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 Sun CTO Conference
Outline

PostPC Motivation
PostPC Microprocessor: IRAM
PostPC Infrastructure Motivation and Background: Berkeley's Past
PostPC Infrastructure: ISTORE
Hardware Architecture
Software Architecture
Conclusions and Feedback
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 ⇒ ???
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
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, ...
New Architecture Directions

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)
Microprocessor & DRAM on a single chip:

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

PostPC Motivation
PostPC Device Microprocessor: IRAM
PostPC Infrastructure Motivation and Background: Berkeley’s Past
ISTORE Goals
Hardware Architecture
Software Architecture
Conclusions and Feedback
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
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.
Background: Tertiary Disk

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
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!
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
SCSI Time Outs + Hardware Failures (m11)

SCSI Bus 0

- Disk Hardware Failures
- SCSI Time Outs
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
SCSI Bus 2 Parity Errors (m2)

![Graph showing SCSI Bus 2 Parity Errors](image)
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
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
  It's 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
Storage Priorities: Research v. Users

Current Research Priorities
1) Performance
1') Cost
3) Scalability
4) Availability
5) Maintainability

Current Server Customer Priorities
1) Availability
2) Maintainability
3) Scalability
4) Performance
5) Cost

(From Sun marketing presentation, 2/99)
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
Self-maintenance

**Failure management**
- devices must fail fast without interrupting service
- predict failures and initiate replacement
- failures $\Rightarrow$ 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
ISTORE-I Hardware

ISTORE uses intelligent hardware

Intelligent Chassis: scaleable, redundant, fast network + UPS

CPU, memory, NI

Device

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

Continued advance in capacity (60%/yr) and bandwidth (40%/yr)

Slow improvement in seek, rotation (8%/yr)

Time to read whole disk

<table>
<thead>
<tr>
<th>Year</th>
<th>Sequentially (1 sector/seek)</th>
<th>Randomly (1 sector/seek)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>4 minutes</td>
<td>6 hours</td>
</tr>
<tr>
<td>2000</td>
<td>12 minutes</td>
<td>1 week(!)</td>
</tr>
</tbody>
</table>

3.5 form factor make sense in 5-7 years?
ISTORE Hardware Vision

System-on-a-chip enables computer, memory, redundant network interfaces without significantly increasing size of disk

Target for + 5-7 years:

1999 IBM MicroDrive:
- 1.7 x 1.4 x 0.2
- (43 mm x 36 mm x 5 mm)
- 340 MB, 5400 RPM, 5 MB/s, 15 ms seek

2006 MicroDrive?
- 9 GB, 50 MB/s
- (1.6X/yr capacity, 1.4X/yr BW)
2006 ISTORE

ISTORE node
Add 20% pad to MicroDrive size for packaging, connectors
Then double thickness to add IRAM
2.0 x 1.7 x 0.5 (51 mm x 43 mm x 13 mm)

Crossbar switches growing by Moore’s Law
2x/1.5 yrs ⇒ 4X transistors/3yrs
Crossbars grow by N² ⇒ 2X switch/3yrs
16 x 16 in 1999 ⇒ 64 x 64 in 2005

ISTORE rack (19 x 33 x 84)
(480 mm x 840 mm x 2130 mm)
1 tray (3 high) ⇒ 16 x 32 ⇒ 512 ISTORE nodes
20 trays+switches+UPS ⇒ 10,240 ISTORE nodes(!)
Outline

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

ISTORE = Introspective Storage platform

**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
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
From Policy Statements to Adaptation Algorithms

For common adaptation goals, designer can write simple policy statements.

Runtime invariants expressed as 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, ...
Example

**Invariant**: system must maintain 3 replicas of all data objects

**view**: disks health status

**trigger**: a disk's health status changes to dead
    - invoke adaptation code, supplying
        - identities of objects that were stored on dead disk
        - view of utilization of all disks in the system

**adaptation code template**: lock object, perform byte copy to least utilized disk, unlock object, update index

**adaptation code can be customized to exploit application semantic information**
    - app-specific selection of new disk to hold replicated objects, data layout on disk, locking policy
How Does the Hardware Help?

Intelligent hardware gathers monitoring data implements views and triggers by filtering and reacting to data as it's gathered.

Intelligent Chassis: switching and power

Highly redundant, scalable, and customizable
Conclusion and Status 1/2

IRAM attractive for both drivers of PostPC Era: Mobile Consumer Electronic Devices and Supporting, Scalable 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

HW Prototype being constructed;
SW Prototype just starting
ISTORE Conclusion 2/2

Qualitative Change for every 10X Quantitative Change

Then what is implications 100X?

Systems no longer Binary ?
(1 perfect, 0 broken)

PostPC infrastructure never perfect, never broken

Based on Probability Theory, not Logic Theory?

Look to Biology for useful models?
Interested in Participating?

Project just getting formed

Contact us if you're interested:
http://iram.cs.berkeley.edu/istore
e-mail: patterson@cs.berkeley.edu

Thanks for support: DARPA

Thanks for advice/inspiration:
Dave Anderson (Seagate),
Greg Papadopolous (Sun), Mike Ziegler (HP)
Backup Slides
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
User Decision Support Demand vs. Processor speed

Database demand: 2X / 9-12 months

“Greg’s Law”

“Moore’s Law”

Database-Proc. Performance Gap:

CPU speed 2X / 18 months
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)

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)
- Cannot use 100% of bus
  - Queuing Theory (< 70%)
  - Command overhead (Effective size = size x 1.2)

Controllers (15 disks)
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, ...)
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
ISTORE-I Software Experiment

Modify Database (e.g., Predator) to send log to mirrored node

Since 1 processor per disk, continuously replay the log on mirrored system

Insert faults in original node to get fail over

Add monitoring, maintenance, fault insertion

Run **ix OS

By running Linix binaries, can get multiple OS with same API: Linix, Free BSD Unix, ...

Increase genetic base of OS software to reduce chances of simultaneous software bugs

Periodic reboot to refresh system
State of the Art: Seagate Cheetah 18

6962 cylinders, 12 platters
18.2 GB, 3.5 inch disk
1MB track buffer (+ 4MB optional expansion)
19 watts
0.15 ms controller time
avg. seek = 6 ms (seek 1 track = 1 ms)
1/2 rotation = 3 ms
21 to 15 MB/s media (=> 16 to 11 MB/s)
  » deliver 75% (ECC, gaps...)
$1647 or 11MB/$ (9¢/MB)

Latency = Queuing Time + Controller time + Seek Time + Rotation Time + Size / Bandwidth

per access + per byte

source: www.seagate.com; www.pricewatch.com; 5/21/98
Review technology trends to help?

Desktop Processor:
- + SPEC performance
  - TPC-C performance, CPU-Memory perf. gap

Embedded Processor: + Cost/Perf

Disk  Memory  Network
- Capacity  +  +
- Bandwidth  +  +  +
- Latency
- Interface
State of the art Cluster: NCR WorldMark

BYNET switched network

Bus bridge

Proc
Mem

Bus bridge

Proc
Mem

Bus bridge

Proc
Mem

bus bridge

Mem

... 1 ...

Mem

... 32 ...

Mem

... 64 ...

TPC-D, TD V2, 10/97

32 nodes ×

4 200 MHz CPUs,

1 GB DRAM, 41 disks

(128 cpus, 32 GB, 1312 disks, 5.4 TB)

CPUs, DRAM, encl., boards, power $5,360k

Disks+cntl$ $2,164k

Disk shelves $674k

Cables $126k

Console $16k

HW total $8,340k

source:

www.tpc.org

Slide 47
State of the Art SMP: Sun E10000

4 address buses
data crossbar switch

TPC-D, Oracle 8, 3/98

SMP 64 336 MHz
CPUs, 64GB dram, 668 disks (5.5TB)
Disks, shelf $2,128k
Boards, encl. $1,187k
CPUs $912k
DRAM $768k
Power $96k
Cables, I/O $69k
HW total $5,161k

source: www.tpc.org
State of the Art Cluster: Tandem/Compaq SMP

ServerNet switched network

Rack mounted equipment

SMP: 4-PPro, 3GB dram, 3 disks (6/rack)

10 Disk shelves/rack @ 7 disks/shelf

Total: 6 SMPs (24 CPUs, 18 GB DRAM), 402 disks (2.7 TB)

TPC-C, Oracle 8, 4/98

<table>
<thead>
<tr>
<th>Component</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPUs</td>
<td>$191k</td>
</tr>
<tr>
<td>DRAM,</td>
<td>$122k</td>
</tr>
<tr>
<td>Disks+cntlr</td>
<td>$425k</td>
</tr>
<tr>
<td>Disk shelves</td>
<td>$94k</td>
</tr>
<tr>
<td>Networking</td>
<td>$76k</td>
</tr>
<tr>
<td>Racks</td>
<td>$15k</td>
</tr>
<tr>
<td>HW total</td>
<td>$926k</td>
</tr>
</tbody>
</table>
ISTORE-1 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?
Grove’s Warning:
Only the Paranoid Survive
(vs. Moore’s Law?)

...a strategic inflection point is a time in the life of a business when its fundamentals are about to change. ... Let's not mince words: A strategic inflection point can be deadly when unattended to. Companies that begin a decline as a result of its changes rarely recover their previous greatness.

Only the Paranoid Survive, Andrew S. Grove, 1996