Stanford HPC Conference

POWERED BY ROCKS

panasas

Panasas Storage System –

Integration into a Cluster

David Yu
Industry Verticals
Panasas Inc.

Steve Jones
Technology Operations Manager
Institute for Computational and Mathematical Engineering
Stanford University

Agenda

- HPC requirements for storage
- Shared storage – promise and challenges
- Need for new storage architecture, object-based storage
- Panasas implementation of object storage architecture
- Panasas integration into a cluster with existing storage
- Panasas integration at Stanford University – Steve Jones, Stanford
- Demo – Alex Krimkevich, Panasas
- Questions and answers

Storage Requirements for HPC, Rocks

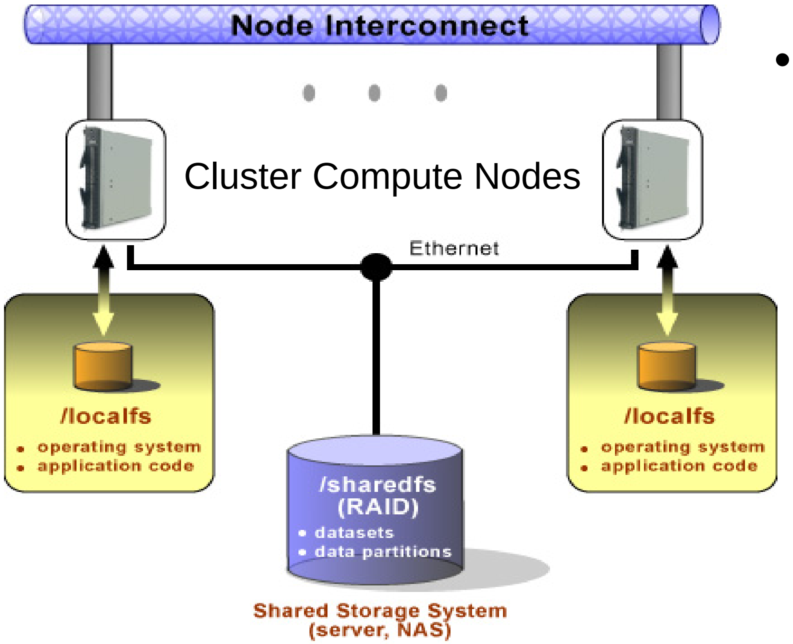
- Performance
  - High read concurrency for parallel application and data sets
  - High write bandwidth for memory checkpointing, interim and final output
- Scalability
  - More difficult problems typically means larger data sets
  - Scaling cluster nodes requires scalable IO performance
- Management
  - Single system image maximizes utility for user community
  - Minimize operations and capital costs (80% of storage TCO)

High demand on storage systems

Shared Storage: The Promise

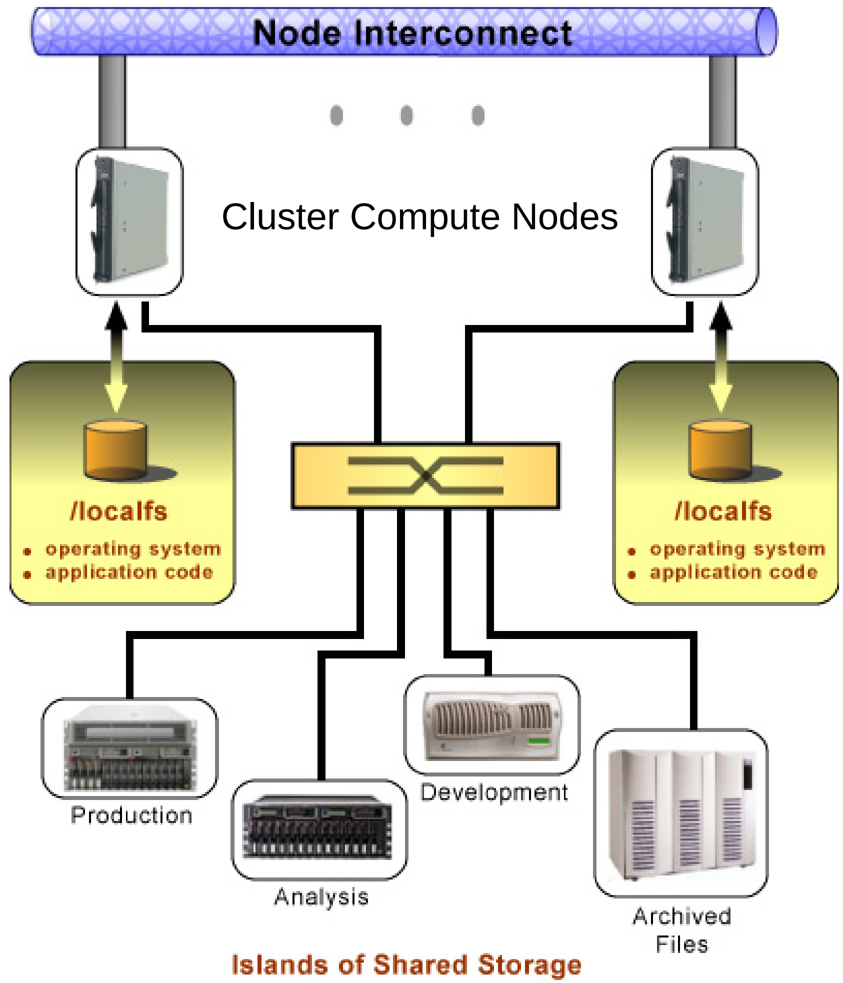
POWERED BY ROCKS

panasas



- Shared storage cluster computing
  - Compute anywhere model
    - Partitions available globally; no replicas required (shared datasets)
    - No data staging required
    - No distributed data consistency issues
    - Reliable checkpoints; application reconfiguration
    - Results gateway
  - Enhanced reliability via RAID
  - Enhanced manageability
    - Policy-based management

Shared Storage Challenges



- Performance, scalability & management
  - Single file system performance limited
  - Multiple volumes and mount points
  - Manual capacity and load balancing
  - Large quantum upgrade costs

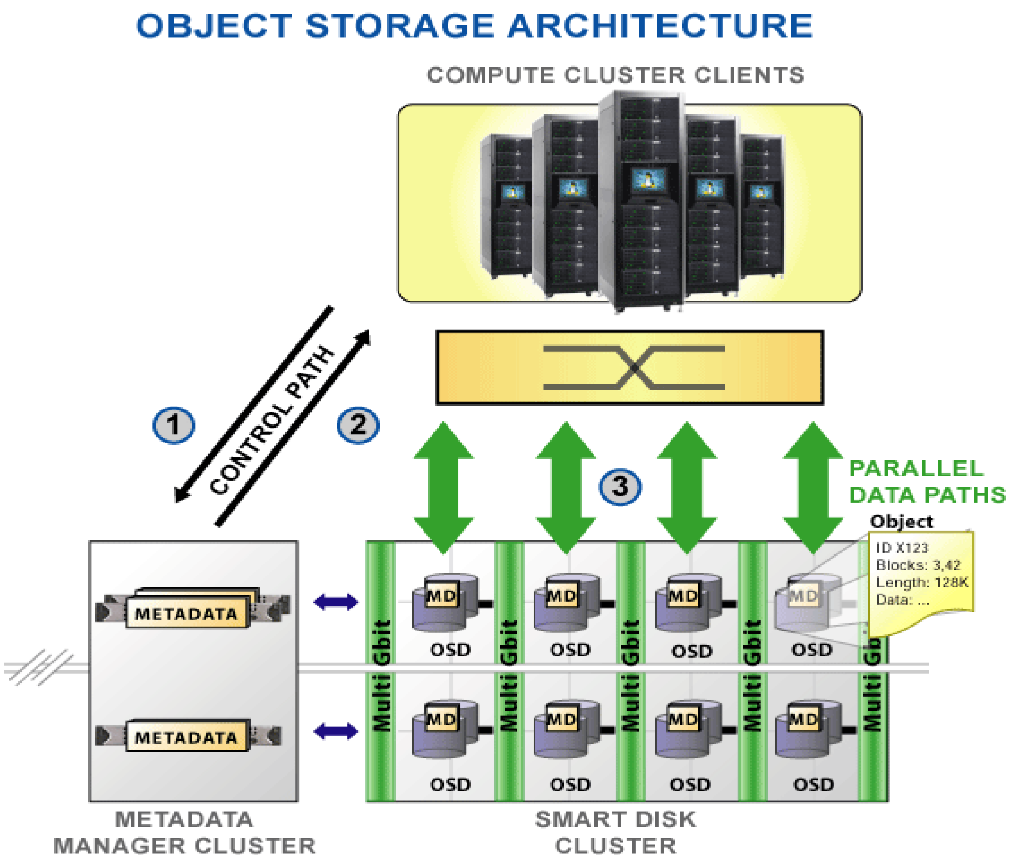
Motivation for New Architecture

- A highly scalable, interoperable, shared storage system
  - Improved storage management
    - Self-management, policy-driven storage (i.e. backup and recovery)
  - Improved storage performance
    - Quality of service, differentiated services
  - Improved scalability
    - Of performance and metadata (i.e. free block allocation)
  - Improved device and data sharing
    - Shared devices and data across OS platforms

Next Generation Storage Cluster

POWERED BY ROCKS

panasas



HOW IT WORKS

Object-based Storage is an emerging networked storage standard that delivers unprecedented performance and scalability within a single management domain.

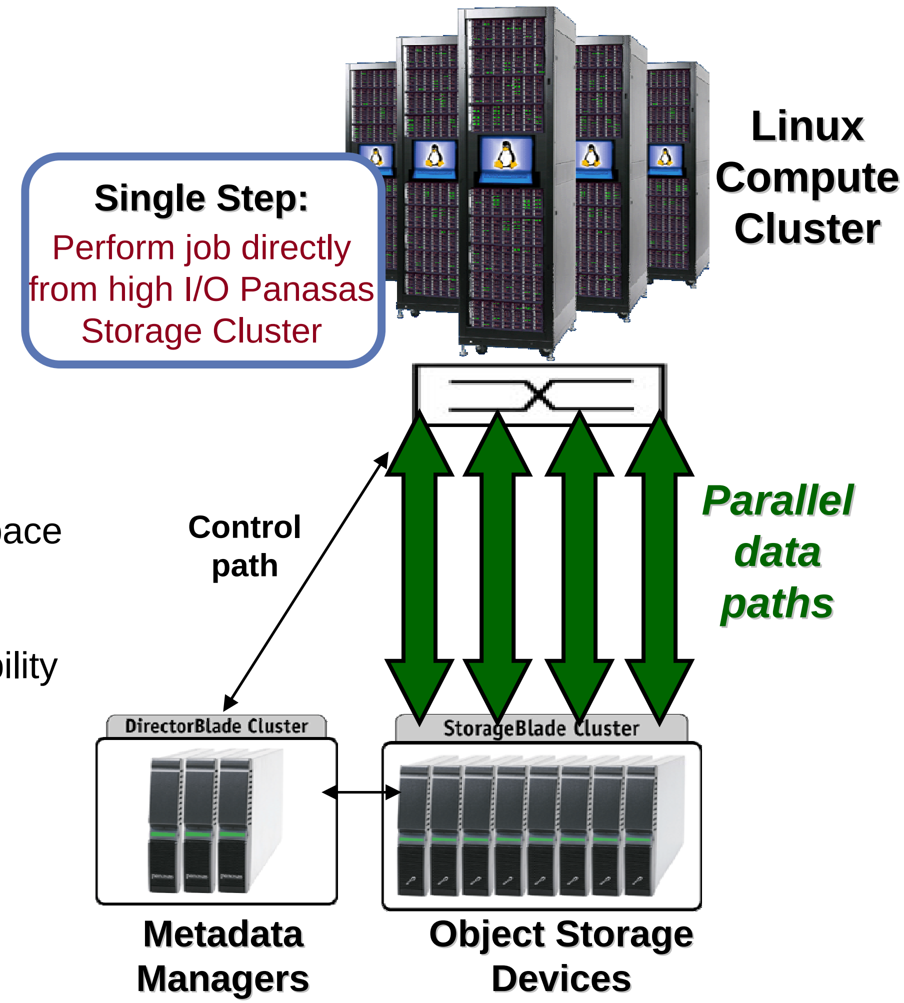
1. Client processes make independent file requests (e.g.,stat, open) of the clustered metadata manager
2. Authenticated clients receive object maps and capabilities
3. Clients use their credentials to perform parallel direct-to-disk requests across the GbE switching network

Panasas ActiveScale File System

POWERED BY ROCKS

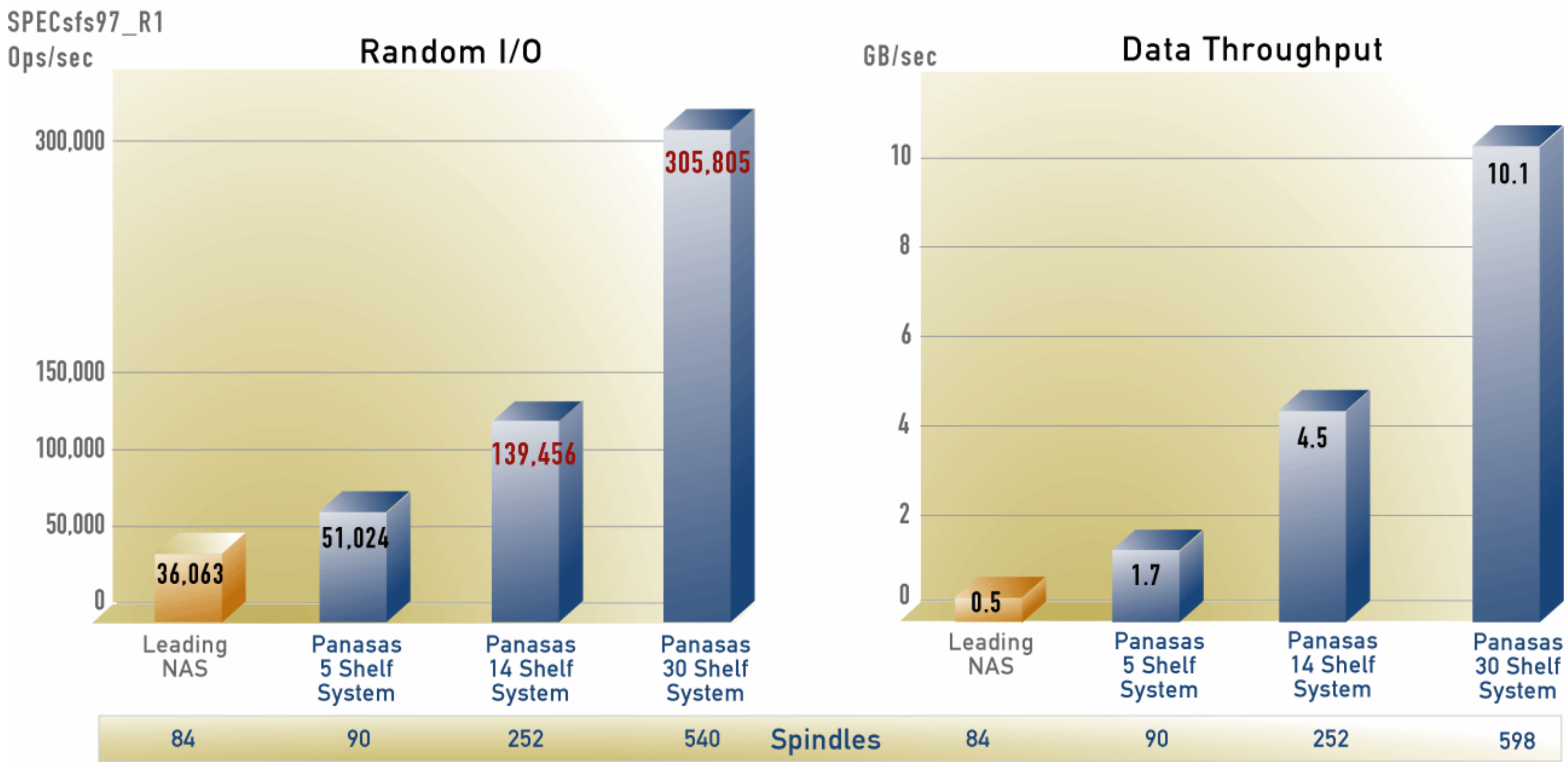
panasas

- Scalable performance
  - Offloaded data path enables nodes to access disks directly
  - Scale nodes, network, and capacity
  - As capacity grows, performance grows
- Simplified and dynamic management
  - Robust, shared file access
  - Seamless growth within single namespace
- Integrated HW/SW solution
  - Optimizes performance and manageability
  - Ease of integration and support



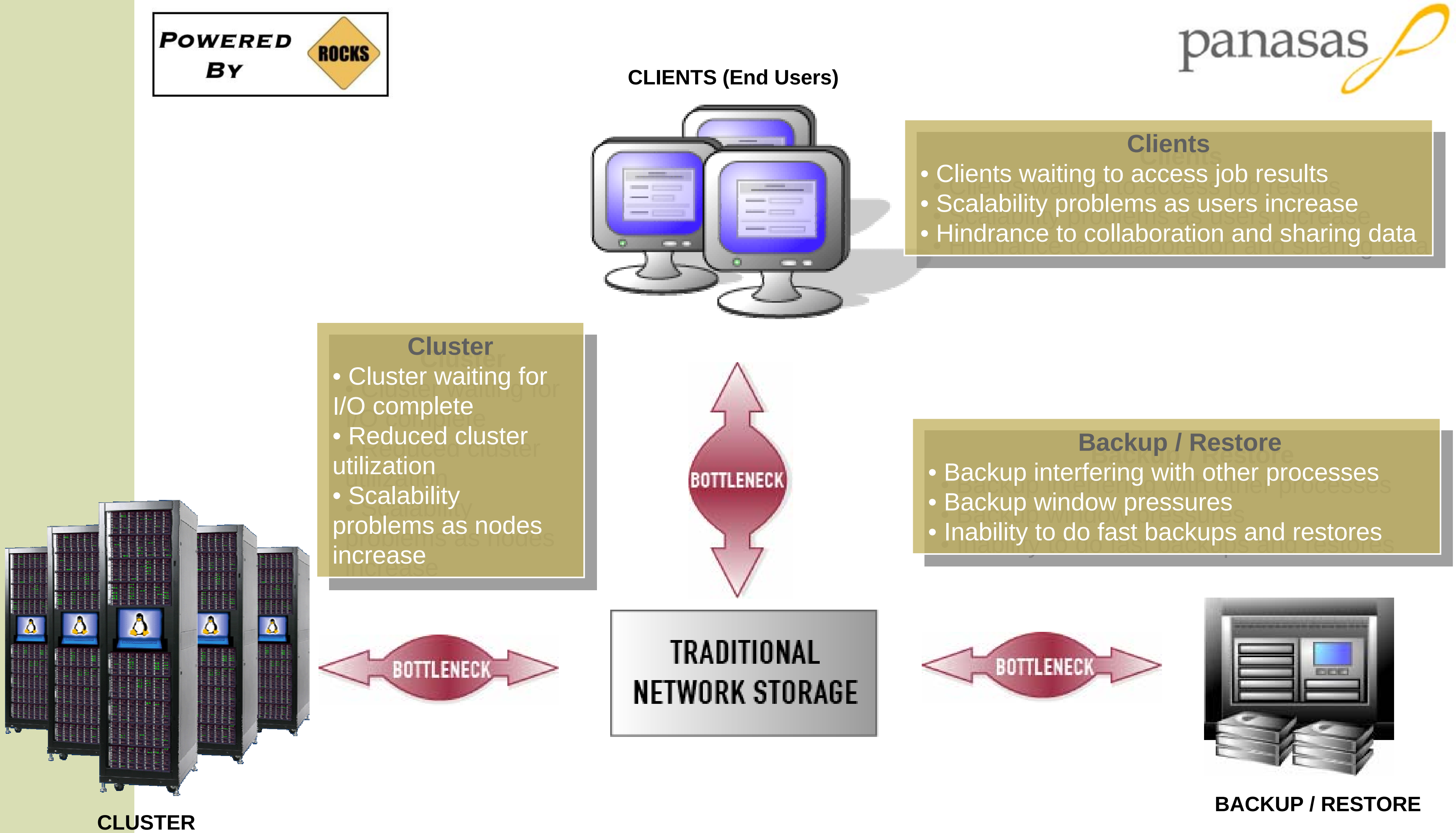
Panasas Storage Cluster

Industry-Leading Performance



Breakthrough data throughput AND random I/O

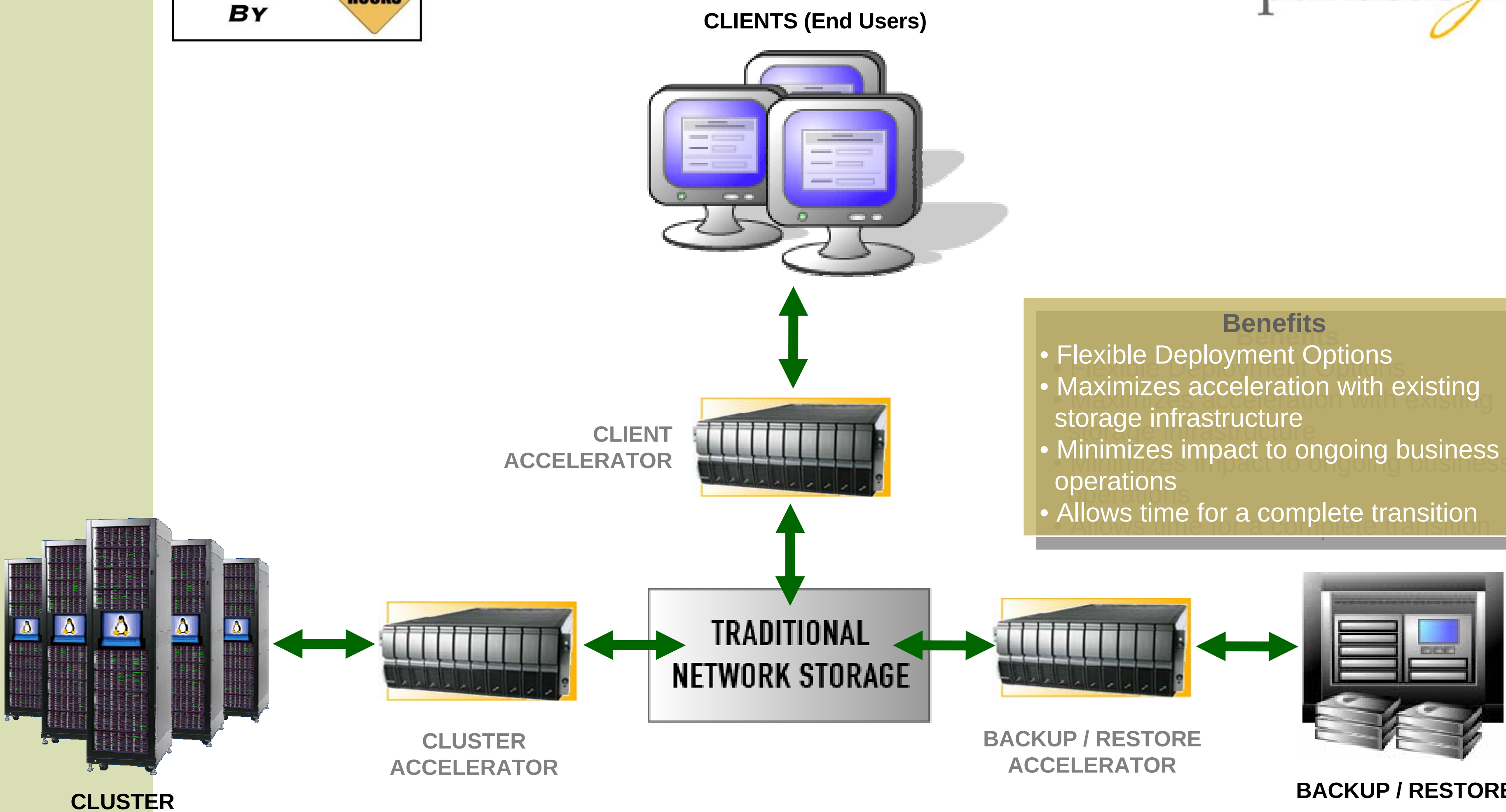
Bottlenecks in the IT Infrastructure



Accelerators To Boost Performance

- Cluster Accelerator
  - Parallel reads, writes
  - Higher cluster utilization, more jobs processed
  - Leverages existing storage assets
- Client Accelerator
  - Fast access to job results for end users
  - Promotes efficient collaboration and independent analysis
- Backup / Restore Accelerator
  - Staging area for fast backups / restores
  - Offloads backup from production storage
  - Works with backup / restore applications

Maximum Acceleration with Existing Storage



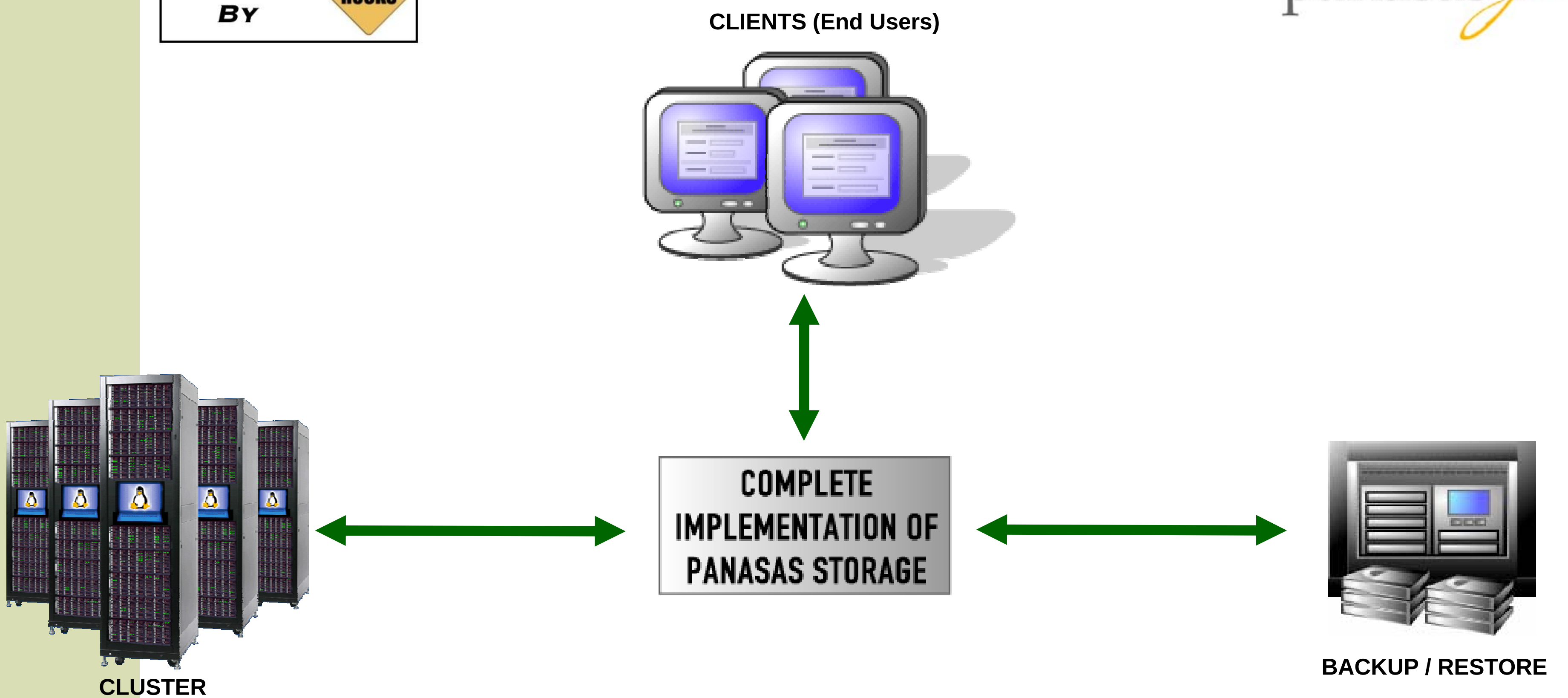
Benefits

- Flexible Deployment Options
- Maximizes acceleration with existing storage infrastructure
- Minimizes impact to ongoing business operations
- Allows time for a complete transition

Complete Acceleration

POWERED BY ROCKS

panasas



Benefits

- Removes bottlenecks of traditional storage
- Minimizes “data motion”
- Compute from home directories
- Output accessible from everywhere
- Data processing pipeline
- Output from one process is input to next process

Summary

- HPC puts unprecedented demands on storage systems
- Object storage architecture fulfills the promise of shared storage
  - Manageability (80% of storage TCO)
  - Scales performance in concert with capacity
  - Implemented by Panasas with industry leading performance
- Integration of Panasas storage with clusters is non-disruptive
  - Deploy with existing storage infrastructure as Accelerators
    - Cluster, Client, Backup / Restore
  - Gain immediate benefits
  - Transition to complete deployment over time
- Key to productivity with HPC – eliminate the bottlenecks to storage systems, simplify management

POWERED BY ROCKS

panasas

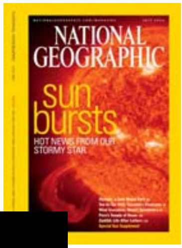
Panasas Storage System – Integration into a Cluster at Stanford University

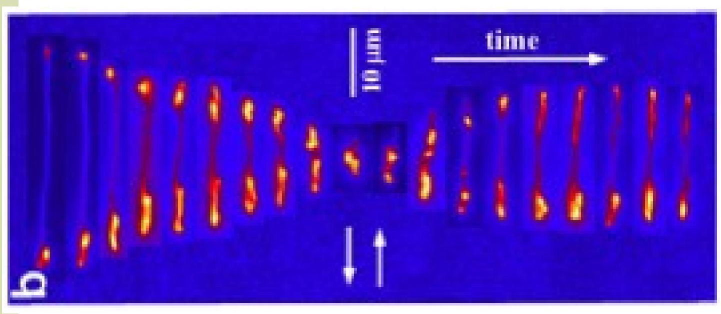
Steve Jones
Technology Operations Manager
Institute for Computational and Mathematical Engineering
Stanford University

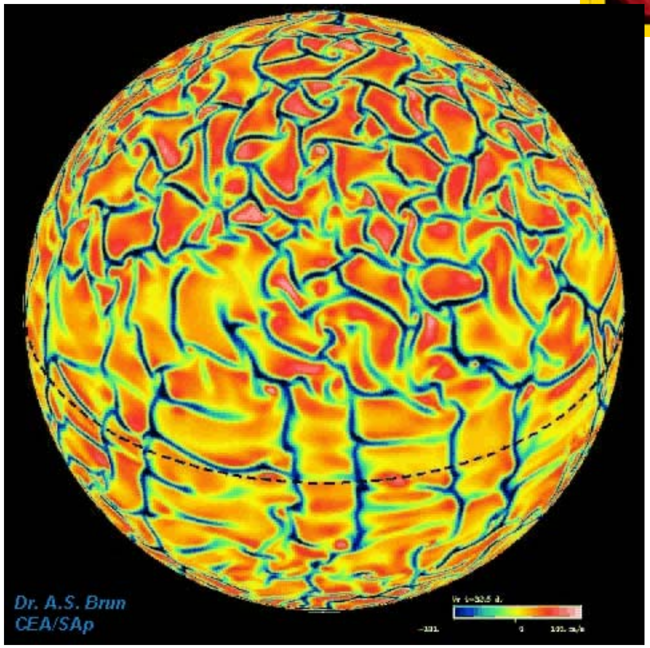
The Research

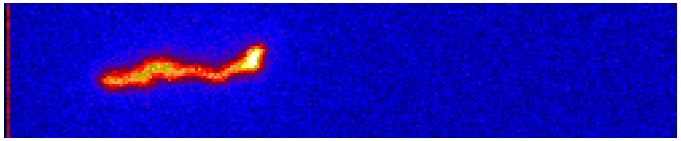
POWERED BY ROCKS

MOLECULES TO PLANETS !









Research Groups

POWERED BY ROCKS

- Flow Physics and Computation
- Aeronautics and Astronautics
- Chemical Engineering
- Center for Turbulence Research
- Center for Integrated Turbulence Simulations
- Thermo Sciences Division

Funding

- Sponsored Research (AFOSR/ONR/DARPA/DURIP/ASC)

Active Collaborations with the Labs

Buoyancy driven instabilities/mixing - CDP for modeling plumes (Stanford/SNL)

LES Technology - Complex Vehicle Aerodynamics using CDP (Stanford/LLNL)

Tsunami modeling - CDP for Canary Islands Tsunami Scenarios (Stanford/LANL)

Parallel I/O & Large-Scale Data Visualization - UDM integrated in CDP (Stanford/LANL)

Parallel Global Solvers - HyPre Library integrated in CDP (Stanford/LLNL)

Parallel Grid Generation - Cubit and related libraries (Stanford/SNL)

Merrimac - Streaming Supercomputer Prototype (Stanford/LLNL/LBNL/NASA)

Affiliates Program

POWERED BY ROCKS

Pratt & Whitney
A United Technologies Company

HONDA

Valeo

GM
General Motors

CERFACS

Honeywell

DOW
The Dow Chemical Compnay

CATERPILLAR®

CIRA
Centro Italiano Ricerche Aerospaziali

Rolls-Royce

FLUENT

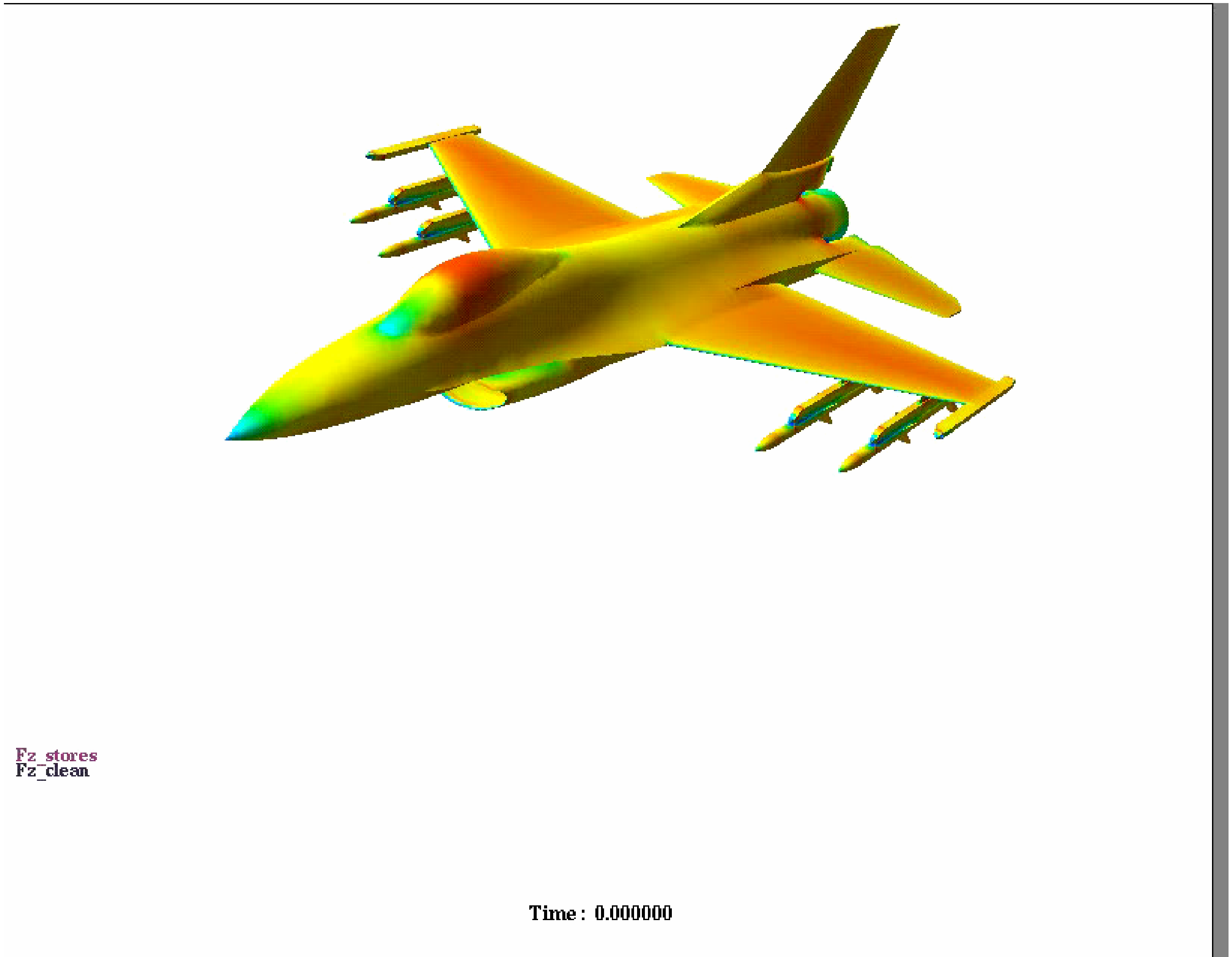
HITACHI

snecma

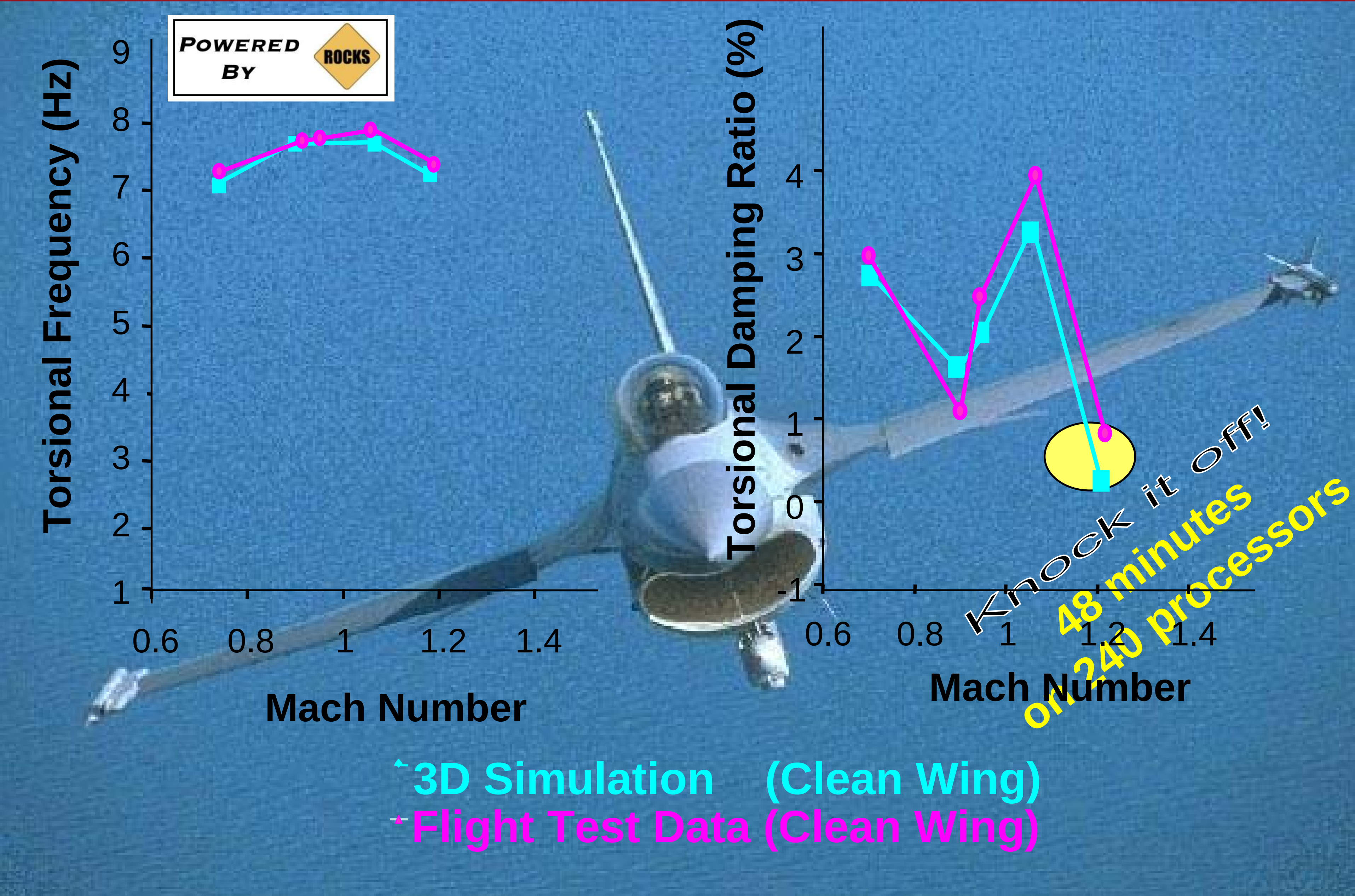
Computer Systems, Inc.
MERCURY

FuelCell Energy

F-16 Simulation



Simulation Versus Flight Data



Databases?

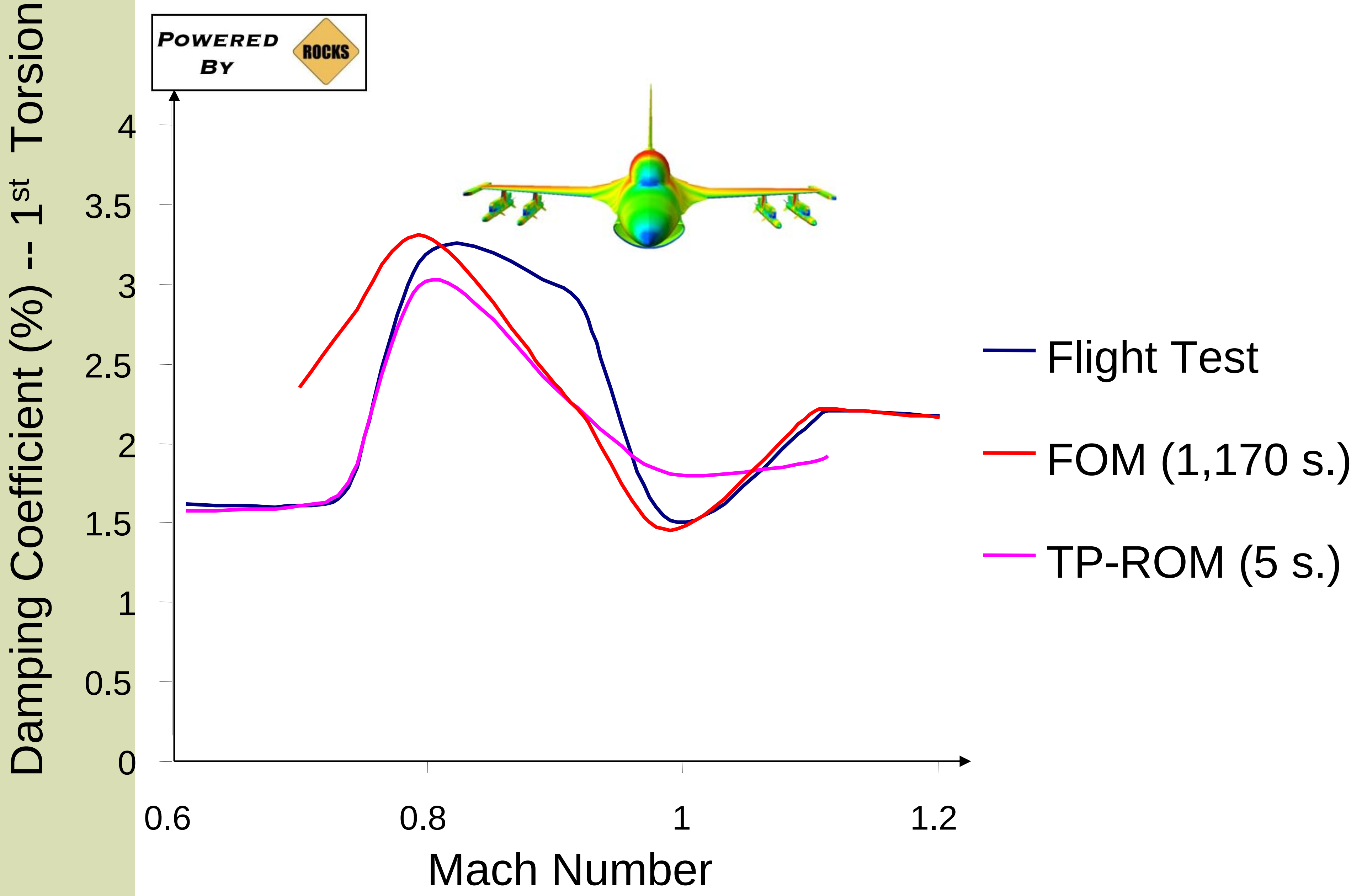
POWERED BY ROCKS

Desert Storm (1991)

Iraq War (2003)

400,000 configurations to be flight tested

Potential



Powered By
ROCKS
Damping Coefficient (%) -- 1st Torsion
4
3.5
3
2.5
2
1.5
1
0.5
0
0.6
0.8
1
1.2
Mach Number
Flight Test
FOM (1,170 s.)
TP-ROM (5 s.)

Nivation

POWERED BY ROCKS

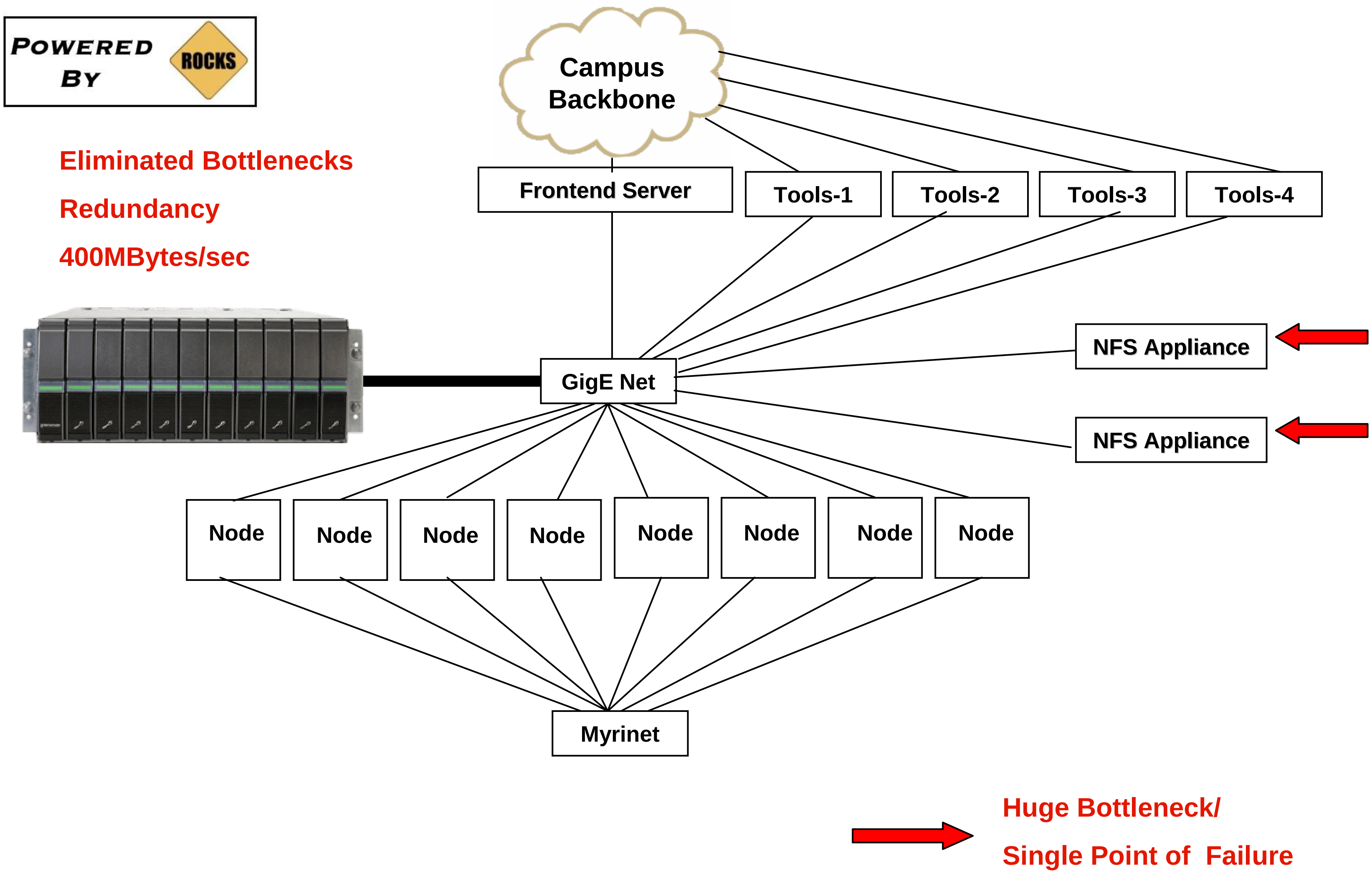
- 164 Processor Intel Xeon 3.0GHz
- 4GB RAM per node
- Myrinet
- Gigabit Ethernet
- Two 1 TB NAS Appliances
- 4 Tools Nodes

Panasas Eliminates Bottlenecks

Eliminated Bottlenecks

Redundancy

400MBytes/sec



Panasas Integration in Less Than 2 Hours

POWERED BY ROCKS

- Installation and configuration of Panasas Shelf - 1 hour
- Switch configuration changes for link aggregation - 10 minutes
- Copy RPM to /home/install/contrib/enterprise/3/public/i386/RPMS - 1 minute
- create/edit extend-compute.xml - 5 minutes

```
# Add panfs to fstab
REALM=10.10.10.10
mount_flags="rw,noauto,panauto"
/bin/rm -f /etc/fstab.bak.panfs
/bin/rm -f /etc/fstab.panfs
/bin/cp /etc/fstab /etc/fstab.bak.panfs
/bin/grep -v "panfs://" /etc/fstab > /etc/fstab.panfs
/bin/echo "panfs://$REALM:global  /panfs  panfs   $mount_flags     0      0" >> /etc/fstab.panfs
/bin/mv -f /etc/fstab.panfs /etc/fstab
/bin/sync

/sbin/chkconfig --add panfs
/usr/local/sbin/check_panfs
LOCATECRON=/etc/cron.daily/slocate.cron
LOCATE=/etc/sysconfig/locate
LOCTEMP=/tmp/slocate.new

/bin/cat $LOCATECRON | sed "s/,proc,/,proc,panfs,/g" > $LOCTEMP
/bin/mv -f $LOCTEMP $LOCATECRON
/bin/cat $LOCATECRON | sed "s/\/afs,/\/afs,\/panfs,/g" > $LOCTEMP
/bin/mv -f $LOCTEMP $LOCATECRON
```

- [root@rockscluster]# rocks-dist dist ; cluster-fork '/boot/kickstart/cluster-kickstart' - 30 minutes
- /etc/auto.home   userX -fstype=panfs panfs://10.x.x.x/home/userX - script it to save time

Benchmarking Panasas Using Bonnie++

```
#!/bin/bash

#PBS -N BONNIE

#PBS -e Log.d/BONNIE.panfs.err

#PBS -o Log.d/BONNIE.panfs.out

#PBS -m aeb

#PBS -M hpcclusters@gmail.com

#PBS -l nodes=1:ppn=2

#PBS -l walltime=30:00:00

PBS_O_WORKDIR='/home/sjones/benchmarks'
export PBS_O_WORKDIR

### ---------------------------------------
### BEGINNING OF EXECUTION
### ---------------------------------------

echo The master node of this job is `hostname`
echo The job started at `date`
echo The working directory is `echo $PBS_O_WORKDIR`
echo This job runs on the following nodes:
echo `cat $PBS_NODEFILE`

### end of information preamble

cd $PBS_O_WORKDIR
cmd="/home/tools/bonnie++/sbin/bonnie++ -s 8000 -n 0 -f -d /home/sjones/bonnie"
echo "running bonnie++ with: $cmd in directory "`pwd`
$cmd >& $PBS_O_WORKDIR/Log.d/run9/log.bonnie.panfs.$PBS_JOBID
```

NFS – 8 Nodes

POWERED BY ROCKS

Version 1.03		Sequential Output				Sequential Input-		Random Seeks	
		-Block-		-Rewrite-		-Block-			
Machine	Size	K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP
compute-3-82	8000M	2323	0	348	0	5119	1	51.3	0
compute-3-81	8000M	2333	0	348	0	5063	1	51.3	0
compute-3-80	8000M	2339	0	349	0	4514	1	52.0	0
compute-3-79	8000M	2204	0	349	0	4740	1	99.8	0
compute-3-78	8000M	2285	0	354	0	3974	0	67.9	0
compute-3-77	8000M	2192	0	350	0	5282	0	46.8	0
compute-3-74	8000M	2292	0	349	0	5112	1	45.4	0
compute-3-73	8000M	2309	0	358	0	4053	0	64.6	0

17.80MB/sec for concurrent write using NFS with 8 dual processor jobs

36.97MB/sec during read process

PanFS – 8 Nodes

Version 1.03		Sequential Output				Sequential Input-		Random Seeks	
		-Block-		-Rewrite-		-Block-			
Machine	Size	K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP
compute-1-18.	8000M	20767	8	4154	3	24460	7	72.8	0
compute-1-17.	8000M	19755	7	4009	3	24588	7	116.5	0
compute-1-16.	8000M	19774	7	4100	3	23597	7	96.4	0
compute-1-15.	8000M	19716	7	3878	3	25384	8	213.6	1
compute-1-14.	8000M	19674	7	4216	3	24495	7	72.8	0
compute-1-13.	8000M	19496	7	4236	3	24238	7	71.0	0
compute-1-12.	8000M	19579	7	4117	3	23731	7	97.1	0
compute-1-11.	8000M	19688	7	4038	3	24195	8	117.7	0

154MB/sec for concurrent write using PanFS with 8 dual processor jobs

190MB/sec during read process

NFS – 16 Nodes

Version 1.03		Sequential Output				Sequential Input-		Random Seeks	
		-Block-		-Rewrite-		-Block-			
Machine	Size	K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP
compute-3-82	8000M	1403	0	127	0	2210	0	274.0	2
compute-3-81	8000M	1395	0	132	0	1484	0	72.1	0
compute-3-80	8000M	1436	0	135	0	1342	0	49.3	0
compute-3-79	8000M	1461	0	135	0	1330	0	53.7	0
compute-3-78	8000M	1358	0	135	0	1291	0	54.7	0
compute-3-77	8000M	1388	0	127	0	2417	0	45.5	0
compute-3-74	8000M	1284	0	133	0	1608	0	71.9	0
compute-3-73	8000M	1368	0	128	0	2055	0	54.2	0
compute-3-54	8000M	1295	0	131	0	1650	0	47.4	0
compute-2-53	8000M	1031	0	176	0	737	0	18.3	0
compute-2-52	8000M	1292	0	128	0	2124	0	104.1	0
compute-2-51	8000M	1307	0	129	0	2115	0	48.1	0
compute-2-50	8000M	1281	0	130	0	1988	0	92.2	1
compute-2-49	8000M	1240	0	135	0	1488	0	54.3	0
compute-2-47	8000M	1273	0	128	0	2446	0	52.7	0
compute-2-46	8000M	1282	0	131	0	1787	0	52.9	0

20.59MB/sec for concurrent write using NFS with 16 dual processor jobs
27.41MB/sec during read process

PanFS – 16 Nodes

POWERED BY ROCKS

Version 1.03		Sequential Output				Sequential Input-		Random Seeks	
		-Block-		-Rewrite-		-Block-			
Machine	Size	K/sec	%CP	K/sec	%CP	K/sec	%CP	K/sec	%CP
compute-1-26	8000M	14330	5	3392	2	28129	9	54.1	0
compute-1-25	8000M	14603	5	3294	2	30990	9	60.3	0
compute-1-24	8000M	14414	5	3367	2	28834	9	55.1	0
compute-1-23	8000M	9488	3	2864	2	17373	5	121.4	0
compute-1-22	8000M	8991	3	2814	2	21843	7	116.5	0
compute-1-21	8000M	9152	3	2881	2	20882	6	80.6	0
compute-1-20	8000M	9199	3	2865	2	20783	6	85.2	0
compute-1-19	8000M	14593	5	3330	2	29275	9	61.0	0
compute-1-18	8000M	9973	3	2797	2	18153	5	121.6	0
compute-1-17	8000M	9439	3	2879	2	22270	7	64.9	0
compute-1-16	8000M	9307	3	2834	2	21150	6	99.1	0
compute-1-15	8000M	9774	3	2835	2	20726	6	77.1	0
compute-1-14	8000M	15097	5	3259	2	32705	10	60.6	0
compute-1-13	8000M	14453	5	2907	2	36321	11	126.0	0
compute-1-12	8000M	14512	5	3301	2	32841	10	60.4	0
compute-1-11	8000M	14558	5	3256	2	33096	10	62.2	0

187MB/sec for concurrent write using PanFS with 8 dual processor jobs
405MB/sec during read process
Capacity imbalances on jobs - 33MB/sec increase from 8 to 16 job run

Panasas Statistics During Write Process

[pancli] sysstat storage

IP	CPU	Disk	Ops/s	KB/s		Capacity (GB)		
	Util	Util		In	Out	Total	Avail	Reserved
10.10.10.250	55%	22%	127	22847	272	485	367	48
10.10.10.253	60%	24%	140	25672	324	485	365	48
10.10.10.245	53%	21%	126	22319	261	485	365	48
10.10.10.246	55%	22%	124	22303	239	485	366	48
10.10.10.248	57%	22%	134	24175	250	485	369	48
10.10.10.247	52%	21%	124	22711	233	485	366	48
10.10.10.249	57%	23%	135	24092	297	485	367	48
10.10.10.251	52%	21%	119	21435	214	485	366	48
10.10.10.254	53%	21%	119	21904	231	485	367	48
10.10.10.252	58%	24%	137	24753	300	485	366	48
Total "Set 1"	**55%**	**22%**	**1285**	**232211**	**2621**	**4850**	**3664**	**480**

Sustained BW 226 MBytes/Sec during 16 1GB concurrent writes

Panasas Statistics During Read Process

[pancli] sysstat storage

IP	CPU Util	Disk Util	Ops/s	KB/s In	KB/s Out	Capacity (GB) Total	Avail	Reserved
10.10.10.250	58%	95%	279	734	21325	485	355	48
10.10.10.253	60%	95%	290	727	22417	485	353	48
10.10.10.245	54%	92%	269	779	19281	485	353	48
10.10.10.246	59%	95%	290	779	21686	485	354	48
10.10.10.248	60%	95%	287	729	22301	485	357	48
10.10.10.247	52%	91%	256	695	19241	485	356	48
10.10.10.249	57%	93%	276	708	21177	485	356	48
10.10.10.251	49%	83%	238	650	18043	485	355	48
10.10.10.254	45%	82%	230	815	15225	485	355	48
10.10.10.252	57%	94%	268	604	21535	485	354	48
Total "Set 1"	**55%**	**91%**	**2683**	**7220**	**202231**	**4850**	**3548**	**480**

Sustained BW 197 MBytes/Sec during 16 1GB concurrent sequential reads

Typical Storage Utilization with Cluster at 76%

Powered By ROCKS

[pancli] sysstat storage

IP	CPU Util	Disk Util	Ops/s	KB/s In	KB/s Out	Capacity (GB) Total	Capacity (GB) Avail	Capacity (GB) Reserved
10.10.10.250	6%	5%	35	292	409	485	370	48
10.10.10.253	5%	4%	35	376	528	485	368	48
10.10.10.245	4%	3%	29	250	343	485	368	48
10.10.10.246	6%	4%	28	262	373	485	369	48
10.10.10.248	5%	3%	27	234	290	485	372	48
10.10.10.247	3%	3%	1	1	2	485	370	48
10.10.10.249	5%	3%	48	258	365	485	371	48
10.10.10.251	4%	3%	46	216	267	485	369	48
10.10.10.254	4%	3%	32	256	349	485	370	48
10.10.10.252	4%	3%	34	337	499	485	370	48
Total	**4%**	**3%**	**315**	**2482**	**3425**	**4850**	**3697**	**480**

sustained BW 2.42 Mbytes/sec in - 3.34 Mbytes/sec out

[root@frontend-0 root]# showq

ACTIVE JOBS--------------------

JOBNAME	USERNAME	STATE	PROC	REMAINING	STARTTIME
8649	sjones	Running	2	1:04:55:08	Sun May 15 18:20:23
8660	user1	Running	6	1:23:33:15	Sun May 15 18:58:30
8524	user2	Running	16	2:01:09:51	Fri May 13 20:35:06
8527	user3	Running	16	2:01:23:19	Fri May 13 20:48:34
8590	user4	Running	64	3:16:42:50	Sun May 15 10:08:05
8656	user5	Running	16	4:00:55:36	Sun May 15 18:20:51
8647	user6	Running	5	99:22:50:42	Sun May 15 18:15:58

7 Active Jobs 125 of 164 Processors Active (76.22%)

65 of 82 Nodes Active (79.27%)

Conclusion

- Summary
- Demo
- Questions and Answers

Thank you

For more information about Steve Jones:
High Performance Computing Clusters
http://www.hpcclusters.org

For more information about Panasas:
http://www.panasas.com

For more information about Rocks:
http://www.rocksclusters.org