Commoditisation of the High-End Research Storage Market with the Dell MD3460 & Intel Enterprise Edition Lustre

University of Cambridge, UIS, HPC Service

Authors: Wojciech Turek, Paul Calleja, John Taylor

Table of Contents

Abstract

This paper clearly demonstrates that once optimised for large I/O throughput the Dell MD3460 / Intel Enterprise Edition Lustre (IEEL) solution provides storage density and performance characteristics that are very well aligned to the requirements of the mid-to-high end research storage market. After the throughput tuning had been applied the I/O performance of the Dell storage brick doubled, producing single brick IOR client performance maxima of 4.5GB/s R/W. Single rack configurations can thus be implemented that provide 2.1 PB of usable storage and 36 GB/s R/W performance. A capacity optimised configuration is also illustrated providing a solution with a cost reduction of ~35% relative to the performance optimised solution. These bulk performance and density metrics place the Dell / IEEL solution at the high end of the solution space but within the commodity IT supply chain model. This will provide the price performance step change that the scientific, technical and medical research computing communities need to help close the demand vs. budget gap that has emerged due to huge growth in demand seen within the research community for both storage capacity and performance. This marks a turning point in commoditisation of research storage solutions echoing the commodity revolution that was seen in research computing market with the advent of HPC clusters. Many large scale HPC customers are finding it difficult to architect HPC and data analysis system with the required capacity, performance and cost parameters. Commodity high end parallel files system as described in this paper dramatically improve this situation.

Introduction

The scientific, technical and medical research computing domains are currently undergoing a data explosion driving rapid growth in demand for storage capacity and performance. Growth in research computing budgets are not keeping pace with increasing storage demands. Thus we are seeing the emergence of a ***research storage demand vs. budget gap.*** A large step change improvement in research storage price-performance ratio is required to close this demand-budget gap and enable the research community to meet its increasing data storage demands within a world of static or slow growing budgets.

The multi-petabyte multi-10GB/s throughput research storage solution space has yet to undergo mainstream commoditisation akin to what has already happened in research computing market. Mainstream commoditisation of the HPC "compute" market in the late 90's with the advent of HPC clusters transformed the price-performance of large scale compute solutions, but the storage systems they depend on are still largely met by proprietary vendor solution silos. Thus the price performance gains seen with HPC clusters has not been seen with research storage leading to the current day demand-budget gap. What is needed is mainstream commoditisation of the research storage market.

The combination of the Dell MD4360 storage array with Intel Enterprise Edition Lustre provides the first commodity research storage solution with the performance, features and full OEM support needed to satisfy the mainstream mid-high end research computing market.

This paper examines I/O throughput performance optimisation for the Dell/Intel commodity lustre solution demonstrating how to unlock the full performance of the system. The paper then illustrates a number of different Petabyte scale single rack configurations that are optimised for either performance or capacity, highlighting the overall fit of the solution within the research computing space.

The paper starts by analysing performance on the system with default settings and then describes tuning methods for the Power Vault MD3460 storage system focused on optimising I/O for the Intel Enterprise edition Lustre file system. This paper is focused on Dell/Intel Lustre I/O throughput, future papers in the series will look at Dell/Intel Lustre metadata/IOPS performance and Dell/Intel Lustre features and functionality.

Lustre file system

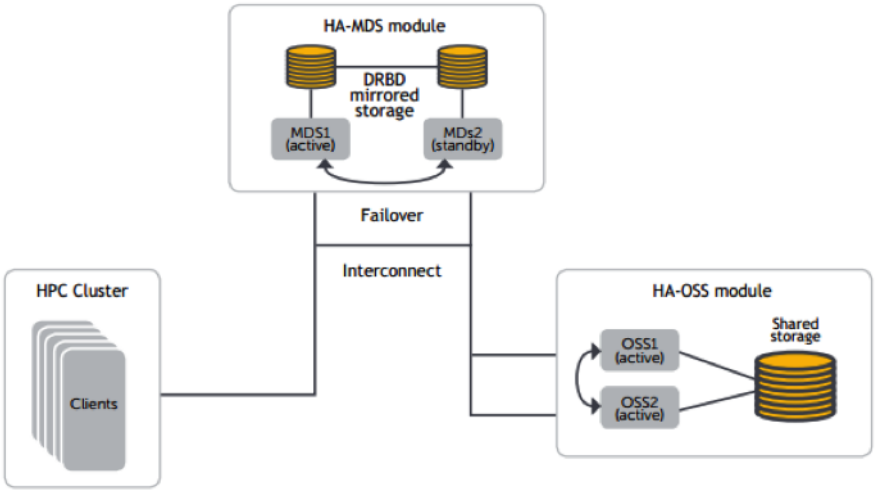

Figure1 Lustre file system

Lustre provides a storage architecture for clusters which allows significant freedom in hardware implementation. At the user level the Lustre filesystem provides a POSIX-compliant UNIX filesystem interface. The main components of Lustre are the Management server (MGS), Metadata Server (MDS), Object Storage Server (OSS) and the Lustre client. The Lustre file system uses an object-based storage model and provides several abstractions designed to improve both performance and scalability. At the file system level, Lustre treats files as objects which are located through the MDS. Metadata Servers support all file system name space operations, such as file lookups, file creation, and file and directory attribute manipulation. This metadata information is physically stored on the metadata target device (MDT). Multiple MDT devices can be used per filesystem to improve the performance and scalability of the metadata operations. The Management Target is a registration point for all the devices (MDT, OST, clients) in the Lustre file system. The Management Server and Target have a central role in the new recovery model (Imperative Recovery) introduced in lustre 2.2. Because of the increased importance of the MGS in recovery, it is strongly recommended that the MGS node be separate from the MDS. If the MGS is co-located on the MDS node, then in case of MDS/MGS failure there will be no IR notification for the MDS restart, and clients will always use timeout-based recovery for the MDS. IR notification would still be used in the case of OSS failure and recovery.

File data is stored in objects on the object storage targets (OST) which are managed by OSSs. The MDS directs actual file I/O requests from a Lustre client to the appropriate OST, which manages the objects that are physically located on the underlying storage block devices. Once the MDS identifies the storage location of a file, all subsequent file I/O is performed between the client and the OSSs. The Lustre clients are typically HPC cluster compute nodes which run Lustre client software and communicate with Lustre servers over Ethernet or Infiniband. The Lustre client software consists of an interface between the Linux virtual filesystem and the Lustre servers. Each server target has a client counterpart: Metadata Client (MDC), Object Storage Client (OSC), and a Management Client (MGC). OSCs are grouped into a single Logical Object Volume (LOV), which is the basis for transparent access to the file system. Also the MGCs are grouped into a single Logical Metadata Volume (LMV) in order to provide transparent scalability.

Clients mounting the Lustre file system see a single, coherent, synchronised namespace at all times. Different clients can write to different parts of the same file at the same time, while other clients read from the file. This design divides file system operation into two distinct parts: file system metadata operations on the MDS and file data operations on the OSSs. This approach not only improves filesystem performance but also other important operational aspects such as availability and recovery times. As shown in Figure 1, the Lustre file system is built on scalable modules and can support a variety of hardware platforms and interconnects.

Test System Overview

This technical paper focuses on a single OSS storage block and optimising its throughput performance when subjected to a sequential Lustre IO. Typically production configurations deploy two blocks of OSS storage to provide failover capability. Since this paper focuses mainly on performance capabilities of the Dell storage, a single OSS only configuration will be used. The test platform consists of one R620 OSS server and two disk enclosures, MD3460 with one expansion enclosure MD3060E.

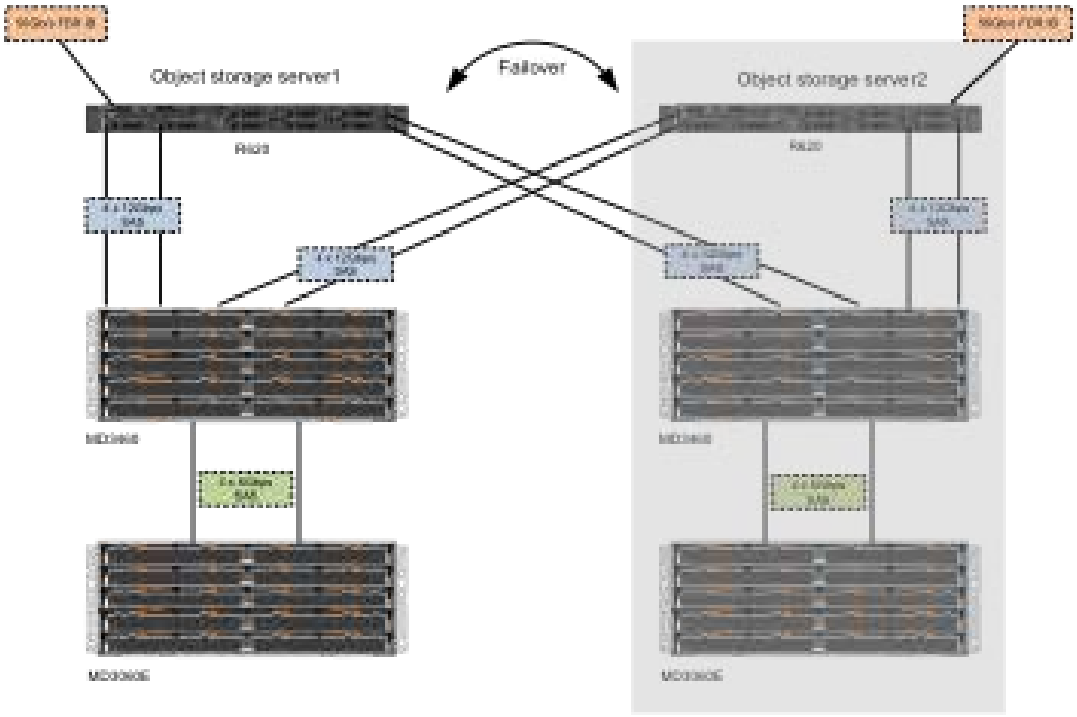

Figure2 Dell Lustre Storage Test System

Dell Lustre Storage	
Component	Description
Lustre server version	IEEL 2.0.1
OSS Nodes	R620
OSS Memory	32 GB 1600Mhz
OSS Processors	CPU E5-2420 v2 @ 2.20GHz
OSS SAS HBA	2 x 12Gbps HBA
OSS IB HCA	Mellanox 56Gb/s FDR HCA
OSS Storage Arrays	Dell MD3460 and Dell MD3060E
Storage	120 x 4TB NL SAS

Table1: Lustre OSS storage specification

Lustre Clients	
Component	Description
Lustre server version	IEEL 2.0.1
OSS Nodes	C6220
OSS Memory	64 GB 1600Mhz
OSS Processors	Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz
OSS IB HCA	Mellanox 56Gb/s FDR HCA

Table2: Lustre client specification

Figure3 OSS server R620

The Dell R620 server provides 3 x PCIe ports, allowing 2 SAS HBAs and, in this case, an (FDR) Infiniband card. This provides a good match for the backend storage and client side throughput. The Object Storage servers are the basic building blocks of the solution and provide an easy way to scale the storage with the demand. In production configuration storage system would use 2 of the OSS server redundantly connected to the high density Dell MD3460 storage arrays.

The MD3460 and MD3060E are high density disk arrays and deliver 60 HDDs in per 4U of rack space. The MD3460 disk enclosure is equipped with dual redundant RAID controllers with BBU cache. MD3460 provides 4 x 12Gbps SAS host ports and each host port consists of 4 x 12Gbps SAS lanes giving 48Gbps per host port. Each storage array is divided into 6 RAID virtual disks consisting of 8 data and 2 parity disks. Raid configuration is optimised for 1MB I/O request size. Each OST when formatted with Lustre file system provides 29TB of usable capacity. Using expansion enclosure allows doubling the capacity of the solution without doubling the cost.

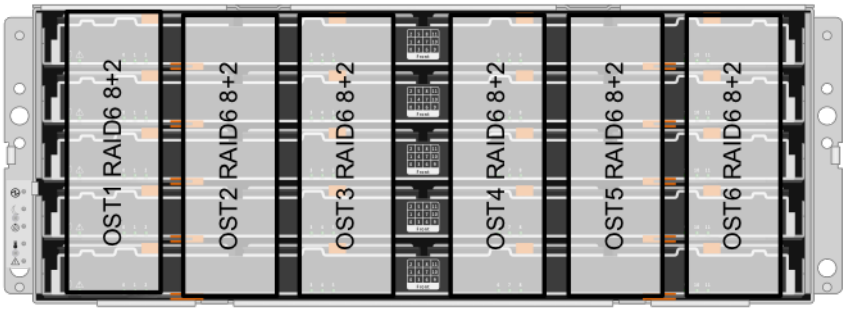

Figure4 Lustre storage disks enclosure

Linux large I/O tuning

Typically the Linux kernel is tuned to work with a range of I/O workloads and is not optimised for very large I/Os. In HPC the typical I/O requests are large and therefore the storage servers performing I/O workload should be optimised accordingly. The test system uses the CentOS 6.5 Linux distribution with a Lustre patched kernel. Lustre by default is tuned to work with 1MB RPCs and ideally it should be avoided to split those when submitting to disk. Therefore the entire storage system should be tuned and aligned with 1MB I/O request size. The major problem for SAS connected storage systems is that by default the *mpt2sas* and *mpt3sas* drivers which handle the MD storage devices are by default limited to maximum 512KB I/O request size. That in turn causes fragmentation of 1MB Lustre RPC. This device limit however can be raised to 1MB with little effort. The parameter responsible for allowing the SAS driver to carry out large I/O requests is called SCSI_MPT2SAS_MAX_SGE the LSI MPT Fusion Max number of SG Entries. Most mainstream Linux distributions still provide kernel with SAS HBA driver configured and compiled with SCSI_MPT2SAS_MAX_SGE=128. That enforces max 128 segments per I/O, that in turn with segment size of 4Kb results in max 512Kb I/O requests. The value for the SCSI_MPT2SAS_MAX_SGE is set in the kernel config. It is safe to change that value to 256 and recompile the SAS HBA module. When loading the *mpt2sas* or *mpt3sas* module the module option MAX_SGL_ENTRIES should be set to 256 to ensure that correct parameter value is set. This will allow the SCSI device parameters to be tuned to allow for 1MB I/O requests to be committed to the disk without fragmentation. Also on the newer 12Gbps SAS cards the maximum queue depth size is bigger than the default value and also could be increased. Table3 lists the Linux parameters that need to be changed to obtain optimal IO performance. Each Lustre mount operation may change some of the parameters. This parameters should be reset to their optimal value after mounting Lustre.

Linux tuning for large I/O	
Parameter name	**Value**
scheduler	deadline
max_sgl_entries	256
max_queue_depth	600
max_sectors_kb	1024
nr_requests	1024
read_ahead_kb	8192
rq_affinity	2
redhat_transparent_hugepage	never
vfs_cache_pressure	50

Table 3

MD3460 large I/O tuning

The Power Vault MD3460 comes equipped with two redundant RAID controllers that typically work in active-active mode. Both RAID controllers need to be configured and tuned to handle the large I/O requests efficiently. The 60 disks are divided into six RAID6 groups. Each RAID6 group consists of ten disks in 8+2 configuration. Disks groups are tuned for a 1MB stripe size by creating Virtual Disks with a segment size parameter set to 128KB. This enables full alignment with 1MB I/O requests. In addition the cache block size is set to the maximum 32KB which enables faster cache operations on bigger blocks. There is no benefit from read cache if the read I/O requests are aligned with 1MB stripe size. Therefore it is recommended to disable read cache and use all of the available cache for writes. Write cache with mirroring should always be enabled to ensure data consistency.

MD3460 RAID controller configuration	
Parameter name	**Value**
RAID6	8+2
Segment size	128KB
cache block size	32KB
cache flush	98%
Write cache with mirroring	Enabled
Read cache	Disabled

Table 4

Lustre I/O tuning

The Lustre filesystem can further be tuned on both server and client. The server end tuning is somewhat limited, as by default Lustre is already optimised to work with large I/O sizes. The relevant parameter that needs to be correctly set is called threads_max and threads_min. This parameter decides how many I/O threads will be started on the OSS server to perform I/O operations. The best way to determine the optimal value for this parameter is by running the obdfiler-survey test, which evaluates the storage hardware performance capability. The Power Vault MD3460 storage array is capable of running with the maximum number of OSS threads enabled. At the Lustre client-side the default setting is tuned for moderate I/O sizes and loads and can be further optimised to give better performance numbers. The table below shows the parameters names and their recommended values when optimising for large I/O.

Lustre OSS tuning	
Parameter name	**Value**
threads_max	512
threads_min	512

Table 5

Lustre client parameters tuning	
Parameter name	**Value**
max_rpcs_in_flight	256
max_dirty_mb	1024

Table 6

The purpose of the tests performed in this study is to profile the performance of the Dell HPC storage optimised for Lustre. In the case where the I/O block size of the applications is very high, Lustre can be tuned to support 4MB RPC size.

System Performance evaluation and analysis

Using obdfilter-survey tool for storage tuning and analysis

The Lustre IOKit provides a range of I/O evaluation tools of which one of them is obdfilter-survey. The script profiles the overall throughput of the storage hardware by applying a range of workloads to the OSTs. The main purpose of running *obdfilter-survey* is to measure the maximum performance of a storage system and to find the saturation points which cause performance drops. Test is run from a command line.

obdfilter-survey command line 6 OST run
nobjlo=1 thrlo=24 thrhi=144 size=32768 targets="testfsOST0000 testfs-OST0001 testfs-OST0002 testfs-OST0003 testfs-OST0004 testfsOST0005" ./obdfilter-survey

Table 7

obdfilter-survey command line 16 OST run
nobjlo=1 thrlo=24 thrhi=144 size=32768 targets="testfsOST0000 testfs-OST0001 testfs-OST0002 testfs-OST0003 testfs-OST0004 testfsOST0005 testfsOST0006 testfs-OST0007 testfs-OST0008 testfs-OST0009 testfs-OST000a testfsOST000b" ./obdfilter-survey

Table 8

obj *(Lustre objects) - describes how many Lustre objects are written or read. This parameter simulates multiple Lustre clients accessing the OST and reading/writing multiple objects.*

thr *(number of threads) - this parameter simulates Lustre OSS threads. More OSS threads can do more I/O, but if too many threads are in use and the storage system is not being able to process them the performance will drop.*

The obdfilter-survey benchmark is intended for sequential performance testing throughput capability of the Lustre storage hardware. The test runs on the Lustre OSS storage server itself thus only testing the performance of the storage arrays and not the interconnect.

Obdfilter performance before large I/O optimisation

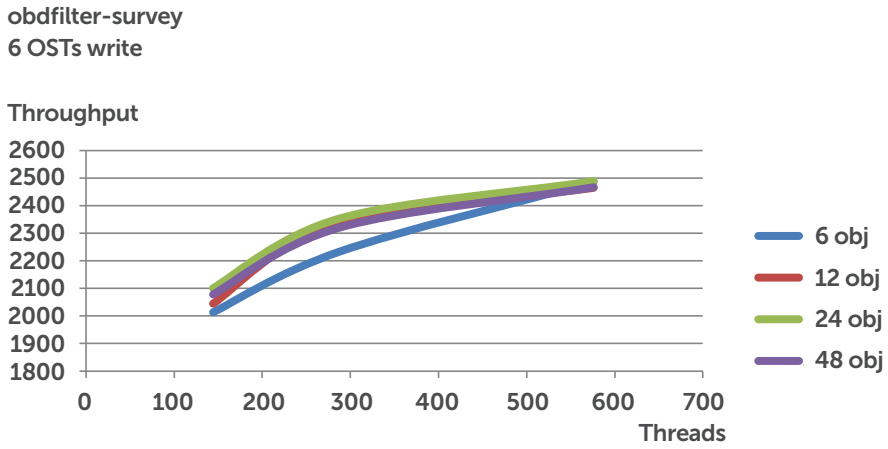

Figure 7 obdfilter-survey write throughput MD3460 only

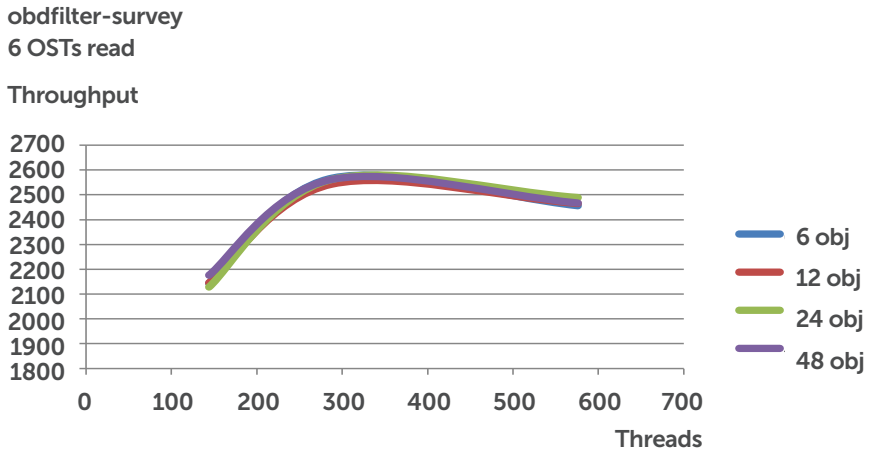

Figure6 obdfilter-survey read throughput MD3460 only

Obdfilter performance after large I/O optimisation

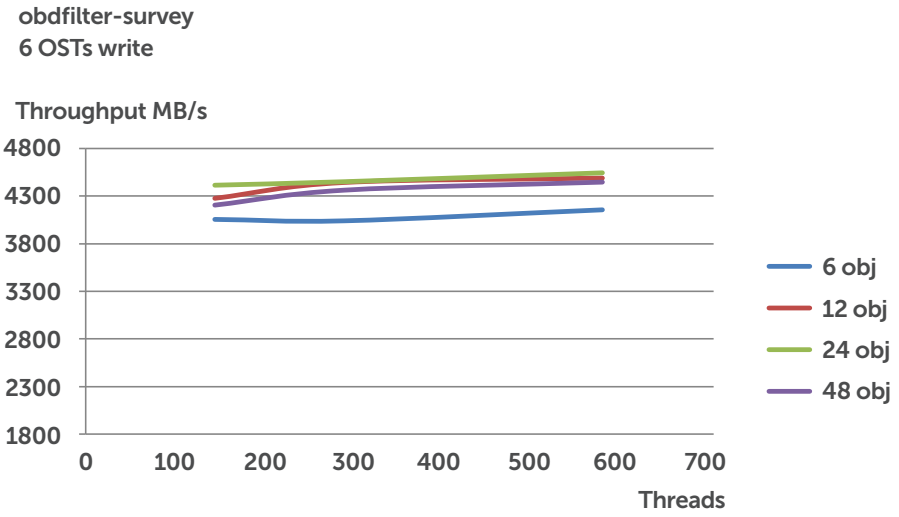

Figure 7 obdfilter-survey write throughput MD3460 only

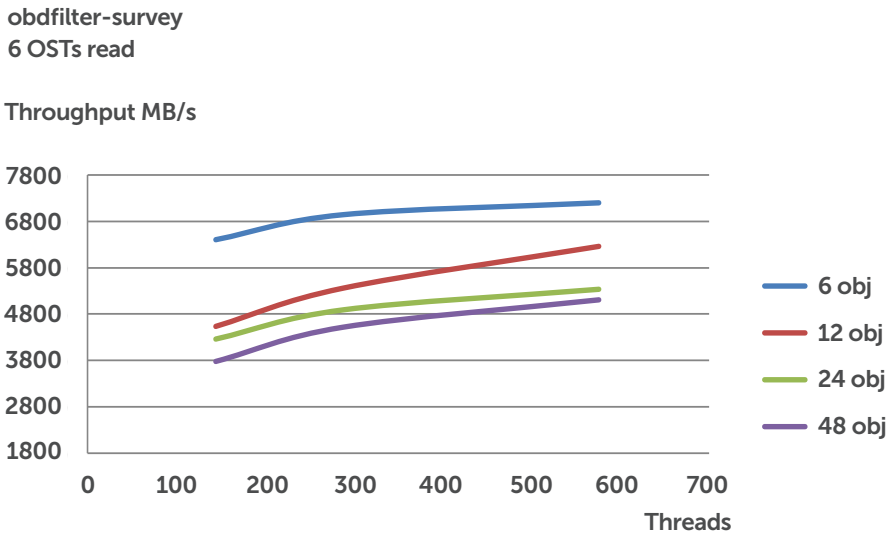

Figure8 obdfilter-survey read throughput MD3460 only

Obdfilter performance after large I/O optimisation

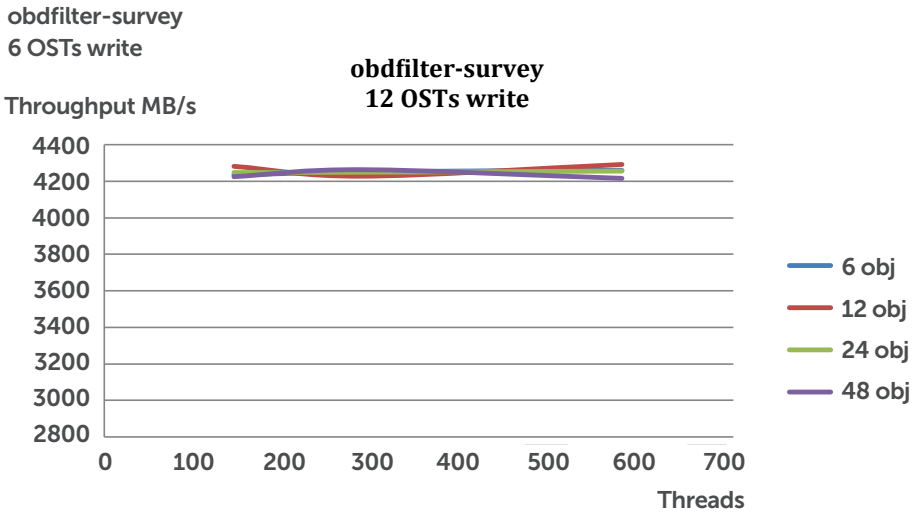

Figure 9 obdfilter-survey write throughput MD3460+MD3060E

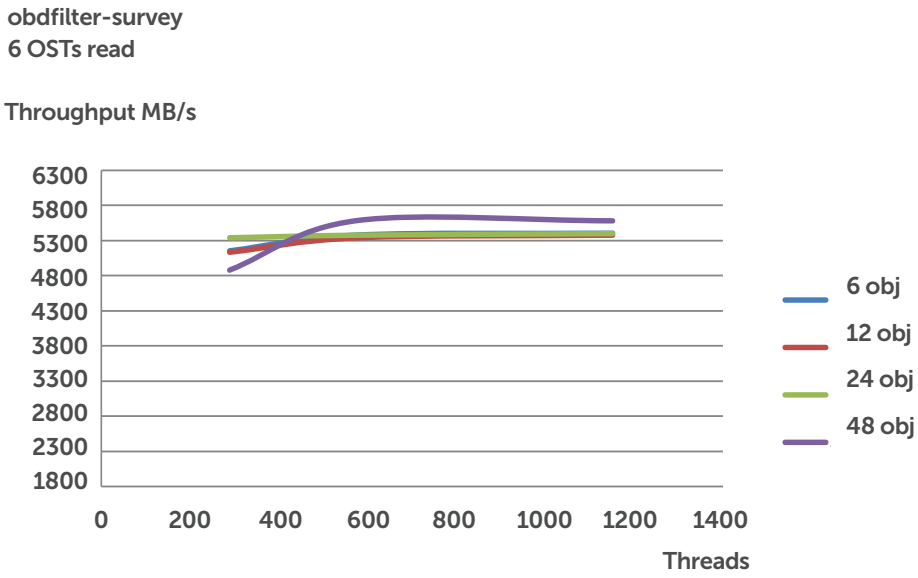

Figure 10 obdfilter-survey read throughput MD3460+MD3060E

The above obdfilter-survey tests were run using on two different storage brick configurations (a) a single storage block consisting of one OSS server and (b) a two disks enclosure solution. The 6 OST charts represent the performance of the single MD3460 disk array. The 12 OST charts represent combine performance of the MD3460 and MD3060E.

Figures 5 and 6 represent benchmark results before applying large I/O optimisation described in this paper. It is clear that the default settings are not optimal for the HPC workloads and Lustre filesystem. This is mainly down to the fact that the I/O request size is limited to maximum 512KB requests and is not aligned with the Lustre I/O size and RAID configuration.

Figures 7, 8, 9 and 10 demonstrate performance of the storage system with the optimal settings applied. Figures 7, 8 show performance of a single MD3460 enclosure (6 OSTs) and figures 9, 10 represent performance of both, MD3460 and MD3060E (12 OSTs).

In the numerous tests we have run we concluded that the peak read and write performance can be obtained with a single MD3460 disk enclosure. Running obdfilter-survay across the two disk enclosures (using all 12 OSTs) does not yield more performance. The MD3060E expansion enclosure connects to the MD3460 disk enclosure via 6Gbps SAS links. This results in slower access to disks in the expansion enclosures bringing the overall performance down.

When I/O is aligned with the RAID stripe size, disabling write cache can improve write performance for large sequential I/O workloads. This is because the I/O is done in a write through mode, which results in less RAID controller operations. Enabling write cache can be beneficial for workloads with short, large I/O bursts. If write cache is enabled it is mandatory to enable cache mirroring to ensure data consistency when using failover software on the storage servers.

IOR benchmark

The Lustre client performance results were obtained using the IOR benchmark. IOR is a popular HPC I/O benchmark providing a wide range of useful functions and features. Running IOR in multi node tests allows clients to first write data and then when reads are performed, clients read data written by another client hence avoiding their own buffer cache. This completely eliminates the client read cache effect, so avoiding the problem of having to flush the client cache after write. IOR uses MPI for multi node communication and thread synchronisation which helps to provide very accurate results from large scale multi node tests.

IOR command line
IOR –vv –w –r –C –b 16g –t 1m –i 3 –m –k –o –F /ltestfs2/wjt27/FPP

Table 9

Obdfilter performance after large I/O optimisation

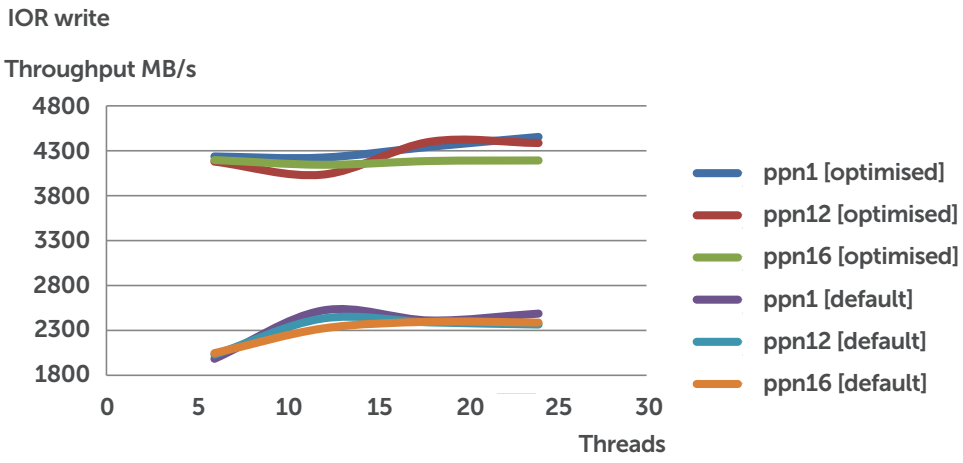

Figure 11 Lustre clients IOR write test

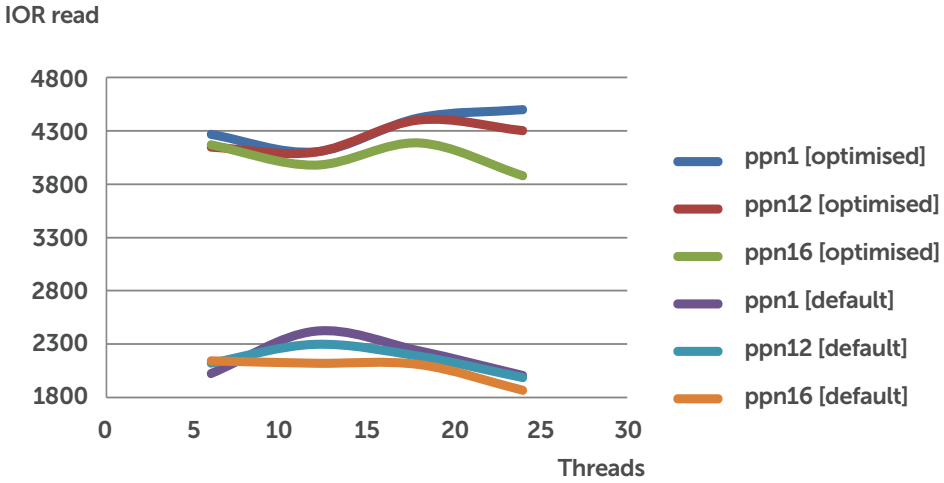

Figure 12 Lustre clients IOR read test

Figures 11 and 12 show both the pre and post optimisation results of IOR benchmark. The performance gain achieved by optimisation methods described in this paper is very high. The optimised performance is almost double of the performance results obtained with default settings. The main reason why default performance is so much lower is mostly down to the SAS driver fragmenting the I/O, hence causing misalignment with the RAID configuration. Because the I/O requests are fragmented and not aligned with the RAID segment size, RAID controllers have to perform read-modify-write operations which are resource expensive and create large overhead. Additionaly, if the write cache with cache mirroring is enabled (mandatory for high availability production enviornments that use write cache) in order to keep the cache coherent across controllers additional data processing is required which results in even more controller overhead.

After optimisation the storage array shows consistent perfomance throughout all tests for both read and write I/O operations. Achieving this performance result was not possible before optimisation because the SAS driver was limiting the I/O request size to 512kB and the I/O arriving to the controller was fragmented and not aligned with the 1MB stripe size. After removing that limitation system can unlock its full I/O potential. The benchmark results were verified by monitoring I/O performance directly on the storage hardware using Dell SMcli tool. The captured I/O profile confirmed the throughput values produced by the benchmark were in ageement with I/O seen on the hardware itself.

Petabyte scale solutions optimised for performance or capacity

The Dell MD3460 RAID enclosure combined with the MD3060E expansion enclosure allows a wide range of spindle to RAID controller configurations to be constructed which change the capacity per rack, performance per rack and cost of the solution. Two contrasting configurations have been illustrated in the figure below

Capacity Configuration
2 * capacity I/O bricks
1PB usable capacity
9 GB/s RW performance
Relative price 0.65

Performance Configuration
6 * performance I/O bricks
1PB usable capacity
26 GB/s RW performance
Relative price 1.0

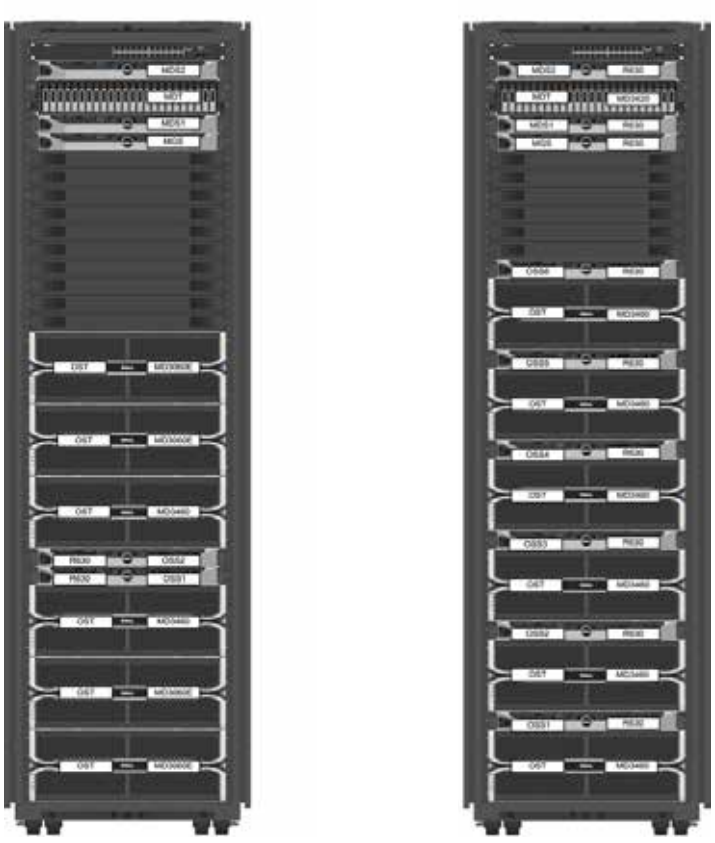

Figure14. PetaByte in a rack - capcity optimised and performance optimised solutions

It can be seen that the capacity optimised rack provides 1PB of storage in a single rack with a performance of 9GB/s RW and a relative cost of 0.65 compared to the performance optimised solution again with 1PB of storage, 29 GB/s RW and a relative price of 1. Thus we can see that the capacity optimised solution is 35% lower cost than the performance optimised solution.

For installations requiring very high denisity solutions the meta-data component of the solution can be housed in a separate rack normally along with other control elements of the cluster. This allows an additional 1 capacity brick or 2 performance bricks to be added to the solution. This would result in each rack having the attributes below:-

Capacity Configuration	Performance Configuration
3 * capacity I/O bricks	8 * performance bricks
1.5PB usable capacity	1.4PB usable capacity
13.5 GB/s RW performance	36 GB/s RW performance
Relative price 0.65	Relative price 1.0

The above configuration uses 4TB disks, if 6TB disks are used the following configuration is obtained

Capacity Configuration	Performance Configuration
3 * capacity I/O bricks	8 * performance bricks
2.3 PB usable capacity	2.1 PB usable capacity
13.5 GB/s RW performance	36 GB/s RW performance
Relative price 0.65	Relative price 1.0

It should be noted that capacity units used here are based on 1024b/Kb not the 1000b/Kb commonly used within the storage industry. This means that the capacity values used here are actually usable by the user and are the same as described by the default *df* command within Linux. Commonly storage vendors use 1000b/Kb in the calculation of filesystem sizes which results in usable file system estimates smaller than will be shown with the default *df* command.

The two possible configurations have been shown here are a 1 enclosure brick and a 3 enclosure brick. More expansion enclosures can be added reducing the cost per PB even further. Although it is likely that the configurations shown here span the range that most suits high performance use cases.

Discussion

The paper clearly demonstrates that once optimised for large I/O throughput the Dell MD3460 / Intel enterprise edition Lustre solution provides storage density and performance characteristics that are very well aligned to the requirements of the mid-high end research storage market. After the throughput tuning had been applied the I/O performance of the Dell storage brick doubled producing single brick IOR client performance maxima of 4.5GB/s R/W. Single rack configurations can be implemented that provide 2.1 PB of storage and 36 GB/s R/W performance. These bulk performance and density metrics place the Dell / Intel Enterprise Lustre solution at the high end of the HPC storage solution space but within a commodity IT supply chain model.
A capacity optimised configuration is also demonstrated that provides a 35% cost advantage for the same storage capacity as compared to the performance optimised solution

This commodity Dell IEEL parallel file system solution will provide the price performance step change that the scientific, technical and medical research computing communities need to help close the demand vs budget gap that has emerged. This marks a turning point in commoditisation of research storage solutions echoing the revolution that was seen in research computing commoditisation with the advent of HPC clusters.

Future papers in this series will examine metadata and IOPs performance achievable from Dell Intel Lustre solutions. As well as an in-depth review and analysis of deployment, operational and monitoring features of the solution. Future papers will also undertake a detailed analysis of the use and performance of the solution within a busy HPC user environment in an attempt to bridge the gap in understanding seen when translating benchmark storage data to performance under real-world conditions.