Storage Aggregation for Performance & Availability:

The Path from Physical RAID to Virtual Objects

Garth Gibson

Co-Founder & CTO, Panasas Inc.
Assoc. Professor, Carnegie Mellon University

panasas

Changing Computational Architecture

Monolithic Supercomputers

Linux Clusters

- Specialized, but expensive
- Price/performance: often > **$100M/TFLOPS**

- Powerful, scalable, affordable
- Price/performance: often < **$1M/TFLOPS**

Clusters dominating Top500 Supercomputers:	
1998:	**2**
2002:	**94**
2004:	**294**

Source: Top500.org

Matching to Storage Architecture

Traditional Computing

Monolithic Computers

Issues

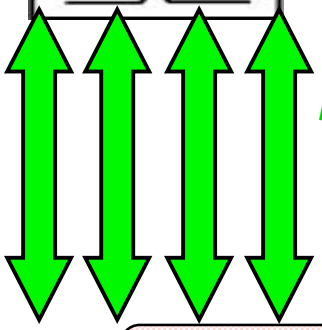
- Complex Scaling
- Limited Bandwidth
- I/O Bottleneck
- Inflexible
- Expensive

Single data path

Monolithic Storage

Cluster Computing

Linux Compute Cluster



Parallel data paths

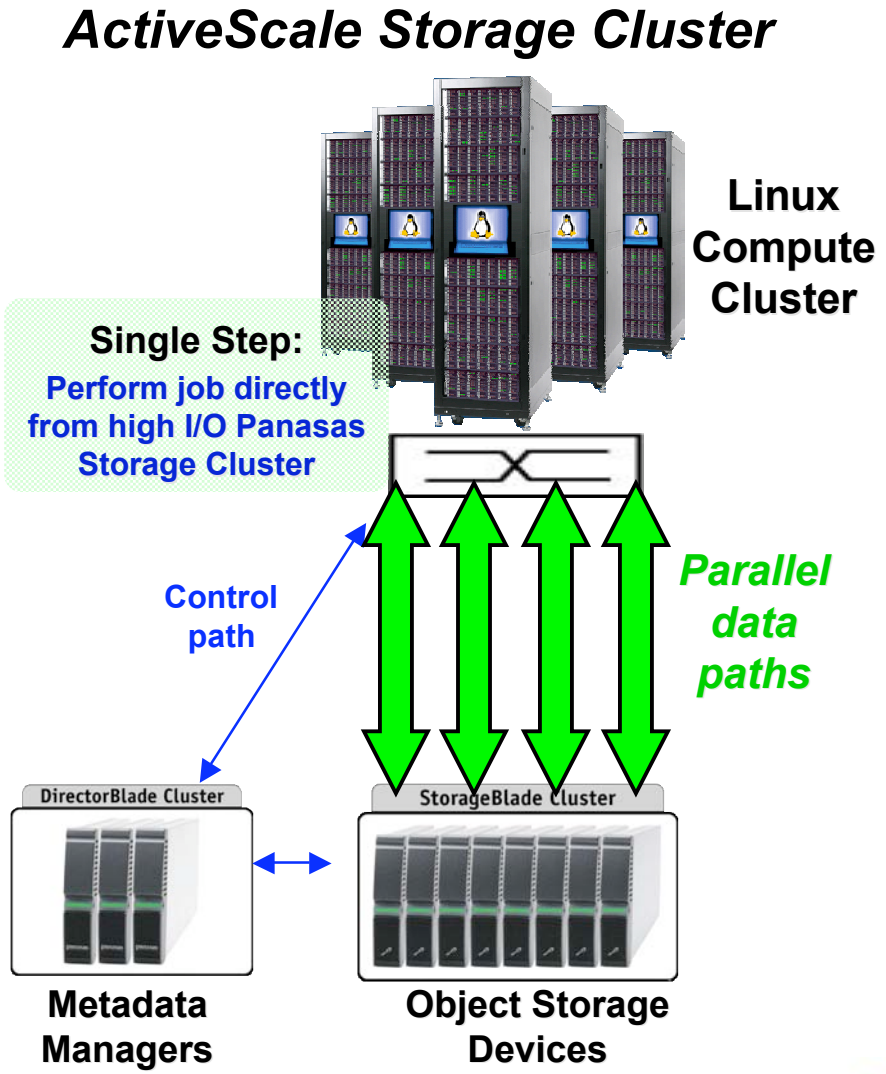
Scale to bigger box?

Scale:
- file & total bandwidth
- file & total capacity
- load & capacity balancing

But lower $ / Gbps

Next Generation Cluster Storage

- **Scalable performance**
  - Offloaded data path enable direct disk to client access
  - Scale clients, network and capacity
  - As capacity grows, performance grows
- **Simplified and dynamic management**
  - Robust, shared file access by many clients
  - Seamless growth within single namespace eliminates time-consuming admin tasks
- **Integrated HW/SW solution**
  - Optimizes performance and manageability
  - Ease of integration and support

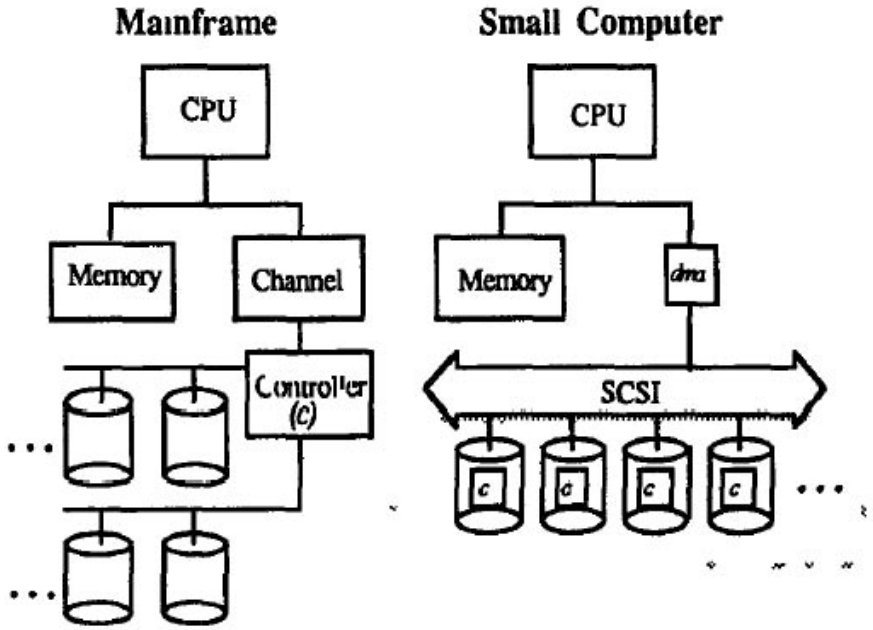


Redundant Arrays of Inexpensive Disks (RAID)

November 24, 2004

Birth of RAID (1986-1991)

- **Member of 4th Berkeley RISC CPU design team (SPUR: 84-89)**
  - Dave Patterson decides CPU design is a “solved” problem
  - Sends me to figure out how storage plays in SYSTEM PERFORMANCE
- **IBM 3380 disk is 4 arms in a 7.5 GB washing machine box**
  - SLED: Single Large Expensive Disk
- **New PC industry demands cost effective 100 MB 3.5” disks**
  - Enabled by new SCSI embedded controller architecture
- **Use many PC disks for parallelism SIGMOD88: A case for RAID**
- **PS. $10-20 per MB (~1000X now)**
  100 MB/arm (~1000X now)
  20-30 IO/sec/arm (5X now)



But RAID is really about Availability

- **Arrays have more Hard Disk Assemblies (HDAs) -- more failures**
  - Apply replication and/or error/erasure detection codes

Block	Disk 0	Disk 1	Disk 2	Disk 3
0	D0	D0	D1	D1
1	D2	D2	D3	D3
2	D4	D4	D5	D5
3	D6	D6	D7	D7
4	D8	D8	D9	D9
5	D10	D10	D11	D11

Mirroring

Block	Disk 0	Disk 1	Disk 2	Disk 3
0	D0	D1	D2	0-2
1	D4	D5	3-5	D3
2	D8	6-8	D6	D7
3	9-11	D9	D10	D11

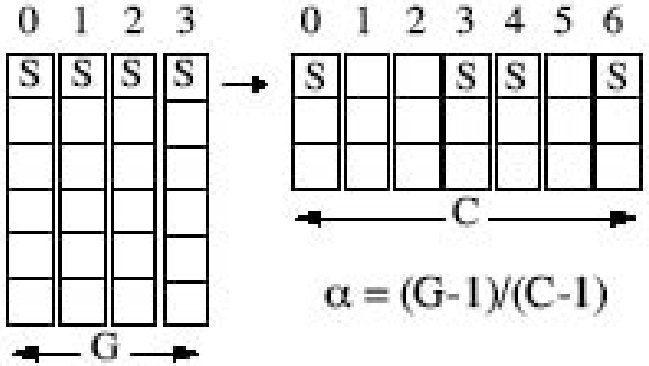
RAID level 5

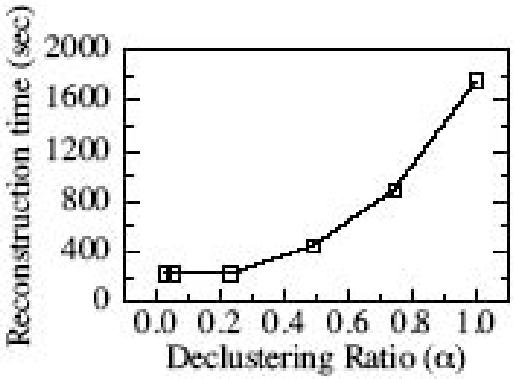
$$MTTF_{RAID} = \frac{(MTTF_{Disk})^2}{(D+C*n_G)*(G+C-1)*MTTR}$$

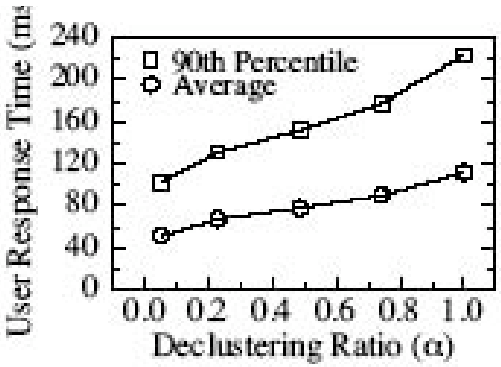
- **Mirroring wastes 50% space; RAID wastes 1/N**
- **Mirroring halves, RAID 5 quarters small write bandwidth**

Off to CMU & More Availability

- **Parity Declustering "spreads RAID groups" to reduce MTTR**
  - Each parity disk block protects fewer than all data disk blocks (C)
  - Virtualizing RAID group lessens recovery work
    - Faster recovery or better user response time during recovery or mixture of both
- **RAID over X?**
  - X = Independent fault domains
  - "Disk" is easiest "X"
  - Parity declustering is my first step in RAID virtualization





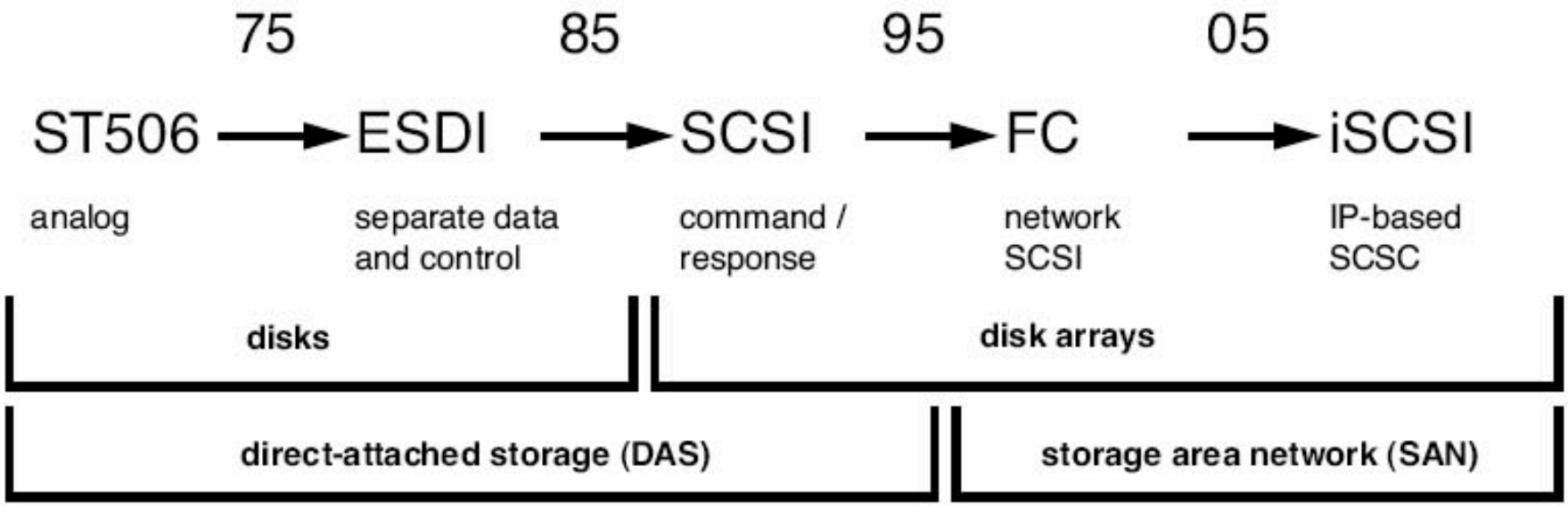


Network-Attached Secure Disks (NASD, 95-99)

November 24, 2004

panasas

Storage Interconnect Evolution

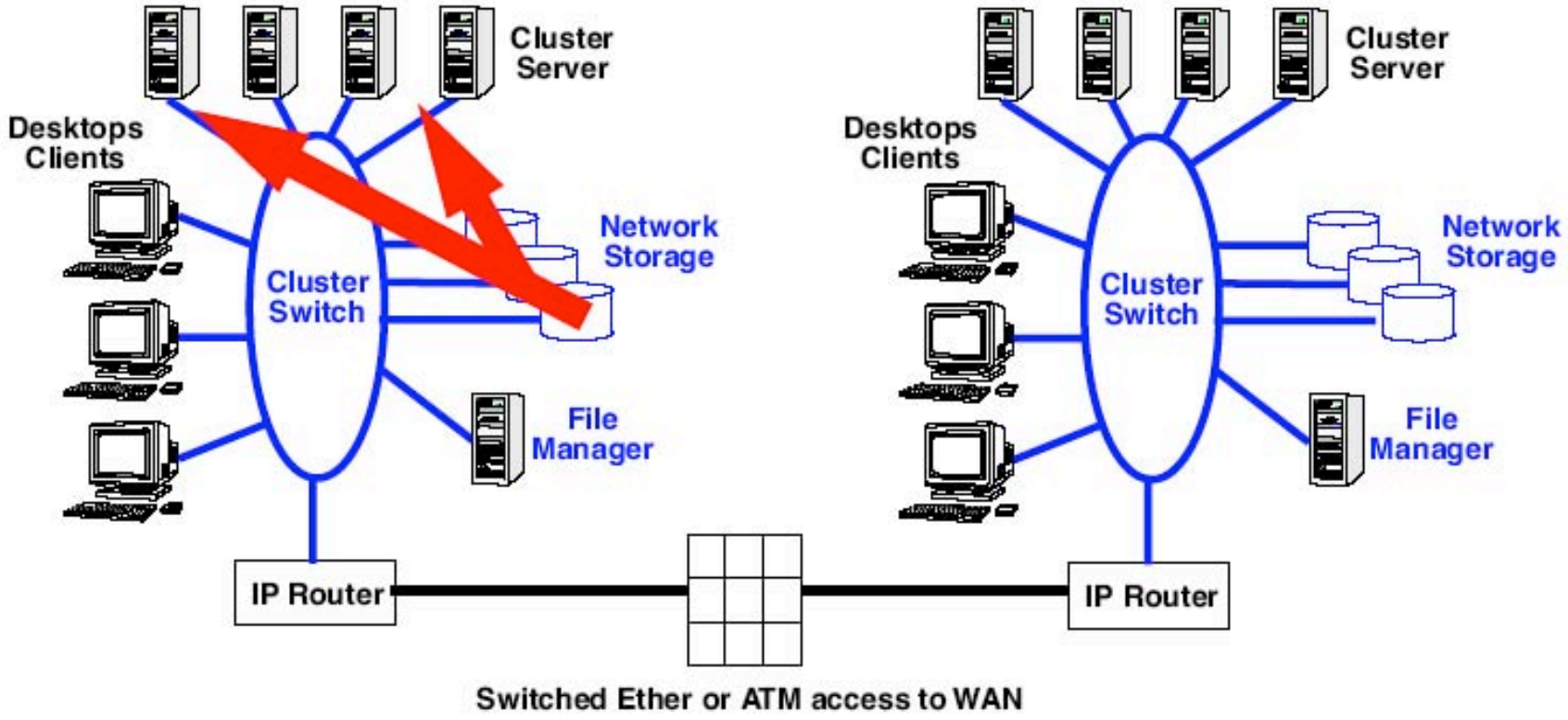


- **Outboard circuitry increases over time (VLSI density)**
- **Hardware (#hosts, #disks, #paths) sharing increases over time**
- **Logical (information) sharing limited by host SW**
- **1995: Fibrechannel packetizes SCSI over a near general network**

panasas

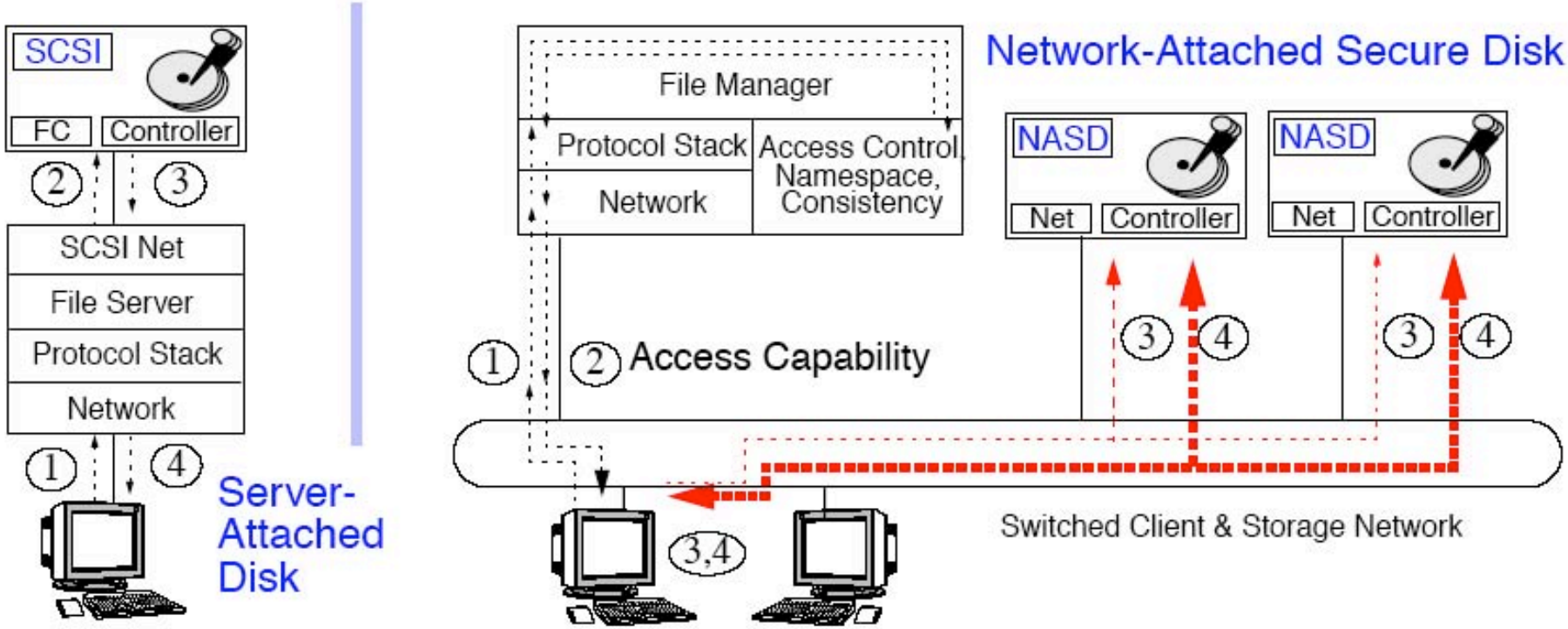
Storage as First Class Network Component

- **Direct transfer between client and storage**
  - Exploit scalable switched cluster area networking
  - Split file service into: primitives (in drive) and policies (in manager)



NASD Architecture

- **Before NASD there was store&forward Server-Attached Disks (SAD)**
- **Move access control, consistency out-of-band and cache decisions**
- **Raise storage abstraction: encapsulate layout, offload data access**



panasas

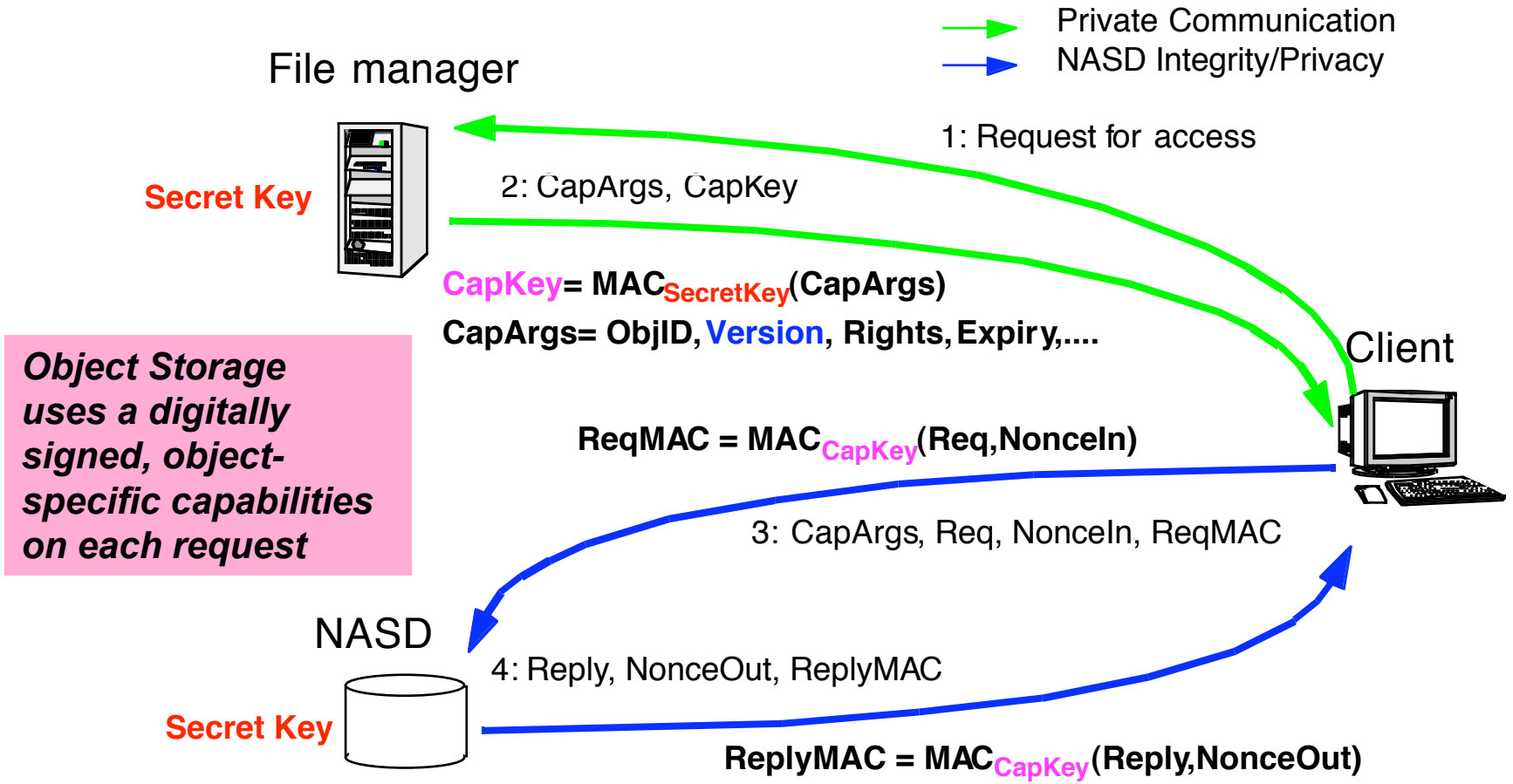
Metadata Performance

- **Command processing of most operations in storage could offload 90% of small file/productivity workload from servers**
  - Key inband attribute updates: size, timestamps etc

NFS Operation	Count in top 2% by work (K)	File Server (SAD)		DMA (NetSCSI)		Object (NASD)	
		Cycles (B)	% of SAD	Cycles (B)	% of SAD	Cycles (B)	% of SAD
Attr Read	792.7	26.4	11.8	26.4	11.8	0.0	0.0
Attr Write	10.0	0.6	0.3	0.6	0.3	0.6	0.3
Data Read	803.2	70.4	31.6	26.8	12.0	0.0	0.0
Data Write	228.4	43.2	19.4	7.6	3.4	0.0	0.0
Dir Read	1577.2	79.1	35.5	79.1	35.5	0.0	0.0
Dir RW	28.7	2.3	1.0	2.3	1.0	2.3	1.0
Delete Write	7.0	0.9	0.4	0.9	0.4	0.9	0.4
Open	95.2	0.0	0.0	0.0	0.0	12.2	5.5
Total	**3542.4**	**223.1**	**100**	**143.9**	**64.5**	**16.1**	**7.2**

Fine Grain Access Enforcement

- **State of art is VPN of all out-of-band clients, all sharable data and metadata**
  - Accident prone & vulnerable to subverted client; analogy to single-address space computing

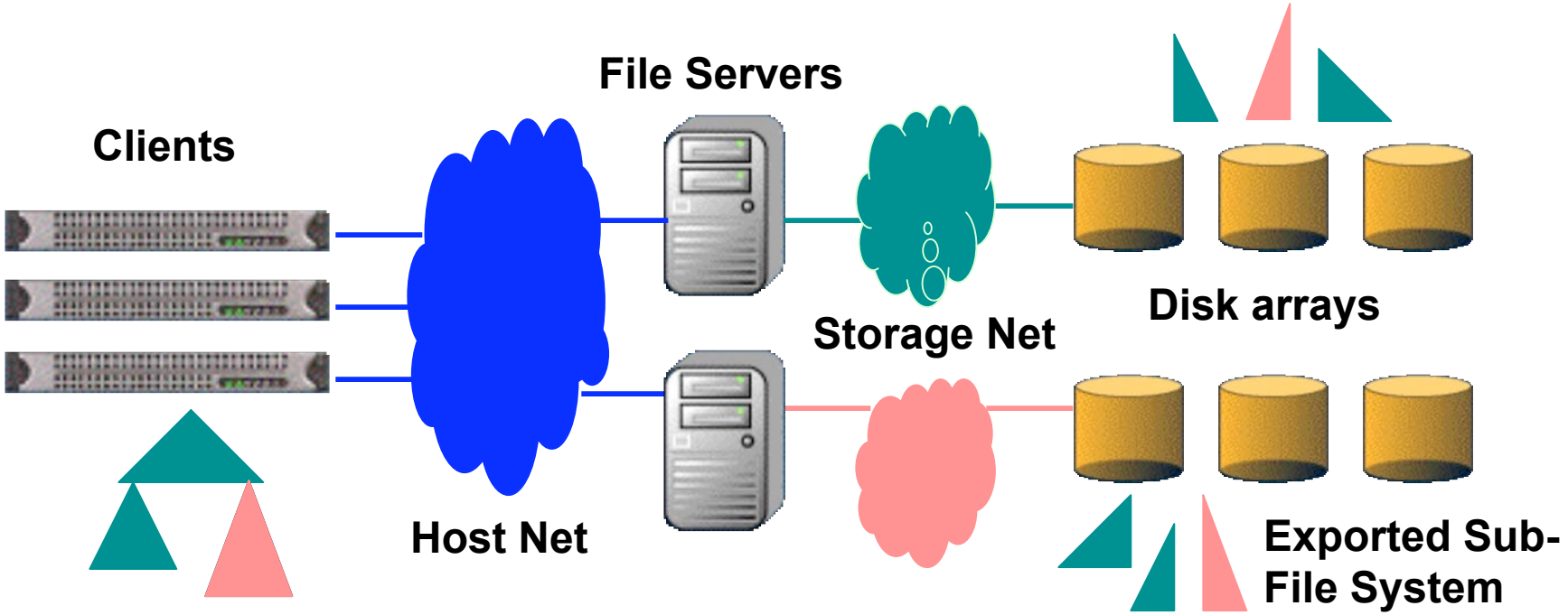


Scalable File System Taxonomy

November 24, 2004

Today's Ubiquitous NFS

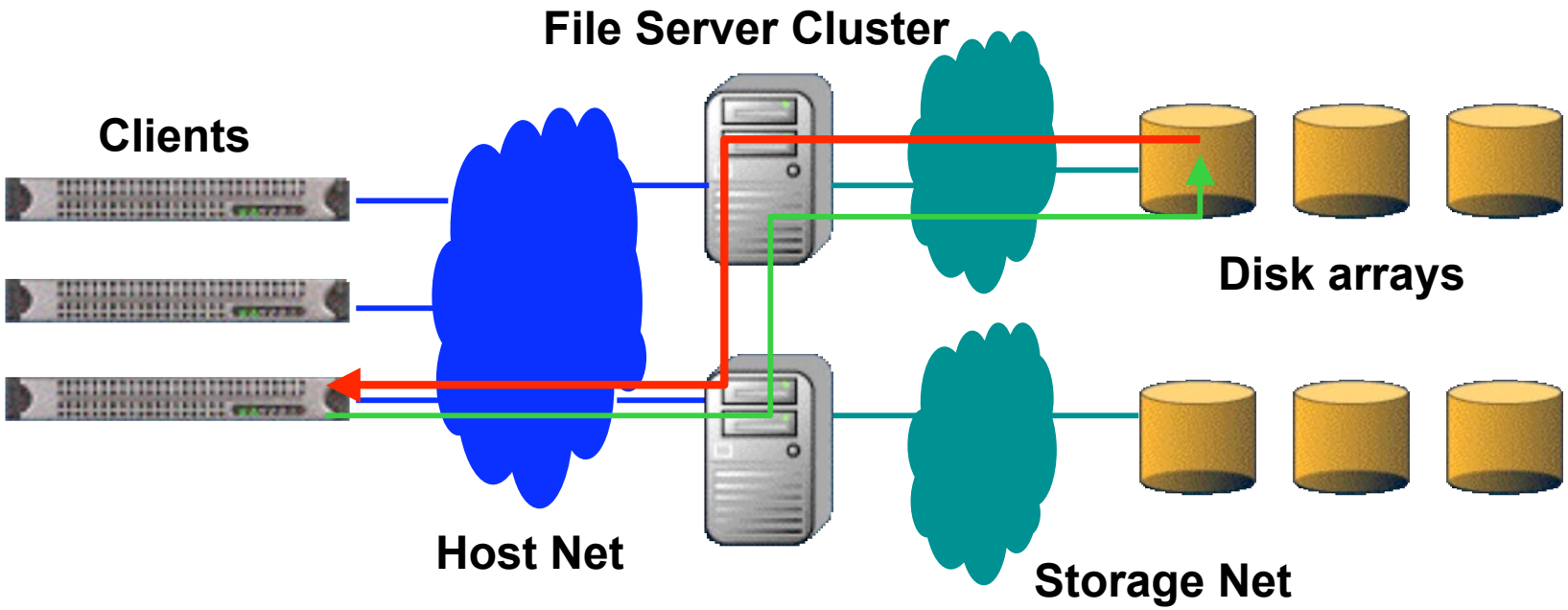
- **ADVANTAGES**
  - Familiar, stable & reliable
  - Widely supported by vendors
  - Competitive market
- **DISADVANTAGES**
  - Capacity doesn't scale
  - Bandwidth doesn't scale
  - Cluster by customer-exposed namespace partitioning



panasas

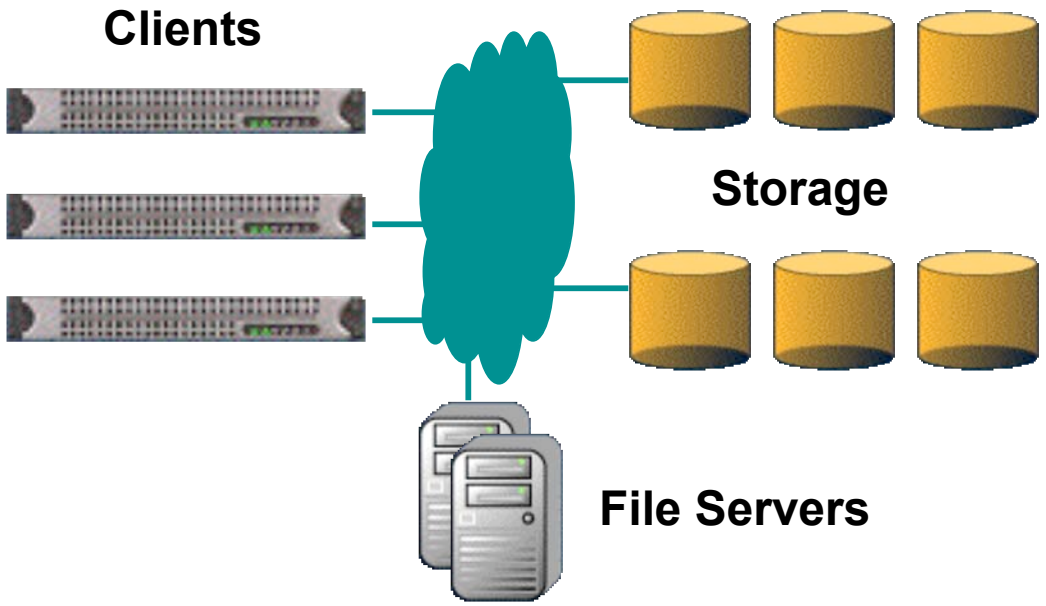
Scale Out w/ Forwarding Servers

- **Bind many file servers into single system image with forwarding**
  - Mount point binding less relevant, allows DNS-style balancing, more manageable
  - Control and data traverse mount point path (in band) passing through two servers
  - Single file and single file system bandwidth limited by backend server & storage
  - Tricord, Spinnaker



Scale Out FS w/ Out-of-Band

- **Client sees many storage addresses, accesses in parallel**
  - Zero file servers in data path allows high bandwidth thru scalable networking
  - E.g.: IBM SanFS, EMC HighRoad, SGI CXFS, Panasas, Lustre, etc
  - Mostly built on block-based SANs where ***servers trust all clients***



Object Storage Standards

November 24, 2004

panasas

Object Storage Architecture

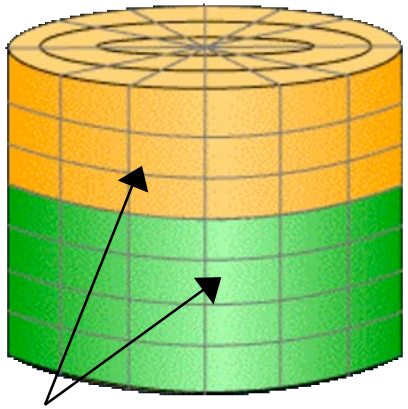
- **An evolutionary improvement to standard SCSI storage interface (OSD)**
- **Offload most data path work from server to intelligent storage**
- **Finer granularity of security: protect & manage one file at a time**
- **Raises level of abstraction: Object is container for "related" data**
  - Storage *understands* how different blocks of a "file" are related -> self-management
  - Per Object Extensible Attributes is key expansion of functionality

Block Based Disk

Operations:
Read block
Write block

Addressing:
Block range

Allocation:
External



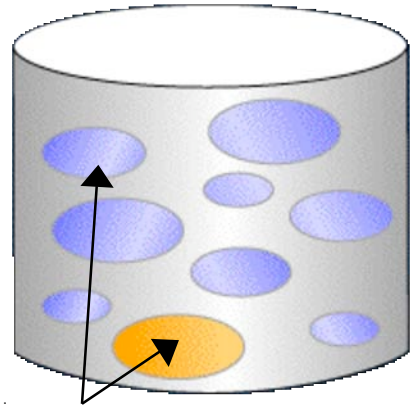
Security At Volume Level

Object Based Disk

Operations:
Create object
Delete object
Read object
Write object

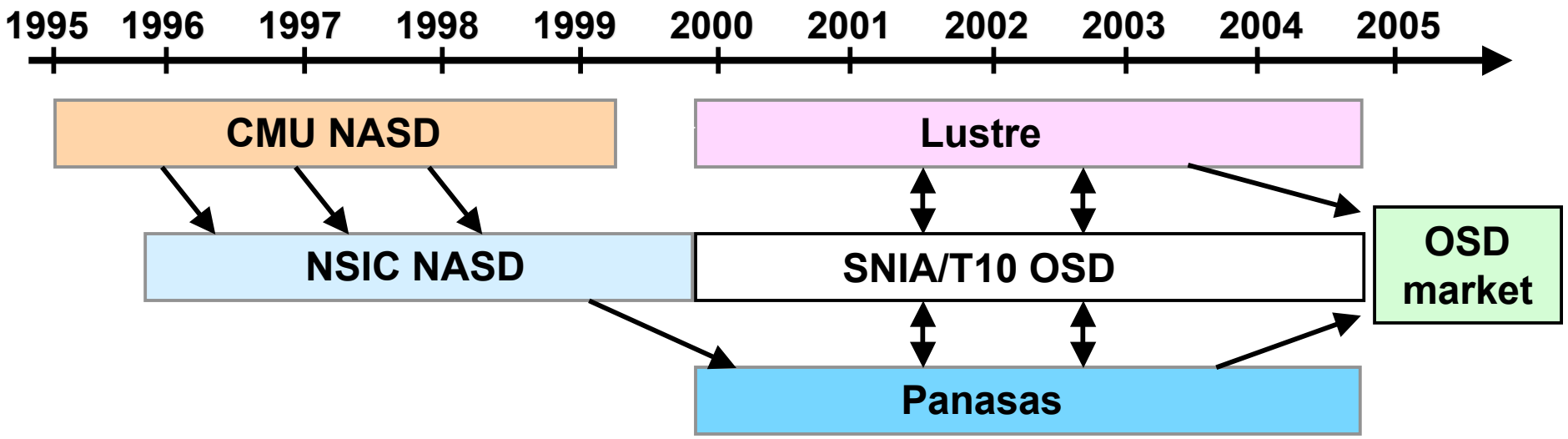
Addressing:
[object, byte range]

Allocation:
Internal



Security At Object Level

OSD is now an ANSI Standard



- **INCITS ratified T10's OSD v1.0 SCSI command set standard, ANSI will publish**
  - Co-chaired by IBM and Seagate, protocol is a general framework (transport independent)
  - Sub-committee leadership includes IBM, Seagate, Panasas, HP, Veritas, ENDL
  - Product plans from HP/Lustre & Panasas; research projects at IBM, Seagate
  - www.snia.org/tech_activities/workgroups/osd & www.t10.org/ftp/t10/drafts/osd/osd-r10.pdf

ActiveScale Storage Cluster

November 24, 2004

Object Storage Systems

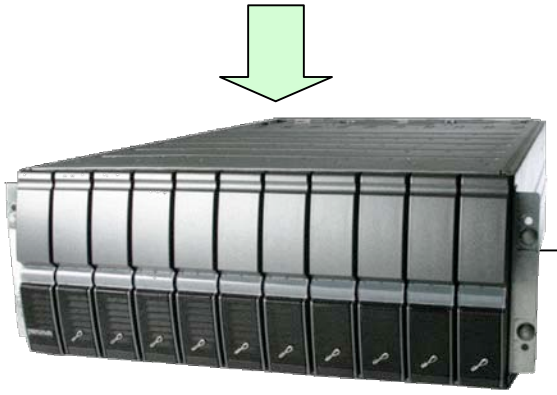
Expect wide variety of Object Storage Devices

- Disk array subsystem
- Ie. LLNL with Lustre

- "Smart" disk for objects
- 2 SATA disks – 240/500 GB

- Prototype Seagate OSD
- Highly integrated, single disk

- Orchestrates system activity
- Balances objects across OSDs



- **Stores up to 5 TBs per shelf**

16-Port GE Switch Blade

- 4 Gbps per shelf to cluster

Scalable Storage Cluster Architecture

Lesson of compute clusters: Scale out commodity components

- **Blade server approach provides**
  - High volumetric density, disk array abstraction
  - Incremental growth, pay-as-you-grow model
  - Needs single system image SW architecture

StorageBlade
2 SATA spindles

Shelf of Blades
5 TB, 4 Gbps

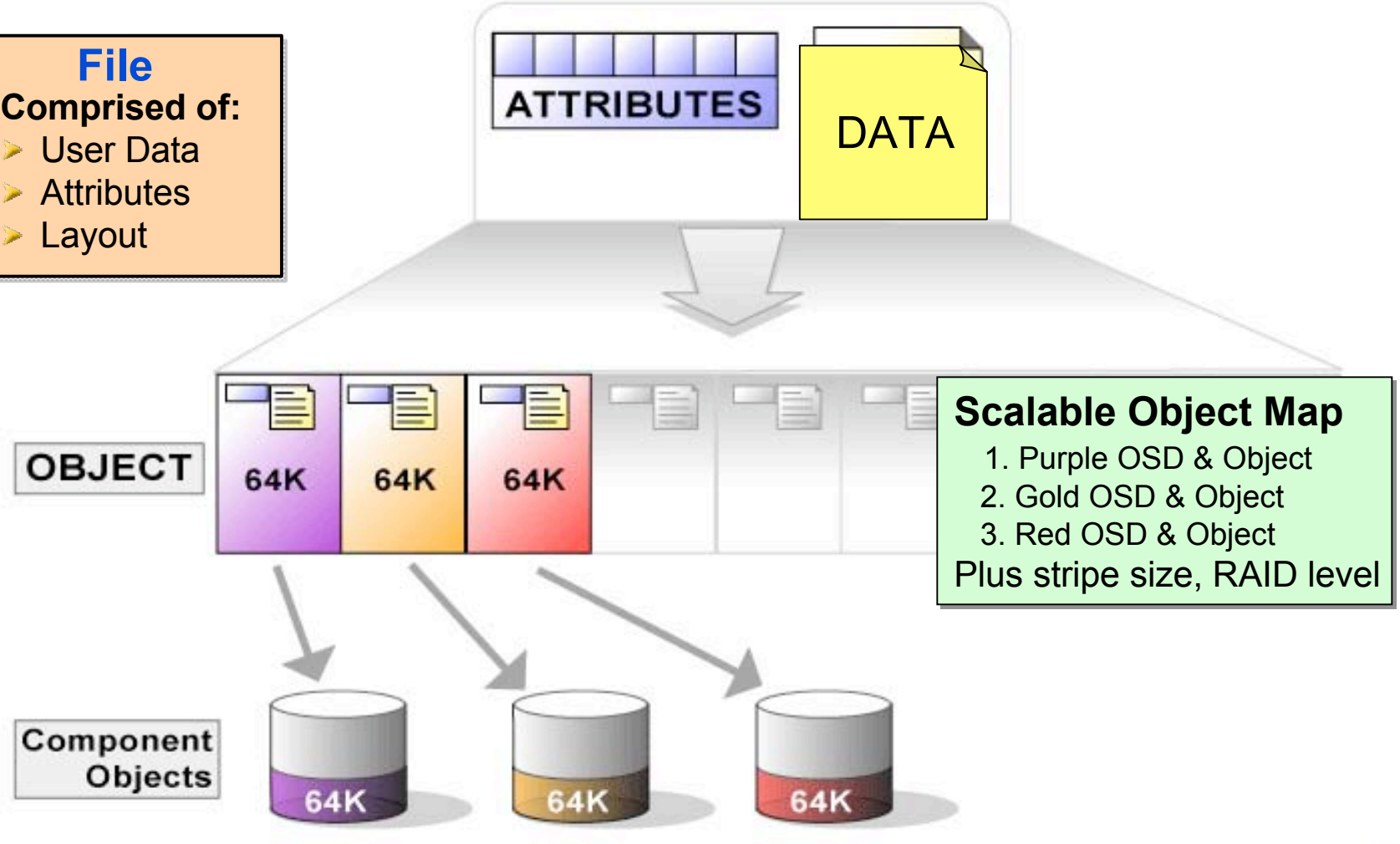
Single System Image
55 TB, 44 Gbps per rack

panasas

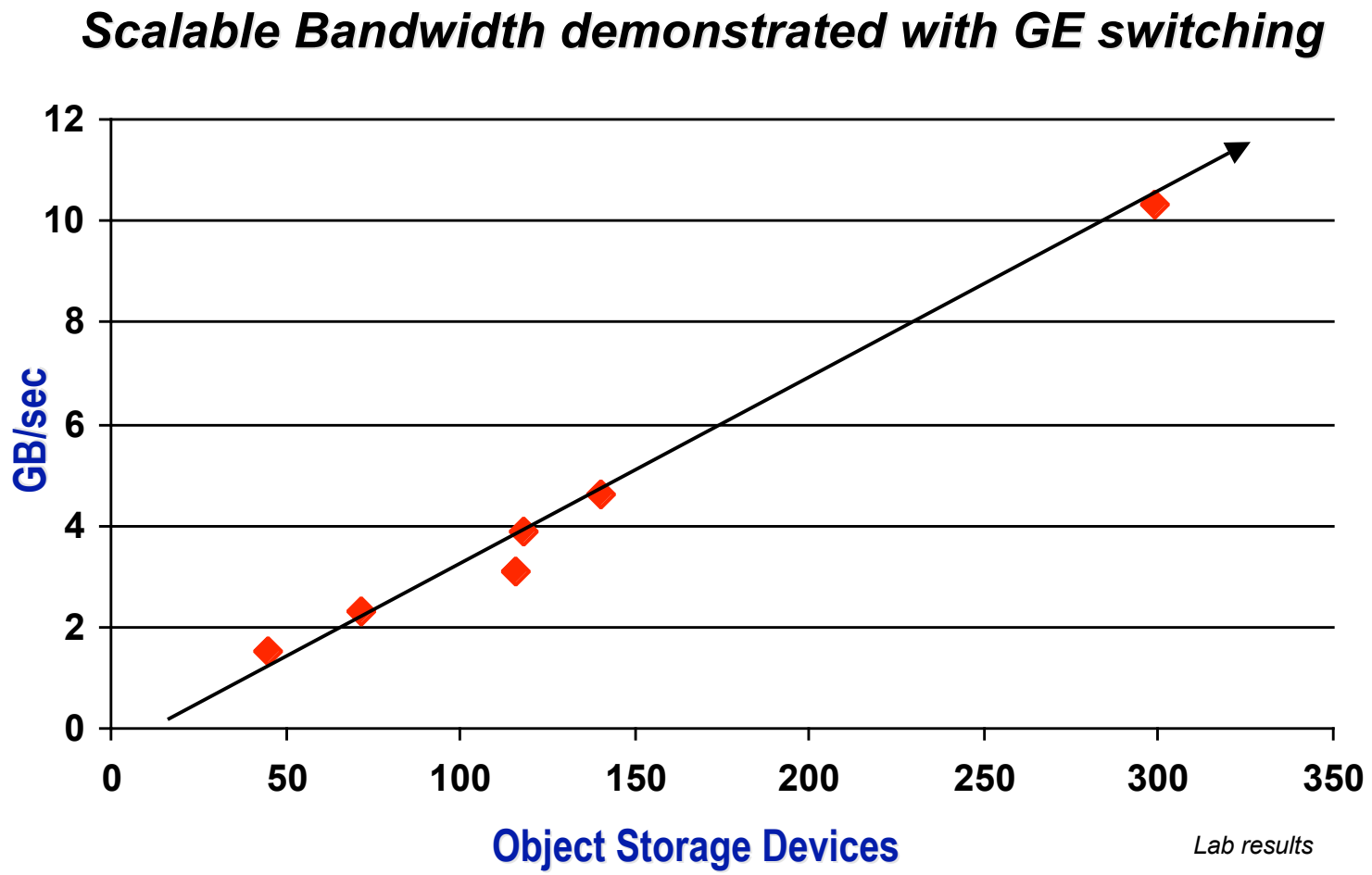
Virtual Objects are Scalable

Scale capacity, bandwidth, reliability by striping according to small map

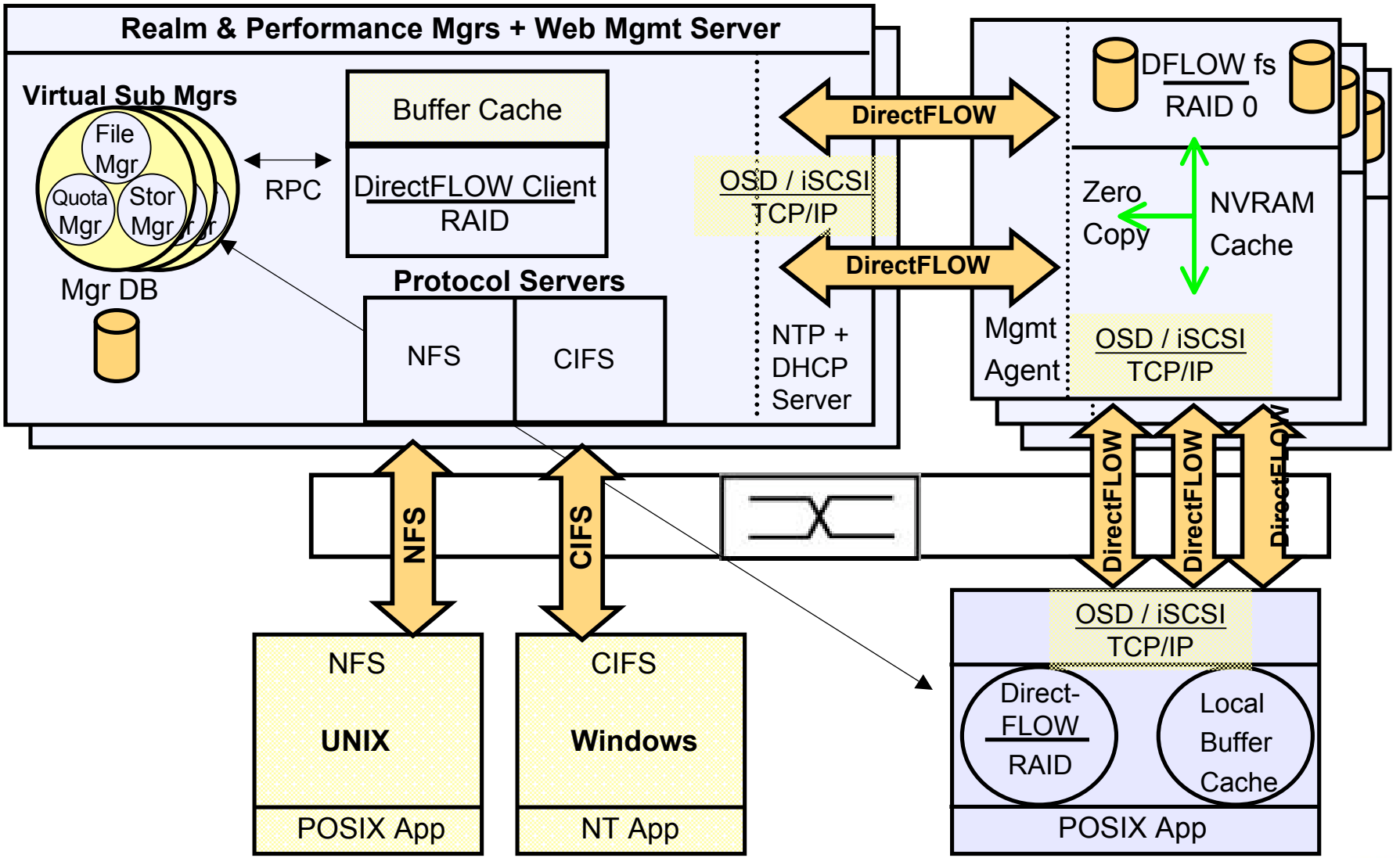
File
Comprised of:
- User Data
- Attributes
- Layout



Object Storage Bandwidth



ActiveScale SW Architecture



panasas

Fault Tolerance

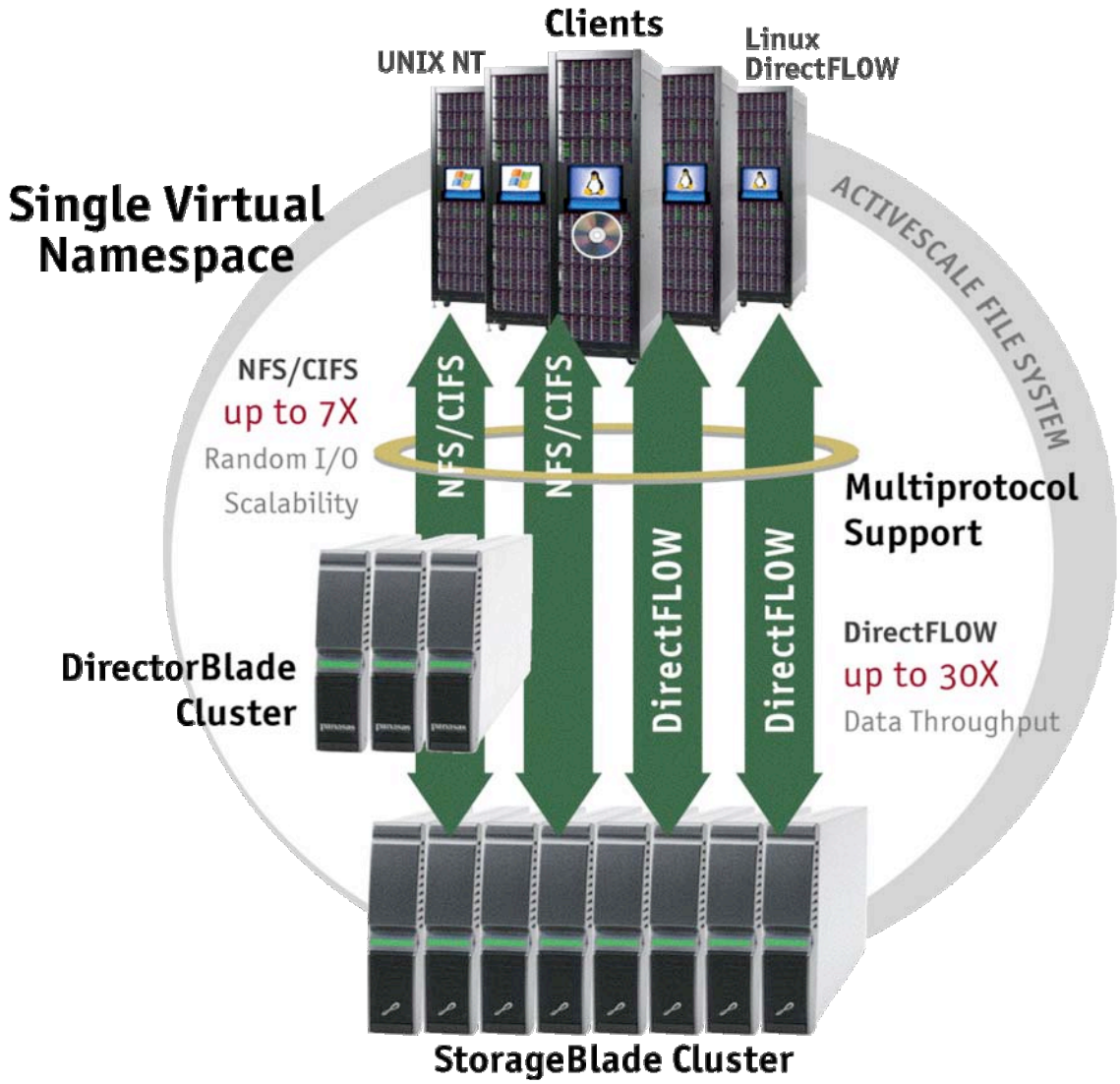
- **Overall up/down state of blades**
  - Subset of managers track overall state with heartbeats
  - Maintain identical state with quorum/consensus
- **Per file RAID: no parity for unused capacity**
  - RAID level per file; small files mirror; RAID5 for large files
  - First step toward policy quality of storage associated w/ data
- **Client based RAID: do XOR where all data sits in memory**
  - Traditional RAID stripes have data of multiple files & metadata
  - Per file RAID covers only data of one file
  - Client computed RAID risks only data client can trash anyway
  - Client memory is most efficient place to compute XOR

panasas

Manageable Storage Clusters

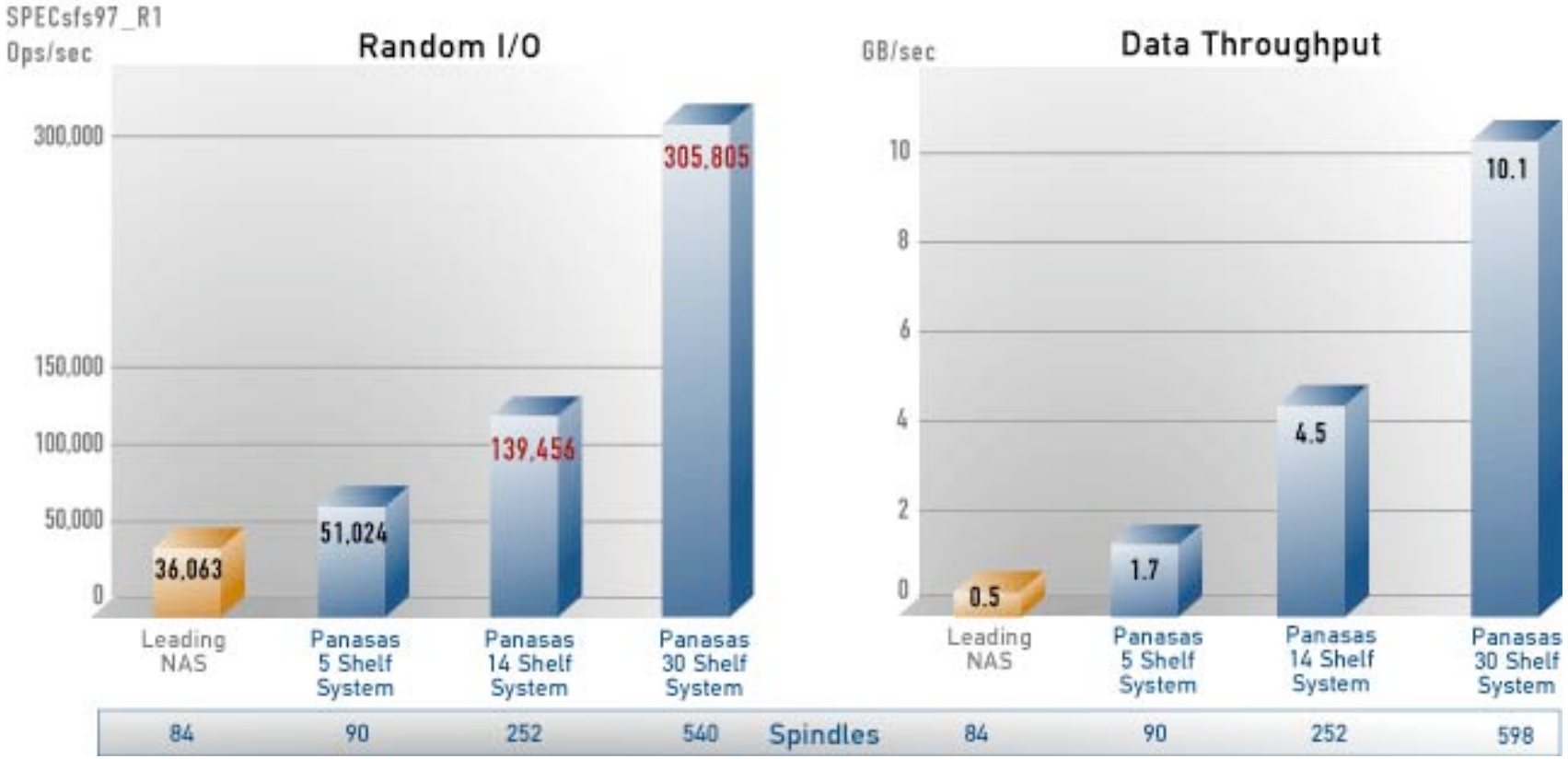
- **Snapshots: consistency for copying, backing up**
  - Copy-on-write duplication of contents of objects
  - Named as "…/.snapshot/JulianSnapTimestamp/filename"
  - Snaps can be scheduled, auto-deleted
- **Soft volumes: grow management without physical constraints**
  - Volumes can be quota bounded, unbounded, or just send email on threshold
  - Multiple volumes can share space of a set of shelves (double disk failure domain)
- **Capacity and load balancing: seamless use of growing set of blades**
  - All blades track capacity & load; manager aggregates & ages utilization metrics
  - Unbalanced systems influence allocation; can trigger moves
  - Adding a blade simply makes a system unbalanced for awhile

Out-of-band & Clustered NAS



Performance & Scalability for All

*Objects: breakthrough data throughput **AND** random I/O*



ActiveScale In Practice

November 24, 2004

Panasas Solution Getting Traction

- **Wins in HPC labs, seismic processing, biotech & rendering**

"We are extremely pleased with the order of magnitude performance gains achieved by the Panasas system…with the Panasas system, we were able to get everything we needed and more."

Tony Katz
Manager, IT
TGS Imaging

"The system is blazing fast, we've been able to eliminate our I/O bottleneck so researchers can analyze data more quickly. The product is 'plug-and-play' at all levels. "

Dr. Terry Gaasterland
Associate Professor
Gaasterland Laboratory
of Computational Genomics

"We looked everywhere for a solution that could deliver exceptional per-shelf performance. Finally we found a system that wouldn't choke on our bandwidth re

Mark Smith
President
MoveDigital

Berkeley University of California

Core Lab

Top Seismic Processing Company

Sandia National Laboratories

MAXUS REPSOL YPF

Leading Animation / Entertainment Company

Los Alamos NATIONAL LABORATORY

Storage Networking World BEST PRACTICES IN STORAGE AWARDS PROGRAM Computerworld SNIA

WINNER INNOVATIVE PRODUCT AWARD nsc network storage conference

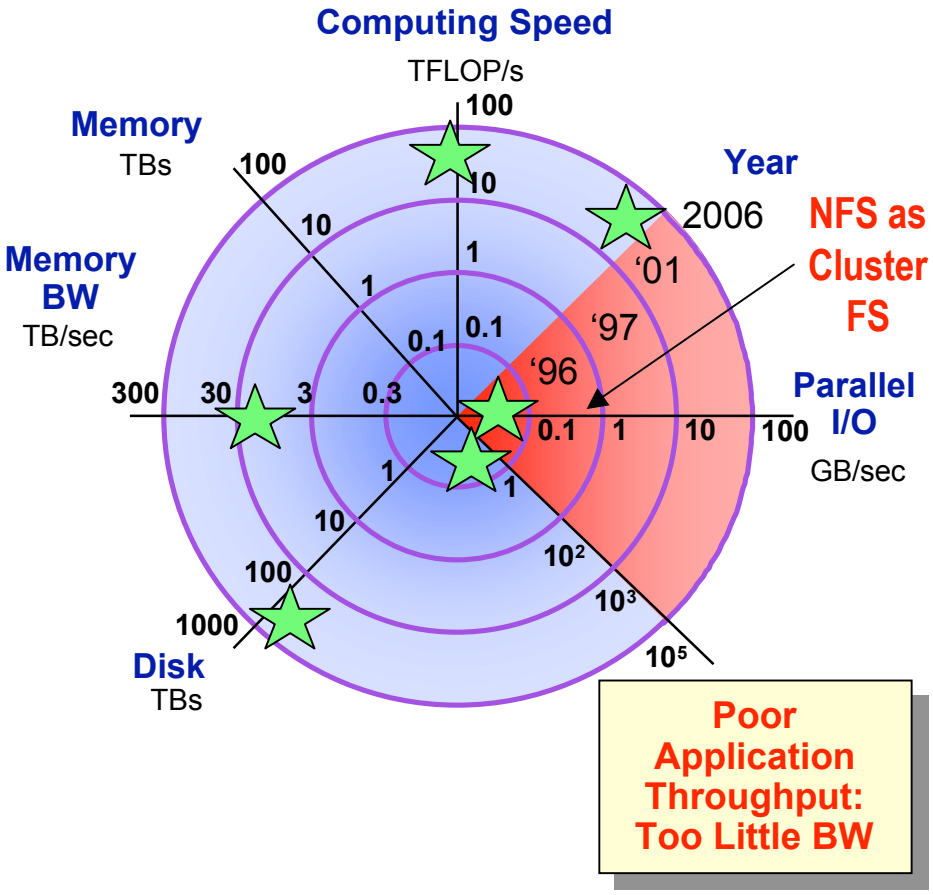
NetworkWorldFusion The leader in network knowledge VENDORS TO WATCH 2004

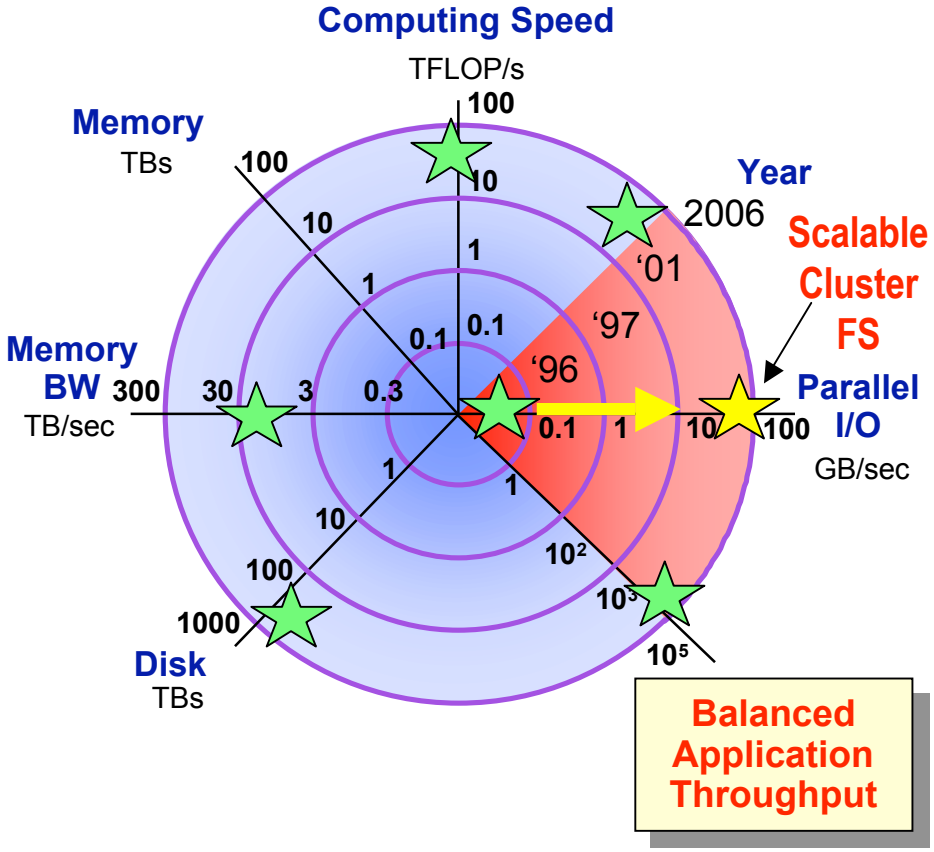
TOP 10 INVESTOR'S CHOICE AWARD DATACENTER VENTURES

Excellence in CLUSTER Technology Storage Solutions WINNER ClusterWorld 2004

Panasas in Action: LANL

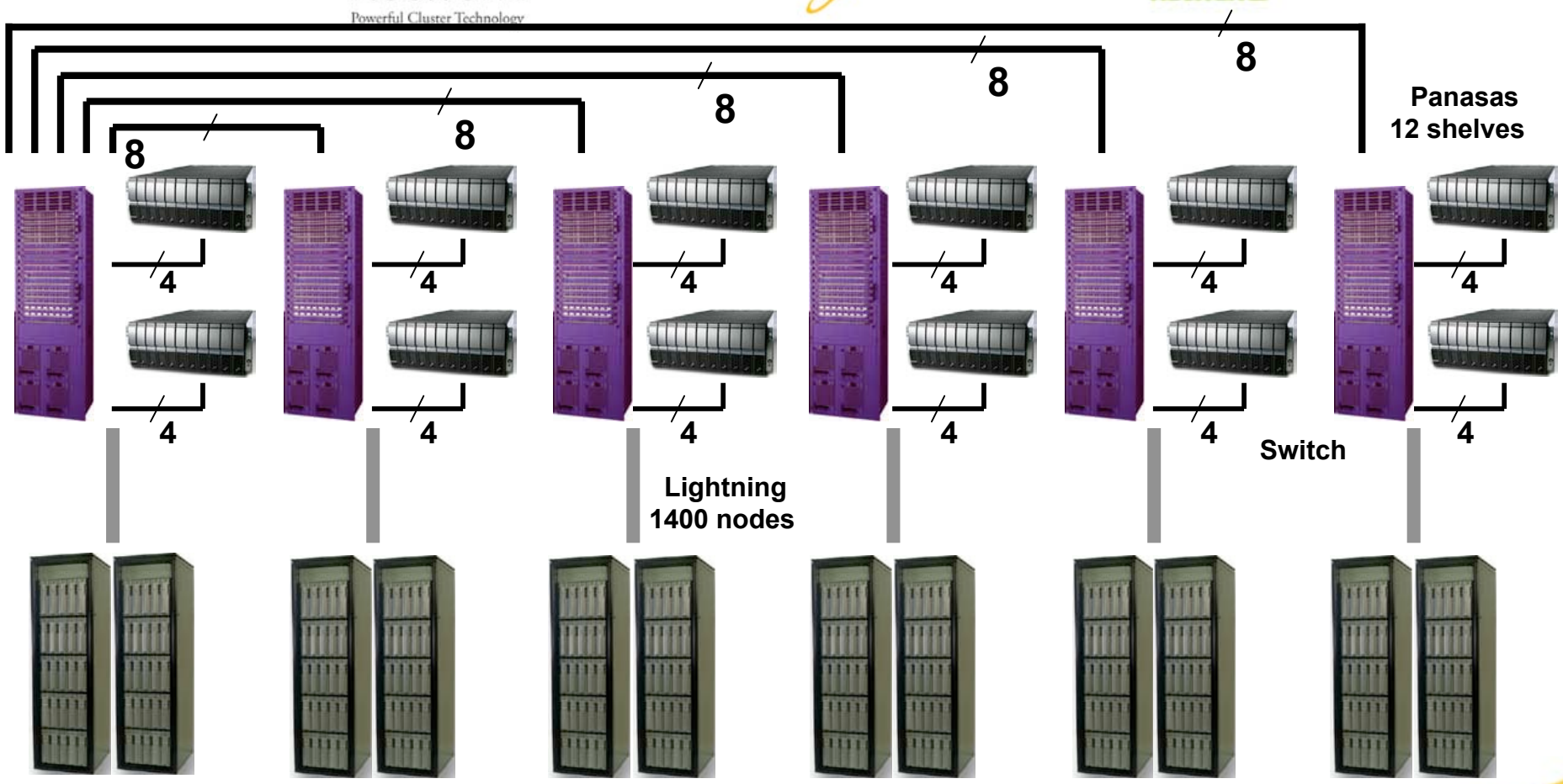
Los Alamos Nat Lab: Seeking a Balanced System





Los Alamos Lightning*

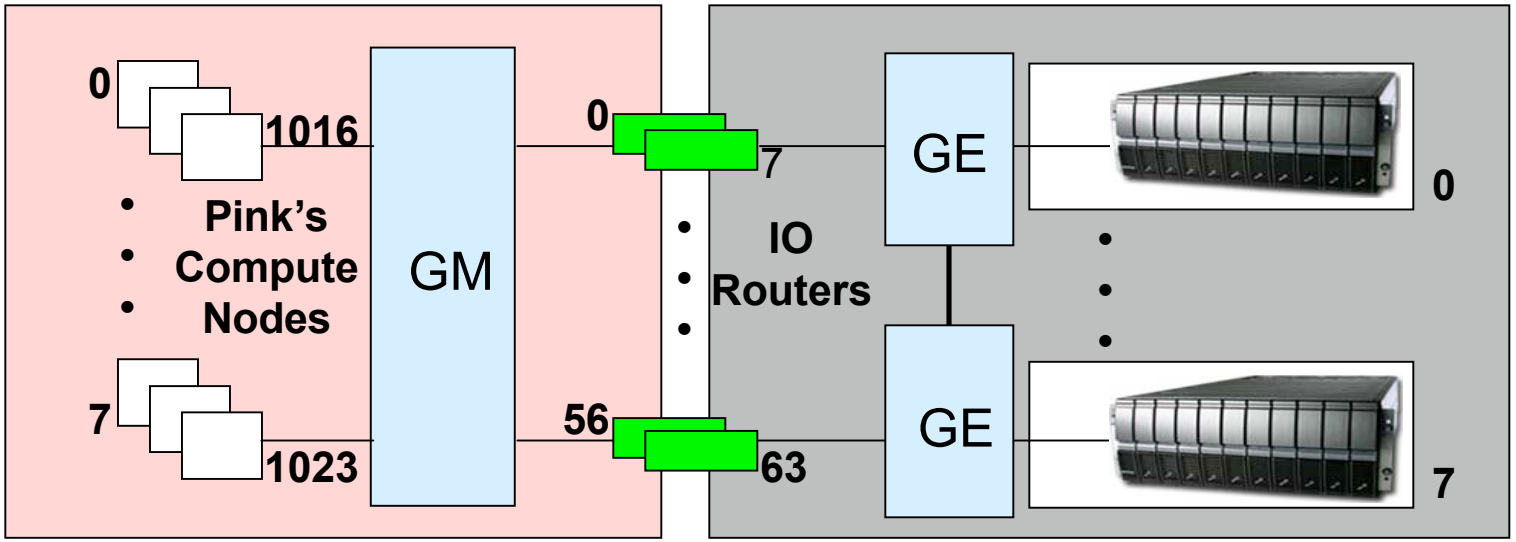
1400 nodes and 60TB (120 TB): Ability to deliver ~ 3 GB/s* (~6 GB/s)



* entering production

Pink: A Non-GE Cluster

- **Non-GE Cluster Interconnects for high bandwidth, low latency**
  - LANL Pink's 1024 nodes use Myrinet; others use Infiniband or Quadrics
- **Route storage traffic (iSCSI) through cluster interconnect**
  - Via IO routers (1 per 16 nodes in Pink)
  - Lower GE NIC & wire costs; Lower bisection BW in GE switches (possibly no GE switches)
  - Linux load balancing, OSPF & Equal Cost Multi-Path for route load balancing and failover
- **Integrate IO node into multi-protocol switch port**
  - E.g. Topspin, Voltaire, Myricom GE line cards head in this direction



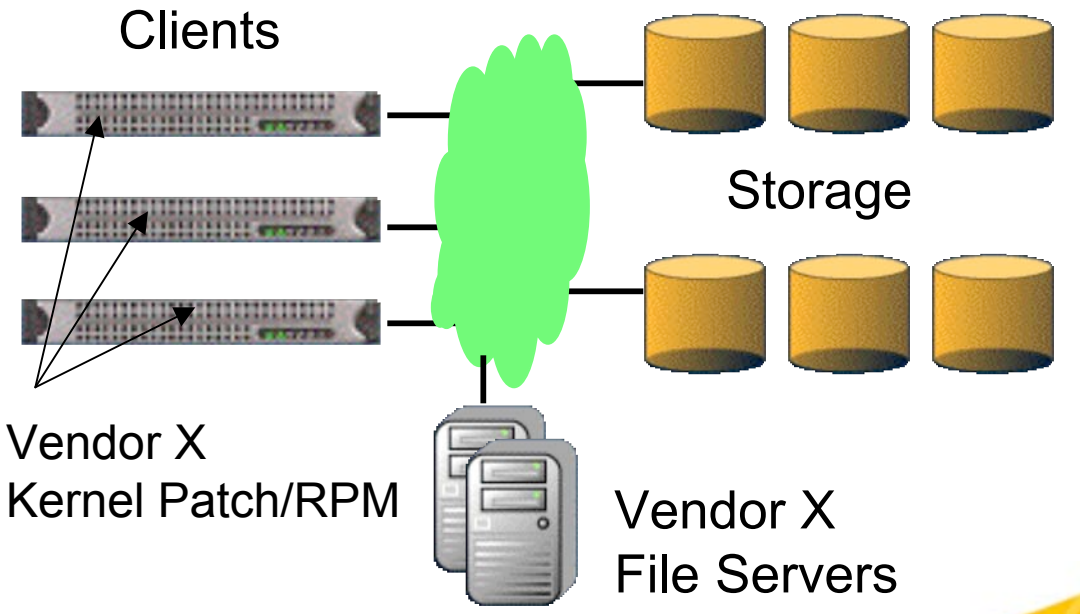
Parallel NFS Possible Future

November 24, 2004

Out-of-Band Interoperability Issues

- **ADVANTAGES**
  - Capacity scales
  - Bandwidth scales
- **DISADVANTAGES**
  - Requires client kernel addition
  - Many non-interoperable solutions
  - Not necessarily able to replace NFS

- **EXAMPLE FEATURES**
  - POSIX Plus & Minus
  - Global mount point
  - Fault tolerant cache coherence
  - RAID 0, 1, 5 & snapshots
  - Distributed metadata and online growth, upgrade



File Systems Standards: Parallel NFS

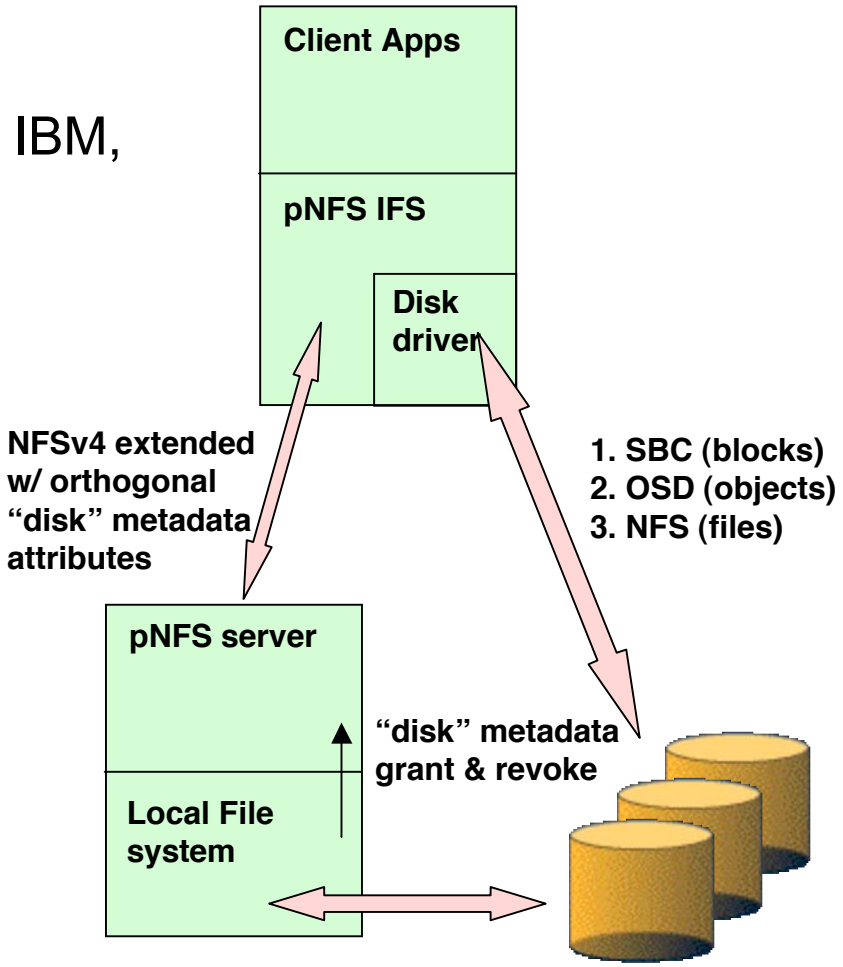
- **IETF NFSv4 initiative**
  - U. Michigan, NetApp, Sun, EMC, IBM, Panasas, ….
  - Enable parallel transfer in NFS

IETF pNFS Documents:

draft-gibson-pnfs-problem-statement-01.txt

draft-gibson-pnfs-reqs-00.txt

draft-welch-pnfs-ops-00.txt



Cluster Storage for Scalable Linux Clusters

Garth Gibson
ggibson@panasas.com
www.panasas.com

panasas

BACKUP

BladeServer Storage Cluster