CSE 153
Design of Operating Systems

Winter 2023

Lecture 4: Processes

Last class

- OS structure, operation, and interaction with user progs
  - **Privileged mode**: To enforce isolation and manage resources, OS must have exclusive powers not available to users
    - » How does the switch happen securely?
  - **OS is not running unless there is an event**:
    - » OS schedules a user process to run then goes to sleep
    - » It wakes up (who wakes it?) to handle events
    - » Many types of events
  - **Program view and system calls:** program asks the OS when it needs a privileged operation

Categorizing Events

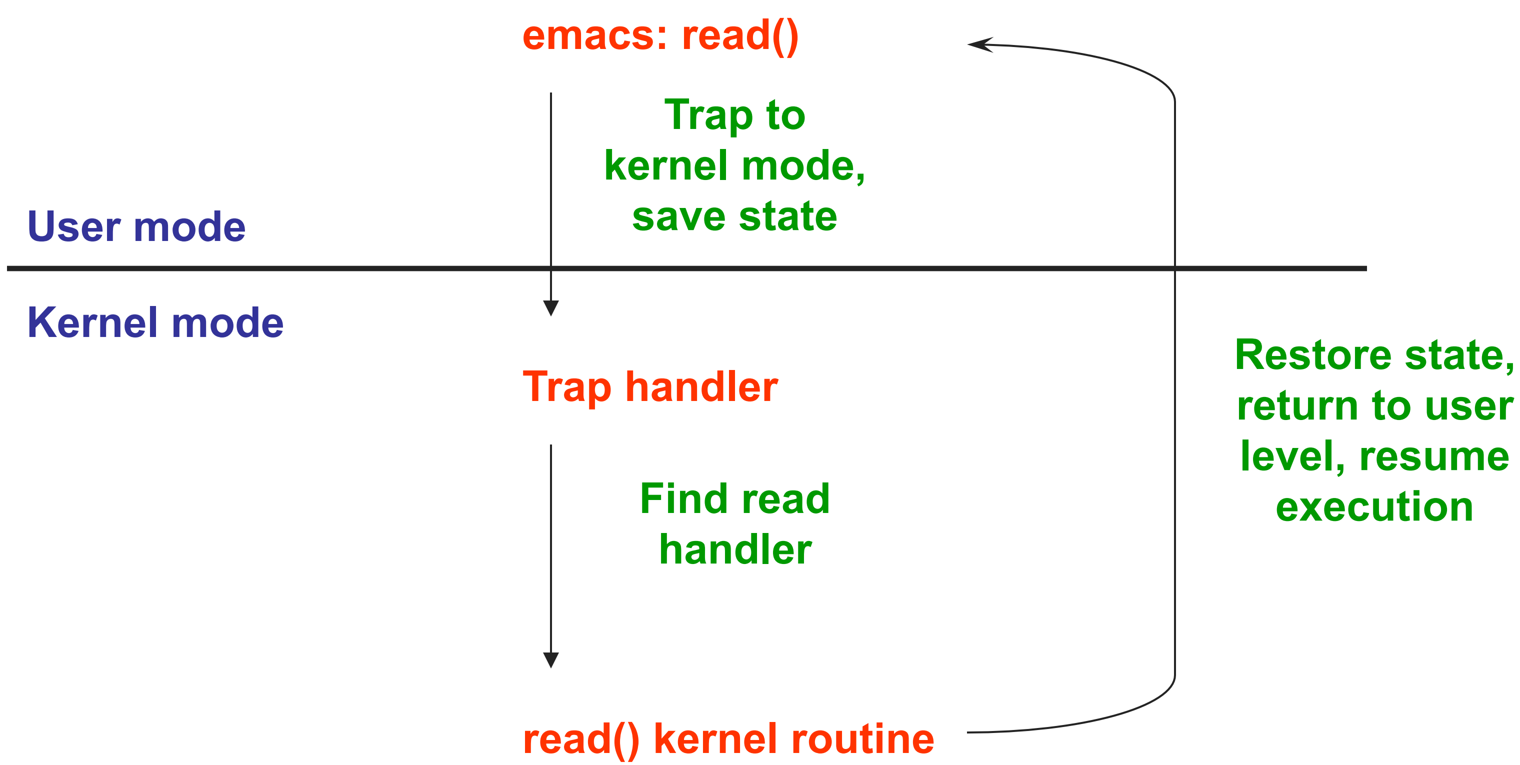
	Unexpected	Deliberate
Synchronous	fault	syscall trap
Asynchronous	interrupt	signal

System Calls

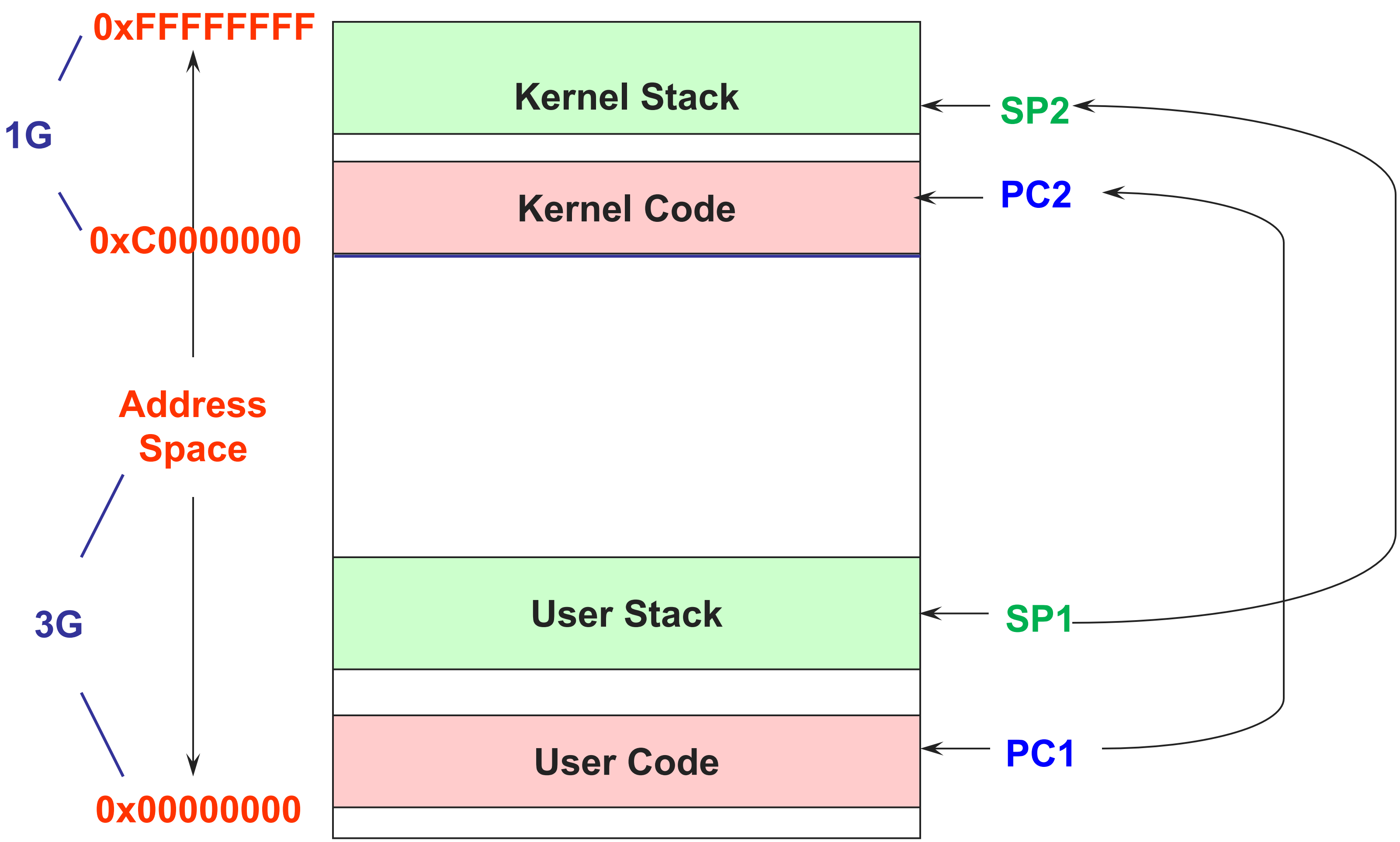
- For a user program to do something "privileged" (e.g., I/O) it must call an OS procedure
  - Known as crossing the protection boundary, or a protected procedure call

- Hardware provides a system call instruction that:
  - Causes an exception, which invokes a kernel handler
    - Passes a parameter determining the system routine to call
  - Saves caller state (PC, regs, mode) so it can be restored
    - Why save mode?
  - Returning from system call restores this state

System Call


emacs: read()
Trap to
kernel mode,
save state
User mode
Kernel mode
Trap handler
Find read
handler
Restore state,
return to user
level, resume
execution
read() kernel routine

Another view (FYI for now)



System Call Questions

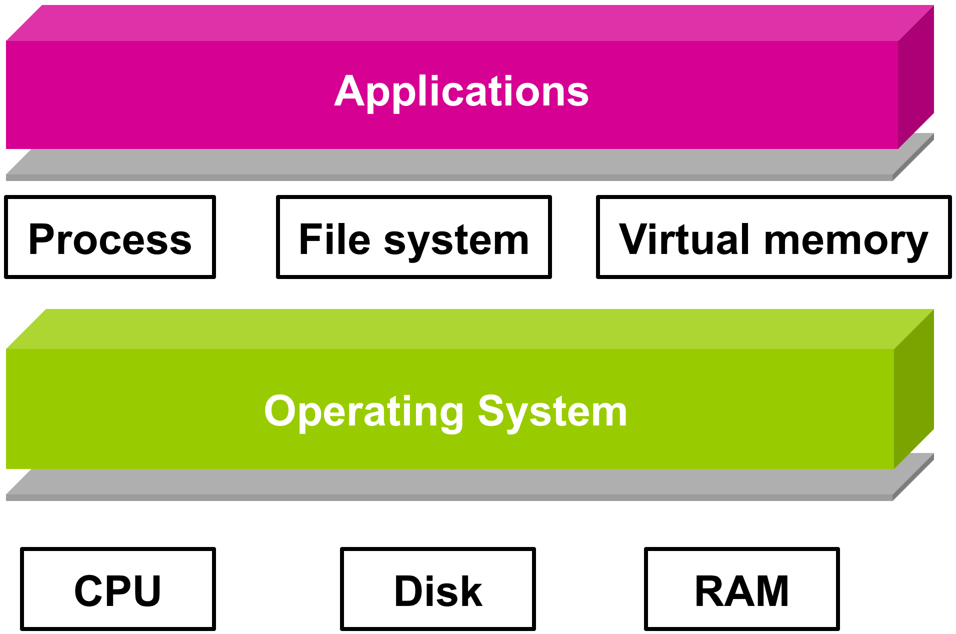
- There are hundreds of syscalls. How do we let the kernel know which one we intend to invoke?
  - Before issuing **int $0x80** or **sysenter**, set **%eax/%rax** with the syscall number

- System calls are like function calls, but how to pass parameters?
  - Just like calling convention in syscalls, typically passed through **%ebx, %ecx, %edx, %esi, %edi, %ebp**

Timer

- The key to a timesharing OS
- The fallback mechanism by which the OS reclaims control
  - 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 the OS, forcing a switch from the user program
    - » Basis for OS scheduler *(more later…)*
- Also used for time-based functions (e.g., *sleep()*)

OS Abstractions



Today, we start discussing the first abstraction that enables us to virtualize (i.e., share) the CPU – processes!

What is virtualization?

- What is a virtual something?
  - Somehow not real? But still functional?

- Provide illusion for each program of own copy of resources
  - Lets say the CPU or memory; every program thinks it has its own
  - In reality, limited physical resources (e.g., 1 CPU)
    - » It must be shared! (in time, or space)

- Frees up programs from worrying about sharing
  - The OS implements sharing, creating illusion of exclusive resources →Virtualization!

- Virtual resource provided as an object with defined operations on it. → Abstraction

Virtualizing the CPU -- Processes

- This lecture starts a class segment that covers processes, threads, and synchronization
  - Basis for Midterm and Project 1

- Today's topics are processes and process management
  - How do we virtualize the CPU?
    - » Virtualization: give each program the illusion of its own CPU
    - » What is the magic? We only have one real CPU
  - How are applications represented in the OS?
  - How is work scheduled in the CPU?

The Process

- The process is the OS abstraction for execution
  - It is the unit of execution
  - It is the unit of scheduling
- A process is a program in execution
  - Programs are static entities with the potential for execution
  - Process is the animated/active program
    - » Starts from the program, but also includes dynamic state
    - » As the representative of the program, it is the "owner" of other resources (memory, files, sockets, …)
- How does the OS implement this abstraction?
  - How does it share the CPU?

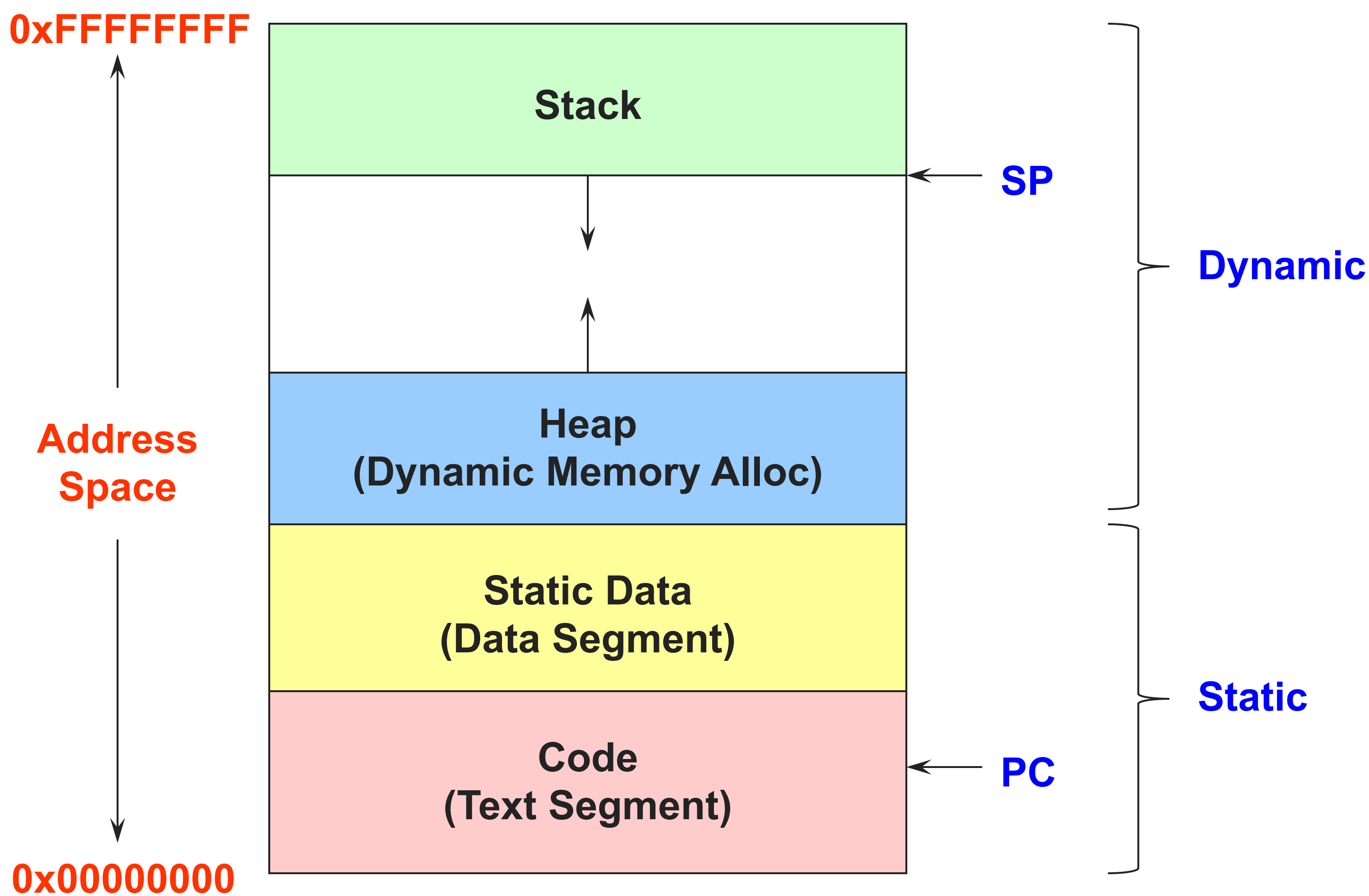
How to support this abstraction?

- First, we'll look at what state a process encapsulates
  - State of the virtual processor we are giving to each program
- Next we will talk about process behavior/CPU time sharing
  - How to implement the process illusion
- Next, we discuss how the OS implements this abstraction
  - What data structures it keeps, and the role of the scheduler
- Finally, we see the process interface offered to programs
  - How to use this abstraction
  - Next class

Process Components

- A process contains all the state for a program in execution
  - An address space containing
    - » Static memory:
      - The code and input data for the executing program
    - » Dynamic memory:
      - The memory allocated by the executing program
      - An execution stack encapsulating the state of procedure calls
  - Control registers such as the program counter (PC)
  - A set of general-purpose registers with current values
  - A set of operating system resources
    - » Open files, network connections, etc.
- A process is named using its process ID (PID)

Address Space (memory abstraction)



How to support this abstraction?

- First, we'll look at what state a process encapsulates
  - State of the virtual processor we are giving to each program
- Next we will talk about process behavior/CPU time sharing
  - How to implement the process illusion
- Next, we discuss how the OS implements this abstraction
  - What data structures it keeps, and the role of the scheduler
- Finally, we see the process interface offered to programs
  - How to use this abstraction
  - Next class

Process Execution State

- A process is born, executes for a while, and then dies
- The process execution state that indicates what it is currently doing
  - Running: Executing instructions on the CPU
    - » It is the process that has control of the CPU
    - » How many processes can be in the running state simultaneously?
  - Ready: Waiting to be assigned to the CPU
    - » Ready to execute, but another process is executing on the CPU
  - Waiting: Waiting for an event, e.g., I/O completion
    - » It cannot make progress until event is signaled (disk completes)

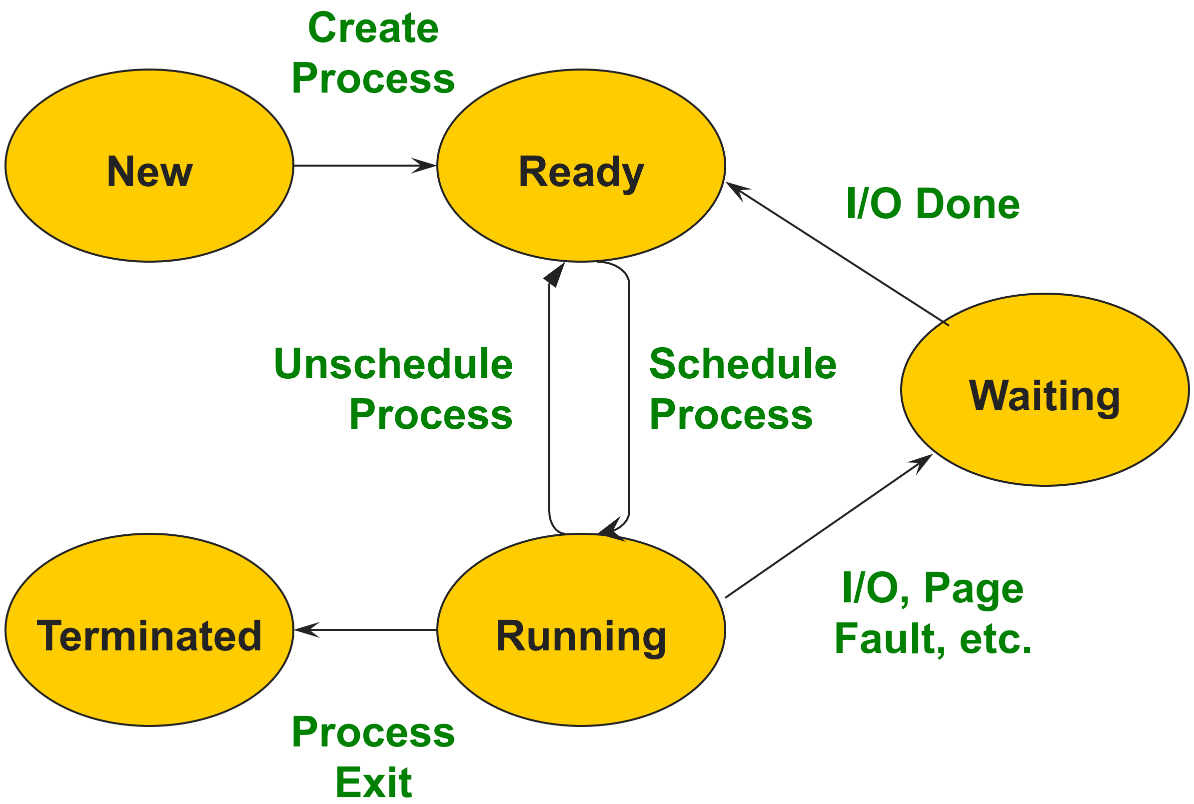
Execution state (cont'd)

- As a process executes, it moves from state to state
  - Unix "ps -x": STAT column indicates execution state
  - What state do you think a process is in most of the time?
  - How many processes can a system support?

```
PROCESS STATE CODES
       Here are the different values that the s, stat and state output specifiers (header "S
       D    uninterruptible sleep (usually IO)
       R    running or runnable (on run queue)
       S    interruptible sleep (waiting for an event to complete)
       T    stopped, either by a job control signal or because it is being traced.
       W    paging (not valid since the 2.6.xx kernel)
       X    dead (should never be seen)
       Z    defunct ("zombie") process, terminated but not reaped by its parent.

       For BSD formats and when the stat keyword is used, additional characters may be displ
       <    high-priority (not nice to other users)
       N    low-priority (nice to other users)
       L    has pages locked into memory (for real-time and custom IO)
       s    is a session leader
       l    is multi-threaded (using CLONE_THREAD, like NPTL pthreads do)
       +    is in the foreground process group.
```

Execution State Graph



How to support the process abstraction?

- First, we'll look at what state a process encapsulates
  - State of the virtual processor we are giving to each program
- Next we will talk about process behavior/CPU time sharing
  - How to implement the process illusion
- Next, we discuss how the OS implements this abstraction
  - What data structures it keeps, and the role of the scheduler
- Finally, we see the process interface offered to programs
  - How to use this abstraction?
  - What system calls are needed?

How does the OS support this model?

We will discuss three issues:

1. How does the OS represent a process in the kernel?
   - The OS data structure representing each process is called the Process Control Block (PCB)

2. How do we pause and restart processes?
   - We must be able to save and restore the full machine state

3. How do we keep track of all the processes in the system?
   - A lot of queues!

PCB Data Structure

- PCB also is where OS keeps all of a process' hardware execution state when the process is not running
  - » Process ID (PID)
  - » Execution state
  - » Hardware state: PC, SP, regs
  - » Memory management
  - » Scheduling
  - » Accounting
  - » Pointers for state queues
  - » Etc.
- This state is everything that is needed to restore the hardware to the same configuration it was in when the process was switched out of the hardware

Xv6 struct proc

```
enum procstate { UNUSED, EMBRYO, SLEEPING, RUNNABLE, RUNNING, ZOMBIE };

// Per-process state
struct proc {
  uint sz;                     // Size of process memory (bytes)
  pde_t* pgdir;                // Linear address of proc's pgdir
  char *kstack;                // Bottom of kernel stack for this process
  enum procstate state;        // Process state
  volatile int pid;            // Process ID
  struct proc *parent;         // Parent process
  struct trapframe *tf;        // Trap frame for current syscall
  struct context *context;     // Switch here to run process
  void *chan;                  // If non-zero, sleeping on chan
  int killed;                  // If non-zero, have been killed
  struct file *ofile[NOFILE];  // Open files
  struct inode *cwd;           // Current directory
  char name[16];               // Process name (debugging)
};
```

struct proc (Solaris)

```
/*
 * One structure allocated per active process.  It contains all
 * data needed about the process while the process may be swapped
 * out.  Other per-process data (user.h) is also inside the proc structure.
 * Lightweight-process data (lwp.h) and the kernel stack may be swapped out.
 */
typedef struct  proc {
        /*
         * Fields requiring no explicit locking
         */
        struct  vnode *p_exec;          /* pointer to a.out vnode */
        struct  as *p_as;               /* process address space pointer */
        struct  plock *p_lockp;         /* ptr to proc struct's mutex lock */
        kmutex_t p_crlock;              /* lock for p_cred */
        struct  cred    *p_cred;        /* process credentials */
        /*
         * Fields protected by pidlock
         */
        int     p_swapcnt;              /* number of swapped out lwps */
        char    p_stat;                 /* status of process */
        char    p_wcode;                /* current wait code */
        ushort_t p_pidflag;             /* flags protected only by pidlock */
        int     p_wdata;                /* current wait return value */
        pid_t   p_ppid;                 /* process id of parent */
        struct  proc    *p_link;        /* forward link */
        struct  proc    *p_parent;      /* ptr to parent process */
        struct  proc    *p_child;       /* ptr to first child process */
        struct  proc    *p_sibling;     /* ptr to next sibling proc on chain */
        struct  proc    *p_psibling;    /* ptr to prev sibling proc on chain */
        struct  proc    *p_sibling_ns;  /* prt to siblings with new state */
        struct  proc    *p_child_ns;    /* prt to children with new state */
        struct  proc    *p_next;        /* active chain link next */
        struct  proc    *p_prev;        /* active chain link prev */
        struct  proc    *p_nextofkin;   /* gets accounting info at exit */
        struct  proc    *p_orphan;
        struct  proc    *p_nextorph;
```

```
*p_pglink;      /* process group hash chain link next */
struct  proc    *p_ppglink;     /* process group hash chain link prev */
struct  sess    *p_sessp;       /* session information */
struct  pid     *p_pidp;        /* process ID info */
struct  pid     *p_pgidp;       /* process group ID info */
/*
 * Fields protected by p_lock
 */
kcondvar_t p_cv;                /* proc struct's condition variable */
kcondvar_t p_flag_cv;
kcondvar_t p_lwpexit;           /* waiting for some lwp to exit */
kcondvar_t p_holdlwps;          /* process is waiting for its lwps */
                                /* to to be held.  */
ushort_t p_pad1;                /* unused */
uint_t  p_flag;                 /* protected while set. */

/* flags defined below */
clock_t p_utime;                /* user time, this process */
clock_t p_stime;                /* system time, this process */
clock_t p_cutime;               /* sum of children's user time */
clock_t p_cstime;               /* sum of children's system time */
caddr_t *p_segacct;             /* segment accounting info */
caddr_t p_brkbase;              /* base address of heap */
size_t  p_brksize;              /* heap size in bytes */
/*
 * Per process signal stuff.
 */
k_sigset_t p_sig;               /* signals pending to this process */
k_sigset_t p_ignore;            /* ignore when generated */
k_sigset_t p_siginfo;           /* gets signal info with signal */
struct sigqueue *p_sigqueue;    /* queued siginfo structures */
struct sigqhdr *p_sigqhdr;      /* hdr to sigqueue structure pool */
struct sigqhdr *p_signhdr;      /* hdr to signotify structure pool */
uchar_t p_stopsig;              /* jobcontrol stop signal */
```

struct proc (Solaris) (2)

```
/*
 * Special per-process flag when set will fix misaligned memory
 * references.
 */
char    p_fixalignment;

/*
 * Per process lwp and kernel thread stuff
 */
id_t    p_lwpid;              /* most recently allocated lwpid */
int     p_lwpcnt;             /* number of lwps in this process */
int     p_lwprcnt;            /* number of not stopped lwps */
int     p_lwpwait;            /* number of lwps in lwp_wait() */
int     p_zombcnt;            /* number of zombie lwps */
int     p_zomb_max;           /* number of entries in p_zomb_tid */
id_t    *p_zomb_tid;          /* array of zombie lwpids */
kthread_t *p_tlist;           /* circular list of threads */
/*
 * /proc (process filesystem) debugger interface stuff.
 */
k_sigset_t p_sigmask;         /* mask of traced signals (/proc) */
k_fltset_t p_fltmask;         /* mask of traced faults (/proc) */
struct  vnode *p_trace;       /* pointer to primary /proc vnode */
struct  vnode *p_plist;       /* list of /proc vnodes for process */
kthread_t *p_agenttp;         /* thread ptr for /proc agent lwp */
struct watched_area *p_warea; /* list of watched areas */
ulong_t p_nwarea;             /* number of watched areas */
struct watched_page *p_wpage; /* remembered watched pages (vfork) */
int     p_nwpage;             /* number of watched pages (vfork) */
int     p_mapcnt;             /* number of active pr_mappage()s */
struct  proc  *p_rlink;       /* linked list for server */
kcondvar_t p_srwchan_cv;
size_t  p_stksize;            /* process stack size in bytes */
/*
 * Microstate accounting, resource usage, and real-time profiling
 */
hrtime_t p_mstart;            /* hi-res process start time */
hrtime_t p_mterm;             /* hi-res process termination time */
hrtime_t p_mlreal;            /* elapsed time sum over defunct lwps */
hrtime_t p_acct[NMSTATES];    /* microstate sum over defunct lwps */
struct lrusage p_ru;          /* lrusage sum over defunct lwps */
struct itimerval p_rprof_timer; /* ITIMER_REALPROF interval timer */
uintptr_t p_rprof_cyclic;     /* ITIMER_REALPROF cyclic */
uint_t  p_defunct;            /* number of defunct lwps */
/*
 * profiling. A lock is used in the event of multiple lwp's
 * using the same profiling base/size.
 */
kmutex_t p_pflock;            /* protects user profile arguments */
struct prof p_prof;           /* profile arguments */

/*
 * The user structure
 */
struct user p_user;           /* (see sys/user.h) */

/*
 * Doors.
 */
kthread_t          *p_server_threads;
struct door_node   *p_door_list;  /* active doors */
struct door_node   *p_unref_list;
kcondvar_t         p_server_cv;
char               p_unref_thread; /* unref thread created */

/*
 * Kernel probes
 */
uchar_t            p_tnf_flags;
```

struct proc (Solaris) (3)

```
        /*
         * C2 Security  (C2_AUDIT)
         */
        caddr_t p_audit_data;           /* per process audit structure */
        kthread_t       *p_aslwptp;     /* thread ptr representing "aslwp" */
#if defined(i386) || defined(__i386) || defined(__ia64)
        /*
         * LDT support.
         */
        kmutex_t p_ldtlock;             /* protects the following fields */
        struct seg_desc *p_ldt;         /* Pointer to private LDT */
        struct seg_desc p_ldt_desc;     /* segment descriptor for private LDT */
        int p_ldtlimit;                 /* highest selector used */
#endif
        size_t p_swrss;                 /* resident set size before last swap */
        struct aio      *p_aio;         /* pointer to async I/O struct */
        struct itimer   **p_itimer;     /* interval timers */
        k_sigset_t      p_notifsigs;    /* signals in notification set */
        kcondvar_t      p_notifcv;      /* notif cv to synchronize with aslwp */
        timeout_id_t    p_alarmid;      /* alarm's timeout id */
        uint_t          p_sc_unblocked; /* number of unblocked threads */
        struct vnode    *p_sc_door;     /* scheduler activations door */
        caddr_t         p_usrstack;     /* top of the process stack */
        uint_t          p_stkprot;      /* stack memory protection */
        model_t         p_model;        /* data model determined at exec time */
        struct lwpchan_data     *p_lcp; /* lwpchan cache */
        /*
         * protects unmapping and initilization of robust locks.
         */
        kmutex_t        p_lcp_mutexinitlock;
        utrap_handler_t *p_utraps;      /* pointer to user trap handlers */
        refstr_t        *p_corefile;    /* pattern for core file */
#if defined(__ia64)
        caddr_t         p_upstack;      /* base of the upward-growing stack */
        size_t          p_upstksize;    /* size of that stack, in bytes */
        uchar_t         p_isa;          /* which instruction set is utilized */
#endif
        void            *p_rce;         /* resource control extension data */
        struct task     *p_task;        /* our containing task */
        struct proc     *p_taskprev;    /* ptr to previous process in task */
        struct proc     *p_tasknext;    /* ptr to next process in task */
        int             p_lwpdaemon;    /* number of TP_DAEMON lwps */
        int             p_lwpdwait;     /* number of daemons in lwp_wait() */
        kthread_t       **p_tidhash;    /* tid (lwpid) lookup hash table */
        struct sc_data  *p_schedctl;    /* available schedctl structures */
} proc_t;
```

How to pause/restart processes?

- When a process is running, its dynamic state is in memory and some hardware registers
  - Hardware registers include Program counter, stack pointer, control registers, data registers, …
  - To be able to stop and restart a process, we need to completely restore this state
- When the OS stops running a process, it saves the current values of the registers (usually in PCB)
- When the OS restarts executing a process, it loads the hardware registers from the stored values in PCB
- Changing CPU hardware state from one process to another is called a context switch
  - This can happen 100s or 1000s of times a second!