

Panasas ActiveStor12 Evaluation Report

Thomas Schoenemeyer and Hussein N. El-Harake and
Swiss National Supercomputing Centre (CSCS), Lugano, Switzerland
schoenemeyer@cscs.ch; hussain@cscs.ch

Abstract: We evaluated the Panasas Active Storage 12 system connected to a small HPC cluster with InfiniBand FDR Fabric. The purpose of this study is to investigate the stability, scalability, manageability and integration of this solution in the CSCS environment. We describe the hardware, the solution architecture and the benchmark procedure. Iozone and mdtest were used for benchmarking. Our InfiniBand cluster (greina) has 16 compute nodes and one head node. The compute nodes deploy the latest Intel Xeon E5-2670 processors. The Panasas Active Storage solution met our expectations in terms of integration and performance.

1 Introduction

In many systems at CSCS we use Lustre or GPFS as parallel file system available as scratch space. We evaluated Panasas ActiveStor12 as a possible alternative to the current solutions at CSCS.

Panasas Active Storage 12 is a complete HPC storage solution and easy to deploy. Since no InfiniBand interface is provided, routers have to be added to connect to the Cluster InfiniBand Fabric. ActiveStor appliances store data as objects, with RAID performed on a per-file basis scale the capacity and performance of the file system as storage requirements grow. According to Panasas the Panasas® PanFS™ parallel file system allows to linearly scale to six petabytes with a performance of up 150GB/s. [1]

CSCS received 3 x ActiveStor 12 shelves; every shelf consists of 2 director blades and 9 storage blades. Each storage blade has 2 x 2 TB HDD, a total of 90TB of usable space were available. Director blades coordinate activity between clients and storage blades. Unlike NAS filers, file system activity is distributed between director blades and storage blades. The concept has some similarity with Lustre, with one director blade as metadata server and storage blades as object storage servers. In contrast to Lustre, PanFS has the ability to use many director blades simultaneously. Up to three *director blades* per shelf might be configured depending on the type of workload.



Active Storage 12 with eleven blades and a single storage blade

From the Panasas documents [3] and [4], a standalone Active Storage 12 should be able to deliver 1600MB/s in write and 1500MB/s in read throughput.

A similar study with two PAS12 shelves was created by Tony Palmer and Ginny Roth [2], they also used 16 Linux clients and measured the throughput performance with IOR and one directory blade per shelf. They measured an aggregate write throughput of 3.1 GB/s and 2.6 GB/s of aggregate read throughput.

2 Description of the test bed

The test bed is shown in figure 1 and deploys the following components

- Three Panasas Storage shelves running PanFs 5.0.0.0b (initially 4.1.3)
- One 24-port 10GbE Switch
- 3 single-socket nodes acting as routers with Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz
- 36-port IB FDR Switch (Mellanox)
- HPC Cluster with 16 dual-socket nodes with Intel Xeon E5-2670 (Sandy Bridge)

Each storage blade has two SATA disks with 2TB of capacity.

In our test bed each shelf has one directory blade and ten storage blades. Configurations are also possible with 2+9 and 3+8. Each director blade connects to the 10GbE switch with two SFP+ ports. For each shelf another two 10GbE ports connect the storage blades to the 24 port GbE switch. In total 12x10GbE links connect the storage and director blades with the three routers. The routers are integrated in the InfiniBand FDR Fabric.

SPECIFICATIONS

ActiveStor Appliance Model	ActiveStor 11	ActiveStor 12
ActiveStor Generation	Fourth	
Max. System Capacity ¹	6 PB	
Max System Throughput ²	115 GB/s	150 GB/s
Max. Shelves per system ¹	100	
Per shelf		
Capacity per Shelf (TB) ²	40 or 60	
Hard Drives (3.5" SATA) ²	20	
ECC Memory (GB of Cache) ²	48	92
Max. Throughput, Write ²	950MB/s	1600MB/s
Max. Throughput, Read ²	1150MB/s	1500MB/s
Supported Blade Configurations (Director Blade + Storage Blade)	1+10, 2+9, or 3+8. Also 0+11 for expansion	
Networking Switch Modules	One (second optional)	Two
Networking Uplinks per Switch Module	1 x 10GbE SFP+/CX4 or 4 x GbE Copper	
Additional Networking per Director Blade	2 x 10GbE SFP+	
Network Failover	Optional	Standard
High Availability Link Aggregation	No	Yes
QDR InfiniBand Router Compatibility	Yes	

¹ No enforced limits. Max tested configuration – 100 shelves
² Based on a 1+10 blade configuration

Table 1: Performance specs provided in the Panasas datasheet

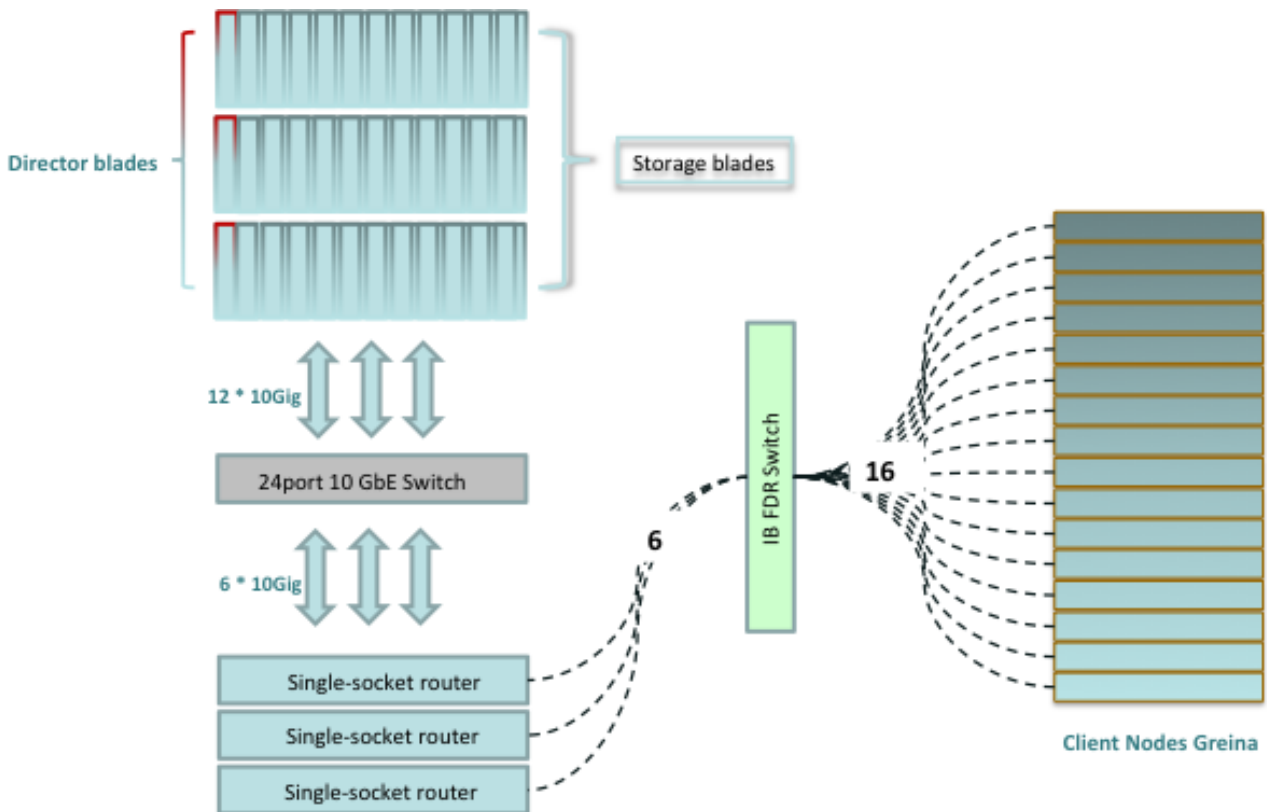


Figure 1: CSCS Test Bed with 1+10 configuration

3 Evaluation Method

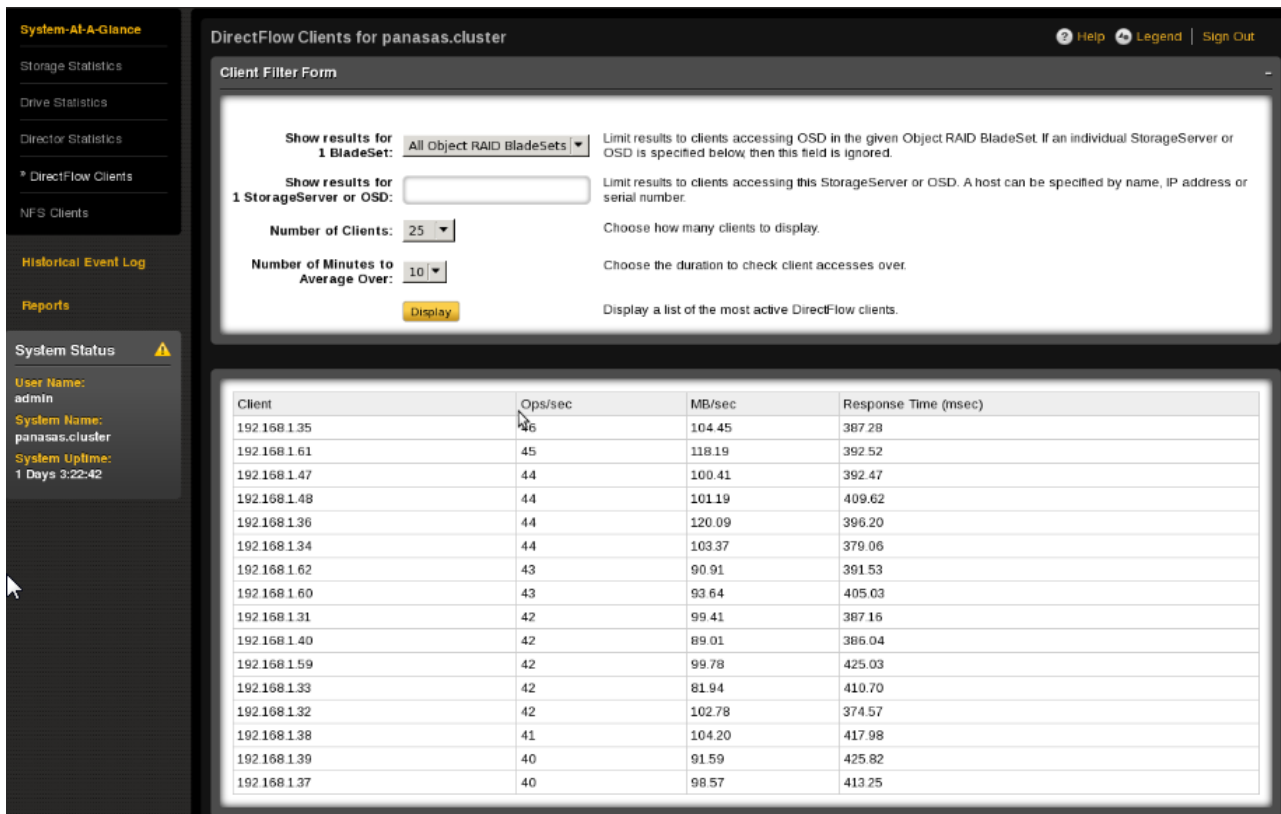
We used Iozone (5) for the throughput measurements, and mdtest version 1.8.3 (6) to measure the IOPs.

We used up to 16 dual-socket nodes of our HPC Cluster as compute clients. Hyperthreading was always enabled, and Turbo Boost was disabled for the first test series and enabled for the second series activated. This allows 3.3 GHz maximum core frequency.

The six raid Volumes were configured as Raid10 for the IOPs test and Raid5 for throughput test. The PanFS storage operating system offers three protocols that can be used simultaneously:

- DirectFlow: high-performance, parallel protocol access for Linux clients (used in this study)
- NFS v3: for Linux and Unix clients
- CIFS: for Windows clients

The PanActive Manager (see below) displays all relevant system information like capacity or disk utilization. In the particular examples below, the individual client performance is collected over a user-specified period.



The screenshot shows the PanActive Manager interface for 'DirectFlow Clients for panasas.cluster'. The 'Client Filter Form' is visible, allowing users to filter results by BladeSet (set to 'All Object RAID BladeSets'), StorageServer, or OSD. The number of clients to display is set to 25, and the duration is set to 10 minutes. A 'Display' button is present. Below the filter form is a table of client performance metrics.

Client	Ops/sec	MB/sec	Response Time (msec)
192.168.1.35	46	104.45	387.28
192.168.1.61	45	118.19	392.52
192.168.1.47	44	100.41	392.47
192.168.1.48	44	101.19	409.62
192.168.1.36	44	120.09	396.20
192.168.1.34	44	103.37	379.06
192.168.1.62	43	90.91	391.53
192.168.1.60	43	93.64	405.03
192.168.1.31	42	99.41	387.16
192.168.1.40	42	89.01	386.04
192.168.1.59	42	99.78	425.03
192.168.1.33	42	81.94	410.70
192.168.1.32	42	102.78	374.57
192.168.1.38	41	104.20	417.98
192.168.1.39	40	91.59	425.82
192.168.1.37	40	98.57	413.25

4 Results

4.1 Iozone Throughput

We decided to select the 10+1 configuration for our testbed, since each storage blade provides roughly 160MB/s in writing. Therefore our expectation is to see 1600 MB/s per shelf in writing or 4.8 GB/s in total for three shelves.

The green bars show the results for 4 nodes with up to 32 threads per node, the red bars show the results for 8 nodes with up to 32 threads per node and the light blue bars show the results for 16 nodes with up to 16 threads per node. (Iozone allows up to 256 threads max.)

As expected, the highest write bandwidth can be measured when running all 16 nodes at 192 threads. The peak bandwidth achieves ~4.5 GB/s, which is 93% of the performance that can be expected from the datasheet.

Running 8 nodes with 32 threads showed lower results 4.1GB/s. When we started the project, we had one 10GbE link per router, at that time routers were the bottleneck at 3.7GB/s, the final layout we used 2 *10GbE links per router, which delivered 7.5GB/s. We believe that 16 nodes are capable to saturate the system, by comparing the results to what every blade delivers, and we almost reached the peak using 16 nodes with 4 threads per node 4.4GB/s.

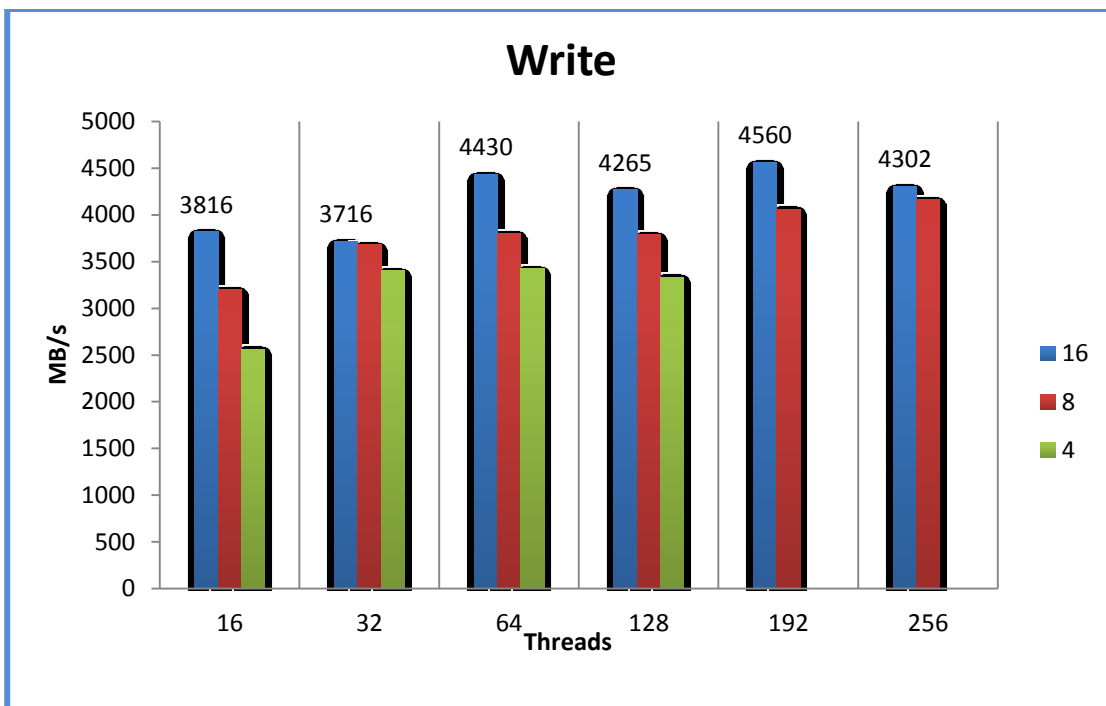


Figure 2: Iozone throughput for 4, 8 and 16 nodes with different number of threads.

The result for the read bandwidth is shown in figure 3. With 16 nodes we see the best-read performance of 4GB/s already at 32 threads which is 89% of the expected read performance for three

shelves. If we further increase the number of threads per node, we observe a performance degradation of about 8%.

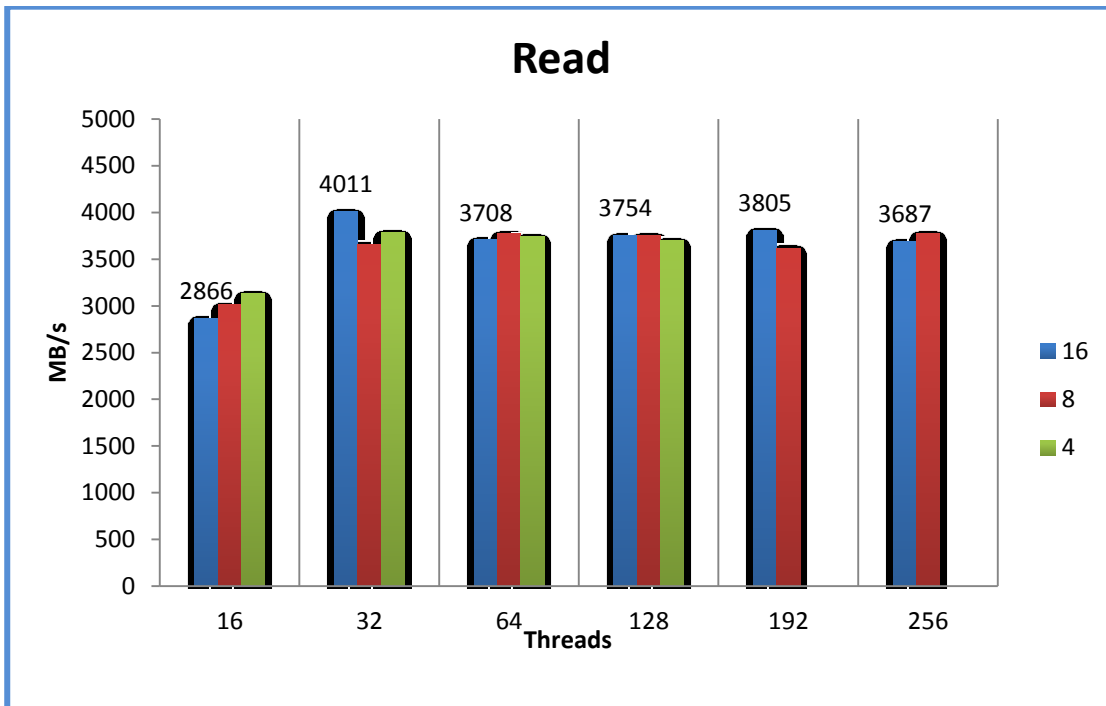


Figure 3: Iozone Read Throughput for 4, 8 and 16 nodes with different number of threads.

4.2 MDTEST

Mdtest measures the performance of multiple tasks creating, stating and deleting files and directories. It is a MPI code, so it can run processes in parallel. Each node run between 1 and 16 threads, each thread was working with 5000 files or directories.

Figures 4 and 5 show the performance numbers for creating files and directories. Both graphs include lines that indicate ideal scaling. It is obvious, that the number of creates scales nearly perfect the ideal line. We did not observe a performance difference between creating files or directories.

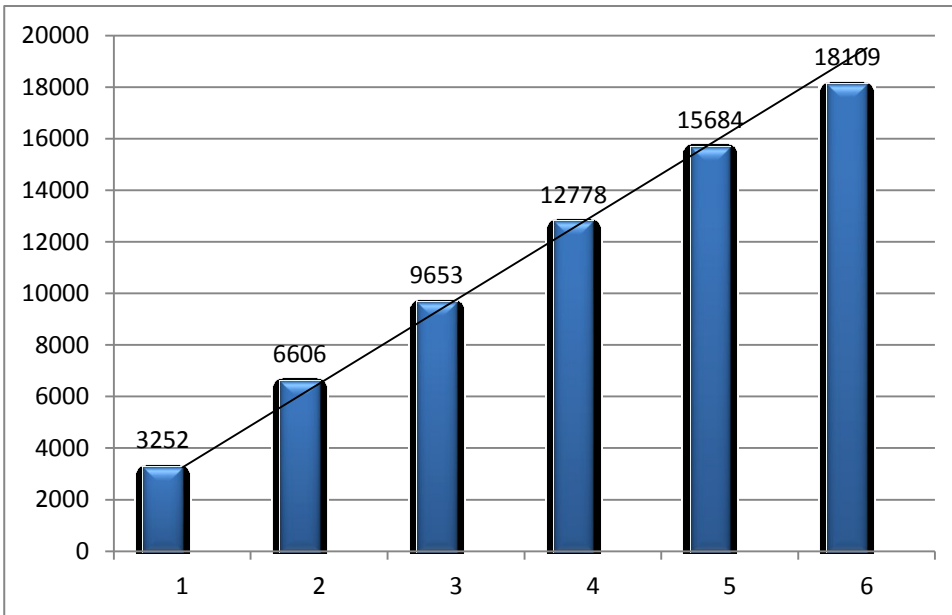


Figure 4: File creates per sec for 1 to 6 director blades

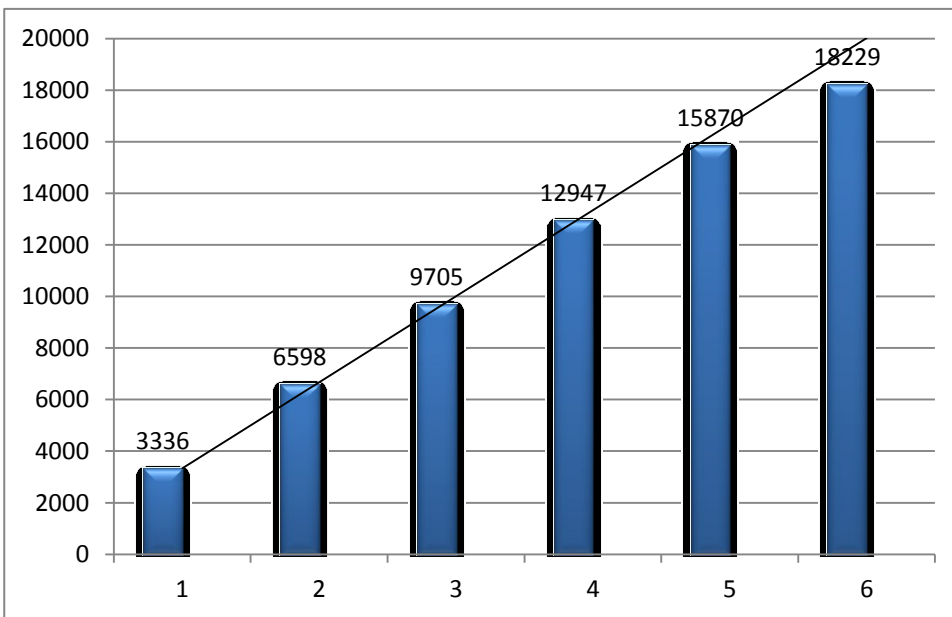


Figure 5: Directory creates per sec for 1 to 6 director blades

Figures 6 and 7 show the performance numbers for removing files and directories. File removals are in general about 10% faster; scalability is even better, especially for file directory removals, where we observed a super linear speedup.

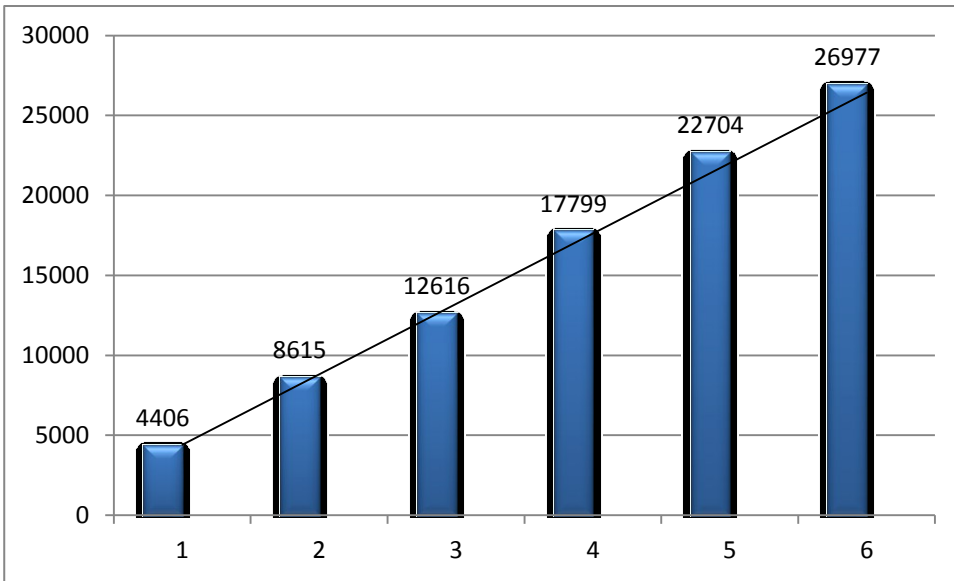


Figure 6: File removals per sec for 1 to 6 director blades

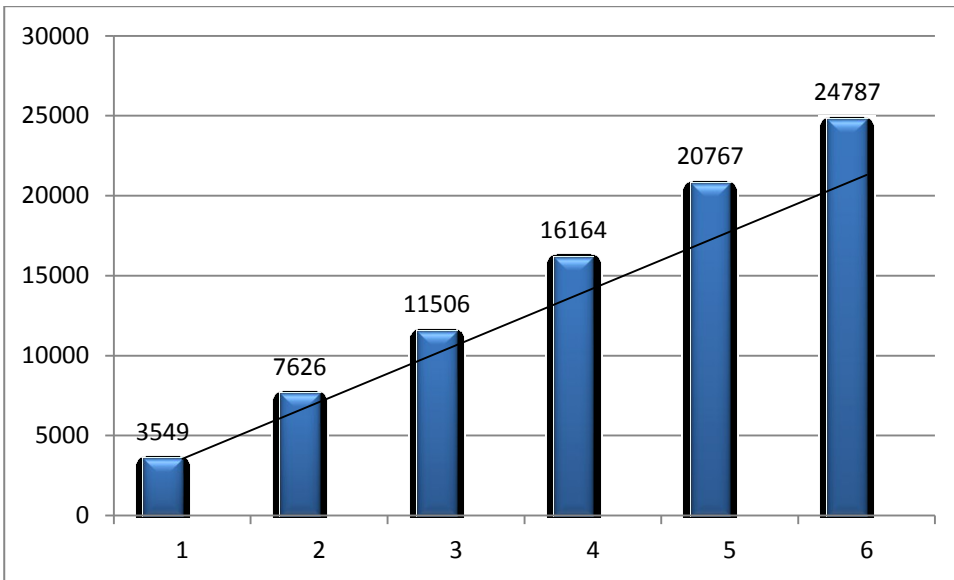


Figure 7: Directory removals per sec for 1 to 6 director blades

All measurements below were done with PanFS 5.0.0.b. Our test bed was initially running PanFS 4.1.3 and we observed scaling problems for the whole mdtest suite. As an example, we compare the result for file creation for up to 6 directory blades for Pan FS 5 and 4.1.3 in figure 8.

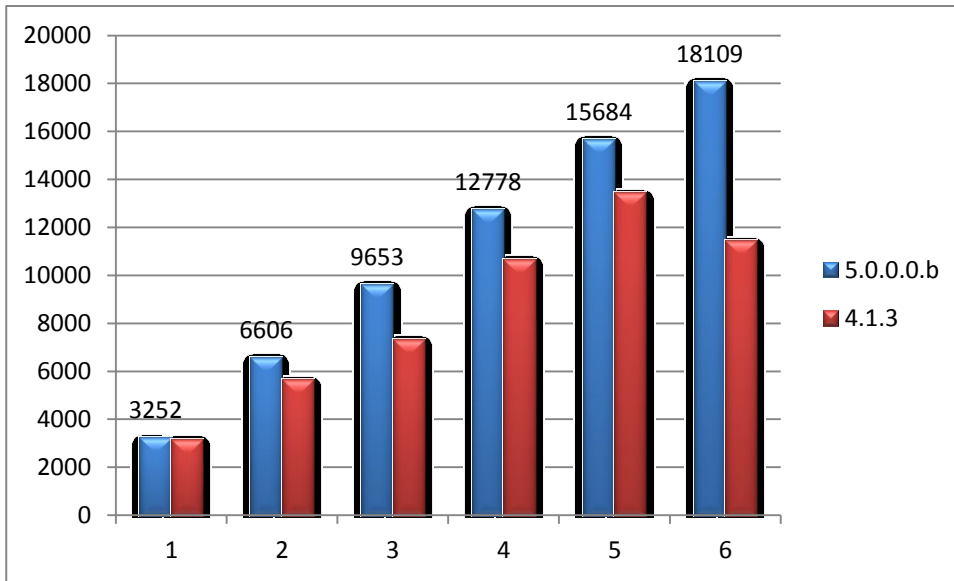


Figure 8: Comparison between Firmware PanFS 4.1.3 and 5.0.0.0.b

5 Impact of directory blade configuration on IO performance

There are different configuration scenarios (1+10, 2+9, 3+8) available with Panasas. We had the chance to investigate also the 2+9 configuration.

Since files are striped across different OSDs it is important to understand how data flows across storage blades. Select the appropriate parameters for `stripe_width`, `group_depth`, `group_width` and `stripe_unit`. (see also man pages for [panasas layout](#)). The default setting for these parameters may not be the best choice for a high throughput performance. During our first test with 2+9 configuration, the write performance did not exceed 3GB/s. Manual tuning of parameters `stripe_width`, `group_width` and `stripe_unit` resulted in 4GB/s. Therefore we recommend to set these numbers according to the number of storage blades and shelves.

For high-throughput environments we believe that 1+10 is the best choice. With 2+9 performance will decrease by 160MB/s per shelf and it would be difficult to map the tuning parameters to such layout ; any miss mapping would result in performance degradation.

The final results in the figures above were retrieved with these values:

```
stripe_unit=128K
stripe_width=10
group_width=30
group_depth=9000
```

6 Resiliency tests

The upgrade procedure of the software stack for our test bed was easy to deploy and took 20 minutes to complete. The actual time might depend on the amount of data and the type of upgrade (minor or major).

We also simulated a failure of a single storage blade; we noticed a drop of 80% performance of the job running when the failure took place, but the file system was stable.

After replacing the failed storage blade, all Raid volumes (one for each director blade) will be rebuilding and rebalancing will take place in background.

In our case it took 10 minutes for integrating a new storage blade and 25 minutes of rebuild time. After that the system was back to normal status.

- ✓ Redundant power supplies and fans in all directors and storage blades
- ✓ Built-in UPS for power failure protection
- ✓ ECC protected memory

- ✓ Mirrored Blade OS – protection against errors in the OS partition
- ✓ Raid5 or Raid10 redundancy per file
- ✓ Proactive monitoring including SMART, heat, fans and battery
- ✓ FreeBSD base operating system

7 Summary

The Panasas ActiveStor solution is a mature alternative for a parallel file system for a HPC cluster.

Our three-shelves PAS12 solution achieved an aggregate write throughput of 4.56GB/s and 4.01GB/s of aggregate read throughput. As a result, this is rather close compared to the theoretical peak throughput for this configuration of 4.8GB/s (write) and 4.5GB/s (read).

The solution includes a couple of high resiliency features; the hardware is hot-swappable and runs very stable. A deep knowledge about the system is not required to make hardware intervention, blades can be easily replace and the system will return to an optimal state without any additional interventions.

Panasas is an impressive complete solution and minimizes deployment complexity. The solution does not require any IO servers and the steps to access the filesystem are as simple as:

- connect the clients to the network
- install panfs packages
- mount the filesystem

The ActiveStor solution has some limitations comparing to alternative solutions:

- The solution is not particular dense. Three shelves are 12U rack units high and deploy 60 drives; alternative solutions are available on the market with up to 240 drives in the same rack space
- The performance of 4.5GB/s aggregate throughout in 12U rack units is rather low compared to other storage solutions with up to 18GB/s aggregate throughput in the same rack space
- PAS12 does not support InfiniBand. Whenever the storage solution is attached to an InfiniBand cluster additional routers and 10GbE switche(s) are required with the appropriate number of GbE and IB ports to saturate the aggregate throughput. The integration and deployment of the routers and the switch was seamless

The management tool called PanActive Manager (GUI or CLI) is clearly arranged and provides all necessary information.

We had to optimize the default PanFS system parameters to achieve good IO throughput numbers, without manual intervention the performance degradation can be significant. The 1+10 configuration provides the highest throughput performance.

8 Outlook

We are planning to investigate the next generation system Panasas ActiveStor14, which provides a significantly higher Metadata Performance realized by SSDs. Each storage blade has one 480 GB SSD that contains PanFS OS, file system metadata and user data.

	ActiveStor 14	ActiveStor 12	ActiveStor 11
ActiveStor Generation	Fifth	Fourth	Fourth
Product Focus	Highest Throughput and High IOPS	Very High Throughput	High Throughput
Capacity (TB)	83 / 81 / 45	60 / 40	60 / 40
SSD Acceleration	Yes	No	No
Throughput, Write/Read (MB/sec)	1,600 / 1,500	1,600 / 1,500	950 / 1,150
4KB File Reads/s	14,150	1,350	1,300
Metadata Performance (Stats/s)	14,150	2,750	2,650
Cache (GB)	92 / 172	92	48
High Availability Network Failover	Standard	Standard	Optional
Link Aggregation	Yes	Yes	No

Note: A single 1+10 shelf configuration is assumed throughout except that the ActiveStor 14 4KB file reads/s and metadata performance numbers are based on a single 2+9 shelf of ActiveStor 14T (a 1+10 shelf of ActiveStor 14 would be lower than shown but still much higher than AS11 or 12). Volume fail-over was on for all configurations.

Figure 9: Taken from [7]

9 Literature

- [1] Panasas [datasheet](#), 2012
- [2] Tony Palmer and Ginny Roth, [ESG Lab Validation Report](#) PAS12 Panasas, February 2011.
- [3] Panasas ActiveStor 11&12, [Product Datasheet](#), 2012
- [4] Panasas PAS12 [Performance Brief](#), 2010
- [5] IOzone file-system performance benchmark utility, www.iozone.org
- [6] Mdtest MPI metadata benchmark. <http://sourceforge.net/projects/mdtest/>
- [7] Derek Burke, [Presentation](#), Panasas, 2012