

CS 110

Computer Architecture

More I/O: DMA, Disks, Networking

Instructor:
Sören Schwertfeger

<http://shtech.org/courses/ca/>

School of Information Science and Technology SIST

ShanghaiTech University

Slides based on UC Berkley's CS61C

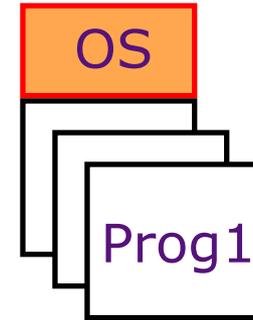
Virtual Memory Review

Modern Virtual Memory Systems

Illusion of a large, private, uniform store

Protection & Privacy

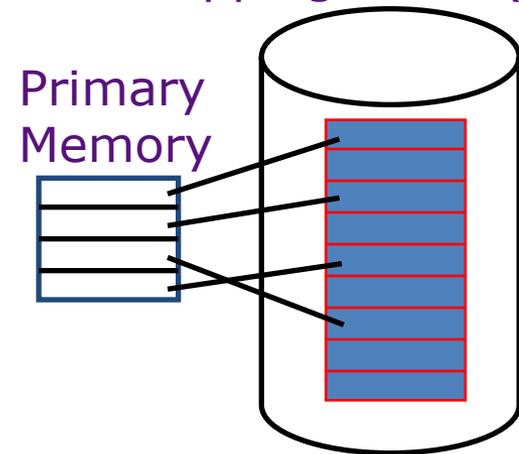
- * Many processes, each with their own private address space and one or more shared address spaces



Demand Paging

- * Many processes share DRAM.
- * Provides ability to run programs with large address space. Pages that aren't yet allocated or pages that don't fit swap to secondary storage.
- * Hides differences in machine configurations

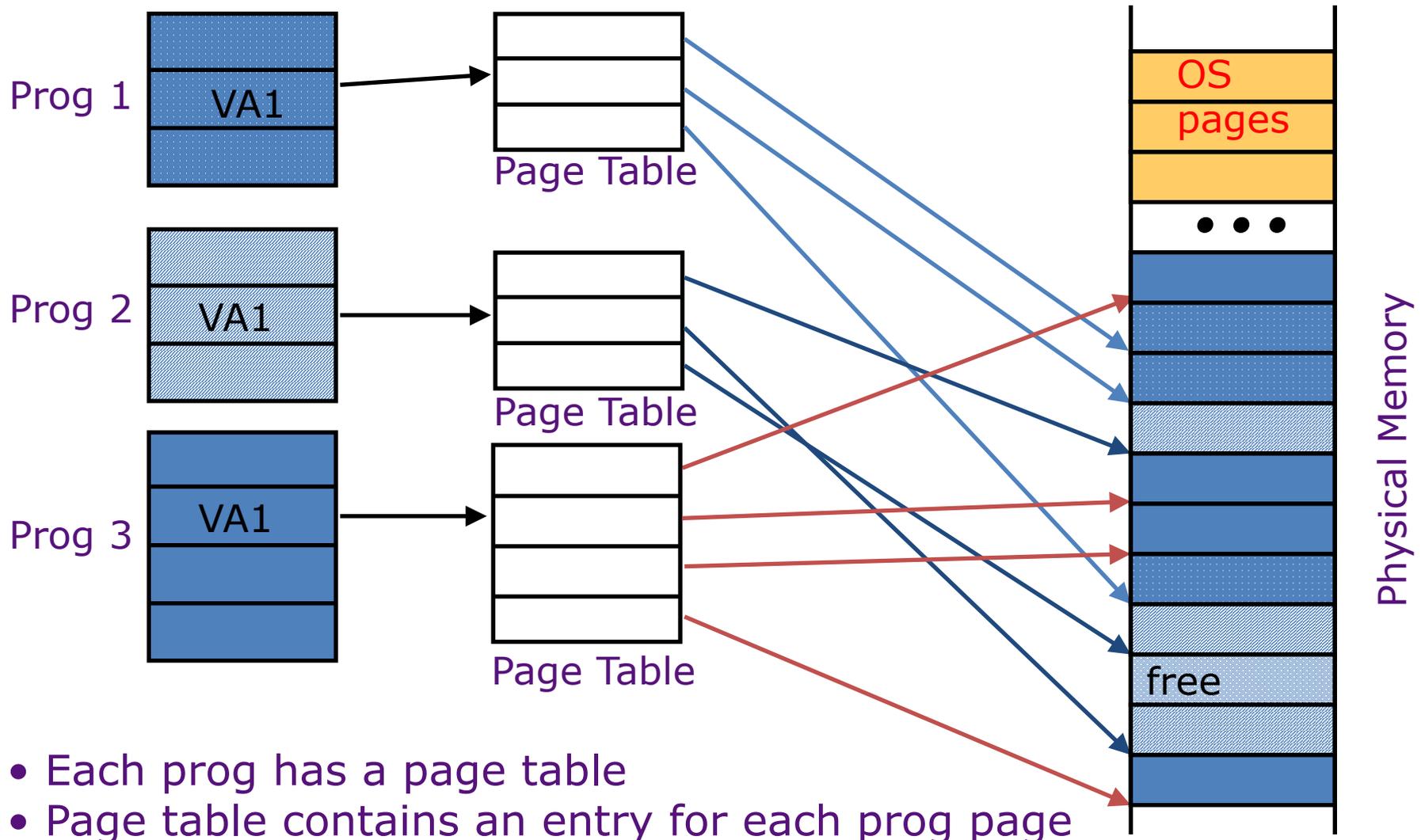
Swapping Store (Disk)



The price is address translation on each memory reference



Private (Virtual) Address Space per Program



Translation Lookaside Buffers (TLB)

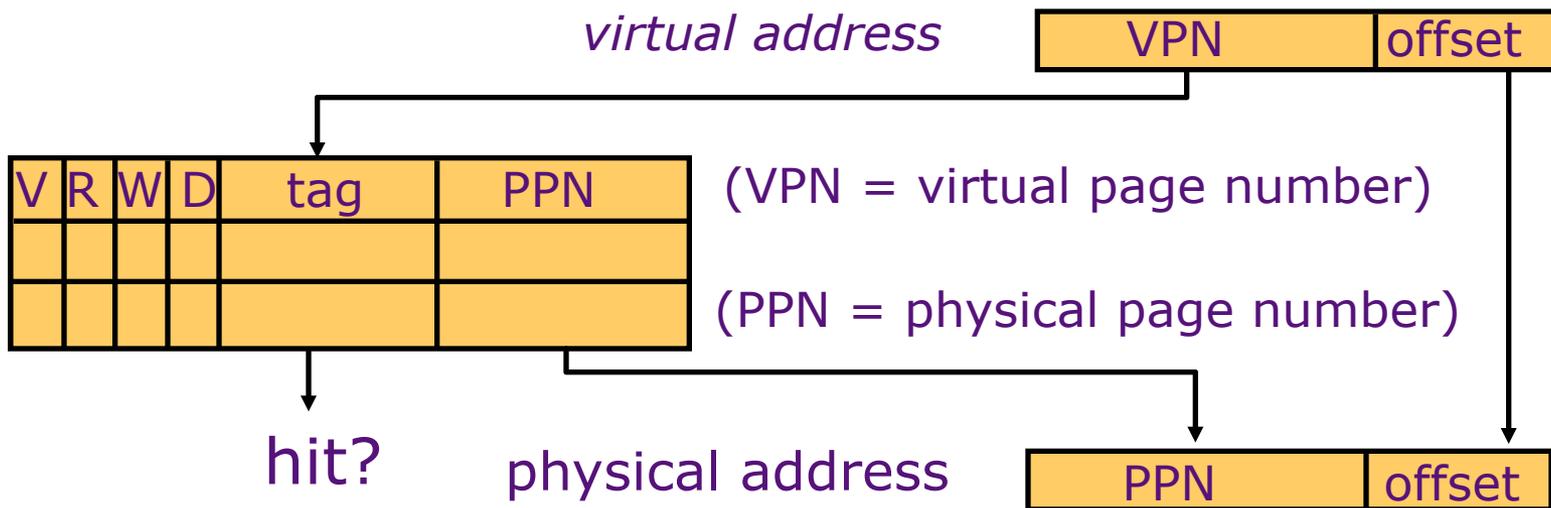
Address translation is very expensive!

In a two-level page table, each reference becomes several memory accesses

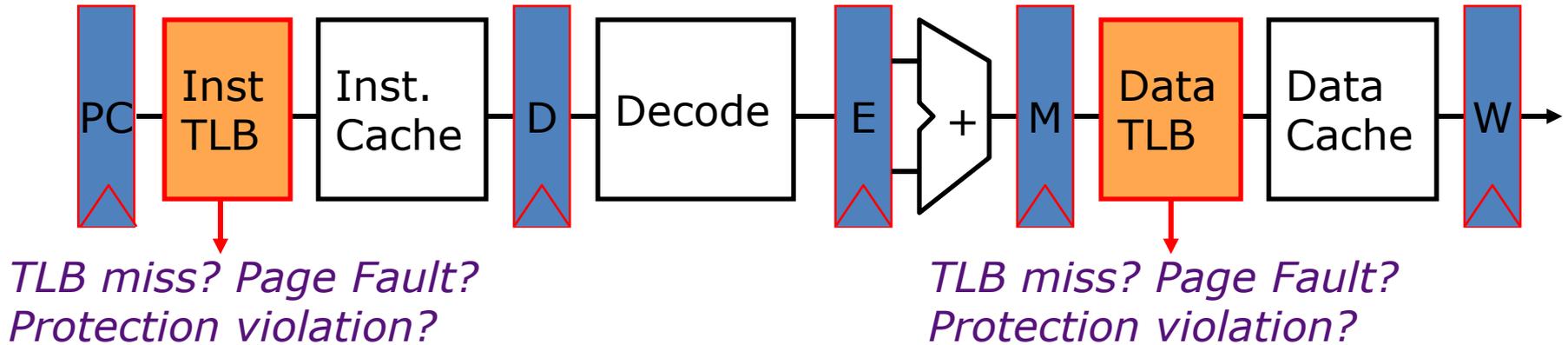
Solution: *Cache translations in TLB*

TLB hit \Rightarrow *Single-Cycle Translation*

TLB miss \Rightarrow *Page-Table Walk to refill*

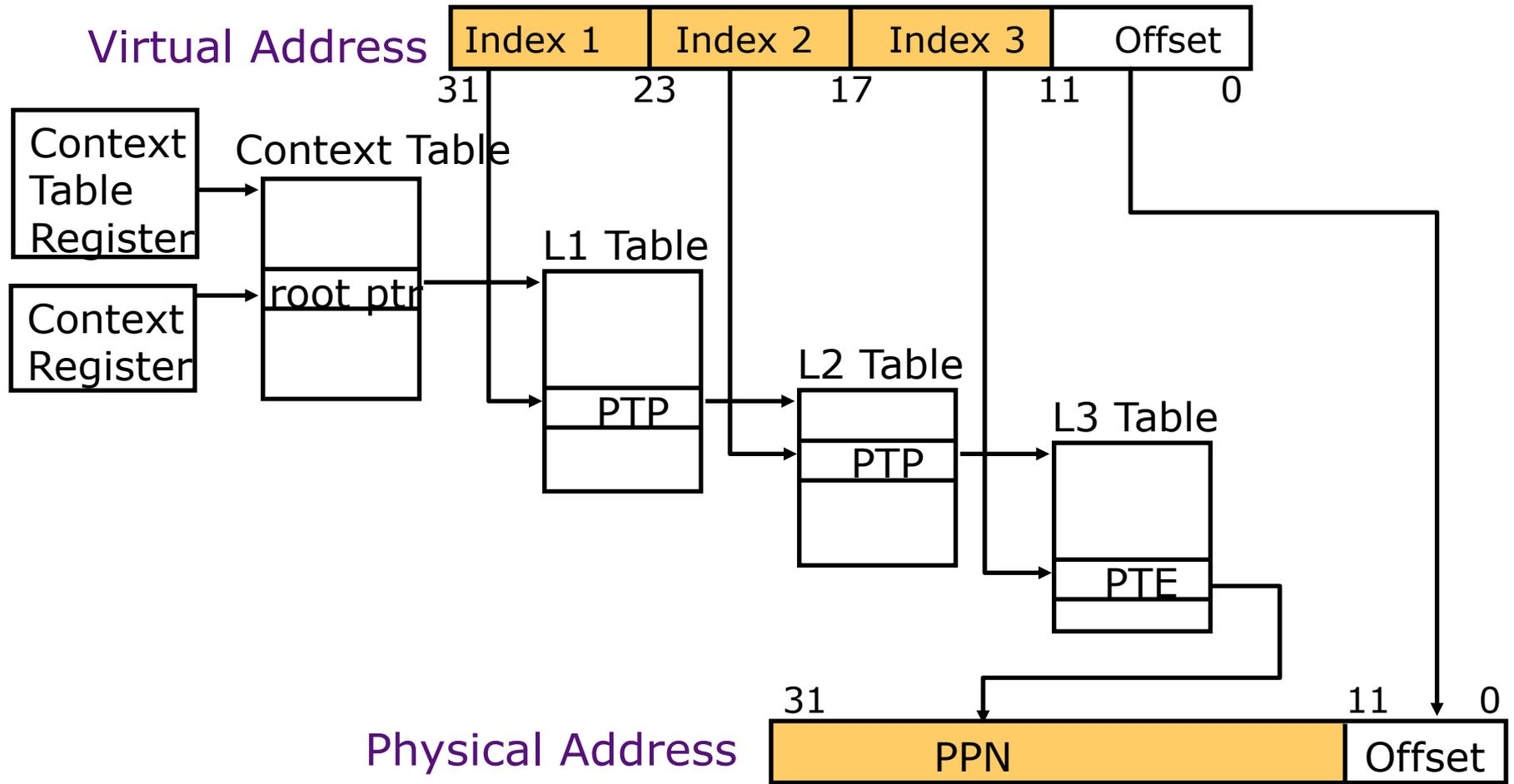


VM-related events in pipeline



- Handling a TLB miss needs a hardware or software mechanism to refill TLB
 - usually done in hardware now
- Handling a page fault (e.g., page is on disk) needs a *precise* trap so software handler can easily resume after retrieving page
- Handling protection violation may abort process

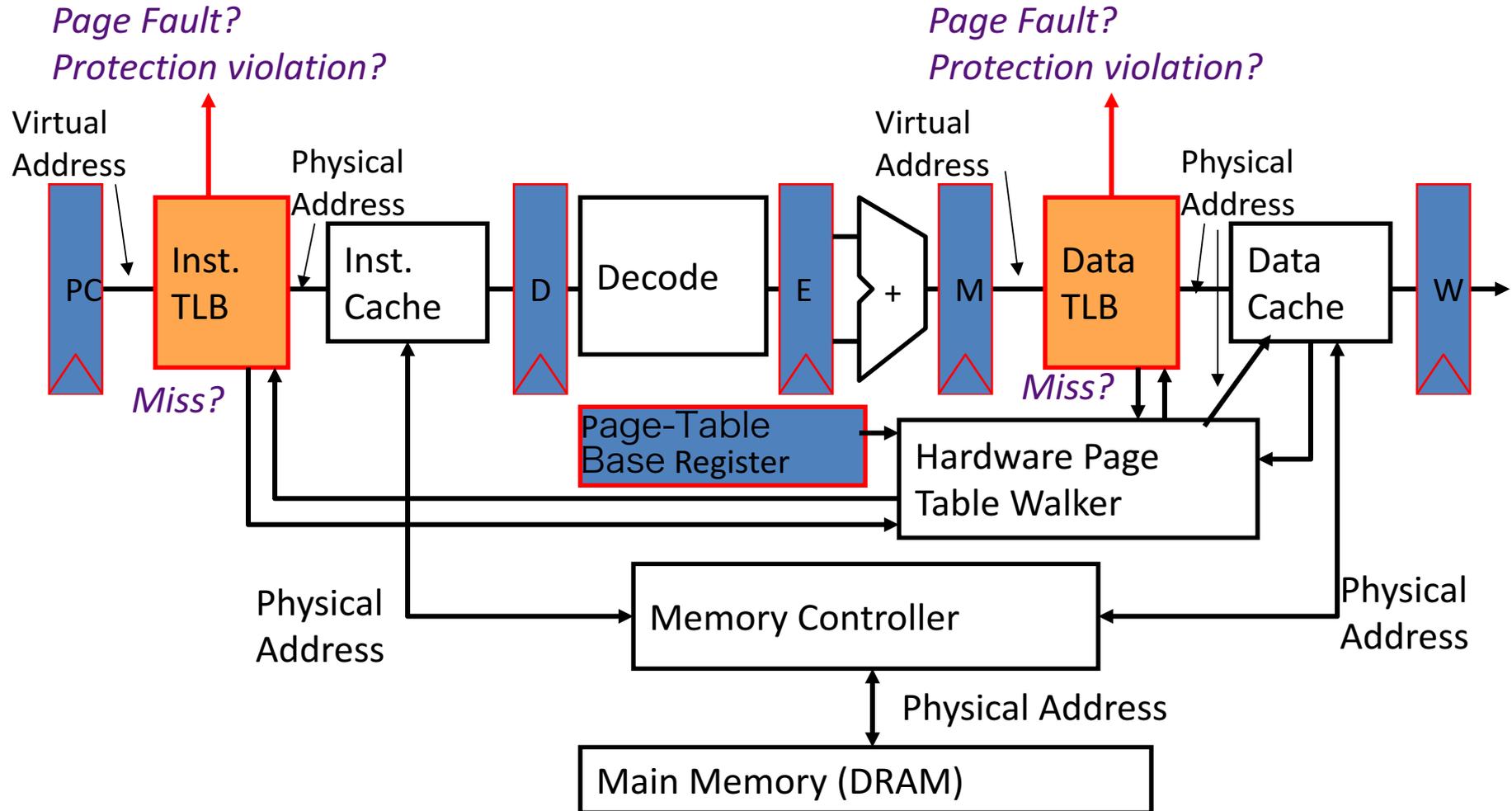
Hierarchical Page Table Walk: SPARC v8



MMU does this table walk in hardware on a TLB miss

Page-Based Virtual-Memory Machine

(Hardware Page-Table Walk)

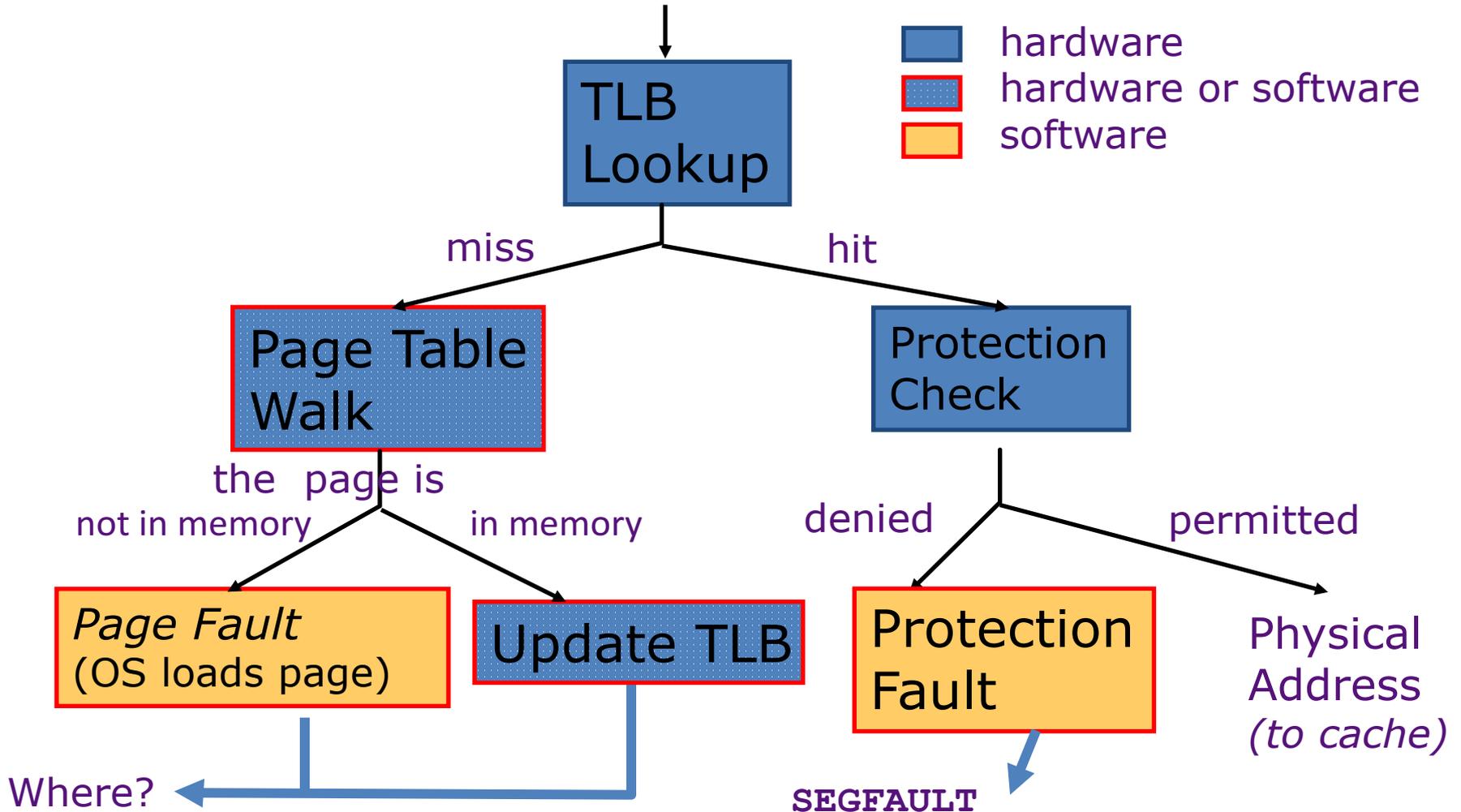


- Assumes page tables held in untranslated physical memory

Address Translation:

putting it all together

Virtual Address

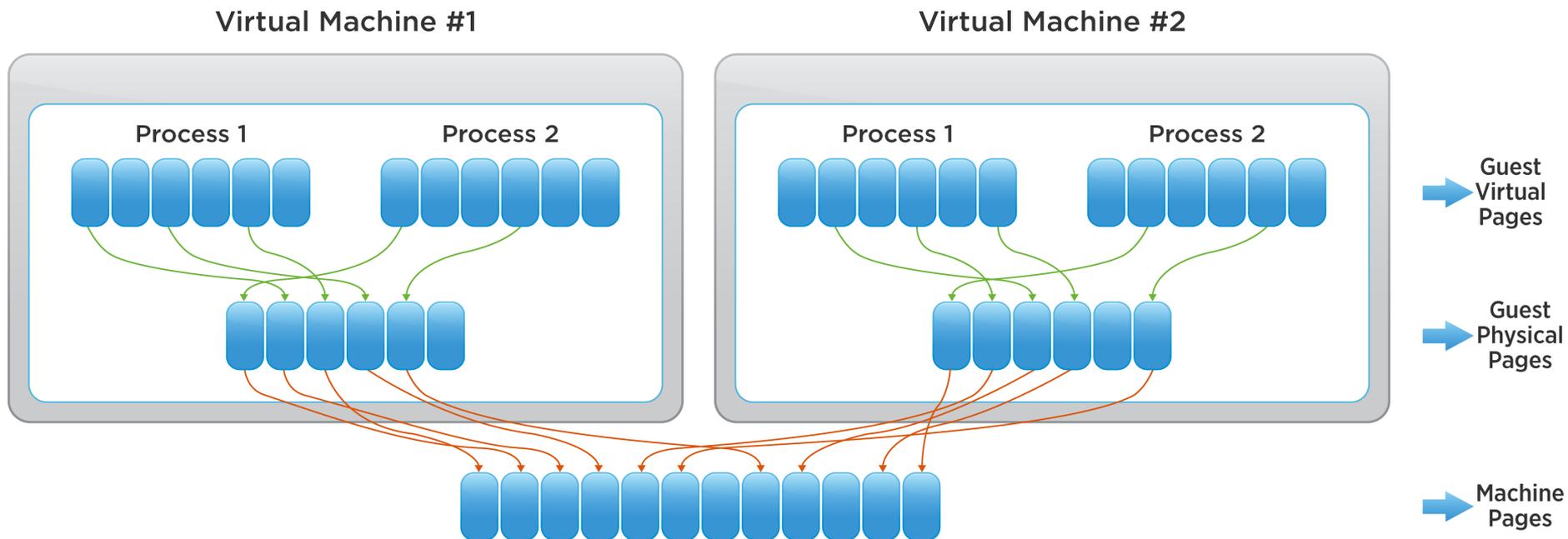


Virtual Machine

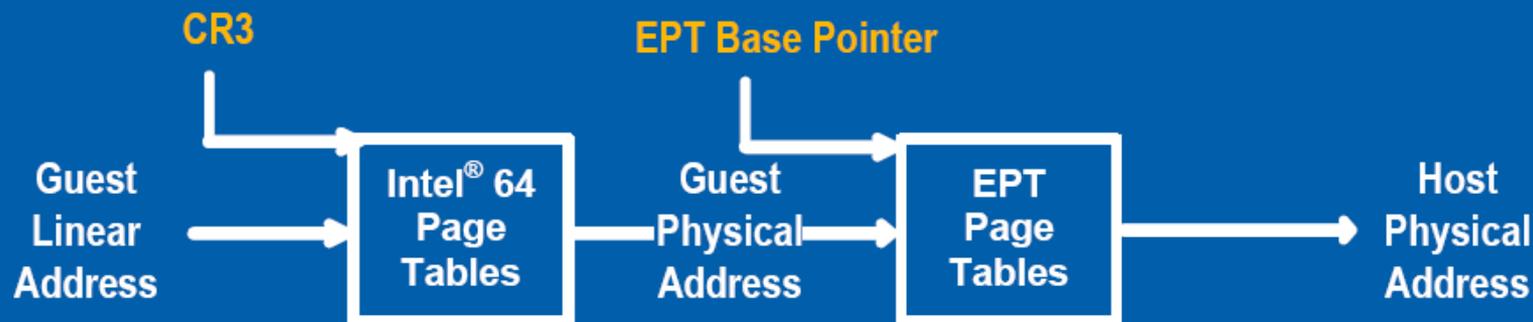
- Virtual Memory (VM) != Virtual Machine (VM) (e.g. Virtual Box)
- Emulation: Run a complete virtual CPU & Memory & ... - a complete virtual machine in software (e.g. MARS)
- Virtual Machine: Run as many instructions as possible directly on CPU, only simulate some parts of the machine.
- Last lecture: Supervisor Mode & User Mode;
now also: Virtual Machine Mode
 - Host OS activates virtual execution mode for guest OS =>
 - Guest OS thinks it runs in supervisor mode, but in fact it doesn't have access to physical memory! (among other limitations)
- CPUs support it (AMD-V, Intel VT-x), e.g. new Intel instructions: VMPTLDR, VMPTLST, VMPCLEAR, VMPCREAD, VMPCWRITE, VMPCCALL, VMPCLAUNCH, VMPCRESUME, VMPCXOFF, and VMPCXON

What about the memory in Virtual Machines?

- Need to translate Guest Virtual Address to Guest Physical address to Machine (Host) Physical address: Earlier the Guest part was done (transparently) in software by the Virtual Machine ... now in hardware!



EPT: Overview



- Intel® 64 page tables
 - Map **guest-linear** to **guest-physical** (translated again)
 - Can be read and written by guest
- New EPT page tables under VMM control
 - Map **guest-physical** to **host-physical** (accesses memory)
 - Referenced by new **EPT base pointer**
- No VM exits due to **page faults**, **INVLPG**, or **CR3** accesses

Intel® VT Roadmap: Overview

Vector 3:
I/O Focus

PCI-SIG

- Standards for I/O-device sharing:**
- Natively sharable I/O devices
 - Endpoint DMA-translation caching

Vector 2:
Platform Focus

VT-d

- Infrastructure for I/O-device virtualization:**
- DMA protection and remapping
 - Interrupt filtering and remapping

Vector 1:
Processor Focus

VT-x

VT-i

Establish foundation for virtualization in the Intel® 64 and Itanium® architectures...

- ... followed by on going evolution of support:
- Microarchitectural (e.g., lower VM entry/exit costs)
 - Architectural (e.g., extended page tables – EPT)

VMM
Software
Evolution

- Software-only VMMs**
- Binary translation
 - Paravirtualization
 - Device Emulation

Simpler and more Secure VMMs through foundation of virtualizable ISAs

Improved CPU and I/O virtualization **Performance and Functionality** as VMMs exploit infrastructure provided by VT-x, VT-i, VT-d

Past
No Hardware Support

Today 
VMM software evolution over time with hardware support



Review: I/O

- “Memory mapped I/O”: Device control/data registers mapped to CPU address space
- CPU synchronizes with I/O device:
 - Polling
 - Interrupts
- “Programmed I/O”:
 - CPU execs lw/sw instructions for all data movement to/from devices
 - CPU spends time doing 2 things:
 1. Getting data from device to main memory
 2. Using data to compute

Working with real devices

- “Memory mapped I/O”: Device control/data registers mapped to CPU address space
- CPU synchronizes with I/O device:
 - Polling
 - Interrupts
- ~~“Programmed I/O”~~: **DMA**
 - ~~CPU execs lw/sw instructions for all data movement to/from devices~~
 - CPU spends time doing ~~2 things~~:
 1. ~~Getting data from device to main memory~~
 2. Using data to compute

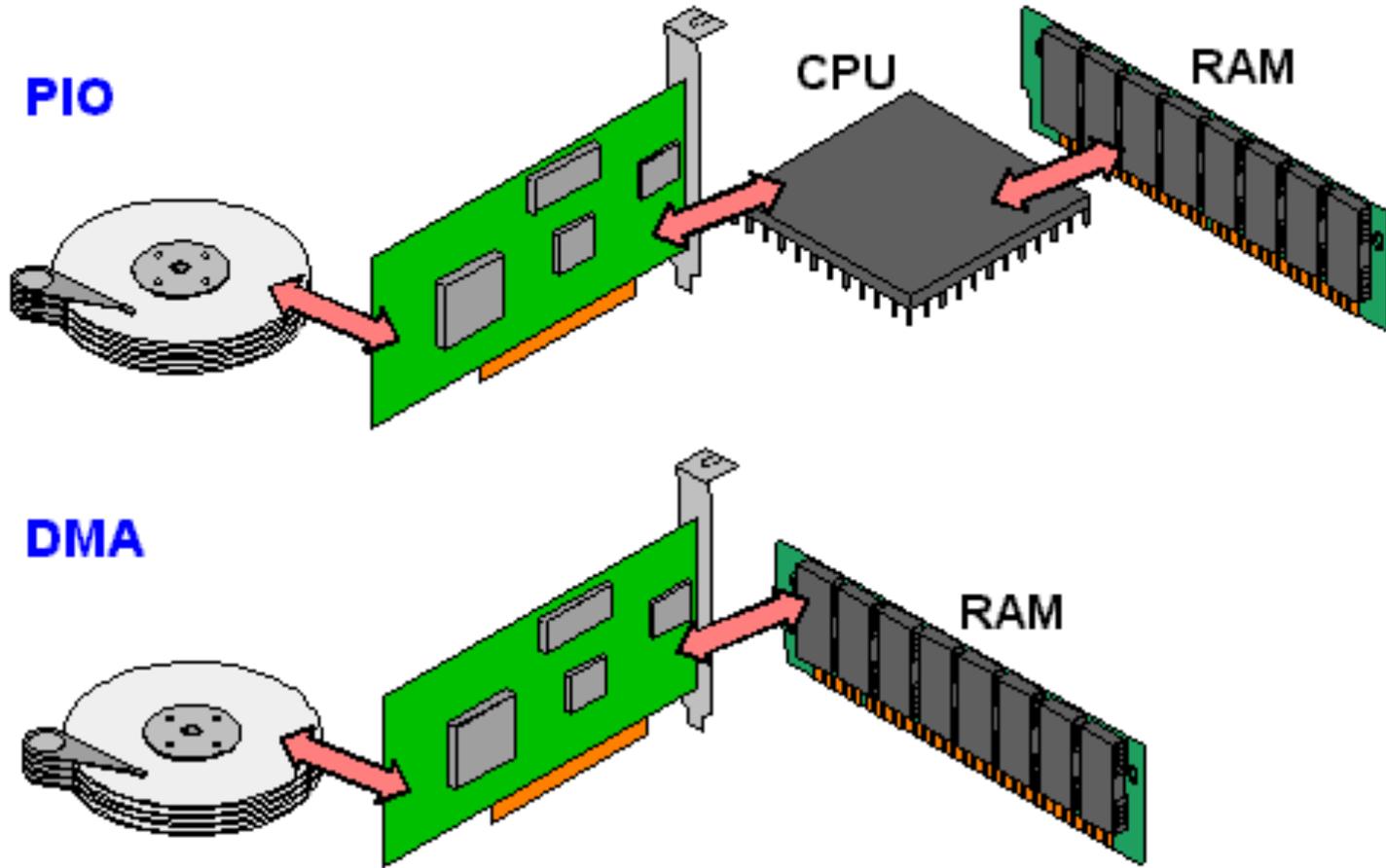
Agenda

- **Direct Memory Access (DMA)**
- Disks
- Networking

What's wrong with Programmed I/O?

- Not ideal because ...
 1. CPU has to execute all transfers, could be doing other work
 2. Device speeds don't align well with CPU speeds
 3. Energy cost of using beefy general-purpose CPU where simpler hardware would suffice
- Until now CPU has sole control of main memory

PIO vs. DMA



Direct Memory Access (DMA)

- Allows I/O devices to directly read/write main memory
- New Hardware: the DMA Engine
- DMA engine contains registers written by CPU:
 - Memory address to place data
 - # of bytes
 - I/O device #, direction of transfer
 - unit of transfer, amount to transfer per burst

Operation of a DMA Transfer

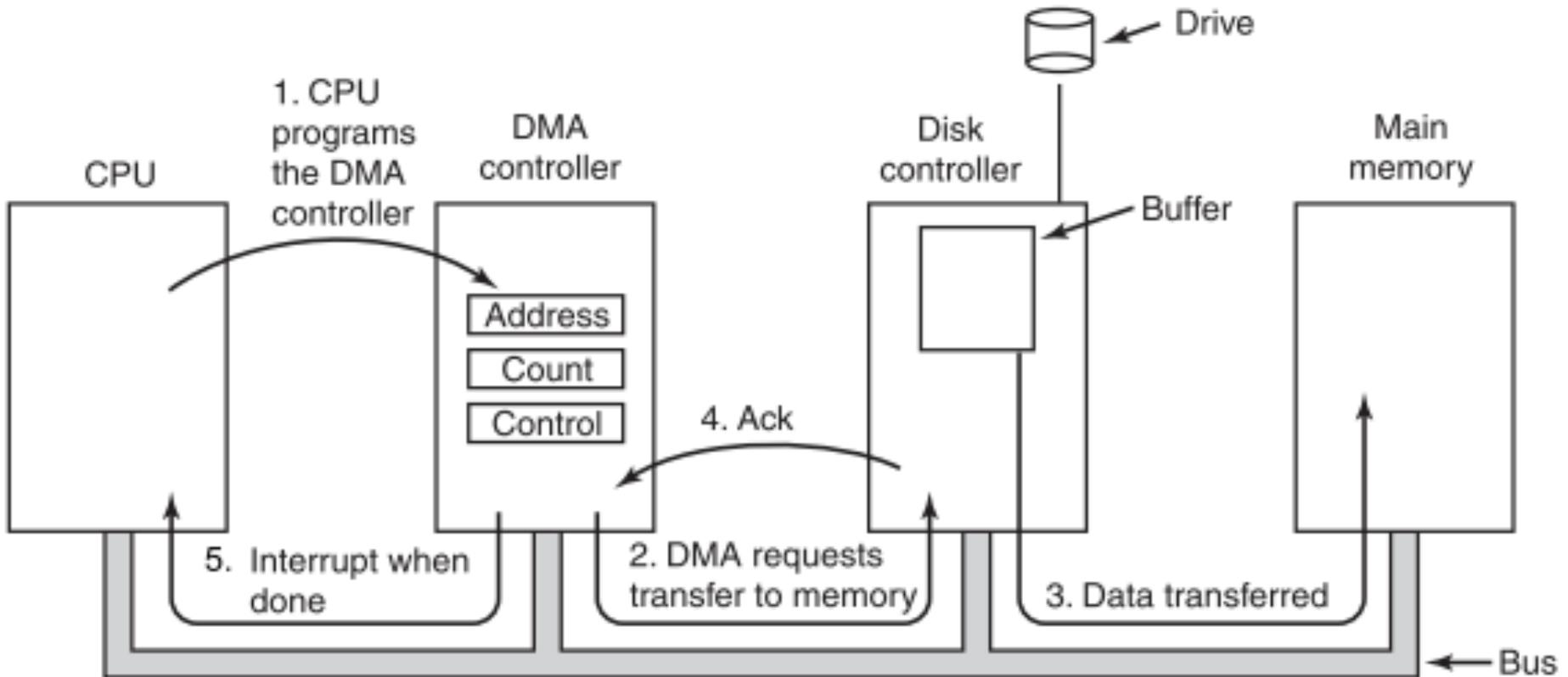


Figure 5-4. Operation of a DMA transfer.

[From Section 5.1.4 Direct Memory Access in *Modern Operating Systems* by Andrew S. Tanenbaum, Herbert Bos, 2014]

DMA: Incoming Data

1. Receive interrupt from device
2. CPU takes interrupt, begins transfer
 - Instructs DMA engine/device to place data @ certain address
3. Device/DMA engine handle the transfer
 - CPU is free to execute other things
4. Upon completion, Device/DMA engine interrupt the CPU again

DMA: Outgoing Data

1. CPU decides to initiate transfer, confirms that external device is ready
2. CPU begins transfer
 - Instructs DMA engine/device that data is available @ certain address
3. Device/DMA engine handle the transfer
 - CPU is free to execute other things
4. Device/DMA engine interrupt the CPU again to signal completion

DMA: Some new problems

- Where in the memory hierarchy do we plug in the DMA engine? Two extremes:
 - Between L1 and CPU:
 - Pro: Free coherency
 - Con: Trash the CPU's working set with transferred data
 - Between Last-level cache and main memory:
 - Pro: Don't mess with caches
 - Con: Need to explicitly manage coherency

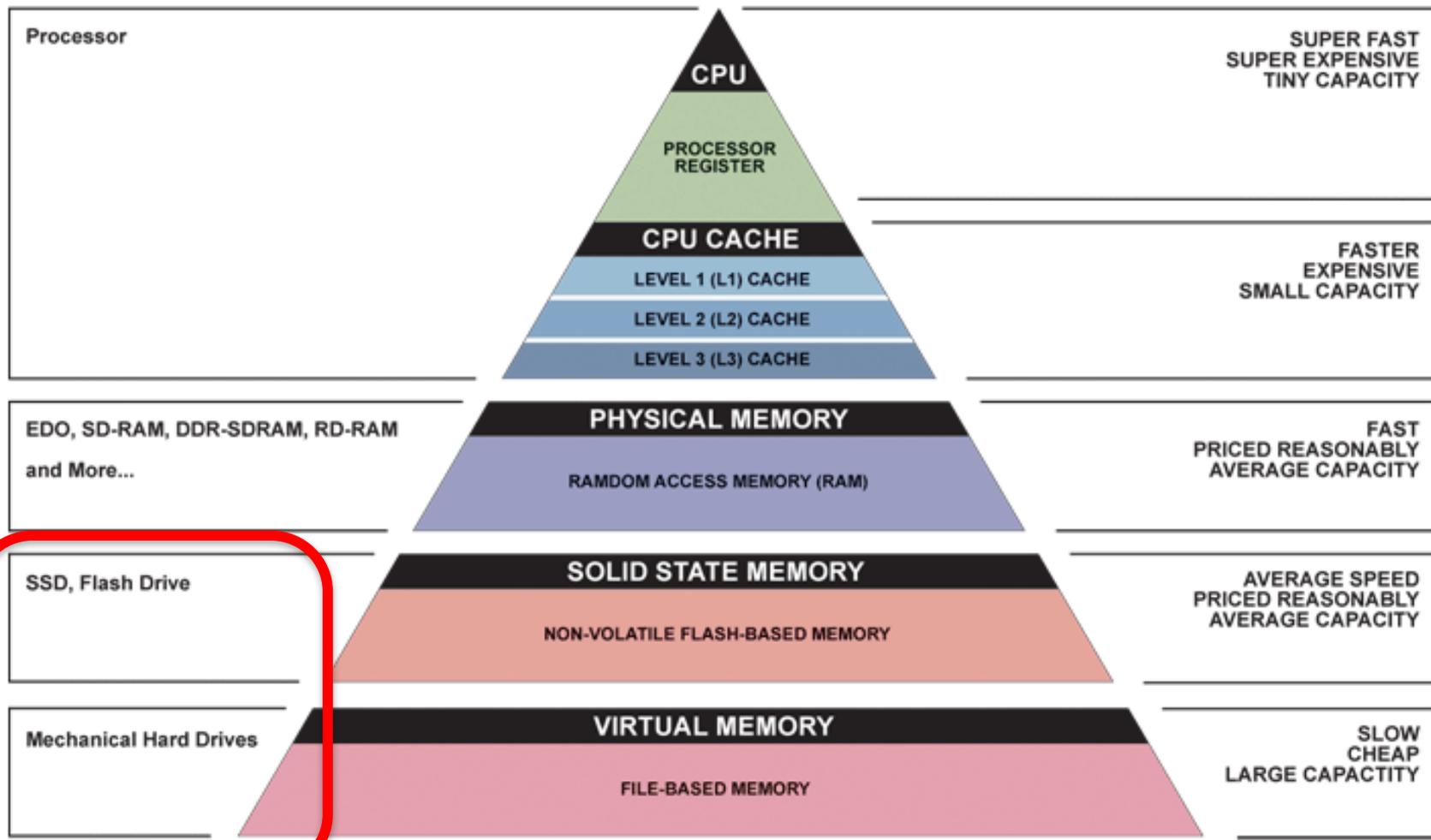
DMA: Some new problems

- How do we arbitrate between CPU and DMA Engine/Device access to memory? Three options:
 - Burst Mode
 - Start transfer of data block, CPU cannot access memory in the meantime
 - Cycle Stealing Mode
 - DMA engine transfers a byte, releases control, then repeats - interleaves processor/DMA engine accesses
 - Transparent Mode
 - DMA transfer only occurs when CPU is not using the system bus

Agenda

- Direct Memory Access (DMA)
- **Disks**
- Networking

Computer Memory Hierarchy



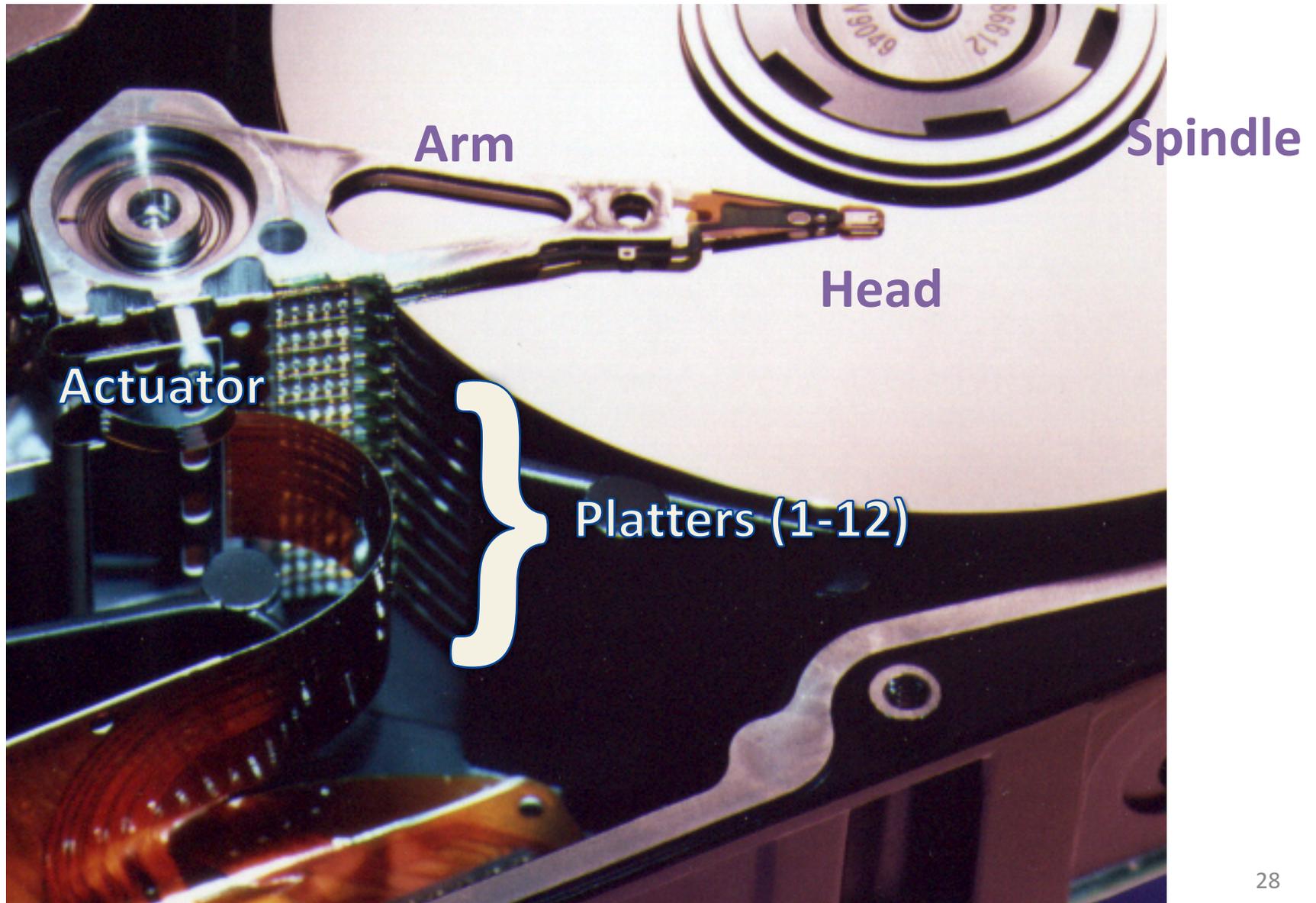
▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

Today

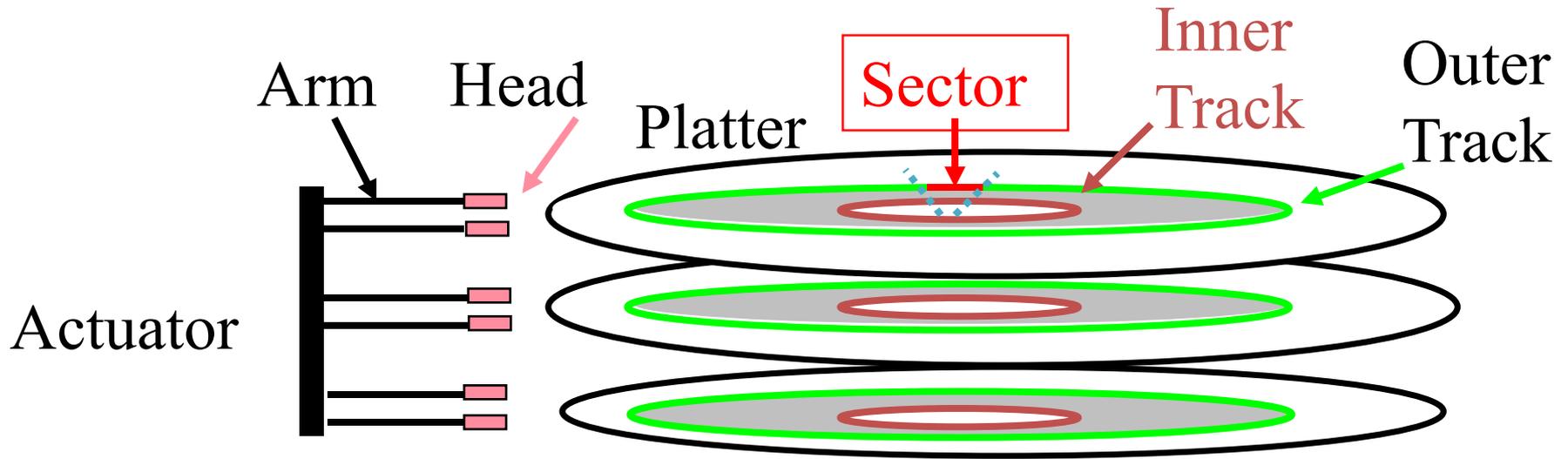
Magnetic Disk – common I/O device

- A kind of computer memory
 - Information stored by magnetizing ferrite material on surface of rotating disk
 - similar to tape recorder except digital rather than analog data
- A type of non-volatile storage
 - retains its value without applying power to disk.
- Magnetic Disk
 1. Hard Disk Drives (HDD) – faster, more dense, non-removable.
- Purpose in computer systems (Hard Drive):
 1. Working file system + long-term backup for files
 2. Secondary “backing store” for main-memory. Large, inexpensive, slow level in the memory hierarchy (virtual memory)

Photo of Disk Head, Arm, Actuator



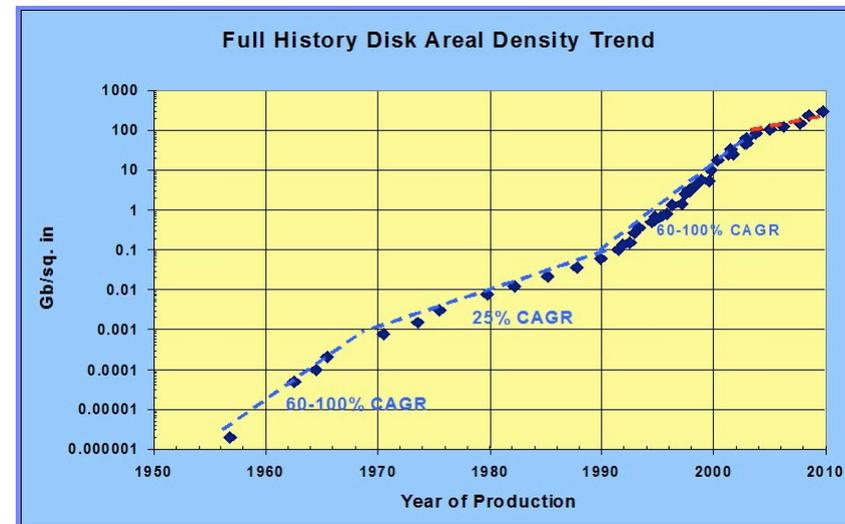
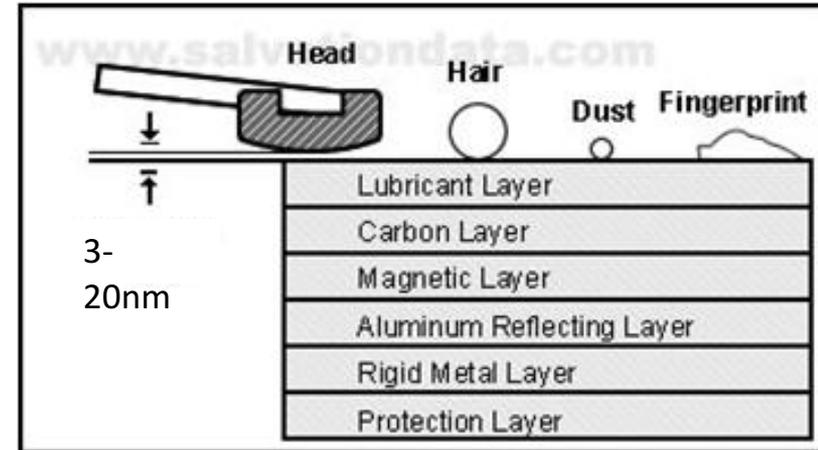
Disk Device Terminology



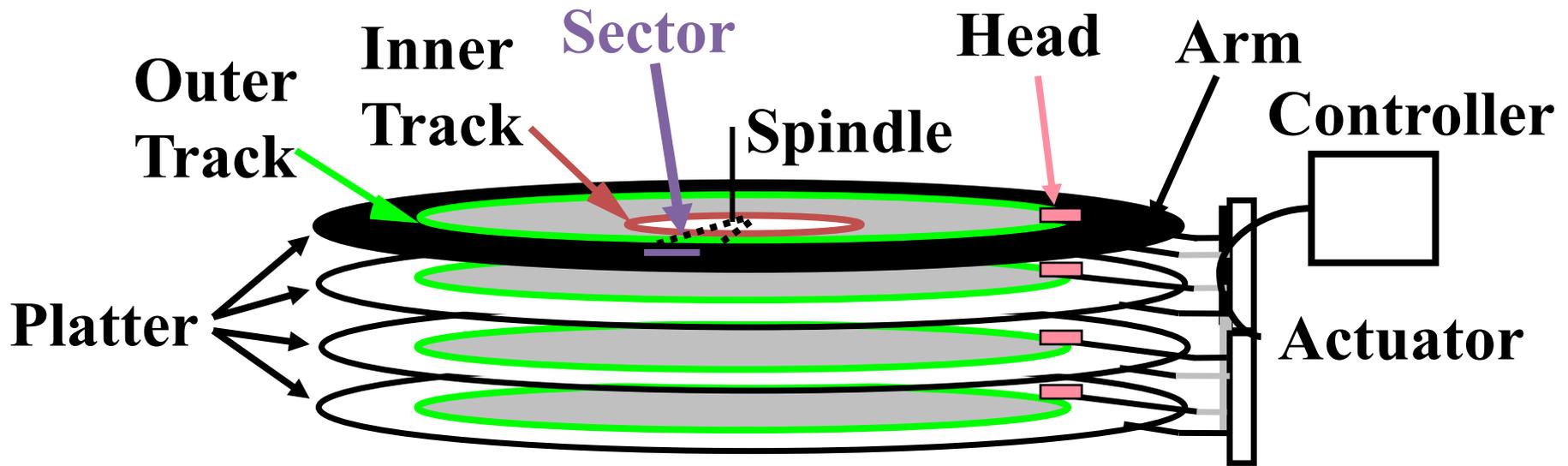
- Several platters, with information recorded magnetically on both surfaces (usually)
- Bits recorded in **tracks**, which in turn divided into **sectors** (e.g., 512 Bytes)
- **Actuator** moves **head** (end of **arm**) over track (**“seek”**), wait for **sector** rotate under **head**, then read or write

Hard Drives are Sealed. Why?

- The closer the head to the disk, the smaller the “spot size” and thus the denser the recording.
 - Measured in Gbit/in²
 - ~900 Gbit/in² is state of the art
 - Started out at 2 Kbit/in²
 - ~450,000,000x improvement in ~60 years
- Disks are sealed to keep the dust out.
 - Heads are designed to “fly” at around 3-20nm above the surface of the disk.
 - 99.999% of the head/arm weight is supported by the air bearing force (air cushion) developed between the disk and the head.



Disk Device Performance (1/2)



- **Disk Access Time = Seek Time + Rotation Time + Transfer Time + Controller Overhead**
 - Seek Time = time to position the head assembly at the proper cylinder
 - Rotation Time = time for the disk to rotate to the point where the first sectors of the block to access reach the head
 - Transfer Time = time taken by the sectors of the block and any gaps between them to rotate past the head

Disk Device Performance (2/2)

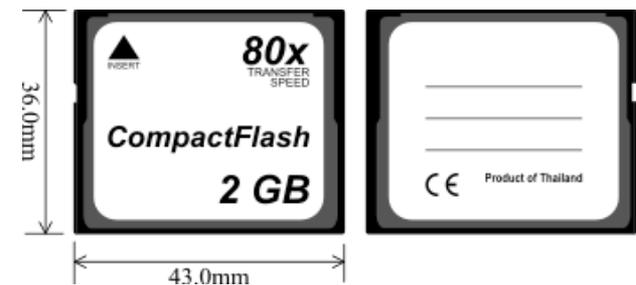
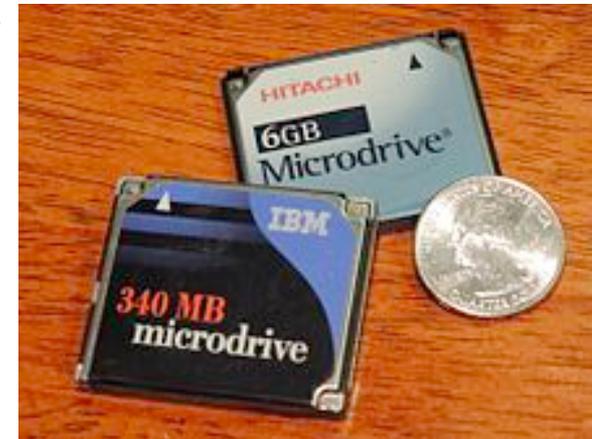
- Average values to plug into the formula:
- Rotation Time: Average distance of sector from head?
 - 1/2 time of a rotation
 - 7200 Revolutions Per Minute => 120 Rev/sec
 - 1 revolution = 1/120 sec => 8.33 milliseconds
 - 1/2 rotation (revolution) => 4.17 ms
- Seek time: Average no. tracks to move arm?
 - Number of tracks/ 3
 - Then, seek time = number of tracks moved × time to move across one track

But wait!

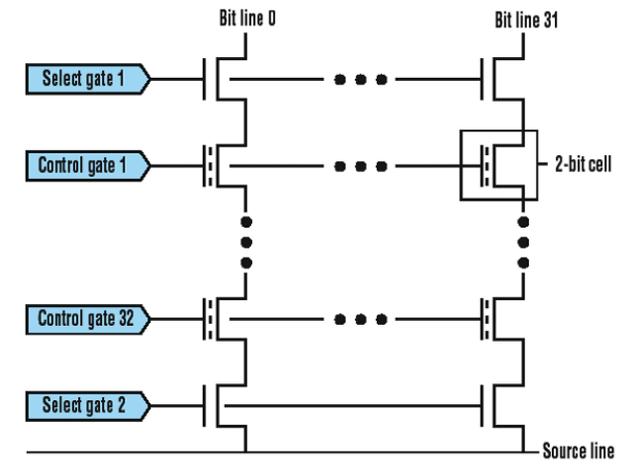
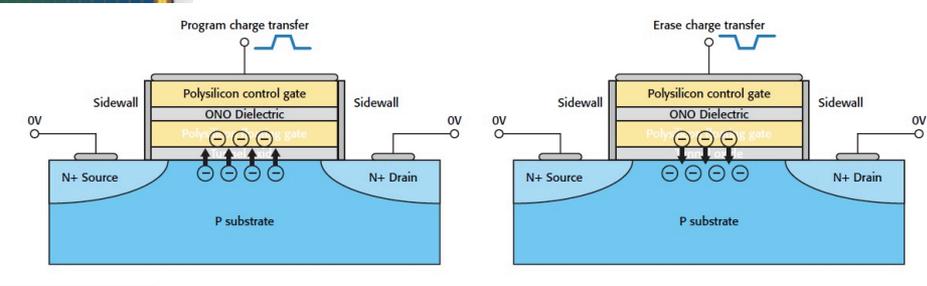
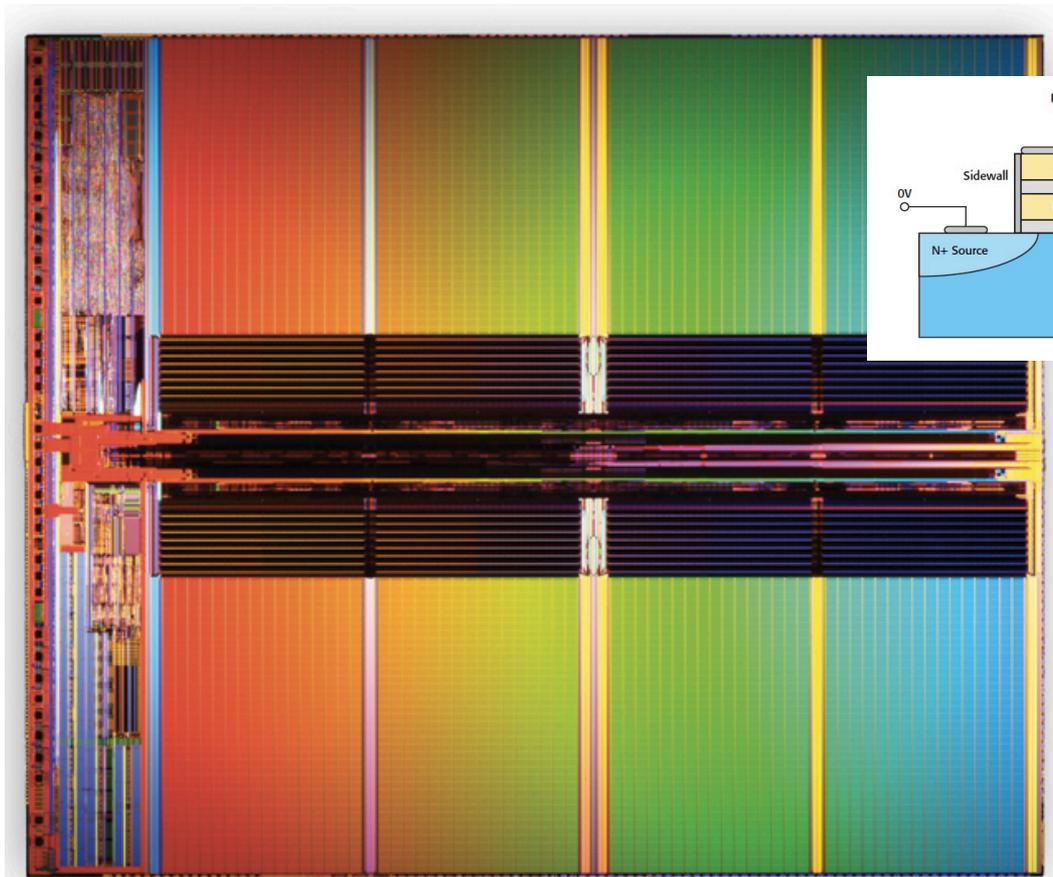
- Performance estimates are different in practice:
- Many disks have on-disk caches, which are completely hidden from the outside world
 - Previous formula completely replaced with on-disk cache access time

Where does Flash memory come in?

- ~15 years ago: Microdrives and Flash memory (e.g., CompactFlash) went head-to-head
 - Both non-volatile (retains contents without power supply)
 - Flash benefits: lower power, no crashes (no moving parts, need to spin μ drives up/down)
 - Disk cost = fixed cost of motor + arm mechanics, but actual magnetic media cost very low
 - Flash cost = most cost/bit of flash chips
 - Over time, cost/bit of flash came down, became cost competitive



Flash Memory / SSD Technology



2. Micron's triple-level cell (TLC) flash memory stores 3 bits of data in each transistor.

In the basic functional block used in multilevel NAND flash memories, 32 rows of bit lines and 32 control-gate lines form a building block that's repeated many times to form the memory array. The select gate lines are used with the control gate lines to control access to the array.

- NMOS transistor with an additional conductor between gate and source/drain which “traps” electrons. The presence/absence is a 1 or 0
- Memory cells can withstand a limited number of program-erase cycles. Controllers use a technique called *wear leveling* to distribute writes as evenly as possible across all the flash blocks in the SSD.

What did Apple put in its iPods?

Toshiba flash
2 GB



Samsung flash
16 GB



Toshiba 1.8-inch HDD
80, 120, 160 GB



Toshiba flash
32, 64, 128 GB



shuffle



nano



classic



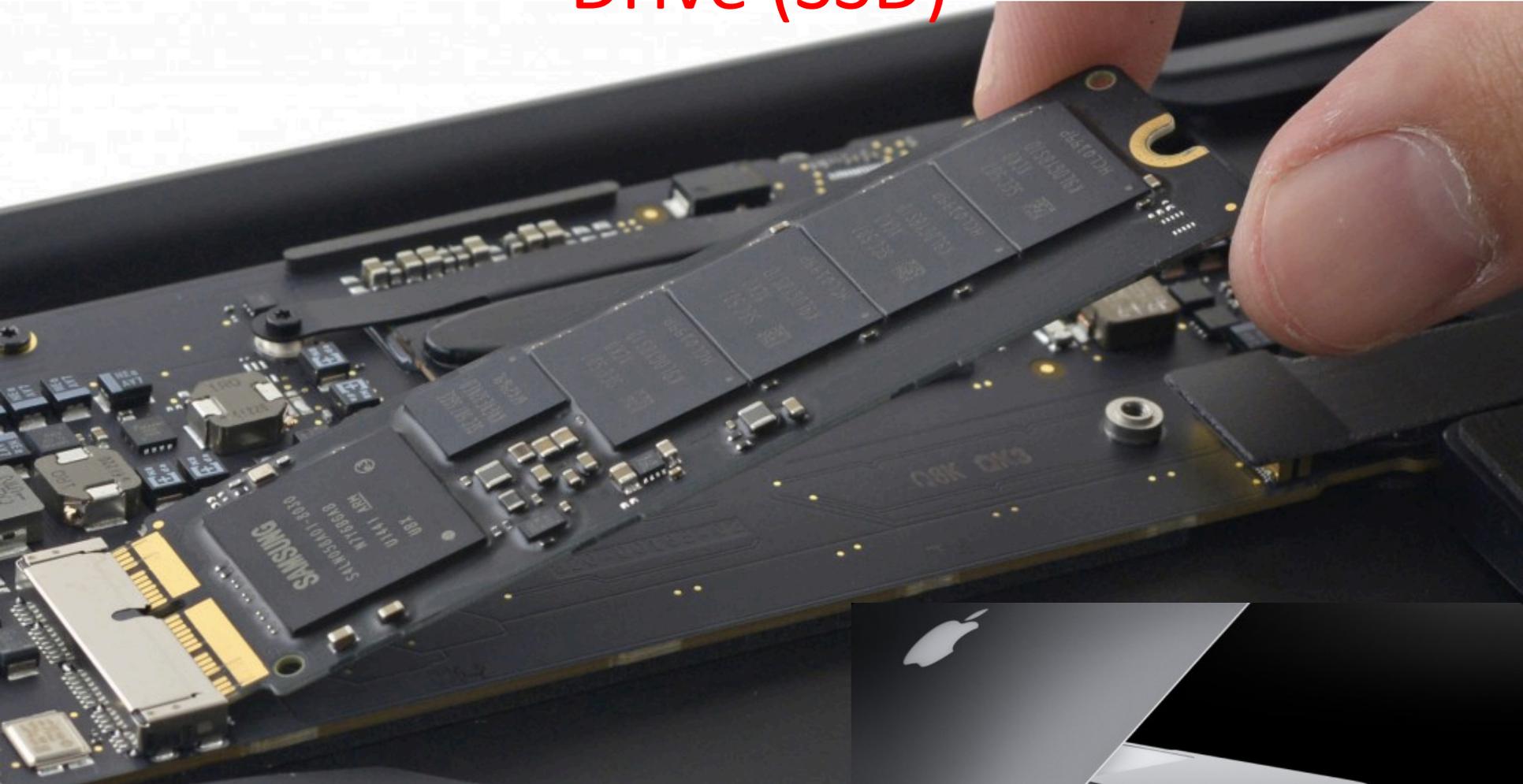
touch

Flash Memory in Smart Phones

iPhone 7: up to 256 GB



Flash Memory in Laptops – Solid State Drive (SSD)

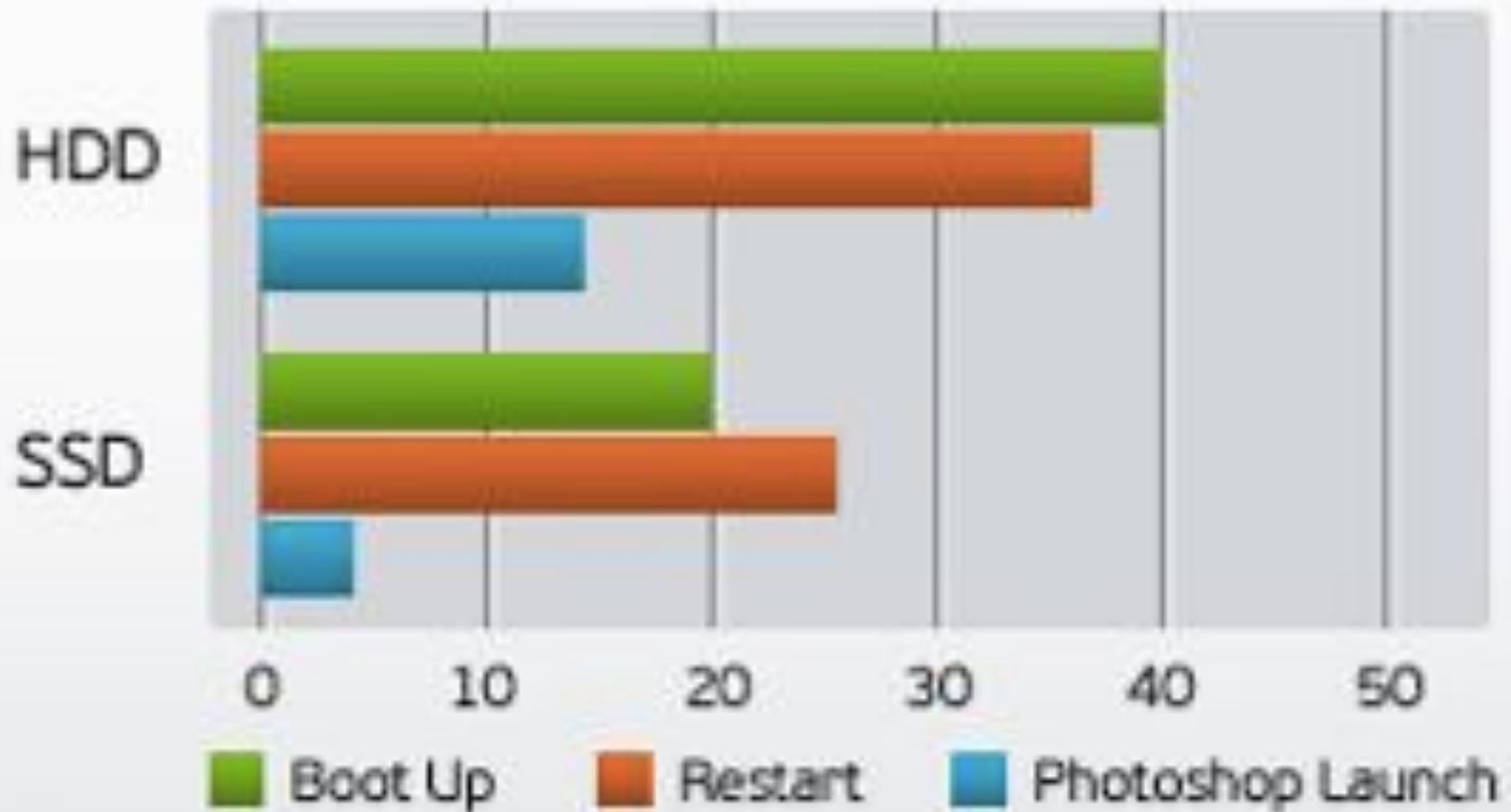


capacities up to 2TB

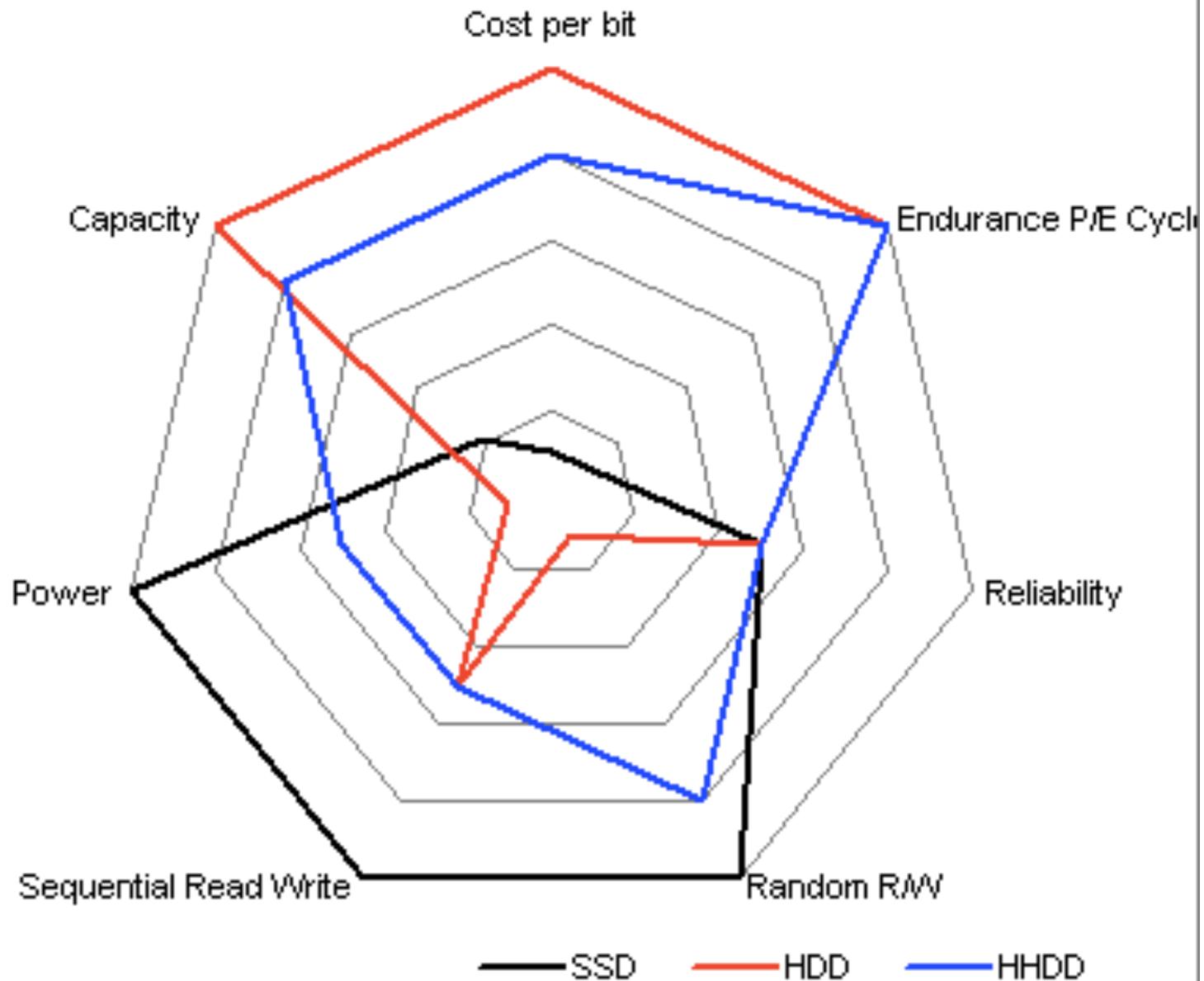


HDD vs SSD speed

Time in Seconds (Lower is Better)



Hard Drive vs. SSD vs. Hybrid HDD



Question

- We have the following disk:
 - 15000 Tracks, 1 ms to cross 1000 Tracks
 - 15000 RPM = 4 ms per rotation
 - Want to copy 1 MB, transfer rate of 1000 MB/s
 - 1 ms controller processing time
- What is the access time using our model?

Disk Access Time = Seek Time + Rotation Time + Transfer Time + Controller Processing Time

A	B	C	D	E
10.5 ms	9 ms	8.5 ms	11.4 ms	12 ms

Question

- We have the following disk:
 - 15000 Cylinders, 1 ms to cross 1000 Cylinders
 - 15000 RPM = 4 ms per rotation
 - Want to copy 1 MB, transfer rate of 1000 MB/s
 - 1 ms controller processing time

- What is the access time?

Seek = # cylinders/3 * time = 15000/3 * 1ms/1000 cylinders = 5ms

Rotation = time for ½ rotation = 4 ms / 2 = 2 ms

Transfer = Size / transfer rate = 1 MB / (1000 MB/s) = 1 ms

Controller = 1 ms

Total = 5 + 2 + 1 + 1 = 9 ms

Agenda

- Direct Memory Access (DMA)
- Disks
- **Networking**

Networks: Talking to the Outside World

- Originally sharing I/O devices between computers
 - E.g., printers
- Then communicating between computers
 - E.g., file transfer protocol
- Then communicating between people
 - E.g., e-mail
- Then communicating between networks of computers
 - E.g., file sharing, www, ...

The Internet (1962)

- History
 - 1963: JCR Licklider, while at DoD's ARPA, writes a memo describing desire to connect the computers at various research universities: Stanford, Berkeley, UCLA, ...
 - 1969 : ARPA deploys 4 "nodes" @ UCLA, SRI, Utah, & UCSB
 - 1973 Robert Kahn & Vint Cerf invent TCP, now part of the Internet Protocol Suite

The image is a composite graphic. In the top left is a black and white portrait of JCR Licklider. To his right is a table titled "ASCII Alphabet" with two columns of letters and their corresponding 7-bit binary codes. Below the portrait is a network diagram showing four nodes: SRI (node 2), UCSB (node 3), UCLA (node 1), and Utah (node 4). Each node is connected to a central hub. Associated with these nodes are computer models: 940 for SRI, PDP 10 for Utah, and Sigma 7 for UCLA. A box labeled "360" is also connected to the UCSB node. Below the network diagram is a portrait of Vint Cerf and a quote: "Revolutions like this don't come along very often".

ASCII Alphabet			
A	1000001	N	1001110
B	1000010	O	1001111
C	1000011	P	1010000
D	1000100	Q	1010001
E	1000101	R	1010010
F	1000110	S	1010011
G	1000111	T	1010100
H	1001000	U	1010101
I	1001001	V	1010110
J	1001010	W	1010111
K	1001011	X	1011000
L	1001100	Y	1011001
M	1001101	Z	1011010

“Lick”

“Revolutions like this don't come along very often”
Vint Cerf

- Internet growth rates
 - Exponential since start!

www.greatachievements.org/?id=3736

en.wikipedia.org/wiki/Internet_Protocol_Suite

The World Wide Web (1989)

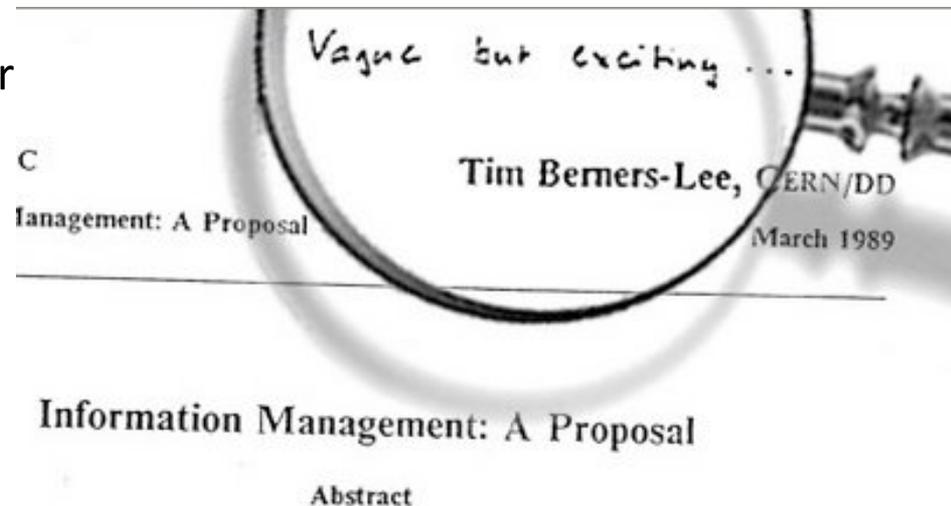
- “System of interlinked hypertext documents on the Internet”
- History
 - 1945: Vannevar Bush describes hypertext system called “memex” in article
 - 1989: Sir Tim Berners-Lee proposed and implemented the first successful communication between a Hypertext Transfer Protocol (HTTP) client and server using the internet.
 - ~2000 Dot-com entrepreneurs rushed in, 2001 bubble burst
- Today : Access anywhere!



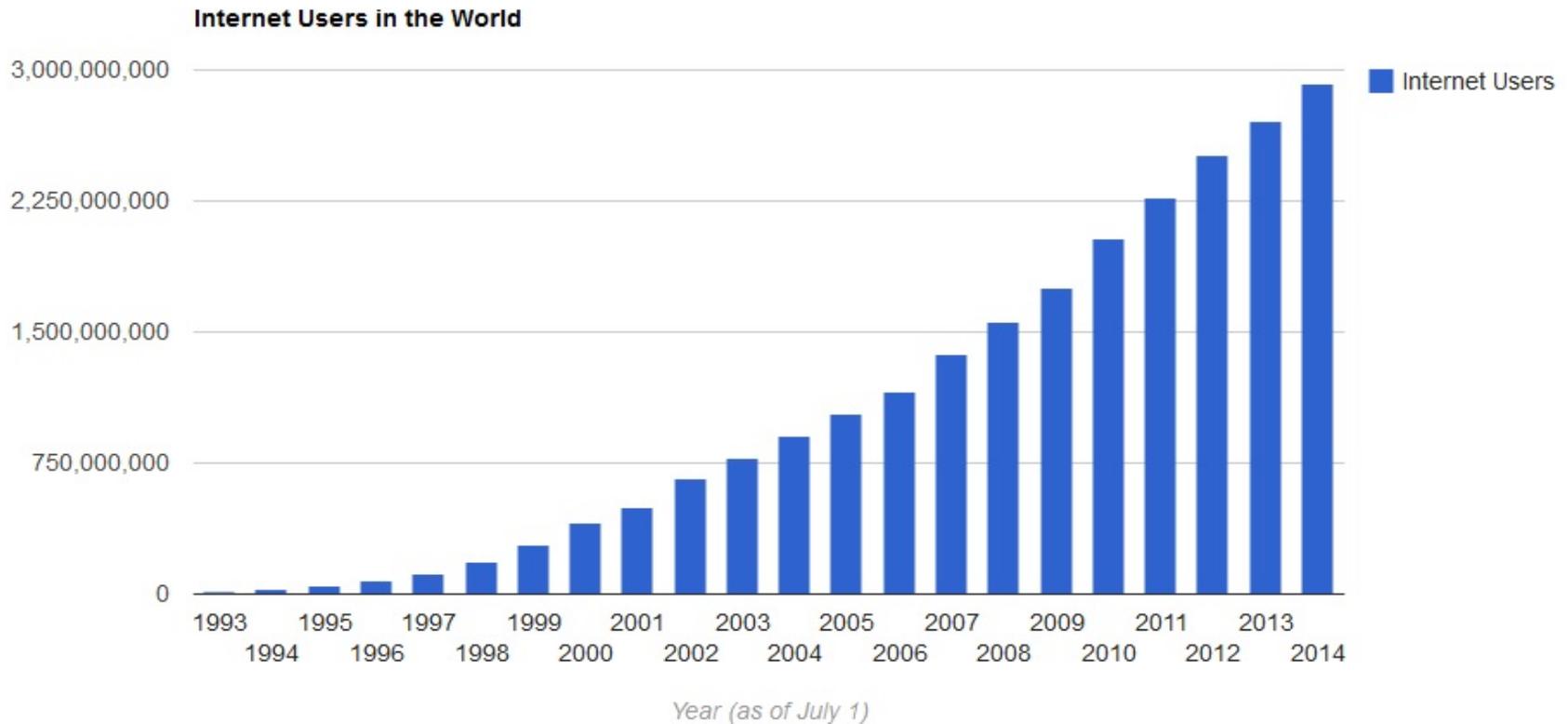
Tim Berners-Lee



World's First web server in 1990

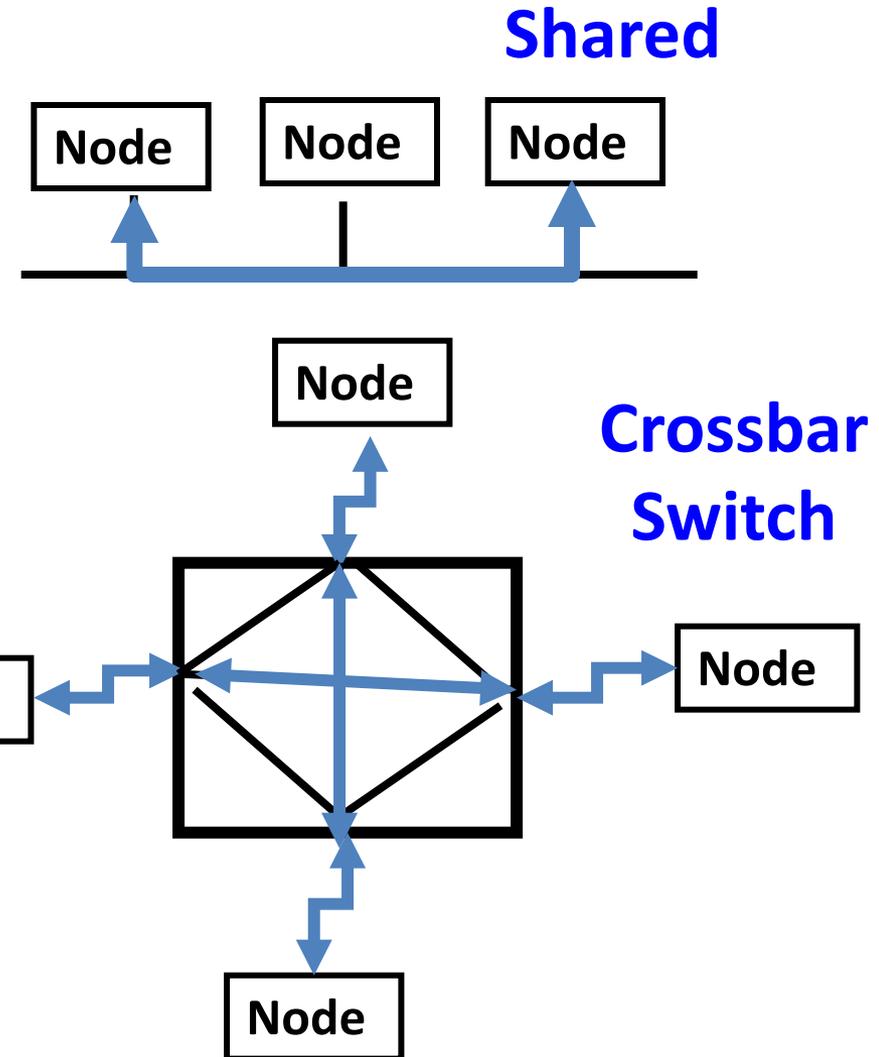


Growth ... exponential



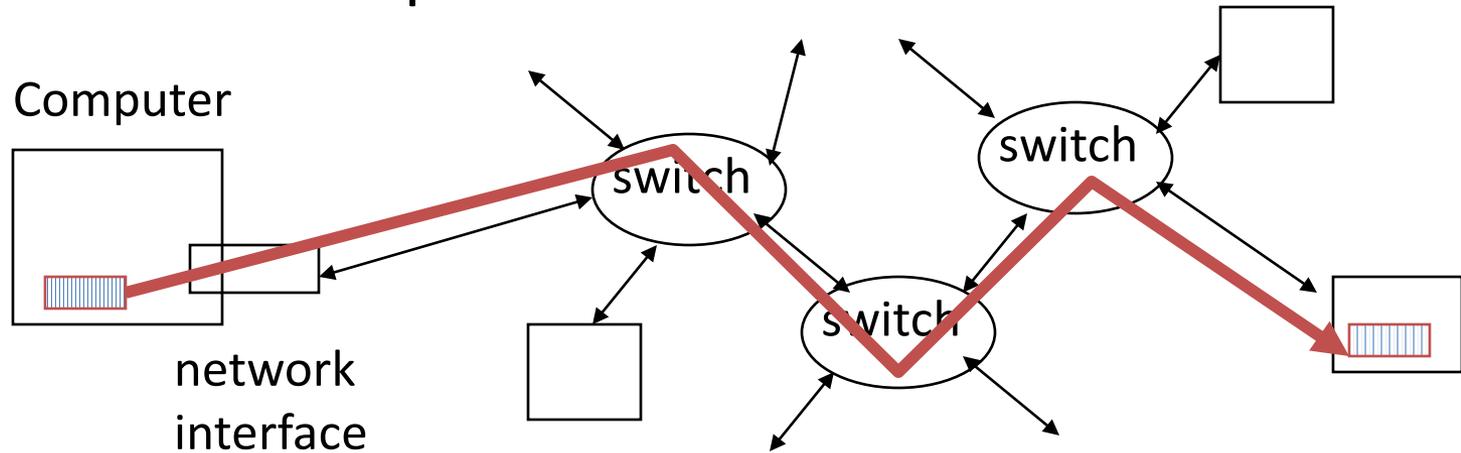
Shared vs. Switch-Based Networks

- Shared vs. Switched:
 - **Shared:** 1 at a time (CSMA/CD)
 - **Switched:** pairs (“point-to-point” connections) communicate at same time
- Aggregate bandwidth (BW) in switched network is many times that of shared:
 - point-to-point faster since no arbitration, simpler interface



What makes networks work?

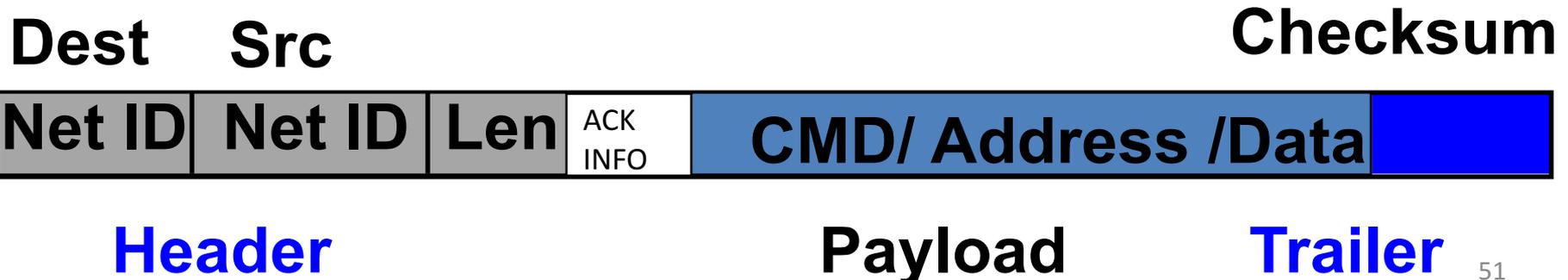
- Links connecting switches and/or routers to each other and to computers or devices



- Ability to name the components and to route packets of information - messages - from a source to a destination
- Layering, redundancy, protocols, and encapsulation as means of abstraction (big idea in Computer Architecture)

Software Protocol to Send and Receive

- SW Send steps
 - 1: Application copies data to OS buffer
 - 2: OS calculates checksum, starts timer
 - 3: OS sends data to network interface HW and says start
- SW Receive steps
 - 3: OS copies data from network interface HW to OS buffer
 - 2: OS calculates checksum, if OK, send ACK; if not, [delete message](#) (sender resends when timer expires)
 - 1: If OK, OS copies data to user address space, & signals application to continue



Protocols for Networks of Networks?

What does it take to send packets across the globe?

- Bits on wire or air
- Packets on wire or air
- Delivery packets within a single physical network
- Deliver packets across multiple networks
- Ensure the destination received the data
- Create data at the sender and make use of the data at the receiver

Protocol for Networks of Networks?

Lots to do and at multiple levels!

Use abstraction to cope with complexity of communication

- Hierarchy of layers:
 - Application (chat client, game, etc.)
 - Transport (TCP, UDP)
 - Network (IP)
 - Data Link Layer (Ethernet)
 - Physical Link (copper, wireless, etc.)

Protocol Family Concept

- *Protocol*: packet structure and control commands to manage communication
- *Protocol families (suites)*: a set of cooperating protocols that implement the network stack
- Key to **protocol families** is that communication occurs **logically** at the same level of the protocol, called **peer-to-peer**...

...but is **implemented via services** at the next lower level
- **Encapsulation**: carry higher level information within lower level “envelope”

Inspiration...

- CEO A writes letter to CEO B
 - Folds letter and hands it to assistant

Dear John,
Assistant:

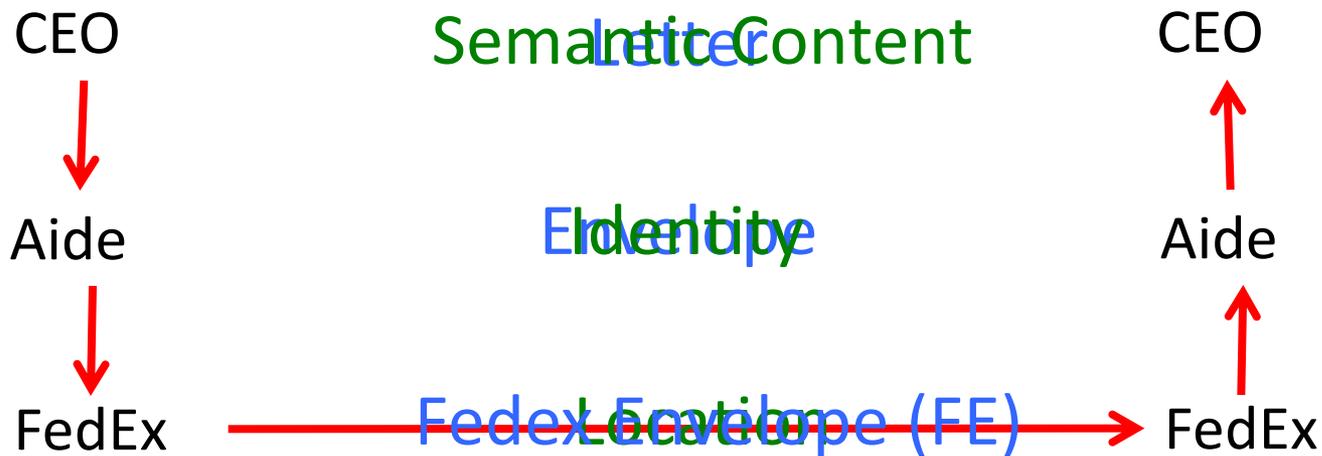
- Puts letter in envelope with CEO B's full name

Your days are numbered.
– Takes to FedEx

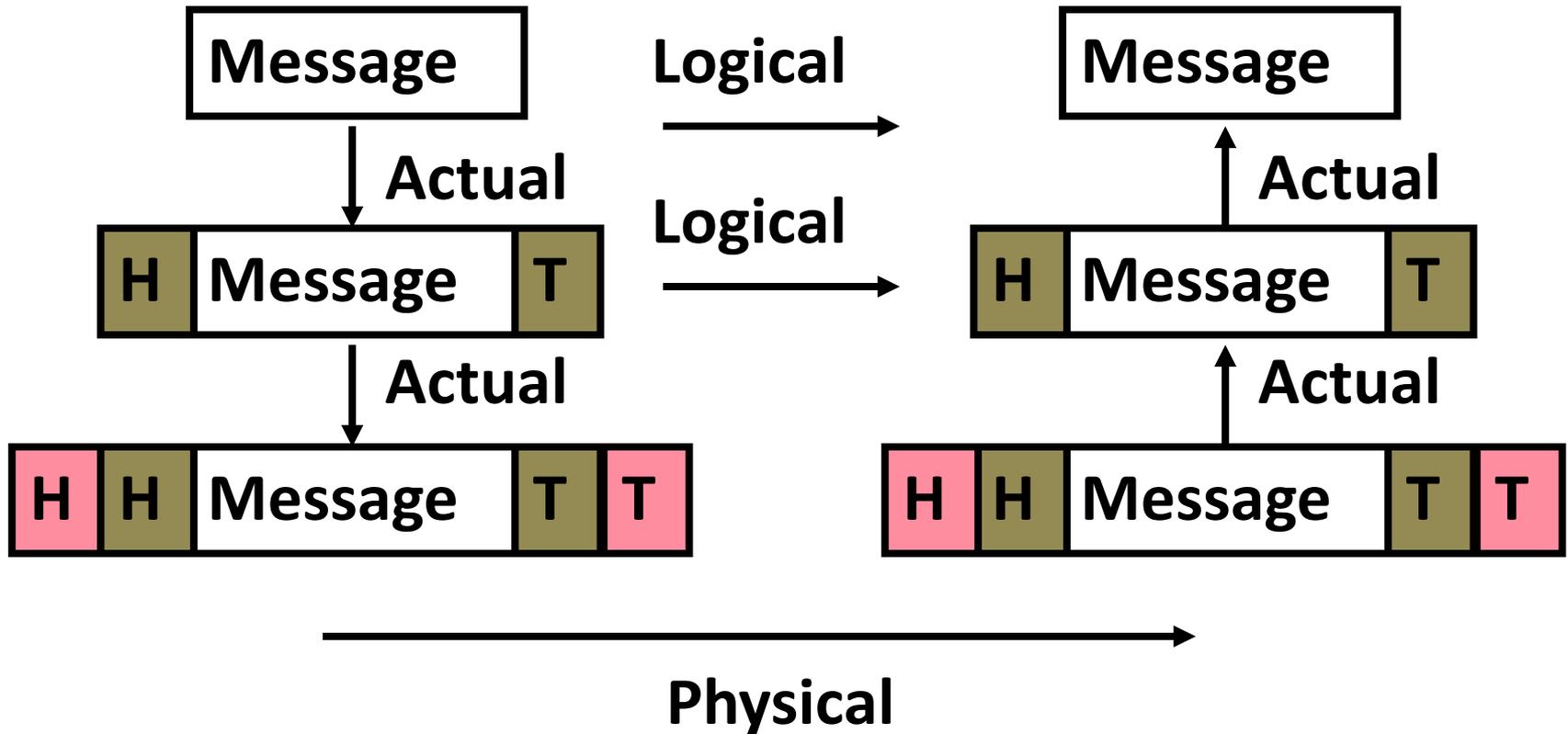
- FedEx Office
 - Puts letter in larger envelope
 - Puts ***Date*** and street address on FedEx envelope
 - Puts package on FedEx delivery truck
- FedEx delivers to other company

The Path of the Letter

“Peers” on each side understand the same things
No one else needs to
Lowest level has most packaging



Protocol Family Concept



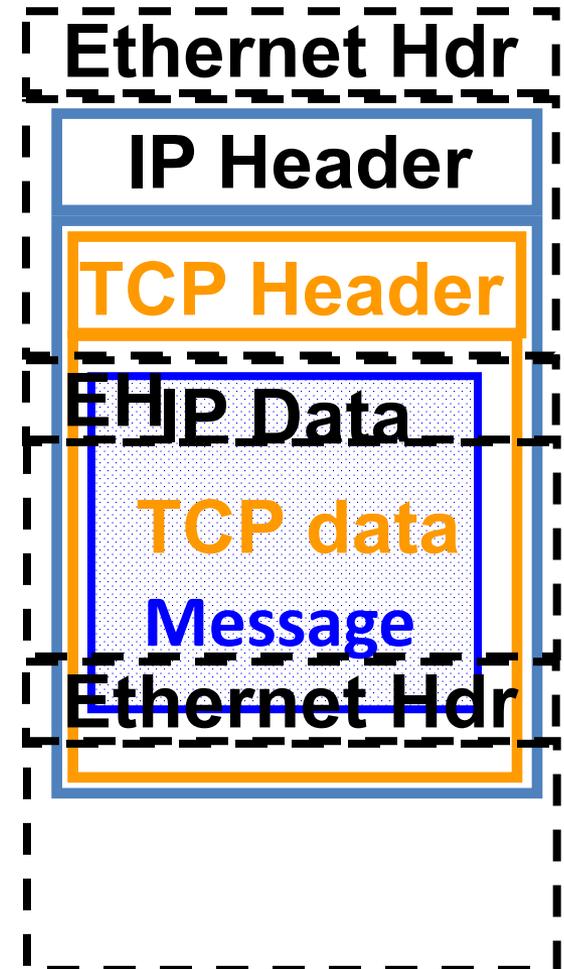
Each lower level of stack “encapsulates” information from layer above by adding header and trailer.

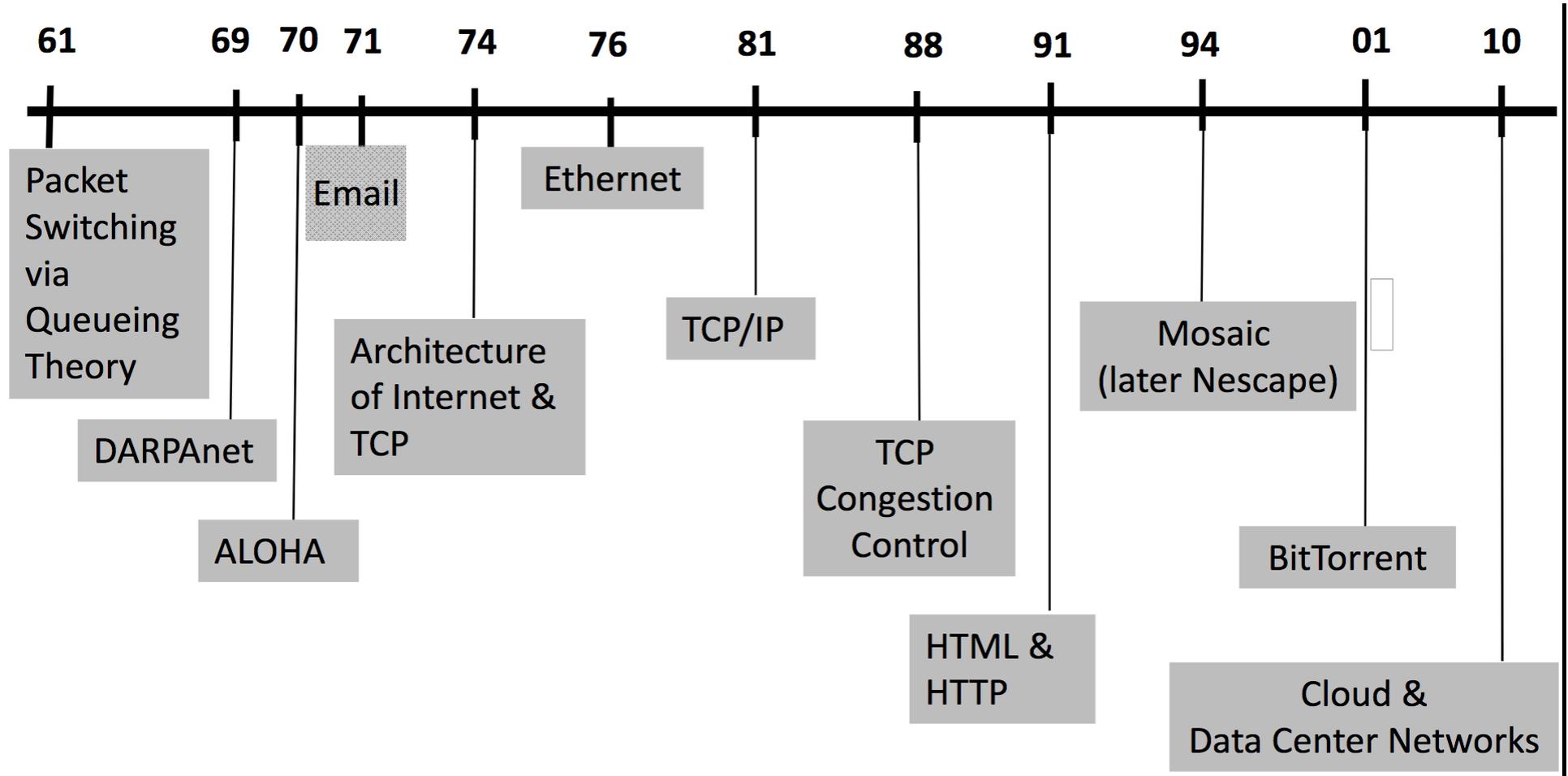
Most Popular Protocol for Network of Networks

- Transmission Control Protocol/Internet Protocol (TCP/IP)
- This protocol family is the **basis of the Internet**, a WAN (wide area network) protocol
 - IP makes best effort to deliver
 - Packets can be lost, corrupted
 - TCP guarantees delivery
 - TCP/IP so popular it is used even when communicating locally: even across homogeneous LAN (local area network)
 - UDP/IP: video or sound streaming; video call....

TCP/IP packet, Ethernet packet, protocols

- Application sends message
- TCP breaks into 64KiB segments, adds 20B header
- IP adds 20B header, sends to network
- If Ethernet, broken into 1500B packets with headers, trailers





“And in conclusion...”

- I/O gives computers their 5 senses
- I/O speed range is 100-million to one
- Polling vs. Interrupts
- DMA to avoid wasting CPU time on data transfers
- Disks for persistent storage, replaced by flash
- Networks: computer-to-computer I/O
 - Protocol suites allow networking of heterogeneous components. Abstraction!!!

“And in conclusion...”

- I/O gives computers their 5 senses
- I/O speed range is 100-million to one
- Polling vs. Interrupts
- DMA to avoid wasting CPU time on data transfers
- Disks for persistent storage, replaced by flash