Evaluation of the NetApp E5560 Storage System

Hussein N. El-Harake, Thomas Schoenemeyer
Swiss National Supercomputing Centre (CSCS)
{hussein, thomas.schoenemeyer}@cscs.ch

Michael Kluge, Wolfgang E. Nagel
Technische Universität Dresden (TUD)
{michael.kluge, wolfgang.nagel}@tu-dresden.de

Abstract: In the frame of the collaboration activities between the Swiss National Supercomputing Center (CSCS) and the Technical University of Dresden (TUD), we evaluated the NetApp E5560 system installed pre GA at the University of Dresden. The purpose of this study was to evaluate the mentioned product, including a comparison between the two storage technologies DDP (Dynamic Disk Pool) and RAID6. We investigated different benchmark scenarios that cover reliability and scalability issues. The benchmarks were run against idle controllers and controllers that had to rebuild one or two failed drives at the same time.

We describe the hardware, infrastructure, the system software stack, and the benchmark tools used. The tools obdfilter-survey, IOR and dd have been used for benchmarking. Most of the test where run multiple times and the best value found is used for this paper. The tests utilized four servers connected to four controller pairs with a total of 120 drives for every controller pair. SAS 2.0 links are used to connect to the DE6600 enclosures and to the I/O servers, while IB FDR is used to connect the file servers with the clients.

1 Introduction

Applications used in High Performance Computing are capable of delivering immense amounts of data. Thus, HPC systems should have access to adequate I/O subsystems which are capable of scaling with such systems. A highly scalable bandwidth and a low latency are the main requirements.

The NetApp E5560 is an HPC storage system that provides block access. For a complete HPC storage solution, additional file servers provide a file system on top of the block storage. The four servers in our case are connected to the controllers using SAS technology; it is also possible to use InfiniBand. The E5560 is a scalable HPC storage solution where the size and the performance capabilities scale with the size of the system. Recently, NetApp started to use Dynamic Disk Pools (DDP) as an alternative storage technology to the traditional RAID5/6 based approach. The system we evaluated is based on units that have two controllers (acting as active/active failover pair) and two enclosures. Each enclosure contains 60 NL-SATA drives of 3TB capacity. A total of 260 TB of usable space are available per controller pair. The controllers reside at the back of the disk chassis and consume no extra rack space.

Figure 1: The NetApp E5560 / DE6600 chassis

Each of the file server has two Sandy Bridge Processors with 8 cores (2.0 GHz), 64 GB RAM, and four dual port SAS cards. From the eight available SAS ports, one cable is drawn to each of the eight E5560 controllers.

From the E5560 datasheet [1], a standalone controller should be able to deliver 6 GB/s in sequential write with enabled cache mirroring. The system scales up to 1,44 PB raw capacity and 12 GB/s in sequential read throughput; while 150.000 in IOPS could be reached with 15K RPM drives. The TUD installed 8 x E5560 controllers (2 x E5560 per chassis) in early 2013. 2 x DE6600 (60 disks per DE6600) enclosures were connected to every pair of controllers of E5560. We executed different benchmark scenarios to compare the numbers on our system with the announced numbers.

2 RAID6 vs. DDP

The experiments described below were performed using Lustre version 2.1.3 (installed by Bull) and Ext4 (as delivered with RHEL6) on the I/O servers. The first objective was to compare DDP and RAID6. Two LUNs were used, the RAID6 LUN was build from 8+2 drives and the DDP LUN was based on 10+1 drives (minimum number of drives required to create a DDP volume). DDP will actually use all of the 11 disks and write data to it. In the mentioned comparison we run dd tests on locally mounted ext4 file-systems. We repeated the same test with ~90% of used capacity on the ext4 file-system to determine an impact of the filling level of the volume on the performance.

Our plan was to run a complete system test for every layer of the storage technology but that was not possible, RAID arrays were already in production mode and we had to live with one array of every storage technology. For the dd tests, we run 4 parallel threads each creating 105 GB (more threads didn't help in scaling).

	cache ON		cache OFF	
Format	**MB/s**	**MB/s 95%**	**MB/s**	**MB/s 95%**
RAID6	1057	1054	380	380
DDP	856	854	438	394

Table 1: Ext4 bandwidth results from DDP and RAID6 using dd; all measurements are done with large block sizes and 4 tasks in parallel; in the "ON" case, cache mirroring was enabled

Results for the single LUN tests are depicted in Table 1. We can see that caching had an impact on the results for both RAID6 and DDP, but almost no difference between RAID6 and DDP. We can see that RAID6 showed better throughput (about 17%) enabled cache. With caching disabled, these numbers changed and DDP showed better performance (impact of number of drives). We didn't see drops in the results if we fill up the raid arrays to 95%. We noticed that parallel dd runs didn't scale well especially with caching disabled, better numbers using obd_filter_survey are shown without caching on Figures 2 to 4.

2.1 Statistical Data

After testing a single LUN we verified that all LUNs (RAID6) in the whole system provide almost the same bandwidth.

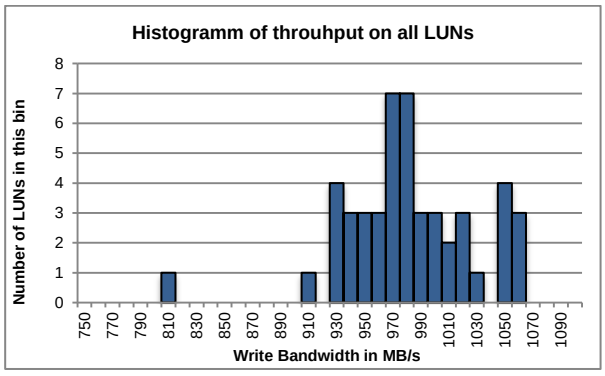

Figure 2: maximum performance observed on all LUNs

In our case (Figure 2) we can conclude that we have a performance problem on one of the LUNs and in addition that the variety in bandwidth is unexpected high (about 150 MB/s between the fastest and the slowest LUN, which is in turn about 15% of the total

performance of the fastest LUN). We will use the NetApp Disk Drive Response Time (DDRT) Utility to investigate this further.

3 Multiple LUN tests

After testing the performance of all individual LUNs the next logical step was to test multiple LUNs from a single server. In the current setup, each server has four cards and eight SAS ports in total. Thus, using two LUNs for performance tests can be actually be done in different ways. It is possible to use one SAS card and two LUNs exported from the same controller (naturally over the same SAS port on this card). Then, one can use two LUNs that are visible over the same card, but over different ports and finally, it is possible to use two LUNs from two different cards (and different controllers). Similar setups can be used for four (and maybe even more) LUNs. The results are summarized in Table 2. Moving from one to two LUNs using the same SAS port gives only a performance increase of about 50%, while writing to two LUNs to the same controller over two different cards almost doubles the performance. What appears a bit strange is the fact that writing to two LUNs, each on a different controller, does result in much less performance that using two LUNs on the same controller. This artifact is probably due to congestion on the PCI-Express bus or to the fact that the controllers mirror their cache. We have not done additional test to verify any of these hypotheses.

After moving to four LUNs we see that the performance observed in the case of two LUNs on two controllers doubles, if four LUNs from two controllers are used. If we use four LUNs on four different controllers, we see better performance of almost four times the performance of a single LUN.

LUNs	Cards	Ports	Controllers	Bandwidth in MB/s
1	1	1	1	1030
2	1	1	1	1548
2	2	2	1	1922
2	2	2	2	1440
4	2	2	2	2859
4	2	4	4	3714

Table 2: Performance data for accessing multiple LUNs in various configurations

4 Full Controller Bandwidth

The next test was a performance throughput test using obdfilter_survey on a complete controller using all 12 LUNs (RAID6). We run this test twice one with controller cache enabled (including cache mirroring) and controller cache disabled. obdfilter_survey is somehow similar to running a raw test on a standard disk, except that the disk is already formatted with ldiskfs. Hence, such tests will help understanding the peak performance of the system. The results are summarized in Figures 3 and 4.

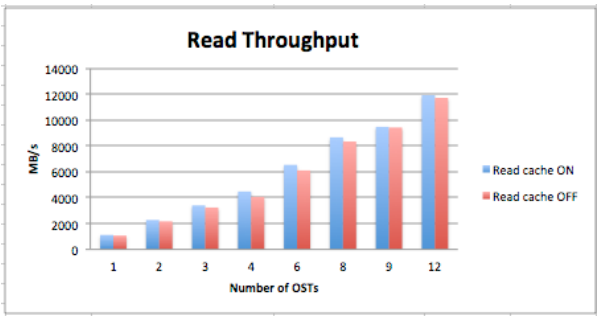

Figure 3: obdfilter_survey read throughput

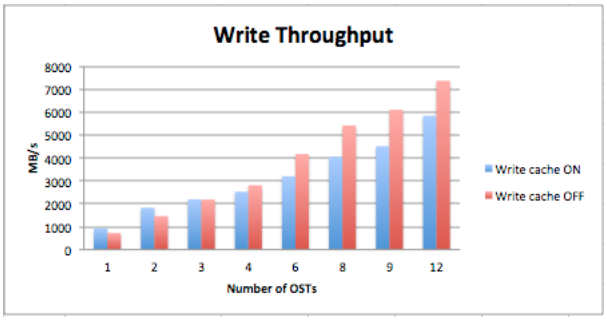

Figure 4: obdfilter_survey write throughput

Using obdfilter_survey showed very interesting numbers on write and read accesses. We used 1MB as block size for both tests. In read perfor-

mance, we saw few improvements if caching is enabled. For writes, we saw about 20% improvements if caching is disabled. This is valid only in cases where block sizes of 1MB or above are used. (NetApp/LSI used to have an interesting feature to bypass the cache for certain block sizes. This feature is still available.) In total we were able to achieve a write bandwidth of about 7.3 GB/s on a single controller pair and a read bandwidth of about 12 GB/s.

5 Lustre Throughput

For the final acceptance tests for the system were done on 450 clients and achieved a maximum throughput of 21 GiB/s for write and 27 GB/s for read accesses. This test has been done using IOR and a 10 MiB block size. With 48 LUNs used in parallel, where each LUN delivers at least 810 MB/s, about 38 GB/s would have been an absolute maximum. To figure out where the performance for a full system tests gets lost, we have conducted a preliminary set of tests. Currently, we cannot get more than about 5,5 GB/s for write accesses with obd_filter_survey from a single server while using all of it eight SAS ports and all attached LUNs. If this issue is investigated further, more global bandwidth will be available to the clients.

6 Conclusion

In this report we evaluated the NetApp 5560 storage subsystem installed at the Technische Universität Dresden for the HRSK-II project. The NetApp 5560 is a solid HPC storage solution that scales with the number of installed controllers and disk shelves. We have seen 7.3GB/s for write with large block sizes and 12GB/s for read accesses for a single controller pair with 120 disks, matching resp. exceeding the numbers announced by NetApp.

For stability we tried different scenarios that might occur in real environment like removing one or more drives during intensive I/O phases and rebuilding failed drives during high I/O load. The system was very stable during these tests and almost didn't show any performance impact. For scalability we will continue our investigation to improve the 5.5GB/s for write accesses as described in Section 5.

We found no performance impact for single LUNs between the traditional RAID6 LUNs and DDP LUNs. Caching on the other hand has a big impact on the performance of a single LUN as well as on full controller tests. For single LUNs, caching helps quite a lot. While for full controller tests with large block sizes, it needs to be disabled to reach maximum performance numbers. We want to thank both BULL and NetApp for enabling and supporting these tests.

7 References

[1] NetApp E5560 documentation
https://communities.netapp.com/docs/DOC-23812

[2] Lustre HPC Parallel File System Wiki
http://wiki.lustre.org/index.php/Main_Page

[3] IOR file-system performance benchmark utility,
http://sourceforge.net/projects/ior-sio

[4] Man page of the dd utility
http://linux.die.net/man/1/dd