

Computers for the Post-PC Era

Aaron Brown, Jim Beck, Rich Martin,
David Oppenheimer, Kathy Yelick, and
David Patterson

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

1999 Grad Visit Day

Slide 1

Berkeley Approach to Systems

- Find an important problem crossing HW/SW Interface, with HW/SW prototype at end
- Assemble a band of 3-6 faculty, 12-20 grad students, 1-3 staff to tackle it over 4 years
- Meet twice a year for 3-day retreats with invited outsiders
 - Builds team spirit
 - Get advice on direction, and change course
 - Offers milestones for project stages
 - Grad students give 6 to 8 talks Great Speakers
- Write papers, go to conferences, get PhDs, jobs
- End of project party, reshuffle faculty, go to .1

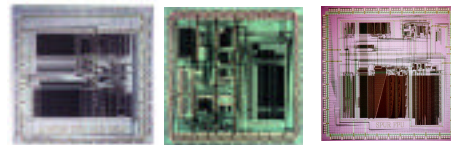
For Example, Projects I Have Worked On

- RISC I, II
 - Sequin, Ousterhout (CAD)
- SOAR (Smalltalk On A RISC) Ousterhout (CAD)
- SPUR (Symbolic Processing Using RISCs)
 - Fateman, Hilfinger, Hodges, Katz, Ousterhout
- RAID I, II (Redundant Array of Inexp. Disks)
 - Katz, Ousterhout, Stonebraker
- NOW I, II (Network of Workstations), (TD)
 - Culler, Anderson
- IRAM I (Intelligent RAM)
 - Yelick, Kubiawicz, Wawrzynek
- ISTORE I, II (Intelligent Storage)
 - Yelick, Kubiawicz

Slide 3

Symbolic Processing Using RISCs: '85- '89

- Before Commercial RISC chips
- Built Workstation Multiprocessor and Operating System from scratch(!)
- Sprite Operating System
- 3 chips: Processor, Cache Controller, FPU
 - Coined term "snopping cache protocol"
 - 3C's cache miss: compulsory, capacity, conflict



Slide 4

SPUR 10 Year Reunion, January '99

- Everyone from North America came!
- 19 PhDs: 9 to Academia
 - 8/9 got tenure, 2 full professors (already)
 - 2 Romme fellows (3rd, 4th at Wisconsin)
 - 3 NSF Presidential Young Investigator Winners
 - 2 ACM Dissertation Awards
 - They in turn have produced 30 PhDs (so far)
- 10 to Industry
 - Founders of 4 startups, (1 failed)
 - 2 Department heads (AT&T Bell Labs, Microsoft)
- Very successful group; SPUR Project "gave them a taste of success, lifelong friends",

Slide 5

Group Photo (in souvenir jackets)



- See www.cs.berkeley.edu/Projects/ARC to learn more about Berkeley Systems

Slide 6

Outline

- Background: Berkeley Approach to Systems
- PostPC Motivation
- PostPC Microprocessor: IRAM
- PostPC Infrastructure Motivation
- PostPC Infrastructure: ISTORE
- Hardware Architecture
- Software Architecture
- Conclusions and Feedback

Slide 7

Perspective on Post-PC Era

- PostPC Era will be driven by two technologies:
 - 1) Mobile Consumer Electronic Devices
 - e.g., successor to PDA, Cell phone, wearable computers
 - 2) Infrastructure to Support such Devices
 - e.g., successor to Big Fat Web Servers, Database Servers

Slide 8

Intelligent PDA (2003?)

Pilot PDA

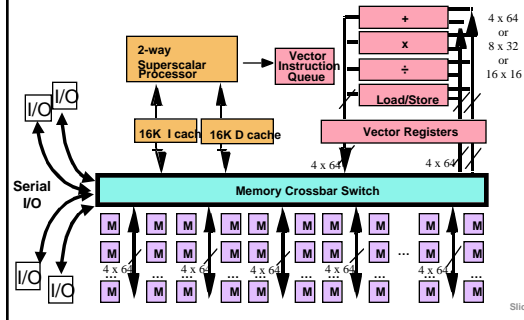
- + gameboy, cell phone, radio, timer, camera, TV remote, am/fm radio, garage door opener, ...
- + Wireless data (WWW)
- + Speech, vision recog.
- + Voice output for conversations



Speech control
+ Vision to see, scan documents, read bar code, ...

Slide 9

V-IRAM1: 0.18 μm , Fast Logic, 200 MHz 1.6 GFLOPS(64b)/6.4 GOPS(16b)/32MB

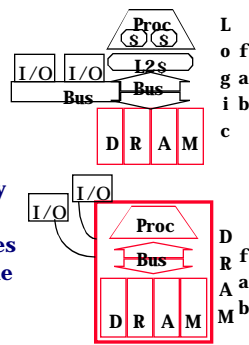


Slide 10

IRAM Vision Statement

Microprocessor & DRAM on a single chip:

- 10X capacity vs. DRAM
- on-chip memory latency 5-10X, bandwidth 50-100X
- improve energy efficiency 2X-4X (no off-chip bus)
- serial I/O 5-10X v. buses
- smaller board area/volume
- adjustable memory size/width



Slide 11

Outline

- PostPC Infrastructure Motivation and Background: Berkeley's Past
- PostPC Motivation
- PostPC Device Microprocessor: IRAM
- PostPC Infrastructure Motivation
- ISTORE Goals
- Hardware Architecture
- Software Architecture
- Conclusions and Feedback

Slide 12

Background: Tertiary Disk (part of NOW)

Tertiary Disk (1997)

- cluster of 20 PCs hosting 364 3.5" IBM disks (8.4 GB) in 7 19"x 33" x 84" 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 13

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 SCSI cables!

Difference between simulation and prototypes

Slide 14

Saw 2 Error Messages per Day

SCSI Error Messages:

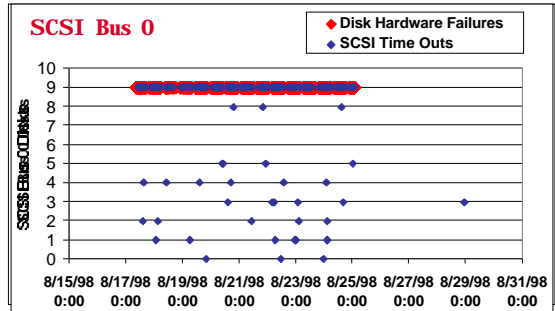
- **Time Outs:** Response: a BUS RESET command
- **Parity:** Cause of an aborted request

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 15

SCSI Time Outs + Hardware Failures (m11)



Slide 16

Can we predict a disk failure?

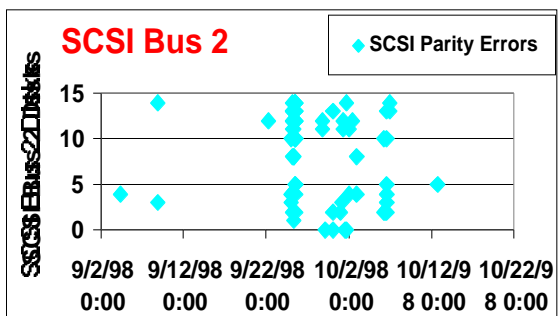
Yes, look for Hardware Error messages

- 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 17

SCSI Bus 2 Parity Errors (m2)



Slide 18

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 19

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"
- Most failures may be predicted

Slide 20

Outline

- Background: Berkeley Approach to Systems
- PostPC Motivation
- PostPC Microprocessor: IRAM
- PostPC Infrastructure Motivation
- PostPC Infrastructure: ISTORE
- Hardware Architecture
- Software Architecture
- Conclusions and Feedback

Slide 21

Storage Priorities: Research v. Users

Current Research Priorities

- 1) Performance
- 1') Cost
- 3) Scalability
- 4) Availability
- 10) Maintainability

} easy to measure

ISTORE Priorities

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

Slide 22

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 23

Self-maintenance

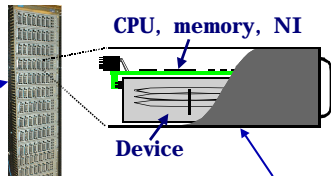
- Failure management
 - devices must fail fast without interrupting service
 - predict failures and initiate replacement
 - failures ∇ immediate human intervention
- System upgrades and scaling
 - new hardware automatically incorporated without interruption
 - new devices immediately improve performance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

Slide 24

ISTORE- I Hardware

- ISTORE uses “intelligent” hardware

Intelligent Chassis: scaleable, redundant, fast network + UPS



Intelligent Disk “Brick”: a disk, plus a fast embedded CPU, memory, and redundant network interfaces

Slide 25

ISTORE- I: 2H99?

- **Intelligent disk**
 - Portable PC Hardware: Pentium II, DRAM
 - Low Profile SCSI Disk (9 to 18 GB)
 - 4 100-Mbit/s Ethernet links per node
 - Placed inside Half-height canister
 - Monitor Processor/path to power off components?
- **Intelligent Chassis**
 - 64 nodes: 8 enclosures, 8 nodes/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 26

Disk Limit

- 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 |
| 1999 | 35 minutes | 1 week(!) |
- 3.5” form factor make sense in 5- 7 years?

Slide 27

2006 ISTORE

- **IBM MicroDrive**
 - 1.7” x 1.4” x 0.2”
 - 1999: 340 MB, 5400 RPM, 5 MB/s, 15 ms seek
 - 2006: 9 GB, 50 MB/s?
- **ISTORE node**
 - MicroDrive + IRAM
- **Crossbar switches growing by Moore’s Law**
 - 16 x 16 in 1999 64 x 64 in 2005
- **ISTORE rack (19” x 33” x 84”)**
 - 1 tray (3” high) 16 x 32 512 ISTORE nodes
 - 20 trays+switches+UPS **10,240 ISTORE nodes(!)**



Slide 28

Software Motivation

- Data-intensive network-based services are becoming the most important application for high-end computing
- But servers for them are too hard to manage!
- **We need single-purpose, introspective storage appliances**
 - single-purpose: customized for one application
 - introspective: self-monitoring and adaptive
 - » with respect to component failures, addition of new hardware resources, load imbalance, workload changes, ...
- **But introspective systems are hard to build!**

Slide 29

Introspective Storage Service

- **Single-purpose, introspective storage**
 - single-purpose: customized for one application
 - introspective: self-monitoring and adaptive
- **Software:** toolkit for defining and implementing application-specific monitoring and adaptation
 - **base layer** supplies repository for monitoring data, mechanisms for invoking reaction code
 - for common adaptation goals, appliance designer’s **policy statements** guide automatic generation of adaptation algorithms
- **Hardware:** intelligent devices with integrated self-monitoring

Slide 30

Base Layer: Views and Triggers

- Monitoring data is stored in a dynamic system database
 - device status, access patterns, perf. stats, ...
- System supports **views** over the data ...
 - applications select and aggregate data of interest
 - defined using SQL-like declarative language
- ... as well as application-defined **triggers** that specify interesting situations as **predicates** over these views
 - triggers invoke application-specific reaction code when the predicate is satisfied
 - defined using SQL-like declarative language

Slide 31

From Policy Statements to Adaptation Algorithms

- For common adaptation goals, designer can write simple policy statements
- Runtime **integrity constraints** over data stored in the DB
- System automatically generates appropriate views, triggers, & adaptation code templates
- **claim:** doable for common adaptation mechanisms needed by data-intensive network services
 - component failure, data hot-spots, integration of new hardware resources, ...

Slide 32

Conclusion and Status 1/2

- IRAM attractive for **both** drivers of PostPC Era: Mobile Consumer Electronic Devices and Scaleable Infrastructure
 - Small size, low power, high bandwidth
- ISTORE: hardware/software architecture for single-use, introspective storage appliances
- Based on
 - intelligent, **self-monitoring hardware**
 - a virtual **database** of system status and statistics
 - a **software toolkit** that uses a domain-specific declarative language to specify integrity constraints
- 1st HW Prototype being constructed;
1st SW Prototype just starting

Slide 33

ISTORE Conclusion 2/2

- Qualitative Change for every factor 10X
Quantitative Change
 - Then what is implication of 100X?
- PostPC Servers no longer "Binary" ?
(1 perfect, 0 broken)
 - infrastructure never perfect, never broken
- PostPC Infrastructure Based on Probability Theory (>0, <1), not Logic Theory (true or false)?
- Look to Biology, Economics for useful models?
<http://iram.cs.berkeley.edu/istore>

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 35

Backup Slides

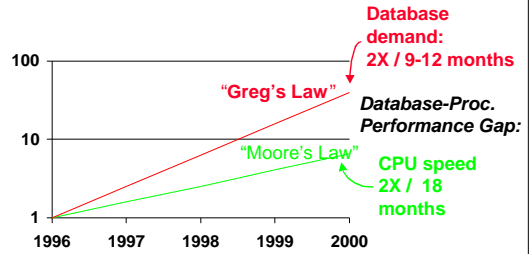
Slide 36

Post PC 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

Slide 37

User Decision Support Demand vs. Processor speed



Slide 38

State of the Art: Seagate Cheetah 36

- 36.4 GB, 3.5 inch disk
- 12 platters, 24 surfaces
- 10,000 RPM
- 18.3 to 28 MB/s internal media transfer rate
- 9772 cylinders (tracks), (71,132,960 sectors total)
- Avg. seek: read 5.2 ms, write 6.0 ms (Max. seek: 12/13, 1 track: 0.6/0.9 ms)
- \$2100 or 17MB/S (6¢/MB) (list price)
- 0.15 ms controller time



Slide 39

Disk Limit: I/O Buses

- Multiple copies of data, SW layers
 - Cannot use 100% of bus
 - Queuing Theory (< 70%)
 - Command overhead (Effective size = size x 1.2)
- Diagram illustrating I/O buses: CPU, Memory bus, Internal I/O bus (PCI), External I/O bus (SCSI), and Controllers (15 disks).
- Bus rate vs. Disk rate
 - SCSI: Ultra2 (40 MHz), Wide (16 bit): 80 MByte/s
 - FC-AL: 1 Gbit/s = 125 MByte/s (single disk in 2002)

Slide 40

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 41

Related Work

- ISTORE adds to 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 42

New Architecture Directions for PostPC Mobile Devices

- “...media processing will become the dominant force in computer arch. & MPU design.”
- “... new media- rich applications... involve significant real- time processing of continuous media streams, & make heavy use of **vectors of packed 8-, 16-, and 32- bit integer and Fl.Pt.**”
- Needs include real- time response, continuous media data types, fine grain parallelism, coarse grain parallelism, memory BW
 - “How Multimedia Workloads Will Change Processor Design”. Diefendorff & Dubey. IEEE Computer (9/97)

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 44