ECE 485/585
Microprocessor System Design

Prof. Mark G. Faust

Maseeh College of Engineering
and Computer Science

PORTLAND STATE UNIVERSITY

Disks and Other I/O Devices

ECE 485/585

Mark G. Faust

Disks and other I/O Devices

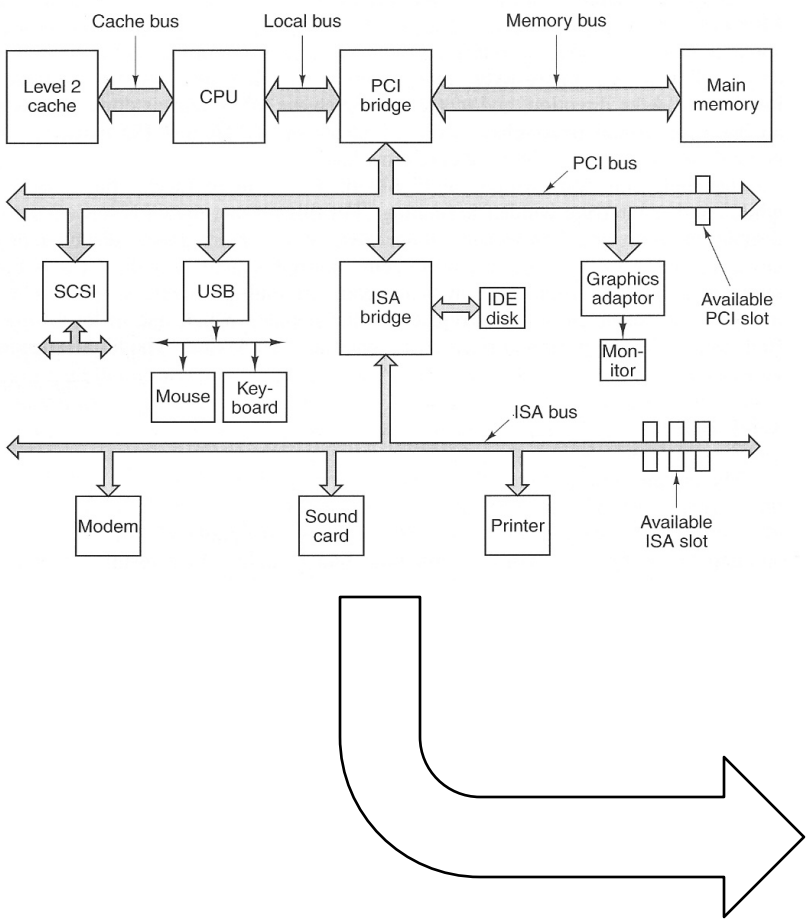
- *Handouts: none (see readings on web site)*
- Topics
  - I/O Devices
    - Keyboards, Mice, etc
    - Hard Disk Drives
      - The Memory Hierarchy (again…)
      - Disk Drive Basics
      - Disk Geometry
      - Fragmentation and Bad block handling
      - Transition-based recording and RLL codes
      - Performance
      - Disk Interfaces
        - » IDE/ATA, S-ATA
        - » SCSI, SAS
        - » FibreChannel
    - CDs and DVDs

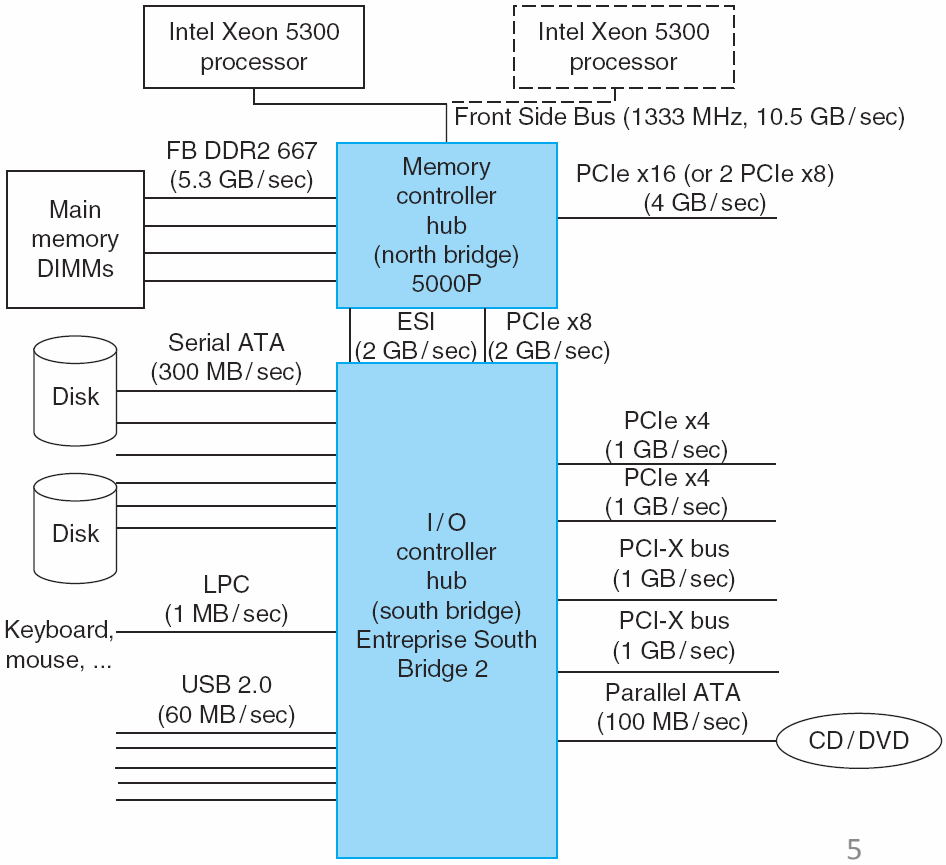
Diversity of I/O Devices

Device	Behavior	Partner	Data rate (Mbit/sec)
Keyboard	Input	Human	0.0001
Mouse	Input	Human	0.0038
Voice input	Input	Human	0.2640
Sound input	Input	Machine	3.0000
Scanner	Input	Human	3.2000
Voice output	Output	Human	0.2640
Sound output	Output	Human	8.0000
Laser printer	Output	Human	3.2000
Graphics display	Output	Human	800.0000–8000.0000
Cable modem	Input or output	Machine	0.1280–6.0000
Network/LAN	Input or output	Machine	100.0000–10000.0000
Network/wireless LAN	Input or output	Machine	11.0000–54.0000
Optical disk	Storage	Machine	80.0000–220.0000
Magnetic tape	Storage	Machine	5.0000–120.0000
Flash memory	Storage	Machine	32.0000–200.0000
Magnetic disk	Storage	Machine	800.0000–3000.0000

Latency vs. Bandwidth

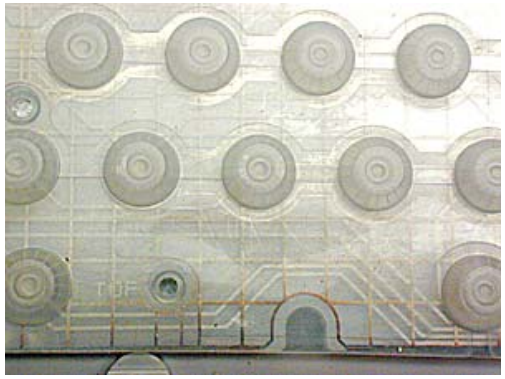
Typical PC Peripheral Architecture





Keyboards

- Matrix of switches, one per key
- Keyboard controller in keyboard
  - Continuously scans matrix for press/release events
  - Pressing key generates its “make code”
  - Releasing key generates its “break code”
  - Enables controller to recognize when two or more keys pressed simultaneously (^C, CTR ALT DEL)
  - Look-up table maps switch coordinates to ASCII code
  - Controller sends “scan code” to controller in PC
- Switch Types
  - Mechanical
    - spring-loaded momentary on switches
    - “plunger” makes contact
  - Capacitive (similar to mechanical)
  - Membrane (three layer membrane, dome)



Keyboards

- Three connectors still in use
  - AT connector (DIN) (obsolete)
  - PS/2
  - USB
- All provide
  - keyboard power
  - data
  - [clock]
- Also
  - Bluetooth
  - Infrared

PS/2 Connector

USB Connector

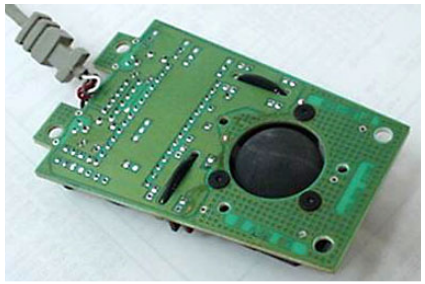
Mice

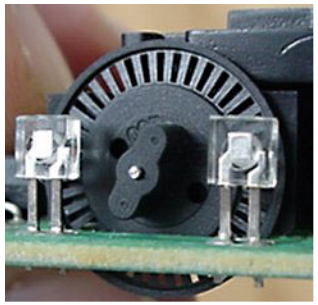
- "Graphical positioning" or "pointing device"
  - Light pens
  - Track balls
  - Joysticks
  - Mice
- Provide
  - X,Y position
  - 1 to 3 buttons
  - Scroll wheel

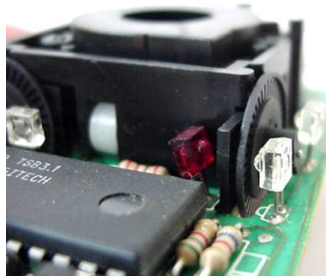
- Unlike keyboard, provides *stream* of X,Y coordinates (tracking cursor)

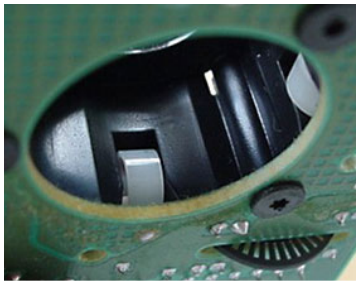
Mice

- Mechanical
  - Ball
  - Wheels and shafts
  - Optical shaft motion detectors
- Optical
  - LED or Laser
  - CMOS sensor
  - DSP
- Interfaces
  - PS/2 (obsolete)
  - USB
  - Bluetooth
  - Infrared



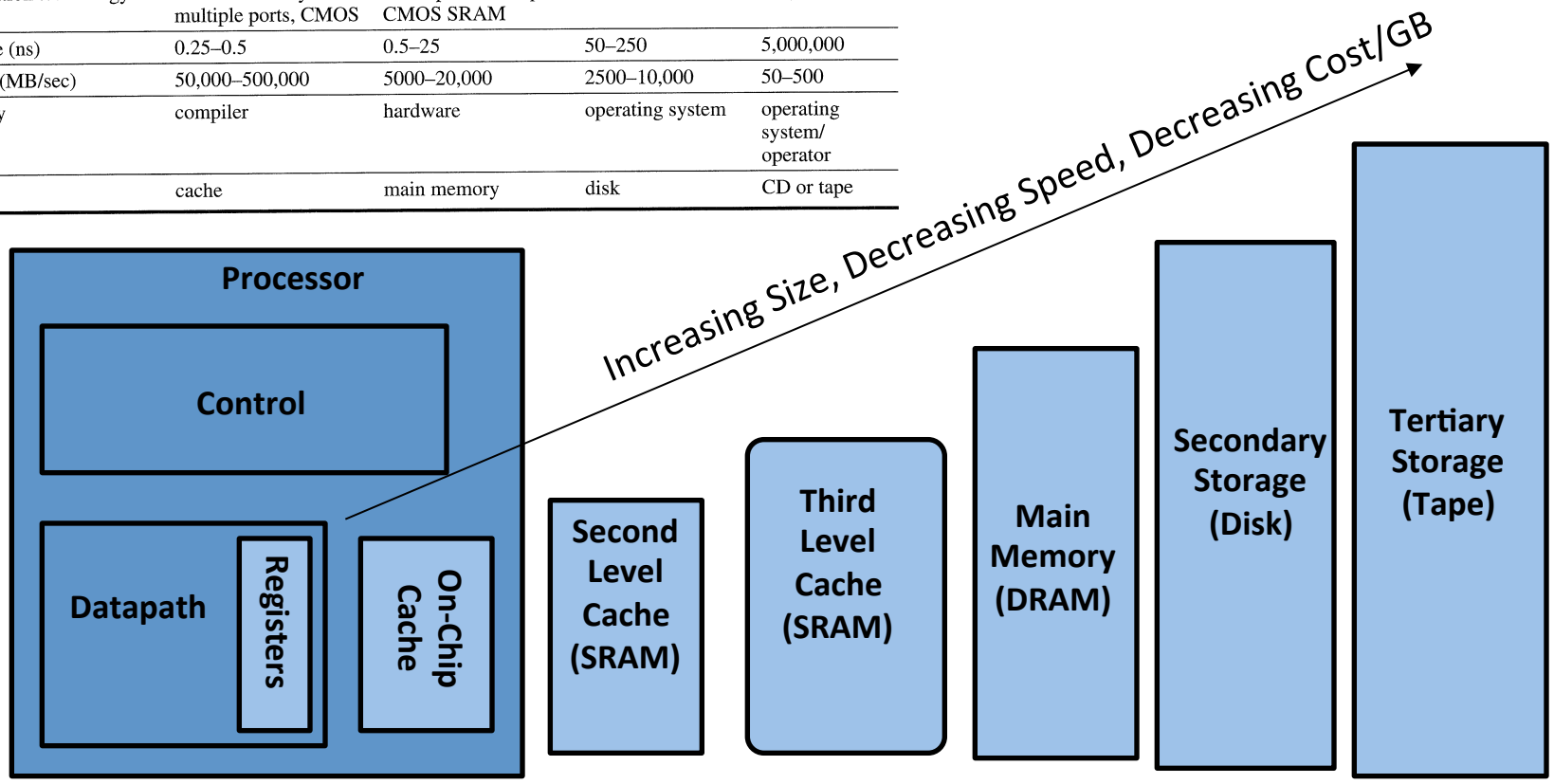




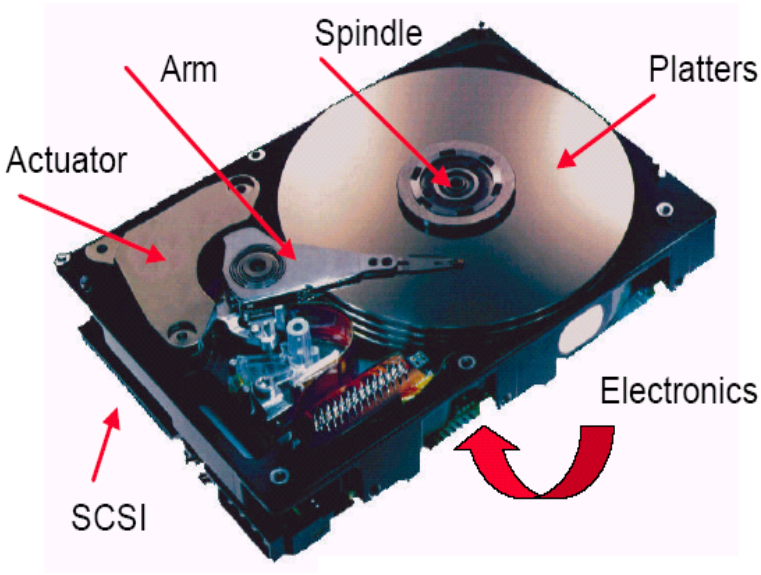


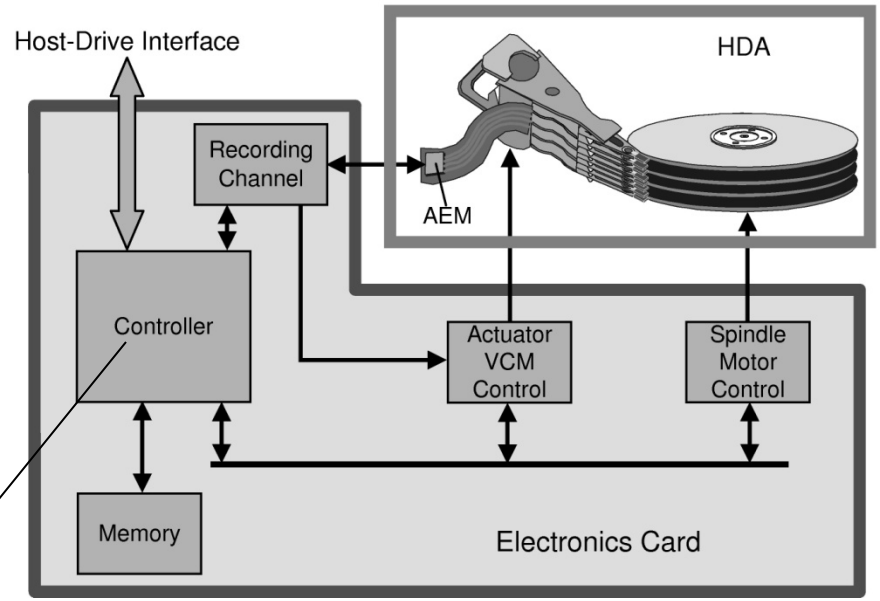
Computer Memory Hierarchy: Disks

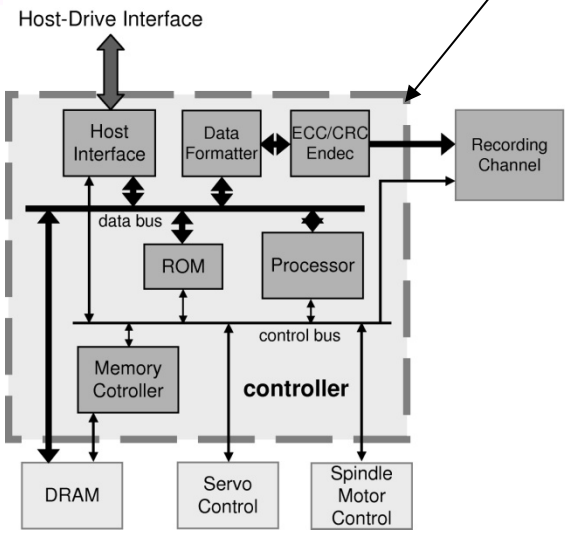
Level	1	2	3	4
Name	registers	cache	main memory	disk storage
Typical size	< 1 KB	< 16 MB	< 512 GB	> 1 TB
Implementation technology	custom memory with multiple ports, CMOS	on-chip or off-chip CMOS SRAM	CMOS DRAM	magnetic disk
Access time (ns)	0.25–0.5	0.5–25	50–250	5,000,000
Bandwidth (MB/sec)	50,000–500,000	5000–20,000	2500–10,000	50–500
Managed by	compiler	hardware	operating system	operating system/ operator
Backed by	cache	main memory	disk	CD or tape



Typical Disk Drive

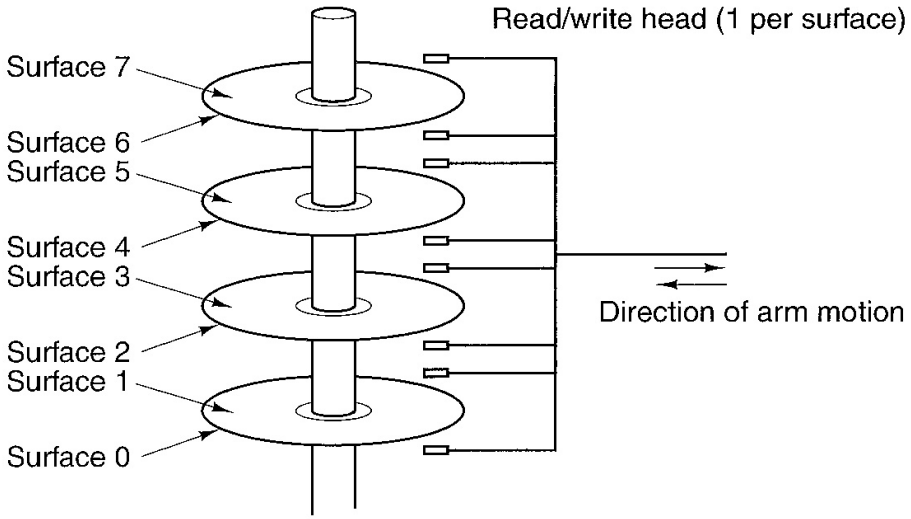


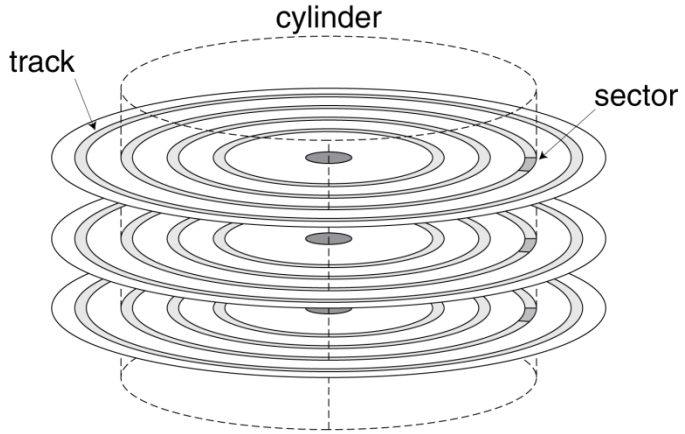


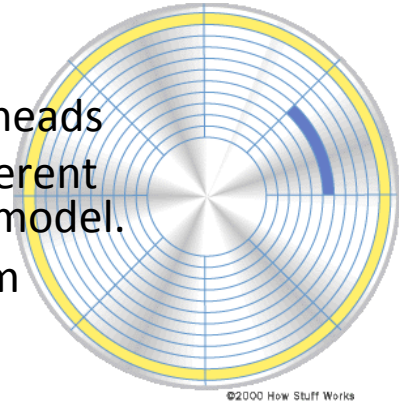


HDA – hard disk assembly
VCM – voice coil motor
AEM – arm-mounted electronics module

Typical Disk Drive





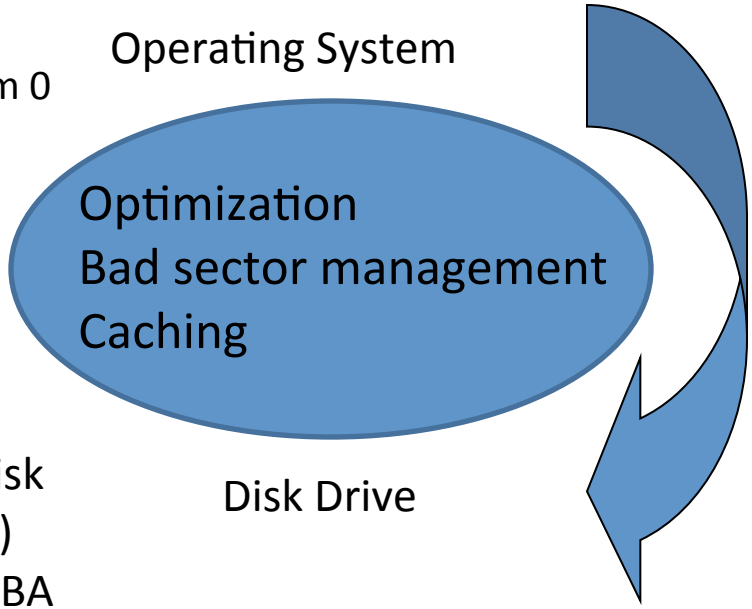


- Cost of coating surfaces (platters) is much less than cost of read/write heads
- Often manufacturers include the same platter configuration in two different models but omit some of the read/write heads in the smaller capacity model.
- Cylinder refers to "stack" of tracks on all surfaces at same distance from spindle
- Sectors are fixed-length (typically standardized on 512 bytes)*
- Need addressing scheme to uniquely identify sectors

* Windows Vista has support for 4KB sectors

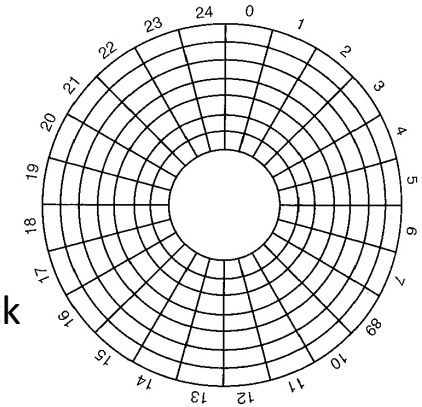
Disk Geometry

- Cylinder/Head/Sector Addressing (CHS)
  - Limitations due to PC BIOS bit field sizes
    - 1024 cylinders, 256 heads, 64 sectors
    - Cylinders, sectors numbered from 1, heads from 0
    - 8 GB
  - Limitations due to ATA
    - 65536 cylinders, 16 heads, 256 sectors
    - 137 GB
- Logical Block Addressing (LBA)
  - Sectors just numbered consecutively from 0
  - Drive maps from LBA to physical location on disk
  - Initially 28-bit LBA (so same 137 GB restriction)
  - Big Drive Initiative (ATA/ATAPI-6) uses 48-bit LBA

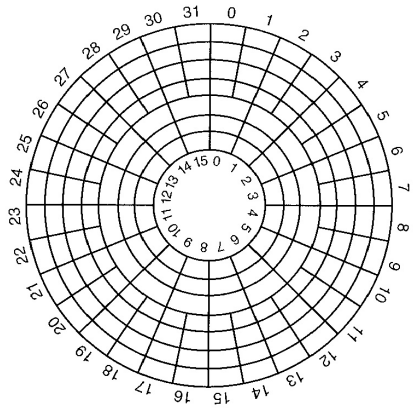


Disk Geometry and Recording Density

- Areal density = tracks/inch x bits/inch = (tpi x bpi)
- Early disks had same number of sectors in all tracks
- More space along outer diameter (OD) tracks
  - Can store more sectors in OD tracks
- Disk spins at constant (angular) velocity (CAV)
  - Must tolerate different transfer (read/write) rates for different track
- Why not variable speed (slow down disk when accessing OD)?
  - Moving mass – takes time to speed up/slow down
  - Rotational delay greater at OD and most data stored at OD
  - Easier to deal with different transfer rates (buffer memory in drive)
- Difficult if <u>every</u> track has different number of sectors
- Zoned Bit Recording
- Drive maps LBA to physical location

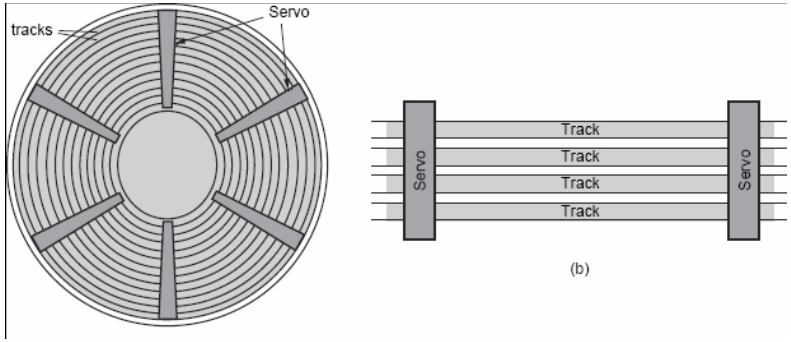


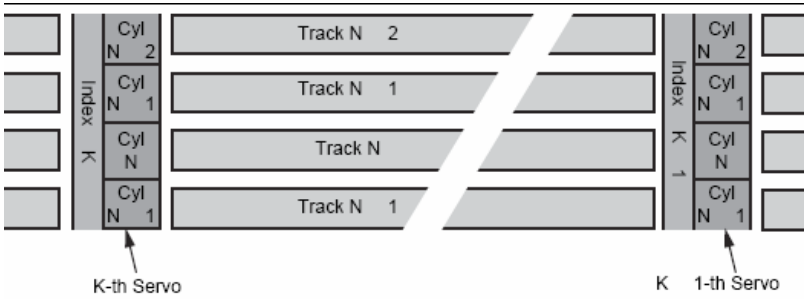
Fixed sectors/track



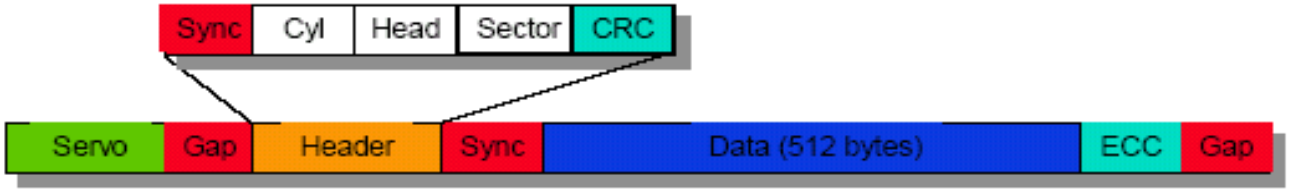
Variable sectors/track
(Zoned Bit Recording)

Sectors up close…





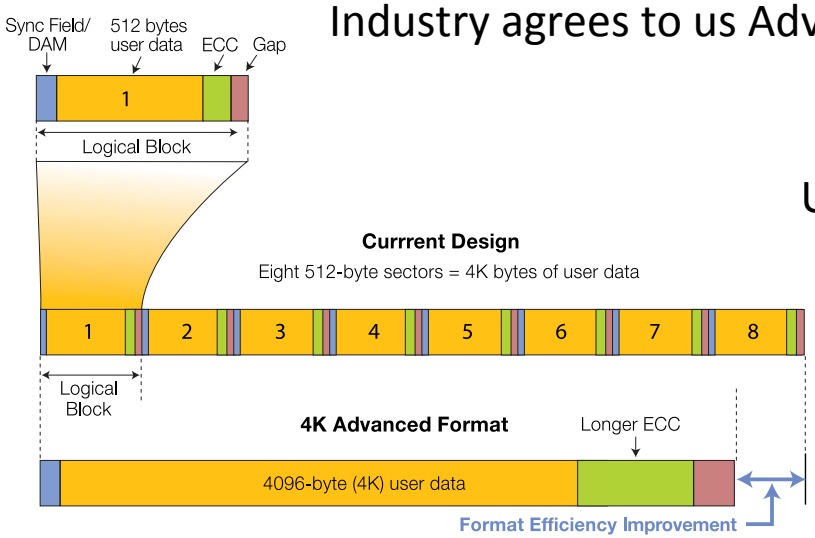
Servo sectors interspersed with data sectors
Written once during manufacturing
Allows head to accurately position and follow track



Gaps prevent write head (which must be turned off after writing but take time) from overwriting beginning of following sector
Process of low level disk formatting writes these headers
Higher level formatting does free storage data structures, root directory, empty file system
New drives omit ID/headers (saves space)

Advanced Format

Industry agrees to us Advanced Format in all new products after January 2011



Uses 4K sectors

- Denser data
  - Fewer sync fields, data address marks, gaps
- Better error correction
  - More ECC bits

Most OS file systems use 4K file blocks, mapping a file system block to 8 logical disk blocks where each logical block is one 512-byte disk sector

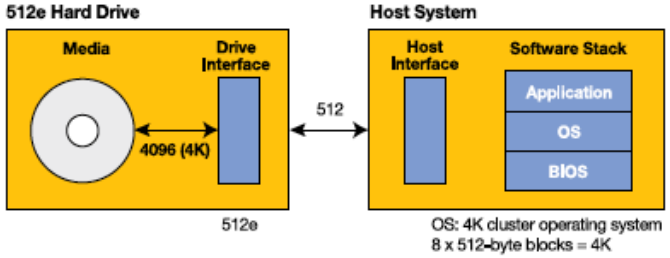
4K block = 8 x 512-byte blocks								OS File System
0	1	2	3	4	5	6	7	Logical Blocks

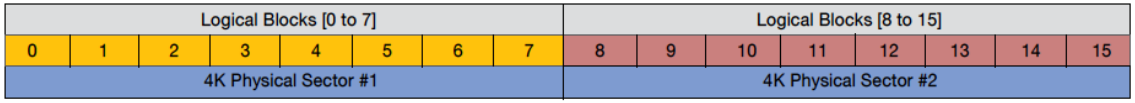
If file system aligns partitions to (new) 4K sector then file block can be sector

4K block = 8 x 512-byte blocks								OS File System
0	1	2	3	4	5	6	7	Logical Blocks
4K Physical Sector #1								Physical Sectors

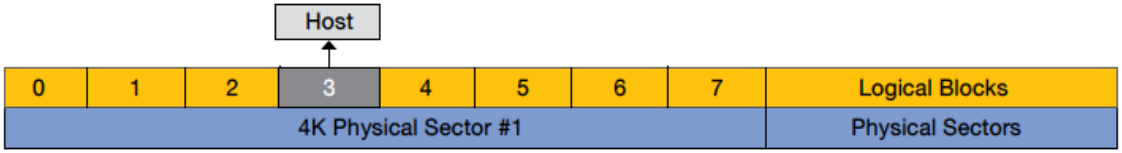
Advanced Format

Drives offer 512-byte sector emulation (512e) mode for legacy systems and programs that read/write older 512-byte sectors.

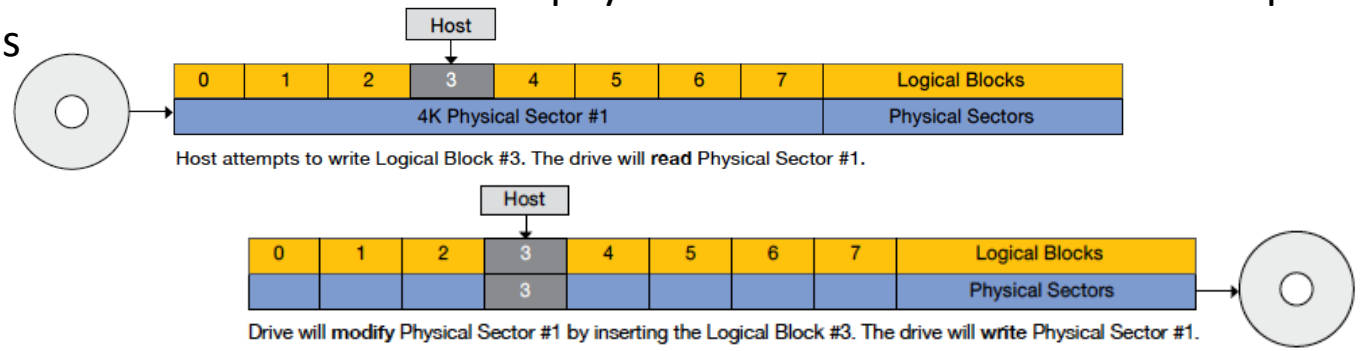




Program reads 512-byte sector and drive reads 4K physical sector containing it:

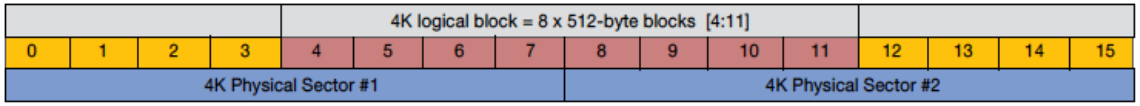


Program writes 512-byte sector and drive reads 4K physical sector containing it, modifies the 512-byte sector and write the entire physical sector back. Note: could require multiple disk rotations



Advanced Format

If file system logical blocks are not aligned to 4K physical sectors multiple reads/writes are required for single file system block access.



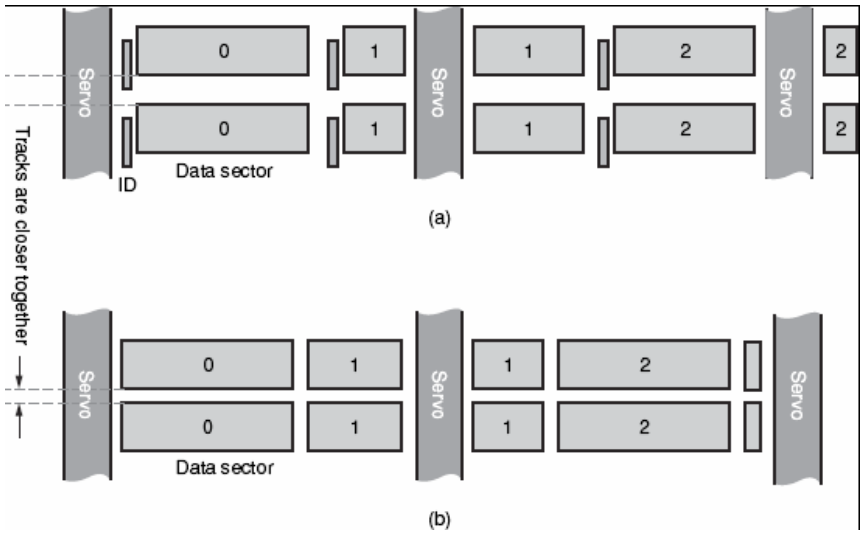
Modern OS file systems use 4K file systems and align partitions to physical sectors, permitting 4K native mode (4Kn)

Operating System	4K file systems	Automatically aligns partitions during installation
Microsoft® Windows® Vista SP1 or later	Yes	Yes
Microsoft Windows 7	Yes	Yes
Microsoft Server 2008	Yes	Yes
Mac OS® X 10.4+	Yes	Yes
Linux Ubuntu 8.04+, SUSE, Linux kernel 2.6.34+	Yes	Use Linux Partitioning Utility

Individual legacy programs might still read/write 512-byte blocks

Servo Information

ZBR means that servo data may require data sectors be split



Omitting headers (IDs) allows tracks to be closer together
(Were offset because of geometry of read/write heads)

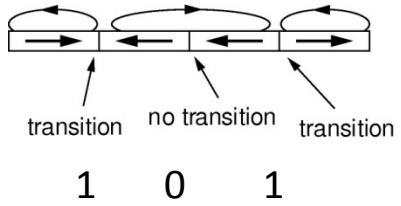
Issues

- Fragmentation
  - Internal Fragmentation
    - Files unlikely to be exact multiples of sector size, so there will be unused space in last sector of every file (natural consequence of fixed size sectors)
  - External Fragmentation
    - Over time, with additions and deletions, there will be gaps in allocated sectors. These must be managed by the file system and unused sectors allocated to new files (often allocated multiple sectors at a time (blocks)).
- Bad block handling
  - Increasing density, lower signal-to-noise ratio
  - Increase in SER (soft error rate)
  - SER ~ 1 in 10^5 bits or about 1 in 24 sectors
  - ECC handles most of these
  - For those which can't be corrected, drive maintains internal tables which mark them as bad and never uses them
  - Drive still presents a contiguous LBA address space to the user
  - Two techniques
    - Sector Slipping
      - If sector is bad, find next non-defective sequential sector to be that LBA
    - Sector Sparing
      - Reserve some spare sectors at one or more locations on the disk

Transition-Based Data and RLL

- Reading involves sensing the magnetic field in an area on the disk
- Not the direction of the field's orientation that determines 1s and 0s
- It's changes in the direction

- Reversal of field represents a 1
- No reversal of field represents 0



- This is one reason we can't write individual bits but must write entire sectors
- Problems arise with long sequence of no reversals (e.g. 0s) -- lose sync: missing bits
- Frequent reversals (e.g. 11111) can be difficult, too – maximum density of flux reverals/area: bit errors

RLL: Run Length Limited Codes

- Encodes data to prevent long sequences of repeated bits
- (d,k) RLL code
  - Minimum run length of d (0s between 1s)
  - Maximum run length of k

(1,7) RLL Coding Example

```
Data:     0 0 1 0 1 1 0 1 0 0 0 1 1 0
Encoded: 101 001 010 100 101 100 001
```

(2,7) RLL Coding Example

```
Data:     1 1 0 1 1 0 0 1 1
Encoded: 100000100000001000
```

(1,7) RLL

Data	Encoded
00 00	101 000
00 01	100 000
10 00	001 000
10 01	010 000
00	101
01	100
10	001
11	010

Maps 2 bits of data to 3 bits using groups of 2 or 4 bits

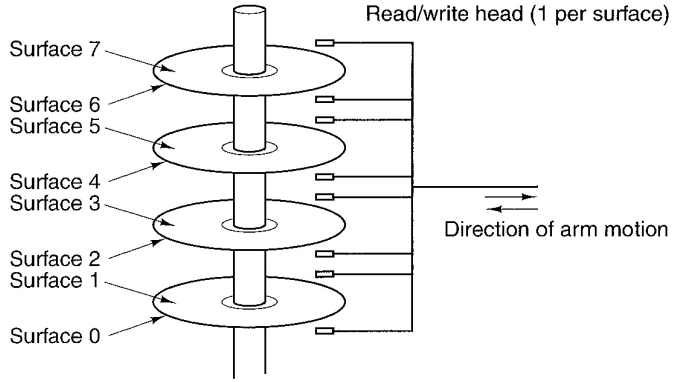
(2,7) RLL

Data	Encoded
11	1000
10	0100
000	000100
010	100100
011	001000
0011	00001000
0010	00100100

Maps n bits of data to 2n bits using groups of 2, 3, or 4 bits

Key Disk Performance Parameters

- Seek time
  - Time required to move head to the desired track
  - Depends on where the head is currently (less for adjacent or nearby tracks)
  - 5ms to 15ms typical
- Rotational Latency
  - Time required for sector to rotate under head
  - 1/rotational speed of disk
  - Average is ½ rotational latency
  - 5,400 to 12,000 RPM typical
- Transfer Time
  - Time to transfer a sector
  - 20 to 160 MB/s typical
  - Sometimes see
    - Media Transfer Rate (related to rotational speed of disk and sectors/track)
    - Interface Transfer Rate (determined by interface, e.g. SCSI Ultra 3)
    - Sustained Data Rate (affected by ability to buffer/overlap rotational delay)
- Controller Time
  - Controller overhead

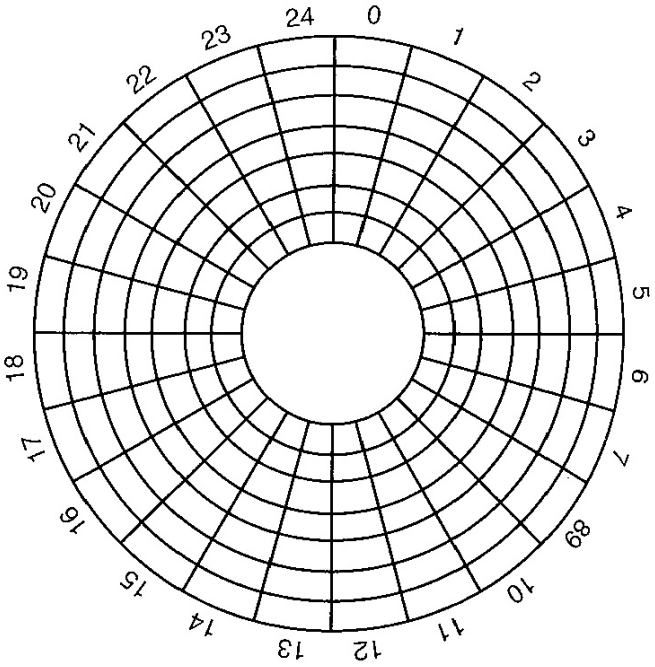


Disk Performance

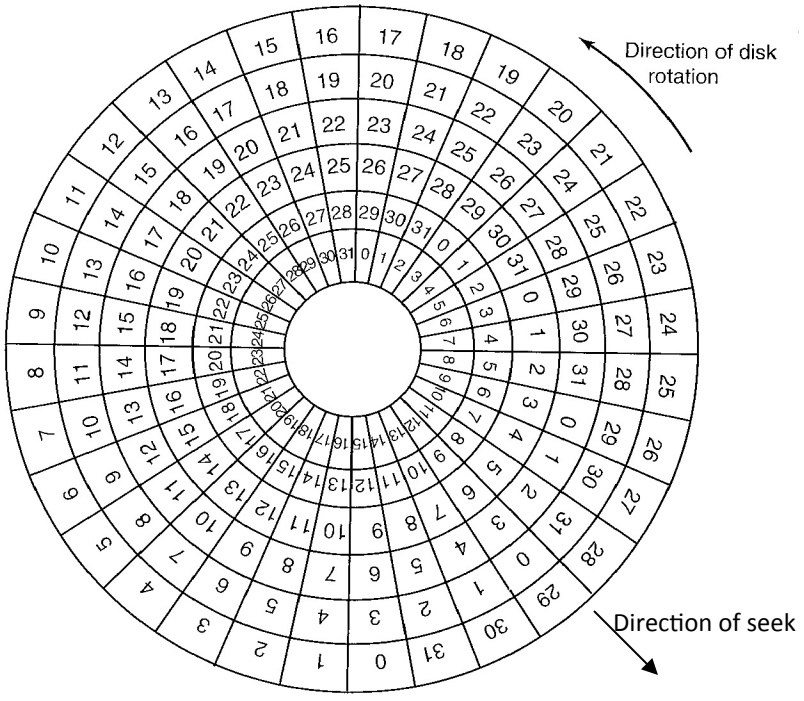
- Average access time
  - Seek time + rotational latency + transfer time + controller time
- Track and Cylinder Skew
  - Cylinder switch
    - May have to wait an additional rotation
      - Desired sector passes while repositioning head
    - Solution: offset sectors (sector skew)
  - Buffer copy time
  - Head switch
    - Some settling time
- Buffers
  - DRAM
  - Write buffer allows actual disk write to occur later
  - Read buffer necessary because media transfer rate varies (greater at OD than ID)
- Pre-fetching
  - Anticipate reads beyond current sector
  - Read and cache sectors during rotational latency
    - Exploit spatial locality
    - No additional latency for sectors in same track
- Caching (on disk and in OS)
- Command Queuing and Scheduling

Disk Layout and Optimization

Consider reading from two consecutive tracks (inner to outer). After reading last sector of first track, head must be re-positioned to next track. During this seek, the disk is still spinning and the first sector of the track will have passed under the head. So, offset or skew, sectors in adjacent tracks…

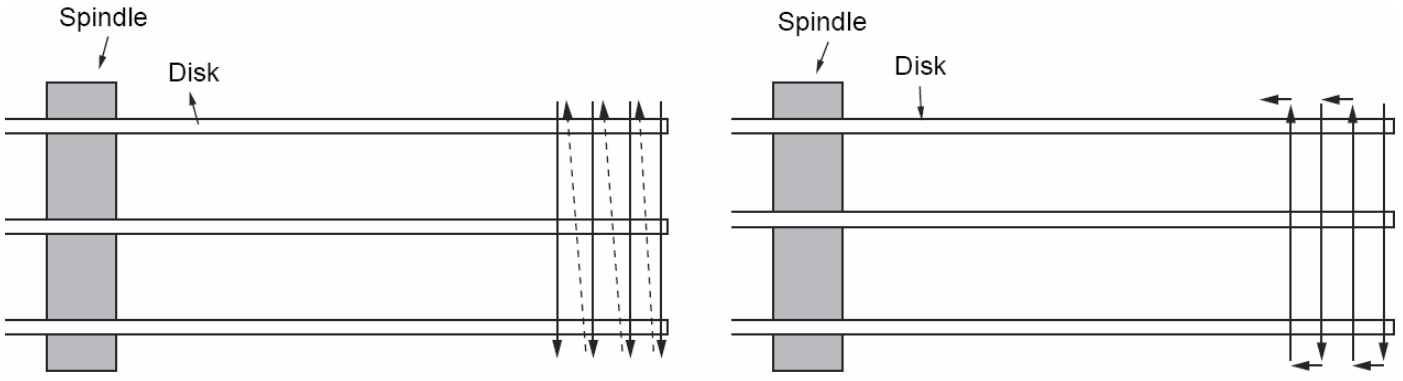


No sector skew

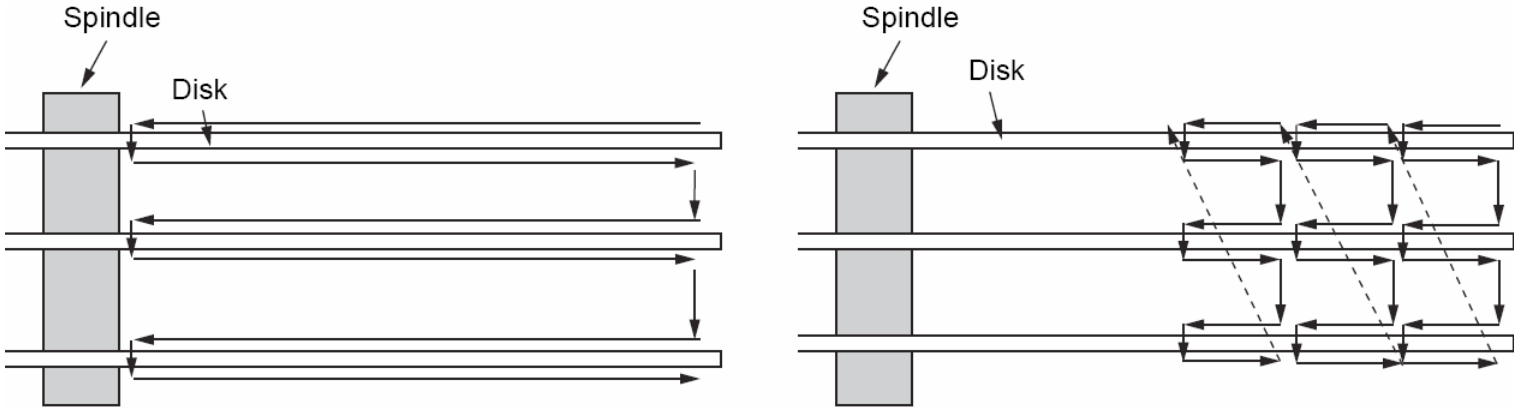


Sector skew

Logical to Physical Address Mapping

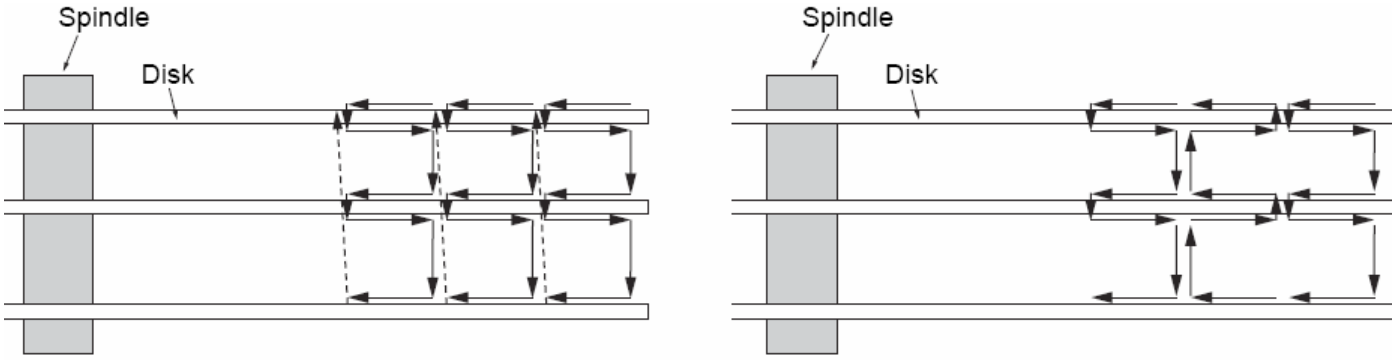


Cylinder mode formatting: After reading all sectors in track switch heads before repositioning heads to adjacent cylinder (two alternatives)



Serpentine and banded serpentine formatting

Logical to Physical Address Mapping



Banded serpentine with odd number of read heads

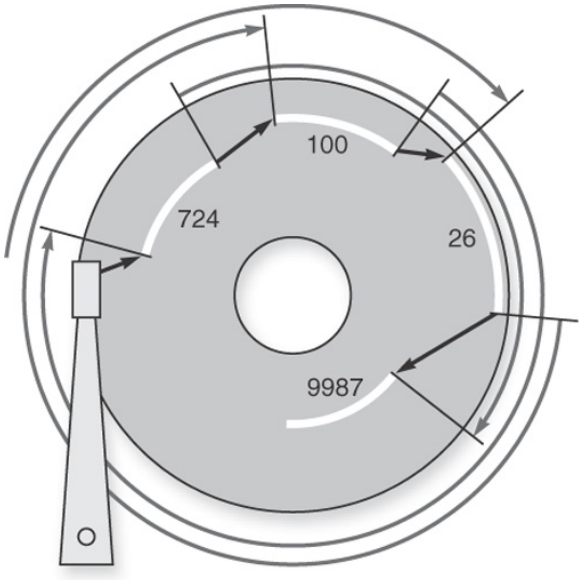
Command Queues and Optimization

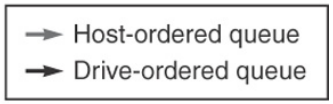
Operation	Starting LBA	Length
Read	724	8
Read	100	16
Read	9987	1
Read	26	128

Original sequence

Operation	Starting LBA	Length
Read	26	128
Read	100	16
Read	724	8
Read	9987	1

Host-ordered sequence



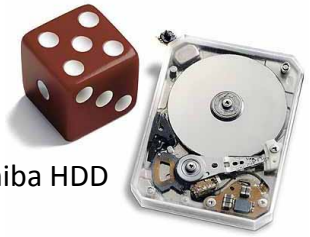


Host-ordered queue takes 3 revolutions
Disk-ordered queue takes ¾ of a revolution

Key Parameters: A Comparison

First commercial computer with HDD

- 1957 IBM 305 RAMAC
- Disk (IBM 350) shown
  - 1 ton
  - 5 MB



4 GB Toshiba HDD

Parameter	IBM 360-KB floppy disk	WD 18300 hard disk
Number of cylinders	40	10601
Tracks per cylinder	2	12
Sectors per track	9	281 (avg)
Sectors per disk	720	35742000
Bytes per sector	512	512
Disk capacity	360 KB	18.3 GB
Seek time (adjacent cylinders)	6 msec	0.8 msec
Seek time (average case)	77 msec	6.9 msec
Rotation time	200 msec	8.33 msec
Motor stop/start time	250 msec	20 sec
Time to transfer 1 sector	22 msec	17 μsec
Zones	1	16

Common Contemporary Physical Parameters

Depend on application

- Server, desktop, laptop
- Specialty (embedded systems)

Form factor

- 1.8" 2.5" 3.5" diameter (platter, not enclosure)

Rotational speed

- 3600 4200 5400 7200 10000 15000 RPM

Seagate Cheetah15K.4

Attribute	Value
Platters	4
Surfaces	8
Diameter	3.5 in
Sector size	512 bytes
Zones	15
Cylinders	50,864
Recording density	628,000 bits/in
Track density	85,000 tracks/in
Areal density	53.5 Gb/in^2
Formatted capacity	146.8 GB
Rotational rate	15,000 RPM
Average rotational latency	2 ms
Average seek time	4 ms
Sustained data rate	58-96 MB/s

From published specifications www.seagate.com

Zone	Sectors /Track	Cylinders /Zone	Logical Blocks /Zone
0	864	3201	22,076,928
1	844	3200	21,559,136
2	816	3400	22,149,504
3	806	3100	19,943,664
4	795	3100	19,671,480
5	768	3400	20,852,736
6	768	3450	21,159,936
7	725	3650	21,135,200
8	704	3700	20,804,608
9	672	3700	19,858,944
10	640	3700	18,913,280
11	603	3700	17,819,856
12	576	3707	17,054,208
13	528	3060	12,900,096
14	---	---	---

From Schindler & Ganger (DIXtrac)

Characteristics	Seagate Cheetah ST173404LC Ultra160 SCSI Drive	IBM Travelstar 32GH DJSA-232 ATA-4 Drive	IBM 1 GB Microdrive DSCM-11000
Disk diameter (inches)	3.5	2.5	1.0
Formatted data capacity (GB)	73.4	32.0	1.0
Cylinders	14,100	21,664	7,167
Disks	12	4	1
Recording surfaces (or heads)	24	8	2
Bytes per sector	512–4,096	512	512
Average sectors per track (512 byte)	≈ 424	≈ 360 (256–469)	≈ 140
Maximum areal density (Gb/sq.in.)	6.0	14.0	15.2
Rotation speed (RPM)	10,033	5,411	3,600
Average seek random cylinder to cylinder (read/write) (ms)	5.6/6.2	12.0	12.0
Minimum seek (read/write) (ms)	0.6/0.9	2.5	1.0
Maximum seek (ms)	14.0/15.0	23.0	19.0
Data transfer rate (MB/sec)	27–40	11–21	2.6–4.2
Link speed to disk buffer (MB/sec)	160	67	13
Power idle/operating (W)	16.4/23.5	2.0/2.6	0.5/0.8
Buffer size (MB)	4.0	2.0	0.125
Size: height × width × depth (inches)	1.6 × 4.0 × 5.8	0.5 × 2.7 × 3.9	0.2 × 1.4 × 1.7
Weight (pounds)	2.00	0.34	0.035
Rated MTTF (powered-on hours)	1,200,000	(see caption)	(see caption)
Percentage of powered-on hours (POH) per month	100%	45%	20%
Percentage of POH seeking, reading, writing	90%	20%	20%
Load/unload cycles (disk powered on/off)	250 per year	300,000	300,000
Nonrecoverable read errors per bits read	< 1 per 10^{15}	< 1 per 10^{13}	< 1 per 10^{13}
Seek errors	< 1 per 10^{7}	not available	not available
Shock tolerance: operating, not operating	10 G, 175 G	150 G, 700 G	175 G, 1500 G
Vibration tolerance: operating, not operating (sine swept, 0 to peak)	5–400 Hz @ 0.5 G, 22–400 Hz @ 2 G	5–500 Hz @ 1G, 2.5–500 Hz @ 5 G	5–500 Hz @ 1 G, 10–500 Hz @ 5 G

Figure 7.2 Characteristics of three magnetic disks of 2000. To help the reader gain intuition about disks, this table gives typical values for disk parameters. The 2.5-inch drive is a factor of 6 to 9 better in weight, size, and power than the 3.5-inch drive. The 1.0-inch drive is a factor of 10 to 11 better than the 2.5-inch drive in weight and size, and a factor of 3–4 better in power. Note that 3.5-inch drives are designed to be used almost continuously, and so are rarely turned on and off, while the smaller drives spend most of their time unused and thus are turned on and off repeatedly. In addition, these mobile drives must handle much larger shocks and vibrations, especially when turned off. These requirements affect the relative cost of these drives. Note that IBM no longer quotes MTBF for 2.5-inch drives, but when they last did it was 300,000 hours. IBM quotes the service life as 5 years or 20,000 powered-on hours, whichever is first. The service life for the 1.0-inch drives is 5 years or 8800 powered-on hours, whichever is first.

Seagate Cheeatah 73GB

IBM Microdrive

Hitachi Microdrive

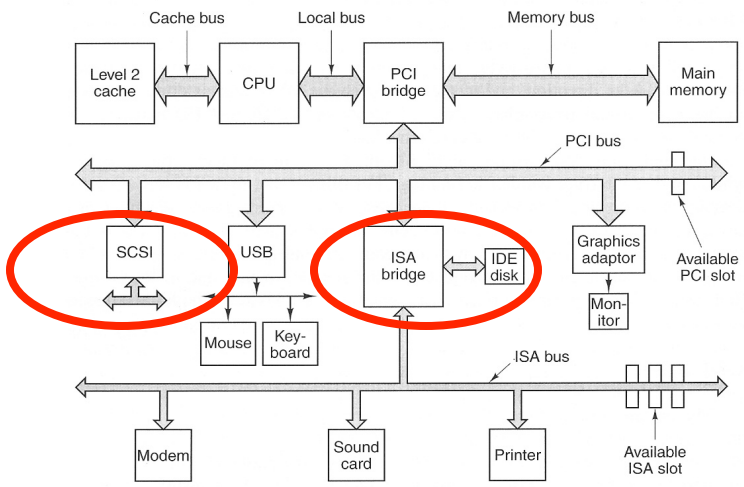
Characteristics	Seagate ST33000655SS	Seagate ST31000340NS	Seagate ST973451SS	Seagate ST9160821AS
Disk diameter (inches)	3.50	3.50	2.50	2.50
Formatted data capacity (GB)	147	1000	73	160
Number of disk surfaces (heads)	2	4	2	2
Rotation speed (RPM)	15,000	7200	15,000	5400
Internal disk cache size (MB)	16	32	16	8
External interface, bandwidth (MB/sec)	SAS, 375	SATA, 375	SAS, 375	SATA, 150
Sustained transfer rate (MB/sec)	73–125	105	79–112	44
Minimum seek (read/write) (ms)	0.2/0.4	0.8/1.0	0.2/0.4	1.5/2.0
Average seek read/write (ms)	3.5/4.0	8.5/9.5	2.9/3.3	12.5/13.0
Mean time to failure (MTTF) (hours)	1,400,000 @ 25°C	1,200,000 @ 25°C	1,600,000 @ 25°C	—
Annual failure rate (AFR) (percent)	0.62%	0.73%	0.55%	—
Contact start-stop cycles	—	50,000	—	>600,000
Warranty (years)	5	5	5	5
Nonrecoverable read errors per bits read	<1 sector per 10^{16}	<1 sector per 10^{15}	<1 sector per 10^{16}	<1 sector per 10^{14}
Temperature, shock (operating)	5°–55°C, 60 G	5°–55°C, 63 G	5°–55°C, 60 G	0°–60°C, 350 G
Size: dimensions (in.), weight (pounds)	1.0" × 4.0" × 5.8", 1.5 lbs	1.0" × 4.0" × 5.8", 1.4 lbs	0.6" × 2.8" × 3.9", 0.5 lbs	0.4" × 2.8" × 3.9", 0.2 lbs
Power: operating/idle/ standby (watts)	15/11/—	11/8/1	8/5.8/—	1.9/0.6/0.2
GB/cu. in., GB/watt	6 GB/cu.in., 10 GB/W	43 GB/cu.in., 91 GB/W	11 GB/cu.in., 9 GB/W	37 GB/cu.in., 84 GB/W
Price in 2008, $/GB	~ $250, ~ $1.70/GB	~ $275, ~ $0.30/GB	~ $350, ~ $5.00/GB	~ $100, ~ $0.60/GB

Disk Interfaces

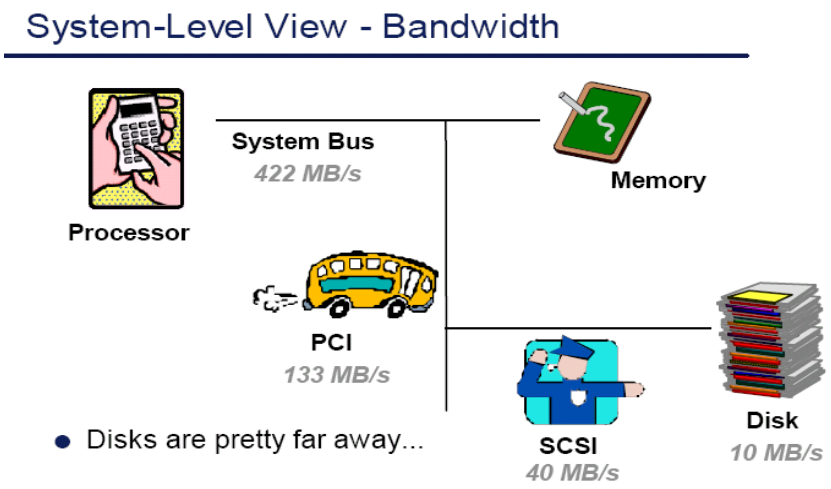
- Some Terminology
  - Drives: HDA (Hard Disk Assembly)
    - Platters, Heads, Servo
  - Controller: Logical interface to drive
    - Commands to seek, read, write
  - Adapter: Interfaces controller to PC via standard bus
    - PCI to IDE

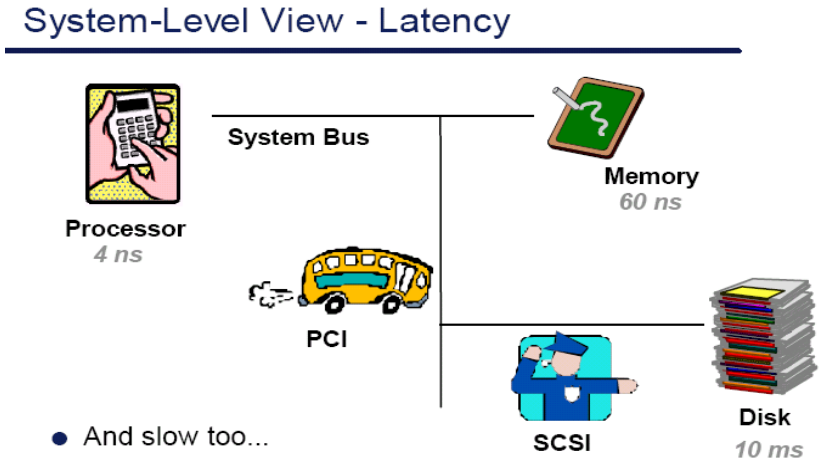
IDE
Integrated Drive Electronics

- Personal storage
  - IDE/ATA
  - SATA
- Enterprise (server) storage
  - SCSI



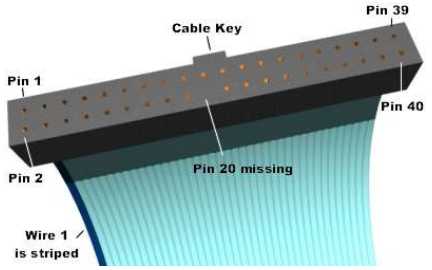
System Level View

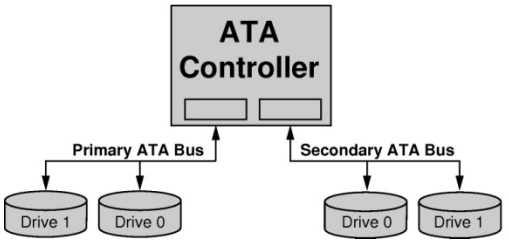




Disk Interfaces

- IDE
  - Integrated Drive Electronics (IDE/ATA)
    - Put the controller on the drive
    - Connected to adapter via ribbon cable
      - 40-pin connector
  - AT Attachment (ATA) as in IBM PC/AT
  - Combined Drive/Controller interfaced to ISA bus
  - Each IDE Interface Supports Two Drives
  - Many PC motherboards have two IDE Interfaces
  - Programmed IO (PIO) and DMA (later)
  - S-ATA (Serial ATA) in 2003
    - ATA then became known as PATA (Parallel ATA)





ATA

- ATA-1
  - Original IDE Specification
  - 5V TTL I/F
  - 8/16-bit data width
  - 8.3 MB/s DMA
- ATAPI (ATA Packet Interface)
  - Adapted ATA for CD-ROM, tape
- ATA-2
  - EIDE (Enhanced IDE)
  - Added Logical Block Addressing
  - 16.67 MB/s DMA
- ATA-3
  - 16-bit data width
  - Improved Power Management

ATA/ATAPI-4
- ATA & ATAPI Integrated into same standard
- Ultra DMA, Ultra ATA and Ultra ATA/33
- 33.33 MB/s DMA

ATA/ATAPI-5
- 66.67 MB/s DMA

ATA/ATAPI-6
- 3.3V I/F
- 100 MB/s DMA
- Big Drive Initiative (48-bit LBA)

ATA/ATAPI-7
- 133 MB/s

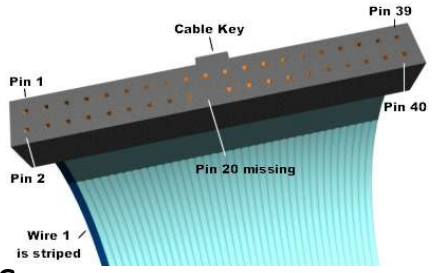
ATA/ATAPI-8
- Hybrid drives (incorporate NVM with HDD)

ATA/ATAPI-9
- Advanced Format Technology

IDE/ATA Signals

40 Pin IDE Connector Pin-Out

Pin	Function	Pin	function
1	Reset#	2	Ground
3	Data 7	4	Data 8
5	Data 6	6	Data 9
7	Data 5	8	Data 10
9	Data 4	10	Data 11
11	Data 3	12	Data 12
13	Data 2	14	Data 13
15	Data 1	16	Data 14
17	Data 0	18	Data 15
19	Ground	20	Key
21	DMARQ	22	Ground
23	DIOW#	24	Ground
25	DIOR#	26	Ground
27	IORDY	28	CSEL
29	DMACK#	30	Ground
31	INTRQ	32	IOCS16#
33	DA1	34	PDIAG#
35	DA0	36	DA2
37	CS0#	38	CS1#
39	DASP#	40	Ground



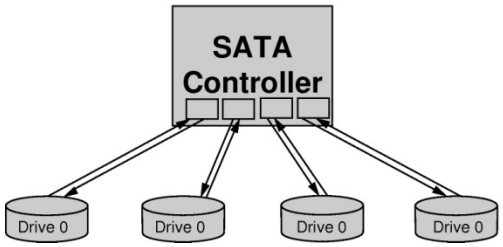
Data: 16 bits
DIOW: IO Write Strobe
DIOR: IO Read Strobe
DA2, DA1, DA0: Device Address
CS0 (CS1FX), CS1 (CS3FX): Chip Select
IDE addressed via IN/OUT instructions

- I/O addresses 37Xh and 3FXh
  - Control Block Registers
- I/O addresses 17Xh and 1Fxh
  - Command Block Registers
- Controller maps these to bit combinations on {DIOW, DIOR, CS0, CS1, DA2, DA1, DA0}
  - To form addresses for drive's control/command registers

SATA

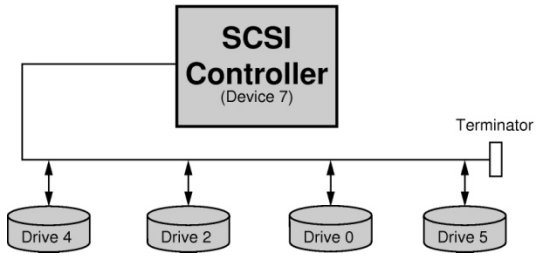
- SATA-1 (Serial ATA) 2001
  - 7-wire cable
  - 250mV LVDS (Tx, Rx pairs, 3 Ground)
  - 1.5 Gbps, 8b/10b encoding
  - 150 MB/s
  - Motherboards with SATA-1 in Spring 2003
- SATA-2
  - Additional features
  - 300 MB/s
- SATA-3
- Advantages over ATA (IDE)
  - Arguably higher bandwidth
  - Smaller connector, cable (space, airflow)
  - Lower power
  - Higher reliability

Parallel and Serial ATA Connectors



SCSI

- Small Computer Systems Interface
- Pronounced “scuzzy”
- Higher Performance
  - Enterprise Storage: Servers
  - High end workstations
  - Earlier Macintosh
  - Evolved range of performance over time
- Bus architecture
  - Permits daisy chaining drives (7, 15)
- Commands, Messages, Status
  - Asynchronously
  - Request/Acknowledge handshaking (slower)
- Data
  - Synchronously (fast!)



Name	Clock (MHz)	Width (Bytes)	Speed (MB/s)
(Narrow) SCSI-1	5	1	5
Fast (Narrow) SCSI	10	1	10
Fast Wide SCSI	10	2	20
(Narrow) Ultra SCSI	20	1	20
Wide Ultra SCSI	20	2	40
(Narrow) Ultra2 SCSI	40	1	40
Wide Ultra2 SCSI	40	2	80
Ultra3 SCSI (Ultra 160)	80	2	160
Ultra 320 SCSI	160	2	320

SCSI

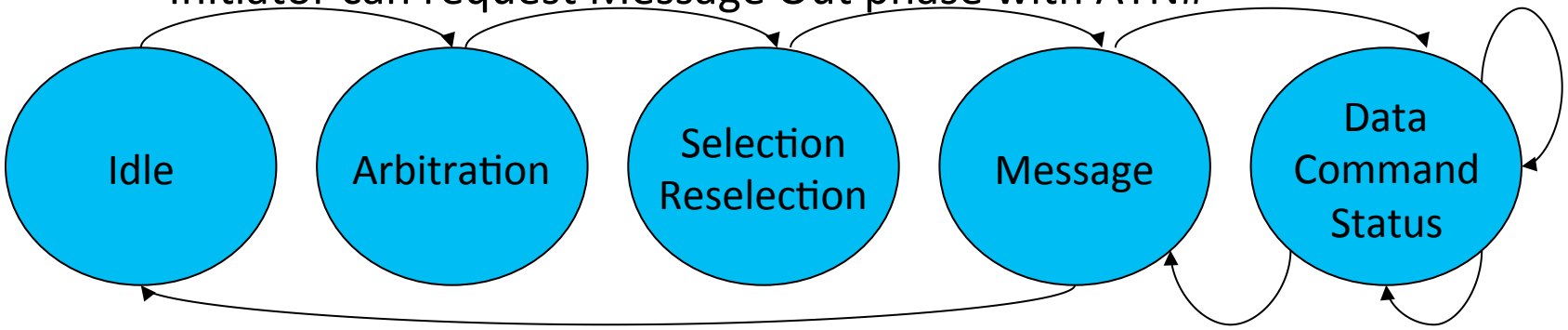
- Connectors
  - Wide variety in use
- 8 /16 devices per bus
  - Narrow – 8
  - Wide – 16
- Contemporary versions use LVDS
  - Older versions used SE, differential (high voltage)
- Daisy chained topology
- Each device on bus has a unique SCSI ID (1-of-m code)
  - 8-bit version
    - 00000001 is SCSI ID 0; 10000000 is SCSI ID 7
- Initiators and Targets
  - Beyond bus mastering
    - E.g. device-to-device transfer, command chaining

SCSI Signals

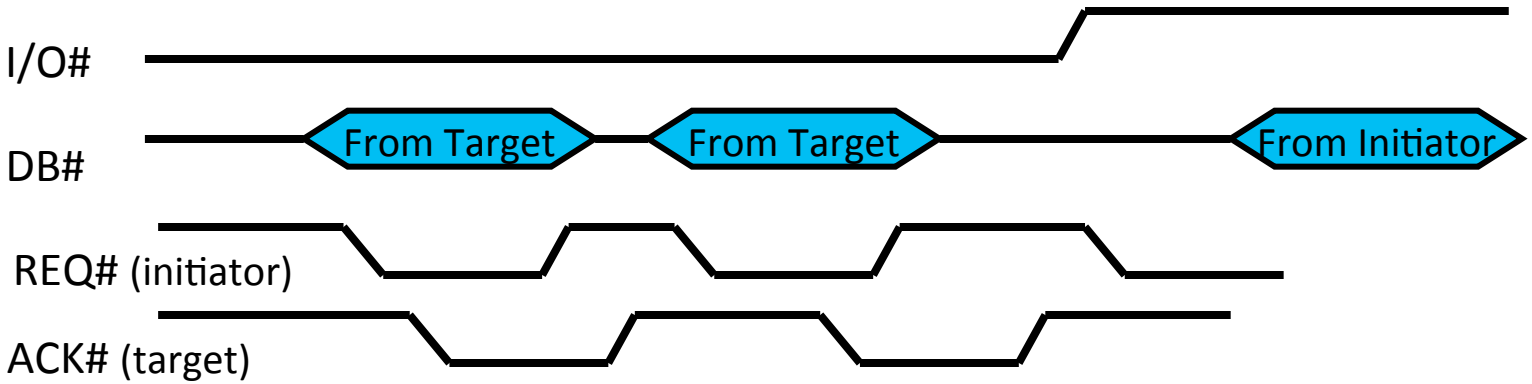
BSY#	Busy. Bus is busy
SEL#	Select (of target by initiator or initiator by target)
C/D#	Control/Data [L = Control]
I/O#	Input/Output (relevant to initiator) [L= Input]
MSG#	Message. Sending a message
DB[0:31]#	Data Bus. Inverted!
DP[0:3]#	Data Parity by byte (odd). E.g. DP[3] for bits DB[24:31]
ATN#	Attention. Indicated initiator's intention to send message
REQ#	Request. Request from target device to send data
ACK#	Acknowledge. Response to REQ#
RST#	Reset
DIFFSENS	Indicates type of interface (<0.7V = SE, 0.9 to 1.9V LVD, >2.4V HVD)

SCSI

- Transactions involve 8 distinct phases
  - Arbitration (decentralized)
    - Initiators signal desire to become bus master by asserting their own SCSI ID
    - Device with highest ID wins in case of conflict
  - Information transfer phases determined by MSG#, I/O#, C/D# signals (controlled by Target)
    - Data In/Out
    - Command
    - Status
    - Message In/Out
  - Initiator can request Message Out phase with ATN#



- Idle
  - SEL# and BSY# de-asserted (high)
- Arbitration
  - Initiator asserts BSY# and places its SCSI ID on data bus
  - If it's the highest priority requestor, asserts SEL#
- Selection
  - Initiator puts OR of its SCSI ID and target's SCSI ID on data bus
  - Target detects its SCSI ID and asserts BSY#
- Reselection
  - Sometimes target may take a while to respond, so it reselects initiator
- Command
  - Target requests command information from initiator
  - Asserts C/D#; I/O# and MSG# de-asserted during REQ#/ACK# handshake
- Data
  - Data In: from target initiator
  - Target asserts I/O#; C/D# and MSG# de-asserted during REQ#/ACK# handshake
  - Data Out: from initiator to target
  - Target negates I/O#, C/D#, and MSG# de-asserted during REQ#/ACK# handshake
- Message
  - Message In and Message Out
- Status
  - Initiator requests status of target
  - Target asserts C/D# and I/O# and de-asserts MSG# during REQ#/ACK# handshake



SCSI Handshake for Asynchronous Transfer

SCSI Synchronous Mode

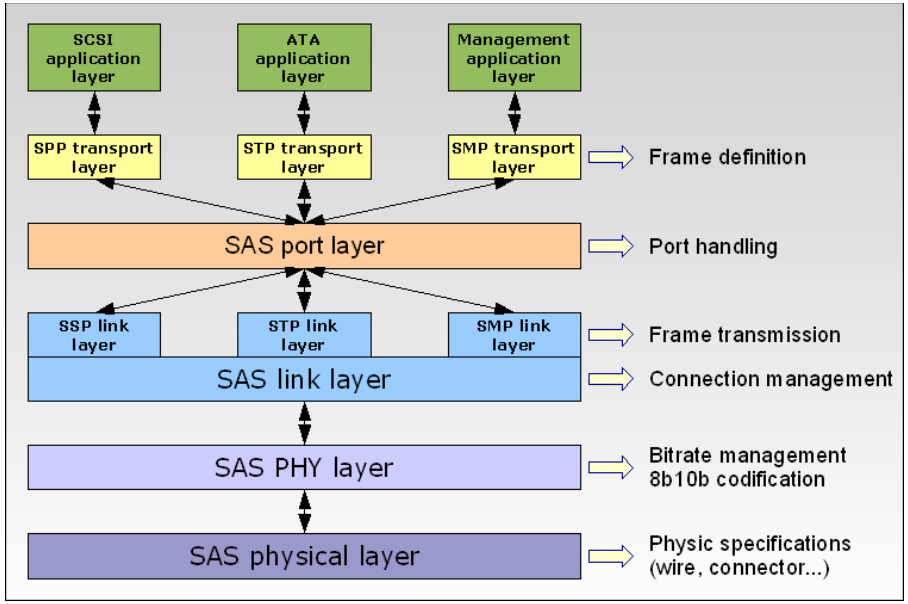
- Negotiation between devices
  - Using Message system
    - Synchronous Data Transfer Request
  - Support synchronous mode?
  - Minimum lengths and periods of ACK# and REQ# determined
  - Permissible time lags between REQ#/ACK#
- Similar Message-based inquiry for 8- and 16-bit modes

Example: Read

- Idle
- Arbitration
  - Initiator arbitrates for bus ownership
- Selection
  - Initiator selects target device
  - Initiator asserts ATN# (intention to send Identify message indicating the logical device)
- Command
  - Target device goes into command phase and receives descriptor block of Read command
- Data
  - Target devices interprets command and goes to Data In phase, sending the data
- Status
  - Upon completion, target goes into Status phase and sends that status Good
- Message
  - Target goes to Message In phase, sending a Command Complete message and releases the bus

SAS

- Serial Attached SCSI
  - Borrowed from SATA work
    - Similar cables, connectors
    - Released in 2002
  - Retains SCSI command set
  - Uses underlying serial layer
  - SATA Tunneling Protocol
    - SAS controllers can support SATA drives



FibreChannel

- Predates SATA and SAS
- Interface used in SAN (Storage Area Network)
  - Remote storage devices (disk arrays, tape libraries, optical jukeboxes)
  - Appear local to operating system
- Designed for operation over fiber optic physical layer
- Evolved to use copper wire
- Originally 1Gbps (100 MB/s)[1]
- Now 200 MB/s, 400 MB/s, 800 MB/s
- 10 Gbps under development (not backward compatible) to yield 1GB/s

[1] Serial links for storage devices use 8b/10b encoding so 1Gbps yields 1MB/s of actual data

A Comparison

Interface	PATA	SATA	SCSI	SAS	Fibre Channel	USB 2.0	IEEE1394 (FireWire)
Internal	Yes	Yes	Yes	Yes	Yes	No	Yes
External	No	No	Yes	Yes	Yes	Yes	Yes
Storage	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other I/O	No	No	Yes	Yes	No	Yes	Yes
Speed MB/s	133	150 300[1]	160 320	300[3] 600[4]	100- 800	60	50 100[2]

[1] SATA-2
[2] FireWire 800
[3] Full duplex, 300MB/s each direction
[4] Second generation SAS

A Comparison

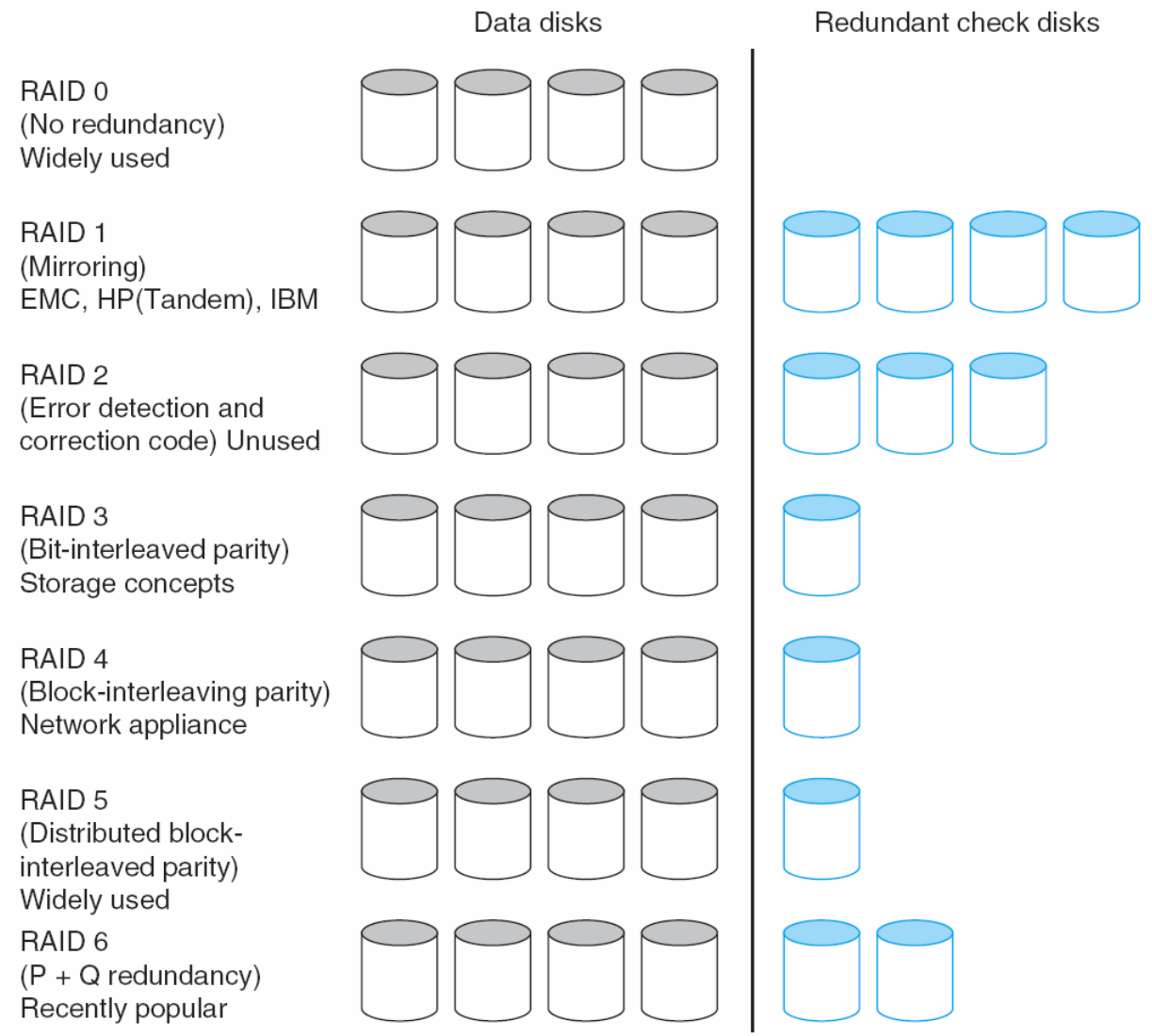
Characteristic	Firewire (1394)	USB 2.0	PCI Express	Serial ATA	Serial Attached SCSI
Intended use	External	External	Internal	Internal	External
Devices per channel	63	127	1	1	4
Basic data width (signals)	4	2	2 per lane	4	4
Theoretical peak bandwidth	50 MB/sec (Firewire 400) or 100 MB/sec (Firewire 800)	0.2 MB/sec (low speed), 1.5 MB/sec (full speed), or 60 MB/sec (high speed)	250 MB/sec per lane (1x); PCIe cards come as 1x, 2x, 4x, 8x, 16x, or 32x	300 MB/sec	300 MB/sec
Hot pluggable	Yes	Yes	Depends on form factor	Yes	Yes
Maximum bus length (copper wire)	4.5 meters	5 meters	0.5 meters	1 meter	8 meters
Standard name	IEEE 1394, 1394b	USB Implementors Forum	PCI-SIG	SATA-IO	T10 committee

RAID

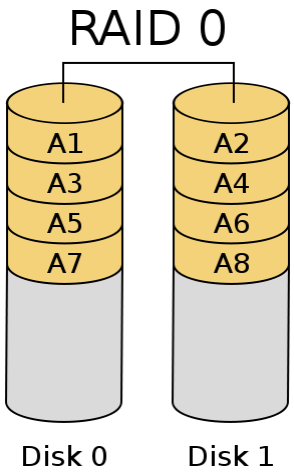
- Redundant Array of Inexpensive/Independent Drives/Disks
- Idea: Use many smaller, cheaper disks to improve I/O operations/sec and data transfer rates
- Weakness: Lower reliability (more disks → more failures)
- Solution: Add redundancy
- Important: Concerned with entire disk failure not with errors within a sector (ECC on disk)
- Ken Ouichi, IBM (1978) applied parity scheme to disks
- UC Berkeley RAID paper (1988) proposed taxonomy and described RAID 1 through RAID 5
- Widely adopted (and extended) by industry

RAID

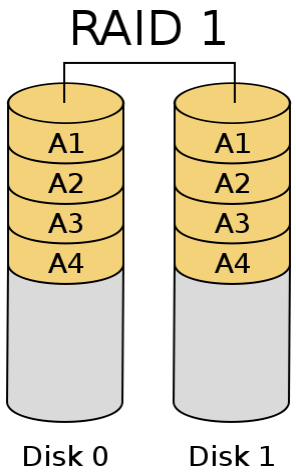
"striping"



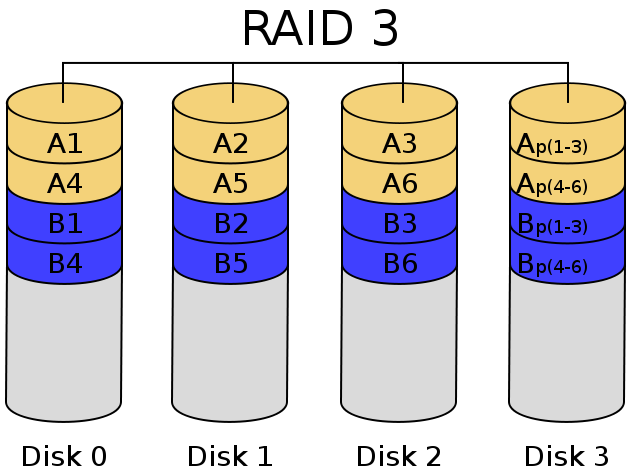
RAID



Striping (block level)
No redundancy



Mirroring
Requires 2x disks



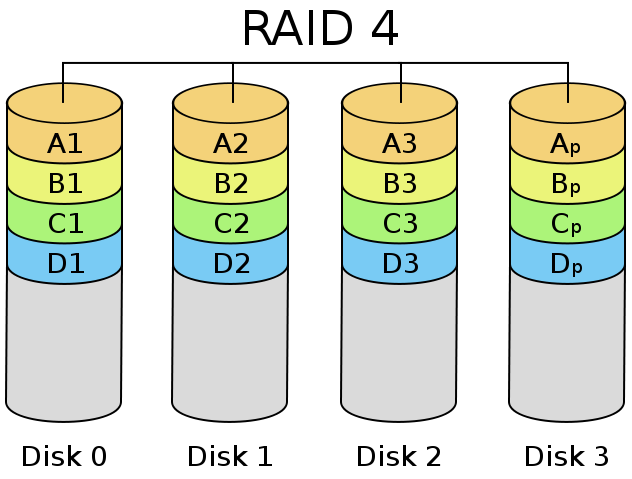
Striping (bit or byte level)
Parity
Inexpensive (1 redundant disk)
Read all disks to recover

Issue:
All reads/writes access all disks

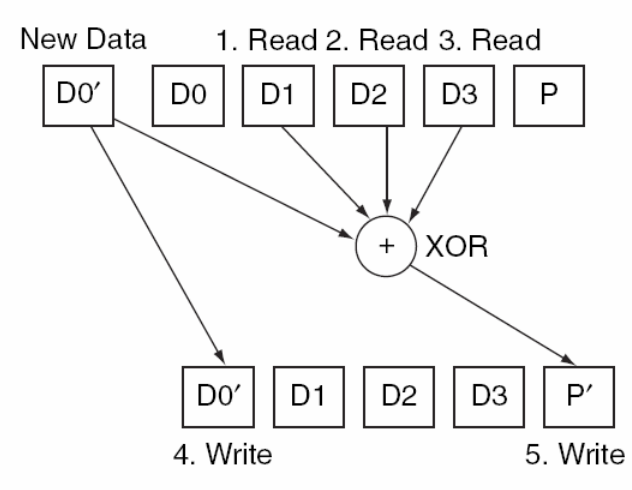
RAID 2

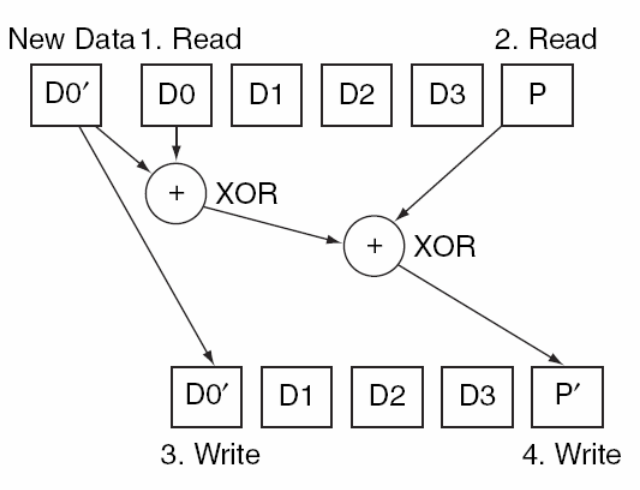
Striping at bit level
DRAM style ECC
Synchronized drives
Not used
- ECC incorporated in drives
- Large number of drives for check bits

RAID

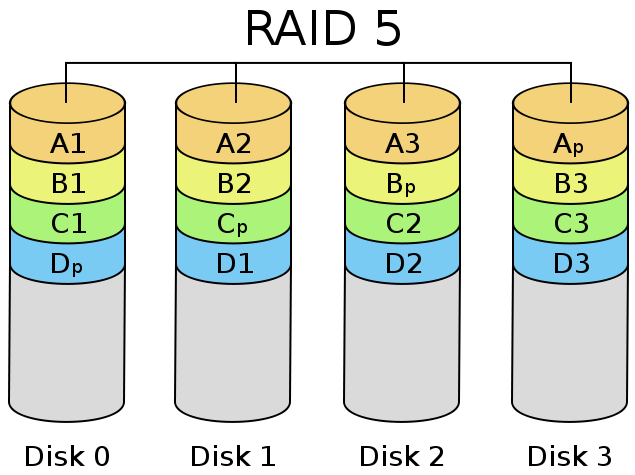


Small reads (block or smaller) require only one disk read
Small reads can occur in parallel
Small writes more complicated
But optimization possible
  instead of reading all disks and two writes
  read two disks and two writes
Problem: contention for disk with parity

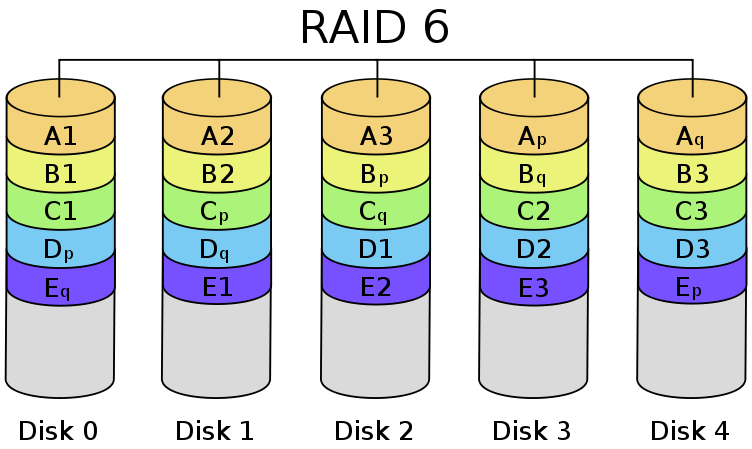




RAID

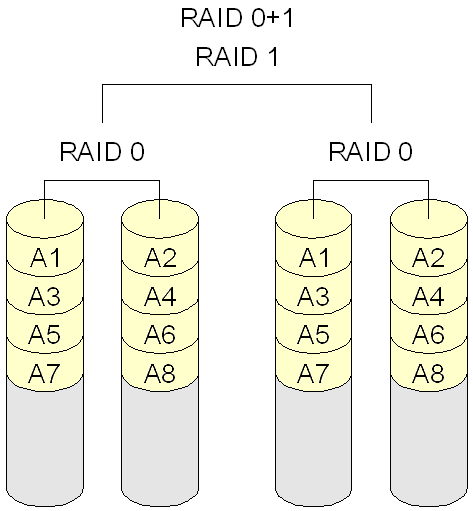


Stripe parity blocks across disks
Writes proceed in parallel if parity on different disks

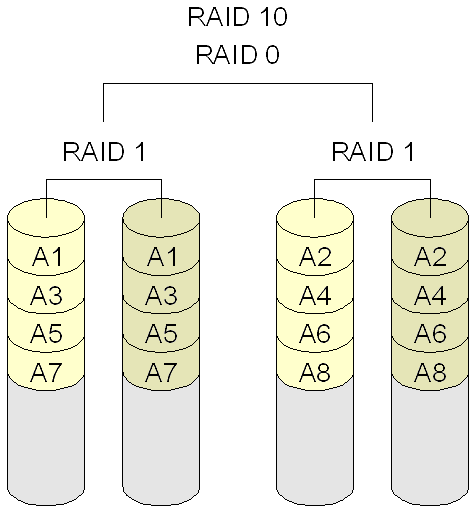


If two drives fail previous schemes fail
- Two (striped) parity checks
- Row and diagonal

RAID



Mirror of stripes



Stripe of mirrors

Others: RAID 50, RAID 60, RAID15, …

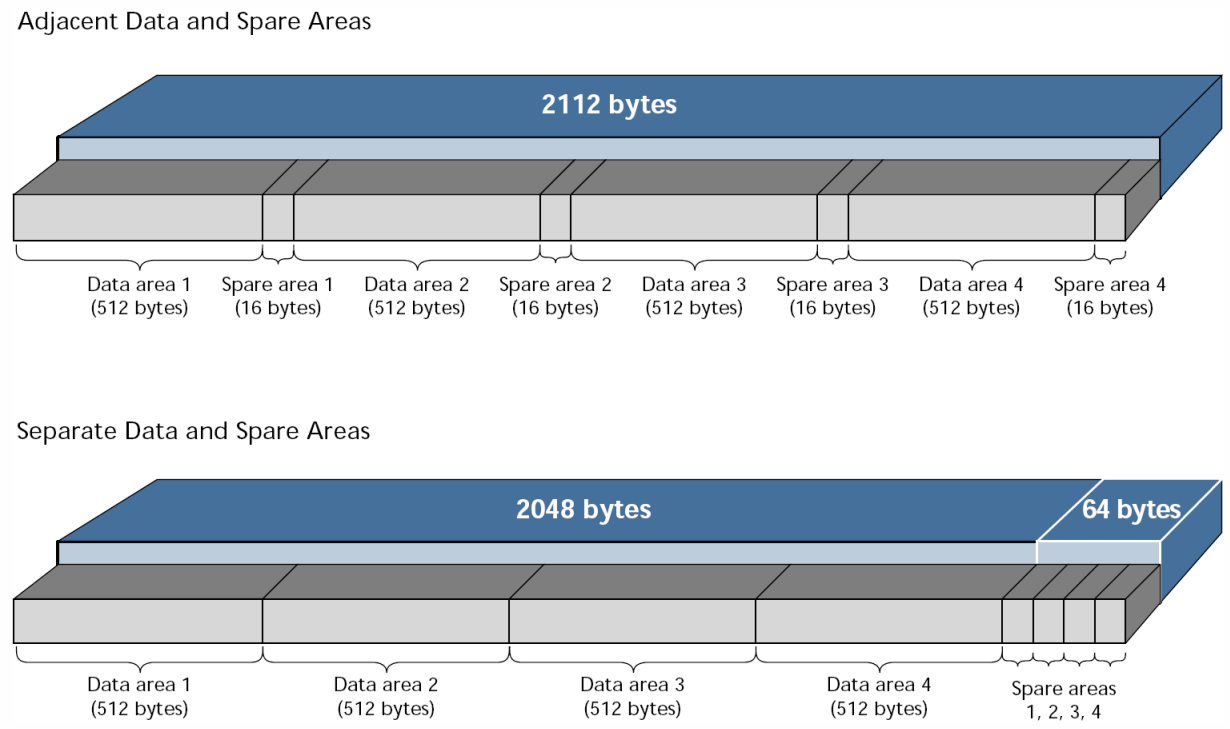
Flash Technology

- Non-volatile
- Used in cell phones, digital cameras, digital music players, USB thumb drives
- NAND vs. NOR Flash Technology
  - NOR
    SRAM-like pin interface
    Read/write bytes, erase blocks
    Random access suitable for program storage (e.g. BIOS)
  - NAND
    Dispenses with some addressing logic for increased density but page read
    Used in cameras, USB "thumb drives"
    Read/write pages, erase blocks
      Blocks comprised of pages (32-128)
      Pages of 512B-4 KB
    Spare blocks
- Both suffer from wearout
  - Maximum writes/cell

Characteristics	NOR Flash Memory	NAND Flash Memory
Typical use	BIOS memory	USB key
Minimum access size (bytes)	512 bytes	2048 bytes
Read time (microseconds)	0.08	25
Write time (microseconds)	10.00	1500 to erase + 250
Read bandwidth (MBytes/second)	10	40
Write bandwidth (MBytes/second)	0.4	8
Wearout (writes per cell)	100,000	10,000 to 100,000
Best price/GB (2008)	$65	$4

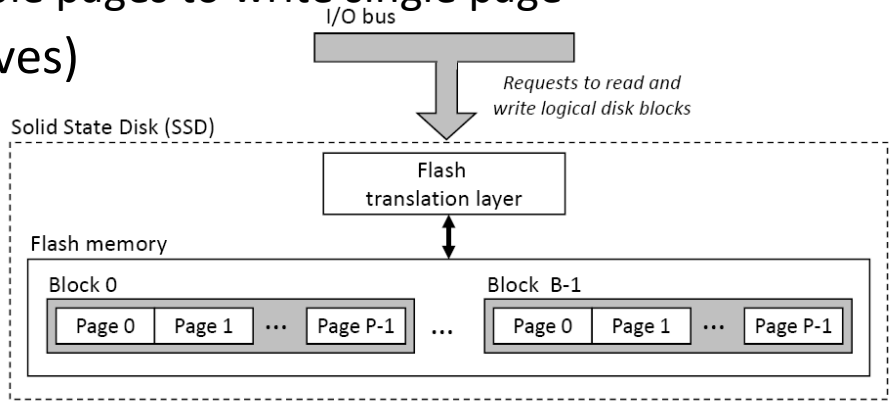
NAND Flash

2K page has spare bytes for ECC and other uses (e.g. write counts)



SSD: Solid State Disks

- Use standard HDD interface (USB, SATA)
- Translation layer (SSD controller)
  - Map disk "blocks" to block/page
  - Eliminates need to write entire block when disk block written
  - Facilitates wear leveling by writing to new block/page
  - Atomicity of writes
    - Don't have to read/write multiple pages to write single page
- Advantages over HDD (Hard Disk Drives)
  - Lower power
  - Fast boot times (low latency)
  - Greater reliability (no moving parts)
  - Faster random access times
  - Density increasing faster
    - HDD 3.5" 2000-2009: 5x (180GB to 2TB) while SSD 2001-2009: 71x
- Disadvantages
  - Cost, Data rate (improving)

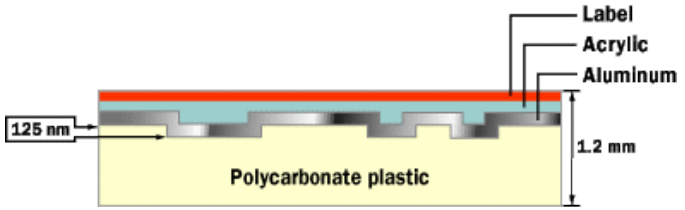


SSD: Solid State Disks

Characteristics	Kingston SecureDigital (SD) SD4/8 GB	Transend Type I CompactFlash TS16GCF133	RiDATA Solid State Disk 2.5 inch SATA
Formatted data capacity (GB)	8	16	32
Bytes per sector	512	512	512
Data transfer rate (read/write MB/sec)	4	20/18	68/50
Power operating/standby (W)	0.66/0.15	0.66/0.15	2.1/—
Size: height × width × depth (inches)	0.94 × 1.26 × 0.08	1.43 × 1.68 × 0.13	0.35 × 2.75 × 4.00
Weight in grams (454 grams/pound)	2.5	11.4	52
Mean time between failures (hours)	> 1,000,000	> 1,000,000	> 4,000,000
GB/cu. in., GB/watt	84 GB/cu.in., 12 GB/W	51 GB/cu.in., 24 GB/W	8 GB/cu.in., 16 GB/W
Best price (2008)	~ $30	~ $70	~ $300

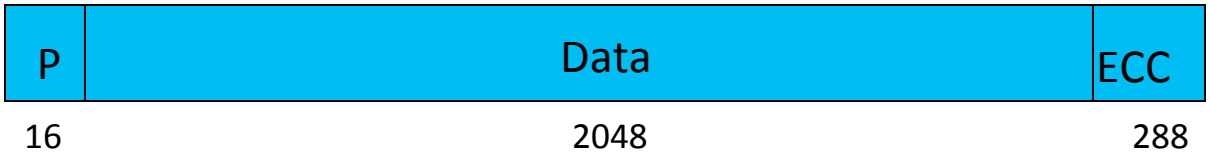
CD-ROM

- Polycarbonate plastic disc
  - Microscopic bumps/pits in spiral
  - Covered with thin layer of aluminum
  - Layer of acrylic for protection
- Single spiral track instead of concentric tracks
  - Starts at center and spirals out
  - Larger circumference at outside
  - Audio required constant stream
  - Rotational velocity decreases as head moves out
    - CLV (Constant Linear Velocity) CDs
    - CAV (Constant Angular Velocity) disk drives
  - 200 RPM to 530 RPM



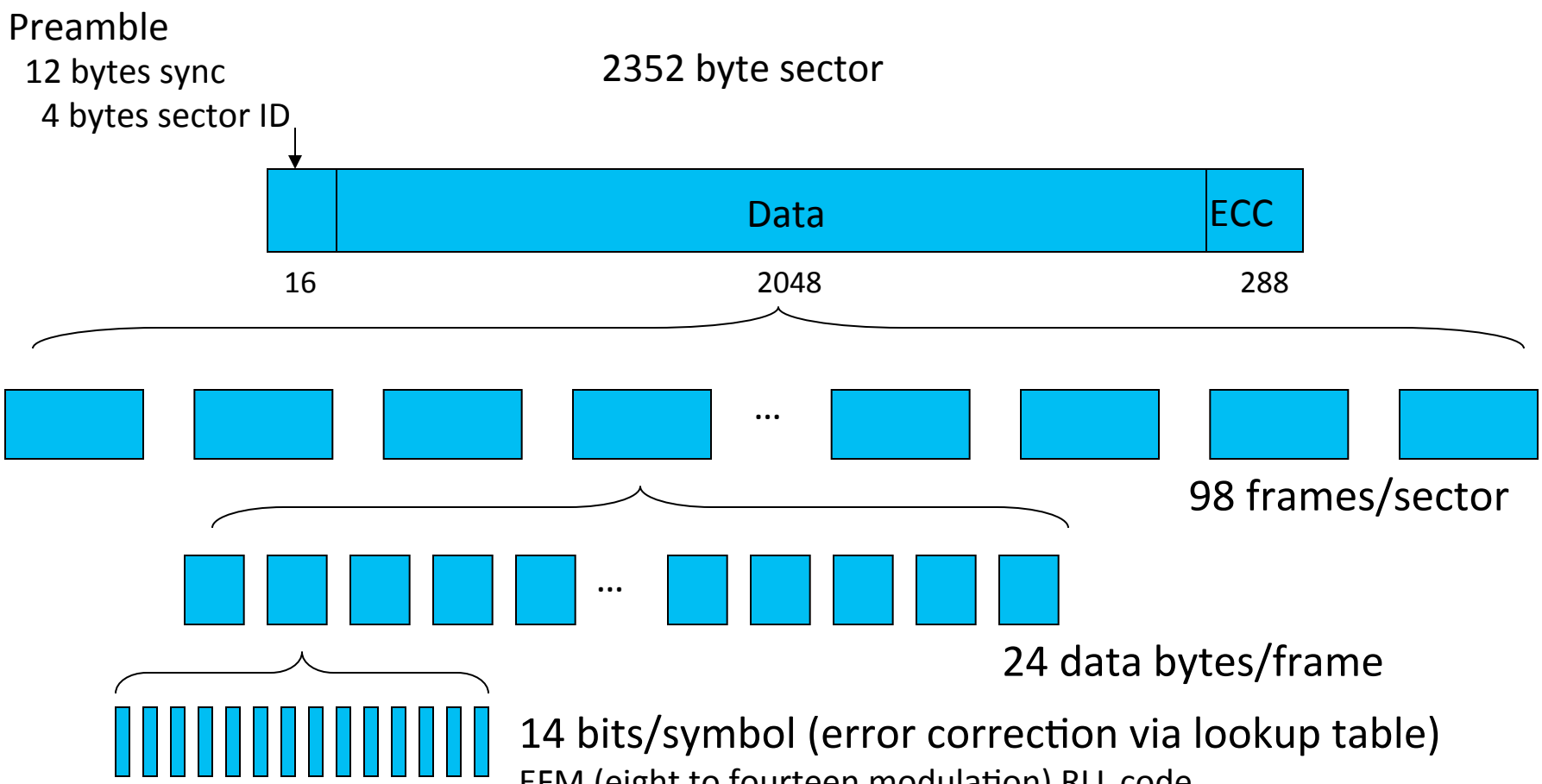
CD-ROM

- Initial Audio CDs
  - 74 minute CD
  - 333,000 sectors of 2,352 bytes/sector
  - 44,100 samples/second x 16-bit samples x 2 (stereo)
  - 176,400 bytes/second (75 sectors with 2,352 bytes/sector)
  - Entire size used (no error checking/correcting)
- Data CDs
  - Use 2,048 data bytes/sector (2^{11})
  - Remaining 304 bytes for preamble, error detection/correction
    - Music listeners unlikely to notice an occasional lost bit or two
    - PC users pretty picky about it
  - If running at same speed as audio CD
    - 75 sectors x 2,048 bytes/sector per second
    - 153.6 KB/s (1X)
  - 650MB capacity



Payload efficiency

Logical CD-ROM Layout



14 bits/symbol (error correction via lookup table)

EFM (eight to fourteen modulation) RLL code

Employs NRZI

Trade-off

Ensure minimum length of bump/land (3 clock times)

Ensure edge density

CD-ROM and Hard Disk Comparison

- 32X Speed
  - 32 x 153.6 KB/s = 4.9152 MB/s (650MB capacity)
  - Compare with magnetic hard disk drives (250GB)
    - “slower” SCSI @ 40MB/s (much less 160MB/s)
    - 60 MB/s USB 2.0
    - 50 MB/s FireWire

DVDs

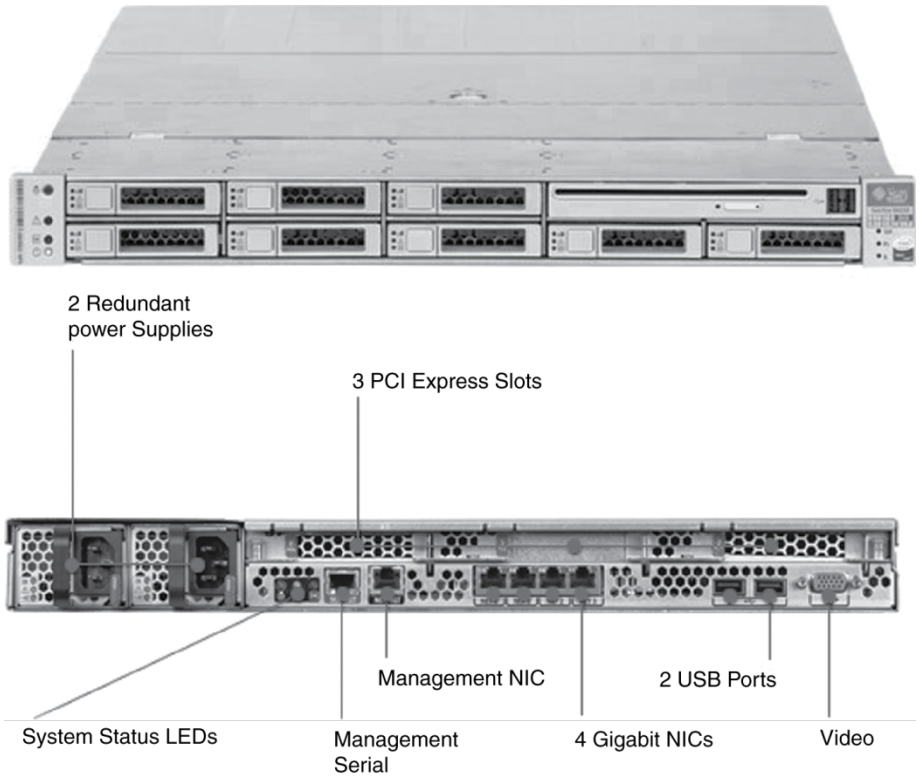
- Digital Video Disk (Digital Versatile Disk)
- Very similar to CD technology
  - Smaller pits
  - Tighter spiral
  - Shorter laser wavelength (red)
  - Faster rotational velocity
    - 1600 RPM to 570 RPM on outside
- Capacity increases to 4.7GB
- 1x DVD drive operates at 1.4MB/s

- New
  - HD-DVD (15 GB single layer, 30 GB dual layer) RIP 2/20/2008
    - Toshiba, Hitachi
  - Blu-Ray (25 GB single layer, 50 GB dual layer)
    - Sony, Philips et.al.

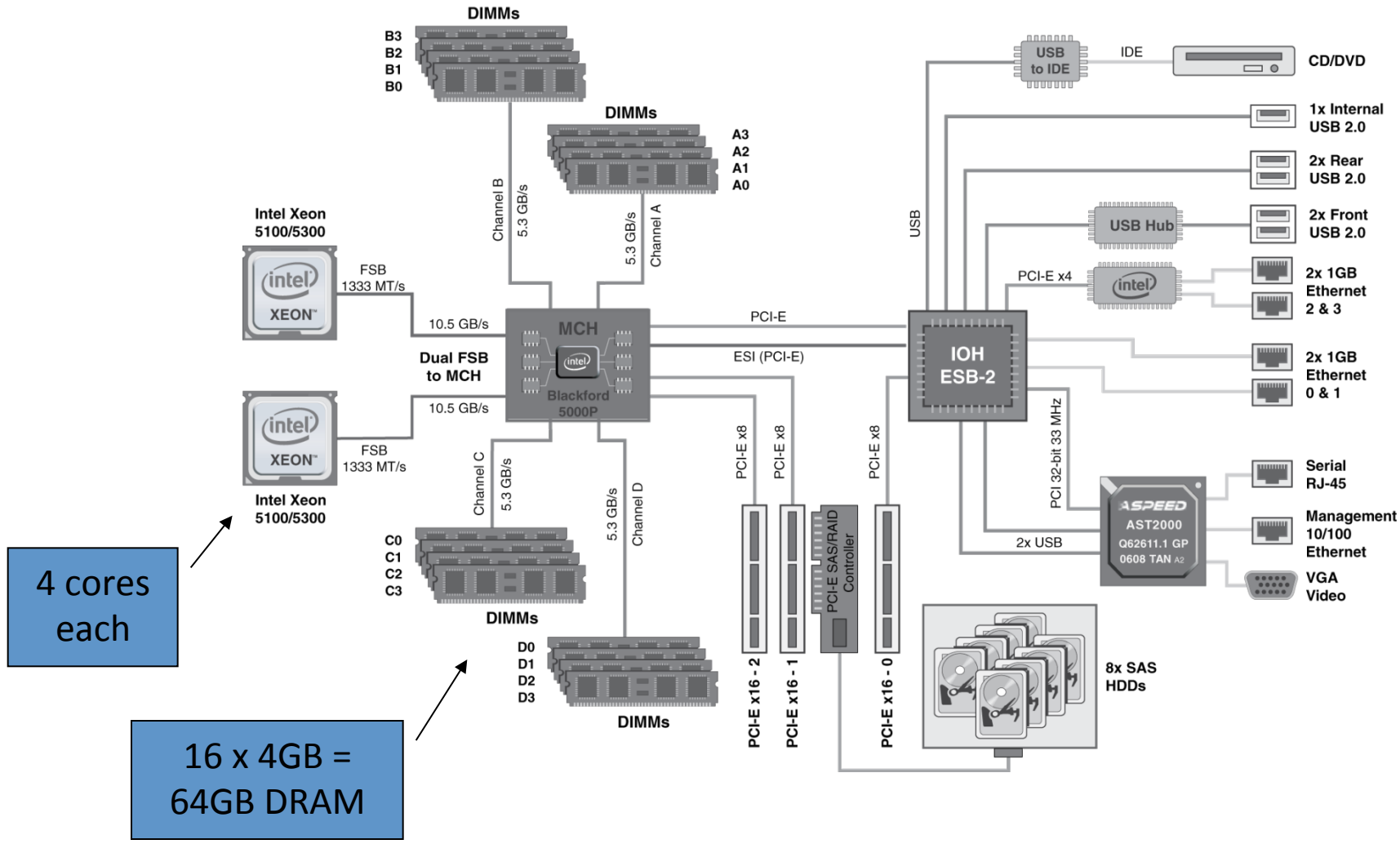
Rack-mounted Servers

19” (width) racks
multiples of 1.75” (1U) high

Sun Fire x4150 1U server



Sun Fire x 4150 Server



4 cores each

16 x 4GB = 64GB DRAM

I/O Subsystem Design Example

- Given a Sun Fire x4150 system with
  - Workload: 64KB disk reads
    - Each I/O op requires
      - 200,000 user-code instructions
      - 100,000 OS instructions
  - Each CPU: 10^9 instructions/s
  - FSB: 10.6 GB/s peak
  - DRAM DDR2 667MHz: 5.336 GB/s
  - PCI-E 8× bus: 8 × 250MB/sec = 2GB/s
  - Disks: 15,000 rpm, 2.9ms avg. seek time, 112MB/s transfer rate
- What I/O rate can be sustained?
  - For random reads
  - For sequential reads

Determine Operations/Second

- I/O rate for CPUs
  - Per core: 10^9/(100,000 + 200,000) = 3,333
  - 8 cores: 26,667 ops/sec
- Random reads, I/O rate for disks
  - Assume actual seek time is average/4
  - Time/op = seek + latency + transfer
    = 2.9ms/4 + 4ms/2 + 64KB/(112MB/s) = 3.3ms
  - 303 ops/s per disk, 2424 ops/s for 8 disks
- Sequential reads
  - 112MB/s / 64KB/op = 1750 ops/s per disk
  - 14,000 ops/s for 8 disks

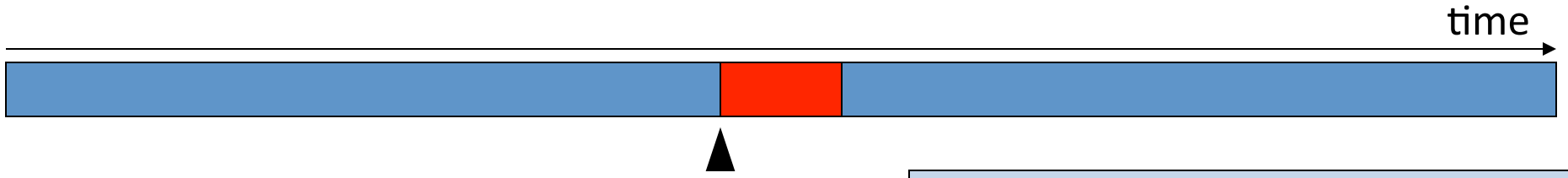
Determine Operations/Second

- PCI-E I/O rate
  - 2GB/s / 64KB/op = 31,250 ops/s
- DRAM I/O rate
  - 5.336 GB/s / 64KB/op = 83,375 ops/s
- FSB I/O rate
  - Assume we can sustain half the peak rate
  - 5.3 GB/s / 64KB/op = 81,540 ops/s per FSB
  - 163,080 ops/s for 2 FSBs
- Weakest link: disks
  - 2424 ops/s random, 14,000 ops/sec sequential
  - Other components have ample headroom to accommodate these rates

Reliability

Service accomplishment vs. Service interruption

- System operating/available according to service level agreement (SLA)
- System unavailable or performance not as agreed
- Failure which causes disruption of service



MTTF	Mean Time To Failure
MTTR	Mean Time To Repair
MTBF	Mean Time Between Failures
FIT	Failures in Time (10^9 hours operation)

$$FIT = \frac{10^9}{MTTF} \qquad MTBF = MTTF + MTTR$$

$$Availability = \frac{\text{time available}}{\text{total time}} = \frac{MTTF}{MTTF + MTTR} = \frac{MTTF}{MTBF}$$

Averages!
Modules → System

Reliability: An Example

Assume a disk subsystem with the following components and MTTF:

10 disks, each rated at 1,000,000-hour MTTF
1 SCSI controller, 500,000-hour MTTF
1 power supply, 200,000-hour MTTF
1 fan, 200,000-hour MTTF
1 SCSI cable, 1,000,000-hour MTTF

Component lifetimes are exponentially distributed, failures independent, compute MTTF of the system as a whole

$$\text{Failure rate}_{\text{system}} = 10 \times \frac{1}{1{,}000{,}000} + \frac{1}{500{,}000} + \frac{1}{200{,}000} + \frac{1}{200{,}000} + \frac{1}{1{,}000{,}000}$$

$$= \frac{10 + 2 + 5 + 5 + 1}{1{,}000{,}000} = \frac{23}{1{,}000{,}000} = \frac{23{,}000}{1{,}000{,}000{,}000}$$

$$\text{MTTF} = \frac{1}{\text{Failure rate}} = \frac{1{,}000{,}000{,}000}{23{,}000} = 43{,}500 \text{ hours } (< 5 \text{ years})$$

Reliability – Exploiting Redundancy

Assume example from before but we add a second (redundant) power supply.
Maintain assumption about independence of failures!
MTTF for redundant supplies is mean time until one power supply fails divided by the chance that the second fails before the first is replaced.

$$\text{MTTF of single supply failure} = \frac{\text{MTTF}_{\text{power supply}}}{2}$$

$$\text{Probability of a second failure before first is repaired} = \frac{\text{MTTR}_{\text{power supply}}}{\text{MTTF}_{\text{power supply}}}$$

$$\text{MTTF}_{\text{power supply pair}} = \frac{\frac{\text{MTTF}_{\text{power supply}}}{2}}{\frac{\text{MTTR}_{\text{power supply}}}{\text{MTTF}_{\text{power supply}}}} = \frac{\text{MTTF}^2_{\text{power supply}} / 2}{\text{MTTR}_{\text{power supply}}} = \frac{\text{MTTF}^2_{\text{power supply}}}{2 \times \text{MTTR}_{\text{power supply}}}$$

assuming it takes 24 hours to detect and replace failed power supply...

$$= \frac{200{,}000^2}{2 \times 24} \cong 830{,}000{,}000$$

making the pair about 4150 times more reliable than a single power supply