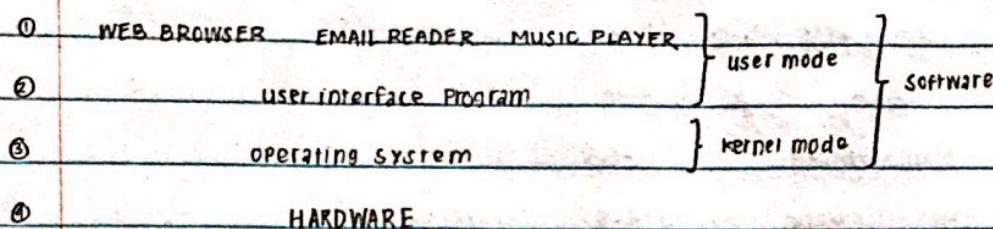


OPERATING SYSTEMS Provide user programs with a better/simpler/cleaner model

of the computer & to handle managing resources throughout the machine.



the "levels" of a machine

They are built to evolve due to how hefty they are,

and they manage/protect memory, I/O devices, and resources: keeps track of

which program is using which resource, to track resources, grant resource

requests, account for usage, and mediate conflicting requests from different

programs & users.

Resource management includes multiplexing: sharing resources in time & space

Time multiplexing: programs/users take turns using it (issues of scheduling and fairness)

Space multiplexing: resources are divided into parts (issues of fairness and protection)

HISTORY OF OPERATING SYSTEMS (BRIEFLY)

VACUUM TUBES

just to get a

TRANSISTORS AND BATCH SYSTEMS

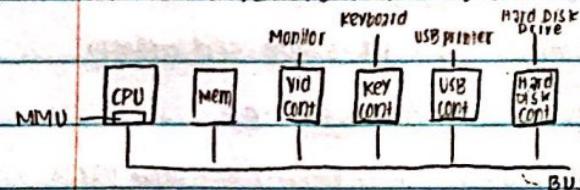
fun idea of how things progressed

ICs AND MULTIPROGRAMMING (INTEGRATED CIRCUITS)

PERSONAL COMPUTERS

MOBILE COMPUTERS

COMPUTER HARDWARE REVIEW



CPU: fetches instr from mem, executes them

has general registers (to hold vars / temp results)

& special registers: program counter (mem addr of next instr)

stack pointer (top of curr stack in mem)

PSW (comparison instr for I/O & sys calls)

NOW WE INTRODUCE MULTITHREADING (ALSO KNOWN AS HYPERTHREADING)

It allows the CPU to hold the states of different threads and switch back and forth

on a nanosecond time scale. (it doesn't offer true parallelism, only pseudoparallelism)

Memory Overview

TYPICAL ACCESS TIME

TYPICAL CAPACITY

1 nsec	Registers	< 1KB	smallest
2 nsec	Cache	4MB	
10 nsec	Main Memory	1-8GB	
10 msec	Magnetic Disk	1-4TB	larger

Registers: internal to CPU, programs must manage them

Cache Mem: mostly controlled by hardware, most heavily used cache lines are either located inside or close to the CPU. Cache miss goes to memory, so most OS's keep (pieces of) heavily used files in main memory to avoid fetching from disk
BUT, when do we put a new item into the cache? which cache line do we put it on? which item do we remove from the cache when out of space? and where do we put this newly evicted item in memory?

Disks (magnetic/hard disks) are mechanical devices and one read/write head per surface

They have two surfaces per platter, and information is written onto it in a series of concentric circles.

At any given arm position, heads can read tracks, which are divided into sectors (typically 512 bytes/sector). Outer disks = usually more sectors.

one cylinder to next: 1msec

random cylinder: 5-10 msec

once on correct track, sector to rotate under head: 5msec - 10 msec

once sector under head, read/write: 50MB/sec - 160 MB/sec

SSDs store in flash memory (no moving parts), data not lost when powered off.

Virtual memory: can run programs larger than physical mem by placing them on the disk and using main mem as a kind of cache for most heavily executed parts

It requires mapping mem addrs on the fly to convert the address the program generated to the physical address in RAM where the word is located.
done by the MMU

And as a note, switching from one program to another is called a context switch, and sometimes it is necessary to flush all modified blocks from the cache and change the mapping registers in the MMU while doing so (but it's expensive, so we try to avoid this)

I/O usually consists of: the device itself

- it's controller (chipset of chips that controls device)
- accepts commands from OS and carries said commands out

The software that talks to a controller (gives commands/accepts responses) is a device driver.

Each controller has a small # of registers for possibly specifying the disk address, memory address, sector count, and direction (read or write). These registers can be mapped into the OS's address space. If this happens, no special I/O instructions are required, and user progs can be kept away from hardware using base and limit registers.

I/O can be done using:

Busy Waiting - User prog issues sys call, kernel translates to procedure call, ^{to driver}

driver starts I/O & sits in tight loop polling device to see if done.

When I/O = complete, driver puts data (if any) where needed, and returns.

OS then returns control to caller.

Interrupts - Starts device & asks to give interrupt when finished. Driver then returns, and OS blocks caller if needed/looks for other work to do.

When controller detects end of transfer, interrupt is generated to signal completion.

DMA (DIRECT MEMORY ACCESS) - special hardware in form of chip that can control the flow of bits b/w mem and some controller w/o constant CPU intervention.

CPU sets up DMA by telling it how many bytes to transfer, the device/mem addrs involved, and the direction. When DMA chip is done, it causes interrupt (process discussed later).

OVERVIEWS OF OPERATING SYSTEM CONCEPTS

Processes are programs in execution. They have their own address space (contains executable program, program's data, and stack), and registers.

Related processes cooperating to get a job done often need to communicate and synchronize activities (this is called IPC, Interprocess communication).

Address space (typically 2^{32} or 2^{64} bytes each), further discussed later

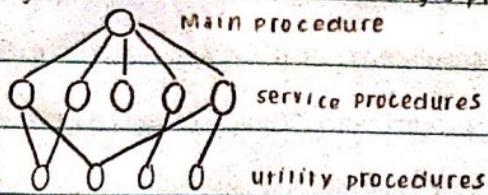
i.e. talks about different calls but it's a straightforward

important ones to note are link, fork, and the rest are common sense

We'll go more into files and file systems later

OPERATING SYSTEM STRUCTURES

MONOLITHIC: (most common organization) OS runs as a single program in kernel mode



LAYERED: organize the OS as a hierarchy of layers, each one constructed upon the one below it

layer	function
5	the operator
4	user programs
3	input/output management
2	operator-process communication
1	memory and drum management
0	processor allocation and multiprogramming

Microkernel: put as little as possible in kernel mode. let microkernel handle
interrupts, processes, scheduling, and IPC.

Client-Server Model: servers provide some service each, and clients use
these services. communication occurs w/ message passing.

Virtual Machines: exactly that, virtual machines that can each run
different operating systems, all on the same computer

Exokernels: instead of cloning the actual machine (like VMs), you can
partition it (give each user a subset of the resources). Exokernel
allocates resources to VMs and checks attempts to use them to
make sure no machine is trying to use another's resources.

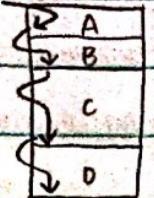
END OF REVIEW/GENERALIZATIONS/OVERVIEWS, let's jump in (i)

processes and threads

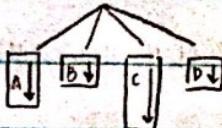
A **process** is the abstraction of a running program, and in any multiprogramming
system, the CPU switches from process to process quickly. The CPU is only ever
running one process, but these speedy switches give the illusion of the CPU
running multiple processes at a time.

There's only one physical program counter, so when each process runs, its
logical PC is loaded into the physical one. When finished (for time being), it's
stored in the process' stored logical PC in memory.

One program counter



Four program counters



please note, only one PC/program

is active at once

processes may be created due to four principal events:

System initialization

Execution of a process-creation system call by running a process

A user request to create a new process

Initiation of a batch job

and processes that stay in the background to handle some activity are called daemons

And on the opposite end, processes will terminate due to:

Normal exit (voluntary)

Error exit (voluntary)

1) processor blocks for input

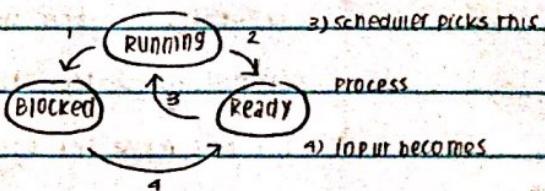
Fatal error (involuntary)

2) scheduler picks another process

Killed by another process (involuntary)

processes may be in these three states

1) Running (actually using the CPU at that instant)



2) Ready (runnable, temporarily stopped to let another process run)

available

3) Blocked (unable to run until some external event happens)

To maintain the process model, the OS maintains a process table (w/ one entry per process)

Each entry contains that process' state, program counter, stack pointer, memory allocation,

status of its open files, accounting/scheduling information, & everything else

about the process that must be saved when switched from running to ready/blocked state.

When a disk interrupt happens, the running process' PC, program status word,

and sometimes register(s) are pushed onto the (current) stack by the interrupt hardware.

The computer then jumps to the address of the interrupt service procedure which decides what to do next.

processes are interrupted thousands of times during execution, but after each

interrupt, the interrupted process returns to precisely the same state it was

in before the interrupt occurred.

Skeleton interrupt - 1) hardware stacks PC, etc 2) hardware loans new PC from int vector

3) assembly saves registers, sets up new stack 4) interrupt service reads/buffers input

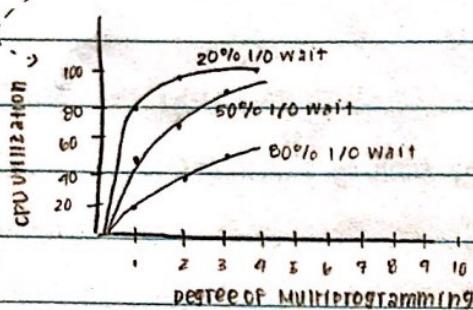
5) scheduler decides which process to run next 6) assembly starts up new curr. process

When multiprogramming is used, CPU utilization can be improved.

Suppose that a process spends a fraction "p" of its time waiting for I/O to complete.

With "n" processes in memory at once, the prob that all processes are waiting for I/O is " p^n "

$$\text{so CPU utilization} = 1 - p^n$$



Threads (aka, a kind of process within a process)

Threads are a miniprocess that help make the programming model simpler.

(aka they share common memory)

They add the ability for parallel entities to share an address space and all its data

amongst themselves. Not only that though. They're lighter weight than

processes, so easier/faster to create and destroy. Threads yield no

performance gain when they're CPU-bound, but when there's substantial

computing & I/O, they speed up the application.

All threads have the same address space, which means they also share the

same global variables. Every thread can access every mem addr within

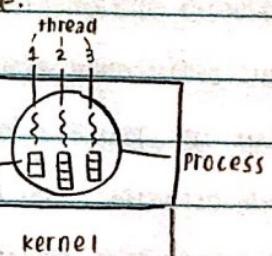
the process' addr space, so one thread could read/write/overwrite another

thread's stack. No protection is necessary b/c threads though b/c it's

impossible and unnecessary. They're built to cooperate.

per-process items

per-thread items



addr space

program counter

global vars

registers

open files

stack

child processes

state

Pending alarms

signals/signal handlers

accounting information

so... threads... do they go in the kernel?

At the user-level?

BOTH???

YOU CAN PUT THE THREADS PACKAGE ENTIRELY IN USER SPACE.

The kernel will know nothing about them, and will see them as one single-threaded process, and this can be good for implementing threads on an OS that doesn't support them.

If managed in user space, each process needs its own thread table to keep track of the threads within it.
keeps track of each thread's PC, stack pointer, registers, state, etc

Thread scheduling/switch is quite fast, & user level threads allow each process to have its own custom scheduling alg. User level threads are also more scalable.

User level threads interfere with the implementation of blocking system calls though, because one blocked thread shouldn't affect another.

Code placed around the system call to do checking (a jacket/wrapper) is inefficient, but helps to see if a call will block.

User level threads: pointless for CPU bound applications

You can also put the threads package in the kernel. (this tends to be the better/more popular option)

The kernel has a thread table that keeps track of all threads. If a process wants to make or destroy a thread, it makes a kernel call. The cost is more expensive, so thread recycling is utilized.

Kernel threads: don't require any new, nonblocking system calls, & page faults are more efficient to work around, BUT overhead is incurred b/c of high-cost system calls.

Hybrid approaches also exist. Kernel-level threads can have user-level threads multiplexed onto them.

Scheduler activations mimic the functionality of kernel-level threads, but with better performance & flexibility like those in user space.

They use upcalls so that the run-time system can reschedule threads when a thread has blocked.

Pop-up threads can be created to handle incoming messages.

They're quick to create, and reduce the latency b/c message arrival and the start of processing.

Issues with making single-threaded code multithreaded:

- stack management
- variables global to a thread but not global to the entire program, because other threads should leave them alone
- many library procedures are not reentrant (not designed to have 2nd call made while prev has not finished)
- some signals are thread-specific (where to direct by kernel? who catches what?)

IPC, Interprocess communication

Processes need to communicate with other processes, so (in a well-structured way not using interrupts)

How can one process pass information to another

How can we prevent two processes from getting into each other's way

and How do we sequence processes if dependencies are present?

Race conditions

Processes that are working together may share common storage, so race conditions can arise when two or more processes are reading or writing some shared data and the final result depends on who runs precisely when.

To prevent trouble, we will find ways to prohibit more than one process from reading and writing the shared data at the same time.

This is called mutual exclusion (making sure that if one process is using a shared variable/file, other processes will be excluded from doing the same thing.)

Critical regions/sections are the part of the program where shared memory is accessed. To avoid races, we will prevent processes from ever being in their critical regions at the same time. conditions:

- ① No two processes may be simultaneously inside their critical regions
- ② No assumptions can be made about speed / # CPUs
- ③ No process running outside critical region may block any process
- ④ No process should have to wait forever to enter its critical region

We can achieve mutual exclusion by:

Disabling interrupts: disable interrupts just after entering crit region, re-enable just before leaving it. This ensures no process switching will occur, but this can be dangerous, or difficult with multiprocessors.

Lock variables: test lock, if lock = 0, set to 1 & enter region. If lock = 1, wait until lock = 0 to enter. But what if two processes see lock = 0 before first sets it to 1? Race occurs.

Strict Altercation: turn = 0, p0 sees this and enters crit. p1 also sees this, and sits in a tight loop until it's 1. Continuously testing a variable until a value occurs is busy waiting, and a lock that uses busy waiting is a spin lock.
It can violate the third condition above, though.

The Producer-consumer problem / bounded-buffer problem

TWO PROCESSES SHARE A BUFFER (fixed size). PRODUCER PUSHES INFO INTO IT, AND

CONSUMER TAKES INFO OUT. IF PRODUCER WANTS TO PUT IN ITEM, BUT BUFFER IS
EMPTY
FULL, IT CAN GO TO SLEEP, AND BE AWAKENED WHEN THE CONSUMER HAS
PRODUCED MORE
REMOVED ITEMS. VARIABLE COUNT KEEPS TRACK OF ITEMS IN BUFFER. BUT A RACE CONDITION CAN OCCUR BECAUSE COUNT ACCESS IS UNCONSTRAINED.

A WAKEUP BIT COULD BE SET TO PREVENT MISUNDERSTANDINGS WHEN IT COMES TO PROCESSES BEING "LOGICALLY" ASLEEP (IN A SIMPLE SCENARIO).

SEMAPHORES (USUALLY CALLED DOWN & UP), CHANGE THEIR VALUES AS A SINGLE, INDIVISIBLE ATOMIC ACTION (SO ONCE A SEMAPHORE OPERATION HAS STARTED, NO OTHER PROCESS CAN ACCESS THE SEMAPHORE UNTIL THE ACTION IS COMPLETED/BLOCKED). THEY CAN SOLVE OUR BOUNDED BUFFER PROBLEM WITH 3 SEMAPHORES:

empty: # of slots that are empty

full: # of slots that are full

mutex: stops prod & cons from entering buffer at same time

THEY ALSO HELP WITH SYNCHRONIZATION.

A MUTEX (SIMPLIFIED VERS OF A SEMAPHORE) IS USEFUL WHEN A SEMAPHORE'S

ABILITY TO COUNT IS NOT NEEDED. IT'S A VARIABLE THAT CAN EITHER BE IN A LOCKED OR UNLOCKED STATE. A FUTEX IS A FEATURE OF LINUX THAT IMPLEMENTS BASIC LOCKING, BUT AVOIDS DROPPING INTO THE KERNEL UNLESS IT REALLY HAS TO, AND PTHREADS NOT ONLY OFFER MUTEXES, BUT ALSO CONDITION VARIABLES, WHICH ALLOW THREADS TO BLOCK DUE TO A CONDITION NOT BEING MET.

MONITORS ARE A HIGHER-LEVEL SYNCHRONIZATION PRIMITIVE. THEY'RE A COLLECTION OF

PROCEDURES, VARIABLES, AND DATA STRUCTURES ALL GROUPED TOGETHER IN A SPECIAL

KIND OF MODULE/PACKAGE. ONLY ONE PROCESS CAN BE IN A MONITOR AT ANY INSTANT.

THEY ARE AN EASY WAY TO ACHIEVE MUTUAL EXCLUSION, AND THE CONDITION

VARIABLES WAIT AND SIGNAL ALLOWS PROCESSES TO BLOCK WHEN THEY CANNOT PROCEED, AND THE WAIT MUST COME BEFORE THE SIGNAL TO PREVENT LOST SIGNALS.

MESSAGE PASSING ALLOWS INFORMATION EXCHANGE BETWEEN MACHINES, USING

THE TWO PRIMITIVES SEND AND RECEIVE. IF A MESSAGE IS RECEIVED, AN ACKNOWLEDGE MESSAGE IS SENT BACK.

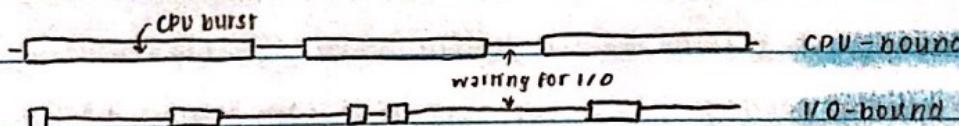
The last synchronization mechanism, barriers, is intended for process groups.

No processes may be allowed to proceed until they all reach a specific phase. When a process reaches a barrier, it is blocked until all processes reach the same barrier. This allows groups to synchronize.

SCHEDULING

Multiple processes or threads may be competing for the CPU at the same time, and it's up to the OS's scheduler to use scheduling algorithms to determine which ready process will run next. The scheduler must make efficient use of the CPU because process switching is expensive. A switch from user → kernel mode occurs, & the state of the curr process is saved (along w/ storing registers and process table) and sometimes memory map (page table). Next process is selected by scheduler, MMU is reloaded w/ mem map of new process, and mem cache & related tables may now be invalid, so those will be dynamically reloaded from main mem twice.

Processes alternate b/w bursts of computing with I/O requests:



more common as CPUs get faster

When to schedule?

① When new process created ② Process exits

③ Process blocks on I/O ④ I/O interrupt occurs

~ only option if no clock available

Nonpreemptive scheduling algos let a process run until it blocks or releases CPU, while preemptive only lets it run for a maximum of fixed time.

And scheduling algorithm goals differ for each system

Batch systems want to maximize throughput, minimize turnaround time,

and keep the CPU busy at all times

Interactive systems want to minimize response time & meet expectations

Real-time systems want to meet deadlines & avoid quality degradation

& All systems want fairness, balance, and policy enforcement

SCHEDULING ALGORITHMS:

EOPES: processes assigned to CPU in order they request it, and are allowed to run for as long as it wants to.

SJF: scheduler picks the shortest jobs in the input queue to run first. Optimal only when all jobs available simultaneously.

Shortest Remaining Time Next: Again, optimal when all run times known in advance. Allows new short jobs to get good service.

Round Robin: each process assigned time interval (quantum) during which it is allowed to run. Too short of a quantum causes too many switches & lowers CPU efficiency, but too long causes poor response to interactive requests.

PRIORITY SCHEDULING: Runnable process w/ highest priority allowed to run.

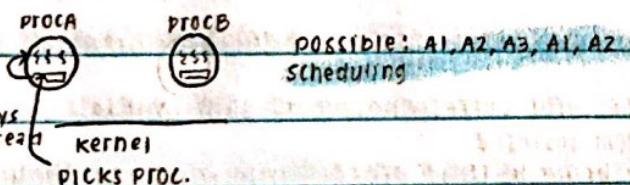
Might want to incorporate RR to prevent starvation.

Lottery Scheduling: give processes lottery tickets for various resources.

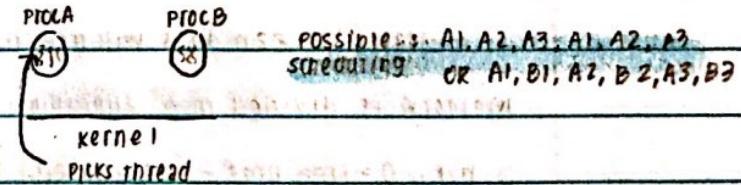
It's highly responsive, and can be used to solve otherwise difficult scheduling problems.

Thread scheduling differs depending on user-level vs kernel-level.

user level:



kernel level:



classical IPC problems include:

Dining Philosophers → (watch videos online for clarity)

Readers and Writers → use semaphores!

→ see Courtois et al.

MEMORY MANAGEMENT

Two problems have to be solved to allow multiple applications to be in mem at the same time without interfering with each other: protection & relocation (and each process has one)

An address space (set of addresses that a process can use to address memory) is a good solution for this.

It's a bit more difficult to give each program an address space. Let's start with a simpler (but dated) implementation.

We will use a simple version of dynamic relocation to accomplish this, for our solution will map each process' address space onto a different part of physical memory using base and limit registers. Programs will be loaded into consecutive memory locations whenever there is room.

The base register is loaded with the phys addr where the program begins, the limit register is loaded w/ the length of the program.

Base + Physical Addr → sent on memory bus → memory reference reached
We can't fit ALL of the processes into physical memory, and we have two strategies to help deal with this:

SWAPPING (bringing in each process, running it for a while, then putting it back to disk) ^{r in its entirety}

& virtual memory (allows programs to run even when they're partially in main mem)

SWAPPING:

It can create holes in memory, and we could either swap into those holes (if space allowed us to), or we could compact by moving all processes downward.

We should leave room for growth, though, due to a program's data segment/stack.
We can manage free memory with bitmaps and free lists.

BITMAPS

A memory of $32n$ bits will use n map bits. Large allocation = smaller bitmap.

Memory is divided into allocation units, and corresponding to each unit is a bit. 0 = free unit, 1 = occupied. If we bring k -unit process into memory though, we'd have to find a run of k consecutive 0s in the map.

LINKED LISTS

We can also keep a list of allocated and free segments, where a segment either contains a process, or a hole between two processes.

There's algorithms used to allocate memory for processes, listed below:

FIRST FIT: fast, because it searches as little as possible

NEXT FIT: finds next available hole, slightly worse performance than

BEST FIT: slower, and lots of wasted mem b/c tiny holes

WORST FIT: not a good idea, but new hole kinda useful

VIRTUAL MEMORY

We need to be able to run programs too big to fit in MM, and have systems that can support multiple programs running simultaneously (that might collectively exceed memory).

Virtual memory! the idea:

each program has its own address space, which is broken up into chunks called pages.

Page: contiguous range of addresses

the pages are mapped onto physical memory, but they don't all have to be in physical memory. When the program references a part of its address space that's not in physical memory, the OS is alerted to go get the missing piece and reexecute.

BUT how does it work?

It uses a technique called paging.

Program-generated addresses are called virtual addresses, and form

the virtual address space. When virtual memory is used, virtual

addresses go to an MMU that maps virtual addresses onto physical mem addrs.

Virtual address space consists of fixed-size units called pages. Corresponding units in phys mem are called page frames.

Virtual addr → MMU → finds page frame → outputs phys addr to bus

A present/absent bit keeps track of which pages are physically in memory.

If a program references an unmapped address, the MMU notices the page is

unmapped, and causes the CPU to trap to the operating system (page fault).

OS then picks little used page frame, writes its contents back to disk,

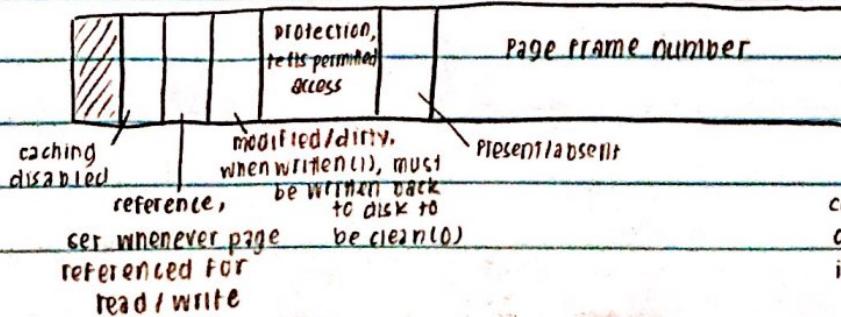
then fetches page just referenced into the freed page frame & restarts trapped instr.

Virtual page # & offset

A virtual address is a series of bits, and a portion of these is the page number,

which is used as an index into the page table (yields # of page frame of that virt page)

A Page table entry (though it varies) can look like this:



and note: Page table doesn't contain info about disk address or unreferenced page, that's kept in software tables.

TWO major issues must be faced in any paging system:

- 1) Virtual-to-physical must be done on every memory reference
- 2) The mapping from virtual to physical address must be fast
 - ↳ virtaddr ≥ 32 bits. ex: 4KB page size → 32-bit addr = 1 million pages!
- 3) If the virtual address space is large, the page table must be large

MOST programs make many references to a small number of pages,

so let's equip computers with TLBs (translation lookaside buffer)

to map virtual to physical addresses without going through a page table.

They contain a small # entries, & contain information about most recently referenced pages.

soft miss: page not in TLB, but in mem

hard miss: not in TLB, nor in mem, in disk

Multilevel Page Tables

contains PT1 field as an index into first PT, and entry located there

yields address/page frame number of a second page table, indexed

into with the PT2 field. Page frame number there + offset = phys. addr.

Inverted Page Tables

one entry per page frame in real memory, instead of one entry

per page of virtual addr space. The entry keeps track of which

(process, virtual page) is located in the page frame. (hashtable)

they save space, but virtual-to-physical translation is harder.

Page Replacement Algorithms

pages are chosen to be evicted when faults occur, but how should

we determine this? The optimal alg is to remove the page that will

be referenced latest in the future, but this is unrealizable.

realizable ones are:

NOT recently used: status R = referenced, M = modified, and periodically,

R bit is cleared. When page fault occurs, OS inspects pages.

They will be: 0) not referenced, not modified 1) not referenced, modified

2) referenced, not modified 3) referenced, modified

The alg then removes a page at random from the lowest numbered (nonempty) class.

Easy to understand, moderately efficient

to implement, adequate performance.