

Storage Architecture Research at Intel

Rumi Zahir, Don Cameron, Mike Mesnier, John Litvin
rumi.zahir@intel.com

Storage Architecture Group
Enterprise Architecture Lab
Intel Corporation

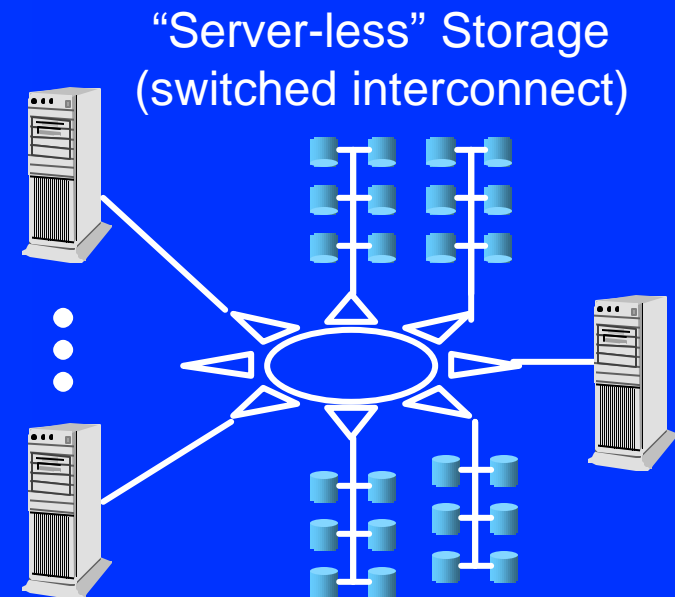
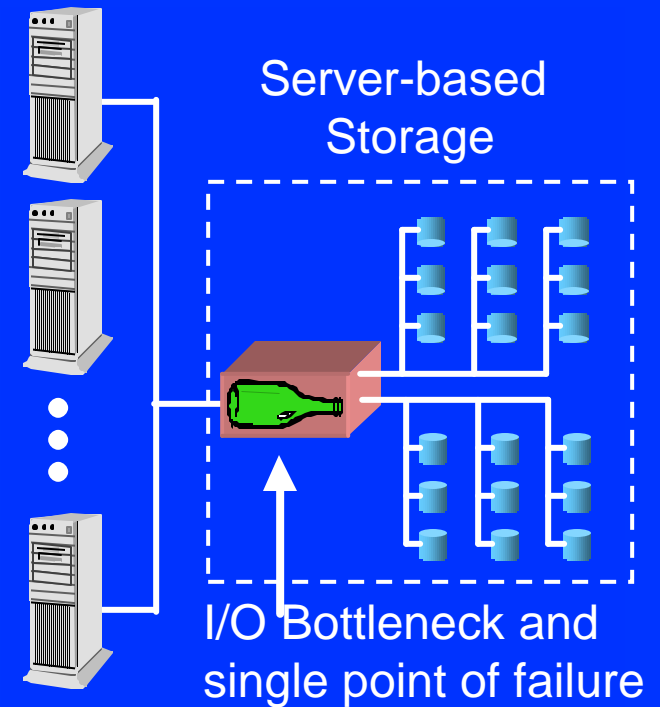
October 2000

Agenda

- Enterprise Storage
- Research Goals
- Hardware Configuration
- Projects
 - Scalable Native Disks, iSCSI, DAFS, and...
 - Self-Managed Storage (new project)
- Related Cool Work

Enterprise Storage

- **Many small servers**
 - Management nightmare
- **Management issues drive storage consolidation**
 - Ⓜ **fewer big servers**
 - Server scaling I/O limited
 - Availability issues
- **Server-less Storage**
 - LAN-attached storage at lower-end (aka NAS)
 - SAN-attached storage at higher-end



Research Goals

- **Server-based vs. Server-less:** compare configurations in terms of performance, CPU compute requirements and cost
- **Software Stacks:** comprehend & validate scalability and performance of various stacks, e.g. VI-architecture vs. TCP/IP, centralized NFS block-based vs. distributed NASD object-based
- **Storage Management:** develop & evaluate novel storage management techniques that take advantage of intelligence on the disk

Hardware Configuration



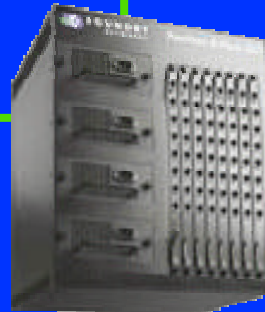
4 Application Servers:
2x500 MHz Pentium® III
Processor, 128 MB RAM,
Intel® Pro 1000 Enet,
Red Hat* Linux* 6.1



Metadata Server:
266 MHz Pentium® II
Processor, 32 MB RAM,
Intel® Pro 100 Enet,
Red Hat* Linux* 6.1



64 Native Disks:
233 MHz SA-110,
32MB RAM,
1 13 GB IDE Disk,
100 Mb Enet,
Linux* 2.2.7

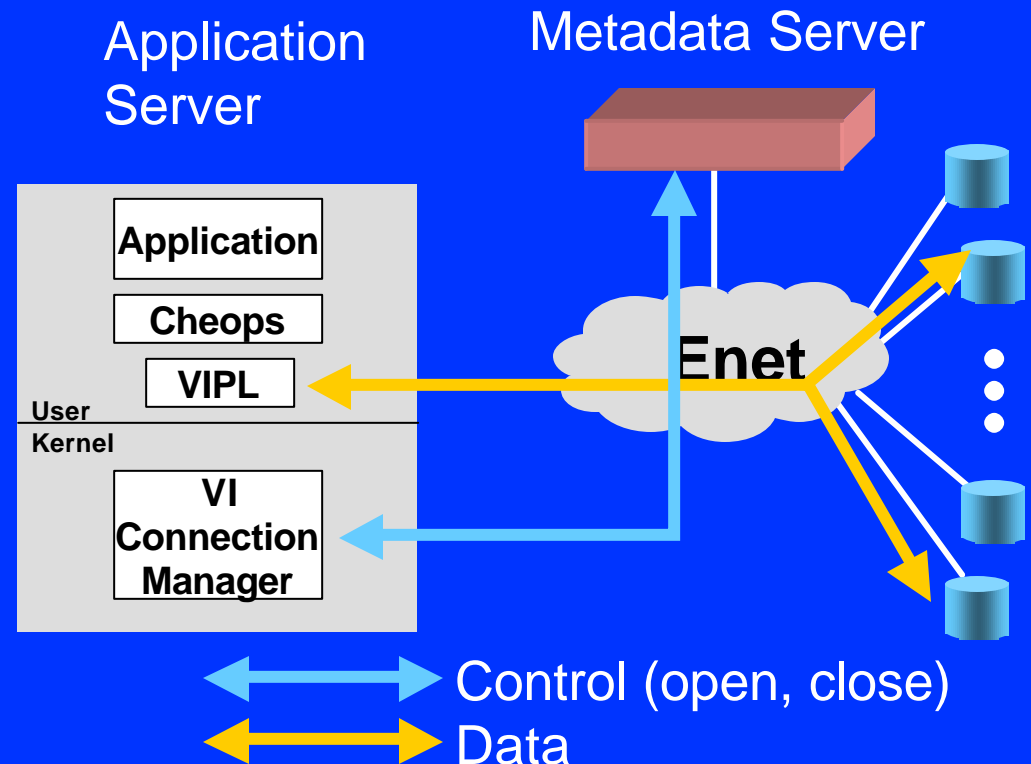


Ethernet Switch:
Foundry Networks* FastIron* II+

*All other names and brands are the property of their respective owners.

Scalable Native Disks Project

- Object based storage model
 - CMU NASD code
 - NASD over TCP/IP
 - NASD over virtual interface arch. (M-VIA from LBNL) [user-level networking]
- Demonstrated
 - Linear data bandwidth scaling
 - CPU utilization of VI <50% of TCP/IP



Details at:

<http://developer.intel.com/design/ldf/spr2000/presentations/SERS04PS.htm>

On-going Work

- iSCSI prototype in development
 - Encapsulates SCSI commands in TCP/IP
 - IETF proposal
 - <http://www.ietf.org/internet-drafts/draft-satran-iscsi-01.txt>
- DAFS (Direct Access File System) in development
 - Implement NFS using client-side VI-A
 - <http://www.dafscollaborative.org>
- Self-managed storage prototype (new project)
 - Use intelligence on storage node for **continuous and pro-active** backup, versioning, dynamic load balancing, and
 - Peer-to-peer communication between storage nodes
 - Needs
 - Unified namespace crucial: clustered/distributed file system, multi-node LVM, or Intermezzo style replication
 - On-line add, remove, resize, migrate

Self-Managed Storage

- **Continuous Backup/Versioning**
 - Goal
 - Recover from device failures **and end-user errors** (accidental deletions)
 - Automatic and deliberate **replication** of data to other storage nodes (to recover from device failures)
 - Automatic **versioning** of data (to recover from user errors)
[Continuous snapshotting, allows ALL versions to be reconstructed]
- **Automatic Load Balancing**
 - In a non-uniform-latency network **transparently migrate data closer to consumer** (migration is between storage nodes)
- **No need for synchronous operation!**
 - Instead, a **pro-active continuous background service** that utilizes otherwise unused compute, network and storage bandwidth.
 - **Weak consistency model OK, use branches for conflicts.**

Related Cool Work

- Self-Managed / Adaptive / Replicating Storage
 - Oceanstore [Kubi, ASPLOS, Nov'00]
 - Intermezzo [Braam, Linux Storage Workshop, Oct'00]
 - Farsite [Boloski, Sigmetrics, Jun'00]
 - Freenet [<http://freenet.sourceforge.net>]
 - Gnutella [<http://www.gnutella.wego.com>]
- Clustered/Distributed File Systems
 - GFS [Preslan, Mass Storage Systems Conf., Mar'99]
 - Others: Frangipani'97, Serverless File System'95, Coda'90
- Versioning
 - Elephant File System [Feely, SOSP, Dec'99]
 - Restore-o-Mounter [Moran, Usenix, Jun'93]

**Interested ? Talk to me, or
E-mail me at <rumi.zahir@intel.com>**