EECS 252 Graduate Computer Architecture

Lec 18 – Storage

David Patterson
Electrical Engineering and Computer Sciences
University of California, Berkeley

http://www.eecs.berkeley.edu/~pattrsn
http://vlsi.cs.berkeley.edu/cs252-s06

Review

- Virtual Machine Revival
  - Overcome security flaws of modern OSes
  - Processor performance no longer highest priority
  - Manage Software, Manage Hardware
- "… VMMs give OS developers another opportunity to develop functionality no longer practical in today's complex and ossified operating systems, where innovation moves at geologic pace."
  [Rosenblum and Garfinkel, 2005]
- Virtualization challenges for processor, virtual memory, I/O
  - Paravirtualization, ISA upgrades to cope with those difficulties
- Xen as example VMM using paravirtualization
  - 2005 performance on non-I/O bound, I/O intensive apps: 80% of native Linux without driver VM, 34% with driver VM
- Opteron memory hierarchy still critical to performance

Case for Storage

- Shift in focus from computation to communication and storage of information
  - E.g., Cray Research/Thinking Machines vs. Google/Yahoo
  - "The Computing Revolution" (1960s to 1980s)
    ⇨ "The Information Age" (1990 to today)
- Storage emphasizes reliability and scalability as well as cost-performance
- What is "Software king" that determines which HW acually features used?
  - Operating System for storage
  - Compiler for processor
- Also has own performance theory—queuing theory—balances throughput vs. response time

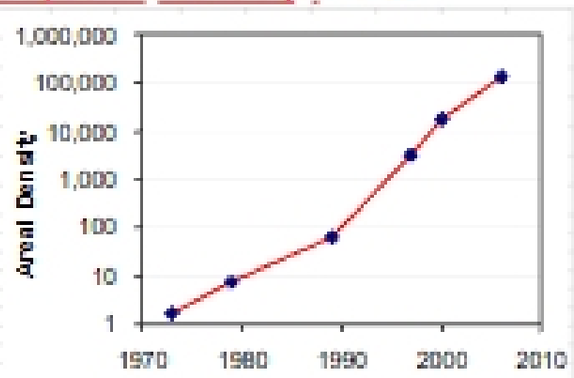
Outline

- Magnetic Disks
- RAID
- Administrivia
- Advanced Dependability/Reliability/Availability
- I/O Benchmarks, Performance and Dependability
- Intro to Queueing Theory (if we have time)
- Conclusion

Disk Figure of Merit: 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: Areal Density = BPI x TPI

Year	Areal Density
1973	2
1979	8
1989	63
1997	3,090
2000	17,100
2006	130,000



Historical Perspective

- 1956 IBM Ramac — early 1970s Winchester
  - Developed for mainframe computers, proprietary interfaces
  - Steady shrink in form factor: 27 in. to 14 in.
- Form factor and capacity drives market more than performance
- 1970s developments
  - 5.25 inch floppy disk formfactor (microcode into mainframe)
  - Emergence of industry standard disk interfaces
- Early 1980s: PCs and first generation workstations
- Mid 1980s: Client/server computing
  - Centralized storage on file server
    - accelerates disk downsizing: 8 inch to 5.25
  - Mass market disk drives become a reality
    - industry standards: SCSI, IPI, IDE
    - 5.25 inch to 3.5 inch drives for PCs, end of proprietary interfaces
- 1900s: Laptops => 2.5 inch drives
- 2000s: What new devices leading to new drives?

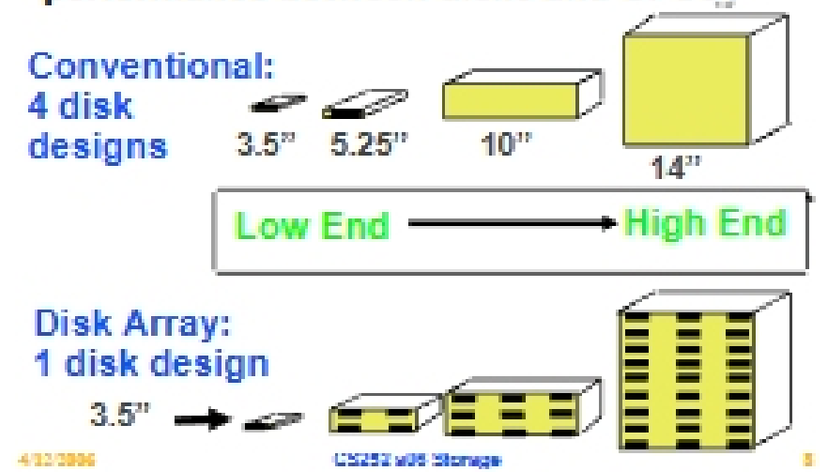
Future Disk Size and Performance

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

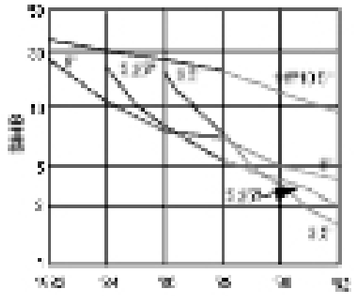
Year	Sequentially	Randomly (1 sector/seek)
1990	4 minutes	6 hours
2000	12 minutes	1 week(!)
2006	56 minutes	3 weeks (SCSI)
2006	171 minutes	7 weeks (SATA)

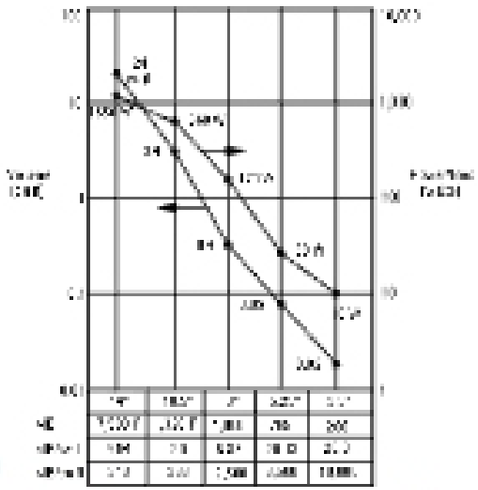
Use Arrays of Small Disks?

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



Advantages of Small Formfactor Disk Drives





Low cost/MB
High MB/volume
High MB/watt
Low cost/Actuator

Cost and Environmental Efficiencies

Replace Small Number of Large Disks with Large Number of Small Disks! (1988 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