Coping with Petabyte Files at Petascale Performance

The Salishan Conference on High-Speed Computing

Garth Gibson
Founder & CTO, Panasas, and Assoc Prof., Carnegie Mellon University
garth@panasas.com and garth@cs.cmu.edu

April 27, 2006

SGPFS
Challenges and Hurdles

Garth Gibson
garth.gibson@cs.cmu.edu

PARALLEL DATA LABORATORY
Carnegie Mellon University
www.pdl.cs.cmu.edu

SGPFS requirements summary

Scalable: bandwidth scales with capacity (10,000+ devices)

Global: shared, heterogeneous OS and SAN/WAN support

Parallel: multiple concurrent readers and writers in a file

FileSystem: manageable, persistent, familiar

(Secure): stored and transmitted data safe from tampering

Distributed FS = (Secure) Global FileSystem

SGPFS = High-Bandwidth Concurrent-Writers DFS

What is the problem?

Financial realities compel use of COTS technology

COTS products respond to size of market

High-Bandwidth, Concurrent-Writers is small market

So, Big science DFS systems are not COTS

Not COTS DFS require “improvements” - costly, fragile

Expenditure not persistent SW development investment

Carnegie Mellon

Alternative solution philosophy

Make non-COTS features “easy” for DFS to provide

- depend only on big market features: large capacity, manageability

High-bandwidth: direct transfer between app and device

- network-attached storage on scalable storage area networks
- server machine specs do not define peak storage bandwidth

Concurrent-writers: middleware in app, little in DFS

- MPI-IO

Carnegie Mellon

NASD and PFS (SIO LLAPI)

Example: weakly consistent caching

Consistency best known to application is left to application

Simple system support

byte range caching, propagate/refresh primitives

Client A	Client B
write(FD, Fshared, Ma)	
	unsafe read(FD, Fshared, Mb)
propagate(FD, Fshared)	
	unsafe read(FD, Fshared, Mb)

- - - - - - - Synchronization Event -

Client A	Client B
	refresh(FD, Fshared)
	safe read(FD, Fshared, Mb)

Carnegie Mellon

Alternative solution philosophy

Make non-COTS features "easy" for DFS to provide

- **depend only on big market features: large capacity, manageability**

Revise: simple BW "easy"; increasing async & failure scope are not

High-bandwidth: direct transfer between app and device

- **network-attached storage on scalable storage area networks**
- **server machine specs do not define peak storage bandwidth**

We're good here

Concurrent-writers: middleware in app, little in DFS

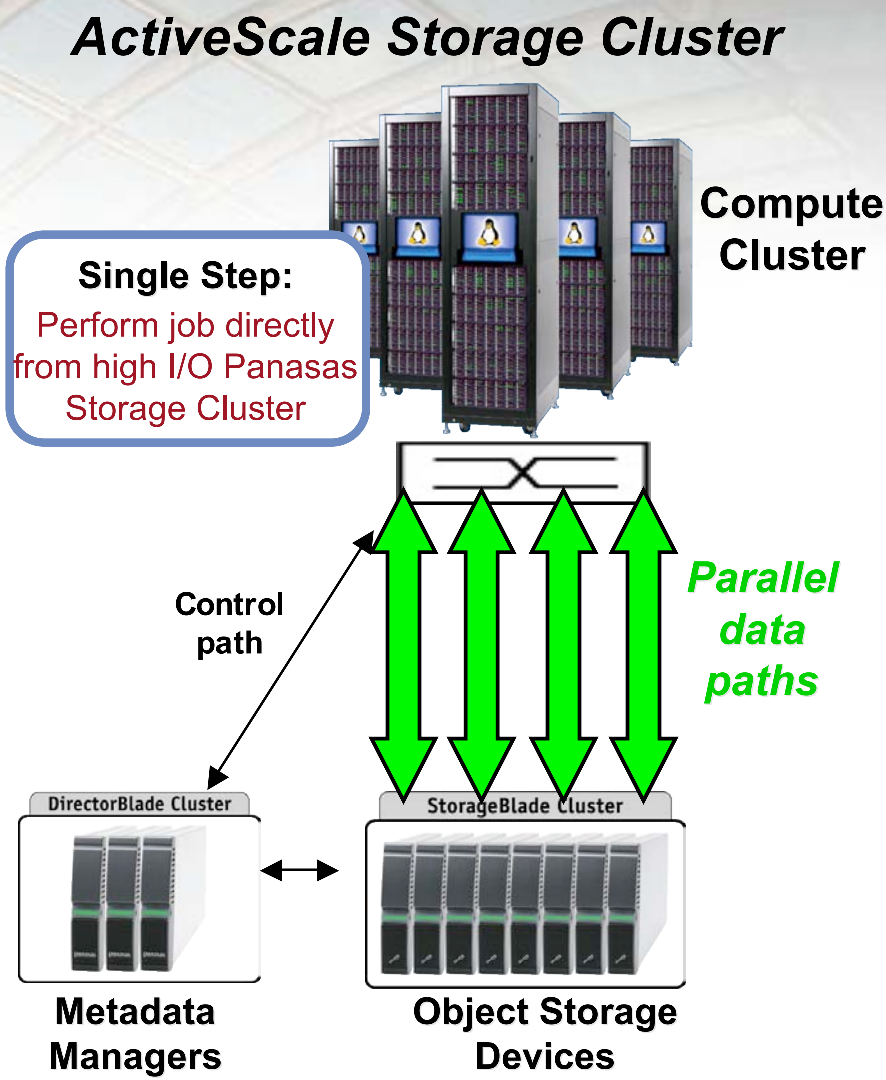
- **MPI-IO**

Revise: programmers weren't listening and may not until FS fails :-(

Carnegie Mellon

High Performance Cluster Storage

- Scalable performance
  - Parallel data paths to compute nodes
  - 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



panasas

Panasas in the Field

HPC wire
Editors' Choice Awards
Best Price / Performance Storage Solution
Panasas
Panasas ActiveScale File System
2005

Seismic Processing

TOTAL
bp
PGS
TGS
PetroChina 中国石油
MAXUS
Paradigm THE GEOSCIENCE KNOWLEDGE COMPANY
REPSOL YPF
SPINNAKER EXPLORATION

Life Sciences

Berkeley University of California
UC SANTA CRUZ
THE ROCKEFELLER UNIVERSITY
NOVARTIS
EMBL-EBI European Bioinformatics
UCSF
MERCK

HPC Simulations

ECMWF
NASA
Los Alamos NATIONAL LABORATORY
NW-GRID
BROOKHAVEN NATIONAL LABORATORY
Sandia National Laboratories
CCLRC
NORTHROP GRUMMAN

Rendering

Walt Disney FEATURE ANIMATION
School of VISUAL ARTS
Weta Digital

Fluid Dynamics

BOEING
FLUENT
Ferrari
Princeton University

BNP PARIBAS
Stephens Inc. Investment Bankers

Excellence in CLUSTER Technology
Storage Solutions
WINNER
ClusterWorld
2004

COMPUTERWORLD INNOVATIVE TECHNOLOGY AWARDS 2004
STORAGE NETWORKING WORLD BEST PRACTICES IN STORAGE AWARDS PROGRAM COMPUTERWORLD SNIA
WALL STREET JOURNAL
WINNER INNOVATIVE PRODUCT AWARD NSC NETWORK STORAGE CONFERENCE
InformationWeek BUSINESS INNOVATION POWERED BY TECHNOLOGY
Bio·IT World
NetworkWorldFusion The leader in network knowledge VENDORS TO WATCH 2004

ActiveScale Storage Cluster

April 27, 2006

G. Gibson, Panasas

Object Storage Systems

Expect wide variety of Object Storage Devices

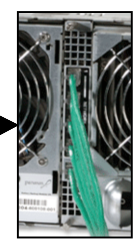
- Disk array subsystem
- Ie. LLNL with Lustre

- “Smart” disk for objects
- 2 SATA disks – 500/800 GB

- Prototype Seagate OSD
- Highly integrated, single disk

- Orchestrates system activity
- Balances objects across OSDs

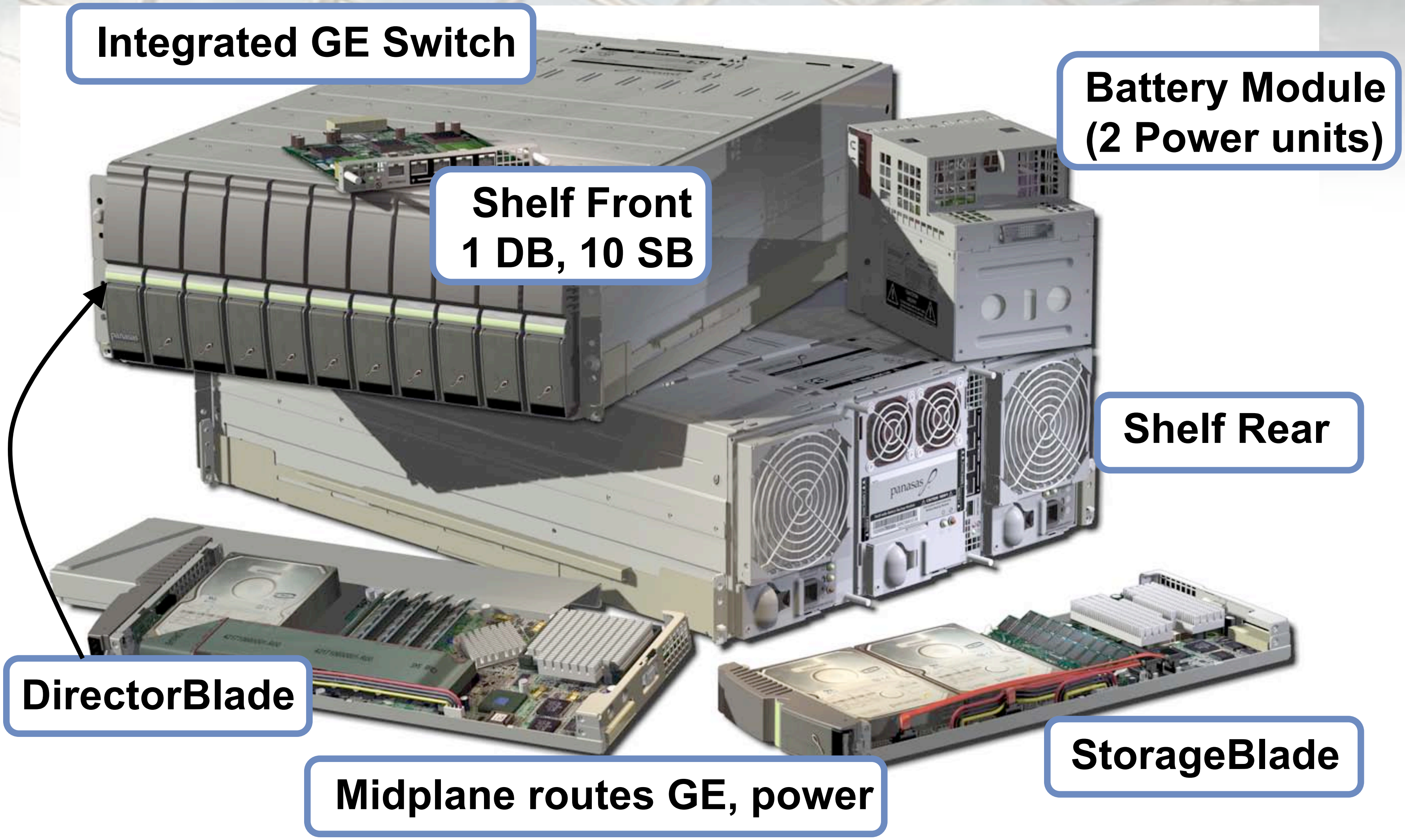
- **Stores up to 8 TBs per shelf**



16-Port GE Switch Blade

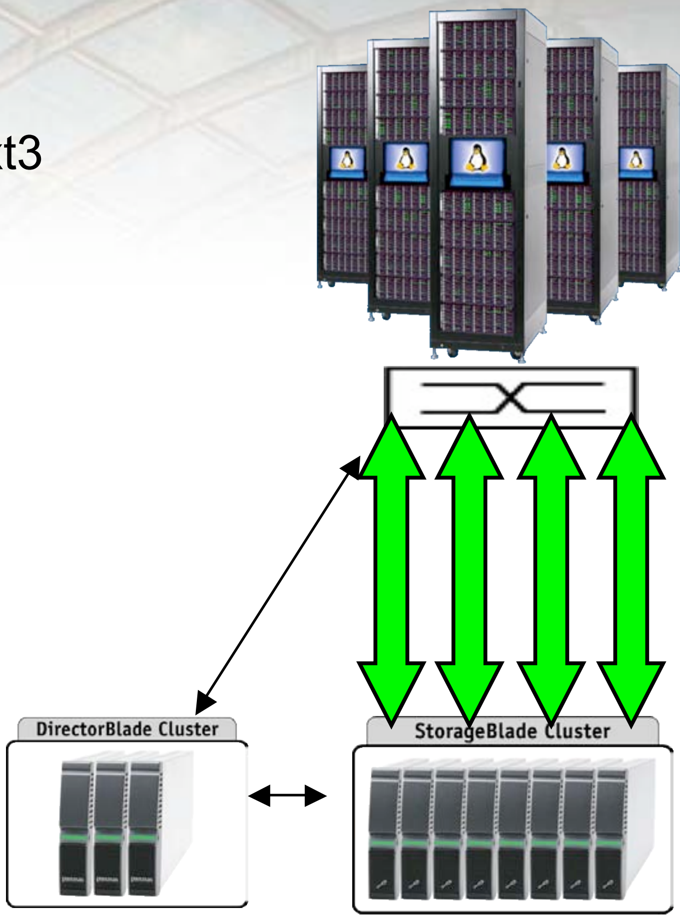
- 4 Gbps per shelf to cluster

BladeServer Storage Cluster



DirectFLOW Linux Client

- Installable File System
  - Uses standard Linux VFS interface, like ext3
- Kernel Loadable Module
  - No kernel modifications required
- Presents a POSIX Interface
  - No Application modifications required
- Uses iSCSI with OSD command set
- Major Linux Distributions are supported
  - RedHat, SLES, Fedora
  - Custom ports available for customised kernels.



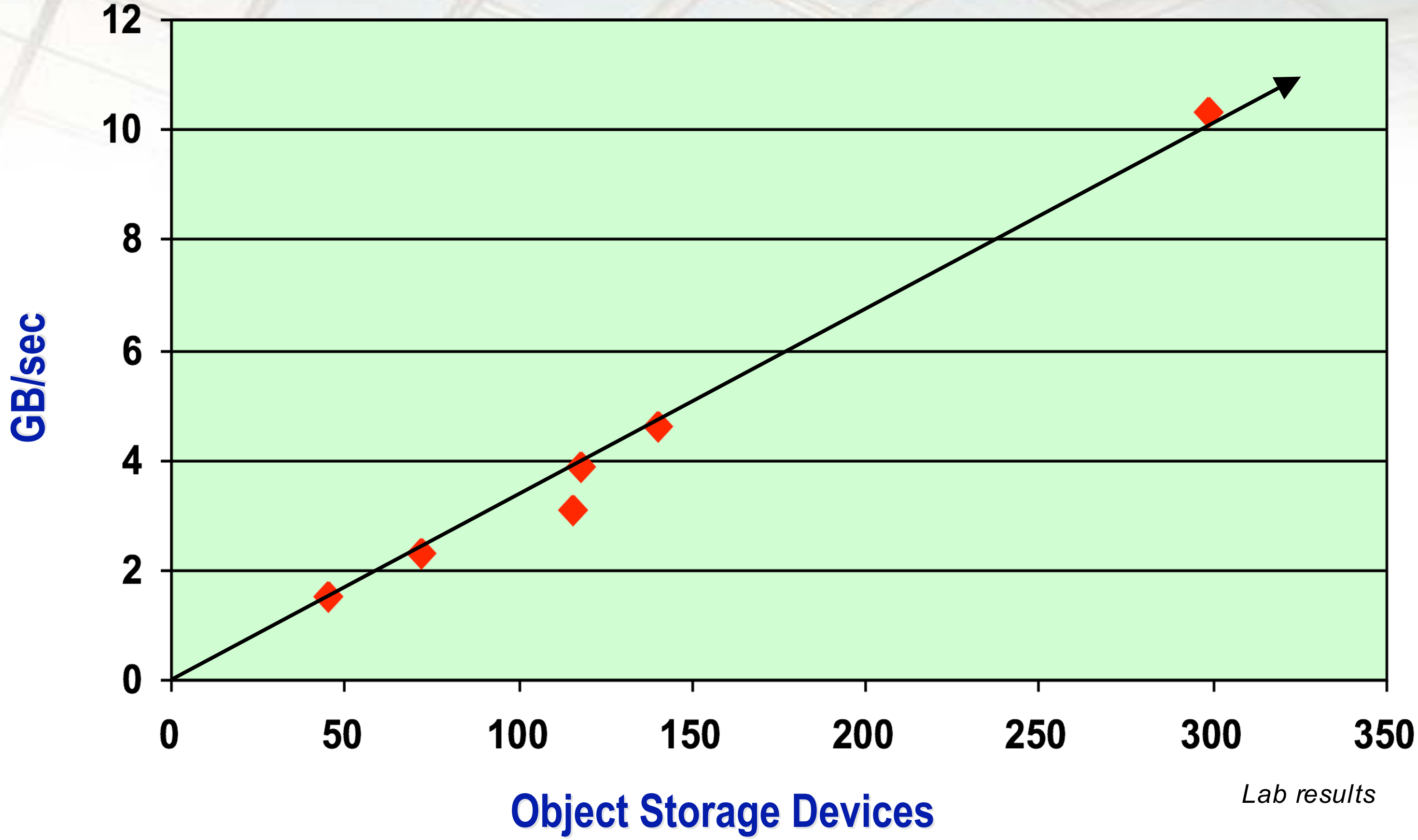
Striping, PanRAID & Reliability

April 27, 2006

G. Gibson, Panasas

Object Storage Bandwidth

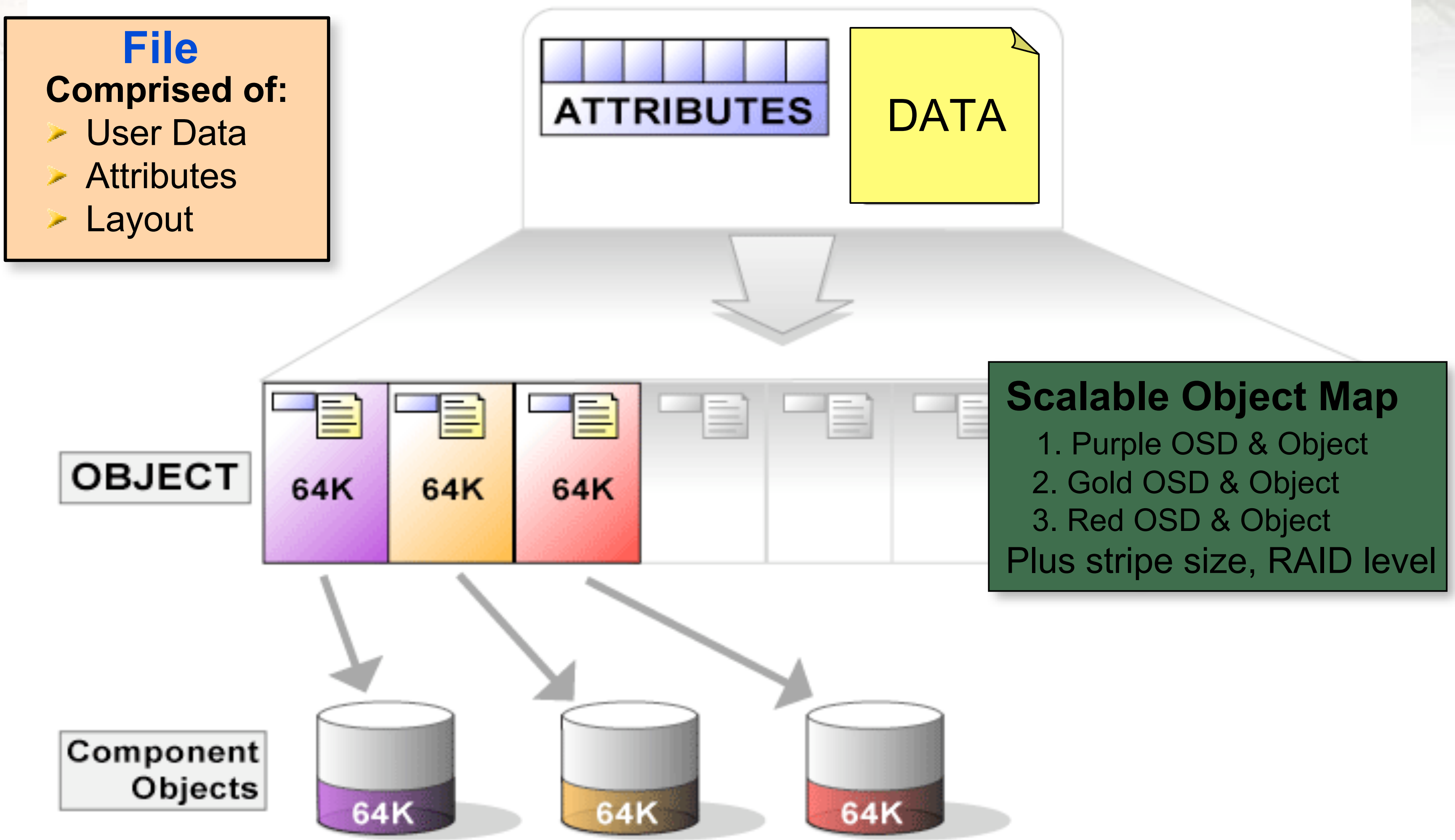
Scalable Bandwidth demonstrated with GE switching



Lab results

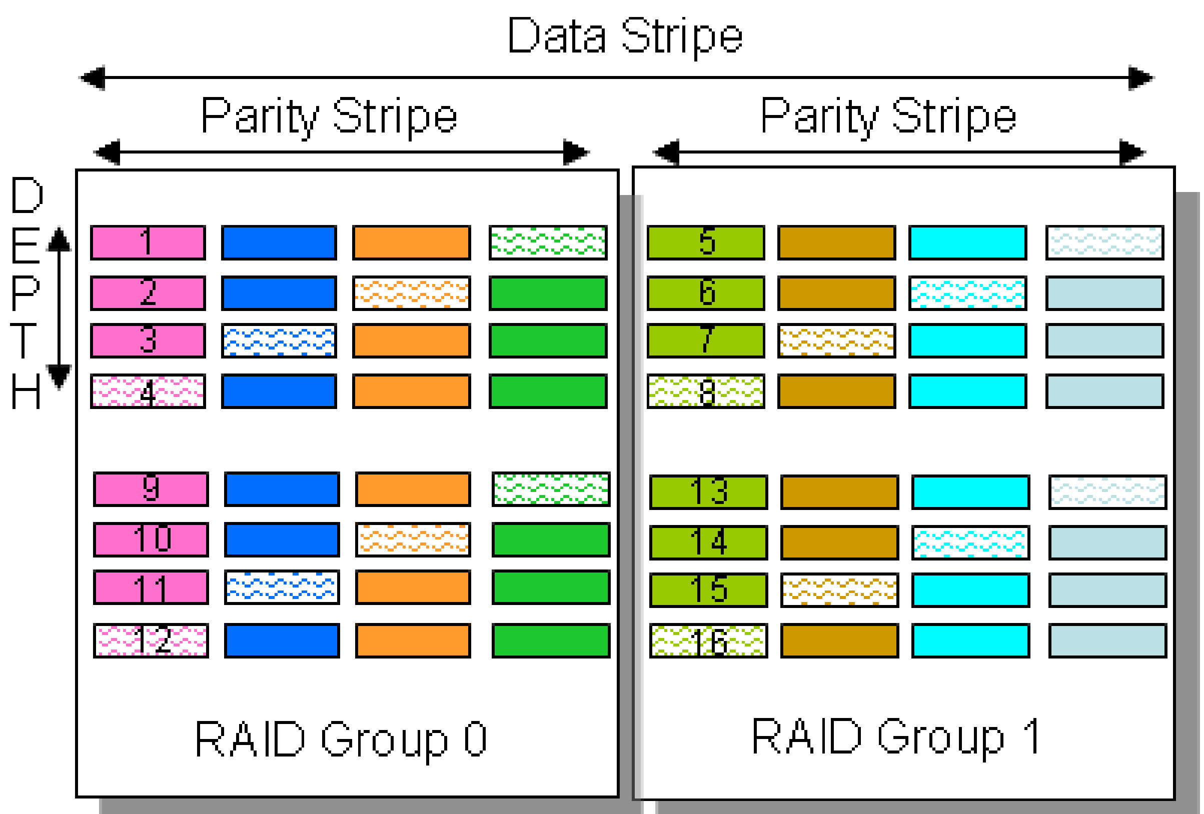
How Do Panasas Objects Scale?

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



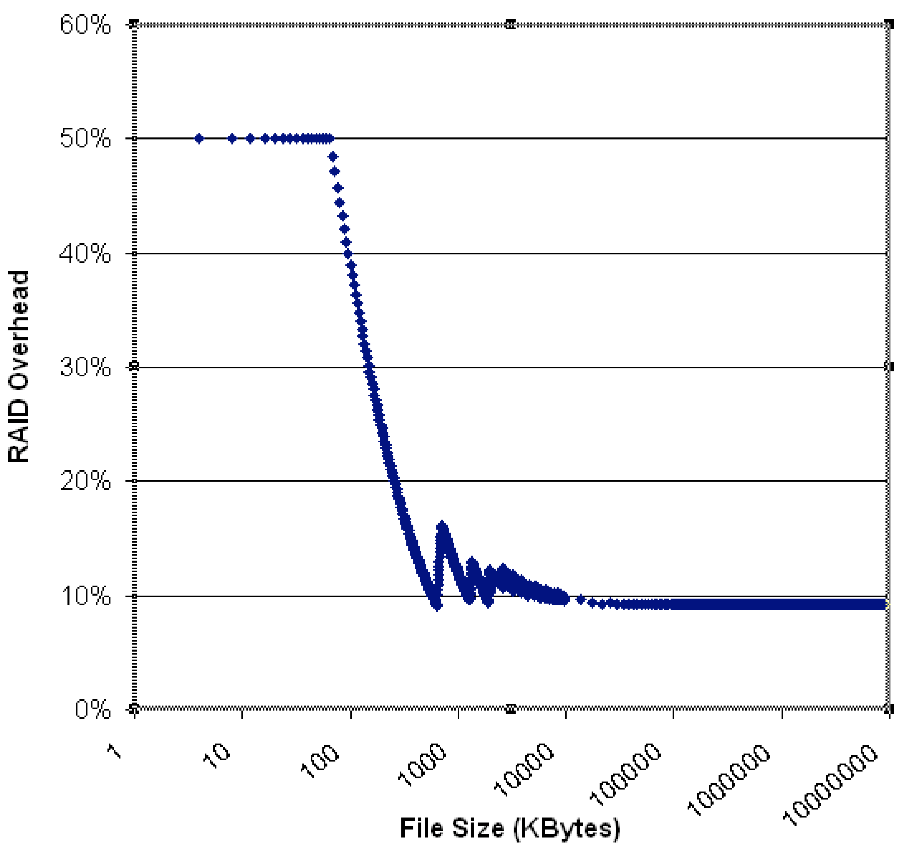
Huge Files Support Huge Bandwidth

- Two-level map spreads huge files over lots of disks efficiently
  - Separate parity OV from depth under disk head & total disks sharing file
  - Controls # of disks streaming at 1 client, limits network backup [Nagle, SC04]



Mixing Small Files with Large

- Small files are mirrored; larger files have lower RAID5 overhead
- Mixed file systems: most files are small, most space is in large files
- Combined parity overhead follows large files more than small
- Panasas /build & /home: 12.5%
- Five volumes from 2 customers:
  - 14%, 12%, 12%, 19%, 21%

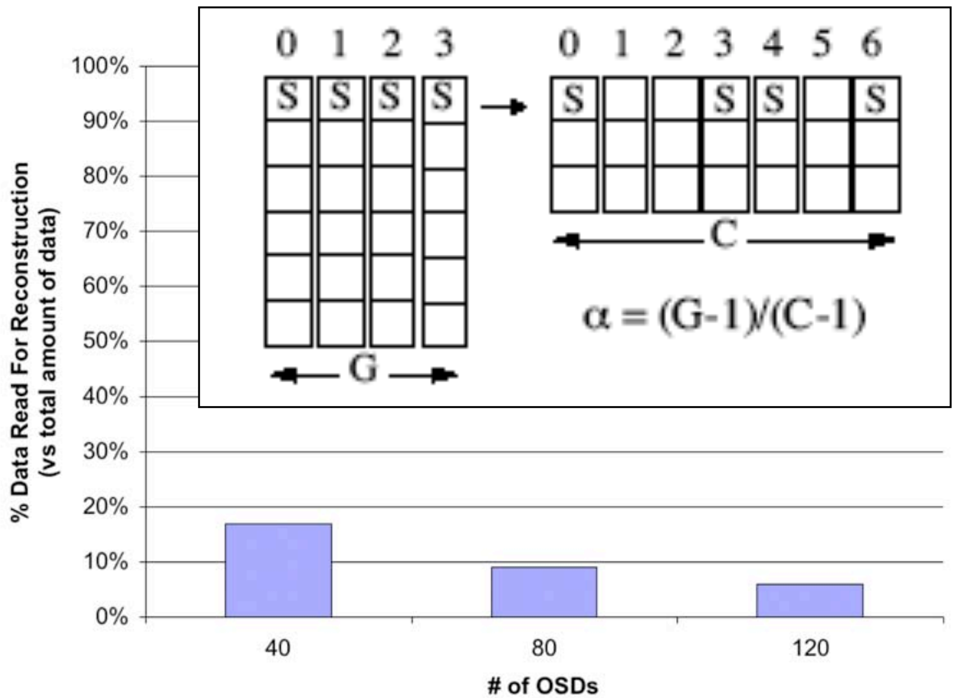


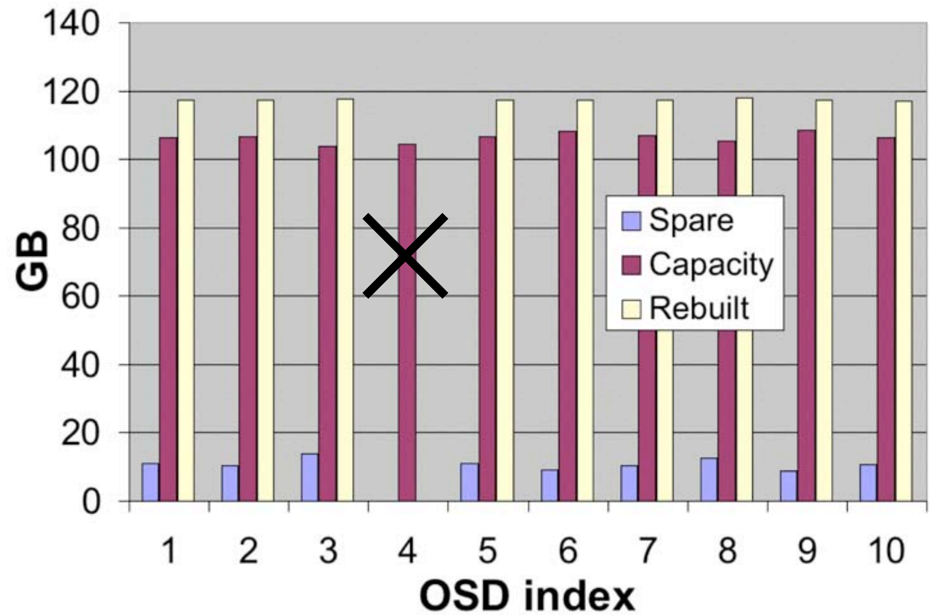
Physical Capacity Used	1,144 GB	(100.0%)
Data Capacity Used	956 GB	(83.6%)
Parity Capacity Used	143 GB	(12.5%)
Metadata Overhead	30 GB	(2.6%)
Internal Fragmentation	15 GB	(1.3%)
Total Number of Files	3,459,752 files	

Table 1: Selfhost observed capacity overhead.

Per-file Map Declusters RAID

- Each file has its own map, drawing on different OSDs evenly
- Fraction of each OSD read to rebuild a failure decreases with # OSDs
- Rebuilt data spread over all surviving OSDs evenly
- All disk arms available for reading & writing during reconstruction

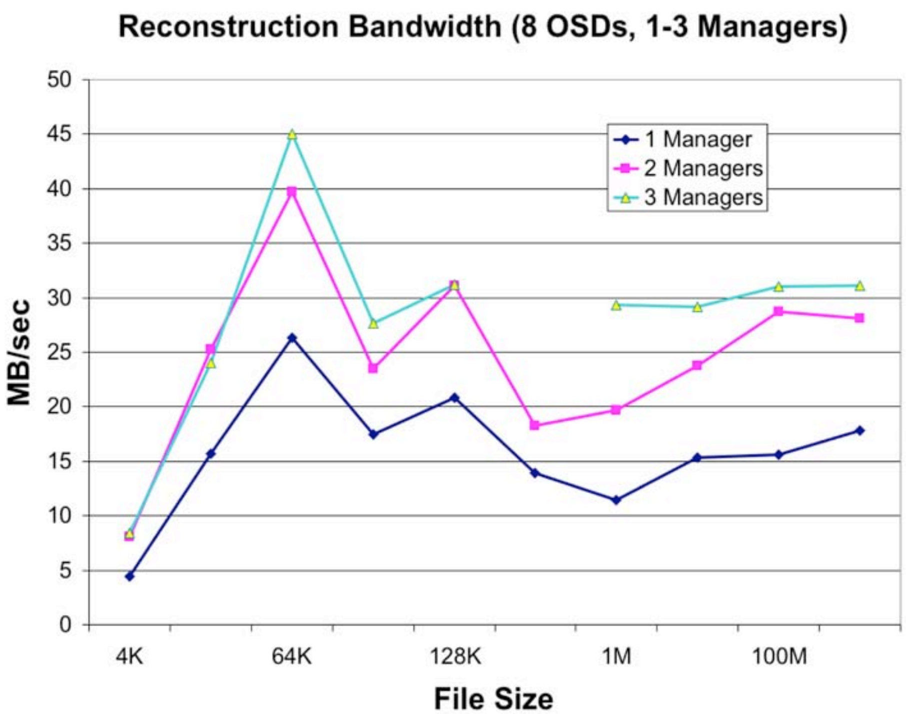


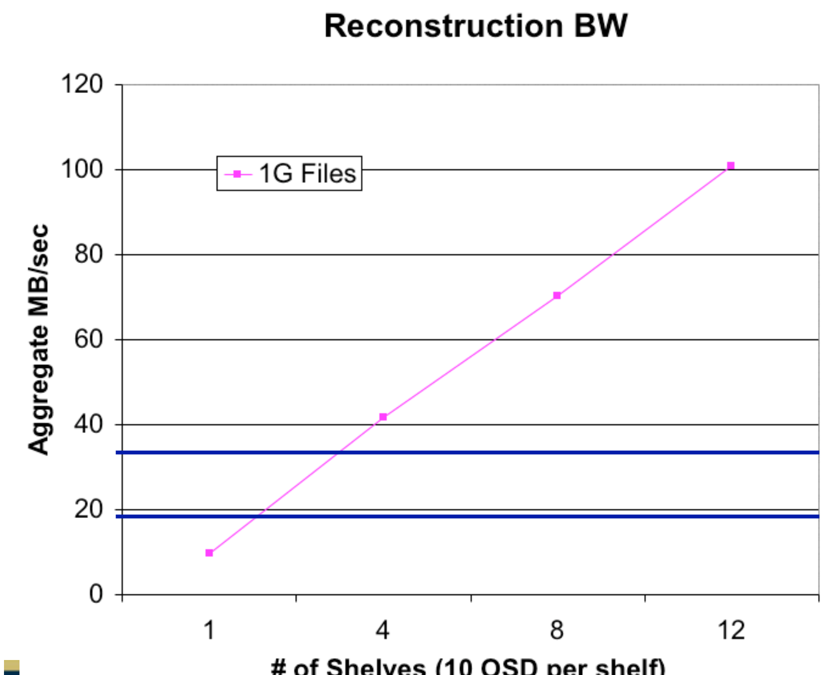


Scaling Shelves Scales Repair Rate

- Compare n RAIDs of C+G disks to 1 declustered array of n*(C+G) disks
- Use managers of all shelves to scale reconstruction/repair rate
- Adding shelves increase repair rate linearly
- And shorter degraded periods!

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



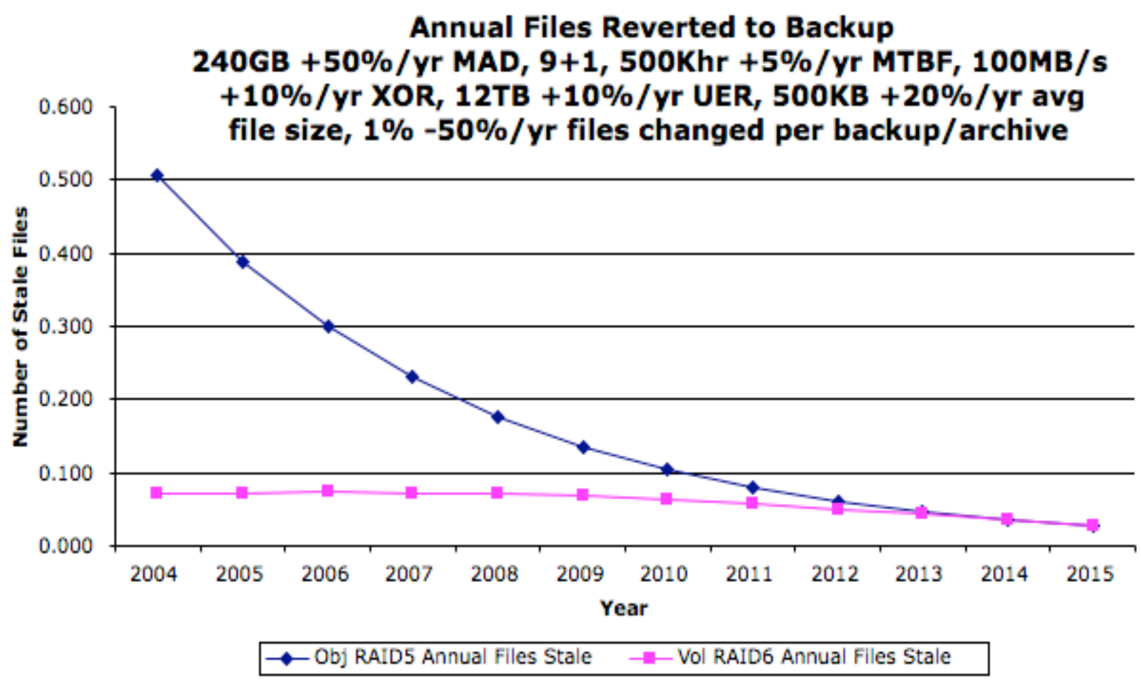


Emphasis on Data Reliability

- Reliability designed into Panasas Hardware:
  - Redundant power supplies and fans
  - Redundant network connections to each blade
  - Built in UPS for power fail protection
  - ECC memory
  - Backup network built into shelf
- Reliability built into Panasas Software:
  - RAID 1 & 5 data redundancy with scalably fast reconstruction
  - Background, file-aware media, parity & attributes scrubbing and recovery
  - Proactive monitoring including disk SMART, heat, fans, battery
  - Scalable, high performance Backup and Restore
  - Proven FreeBSD base operating system
  - Mirrored Blade OS – protection against errors & repair in the OS partition (beta)
  - Systems services failover; file service metadata manager failover (beta)
  - Media + Disk failure => rebuild succeeds w/ loss of one file (fenced), not millions of files

Unrecoverable Read Errors (URE)?

- 1 bit in ~12 TB read unreadable, so reconstruction will see drop outs
  - Today loss of a sector during reconstruction of RAID5 “loses the volume”
  - NetApp developed RAID-DP (EvenOdd variant) to tolerate all disk+URE failures
- Object-RAID, RAID 5 inside a file loses 1 file on URE
  - Lost file often in backup or archive
  - Low annual rate of lost files
  - Vol-RAID 6 can’t survive double disk failures b/c of UREs
- Example trends shown
  - Lines converge unless URE rate decreases a lot
  - Trend is up or down with AvgFileSize/MAD trends



panasas

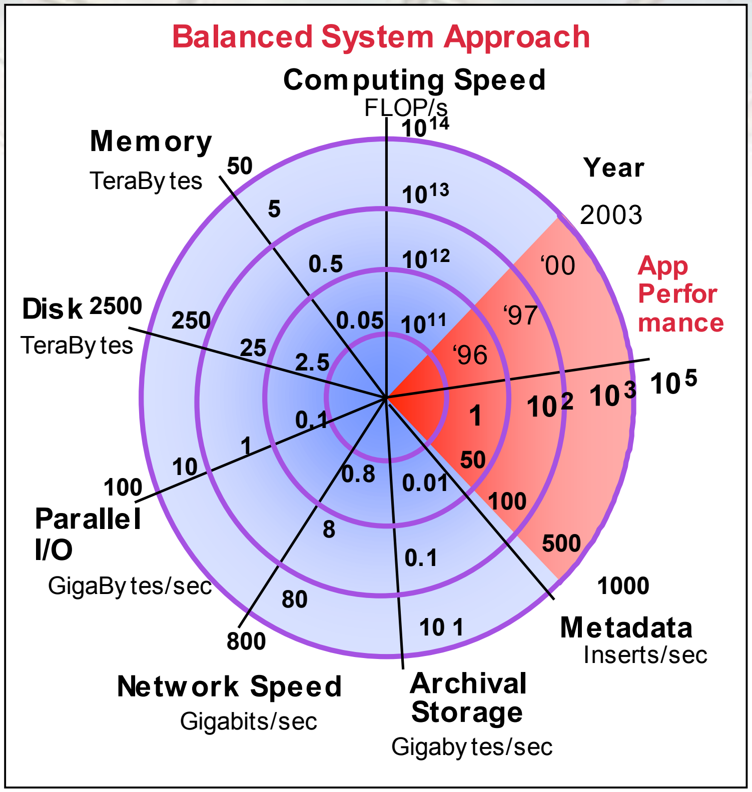
Los Alamos Case Study

April 27, 2006

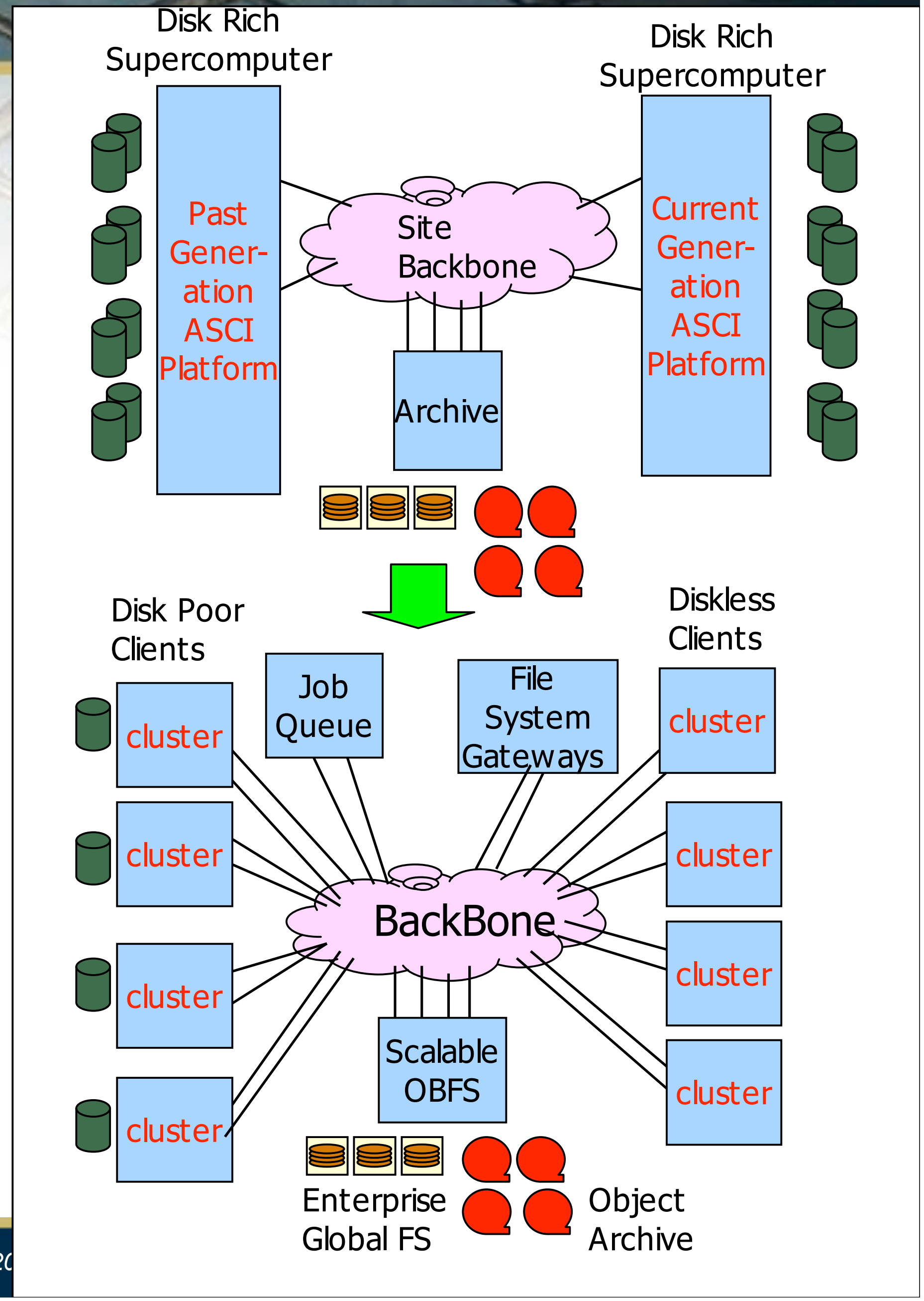
G. Gibson, Panasas

Panasas in Action: LANL

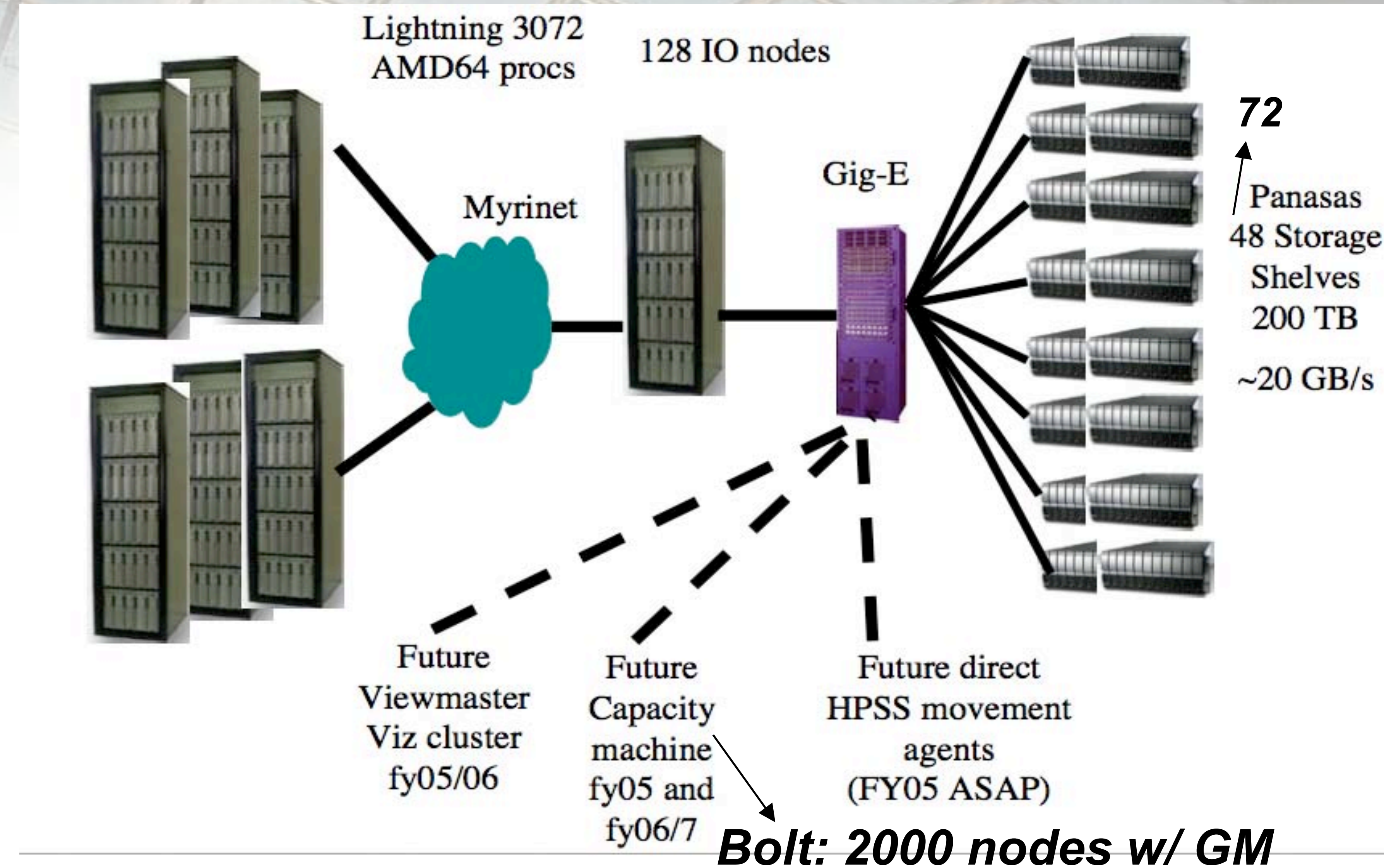
1 GB/s per TFLOPS



- 750 TB Panasas in many clusters
  - GM: 5600 nodes, 11000+ procs, Lightning, Bolt, Pink, TLC, Flash, Gordon
  - IB: 1856 nodes, 3700+ procs, Blue Steel, Coyote

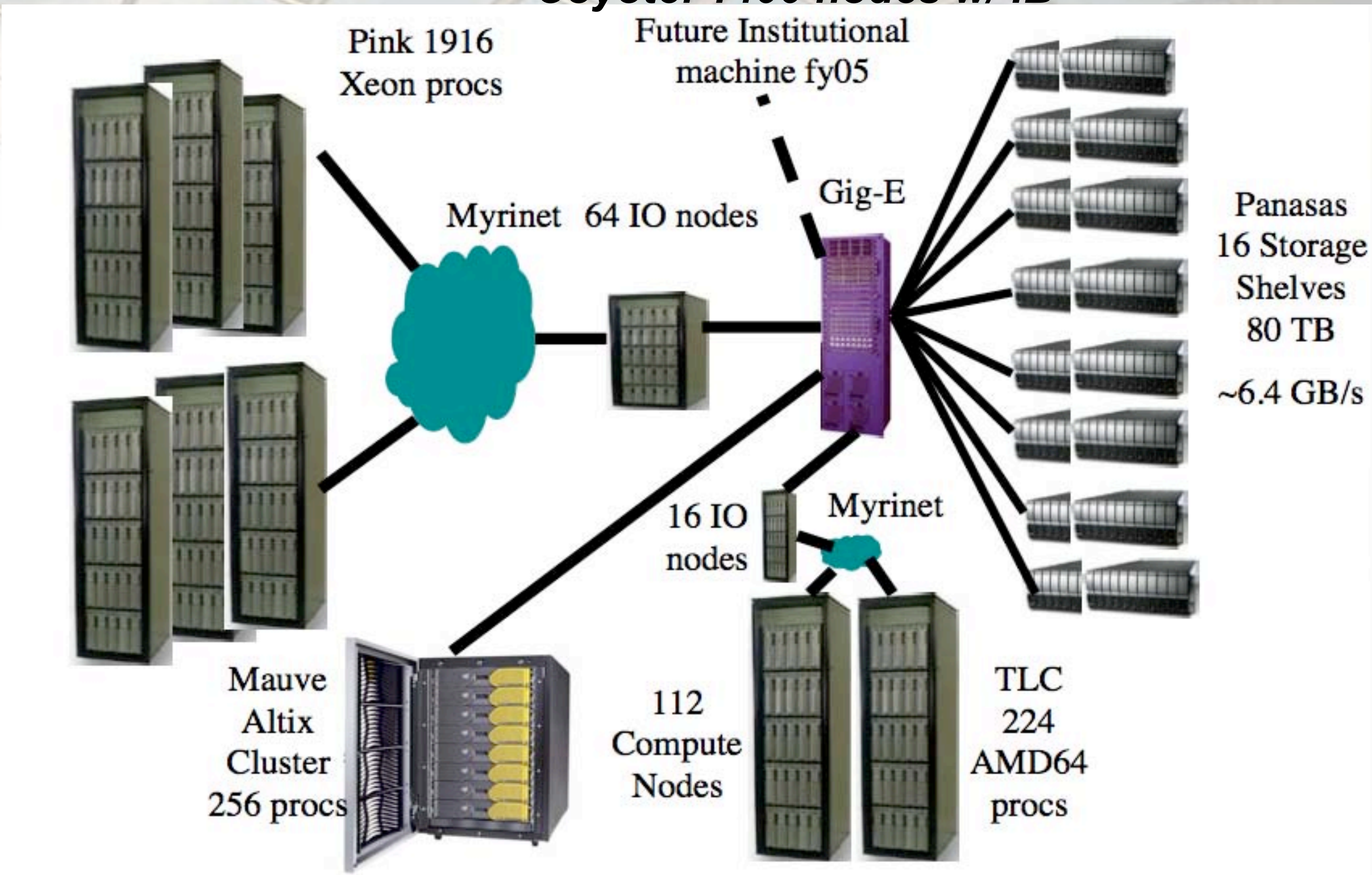


LANL's Lightning Cluster



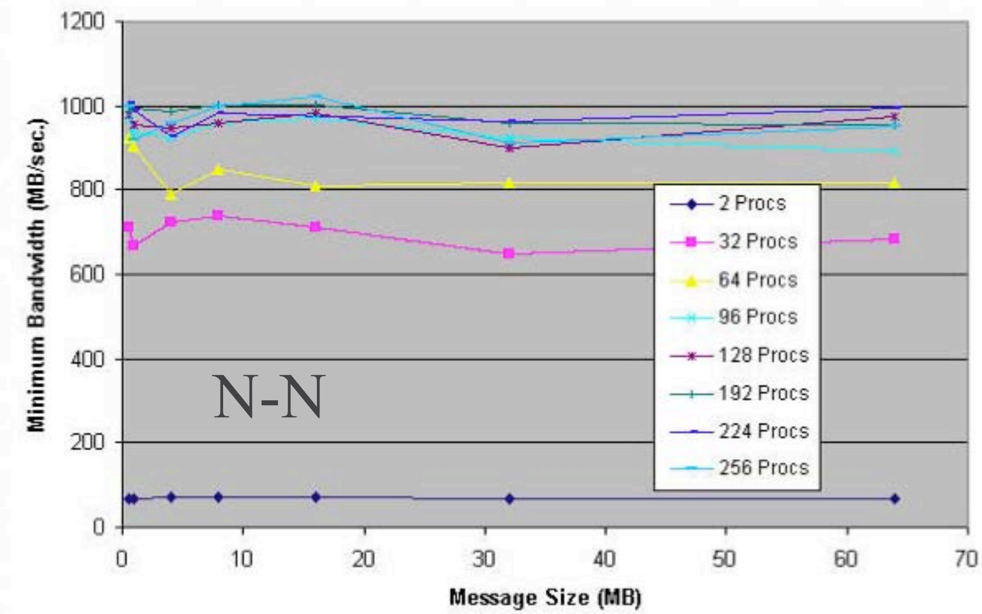
Pink: Single-Image (Diskless)

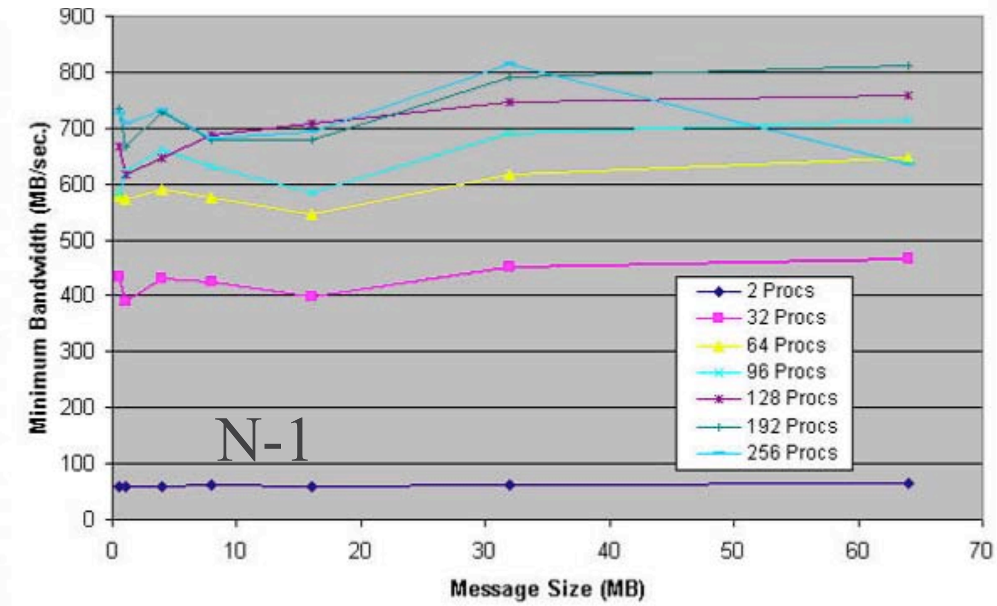
Coyote: 1400 nodes w/ IB



Pink MPI-IO Write BW

- MPI-IO assessment benchmark emulates scientific simulation codes at LANL
- Writes 4GB sequentially in “message” size chunks
  - N-N: 1 file per proc (2 per node)
  - N-1: 1 file shared by all procs
- Minimum BW is slowest proc, including file open/close
- Panasas storage was 4 shelves (20TB) w/ raw speed 1600 MB/s, 1200 MB/s average (not min)
- Performance stable across chunk size
- Grider, Chen et al, LAUR-05-7620, Int. Performance, Computing & Comm. Conf., Pheonix AZ, Apr 10-12, 2006.





NASD and PFS (SIO LLAPI)

Example: weakly consistent caching

Consistency best known to application is left to application

Simple system support

byte range caching, propagate/refresh primitives

Client A	Client B
write(FD, Fshared, Ma)	
	unsafe read(FD, Fshared, Mb)
propagate(FD, Fshared)	
	unsafe read(FD, Fshared, Mb)
------- Synchronization Event -----------------------	
	refresh(FD, Fshared)
	safe read(FD, Fshared, Mb)

Carnegie Mellon

Small Strided Concurrent Write (N-1)

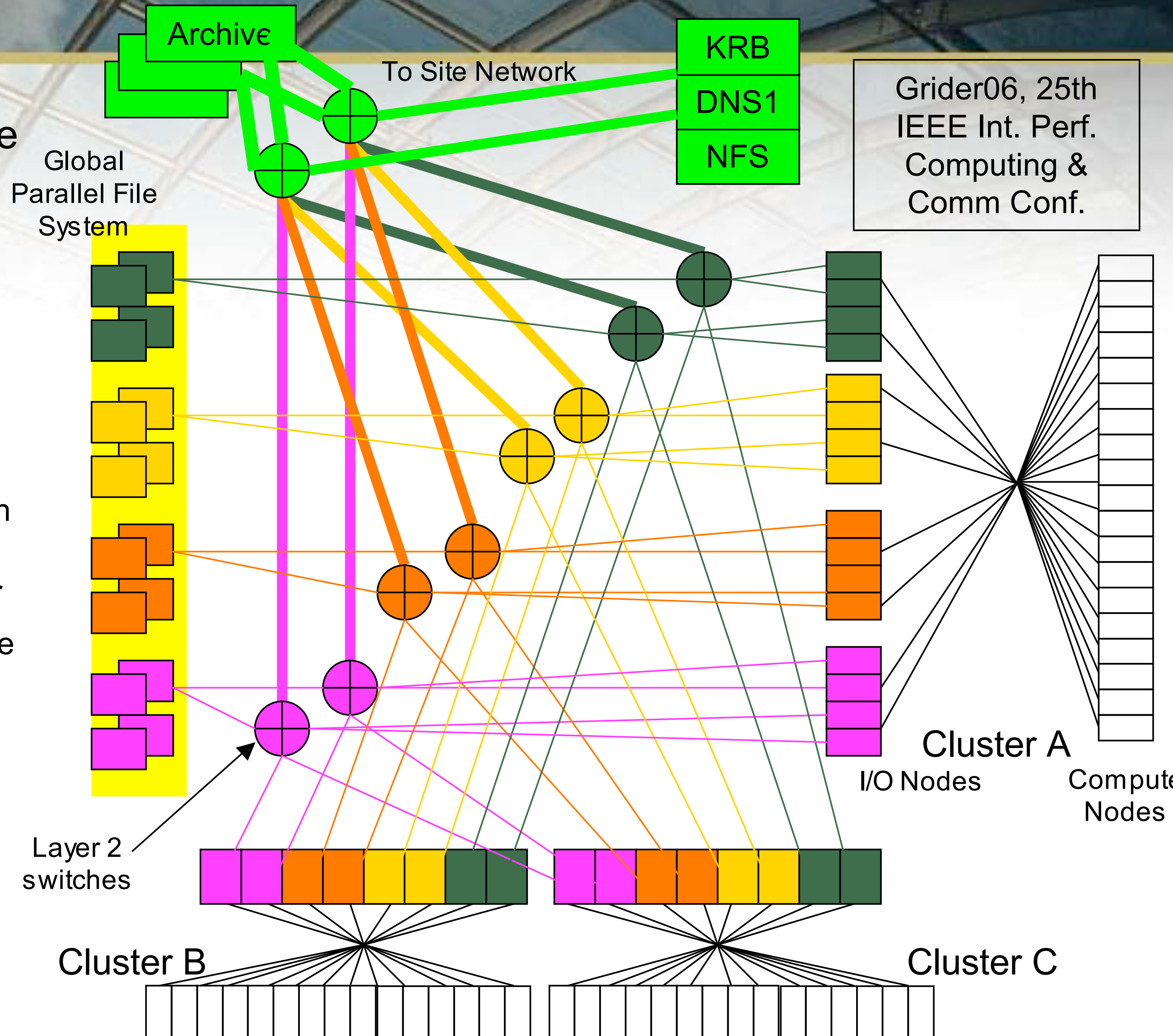
- Some checkpoints write small per proc records adjacent & unaligned then stride down & repeat
  - Kills 2/3rds of achievable BW
  - Was much worse b/c RAID locks
  - Number is data MB/s w/ LANL MPIIO test (min client speed, incl create/sync/close) 90 clients, 1 process per client
- Users rejected middleware lib, so …
- Supporting "tight & unaligned" N-1
  - Per-file RAID 10: 2 IO writes vs 4 IO writes
  - Trust the apps (if opened in CW mode):
    - No locking on redundant data
      - Exploit byte addressable OSD
    - Huge overlapping escrow/maps
  - Page unaligned: stay out of Linux buffer $
    - Write exact byte range immed but asynch

writesz/ RAID5 4 shelves	4096KB	64KB	63KB	65KB
N-N	1167	1190	1109	1138
N-1 contig	688	652	389	457
N-1 strided	681	442	402	**397**

writesz/ RAID10 8 shelves	4096KB	64KB	63KB	65KB
N-N	959	885	908	1099
N-1 contig	849	852	839	838
N-1 strided	843	808	820	**820**

Multi-Cluster : scalable BW w/ failover

- "Lane" Architecture
  - Share storage over many clusters
  - Multi-subnet IP routing
    - Parallel routing
  - Load balancing
    - Client multi-path routes (Linux)
  - Network Failover
    - Dual net storage
    - Multiple IO Node & switch per Lane
  - Incremental growth
    - Storage
    - Lane switches
    - I/O Nodes



Grider06, 25th IEEE Int. Perf. Computing & Comm Conf.

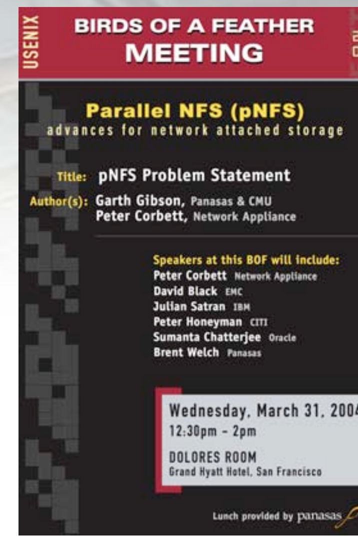
pNFS - Parallel NFS

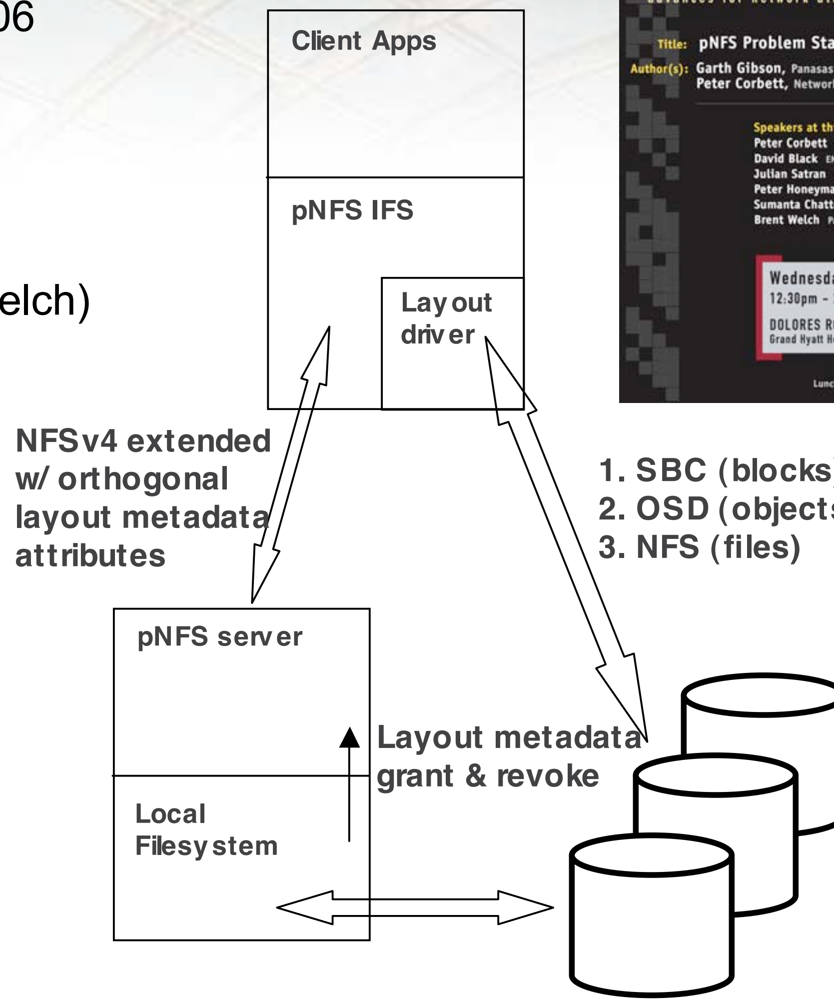
April 27, 2006

G. Gibson, Panasas

Parallel NFS: Scalability for Mainstream

- IETF NFSv4.1
  - draft-ietf-nfsv4-minorversion1-02.txt 3/06
  - Includes pNFS, stronger security, sessions/RDMA, directory delegations
  - U.Mich/CITI impl'g Linux client/server
  - www.panasas.com/webinar.html (B. Welch)
- Three (or more) flavors of out-of-band metadata attributes:
  - FILES: NFS/ONCRPC/TCP/IP/GE for files built on subfiles NetApp, Sun, IBM, U.Mich/CITI
  - BLOCKS: SBC/FCP/FC or SBC/iSCSI for files built on blocks EMC (-pnfs-blocks-00.txt)
  - OBJECTS: OSD/iSCSI/TCP/IP/GE for files built on objects Panasas, Sun (-pnfs-obj-00.txt)





Summary

April 27, 2006

G. Gibson, Panasas

Alternative solution philosophy

Make non-COTS features "easy" for DFS to provide

- depend only on big market features: large capacity, manageability

Revise: simple BW "easy"; increasing async & failure scope are not

High-bandwidth: direct transfer between app and device

- network-attached storage on scalable storage area networks
- server machine specs do not define peak storage bandwidth

We're good here

Concurrent-writers: middleware in app, little in DFS

- MPI-IO

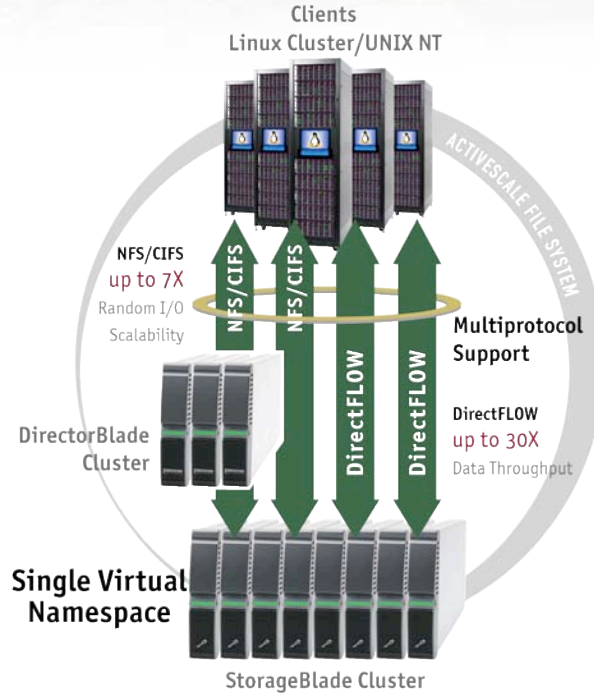
Revise: programmers weren't listening and may not until we fail :-(

Carnegie Mellon

Scale Out Clustering is Actually Many Things

- Applies HPC Clustering concepts to storage in many dimensions
  - Achieves new levels of Performance, Reliability and Manageability
  - Delivers on the “Scale-Out” promise

Benefit	Technology	Scale
Performance	Bandwidth Clustering (Parallel, direct access)	10GB/s
	NAS Clustering (N filers export same files)	70 + servers
	Cache Clustering (Support for large data sets)	unlimited
Reliability	Failover Clustering (N+1 active-active)	in Beta
	Recovery Clustering (Faster rebuild with scale)	10x faster
Manageability	Utilization Clustering (Balancing utilization)	file level
	Cluster Management (integrated h/w and s/w)	Petascale



Contact: garth@panasas.com

HPC wire
Editors' Choice Awards
Best Price / Performance Storage Solution
Panasas
Panasas ActiveScale File System
2005

Seismic Processing

TOTAL, bp, PGS, TGS, PetroChina 中国石油, MAXUS, Paradigm THE GEOSCIENCE KNOWLEDGE COMPANY, REPSOL YPF, SPINNAKER EXPLORATION

Life Sciences

Berkeley University of California, UC SANTA CRUZ, THE ROCKEFELLER UNIVERSITY, NOVARTIS, EMBL-EBI European Bioinformatics, UCSF, MERCK

HPC Simulations

ECMWF, NASA, Los Alamos NATIONAL LABORATORY, NW-GRID, BROOKHAVEN NATIONAL LABORATORY, Sandia National Laboratories, CCLRC, NORTHROP GRUMMAN

Rendering

Walt Disney FEATURE ANIMATION, School of VISUAL ARTS, Weta Digital

Fluid Dynamics

BOEING, FLUENT, Ferrari, Princeton University

Stephens Inc. Investment Bankers, BNP PARIBAS

COMPUTERWORLD INNOVATIVE TECHNOLOGY AWARDS 2004

STORAGE NETWORKING WORLD BEST PRACTICES IN STORAGE AWARDS PROGRAM COMPUTERWORLD SNIA

WALL STREET JOURNAL

WINNER INNOVATIVE PRODUCT AWARD NSC NETWORK STORAGE CONFERENCE

InformationWeek BUSINESS INNOVATION POWERED BY TECHNOLOGY

Bio·IT World

NetworkWorldFusion The leader in network knowledge VENDORS TO WATCH 2004

Excellence in CLUSTER Technology Storage Solutions WINNER ClusterWorld 2004

SOS10 Debrief: Predicting the Future

April 27, 2006

G. Gibson, Panasas

Back to the Panel on Complexity

- Clusters get bigger, applications get bigger, so why would storage getting bigger be any harder?
- Could it be that having every byte of tera- and petabyte stores available to all nodes with good performance for all but minutes a year, when files & volumes are parallel apps on the storage servers, might be a higher standard than compute nodes are held to? (failure…)
- Or perhaps it is deeper and deeper writebehind and readahead, and more and more concurrency, needed to achieve the ever larger contiguous blocks that are needed to minimize seeks in ever wider storage striping. (failure…)
- Or maybe Amdahl's law is hitting us with the need to parallelize more and more of the metadata work which has been serial and synchronous for correctness and error code simplicity in the past. (failure…)
- Or maybe parallel file systems developers have inadequate development tools in comparison to parallel app writers. (test…)
- Or perhaps storage system developers are just wimps. (nerds instead of geeks…)

BANDWIDTH

- 1) In the next decade is the bandwidth transferred into or out of one "high end computing file system"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d)"your answer here",
- as a result of the expected increase in computational speed in its client clusters/MPPs, and why?

- Garth (c): 30 GB/s to 1 TB/s is at least 10X
  - But in and of itself this is OK – Object storage scales

SPINDLE COUNT

- 2) In the next decade is the number of magnetic disks in one "high end computing file system"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d) "your answer here",
- as a result of the expected increase in computational speed in its client clusters/MPPs, and why?

- Garth (c): 10 year data rate increases (SQRT(MAD))^10
  - This is 8X to 10X based on MAD of 50-60%/yr
  - But if demand goes up 100X, spindle count is still up 10X

CONCURRENCY

- 3) In the next decade is the number of concurrent streams of requests applied to one "high end computing filesystem"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d) "your answer here",
- as a result of the expected increase in concurrency in client clusters/MPPs, and why?
- Garth (c): many cores*sockets instead of faster cores
  - Lots more threads, concurrent accesses to storage
  - Seq. data access OK, but metadata concurrency harder

SEEK EFFICIENCY

- 4) In the next decade is the number of bytes moved per magnetic disk seek in one "high end computing file system"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d) "your answer here",
- as a result of the expected increase in computational speed in its client clusters/MPPs, and why?

- Garth (b): Possible but not obvious for read/write calls to move more data each, while the cry for 32,000 small file creates/sec means lots more tiny writes
  - Mechanical positioning may continue to hurt big time
  - But file systems still may be faster than DBs for this :-(

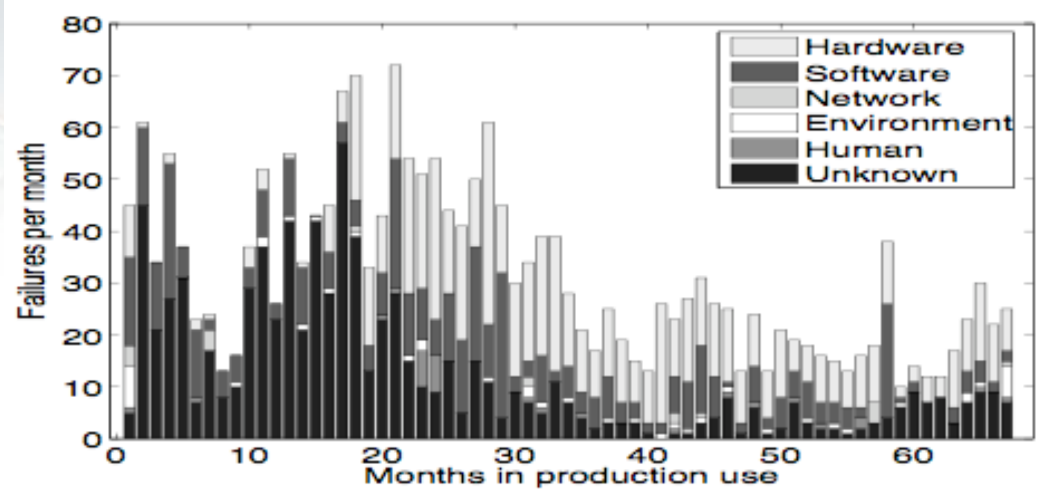
FAILURES

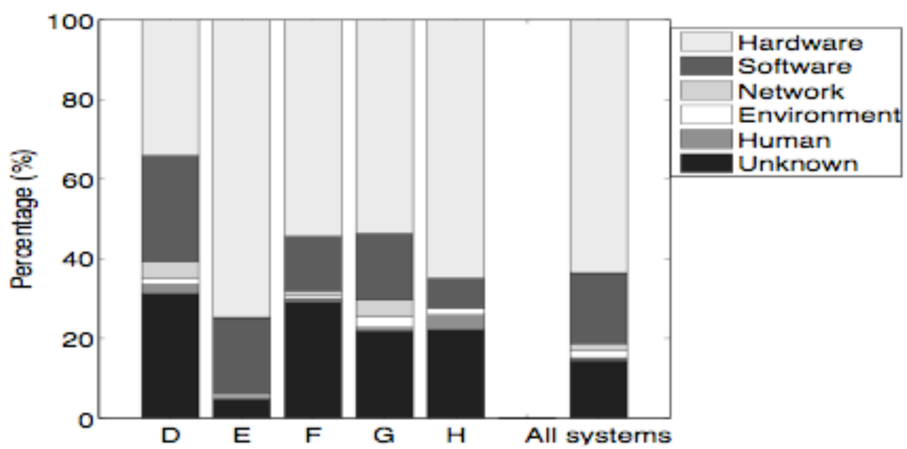
- 5) In the next decade is the number of independent failure domains in one "high end computing file system"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d)"your answer here",
- and why?

- Garth (c): as a direct result of all those spindles and and cables
  - All the hard problems come down to the failure cases
  - An now for some interesting data ……

9 YEARS of LANL HEC FAULT DATA

- Failure characteristics differ system to system in rates, causes, and are not stationary over time
- Virtual no widely shared hard data on how HEC computers fail
- Schroeder, DSN06





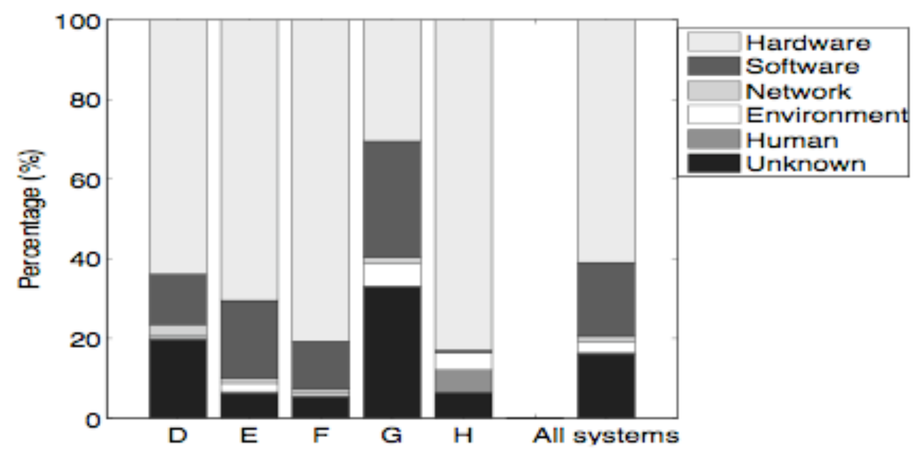


Figure 1: *The breakdown of failures into root causes (left) and the breakdown of downtime into root causes*

COPING WITH COMPLEXITY

- 6) If you have answered (c) one or more times,
  - please explain why these large increases are not going to increase the complexity of storage software significantly?
  - Are you relying on the development of any currently insufficient technologies, and if so, which?

- Garth: Storage developers are at risk here
  - Scaling BW I think we can do
  - Doing that without loss of 9s is hard
  - But scaling metadata rates w/ POSIX consistency is hard
  - Interesting technology: Model checking, for protocol correctness

DEVELOPMENT TIME TRENDS

- 7) If complexity is increasing in high end computing file systems, is the time and effort required to achieve acceptable 9s of availability at speed
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d) "your answer here",
- and why? Are you relying on the development of any currently insufficient technologies, and if so, which?

- Garth (b-c): Can't face 10X up, but it is increasing
  - Testing can be a big drag with rapidly changing OS/platform
  - To repeat: model checking is interesting

Contact: garth@panasas.com

Seismic Processing

Life Sciences

HPC Simulations

Rendering

Fluid Dynamics

Next Generation Network Storage

Garth Gibson
garth@panasas.com

April 27, 2006

Panasas in Oil and Gas

TGS Imaging

- Customer Profile
  - Seismic processing outsource company for the energy industry
  - Delivers massively parallel systems to accelerate solutions for scientific discovery
- Challenge
  - Find storage compliment to recent Linux cluster purchase
  - Maximize price-performance and simplify management
- Results
  - 10X performance improvement in seismic analysis
  - 225 TB in production to date
  - Integrated HW/SW solution simplifies management
  - Commodity components over GE maximum price -performance

"We are extremely pleased with the order of magnitude performance gains achieved by the Panasas system. With other products, we were forced to make trade-offs, but with the Panasas system, we were able to get everything we needed and more."

Tony Katz
Manager, Information Technology
TGS Imaging

Panasas in Action: Oil and Gas

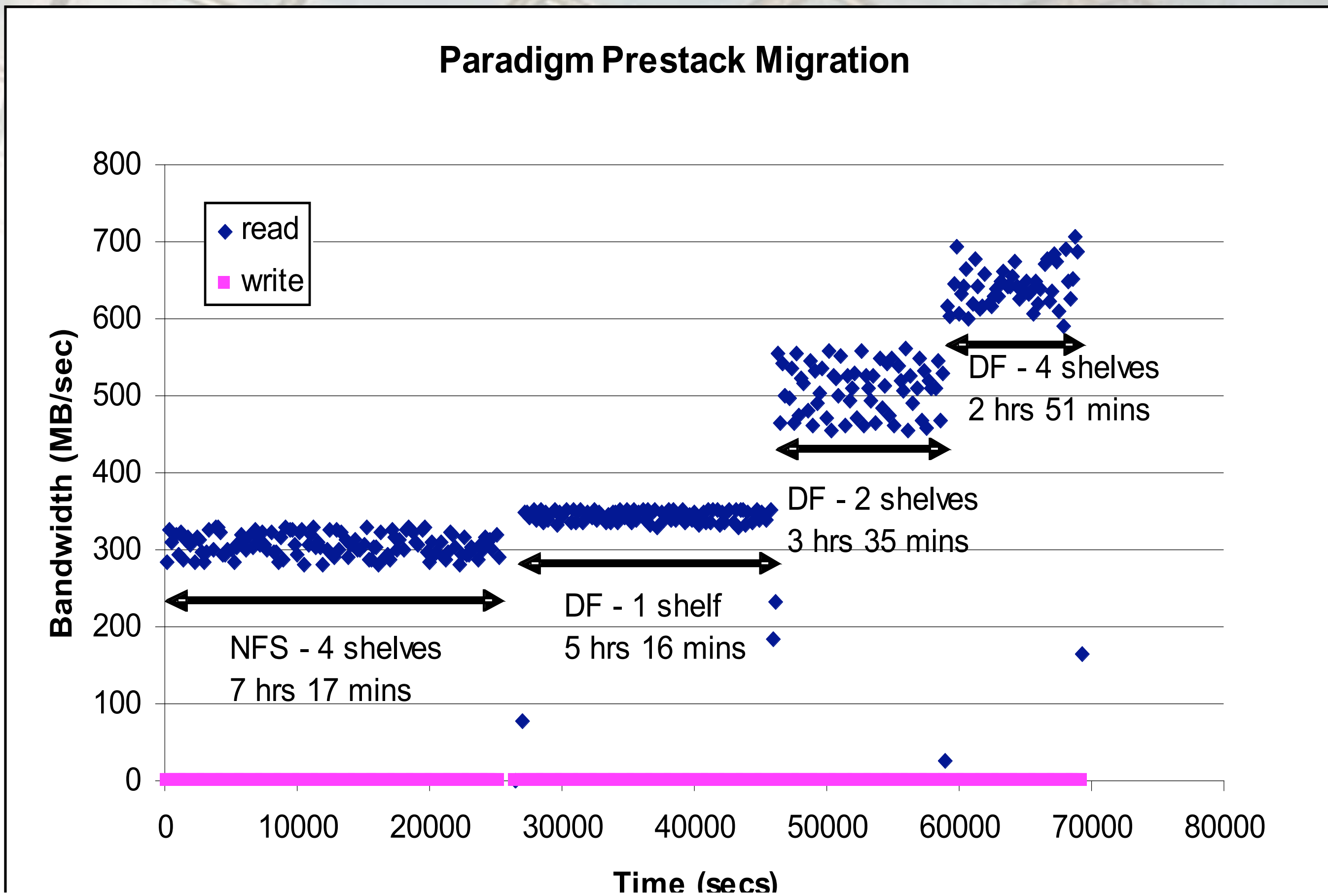
Petroleum Geo-Services Corporation (PGS)

- The customer
  - Seismic processing outsource company with offices around the world
  - Delivers massively parallel systems to accelerate solutions for Oil and Gas discovery
- The challenge
  - Deliver higher performance storage solution for worldwide seismic processing operations
  - Simplify storage management to minimize IT resources in remote processing offices
- The solution
  - Over 200 TB worldwide
  - Installations in Houston, Walton on Thames, Kuala Lumpur, Cairo, Lagos, Nigeria, Azerbaijan, Perth
    - More worldwide sites planned, some on ships
- The value
  - Very high performance for parallel IO in seismic analysis
  - Integrated HW/SW solution simplifies management
  - Commodity components over GE maximize price-performance

"The large data sets with which we work require very high bandwidth in order to process data as fast as possible. After evaluating several storage products, none offered the compelling performance and ease-of-management that we receive with Panasas. The Panasas DirectFLOW data path allows us to avoid partitioning the cluster with expensive connections in order to keep up with our heavy bandwidth requirements.

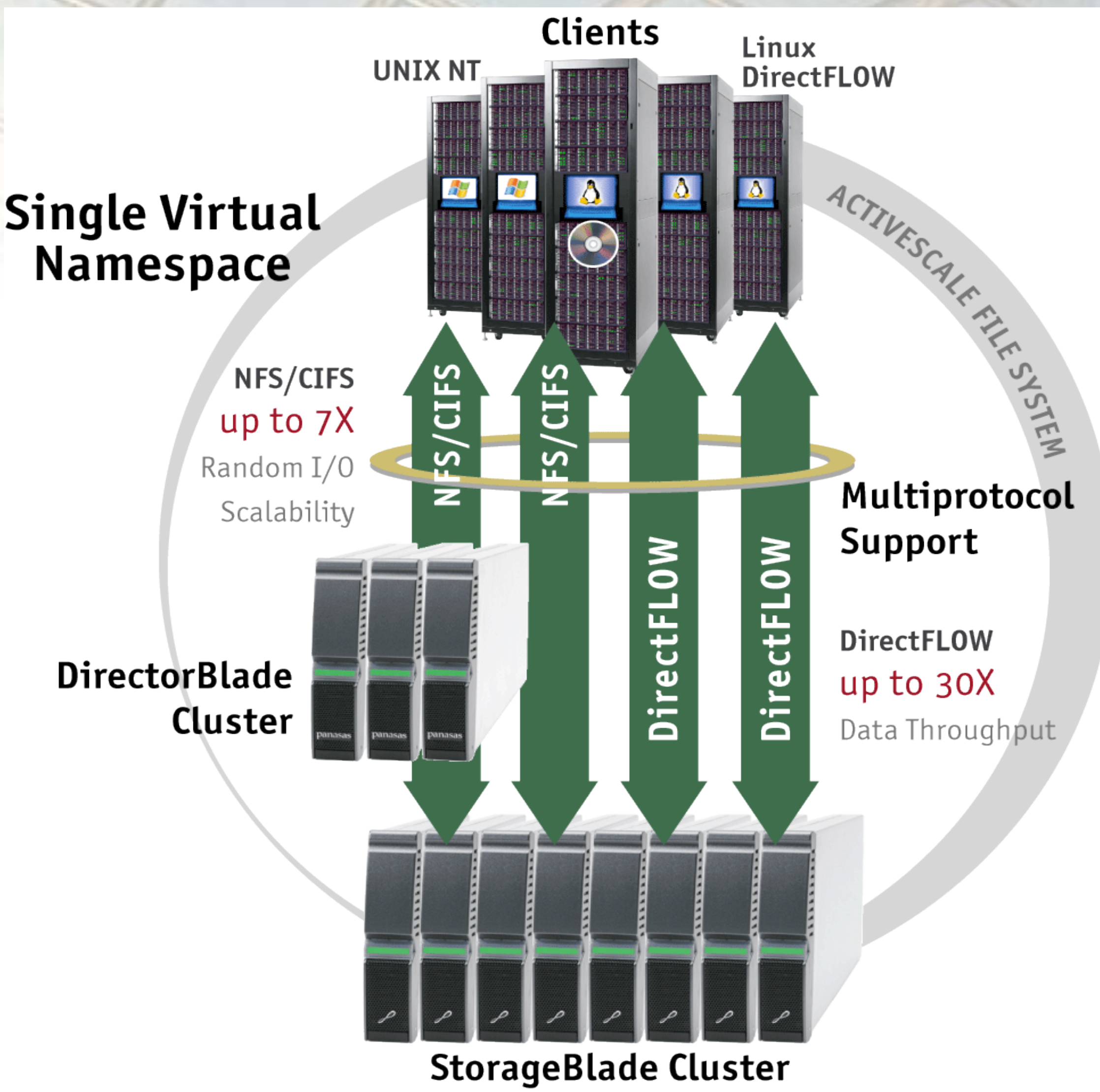
Andy Wrench
DP Computer Systems Manager
PGS Global Computer Resources

Seismic Performance



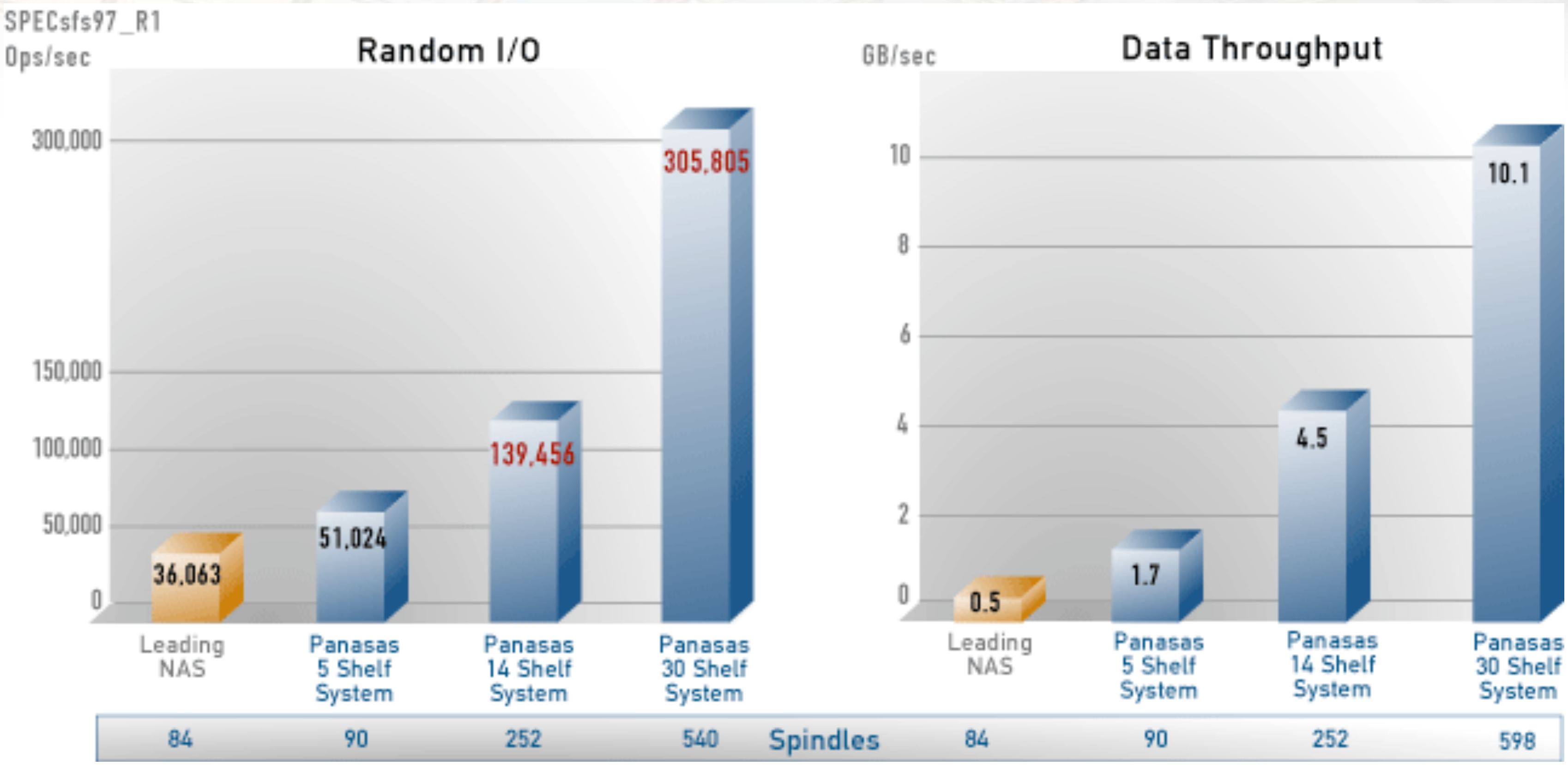
- Testing Results
  - GeoDepth has a parallel I/O architecture that takes advantage of our DF scalability
  - DirectFLOW is 3x faster than our own large scalable NFS configuration

Out-of-band DF & Clustered NAS



Performance & Scalability for all Workloads

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



Panasas in Action: Media

Walt Disney Feature Animation

- The customer
  - Creative unit of The Walt Disney Studios producing animated films
  - 45 films, 106 Oscar nominations, 31 Academy Awards
- The challenge
  - Production going all CGI: 700M files, 30TB, 1K render nodes
  - Maximize performance and simplify management
- The solution
  - Twenty seven 5 TB Panasas Storage Cluster shelves (135 TB)
  - First all-CGI, all-Panasas film, Chicken Little, $125M US revenue
    - Four more animated films in the pipeline
- The value
  - Lowered time to market for computer generated animated films
  - 150,000+ ops/sec, 500+ MB/s over scalable NFS, 3-14X predecessor
  - Simplified operations by consolidating NFS servers, 30% less mgmt OV