

Brent Welch
Director, Architecture

Panasas Technology

HPC Advisory Council

Lugano, March 2011



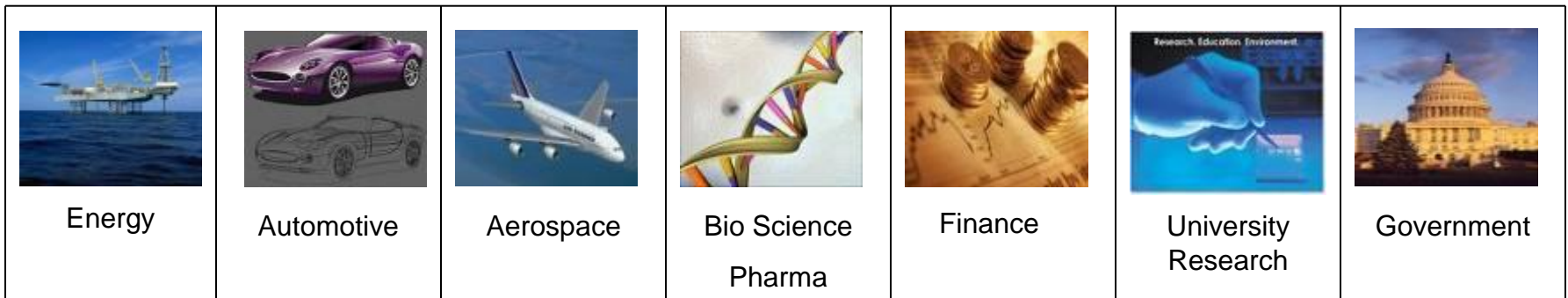
Panasas Background

- Technology based on Object Storage concepts invented by company founder, Garth Gibson, a pioneer in RAID technology
 - 1990's research on Network Attached Secure Disk (NASD) lead to (Object Storage Device) OSD standards for iSCSI.
- Storage Systems Company
 - Integrated hardware and software storage systems
 - Hardware, FS, RAID
 - 4th generation blade platform
 - HC storage platform
- Customer base
 - 2/3 Commercial
 - 1/3 Research and Higher Ed



Customer Success Highlights

- Half the Formula One teams use Panasas
- Five of the top six O&G companies in the world use Panasas
- The world's first Petascale system, RoadRunner, uses Panasas
- The world's three largest genomic data centres use Panasas
- The largest aircraft manufacturer in the world uses Panasas
- Leading Universities including Cambridge, Oxford, Stanford & Yale use Panasas
- The world's largest Hedge Fund uses Panasas



Technical Differentiation



- Scalable Performance
 - start with 12 servers -> grow to 12,000
- Novel Data Integrity Protection
 - File system and RAID are integrated
 - highly reliable data w/ novel data protection systems
- Maximum Availability
 - built-in distributed system platform, tested under fire
 - LANL measured > 99.9% availability across its Panasas systems for all reasons
- Simple to Deploy and Maintain
 - integrated storage system with appliance model
- Application Acceleration
 - customer proven results

Data Integrity

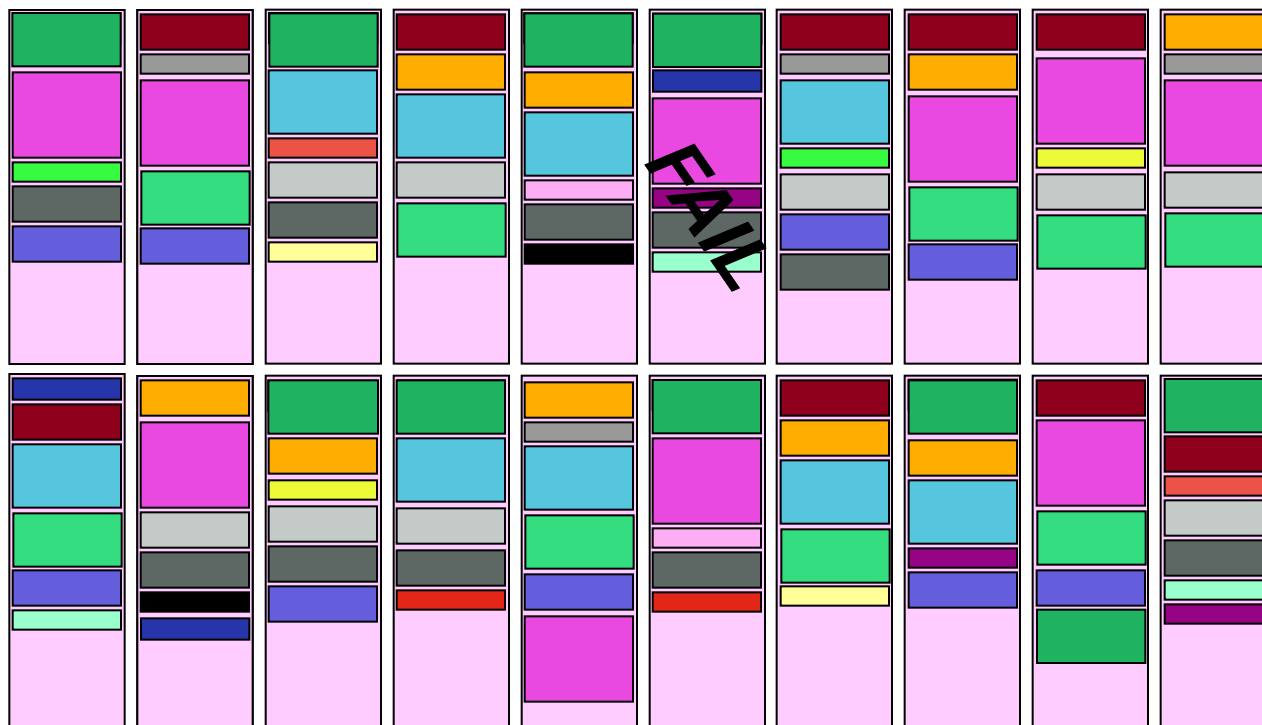
- Object RAID
 - Horizontal striping with redundant data on different OSDs
 - Per-file RAID equation allows multiple layouts
 - Small files are mirrored RAID-1
 - Large files are RAID-5 or RAID-10
 - Very large files use two level striping scheme to counter network incast
- Vertical Parity
 - RAID across sectors to catch silent data corruption
 - Repair single sector media defects
- Network Parity
 - Read back per-file parity to achieve true end-to-end data integrityg
- Background scrubbing
 - Media, RAID equations, distributed file system attributes

Declustered Object RAID

- Each file striped across different combination of StorageBlades
- Component objects include file data and file parity
- File attributes replicated on first two component objects
- Components grow & new components created as data written
- Declustered, randomized placement distributes RAID workload

**20 OSD
Storage Pool**

**Mirrored
or 9-OSD
Parity
Stripes**



**Read
about
half of
each
surviving
OSD**

**Write a
little
to each
OSD**

**Scales up
in larger
Storage
Pools**

Panasas Manageability

- Appliance-like quick setup and ease of use
- Automatic discovery of new equipment with transparent load balancing and capacity balancing
- A “no knobs” attitude towards performance tuning
- Detailed performance charts and reporting
- Transparent software and hardware upgrades
- Fully integrated automatic fail over
 - Realm, MetaData, Data, Network, NFS/CIFS
- Continuous Improvement
 - This job is never done: systems get larger, new problems emerge
 - Tight feedback loop between experienced service organization and product development

panwest System-At-A-Glance


[? Help](#) |
 [Legend](#) |
 [Sign Out](#)

System-At-A-Glance

- StorageBlades
- Drive Statistics
- DirectorBlades
- DirectFLOW Clients
- NFS Clients

Event Viewer

Reports

System Status: 

System Is Locked

User Name:
admin

System Name:
panwest

System Uptime:
3 Days 19:09:56

Status

[Download System Summary](#)

Blades



No errors.

Storage



1 warning.

BladeSet capacity imbalance detected

Events

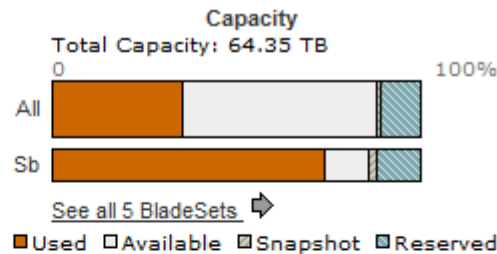


No errors.

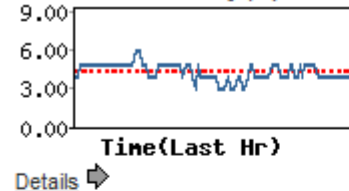
System Tasks

No tasks in progress.

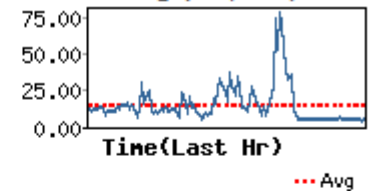
StorageBlades



Disk Activity (%)

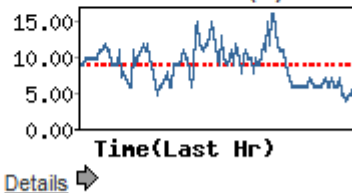


Throughput (MB/s)

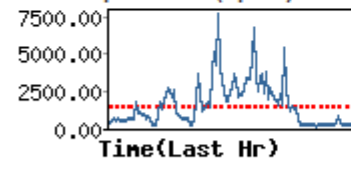


DirectorBlades

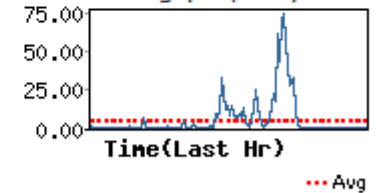
CPU Utilization (%)



Operations (ops/s)



Throughput (MB/s)








Page Refresh Rate:

Panasas Distributed System Platform

- Problem: managing large numbers of hardware and software components in a highly available system
 - What is the system configuration?
 - What hardware elements are active in the system?
 - What software services are available?
 - What is the desired state of the system?
 - What components are failed?
- Answer: The Panasas Realm Manager
 - 3-way or 5-way Redundant “Master Control Program”
 - Distributed file system one of several services managed by the RM
 - Configuration management
 - Software upgrade
 - Failure Detection
 - GUI/CLI management
 - Hardware monitoring

Configuring the RM Replication Set

System Clustering

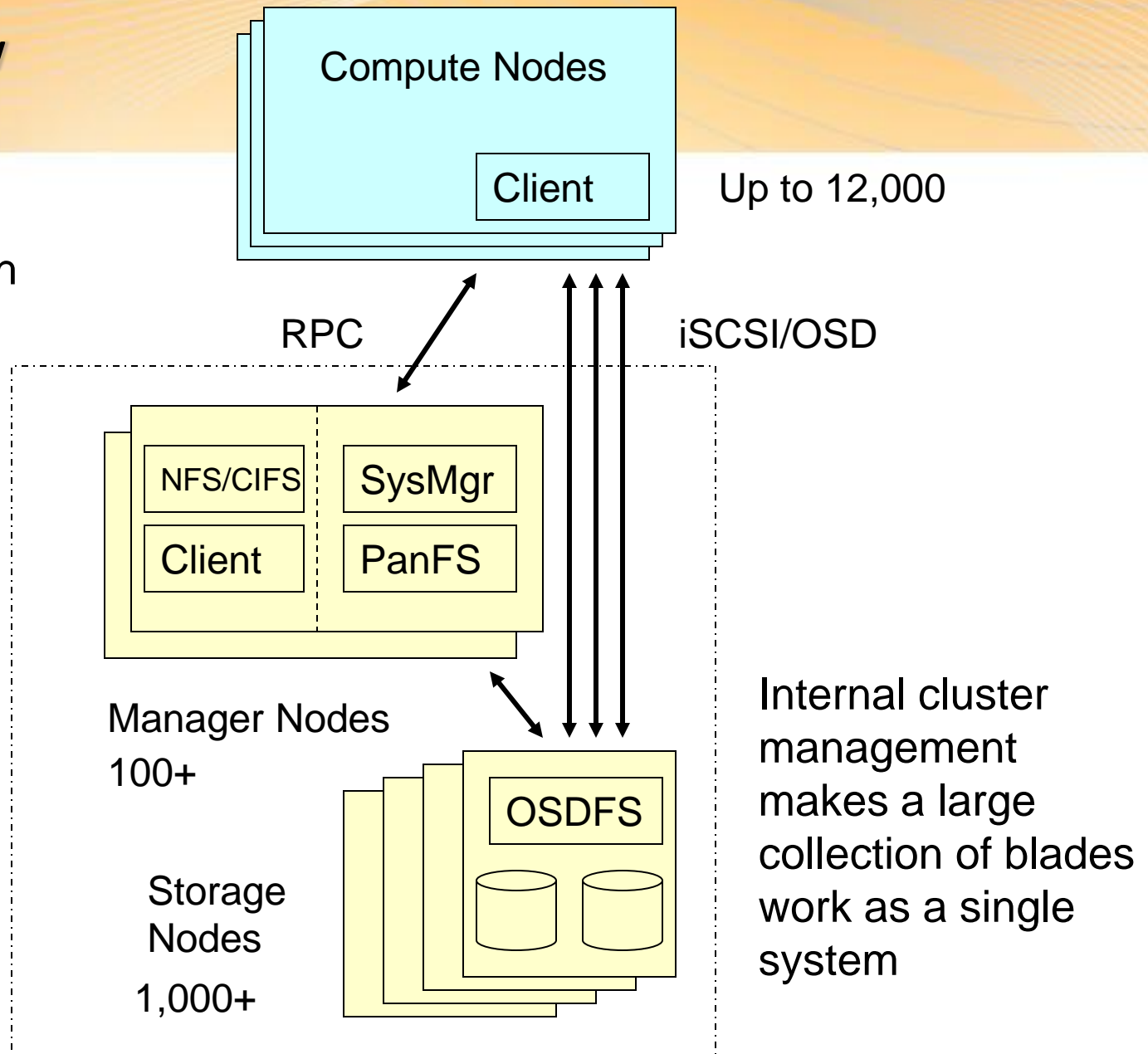
	Shelf	Slot	IP Address	Role	Status
	panwest-6	1	10.0.8.88	RM President, DNS	Online
	panwest-2	1	10.0.8.2	RM, DNS	Online
	panwest-4	1	10.0.8.5	RM, DHCP, DNS	Online
	panwest-3	1	10.0.8.4	RM, DNS	Online
	panwest-5	1	10.0.8.41	RM, MConsole, DNS	Online

IP Address

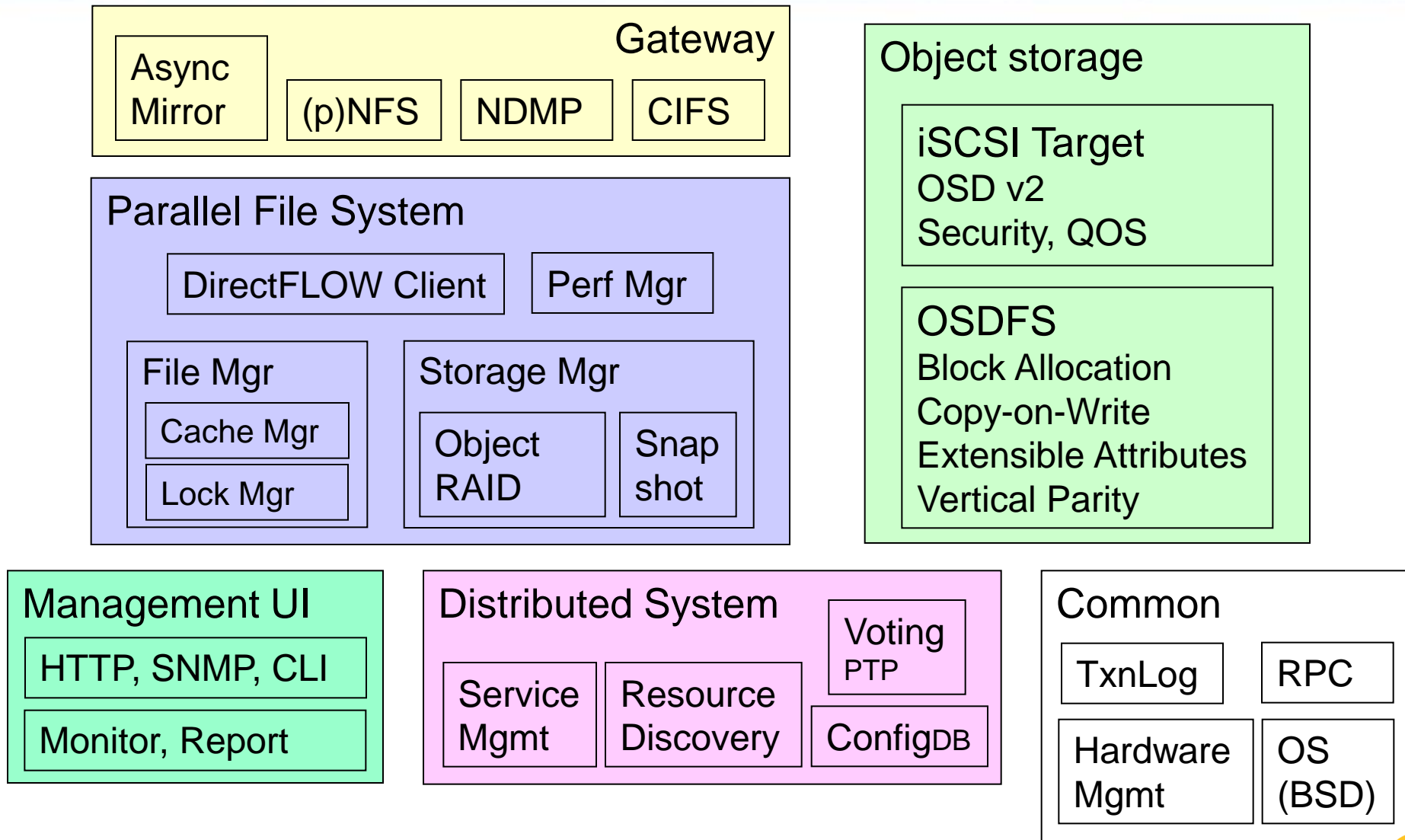
A set of DirectorBlades can be configured to host critical services and replicate system configuration data. For service, the management console, and supporting services like DHCP and delegated DNS. An odd number of blades provides redundancy.

System View

Out of Band architecture with direct, parallel paths from clients to storage nodes






















Functional Architecture



Managing metadata services

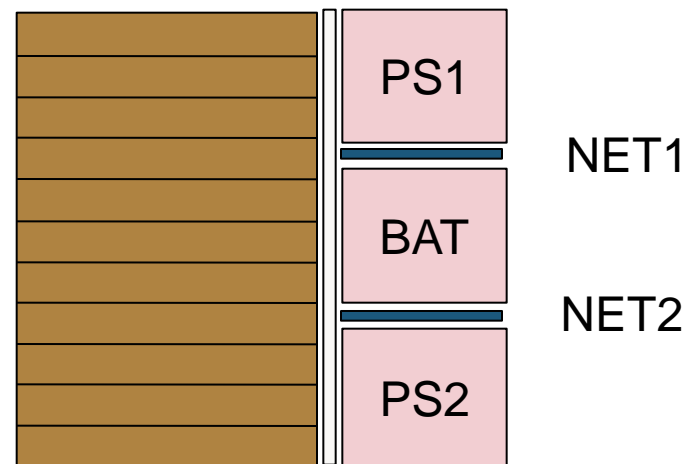
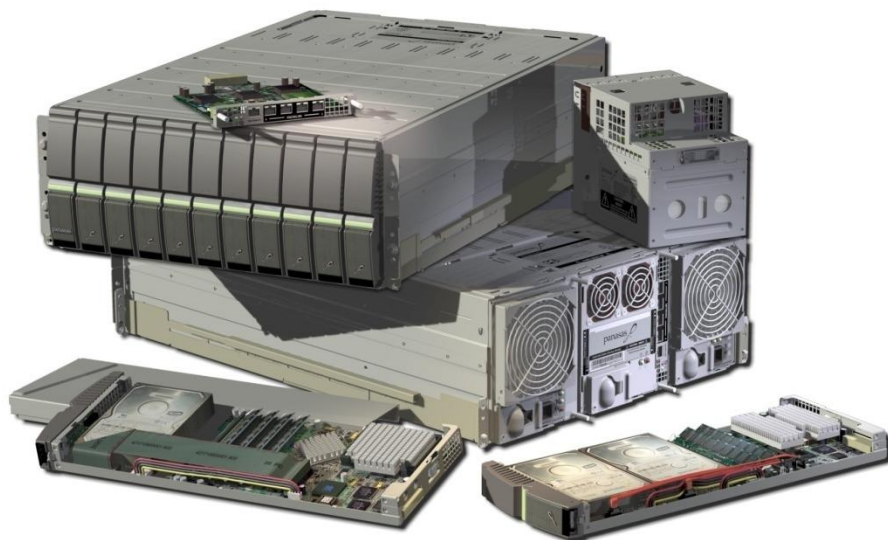
Volume Services Listing

Blade	Primary Service Location	Service ID	Backup Service Location	Volume Count	Volumes
	panwest-6 Slot 1 10.0.8.88	 0x040013670b0f001d(FM)	panwest-2 Slot 1 10.0.8.2	6	/archive /db /home /http /perforce /rsync_data
		 0x040013670b0f001e(FM)	panwest-5 Slot 1 10.0.8.41	1	/
	panwest-2 Slot 1 10.0.8.2	 0x0400247b0e7f0002(FM)	panwest-7 Slot 1 10.0.8.89	1	/sb
	panwest-2 Slot 2 10.0.8.76	 0x0400000000000005e(FM)	panwest-6 Slot 1 10.0.8.88	0	
	panwest-4 Slot 1 10.0.8.5	 0x04007140158f0281(FM)	panwest-4 Slot 2 10.0.8.12	1	/scratch1
	panwest-4 Slot 2 10.0.8.12	 0x04007164158f0286(FM)	panwest-4 Slot 1 10.0.8.5	1	/sb4
	panwest-3 Slot 1 10.0.8.4	 0x04002cb40f7f011f(FM)	panwest-3 Slot 2 10.0.8.43	2	/pansouth /sb1
	panwest-3 Slot 2 10.0.8.43	 0x040016dd0c1f0293(FM)	panwest-3 Slot 1 10.0.8.4	2	/backup /sb2
	panwest-5 Slot 1 10.0.8.41	 0x0400711a157f0070(FM)	panwest-7 Slot 1 10.0.8.89	4	/qa-daily /sb3 /software /testzilla
	panwest-7 Slot 1 10.0.8.89	 0x040025520e7f0003(FM)	panwest-2 Slot 1 10.0.8.2	1	/zilla

PAS-12 4th Generation Blade

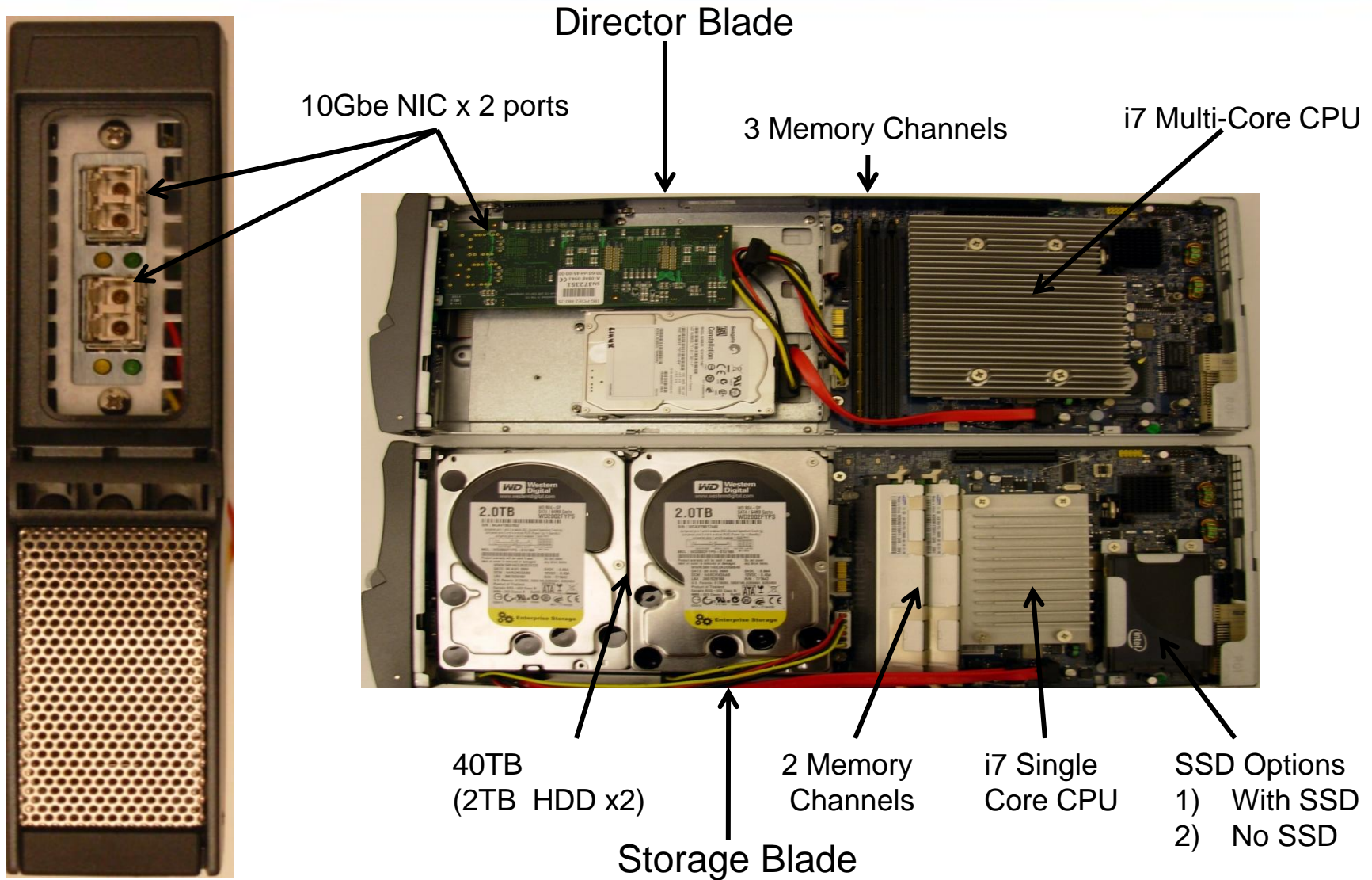
Panasas Active Scale

2002	850 MHz / PC 100	80 GB PATA	
2004	1.2 GHz / PC 100	250 GB SATA	330 MB/sec
2006	1.5 GHz / DDR 400	500 GB SATA	400 MB/sec
2008	10 GE shelf switch	750 GB SATA	600 MB/sec
2009	SSD Hybrid	1000 GB SATA, 32GB SSD	600 MB/sec
2010	1.67 GHz / DDR3 800	2000 GB SATA, (64GB SSD)	1.5 GB/sec



11x Blades

PAS-12 Blade Configurations



Managing Hardware

Total DirectorBlades: 9		Total StorageBlades: 57	Total Shelves: 6
Status	Shelf		
	Shelf: panwest-2 BladeSet: Sb Vertical Parity: disabled		
	Shelf: panwest-3 BladeSet: Sb1 Sb2 Vertical Parity: disabled		
	Shelf: panwest-4 BladeSet: Sb Vertical Parity: disabled		
	Shelf: panwest-5 BladeSet: Sb3 Vertical Parity: disabled		
	Shelf: panwest-6 BladeSet: Home Vertical Parity: enabled		
	Shelf: panwest-7 BladeSet: hybrid Vertical Parity: enabled		

Controls

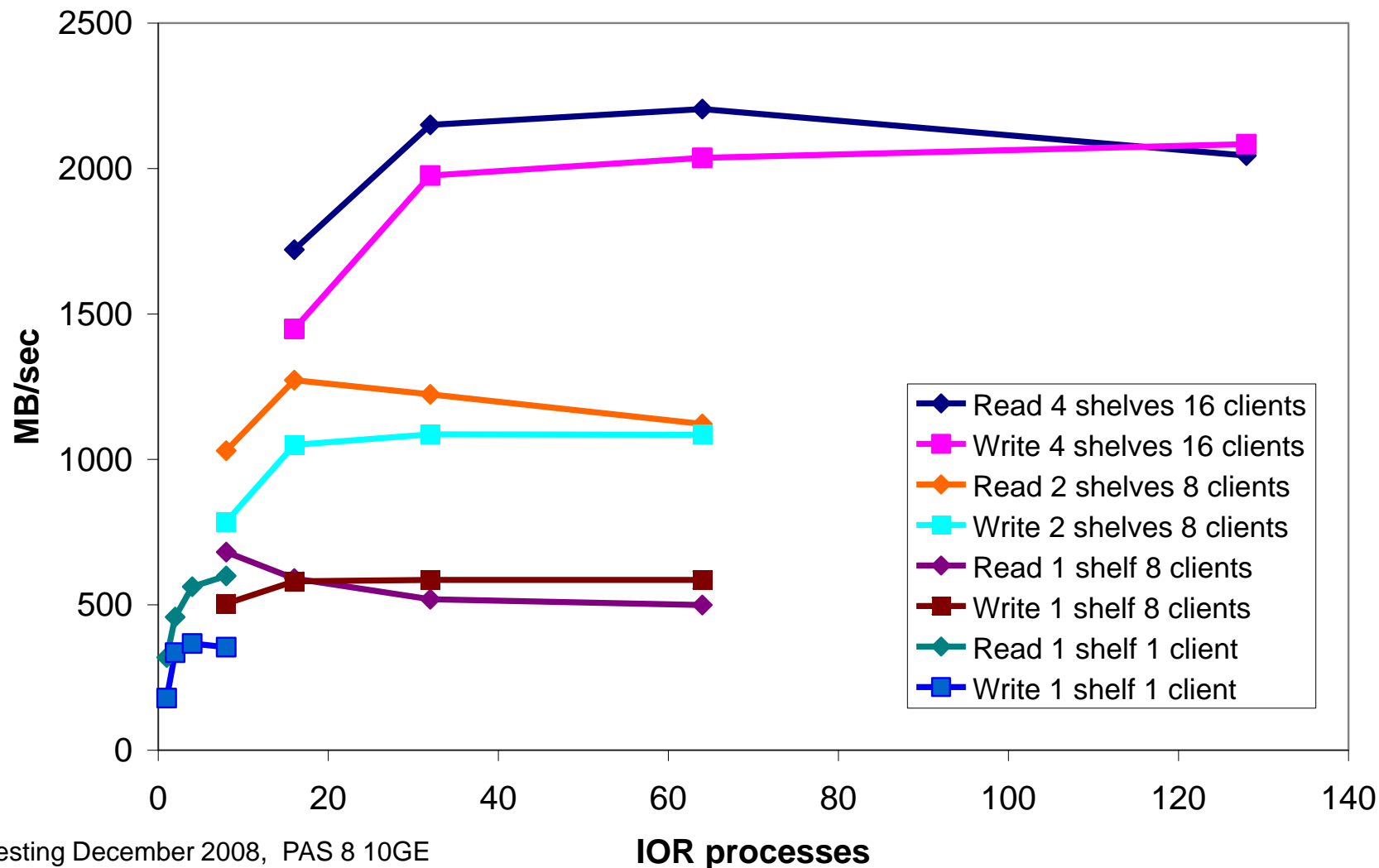
Panasas Features

- Object RAID (2003-2004)
 - NFS w/ multiprotocol file locking (2005)
 - Replicated cluster management (2006)
 - Declustered Parallel Object RAID rebuild (2006)
 - Metadata Fail Over (2007)
 - Snapshots, NFS Fail Over, Tiered Parity (2008)
 - Async Mirror, Data Migration (2009)
 - SSD/SATA Hybrid Blade (2009)
 - 64-bit multicore (2010)
 - User Group Quota (2010)
- Dates are the year the feature shipped in a production release



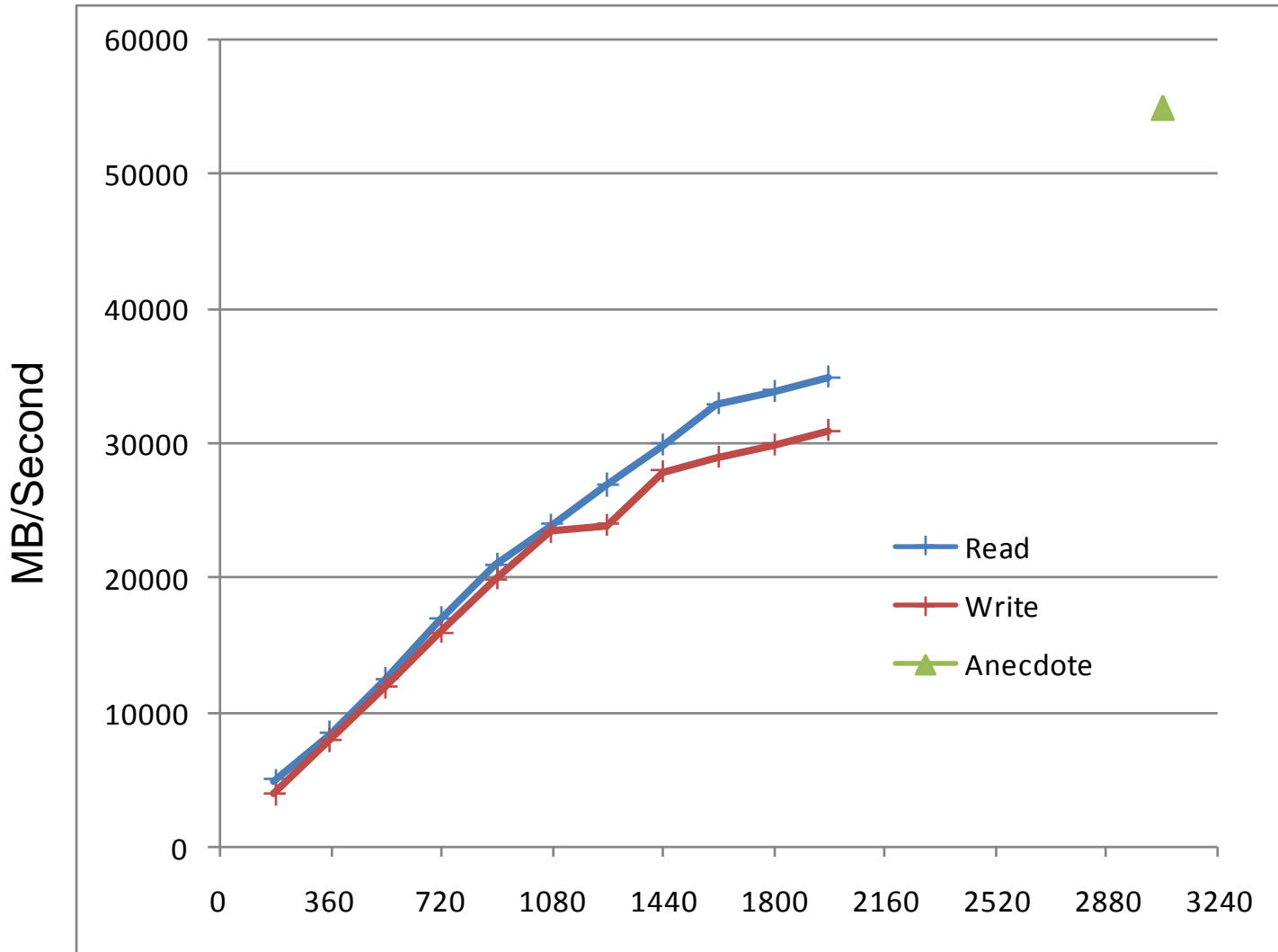
Scaling Performance with Capacity

Shelf Scaling



3.4 testing December 2008, PAS 8 10GE

Scaling Clients (100 shelves, 1000 Blades)

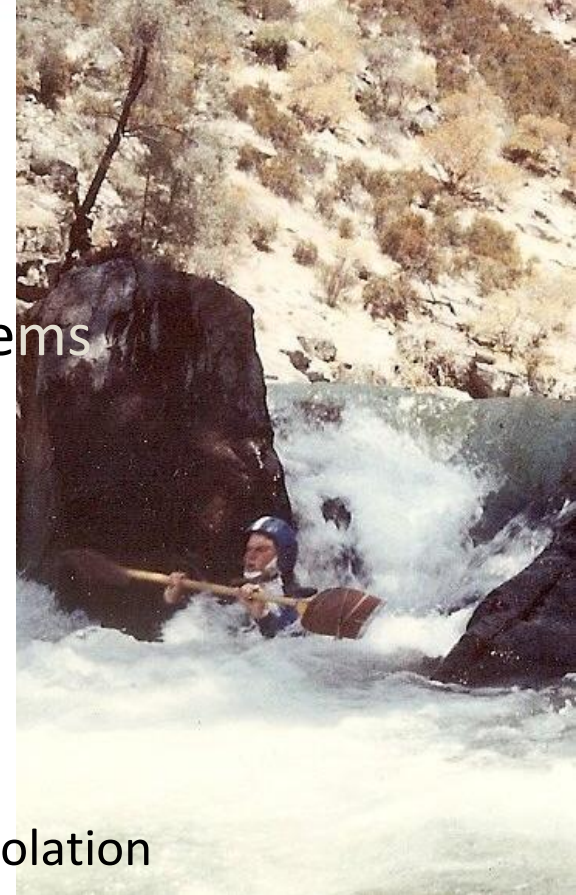


“Anecdote” is a 55 GB/sec observation during a full machine checkpoint restart

The read/write results are from early tests when about half the machine was available

Technical Philosophy

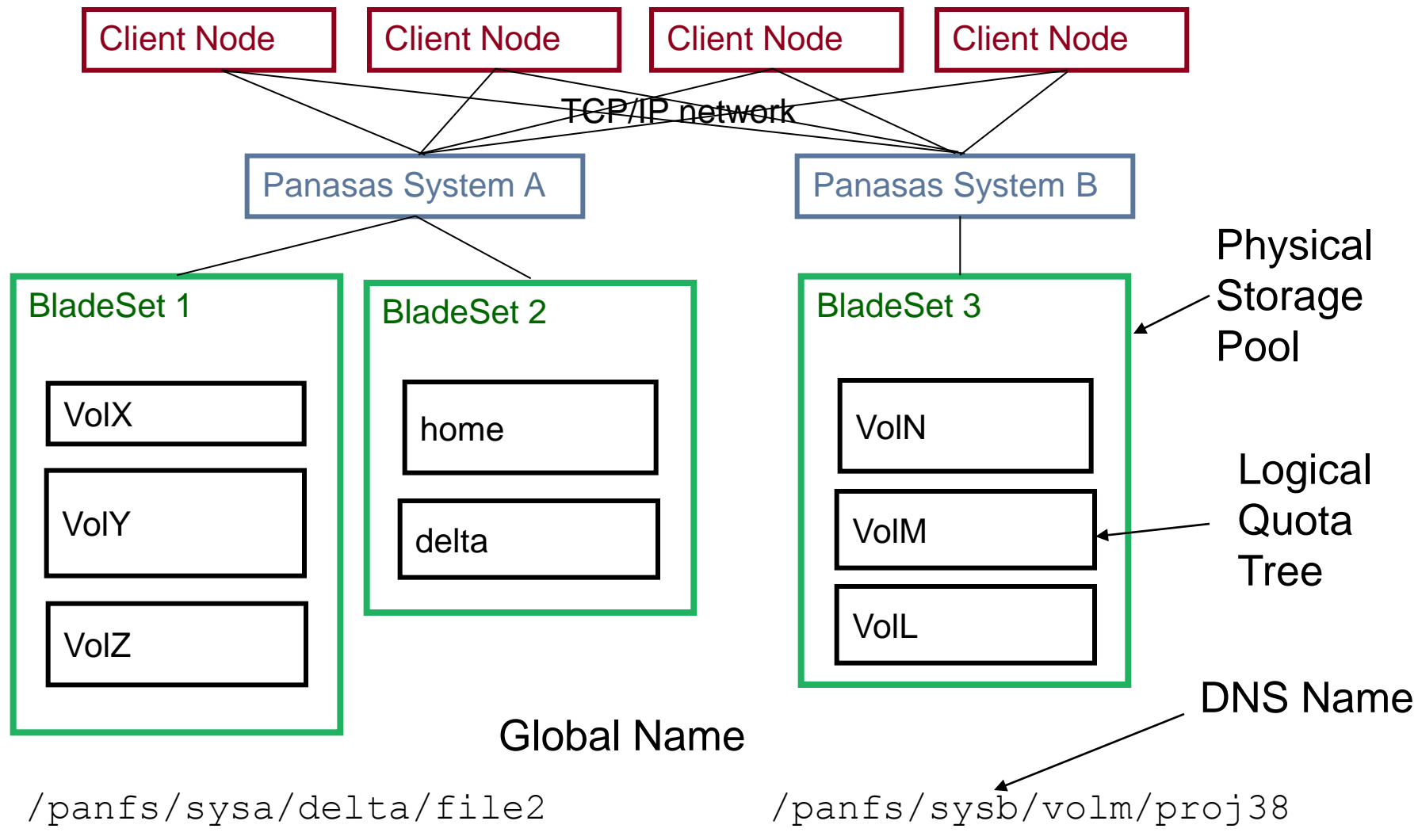
- Storage is Hard
 - Never fail, ever scale, wire-speed goals
 - Built from low-cost, flakey hardware
- Fault handling is the key to building large systems
 - Performance comes naturally if you can scale up
- Panasas layers its parallel file system on top of its distributed system platform
- Data integrity becomes more critical in today's very large systems
 - Object-RAID provides per-file protection and fault isolation
 - Vertical parity provides protection from media defects and firmware bugs
 - Network parity provides end-to-end data integrity



Thank You

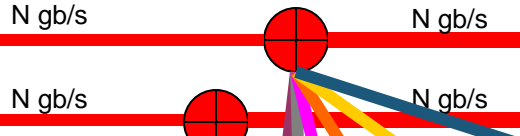


Panasas Global Storage Model



LANL Petascale Network (Red) FY08

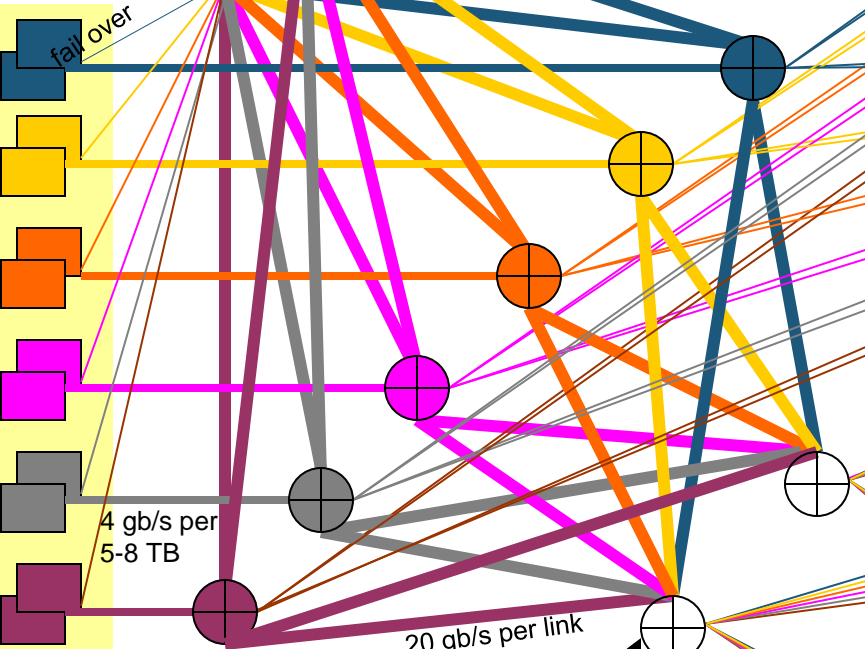
NFS complex and other network services, WAN



Archive



Site wide Shared Global Parallel File System
650-1500 TB
50-200 GB/s (spans lanes)



4 gb/s per 5-8 TB

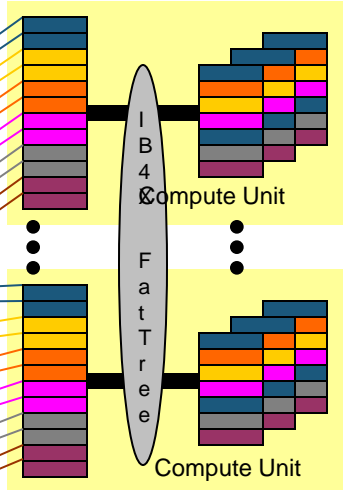
20 gb/s per link

Lane switches, 6 X 105 = 630 GB/s

Lane passthru switches, to connect legacy Lightning/Bolt

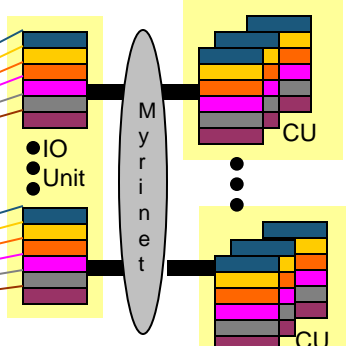
If more bandwidth is needed we just need to add more lanes and add more storage, scales nearly linearly, not N^2 like a fat tree Storage Area Network.

156 I/O nodes, 1 - 10gbit link each, 195 GB/sec, planned for Accelerated Road Runner 1PF sustained



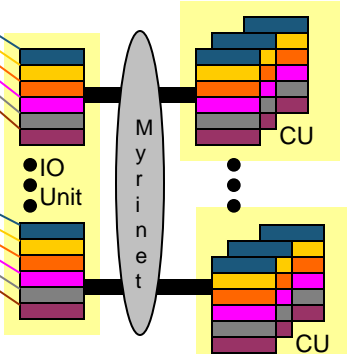
Road Runner Base, 70TF, 144 node units, 12 I/O nodes/unit, 4 socket dual core AMD nodes, 32 GB mem/node, full fat tree, 14 units, Acceleration to 1 PF sustained

96 I/O nodes, 2 - 1gbit links each, 24 GB/sec



Bolt, 20TF, 2 socket, single/dual core AMD, 256 node units, reduced fat tree, 1920 nodes

64 I/O nodes, 2 - 1gbit links each, 16 GB/sec



Lightning, 14 TF, 2 socket single core AMD, 256 node units, full fat tree, 1608 nodes

Storage Server Platform (PAS-HC)

- Panasas architecture mapped from blades to servers
 - Director function still runs on DirectorBlades
 - Integrated battery provides NVRAM for fast in-memory journals
 - Number of DirectorBlades is flexible and orthogonal to data storage
 - Object storage (OSDFS) stored in LUNs managed by dual-redundant block-based RAID controllers (e.g., LSI 9700)
 - Front end servers run OSDFS and implement iSCSI/OSD target
 - Each LUN stores one instance of OSDFS (i.e., OSD or OST)
 - Any server attached to controller can manage the OSDs, as directed by Realm
 - Data Striping
 - Panasas files are striped RAID-0 across different OSDs (on any/all controllers)
 - Panasas directories are replicated RAID-1 on two different OSDs
 - Can scavenge the file system if a LUN is lost

Storage Server Platform

