
    .ent name           ; \  
name:                   ; \  
    addiu $2,$0,number  ; \  
    syscall             ; \  
    j $31               ; \  
    .end name
```

# System Call



# LINUX System Call Quick Reference

Jialong He

[Jialong\\_he@bigfoot.com](mailto:Jialong_he@bigfoot.com)

[http://www.bigfoot.com/~jialong\\_he](http://www.bigfoot.com/~jialong_he)

## Introduction

System call is the services provided by Linux kernel. In C programming, it often uses functions defined in **libc** which provides a wrapper for many system calls. Manual page section 2 provides more information about system calls. To get an overview, use "man 2 intro" in a command shell.

It is also possible to invoke **syscall()** function directly. Each system call has a function number defined in **<syscall.h>** or **<unistd.h>**. Internally, system call is invoked by software interrupt 0x80 to transfer control to the kernel. System call table is defined in Linux kernel source file "arch/i386/kernel/entry.S".

## System Call Example

```
#include <syscall.h>
#include <unistd.h>
#include <stdio.h>
#include <sys/types.h>

int main(void) {

    long ID1, ID2;
    /*-----*/
    /* direct system call      */
    /* SYS_getpid (func no. is 20) */
    /*-----*/
    ID1 = syscall(SYS_getpid);
    printf ("syscall(SYS_getpid)=%ld\n", ID1);

    /*-----*/
    /* "libc" wrapped system call */
    /* SYS_getpid (Func No. is 20) */
    /*-----*/
    ID2 = getpid();
    printf ("getpid()=%ld\n", ID2);

    return (0);
}
```

## System Call Quick Reference

| No | Func Name               | Description                   | Source                            |
|----|-------------------------|-------------------------------|-----------------------------------|
| 1  | <a href="#">exit</a>    | terminate the current process | <i>kernel/exit.c</i>              |
| 2  | <a href="#">fork</a>    | create a child process        | <i>arch/i386/kernel/process.c</i> |
| 3  | <a href="#">read</a>    | read from a file descriptor   | <i>fs/read_write.c</i>            |
| 4  | <a href="#">write</a>   | write to a file descriptor    | <i>fs/read_write.c</i>            |
| 5  | <a href="#">open</a>    | open a file or device         | <i>fs/open.c</i>                  |
| 6  | <a href="#">close</a>   | close a file descriptor       | <i>fs/open.c</i>                  |
| 7  | <a href="#">waitpid</a> | wait for process termination  | <i>kernel/exit.c</i>              |

|    |                            |   |                                    |
|----|----------------------------|---|------------------------------------|
| 8  | <a href="#">creat</a>      | create a file or device ("man 2 open" for information)                      | <i>fs/open.c</i>                   |
| 9  | <a href="#">link</a>       | make a new name for a file  | <i>fs/namei.c</i>                  |
| 10 | <a href="#">unlink</a>     | delete a name and possibly the file it refers to                            | <i>fs/namei.c</i>                  |
| 11 | <a href="#">execve</a>     | execute program   | <i>arch/i386/kernel/process.c</i>  |
| 12 | <a href="#">chdir</a>      | change working directory  | <i>fs/open.c</i>                   |
| 13 | <a href="#">time</a>       | get time in seconds   | <i>kernel/time.c</i>               |
| 14 | <a href="#">mknod</a>      | create a special or ordinary file   | <i>fs/namei.c</i>                  |
| 15 | <a href="#">chmod</a>      | change permissions of a file  | <i>fs/open.c</i>                   |
| 16 | <a href="#">lchown</a>     | change ownership of a file  | <i>fs/open.c</i>                   |
| 18 | <a href="#">stat</a>       | get file status   | <i>fs/stat.c</i>                   |
| 19 | <a href="#">lseek</a>      | reposition read/write file offset   | <i>fs/read_write.c</i>             |
| 20 | <a href="#">getpid</a>     | get process identification  | <i>kernel/sched.c</i>              |
| 21 | <a href="#">mount</a>      | mount filesystems   | <i>fs/super.c</i>                  |
| 22 | <a href="#">umount</a>     | unmount filesystems   | <i>fs/super.c</i>                  |
| 23 | <a href="#">setuid</a>     | set real user ID  | <i>kernel/sys.c</i>                |
| 24 | <a href="#">getuid</a>     | get real user ID  | <i>kernel/sched.c</i>              |
| 25 | <a href="#">stime</a>      | set system time and date  | <i>kernel/time.c</i>               |
| 26 | <a href="#">ptrace</a>     | allows a parent process to control the execution of a child process         | <i>arch/i386/kernel/ptrace.c</i>   |
| 27 | <a href="#">alarm</a>      | set an alarm clock for delivery of a signal                                 | <i>kernel/sched.c</i>              |
| 28 | <a href="#">fstat</a>      | get file status   | <i>fs/stat.c</i>                   |
| 29 | <a href="#">pause</a>      | suspend process until signal  | <i>arch/i386/kernel/sys_i386.c</i> |
| 30 | <a href="#">utime</a>      | set file access and modification times                                      | <i>fs/open.c</i>                   |
| 33 | <a href="#">access</a>     | check user's permissions for a file   | <i>fs/open.c</i>                   |
| 34 | <a href="#">nice</a>       | change process priority   | <i>kernel/sched.c</i>              |
| 36 | <a href="#">sync</a>       | update the super block  | <i>fs/buffer.c</i>                 |
| 37 | <a href="#">kill</a>       | send signal to a process  | <i>kernel/signal.c</i>             |
| 38 | <a href="#">rename</a>     | change the name or location of a file                                       | <i>fs/namei.c</i>                  |
| 39 | <a href="#">mkdir</a>      | create a directory  | <i>fs/namei.c</i>                  |
| 40 | <a href="#">rmdir</a>      | remove a directory  | <i>fs/namei.c</i>                  |
| 41 | <a href="#">dup</a>        | duplicate an open file descriptor   | <i>fs/fcntl.c</i>                  |
| 42 | <a href="#">pipe</a>       | create an interprocess channel  | <i>arch/i386/kernel/sys_i386.c</i> |
| 43 | <a href="#">times</a>      | get process times   | <i>kernel/sys.c</i>                |
| 45 | <a href="#">brk</a>        | change the amount of space allocated for the calling process's data segment | <i>mm/mmap.c</i>                   |
| 46 | <a href="#">setgid</a>     | set real group ID   | <i>kernel/sys.c</i>                |
| 47 | <a href="#">getgid</a>     | get real group ID   | <i>kernel/sched.c</i>              |
| 48 | <a href="#">sys_signal</a> | ANSI C signal handling  | <i>kernel/signal.c</i>             |
| 49 | <a href="#">geteuid</a>    | get effective user ID   | <i>kernel/sched.c</i>              |
| 50 | <a href="#">getegid</a>    | get effective group ID  | <i>kernel/sched.c</i>              |

# Referencing OS Data

---

- Processes and the OS are in different address spaces
  - ♦ How can the OS return references to kernel data structures?
- A naming issue
  - ♦ Use integer object handles or descriptors
    - » Unix file descriptors, Windows HANDLEs
    - » Only meaningful as parameters to other system calls
  - ♦ Also called capabilities (more later when we do protection)

# Interrupts

---

- Interrupts signal asynchronous events
  - ◆ I/O hardware interrupts
  - ◆ Software and hardware timers
- Interrupts on modern CPUs are **precise**
  - ◆ CPU transfers control only on instruction boundaries
  - An interrupt can occur in between any two instructions



# Timer

---

- **The timer is critical for an operating system**
- It is the fallback mechanism by which the OS reclaims control over the machine
  - ◆ Timer is set to generate an interrupt after a period of time
    - » Setting timer is a privileged instruction
  - ◆ When timer expires, generates an interrupt
  - ◆ Handled by kernel, which controls what runs next
    - » Basis for OS **scheduler** (*more later in quarter...*)
- Prevents infinite loops
  - ◆ OS can always regain control from erroneous or malicious programs that try to hog CPU
- Also used for time-based functions (e.g., `sleep`)

# I/O Completion

---

- Interrupts are the basis for asynchronous I/O
  - ◆ OS initiates I/O
  - ◆ Device operates independently of rest of machine
  - ◆ Device sends an interrupt signal to CPU when done
  - ◆ OS maintains a vector table containing a list of addresses of kernel routines to handle various events
  - ◆ CPU looks up kernel address indexed by interrupt number, context switches to routine
- If you have ever installed early versions of Windows, you now know what IRQs are for

# Synchronization

---

- Interrupts cause difficult problems
  - ◆ An interrupt can occur at any time
  - ◆ A handler can execute that interferes with code that was interrupted
- OS must be able to synchronize concurrent execution
- Need to guarantee that short instruction sequences execute atomically
  - ◆ Disable interrupts – turn off interrupts before sequence, execute sequence, turn interrupts back on
  - ◆ Special atomic instructions – read/modify/write a memory address atomically
    - » XCHG instruction on x86

# Next Time...

---

- Read Chapters 3-5 (Processes)
- Start on Project 0, Homework 1
- Submit group info on Google form