CMPE 411
Computer Architecture

Lecture 25

I/O Systems

November 28, 2017

www.csee.umbc.edu/~younis/CMPE411/CMPE411.htm

Lecture's Overview

❑ ***Previous Lecture:***

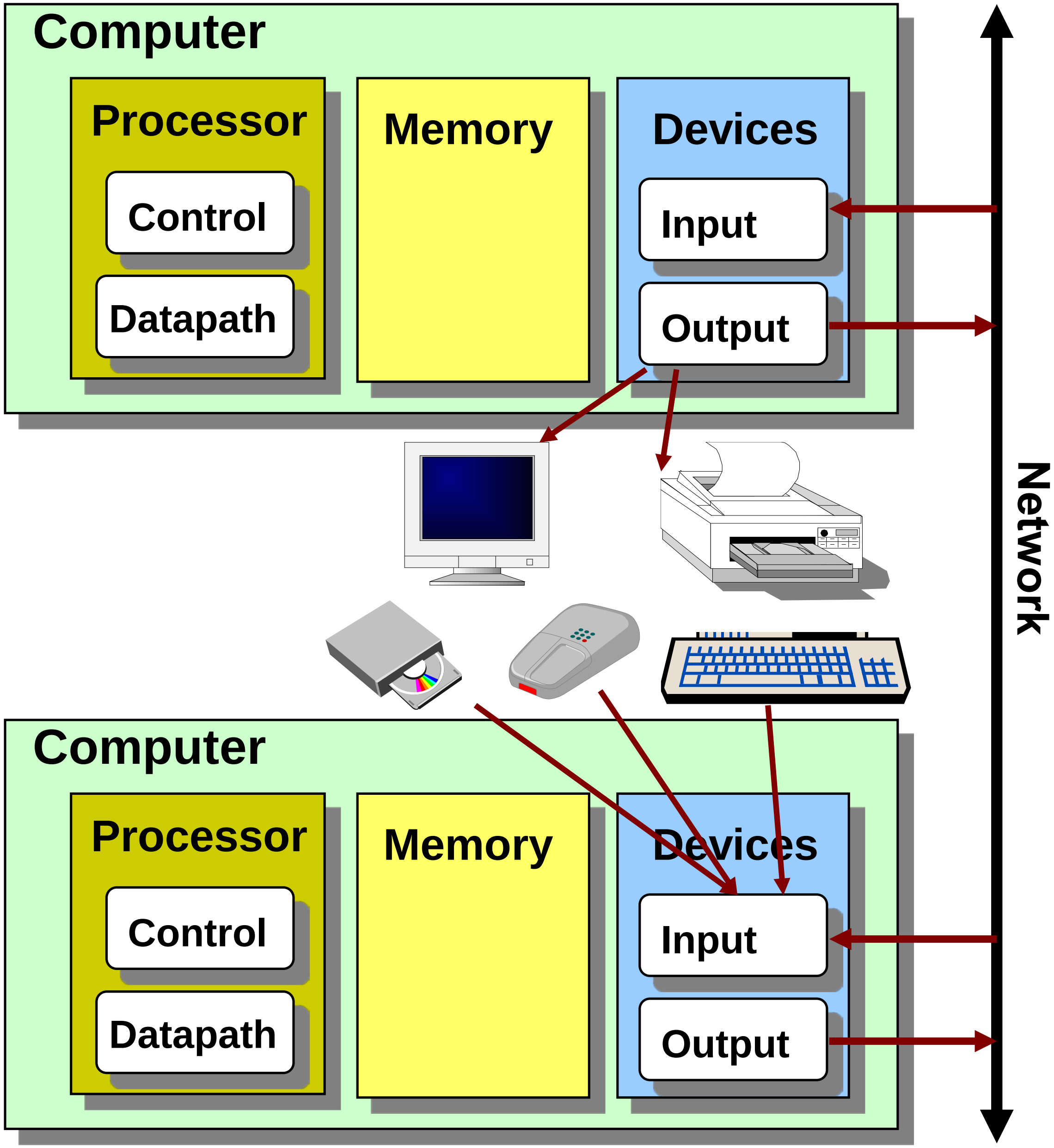
- Virtual Memory
  - ➔ Virtual addressing
  - ➔ Address translation
- Memory paging
  - ➔ Page table
  - ➔ Page faults
  - ➔ Translation look-aside buffer
- Memory-related exceptions
  - ➔ Relationship between TLB, cache miss and page fault exceptions
  - ➔ Handling of memory-related exceptions

❑ ***This Lecture:***

- I/O systems architecture
- Types and characteristics of I/O devices

Computer Input/Output

- ❑ I/O Interface
  - ➔ Device drivers
  - ➔ Device controller
  - ➔ Service queues
  - ➔ Interrupt handling
- ❑ Design Issues
  - ➔ Performance
  - ➔ Expandability
  - ➔ Standardization
  - ➔ Resilience to failure
- ❑ Impact on Tasks
  - ➔ Blocking conditions
  - ➔ Priority inversion
  - ➔ Access ordering



Impact of I/O on System Performance

Suppose we have a benchmark that executes in 100 seconds of elapsed time, where 90 seconds is CPU time and the rest is I/O time. If the CPU time improves by 50% per year for the next five years but I/O time does not improve, how much faster will our program run at the end of the five years?

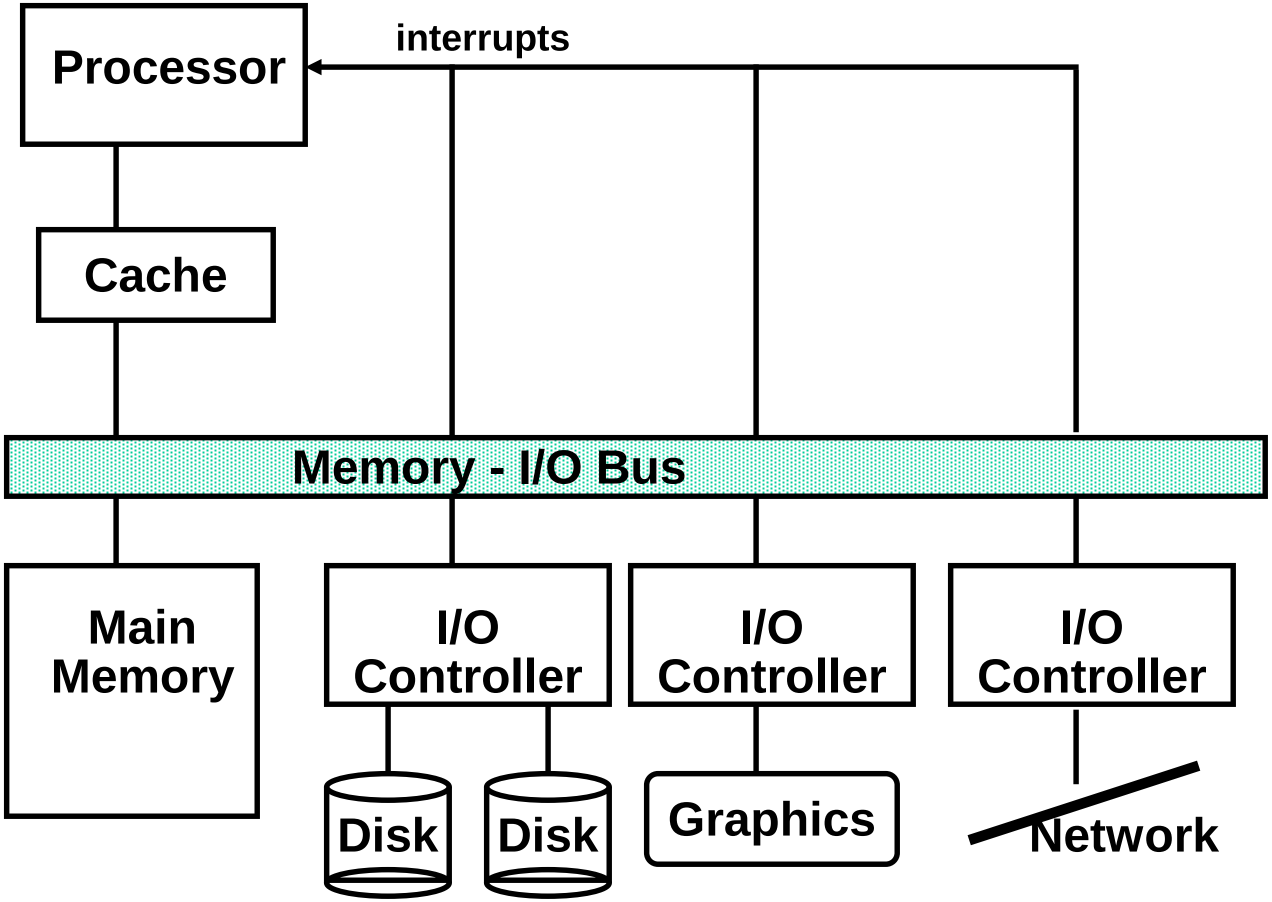
Answer: Elapsed Time = CPU time + I/O time

After n years	CPU time	I/O time	Elapsed time	% I/O time
0	90 Seconds	10 Seconds	100 Seconds	10%
1	$\frac{90}{1.5} = 60$ Seconds	10 Seconds	70 Seconds	14%
2	$\frac{60}{1.5} = 40$ Seconds	10 Seconds	50 Seconds	20%
3	$\frac{40}{1.5} = 27$ Seconds	10 Seconds	37 Seconds	27%
4	$\frac{27}{1.5} = 18$ Seconds	10 Seconds	28 Seconds	36%
5	$\frac{18}{1.5} = 12$ Seconds	10 Seconds	22 Seconds	45%

Over five years:

CPU improvement = 90/12 = 7.5 **BUT** System improvement = 100/22 = 4.5

Typical I/O System



- The connection between the I/O devices, processor, and memory are usually called (local or internal) *buses*

- Communication among the devices and the processor use both protocols on the bus and interrupts

I/O Device Examples

Device	Behavior	Partner	Data Rate (KB/sec)
Keyboard	Input	Human	0.01
Mouse	Input	Human	0.02
Line Printer	Output	Human	1.00
Floppy disk	Storage	Machine	50.00
Laser Printer	Output	Human	100.00
Optical Disk	Storage	Machine	500.00
Magnetic Disk	Storage	Machine	5,000.00
Network-LAN	Input or Output	Machine	20 – 1,000.00
Graphics Display	Output	Human	30,000.00

* Slide is courtesy of Dave Patterson

I/O System Performance

❑ I/O System performance depends on many aspects of the system ("limited by weakest link in the chain"):

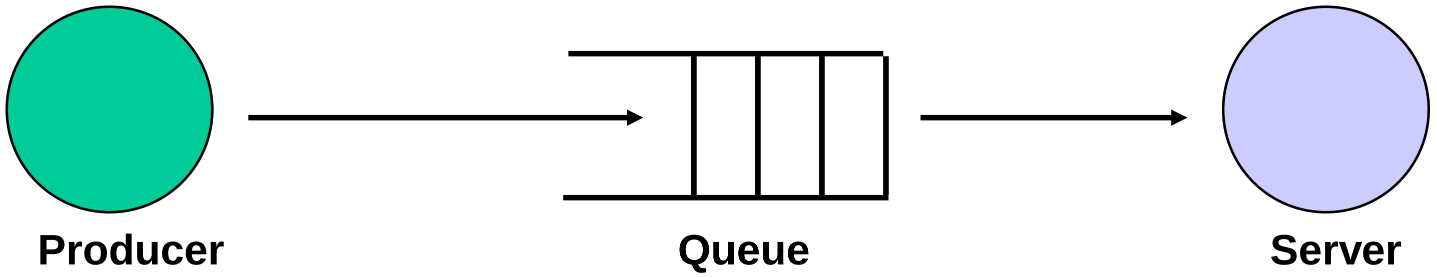
- ➔ The CPU
- ➔ The memory system:
  - Internal and external caches
  - Main Memory
- ➔ The underlying interconnection (buses)
- ➔ The I/O controller
- ➔ The I/O device
- ➔ The speed of the I/O software (Operating System)
- ➔ The efficiency of the software's use of the I/O devices

❑ Two common performance metrics:

- ➔ *Throughput*: I/O bandwidth
- ➔ *Response time*: Latency

* Slide is courtesy of Dave Patterson

Simple Producer-Server Model



❑ *Throughput*:

➔ The number of tasks completed by the server in unit time

➔ In order to get the highest possible throughput:

- The server should never be idle
- The queue should never be empty

❑ *Response time*:

➔ Begins when a task is placed in the queue

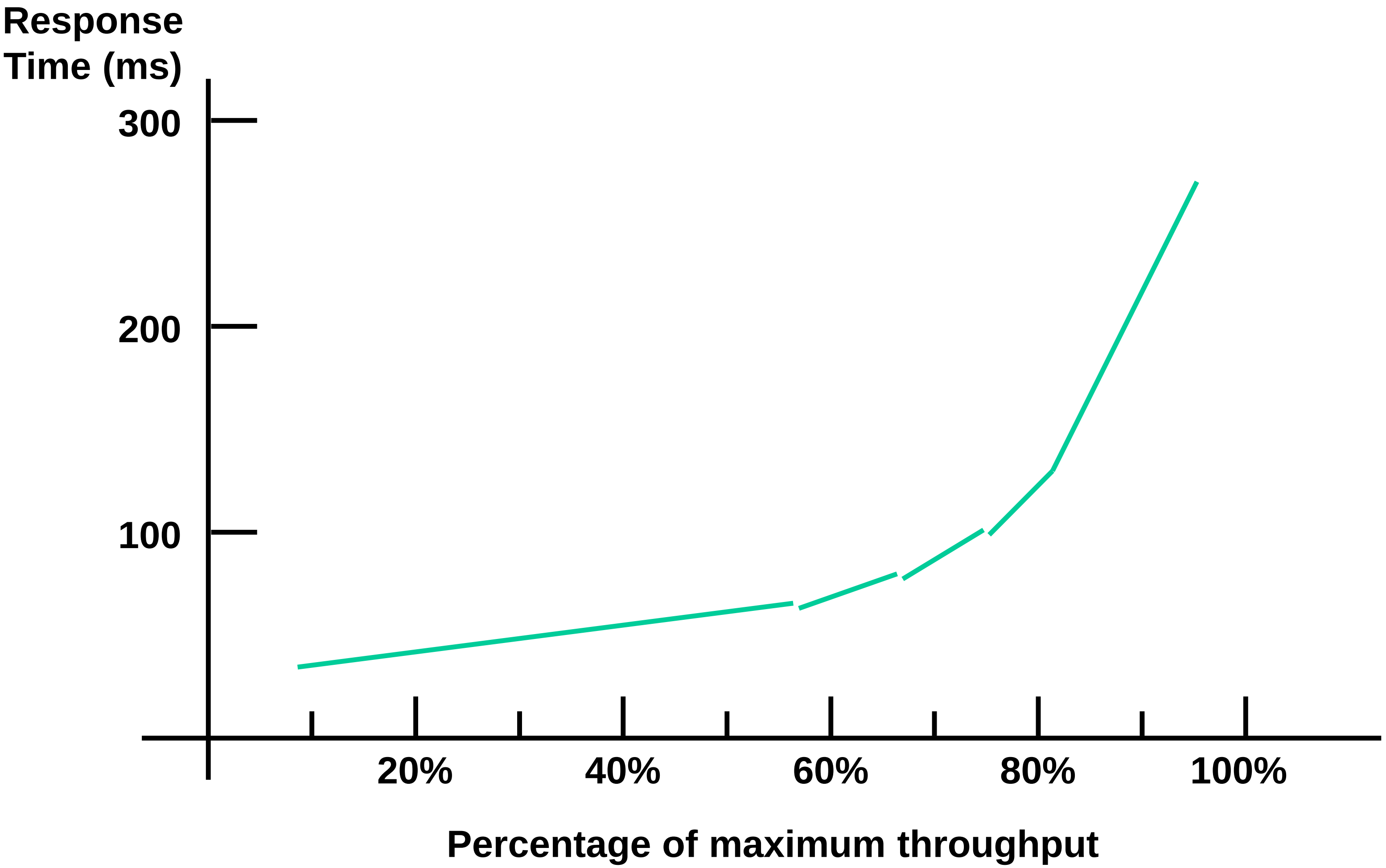
➔ Ends when it is completed by the server

➔ In order to minimize the response time:

- The queue should be empty
- The server will be idle

* Slide is courtesy of Dave Patterson

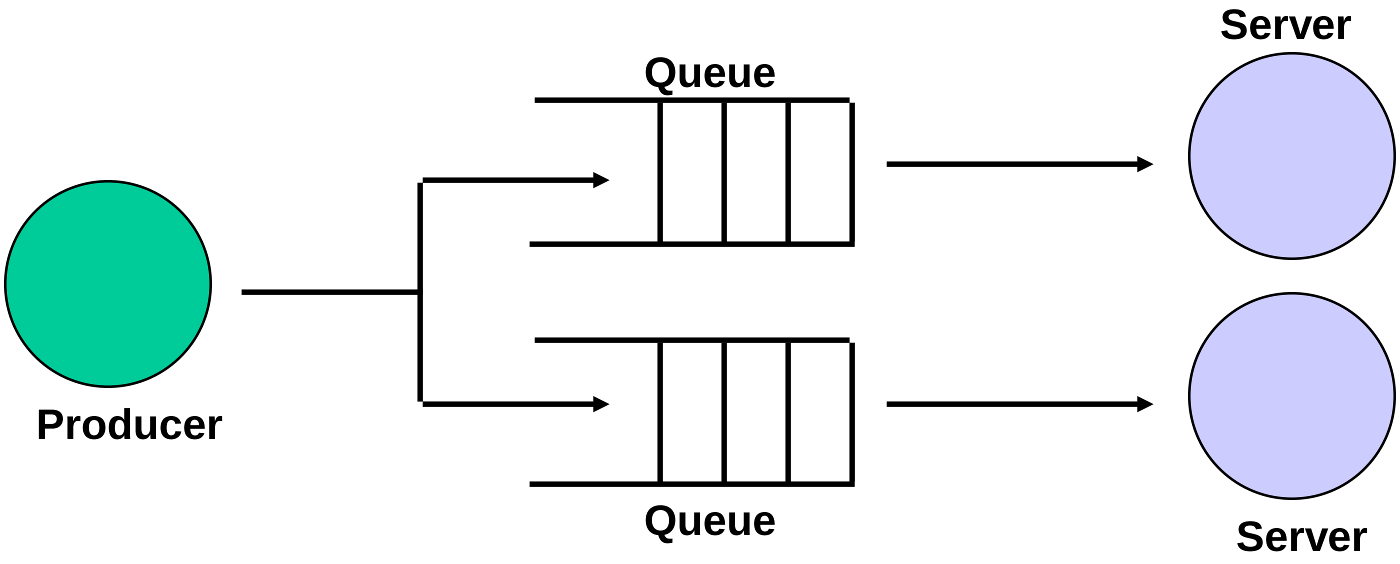
Throughput versus Respond Time



Low response time is user-desirable but leads to low throughput that is system-*Un*desirable (low device utilization)

* Slide is courtesy of Dave Patterson

Throughput Enhancement



- ❑ In general throughput can be improved by:
  - ➔ Throwing more hardware at the problem
  - ➔ reduces load-related latency
- ❑ Response time is much harder to reduce:
  - ➔ Ultimately it is limited by the mechanical subsystems

* Slide is courtesy of Dave Patterson

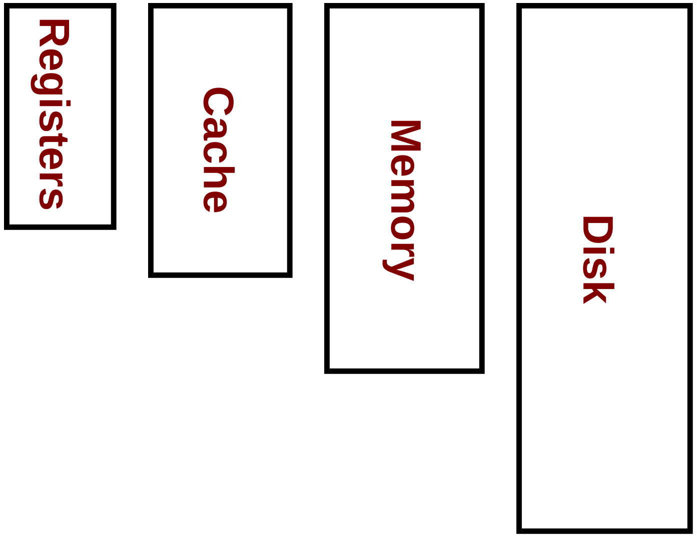
I/O Benchmarks for Magnetic Disks

- *Supercomputer application*:
  - ➔ Large-scale scientific problems => large files
  - ➔ One large read and many small writes to snapshot computation
  - ➔ Data Rate: MB/second between memory and disk
- *Transaction processing*:
  - ➔ Examples: Airline reservations systems and bank ATMs
  - ➔ Small changes to large shared database
  - ➔ I/O Rate: Number of disk accesses / second given upper limit for latency
- *File system*:
  - ➔ Measurements of UNIX file systems in an engineering environment:
    - 80% of accesses are to files less than 10 KB
    - 90% of all file accesses are to data with sequential disk addresses
    - 67% of the accesses are reads, 27% writes, 6% read-write
  - ➔ I/O Rate & Latency: Number of accesses /second and response time

* Slide is courtesy of Dave Patterson

Magnetic Disk

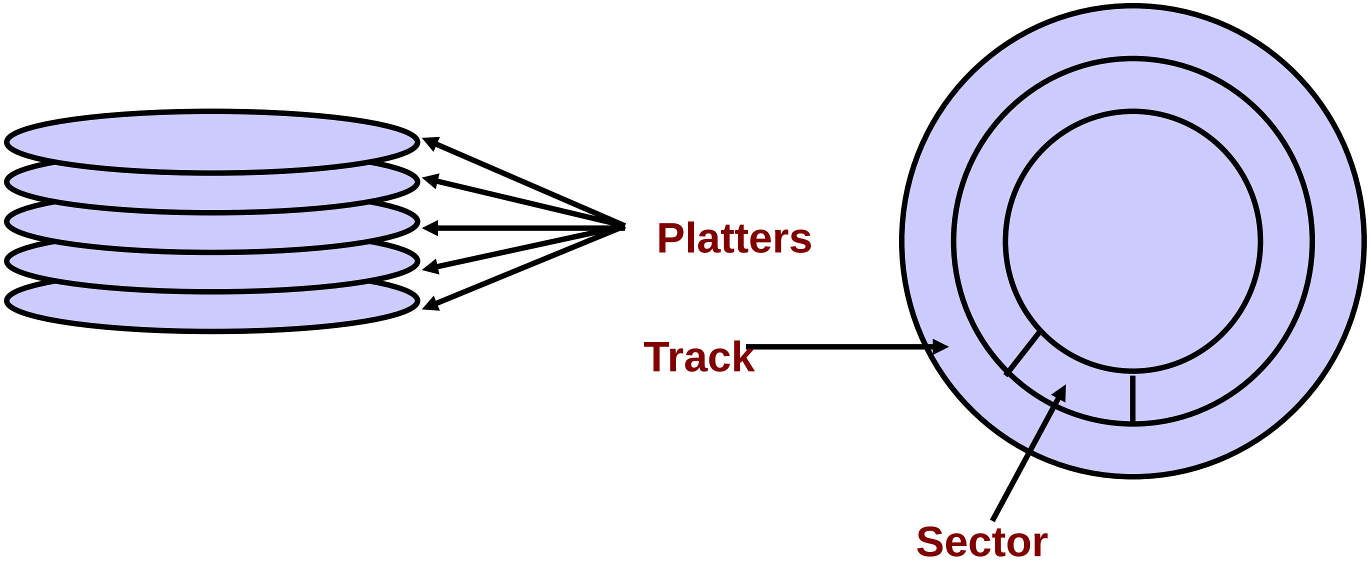
- Purpose:
  - Long term, nonvolatile storage
  - Large, inexpensive, and slow
  - Low level in the memory hierarchy



- Two major types:
  - Floppy disk
  - Hard disk
- Both types of disks:
  - Rely on a rotating platter coated with a magnetic surface
  - Use a moveable read/write head to access the disk
- Advantages of hard disks over floppy disks:
  - Platters are more rigid (metal or glass) so they can be larger
  - Higher density because it can be controlled more precisely
  - Higher data rate because it spins faster
  - Can incorporate more than one platter

* Slide is courtesy of Dave Patterson

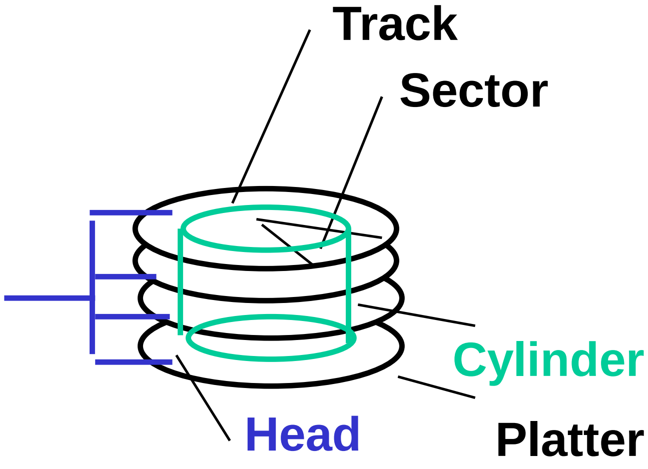
Organization of a Hard Magnetic Disk



- ❑ Typical numbers (depending on the disk size):
  - ➔ 500 to 2,000 tracks per surface
  - ➔ 32 to 128 sectors per track
    - A sector is the smallest unit that can be read or written to
- ❑ Traditionally all tracks have the same number of sectors:
  - ➔ Constant bit density: record more sectors on the outer tracks
  - ➔ Recently relaxed: constant bit size, speed varies with track location

* Slide is courtesy of Dave Patterson

Magnetic Disk Operation

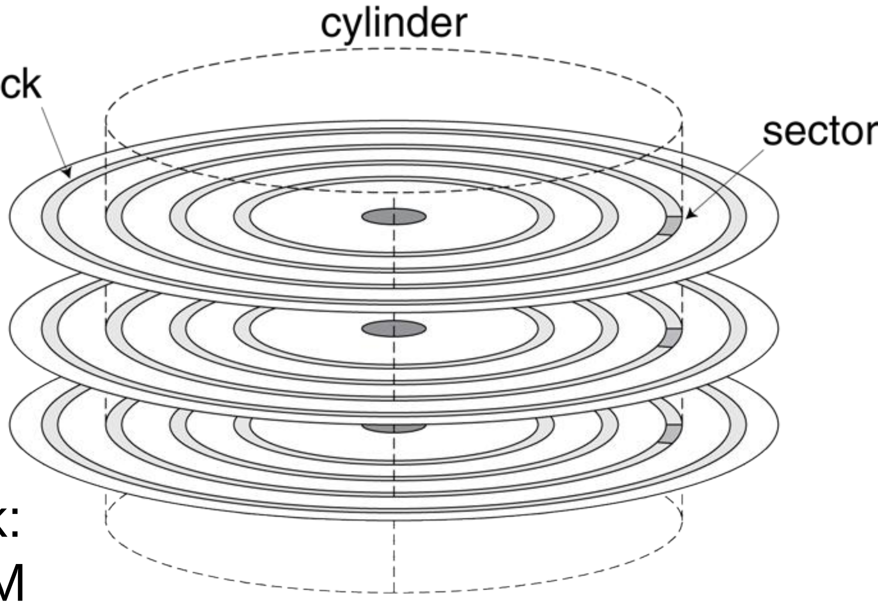


- ❑ Cylinder: all the tacks under the head at a given point on all surface
- ❑ Read/write data is a three-stage process:
  - ➔ Seek time: position the arm over proper track
  - ➔ Rotational latency: wait for the desired sector to rotate under the read/write head
  - ➔ Transfer time: transfer a block of bits (sector) under the read-write head
- ❑ Average seek time
  - ➔ (Sum of the time for all possible seek) / (total # of possible seeks)
  - ➔ Typically in the range of 8 ms to 12 ms (as reported by the industry)
  - ➔ Due to locality of disk reference, actual average seek time may only be 25% to 33% of the advertised number

* Slide is courtesy of Dave Patterson

Magnetic Disk Characteristic

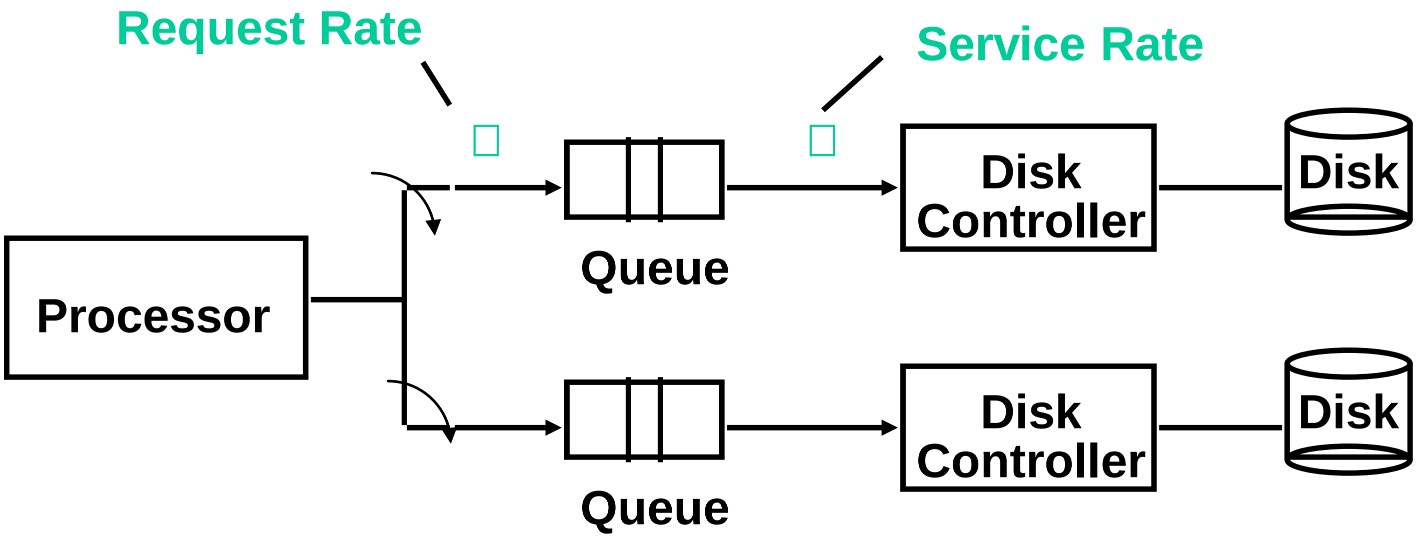
- ❑ *Rotational Latency:*
  - ➔ Most disks rotate at a speed of 3,600 to 7,200 RPM
  - ➔ Approximately 16 ms to 8 ms per revolution, respectively
  - ➔ An average latency to the desired information is halfway around the disk: 8 ms at 3600 RPM, 4 ms at 7200 RPM



- ❑ *Transfer Time* is a function of :
  - ➔ Transfer size (usually a sector): 1 KB / sector
  - ➔ Rotation speed: 3600 RPM to 7200 RPM
  - ➔ Recording density: bits per inch on a track
  - ➔ Diameter: typical diameter ranges from 2.5 to 5.25 in
  - ➔ Typical values: 2 to 12 MB per second

* Slide is courtesy of Dave Patterson

Disk I/O Performance



- Disk Access Time = Seek time + Rotational Latency + Transfer time + Controller Time + Queuing Delay
- Estimating Queue Length:
  - ➔ Utilization = U = Request Rate / Service Rate
  - ➔ Mean Queue Length = U / (1 - U)
  - ➔ Request Rate grows ⇒ Service Rate diminishes
    - Mean Queue Length ⇒ Infinity

* Slide is courtesy of Dave Patterson

Example

Calculate the access time for a disk with 512 byte/sector and 12 ms advertised seek time. The disk rotates at 5400 RPM and transfers data at a rate of 4MB/sec. The controller overhead is 1 ms. Assume that the queue is idle (so no service time)

Answer:

Disk Access Time = Seek time + Rotational Latency + Transfer time + Controller Time + Queuing Delay

= 12 ms + 0.5 / 5400 RPM + 0.5 KB / 4 MB/s + 1 ms + 0

= 12 ms + 0.5 / 90 RPS + 0.125 / 1024 s + 1 ms + 0

= 12 ms + 5.5 ms + 0.1 ms + 1 ms + 0 ms

= 18.6 ms

If real seeks are 1/3 the advertised seeks, disk access time would be 10.6 ms, with rotation delay at 50% of the time!

* Slide is courtesy of Dave Patterson

Historical Trend

Characteristics	IBM 3090	IBM UltraStar	Integral 1820
Disk diameter (inches)	10.88	3.50	1.80
Formatted data capacity (MB)	22,700	4,300	21
MTTF (hours)	50,000	1,000,000	100,000
Number of arms/box	12	1	1
Rotation speed (RPM)	3,600	7,200	3,800
Transfer rate (MB/sec)	4.2	9-12	1.9
Power/box (watts)	2,900	13	2
MB/watt	8	102	10.5
Volume (cubic feet)	97	0.13	0.02
MB/cubic feet	234	33000	1050

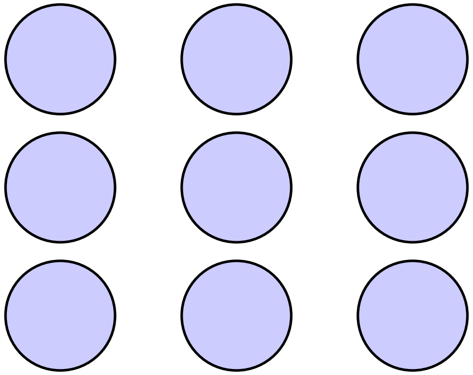
* Slide is courtesy of Dave Patterson

Reliability and Availability

- ❑ Two terms that are often confused:
  - ➔ Reliability: Is anything broken?
  - ➔ Availability: Is the system still available to the user?
- ❑ Availability can be improved by adding hardware:
  - ➔ Example: adding ECC on memory
- ❑ Reliability can only be improved by:
  - ➔ Enhancing environmental conditions
  - ➔ Building more reliable components
  - ➔ Building with fewer components
    - Improve availability may come at the cost of lower reliability

* Slide is courtesy of Dave Patterson

Disk Arrays

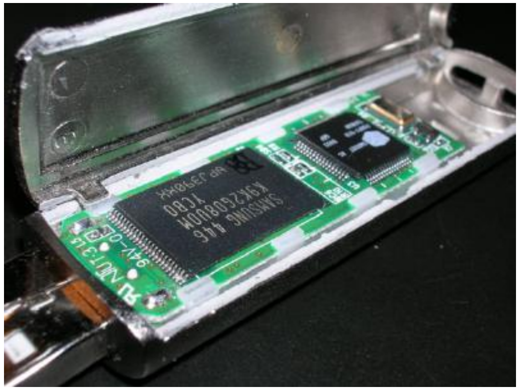


- ❑ A new organization of disk storage:
  - ➔ Arrays of small and inexpensive disks
  - ➔ Increase potential throughput by having many disk drives:
    - ➢ Data is spread over multiple disk
    - ➢ Multiple accesses are made to several disks
- ❑ Redundant Array of Inexpensive Disks (RIAD)
  - ➔ Widely available and used in today's market
  - ➔ Different levels based on the number of replicas
- ❑ Reliability is lower than a single disk:
  - ➔ But availability can be improved by adding redundant disks (RAID): Lost information can be reconstructed from redundant information
  - ➔ MTTR: mean time to repair is in the order of hours
  - ➔ MTTF: mean time to failure of disks is tens of years

* Slide is courtesy of Dave Patterson

Flash Storage

- ❑ Non-volatile semiconductor storage
  - ➔ 100× – 1000× faster than disk
  - ➔ Smaller, lower power, more robust
  - ➔ But more $/GB (between disk and DRAM)



Flash Type:

- ❑ NOR flash: bit cell like a NOR gate
  - ➔ Random read/write access
  - ➔ Used for instruction memory in embedded systems
- ❑ NAND flash: bit cell like a NAND gate
  - ➔ Denser (bits/area), but block-at-a-time access
  - ➔ Cheaper per GB
  - ➔ Used for USB keys, media storage, …
- ❑ Flash bits wears out after 1000's of accesses
  - ➔ Not suitable for direct RAM or disk replacement
  - ➔ Wear levelling: remap data to less used blocks

* Slide is courtesy of Morgan Kaufmann Publishers

Conclusion

- ***Summary***
  - I/O systems architecture
    - I/O role and interface to the processor
    - I/O design issues
  - I/O devices
    - Types and characteristics
    - Performance metrics and optimization factors
    - I/O benchmarks
  - Magnetic Disk
    - Access time and performance characteristics
    - Theory of operation and historical trend
    - Disk non-functional attributes (reliability and availability)
- ***Next Lecture***
  - Memory to processor interconnect

Read Sections 6.1-6.3 in 4th Ed. of the textbook