NSIC/NASD workshop:
What do we do with excess computational power in storage devices?

Garth Gibson
Computer Science and Computer Engineering, CMU

CMU perspective on the path to smart storage

Quick summary of NSIC/NASD positions

What are we doing here today?

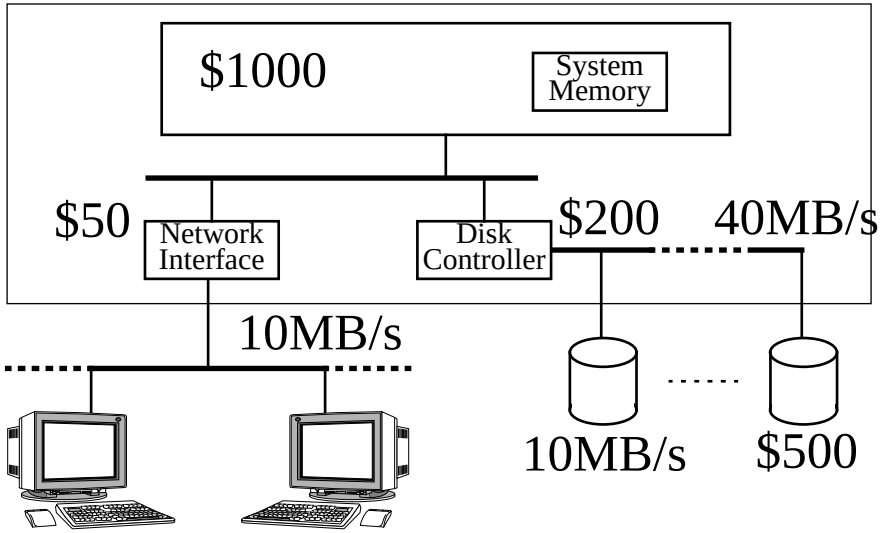
What should we do here tomorrow?

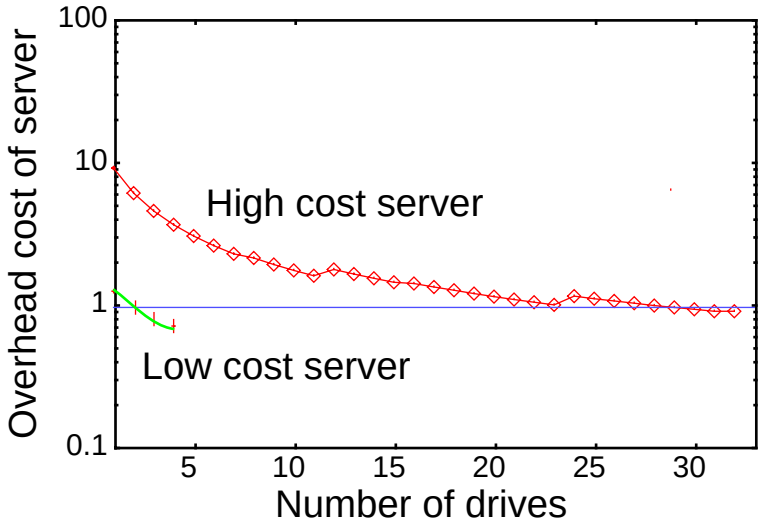
CMU's work is sponsored by DARPA/ITO Quorum/Scalable Systems and HP, Quantum, Seagate, STK, Symbios, Clariion, Compaq, Wind River, Intel, 3Com

Server-Attached Disks don't deliver cost-effective bandwidth

Cheap server workstation, 100Mb ether, UltraSCSI

- server often limited by cycles, PCI bandwidth or PCI slots
- one net, one drive with server overhead cost of > 100%
- AMORTIZE: 4 nets and drives > 70% overhead
- real servers usually much beefier: $7,000 on PC web pages
- AMORTIZE: 24 drives and 3 giga-ether > 100% overhead





CMU's Answer: NASD and SCSI-4

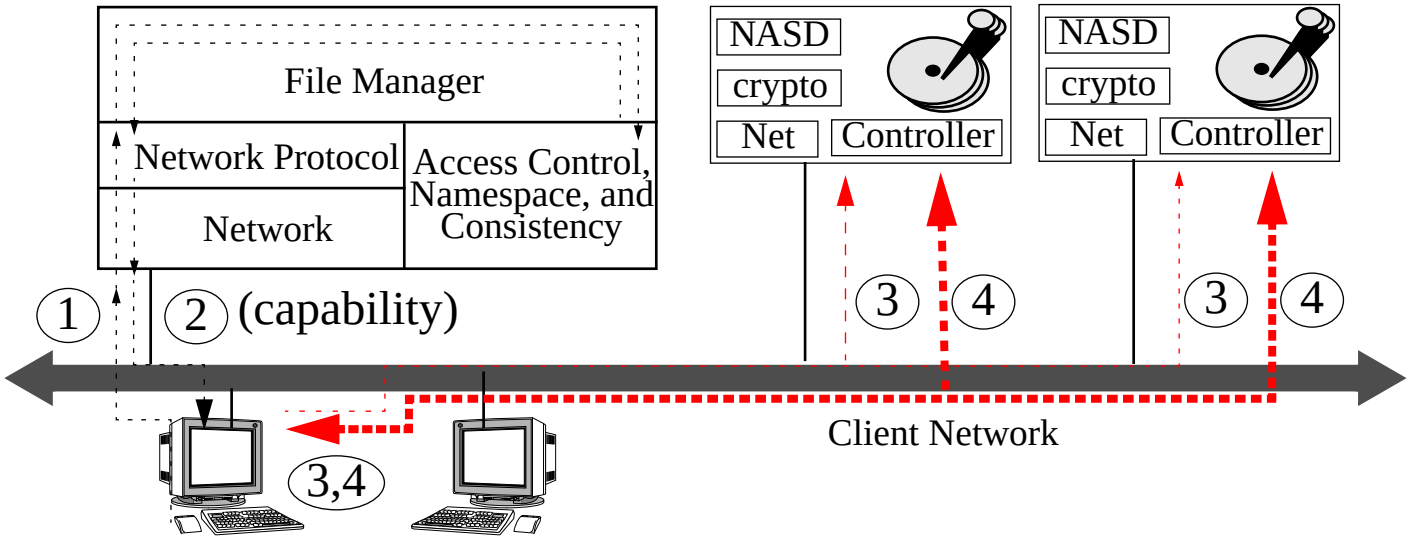
Enable direct transfer between client & storage device

Policy manager (names, access control, consistency, atomicity)

Device understood cryptographic capabilities

Object-oriented devices, datastores, persistent objects

Client-based libraries execute managed storage actions



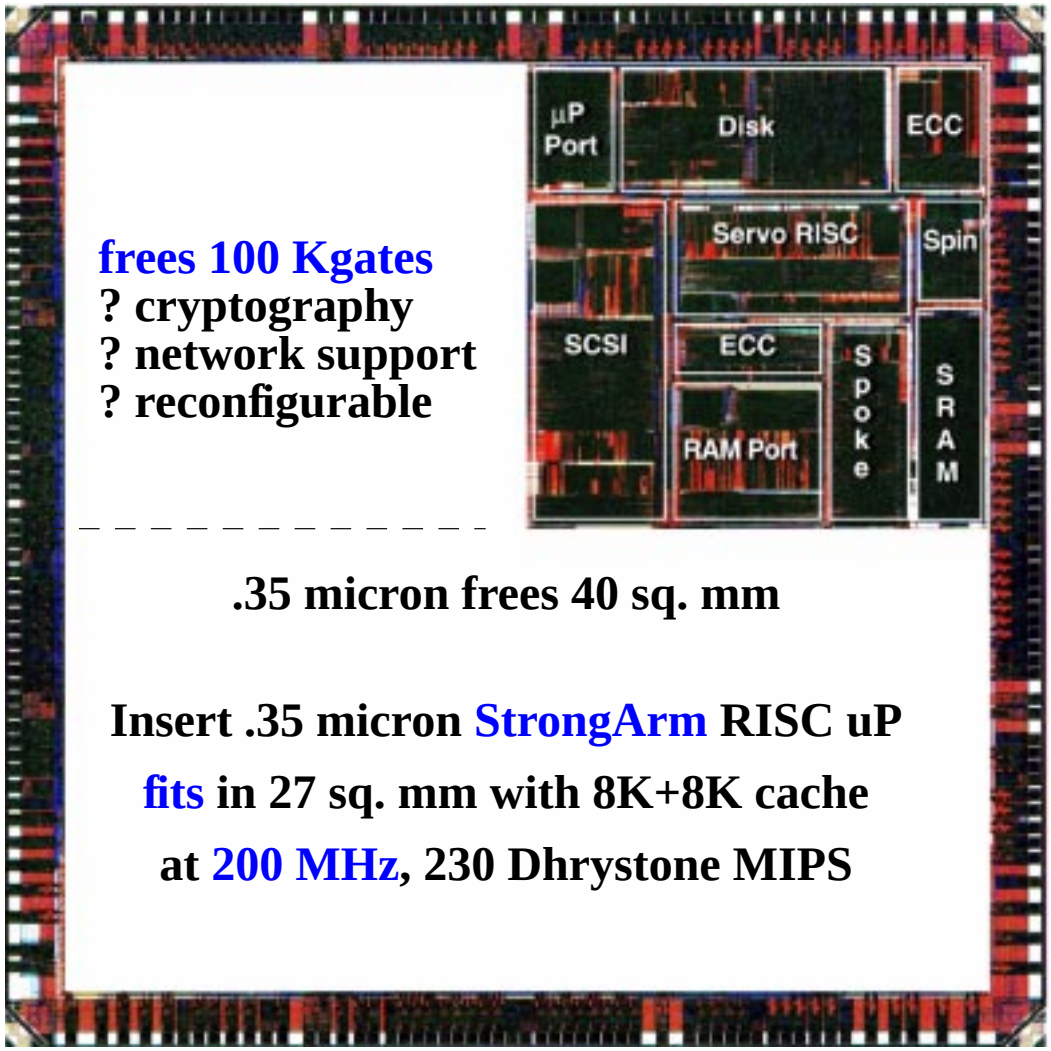
Are Device Cycles Really Available?

Quantum Trident drive

- Control: M68020
- Datapath ASIC
- .68 micron in 1997
- 4 indep clock domains, each 40 MHz
  SCSI processor
  disk R/W channel
  uP control port
  DRAM port
- ~ 110 Kgates + 22Kb
- .35 micron next gen. enables integration of control uP onto ASIC

Also Siemens TriCore

Current .68 micron chip is 74 sq. mm



frees 100 Kgates
? cryptography
? network support
? reconfigurable

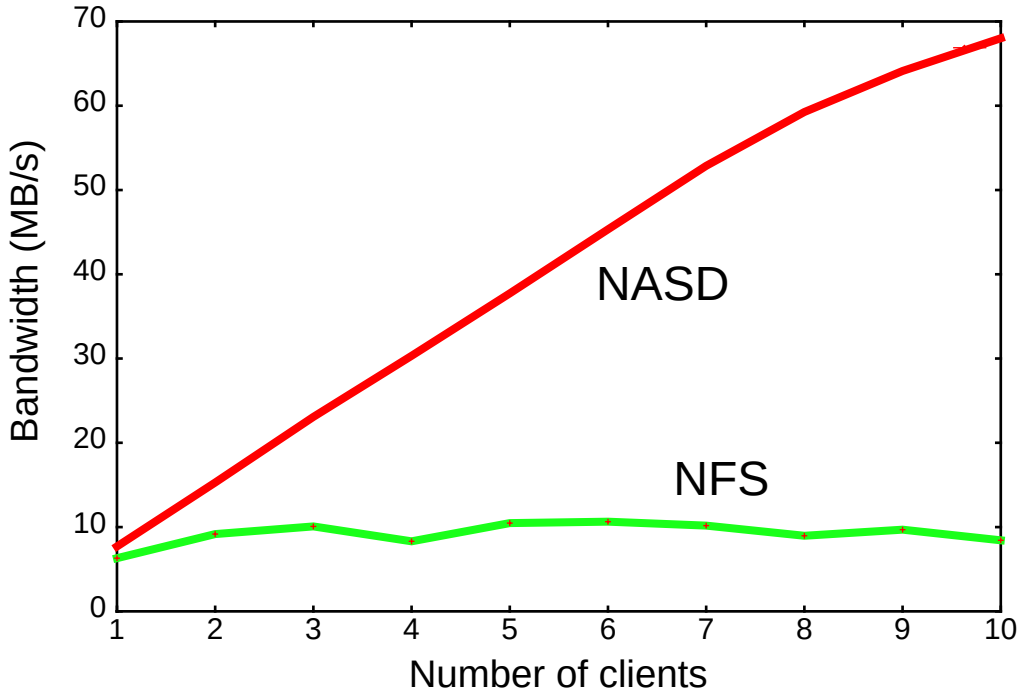
.35 micron frees 40 sq. mm

Insert .35 micron StrongArm RISC uP fits in 27 sq. mm with 8K+8K cache at 200 MHz, 230 Dhrystone MIPS

Demo'd Scalable Bandwidth for Parallel Applications

Client library implementation enables aggressiveness

- **parallel file system accommodates parallel nature of scaling**
- **NASD middleware fetches large blocks in parallel**



Parallel Data Mining

- **13 NASDs (133Mhz)**
- **1-10 clients (233Mhz)**
- **MPI + SIO LLAPI**
- **Cheops + NASD**
- **7.4 MB/s per client until drive limits**
- **switched ATM LAN**

March 5, 1998 - SNIA/NSIC joint NASD workshop

8:00 Conference Introduction, Paul Borrill, Quantum & SNIA Chairman

NSIC Technical program introduction, Barry Schechtman, NSIC

8:15 Introduction to Network-Attached Storage Devices, Garth Gibson, CMU

8:45 NASD and OOD: Seagate's View, Dave Anderson, Seagate

10 Attribute-based Storage Management, Liz Borowsky, Hewlett Packard

10:45 CMU's NASD: Network-Attached Secure Disks, Garth Gibson, CMU

11:30 Secure, Widely Distributed Filesystems, Jim Hughes, STK

1:30 An Object-Oriented Approach to NASD, Geoff Peck, Quantum

2:15 ISI's NASD: Derived Virtual Devices, Rod Van Meter, Quantum

3:00 The Swarm Scalable Storage System, John Hartman, Arizona

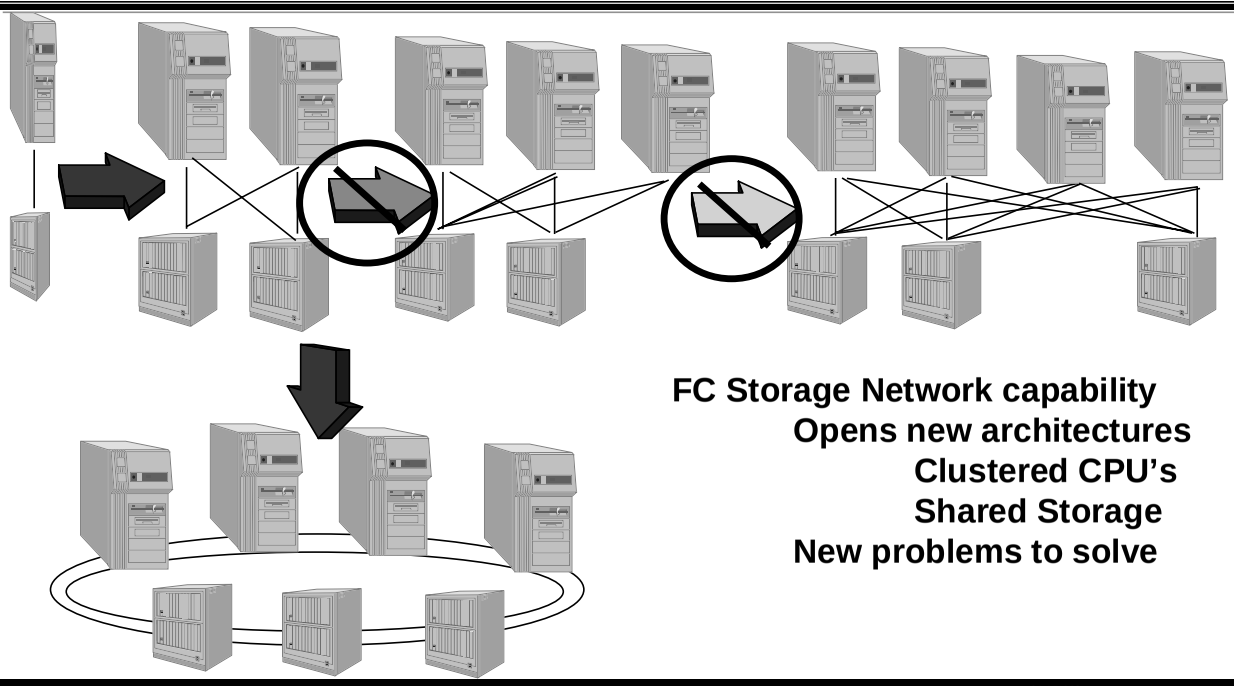
4:15 Petal: Distributed Virtual Disks, Ed Lee, DEC-SRC

5:00 Network Storage Manager, Greg VanHise, IBM

8:00 NASD Panel Discussion Including speakers plus Percy Tzelnic, EMC, Gene Freeman, Compaq, Dave Hitz, Network Appliance, Don Cameron, Intel, Jerry Fredin, Symbios, Ed Zayas, Novell, Kim Minuzzo, Lawrence Livermore National Lab

Slides available at www.nsic.org/nasd or www.snia.org

Scalable Clusters D. Anderson

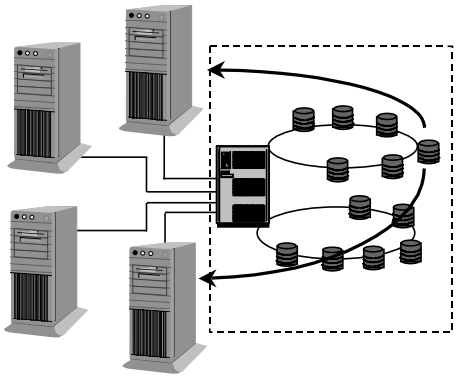


FC Storage Network capability
Opens new architectures
Clustered CPU's
Shared Storage
New problems to solve

Seagate
Information, the way you want it.

NAS Research Objectives: Find Solutions

- Scalable Computing
  - Need to share access to data
  - Heterogeneous computing
  - Dynamic scaling w/o interruption
  - Scalable resiliency & security
- Storage Management
  - Today more expensive that storage itself
  - Manual management proven impossible
  - Need more automated management
  - Goal is self managed storage
    - Scales with storage
    - Managed by policies & attributes



Seagate
Information, the way you want it.

Attribute-managed storage

Elizabeth Borowsky, John Wilkes, et al

Say what you want not how to do it!

RAID 3 data layout, across 5 of the disks on disk array F, using 64KB stripe size, 3MB dedicated buffer cache with 128KB sequential readahead buffer, delayed write-back with 1MB NVRAM buffer and max 10s residency time, dual 256Kb/s links via host interfaces 12.4.3 and 16.0.4, 1Gb/s trunk links between FibreChannel switches A-3 and B-1, ...

- business-critical availability
- 100 IOs/sec
- 200ms response time

Attribute-managed storage

The key

Attributes!

Workload Unit

Requirements:

capacity,
response time,
availability,
throughput...

Use patterns:

location pattern,
hot spots,
request rate...

Device

Capabilities:

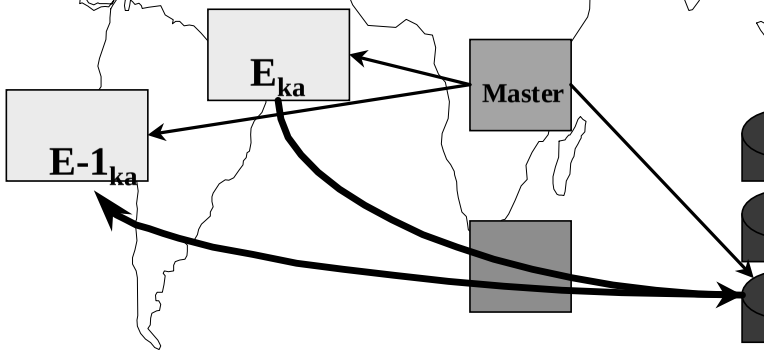
transfer rate,
positioning time,
capacity,
annualized failure rate...

Information Security for Network Attached Storage

James Hughes
StorageTek
hughes@network.com
http://www.network.com/~hughes

Secure File System (STK)

- Separate information access rights
  - from object access rights
- User has the keys
  - File system need not be trusted to protect Information
- NA*SD* "modification rights" model to protect object



Storage Object — Geoffrey Peck

- *Object* is an instance of a class
- *Class* is identified by a UUID
- Class supports *methods* (operations), identified by UUIDs
  - Note: method UUIDs are not unique to a single class – UUID for *read* operation is same across all classes
- Method invocation on a given object is permitted or denied based on the requestor's identity (*principal*)

Quantum

Summary

- Storage Objects are the next generation in storage systems
- Quantum and others are working on developing and standardizing this technology through NSIC and SNIA
- Storage product vendors should start to think about what they'll do with this technology

Quantum

ISI's NASD: Derived Virtual Devices

Rod Van Meter
rdv@isi.edu
Storage Networking Industry Association
March 5, 1998

The Netstation Project

- Replace I/O bus with a gigabit network
- Buses not scaling in:
  - # devices connected
  - aggregate bandwidth
  - distance

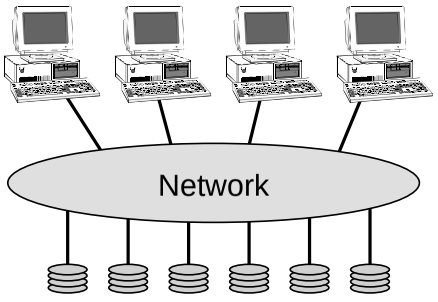
The Swarm Scalable Storage System

John H. Hartman

Department of Computer Science
The University of Arizona
jhh@cs.arizona.edu

Swarm

- Direct client/disk transfers
- Log-based striping
- RAID-style fault-tolerance
- Multiple access protocols
  - NFS, SIO, Swarm, HTTP

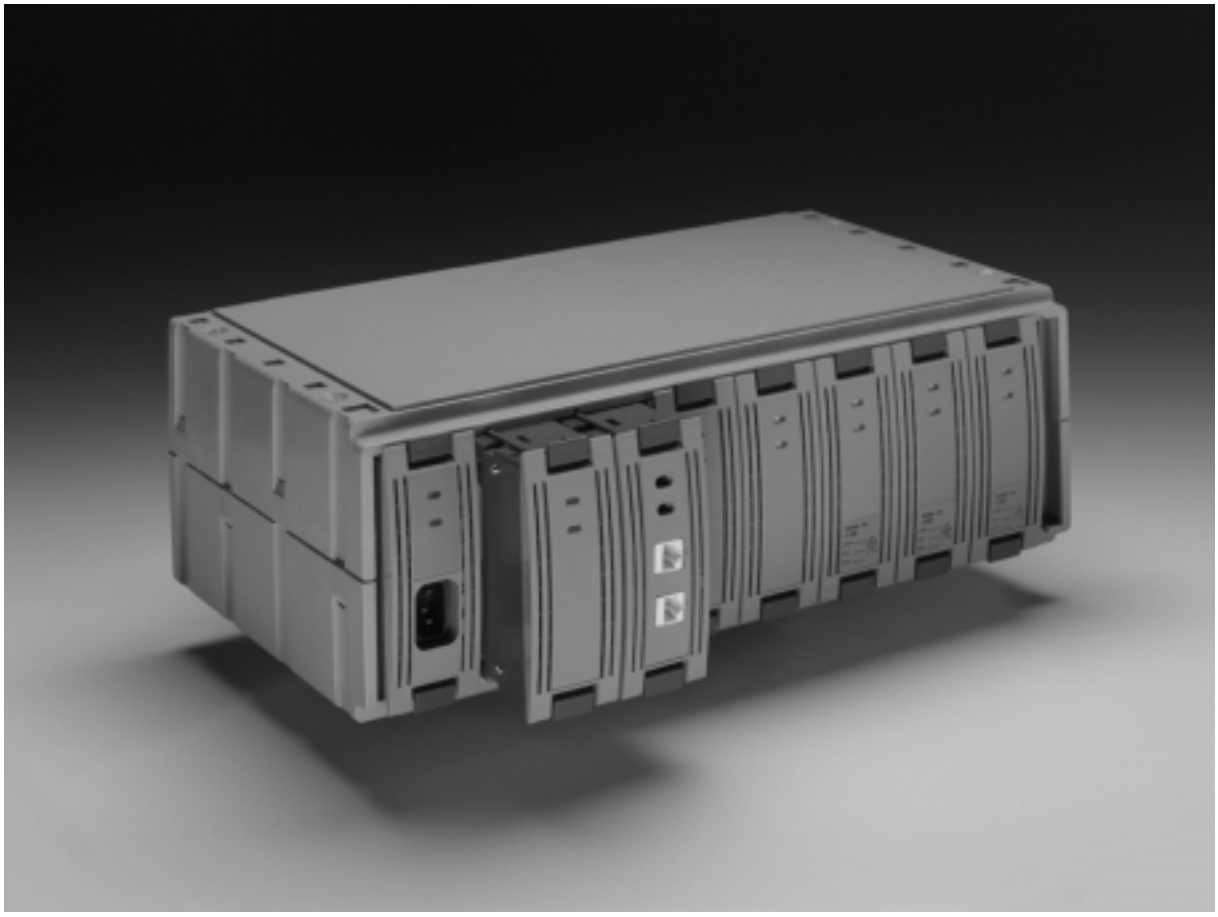


digital™

Petal: Distributed Virtual Disks

Systems Research Center
Digital Equipment Corporation

Edward K. Lee
Chandramohan A. Thekkath

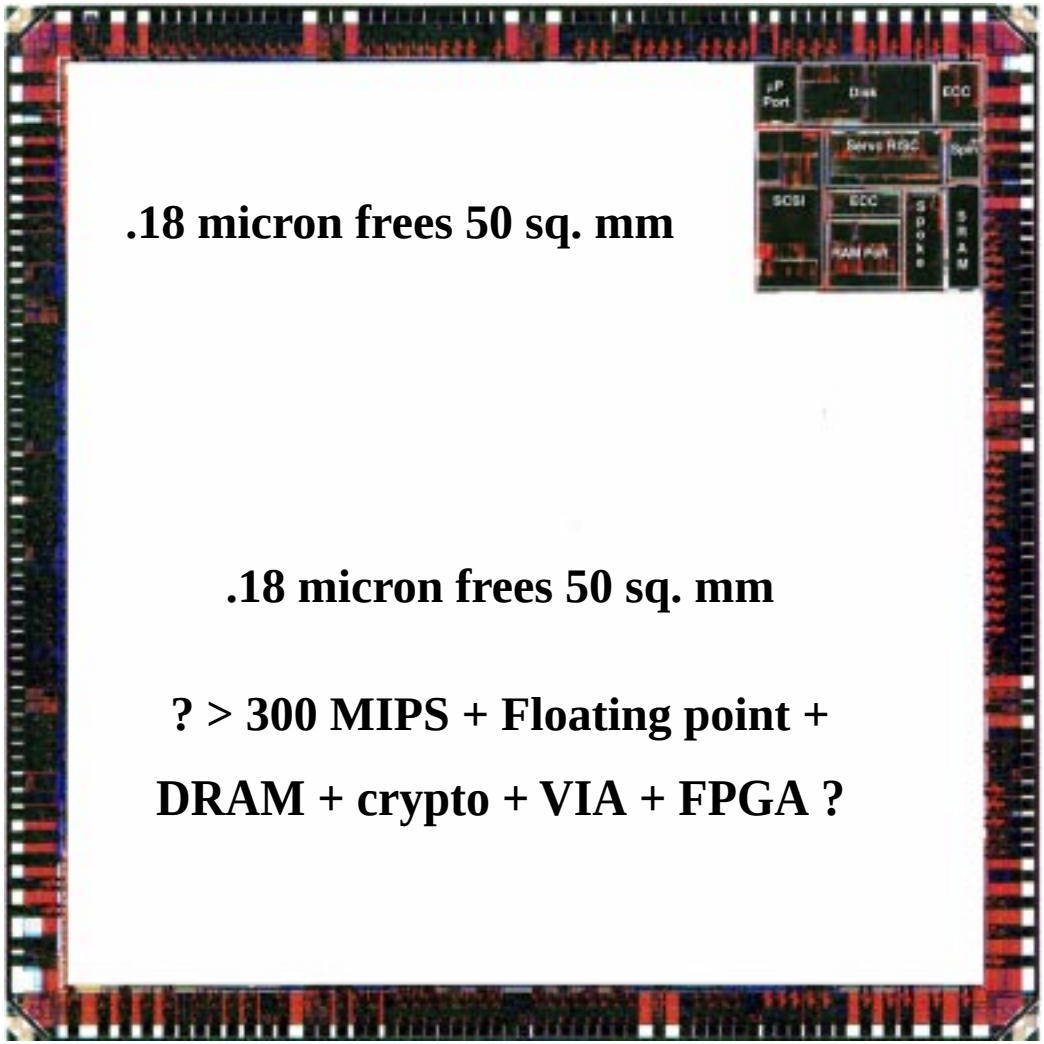


How Many Cycles Can Be Available?

VLSI trends continue!

What to do with it?

1) smaller chips lower cost?
2) on-chip track buffers lower cost?
3) support for system code (FS, DB)
4) application programmability?



Workshop Theme: What impact from this chip area?

Is it real? What does it cost? What can't it do?

- How is this different from a compute node + disk?

What things might it enable us to do?

- super optimized SCSI disks?
  file system in the disk? persistent object store?
  embedded database acceleration primitives?
  JAVA server? MPP I/O node?

What are research and pragmatic obstacles?

- recovery? programming model? programming environment?
  automatic functional partitioning? resource management?
  business model? business partnerships? deployment?
  secure & trustworthy? storage, security, rights management?
  cost-effective design? which network?

NSIC/NASD June 8-9 98 Meeting Agenda

Morning sessions: Application code in the disk

8:30 What to do with lots more computing inside storage?, Garth Gibson, CMU

9:00 Put EVERYTHING in the Storage Device, Jim Gray, Microsoft Research

9:35 Active Disks for Data Mining and Multimedia, Erik Riedel, CMU

10:25 Intelligent Disks: A New Computing Infrastructure for Decision Support Databases, Kimberly Keeton, UC Berkeley

11:00 Active Disk Architectures for Rapidly Growing Datasets, Anurag Acharya, UC Santa Barbara

11:35 Panel Discussion

Afternoon sessions: Storage and file systems support in the disk

1:45 Consideration for smarter storage device, David Anderson, Seagate

2:20 SCSI Disk Requirements for Shared Disk File Systems, Matthew O'Keefe, Univ of Minnesota

3:15 NFS v4 and Compound Requests, Brent Callaghan, Sun Microsystems

3:50 A File system for Intelligent Disks, Randy Wang, UC Berkeley

4:25 Panel Discussion

June 9 - groups construct white paper outlining opportunities & challenges

Plan for Tuesday June 9

Collaborate on a "Storage Computing" report

- community collaboration to "make it happen"
- arguments to funders, management, marketing

Propose a nine group approach to three topics:

1) Why: Storage technology are market trends
2) Why: Networking trends impacting storage
3) What: Storage Management
4) What: File Systems
5) What: Database and new applications
6) How: Robust correctness - security, reliability?
7) How: Robust performance - resource management?
8) How: Computational model - how to program storage?
9) How: Business model - how to deploy changes?