

Operating Systems Notes

Anthony Catterwell

April 17, 2019

Contents

1	Introduction	3
2	Operating System Structure	3
2.1	Architectural impact	3
2.2	User operating interaction	3
2.2.1	User v.s. kernel	3
2.2.2	Syscall	4
2.3	Operating System structure	5
2.3.1	Layers	5
2.3.2	Examples	7
2.4	Summary	10
3	Processes	11
3.1	Process	11
3.2	Process control block	12
3.3	Process state & context switch	13
3.4	Process creation and termination	15
3.5	Summary	18
4	Threads	19
4.1	Process vs Threads	19
4.2	Concurrency	20
4.3	Design space of process/threads	22
4.4	Kernel threads	22
4.5	User-level threads	23
4.6	Summary	25
5	Synchronisation	27
5.1	Locks	28
5.2	Spinlocks	29
6	Semaphores, Condition Variables, and Monitors	31
6.1	Semaphore	31
6.2	Monitors	33
7	Deadlock	37
7.1	Graph reduction	37
7.2	Banker's Algorithm	39
8	Scheduling	43
8.1	Scheduling Goals	44
8.2	Laws and properties	44

8.3 Algorithms	45
8.4 Summary	47
9 Memory Management	48
9.1 Background	48
9.2 Logical/Virtual address space v.s. Physical address space	48
9.3 Swapping	49
9.4 Contiguous Memory Allocation	50
9.5 Segmentation	52
9.6 Summary	54
10 Paging	55
10.1 Paging	55
10.2 Page Tables	55
10.3 TLB	56
10.4 Shared Pages	57
10.5 Hierarchical Pages	58
10.6 Hashed Pages	60
10.7 Inverted Pages	60
10.8 Uses	61
11 Virtual Memory	63
11.1 Virtual memory	63
11.2 Page Fault	63
11.3 Demand Paging	64
11.4 Page replacement	65
11.5 Page replacement algorithms	66
11.6 Working set	67
11.7 Thrashing	68
12 File Systems	69
12.1 Roles	69
12.2 Files	69
12.3 Directories	71
12.4 File Protection	72
12.5 Unix	73
13 Virtualisation	78
14 Secondary Storage	83
14.1 Disk trends	83
14.2 Memory hierarchy	84
14.3 Performance	85
14.4 Scheduling	86
14.5 SSDs	87
14.5.1 Read	88
14.5.2 Write	88
14.5.3 Cost	89

1 Introduction

2 Operating System Structure

2.1 Architectural impact

Architectural features affecting OSs

- These features were built primarily to support OSs:
 - timer (clock) operations
 - synchronisation instructions
 - memory protection
 - I/O control operations
 - interrupts and exceptions
 - protected modes of operation (kernel vs. user mode)
 - privileged instructions
 - system calls (including software interrupts)
 - virtualisation architectures
- ASPLOS

2.2 User operating interaction

2.2.1 User v.s. kernel

Privileged instructions

- Some instructions are restricted to the OS
 - known as *privileged* instructions
- Only the OS can:
 - directly access I/O devices
 - manipulate memory state management (page table pointers, TLB loads, etc.)
 - manipulate special *mode bits* (interrupt priority level)
- Restrictions provide safety and security

OS protections

- So how does the process know if a privileged instruction should be executed?
 - the architecture must support at least two modes of operation: kernel mode, and user mode
 - mode is set by status bit in a protected processor register.
 - * user programs execute in user mode
 - * OS executes in kernel (privileged) mode (OS == kernel)
 - Privileged instructions can only be executed in kernel (privileged) mode
 - * if code running in user mode attempts to execute a privileged instruction, the illegal execution trap.

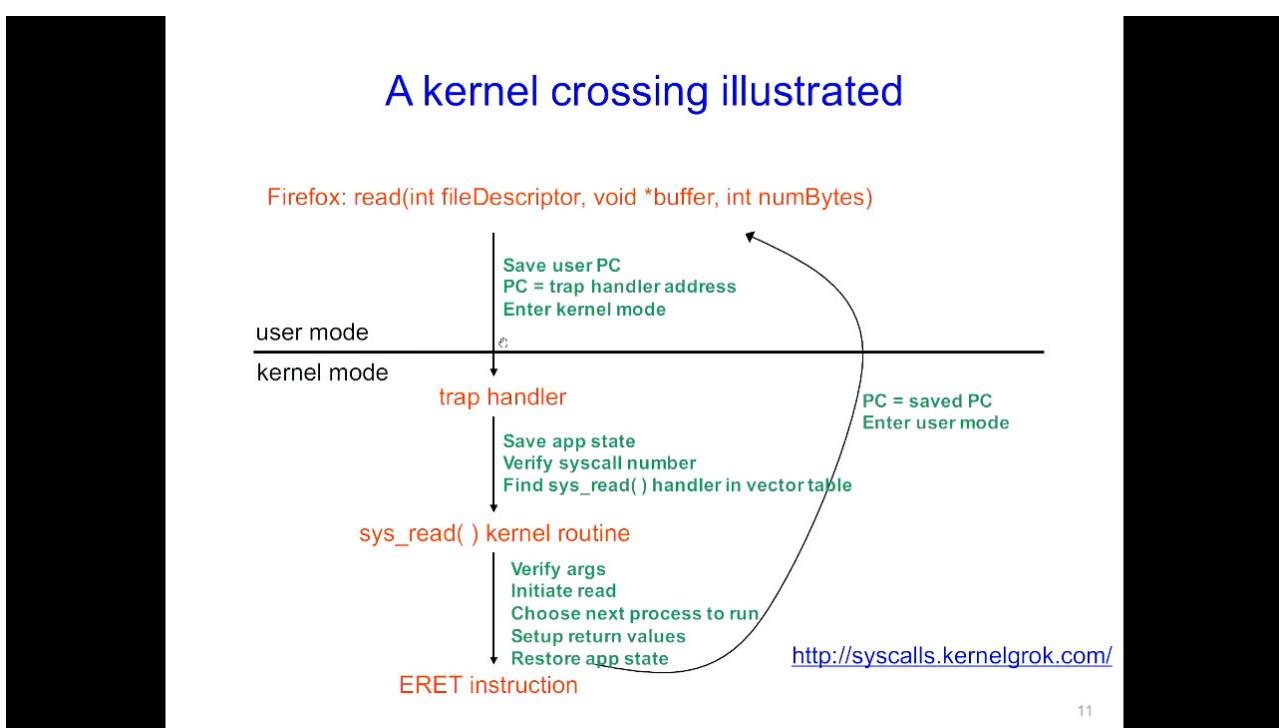
Crossing protection boundaries

- So how do user programs do something privileged?
 - e.g. how can you write to a disk if you can't execute any I/O instructions?
- User programs must call on OS procedure — that is to ask the OS to do it for them.
 - OS defines a set of system calls
 - User-mode program executes system call instruction
- Syscall instruction
 - like a protected procedure call

2.2.2 Syscall

Syscall

- The syscall instruction *atomically*:
 - saves the current PC
 - sets the execution mode to privileged
 - sets the PC to a handler address
- Similar to a procedure call
 - Caller puts arguments in a place the callee expects (registers, or stack)
 - * One of the args is a syscall number, indicating which OS function to invoke
 - Callee (OS) saves caller's state (registers, other control states) so it can use the CPU
 - OS function code runs
 - * OS must verify caller's arguments (e.g. pointers)
 - OS returns using a special instruction
 - * Automatically sets PC to return address and sets execution mode to user.



- A syscall is not a subroutine call, with the caller specifying the next PC.
 - the caller knows where the subroutines are located in memory; therefore they can be the target of an attack.
- The kernel saves state?
 - Prevents overwriting of values
- The kernel verify arguments
 - Prevents buggy code crashing the system
- Referring to kernel objects as arguments
 - Data copied between user buffer and kernel buffer.

Exception handling and protection

- All entries to the OS occur via the mechanism just shown
 - Acquiring privileged mode and branching to the trap handler are inseparable
- Terminology
 - *Interrupt*: asynchronous; caused by an external device
 - *Exception*: synchronous; unexpected problem with instruction
 - *Trap*: synchronous; intended transition to OS due to an instruction

In all three cases, they are instances of where something strange happens, and the OS takes control: whether by accident, or by intention.
- Privileged instructions and resources are the basis for most everything: memory protection, protected I/O, limiting user resource consumption.

2.3 Operating System structure

2.3.1 Layers

Operating System structure

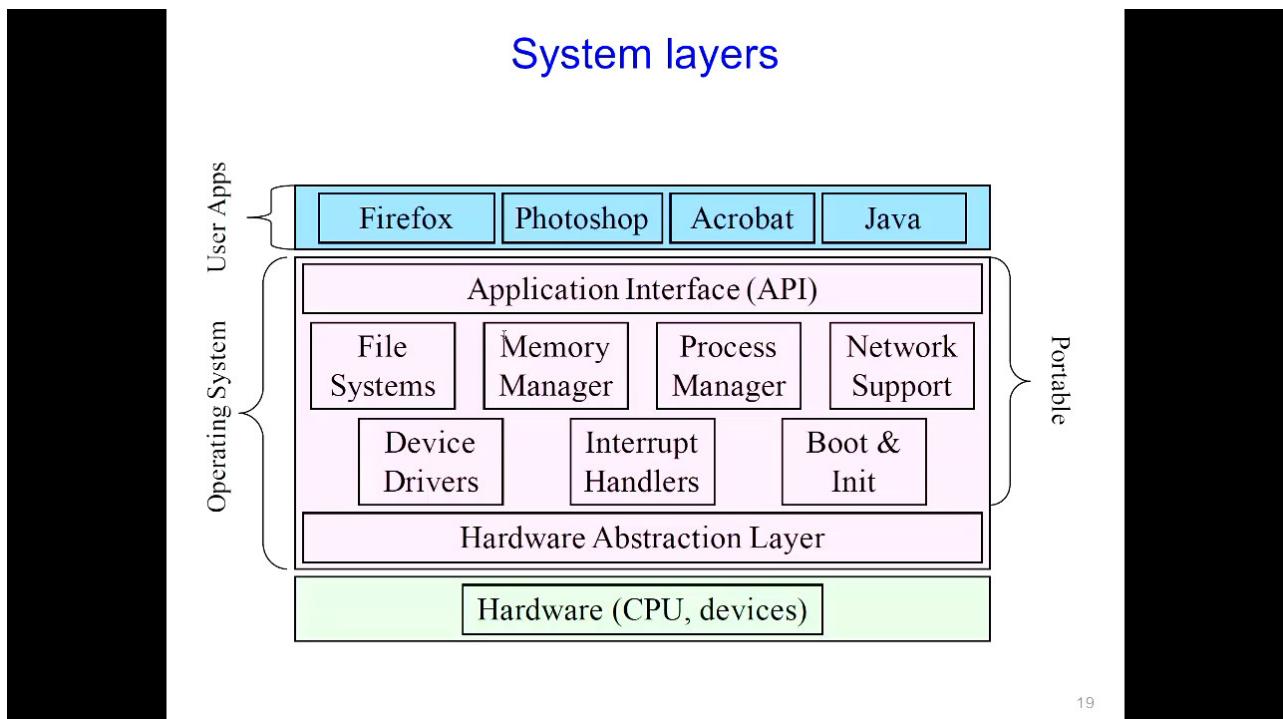
- The OS sits between application programs and the hardware
 - it mediates access and abstracts away ugliness
 - programs request services via traps or exceptions
 - devices request attention via interrupts

Operating system design and implementation

- Design and implementation of OS not “solvable”, but some approaches have proven successful.
- Internal structure of different OSs can vary widely.
- Start the design by defining goals and specifications.
- Affected by choice of hardware, type of system.
- *User* goals, and *system* goals
 - User goals: OS should be convenient to use, easy to learn, reliable, safe, and fast
 - System goals: OS should be easy to design, implement, and maintain, as well as flexible, reliable, error-free, and efficient.
- Important principle to separate

- **Policy:** *What* will be done?
- **Mechanism:** *How* to do it?
- Mechanisms determine how to do something, policies decide what will be done.
- The separation of policy from mechanism is a very important principle, it allows maximum flexibility if policy decisions are to be changed later (e.g. timer).
- Specifying and designing an OS is a highly creative task of *software engineering*.

System layers



19

Major OS components

- processes
- memory
- I/O
- secondary storage
- file systems
- protection
- shells
- GUI
- networking

OS structure

- There's no clear hierarchy within an OS — each of them needs access to different things.
- An OS consists of all these components, plus:
 - many other components
 - system programs (privileged, and non-privileged)

- Major issue:
 - how do we organize all this?
 - what are all of the code modules, and where do they exist?
 - how do they cooperate?
- Massive software engineering and design problem
 - design a large, complex program that: performs well, is reliable, is extensible, and is backwards compatible.

2.3.2 Examples

Monolithic design

- Traditionally, OSs (like Unix) were built as a *monolithic* entity User programs — OS (everything)
 - hardware
 - Major advantage: cost of module interactions is low (procedure call)
 - Disadvantages:
 - hard to understand
 - hard to modify
 - unreliable (no isolation between system modules)
 - hard to maintain
 - What is the alternative?
- Find a way to organise the OS in order to simplify its design and implementation.

Layering

- The traditional approach is layering
 - implement OS as a set of layers
 - each layer presents an enhanced *virtual machine* to the layer above
- The first description of this approach was Dijkstra's THE system
 - Layer 5: *Job managers* execute users' programs
 - Layer 4: *Device managers* handle devices and provide buffering
 - Layer 3: *Console manager* implements virtual consoles
 - Layer 2: *Page manager* implements virtual memories for each process
 - Layer 1: *Kernel* implements a virtual processor for each process
 - Layer 0: *Hardware*
- Each layer can be tested and verified independently
- Imposes a hierarchical stricture
 - but real systems are more complex: file systems require VM services (buffer); VM would like to use files for its backing store
 - strict layering isn't flexible enough
- Poor performance: each layer crossing has *overhead* associated with it

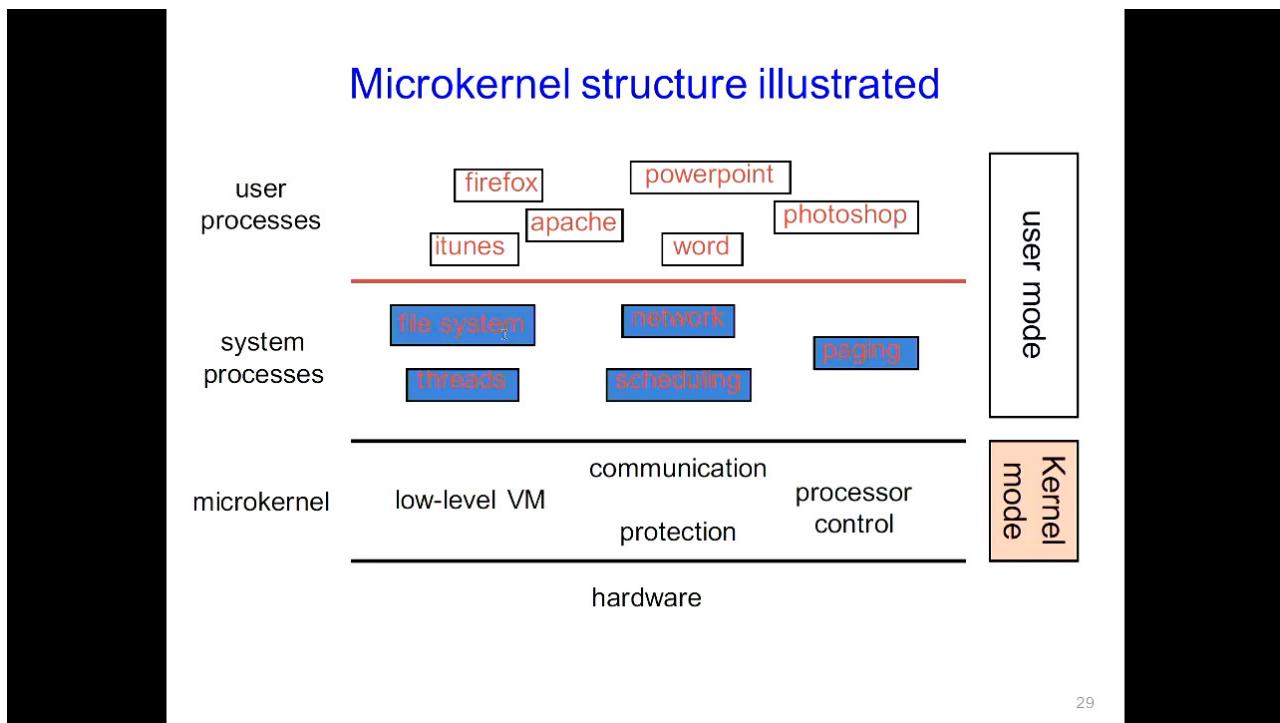
- Disjunction between model and reality: systems modelled as layers, but not really built that way.

Hardware abstraction layer

- An example of layering in modern operating systems
- Goal: separates hardware-specific routines from the *core* OS
 - Provides portability
 - Improves readability

Microkernels

- Popular in the late 80s, early 90s
- Goal: minimize what happens in kernel; item organize rest of OS as user-level processes.
- This results in:
 - better reliability (isolation between components)
 - easy of extension and customisation
 - poor performance (user/kernel boundary crossings)
- First microkernel system was Hydra (CMU, 1970)
 - Contemporaries: Mach (CMU), Chorus (French Unix-like OS), OS X (Apple), in some ways NT (Microsoft)

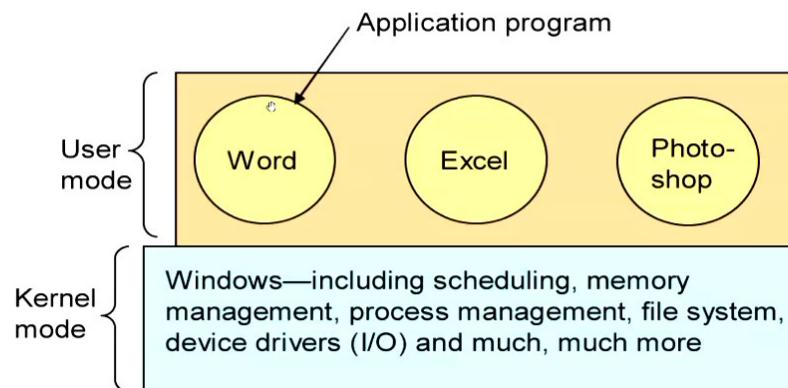


Comparison of OS structures

Windows

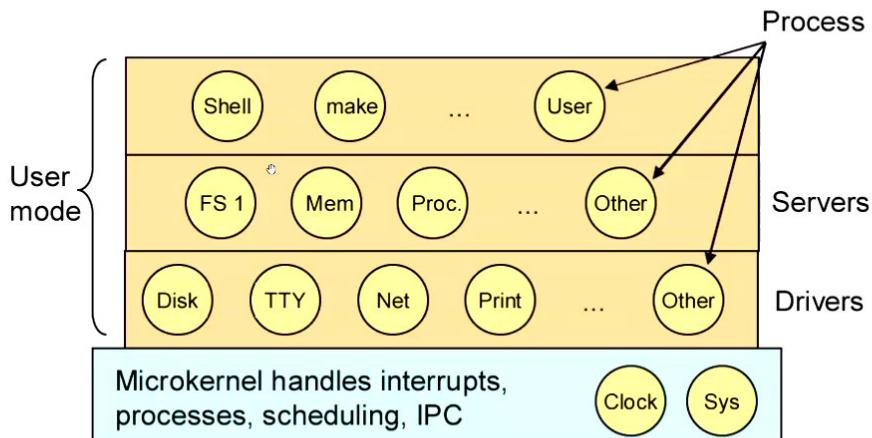
Monolithic

EXAMPLE: WINDOWS



MINIX 3

ARCHITECTURE OF MINIX 3



Loadable kernel modules

- (Perhaps) the best practice for OS design
- Core services in the kernel, and others dynamically loaded
- Common implementations include: Solaris, Linux, etc.
- Advantages
 - convenient: no need for rebooting for newly added modules
 - efficient: no need for message passing unlike micro-kernel

- flexible: any module can call any other module unlike layered model

2.4 Summary

- Fundamental distinction between user and privileged mode supported by most hardware
- OS design has been an evolutionary process of trial and error.
- Successful OS designs have run the spectrum from monolithic, to layered, to micro-kernels
- The role and design of an OS are still evolving
- It is impossible to pick one “correct” way to structure an OS

3 Processes

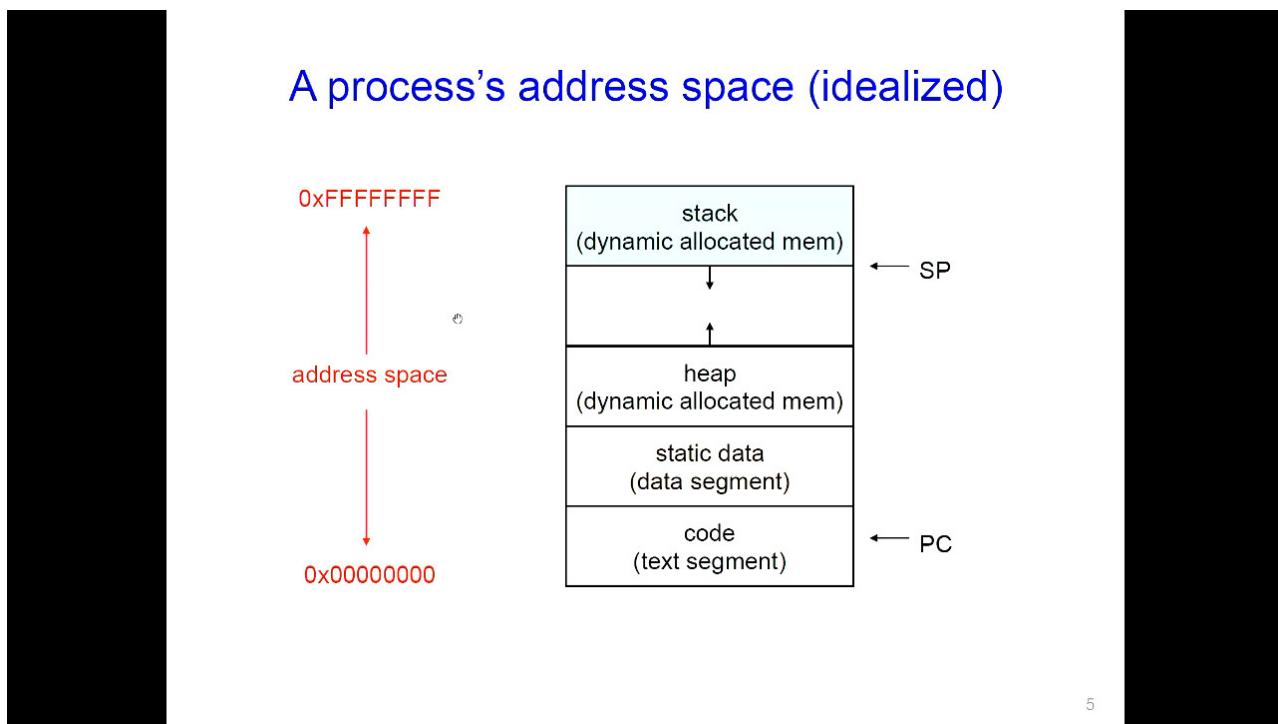
3.1 Process

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 the CPU)
- A sequential process is:
 - The unit of execution
 - The unit of scheduling
 - The dynamic (active) execution context (as opposed to 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;
 - * other general purpose register values.
 - A set of *OS resources*
 - * open files, network connections, sound channels, ...
 - In other words, everything needed to run the program (or to restart, if interrupted).



The OS process namespace

- 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;
- Operations that create processes return a PID (e.g. fork);
- Operations on processes take PIDs as an argument (e.g. kill, wait, nice).

3.2 Process control block

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
 - PC, SP, registers, etc.
 - when a process is unscheduled, the state is transferred out of the hardware into the PCB
 - (when a process is running, its state is spread between the PCB and the CPU).

The PCB

- The PCB is a data structure with many, many fields
 - PID
 - parent PID
 - execution state
 - PC, SP, registers

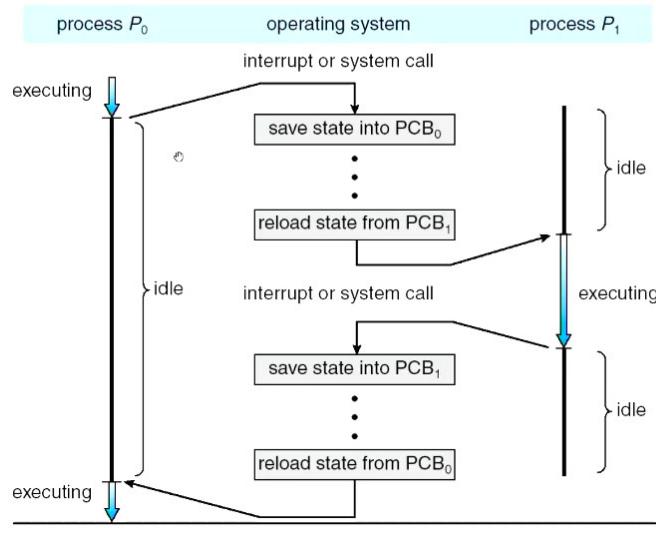
- address space info
 - Unix user id, group id
 - scheduling priority
 - accounting info
 - pointers for state queues
- In Linux:
 - defined in `task_struct` (`include/linux/sched.h`)
 - Over 95 fields!

3.3 Process state & context switch

PCBs and CPU state

- When a process is running, its CPU state is inside 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 in that process's PCB.
- When the OS returns the process to the running state
 - it loads the hardware registers with values from that process's PCB
 - e.g. general purpose registers, SP, instruction pointer
 - This act of switching the CPU from one process to another is called a *context switch*
 - systems may do 100s or 1000s of switches per second;
 - takes a few microseconds on today's hardware;
 - still expensive relative to thread-based context switches.
 - Choosing which process to run next is called *scheduling*.

Process context switch

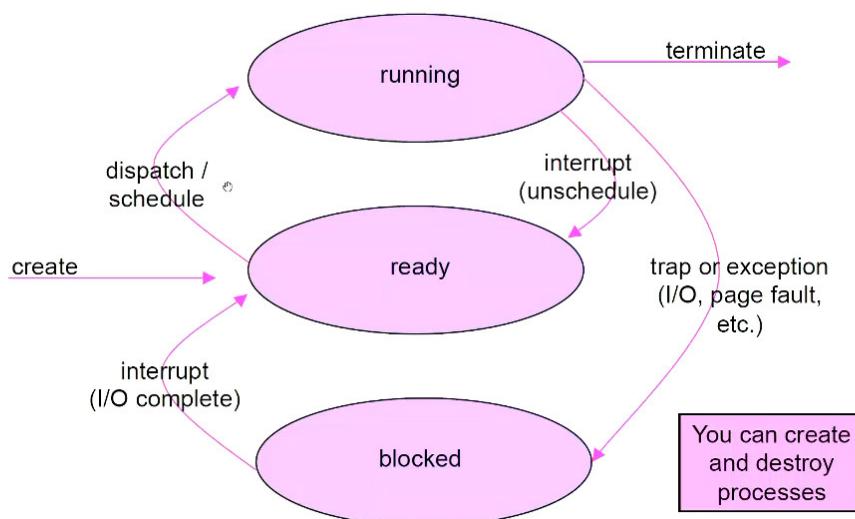


.2

Process execution states

- Each process has an *execution state*, which indicates what it's currently doing
 - *ready*: waiting to be assigned to a CPU — could run, but another process has the CPU;
 - *running*: executing on a CPU — it's the process that currently controls the CPU;
 - *waiting* (aka “blocked”): waiting for an event, e.g. I/O completion, or a message from (or the completion of) another process — cannot make progress until the event happens.
- As a process executes, it moves from state to state
 - Unix: run `top`, STAT column shows current state
 - which state is a process most of the time?

Process states and state transitions



14

State queues

- The OS maintains a collection of queues that represent the state of all processes in the system
 - typically one queue for each state (e.g. ready, waiting, ...);
 - 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.
- The PCBs are moved between queues, which are represented as linked lists.
- There may be many wait queues, one for each type of wait (particular device, timer, message, ...).

PCBs and state queues

- PCBs are data structures
 - dynamically allocated inside 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 (to receive signals, etc.)
 - eventually, OS deallocates the PCB.

3.4 Process creation and termination

Process creation

- New processes are created by existing processes
 - creator is called the *parent*;
 - created process is called the *child*;
 - Unix: do `ps -ef`, look for PPID field
 - what creates the first process, and when?
 - on Unix, this first process is init;
 - on many Linux distributions, this is SystemD or Runit (on Void).

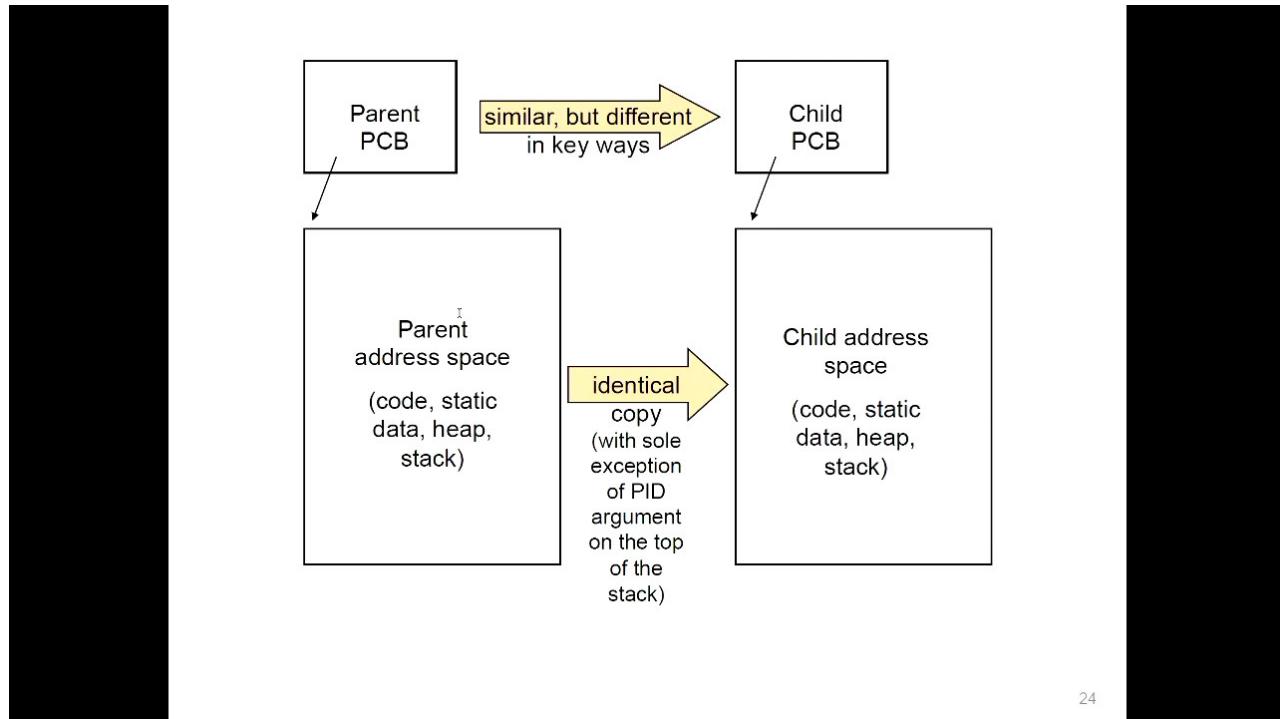
Process creation semantics

- (Depending on the OS) child processes inherit certain attributes of the parent. E.g.
 - 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.

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
 - * initialises new address space with a copy of the entire contents of the address space of the parent
 - places new PCB on the ready queue.
- the `fork` system call “returns twice”
 - once into the parent, and once into the child
 - * returns the child’s PID to the parent
 - * returns 0 to the child
- `fork` = “clone me”.

The return value is used to determine whether we’re the clone or the original.

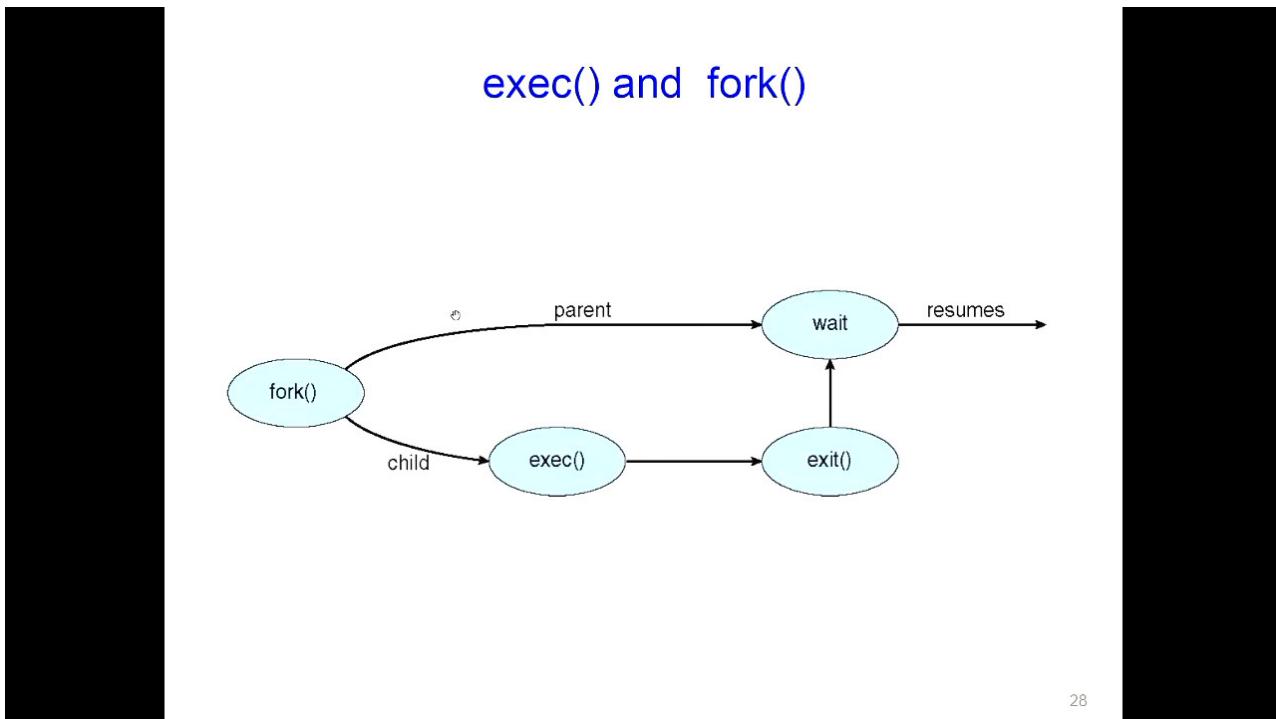


24

`exec` v.s. `fork`

- Q: So how do we start a new program, instead of just forking the old program?
- A: First `fork`, then `exec`.
- `exec`
 - stops the current process
 - loads program ‘prog’ into the address space (i.e. overwrites the existing process image)
 - initialises hardware context, args for new program
 - places PCB onto ready queue

- does not create a new process!



28

Method 1: vfork

- **vfork** is the older (now uncommon) of the two approaches.
- Instead of “child’s address space is a copy of the parent’s”, the semantics are “child’s address space *is* the parent’s”,
 - with a “promise” that the child won’t modify the address space before doing an **execve**.
 - When **execve** is called, a new address space is created and it’s loaded with the new executable.
 - Parent is blocked until **execve** is executed by child.
 - Saves wasted effort of duplicating parent’s address space.

Method 2: copy-on-write

- Retains the original semantics, but copies “only what is necessary” rather than the entire address space.
- On **fork**:
 - Create a new address space
 - Initialise 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, etc.

3.5 Summary

- Process
- PCB
- Process state
- Context switch
- Process creation and termination

4 Threads

4.1 Process vs Threads

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
 - *Thread state*, consisting of
 - * The PC, indicating the next instruction
 - * The SP, indicating the position on the stack
 - * Other general purpose registers
 - A set of *OS resources*
 - * Open files, network connections, sound channels, ...
- Decompose ...
 - address space
 - *thread of control* (stack, SP, PC, registers)
 - OS resources

Motivation

- Threads are about *concurrency* and *parallelism*
- One way to get concurrency and parallelism is to use multiple processes
 - The programs (code) of distinct processes are isolated from each other
- Threads are another way to get concurrency and parallelism
 - Threads *share a process* — same address space, same OS resources
 - Threads have private stack, CPU state — are schedulable

What's needed?

- In many cases
 - Everybody wants to run the same code
 - Everybody wants to access the same data
 - Everybody has the same privileges
 - Everybody uses the same resources (open files, network connections, etc.)
- But you'd like to have multiple hardware execution states:
 - an execution stack and SP
 - * traces state of procedure calls made
 - the PC, indicating the next instruction
 - a set of general-purpose processor registers and their values

How could we achieve this?

- Given the process abstraction as we know it:
 - for several processes
 - cause each to *map* to the *same* physical memory to share data (`shmget`),
- This is really inefficient
 - space: PCB, page tables, etc.
 - time: creating OS structures, fork/copy address space, etc.

Can we do better?

- Key idea:
 - separate the concept of a *process* (address space, OS resources)
 - ... from that of a minimal *thread of control* (execution state: stack, SP, PC, registers),
- This execution state is usually called a *thread*, or a *lightweight process*.

Threads and processes

- Most modern OSs support two entities:
 - the *process*, which defines the address space and general process attributes (such as open files, etc.)
 - the *thread*, which defines a sequential execution stream within a process.
- A thread is bound to a single process / address space
 - address spaces, however, can have multiple threads executing within them
 - sharing data between threads is cheap: all see the same address space
 - creating threads is cheap, too!
- *Threads become the unit of scheduling*
 - processes / address spaces are just *containers* in which threads execute.

Single and Multi-threaded Processes

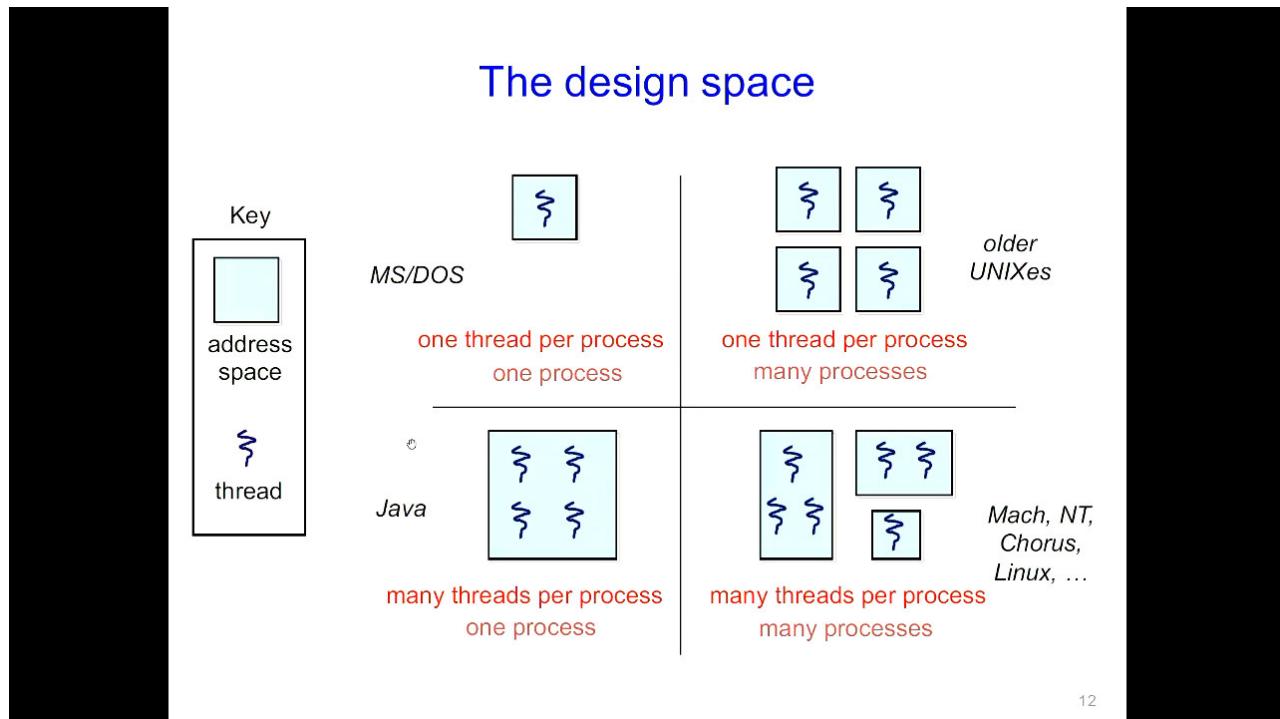
- Different threads in the same process have separate registers and stacks.
- This is cheaper than duplicating the instructions and PCB etc., as required by having multiple processes.

4.2 Concurrency

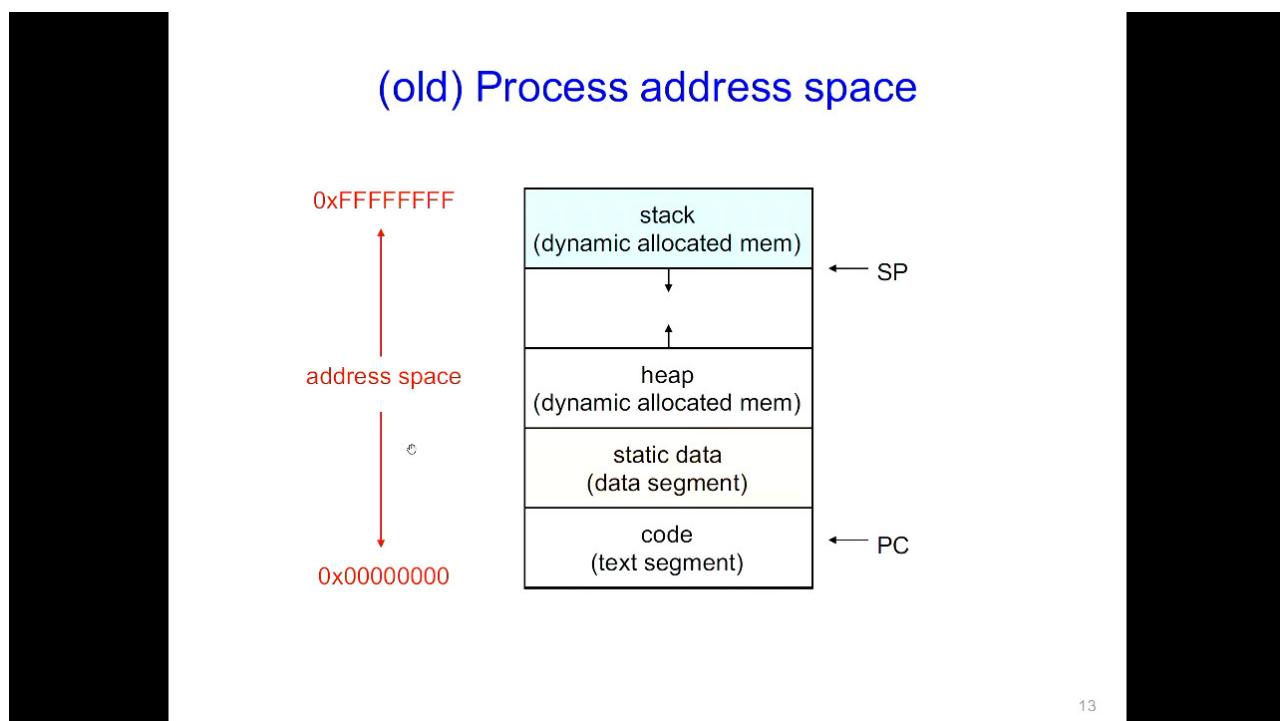
Communication

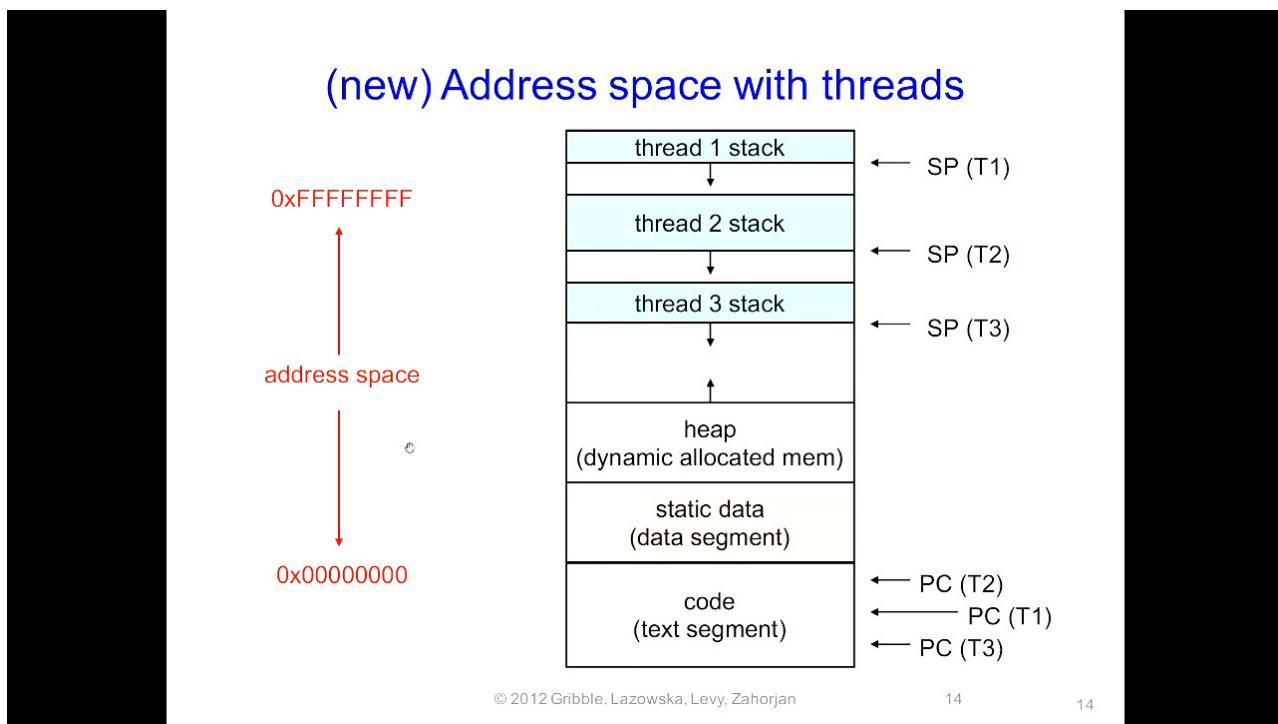
- Threads are concurrent executions sharing an address space (and some OS resources)
- Address spaces provide isolation
 - If you can't name an object, you can't read or write to it
- Hence, communicating between processes is expensive
 - Must go through the OS to move data from one address space to another
- Because threads are in the same address space, communication is simple/cheap
 - Just update a shared variable!

The design space



Process address space





4.3 Design space of process/thread

Process/thread separation

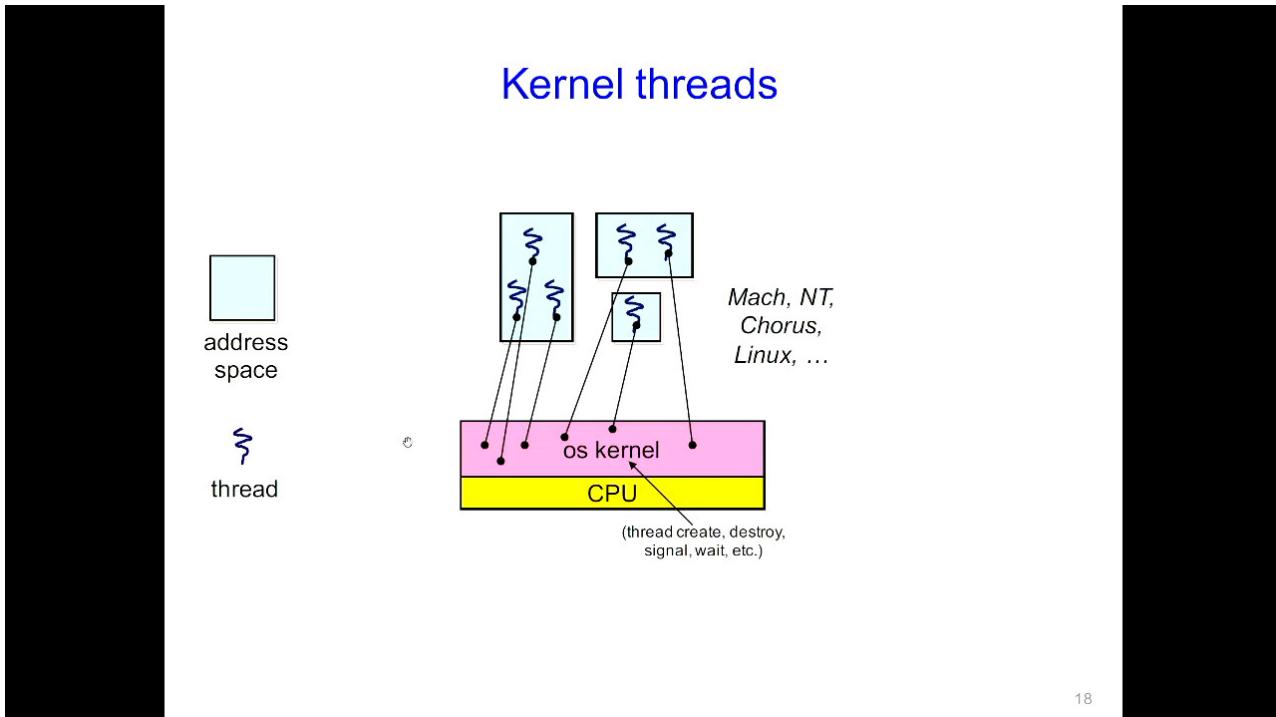
- Concurrency (multi-threading) is useful for:
 - handling concurrent events (e.g. web servers and clients)
 - building parallel programs (e.g. matrix multiply, ray tracing)
 - improving program structure (the Java argument),
- Multi-threading is useful even on a uniprocessor
 - even though only one thread can run at a time
- Supporting multi-threading — that is, separating the concept of a *process* (address space, files, etc.) from that of a minimal *thread of control* (execution state), is a big win
 - creating concurrency does not require creating new processes
 - “faster / better / cheaper”

4.4 Kernel threads

Where do threads come from?

- Natural answer: the OS is responsible for creating/managing threads
For example, the kernel call to create a new thread would
 - allocate an execution stack within the process address space
 - create and initialize a *Thread Control block* (SP, PC, register values)
 - stick it on the ready queue
- We call these *kernel threads*
There is a “thread name space”
 - Thread IDs (TIDs)

- TIDs are integers



18

Kernel Threads

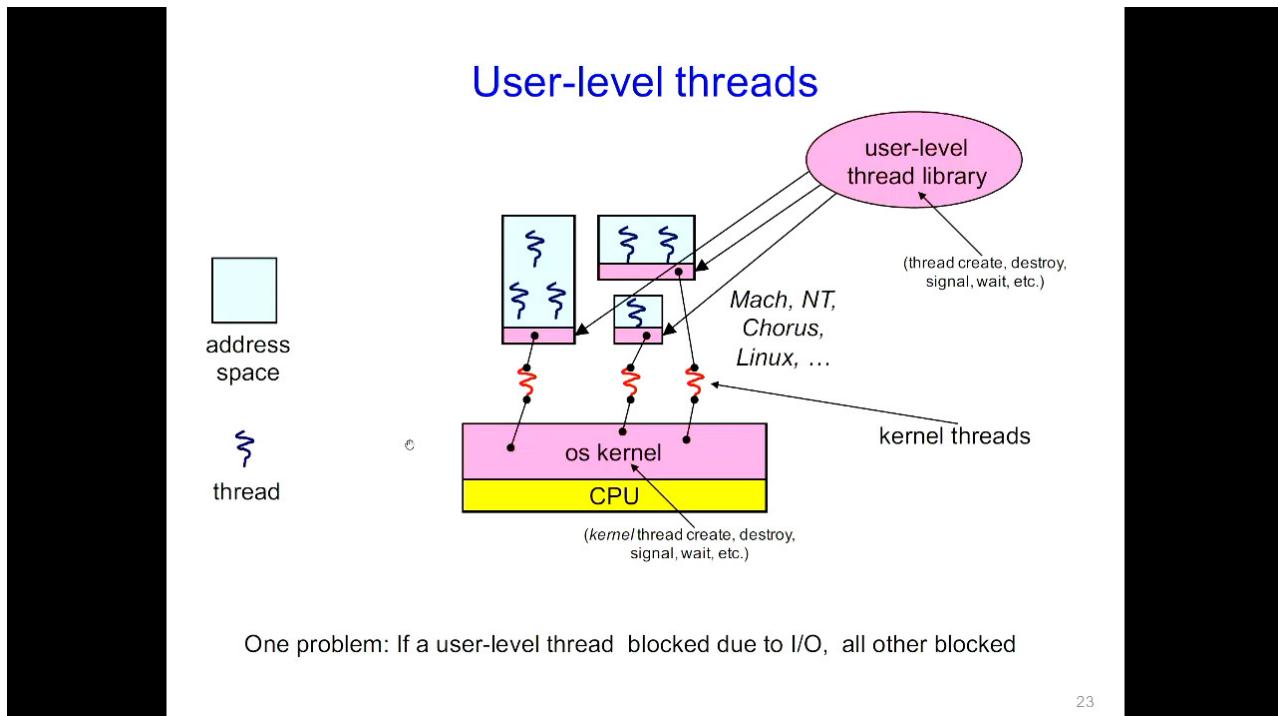
- OS now manages threads *and* processes / address spaces
 - all thread operations are implemented in the kernel
 - OS schedules all of the threads in a system
 - * if one thread in a process blocks (e.g. on I/O), the OS knows about it, and can run other threads from that process
 - * possible to overlap I/O and computation *inside* a process
- Kernel threads are cheaper than processes
 - less state to allocate and initialise
- But, they're still pretty expensive for fine-grained use
 - orders of magnitude more expensive than a procedure call
 - thread operations are all *system calls*
 - * context switch
 - * argument checks
 - must maintain kernel state for each thread

4.5 User-level threads

Cheaper alternative

- There is an alternative to kernel threads
- Threads can also be managed at the user level (within the process)
 - a library linked into the program manages the threads

- * the thread manager doesn't need to manipulate address spaces (which only the kernel can do)
- * threads differ (roughly) only in hardware contexts (PC, SP, registers), which can be manipulated by user-level code
- * the *thread package* multiplexes user-level threads on top of kernel threads
- * each kernel thread is treated as a *virtual processor*
- we call these *user-level threads*



User-level threads

- User-level threads are small and fast
 - managed entirely by user-level library (e.g. `pthreads`)
 - each thread is represented by a PC, registers, a stack, and a small *thread control block* (TCB)
 - creating a thread, switching between threads, and synchronising threads are done *via procedure calls*
 - * no kernel involvement necessary!
- User-level thread operations can be 10–100x faster than kernel threads as a result.

User-level thread implementation

- The OS schedules the kernel thread
- The kernel thread executes user code, including the thread support library and its associated thread scheduler
- The thread scheduler determines when a user-level thread runs
 - it uses queues to keep track of what threads are doing: run, ready, wait
 - * just like the OS and processes
 - * but, implemented at user-level as a library

Thread context switch

- Very simple for user-level threads:
 - save context of currently running thread
 - * push CPU state onto thread stack
 - restore context of the next thread
 - * pop CPU state from next thread's stack
 - return as the new thread
 - * execution resume at PC of next thread
 - Note: no changes to memory mapping required
- This is all done in assembly language
 - it works at the level of the procedure calling convention

How to keep a user-level thread from hogging the CPU?

- Strategy 1: force everyone to cooperate
 - a thread willingly gives up the CPU by calling `yield`
 - `yield` calls into the scheduler, which context switches to another ready thread
 - what happens if a thread never calls `yield`?
- Strategy 2: use presumption
 - scheduler requests that a timer interrupt be delivered by the OS periodically
 - * usually delivered as a Unix signal (`man signal`)
 - * signals are just like software interrupts, but delivered to user-level by the OS instead of delivered to the OS by hardware
 - at each timer interrupt, scheduler gains control and context switches as appropriate.

What if a thread tries to do I/O

- The kernel thread “powering” it is lost for the duration of (synchronous) I/O operation!
 - The kernel thread blocks in the OS, as always
 - It maroons with it the state of the user-level thread
- Could have one kernel thread “powering” each user-level thread
 - “common case” operations (e.g. synchronisation) would be quick
- Could have a limited-size “pool” of kernel threads “powering” all the user-level threads in the address space
 - the kernel will be scheduling these threads, obviously to what's going on at user-level.

4.6 Summary

- Multiple threads per address space
- Kernel threads are much more efficient than processes, but still expensive
 - all operations require a kernel call and parameter validation
- User-level threads are:

- much cheaper and faster
- great for common-case operations
 - * creation, synchronisation, destruction
- can suffer in uncommon cases due to kernel obliviousness
 - * I/O
 - * pre-emption of a lock-holder

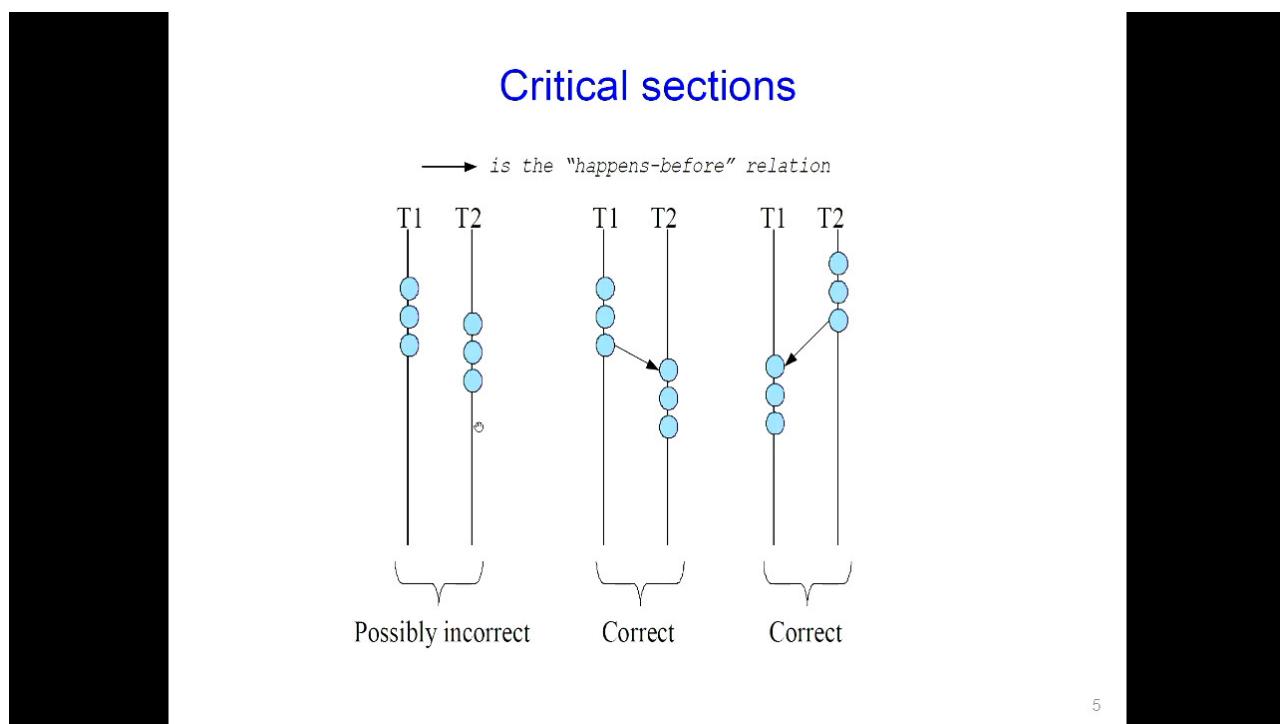
5 Synchronisation

Temporal relations

- User view of parallel threads
 - Instructions executed by a single thread are totally ordered
 - * $A < B < C < \dots$
 - In absence of *synchronisation*:
 - * instructions executed by distinct threads must be considered unordered / simultaneous
 - * Not $X < X'$, and not $X' < X$
- Hardware largely supports this

Critical sections / mutual exclusion

- Sequences of instructions that may get incorrect results if executed simultaneously are called *critical sections*.
- *Race condition* results depend on timing
- *Mutual exclusion* means “not simultaneously”
 - $A < B$ or $B < A$
 - We don’t care which
- Forcing mutual exclusion between two critical section executions
 - is sufficient to ensure correct execution
 - guarantees ordering.



When do critical sections arise?

- One common pattern:
 - read-modify-write of

- a shared value (variable)
- in code that can be executed by concurrent threads
- Shared variable:
 - Global and heap-allocated variables
 - NOT local variables (which are on the stack)

Race conditions

- A program has a *race condition* (data race) if the result of an execution depends on timing (i.e. it is non-deterministic)
- Typical symptoms
 - I run it on the same data, and sometimes it prints 0 and sometimes 4
 - I run it on the same data, and sometimes it prints 0 and sometimes crashes

Correct critical section requirements

- *Mutual exclusion*
At most one thread is in the critical section.
- *Progress*
If thread T is outside the critical section, then T cannot prevent thread S from entering the critical section.
- *Bounded waiting (no starvation)*
If thread T is waiting on the critical section, then T will eventually enter the critical section (assumes threads eventually leave critical sections).
- *Performance*
The overhead of entering and exiting the critical section is small with respect to the work being done within it.

Mechanisms for building critical sections

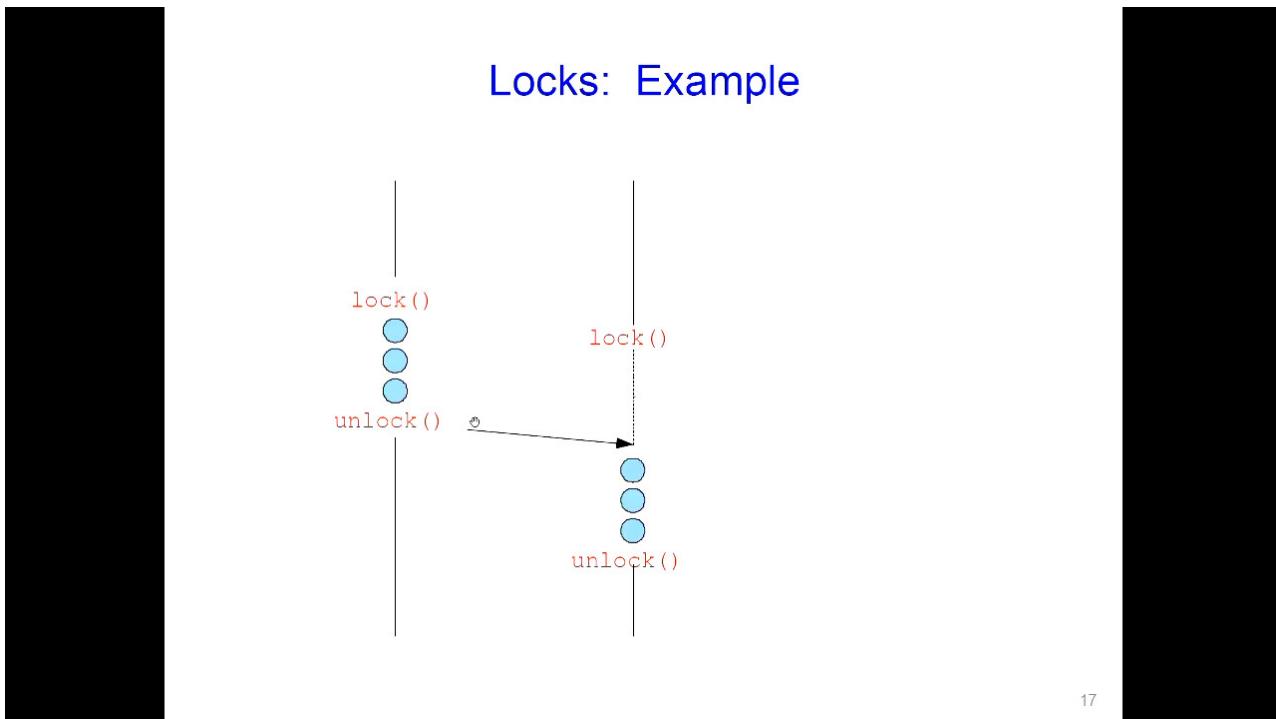
- Spinlocks
 - primitive, minimal semantics — used to build others
- Semaphores (and non-spinning locks)
 - basic, easy to understand, somewhat hard to program with
- Monitors
 - higher level, requires language support, implicit operations
 - easier to program with; Java “**synchronised**”, for example
- Messages
 - Simple model of communication and synchronisation based on (atomic) transfer of data across a channel
 - direct application to distributed systems

5.1 Locks

Locks

- A lock is a memory object with two operations:

- **acquire**: obtain the right to enter the critical section
- **release**: give up the right to be in the critical section
- **acquire** prevents the progress of the thread until the lock can be acquired.
- Note: terminology varies: acquire/release, lock/unlock



17

Acquire/release

- Threads pair up calls to **acquire** and **release**
 - between **acquire** and **release**, the thread *holds* the lock
 - **acquire** does not return until the caller “owns” (holds) the lock
 - * at most one thread can hold a lock at a time
- What happens if the calls aren’t paired
 - I acquire, but neglect to release?
- What happens if the two threads acquire different locks
 - I think that access to a particular shared data structure is mediated by lock A, and you think it’s mediated by lock B?
- What is the right granularity of locking?

5.2 Spinlocks

Spinlocks

- How do we implement spinlocks? Here’s one attempt:

```
struct lock_t {
    int held = 0;
}
void acquire(lock) {
    while (lock->held);
    lock->held = 1;
```

```

    }
    void release(lock) {
        lock->held = 0;
    }
}

```

- Race condition in acquire.

Implementing spinlocks

- Problem is that implementation of spinlocks has critical sections, too!
 - the acquire/release must be *atomic*
 - compiler can hoist code that is invariant
- Need help from the hardware
 - atomic instructions
 - test-and-set, compare-and-swap, ...

Spinlocks: Hardware Test-and-Set

- CPU provides the following as *one atomic instruction*:

```

bool test_and_set(bool *flag) {
    bool old = *flag;
    *flag = True;
    return old;
}

```

- This is a single *atomic* instruction

Implementing spinlocks using Test-and-Set

- So, to fix our broken spinlocks:

```

struct lock{
    int held = 0;
}
void acquire(lock) {
    while (test_and_set(&lock->held));
}
void release(lock) {
    lock->held = 0;
}

```

- *mutual exclusion?* (at most one thread in the critical section)
- *progress?* (T outside cannot prevent S from entering)
- *bounded waiting?* (waiting T will eventually enter)
- *performance?* (low overhead (modulo the spinning part...))

6 Semaphores, Condition Variables, and Monitors

6.1 Semaphore

Semaphore

- More sophisticated synchronisation mechanism
- Semaphore S — integer variable
- Can only be accessed via two atomic operations: `wait` and `signal` (originally called `P` and `V`).
- Definitions

```
wait(S) {
    while (S <= 0); // busy wait
    S--;
}
signal(S) {
    S++;
}
```

- These are performed *atomically*

Semaphore Usage

- *Counting semaphore*: integer value can range over an unrestricted domain
- *Binary semaphore*: integer value can range only between 0 and 1 (same as `lock`)
- Can solve various synchronisation problems
- Consider P_1 and P_2 that require S_1 to happen before S_2
Create a semaphore “synch” initialised to 0

```
P1:
    S_1;
    signal(synch);
P2:
    wait(synch);
    S_2;
```

- Can implement a counting semaphore S as a binary semaphore.

Implementation with no Busy waiting

Each semaphore has an associated queue of threads

```
wait(semaphore *S) {
    S->value--;
    if (S->value < 0) {
        add this thread to S->list;
        block();
    }
}
signal(semaphore *S) {
    S->value++;
    if (S->value <= 0) {
        remove a thread T from S->list;
        wakeup(T);
    }
}
```

```

    }
}

```

Examples

Bounded buffer using semaphores (both binary and counting)

```

var mutex: semaphore = 1      ; mutual exclusion to shared data
empty: semaphore = n        ; count of empty slots (all empty to start)
full: semaphore = 0         ; count of full slots (none full to start)

```

producer:

```

P(empty)   ; block if no slots available
P(mutex)   ; get access to pointers
<add item to slot, adjust pointers>
V(mutex)   ; done with pointers
V(full)    ; note one more full slot

```

consumer:

```

P(full)    ; wait until there's a full slot
P(mutex)   ; get access to pointers
<remove item from slot, adjust pointers>
V(mutex)   ; done with pointers
V(empty)   ; note there's an empty slot
<use the item>

```

7

Readers/Writers using semaphores

```

var mutex: semaphore = 1      ; controls access to readcount
wrt: semaphore = 1           ; control entry for a writer or first reader
readcount: integer = 0        ; number of active readers

```

writer:

```

P(wrt)       ; any writers or readers?
<perform write operation>
V(wrt)       ; allow others

```

reader:

```

P(mutex)           ; ensure exclusion
readcount++       ; one more reader
if readcount == 1 then P(wrt) ; if we're the first, synch with writers
V(mutex)
<perform read operation>
P(mutex)           ; ensure exclusion
readcount--       ; one fewer reader
if readcount == 0 then V(wrt) ; no more readers, allow a writer
V(mutex)

```

9

Semaphores v.s. Spinlocks

- Threads that are blocked at the level of program logic (that is, by the semaphore P operation) are placed on queues, rather than busy-waiting.
- Busy-waiting may be used for the “real” mutual exclusion required to implement P and V
 - but these are very short critical sections — totally independent of program logic
 - and they are not implemented by the application programmer.

Abstract implementation

- **P (`sem`)**
 - acquire “real” mutual exclusion
 - * if `sem` is “available” (≥ 0), decrement sum; *release “real” mutual exclusion*; let thread continue
 - * otherwise, place thread on associated queue; *release “real” mutual exclusion*; run some other thread.
- **V (`sem`)**
 - *acquire “real” mutual exclusion*
 - * if threads are waiting on the associated queue, unblock one (place it on the ready queue)
 - * if no threads are on the queue, `sem` is incremented
the signal is “remembered” for the next time P (`sem`) is called
 - release “real” mutual exclusion
 - the “V-ing” thread continues execution.

Problems with semaphores, locks

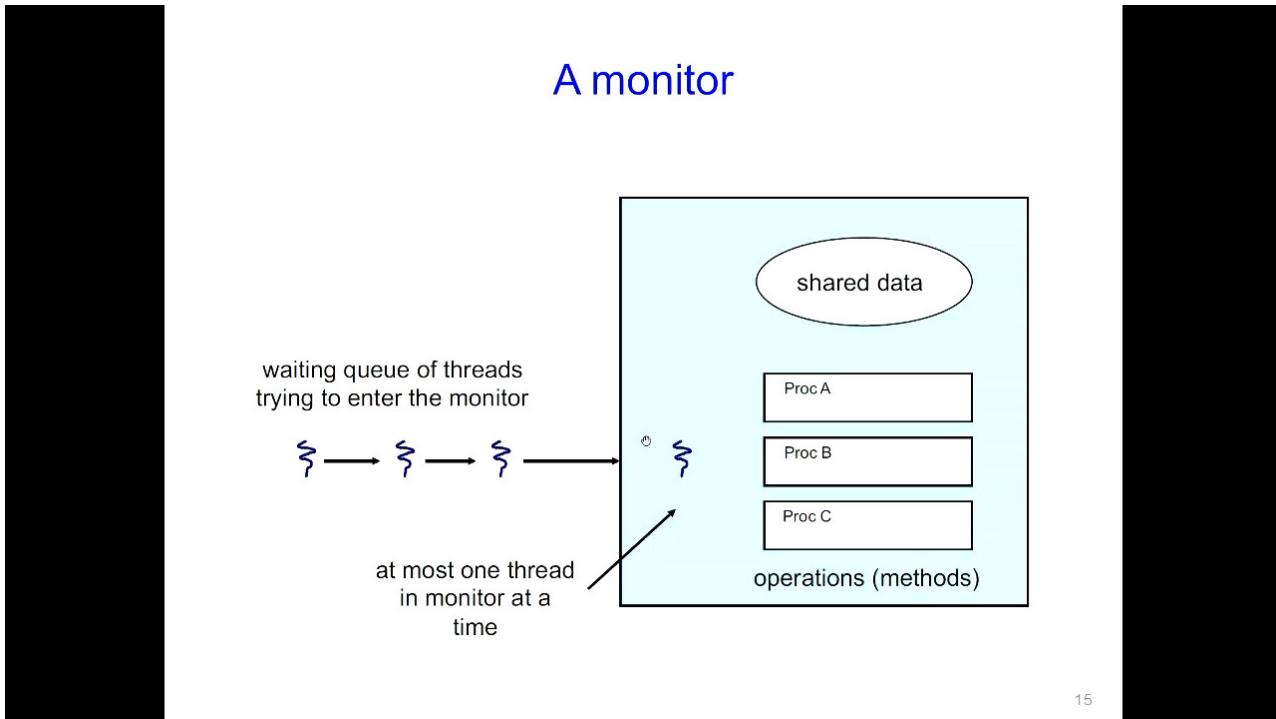
- They can be used to solve any of the traditional synchronisation problems, but it’s easy to make mistakes
 - they are essentially shared global variables
 - * can be accessed from anywhere (bad software engineering)
 - there is no connection between the synchronisation variable and the data being controlled by it
 - no control over their use, no guarantee of proper usage
 - * Semaphores: will there ever be a V ()?
 - * Locks: did you lock when necessary? Unlock at the right time? At all?
- Thus, they are prone to bugs
 - We can reduce the chance of bugs by “stylising” the use of synchronisation
 - Language help is useful for this.

6.2 Monitors

Monitors

- A programming language construct supports controlled shared data access
 - synchronisation code is added by the compiler.
- A class in which every method automatically acquires a lock on entry, and releases it on exit — it combines:
 - *shared data structures* (object);
 - *procedures* that operate on the shared data (object methods);
 - *synchronisation* between concurrent threads that invoke those procedures.
- Data can only be accessed from within the monitor

- protects the data from unstructured access;
- prevents ambiguity about what the synchronisation variable protects.
- Addresses the key usability issues that arise with semaphores.



15

Monitor facilities

- “Automatic” mutual exclusion
 - only one thread can be executing inside at any time
 - * thus, synchronisation is implicitly associated with the monitor — it “comes for free”;
 - if a second thread tries to execute a monitor procedure, it blocks until the first has left the monitor;
 - * more restrictive than semaphores,
 - * but easier to use (most of the time).
- But, there’s a problem...
 - Bounded buffer scenario.

Bounded Buffer scenario

- Monitors require condition variables
- Operations on condition variables
 - `wait (c)`
 - * release monitor lock, so somebody else can get in
 - * wait for somebody else to signal condition
 - * thus, condition variables have associated wait queues
 - `signal (c)`
 - * wake up at most one waiting thread
 - “Hoare” monitor: wakeup immediately, signaller steps outside

- * if no waiting threads, signal is lost
 - this is different from semaphores — no history!
- broadcast (c)
 - * wake up all waiting threads.

Bounded buffer using (Hoare) monitors

```

Monitor bounded_buffer {
    buffer resources[];
    condition not_full;
    condition not_empty;

    produce(resource x) {
        if (array "resources" is full, determined maybe by a count) {
            wait(not_full);
        }
        insert "x" in array "resources";
        signal(not_empty);
    }

    consume(resource *x) {
        if (array "resources" is empty, determined maybe by a count) {
            wait(not_empty);
        }
        *x = get resource from array "resources";
        signal(not_full);
    }
}

```

Runtime system calls for (Hoare) monitors

- EnterMonitor (m) {guarantee mutual exclusion}
- ExitMonitor (m) {hit the road, letting someone else run}
- Wait (c) {step out until condition satisfied}
- Signal (c) {if someone's waiting, step out and let them run}
- EnterMonitor and ExitMonitor are inserted automatically by the compiler.
- This guarantees mutual exclusion for code inside of the monitor.

Monitor Summary

- Language supports monitors
- Compiler understands them
 - Compiler inserts calls to runtime routines for
 - * monitor entry
 - * monitor exit
 - Programmer inserts calls to runtime routines for
 - * signal
 - * wait

- Language/object encapsulation ensures correctness
 - * Sometimes! With conditions, you *still* need to think about synchronisation
- Runtime system implements these routines
 - moves threads on and off queues
 - *ensures mutual exclusion!*

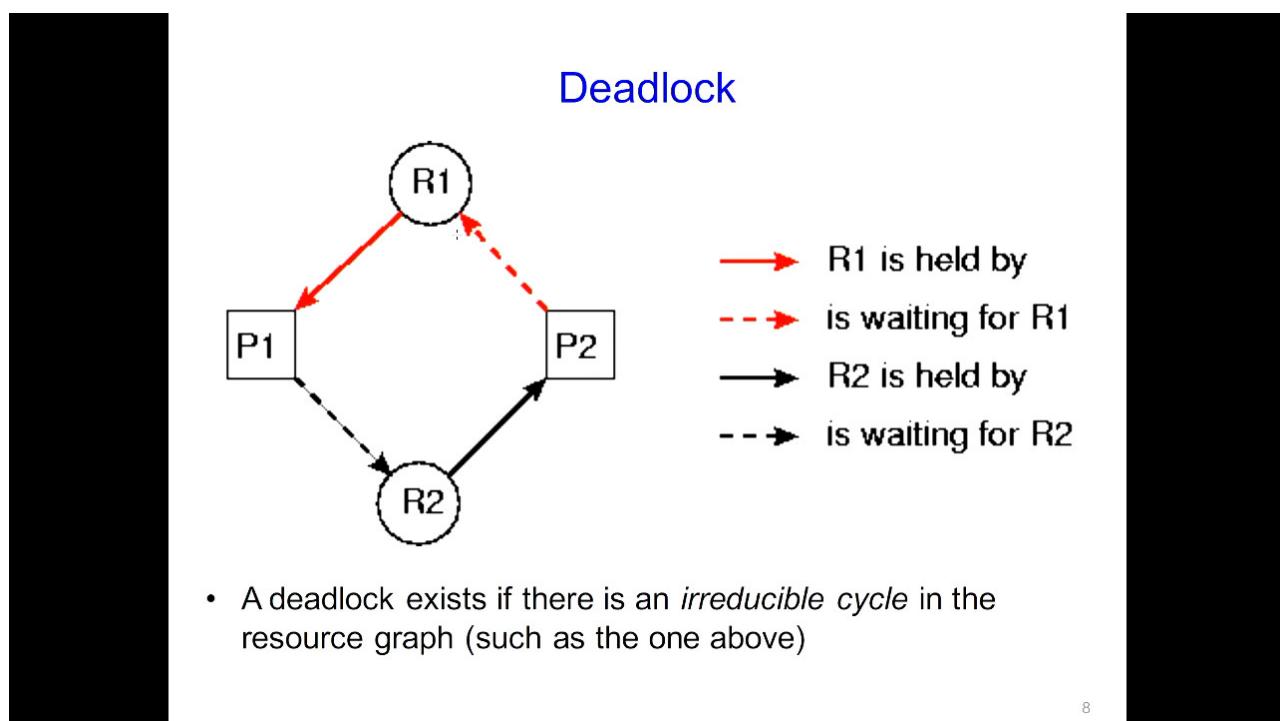
7 Deadlock

Definition

- A thread is deadlocked when it's waiting for an event that can never occur
- Thread A is in critical section 1
waiting for access to critical section 2;
- Thread B is in critical section 2
waiting for access to critical section 1

Four conditions must exist for deadlock to be possible

1. Mutual exclusion
2. Hold and wait
3. No pre-emption
4. Circular wait



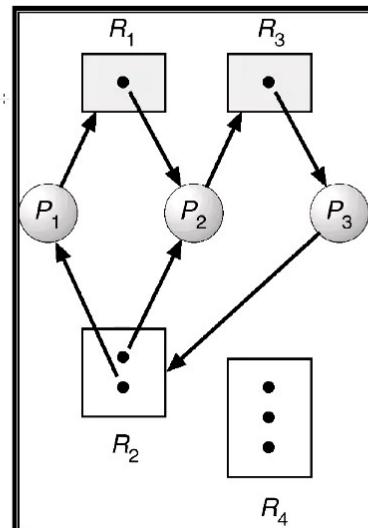
8

7.1 Graph reduction

Graph reduction

- A graph can be *reduced* by a thread if all of that thread's requests can be granted
 - in this case, the thread eventually will terminate — all resources are freed — all arcs (allocations) to/from it in the graph are deleted.
- Miscellaneous theorems (Holt, Havender):
 - There are no deadlocked threads if and only if the graph is completely reducible.
 - The order of reductions is irrelevant.

Resource allocation graph with a deadlock



11

Handling deadlock

- Eliminate one of the four required conditions
 - Mutual exclusion
 - Hold and Wait
 - No pre-emption
 - Circular wait
- Broadly classified as:
 - Prevention, or
 - Avoidance, or
 - Detection (and recovery)

Deadlock prevention

Restrain the ways requests can be made

- Mutual exclusion
not required for sharable resources (e.g. read-only files); must hold for non-sharable resources.
- Hold and wait
must guarantee that whenever a process requests a resources, it does not hold any other resources.
 - Low resources utilisation; starvation is possible.
- No (resource) Pre-emption
 - If a process holding some resources requests another unavailable resource all resources currently held are released.
 - Process will be restarted only when it can regain its old resources, as well as the new ones that it is requesting.
- Circular wait

- impose a total ordering of all resource types, and require that each process requests resources in an increasing order of enumeration.

Avoidance

Less severe restrictions on program behaviour.

- Eliminating circular wait
 - each thread states its maximum claim for every resource type;
 - system runs the Banker's Algorithm at each allocation request
 - Banker \implies highly conservative

7.2 Banker's Algorithm

Banker's Algorithm example

- Background
 - The set of controlled resources is known to the system.
 - The number of units of each resource is known to the system.
 - Each application must declare its maximum possible requirement of each resource type.
- The, the system can do the following:
 - When a request is made:
 - * pretend you granted it;
 - * pretend all other legal requests were made;
 - * can the graph be reduced?
 - If so: allocate the requested resource.
 - If not, block the thread until some thread releases resources, and then try pretending again.

Safe state

- When requesting an available resource decide if allocation leaves the system in a safe state
- We're in a *safe state* if there exists a sequence $\langle P_1, P_2, \dots, P_n \rangle$ of *all* the processes in the systems
 - such that for each P_i , the resources that P_i can still request can be satisfied by currently available resources + resources held by all the P_j , with $j < i$.
- That is:
 - If P_i resource needs are not immediately available, then P_i can wait until all P_j have finished.
 - When P_j is finished, P_i can obtain needed resources, execute, return allocated resources, and terminate.
 - When P_i terminates, P_{i+1} can obtain its needed resources, and so on.

Safe \implies no deadlock; deadlock \implies unsafe

Data Structures for the Banker's Algorithm Let n = number of processes, and m = number of resource types.

- **Available:** Vector of length m . If $\text{Available}[j] = k$, there are k instances of resource type R_j available.

- Max $n \times m$ matrix. If $\text{Allocation}[i, j] = k$, then P_i is currently allocated k instances of R_j
- Allocation: $n \times m$ matrix. If $\text{Need}[i, j] = k$, then P_i may need k more instances of R_j to complete its task.

$$\text{Need}[i, j] = \text{Max}[i, j] - \text{Allocation}[i, j]$$

Safety Algorithm

1. Let Work and Finish be vectors of length m and n , respectively. Initialise:

```
Work = Available
Finish[i] = false for i = 0..n-1
```

2. Find an i such that both:

- (a) $\text{Finish}[i] == \text{false}$
- (b) $\text{Need}_i \leq \text{Work}$

If no such i exists, go to step 4

3. $\text{Work} = \text{Work} + \text{Allocation}$

```
Finish[i] = true
go to step 2
```

4. If $\text{Finish}[i] == \text{true}$, for all i , then the system is in a safe state.

Resource-Request Algorithm for Process P_i

Request_i = request vector for process P_i . If $\text{Request}_i[j] = k$ then process P_i wants k instances of resource type R_j .

1. If $\text{Request}_i \leq \text{Need}_i$ go to step 2. Otherwise raise error condition, since process has exceeded its maximum claim.

2. If $\text{Request}_i \leq \text{Available}$, go to step 3. Otherwise P_i must wait, since resources are not available.

3. Pretend to allocate requested resources to P_i by modifying the state as follows:

```
Available = Available - Request
Allocation_i = Allocation_i + Request_i
Need_i = Need_i - Request_i
```

- (a) If safe, then resources allocated to P_i

- (b) If unsafe, then P_i must wait, and the old resource-allocation state is restored.

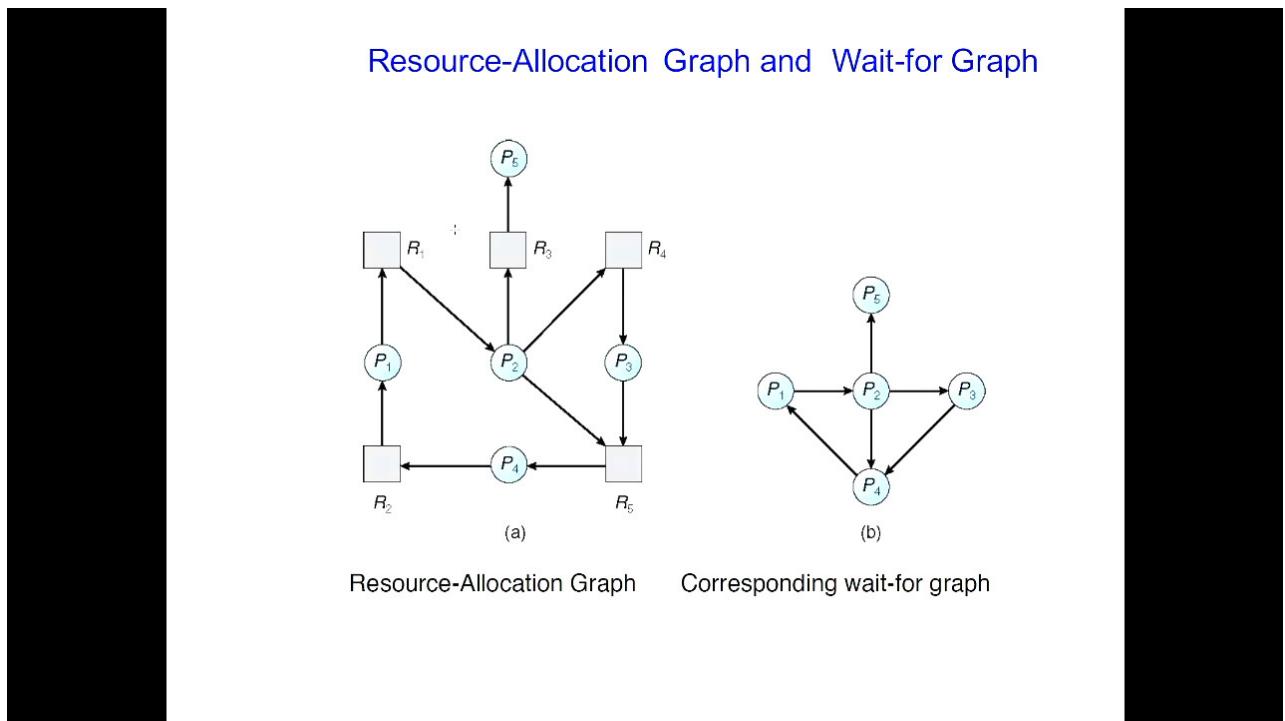
Deadlock Detection

1. Allow system to enter deadlock state
2. Detection algorithm
3. Recovery scheme

Single instance of each resource type

- Maintain a *wait-for* graph
 - Nodes are processes
 - $P_i \rightarrow P_j$ if P_i is waiting for P_j
- Periodically invoke an algorithm that searches for a cycle in the graph.
 - If there is a cycle, there exists a deadlock.
- An algorithm to detect a cycle in a graph

- has runtime complexity $\mathcal{O}(n^2)$ with n being the number of vertices in the graph.



Detection-Algorithm usage

- When, and how often to invoke depends on:
 - How often a deadlock is likely to occur?
 - How many processes will need to be rolled back?
 - * One for each disjoint cycle.
- If detection algorithm is invoked arbitrarily:
 - there may be many cycles in the resource graph
 - we would not be able to tell which deadlocked processes “caused” the deadlock.

Recovery from deadlock

- Process termination
 - Abort all deadlocked processes
 - Abort one process at a time until the deadlock cycle is eliminated
 - In which order should we choose to abort?
- Resource pre-emption
 - *Select a victim* — minimise cost
 - *Rollback* — return to some safe state, restart process for that state.
 - *Starvation* — same process may always be picked as victim, include number of rollback in cost factor.

Summary

- Deadlock is bad!
- We can deal with it either statically (prevention) or dynamically (avoidance and/or detection)

- In practice, you'll encounter lock ordering, periodic deadlock detection/correction, and mine-fields.

8 Scheduling

Scheduling

- We have talked about *context switching*
 - an interrupt occurs (device completion, timer interrupt)
 - a thread causes a trap or execution
 - may need to choose a different thread/process to run
- Glossed over which process or thread to run next
 - “some thread from the ready queue”
- This decision is called *scheduling*
 - scheduling is a *policy*
 - context switching is a *mechanism*

Classes of Schedulers

- Batch
 - Throughput / utilisation oriented
 - Example: audit inter-bank funds transfers each night, Pixar rendering, Hadoop/MapReduce jobs.
- Interactive
 - Response time oriented
- Real time
 - Deadline driven
 - Example: embedded systems (cars, aeroplanes, etc.)
- Parallel
 - Speedup-driven
 - Example: “space-shared” use of a 1000-processor machine for large simulations.

Multiple levels of scheduling decisions

- Long term
 - Should a “job” be “initiated”, or should it be held?
 - Typical of batch systems.
- Medium term
 - Should a running program be temporarily marked as non-runnable (e.g. swapped out)?
- Short term
 - Which thread should be given to the CPU next? For how long?
 - Which I/O operation should be sent to the disk next?
 - On a multiprocessor:
 - * Should we attempt to coordinate the running of threads from the same address space in some way?

- * Should we worry about cache state (processor affinity)?

8.1 Scheduling Goals

Scheduling Goals I: Performance

Many possible metrics / performance goals (which sometimes conflict)

- maximise *CPU utilisation*
- maximise *throughput* (requests completed per second)
- minimise *average response time* (average time from submission of request to completion of response)
- minimise *average waiting time* (average time from submission of request to start of execution)
- minimise *energy* (joules per instruction) subject to some constraint (e.g. frames per second)

Scheduling Goals II: Fairness

- No single, compelling definition of “fair”
 - How to measure fairness?
 - Fair per-user? Per-process? Per-thread?
 - What if one process is CPU bound, and one is I/O bound?
- Sometimes the goal is to be unfair:
 - Explicitly favour some particular class of requests (priority system), but...
 - avoid starvation (be sure everyone gets at least some service).

When to assign?

Pre-emptive v.s. non-pre-emptive schedulers

- Non pre-emptive
 - once you give somebody the green light, they've got it until they relinquish it
 - an I/O operation
 - allocation of memory in a system without swapping
- Pre-emptive
 - you can re-visit a decision
 - * setting the timer allows you to pre-empt the CPU from a thread even if it doesn't relinquish it voluntarily.
 - Re-assignment always involves some overhead
 - * Overhead doesn't contribute to the goal of any scheduler.

We'll assume “work conserving” policies

- Never leave a resource idle when someone wants it

8.2 Laws and properties

The Utilisation Law: $U = X \times S$

- U utilisation
- X throughput (requests per second)
- S average service time

- Utilisation is constant, independent of the schedule, so long as the workload can be processed

Little's Law: $N = X \times R$

- N average number in system
- X throughput
- R average response time
- a better average response time implies fewer in system, and vice versa.

Response Time R at a single server under FCFS scheduling:

- $R = \frac{S}{1-U}$
- $N = \frac{U}{1-U}$

8.3 Algorithms

Algorithm 1: First-come first-served (FCFS)

- schedule in the order that they arrive
- “real-world” scheduling of people in (single) lines
- jobs treated equally, no starvation
- Drawbacks:
 - Average response time can be poor: *convoy effect*
 - May lead to poor utilisation of other resources
 - * if you send me on my way, I can go keep another resource busy
 - * FCFS may result in poor overlap of CPU and I/O activity
 - The more copies of the resource there are to be scheduled
 - * the less dramatic the impact of occasional very large jobs (so long as there is a single waiting line)
 - * e.g. multiple cores v.s. single core

Algorithm 2: Shortest-job-first (SJF)

- Associate with each process the length of its next CPU burst
 - use these lengths to schedule the process with the shortest time
- SJF is optimal — gives minimum average waiting time for a given set of processes
 - the difficulty is knowing the length of the next CPU request
 - could ask the user.
- Determining the length of next CPU burst
 - Can only estimate the length — should be similar to the previous one
 - * then pick process with shortest predicted next CPU burst.
 - Can be done by using the length of previous CPU bursts, using exponential averaging
 1. t_n actual length of n th CPU burst
 2. τ_{n+1} predicted value for the next CPU burst

- 3. $\alpha, 0 \leq \alpha \leq 1$
- 4. Define: $\tau_{n+1} = \alpha t_n + (1 - \alpha)\tau_n$
- Commonly, set $\alpha = 0.5$
- Pre-emptive version called *shortest-remaining-time-first*

Algorithm 3: Round Robin (RR)

- Each process gets a small unit of CPU time (*time quantum q*), usually 10–100 milliseconds.
 - After this time has elapsed, the process is pre-empted and added to the end of the ready queue.
- If there are n processes in the ready queue and the time quantum is q ,
 - then each process gets $\frac{1}{n}$ of the CPU time in chunks of at most q time units at once.
 - No process waits more than $(n - 1)q$ time units.
- Timer interrupts every quantum to schedule next process
- Performance
 - q large \implies FIFO
 - q small \implies q must be large with respect to context switch, otherwise overhead is too high.
- Drawbacks:
 - What if all jobs are exactly the same length?
 - What do you set the quantum to be?
 - * no value is “correct”
 - * if small, then context switch often, incurring high overhead
 - * if large, then the response time degrades.
 - Treats all jobs equally

Algorithm 4: Priority Scheduling

- A priority number (integer) is associated with each process
- The CPU is allocated to the process with the highest priority
- SJF is priority scheduling where priority is the inverse of predicted next CPU burst time.
- Problem: *starvation* — low priority processes may never execute.
- Solution: *ageing* — as time progresses, increase the priority of the process.

Multi-level Feedback Queues (MLFQ)

- It’s been observed that workloads tend to have increasing residual life — “if you don’t finish quickly, you’re probably a lifer”
- This is exploited in practice by using a policy that discriminates against the old.
- MLFQ:
 - there is a hierarchy of queues
 - there is a priority ordering among the queues
 - new requests enter the highest priority queue

- each queue is scheduling RR
- requests move between queues based on execution history.

Unix scheduling

- Canonical scheduler is pretty much MLFQ
 - 3–4 classes spanning ~ 170 priority levels
 - * time-sharing: lowest 60 priorities
 - * system: middle 40 priorities
 - * real-time: highest 60 priorities
 - priority scheduling across queues, RR within
 - * process with highest priority always run first
 - * processes with same priority scheduled RR
 - processes dynamically change priority
 - * increases over time if process blocks before end of quantum
 - * decreases if process uses entire quantum
- Goals:
 - reward interactive behaviour over CPU hogs
 - * interactive jobs typically have short bursts of CPU

8.4 Summary

- Scheduling takes place at many levels
- It can make a huge difference in performance
 - this difference increases with the variability in service requirements
- Multiple goals, sometimes conflicting
- There are many “pure” algorithms, most with some drawbacks in practice — FCFS, SJF, RR, Priority
- Real system use hybrids that exploit observed program behaviour
- Scheduling is important

9 Memory Management

9.1 Background

Goals and Tools of memory management

- Allocate memory resources among competing processes
 - maximising memory utilisation and system throughput
- Provide isolation between processes
 - Addressability and protection: orthogonal
- Convenient abstraction for programming
 - and compilers, etc.
- Tools
 - Base and limit registers
 - Swapping
 - Segmentation
 - Paging, page tables, and TLB
 - Virtual memory

Background

- Program must be brought (from disk) into memory and placed within a process for it to be run.
- Main memory and registers are only storage CPU can access directly.
- Memory unit only sees a stream of address + read requests, or addresses + data and write requests
- Register access in one CPU clock (or less)
- Main memory can take many cycles, causing a *stall*
- *Cache* sits between main memory and CPU registers.
- Protection of memory required to ensure correct operation.

Base and Limit Registers

- A pair of *base* and *limit* registers define the logical address space.
- CPU must check every memory access generated in user mode to be sure it is between base and limit for that user.

9.2 Logical/Virtual address space v.s. Physical address space

Virtual address for multiprogramming

- To make it easier to manage memory of multiple processes, make processes *use logical or virtual address*
 - Logical/virtual addresses are independent of location in physical memory data lives
 - * OS determines location in physical memory.
- Instructions issued by CPU reference logical/virtual addresses
 - e.g. pointers, arguments to load/store instructions, PC, etc.

- Logical/virtual addresses are translated by hardware into physical addresses (with some setup from OS).

Logical/Virtual Address Space

- The set of logical/virtual addresses a process can reference is its *address space*
 - many different possible mechanisms for translating logical/virtual addresses to physical addresses.
- Program issues addresses in a logical/virtual address space
 - must be *translated* to physical address space;
 - think of the program as having a contiguous logical/virtual address space that starts at 0; and a contiguous physical address space that starts somewhere else.
- *Logical/virtual address space* is the set of all logical addresses generated by a program.
- *Physical address space* is the set of all physical addresses generated by a program.

Memory-Management Unit (MMU)

- Hardware device
 - at runtime maps virtual to physical addresses
- Many methods possible
- Simple scheme: value in relocation register is added to every address generated by a user process at the time it is sent to memory.
 - Base register now called *relocation register*
- The user program deals with *logical* addresses — it never sees the *real* physical addresses.
 - Execution-time binding occurs when reference is made to location in memory.
 - Logical address bound to physical addresses.

9.3 Swapping

Swapping

- What if not enough memory to hold all processes?
- A process can be *swapped* temporarily
 - out of memory to a backing store,
 - brought back into memory for continued execution
 - total physical memory space of processes can exceed physical memory.
- *Backing store* — fast disk
 - large enough to accommodate copies of all memory images for all users;
 - must provide direct access to these memory images.
- *Roll out, roll in* — swapping variant
 - used for priority-based scheduling algorithms;
 - lower-priority process is swapped out so higher-priority processes can be loaded and executed.
- Major part of swap time is transfer time

- total transfer time is directly proportional to the amount of memory swapped.
- System maintains a *ready queue*
 - ready-to-run processes which have memory images on disk.

Context Switch Time including Swapping

- If next processes to be put on CPU is not in memory
 - need to swap out a process and swap in target process.
- Context switch time can then be very high
- Can reduce cost
 - reduce size of — by knowing how much memory really being used;
 - inform OS of memory use via `request_memory()` and `release_memory()`.
- Other constraints as well on swapping
 - Pending I/O — can't swap out as I/O would occur to wrong process.
- Or always transfer I/O to kernel space, then I/O device
 - known as *double buffering*, adds overhead.
- Standard swapping not used in modern operating systems
 - But modified version common
 - * Swap only when free memory extremely low.

9.4 Contiguous Memory Allocation

Contiguous Allocation

- Main memory must support both OS and user processes
- Limited resource, must allocate efficiently
- Contiguous allocation is one early method
- Main memory usually into two *partitions*:
 - Resident OS, usually held in low memory with interrupt vector;
 - User processes then held in high memory;
 - Each process contained in single contiguous section of memory.
- Relocation registers
 - used to protect user processes from each other, and from changing OS code and data;
 - base register contains value of smallest physical address;
 - limit register contains range of logical addresses — each logical address must be less than the limit register.
- MMU maps logical address *dynamically*
 - can then allow actions such as kernel code being *transient* and kernel changing size.

Multiple-partition allocation

- Degree of multiprogramming limited by number of partitions
- 2 approaches

- Fixed partition
- Variable partition

Old technique 1: Fixed partitions

- Physical memory is broken up into fixed partitions
 - partitions may have different sizes, but partitioning never changes
 - hardware requirement: *base/relocation register, limit register*
 - * physical address = logical address + base register
 - * base register loaded by OS when it switches to a process
- Advantages:
 - Simple
- Problems
 - *internal fragmentation*: the available partition is larger than what was requested.

Old technique 2: Variable partitions

- Obvious next step: physical memory is broken up into partitions dynamically — partitions are tailored to programs
 - hardware requirements: *base register, limit register*
 - physical address = logical address + base register
- Advantages
 - no internal fragmentation
 - * simply allocate partition size to be just big enough for process (assuming we know what that is!)
- Problems
 - *external fragmentation*
 - * as we load and unload jobs, holes are left scattered throughout physical memory.

Multiple-partition allocation

- *Variable-partition* sizes for efficiency (sized to a given process' needs).
- *Hole* — block of available memory; holes of various sizes are scattered throughout memory.
- When a process arrives, allocated memory from a hole large enough to accommodate it.
- Process exiting frees its partition, adjacent free partitions combined.
- OS maintains information about:
 1. allocated partitions,
 2. free partitions (hole)

Dynamic Storage-Allocation Problem

- *First-fit*: Allocate the *first* hole that is big enough
- *Best-fit*: Allocate the *smallest* hole that is big enough; must search entire list, unless ordered by size
 - produces the smallest leftover hole.

- *Worst-fit*: Allocates the *largest* hole; must also search the entire list
 - produces the largest leftover hole

First-fit and best-fit better than worst-fit in terms of speed and storage utilisation.

Fragmentation

- *External fragmentation*: total memory space exists to satisfy a request, but it is not contiguous;
- *Internal fragmentation*: allocated memory may be slightly larger than requested memory;
- First fit analysis reveals that given N blocks allocated, $0.5N$ blocks lost to fragmentation
- $\frac{1}{3}$ may be unusable $\implies 50\text{ percent rule.}$

Dealing with fragmentation

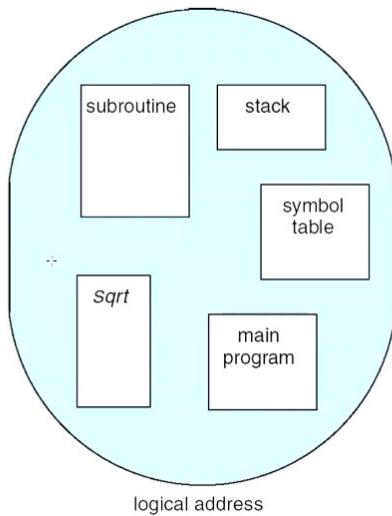
- Compact memory by copying
 - Swap a program out
 - Reload it, adjacent to another
 - Adjust its base register
 - Compaction is possible *only* if relocation is dynamic
 - I/O problem:
 - * Latch job in memory while it is involved in I/O
 - * Do I/O only into OS buffers

9.5 Segmentation

Segmentation

- Dealing with fragmentation
 - Why not remove the need for continuous addresses?
- Segmentation
 - partition an address space into *logical* units
 - * stack, code, heap, subroutines
 - a virtual address is <segment#, offset>
- Facilitates sharing and reuse
 - a segment is a natural unit of sharing — a subroutine or function
- A natural extension of variable-sized partitions
 - variable-sized partition = 1 segment per process
 - segmentation = many segments per process.

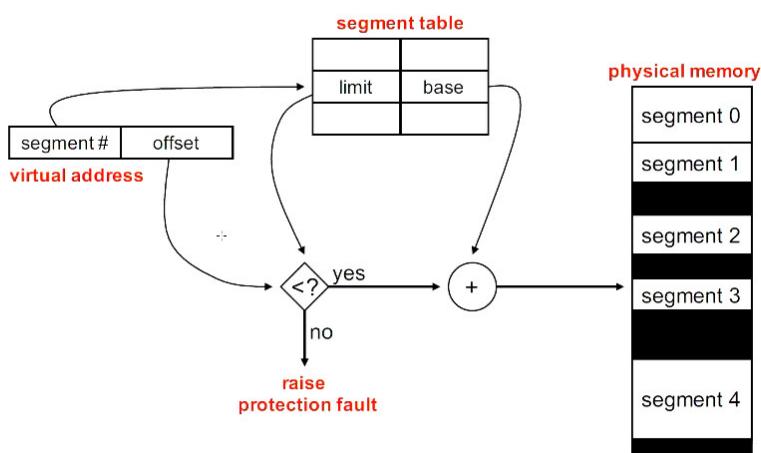
User's View of a Program



Hardware support

- Segment table
 - multiple base/limit pairs, *one per segment*
 - segments named by segment#, used as index into table
 - * a virtual address is <segment#, offset>
 - offset of virtual address added to base address of segment to yield physical address

Segment lookups



30

Pros and Cons

- Logical and it facilitates sharing and reuse

- Allows non-contiguous physical addresses
 - Helps exploit varying sized holes
- But it has the complexity of a variable partition system
 - except that linking is simpler, and the “chunks” that must be allocated are smaller than a “typical” linear address space.
- Segmentation rarely used alone
 - Paging is the basis for modern memory management

9.6 Summary

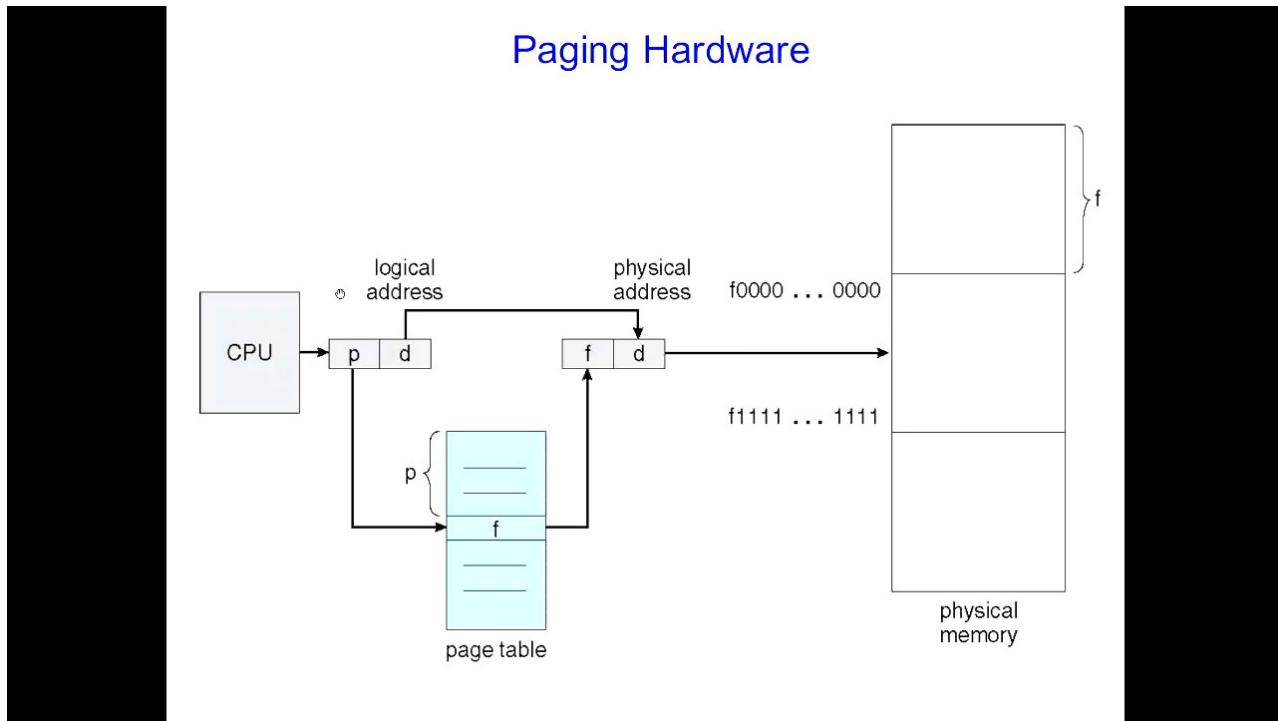
- Logical/Virtual Address Space v.s. Physical Address Space
- Swapping
- Contiguous memory allocation
- Fragmentation
- Segmentation
- Paging
 - A better solution

10 Paging

10.1 Paging

Address translation scheme

- Address generated by CPU is divided into:
 - *Page number (p)* — used as an index into a *page table* which contains base address of each page in physical memory
 - *Page offset (d)* — combined with base address to define the physical memory address that is sent to the memory unit



10.2 Page Tables

Implementation of Page Table

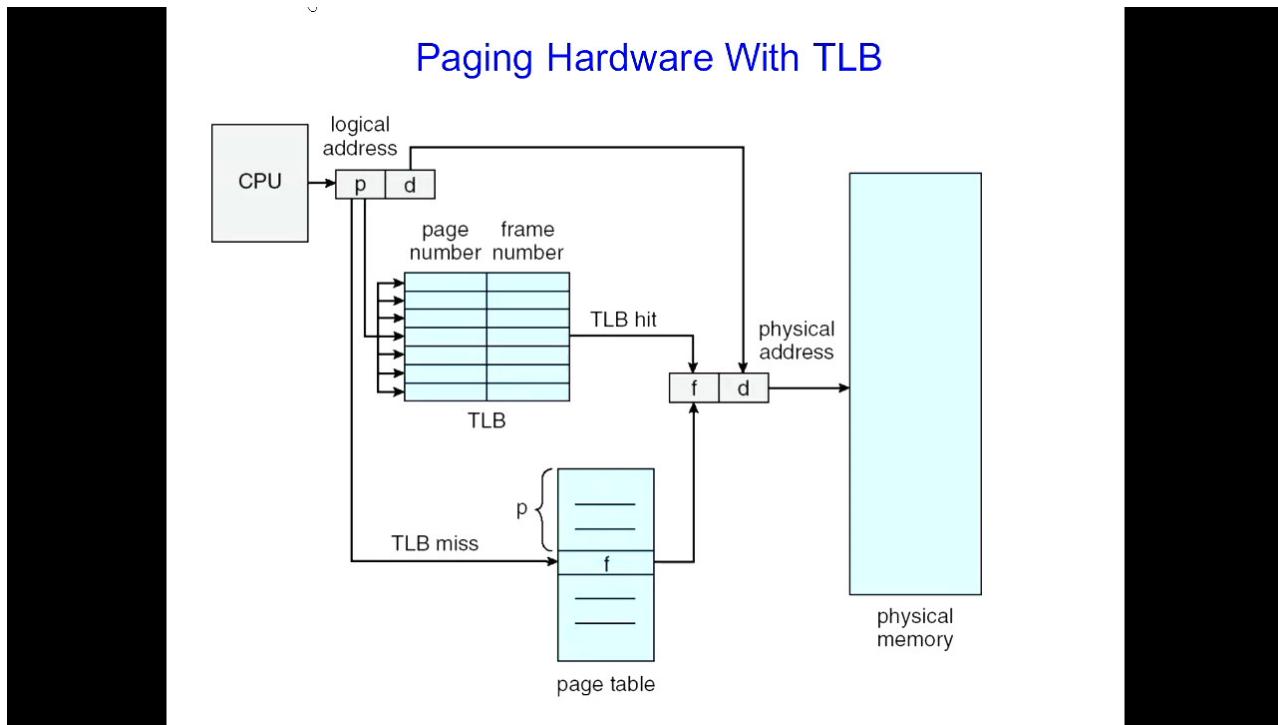
- Page table is kept in main memory
- Page-table base register (PTBR) points to the page table
- Page-table length register (PTLR) indicates the size of the page table
- In this scheme, every data/instruction access requires two memory access
 - one for the page table, and one for the data/instruction
- The two memory access problem can be solved
 - by the use of a special fast-lookup hardware cache
 - called *associative memory* or *translation look-aside buffers* (TLBs)
- Some TLBs store *address-space identifiers* (ASIDs) in each TLB entry
 - uniquely identifies each process;
 - provides address-space protection for that process;
 - otherwise need to flush at every context switch.

- TLBs typically small (64–1,024 entries)
- On a TLB miss, value is loaded into the TLB for faster access next time.
 - Replacement policies must be considered.
 - Some entries can be *wired down* for permanent fast access.

10.3 TLB

Associative Memory

- Associative memory — parallel search
- Address translation (p, d)
 - if p is in associative register, get frame# out;
 - otherwise get frame# from page table in memory.



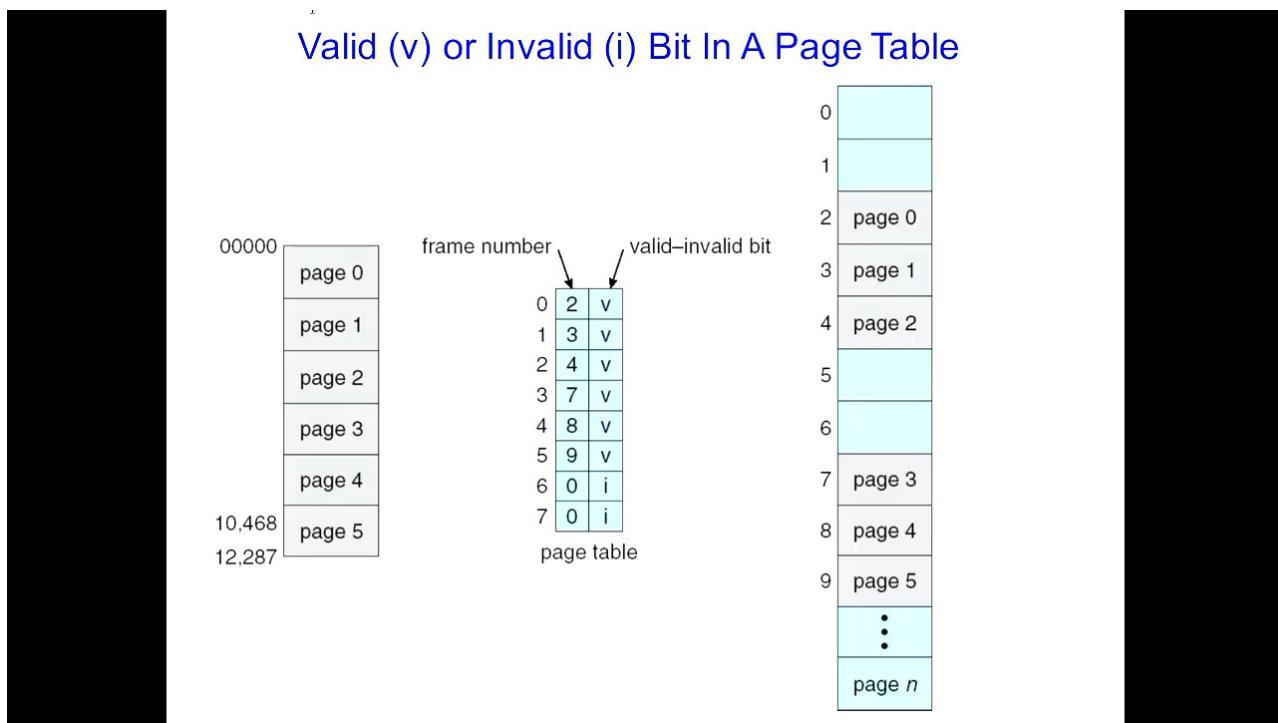
Effective Access Time

- Associative lookup
 - Extremely fast
- Hit ratio = α
 - Hit ratio — percentage of times that a page number is found in the associative memory;
- Consider $\alpha = 80\%$, 100ns for memory access
 - EAT = $0.80 \times 100 + 0.20 \times 200 = 120\text{ns}$
- Consider hit ratio $\alpha = 99$, 100ns for memory access
 - EAT = $0.99 \times 100 + 0.01 \times 200 = 101\text{ns}$

Memory Protection

- Memory protection implemented
 - by associating protection bit with each frame

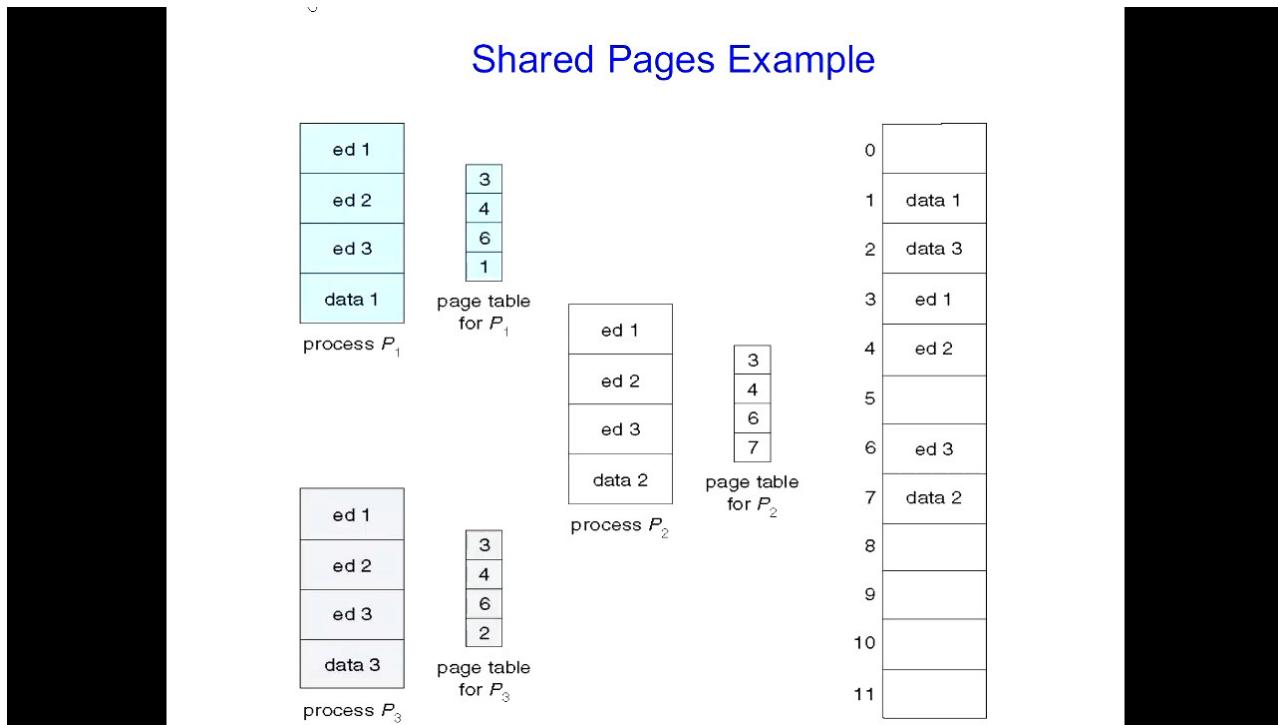
- to indicate if read-only or read-write access is allowed.
- Can also add more bits to indicate page execute-only, and so on.
- *Valid-invalid* bit attached to each entry in the page table:
 - *valid* indicates that the associated page
 - * is in the process' logical address space, and is thus a legal page
 - *invalid* indicates that the page
 - * is not in the process' logical address space.
 - Or use Page-Table Length Register (PTLR)
 - Page Table Entries (PTEs) can contain more information.
- Any violations result in a trap to the kernel.



10.4 Shared Pages

Shared Pages

- *Shared Code*
 - One copy of read-only (*re-entrant*) code shared among processes (i.e. text editors, compilers, window systems).
 - Similar to multiple threads sharing the same process space.
 - Also useful for interprocess communication if sharing or read-write pages is allowed.
- *Private code and data*
 - Each process keeps a separate copy of the data.
 - The pages for the private code and data can appear anywhere in the logical address space.



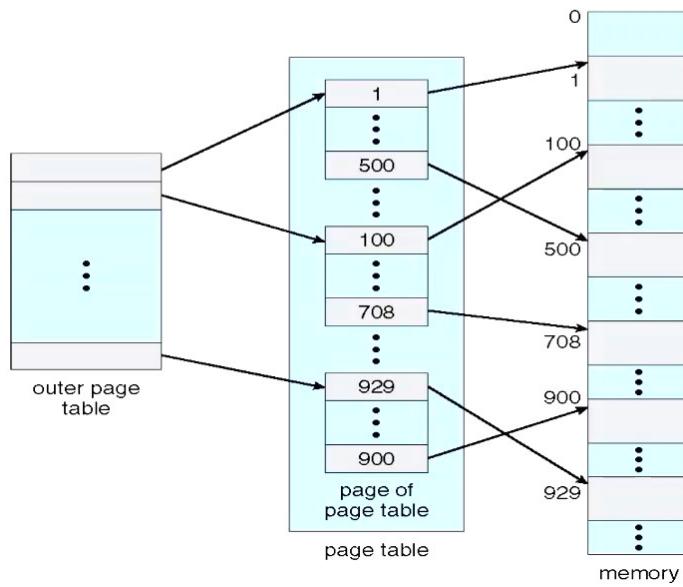
Structure of the Page Table

- Memory structures for paging can get huge using straightforward methods
 - Consider 32-bit logical address space as on modern computers
 - Page size of 4KB (2^{12})
 - Page table would have 1 million entries ($\frac{2^{32}}{2^{12}}$)
 - If each entry is 4 bytes \Rightarrow 4MB of physical address space / memory for page table alone.
 - * That amount of memory used to cost a lot
 - * Don't want to allocate that contiguously in main memory.
- Hierarchical paging
- Hashed page tables
- Inverted page tables

10.5 Hierarchical Pages

- Break up the logical address space into multiple page tables
- A simple technique is a two-level page table
- We then page the page table

Two-Level Page-Table Scheme



Two-level paging example

- A logical address (on 32-bit machines with 1K page size) is divided into:
 - a page number consisting of 22 bits
 - a page offset consisting of 10 bits
- Since the page table is paged, the page number is further divided into:
 - a 12-bit page number
 - a 10-bit page offset
- Thus a logical address is as follows

page number	page offset
p_1	p_2
12	10

where p_1 is an index into the outer page table, and p_2 is the displacement within the page of the inner page table.

- Known as *forward-mapped page table*.

64-bit logical address space

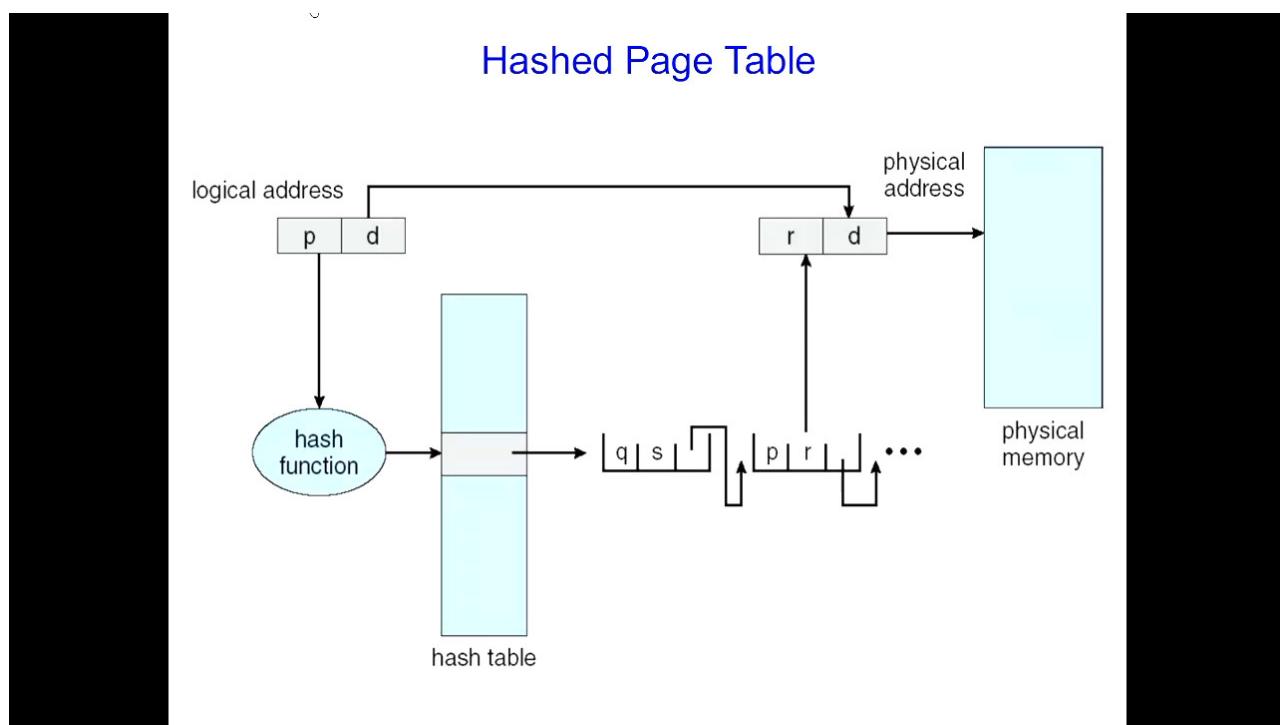
- Even two-level paging scheme not sufficient
- If page size is 4KB (2^{12})
 - Then page table has 2^{52} entries
 - If two level scheme, inner page tables could be 2^{10} 4-byte entries
 - Addresses would look like

outer page	inner page	page offset
p_1	p_2	d
42	10	12

- Outer page table has 2^{42} entries or 2^{44} bytes
- One solution is to add a second outer page table
- But in the following example, the second outer page table is still 2^{34} bytes in size,
 - * And possibly 4 memory access to get one physical memory location.

10.6 Hashed Pages

- Common in address spaces > 32 bits
- The virtual page number is hashed into a page table
 - This page table contains a chain of elements hashing to the same location
- Each element contains
 1. the virtual page number
 2. the value of the mapped page frame
 3. a pointer to the next element
- Virtual page numbers are compared in this chain searching for a match
 - If a match is found, the corresponding physical frame is extracted.
- Variation for 64-bit addresses is *clustered page tables*
 - Similar to hashed but each entry refers to several pages (such as 16) rather than 1.
 - Especially useful for *sparse* address spaces (where memory references are non-contiguous and scattered).



10.7 Inverted Pages

- Rather than each process having a page table and keeping track of all possible logical pages,
 - track all physical pages.
- One entry for each real page of memory.

- Entry consists of
 - the virtual address of the page stored in that real memory location,
 - information about the process that owns that page.
- Decreases memory needed to store each page table
 - but increases time needed to search the table when a page reference occurs.
- Use hash table to limit the search to one/few page-table entries
 - TLB can accelerate access.
- But how to implement shared memory?
 - One mapping of a virtual address to the shared physical address.

10.8 Uses

Functionality enhanced by page tables

- Code (instructions) is read-only
 - A bad pointer can't change the program code
- Dereferencing a null pointer is an error caught by hardware
 - Don't use the first page of the virtual address space — mark it as invalid — so references to address 0 cause an interrupt.
- Inter-process memory protection
 - My address XYZ is different to your address XYZ
- Shared libraries
 - All running C programs use libc
 - Have only one (partial) copy in physical memory, not one per process
 - All page table entries mapping libc point to the same set of physical frames
 - * DLLs in Windows
- Generalising the use of “shared memory”
 - Regions of two separate processes' address spaces map to the same physical frames
 - Faster inter-process communication
 - * just read/write from/to shared memory Don't have to make a syscall
 - Will have separate Page Table Entries (PTEs) per process, so can give different processes different access rights
 - * E.g. one reader, one writer.
- Copy-on-write (CoW), e.g. on `fork()`
 - Instead of copying all pages, create shared mappings of parent pages in child address space
 - * Make shared mappings read-only for both processes
 - * When either process writes, fault occurs, OS “splits” the page

Less familiar uses

- Memory-mapped files

- instead of using open, read, write, close
 - * “map” a file into a region of the virtual address space
 - e.g. into region with base X
 - * accessing virtual address $X + N$ refers to offset N in file
 - * initially, all pages in mapped region marked as invalid
- OS reads a page from file whenever invalid page accessed
- OS writes a page to file when evicted from physical memory
 - * only necessary if page is dirty.

More unusual use

- Use “soft faults”
 - faults on pages that are actually in memory
 - but whose PTE entries have artificially been marked as invalid
- That idea can be used whenever it would be useful to trap on a reference to some data item.
 - Example: debugger watchpoints
- Limited by the fact that the granularity of detection is the page.

11 Virtual Memory

11.1 Virtual memory

Paged virtual memory

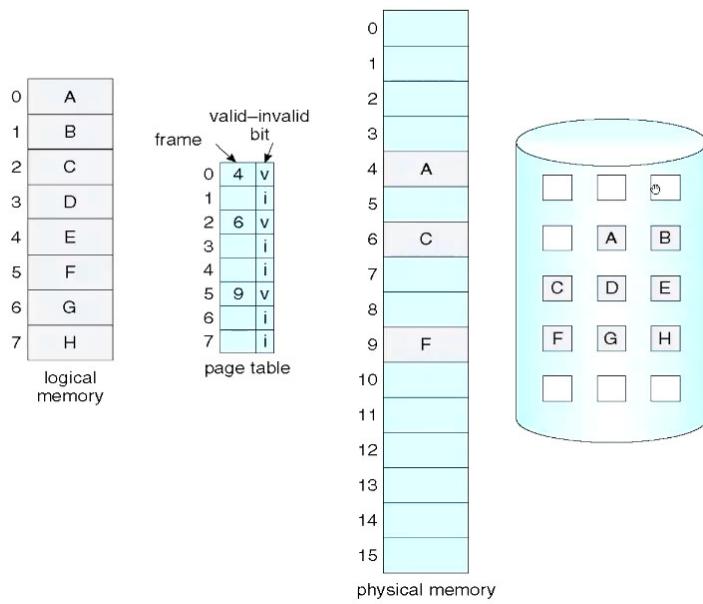
- Allows a larger logical address space than physical memory
- All pages of address space do not need to be in memory
 - the full (used) address space on disk in page-sized blocks
 - main memory used as a (page) cache
- Needed page transferred to a free page frame
 - if no free page frames, evict a page
 - * evicted pages go to disk only if *dirty*
 - Transparent to the application, except for performance
 - Managed by hardware and OS
- Traditionally called *paged virtual memory*

11.2 Page Fault

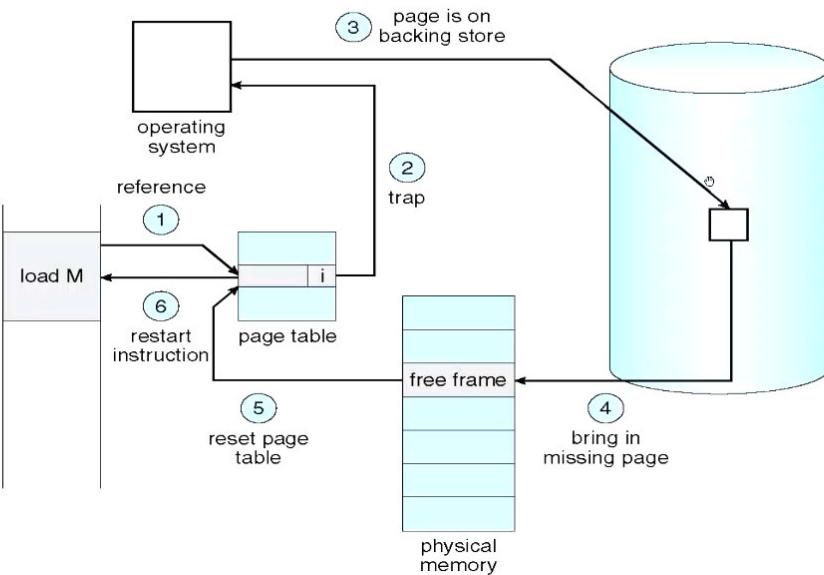
Page Fault

- If there is a reference to a page, first reference to that page will trap to OS:
page fault
1. OS looks at another table to decide
 - invalid reference \implies abort
 - just not in memory
 2. Find free frame
 3. Swap page into frame via scheduled disk operation
 4. Reset tables to indicate page now in memory
Set validation bit = v
 5. Restart the instruction that caused the page fault

Page Table When Some Pages Are Not in Main Memory



Steps in Handling a Page Fault



11.3 Demand Paging

Demand paging

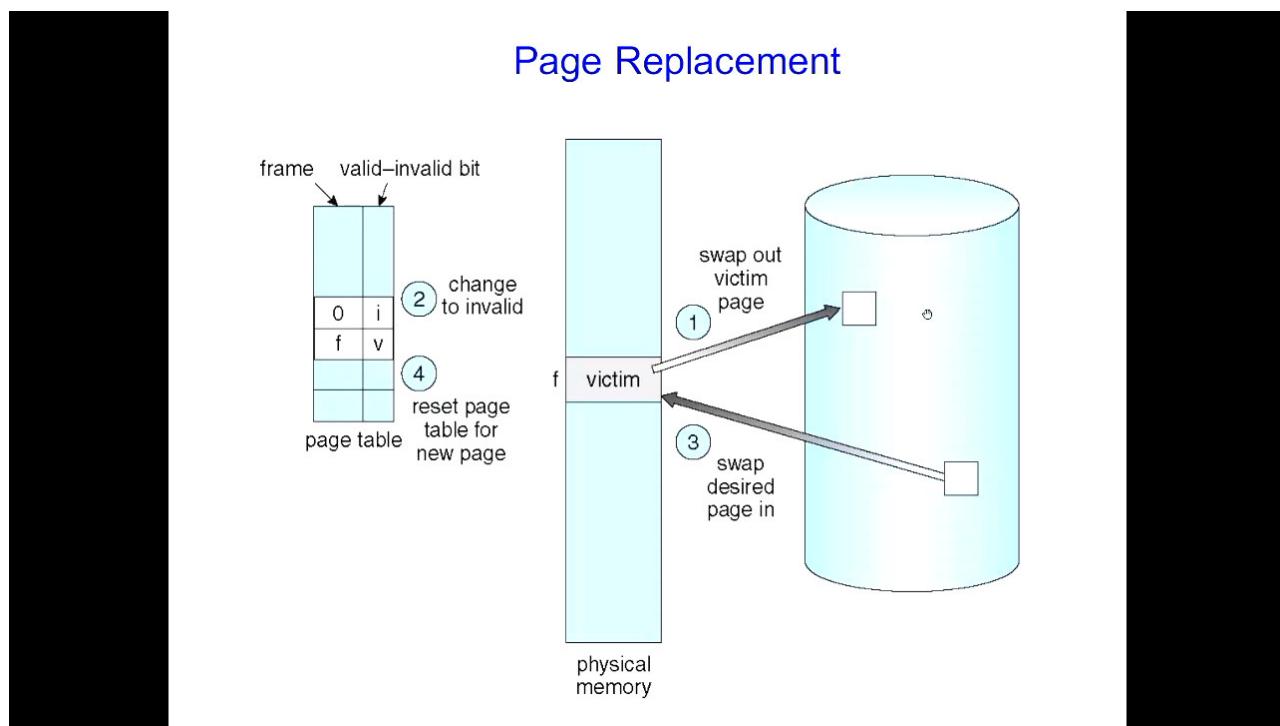
- Pages only brought into memory when referenced
 - Only code/data that is needed by a process needs to be loaded
 - * What's needed changes over time
 - Hence, it's called *demand paging*
- Few systems try to anticipate future needs
- But sometimes cluster pages

- OS keeps track of pages that should come and go together
- bring in all when one is referenced
- interface may allow programmer or compiler to identify clusters

11.4 Page replacement

Page replacement

- When you read in a page, where does it go?
 - if there are free page frames, grab one
 - if not, must evict something else
 - * this is called *page replacement*
- Page replacement algorithms
 - try to pick a page that won't be needed in the near future
 - try to pick a page that hasn't been modified (thus saving the disk write)
- OS tries to keep a pool of free pages around
 - so that allocations don't inevitably cause evictions
- OS tries to keep some “clean” pages around
 - so that even if you have to evict a page, you won't have to write it



Evicting the best page

- The goal of the page replacement algorithm:
 - reduce fault rate by selecting best victim page to remove
 - the best page to evict is one that will never be touched again
 - Belady's proof:

- * evicting the page that won't be used for the longest period of time minimises the page fault rate
- Examine *page replacement algorithms*
 - assume that a process pages against itself
 - using a fixed number of page frames
- Number of frames available impacts page fault rate
 - Note Belady's anomaly

11.5 Page replacement algorithms

First-In-First-Out (FIFO) Algorithm

- Not always the best page replacement behaviour.

Belady's Optimal Algorithm

- Replace page that will not be used for longest period of time
- How do you know this?
 - Can't predict the future
- Used for measuring how well your algorithm performs

Least Recently Used (LRU) algorithm

- Use past knowledge rather than future
- Replace page that has not been used in the most amount of time
- Associate time of last use with each page
- 12 pages — better than FIFO but works than Belady's/OPT
- Generally good algorithm and frequently used
- But how to implement?

Approximating LRU

- Many approximations, all use the PTE's referenced bit
 - keep a counter for each page
 - at some regular interval, for each page, do:
 - * if reference bit = 0, increment the counter (hasn't been used)
 - * if reference bit = 1, zero the counter (hasn't been used)
 - * regardless, zero bit ref
 - the counter will contain the number of intervals since the last reference to the page
 - * page with largest counter is least recently used
- Some architectures don't have PTE reference bits
 - can simulate reference bit using the valid bit to induce faults

Second-chance clock

- Not Recently Used (NRU) or Second Change
 - replace page that is “old enough”

- logically, arrange all physical page frames in a big circle (clock)
 - * just a circular linked list
- A “clock hand” is used to select a good LRU candidate
 - sweep through the pages in circular order like a clock
- If reference bit is off, it hasn’t been used recently, and we have a victim
- If reference bit is on, turn it off and go to next page
 - arm moves quickly when pages are needed.

Allocation of frames among processes

- FIFO and LRU Clock each can be implemented as either *local* or *global* replacement algorithms
 - local
 - * each process is given a limit of pages it can use
 - * it “pages against itself” (evicts its own pages)
 - global
 - * the “victim” is chosen from among all page frames, regardless of owner
 - * processes’ page frame allocation can vary dynamically
- Issues with local replacement?
 - poor utilisation of free page frames, long access time
- Issues with global replacement
 - Linux uses global replacement: global thrashing.

11.6 Working set

The *working set model of program behaviour*

- *Working set* of a process is used to model the dynamic locality of its memory usage
 - works set = set of pages a process currently “needs”
 - formally defined by Peter Denning in the 1960s.
- Definition:
 - $WS(t, w) = \{ \text{pages } P \text{ such that } P \text{ was referenced in the time interval } (t, t - w) \}$
 - * t time
 - * w working set *window* (measured in page references)
 - * a page in WS only if it was referenced in the last w references
- Working set varies over the life of the program
 - so does the *working set size*

Working set size

- The working set size, $|WS(t, w)|$
 - changes with program locality
- During periods of poor locality

- more pages are referenced
- Within that period of time
 - the working set size is larger
- Intuitively, the working set must be in memory
 - otherwise you'll experience heavy faulting
 - *thrashing*

Hypothetical Working Set Algorithm

- Estimate $|\text{WS}(0, w)|$ for a process
 - Allow that process to start only if you can allocate it that many page frames
- Use a local replacement algorithm (LRU Clock?)
 - make sure that the working set are occupying the process's frames
- Track each process's working set size,
 - and re-allocate page frames among processes dynamically
- Problem
 - keep track of working set size
- Use reference bit with a fixed-interval timer interrupt.

Working Sets and Page Fault Rates

- Direct relationship between working set of a process and its page-fault rate
- Working set changes over time
- Peaks and valleys over time

Page-Fault frequency

- More direct approach than WSS
- Establish “acceptable” *page-fault frequency (PFF)* rate and use local replacement policy
 - If actual rate too low, process loses frame
 - If actual rate too high, process gains frame

11.7 Thrashing

Thrashing

- Thrashing
 - when the system spends most of its time servicing page faults, little time doing useful work
- Could be that there is enough memory
 - but a poor replacement algorithm — incompatible with program behaviour
- Could be that memory is over-committed
 - OS sees CPU poorly utilised and adds more processes
 - * too many active processes
 - Makes problem works

12 File Systems

12.1 Roles

Primary Roles of the OS (file system)

1. Hide hardware specific interface
2. Allocate disk blocks
3. Check permissions
4. Understand directory file structure
5. Maintain *metadata*
6. Performance
7. Flexibility

File systems

- The concept of a file system is simple
- The implementation of the abstraction for secondary storage
 - abstraction = files
- Logical organisation of files into directories
 - the directory hierarchy
- Sharing of data between
 - processes, people, and machines
- Access control, consistency, etc.

12.2 Files

Files

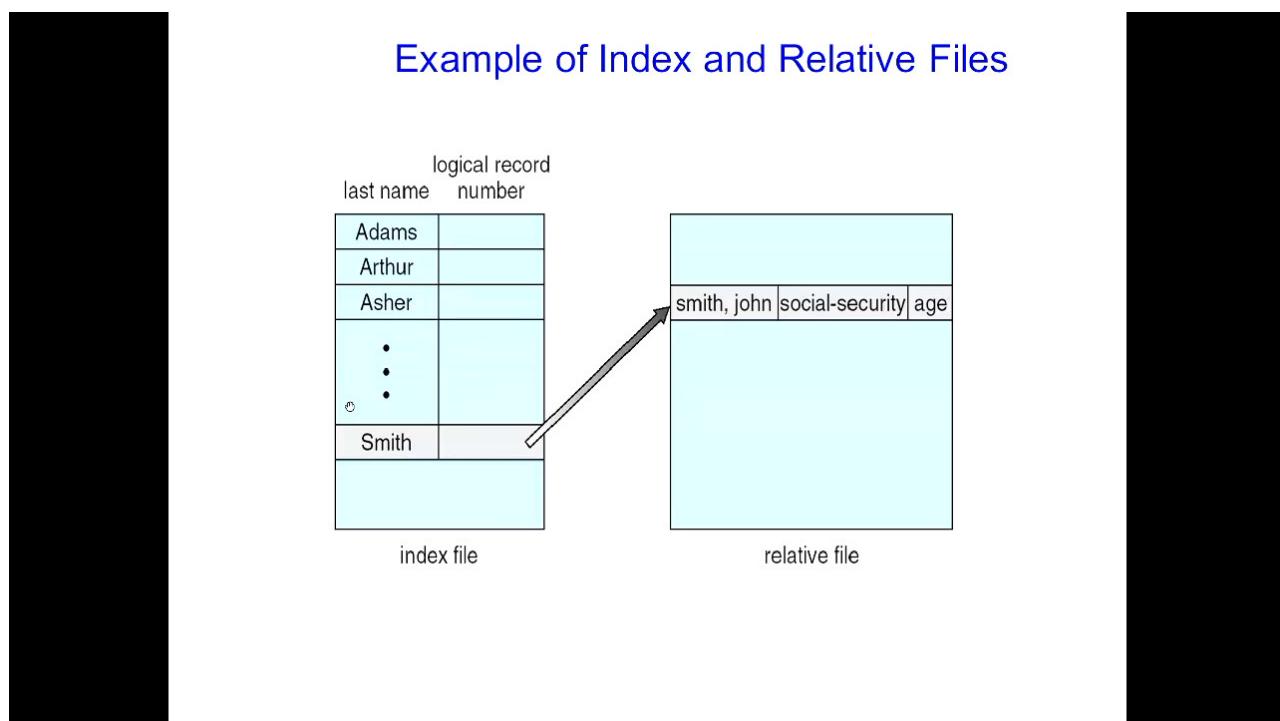
- A collection of data with some properties
 - contents, size, owner, last read/write time, protection, etc.
- File types
 - understood by file system
 - * device, directory, symbolic link
 - understood by other parts of OS or by runtime libraries
 - * executable, dll, source code, object code, text file, etc.
- Type can be encoded in the file's name or contents
 - windows encodes type
 - * com, exe, bat, dll, jpg, mov, mp3, ...
- Old Mac OS stored the name of the creating program along with the file
- Unix — initial characters (e.g.)

Basic operations

<i>Unix</i>	<i>Windows</i>
create (name)	CreateFile (name, CREATE)
open (Name, mode)	CreateFile (name, OPEN)
read (fd, buf, len)	ReadFile (handle, ...)
write (fd, buf, len)	WriteFile (handle, ...)
sync (fd)	FlushFileBuffers (handle, ...)
seek (fd)	SetFilePointer (handle, ...)
close (fd)	CloseHandle (handle, ...)
unlink (name)	DeleteFile (name)
rename (old, new)	CopyFile (name) MoveFile (name)

File access methods

- File systems provide different *access methods*
 - sequential access
 - * read bytes one at a time, in order
 - direct access
 - * random access given a block/byte number
 - record access
 - * file is array of fixed- or variable-sized records
 - indexed access
 - * one file contains an index to a record in another file
 - * apps can find a file based on the value in that record (similar to DB)
- Why do we care about distinguishing sequential from direct access?
 - what might the FS do differently in these cases?



12.3 Directories

Directories

- Directories provide:
 - a way for users to organise their files
 - a convenient file name-space for both users and file-systems
- Most file-systems support multi-level directories
 - naming hierarchies (`/`, `/usr`, `/usr/local`, `/usr/local/bin`, ...)
- Most file-systems support the notion of current directory
- Absolute names: fully-qualified starting from root of file-system
- Relative names: specified with respect to current directory

Directory internals

- A directory is typically just a file that happens to contains special meta-data
- Directory
 - list of (name of file, file attributes)
- Attributes include such things as:
 - size, protection, location on disk, creation time, access time, ...
- The directory list is usually unordered (effectively random),
 - the `ls` command sorts the results for you

Path name translation

- You want to open `/one/two/three`

```
fd = open("one/two/three", O_RDWR)
```
- Inside the file system
 - open directory `/` (well known, can always find)
 - search the directory for `one`, get location of `one`
 - open directory `one`, search for `two`, get location of `two`
 - open directory `two`, search for `three`, got location of `three`
 - open file `three`
 - (of course, permissions are checked at each step)
- File-system spends much time walking down directory paths
 - this is why open is separate from read/write (session state)
 - OS will cache prefix lookups to enhance performance
 - * `/a/b`, `/a/bb`, `/a/bbb` all share the `/a` prefix

12.4 File Protection

- FS implements a protection system
 - to control who can access a file (user)
 - to control how they can access it (e.g. read, write, or exec)
- Often generalised
 - generalise file to *objects* (the “what”)
 - generalise users to *principals* (the “who”, user or program)
 - generalise read/write to *actions* (the “how,” or operations)
- Protection system dictates
 - whether a given action
 - performed by a given principal
 - on a given object should be allowed
- E.g. you can read or write your files, but others cannot
- E.g. you can read `/group/teaching/cs3` but you cannot write to it

Model for representing protection

- Two different ways of thinking about it:
 - access control lists (ACLs)
 - * for each object, keep lists of principals and principals’ allowed actions
 - capabilities
 - * for each principal, keep lists of objects and principals’ allowed actions
- Both can be represented with the following matrix:

	<code>/etc/passwd</code>	<code>/home/gribble</code>	<code>/home/guest</code>
<code>root</code>	<code>rw</code>	<code>rw</code>	<code>rw</code>
<code>gribble</code>	<code>r</code>	<code>rw</code>	<code>r</code>
<code>guest</code>			<code>r</code>

ACLs v.s. Capabilities

- Capabilities are easy to transfer
 - they are like keys: can hand them off
 - they make sharing easy
- ACLs are easier to manage
 - object-centric, easy to grant and revoke
 - * to revoke capability, need to keep track of principals that have it
 - * hard to do, given that principals can hand off capabilities
- ACLs grow large when object is heavily shared
 - can simplify by using *groups*
 - * put users in groups, put groups in ACLs
- Additional benefit

- change group membership, affects *all* objects that have this group in its ACL

12.5 Unix

The original Unix file system

- Dennis Ritchie and Ken Thompson, Bell Labs, 1969
- “Unix rose from the ashes of a multi-organisational effort in the early 1960s to develop a dependable timesharing operating system” — Multics
- Designed for a “workgroup” sharing a single system
- Did its job well
 - Although it has been stretched in many directions

(Old) Unix disks are divided into five parts...

1. Boot block
can boot the system by loading from this block
2. Superblock
specifies boundaries of next 3 areas, and contains head of freelists of inodes and file blocks
3. inode area
contains descriptors (inodes) for each file on the disk; all inodes are the same size; head of freelist is in the superblock.
4. File contents area
fixed-size blocks; head of freelist is in the superblock
5. Swap area
holds processes that have been swapped out of memory

Basic file system structures

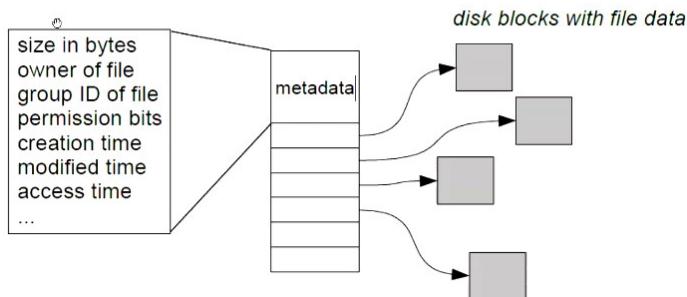
- Every file and directory is represented by an inode
 - stands for “index node”
- Contains two kinds of information
 1. Metadata describing the file’s owner, access rights, etc.
 2. Location of the file’s blocks on disk

inode format

- User number
- Group number
- Protection bits
- Times (file last read, file last written, inode last written)
- File code: specifies if the inode represents a directory, and ordinary user file, or a “special” file (typically an I/O device)
- Size: length of file in bytes
- Block list: locates contents of file (in the file contents area)
- Link count: number of directories referencing this inode

Basic file system structures

- Every file and directory is represented by an inode
 - Stands for “index node”
- Contains two kinds of information:
 - 1) Metadata describing the file's owner, access rights, etc.
 - 2) Location of the file's blocks on disk



19

The flat (inode) file system

- Each file is known by a number, which is the number of the inode
 - 1, 2, 3, etc.!
- Files are created empty, and grow when extended through writes

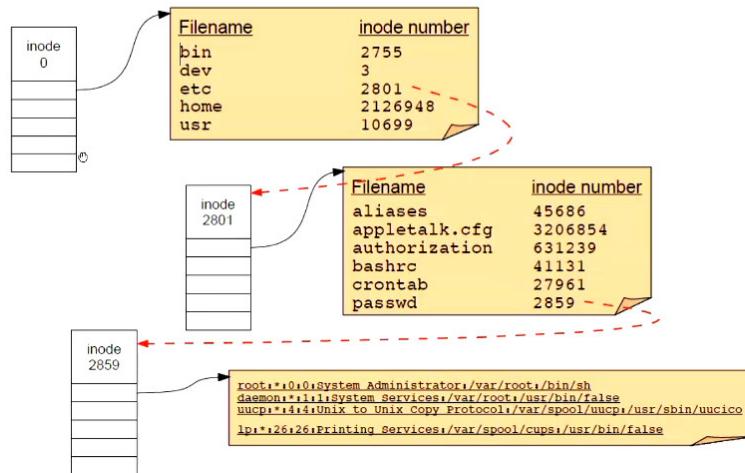
The tree (directory, hierarchical) file system

- A directory is a flat file of fixed-size entries
- Each entry consists of an inode number and a file name
 - These are the contents of the directory “file data” itself — *not* the directory’s inode
 - Filenames (in Unix) are not stored in the inode at all.
- Special inodes:
 - inode 2 is the root directory
 - inode 1 — a hidden file containing all bad blocks

Pathname resolution

Pathname resolution

- To look up a pathname “/etc/passwd”, start at root directory and walk down chain of inodes...



23

Locating inodes on disk

- Directories give the *inode number* of a file.
 - How do we find the inode itself on disk?
 - Basic idea: Top part of filesystem contains *all* of the inodes
 - inode number is just the “index” of the inode
 - Easy to compute the block address of a given inode:

```
block_addr(inode_num) = block_offset_of_first_inode + (inode_num * inode_size)
```

- This implies that a filesystem has a fixed number of potential inodes
 - this number is generally set when the filesystem is created
 - The superblock stores important metadata on filesystem layout, list of free blocks, etc.

Directory issues

- Directories map filenames to inode numbers
 - We can create multiple pointers to the same inode in different directories
 - or even the same directory with different filenames.
 - In Unix, this is called a *hard link* and can be done using `ln`
 - `/home/foo` and `/tmp/foo` now refer to the same file on disk

How should we organise blocks on a disk?

- Very simple policy: A file consists of linked blocks
 - inode points to the first block of the file;
 - each block points to the next block in the file (just a linked list on disk)
 - Indexed files

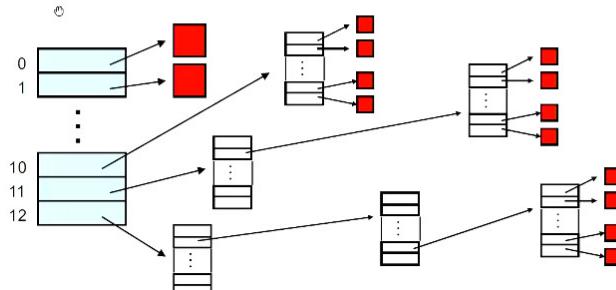
- inode contains a list of block numbers containing the file:
- array is allocated when the file is created.

The “block list” portion of the inode (Unix Version 7)

- Points to blocks in the file contents area
- Must be able to represent very small and very large files.
- Each inode contains 13 block pointers
 - first 10 are *direct pointers* (pointers to 512B blocks of file data)
 - then single, double, and triple indirect pointers

The “block list” portion of the inode (Unix Version 7)

- Points to blocks in the file contents area
- Must be able to represent very small and very large files.
- Each inode contains 13 block pointers
 - first 10 are “direct pointers” (pointers to 512B blocks of file data)
 - then, single, double, and triple indirect pointers



27

File Size

- Only occupies $13 \times 4B$ in the inode
- Can get to $10 \times 512B = a 5120B$ file directly
 - (10 direct pointers, blocks in the file contents area are 512B)
- Can get to $128 \times 512B =$ an additional 65KB with a single indirect reference
 - (the 11th pointer in the inode gets you to a 512B block in the file contents area that contains 128 4B pointers to blocks holding file data)
- Can get to $128 \times 128 \times 512B =$ an additional 8MB with a double indirect reference
 - (the 12th pointer in the inode gets you to a 512B block in the file contents area that contains 128 4B pointers to 512B blocks in the file contents area that contains 128 4B pointers to 512B blocks holding file data)
- Can get to $128 \times 128 \times 128 \times 512B =$ an additional 1GB with a triple indirect reference
- Maximum file size is a little over 1GB+

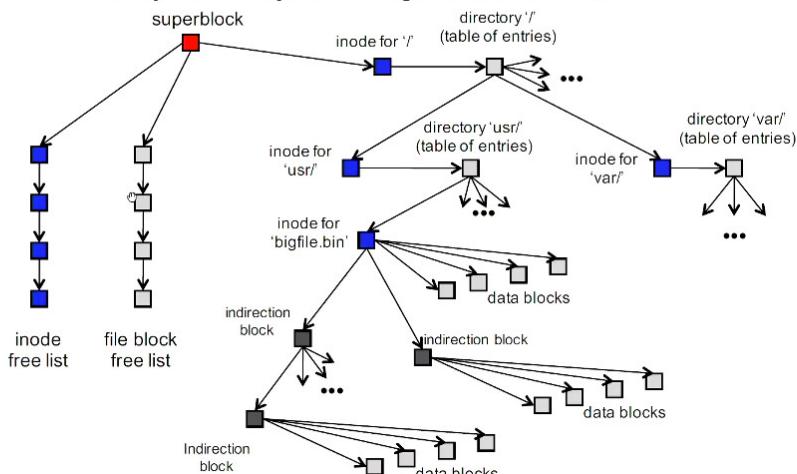
Update File Size

- A later version of Bell Labs Unix utilised 12 direct pointers, rather than 10

- Berkeley Unix went to 1KB block sizes
 - What's the effect on the maximum file size?
* $256 \times 256 \times 256 \times 1\text{K} = 17\text{GB} + \text{a smidge}$
- Subsequently went to 4KB blocks
 - $1\text{K} \times 1\text{K} \times 1\text{K} \times 4\text{K} = 4\text{TB}$

Putting it all together

- The file system is just a huge data structure



31

File system layout

- One important goal of a filesystem is to lay this data structure out on disk
 - keep in mind the physical characteristics of the disk
 - seeks are expensive
 - characteristics of the workload
 - * locality across files within a directory
 - * sequential access to many files
- Old Unix's layout is very inefficient
 - constantly seeking
 - * between inode area and data block area
 - * as traverse filesystem, or sequentially read files
- Newer file systems are smarter
- Newer storage devices (SSDs) change the constraints

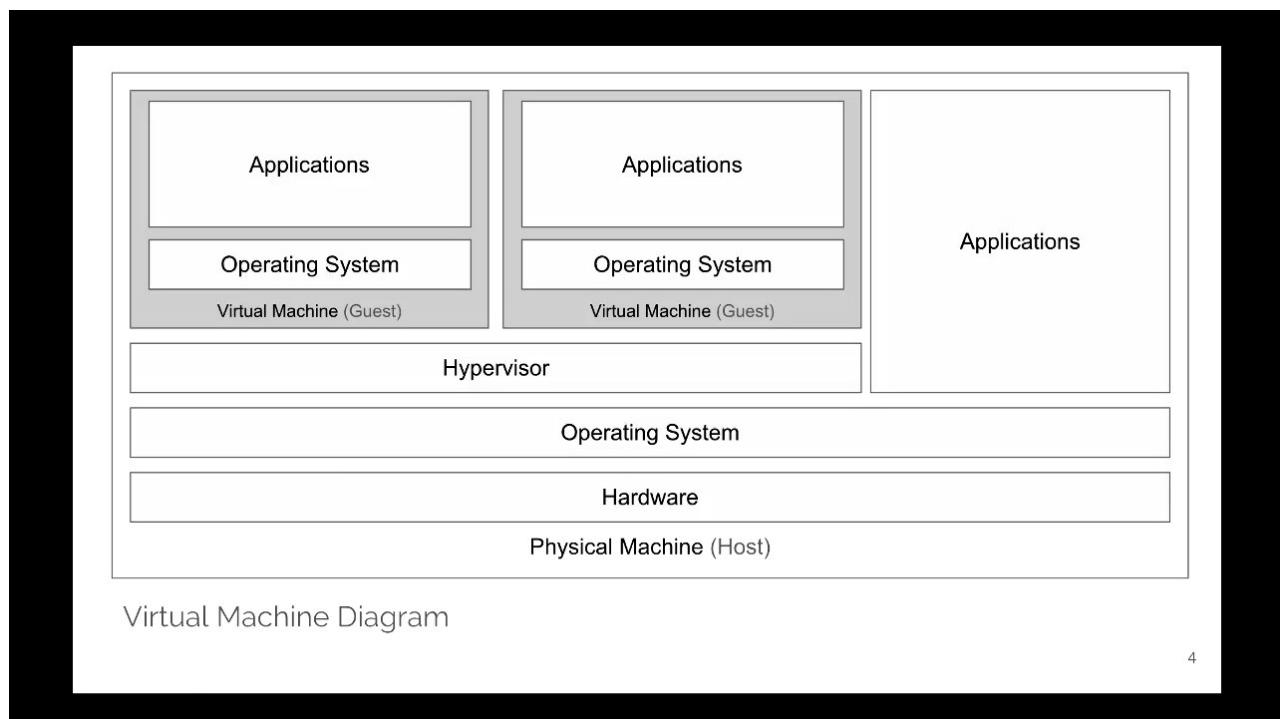
13 Virtualisation

Introduction

- Virtualisation is the process of creating a *virtual* version of a *physical* object.
- In computing, *hardware virtualisation* is the process of creating a virtual version of real hardware.
- This virtual hardware can be used to run a complete OS.

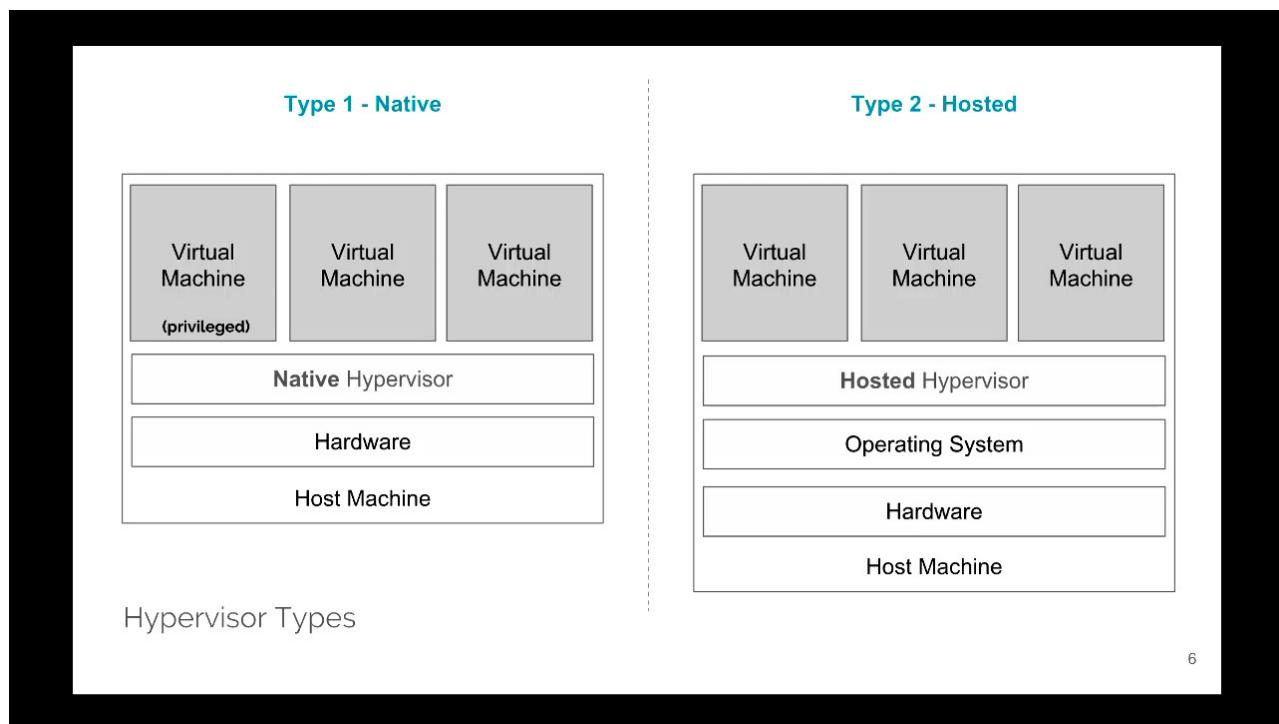
Terminology

- *Virtual Machine*: A *virtual* representation of a *physical* machine
 - Not to be confused with a *Java Virtual Machine* or the *CLR* (.NET)
- *Virtual Machine Monitor* or *Hypervisor*: A software application that monitors and manages running virtual machines.
- *Host Machine*: The *physical* machine that a *virtual* machine is running on.
- *Guest Machine*: The *virtual* machine, running on the *host* machine.



Virtual Machine Monitor (Hypervisor)

- The VMM is in charge of running the *virtual machines*.
- There are two main types of VMM:
 - Type 1: *Native*
run directly on the host machine, and share out resources (such as memory and devices) between guest machines. (E.g. XEN, Oracle VM Server.)
 - Type 2: *Hosted*
run as an application inside an OS, and support virtual machines running as individual processes. (E.g. VirtualBox, Parallels Desktop, QEMU.)



Uses of Virtualisation

- *Personal* (e.g. Parallels Desktop / VirtualBox)
 - Running multiple OSs on one host, without the inconvenience of rebooting.
 - E.g. running Windows inside OS X.
 - Some hypervisors support *seamless integration*.
- *Technical* (e.g. QEMU as used in the coursework)
 - OS / Hardware Design
 - Kernel Debugging/Testing
 - Prototyping new architectures / architectural features
- *Commercial* (e.g. XEN/VMWare)
 - Data centre server consolidation
 - High availability / Migration

Types of Virtualisation

- Software emulation
 - Maximum flexibility for virtualisation, but very slow to run (high overhead).
 - Each guest instruction is emulated (can use *binary translation* for speed-up)..,
- Containers/Namespace
 - Isolate processes / groups of processes within a single OS, e.g. Docker.
- Full System or Hardware Virtualisation
 - Guest machine is the *same* architecture as the host machine, e.g. Intel x86 on Intel x86.
- Cross-architecture Virtualisation
 - Guest machine has a *different* architecture to the host machine, e.g. ARM on Intel X86.

- Must use *software emulation* to do this.

Popek and Goldberg Requirements for virtualisation

Paper published in 1974 that laid the foundations for hardware virtualisation, and formalised the requirements for an architecture to be *virtualisable*.

Three main properties for a virtual machine

1. Efficiency
 - The majority of guest instructions are executed directly on the host machine.
2. Resource control
 - The virtual machine monitor must remain in control of all machine resources.
3. Equivalence
 - The virtual machine must behave in a way that is indistinguishable from if it was running as a physical machine.

Efficiency

- “All innocuous instructions are executed by the hardware directly, with no intervention at all on the part of the control program.”
- Normal guest machine instructions should be executed directly on the processor. System instructions need to be emulated by the VMM.

Resource control

- “It must be impossible for that arbitrary program to affect the system resources, i.e. memory available to it; the allocator of the control program is to be invoked upon any attempt.”
- The virtual machine should not be able to affect the host machine in any adverse way. The host machine should remain in control of all physical resources, sharing them out to guest machines.

Equivalence

- “Any program K executing with a control program resident, with two possible exceptions, performs in a manner indistinguishable from the case when the control program did not exist and K had whatever freedom of access to privileged instructions that the programmer had intended.”
- A formal way of saying that the OS running on a virtual machine should believe it is running on a physical machine, i.e. the behaviour of the virtual machine (from the guest OS’ point of view) is identical to that of the corresponding physical machine.
- The two exceptions mentioned are:
 - *temporal latency* (some instruction sequences will take longer to run);
 - *resource availability* (physical machine resources are shared between virtual machines).

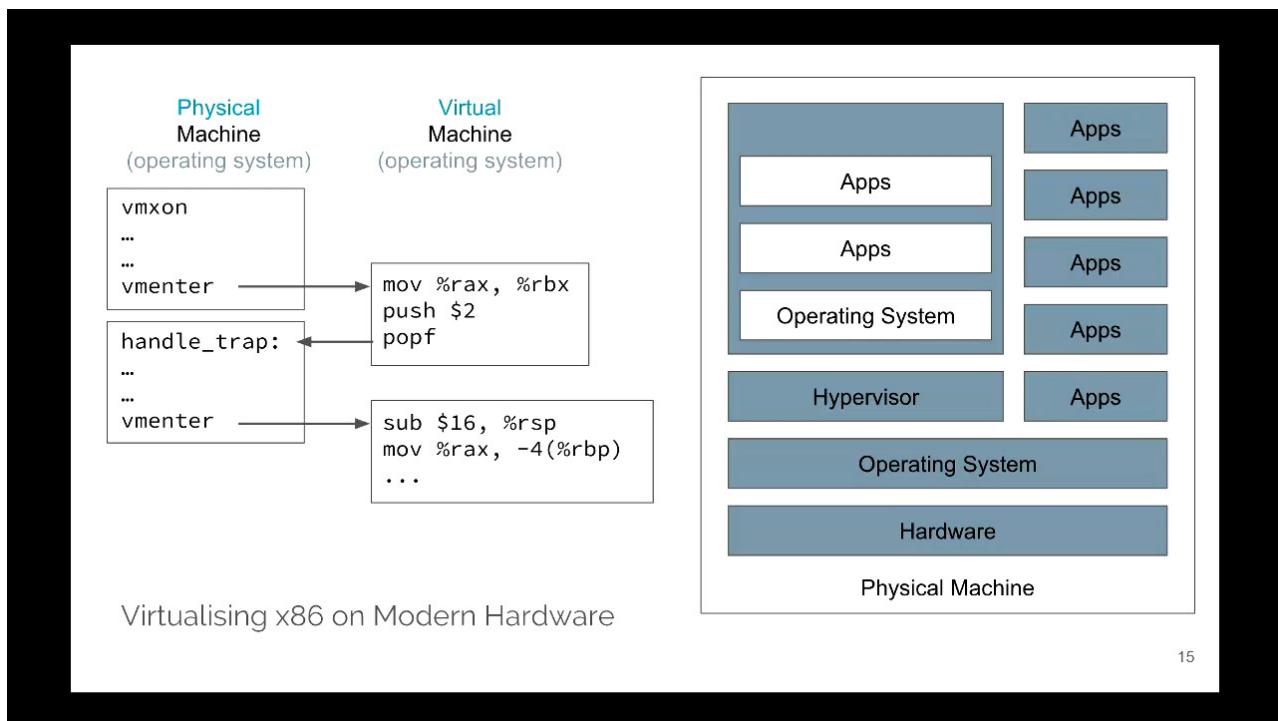
Methods of virtualisation

- Full software emulation
 - Not permitted by *Popek ad Goldberg* because it violates the *efficiency* property.
 - * Although, this no longer holds due to the advent of *efficient binary translation*.
 - Required for *cross-architecture virtualisation*, as guest instructions cannot execute natively on the host.
- Trap-and-emulate

- The guest OS runs *de-privileged*, all non-privileged instructions execute natively on the host.
- All privileged instructions *trap* to the VMM.
- VMM *emulates* these privileged operations.
- Guest resumes execution after emulation.

Virtualising x86

- Originally x86 was not “classically” virtualisable.
 - Some privileged instructions did not *trap*, and so could not be emulated correctly.
- *Interpretation* is too slow (violates *efficiency*)
- *Code Patching* leaves traces of virtualisation (violates *equivalency*)
- *Binary Translation* is better, but still incurs overhead.
- Since 2005, x86 processors now support virtualisation in hardware,
 - Intel-VT
 - AMD-V
- This enables *trap-and-emulate* style virtualisation.
- Unmodified OSs can run natively on host machines.



Hardware acceleration for virtualisation

- Modern processors include hardware support for running virtual machines
 - Intel VT-X and AMD-V for x86 processors.
 - ARM virtualisation extensions for ARM processors.
- Hardware extensions allow all guest instructions (including system instructions) to run natively on the processor.
- This works by providing an isolated view of the processor to virtual machines.

- OSs can then run *directly* on the processor, believing they are running on physical hardware.
- Certain privileged operations *trap* back to the hypervisor.

Virtual machine access to resources

- Virtual machines need to be given access to resources such as:
 - Memory
 - Storage
 - Networking
 - Graphics
- It is the responsibility of the VMM to share out these resources.
- Access to physical memory is managed by the VMM.
- For an *unmodified* OS, expecting a “real” storage device (e.g. a hard disk), the VMM must provide an *emulation* of that device.
- Some devices may be *passed straight through* to the virtual machine, e.g. dedicated *network cards*.

Paravirtualisation

- Guest OSs are *aware* they are being virtualised.
- They *co-operate* with the hypervisor to enable increased *memory* and *device* performance.
- They no longer *trap-and-emulate*, but instead request privileged operations directly from the hypervisor.
- They can *co-operate* with the hypervisor so that host memory can be more efficiently distributed. Instead of providing an *emulated storage device*, the hypervisor can provide a *paravirtualised* implementation.
- Typically used in *data centres* for large-scale virtualisation.

14 Secondary Storage

Secondary storage

- Secondary storage:
 - anything outside “primary memory”
 - direct execution of instructions/data retrieval via machine load/store
 - * not permitted
- Characteristics:
 - it’s large: 250–2000GB and more
 - it’s cheap: \$0.05/GB for hard drives
- Persistent: data survives power loss
 - it’s slow: milliseconds to access
- It *does* fail, if rarely
- Big failures:
 - drive dies: Mean Time Between Failure ~ 3 years
 - 100K drives and MTBF is 3 years
 - * that’s 1 “big failure” every 15 minutes!
- Little failures (read/write errors, one byte in 10^{13})

14.1 Disk trends

Disk trends

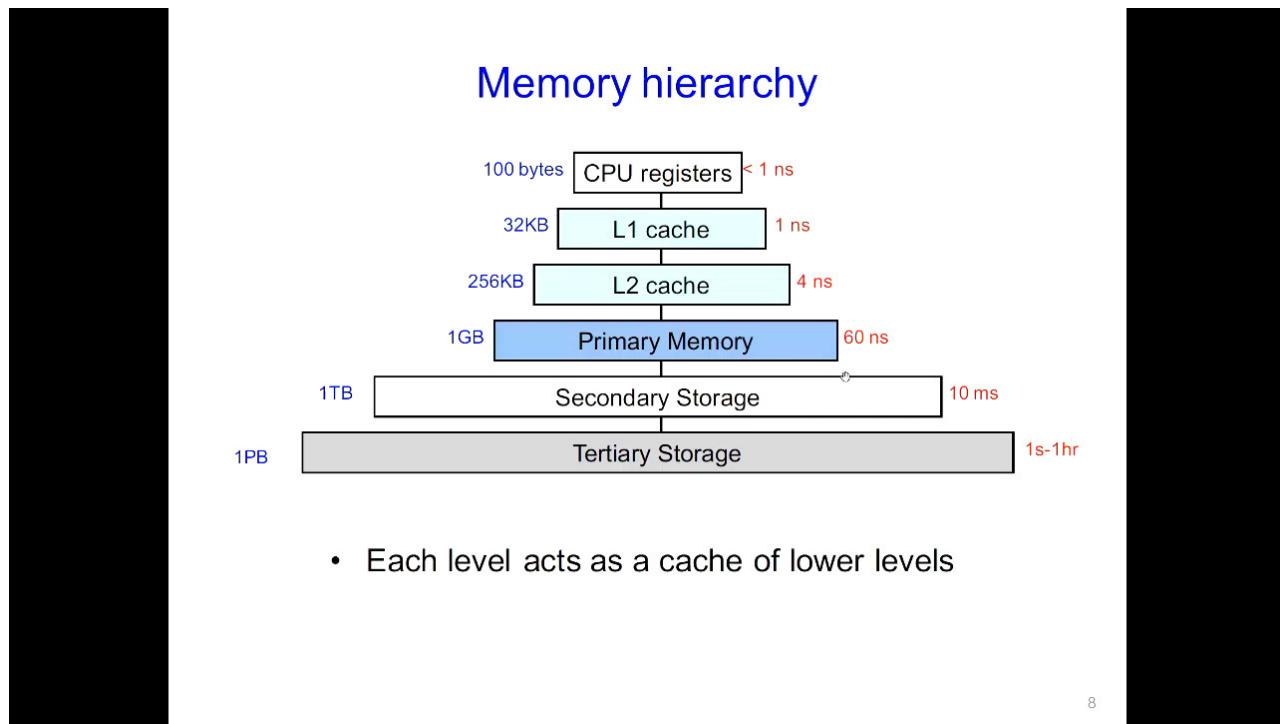
- Disk capacity, 1975–1989
 - doubled every 3+ years
 - 25% improvement each year
 - factor of 10 every decade
 - *Still exponential, but far less rapid than processor performance*
- Disk capacity, 1990–recently
 - doubling every 12 months
 - 100% improvement each year
 - factor of 1000 every decade
 - *Capacity growth 10× as fast as processor performance*

Disk cost

- Only a few years ago, disks purchased by the megabyte
- Today, 1GB costs \$0.05 from Dell
 - (except you have to buy in increments of 1000GB)
 - \Rightarrow 1TB costs \$50, 1PB costs \$50K
- Performance analogy

- Flying an aircraft at 600mph 6 inches above the ground
- Reading/write a strip of postage stamps

14.2 Memory hierarchy



8

Disks and the OS

- Disks are difficult devices
 - errors, bad blocks, missed seeks, etc.
- OS abstracts this for higher-level software
 - lower-level device drivers (initiate a disk read, etc.)
 - higher-level abstractions (file, databases, etc.)
 - disk hardware increasingly helps with this
- OS provide different levels of disk access to different clients
 - physical disk block (surface, cylinder, sector)
 - disk logical block (disk block number)
 - file logical (filename, block or record or byte number)

Physical disk structure

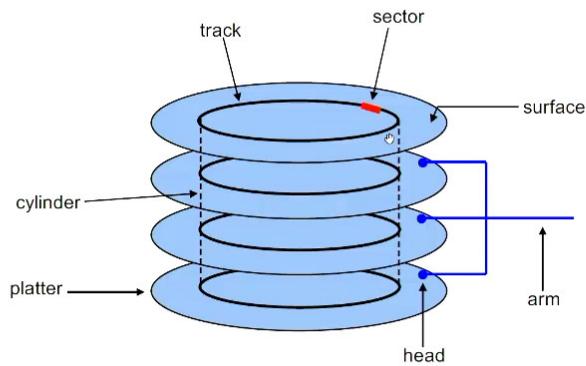
- Disk components
 - platters
 - surfaces
 - tracks
 - sectors
 - cylinders
 - arm

- heads

Physical disk structure

- Disk components

- platters
- surfaces
- tracks
- sectors
- cylinders
- arm
- heads



10

Disk structure

- Disk drives are addressed as
 - large 1-dimensional arrays of *logical blocks*
 - the logical block is the smallest unit of transfer
 - low-level formatting creates *logical blocks* on physical media
- The 1-dimensional array of logical blocks
 - is mapped onto the sectors of the disk sequentially
- Sector 0 is the first sector of the first track on the outermost cylinder
 - mapping proceeds in order through that track
 - then the rest of the tracks in that cylinder
 - then through the rest of the cylinders from outermost to innermost
- Logical to physical address should be easy
 - except for bad sectors
 - non-constant number of sectors per track via constant angular velocity

14.3 Performance

Disk performance

- Performance depends on a number of steps
- *Seek*: moving the disk arm to the correct cylinder
 - depends on how fast disk arm can move
 - * not diminishing quickly due to physics
- *Rotation (latency)*: waiting for the sector to rotate under head

- depends on rotation rate of disk
 - * rates are slowly increasing
- *Transfer*: transferring data from surface to disk controller
 - then sending it back to host
 - depends on density of bytes on disk
 - * increasing, relatively quickly
- When the OS uses the disk, it tries to minimise the cost of all these steps
 - particularly seeks and rotation

Performance

- OS may increase file block size
 - in order to reduce seeking
- OS may seek to co-locate “related” items
 - in order to reduce seeking
 - * blocks of the same file
 - * data and metadata for a file
- Keep data or metadata in memory to reduce physical disk access
 - Waste valuable physical memory:
- If file access is sequential
 - fetch blocks into memory before requested

14.4 Scheduling

Performance via disk scheduling

- Seek times are very expensive, so the OS attempts to schedule disk requests that are queued waiting for the disk
 - FCFS (do nothing)
 - * reasonable when load is low
 - * long waiting time for long request queues
 - SSST (shortest seek time first)
 - * minimise arm movement (seek time), maximise request rate
 - * unfairly favours middle blocks
 - SCAN (elevator algorithm)
 - * service requests in one direction until done, then reverse
 - * skews wait times non-uniformly
 - C-SCAN
 - * like scan, but only go in one direction (typewriter)
 - * uniform wait times
 - C-LOOK

- * similar to C-SCAN
- * The arm goes only as far as the final request in each direction

Selecting a Disk-Scheduling Algorithm

- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk
 - Less starvation
- Performance depends on the number and types of requests
- Requests for disk service can be influenced by the file-allocation method
 - And metadata layout
- The disk-scheduling algorithm should be
 - written as a separate module of the OS
 - allowing it to be replaced with a different algorithm if necessary
- Either SSTF or C-LOOK is a reasonable choice for the default algorithm
- What about rotational latency?
 - Difficult for OS to calculate

Interacting with disks

- Previously
 - OS would specify cylinder number, sector number, surface number, transfer size
 - * i.e. OS needs to know all the disk parameters
- Modern disks more complex
 - not all sectors are the same size, sectors are remapped, ...
- Disk provides a higher-level interface e.g. SCSI
 - exports data as a logical array of blocks [0 ... N]
 - maps *logical blocks* to cylinder/surface/sector
- OS only names logical block number
 - disk maps this to cylinder/surface/sector
 - on-board cache
 - as a result, physical parameters are hidden from OS

14.5 SSDs

Solid state drives: ongoing disruption

- Hard drives are based on spinning magnetic platters
 - *mechanics* of drives determine performance characteristics
 - * sector addressable, not byte addressable
 - * capacity improving exponentially
 - * sequential bandwidth improving reasonably

- * random access latency improving very slowly
- Cost dictated by
 - massive economies of scale
 - and many decades of commercial development and optimisation

SSD

- Solid state drives are based on NAND flash memory
 - no moving parts; performance characteristics driven by electronics and physics — more like RAM than spinning disk
 - relative technological newcomer, so costs are still quite high in comparison to hard drives, but dropping fast

14.5.1 Read

SSD performance: reads

- Reads
 - unit of read is a *page*, typically 4KB large
- Today's SSDs can typically handle
 - 10,000–100,000 reads/s
- 0.01–0.1 ms read latency
 - 50–1000× better than disk seeks
- 40–400 MB/s read throughput
 - 1–3× better than disk sequential throughput

14.5.2 Write

SSD performance: writes

- Writes
 - flash media must be *erased* before it can be written to
 - unit of erase is a block, typically 64–256 pages long
 - * usually takes 1–2ms to erase a block
 - * blocks can only be erased a certain number of times before they become unusable — typically 10,000–1,000,000 times
 - unit of write is a page
 - * writing a page can be 2–10× slower than reading a page
- Writing to an SSD is complicated
 - random write to existing block: read block, erase block, write back modified block
 - * leads to hard-drive like performance (300 random writes / s)
 - sequential writes to erased blocks: fast!
 - * SSD-read like performance (100–200 MB/S)

SSDs: dealing with erases, writes

- Lots of higher-level strategies can help hide the warts of an SSD
- Many of these work by
 - exposing logical pages, not physical pages
- Wear-levelling
 - when writing
 - try to spread erases out evenly across physical blocks of the SSD
 - * Intel promises 100GB/day × 5 years for its SSD drives
- Log-structured filesystems
 - convert random writes within a filesystem
 - to log appends on the SSD

14.5.3 Cost

SSD cost

- Capacity
 - today, flash SSD costs ~\$2.50/GB
 - * 1TB ssd costs around \$2,500
 - * 1TB hdd costs around \$50
 - Data on cost trends is volatile/preliminary
- Energy
 - SSD is typically more energy efficient than a HDD
 - * 1–2 watts to power an SSD
 - * ~10 watts to power a higher performance HDD
 - (can also buy a 1 watt lower-performance drive)