

Computers for the Post-PC Era

Aaron Brown, Jim Beck, Kimberly Keeton,
Rich Martin, David Oppenheimer, Randi
Thomas, John Kubiawicz, Kathy Yelick,
and David Patterson

<http://iram.cs.berkeley.edu/istore>

1999 Industrial Relations Conference

Slide 1

Outline

- Motivation and Background: Berkeley's Past
- ISTORE Goals
- Hardware Architecture
- Software Architecture
- Discussion and Feedback

Slide 2

Motivation

- Next generation fixes problems of last gen.
- 1960s: batch processing + slow turnaround
⇒ Timesharing
 - 15-20 years of performance improvement, cost reduction (minicomputers, semiconductor memory)
- 1980s: Time sharing + inconsistent response times ⇒ Workstations/Personal Computers
 - 15-20 years of performance improvement, cost reduction (microprocessors, DRAM memory, disk)
- 2000s: PCs + difficulty of use/high cost of ownership ⇒ ???

Slide 3

Perspective on Post-PC Era

- PostPC Era Divides built on two technologies:
 - 1) Mobile Consumer Electronic Devices
 - e.g., successor to PDA, Cell phone
 - Prior talks in this session
 - See Posters on Ninja, Iceberg, IRAM (12-1:30) and Post PC session 1:30-3:30 in 306 Soda
 - 2) Infrastructure to Support such Devices
 - e.g., successor to Big Fat Web Servers, Databases
 - This talk and Posters on ISTORE (12-1:30)

Slide 4

Background for ISTORE: 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



Slide 5

Background for ISTORE: RAID-II

• RAID-II (1993)

- A network attached storage device. 2 outer racks contained 144 disks (3.5" IBM 320 MB SCSI) & power supplies. Center rack in 3 parts: top chassis holds VME disk controller boards, center chassis contains custom crossbar switch and HIPPI network (1Gb/s) interface boards; bottom chassis contains the Sun 4/280 workstation.



Slide 6

Background: Tertiary Disk

• Tertiary Disk (1997)

- cluster of 20 PCs hosting 364 3.5" IBM disks (8.4 GB) in 7 7'x19" racks, or 3 TB. The 200MHz, 96 MB P6 PCs run FreeBSD and a switched 100Mb/s Ethernet connects the hosts. Also 4 UPS units.



- Hosts world's largest art database: 72,000 images in cooperation with San Francisco Fine Arts Museum: Try www.thinker.org

Slide 7

Tertiary Disk HW Failure Experience

Reliability of hardware components (20 months)

7 IBM SCSI disk failures	(out of 364, or 2%)
6 IDE (internal) disk failures	(out of 20, or 30%)
1 SCSI controller failure	(out of 44, or 2%)
1 SCSI Cable	(out of 39, or 3%)
1 Ethernet card failure	(out of 20, or 5%)
1 Ethernet switch	(out of 2, or 50%)
3 enclosure power supplies	(out of 92, or 3%)
1 short power outage	(covered by UPS)

Did not match expectations:
SCSI disks more reliable than cables!

Slide 8

Saw 2 Error Messages per Day

• SCSI Error Messages:

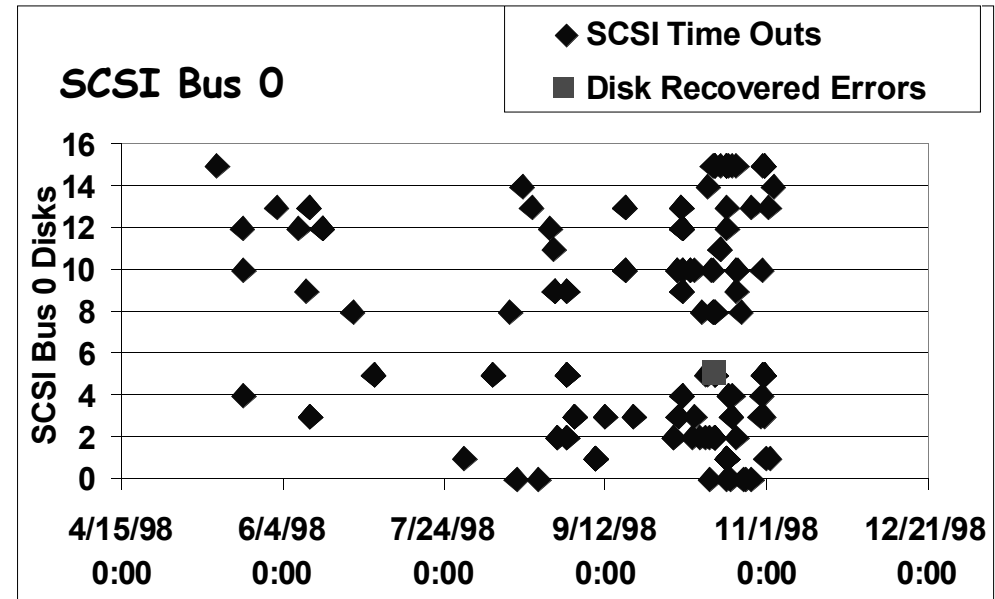
- Time Outs: Response: a BUS RESET command
- Parity: Cause of an aborted request

• Data Disk Error Messages:

- Hardware Error: The command unsuccessfully terminated due to a non-recoverable HW failure.
- Medium Error: The operation was unsuccessful due to a flaw in the medium (try reassigning sectors)
- Recovered Error: The last command completed with the help of some error recovery at the target
- Not Ready: The drive cannot be accessed

Slide 9

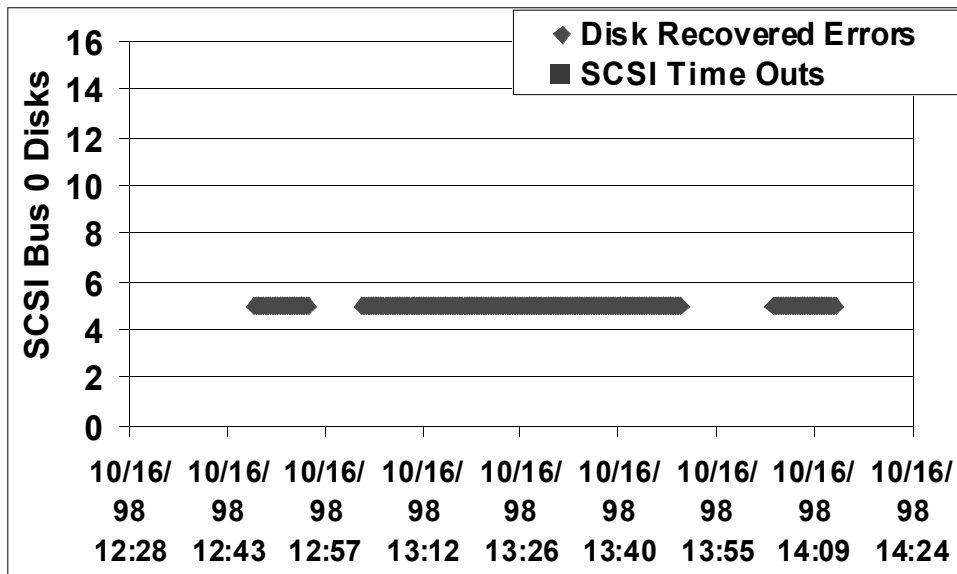
SCSI Time Outs+Recovered Errors (m0)



Slide 10

Zoom In: Disk Recovered Errors

SCSI Bus 0



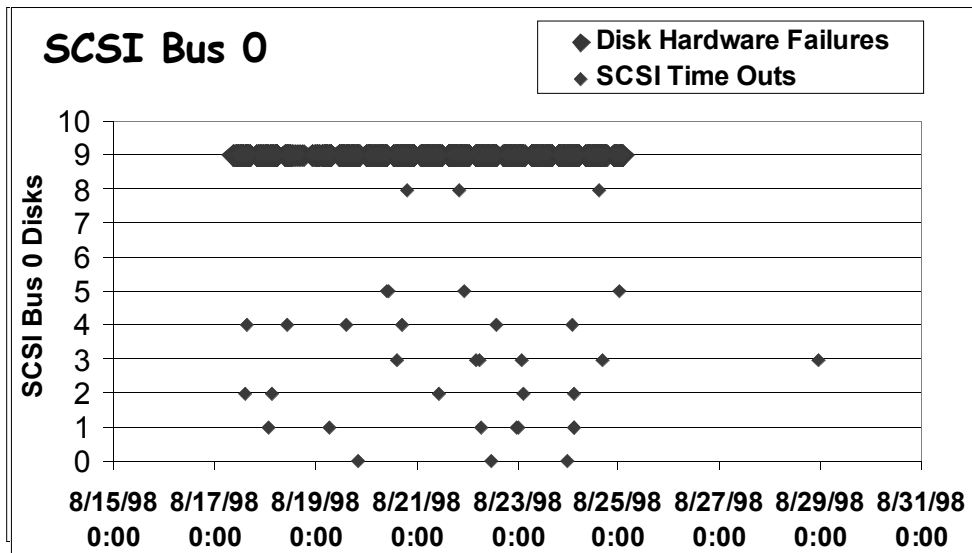
Slide 11

Can we predict a disk failure?

- Yes, we can look for Recovered Error messages \Rightarrow on 10-16-98:
 - There were 433 Recovered Error Messages
 - These messages lasted for slightly over an hour between: 12:43 and 14:10
- On 11-24-98: Disk 5 on m0 was "fired", i.e. it appeared to operator it was about to fail, so it was swapped

Slide 12

SCSI Time Outs + Hardware Failures (m11)



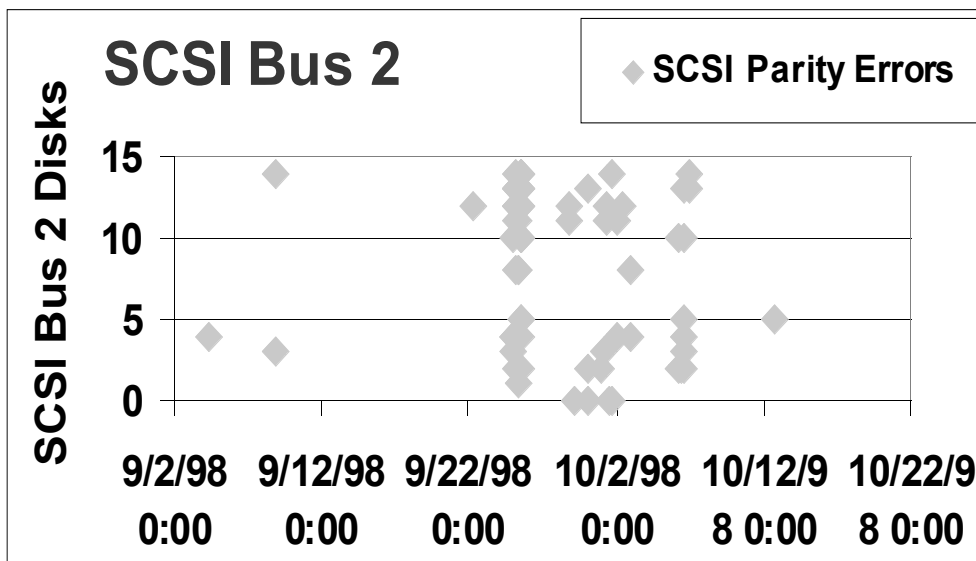
Slide 13

Can we predict a disk failure?

- Yes, look for Hardware Error messages
 - These messages lasted for 8 days between:
 - » 8-17-98 and 8-25-98
 - On disk 9 there were:
 - » 1763 Hardware Error Messages, and
 - » 297 SCSI Timed Out Messages
- On 8-28-98: Disk 9 on SCSI Bus 0 of m11 was "fired", i.e. appeared it was about to fail, so it was swapped

Slide 14

SCSI Bus 2 Parity Errors (m2)



Slide 15

Can We Predict Other Kinds of Failures?

- Yes, the flurry of parity errors on m2 occurred between:
 - 1-1-98 and 2-3-98, as well as
 - 9-3-98 and 10-12-98
- On 11-24-98
 - m2 had a bad enclosure
 - ⇒ cables or connections defective
 - The enclosure was then replaced

Slide 16

Lessons from Tertiary Disk Project

- **Maintenance is hard on current systems**
 - Hard to know what is going on, who is to blame
- **Everything can break**
 - Its not what you expect in advance
 - Follow rule of no single point of failure
- **Nothing fails fast**
 - Eventually behaves bad enough that operator fires poor performer, but it doesn't quit
- **Many failures may be predicted**

Slide 17

Outline

- **Motivation and Background: Berkeley's Past**
- **ISTORE Goals**
- **Hardware Architecture**
- **Software Architecture**
- **Discussion and Feedback**

Slide 18

Storage Priorities: Research v. Users

Current Research Priorities	Current Server Customer Priorities
------------------------------------	-------------------------------------------

- | | |
|--------------------|--------------------|
| 1) Performance | 1) Availability |
| 1') Cost | 2) Maintainability |
| 3) Scalability | 3) Scalability |
| 4) Availability | 4) Performance |
| 5) Maintainability | 5) Cost |

(From Sun marketing presentation, 2/99)

Slide 19

Intelligent Storage Project Goals

- **ISTORE: a hardware/software architecture for building scaleable, self-maintaining storage**
 - An *introspective* system: it monitors itself and acts on its observations
- **Self-maintenance: does not rely on administrators to configure, monitor, or tune system**

Slide 20

Self-maintenance

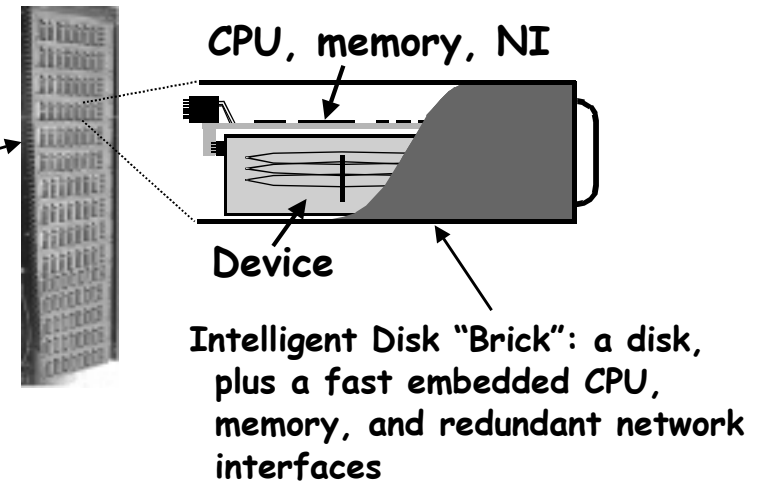
- **Failure management**
 - devices must fail fast without interrupting service
 - predict failures and initiate replacement
 - failures \neq immediate human intervention
- **System upgrades and scaling**
 - new hardware automatically incorporated without interruption
 - new devices immediately improve performance or repair failures
- **Performance management**
 - system must adapt to changes in workload or access patterns

Slide 21

ISTORE-I Hardware

- ISTORE uses "intelligent" hardware

Intelligent
Chassis:
scalable,
redundant,
fast
network +
UPS



Slide 22

ISTORE-I: Summer 99?

- **Intelligent disk**
 - Portable PC Hardware: Pentium II, DRAM
 - Low Profile SCSI Disk (9 to 18 GB)
 - 4 100-Mbit/s Ethernet links per Idisk
 - Placed inside Half-height canister
 - Monitor Processor/path to power off components?
- **Intelligent Chassis**
 - 64 IDisks: 8 enclosures, 8 IDisks/enclosure
 - » 64 x 4 or 256 Ethernet ports
 - 2 levels of Ethernet switches: 14 small, 2 large
 - » Small: 20 100-Mbit/s + 2 1-Gbit; Large: 25 1-Gbit
 - Enclosure sensing, UPS, redundant PS, fans, . . .

Slide 23

ISTORE Hardware Vision

- System-on-a-chip enables computer, memory, redundant network interfaces without increasing size of disk canister
- Disk enclosure includes (redundant) 1st-level switches as well as redundant power supplies, fans
- Rack includes 2nd-level switches, UPS

Slide 24

ISTORE-I Software Plan

- **Modify Database (e.g., Predator) to send log to mirrored Idisk**
 - Since 1 processor per disk, continuously replay the log on mirrored system
- **Insert faults in original Idisk to get fail over**
- **Add monitoring, maintenance, fault insertion**
- **Run ****ix OS****
 - By running Linux binaries, can get multiple OS with same API: Linux, Free BSD Unix, ...
 - Increase genetic base of OS software to reduce chances of simultaneous software bugs
 - Periodic reboot to "refresh" system

Slide 25

Benefits of ISTORE

- **Decentralized processing (shared-nothing)**
 - system can withstand partial failure
- **Monitor their own "health," test themselves, manage failures, collect application-specified performance data, and execute applications**
 - fault insertion to test availability
 - provides the foundation for self-maintenance and self-tuning
- **Plug & play, hot-swappable bricks ease configuration, scaling**
 - hardware maybe specialized by selecting an collection of devices: DRAMs, WAN/LAN interfaces

Slide 26

Other (Potential) Benefits of ISTORE

- **Scalability: add processing power, memory, network bandwidth as add disks**
- **Smaller footprint vs. traditional server/disk**
- **Less power**
 - embedded processors vs. servers
 - spin down idle disks?
- **For decision-support or web-service applications, potentially better performance than traditional servers**

Slide 27

Related Work

- **ISTORE adds several recent research efforts**
 - Active Disks, NASD (UCSB, CMU)
 - Network service appliances (NetApp, Snap!, Qube, ...)
 - High availability systems (Compaq/Tandem, ...)
 - Adaptive systems (HP AutoRAID, M/S AutoAdmin, M/S Millennium)
 - Plug-and-play system construction (Jini, PC Plug&Play, ...)

Slide 28

Interested in Participating?

- Project just getting formed
- Contact us if you're interested:
<http://iram.cs.berkeley.edu/istore>
email: patterson@cs.berkeley.edu
- Thanks for support: DARPA
- Thanks for advice/inspiration:
Dave Anderson (Seagate),
Greg Papadopolous (Sun), Mike Ziegler (HP)

Slide 29

Backup Slides

Slide 30

ISTORE Cluster?

- 8 -12 disks / enclosure
- 12 enclosures / rack = 96-144 disks/rack



Cluster of PCs?

- 2 disks / PC
- 10 PCs /rack = 20 disks/rack
- Reliability?
- Ease of Repair?
- System Admin.?
- Cost only plus?



Slide 31

ISTORE and IRAM

- ISTORE relies on intelligent devices
- IRAM is an easy way to add intelligence to a device
 - embedded, low-power CPU meets size and power constraints
 - integrated DRAM reduces chip count
 - fast network interface (serial lines) meets connectivity needs
- Initial ISTORE prototype won't use IRAM
 - will use collection of commodity components that approximate IRAM functionality, not size/power

Slide 32