16.480/552 Micro II

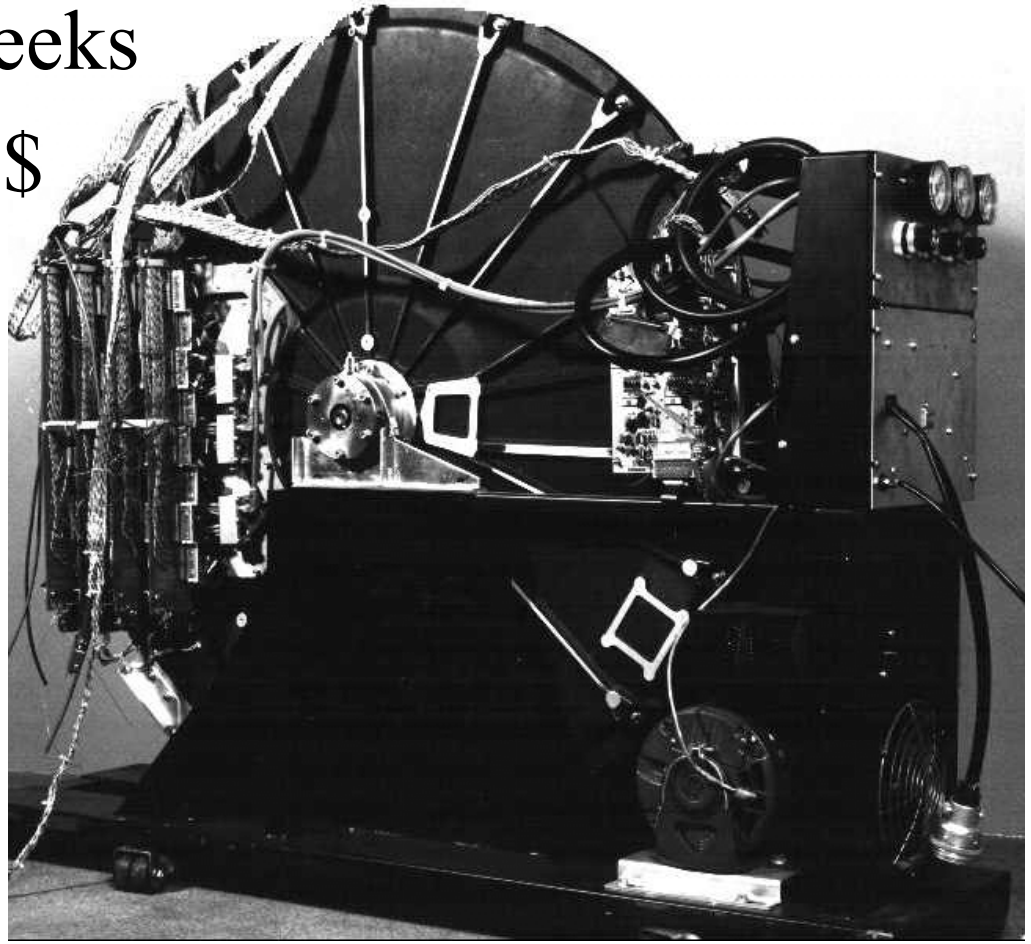
Hard Disks, ATA, PCMCIA etc.

Prof. Yan Luo

UMass Lowell

Disks of 30 Years Ago

- 10 MB
- Failed every few weeks
- Cost more than 400$



Disk Arrays

- 24 cpus
- 384 disks
- More mips in the disks than in the cpus

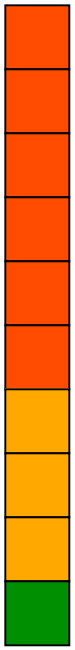
Reference: Bitton & Gray: The Rebirth of Database Machines.
http://research.microsoft.com/~Gray/talks/vldb98.ppt

Disk Terms

- Disks are called *platters*
- Data is recorded on *tracks* (circles) on the disk.
- Tracks are formatted into fixed-sized *sectors*.
- A pair of *Read/Write heads* for each platter
- Mounted on a *disk arm*
- Client *addresses* logical blocks (cylinder, head, sector)
- Disk arm Seeks to a cylinder,
  selects a head
  waits for data to rotate under head
  transfers data to or from
- Bad blocks are *remapped* to spare good blocks.

Disk Access Time

- Access time = SeekTime 6 ms
  + RotateTime 3 ms
  + ReadTime 1 ms



- Rotate time:
  - 5,000 to 10,000 rpm
    - ~ 12 to 6 milliseconds per rotation
    - ~ 6 to 3 ms rotational latency

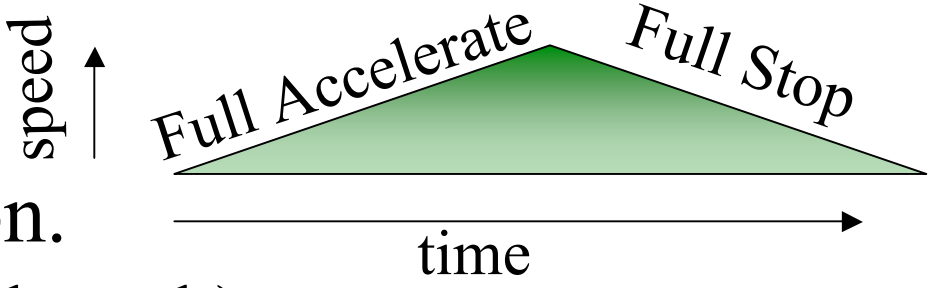
Disk Access Time Improves Slowly

• Access time =	SeekTime	6 ms	8%/y
	+ RotateTime	3 ms	8%/y
	+ ReadTime	1 ms	40%/y

- Other useful facts:
  - Power rises more than size^3 (small is indeed beautiful)
  - Small devices are more rugged
  - Small devices can use plastics (forces are much smaller) e.g. bugs fall without breaking anything

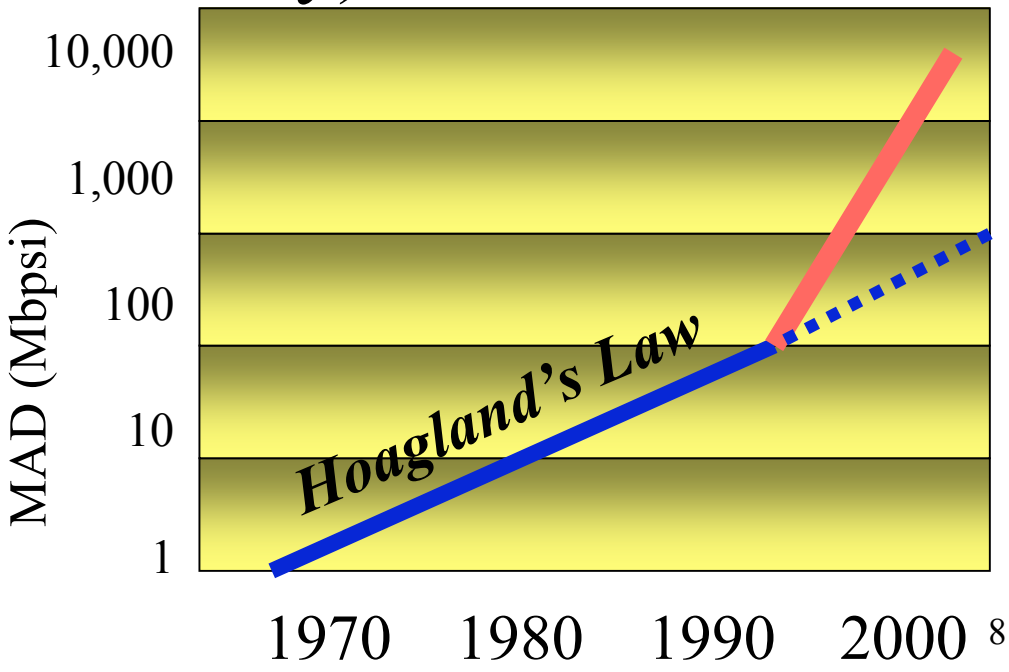
Disk Seek Time

- Seek time is ~ Sqrt(distance)
  (distance = 1/2 acceleration x time^2)
- Specs assume seek is
  1/3 of disk
- Short seeks are common.
  (over 50% are zero length)
- Typical 1/3 seek time: 6 ms
- 4x improvement in 20 years.



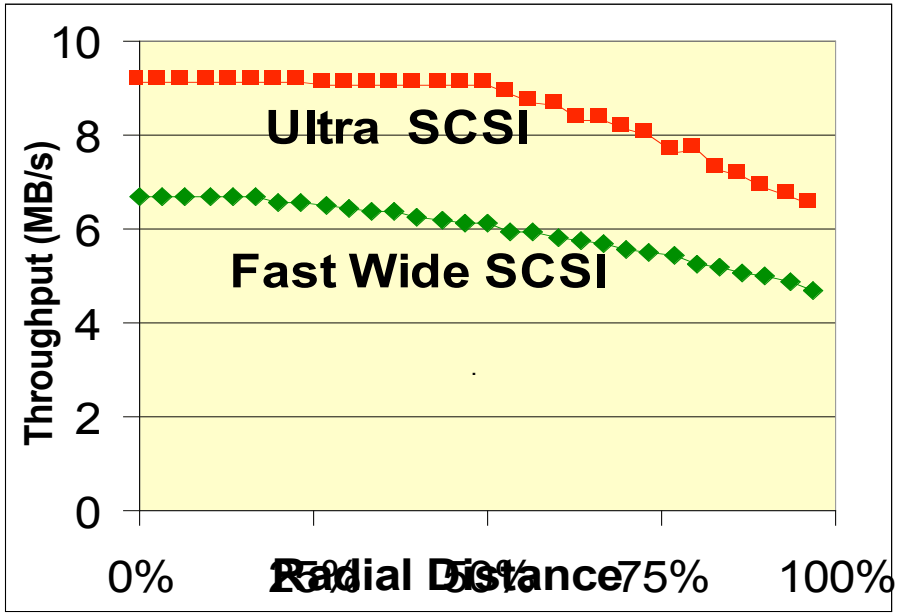
Read/Write Time: Density

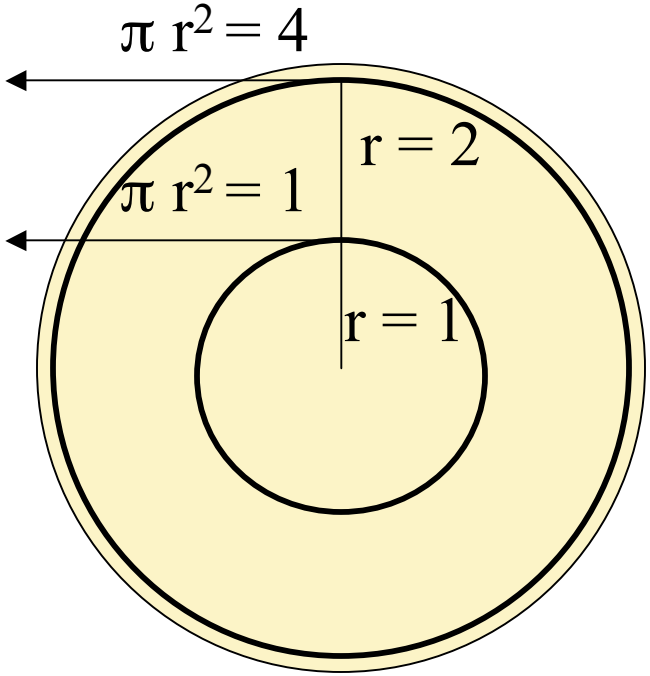
- Time = Size / BytesPerSecond
- Bytes/Second = Speed * LinearDensity
  - 5 to 15 MBps
- MAD (Magnetic Aerial Density)
  - Today 3 Gbits/inch^2
    11 gbpsi in lab
  - Rising > 60%/year
  - ParaMagnetic Limit:
    ??? ~ 60 Gb/inch^2
  - linear density is sqrt
    10x per decade



Read/Write Time: Rotational Speed

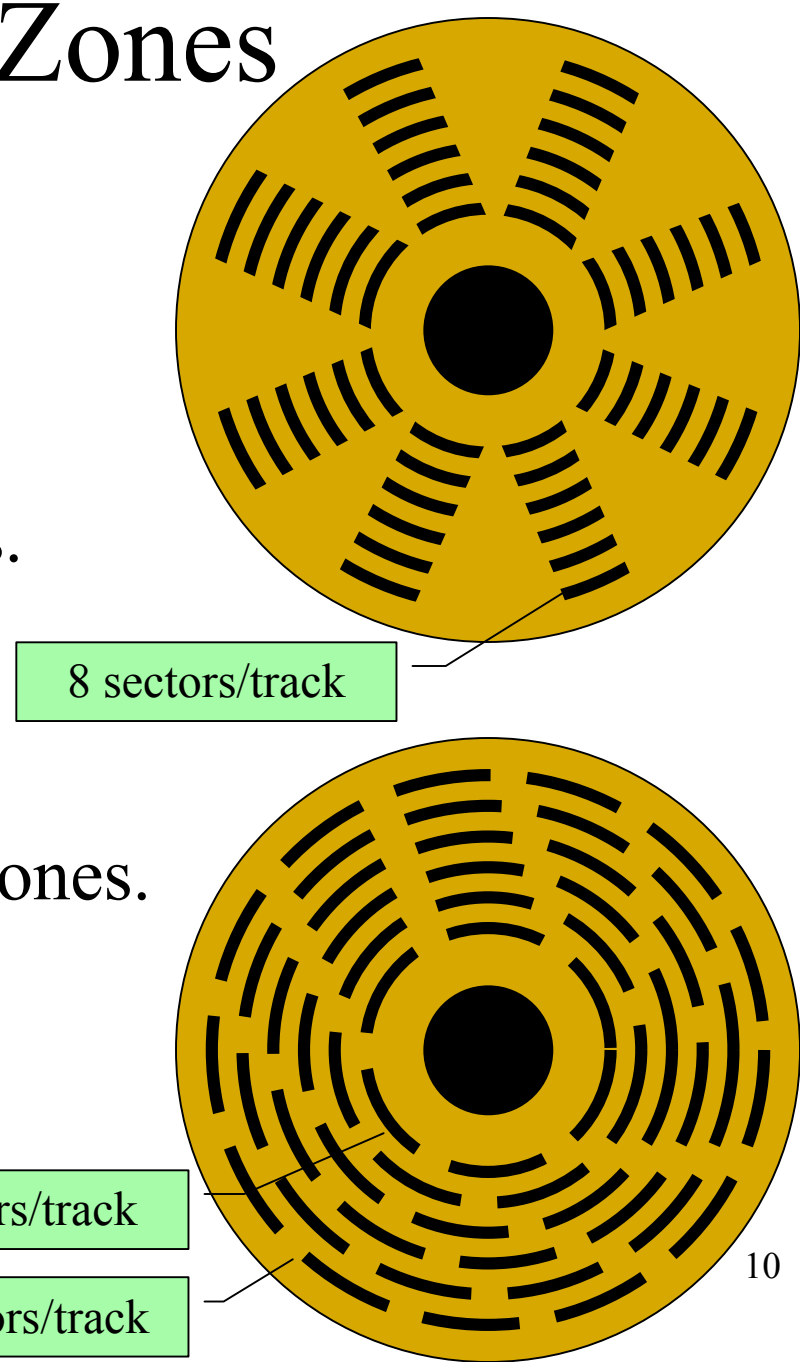
- Bytes/Second = Speed * Density
- Speed greater at edge of circle
- Speed 3600 -> 10,000 rpm
  - 5%/year improvement
- bit rate varies by ~1.5x today





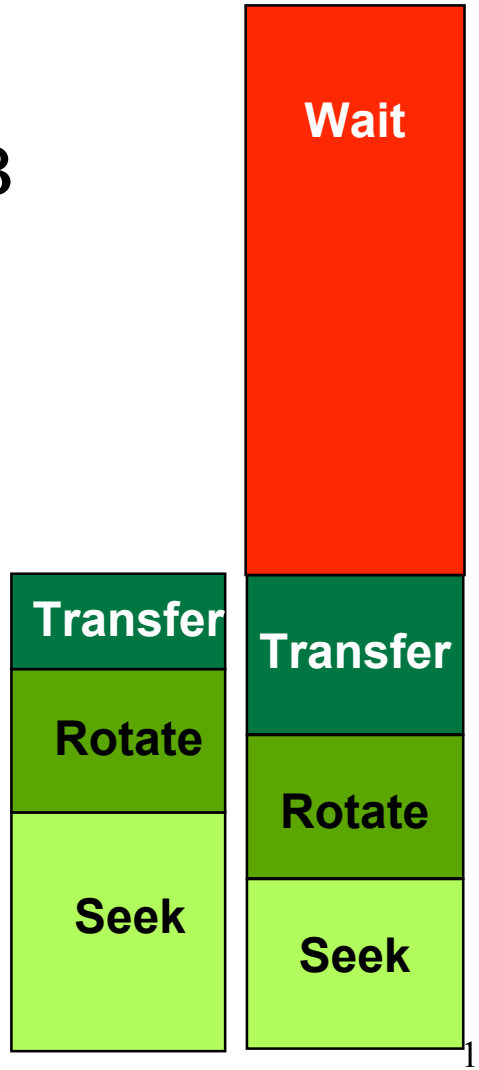
Read/Write Time: Zones

- Disks are sectored
  - typical: 512 bytes/sector
  - Sector is read/write unit
  - Failfast: can detect bad sectors.
- Disks are zoned
  - outer zones have more sectors
  - Bytes/second higher in outer zones.



The Access Time Myth

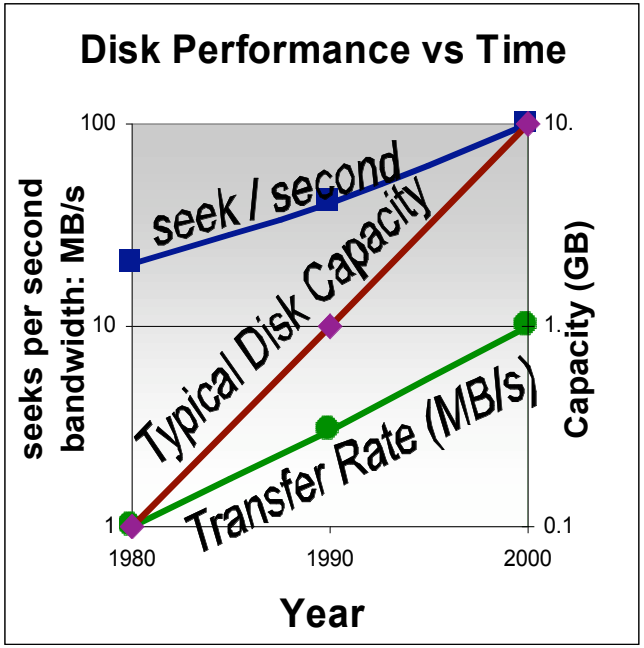
- The Myth: seek time dominates
- The Reality: (1) Queuing dominates
  (2) Transfer dominates BLOB
  (3) Disk seeks often short
- Implication: many cheap servers
  better than one fast expensive server
  - shorter queues
  - parallel transfer
  - lower cost/access and cost/byte
- This is now obvious for disk arrays
- This will be obvious for tape arrays

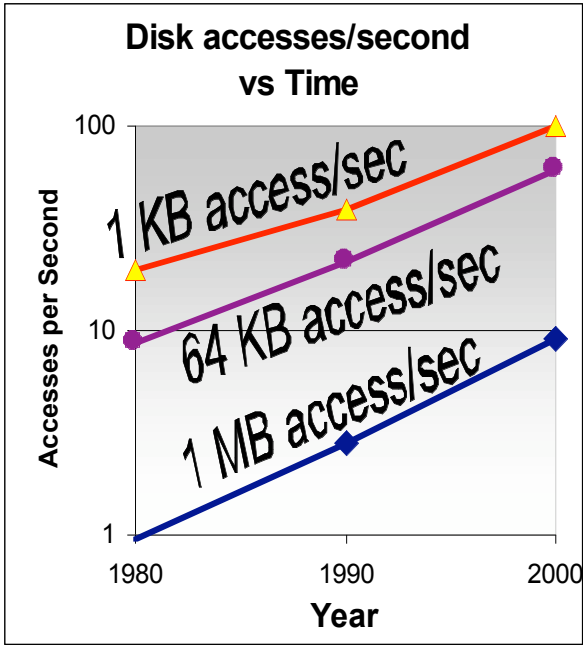


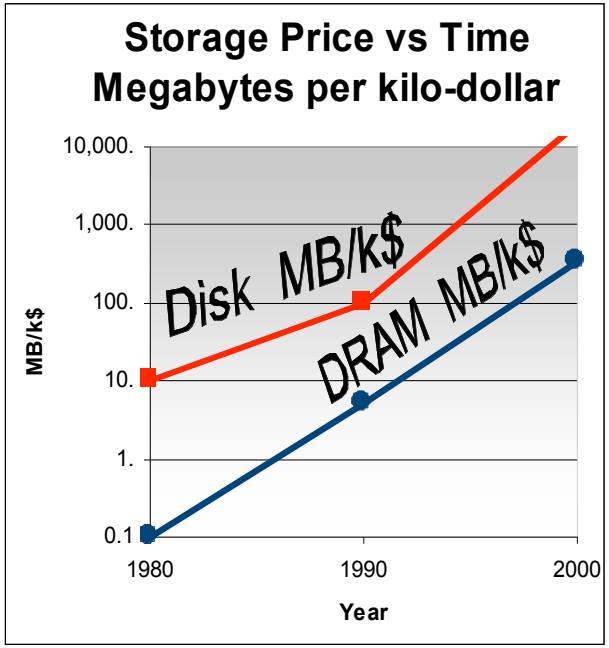
Storage Ratios Changed

- 10x better access time
- 10x more bandwidth
- 4,000x lower media price

- DRAM/disk media price ratio changed
  - 1970-1990 100:1
  - 1990-1995 10:1
  - 1995-1997 50:1
  - today ~ .2$pMB disk
    10$pMB dram







Disk Access Ratios Have Changed

- Key metrics:
  - $/GB
  - Kaps/GB (KB accesses per second per GB)
  - SCAN: time to scan the disk
- Scan going from minutes to days
- Disk arms are precious resource

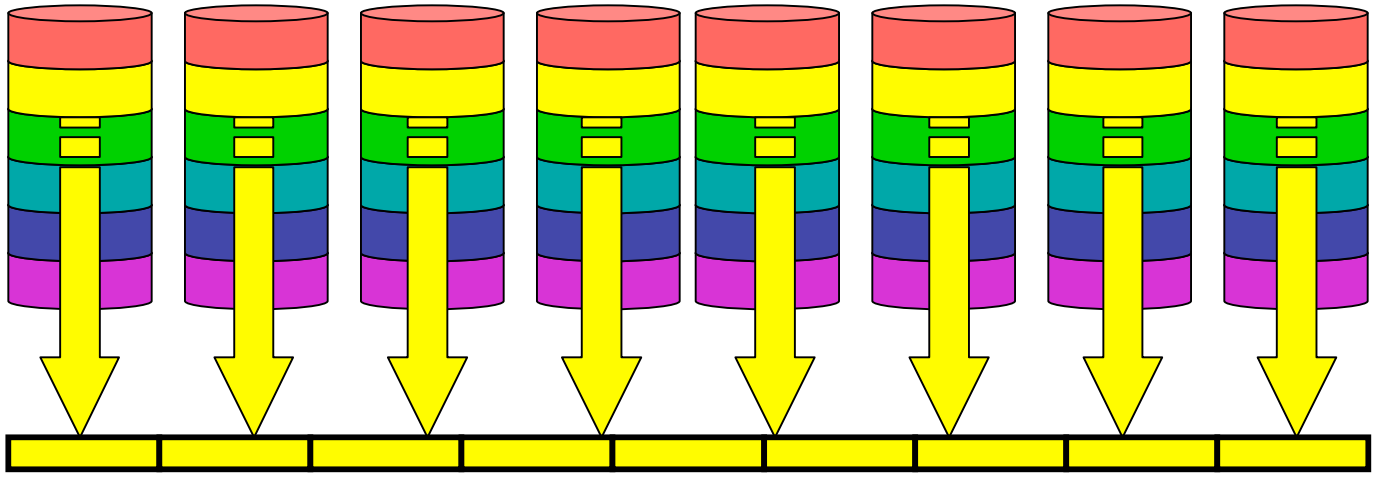
  (disk capacity is no longer the precious resource)

Kaps/GB went from 500 to 7 and going to 1

year	Capacity GB	$/GB	kaps	kaps/ GB	Scan Sequential	Scan Random
1988	0.25	20,000	30	120	2 minutes	20 minutes
1998	18	50	120	7	20 minutes	5 hrs
2003	200	5	200	1	2 hrs	1.2 days

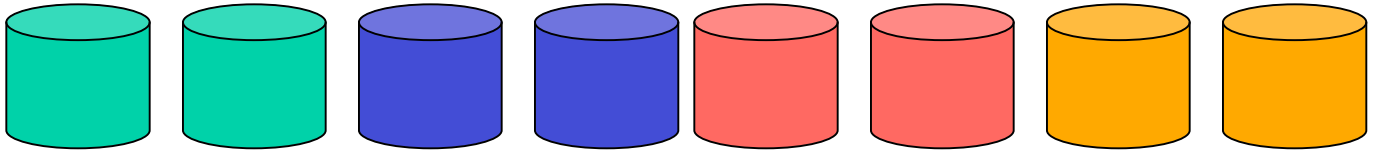
Stripe For More Bandwidth

- N-stores have N-times the bandwidth
- Works great!
- Supported by most file systems

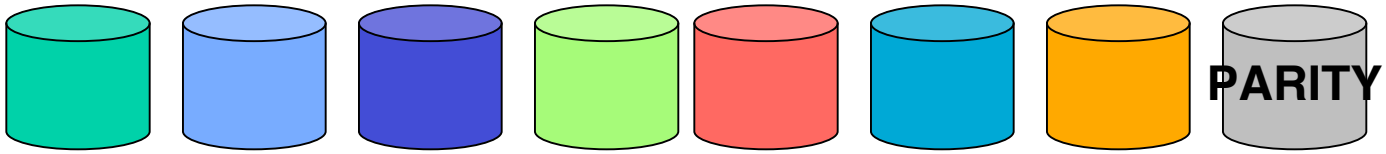


RAID

- redundant array of independent disks
- RAID 0: Striped Set
- RAID 1: Mirrors: 50% storage overhead
  - read one, write both



- RAID5: 12% Storage overhead:
  - read one, write one plus parity



ATA, SATA, etc.

- **ATA or PATA** - Parallel ATA - the traditional parallel ATA interface - used by disk drives.
  - ATA is synonym of IDE or EIDE
- **ATAPI or PATAPI** - ATAPI using PATA - used by CD, DVD and tape devices.
- **SATA** - Serial ATA - serial version of parallel ATA - mostly used by disk drives.
- **SATAPI** - Serial ATAPI - ATAPI using SATA - used by CD, DVD and tape devices.

Parallel ATA (PATA/PATAPI)

- PATA and PATAPI is the traditional ATA interface that has been standardized for 10+ years by the ANSI/INCITS ATA-1 through ATA/ATAPI-6 standards.
- This interface is widely used by PATA disk drives and PATAPI CD/DVD and tape drives.
- PCMCIA PC Card ATA devices and Compact Flash (CF) devices use a PATA based interface.

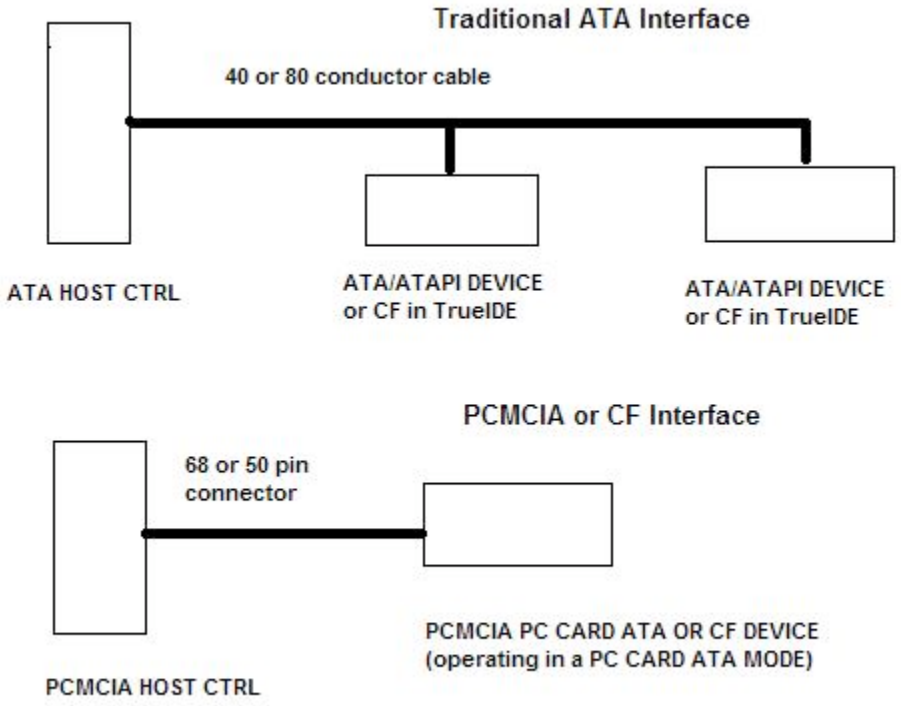
ATA device

- A mass storage device that stores data in 512-byte chunks called sectors.
- sectors are transferred to/from the device in contiguous blocks of data that are a multiple of 512 bytes.
- Each sector stored by a device has a unique sector address that is called a Logical Block Address (LBA).
  - The first sector on a device is at LBA 0, the next at LBA 1, etc.

ATAPI device

- An ATAPI device is really a SCSI device that uses the ATA interface.
- ATA/ATAPI is one of many SCSI physical interfaces.
  - simplest SCSI physical interfaces.
- An ATAPI device
  - implement all the ATA signals and most of the ATA command protocols
  - an ATAPI device uses SCSI commands.
  - store data in a variety of formats and block sizes
    - 512 bytes used by many SCSI disk drives, 2048 bytes used for CD/DVD data, 2352 bytes used by CD-DA (digital audio) data (music)
    - Many SCSI commands, such as Request Sense, transfer only a few bytes of data.

ATA Physical Interface



ATA (IDE) Signals

- DD0-15: data bus
- DA0-2: address
- CS0-, CS1-: address
- DIOR-, DIOW- : read/write for PIO
- DMARQ, DMACK-, INTQ, IORDY
  - DMA control signals

Signal name	Connector contact
RESET-	1
Ground	2
DD7	3
DD8	4
DD6	5
DD9	6
DD5	7
DD10	8
DD4	9
DD11	10
DD3	11
DD12	12
DD2	13
DD13	14
DD1	15
DD14	16
DD0	17
DD15	18
Ground	19
(keypin)	20
DMARQ	21
Ground	22
DIOW-	23
Ground	24
DIOR-	25
Ground	26
IORDY	27
CSEL	28
DMACK-	29
Ground	30
INTRQ	31
Reserved	32
DA1	33
PDIAG-	34(see note)
DA0	35
DA2	36
CS0-	37
CS1-	38
DASP-	39
Ground	40

IDE Command and Control Registers

HEX	BINARY	DESCRIPTION
1FX	0001 1111 XXXX	Primary Command Registers
1F0		Data Port
1F1		Error
1F2		Sector Count
1F3		Sector Number
1F4		Cylinder Low
1F5		Cylinder High
1F6		Drive/Head
1F7		Status (read), Command (write)
3FX	0011 1111 XXXX	Primary Control Registers
3F6		Alternative Status
3F7		Driver address
17X	0001 0111 XXXX	Secondary Command Registers
37X	0011 0111 XXXX	Secondary Control Registers

A Typical ATA Commanding

- IDENTIFY_DEVICE (ECh)
  - Host writes command (ECh) to 1F7h
  - Host writes device number (A0) to 1F6h
  - Drive prepares identification information (128 words)
  - Drive sets DRDY, DRQ bits in status register (1f7)
  - Host checks status register and read identification information.

DMA read

- Host writes DMA read command (C8h) to 1F7h
- Host writes device number (A0) to 1F6h, LBA to 1F3h, 1F4h, 1F5h, and sector count to 1F2h
- Host writes DMA controller to initiate DMA
- Host waits for interrupt
- Drive performs read from disk
- DMA controller transfer data from drive to memory
- DMA controller triggers interrupt to host CPU

PCMCIA

- ***P**ersonal **C**omputer **M**emory **C**ard **I**nternational **A**ssociation*
  - A standard for small, credit card-sized devices, called *PC Cards*
- Type I cards
  - up to 3.3 mm thick
  - for adding additional ROM or RAM to a computer.
- Type II cards
  - up to 5.5 mm thick. The slot an hold Type I card.
  - E.g modem cards.
- Type III cards
  - up to 10.5 mm thick.
  - The slot can hold up to 2 Type I or II cards
  - sufficiently large for portable disk drive.

Compact Flash Cards

- CompactFlash is a small, removable mass storage device.
  - First introduced in 1994
  - CF cards weigh a half ounce and are the size of a matchbook.
  - They provide complete PCMCIA-ATA functionality and compatibility.
  - Type I and II (thicker)
- Based on flash technology
  - a nonvolatile storage solution that does not require a battery to retain data indefinitely.
  - Erased and programmed in blocks
  - supports data rates up to 66MB/sec and capacities up to 137GB.
  - Dual voltage 3.3V or 5V