Technical Bulletin: High Performance Lustre Filesystems using Dell PowerVault MD Storage

Dell / Cambridge HPC Solution Centre
University of Cambridge

Wojciech Turek, Dr Paul Calleja

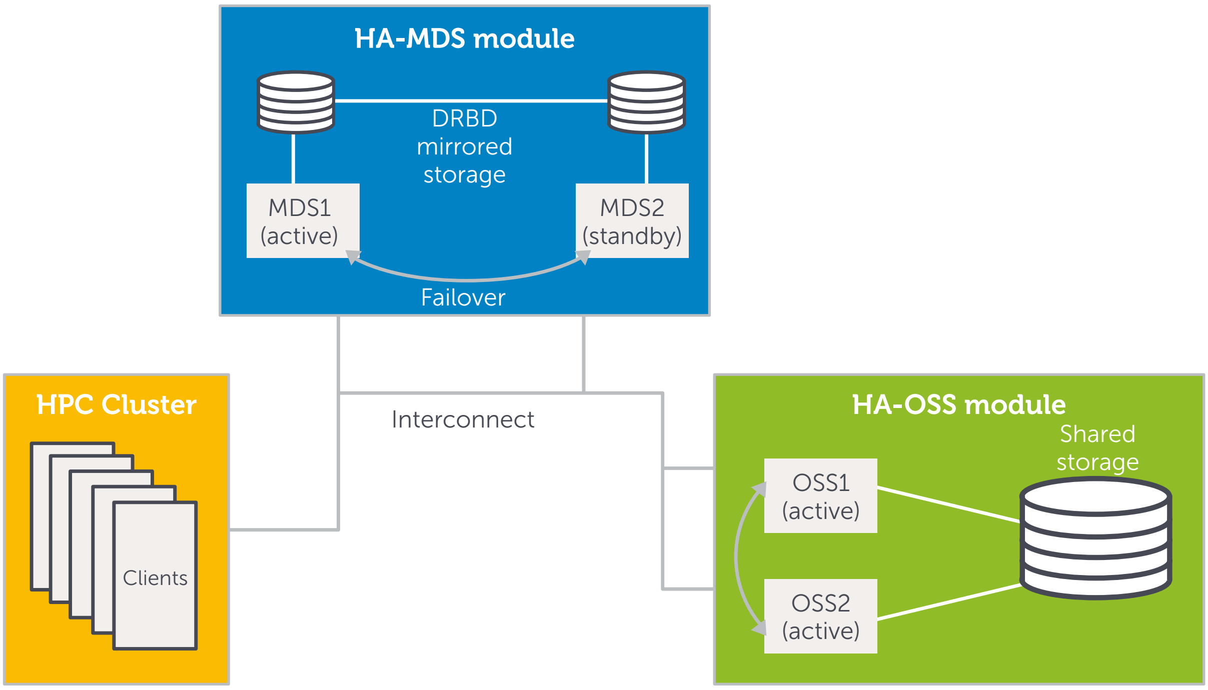
Introduction

This technical bulletin provides an overview of a high performance Lustre solution built using the PowerVault MD3200 and MD1200 storage products, with details of hardware and software configuration and corresponding benchmark results. This paper primarily looks at the Object Storage Server (OSS) component of the Lustre system. The Dell MDxxxx Lustre Module discussed in this paper is designed to optimise I/O throughput while maintaining high availability, and it delivers a maximum read/write performance of 2GB/s per brick. The Lustre file system performance scales almost linearly with the addition of successive modules, so a single 42U rack, housing four such modules, would provide ~200TB of useable storage and around 8GB/s read/write bandwidth. A previous paper provides a full treatment of an earlier version of the hardware and can be used as background reading for this technical bulletin. (http://www.dell.com/downloads/global/solutions/200-DELL-CAMBRIDGE-SOLUTIONS-WHITEPAPER-20072010b.pdf)

Dell PowerVault MD Lustre System Overview

The PowerVault Dell MD Lustre based storage system is designed with price, performance, scalability and availability in mind. These goals are mainly achieved by:

- choosing standard enterprise-class hardware, with 2-socket Intel servers and SAS-attached storage arrays
- using open source software
- adopting a modular approach that provides scalability, allowing further growth of capacity and throughput while remaining cost-effective.

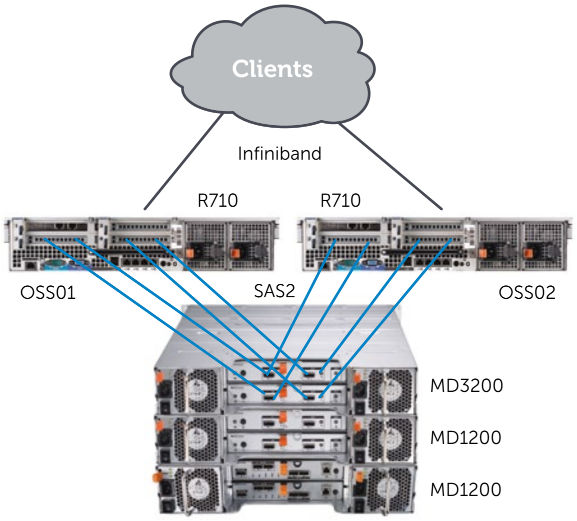


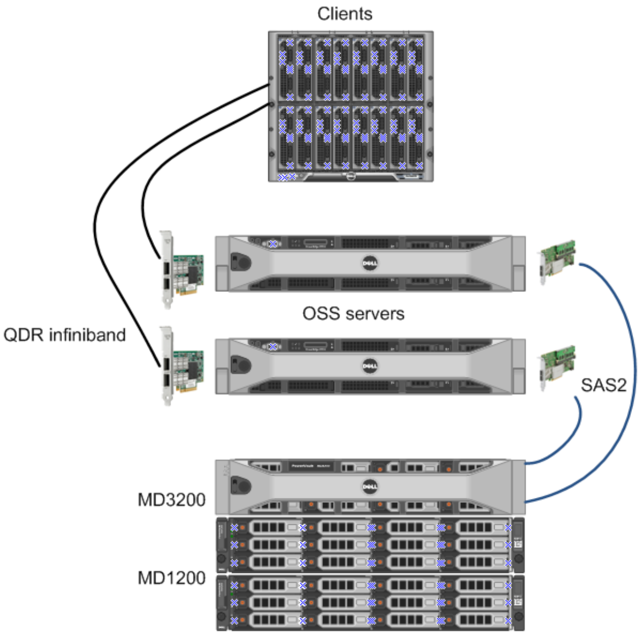
Hardware components

The DellTM PowerEdge™ and Dell™ PowerVault™ product lines are the foundation of the Dell Lustre Storage System. Specifically, PowerEdge R710 servers and PowerVault MD3200 and MD1200 disk arrays are used as the building blocks of the Dell Lustre Storage System modules.
The OSS Dell Lustre module comprises of:

- OSS servers - PowerEdge R710: Intel Xeon X5560 @ 2.8GHz, 24GB RAM, equipped with two 6Gbps SAS HBA cards, one Mellanox QDR ConnectX2 Infiniband card, running 64-bit CentOS 5.5 with Lustre patched Linux kernel.
- OST storage - MD3200: two 6Gbps SAS RAID controllers, each controller equipped with four host ports and 2GB battery-backed cache, 12 x 1TB nearline SAS 7.2K disk drives, and the High Performance Tier feature.

  Expansion storage - two MD1200 expansion disk enclosures equipped with 12 x 1TB nearline SAS 7.2K disks.

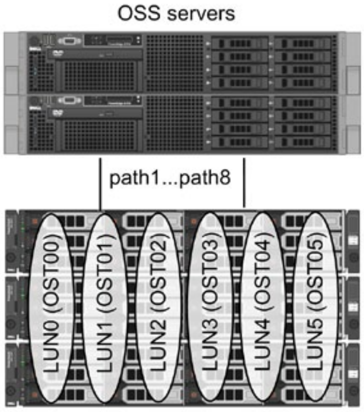




OSS module configuration

The hardware configuration is shown above for the OSS module. Note that each OSS (PowerEdge R710) maintains four physical SAS links to the dual MD3200 controllers. This is achieved using two dual-port SAS2 cards in the servers. While the MD storage in this case uses 1TB drives, 2TB drives are also available.

It is good practice to configure an even number of disk groups per MD3200 enclosure, equipped with two RAID controllers. This allows the I/O load to be spread symmetrically across controllers. The test storage system is equipped with 36 disk drives that were grouped into six RAID6 disk groups. Each disk group consists of four data disks and two parity disks. This layout allows the most efficient use of available disks while providing high data rates.



Virtual Disk 1 = LUN0 = Lustre OST00

Virtual Disk 2 = LUN1 = Lustre OST01

Virtual Disk 3 = LUN2 = Lustre OST02

Virtual Disk 4 = LUN3 = Lustre OST03

Virtual Disk 5 = LUN4 = Lustre OST04

Virtual Disk 6 = LUN5 = Lustre OST05

The above configuration provides six Lustre OST devices, each 4TB in size. Thus the overall usable disk capacity of the OST Lustre module is 24 TB, which provides 66.7% of storage capacity efficiency. To increase the storage capacity efficiency of the system, larger RAID6 disk groups would need to be configured. This can be done by increasing the number of disk drives by adding more expansion enclosures. For example a system with 60 disk drives (five disk enclosures) can be configured into six RAID6 (8+2) disk groups, while providing 80% storage capacity efficiency.

OSS module tuning

Prior to benchmarking, it is crucial to tune some parameters at various layers of the system.

Storage array tuning

- **Segment size -** Starting at the bottom, that is, at a RAID level, it is important to make sure that the RAID stripe size is aligned with the Lustre 1MB I/O request size. This is done by using the correct segment size while configuring a disk group.

disk group segment size= (Lustre I/O request size) / (number of data disks in a disk group)

In our case disk group segment size is set to 256KBytes.

- **Cache block size -** An important parameter that may have significant impact on performance is cache block size. For sequential I/O patterns the cache block size should be set to at least 16KB. We set it to its maximum value of 32KB.
- **Write Cache -** To enable write caching, Write Cache is recommended as the cache is protected by a battery backup system.
- **Cache mirroring -** For best performance, cache mirroring is disabled.
- Read Cache pre-fetch - For sequential I/O patterns Read Cache pre-fetch is recommended.

OSS tuning

- **Linux block device parameters:**

  Linux I/O scheduler - parameter set to "deadline"

  echo deadline > /sys/block/<blk_dev>/queue/scheduler

  max_sectors_kb - parameter set to 2048

  echo 2048 > /sys/block/<blk_dev>/queue/max_sectors_kb

  read_ahead_kb - parameter set to 2048

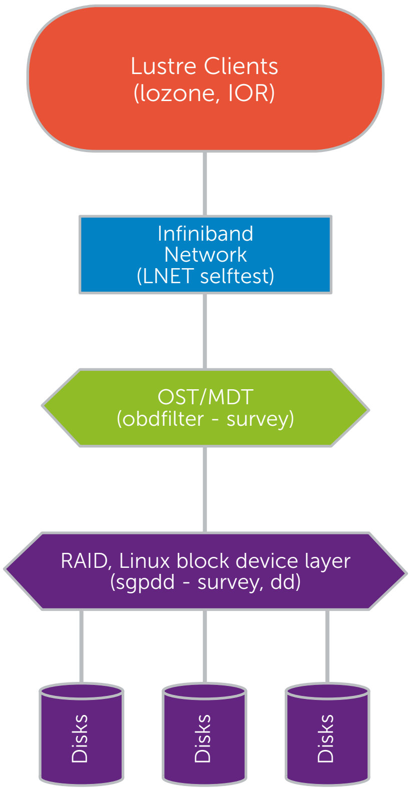
  echo 2048 > /sys/block/<blk_dev>/queue/read_ahead_kb

  nr_requests - parameter set to 512

  echo 512 > /sys/block/<blk_dev>/queue/nr_requests

Benchmarking Dell Lustre OSS module

Benchmarking stack



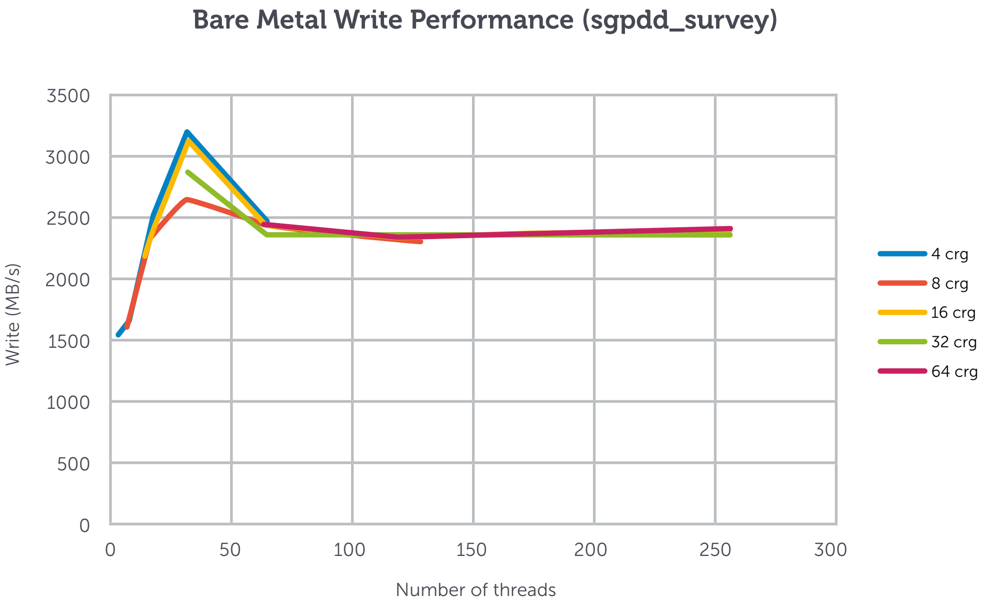
Testing individual disks with dd

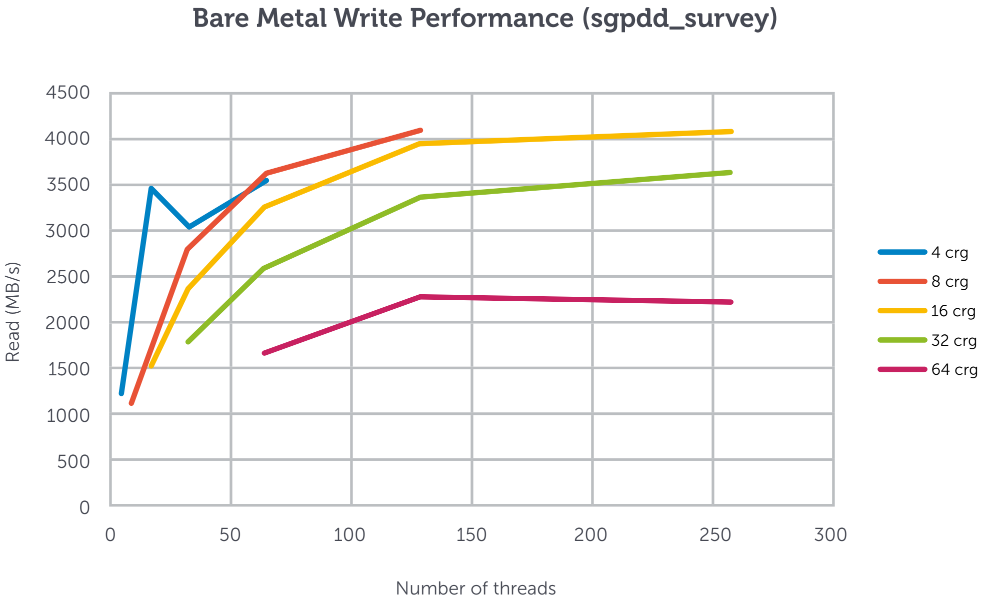
Each disk needs to be measured in order to obtain a baseline disk throughput and to identify any poorly performing disks. The following test is run at least three times per disk. A variation in the result of +/- 5% is satisfactory.

```
dd if=/dev/zero of=<disk name> bs=1M count=49152 oflag=direct
```

Testing Dell Lustre OSS module raw performance with sgpdd survey

sgpdd-survey is a shell script included in the Lustre IOKit toolkit. It uses the sgp_dd tool provided by the sg3 utilities package, which is a scalable version of the dd command and adds the ability to do multi-threaded I/O with various block sizes and to multiple disk regions. The script uses sgp_dd to carry out raw sequential disk I/O. It runs with variable numbers of sgp_dd threads to show how performance varies with different request queue depths. The script spawns variable numbers of sgp_dd instances, each reading or writing to a separate area of the disk to demonstrate performance variance within a number of concurrent stripe files. The purpose of running sgppdd_survey is to obtain a performance baseline and to set expectations for the Lustre filesystem. Lustre should deliver 85-90% of the raw device performance.





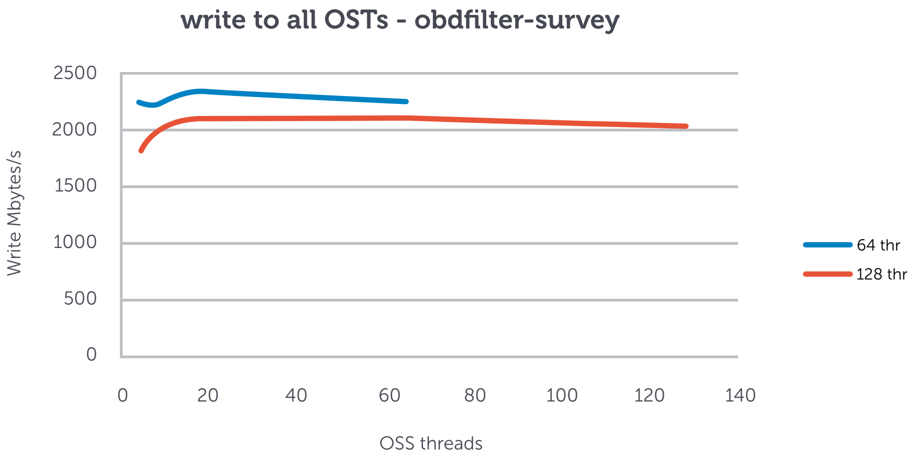
Understanding sgpdd-survey

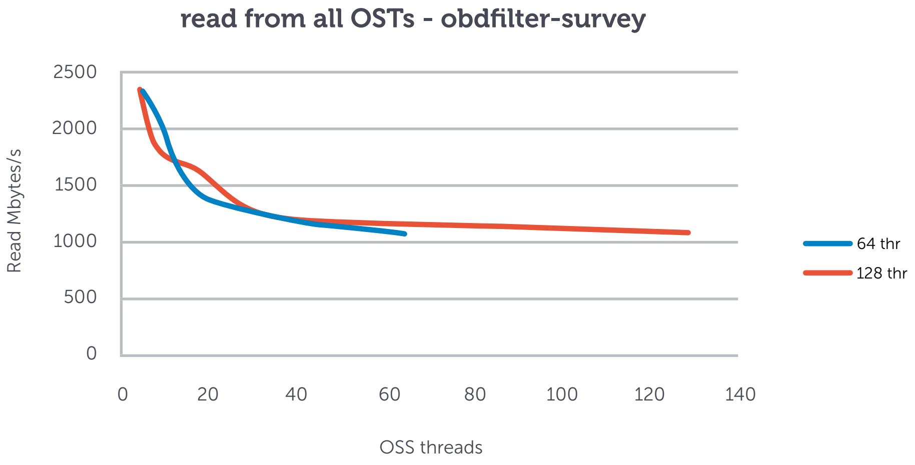
crg (concurrent regions) - describes how many regions on the disk are written to or read by sgp_dd. crg simulates multiple lustre clients accessing an OST. As the number of crg grows, performance will drop due to more disk seeks (especially for reads).

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, the hardware will not be able to process them and performance will drop.

Testing Dell Lustre OSS module with obdfilter-survey

Once the block devices are formatted with a Lustre filesystem, the next step is to benchmark the OSTs and their underlying ldiskfs filesystem. This can be done using a tool provided by Lustre IOKit called obdfilter-survey. This script profiles overall throughput of 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 find the saturation points which cause performance drops.





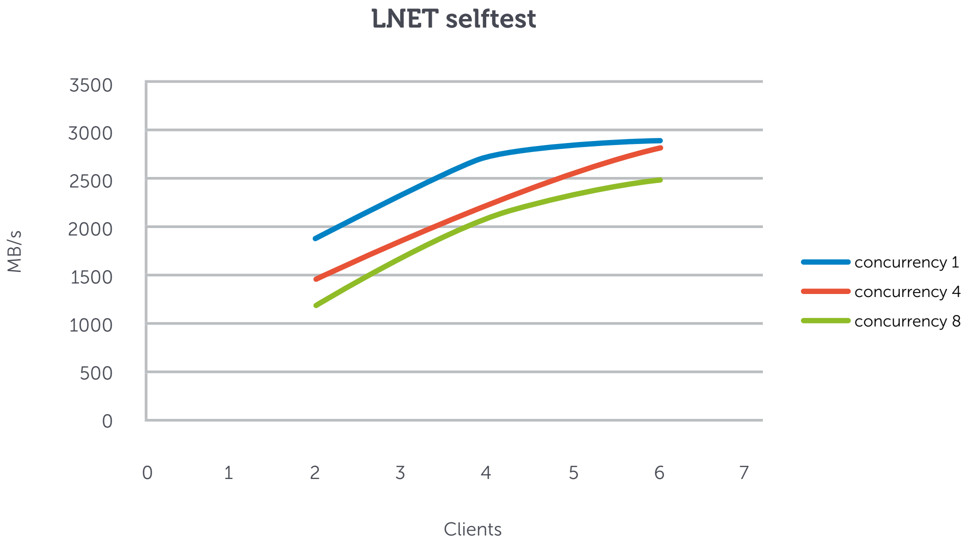
Understanding obdfilter-survey

obj (Lustre objects) - describes how many Lustre objects are written or read. obj simulates multiple lustre clients accessing the OST and reading/writing multiple objects. As the number of obj grows, performance will drop due to more disk seeks (especially for reads).

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, the hardware will not be able to process them and performance will drop.

Measuring Lustre Infiniband network performance with LNET selftest

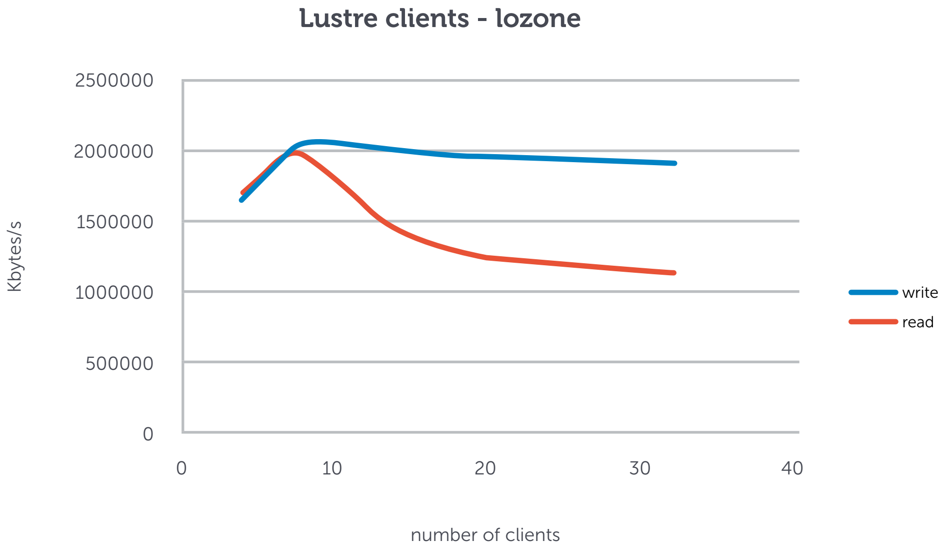
LNET selftest (lst) utility helps to test the operation of a Lustre network between servers and clients. It allows for verification that the Lustre network was properly configured and that performance meets expected values. LNET selftest runs as a kernel module and has to be loaded on all nodes that take part in the test. A utility called lst is used to configure and launch Lustre network tests. Below are shown results from tests run between six Lustre clients and two OSS servers with variable numbers of concurrency (RPCs in flight). More RPC in flight causes more overhead for the LNET protocol, which decreases performance for bulk sequential I/O. However if clients are performing operations on files smaller than RPC size (1MB) it is recommended that the parameter is tuned by increasing it slightly.



Concurrency - determines how many requests each client node in a test keeps on the wire.

Measuring Lustre client performance with Iozone

Iozone is a benchmark that measures read/write performance across multiple Lustre clients. In cluster mode it generates a number of sequential writes and then reads per client and it can run on many clients simultaneously, providing aggregate performance results. Below are the results of a run of Iozone on number of clients from four - 32 where each client was writing/reading a single file. The chart below shows that overall write performance is better than read performance. This behaviour is to be expected with a Lustre filesystem. When doing writes, clients send RPCs asynchronously and those RPCs are allocated and written to disks in the order they arrive, and the back-end storage can aggregate those writes efficiently. In the case of reads, the read requests from clients may come in a different order, which requires a lot of seeking and which throttles overall read throughput.



Conclusions

The Dell Lustre solution described in this technical bulletin, using commodity hardware and open source software, allows administrators of HPC clusters and Data Centres to build their storage systems in a very cost effective way. The modular design of the Dell Lustre OSS module enables administrators to scale the storage capacity and throughput in tune with growing demand. Tests performed on the Lustre OSS module demonstrate that this solution provides significant performance all the way from back-end disk system to the end client devices. The use of redundancy hardware and RAID with battery-backed cache provides a highly reliable model that can satisfy strong data safety requirements and performance demands. A single rack cabinet with four Dell MD3200 Lustre bricks would provide ~200TB of usable storage and 8GB/s read/write bandwidth.