Providing Australian researchers with world-class computing services

# QUESTnet 2016

*HPC Ready -*
*Building a Storage Platform for Research Datasets*

Daniel Rodwell
Manager, Data Storage Services

July 2016

- **What is NCI**
  - Who uses NCI
- **Petascale HPC at NCI**
  - Raijin High Performance Compute
  - Tenjin High Performance Cloud
- **Storage and Data at NCI**
  - Data Challenge
  - Data Storage
  - Lustre
- **Gdata3**
  - Requirements & Design
  - Validation

# What is NCI?

- NCI is Australia's national high-performance computing service
  - comprehensive, vertically-integrated research service
  - providing national access on priority and merit
  - driven by research objectives
- Operates as a formal collaboration of ANU, CSIRO, the Australian Bureau of Meteorology and Geoscience Australia
- As a partnership with a number of research-intensive Universities, supported by the Australian Research Council.

# Where are we located?

- Canberra, ACT
- The Australian National University (ANU)

**Research focus areas**

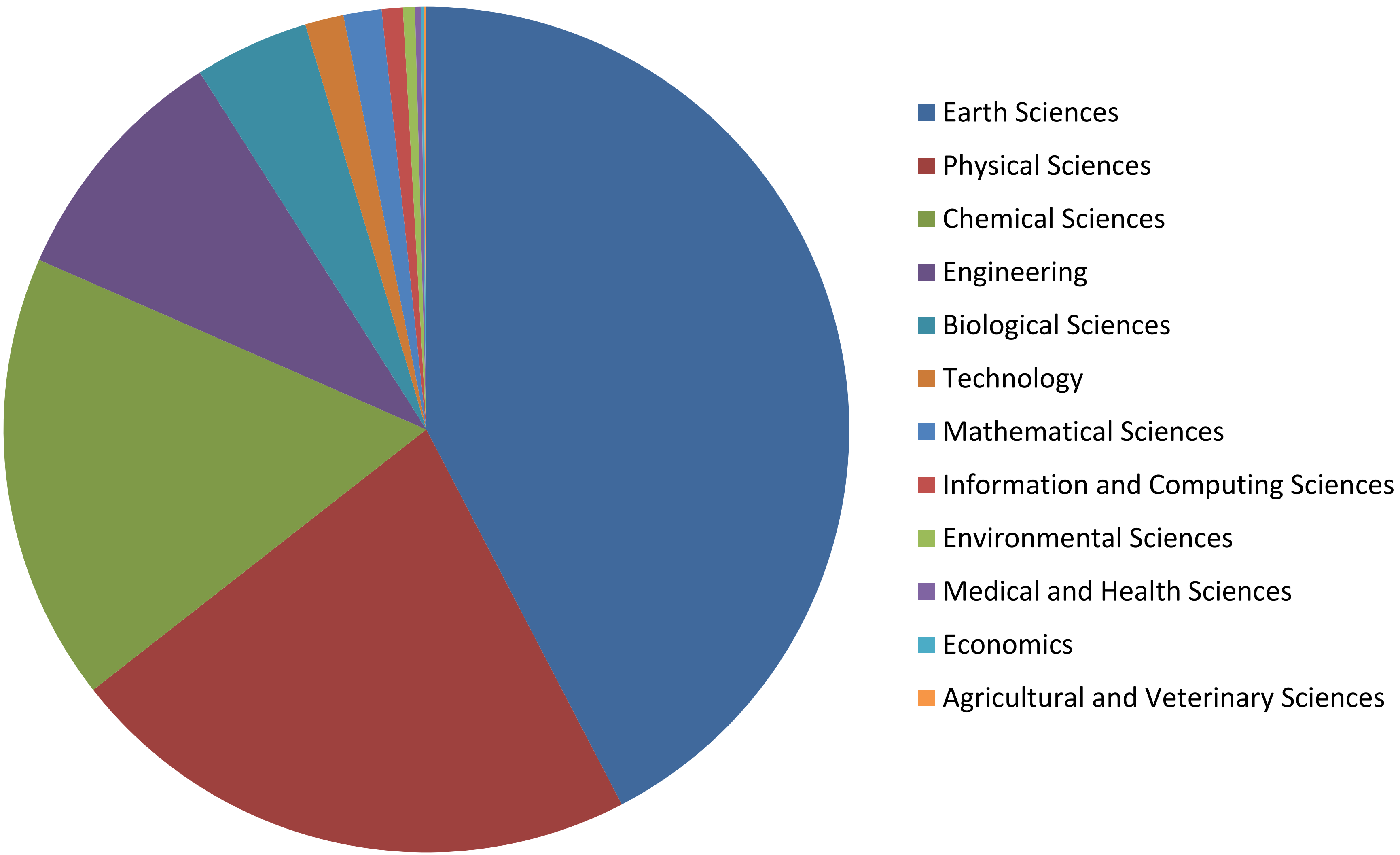
- Climate Science and Earth System Science
- Astronomy (optical and theoretical)
- Geosciences: Geophysics, Earth Observation
- Biosciences & Bioinformatics
- Computational Sciences
  - Engineering
  - Chemistry
  - Physics
- Social Sciences

- Growing emphasis on data-intensive computation
  - Cloud Services
  - Earth System Grid

- 3,000+ users
- 10 new users every week
- 600+ projects

Astrophysics, Biology, Climate & Weather, Oceanography, particle Physics, fluid dynamics, materials science, Chemistry, Photonics, Mathematics, image processing, Geophysics, Engineering, remote sensing, Bioinformatics, Environmental Science, Geospatial, Hydrology, data mining

# What do they use it for ?



## The greatest map ever made

Led by Nobel Laureate, Professor Brian Schmidt, Australian astronomers are using NCI to carry our the most detailed optical survey yet of the southern sky. The project involves processing and storing of many terabytes of optical telescopic images, and has led to the discovery of the oldest star in the universe.

## Unlocking the Landsat Archive

NCI is enabling researchers at Geoscience Australia to 'unlock' decades of Landsat earth observation satellite images of Australia since 1979.  A one petabyte *data cube* has been generated by processing and analysing hundreds of thousands of images, yielding important insights for water/land management decision making and policy, with benefits for the environment and agriculture.

## Predicting the unpredictable

Australia's weather and future climate are predicted using the ACCESS model—developed by BoM, CSIRO, and ARCCSS—and operating on time spans ranging from hours/days, to centuries. Collaborating with NCI and Fujitsu, BoM, using NCI as its research system, is increasing the scalability of ACCESS to many 1000s of cores, to prepare for its next-gen system, and more accurate predictions of extreme weather.

‘Raijin’ – 1.2 PetaFLOP Fujitsu Primergy Cluster

# Petascale HPC at NCI

**Raijin** Fujitsu Primergy cluster, June 2013**:**

- 57,472 cores (Intel Xeon Sandy Bridge, 2.6 GHz) in 3592 compute nodes;
- 157TBytes of main memory;
- Infiniband FDR interconnect; and
- 7.6 Pbytes of usable fast filesystem (for short-term scratch space)

– Accelerator Nodes
  - 14x Dell C4310 GPU nodes, with 56 Nvidia K80 GPUs
  - 32x SGI Nodes with Intel Xeon Phi ‘Knights Landing’ processors

– 24$^{th}$ fastest in the world on debut (November 2012); first petaflop system in Australia
  - 1195 Tflops, 1,400,000 SPECFPrate
  - Custom monitoring and deployment
  - Custom Kernel, CentOS 6.7 Linux
  - Highly customised PBS Pro 13 scheduler.
  - FDR interconnects by Mellanox
    - ~52 KM of IB cabling.
  - 1.5 MW power; 100 tonnes of water in cooling


## **Tenjin** Dell C8000 High Performance Cloud

- 1,600 cores (Intel Xeon Sandy Bridge, 2.6 GHz), 100 nodes;
- 12+ TBytes of main memory; 128GB per node
- 800GB local SSD per node

- 56 Gbit Infiniband/Ethernet FDR interconnect
- 650TB CEPH filesystem

- Architected for strong computational and I/O performance needed for “big data” research.

- On-demand access to GPU nodes.

- Access to over 21PB Lustre storage.

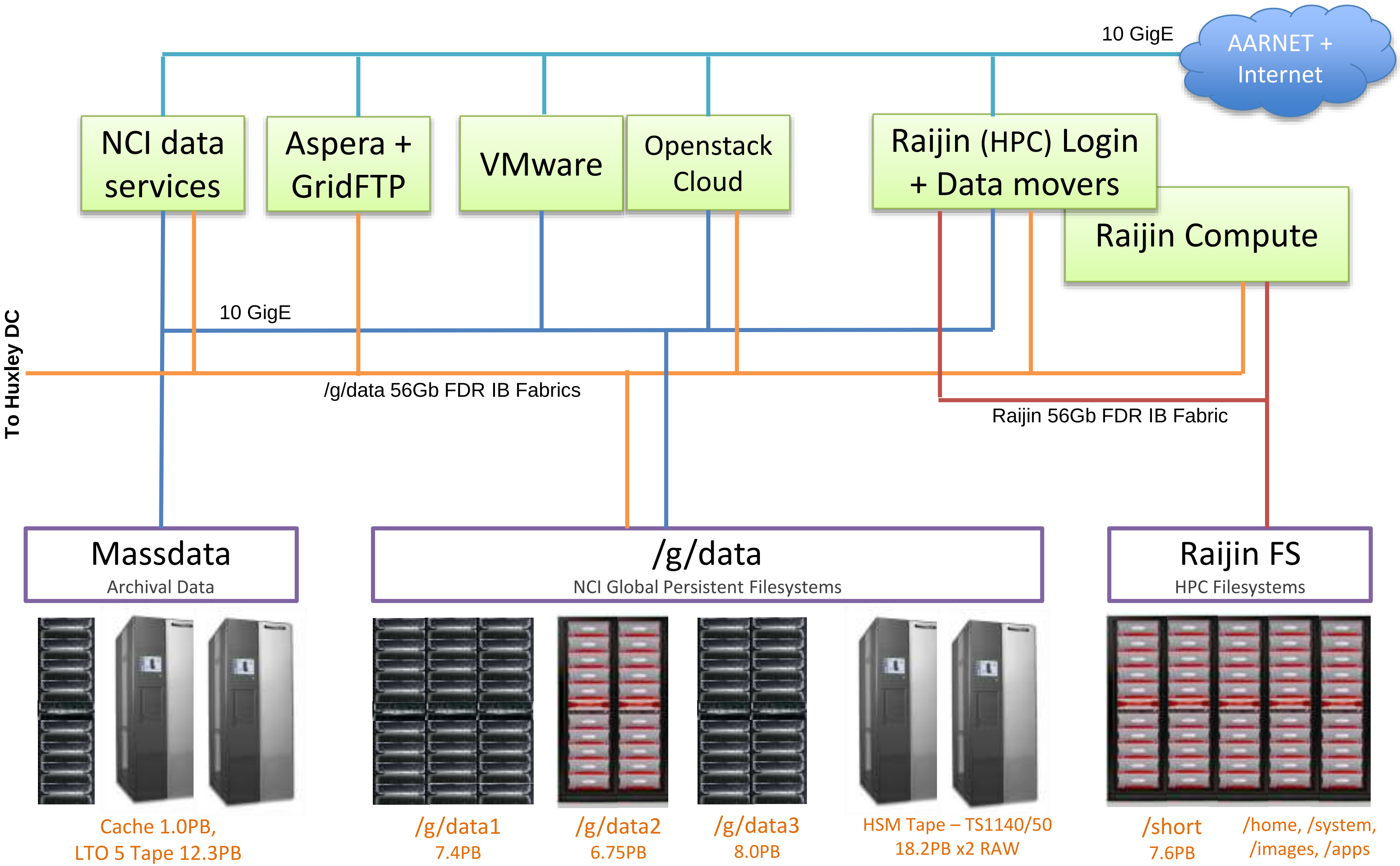
30PB High Performance Storage

# Storage at NCI

- Lustre Systems
  - Raijin Lustre – HPC Filesystems: includes /short, /home, /apps, /images, /system
    - 7.6PB @ 150GB/Sec on /short (IOR Aggregate Sequential Write)
    - Lustre 2.5.23 + Custom patches (NCI + DDN)
  - Gdata1 – Persistent Data: /g/data1
    - 7.4PB @ 21GB/Sec (IOR Aggregate Sequential Write)
    - Lustre 2.3.11 (IEEL v1)
  - Gdata2 – Persistent Data: /g/data2
    - 6.75PB @ 65GB/Sec (IOR Aggregate Sequential Write)
    - Lustre 2.5.42.8 (IEEL v2)
  - Gdata3 – Persistent Data: /g/data3 –
    - Stage 1: 5.7PB @ 92GB/sec
    - Stage 2: 8.0PB @ 120GB/Sec+
    - (Lustre 2.5.42.8, IEEL v2)

- Other Systems
  - Massdata – Archive Data: Migrating CXFS/DMF, 1PB Cache, 6PB x2 LTO 5 dual site tape
  - OpenStack – Persistent Data: CEPH, 1.1PB over 2 systems
    - Nectar Cloud, v0.72.2 (Emperor), 436TB
    - NCI Private Cloud, 0.80.5 (Firefly), 683TB
  - HA Data – Persistent High Availability Data, Netapp Clustered DataONTAP v8.3
    - High Security, Isolated tenancy, SAN (Block) and NAS (NFS/CIFS)
    - 200 TB per site at 2 sites (NCIDC, LHDC)
    - Single site or full dual site replication
    - Flash cache tiers – High Performance database and small I/O NFS.
    - Available on VMware (default) and Tenjin (use case dependent)
    - Global Home (/g/home) across both Cloud and HPC Systems – Q3 2016.

# Systems Overview



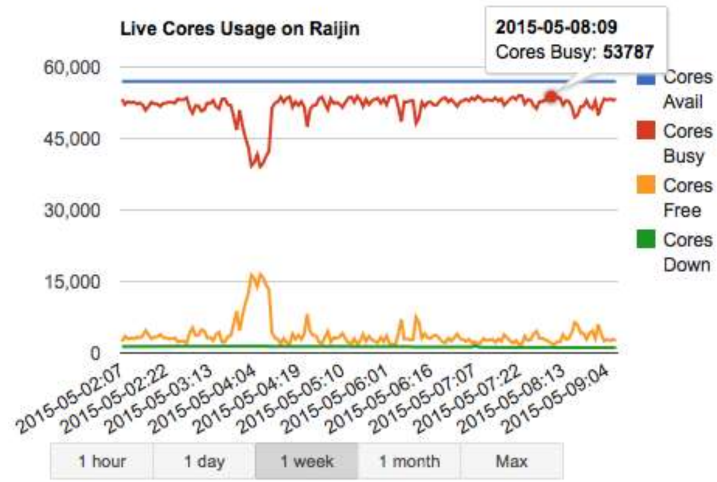
# What do we store?

- How big?
  - Very.
  - Average data collection is 50-100+ Terabytes
  - Larger data collections are multi-Petabytes in size
  - Individual files can exceed 2TB or be as small as a few KB.
  - Individual datasets consist of tens of millions of files
  - Next Generation likely to be 6-10x larger.
    - Gdata1+2 = 380 Million inodes stored
    - 1% of /g/data1 capacity = 74TB
- What ?
  - High value, cross-institutional collaborative scientific research collections.
  - Nationally significant data collections such as:
    - Australian Community Climate and Earth System Simulator (ACCESS) Models
    - Australian & international data from the CMIP5 and AR5 collection
    - Satellite imagery (Landsat, INSAR, ALOS)
    - Skymapper, Whole Sky Survey/ Pulsars
    - Australian Plant Phenomics Database
    - Australian Data Archive

| Collection | TB Approved | TB Ready | Ingested |
| --- | --- | --- | --- |
| Skymapper (Astronomy) | 227.00 | 140.00 | 62% |
| Australian Data Archive (Social Sciences) | 4.00 | 3.00 | 75% |
| BPA Melanoma Dataset (Biosciences) | 129.00 | 123.00 | 95% |
| Plant Phenomics (Biosciences) | 110.00 | 2.00 | 2% |
| Ocean Gen. Circulation Model (Earth Simulator) | 29.00 | 27.00 | 93% |
| Year Of Tropical Convection | 41.00 | 41.00 | 100% |
| CABLE Global Evaluation Datasets | 24.00 | 2.00 | 8% |
| CORDEX Int | 57.00 | 1.00 | 2% |
| Coupled Model Intercomparison Project (CMIP5) 2.6PB → | 2,600.00 | 1,488.00 | 57% |
| Reanalysis | 146.00 | 146.00 | 100% |
| ACCESS Models 2.6PB → | 2,538.00 | 2,098.00 | 83% |
| Seasonal Climate Prediction | 595.00 | 369.00 | 62% |
| Australian Bathymetry and Elevation reference data | 113.00 | 23.00 | 20% |
| Australian Marine Video and Imagery Collection | 7.00 | 7.00 | 100% |
| Global Navigation Satellite System (GNSS) (Geodesy) | 5.00 | 4.00 | 80% |
| Digitised Australian Aerial Survey Photography | 77.00 | 74.00 | 96% |
| Earth Observation (Satellite: Landsat, etc) 1.5PB → | 1,486.00 | 1,413.00 | 95% |
| IMOS+TERN Australasian Satellite Imagery (NOAA/AVHRR, MODIS, VIIRS and AusCover) | 436.00 | 257.00 | 59% |
| Satellite Soil Moisture Products | 5.00 | 1.00 | 20% |
| Synthetic Aperture Radar | 29.00 | 29.00 | 100% |
| BoM Observations | 366.00 | 175.00 | 48% |
| BoM Ocean-Marine Collections | 429.00 | 77.00 | 18% |
| Aust. 3D Geological Models | 3.00 | 1.00 | 33% |
| Aust. Geophysical Data Collection | 330.00 | 7.00 | 2% |
| Aust. Natural Hazards Archive | 27.00 | 3.00 | 11% |
| National CT-Lab Tomographic Collection | 205.00 | 171.00 | 83% |
| TERN eMAST | 90.00 | 15.00 | 17% |
| TERN Phenology Monitoring: Near Surface Remote Sensing | 12.00 | 1.00 | 8% |
| TERN eMAST Data Assimilation | 110.00 | 9.00 | 8% |
| CSIRO/BoM Key Water Assets | 44.00 | 18.00 | 41% |
| Models of Land/Water Dynamics from Space | 22.00 | 11.00 | 50% |
| **Totals** | **10,296** | **6,737** | **65%** |

https://www.rdsi.edu.au/collections-stored

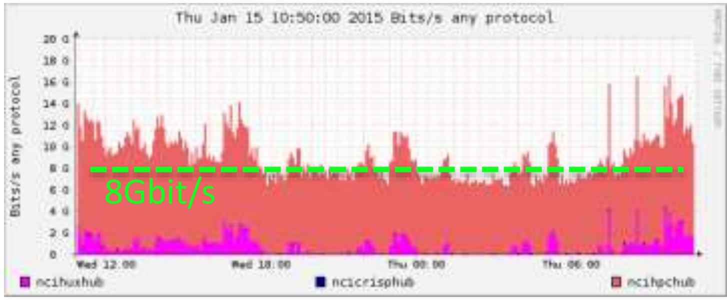
- Raijin - HPC
  - Native Lustre mounts for gdata storage on all 3592 compute nodes (57,472 Xeon cores), 56Gbit per node (each node capable of 5GB/s to fabric)
  - Additional Login nodes + Management nodes also 56GBit FDR IB
  - Scheduler will run jobs as resources become available (semi-predictable, but runs 24/7)
  - A single job may be 10,000+ cores reading (or creating) a dataset.




*53,787 of 56,992 cores in use (94.37% utilisation)*

- Cloud
  - NFS 10 Gbit Ethernet (40GE NFS capable on demand)
  - Unpredictable when load will ramp
  - Typically many small I/O patterns

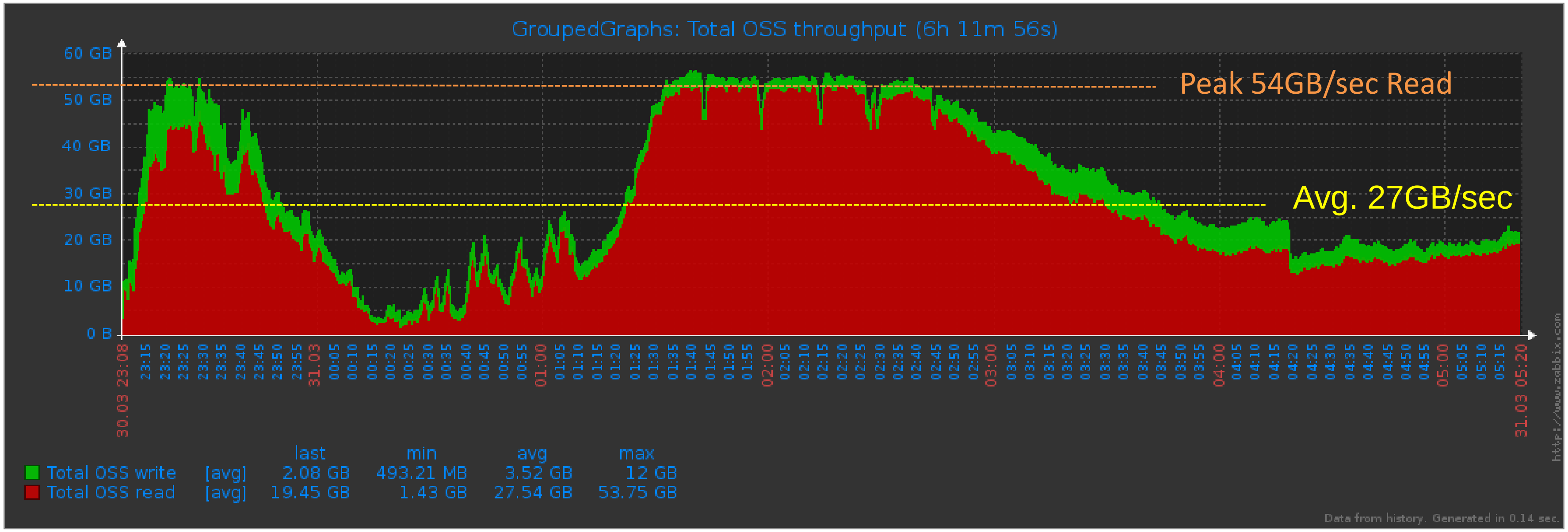
- Datamover Nodes
  - Dedicated datamover nodes connected via 10GE externally and 56Gbit Infiniband internally.
  - Dedicated datamover systems like Aspera, GridFTP, Long Distance IB connected via 10GE, 40Gb IB, optical circuits
  - Data access may be sustained for days or weeks, continual streaming read/write access.




*8Gbit/sec for 24hrs+, inbound transfers*

**Performance (gdata1, HPC User Application)**

Peak 54GB/sec read sustained for 1.5 hrs. Average 27GB/sec sustained for 6 hours

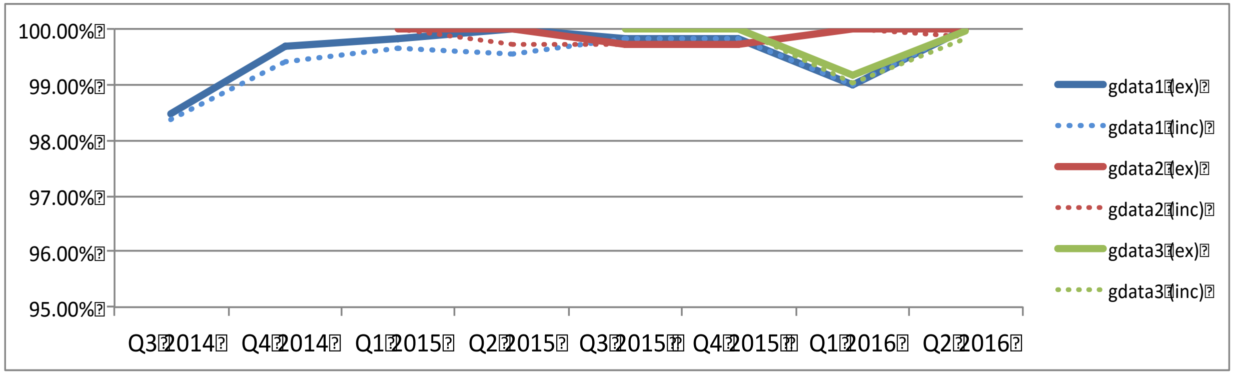



**Availability (Quarterly, 2014-2016)**

Gdata1, 2, 3 filesystems

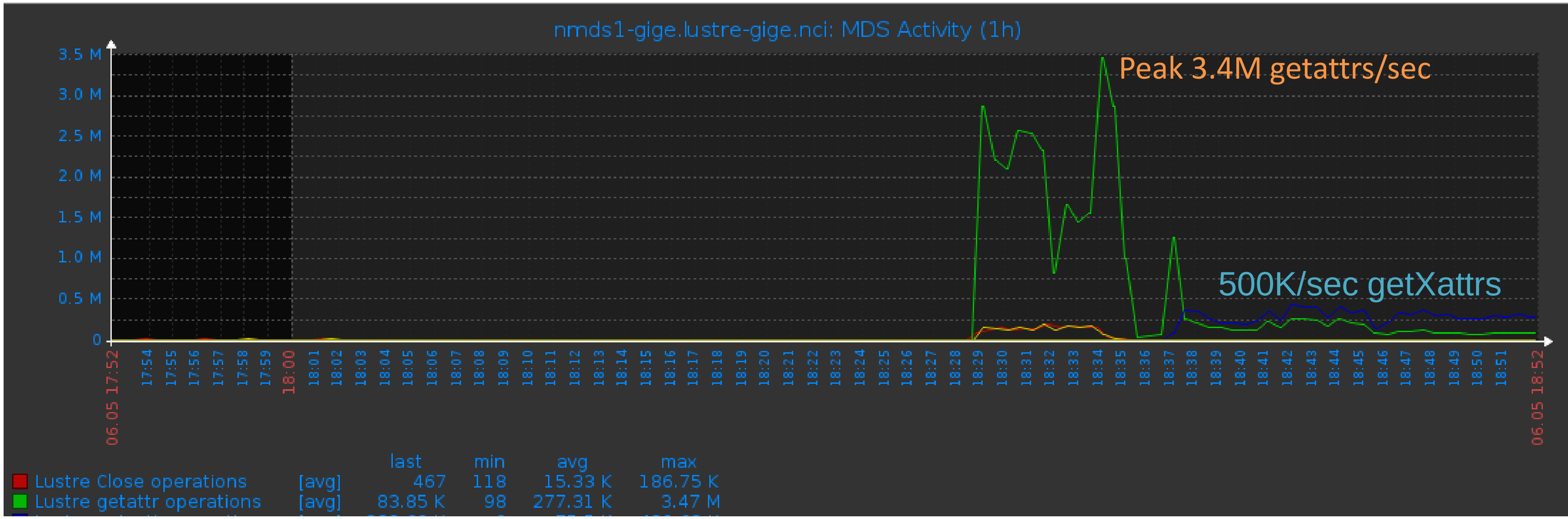
**GdataN** long term availability of **99.7%** over 8 QTRs

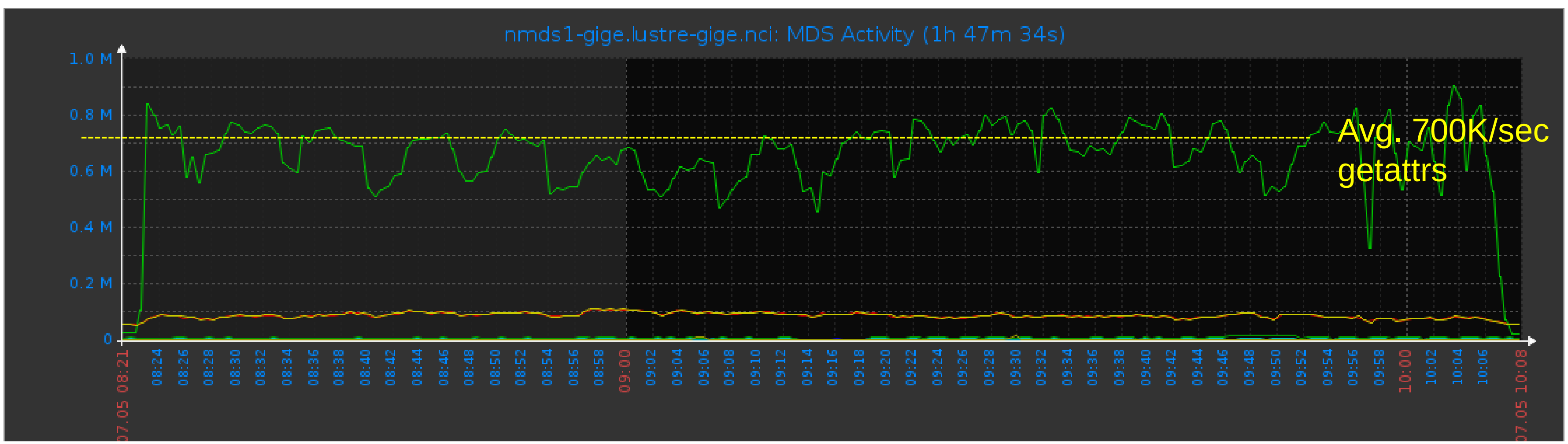
- Ex values – exclusive of published scheduled maintenance events with 24+ hrs notice
- Inc values – including scheduled maintenance events & quarterly maintenance.



## Metadata Performance (gdata1), example applications

Peak 3.5 Million getattrs /sec, . Average 700,000+ getattrs sustained for 1.5 hours



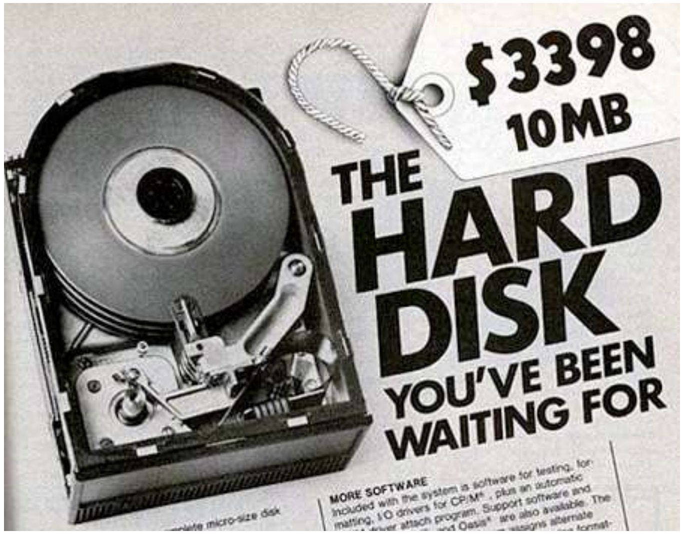



High Performance Persistent Data Store

# Gdata3 – Requirements & Design

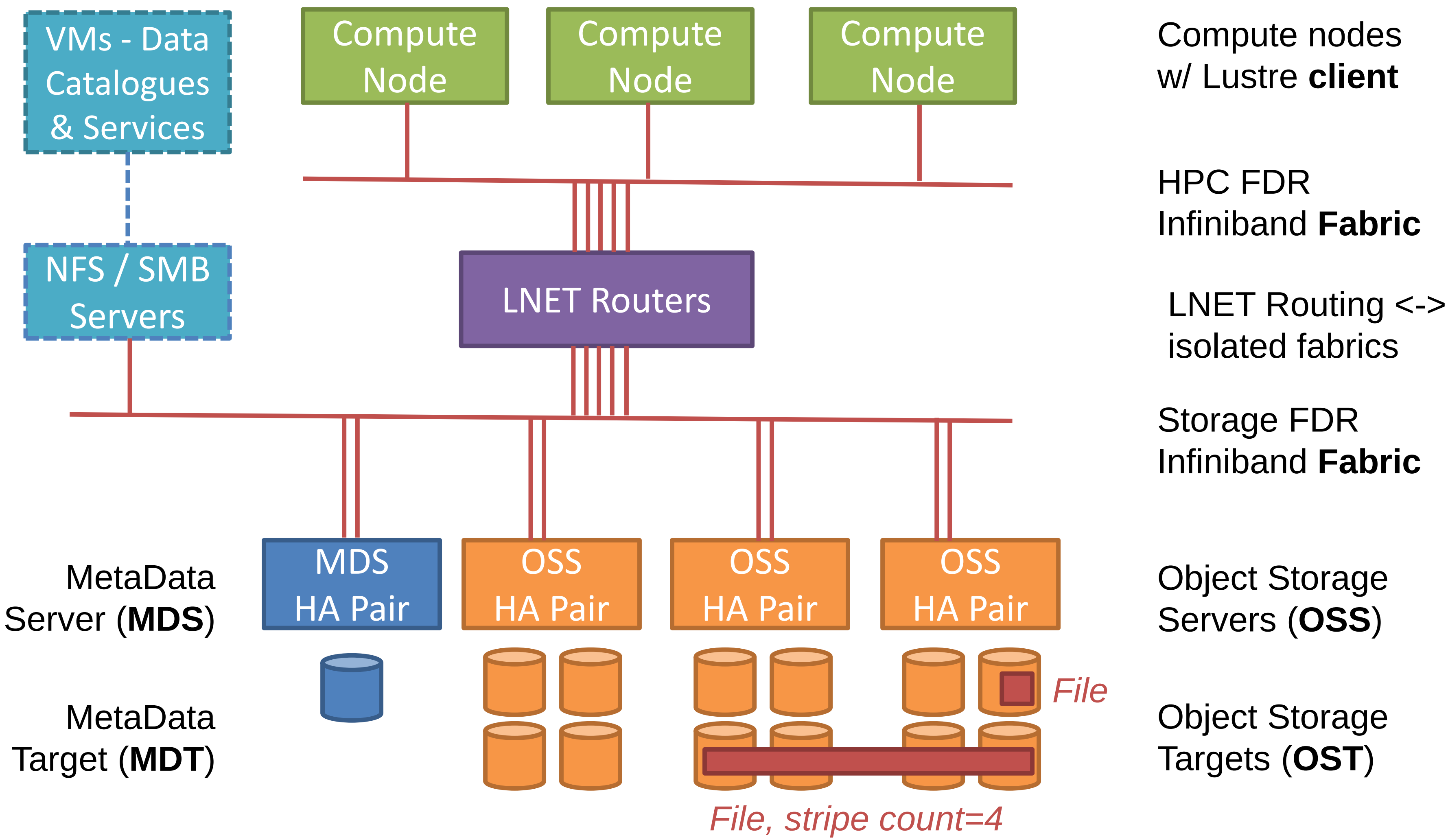
# Requirements

- Data Storage Requirements
  - **8 PB** by Mid 2015, ability to grow to **10PB+.** Additional capacity required for expansion of existing and new data collections.
  - **High Performance, High Capacity** Storage capable of supporting HPC connected workload. **High Availability.**
  - **Persistent Storage** for Active Projects and Reference Datasets, **with 'backup' or HSM capability**.
  - Capable of supporting intense metadata workload of **4 Million+ operations per sec.**
  - **Modular design** that can be scaled out as required for future growth.
  - **120+ GB/sec read** performance**, 80+ GB/sec write** performance. Online, low latency. Mixed workload of stream and IOPS.
  - Available **across all NCI systems** (Cloud, VMWare, HPC) using native mounts and 10/40Gbit NFS.

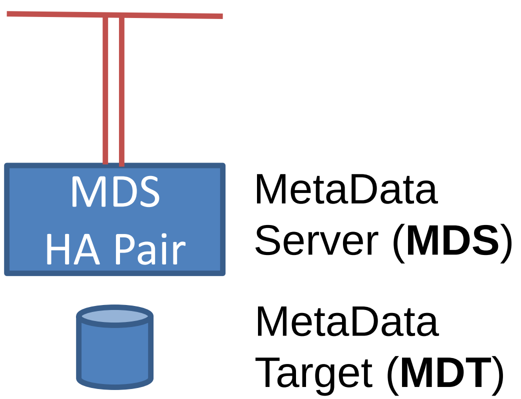


- What is Lustre?
  - Lustre is a high performance parallel distributed filesystem, typically used for large scale compute clusters.
  - Highly scalable for very large and fast filesystems.
  - Is the most widely used filesystem in the top 100 fastest supercomputers world-wide, including Titan (#3), Sequoia (#4, LLNL, 55PB, 1TB/sec).
  - Lustre is used at NCI for Raijin’s HPC filesystems (/short, /apps, /home) and persistent data stores - /g/data1, /g/data2, /g/data3.
  - Can be used with common Enterprise-type server and storage hardware – but will have poor performance and reliability if not correctly specified.

# How Lustre works (NCI context)



- Metadata Design
  - MDT capacity and performance is typically determined for whole filesystem at initial build
  - Need to consider overall capacity of filesystem in initial specification.
  - Must consider MDT Controller + Disk IOPS, MDS Cores + RAM
  - Primarily a Random 4K IO workload
  - Need performance, lots of it.
  - Filesystem performance is heavily dependent on MDS and MDT. Poor metadata performance impacts entire filesystem.
  - Ideally we want to minimise MDT I/O, and have cache hits where possible – very large MDS RAM + params tuning. In Lustre 2.7+, use of Distributed Namespace Entry (Multiple MDT-MDS pairs) is highly recommended
  - Slow filesystem = slow jobs = wasted HPC compute hours.

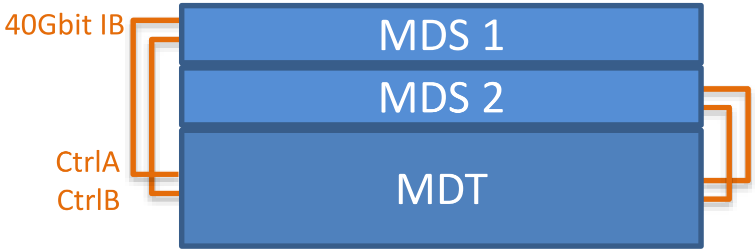


# Netapp EF-550

- MetaData Target – EF550
  - 450,000 IOPS sustained. 900,00 peak.
  - 24x 800GB SAS SSDs (mixed use SLC)
  - Dual Controllers, each with:
    - 12GB Cache
    - 2x 40Gbit Infiniband ports
    - quad-core Intel Xeon E5-2418L (Sandy Bridge)
  - 21KG, 2RU
  - Low power & Thermal loads
  - August 2014 Eval Testing:
    - Fujitsu RX300 S7, each with
    - Dual 2.6GHz E5-2670 8C Xeon *(Sandy Bridge)*
    - 128GB RDIMM DDR3
    - 3x Dual Port Intel X520 10GE NICs for test below
    - Benchmarked up to 320,000 4K IOPS sustained for 2hrs+ with single host, using 6 of 8 available 10GE ports
    - RX300 became CPU limited before maxing out EF550.

EF550 – All Flash Array

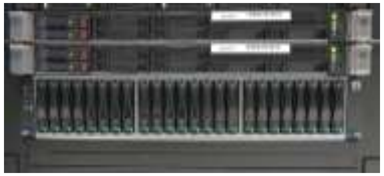
## Gdata 3 Metadata Building Blocks




- MDT storage for Gdata3 is built using a dedicated Netapp EF550 All-Flash block storage array, with 4x MDS-MDT 40Gbit Infiniband interconnects

- Array (MDT)
  - 24 x 800G SAS (SLC mixed use)
  - Dual 40Gbit IB Controllers
  - 2x 10 Disk RAID 10 pools, LVM mirror together, 4 hot spares
  - 1 preferred pool per controller.
  - 1.5 Billion inode capacity (as formatted for MDT)

- Hosts (MDS)
  - 2x Servers as High Availability pair
  - 1RU HP DL 360 Gen 9s, each with
    - 2x Intel Xeon E5-2697v3 'Haswell'
    - 14 Core, 28 Hyperthread, 2.6Ghz Base, 3.6Ghz Turbo Boost max
    - 768GB DDR4 LR-DIMM
    - Single Port FDR connection to Fabric
    - Dual Port FDR connection to EF550



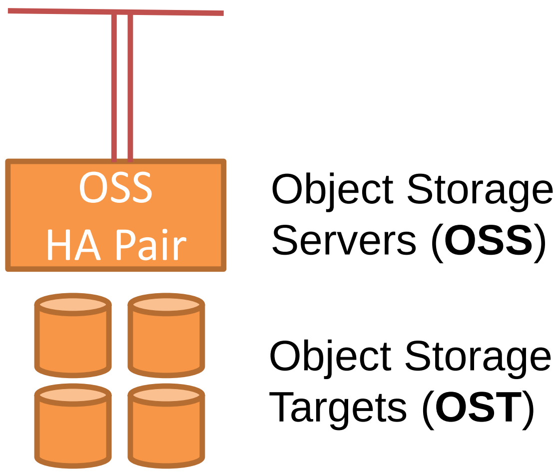
**Gdata3 MDT Array**

- 24x 800GB SAS SSDs
- 2 RU Array
- 2 RU Servers
- 450,000 IOPS Sustain (controller)
- Estimated 240,000 disk IOPS (24 x 10,000)

**Gdata1 + Gdata2 Shared MDT Array**

- 192x 600GB 15K SAS Hard Drives
- 32 RU Array
- 4 RU Servers
- ~50,000 IOPS Sustain (controller)
- Est 38,000 disk IOPS (192 x 200)

- Object Storage Design
  - OST performance is typically determined at initial build by choice of disk array technology (choose carefully if adding incrementally over multiple years).
  - Performance of all OSTs (and OSSes) in the filesystem should be very similar.
  - Mixed OSTs sizes and/or performance will result in hotspotting and inconsistent read/write performance as files are striped across OSTs or allocated in a round-robin / stride.
  - Capacity scales out as you add more building blocks, as does performance*
  - Design building block for your workload – controller to disk to IOPS ratios need to be considered.
  - Mixed 1MB Stream and Random 4K IO workload. Lustre uses 1MB transfers (optimise RAID config for 1MB stripe size).




- More small OSTs preferable to few very large OSTs.
- Loss of a single nnnTB OST = a lot of data gone
- A very large OST (nnnTB) will take a long time to e2fsck.
- Many smaller OSTs can be e2fsck'd in parallel
- Each OST mapping on client requires some memory – fewer are better
- Smaller OSTs can fill quickly with few large files if striping not set by user or default.

*interconnect fabric must scale to accommodate bandwidth of additional OSSes

- Object Storage Target – E5660
  - Latest generation E-Series
  - NCI - 1st Lustre deployment on E5600 series
  - Multi-core optimised Controllers

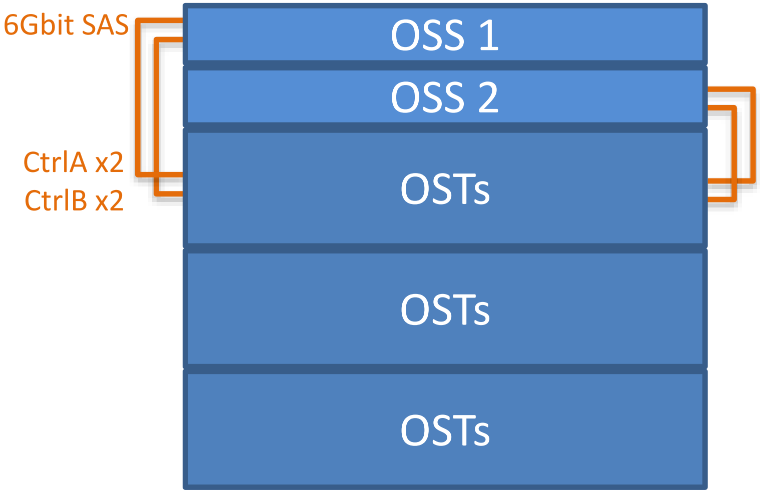
  - 12,000 MB/sec Read Performance (RAW)
  - 180x 4TB NL-SAS 7.2K HDDs (NCI Config)
  - Dual Controllers, each with:
    - 12GB Cache
    - 8x 12Gbit SAS ports
  - 1x E5660 60 Disk Controller shelf
  - 2x DE6600 60 Disk Expansion shelf

NetApp

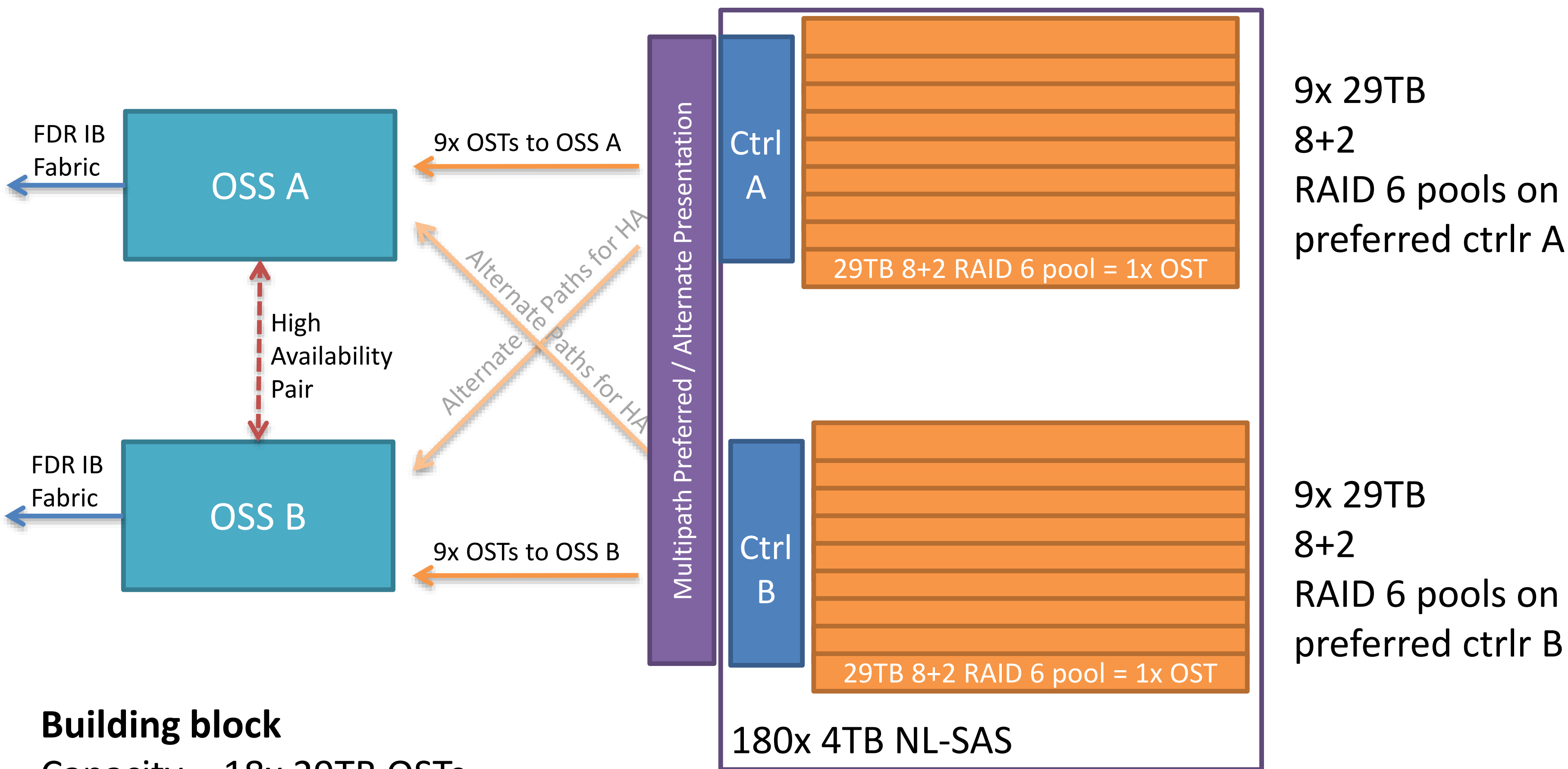
E5660 – 5600 Series

## Gdata 3 Object Storage Building Blocks

- OST storage for Gdata3 is built using Netapp E5660, with 8x OSSS-OST 12Gbit SAS interconnects
- Array (OST)
  - 180 x 4TB NL-SAS, 7.2K
  - Dual 12G SAS Controllers
  - 18x 8+2 RAID 6 Pools
  - 9 Pools per controller
- Hosts (OSS)
  - 2x Servers as High Availability pair
  - 1RU Fujitsu RX2530-M1's each with
    - 2x Intel Xeon E5-2640v3 'Haswell'
    - 8 Core, 16 Hyperthread, 2.6Ghz Base, 3.4Ghz Turbo Boost max
    - 256GB DDR4 RDIMM
    - Single Port FDR connection to Fabric
    - Quad Port 6G SAS connection to E5660



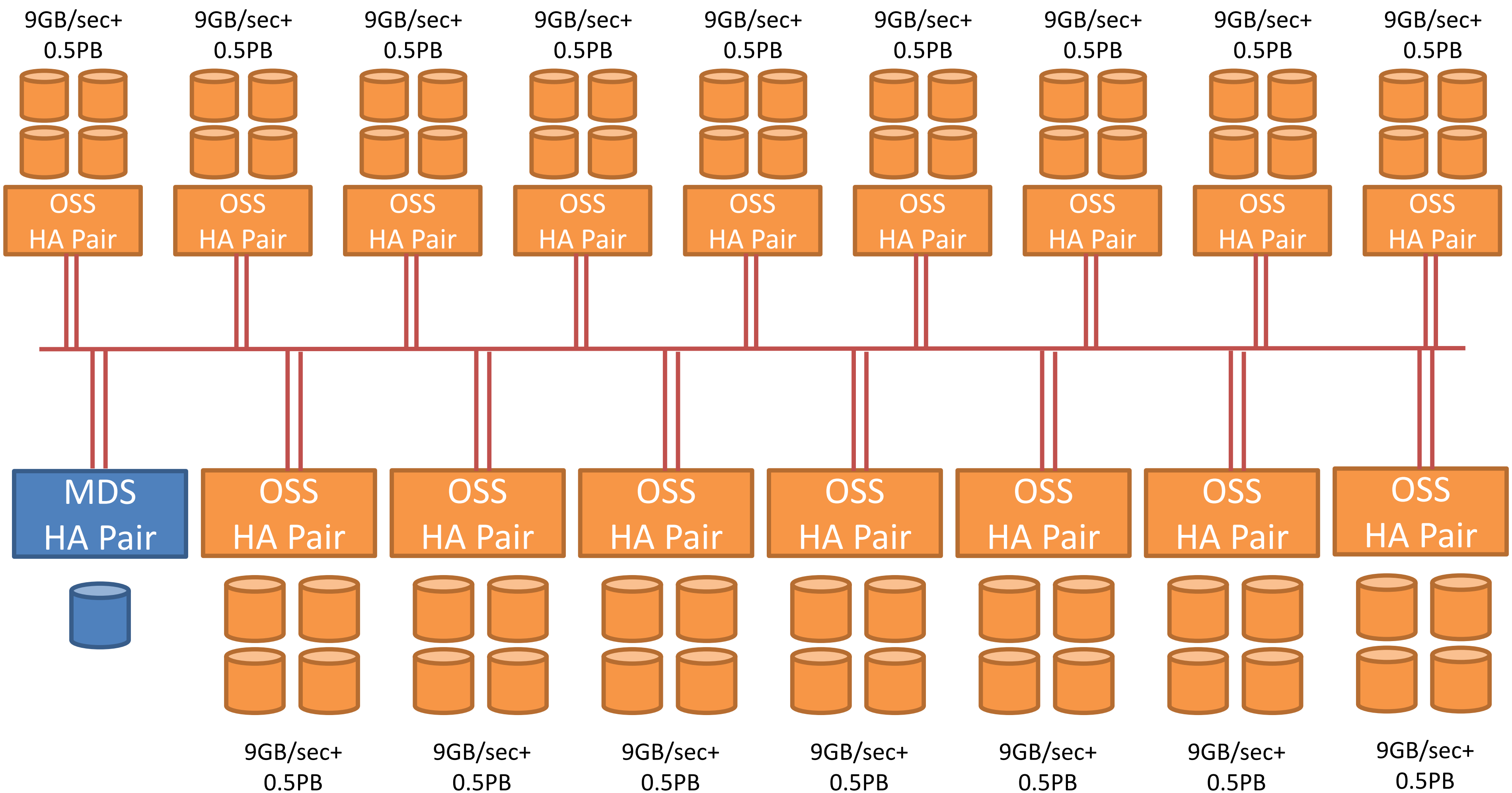

## Gdata 3 RAID 6 configuration




**Building block**
Capacity = 18x 29TB OSTs
= 520TB

**16x Building blocks**

8PB, 144GB/sec+



Gdata 3 Object Storage Building Blocks

**1x Building Block**

- 2x Fujitsu RX2530-M1
- 1x E5660 60 Disk controller shelf
- 2x DE6600 60 Disk expansion shelf

## Gdata 3 Object Storage Building Blocks

**Front View – bezel removed**

- 5x 12 Disk Drawers

**Front View – Tray1, Drawer 5 open**

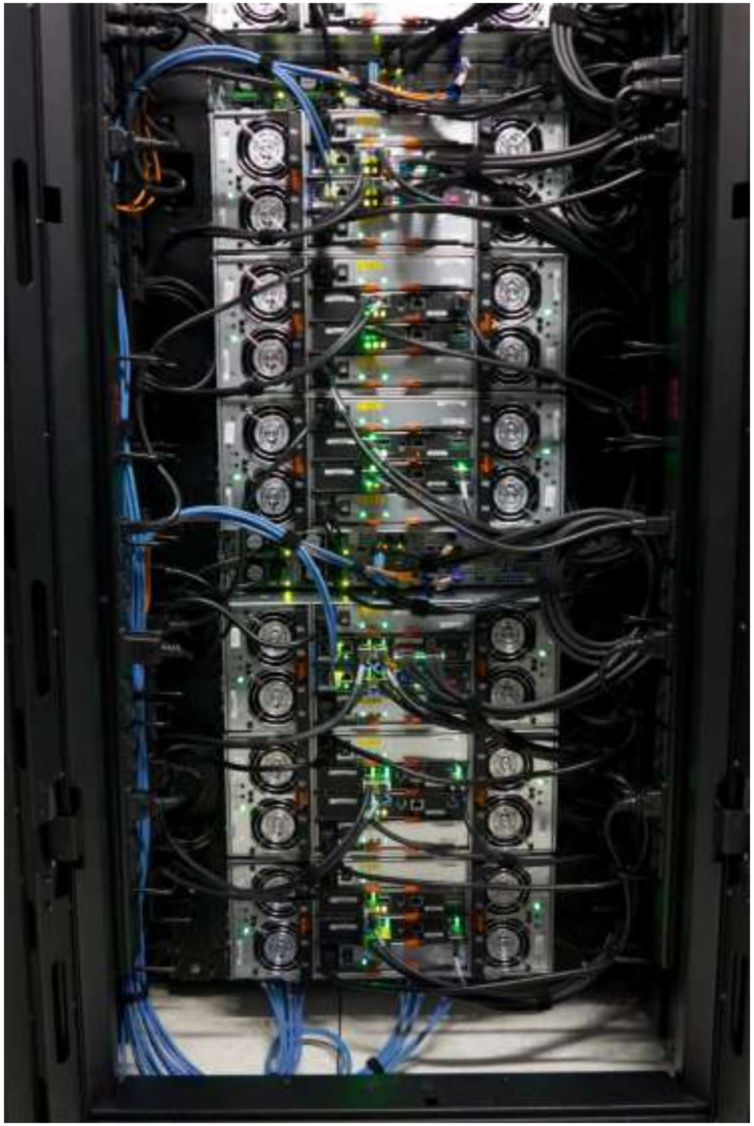
- 12x 4TB NL SAS

## Gdata 3 Object Storage Building Blocks

**Front of Rack**

- 3x Building blocks
- 42 RU Hosts and storage
- 42 RU APC Rack

Gdata 3 Object Storage Building Blocks



**Rear of Rack (as shown)**

- 2x Building blocks
- 1RU in-house custom built UTP Patch panel attachment at RU0 position

High Performance Persistent Data Store

# Gdata3 – Validation & Benchmarking

- Validate all aspects of system prior to production go-live
  - Disk Subsystem through to compute node client
  - Individual drive performance and latency
  - Pool Performance
  - Controller Performance
  - OST & OSS Performance
  - MDT & MDS Performance
  - LNET Routers
  - Interconnect (Infiniband)
  - Whole of Filesystem – Metadata, Small IO, Large IO

  - Establish Baseline for future “health check” run
  - Failover and Failback
  - Hardware Replacement – understand replacement practices before go-live.
  - Security Configuration review – e.g default passwords, IPtables exceptions, services, admin + vendor accounts

- Tools
  - IOR
  - MDTest
  - Bonnie++
  - fio
  - dd

- Information Sources
  - Lustre /proc stat counters - http://wiki.lustre.org/Lustre_Monitoring_and_Statistics_Guide
  - Array side performance counters / tools
  - SNMP / IMPI – sensor data (temp), counters

- Logging & Monitoring
  - Central logging / Logstash
  - Zabbix | Nagios/Icinga

# Individual Drive Latency checks – looking for ‘bad’ disks

Table 2: Performance Monitor data for Storage Arrays g3e5660 - 12, 13 : Parallel DD Writes / Reads

| Drive Locations | | | Media Scan On : Array 12 : WRITE | | | | Media Scan On : Array 12 : READ | | | | Media Scan On: Array 13 : WRITE | | | | Media Scan On : Array 13 : READ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Max Lat ms | Min Lat ms | Avg I/O Lat ms | *Sorted | Max Lat ms | Min Lat ms | Avg I/O Lat ms | *Sorted | Max Lat ms | Min Lat ms | Avg I/O Lat ms | *Sorted | Max Lat ms | Min Lat ms | Avg I/O Lat ms | *Sorted |
| Drive Tray 1 | Drawer 1 | Slot 1 | 41.352 | 0.421 | 11.607 | 31.373 | 40.648 | 0.369 | 9.873 | 18.2 | 40.836 | 0.446 | 12.586 | 31.108 | 32.43 | 0.313 | 9.273 | 17.784 |
| Drive Tray 1 | Drawer 1 | Slot 2 | 35.286 | 0.283 | 12.069 | 31.09 | 1.723 | 0.269 | 1.084 | 18.018 | 34.902 | 0.489 | 13.406 | 31.02 | 1.387 | 0.274 | 0.99 | 17.161 |
| Drive Tray 1 | Drawer 1 | Slot 3 | 32.748 | 0.428 | 11.692 | 30.835 | 10.795 | 0.289 | 9.019 | 18.011 | 37.476 | 0.377 | 14.029 | 30.876 | 10.942 | 0.33 | 8.905 | 17.108 |
| Drive Tray 1 | Drawer 1 | Slot 4 | 34.346 | 0.702 | 11.49 | 30.793 | 37.896 | 0.632 | 9.508 | 17.891 | 28.796 | 0.367 | 11.832 | 30.811 | 27.575 | 0.866 | 8.836 | 17.044 |
| Drive Tray 1 | Drawer 1 | Slot 5 | 36.247 | 0.275 | 11.468 | 30.707 | 30.852 | 0.277 | 2.051 | 17.867 | 34.299 | 0.358 | 12.311 | 30.791 | 20.202 | 0.273 | 1.366 | 17.012 |
| Drive Tray 1 | Drawer 1 | Slot 6 | 31.661 | 0.312 | 11.645 | 30.638 | 10.388 | 0.3 | 8.714 | 17.849 | 36.2 | 0.662 | 12.939 | 30.56 | 10.495 | 0.305 | 8.525 | 16.969 |
| Drive Tray 1 | Drawer 1 | Slot 7 | 32.098 | 0.558 | 11.613 | 30.625 | 9.997 | 0.342 | 8.541 | 17.749 | 33.757 | 0.796 | 12.715 | 30.546 | 10.627 | 0.351 | 8.791 | 16.925 |
| Drive Tray 1 | Drawer 1 | Slot 8 | 37.532 | 0.377 | 11.602 | 30.612 | 31.027 | 0.477 | 2.128 | 17.701 | 32.135 | 0.314 | 12.093 | 30.538 | 23.641 | 0.474 | 1.4 | 16.922 |
| Drive Tray 1 | Drawer 1 | Slot 9 | 38.194 | 0.4 | 12.88 | 30.61 | 31.468 | 0.425 | 9.398 | 17.673 | 32.99 | 0.461 | 12.539 | 30.535 | 26.795 | 0.37 | 8.8 | 16.91 |
| Drive Tray 1 | Drawer 1 | Slot 10 | 25.827 | 0.646 | 14.019 | 30.594 | 13.049 | 0.454 | 9.588 | 17.665 | 37.689 | 0.702 | 12.308 | 30.529 | 10.804 | 0.644 | 8.978 | 16.879 |
| Drive Tray 1 | Drawer 1 | Slot 11 | 27.482 | 0.332 | 11.041 | 30.567 | 2.01 | 0.298 | 1.067 | 17.603 | 36.813 | 0.309 | 12.85 | 30.516 | 2.591 | 0.274 | 1.088 | 16.832 |
| Drive Tray 1 | Drawer 1 | Slot 12 | 35.105 | 0.284 | 12.081 | 30.505 | 36.77 | 0.892 | 9.717 | 17.6 | 30.195 | 0.433 | 12.022 | 30.409 | 30.126 | 0.296 | 8.981 | 16.795 |
| Drive Tray 1 | Drawer 2 | Slot 1 | 40.42 | 0.4 | 11.272 | 30.454 | 44.195 | 0.369 | 9.982 | 17.56 | 38.163 | 0.431 | 12.306 | 30.349 | 32.768 | 0.319 | 9.296 | 16.792 |
| Drive Tray 1 | Drawer 2 | Slot 2 | 33.656 | 0.278 | 11.799 | 30.366 | 1.578 | 0.271 | 1.075 | 17.559 | 39.742 | 0.479 | 13.288 | 30.313 | 1.844 | 0.269 | 1.054 | 16.76 |
| Drive Tray 1 | Drawer 2 | Slot 3 | 31.552 | 0.422 | 11.754 | 30.213 | 10.858 | 0.291 | 9.061 | 17.556 | 34.565 | 0.366 | 13.299 | 30.287 | 11.239 | 0.327 | 8.938 | 16.756 |
| Drive Tray 1 | Drawer 2 | Slot 4 | 38.438 | 0.726 | 11.285 | 29.264 | 41.124 | 0.675 | 9.979 | 17.551 | 28.504 | 0.369 | 11.881 | 30.245 | 30.45 | 1.02 | 9.162 | 16.741 |
| Drive Tray 1 | Drawer 2 | Slot 5 | 31.678 | 0.274 | 11.318 | 29.239 | 30.227 | 0.276 | 2.035 | 17.512 | 32.087 | 0.383 | 12.225 | 30.241 | 23.397 | 0.267 | 1.426 | 16.714 |
| Drive Tray 1 | Drawer 2 | Slot 6 | 28.917 | 0.316 | 11.138 | 29.165 | 10.48 | 0.295 | 8.889 | 17.507 | 34.071 | 0.677 | 12.783 | 30.214 | 10.96 | 0.349 | 8.772 | 16.712 |
| Drive Tray 1 | Drawer 2 | Slot 7 | 43.726 | 0.55 | 11.145 | 28.993 | 10.22 | 0.349 | 8.538 | 17.505 | 38.351 | 0.736 | 12.652 | 30.205 | 10.396 | 0.341 | 8.763 | 16.661 |
| Drive Tray 1 | Drawer 2 | Slot 8 | 38.75 | 0.387 | 11.705 | 28.972 | 34.949 | 0.473 | 2.247 | 17.48 | 34.96 | 0.339 | 12 | 30.193 | 22.836 | 0.495 | 1.399 | 16.656 |
| Drive Tray 1 | Drawer 2 | Slot 9 | 37.985 | 0.382 | 12.635 | 28.951 | 36.227 | 0.376 | 9.484 | 17.415 | 35.587 | 0.441 | 12.341 | 30.14 | 26.619 | 0.383 | 8.762 | 16.652 |
| Drive Tray 1 | Drawer 2 | Slot 10 | 27.287 | 0.595 | 13.2 | 28.835 | 10.622 | 0.49 | 8.892 | 17.39 | 34.694 | 0.636 | 11.984 | 30.132 | 10.766 | 0.615 | 8.758 | 16.644 |
| Drive Tray 1 | Drawer 2 | Slot 11 | 29.791 | 0.327 | 10.767 | 28.815 | 1.724 | 0.286 | 1.065 | 17.384 | 36.161 | 0.3 | 12.406 | 30.088 | 2.128 | 0.27 | 1.072 | 16.627 |
| Drive Tray 1 | Drawer 2 | Slot 12 | 32.43 | 0.281 | 12.069 | 28.75 | 34.195 | 0.947 | 9.421 | 17.33 | 31.694 | 0.426 | 11.929 | 30.018 | 28.707 | 0.288 | 8.725 | 16.608 |
| Drive Tray 1 | Drawer 3 | Slot 1 | 42.299 | 0.396 | 11.647 | 28.676 | 39.916 | 0.358 | 9.875 | 17.323 | 32.836 | 0.399 | 12.461 | 29.846 | 33.101 | 0.31 | 9.312 | 16.601 |
| Drive Tray 1 | Drawer 3 | Slot 2 | 36.915 | 0.42 | 11.772 | 28.674 | 29.61 | 0.266 | 2.081 | 17.307 | 32.867 | 0.266 | 12.264 | 29.795 | 26.298 | 0.265 | 1.457 | 16.596 |
| Drive Tray 1 | Drawer 3 | Slot 3 | 43.221 | 0.418 | 11.923 | 28.641 | 10.915 | 0.29 | 9.044 | 17.301 | 34.582 | 0.366 | 13.347 | 29.72 | 10.803 | 0.313 | 8.929 | 16.589 |
| Drive Tray 1 | Drawer 3 | Slot 4 | 25.979 | 0.5 | 11.282 | 28.63 | 10.475 | 0.298 | 8.618 | 17.29 | 40.544 | 0.753 | 12.748 | 29.71 | 11.015 | 0.325 | 8.924 | 16.587 |
| Drive Tray 1 | Drawer 3 | Slot 5 | 33.154 | 0.276 | 10.844 | 28.515 | 30.249 | 0.271 | 1.993 | 17.228 | 35.856 | 0.455 | 12.152 | 29.504 | 23.89 | 0.267 | 1.41 | 16.538 |
| Drive Tray 1 | Drawer 3 | Slot 6 | 32.876 | 0.305 | 11.463 | 28.501 | 37.817 | 0.336 | 9.858 | 17.174 | 28.957 | 0.84 | 14.254 | 29.3 | 27.536 | 0.448 | 9.304 | 16.528 |
| Drive Tray 1 | Drawer 3 | Slot 7 | 24.899 | 0.57 | 10.933 | 28.491 | 10.231 | 0.347 | 8.638 | 17.017 | 36.355 | 0.762 | 12.72 | 29.191 | 10.472 | 0.346 | 8.838 | 16.525 |
| Drive Tray 1 | Drawer 3 | Slot 8 | 25.956 | 0.283 | 10.727 | 28.476 | 2.065 | 0.281 | 1.037 | 16.968 | 39.457 | 0.518 | 12.484 | 29.067 | 1.624 | 0.267 | 1.041 | 16.509 |
| Drive Tray 1 | Drawer 3 | Slot 9 | 34.476 | 0.375 | 12.275 | 28.415 | 37.599 | 0.411 | 9.603 | 16.879 | 32.683 | 0.467 | 12.218 | 28.975 | 25.395 | 0.384 | 8.807 | 16.505 |
| Drive Tray 1 | Drawer 3 | Slot 10 | 36.468 | 0.416 | 11.565 | 28.394 | 39.782 | 0.338 | 9.633 | 16.857 | 28.566 | 0.416 | 11.687 | 28.926 | 34.739 | 0.38 | 9.109 | 16.501 |
| Drive Tray 1 | Drawer 3 | Slot 11 | 26.791 | 0.292 | 10.88 | 28.372 | 1.855 | 0.289 | 1.082 | 16.813 | 42.448 | 0.295 | 12.668 | 28.847 | 1.629 | 0.277 | 1.096 | 16.499 |
| Drive Tray 1 | Drawer 3 | Slot 12 | 32.352 | 0.649 | 11.303 | 28.34 | 10.362 | 0.413 | 8.809 | 16.807 | 41.932 | 0.391 | 12.519 | 28.786 | 10.382 | 0.682 | 8.754 | 16.492 |
| Drive Tray 1 | Drawer 4 | Slot 1 | 35.649 | 0.405 | 11.441 | 28.252 | 46.162 | 0.359 | 10.077 | 16.799 | 36.215 | 0.422 | 12.736 | 28.621 | 34.354 | 0.311 | 9.287 | 16.484 |
| Drive Tray 1 | Drawer 4 | Slot 2 | 35.971 | 0.456 | 11.74 | 28.25 | 37.801 | 0.266 | 2.322 | 16.767 | 42.407 | 0.266 | 12.671 | 28.611 | 28.383 | 0.265 | 1.529 | 16.476 |
| Drive Tray 1 | Drawer 4 | Slot 3 | 33.043 | 0.418 | 11.768 | 28.249 | 10.682 | 0.282 | 9.042 | 16.694 | 37.416 | 0.367 | 13.46 | 28.583 | 12.476 | 0.321 | 8.914 | 16.447 |
| Drive Tray 1 | Drawer 4 | Slot 4 | 26.815 | 0.504 | 11.15 | 28.246 | 10.689 | 0.302 | 8.683 | 16.656 | 52.268 | 0.811 | 12.737 | 28.573 | 11.015 | 0.324 | 8.952 | 16.438 |
| Drive Tray 1 | Drawer 4 | Slot 5 | 41.301 | 0.273 | 11.462 | 28.22 | 30.037 | 0.269 | 2.059 | 16.635 | 30.385 | 0.354 | 12.315 | 28.562 | 22.829 | 0.267 | 1.4 | 16.435 |
| Drive Tray 1 | Drawer 4 | Slot 6 | 30.82 | 0.298 | 11.656 | 28.123 | 38.541 | 0.335 | 9.859 | 16.592 | 34.382 | 0.838 | 14.848 | 28.549 | 24.71 | 0.446 | 9.424 | 16.435 |
| Drive Tray 1 | Drawer 4 | Slot 7 | 30.104 | 0.573 | 10.809 | 28.091 | 9.945 | 0.367 | 8.589 | 16.576 | 36.733 | 0.784 | 12.536 | 28.539 | 10.469 | 0.341 | 8.813 | 16.409 |

# RAID6 Performance – Pool Testing

– Validate RAID6 Pools prior to use as Lustre OST (array side)

g3e5660-10 - View Real-time Textual Performance Monitor

| Objects | Total IOs | Read % | Current IOs/sec | Maximum IOs/sec | Minimum IOs/sec | Average IOs/sec | SSD Read Cache Hit % | Current MBs/sec | Maximum MBs/sec | Minimum MBs/sec | Average MBs/sec | Current IO Latency(ms) | Maximum IO Latency(ms) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Storage Array g3e5660-10 | 8,006,914 | 99.8 | 3 | 13,230 | 3 | 2,635 | 0.0 | 0.0 | 7,035.9 | 0.0 | 1,601.1 | - | - |
| Controller A | 4,084,737 | 99.8 | 1 | 7,445 | 1 | 1,344 | 0.0 | 0.0 | 3,933.7 | 0.0 | 842.7 | - | - |
| Controller B | 3,922,177 | 99.8 | 1 | 7,155 | 1 | 1,291 | 0.0 | 0.0 | 3,954.4 | 0.0 | 758.4 | - | - |
| Volume Group OST_162 | 1,203,721 | 99.9 | 0 | 2,140 | 0 | 396 | 0.0 | 0.0 | 1,124.6 | 0.0 | 252.6 | - | - |
| Volume Group OST_163 | 928,926 | 99.8 | 0 | 1,917 | 0 | 305 | 0.0 | 0.0 | 976.2 | 0.0 | 168.5 | - | - |
| Volume Group OST_164 | 334,281 | 99.8 | 0 | 1,676 | 0 | 110 | 0.0 | 0.0 | 1,014.2 | 0.0 | 84.3 | - | - |
| Volume Group OST_165 | 335,150 | 99.8 | 0 | 1,663 | 0 | 110 | 0.0 | 0.0 | 968.6 | 0.0 | 84.3 | - | - |
| Volume Group OST_166 | 239,729 | 99.7 | 0 | 1,650 | 0 | 79 | 0.0 | 0.0 | 846.6 | 0.0 | 42.1 | - | - |
| Volume Group OST_167 | 240,259 | 99.7 | 0 | 1,693 | 0 | 79 | 0.0 | 0.0 | 862.4 | 0.0 | 42.1 | - | - |
| Volume Group OST_168 | 241,000 | 99.7 | 0 | 1,931 | 0 | 79 | 0.0 | 0.0 | 879.6 | 0.0 | 42.1 | - | - |
| Volume Group OST_169 | 238,944 | 99.7 | 0 | 1,613 | 0 | 78 | 0.0 | 0.0 | 933.9 | 0.0 | 42.1 | - | - |
| Volume Group OST_170 | 238,025 | 99.7 | 0 | 1,605 | 0 | 78 | 0.0 | 0.0 | 907.2 | 0.0 | 42.1 | - | - |
| Volume Group OST_171 | 907,979 | 99.9 | 0 | 1,740 | 0 | 298 | 0.0 | 0.0 | 983.2 | 0.0 | 168.5 | - | - |
| Volume Group OST_172 | 949,261 | 99.9 | 0 | 1,855 | 0 | 312 | 0.0 | 0.0 | 1,012.0 | 0.0 | 210.7 | - | - |
| Volume Group OST_173 | 385,284 | 99.8 | 0 | 1,463 | 0 | 126 | 0.0 | 0.0 | 805.0 | 0.0 | 84.3 | - | - |
| Volume Group OST_174 | 379,944 | 99.8 | 0 | 1,403 | 0 | 125 | 0.0 | 0.0 | 809.4 | 0.0 | 84.3 | - | - |
| Volume Group OST_175 | 387,340 | 99.8 | 0 | 1,513 | 0 | 127 | 0.0 | 0.0 | 834.0 | 0.0 | 84.3 | - | - |
| Volume Group OST_176 | 249,517 | 99.7 | 0 | 1,662 | 0 | 82 | 0.0 | 0.0 | 861.6 | 0.0 | 42.1 | - | - |
| Volume Group OST_177 | 249,422 | 99.7 | 0 | 1,741 | 0 | 82 | 0.0 | 0.0 | 883.2 | 0.0 | 42.1 | - | - |
| Volume Group OST_178 | 249,259 | 99.7 | 0 | 1,729 | 0 | 82 | 0.0 | 0.0 | 881.0 | 0.0 | 42.1 | - | - |

Start: 7/24/15 4:16:53 PM
Stop:
Time monitored: 00:50:38

Object marked as * will not report any value as it is no more a valid reference.

| Objects | Total IOs | Read % | Current IOs/sec | Maximum IOs/sec | Minimum IOs/sec | Average IOs/sec | SSD Read Cache Hit % | Current MBs/sec | Maximum MBs/sec | Minimum MBs/sec | Average MBs/sec | Current IO Latency(ms) | Maximum IO Latency(ms) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Storage Array g3e5660-11 | 6,294,262 | 99.8 | 5 | 13,564 | 3 | 2,546 | 0.0 | 0.0 | 6,984.2 | 0.0 | 1,602.3 | - | - |
| Controller A | 3,072,381 | 99.8 | 2 | 7,430 | 1 | 1,244 | 0.0 | 0.0 | 3,781.8 | 0.0 | 775.1 | - | - |
| Controller B | 3,211,901 | 99.8 | 2 | 6,384 | 1 | 1,301 | 0.0 | 0.0 | 3,638.6 | 0.0 | 827.2 | - | - |
| Volume Group OST_180 | 560,797 | 99.9 | 0 | 1,545 | 0 | 227 | 0.0 | 0.0 | 918.4 | 0.0 | 153.5 | - | - |
| Volume Group OST_181 | 566,088 | 99.9 | 0 | 1,495 | 0 | 229 | 0.0 | 0.0 | 833.0 | 0.0 | 154.0 | - | - |
| Volume Group OST_182 | 332,462 | 99.8 | 0 | 1,563 | 0 | 134 | 0.0 | 0.0 | 873.5 | 0.0 | 103.7 | - | - |
| Volume Group OST_183 | 331,677 | 99.8 | 0 | 1,485 | 0 | 134 | 0.0 | 0.0 | 873.0 | 0.0 | 103.7 | - | - |
| Volume Group OST_184 | 246,717 | 99.8 | 0 | 1,582 | 0 | 99 | 0.0 | 0.0 | 800.4 | 0.0 | 51.9 | - | - |
| Volume Group OST_185 | 246,073 | 99.8 | 0 | 1,666 | 0 | 99 | 0.0 | 0.0 | 846.8 | 0.0 | 51.9 | - | - |
| Volume Group OST_186 | 246,296 | 99.8 | 0 | 1,626 | 0 | 99 | 0.0 | 0.0 | 826.8 | 0.0 | 51.9 | - | - |
| Volume Group OST_187 | 245,772 | 99.8 | 0 | 1,642 | 0 | 99 | 0.0 | 0.0 | 848.2 | 0.0 | 51.9 | - | - |
| Volume Group OST_188 | 245,666 | 99.8 | 0 | 1,678 | 0 | 99 | 0.0 | 0.0 | 847.8 | 0.0 | 51.9 | - | - |
| Volume Group OST_189 | 563,315 | 99.9 | 0 | 1,592 | 0 | 228 | 0.0 | 0.0 | 895.9 | 0.0 | 154.6 | - | - |
| Volume Group OST_190 | 562,063 | 99.9 | 0 | 1,488 | 0 | 227 | 0.0 | 0.0 | 934.4 | 0.0 | 154.8 | - | - |
| Volume Group OST_191 | 380,491 | 99.9 | 0 | 1,417 | 0 | 154 | 0.0 | 0.0 | 809.4 | 0.0 | 103.7 | - | - |
| Volume Group OST_192 | 384,096 | 99.9 | 0 | 1,584 | 0 | 155 | 0.0 | 0.0 | 903.2 | 0.0 | 103.7 | - | - |
| Volume Group OST_193 | 384,281 | 99.9 | 0 | 1,803 | 0 | 155 | 0.0 | 0.0 | 991.0 | 0.0 | 103.7 | - | - |
| Volume Group OST_194 | 247,003 | 99.8 | 0 | 1,651 | 0 | 100 | 0.0 | 0.0 | 847.2 | 0.0 | 51.9 | - | - |
| Volume Group OST_195 | 247,225 | 99.8 | 0 | 1,690 | 0 | 100 | 0.0 | 0.0 | 869.8 | 0.0 | 51.9 | - | - |
| Volume Group OST_196 | 247,061 | 99.8 | 0 | 1,640 | 0 | 100 | 0.0 | 0.0 | 844.8 | 0.0 | 51.9 | - | - |
| Volume Group OST_197 | 246,779 | 99.8 | 0 | 1,731 | 0 | 99 | 0.0 | 0.0 | 886.2 | 0.0 | 51.9 | - | - |

Start: 7/24/15 4:28:24 PM
Stop:

Object marked as * will not report any value as it is no more a valid reference.

– Each OST is individually tested for performance consistency (client side). Hours and hours in front of Excel, SSH, scripts, CSVs.

Gdata3 DD Benchmarking.xlsx

**RAID 6 - Test 4: Array Contention using both Controller A and Controller B, both OSS**

WRITE TEST -18 OSTs from OSS 21 & OSS 22

| Test 2c - Parallel Test, 128G, 1M block | Pass1 MB/sec | Pass2 MB/sec | Pass3 MB/sec | Avg MB/sec | Santricity Max MB/sec |
|---|---|---|---|---|---|
| OST idx 180, OSS21, CtrlA, r19 | 378 | 373 | 371 | 374.0 | 803 |
| OST idx 181, OSS21, CtrlB, r20 | 463 | 449 | 443 | 451.7 | 847 |
| OST idx 182, OSS21, CtrlA, r21 | 382 | 377 | 369 | 376.0 | 842 |
| OST idx 183, OSS21, CtrlB, r22 | 465 | 445 | 440 | 450.0 | 812 |
| OST idx 184, OSS21, CtrlA, r23 | 381 | 377 | 370 | 376.0 | 850 |
| OST idx 185, OSS21, CtrlB, r24 | 452 | 449 | 439 | 446.7 | 620 |
| OST idx 186, OSS21, CtrlA, r25 | 378 | 374 | 372 | 374.7 | 646 |
| OST idx 187, OSS21, CtrlB, r26 | 454 | 452 | 444 | 450.0 | 576 |
| OST idx 188, OSS21, CtrlA, r27 | 376 | 372 | 368 | 372.0 | 720 |
| OST idx 189, OSS22, CtrlB, r28 | 435 | 434 | 428 | 432.3 | 676 |
| OST idx 190, OSS22, CtrlA, r29 | 369 | 366 | 362 | 365.7 | 864 |
| OST idx 191, OSS22, CtrlB, r30 | 431 | 431 | 425 | 429.0 | 682 |
| OST idx 192, OSS22, CtrlA, r31 | 371 | 369 | 363 | 367.7 | 904 |
| OST idx 193, OSS22, CtrlB, r32 | 432 | 432 | 425 | 429.7 | 690 |
| OST idx 194, OSS22, CtrlA, r33 | 372 | 371 | 364 | 369.0 | 849 |
| OST idx 195, OSS22, CtrlB, r34 | 433 | 430 | 425 | 429.3 | 783 |
| OST idx 196, OSS22, CtrlA, r35 | 371 | 370 | 363 | 368.0 | 869 |
| OST idx 197, OSS22, CtrlB, r36 | 432 | 431 | 425 | 429.3 | 849 |
| average (col) | **409.7** | **405.7** | **399.8** | | **771.2** |
| aggregate (column) | **7375.0** | **7302.0** | **7196.0** | | |
| average all runs, all tests | | | | 405.1 | Array MAX = 6937 |

READ TEST -18 OSTs from OSS 21 & OSS 22

| Test 3c - Parallel Test, 128G, 1M block | Pass1 MB/sec | Pass2 MB/sec | Pass3 MB/sec | Avg MB/sec | Santricity Max MB/sec |
|---|---|---|---|---|---|
| OST idx 180, OSS21, CtrlA, r19 | 528 | 528 | 529 | 528.3 | 985 |
| OST idx 181, OSS21, CtrlB, r20 | 562 | 560 | 557 | 559.7 | 665 |
| OST idx 182, OSS21, CtrlA, r21 | 543 | 547 | 544 | 544.7 | 622 |
| OST idx 183, OSS21, CtrlB, r22 | 542 | 547 | 546 | 545.0 | 662 |
| OST idx 184, OSS21, CtrlA, r23 | 544 | 547 | 541 | 544.0 | 659 |
| OST idx 185, OSS21, CtrlB, r24 | 563 | 654 | 561 | 592.7 | 592 |
| OST idx 186, OSS21, CtrlA, r25 | 534 | 535 | 530 | 533.0 | 904 |
| OST idx 187, OSS21, CtrlB, r26 | 563 | 562 | 561 | 562.0 | 601 |
| OST idx 188, OSS21, CtrlA, r27 | 532 | 533 | 527 | 530.7 | 906 |
| OST idx 189, OSS22, CtrlB, r28 | 534 | 533 | 536 | 534.3 | 804 |
| OST idx 190, OSS22, CtrlA, r29 | 560 | 559 | 558 | 559.0 | 594 |
| OST idx 191, OSS22, CtrlB, r30 | 525 | 526 | 526 | 525.7 | 957 |
| OST idx 192, OSS22, CtrlA, r31 | 582 | 583 | 583 | 582.7 | 661 |
| OST idx 193, OSS22, CtrlB, r32 | 525 | 522 | 522 | 523.0 | 934 |
| OST idx 194, OSS22, CtrlA, r33 | 554 | 556 | 556 | 555.3 | 617 |
| OST idx 195, OSS22, CtrlB, r34 | 550 | 553 | 553 | 552.0 | 728 |
| OST idx 196, OSS22, CtrlA, r35 | 541 | 541 | 541 | 541.0 | 786 |
| OST idx 197, OSS22, CtrlB, r36 | 552 | 557 | 557 | 555.3 | 779 |
| average (col) | **546.3** | **552.4** | **546.0** | | |
| aggregate (coulmn) | **9834.0** | **9943.0** | **9828.0** | | |
| average all runs, all tests | | | | 548.2 | Array MAX = 9727 |

DDPTest 3 – Parallel from Ctrlr | DDPTest 4 – Parallel all | R6 Test 1 – individual OST | R6 Test 2 -OSS | R6 Test 3 -Controller | R6 Test 4 – Parallel All

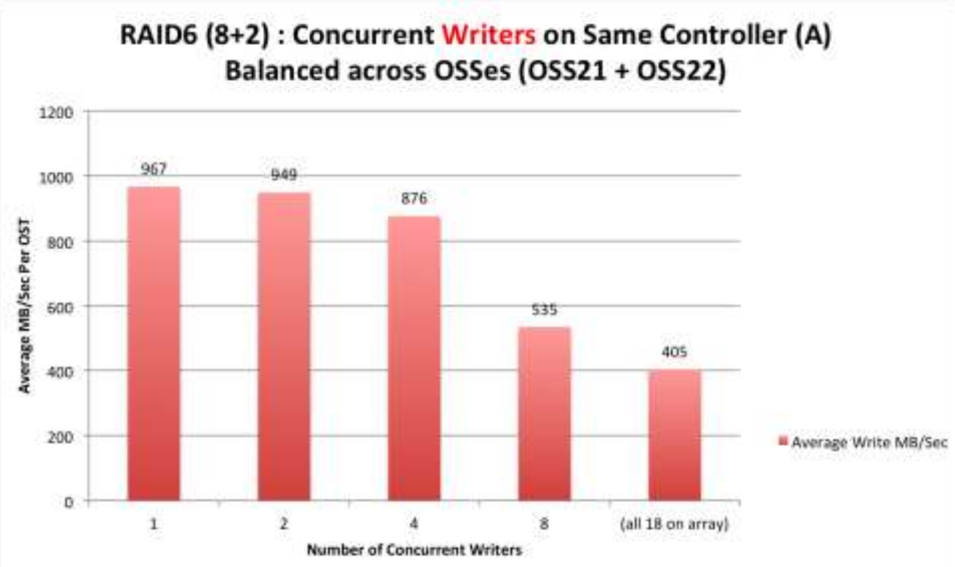
- Another Series of tests, progressively loading up contention on a OSS or an E5600 controller to determine performance decay behavior.

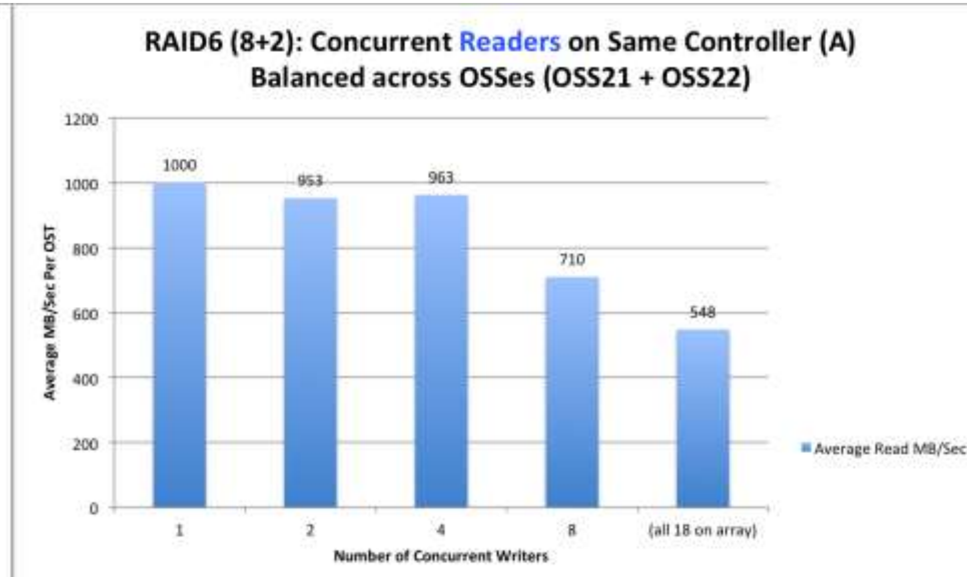
| READ TEST - 8 OSTs from OSS 21 | | | | | |
|---|---|---|---|---|---|
| Test 3c - Parallel Test, 128G, 1M block | Pass1 MB/sec | Pass2 MB/sec | Pass3 MB/sec | Avg MB/sec | Santricity Max MB/sec |
| OST idx 180, OSS21, CtrlA, r19 | 627 | 623 | 627 | 625.7 | 810 |
| OST idx 181, OSS21, CtrlB, r20 | 634 | 632 | 638 | 634.7 | 782 |
| OST idx 182, OSS21, CtrlA, r21 | 640 | 650 | 646 | 645.3 | 685 |
| OST idx 183, OSS21, CtrlB, r22 | 638 | 632 | 636 | 635.3 | 789 |
| OST idx 184, OSS21, CtrlA, r23 | 652 | 655 | 653 | 653.3 | 658 |
| OST idx 185, OSS21, CtrlB, r24 | 651 | 657 | 656 | 654.7 | 671 |
| OST idx 186, OSS21, CtrlA, r25 | 633 | 635 | 636 | 634.7 | 883 |
| OST idx 187, OSS21, CtrlB, r26 | 640 | 643 | 643 | 642.0 | 859 |
| OST idx 188, OSS21, CtrlA, r27 | | | | #DIV/0! | |
| OST idx 189, OSS22, CtrlB, r28 | | | | #DIV/0! | |
| OST idx 190, OSS22, CtrlA, r29 | | | | #DIV/0! | |
| OST idx 191, OSS22, CtrlB, r30 | | | | #DIV/0! | |
| OST idx 192, OSS22, CtrlA, r31 | | | | #DIV/0! | |
| OST idx 193, OSS22, CtrlB, r32 | | | | #DIV/0! | |
| OST idx 194, OSS22, CtrlA, r33 | | | | #DIV/0! | |
| OST idx 195, OSS22, CtrlB, r34 | | | | #DIV/0! | |
| OST idx 196, OSS22, CtrlA, r35 | | | | #DIV/0! | |
| OST idx 197, OSS22, CtrlB, r36 | | | | #DIV/0! | |
| average (col) | **639.4** | **640.9** | **641.9** | | |
| aggregate (coulmn) | **5115.0** | **5127.0** | **5135.0** | | |
| average all runs, all tests | | | | 640.7 | Array MAX = 5074 |

| READ TEST -18 OSTs from OSS 21 & OSS 22 | | | | | |
|---|---|---|---|---|---|
| Test 3c - Parallel Test, 128G, 1M block | Pass1 MB/sec | Pass2 MB/sec | Pass3 MB/sec | Avg MB/sec | Santricity Max MB/sec |
| OST idx 180, OSS21, CtrlA, r19 | 528 | 528 | 529 | 528.3 | 985 |
| OST idx 181, OSS21, CtrlB, r20 | 562 | 560 | 557 | 559.7 | 665 |
| OST idx 182, OSS21, CtrlA, r21 | 543 | 547 | 544 | 544.7 | 622 |
| OST idx 183, OSS21, CtrlB, r22 | 542 | 547 | 546 | 545.0 | 662 |
| OST idx 184, OSS21, CtrlA, r23 | 544 | 547 | 541 | 544.0 | 659 |
| OST idx 185, OSS21, CtrlB, r24 | 563 | 654 | 561 | 592.7 | 592 |
| OST idx 186, OSS21, CtrlA, r25 | 534 | 535 | 530 | 533.0 | 904 |
| OST idx 187, OSS21, CtrlB, r26 | 563 | 562 | 561 | 562.0 | 601 |
| OST idx 188, OSS21, CtrlA, r27 | 532 | 533 | 527 | 530.7 | 906 |
| OST idx 189, OSS22, CtrlB, r28 | 534 | 533 | 536 | 534.3 | 804 |
| OST idx 190, OSS22, CtrlA, r29 | 560 | 559 | 558 | 559.0 | 594 |
| OST idx 191, OSS22, CtrlB, r30 | 525 | 526 | 526 | 525.7 | 957 |
| OST idx 192, OSS22, CtrlA, r31 | 582 | 583 | 583 | 582.7 | 661 |
| OST idx 193, OSS22, CtrlB, r32 | 525 | 522 | 522 | 523.0 | 934 |
| OST idx 194, OSS22, CtrlA, r33 | 554 | 556 | 556 | 555.3 | 617 |
| OST idx 195, OSS22, CtrlB, r34 | 550 | 553 | 553 | 552.0 | 728 |
| OST idx 196, OSS22, CtrlA, r35 | 541 | 541 | 541 | 541.0 | 786 |
| OST idx 197, OSS22, CtrlB, r36 | 552 | 557 | 557 | 555.3 | 779 |
| average (col) | **546.3** | **552.4** | **546.0** | | |
| aggregate (coulmn) | **9834.0** | **9943.0** | **9828.0** | | |
| average all runs, all tests | | | | 548.2 | Array MAX = 9727 |

- RAID 6 Writers and Readers contending on Controller A
- 1 Reader = Single reader on entire array
- 2 readers = 1 reader on OSS 21, 1 on OSS 22.

Controller Contention test

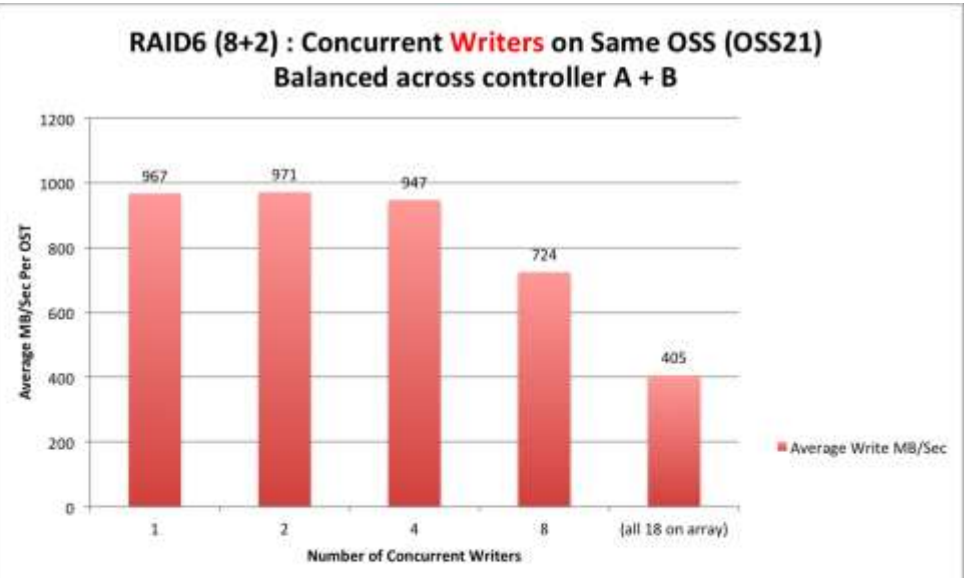


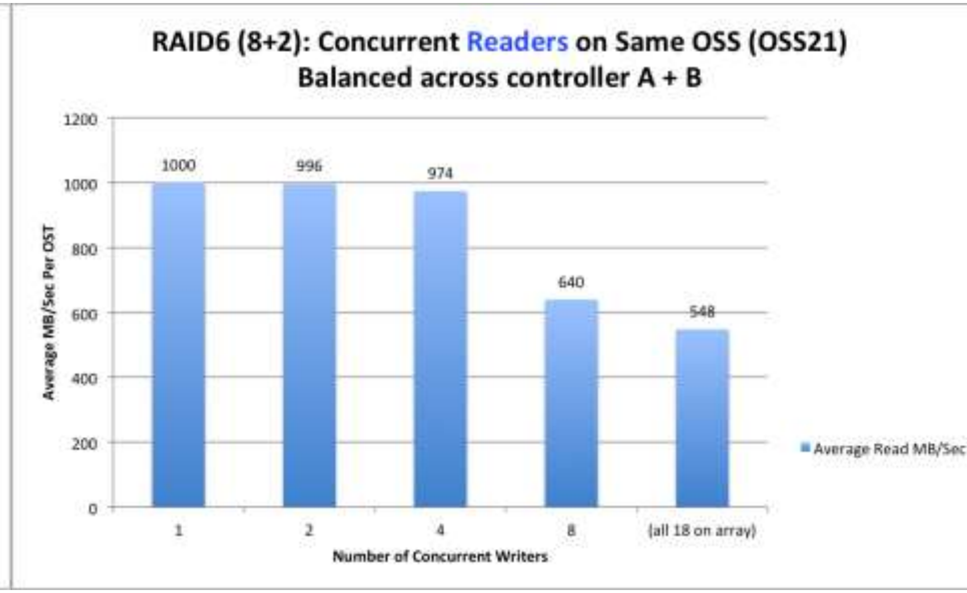




Note: Progressive performance decay as controller (A) reaches fully loaded configuration

- RAID 6 Writers and Readers contending on same OSS
- 1 Reader = Single reader on entire array
- 2 readers = 1 reader on Controller A, 1 on Controller B.

OSS
Contention test

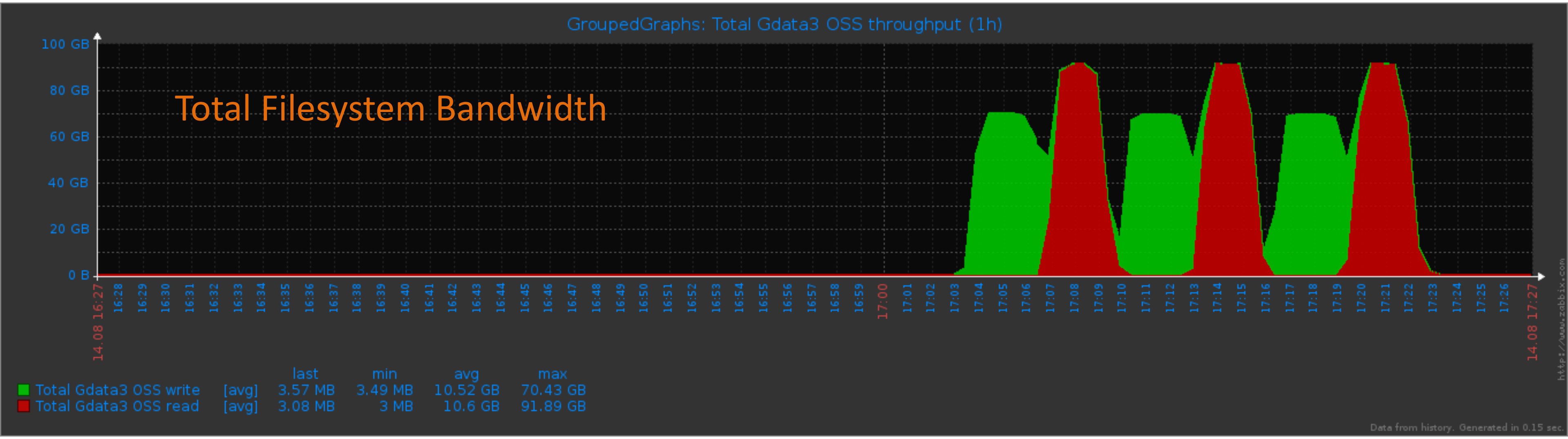



Note: Progressive performance decay as OSS21 reaches fully loaded configuration
(OSS has 1 FDR IB link at 56G, 8x 640MB/sec = 5120MB/sec)

- IOR
  - IOR Benchmark against as-built Lustre filesystem
  - Requires 200-300 clients to fully exercise filesystem
  - Expectations of 150GB+ sec Read, 90GB+ sec Write (sequential aggregate)
  - BUT…
  - LNETs Routers will ultimately cap performance (10GB sec each max, 14x)

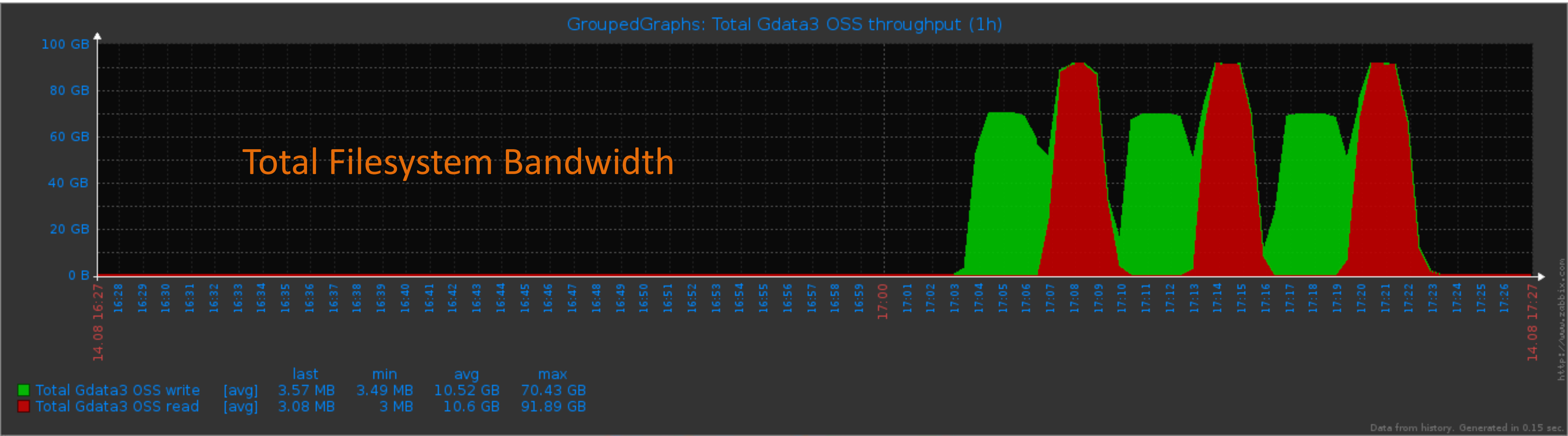
- Gdata3 –IOR. 198 OSTs, 198 Clients (11x Array Configuration)
  - 1 client per OST, 64GB File size.
  - Filesystem empty with exclusive access

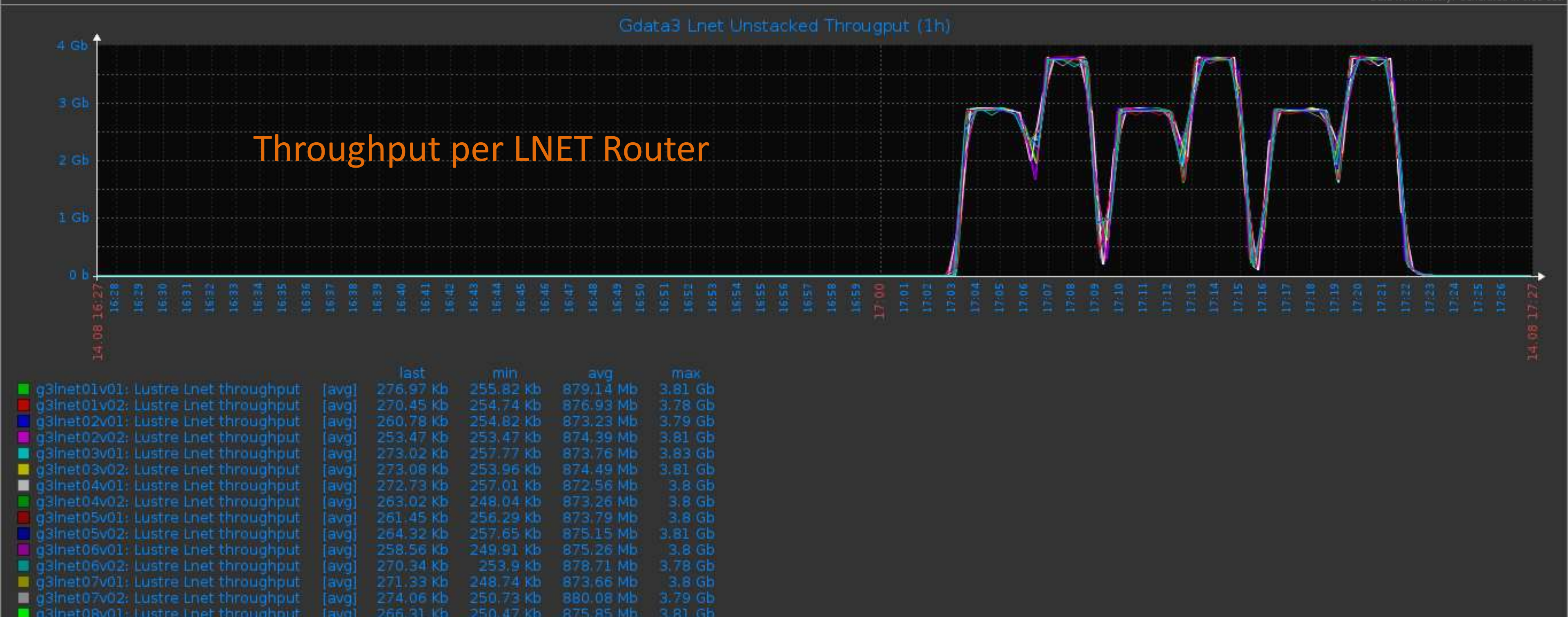



Read Max = 91.89GB/sec

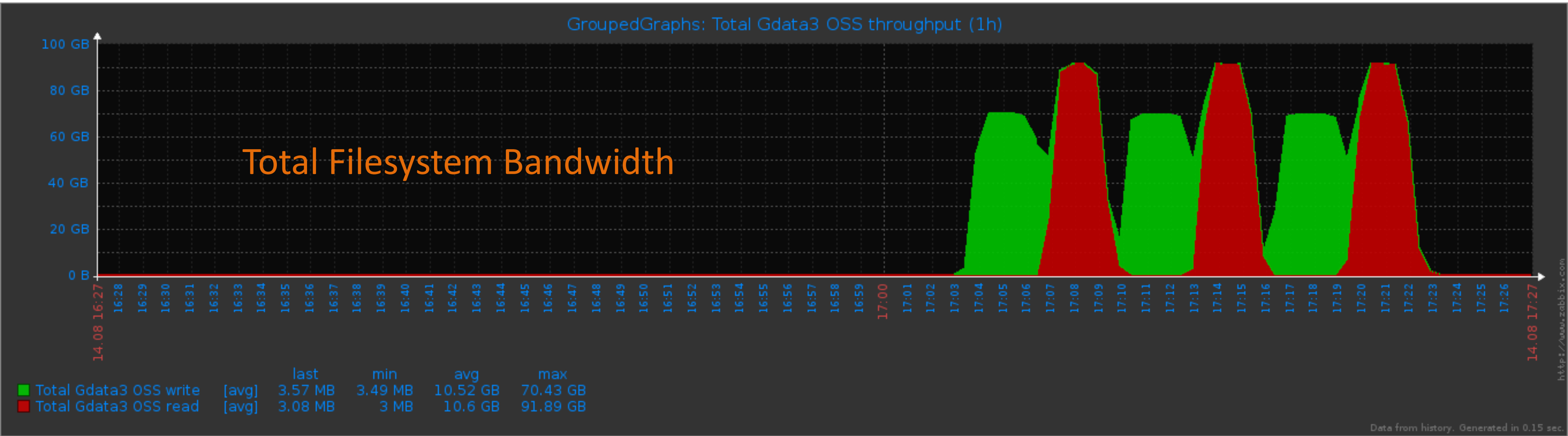
Write Max = 70.43GB/sec

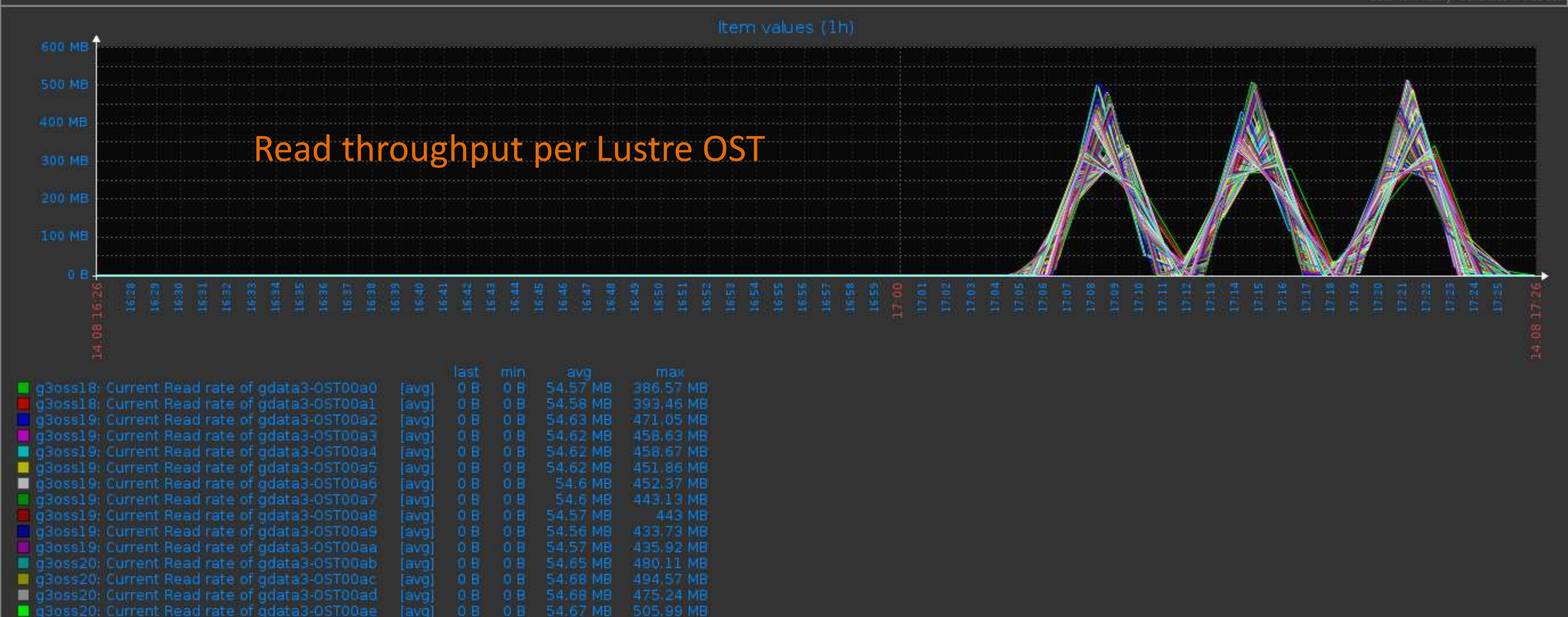
# IOR Performance – what else is happening?



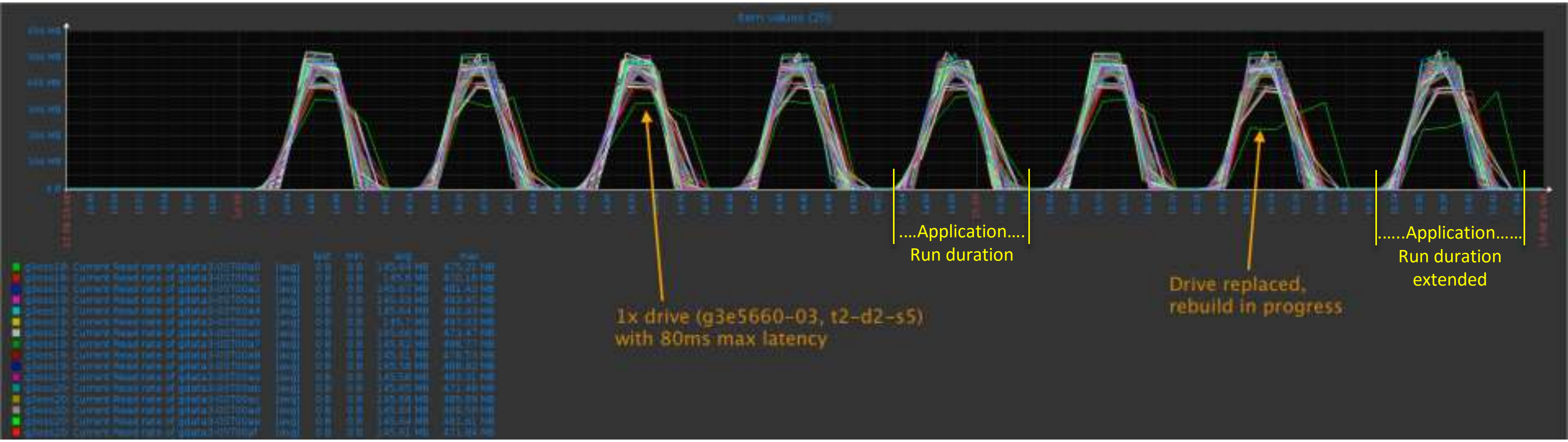



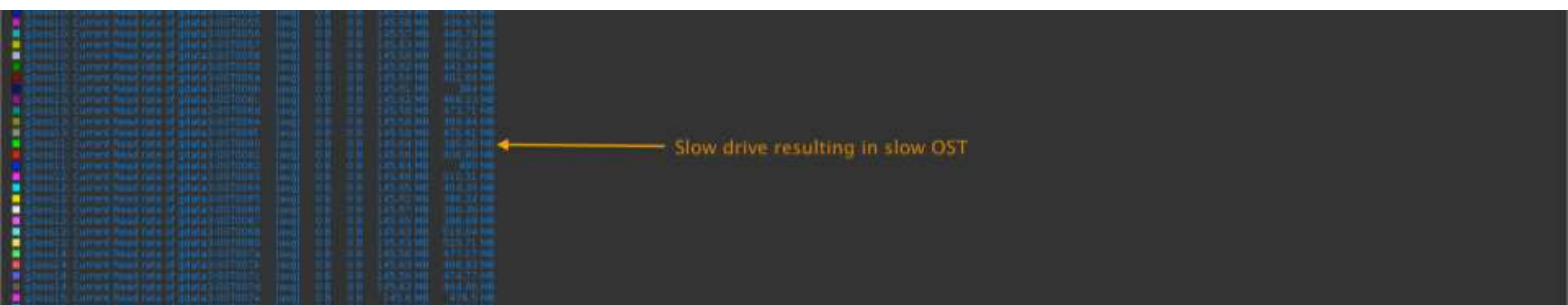
# IOR Performance – what else is happening?



- Good drives go bad – particularly at ends of system lifetime bathtub curve
- Example application – IOR (simulates HPC I/O workload)
- Plot of 198 RAID6 OSTs. Poorly performing OSTs identified
- Slow drive replaced. RAID6 Rebuild time = 17h 14m



# LNET Performance benchmark - Virtual

- Gdata3 – 26x virtual LNETs at scale & balanced, consistent


| | | | last | min | avg | max |
|---|---|---|---|---|---|---|
| g3lnet01v01: | Lustre Lnet throughput | [avg] | 295.55 Kb | 231.28 Kb | 2.44 Gb | 3.74 Gb |
| g3lnet01v02: | Lustre Lnet throughput | [avg] | 245.58 Kb | 234.28 Kb | 2.44 Gb | 3.72 Gb |
| g3lnet02v01: | Lustre Lnet throughput | [avg] | 290.28 Kb | 232.7 Kb | 2.44 Gb | 3.74 Gb |
| g3lnet02v02: | Lustre Lnet throughput | [avg] | 253.79 Kb | 228.67 Kb | 2.46 Gb | 3.75 Gb |
| g3lnet03v01: | Lustre Lnet throughput | [avg] | 294.84 Kb | 230.07 Kb | 2.45 Gb | 3.75 Gb |
| g3lnet03v02: | Lustre Lnet throughput | [avg] | 288.73 Kb | 235.87 Kb | 2.45 Gb | 3.77 Gb |
| g3lnet04v01: | Lustre Lnet throughput | [avg] | 237.05 Kb | 231.71 Kb | 2.45 Gb | 3.74 Gb |
| g3lnet04v02: | Lustre Lnet throughput | [avg] | 234.68 Kb | 231.87 Kb | 2.45 Gb | 3.74 Gb |
| g3lnet05v01: | Lustre Lnet throughput | [avg] | 239.49 Kb | 233.42 Kb | 2.46 Gb | 3.74 Gb |
| g3lnet05v02: | Lustre Lnet throughput | [avg] | 290.4 Kb | 233.84 Kb | 2.45 Gb | 3.74 Gb |
| g3lnet06v01: | Lustre Lnet throughput | [avg] | 234.65 Kb | 233.23 Kb | 2.46 Gb | 3.75 Gb |
| g3lnet06v02: | Lustre Lnet throughput | [avg] | 302.03 Kb | 231.77 Kb | 2.45 Gb | 3.73 Gb |
| g3lnet07v01: | Lustre Lnet throughput | [avg] | 285.2 Kb | 224.87 Kb | 2.45 Gb | 3.74 Gb |
| g3lnet07v02: | Lustre Lnet throughput | [avg] | 314.38 Kb | 227.45 Kb | 2.45 Gb | 3.74 Gb |
| g3lnet08v01: | Lustre Lnet throughput | [avg] | 252.5 Kb | 234.44 Kb | 2.45 Gb | 3.77 Gb |
| g3lnet08v02: | Lustre Lnet throughput | [avg] | 245.86 Kb | 230.55 Kb | 2.46 Gb | 3.87 Gb |
| g3lnet10v01: | Lustre Lnet throughput | [avg] | 251.64 Kb | 223.64 Kb | 2.46 Gb | 3.76 Gb |
| g3lnet10v02: | Lustre Lnet throughput | [avg] | 314.27 Kb | 238.33 Kb | 2.45 Gb | 3.74 Gb |
| g3lnet11v01: | Lustre Lnet throughput | [avg] | 237.1 Kb | 230.37 Kb | 2.46 Gb | 3.74 Gb |
| g3lnet11v02: | Lustre Lnet throughput | [avg] | 240.5 Kb | 236.14 Kb | 2.46 Gb | 3.75 Gb |
| g3lnet12v01: | Lustre Lnet throughput | [avg] | 288.77 Kb | 236.66 Kb | 2.45 Gb | 3.73 Gb |
| g3lnet12v02: | Lustre Lnet throughput | [avg] | 272.59 Kb | 233.03 Kb | 2.45 Gb | 3.72 Gb |
| g3lnet13v01: | Lustre Lnet throughput | [avg] | 305.86 Kb | 228.52 Kb | 2.45 Gb | 3.74 Gb |
| g3lnet13v02: | Lustre Lnet throughput | [avg] | 241 Kb | 232.76 Kb | 2.44 Gb | 3.74 Gb |
| g3lnet14v01: | Lustre Lnet throughput | [avg] | 240.84 Kb | 235.28 Kb | 2.45 Gb | 3.73 Gb |
| g3lnet14v02: | Lustre Lnet throughput | [avg] | 244.66 Kb | 231.93 Kb | 2.43 Gb | 3.73 Gb |

3.74GB/sec sustained peak (read) per virtual LNET Router, 26 of 28 routers active.

G3lnet09v01 and v02 offline for repair

26x 3.74 = 97.24GB/sec (exhausting available array b/w)

# Questions ?

NCI

Providing Australian researchers with world-class computing services


**NCI Contacts**

General enquiries: +61 2 6125 9800  Media enquiries: +61 2 6125 4389

Help desk: help@nci.org.au

**Address:**

NCI, Building 143, Ward Road  The Australian National University  Canberra ACT 0200


Australian Government Department of Education

Australian National University

Australian Government Bureau of Meteorology

CSIRO

Australian Government Geoscience Australia

Australian Government Australian Research Council

nci.org.au

@NCInews