CSE 451: Operating Systems

Spring 2021

Module 4

Processes

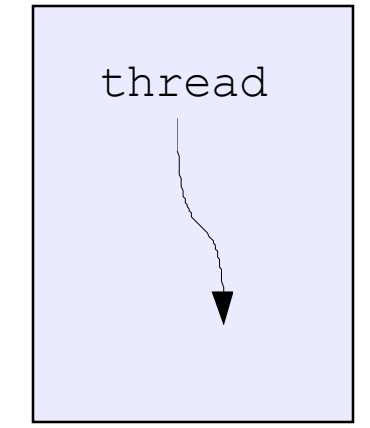
John Zahorjan

Lecture Questions

- What is as process?
- What lower level resources are hidden/simplified by the process abstraction?
- Which aspects of process semantics most commonly important to software design issues? to system management issues?
- Why is process creation broken into fork() and then exec()?
- How do you make fork() less expensive?
- What does a shell do?
- How can processes communicate with each other? why would you want them to?
- Why might you want additional abstraction built above processes? What's the relationship of "user" to "process"? Is "user" the fundamental abstraction?

What is a “process”?

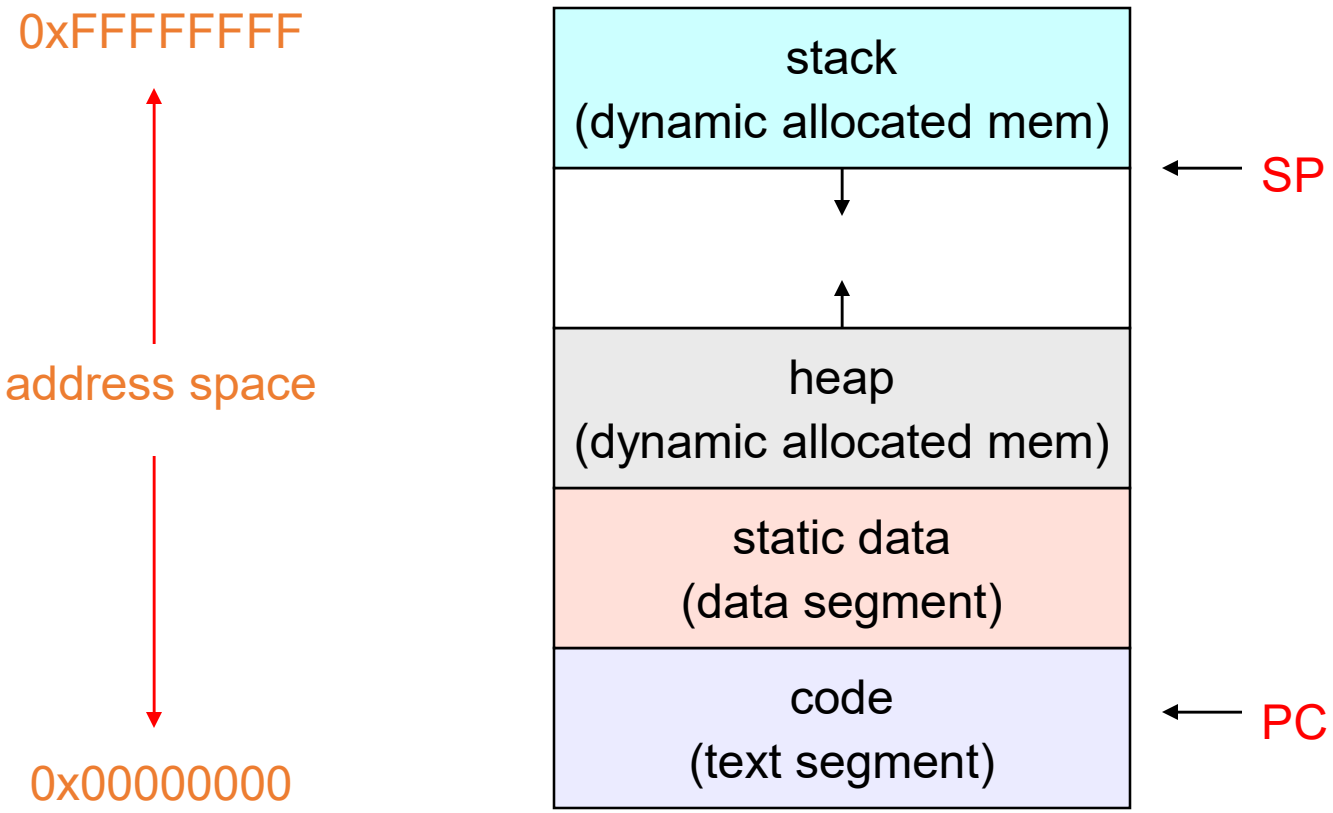
- The process is the OS’s abstraction for execution
  - A process is “a program in execution”
- Simplest (classic) case: a sequential process
  - An address space (an abstraction of memory)
  - A single thread of execution (an abstraction of a single core)
- Traditionally, a sequential process is:
  - The unit of execution
  - The unit of scheduling
  - The unit of failure
  - The dynamic (active) execution context
    - vs. the program – static, just a bunch of bytes



What's "in" a process?

- A process consists of (at least):
  - An address space, containing
    - the code (instructions) for the running program
    - the data for the running program (static data, heap data, stack)
  - CPU state, consisting of
    - The program counter (PC), indicating the next instruction
    - The stack pointer (SP)
    - Other general purpose register (GPR) values
      - Each thread has its own PC, SP, and GPR values
  - A set of OS resources
    - open files, network connections, sound channels, …
- In other words, it's all the stuff you need to run the program
  - or to resume it, if it's interrupted at some point

A process's address space (idealized)



The OS's process namespace

- (Like most things, the particulars depend on the specific OS, but the principles are general)
- The name for a process is called a process ID (PID)
  - an integer
- The PID namespace is global to the system
  - Only one process at a time has a particular PID
- Operations that create processes return a PID
  - E.g., fork()
- Operations on processes take PIDs as an argument
  - E.g., kill(), wait(), nice()

Representation of processes by the OS

- The OS maintains a data structure to keep track of a process's state
  - Called the process control block (PCB) or process descriptor
  - Identified by the PID
- OS keeps all of a process's execution state in (or linked from) the PCB when the process isn't running
  - CPU state (PC, SP, registers, etc.) is transferred out of the hardware registers into the PCB when execution is suspended (transition to blocked or running states)
- When a process is running, its state is spread between the PCB and the CPU
- Note: It's natural to think that there must be some esoteric techniques being used
  - Nope…
    - Except that xk uses some data structures we're pretty sure you wouldn't choose…

The Process Control Block

- The PCB is a data structure with many, many fields:
  - process ID (PID)
  - parent process ID (PPID)
  - execution state
  - program counter, stack pointer, registers
  - address space info
  - user id (uid), group id (gid)
  - scheduling priority
  - accounting info
  - pointers for state queue
  - …

PCBs and CPU state

- When a process is running, its CPU state is on the CPU
  - PC, SP, registers
  - CPU contains current values

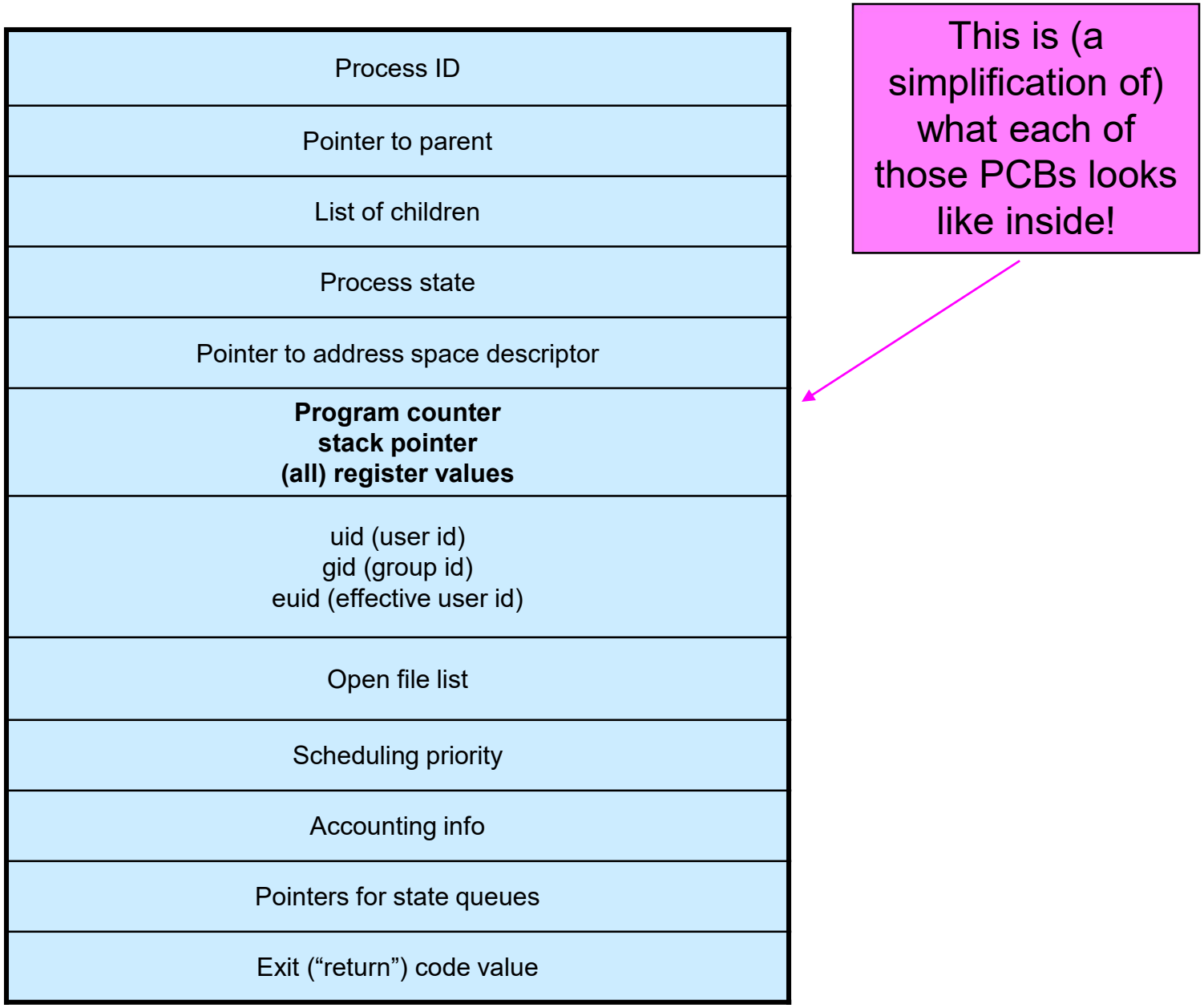
- When the OS gets control because of a …
  - Trap: Program executes a syscall
  - Exception: Program does something unexpected (e.g., page fault)
  - Interrupt: A hardware device requests service

  the OS saves the CPU state of the running process (for instance, in that process's PCB)

  (In xk, the CPU register state is saved at the bottom of the kernel stack associated with the process)

- When the OS returns the process to the running state, it loads the hardware registers with values from that process's PCB – general purpose registers, stack pointer, instruction pointer

- The act of switching the CPU from one process to another is called a context switch
  - systems may do 1000s of switches/sec.
  - takes a few microseconds on today's hardware
    - *See hw 0*
- Choosing which process to run next is called scheduling


Process ID
Pointer to parent
List of children
Process state
Pointer to address space descriptor
Program counter
stack pointer
(all) register values
uid (user id)
gid (group id)
euid (effective user id)
Open file list
Scheduling priority
Accounting info
Pointers for state queues
Exit (“return”) code value
This is (a simplification of) what each of those PCBs looks like inside!

Scope of OS Resources

- OS resources are things like open file tables, network connection points (sockets), pipes, shared memory regions, …
- Each requires its own descriptor block
- If the block is embedded in a PCB, the scope of that resource can be only that one process
- If the block is stored separately and the PCB contains a reference (pointer) to it, the resource can be shared
- What's shared and what isn't is part of OS API design

Allocation of OS Resources

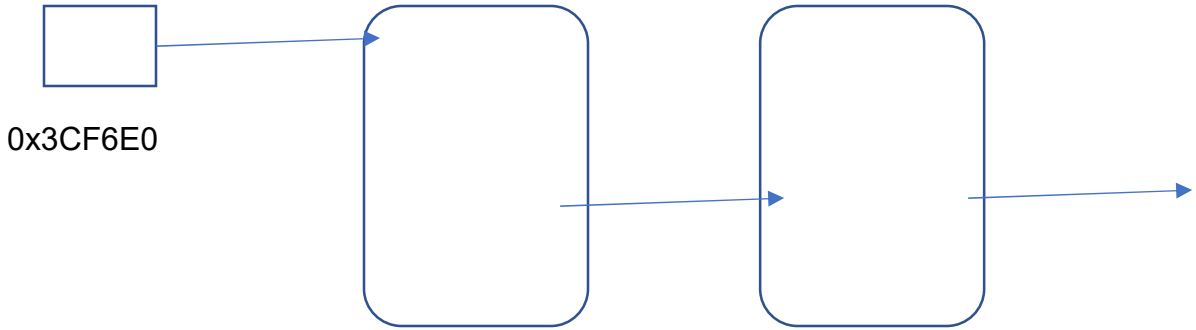
- How should process control blocks be allocated?
  - Statically?
    - A fixed size block of memory is allocated by declaring an array of fixed size in the code
    - A region of memory is allocated during boot
    - Why would you much rather do the latter, if you're going to allocate statically?
  - Dynamically?
    - At the extreme, allocate space for a PCB each time a process is created
    - Free a PCB each time a process terminates
      - Possibly cache free PCBs to avoid allocation/deallocation overheads
  - Which should the OS use?
- You have to worry about running out of memory
  - What to do then?
  - Can't start new process...
    - Means you can't bring up a tool that kills an old process
  - Programmer error can exhaust all system memory...

The OS kernel is not a process

- (The x86 architecture supports switching “task” as part of the hardware action on entry to the OS)
  - But xk (and most OS’s) basically work around that feature
- On entry to the OS, the CPU is still running in the context of the process that was running
  - Address space
  - Registers (to be saved)
  - Current PCB
- On exit back to user level, the context of dispatched process has been re-established
- There’s a brief moment in between when “it’s just code”
  - The CPU doesn’t know anything about processes…

PCB Chaining (except in xk)

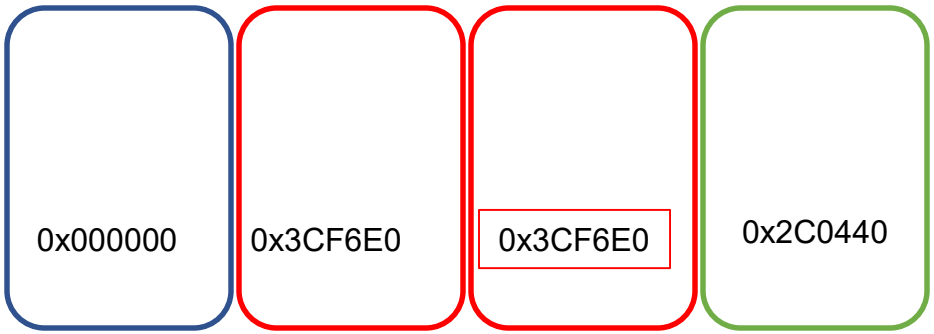
- In “real systems” the PCB contains a pointer field that allows the PCB to be put on (a single, arbitrary) linked list
- For instance, there might be a linked list of processes in the runnable state (i.e., waiting for the cpu)
  - This is often called “the run queue”
- There might be a linked list of processes blocked on disk 2 (either waiting for an operation to complete or else to initiate one)



- Every blocked processes is on some list
- Every blocked process is on exactly one list
  - You can be blocked on only one thing (because when you block on the first thing you can’t execute code to block on the second)
  - As usual, such a simple constraint is too confining, and so there’s a way to work around it (select/poll/epoll)

PCB Chaining in xk

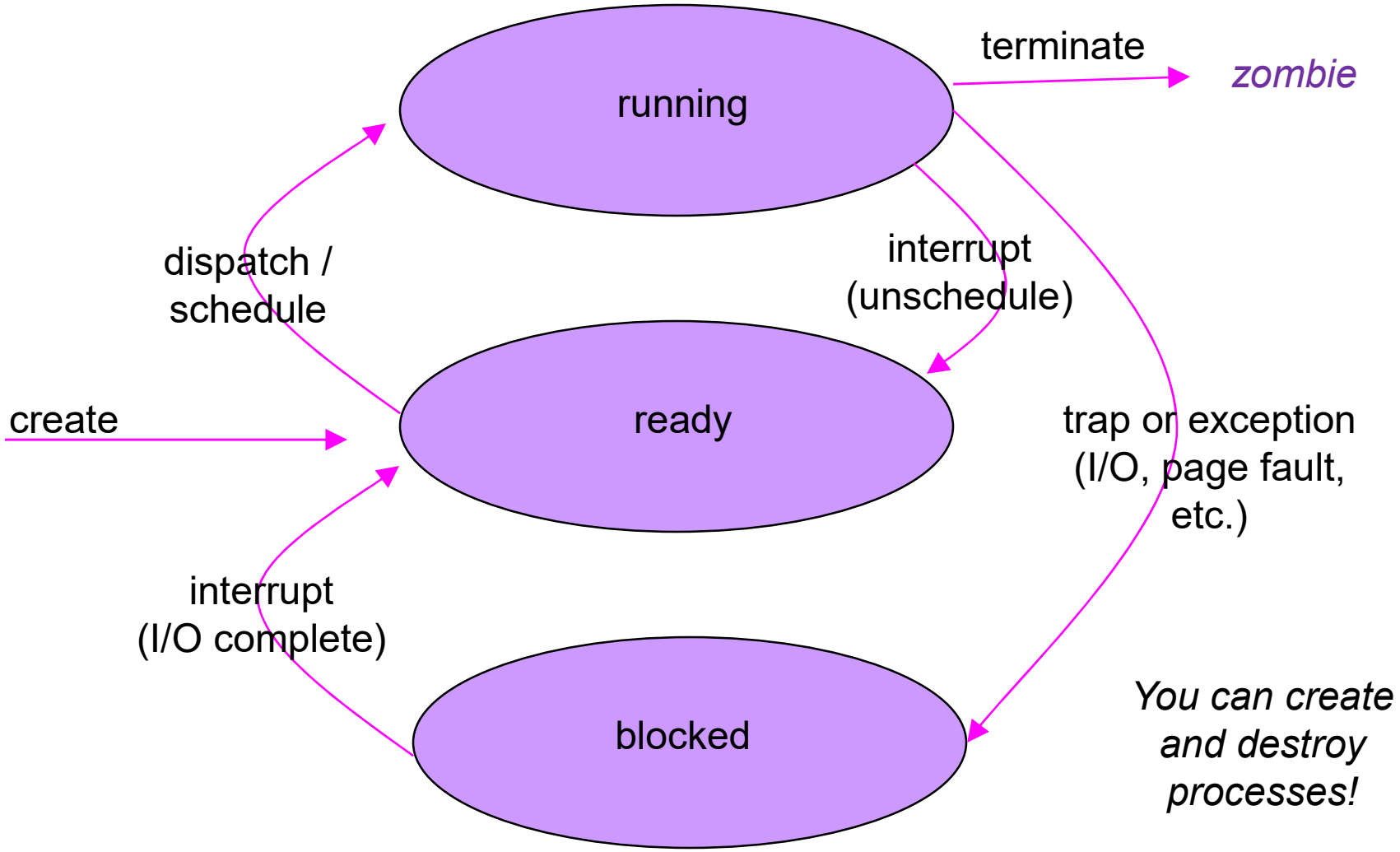
- xk uses a “simpler” scheme
  - A statically allocated array of PCBs (like always)
  - A field that can hold a 32-bit value
    - The value is logically the id of a list
    - We need the id’s to be unique
    - Use the address of something – the thing that would hold the head pointer of a linked list in a normal implementation
  - To understand a “list” of PCBs, xk scans the entire array of PCBs looking for a particular value in the special address field



Process execution states

- Each process has an execution state that indicates what it's currently doing
  - ready: waiting to be assigned a core
    - could run if the OS were to decide it wanted to assign a core
  - running: executing on a core
    - it's the process that currently controls a core
  - blocked (aka "waiting"): waiting for an event
    - e.g., I/O completion, or a message from another process or completion of another process
    - cannot make progress until the event happens (so not eligible to be given a core)
    - Note: blocking is an important construct at the app programming level
      - What's the alternative?
- As a process executes, it moves from state to state
  - UNIX: run `ps`, STAT column shows current state
  - which state is a process in most of the time?

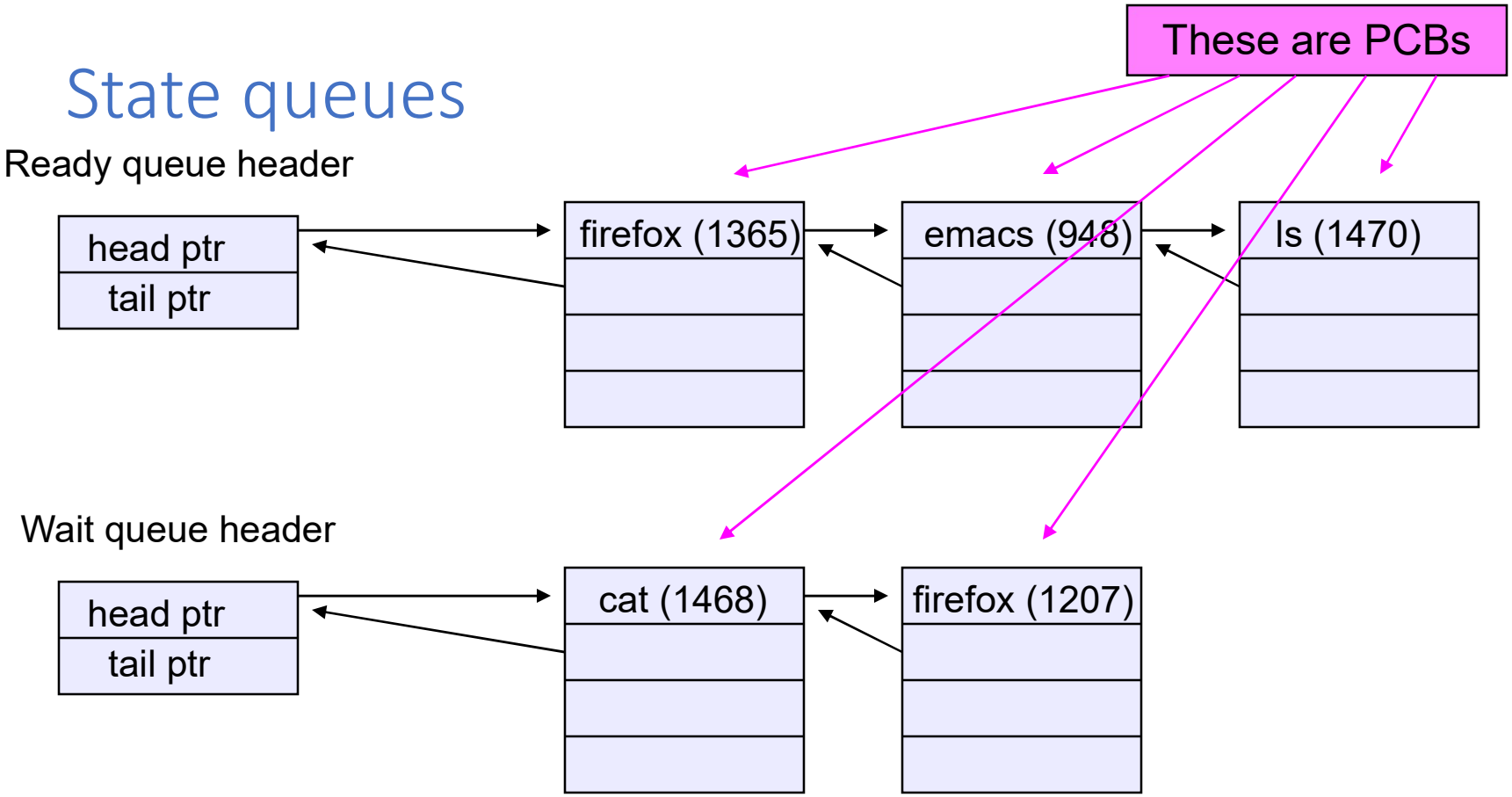
Process states and state transitions



State queues

- The OS maintains a collection of queues that represent the state of all processes in the system
  - each PCB is queued onto a state queue according to the current state of the process it represents
  - as a process changes state, its PCB is unlinked from one queue, and linked onto another

State queues



- There may be many wait queues, one for each type of wait (particular device, timer, message, …)

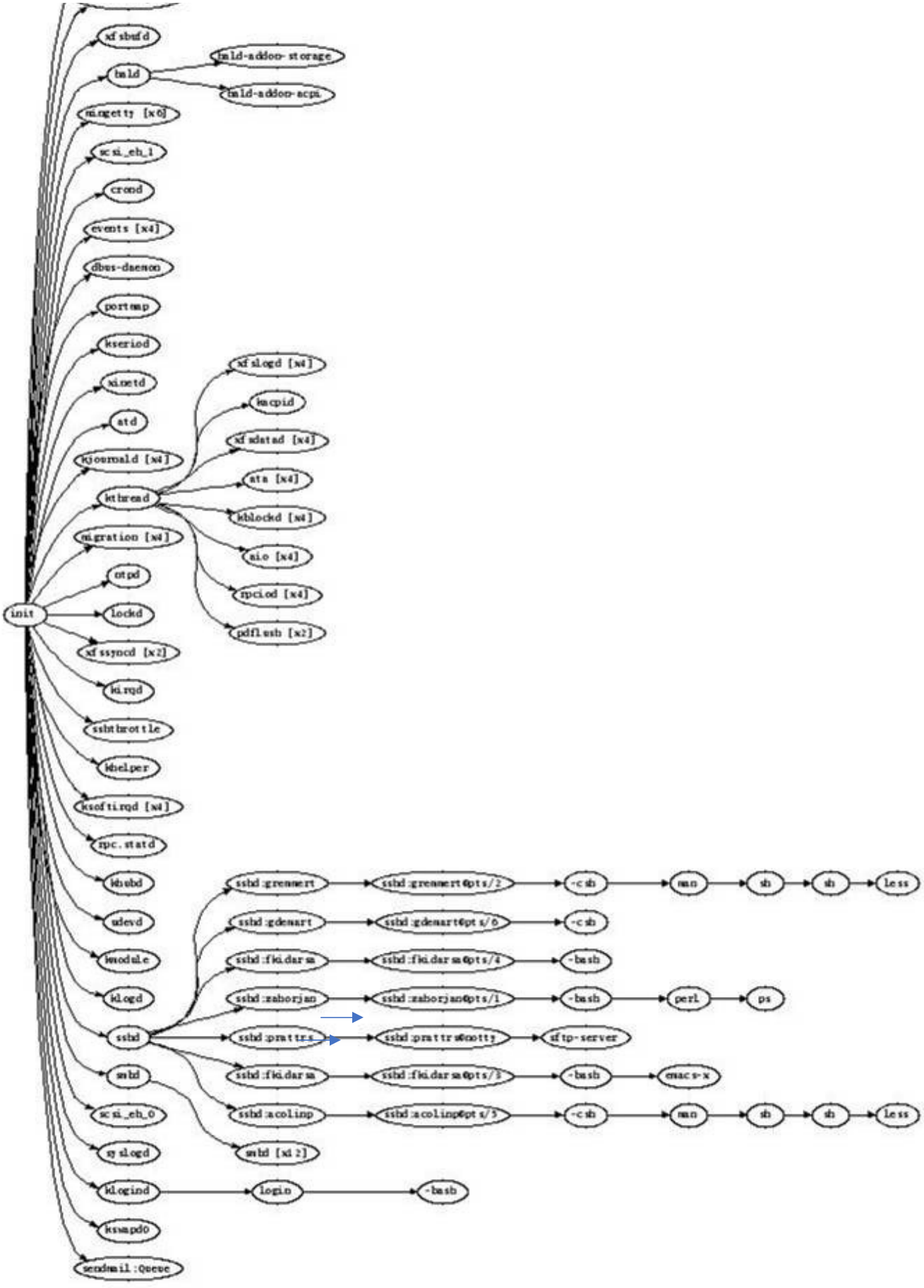
PCBs and state queues

- PCBs are data structures
  - Located in OS memory
- When a process is created:
  - OS allocates a PCB for it
  - OS initializes PCB
  - (OS does other things not related to the PCB)
  - OS puts PCB on the correct queue
- As a process computes:
  - OS moves its PCB from queue to queue
- When a process is terminated:
  - PCB may be retained for a while…
  - eventually, OS deallocates the PCB

Process creation

- New processes are created by existing processes
  - creator is called the parent
  - created process is called the child
    - UNIX: do `ps`, look for PPID field
  - what creates the first process, and when?

- $ ps –ejH
  - prints process tree


xfsbufd
hald
hald-addon-storage
hald-addon-acpi
mingetty [x6]
scsi_eh_1
crond
events [x4]
dbus-daemon
portmap
kseriod
xinetd
atd
kjournald [x4]
kthread
migration [x4]
ntpd
lockd
xfssyncd [x2]
khubd
sshdthrottle
khelper
ksoftirqd [x4]
rpc.statd
khubd
udevd
kmodule
klogd
sshd
smbd
scsi_eh_0
syslogd
klogind
ksmpd0
sendmail:Queue
xfslogd [x4]
kacpid
xfsdatad [x4]
ata [x4]
kblockd [x4]
aio [x4]
rpciod [x4]
pdflush [x2]
sshd:greenert
sshd:greenert@pts/2
-csh
man
sh
sh
less
sshd:gdenart
sshd:gdenart@pts/6
-csh
sshd:fki.darsa
sshd:fki.darsa@pts/4
-bash
sshd:zaborjan
sshd:zaborjan@pts/1
-bash
perl
ps
sshd:prattrs
sshd:prattrs@notty
sftp-server
sshd:fki.darsa
sshd:fki.darsa@pts/3
-bash
emacs-x
sshd:acolinp
sshd:acolinp@pts/5
-csh
man
sh
sh
less
smbd [x2]
klogind
login
-bash
init

Process Creation Semantics

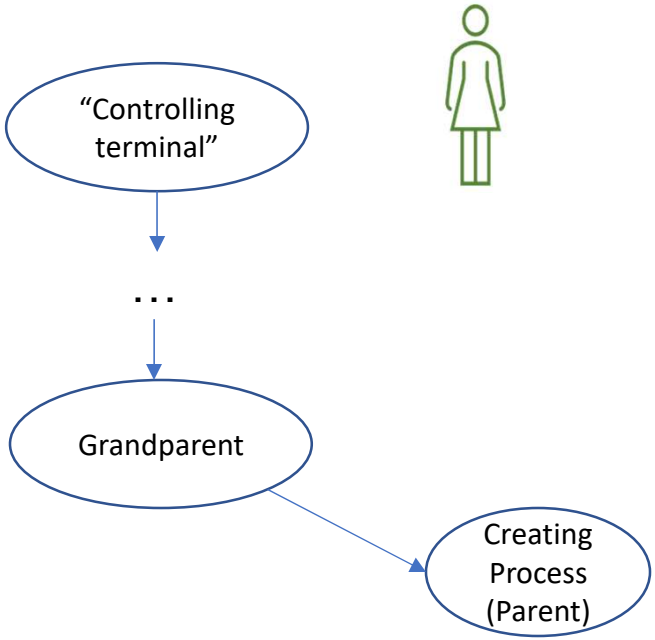
- When we create a process, we have to “communicate” to it what program it should run
  - Name of an executable file
  - *CREATE(* path-to-executable)
- With that information the OS can create and initialize an address space and set the PC to entry point address

- But there’s more to a process than just the address space and PC

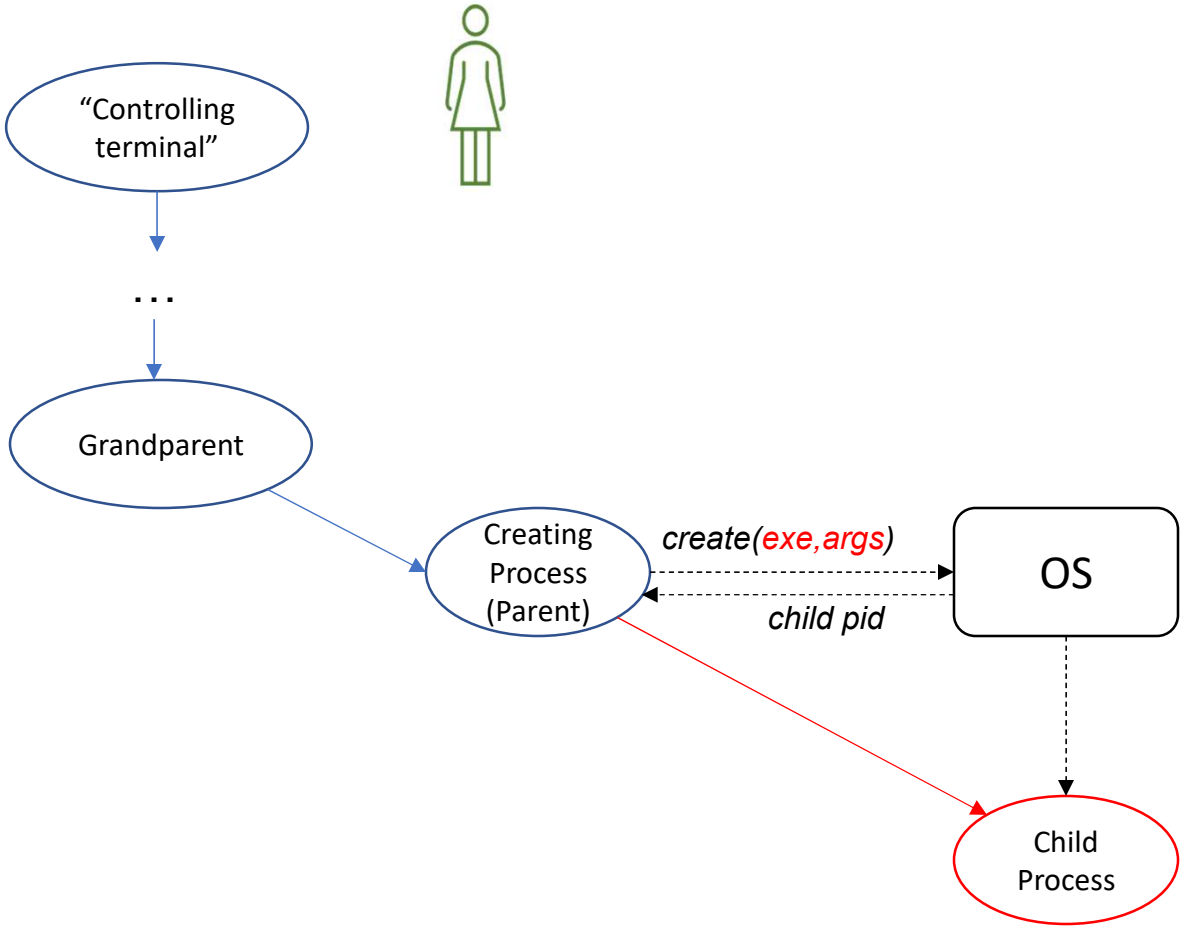
- Examples:
  - If program reads a file, which file should it read?
  - If program writes output, to what device should it write it?
  - As what user should the program run?
  - What “environment” (a map from text key to text value) should it have?
  - …

- Who sets that information?
  - The process, once it starts running?
  - The “parent process” that creates the “child process” ?
  - The parent of the parent process?
  - The user?

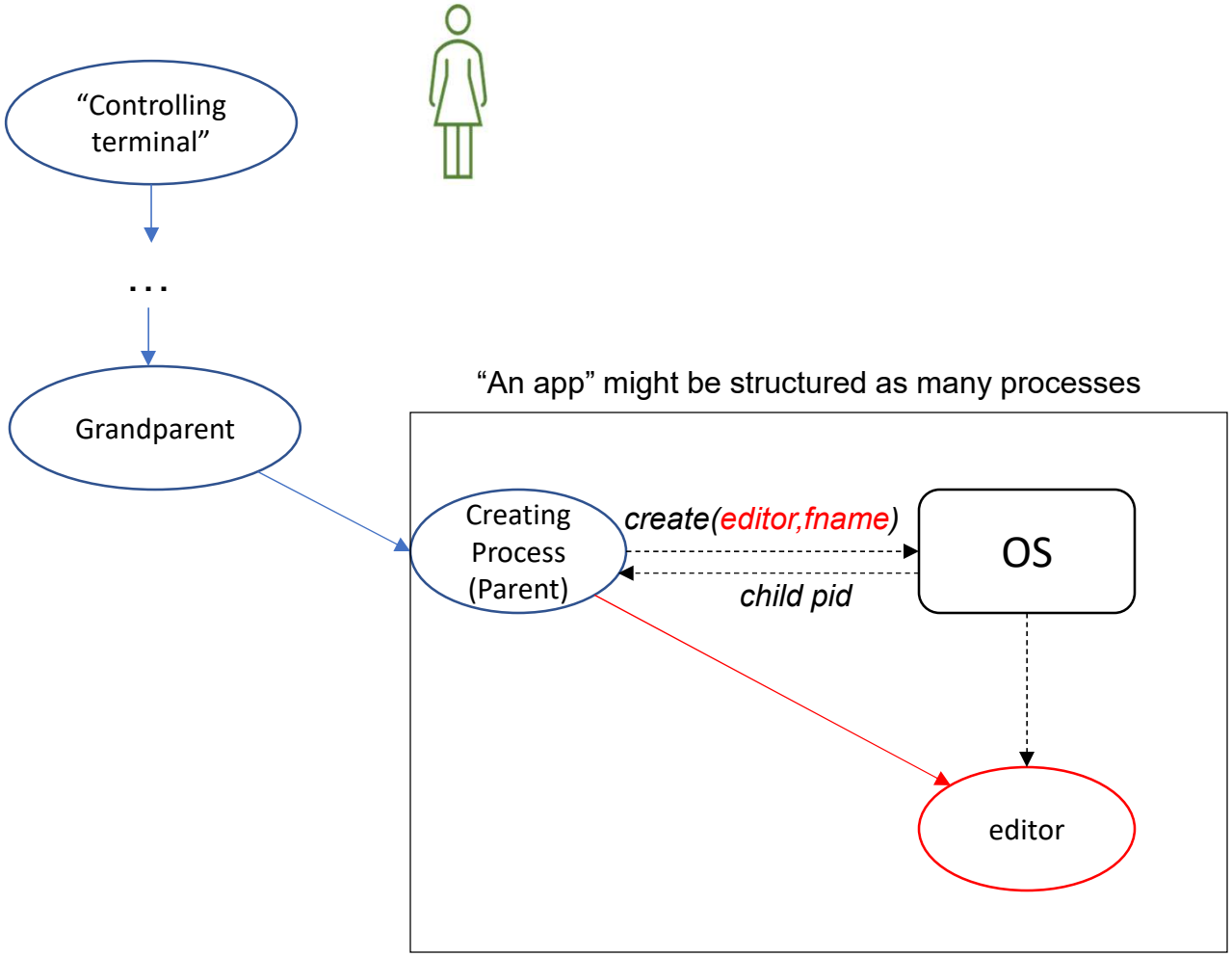
Process Creation: Customization



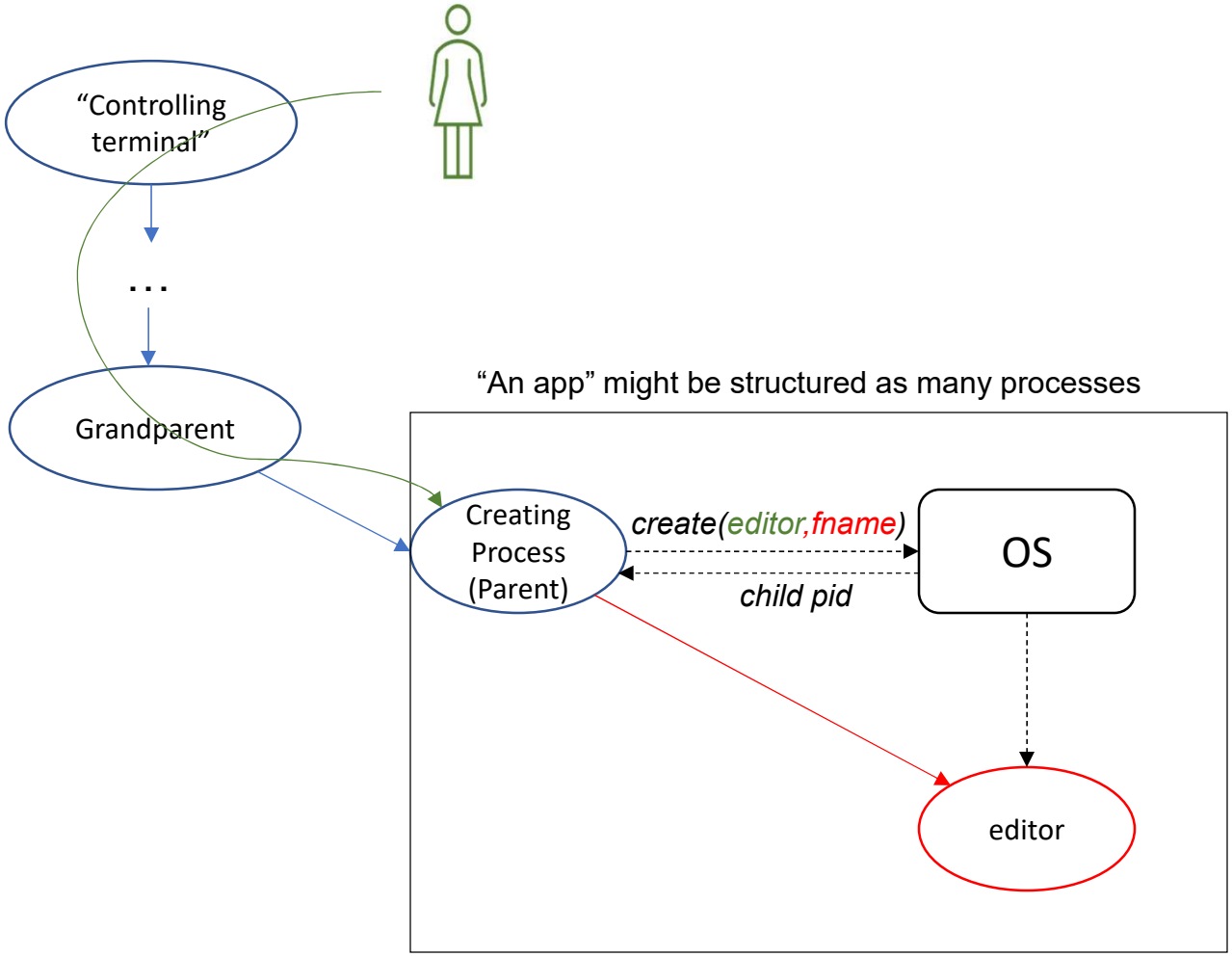
Process Creation: Customization



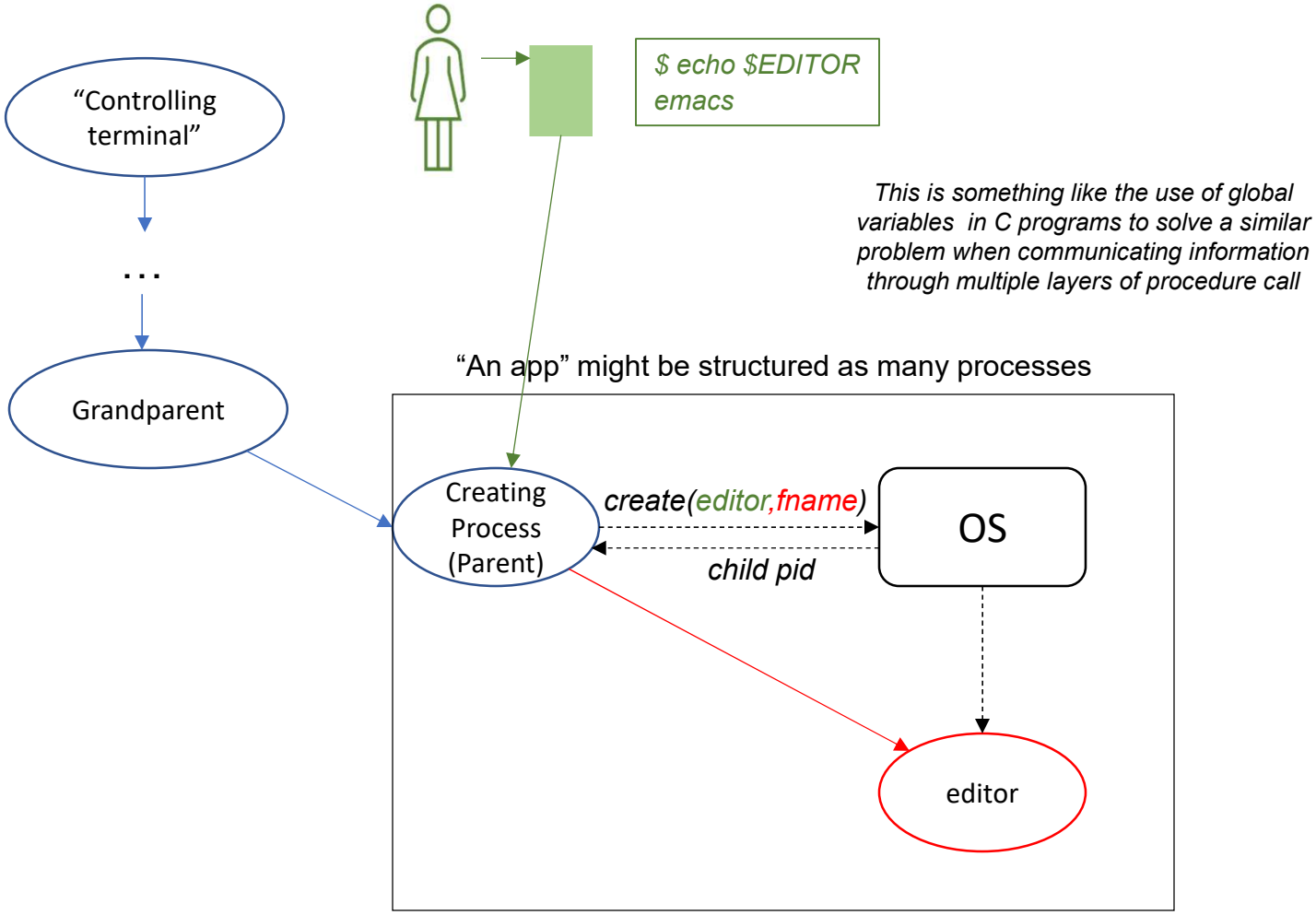
Process Creation: Customization



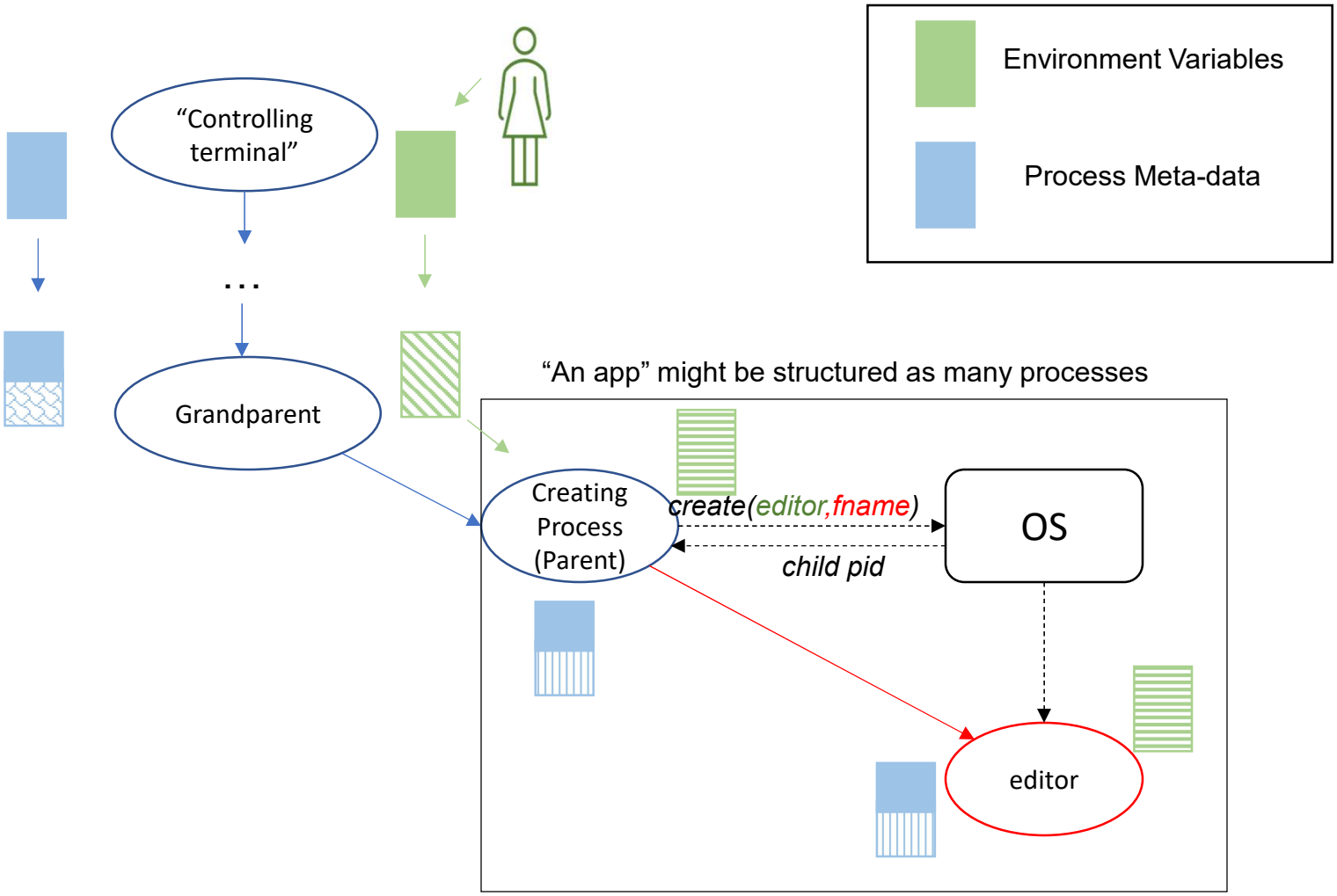
Process Creation: Customization



Process Creation: Customization



Process Creation: Inheritance



Process creation semantics

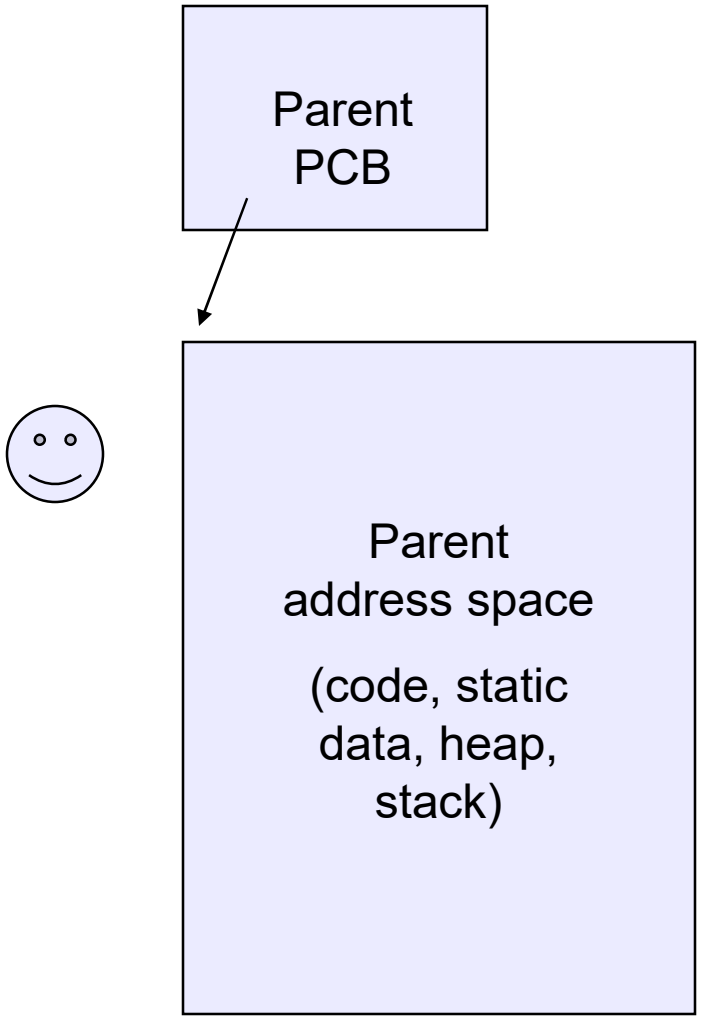
- (Depending on the OS) child processes inherit certain attributes of the parent
  - Examples:
    - Open file table: implies stdin/stdout/stderr
    - On some systems, resource allocation to parent may be divided among children
- (In Unix) when a child is created, the parent may either wait for the child to finish, or continue in parallel
  - This is a “policy” decision
  - Who should make it?
  - How?

UNIX process creation details

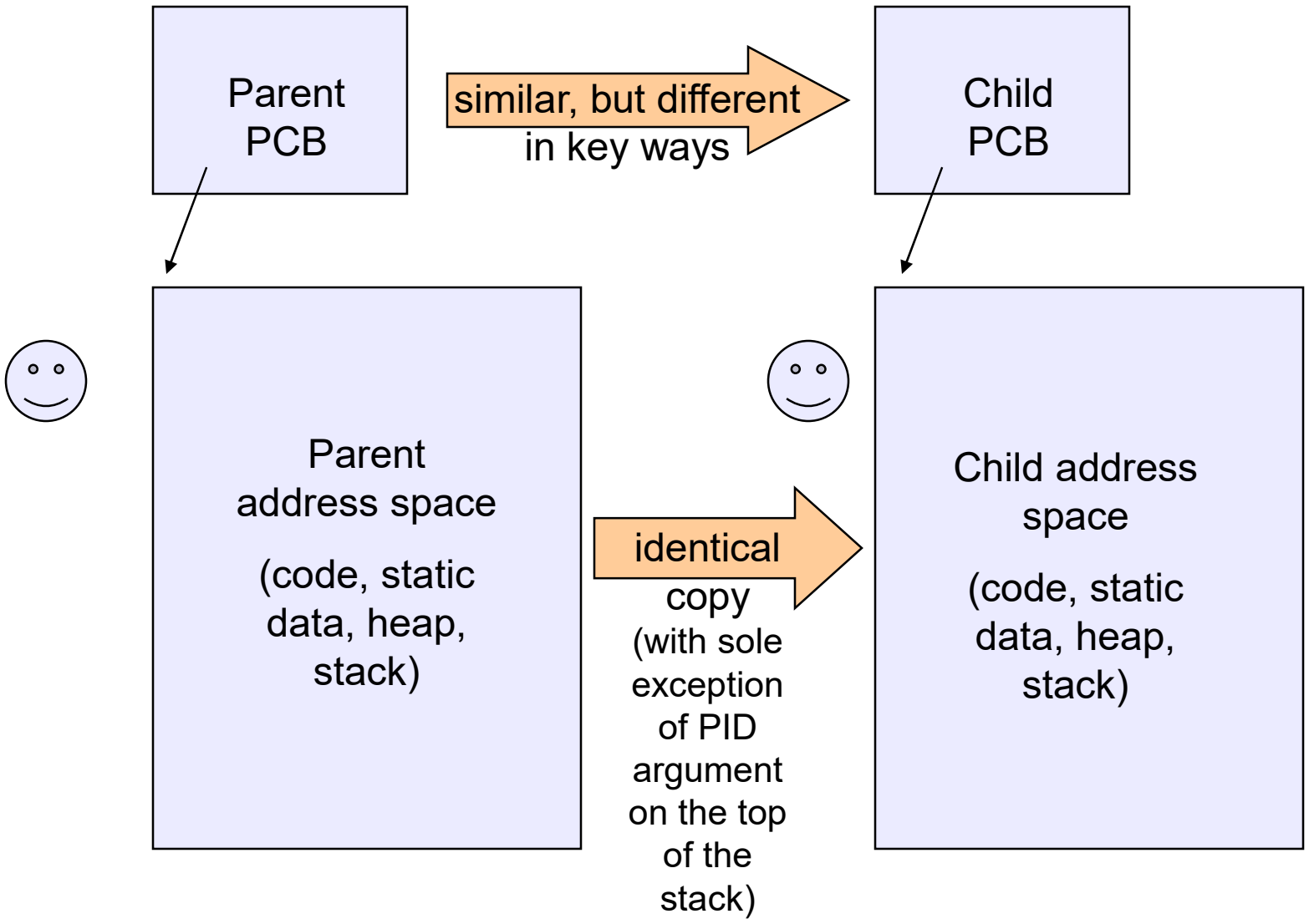
- UNIX process creation through **`fork()`** system call
  - creates and initializes a new PCB
    - initializes kernel resources of new process with resources of parent (e.g., open files)
    - initializes PC, SP to be same as parent
  - creates a new address space
  - initializes new address space with a copy of the entire contents of the parent's address space
  - places new PCB on the ready queue

- the **`fork()`** system call "returns twice"
  - once into the parent, and once into the child
  - value returned from call depends...
    - returns the child's PID to the parent
    - returns 0 to the child

- **`fork()`** = "make a copy of me in my current state"


Parent
PCB
Parent
address space
(code, static
data, heap,
stack)

$ ./myProgram


Parent PCB
similar, but different in key ways
Child PCB
Parent address space
(code, static data, heap, stack)
identical copy
(with sole exception of PID argument on the top of the stack)
Child address space
(code, static data, heap, stack)

$ ./myProgram

testparent – use of fork( )

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

int main(int argc, char **argv)
{
  char *name = argv[0];
  int pid = fork();
  if (pid == 0) {
    printf("Child of %s is %d\n", name, pid);
    return 0;
  } else {
    printf("My child is %d\n", pid);
    return 0;
  }
}
```

testparent output

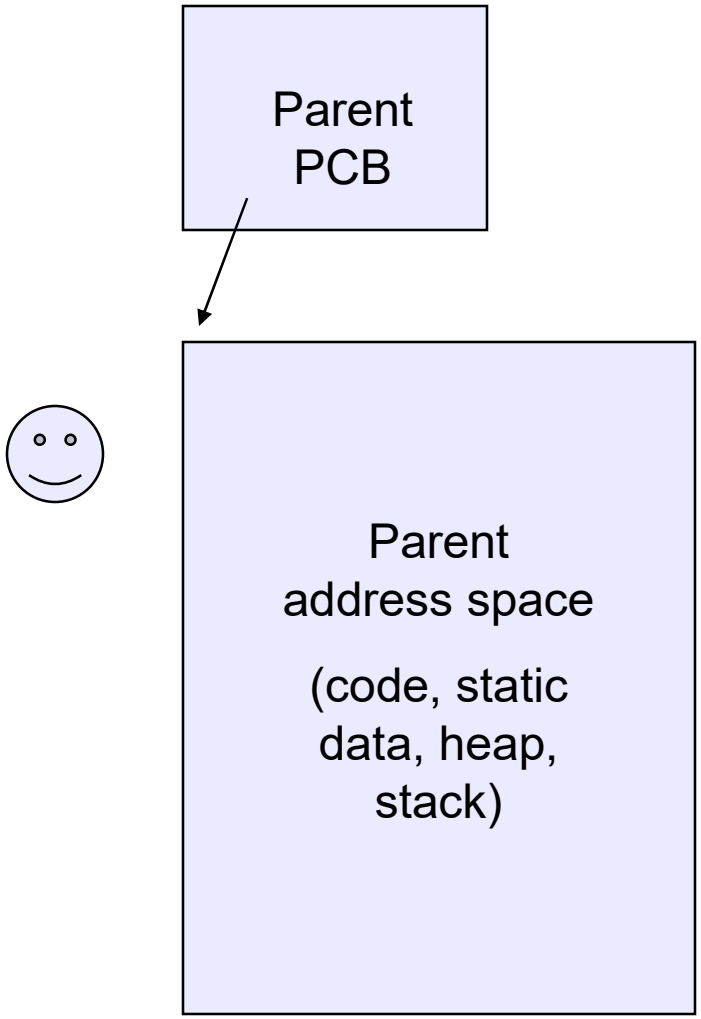
```
spinlock% gcc -o testparent testparent.c
spinlock% ./testparent
My child is 486
Child of testparent is 0
spinlock% ./testparent
Child of testparent is 0
My child is 571
```

fork() then exec()

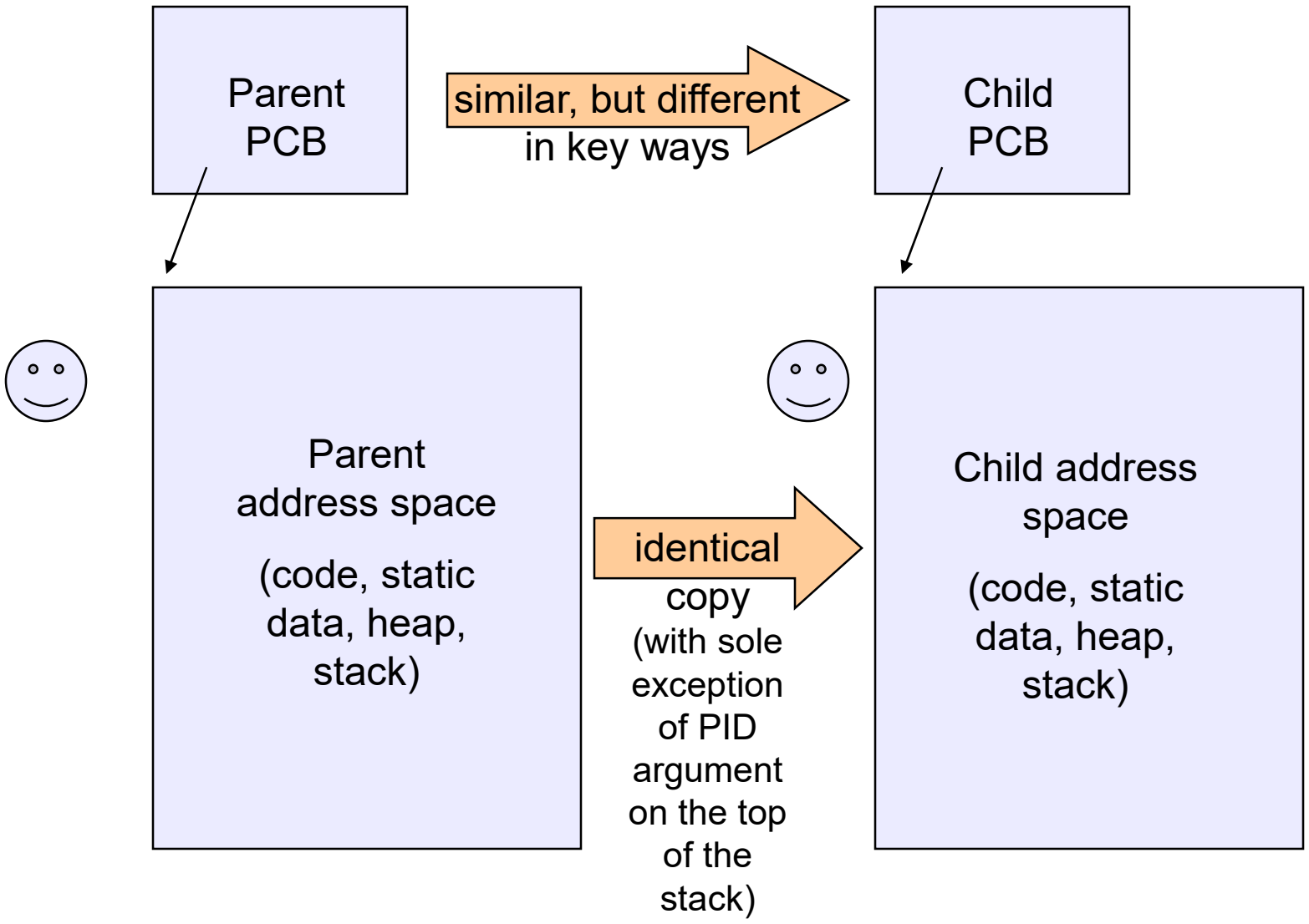
- Q: So how do we start a new program?
- A: First fork, then exec
  - `int exec(char * prog, char * argv[])`

- **`exec()`**
  - note: *exec does not create a new process!*
    - use fork() for that
  - the current process is blocked
    - made non-runnable
    - not on any list (but is "remembered" as part of OS code path functioning)
  - The process address space is replaced by one in which program 'prog' has been loaded
  - initializes hardware context, args for new program
    - sets PC to entry point, sets SP to bottom of empty stack
  - places existing PCB onto ready queue

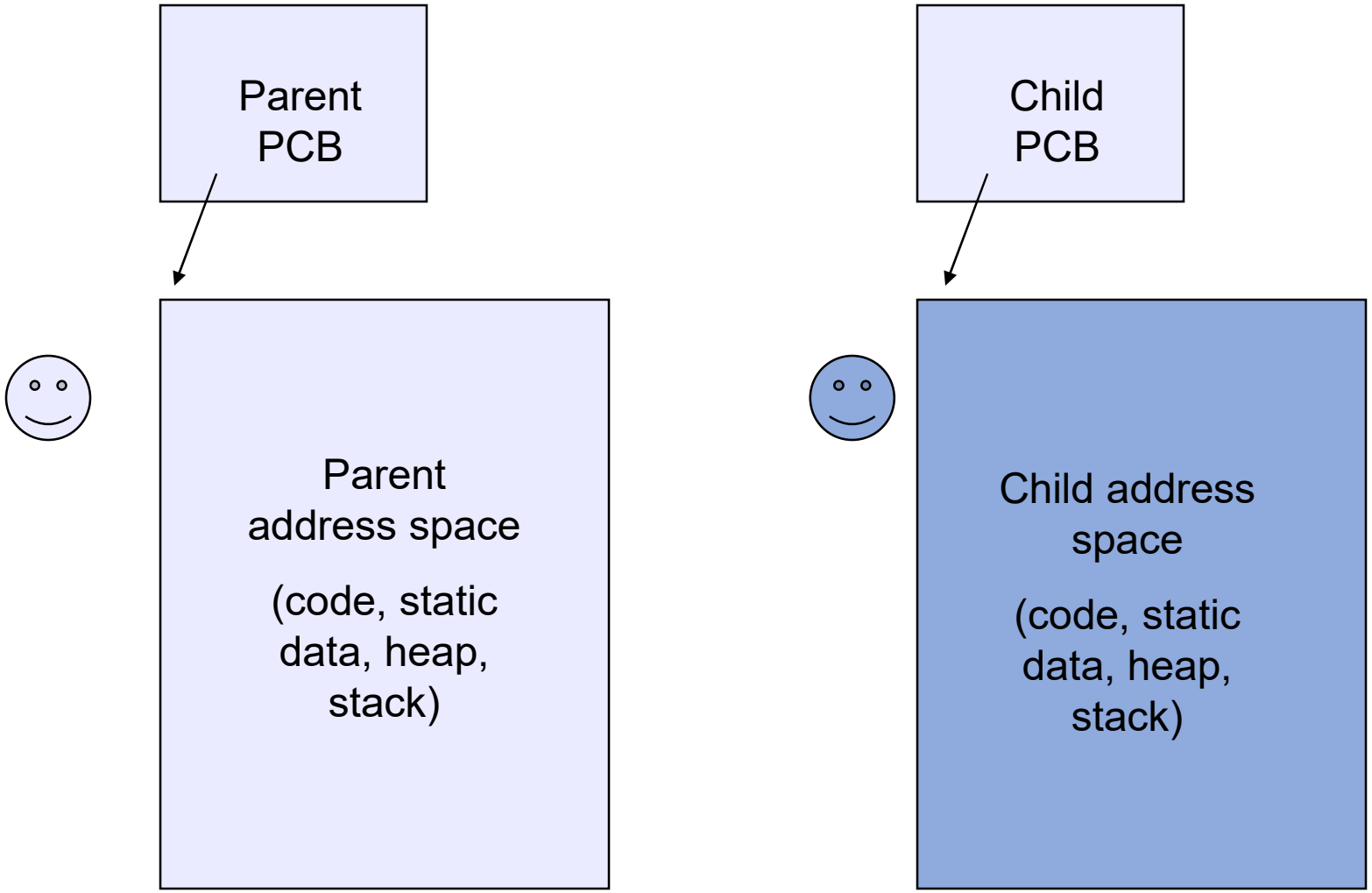
- So, to run a new program:
  - fork()
  - Child process does an exec()
  - Parent either waits for the child to complete, or not


Parent
PCB
Parent
address space
(code, static
data, heap,
stack)

$ ./myProgram


Parent PCB
similar, but different in key ways
Child PCB
Parent address space
(code, static data, heap, stack)
identical copy
(with sole exception of PID argument on the top of the stack)
Child address space
(code, static data, heap, stack)

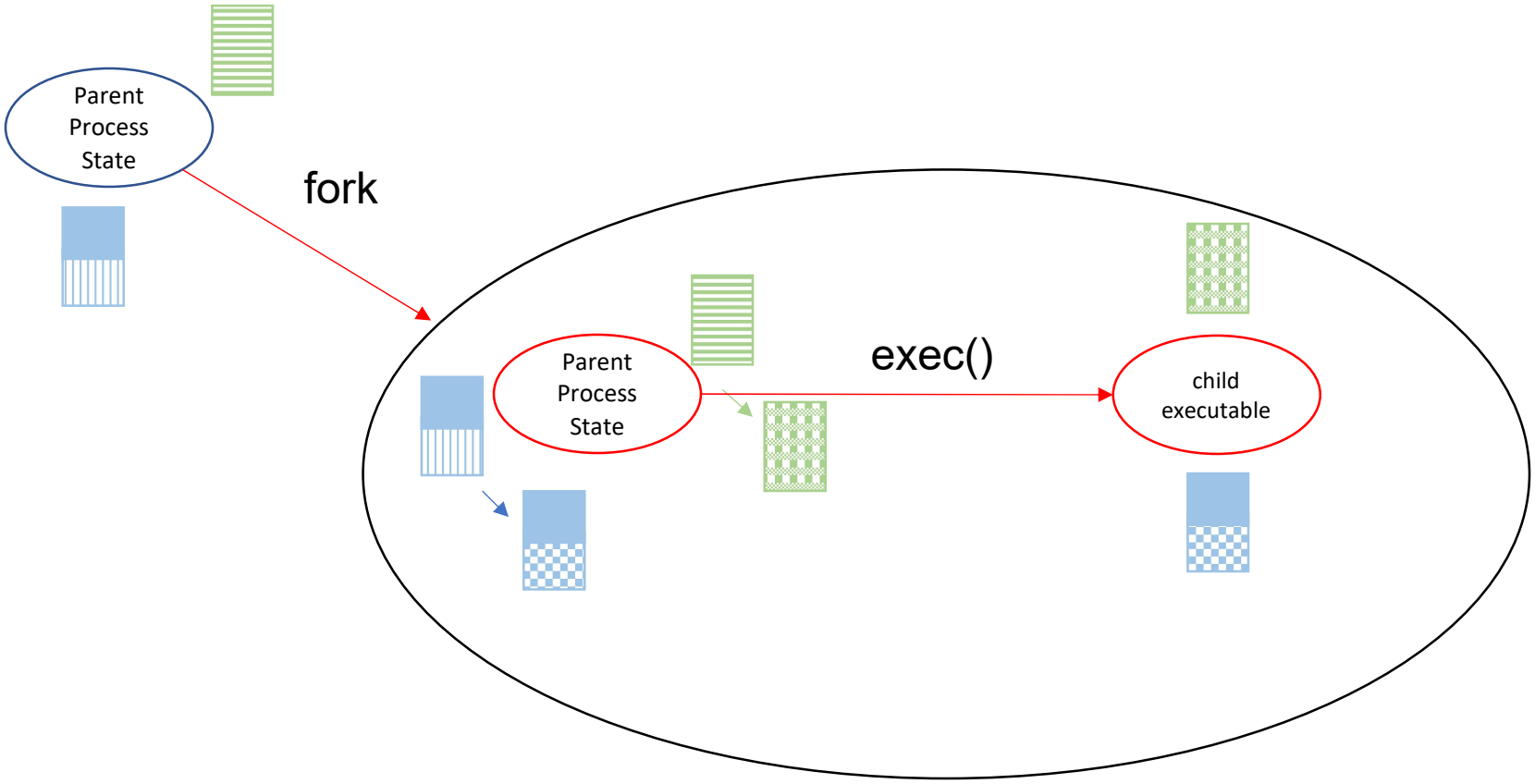
$ ./myProgram



After *exec* of
./myProgram

$ ./myProgram

Why fork/exec?



Making process creation faster

- The semantics of fork() say the child's address space is a copy of the parent's

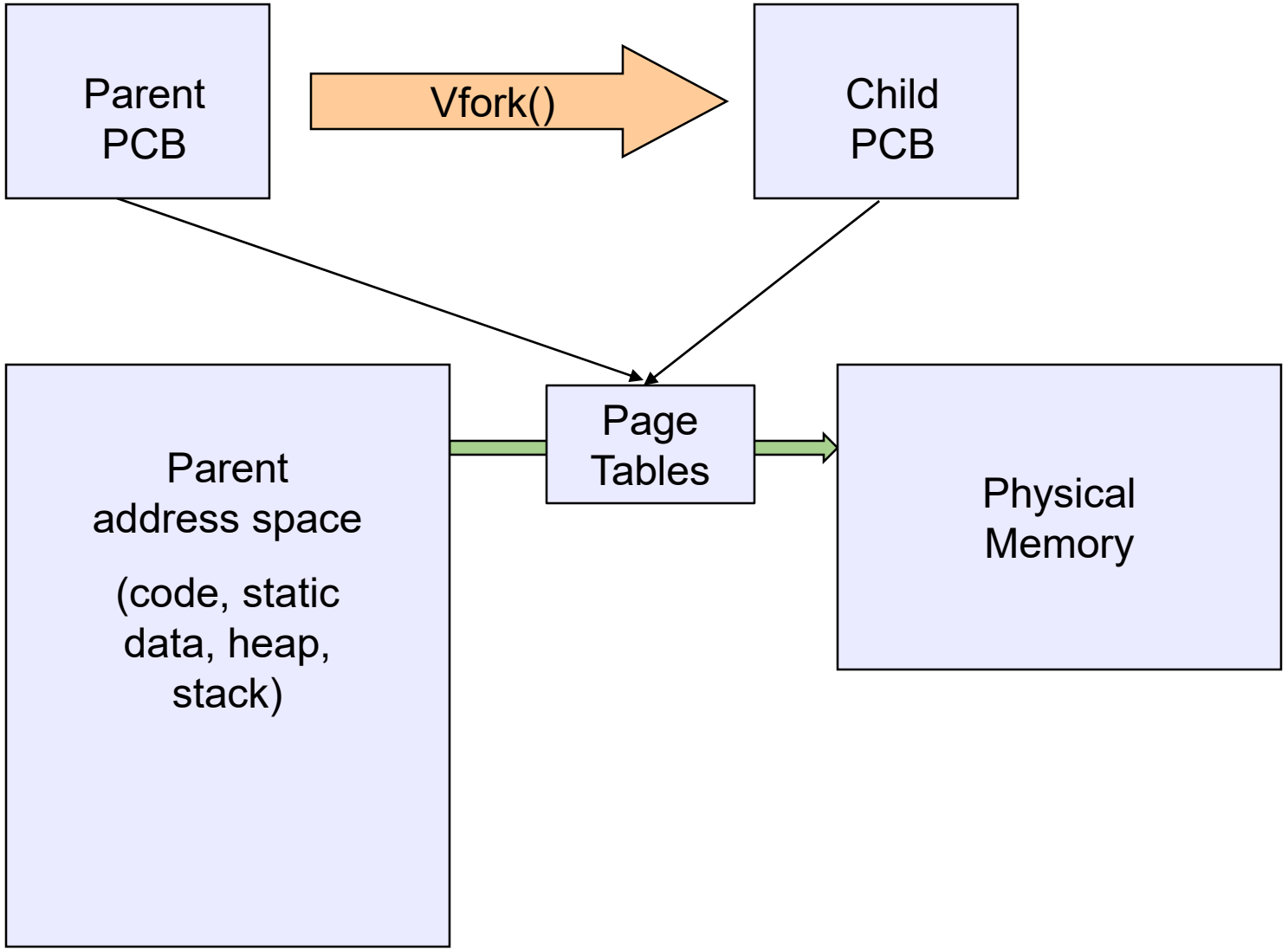
- Implementing fork() that way is slow
  - Have to set up child's page tables to map new address space
  - Have to copy parent's address space contents into child's address space
    - Which you are likely to immediately blow away with an exec()
  - Have to allocate physical memory for the new address space

Method 1: vfork()

- vfork() is the older (now uncommon) of the two approaches we'll discuss

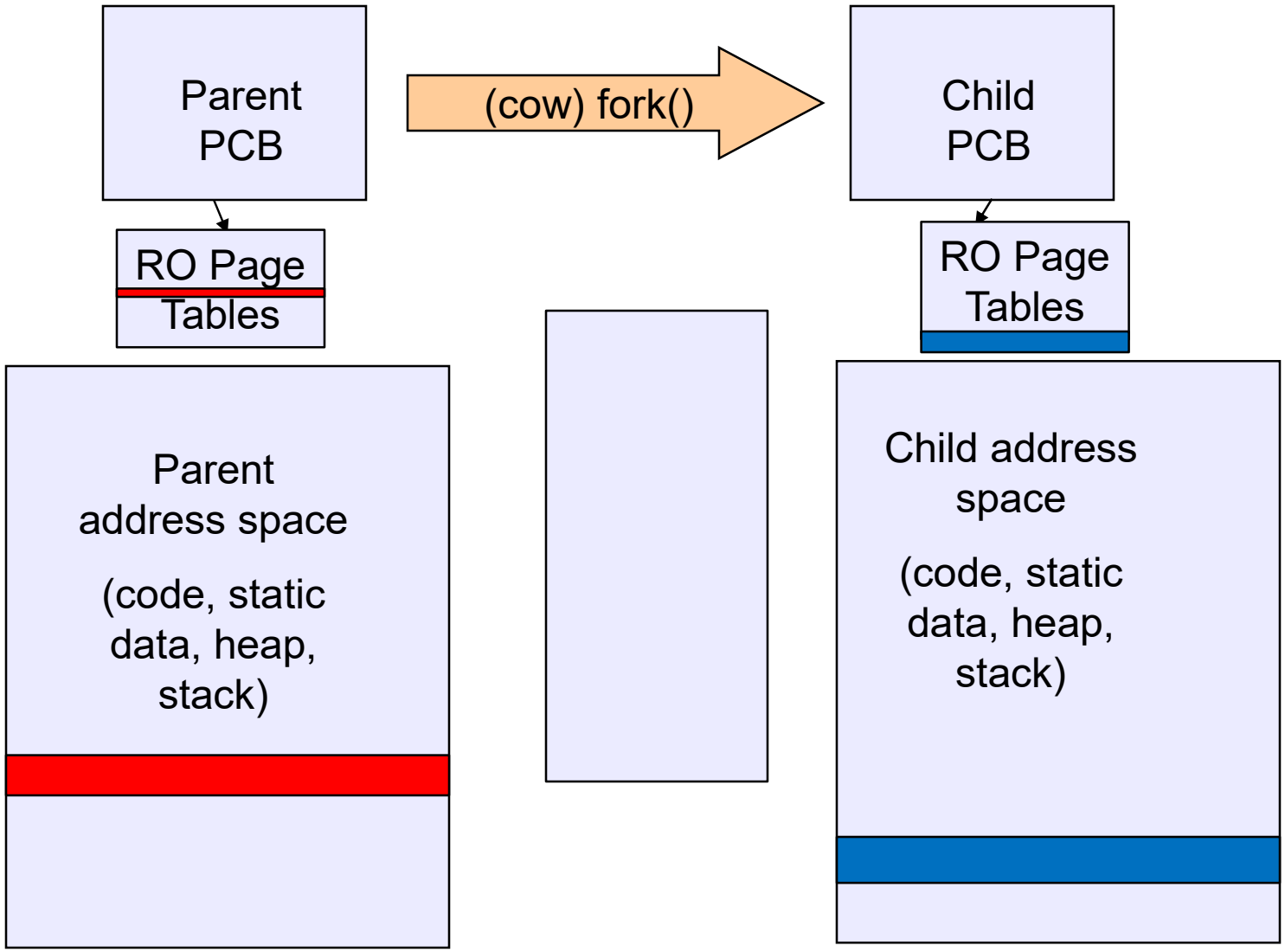
- Instead of "child's address space is a copy of the parent's," the semantics are that the "child's address space *is* the parent's"
  -  There is a terrible race condition!
  - Only sometimes works
  - The sometimes is a pretty common case

- Forking process has to "promise" that its code won't modify the address space in the child before doing before doing an exec()
  - Unenforced! You use vfork() at your own peril
- When exec() is called, a new address space is created and it's loaded with the new executable
  - Saves wasted effort of duplicating parent's address space, just to immediately overwrite it in child
- Parent is blocked until exec() is executed by child


Parent
PCB
Vfork()
Child
PCB
Parent
address space
(code, static
data, heap,
stack)
Page
Tables
Physical
Memory

Method 2: copy-on-write (COW)

- Retains the original semantics, but copies "only what is necessary" rather than the entire address space
- On fork():
  - Create a new address space
  - Initialize new page tables with same mappings as the parent's (i.e., they both point to the same physical memory)
    - No copying of address space contents have occurred at this point – with the sole exception of the top page of the stack
  - Set both parent and child page tables to make all pages read-only
  - If either parent or child writes to memory, an exception occurs
  - When exception occurs, OS copies the page, adjusts page tables, and makes both corresponding pages writable
- It is common for the child process to exec() soon after it is created


Parent
PCB
(cow) fork()
Child
PCB
RO Page
Tables
RO Page
Tables
Parent
address space
(code, static
data, heap,
stack)
Child address
space
(code, static
data, heap,
stack)

Unix Shells

- Shells are just user-level programs
  - What's that mean?
- They're mainly oriented towards launching other programs
  - Using fork() / exec()
- They typically have few "built-in" commands
  - ls, cat, etc. are executables, opaque to the shell
  - (What must be built in?)
- Shells usually offer ways to build "shell scripts"
  - E.g., some looping construct
  - You can view everything you type into a shell as a program that is being simultaneously created and executed
    - This is why programmers have traditional liked shells and users have traditionally liked clicking

UNIX shells – basic operation

```
# this is the main idea, but details are lacking…
int main(int argc, char **argv)
{
  while (1) {
    printf ("$ ");
    char *cmd = get_next_command();
    int pid = fork();
    if (pid == 0) {
       exec(cmd);
       panic("exec failed!");
    } else {
       wait(pid);
    }
  }
}
```

Unix Shells: Jobs / Redirection

- Shells usually offer ways to make “jobs” – assemblages of executions
  - ls | grep *.c | less
  - pushd sub && make && popd
  - pushd sub; make; popd

- One way the shell helps you compose jobs is by input-output redirection
  - You can make the output of one program the input of another, without ever writing to a file

Input/output redirection

- $ ./myprog < input.txt > output.txt
- fork() is key to making this work
  - The input and output files are set by the parent (the shell), not the child (myprog)
- Each process has an open file table
  - by (universal) convention:
    - 0: stdin
    - 1: stdout
    - 2: stderr
    - …
- A child process **inherits** the parent's open file table
- Redirection: the shell code…
  - fork()
  - child (still running shell code) opens input.txt as stdin and output.txt as stdout
  - child exec()'s ./myprog
  - parent calls wait() to suspend execution until child terminates

Example of Deferring Policy

- The basic mechanisms we've seen embed a policy decision:
  - A forked process inherits the user id (uid) of the parent process
- Motivation: security
  - The system "trusts you" to do only the things you're allowed to do, and nothing more
- But there are circumstances where that decision gets in the way
- Example: how do I change my own password?
  - Have to update some privileged data somewhere…
- Sometimes we want to "trust the code" rather than the user
  - Privileges should be associated with the executable, not who is invoking it
  - Example: passwd executable

```
[attu8] ~> ls -l /usr/bin/passwd
-rwsr-xr-x 1 root root 33600 Apr  6  2020 /usr/bin/passwd
```

Linux user id's

- Real uid – launching user
- Effective uid – uid in use right now
- Saved uid – what it was before I changed it

- Set of syscall's to set and get these values
- Set calls fail if you lack permission to change to the uid you're trying to change to

- "setuid executable" – executables can be marked in the file system so that when run their uid is that of the file owner, not the uid of the process that launches them

Example executables…

```
$ ls -l
total 76
-rw-rw-r-- 1 zahorjan zahorjan   236 Oct 19 09:42 makefile
-rwsrwxr-x 1 guest    zahorjan 16960 Oct 19 09:44 suidProgram
-rw-rw-r-- 1 zahorjan zahorjan   488 Oct 19 09:43 suidProgram.c
-rwsrwxr-x 1 root     zahorjan 17096 Oct 19 09:44 suidRootProgram
-rw-rw-r-- 1 zahorjan zahorjan   880 Oct 19 09:44 suidRootProgram.c
-rwxrwxr-x 1 zahorjan zahorjan 17176 Oct 19 09:44 uidProgram
-rw-rw-r-- 1 zahorjan zahorjan   822 Oct 19 09:44 uidProgram.c
```

Example program

- uidprogram.c
  - Prints its real, effective, and saved uids
  - Forks then execs suidProgram
  - Forks then execs suidRootProgram
- suidProgram.c
  - Executable file is owned by account “guest” and is marked setuid
  - Prints its real, effective, and saved uid’s
- suidRootProgram.c
  - Executable file is owned by root and is marked setuid
  - Prints uids on launch
  - Sets euid to user (drops root)
  - Sets euid back to root (recovers root – allowed because saved uid is root)
  - Permanently drops root (by calling setuid with user uid)

Example program output:

```
uidProgram:
real  uid = 1000
effec uid = 1000
saved uid = 1000

suidProgram:
real  uid = 1000
effec uid = 1004
saved uid = 1004

suidRootProgram:
uids as launched:
real  uid = 1000
effec uid = 0
saved uid = 0
uids after switching to user:
real  uid = 1000
effec uid = 1000
saved uid = 0
uids after switching back to root:
real  uid = 1000
effec uid = 0
saved uid = 0
uids after permanently dropping root:
real  uid = 1000
effec uid = 1000
saved uid = 1000
```

Example Program Source

- You can definitely view and compile the source
- You won't be able to set things up to run as intended unless you have (basically) root permission, though
  - You have to chown (change owner) some files and chmod (to mark as setuid) some executables

- https://courses.cs.washington.edu/courses/cse451/21sp/csenetid/uidPrograms/

Linux Job Control

- A “job” is an assemblage of processes
  - $ cat main.c
  - $ cat main.c | grep –w total | less

- Key concepts:
  - Controlling terminal
    - Follow parent PIDs up to “the top”
    - What is that process’s stdin/stdout/stderr connected to?
    - Why does it matter?

  - Session
    - A way to group things that should be terminated if the controlling terminal goes away
    - “Session leader” – process that created the session
    - Sessions are named by integers
      - Must be unique
      - Use PID of the session leader
    - A forked process inherits the session of its parent
    - A process can set its own session id (setsid)
      - Unless it’s a “process group leader”

Linux Job Control

- Process Group
  - $ myprog | myotherprog | grep B
  - A process inherits the process group of its parent
  - A process can set its process group (setpgid)
  - The “process group leader” is the process that created the group
  - The process group’s name is an integer
    - The PID of the creating process
- Why have process groups?
  - Ctrl-C sends a SIGINT
  - A signal can be sent to a process group
    - Sent to each process in the group

Inter-Process Communication (IPC)

- Processes provide isolation (protection) – great!
- But sometimes you want processes to communicate / cooperate
- Inter-process communication (IPC)
- Simple example: How can one process “provide input” to another?

What If Processes Want to Cooperate

- How can one process “provide input” to another?

1. Send command line arguments (argv values)
   - available only to parent process

2. Communicate through files
   - one writes and the other reads
     - synchronization?

3. Optimize that: pipes
   - use memory buffers, not files
   - we’ll see that this works only if the processes are related (usually siblings)

4. Environment variables
   - Why?

IPC (cont).

- Additional mechanisms:
  5. Named pipes
     - like pipes, except that unrelated processes can use them
       - need a namespace
         - use file system names
     - man 3 mkfifo

  6. Named shared memory regions
     - shm_open() followed by mmap()
     - “cut out the middle man”

  7. sockets / Internet protocols
     - robust – prepared to communicate using a heavyweight middle man!
     - optimized when endpoints are on the same machine

IPC: “exception handling”

- Processes can register event handlers
  - Feels a lot like event handlers in Java, which ..
  - Feel sort of like catch blocks in Java programs
  - sigaction()
- When the event occurs, OS causes process to jump to event handler routine
- This is very similar to the exception mechanism used by the OS
  - For the OS, the hardware is the agent that causes the asynchronous jump to the OS’s event handler
  - For the application, the OS is the agent
- Policy/mechanism separation
  - event detection by the OS
  - lets the application do something in response other than the default built into the OS
    - Why is there a default?

IPC: signals

- The “exception mechanism” at this level is usually called “signal handling”
- A “signal is delivered to the application” when an event occurs
- Events (signals) can be generated by code, including code in other processes
- So, signals are an elementary form of IPC
  - signal can be generated by another process
    - send signal using kill (man 2 kill)
  - only argument of the communication is a single int, the signal number

Signals

Signal	Value	Action	Comment
SIGHUP	1	Term	Hangup detected on controlling terminal or death of controlling process
SIGINT	2	Term	Interrupt from keyboard
SIGQUIT	3	Core	Quit from keyboard
SIGILL	4	Core	Illegal Instruction
SIGABRT	6	Core	Abort signal from abort(3)
SIGFPE	8	Core	Floating point exception
SIGKILL	9	Term	Kill signal
SIGSEGV	11	Core	Invalid memory reference
SIGPIPE	13	Term	Broken pipe: write to pipe with no read
SIGALRM	14	Term	Timer signal from alarm(2)
SIGTERM	15	Term	Termination signal
SIGUSR1	30,10,16	Term	User-defined signal 1
SIGUSR2	31,12,17	Term	User-defined signal 2
SIGCHLD	20,17,18	Ign	Child stopped or terminated
SIGCONT	19,18,25		Continue if stopped
SIGSTOP	17,19,23	Stop	Stop process
SIGTSTP	18,20,24	Stop	Stop typed at tty
SIGTTIN	21,21,26	Stop	tty input for background process
SIGTTOU	22,22,27	Stop	tty output for background process

Example use

- You're implementing Apache, a web server
- Apache reads a configuration file when it is launched
  - Controls things like what the root directory of the web files is, what permissions there are on pieces of it, etc.
- Suppose you want to change the configuration while Apache is running
  - If you restart the currently running Apache, you drop some unknown number of user connections
- Solution: send the running Apache process a signal
  - It has registered an signal handler that gracefully re-reads the configuration file

Another Use

- How should ctrl-C be implemented?
- Option 1:
  - OS notices ctrl-C was typed
  - OS determines what process has keyboard focus
  - OS immediately terminates that process
- What can go wrong?
- Option 2:
  - OS notices ctrl-C was typed
  - OS determines what process has keyboard focus
  - OS sends it a SIGINT signal
  - Default behavior is termination, but…
  - An application that needs/wants to can register a handler and do whatever it needs to do (to terminate gracefully)