Terabytes >> Teraflops

(Or why work on processors when I/O is where the action is?)

David A. Patterson

Computer Science Div., EECS Department
University of California, Berkeley

Motivation: Who cares about I/O?

- **CPU Performance:**
  50% to 100% per year

- **Multiprocessor supercomputers:**
  150% per year

- **I/O system performance limited by**
  ***mechanical* delays**

<5% per year
(IO per sec or MB per sec)

Motivation: Who cares about I/O?

- 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

Performance growth in perspective:

	HP 9000/735 Workstation 1992	Cray 1 Vector Supercomputer 1976
MIPS	>130 MIPS	< 10 MIPS
Linpack	41 MFLOPS	27 MFLOPS
Cost	$40,000	$4,000,000 *
Clock	99 MHz	80 MHz
Cache	128 KB	0.25 KB
Memory	400 MB	32 MB

* ($8,000,000 in 1992 dollars)

What have Magnetic Disk Designers been doing?

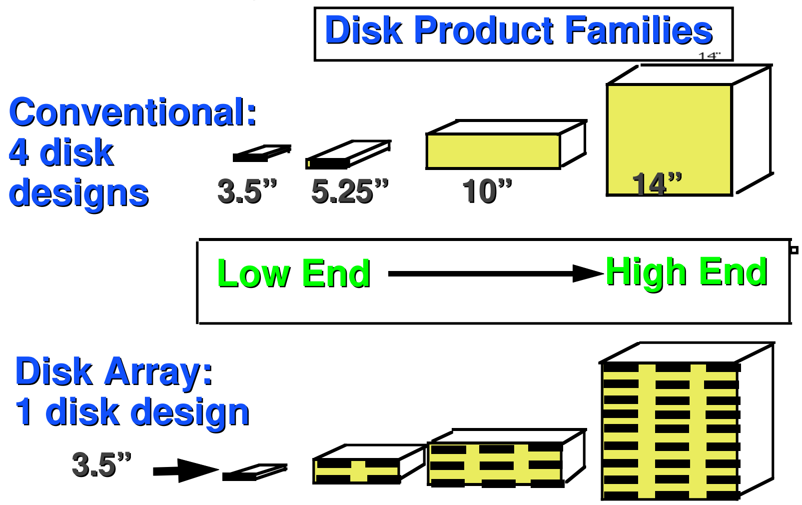
- Capacity: Improving 25% per year
- $/MB: Improving 25% per year
- Evolving to smaller physical size

14" ➡ 10" ➡ 8" ➡ 5.25" ➡
3.5" ➡ 2.5" ➡ 1.8" ➡ 1"?

Smaller and cheaper vs. Larger and faster

Can smaller disks be used
to close gap in performance
between disks and CPUs?

Manufacturing advantages of disk arrays



Arrays of Disks to Close the Performance Gap

Replace Small Number
of Large Disks

With Large Number
of Small Disks!

	IBM 10.4" 3390	IBM 3.5" 0061	x 70
Data Capacity	23 GBytes	320 MBytes	23 GBytes
Volume	97 cu. ft.	0.1 cu. ft.	11 cu. ft.
Power	3 KW	11 W	1 KW
Data Rate	15 MB/s	1.5 MB/s	120 MB/s
I/O Rate	600 I/Os/s	55 I/Os/s	3900 IOs/s
Cost	≈$200K	≈$2K	≈$150K
MTTDL *(hrs)*	250K	50K	???

Disk Arrays have potential for:

- large data and I/O rates
- high MB per cu. ft.
- high MB per KW

But is there a catch?

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 Too Unreliable to be useful!

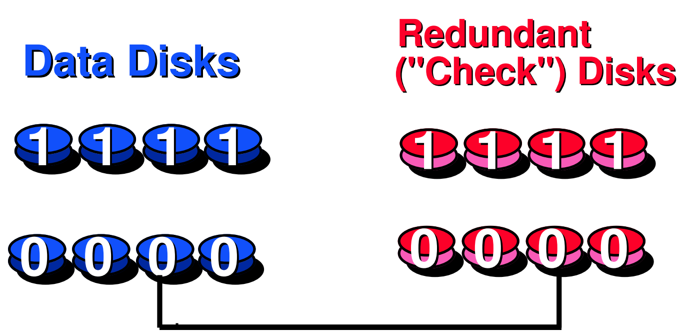
Improving Reliability with Redundancy

- Add redundant drives to handle failures
  Redundant
  Array of
  Inexpensive
  Disks
  RAID
- Redundancy offers 2 advantages:
  1) Data not lost:
     Reconstruct data onto new disk
  2) Continuous operation
     in the presence of failure

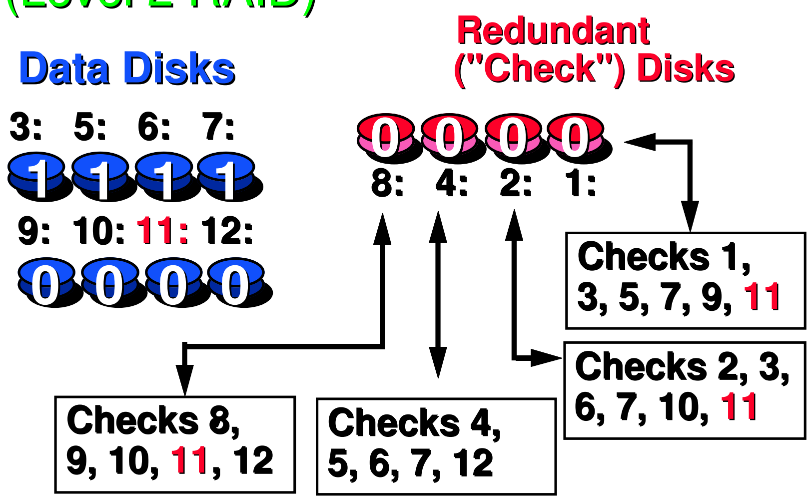
Several RAID organizations

Mirroring/Shadowing	*(Level 1 RAID)*
ECC	*(Level 2 RAID)*
Parity	*(Level 3 RAID)*
Rotated Parity	*(Level 5 RAID)*

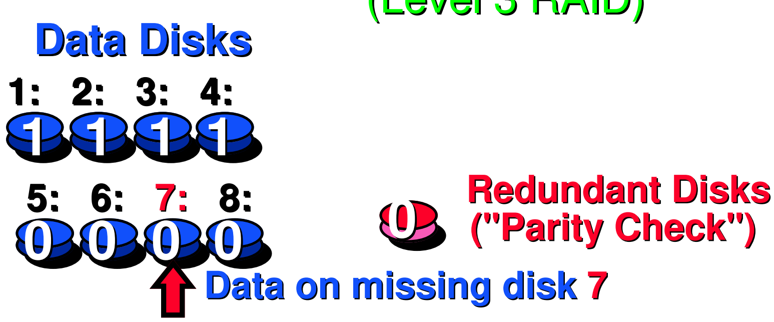
Redundancy via Mirroring (Level 1 RAID)



Redundancy via ECC
(Level 2 RAID)

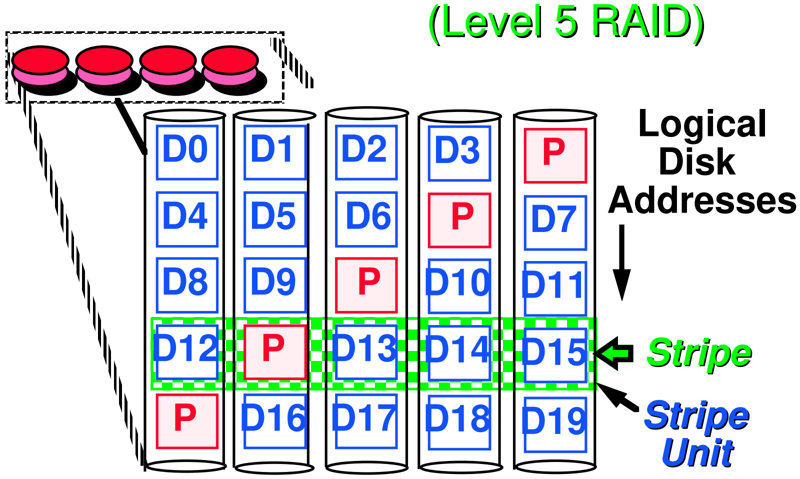


Redundancy via Parity
(Level 3 RAID)

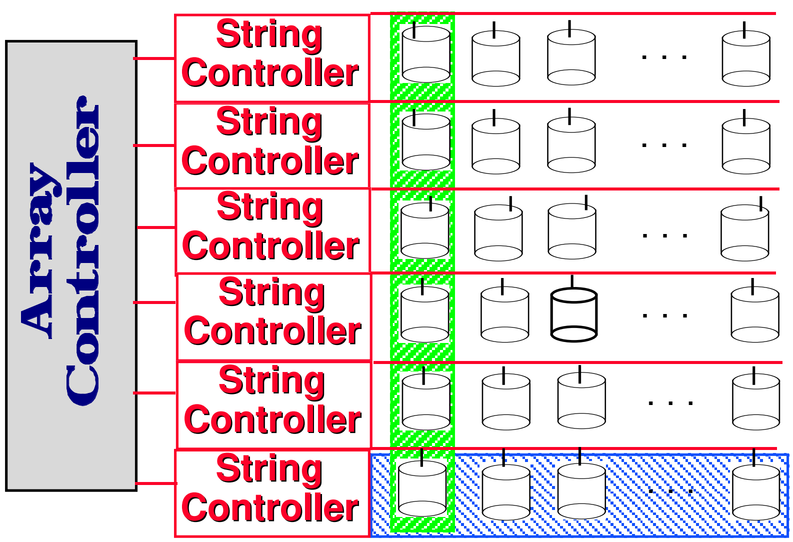


Reed/Solomon code per sector
identifies failed disk

Redundancy via Rotated Parity
(Level 5 RAID)

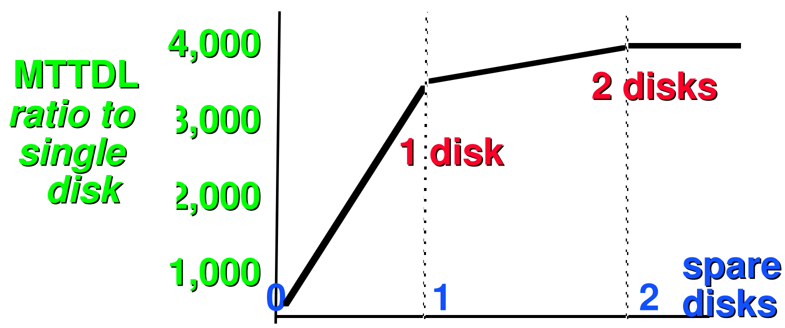


System Availability: Orthogonal RAIDs



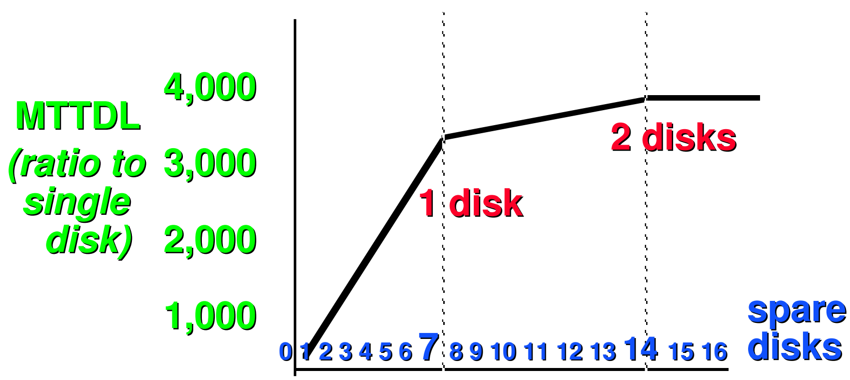
Value of Spares: Independent Failures

- 1 spare disk => 40X MTTDL
- 2 spare disks = ∞ spare disks



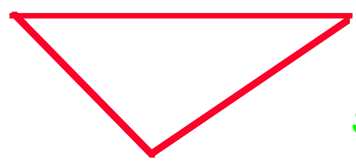
Value of Spares: Dependent Failures

- 1 spare string => 50X MTTDL
- 2 spare string = ∞ spare strings

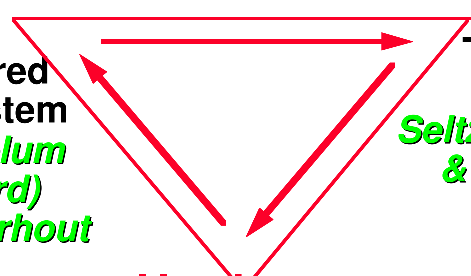


Applications Context for RAID

- *We seek to understand the system demands of I/O intensive applications*
- *Common hardware solution, customized by software*



Multidisciplinary Research Benefits



First Prototype RAID-I: Proof of Concept

Built with off-the-shelf components

Prototype software implementation of RAID algorithms

Status:

Operational late Fall 1989

8 - 10 MB/s sustained transfers

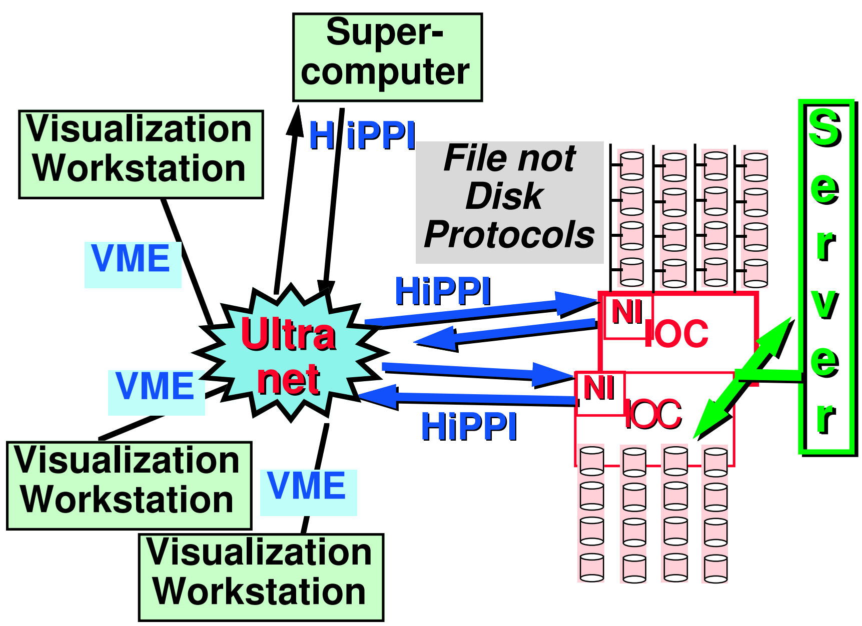
Components:

- ***SUN-4/280*** **w/ 128 MBytes as RAID I/O Processor**
- **4 Interphase Dual SCSI String HBAs (8 strings)**
- **28 *Imprimis* 5.25" Wren-IV Synchronous SCSI Disk Drives**
- **10 GBytes of storage in standard 19" rack**

Berkeley XPRS Project: Supercomputer Application

Concept

- Diskless Supercomputer
- System organization optimized for high performance file service
- Integration of network and disk subsystems
- Called RAID - II

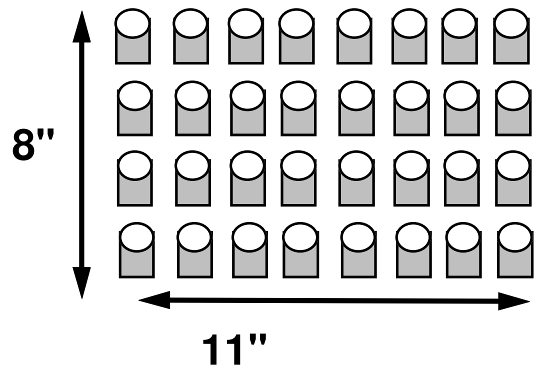


Arrays of Disks to
Close the Performance Gap

Replace a Small number of Large Disks with a Large number of Small Disks!

	IBM 10.4" 3390	IBM 3.5" 0061	*x 70+7+7*
Data Capacity	23 GBytes	320 MBytes	*23 GBytes*
Volume	97 cu. ft.	0.1 cu. ft.	*11 cu. ft.*
Power	3 KW	11 W	*1 KW*
Data Rate	15 MB/s	1.5 MB/s	*120 MB/s*
I/O Rate	600 I/Os/s	55 I/Os/s	*3900 IOs/s*
Cost	≈$200K	≈ $2K	≈*$200K*
MTTDL (hrs)	250K	50K	*6600K*

1995 Disk Array
8 GB, 120 MB/s
high reliability



Technology Transfer

- 200 companies now selling or investigating RAID systems or components
- e.g., NCR, Thinking Machines, Data General, Tandem, Compaq, DEC, STC and several startups selling RAIDs
- e.g., IBM, Sun developing RAID systems
- IBM story

3 Technologies Driving Massive Storage

1) High capacity tape
(e.g., helical scan)

High capacity tape plays the role of disk in RAID

2) Tape Robots
(e.g., storage technologies silo)

3) Data compression
(e.g., LZW,JPEG)

What's all this about tape?

Tape is used for:

- Backup Storage for Hard Disk Data
  Written once, infrequently read
- Software Distribution & Data Interchange
  Written once, read once
- File Retrieval
  Written/Rewritten, files occasionally read
  Near Line or Robo-Line Archive

Tape vs. Disk

Longitudinal tape:
Uses same technology as hard disk

Will track density improvements

Tape vs. Disk

Inherent cost-performance based on geometries:

Disk: fixed rotating platters with gaps
Random access
Limited area
One media per reader

vs.

Tape: removable strips wound on spool
Sequential access
"Unlimited" length,
Multiple media per reader

New technology trend:

Helical Scan
(VCR, Camcoder, DAT)

Spins head at angle to tape to improve density

	Tape					Disk		
	Conventional		Helical Scan			Magnet		M-Op
	.25" Cart	.5" IBM 3490	8mm Video	4mm DAT	1/2" D-3	5.25" Disk	10.5" IBM 3390	5.25" SONY MO
Cap MB	150	800	4600	1300	20k	1200	3800	640
BPI	12k	23k	43k	61k		33k	28k	24k
TPI	104	38	1638	1870		1880	2235	19k
BPI*TPI (million)	1.2	0.9	71	114		63	62	454
Data Xfer Kbytes/s)	92	3000	492	183		3000	4250	88
Access Time	Min.	Sec.	45s	20s	15s	18m	20m	100m

Optical Disk vs. Tape

	Optical Disk	Helical Scan Tape
Type	5.25"	8mm
Capacity	0.75 GB	5 GB
Media Cost	$90 - $175	$8
Drive Cost	$3,000	$3,000
Access	Write Once	Read/Write
Robot Time	10 - 20 s	10 - 20 s

Media cost ratio:
optical disk vs. helical tape
=75 : 1 to 150 : 1

Current Drawbacks to Tape

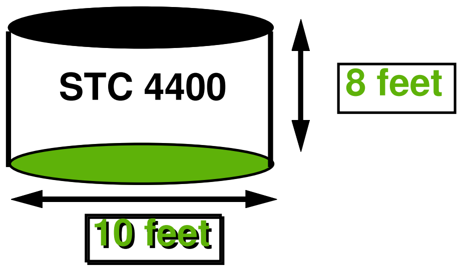
- Tape wear out:
  Helical 100s of passes
  Longitudinal 1000s of passes
- Head wear out: 2000 hours for helical

Economic / reliability model
must consider both factors

- Long rewind, eject, load, spin-up times;

Not inherent, just no need in marketplace
(so far)

Automated Cartridge System



6000 x 0.8 GB 3490 tapes = 5 TBytes in 1992
$500,000 O.E.M. Price

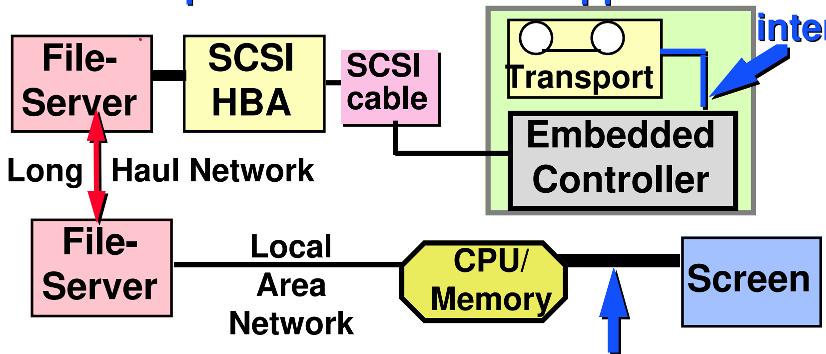
6000 x 20 GB D3 tapes = 120 TBytes in 1994
1 Petabyte (1024 TBytes) in 2000

Data Compression Issues

- 2-3:1 Text Compression (lossless)
- 20:1 Image Compression (lossy)
- Where will compression be used?
- Where should it be used?

Data Compression Location

Peripheral Manufacturer Approach: Standard interface



Systems approach: Just in time decompression

Multiplies storage capacity

Mulitplies network and memory bandwidth

Hints from application pick compression algorithm

Application: Global Change Support

Sequoia 2000

Digital Equipment Corporation

Request for Proposals
"Project Athena of the 1990s"

Problem: Understanding Ocean-Land-Atmospheric Interactions

- Remote Sensing / Satellite Imaging
- 100 Year Climate Simulations
- Correlation with Predicted and Observed
- Sierra Snow Melt, El Nino Weather Patterns

The Team:

Computer Scientists:
UCB, UCSD, SDSC

Earth System Scientists:
UCSB, UCLA, Scripps, NCAR

Real World:
Dept. of Water Resources
Air Resources Board
USGS

GC: They are drowning in data

- **Analyzers** need terabytes of satellite data
- **A modeler** produces about
  0.1 terabyte per year of output data
  which he wants to browse
- **Challenge:**
  Current file systems and DBMS
  do not understand tertiary memory
  very well

Applications of Technology
Robo-Line Library

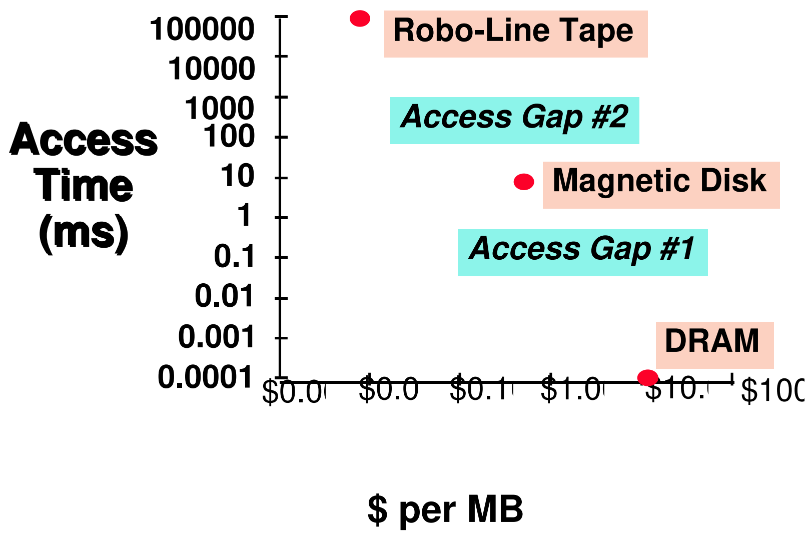
	372,910	Books in Bancroft
x	400	Pages per book
x	4000	Bytes per page
=	0.54TB	Bancroft

Full text Bancroft Near Line = 0.5 TB;
Pages images ≈ 20 TB
4 year, $45,000,000 project
to add 2M books

“RLB” Prediction:
Robo-Line Bancroft = $500,000

Bancroft costs:
- $20 per book to catalogue
- $1 per book to reshelve a book
- 20% new books purchased per year never checked out

Massive Storage: Summary



Open Research Issues

- Hardware/Software Attack on very large storage systems
  - File system extensions:
    - Handle terabyte sized file systems
  - Storage controllers: able to meet
    - Bandwidth / capacity demands
  - Managing storage: 3-4 levels
  - Hiding inherent latency of the technology
  - Guaranteed protection of information
    - No copying?

Open Research Issues

- Compression/decompression
  - Hardware assist
    for on-the-fly compression
  - Application hints
    for data specific compression
  - Image thumbnails:
    lower resolution images used for search

- Reliability Scheme?

- Overcoming / Hiding Latency
  in Robo-Line Storage?

MSS Conclusion and Prediction

- Trend: Minicomputer price will buy
  1000 X storage
  Access time measured in 10s of secs
- Obstacles to adaptation of technology:
  Legal copyrights to interesting information?
  Paper-based Publishers agree to license-based income as do software publishers?
- Patterson predicts by 2000:
  1000X increase in on-line storage
  will have greater impact on society than
  1000X increase in CPU speed

Massive Storage Systems: Foundations

- **Memory is multifaceted:**

  Cost-Capacity ($/MB)

  Cost-Performance (Latency)

  Cost-Performance (Bandwidth)

  Cost-Reliability

  Cost-Non-volatility

Massive Storage Systems: Foundations (continued)

- **Memory is a hierarchy of diffferent technologies**
- **Upper technology must dominate the lower in Cost-Performance;**
- **The lower must dominate the upper in Cost-Capacity**

(Otherwise a level disappears)

Hence must know trends at every level to understand the future of MSS

Abstracts: Overcoming Latency

- Thematic Mapper Image:
  7 x 7000 x 6000 x 1 byte ≈ 300 MB

- Today: T1 + Ethernet + Exabyte - 120

50s	Tape robot unloads the tape & loads new tape
45s	Reader seeks to file
600s	Reads file @ 4 Mbit/s
1600s	UCB to UCSB over T1 @ 1.5 Mbit/s
240s	UCSB FS to WS over Ethernet @10 Mbit/s
2535s	(45 min) Total @ 100% of rated net bandwidth

Sprite's Log Structured File System

Large file caches
effectively reduce disk reads
Disk traffic likely
to be dominated by writes

Write-Optimized File System

- *Only representation on disk is log*
- *Stream out files, directories, maps without seeks*

Advantages:

Speed
Stripes easily across the array
Fast recovery
Temporal locality
Versioning

Problems:

Random access retrieval
Log wrap
Disk space utilization

Performance of RAID using Parity

- Read from all disks works well: "Large Reads"
- Write to all disks works well: "Large Writes"
- Each sector has error detection information

  => could read individual sectors from different disks at same time: "Small Reads"work well
- Small writes?

Small writes

Solution 1:

Write new sector
Read sector from all other disks in group
Write new parity

Write new parity ➡ awful performance:
1 small write ➡ N – 1 reads, 2 writes

Small writes

Solution 2:

- Read old sector
- See which bits change
- Write new sector
- Read old parity sector
- Change the bits that changed with writes

Write new parity sector ➡ better performance:
1 small write ➡ 2 reads, 2 writes

But single parity disk limits number of simultaneous small writes ➡ bottleneck

Automated Tape Library
EXB-120

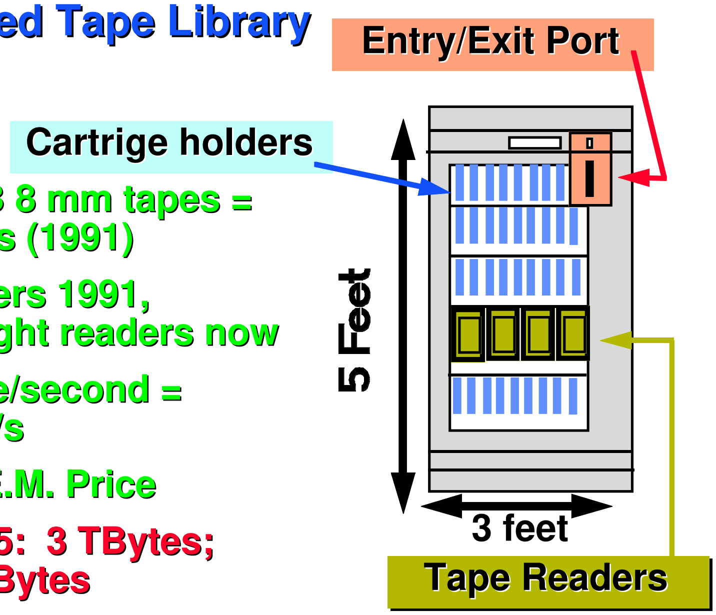
116 x 5 GB 8 mm tapes =
0.6 TBytes (1991)

4 tape readers 1991,
8 half height readers now

4 x .5 MByte/second =
2 MBytes/s

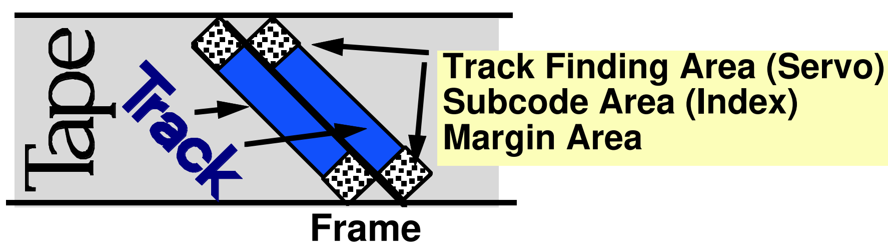
$40,000 O.E.M. Price

Predict 1995: 3 TBytes;
2000: 9 TBytes



R-DAT Technology

DDS ANSI Standard (HP, SONY)

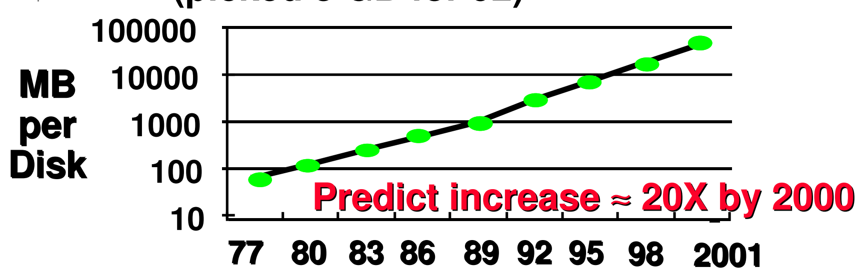


Block
Track (2900 Data Bytes)
Frame (2 Tracks)
Group (22 Frames +
Optional Group ECC, 128K bytes)

65% of Track is Data Area
70% Data Bytes
30% Bytes Parity Plus
Reed-Solomon Codes

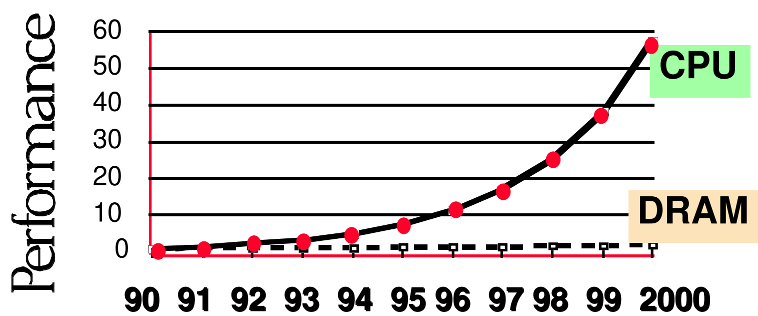
MSS Technology: Magnetic Disks

- Trend: Capacity & $/MB: Improving 27% / year (50X since 1977); since 1990 at 50% / year; now?
- Trend: Evolving to smaller physical size
- Improving areal density (bits/sq.in.) vs. a single target per year like DRAM (picked 3 GB for 92)



DRAM Performance

- Latency & Bandwidth
- Trend: Incredible processor progress creates processor-memory performance gap:

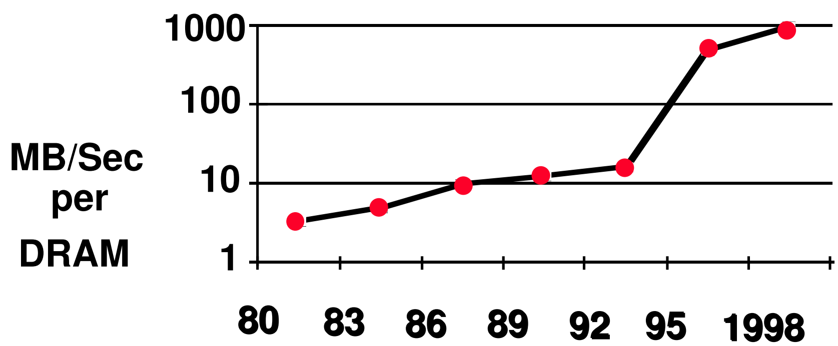


- Predict improvement by 2000: DRAM latency 2X improved; CPU 50X

Future DRAM

- Latency: caches, Prefetching
  caches accesses to use bandwidth
- Bandwidth: new DRAM (RAMbus
  and Synch. DRAM)

Multiple level of Caches to group



I am not sure of the meaning or content of this slide

- **Prediction**

interface for bandwidth;

full DRAM access = 20 to 40 instr.

interface for bandwidth;

full DRAM access = 40 to 80 instr.

1995: new DRAM

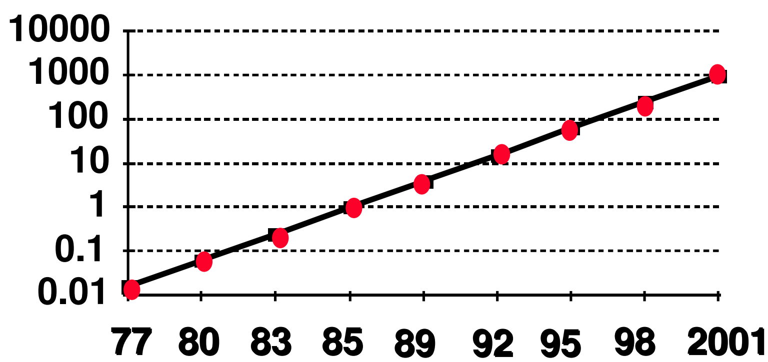
no help latency:

2000: new DRAM

no help latency:

DRAM capacity growth

- Trend: 4X every 3 years:

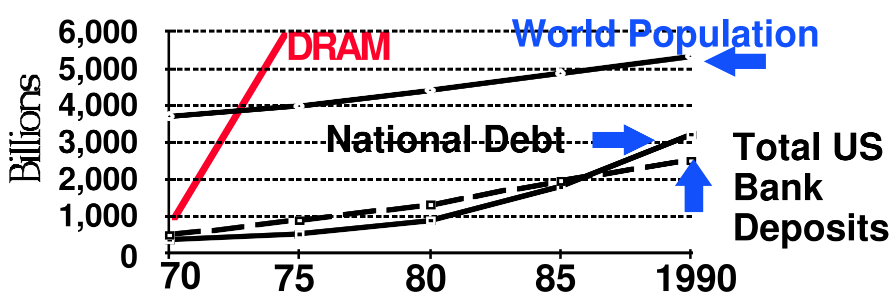


- Feb '93 @ ISSCC: 3 companies unveil 256 Mbits

DRAM Perspective

- > 1000X increase since 1977!
  another > 50X by ≈2001!

Compared to other phenomena:



Since 1970: < 10X Banks, Debt; < 2X population

Prediction: *Traditional* databases all in DRAM

Slide 67
Terabytes >> Teraflops