CS370 Operating Systems

Colorado State University
Yashwant K Malaiya

Fall 2016 Lecture 35

Mass Storage

Slides based on

- **Text by Silberschatz, Galvin, Gagne**
- **Various sources**

Questions For You

- Local/Global replacement policy: what is used more commonly?
- Locality we had seen: was it temporal or spatial?
- Why page size needs to be 2^n?
- Can virtual memory for a process be larger than the entire physical memory?
- If the secondary memory used the same technology as the main memory, would it still be slower?

Virtual Memory

- Address spaces and demand paging
- Page Replacement Algorithms
- Page Buffering
- Frame Allocation
- Page size issues

Windows

- Uses demand paging with **clustering**. Clustering brings in pages surrounding the faulting page
- Processes are assigned **working set minimum** and **working set maximum**
- Working set minimum is the minimum number of pages the process is guaranteed to have in memory
- A process may be assigned as many pages up to its working set maximum
- When the amount of free memory in the system falls below a threshold, **automatic working set trimming** is performed to restore the amount of free memory
- Working set trimming removes pages from processes that have pages in excess of their working set minimum

CS370 Operating Systems

Colorado State University
Yashwant K Malaiya

Fall 2016 Lecture 35+

Mass Storage

Slides based on

- Text by Silberschatz, Galvin, Gagne
- Various sources

Chapter 10: Mass-Storage Systems

- Physical structure of secondary storage devices and its effects on the uses of the devices
- Performance characteristics of mass-storage devices
- Disk scheduling algorithms
- Operating-system services provided for mass storage, including RAID

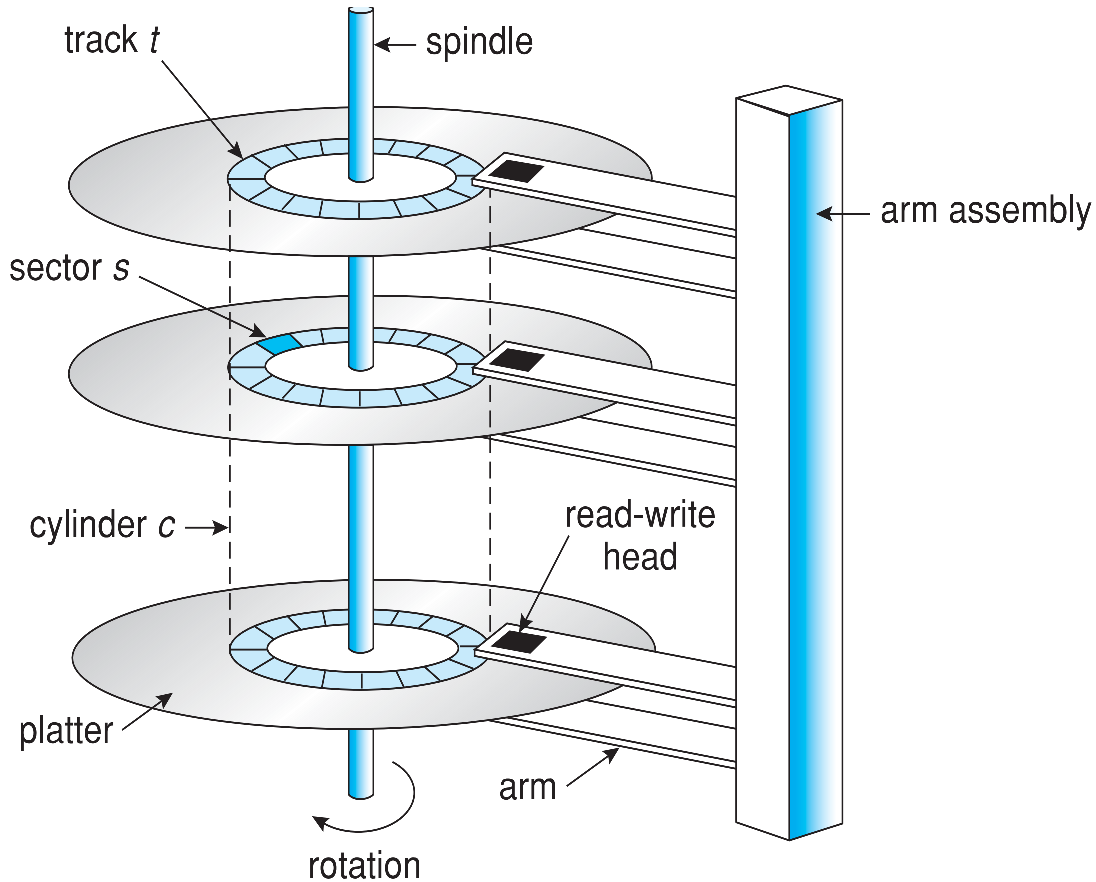
Objectives

- The physical structure of secondary storage devices and its effects on the uses of the devices
- To explain the performance characteristics of mass-storage devices
- To evaluate disk scheduling algorithms
- To discuss operating-system services provided for mass storage, including RAID

Overview of Mass Storage Structure

- **Magnetic disks** provide bulk of secondary storage of modern computers
  - Drives rotate at 60 to 250 times per second
  - **Transfer rate** is rate at which data flow between drive and computer
  - **Positioning time** (**random-access time**) is time to move disk arm to desired cylinder (**seek time**) and time for desired sector to rotate under the disk head (**rotational latency**)
  - **Head crash** results from disk head making contact with the disk surface -- That's bad
- Disks can be removable
- Drive attached to computer via **I/O bus**
  - Busses vary, including **EIDE**, **ATA**, **SATA**, **USB**, **Fibre Channel**, **SCSI, SAS, Firewire**
  - **Host controller** in computer uses bus to talk to **disk controller** built into drive or storage array

Moving-head Disk Mechanism



Hard Disks

- Platters range from .85" to 14" (historically)
  - Commonly 3.5", 2.5", and 1.8"
- Range from 30GB to **10TB** per drive
- Performance
  - Transfer Rate – theoretical – 6 G**b**/sec
  - Effective Transfer Rate – real – 1G**b**/sec (about 150 M**B**/s)
  - Seek time from 3ms to 12ms – 9ms common for desktop drives
  - Average seek time measured or calculated based on 1/3 of tracks
  - Latency based on spindle speed
    - 1 / (RPM / 60) = 60 / RPM
  - Average latency = ½ latency

Spindle [rpm]	Average latency [ms]
4200	7.14
5400	5.56
7200	4.17
10000	3
15000	2

(From Wikipedia)

Hard Disk Performance

- **Average access time** = average seek time + average latency
  - For fastest disk 3ms + 2ms = 5ms
  - For slow disk 9ms + 5.56ms = 14.56ms
- Average I/O time = average access time + (amount to transfer / transfer rate) + controller overhead
- For example to transfer a 4KB block on a 7200 RPM disk with a 5ms average seek time, 1Gb/sec transfer rate with a .1ms controller overhead =
  - 5ms + 4.17ms + 0.1ms + transfer time
  - Transfer time = 4KB / 1Gb/s = 4x8K/G = 0.031 ms
  - Average I/O time for 4KB block = 9.27ms + .031ms = 9.301ms

Question from last time

- Average I/O time = average access time + (amount to transfer / transfer rate) + controller overhead
- 4KB block on a 7200 RPM disk with a 5ms average seek time, 1Gb/sec transfer rate with a .1ms controller overhead =
  - 5ms + 4.17ms + 0.1ms + (4x8K/G) = 9.301ms
- Why are transfer rates usually measure in bits instead of bytes?
- Ans: Convention. Storage often measured in bytes.

The First Commercial Disk Drive

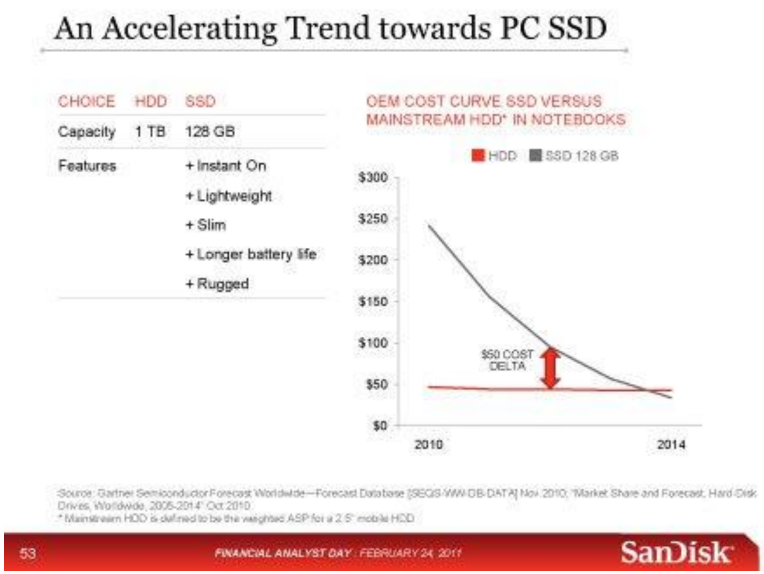
1956
IBM RAMDAC computer included the IBM Model 350 disk storage system

5M (7 bit) characters
50 x 24” platters
Access time = < 1 second

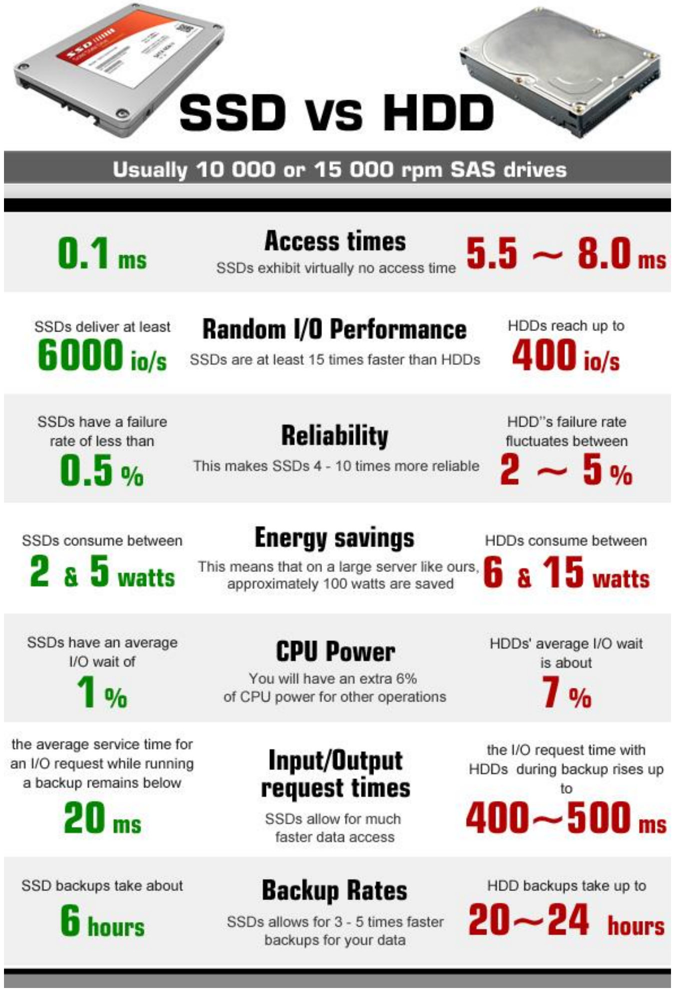
Solid-State Disks

- Nonvolatile memory used like a hard drive
  - Many technology variations
- Can be more reliable than HDDs
- More expensive per MB ($0.30/GB vs $0.05 for HD)
- Shorter life span (100,000 writes?) *Probably not*
- Capacity ? (up to 16 TB vs 8 TB for HD)
- But much faster (access time <0.1 millisec, transfer rate 100MB-GB/s)
- No moving parts, so no seek time or rotational latency
- Lower power consumption
- **"Breakthrough Technology" 3D Xpoint (1000 times faster) expected 2016**

SSD vs HDD

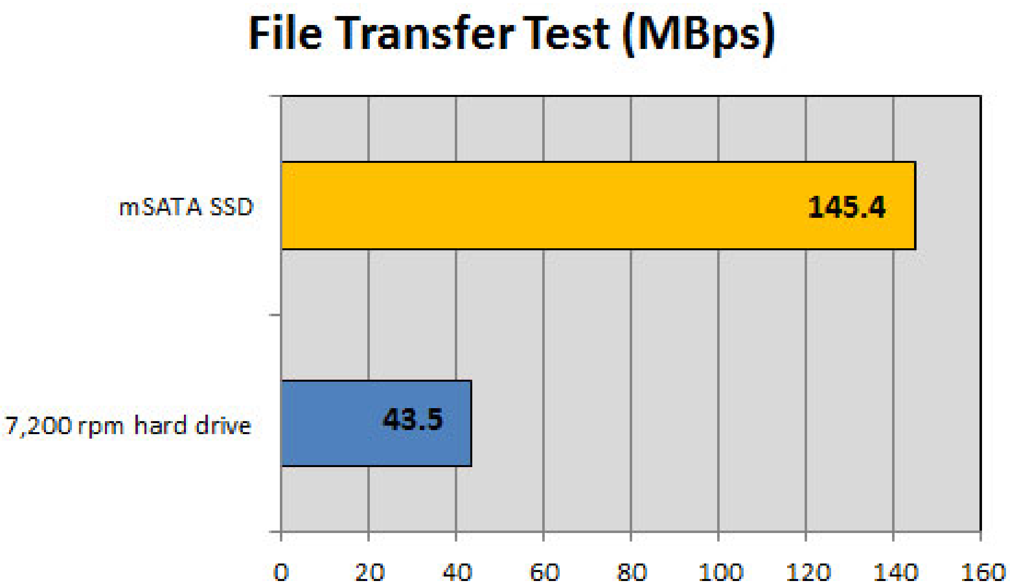


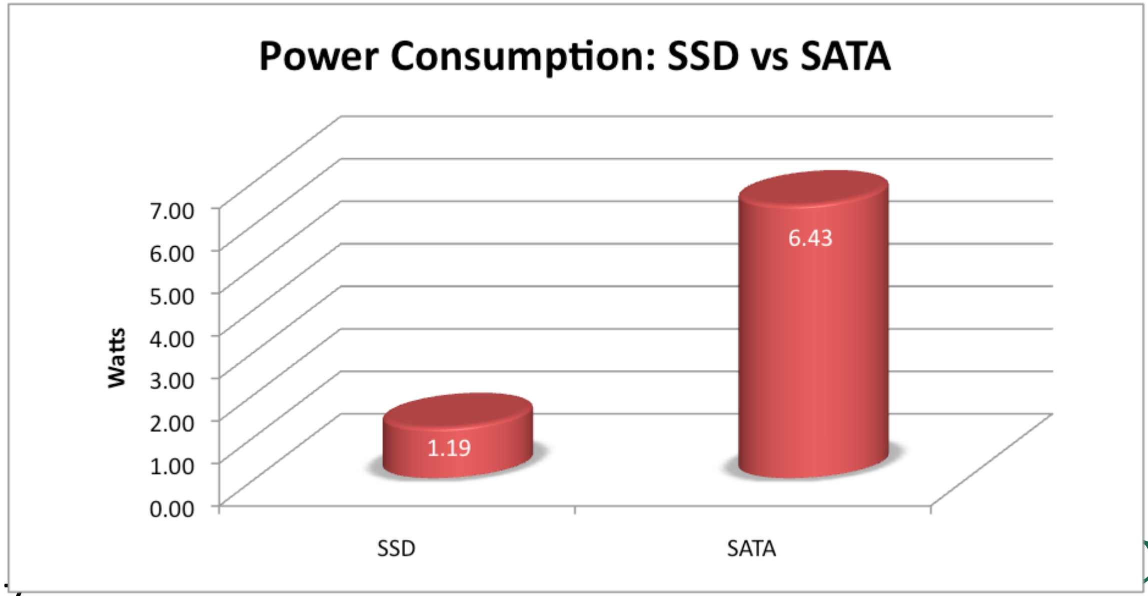
An Accelerating Trend towards PC SSD
CHOICE HDD SSD
Capacity 1 TB 128 GB
Features
+ Instant On
+ Lightweight
+ Slim
+ Longer battery life
+ Rugged
OEM COST CURVE SSD VERSUS MAINSTREAM HDD* IN NOTEBOOKS
HDD
SSD 128 GB
$300
$250
$200
$150
$100
$50
$0
2010
2014
$50 COST DELTA
Source: Gartner Semiconductor Forecast Worldwide—Forecast Database [SEQS-WW-DB-DATA] Nov 2010; "Market Share and Forecast, Hard Disk Drives, Worldwide, 2005-2014" Oct 2010
* Mainstream HDD is defined to be the weighted ASP for a 2.5" mobile HDD
53
FINANCIAL ANALYST DAY : FEBRUARY 24, 2011
SanDisk



SSD vs HDD
Usually 10 000 or 15 000 rpm SAS drives
0.1 ms
Access times
SSDs exhibit virtually no access time
5.5 ~ 8.0 ms
SSDs deliver at least
6000 io/s
Random I/O Performance
SSDs are at least 15 times faster than HDDs
HDDs reach up to
400 io/s
SSDs have a failure rate of less than
0.5 %
Reliability
This makes SSDs 4 - 10 times more reliable
HDD"s failure rate fluctuates between
2 ~ 5 %
SSDs consume between
2 & 5 watts
Energy savings
This means that on a large server like ours, approximately 100 watts are saved
HDDs consume between
6 & 15 watts
SSDs have an average I/O wait of
1 %
CPU Power
You will have an extra 6% of CPU power for other operations
HDDs' average I/O wait is about
7 %
the average service time for an I/O request while running a backup remains below
20 ms
Input/Output request times
SSDs allow for much faster data access
the I/O request time with HDDs during backup rises up to
400~500 ms
SSD backups take about
6 hours
Backup Rates
SSDs allows for 3 - 5 times faster backups for your data
HDD backups take up to
20~24 hours

Solid-State Disks





Magnetic Tape

- Was early secondary-storage medium (now tertiary)
  - Evolved from open spools to cartridges
- Relatively permanent and holds large quantities of data
- Access time slow
- Random access ~1000 times slower than disk
- Mainly used for backup, storage of infrequently-used data, transfer medium between systems
- Kept in spool and wound or rewound past read-write head
- Once data under head, transfer rates comparable to disk
  - 140MB/sec and greater
- 200GB to 1.5TB typical storage Sony: New 185 TB

Disk Structure

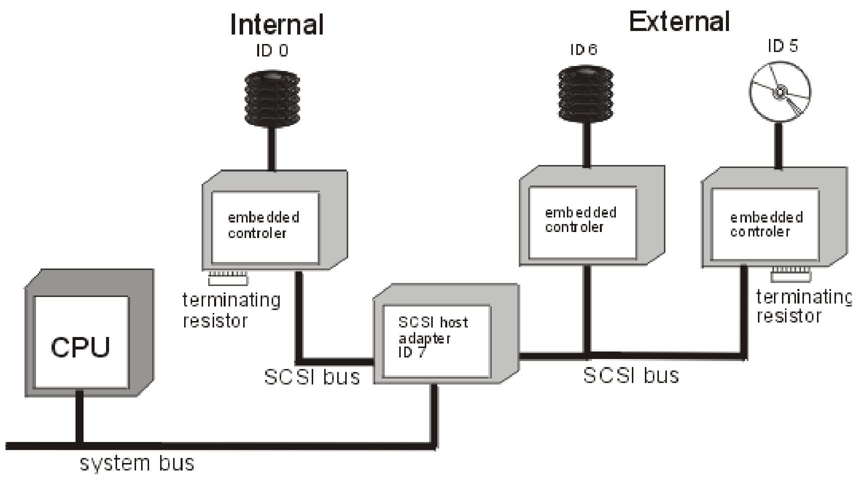
- Disk drives are addressed as large 1-dimensional arrays of **logical blocks**, where the logical block is the smallest unit of transfer
  - Low-level formatting creates **sectors** on physical media (typically 512 bytes)
- The 1-dimensional array of logical blocks is mapped into the sectors of the 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, and then through the rest of the cylinders from outermost to innermost
  - Logical to physical address should be easy
    - Except for bad sectors
    - Non-constant # of sectors per track via constant angular velocity

Disk Formatting

- Low-level formatting marks the surfaces of the disks with markers indicating the start of a recording block (sector markers) and other informationby the disk controller to read or write data.
- Partitioning divides a disk into one or more regions, writing data structures to the disk to indicate the beginning and end of the regions. Often includes checking for defective tracks/sectors.
- High-level formatting creates the file system format within a disk partition or a logical volume. This formatting includes the data structures used by the OS to identify the logical drive or partition's contents.

Disk Attachment: I/O busses

- Host-attached storage accessed through I/O ports talking to I/O busses
- SCSI itself is a bus, up to 16 devices on one cable, **SCSI initiator** (adapter) requests operation and **SCSI targets** (controller) perform tasks
  - Each target can have up to 8 **logical units** (disks attached to device controller)
- FC (fibre channel) is high-speed serial architecture
  - Can be switched fabric with 24-bit address space – the basis of **storage area networks (SANs)** in which many hosts attach to many storage units

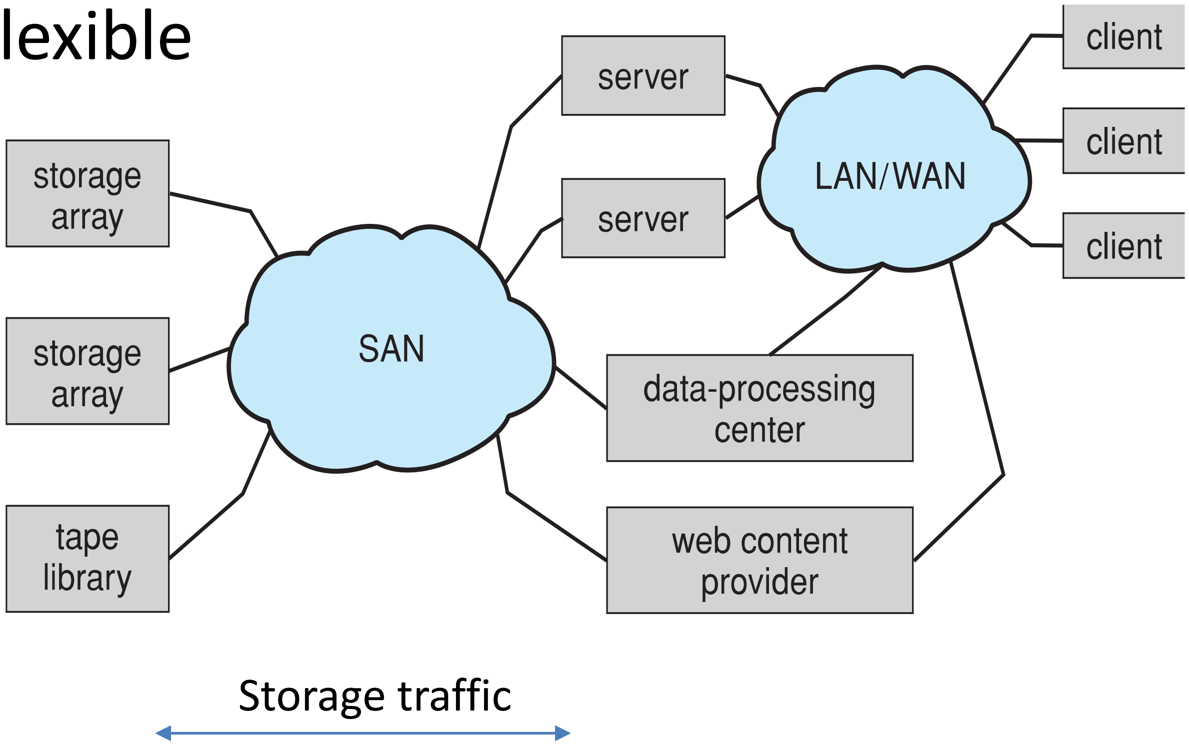


Storage Array

- Can just attach disks, or arrays of disks to an I/O port
- Storage Array has controller(s), provides features to attached host(s)
  - Ports to connect hosts to array
  - Memory, controlling software
  - A few to thousands of disks
  - RAID, hot spares, hot swap
  - Shared storage -> more efficiency

Storage Area Network

- Common in large storage environments
- Multiple hosts attached to multiple storage arrays - flexible



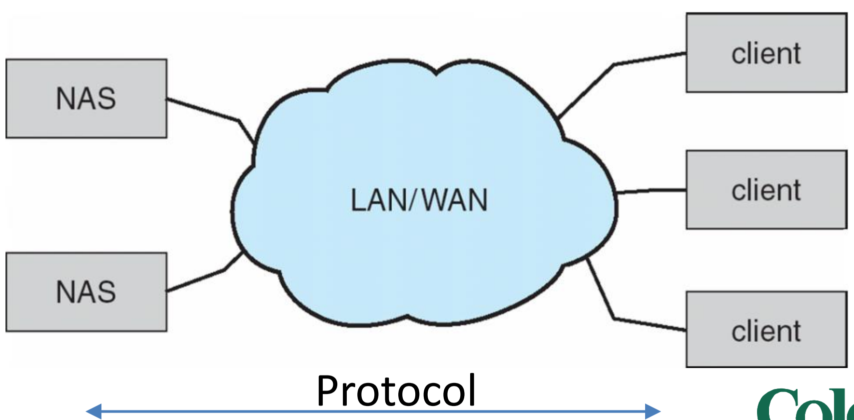
Colorado State University

Storage Area Network (Cont.)

- SAN is one or more storage arrays
- Hosts also attach to the switches
- Storage made available from specific arrays to specific servers
- Easy to add or remove storage, add new host and allocate it storage
  - Over low-latency Fibre Channel fabric

Network-Attached Storage

- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
  - Remotely attaching to file systems
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage over typically TCP or UDP on IP network
- **iSCSI** protocol uses IP network to carry the SCSI protocol
  - Remotely attaching to devices (blocks)



Disk Scheduling

- The operating system is responsible for using hardware efficiently — for the disk drives, this means having a fast access time and disk bandwidth
- Minimize seek time
- Seek time ≈ seek distance (between cylinders)
- Disk **bandwidth** is the total number of bytes transferred, divided by the total time between the first request for service and the completion of the last transfer

Disk Scheduling (Cont.)

- There are many sources of disk I/O request
  - OS
  - System processes
  - Users processes
- I/O request includes input or output mode, disk address, memory address, number of sectors to transfer
- OS maintains queue of requests, per disk or device
- Idle disk can immediately work on I/O request, busy disk means work must queue
  - Optimization algorithms only make sense when a queue exists

Disk Scheduling (Cont.)

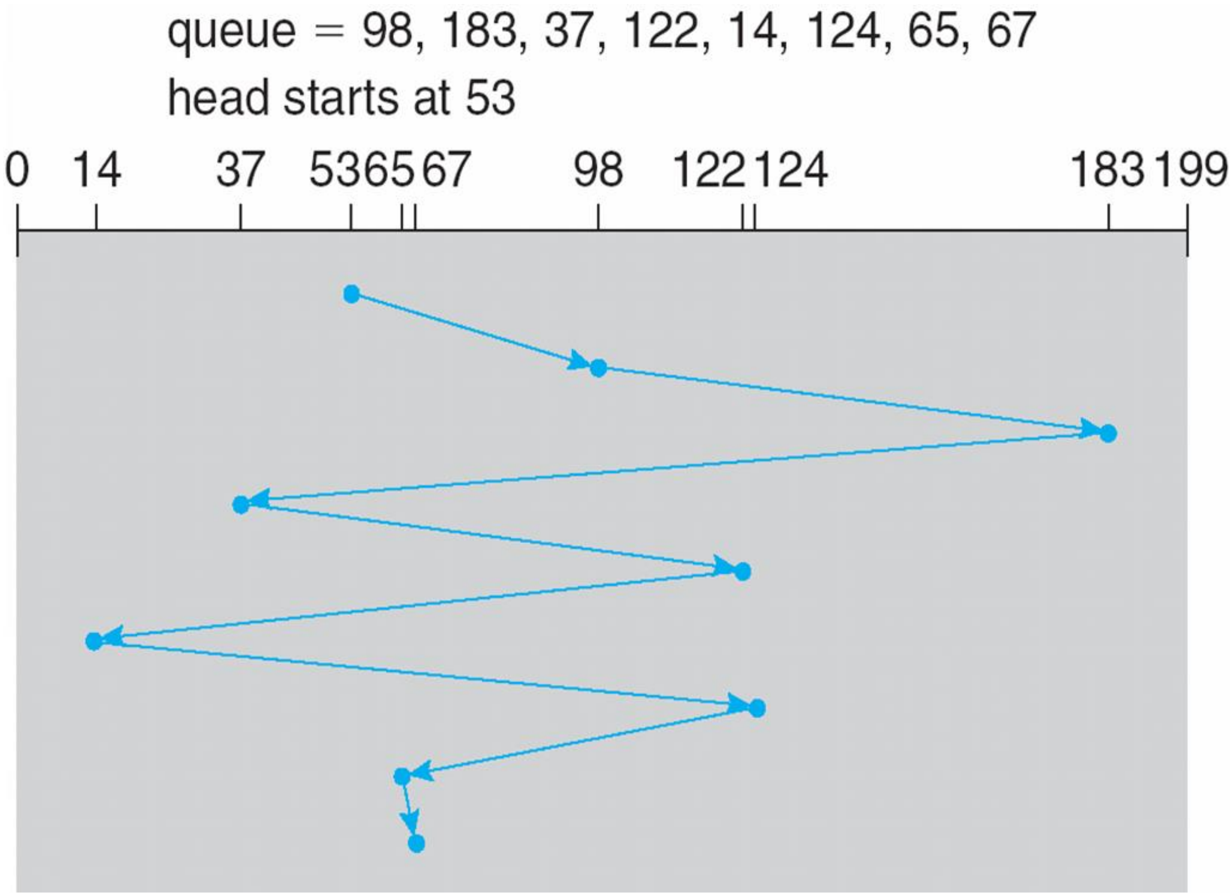
- Note that drive controllers have small buffers and can manage a queue of I/O requests (of varying "depth")
- Several algorithms exist to schedule the servicing of disk I/O requests
- The analysis is true for one or many platters
- We illustrate scheduling algorithms with a request queue (cylinders 0-199)

  98, 183, 37, 122, 14, 124, 65, 67

  Head pointer 53 (head is at cylinder 53)

FCFS (First come first served)

Illustration shows total head movement



Total seek time **640** cylinders

Colorado State University