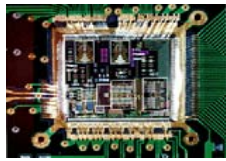
Lecture 8: Memory

CSCE5610 Computer System Architecture
CSCE4610 Computer Architecture

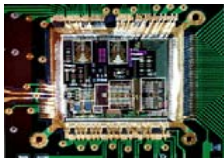
Instructor: Saraju P. Mohanty, Ph. D.

NOTE: The figures, text etc included in slides are borrowed from various books, websites, authors pages, and other sources for academic purpose only. The instructor does not claim any originality. (This slide set is adopted from Dr. Rabi Mahapatra, TAMU.)



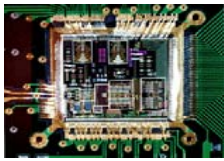
Outline

- Disk Basics
- Disk History
- Disk fallacies and performance
- FLASH
- Tapes
- RAID



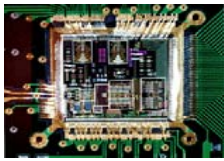
Motivation: Who Cares About I/O?

- CPU Performance: 60% per year
- I/O system performance limited by *mechanical* delays (disk I/O)
  < 10% per year (IO per sec)
- Amdahl's Law: system speed-up limited by the slowest part!
  10% IO & 10x CPU => 5x Performance (lose 50%)
  10% IO & 100x CPU => 10x Performance (lose 90%)
- I/O bottleneck:
  Diminishing fraction of time in CPU
  Diminishing value of faster CPUs

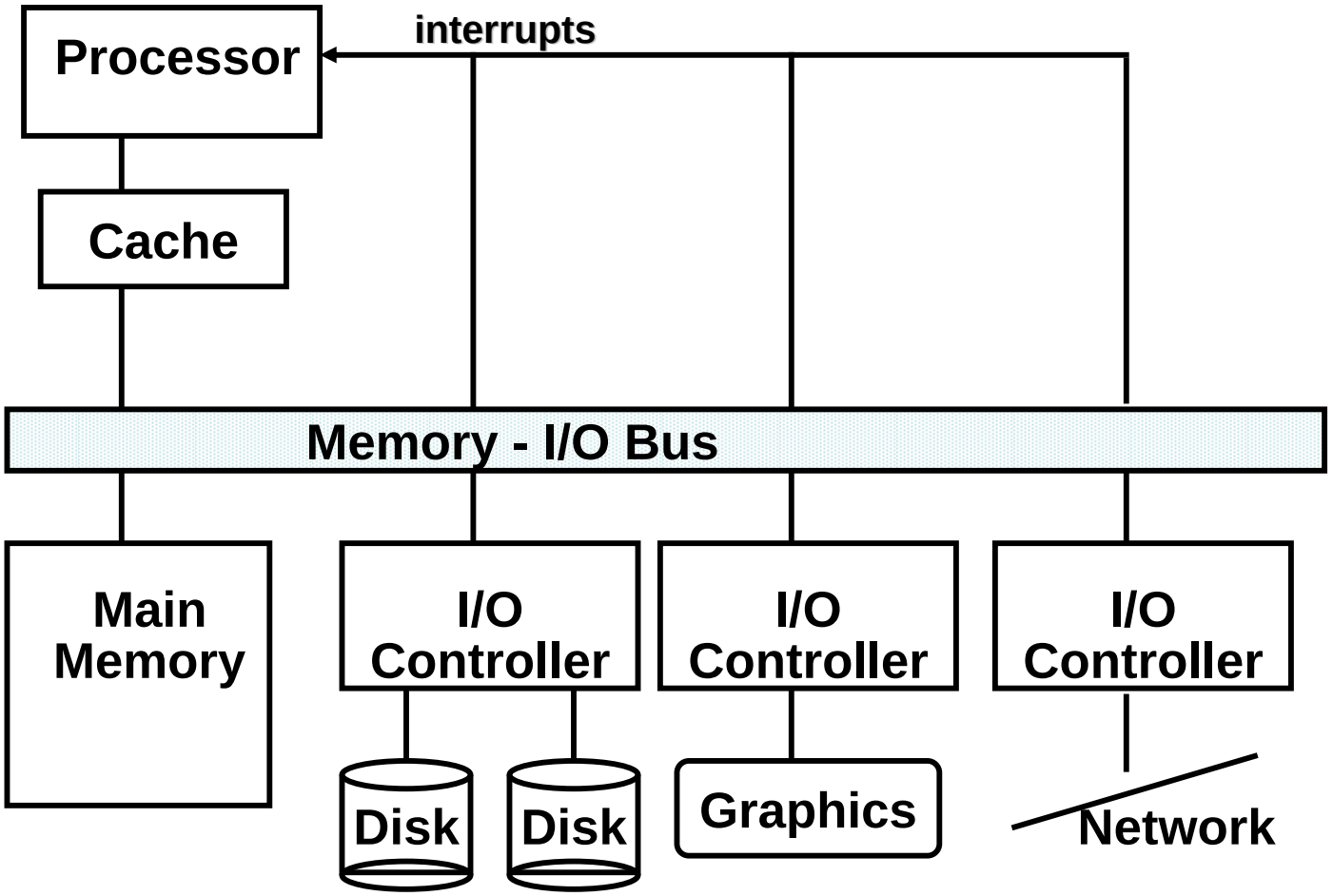


Big Picture: Who cares about CPUs?

- Why still important to keep CPUs busy vs. IO devices ("CPU time"), as CPUs not costly?
  - Moore's Law leads to both large, fast CPUs but also to very small, cheap CPUs
  - 2001 Hypothesis: 600 MHz PC is fast enough for Office Tools?
  - PC slowdown since fast enough unless games, new apps?
- People care more about about storing information and communicating information than calculating
  - "Information Technology" vs. "Computer Science"

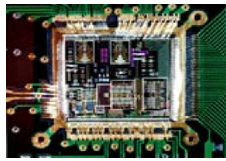


I/O Systems

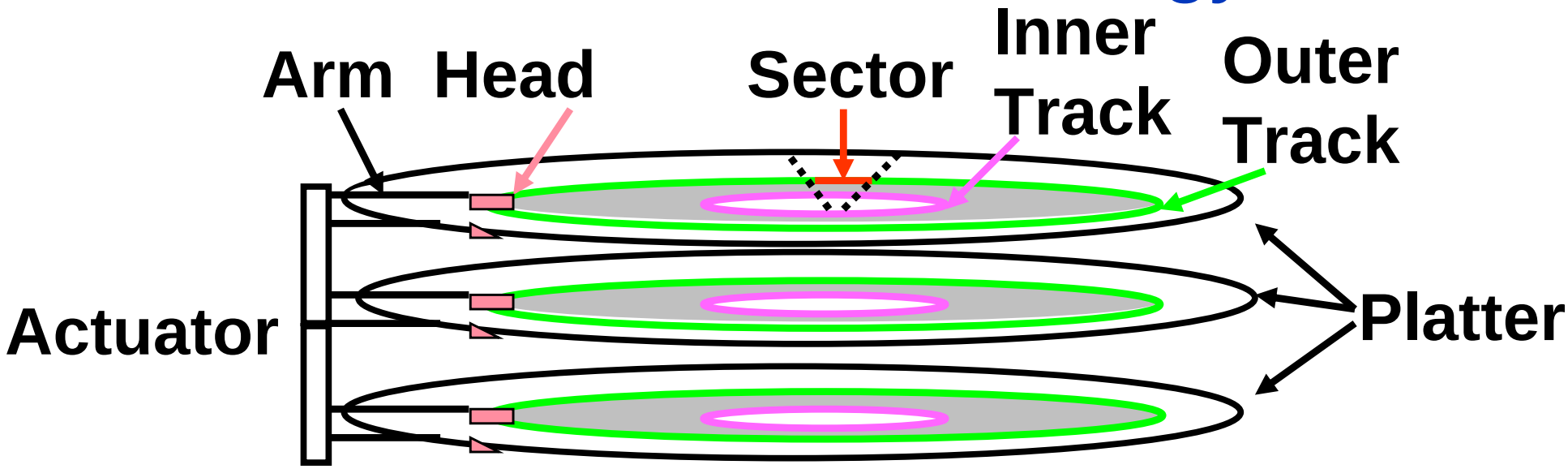

Processor
interrupts
Cache
Memory - I/O Bus
Main Memory
I/O Controller
I/O Controller
I/O Controller
Disk
Disk
Graphics
Network

Storage Technology Drivers

- Driven by the prevailing computing paradigm
  - 1950s: migration from batch to on-line processing
  - 1990s: migration to ubiquitous computing
    - computers in phones, books, cars, video cameras, …
    - nationwide fiber optical network with wireless tails
- Effects on storage industry:
  - Embedded storage
    - smaller, cheaper, more reliable, lower power
  - Data utilities
    - high capacity, hierarchically managed storage



Disk Device Terminology



- Usually several platters, with information recorded magnetically on both surfaces.
- Bits recorded in tracks, which in turn divided into sectors (e.g., 512 Bytes).
- Actuator moves head (end of arm,1/surface) over track (“seek”), select surface, wait for sector rotate under head, then read or write:
  - “Cylinder”: all tracks under heads

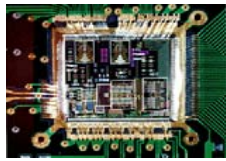
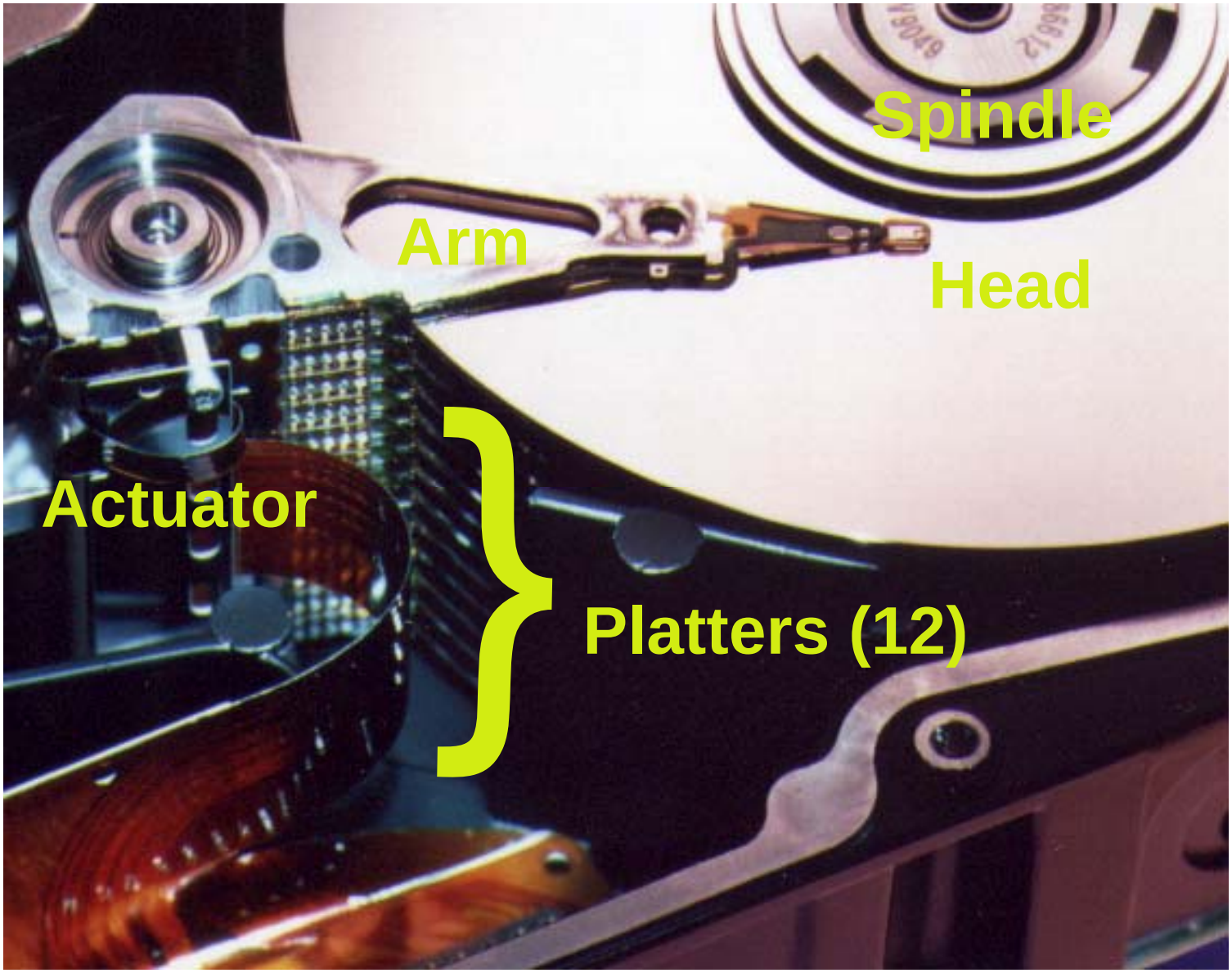
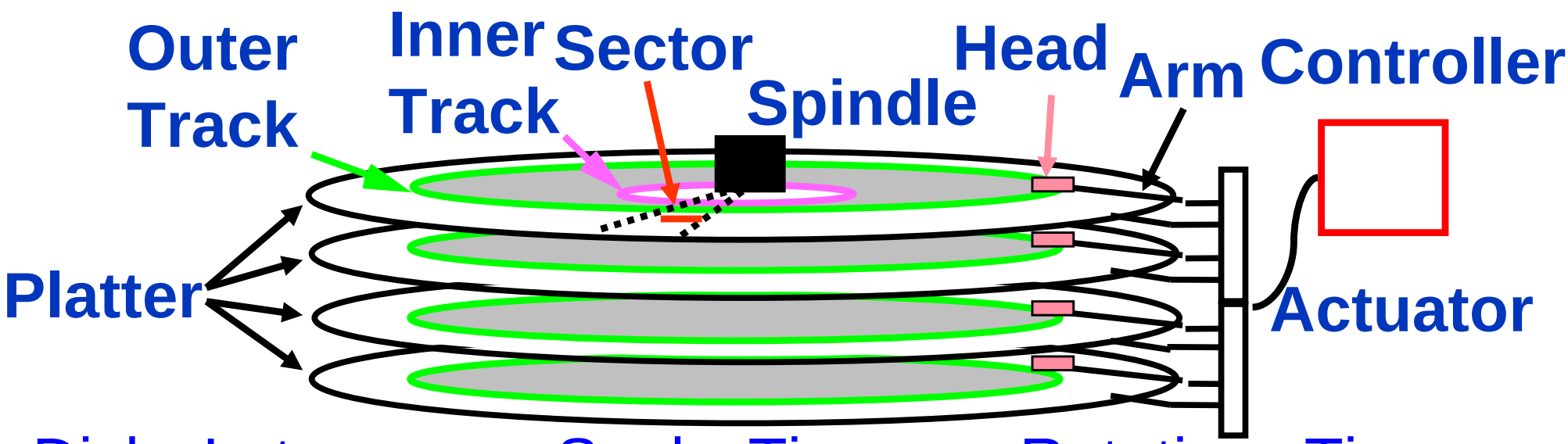


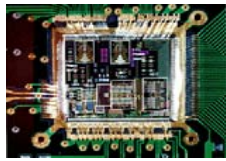
Photo of Disk Head, Arm, Actuator



Disk Device Performance

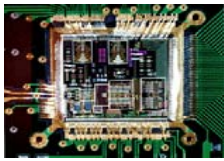


- Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead
- Seek Time? depends no. tracks move arm, seek speed of disk
- Rotation Time? depends on speed disk rotates, how far sector is from head
- Transfer Time? depends on data rate (bandwidth) of disk (bit density), size of request



Disk Device Performance

- Average distance sector from head?
- 1/2 time of a rotation
  - 10000 Revolutions Per Minute $\Rightarrow$ 166.67 Rev/sec
  - 1 revolution = 1/ 166.67 sec $\Rightarrow$ 6.00 milliseconds
  - 1/2 rotation (revolution) $\Rightarrow$ 3.00 ms
- Average no. tracks move arm?
  - Sum all possible seek distances from all possible tracks / # possible
    - Assumes average seek distance is random
  - Disk industry standard benchmark



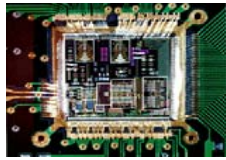
Data Rate: Inner vs. Outer Tracks

- To keep things simple, originally kept same number of sectors per track:
  - Since outer track longer, lower bits per inch
- Competition ⇒ decided to keep BPI the same for all tracks ("constant bit density"):

  ⇒ More capacity per disk

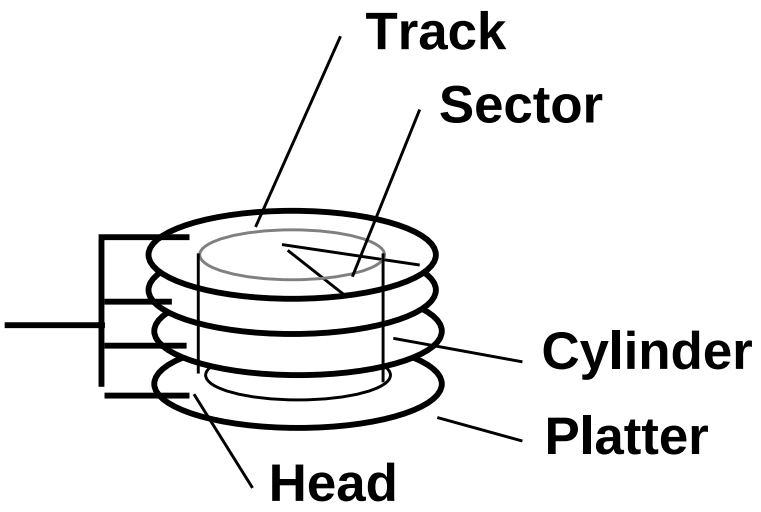
  ⇒ More of sectors per track towards edge

  ⇒Since disk spins at constant speed, outer tracks have faster data rate.
- Bandwidth outer track 1.7X inner track!
  - Inner track highest density, outer track lowest, so not really constant.
  - 2.1X length of track outer / inner, 1.7X bits outer / inner.



Devices: Magnetic Disks

- Purpose:
  - Long-term, nonvolatile storage
  - Large, inexpensive, slow level in the storage hierarchy
- Characteristics:
  - Seek Time (~8 ms avg)
    - positional latency
    - rotational latency
- Transfer rate
  - 10-40 MByte/sec
  - Blocks
- Capacity
  - Gigabytes
  - Quadruples every 2 years (aerodynamics)



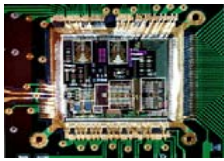
7200 RPM = 120 RPS => 8 ms per rev
ave rot. latency = 4 ms
128 sectors per track => 0.25 ms per sector
1 KB per sector => 16 MB / s

Response time
= Queue + Controller + Seek + Rot + Xfer

Service time

Disk Performance Model /Trends

- Capacity
  + 100%/year (2X / 1.0 yrs)
- Transfer rate (BW)
  + 40%/year (2X / 2.0 yrs)
- Rotation + Seek time
  – 8%/ year (1/2 in 10 yrs)
- MB/$
  > 100%/year (2X / 1.0 yrs)
  Fewer chips + areal density



Disk Performance Example

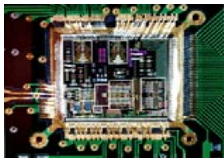
Calculate time to read 64 KB (128 sectors) for Barracuda 180 X using advertised performance; sector is on outer track.

Disk latency = average seek time + average rotational delay + transfer time + controller overhead

= 7.4ms + 0.5*1/(7200RPM) + 64KB / (64MB/s) + 0.1 ms

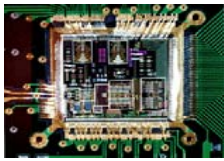
= 7.4ms + 0.5/(7200RPM/(60000ms/M)) + 64 KB / (64 KB/ms) + 0.1 ms

= 7.4 + 4.2 + 1.0 + 0.1 ms = 12.7 ms



Areal Density

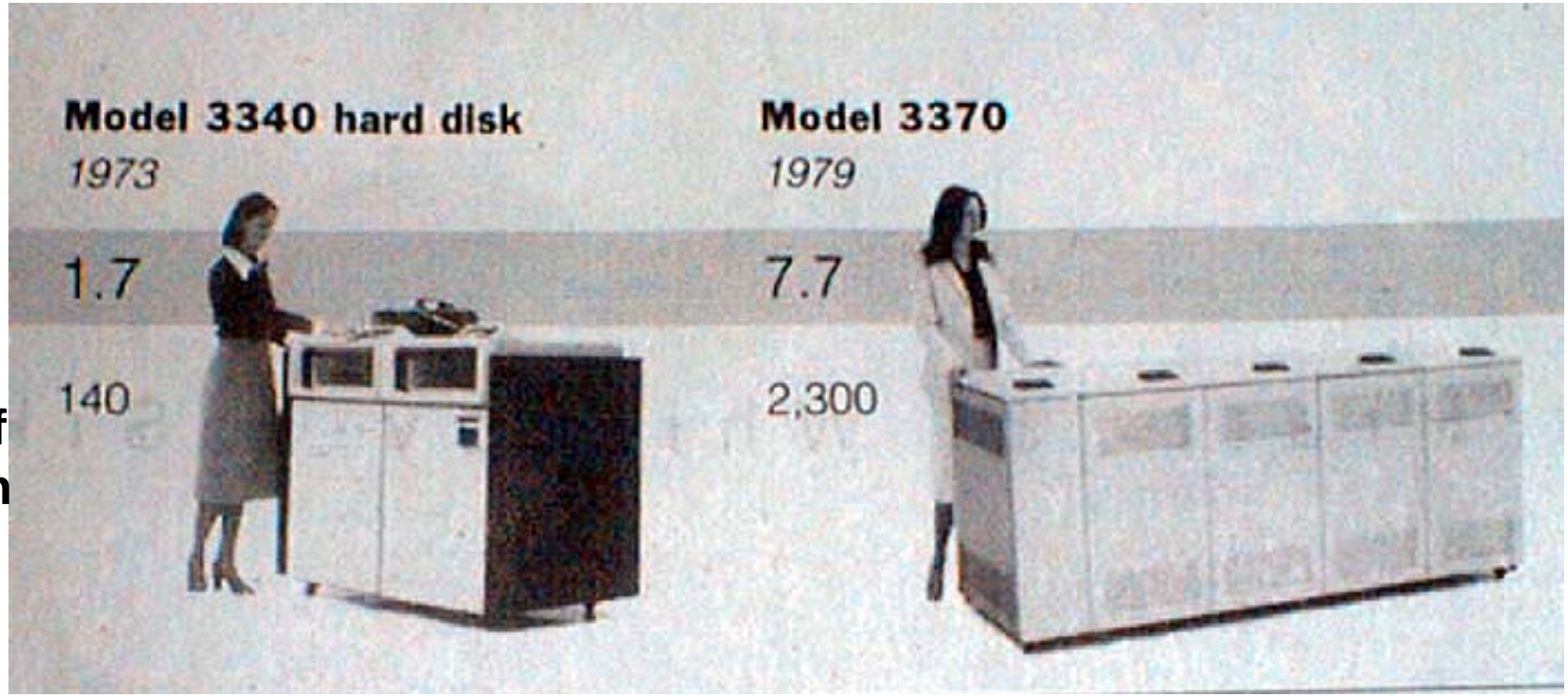
- Bits recorded along a track:
  - Metric is Bits Per Inch (BPI)
- Number of tracks per surface:
  - Metric is Tracks Per Inch (TPI)
- Disk Designs Brag about bit density per unit area:
  - Metric is Bits Per Square Inch
  - Called Areal Density
  - Areal Density = BPI x TPI



Disk History

Data density Mbit/sq. in.

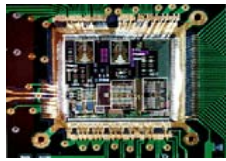
Capacity of Unit Shown Megabytes



1973:
1. 7 Mbit/sq. in
140 MBytes

1979:
7. 7 Mbit/sq. in
2,300 MBytes

source: New York Times, 2/23/98, page C3,
"Makers of disk drives crowd even more data into even smaller spaces"



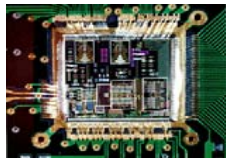
Disk History

1989:
63 Mbit/sq. in
60,000 MBytes

1997:
1450 Mbit/sq. in
2300 MBytes

1997:
3090 Mbit/sq. in
8100 MBytes

source: New York Times, 2/23/98, page C3,
"Makers of disk drives crowd even more data into even smaller spaces"

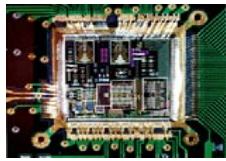


1 inch disk drive!

- 2000 IBM MicroDrive:
  - 1.7” x 1.4” x 0.2”
  - 1 GB, 3600 RPM, 5 MB/s, 15ms seek
  - Digital camera, PalmPC?
- 2006 MicroDrive?
- 9 GB, 50 MB/s!
  - Assuming it finds a niche in a successful product.
  - Assuming past trends continue.

Fallacy: Use Data Sheet "Average Seek" Time

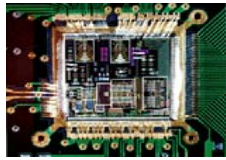
- Manufacturers needed standard for fair comparison ("benchmark")
  - Calculate all seeks from all tracks, divide by number of seeks => "average".
- Real average would be based on how data laid out on disk, where seek in real applications, then measure performance.
  - Usually, tend to seek to tracks nearby, not to random track.
- Rule of Thumb: observed average seek time is typically about 1/4 to 1/3 of quoted seek time (i.e., 3X-4X faster).
  - Barracuda 180 X avg. seek: 7.4 ms $\Rightarrow$ 2.5 ms.



Fallacy: Use Data Sheet Transfer Rate

- Manufacturers quote the speed off the data rate off the surface of the disk
- Sectors contain an error detection and correction field (can be 20% of sector size) plus sector number as well as data
- There are gaps between sectors on track
- Rule of Thumb: disks deliver about 3/4 of internal media rate (1.3X slower) for data
- For example, Barracuda 180X quotes 64 to 35 MB/sec internal media rate

$\Rightarrow$ 47 to 26 MB/sec external data rate (74%)



Disk Performance Example

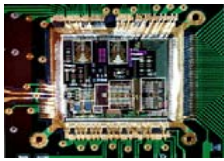
- Calculate time to read 64 KB for UltraStar 72 again, this time using 1/3 quoted seek time, 3/4 of internal outer track bandwidth; (12.7 ms before)

Disk latency = average seek time + average rotational delay + transfer time + controller overhead

= (0.33*7.4ms) + 0.5*1/(7200RPM) + 64KB / (0.75 * 65MB/s) + 0.1ms

= 2.5ms + 0.5/(7200RPM/(60000ms/M)) + 64KB / (47KB/ms) + 0.1ms

= 2.5 + 4.2 + 1.4 + 0.1 ms = 8.2 ms (64% of 12.7)

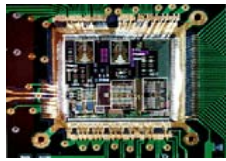


Future Disk Size and Performance

- Continued advance in capacity (60%/yr) and bandwidth (40%/yr)
- Slow improvement in seek, rotation (8%/yr)
- Time to read whole disk

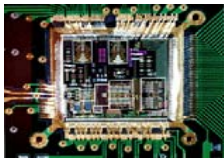
Year	Sequentially	Randomly (1 sector/seek)
1990	4 minutes	6 hours
2000	12 minutes	1 week(!)

- 3.5” form factor make sense in 5 yrs?
  - What is capacity, bandwidth, seek time, RPM?
  - Assume today 80 GB, 30 MB/sec, 6 ms, 10000 RPM



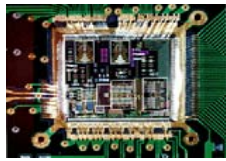
What about FLASH

- Compact Flash Cards
  - Intel Strata Flash
    - 16 Mb in 1 square cm. (.6 mm thick)
  - 100,000 write/erase cycles.
  - Standby current = 100uA, write = 45mA
  - Compact Flash 256MB~=$120 512MB~=$542
  - Transfer @ 3.5MB/s
- IBM Microdrive 1G~370
  - Standby current = 20mA, write = 250mA
  - Efficiency advertised in wats/MB
- VS. Disks
  - Nearly instant standby wake-up time
  - Random access to data stored
  - Tolerant to shock and vibration (1000G of operating shock)



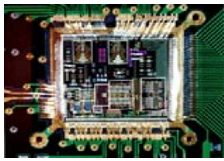
Tape vs. Disk

- Longitudinal tape uses same technology as hard disk; tracks its density improvements.
- Disk head flies above surface, tape head lies on surface.
- Disk fixed, tape removable.
- Inherent cost-performance based on geometries:
  fixed rotating platters with gaps (random access, limited area, 1 media / reader) vs.
  removable long strips wound on spool (sequential access, "unlimited" length, multiple / reader)
- Helical Scan (VCR, Camcoder, DAT)
  Spins head at angle to tape to improve density



Library vs. Storage

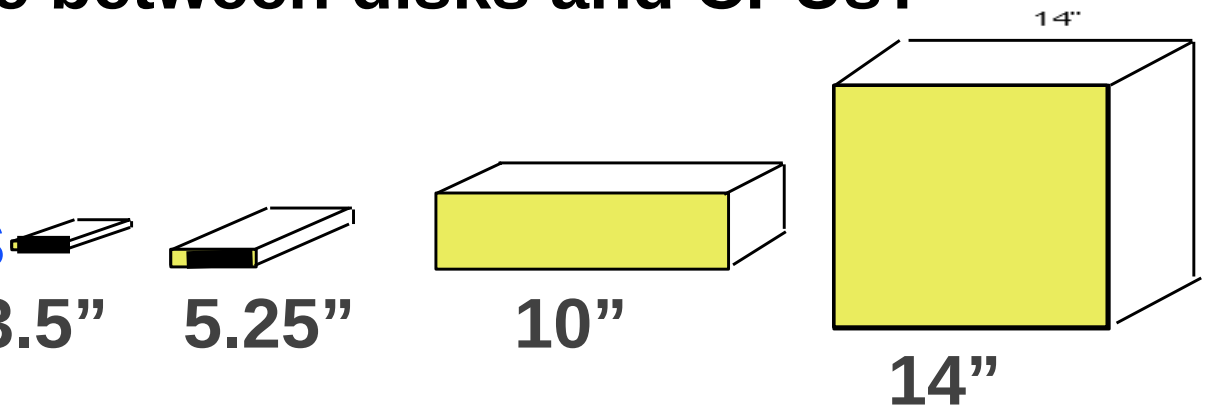
- Getting books today as quaint as the way I learned to program.
  - punch cards, batch processing
  - wander thru shelves, anticipatory purchasing
- Cost $1 per book to check out.
- $30 for a catalogue entry.
- 30% of all books never checked out.
- Write only journals?
- Digital library can transform campuses.



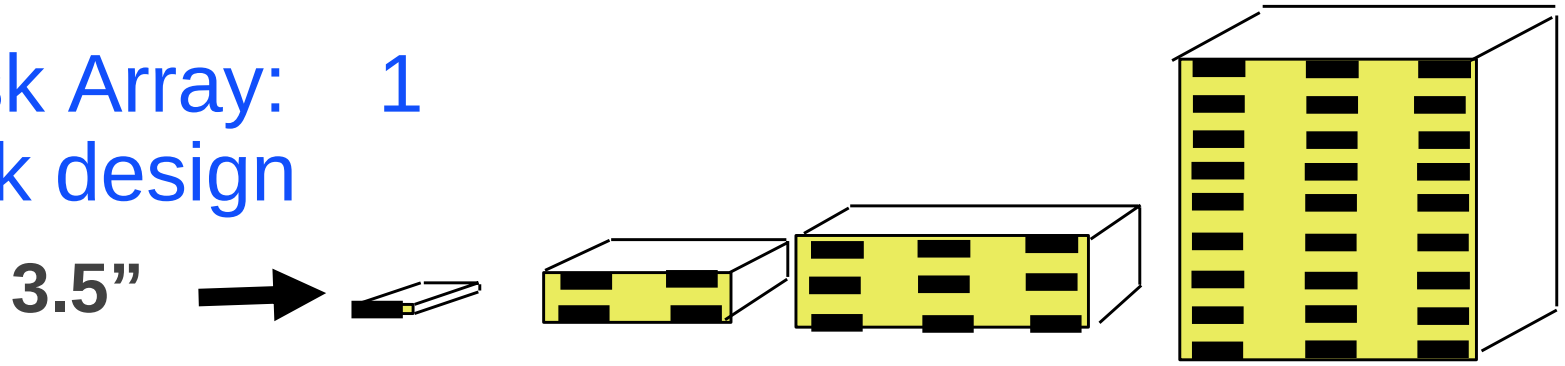
Use Arrays of Small Disks?

- **Katz and Patterson asked in 1987:**
  - **Can smaller disks be used to close gap in performance between disks and CPUs?**

Conventional:
4 disk designs



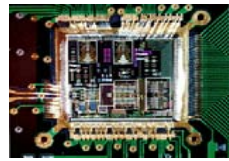
Disk Array: 1
disk design



Replace Small Number of Large Disks with Large Number of Small Disks!

	IBM 3390K	IBM 3.5" 0061	x70	
Capacity	20 GBytes	320 MBytes	23 GBytes	
Volume	97 cu. ft.	0.1 cu. ft.	11 cu. ft.	9X
Power	3 KW	11 W	1 KW	3X
Data Rate	15 MB/s	1.5 MB/s	120 MB/s	8X
I/O Rate	600 I/Os/s	55 I/Os/s	3900 IOs/s	6X
MTTF	250 KHrs	50 KHrs	??? Hrs	
Cost	$250K	$2K	$150K	

Disk Arrays have potential for large data and I/O rates, high MB per cu. ft., high MB per KW, but what about reliability?



Array Reliability

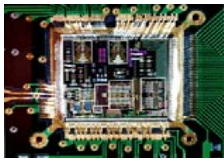
- **Reliability of N disks = Reliability of 1 Disk ÷ N**

  50,000 Hours ÷ 70 disks = 700 hours

  Disk system MTTF: Drops from 6 years to 1 month!

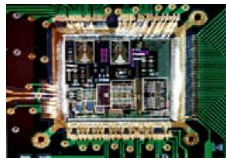
- **Arrays (without redundancy) too unreliable to be useful!**

Hot spares support reconstruction in parallel with access: very high media availability can be achieved

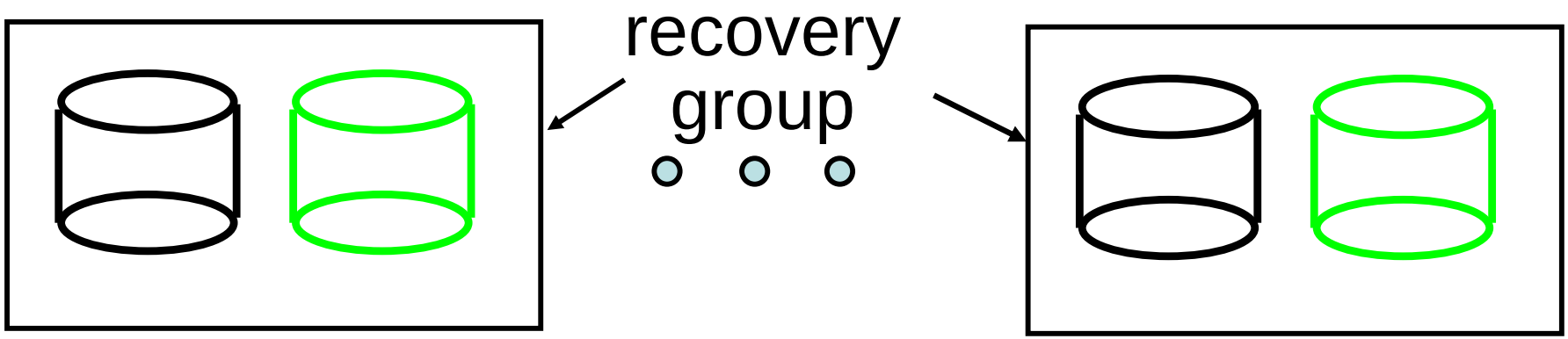


Redundant Arrays of (Inexpensive) Disks

- Files are "striped" across multiple disks.
- Redundancy yields high data availability.
  - Availability: service still provided to user, even if some components failed.
- Disks will still fail.
- Contents reconstructed from data redundantly stored in the array.
  - ⇒ Capacity penalty to store redundant info
  - ⇒ Bandwidth penalty to update redundant info

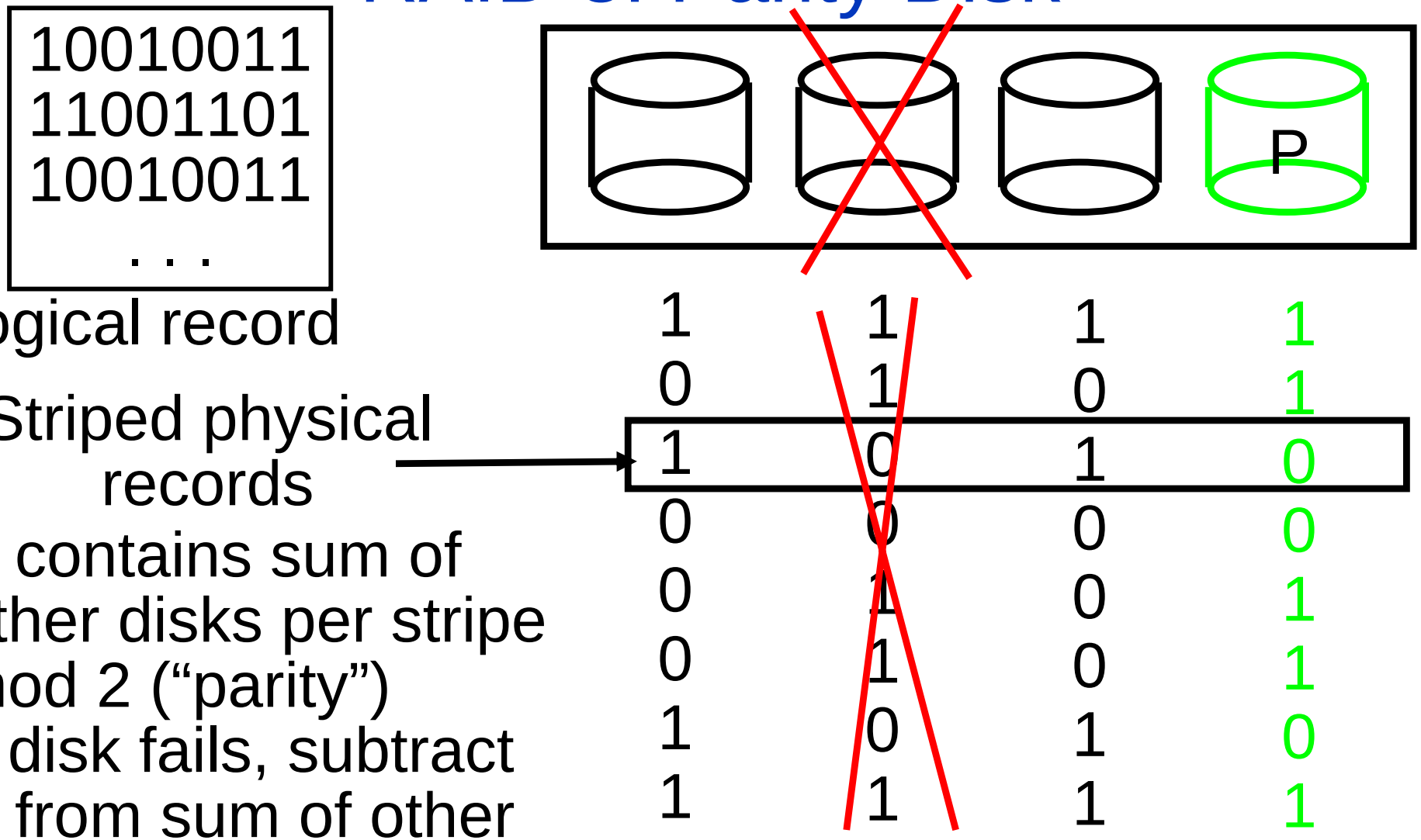


Redundant Arrays of Inexpensive Disks
RAID 1: Disk Mirroring/Shadowing



- Each disk is fully duplicated onto its “mirror”
  - Very high availability can be achieved
- Bandwidth sacrifice on write:
  - Logical write = two physical writes
  - Reads may be optimized
- Most expensive solution: 100% capacity overhead

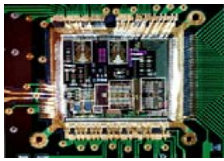
Redundant Array of Inexpensive Disks
RAID 3: Parity Disk



P contains sum of other disks per stripe mod 2 ("parity")
If disk fails, subtract P from sum of other disks to find missing information.

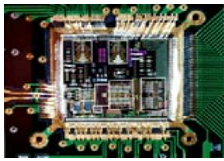
RAID 3

- Sum computed across recovery group to protect against hard disk failures, stored in P disk
- Logically, a single high capacity, high transfer rate disk: good for large transfers
- Wider arrays reduce capacity costs, but decreases availability
- 33% capacity cost for parity in this configuration

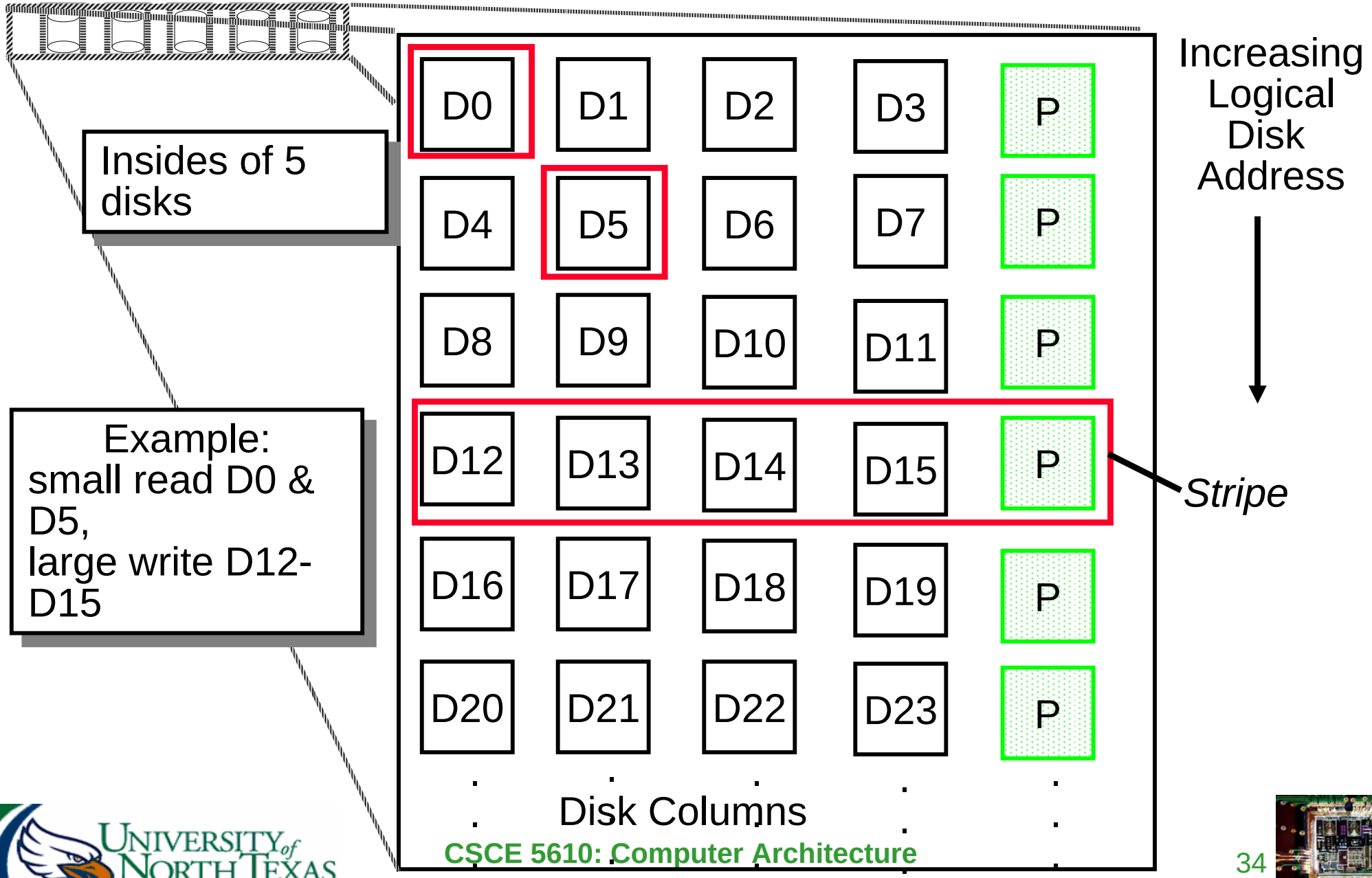


Inspiration for RAID 4

- RAID 3 relies on parity disk to discover errors on Read
- But every sector has an error detection field
- Rely on error detection field to catch errors on read, not on the parity disk
- Allows independent reads to different disks simultaneously

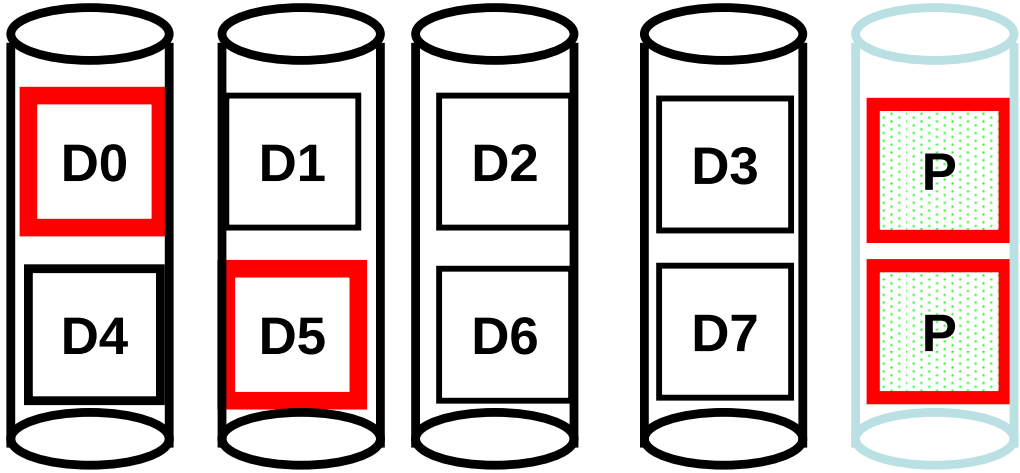


Redundant Arrays of Inexpensive Disks RAID 4: High I/O Rate Parity

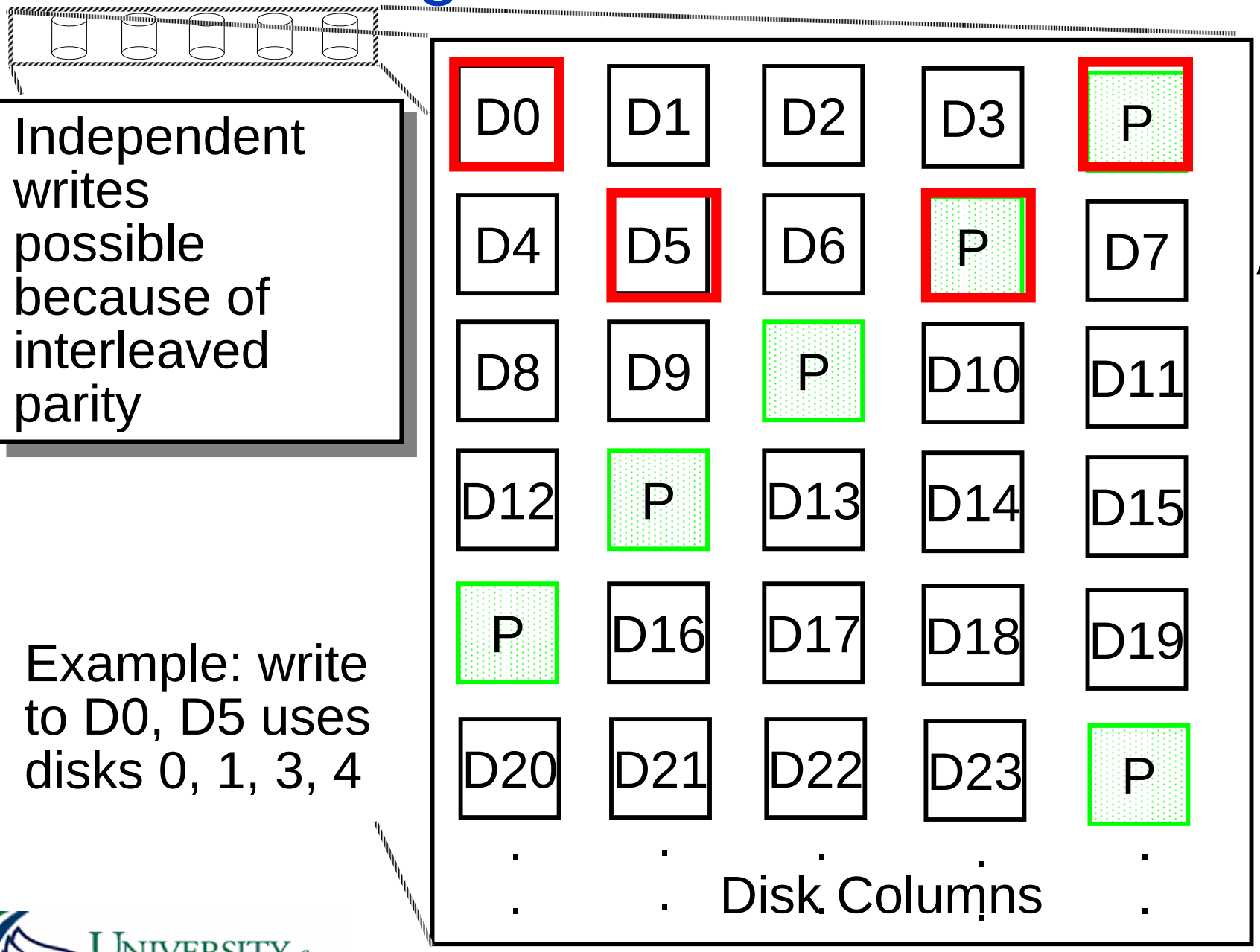


Inspiration for RAID 5

- RAID 4 works well for small reads.
- Small writes (write to one disk):
  - Option 1: read other data disks, create new sum and write to Parity Disk.
  - Option 2: since P has old sum, compare old data to new data, add the difference to P.
- Small writes are limited by Parity Disk: Write to D0, D5 both also write to P disk.



Redundant Arrays of Inexpensive Disks RAID 5: High I/O Rate Interleaved Parity



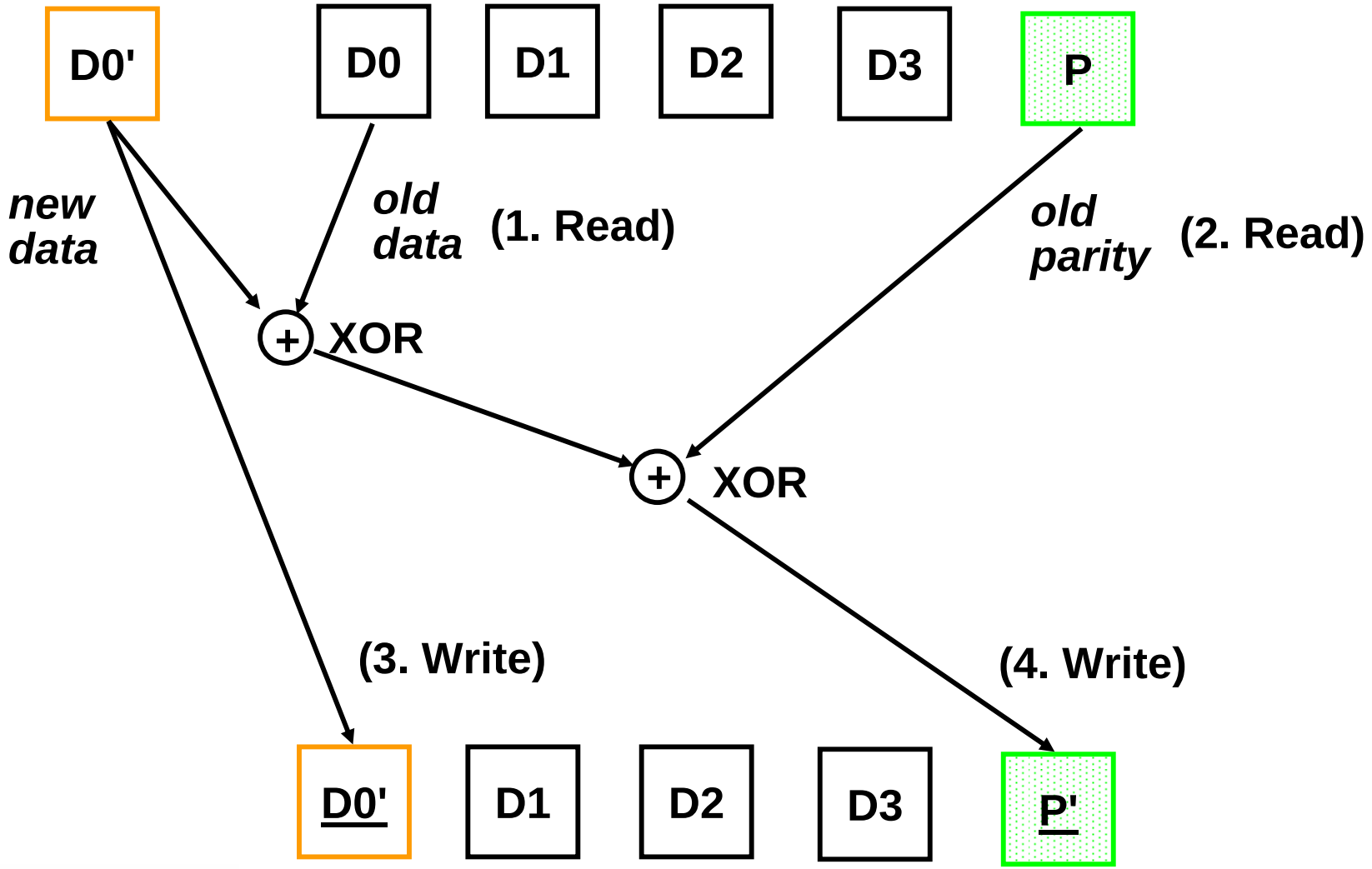
Independent writes possible because of interleaved parity

Example: write to D0, D5 uses disks 0, 1, 3, 4

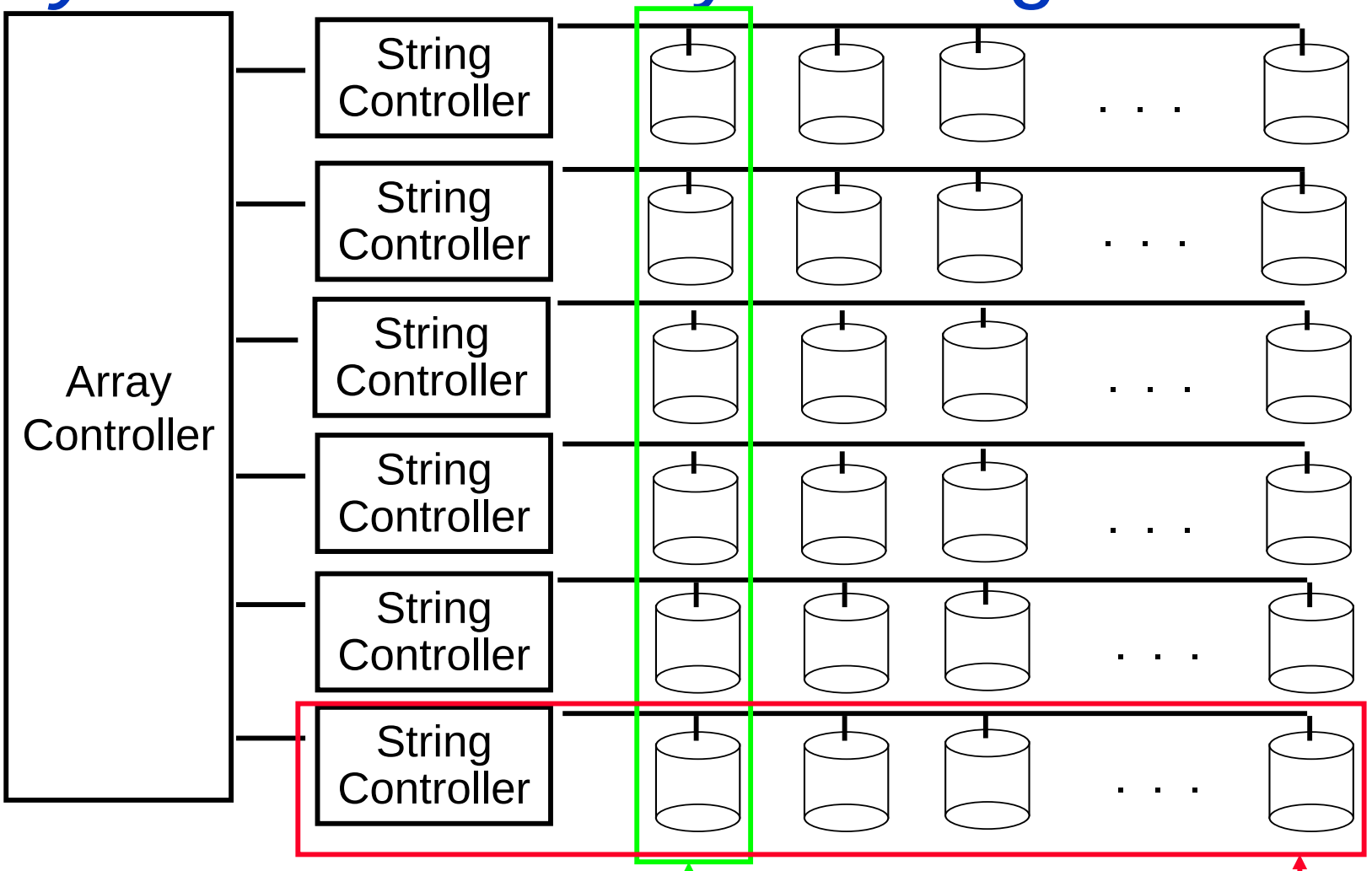
Problems of Disk Arrays: Small Writes

RAID-5: Small Write Algorithm

1 Logical Write = 2 Physical Reads + 2 Physical Writes



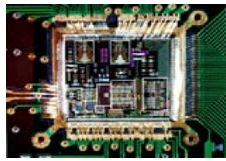
System Availability: Orthogonal RAIDs



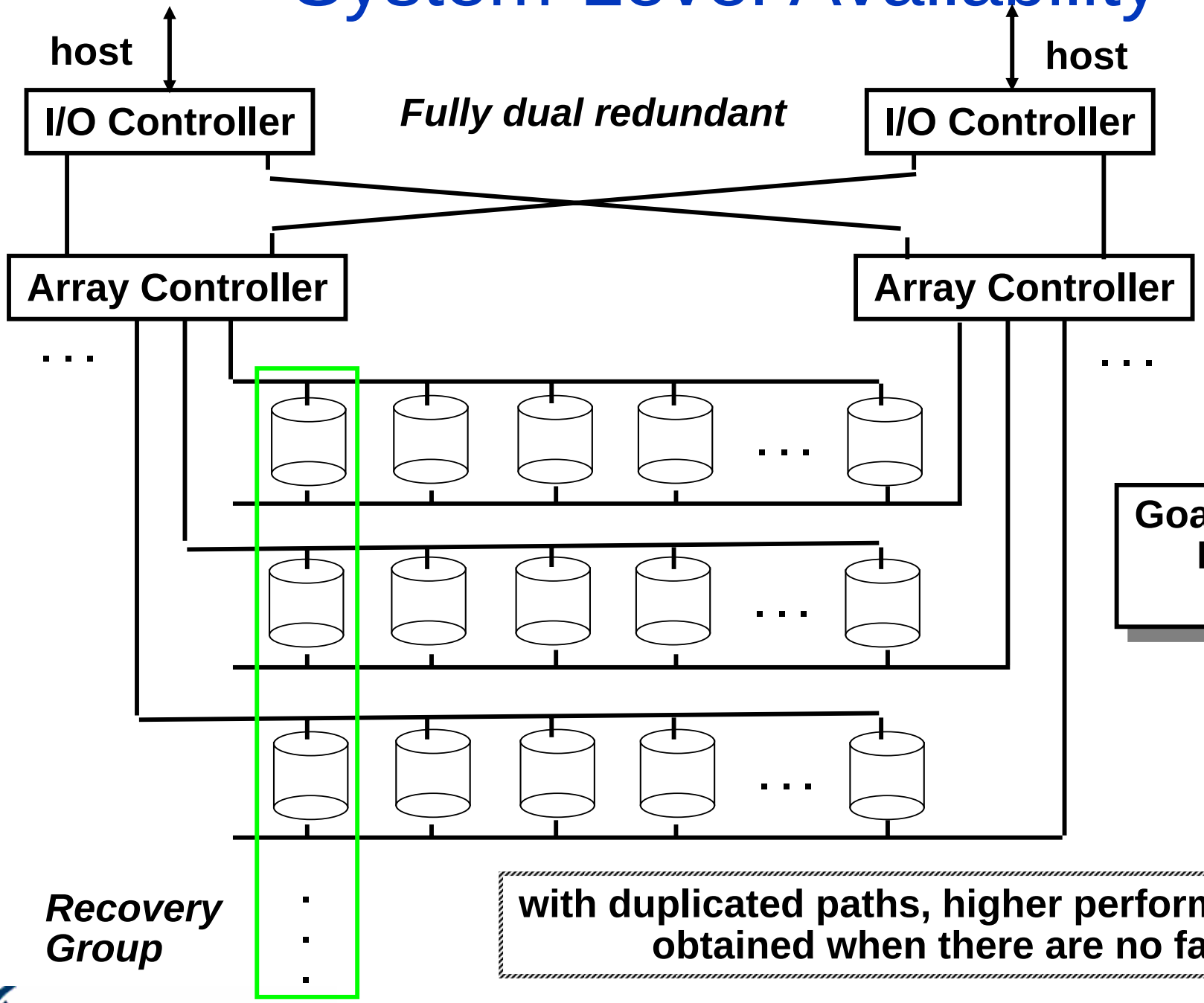
Data Recovery Group: unit of data redundancy

Redundant Support Components: fans, power supplies, controller, cables

End to End Data Integrity: internal parity protected data paths



System-Level Availability



Berkeley History: RAID-I

- RAID-I (1989)
  - Consisted of a Sun 4/280 workstation with 128 MB of DRAM, four dual-string SCSI controllers, 28 5.25-inch SCSI disks and specialized disk striping software
- Today RAID is $19 billion dollar industry, 80% nonPC disks sold in RAIDs

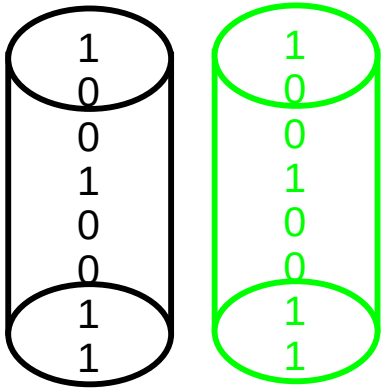
Summary: RAID Techniques: Goal was performance, popularity due to reliability of storage

- *Disk Mirroring, Shadowing (RAID 1)*

  Each disk is fully duplicated onto its "shadow"

  Logical write = two physical writes

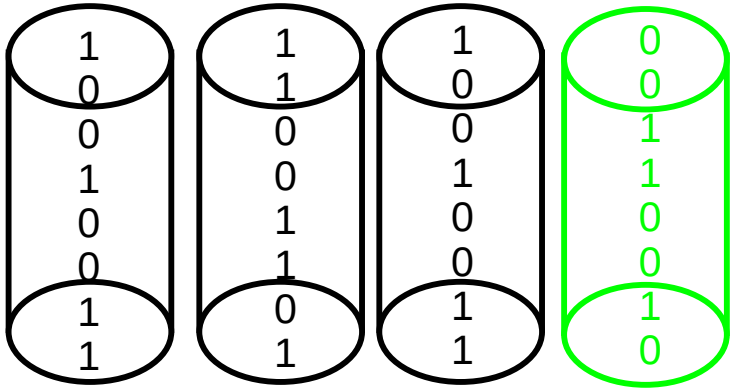
  100% capacity overhead



- *Parity Data Bandwidth Array (RAID 3)*

  Parity computed horizontally

  Logically a single high data bw disk

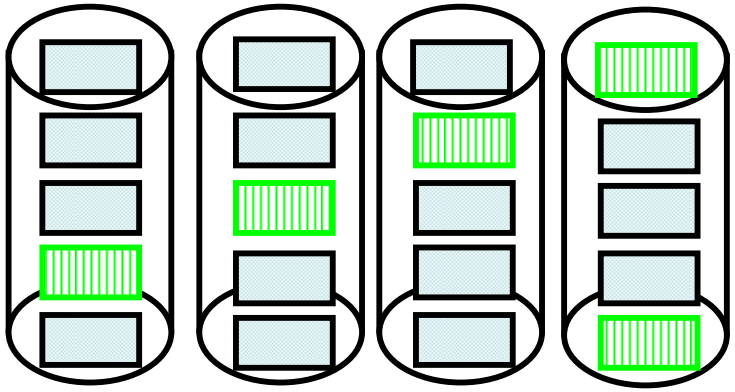


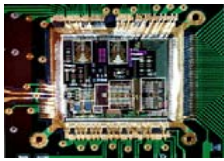
- *High I/O Rate Parity Array (RAID 5)*

  Interleaved parity blocks

  Independent reads and writes

  Logical write = 2 reads + 2 writes





Summary Storage

- Disks:
  - Extraodinary advance in capacity/drive, $/GB
  - Currently 17 Gbit/sq. in. ; can continue past 100 Gbit/sq. in.?
  - Bandwidth, seek time not keeping up: 3.5 inch form factor makes sense? 2.5 inch form factor in near future? 1.0 inch form factor in long term?
- Tapes
  - No investment, must be backwards compatible
  - Are they already dead?
  - What is a tapeless backup system?