NCI

Providing Australian researchers with
world-class computing services

Lustre Admins & Developers Workshop 2018

ZFS or ldiskfs?

12 Months with the NCI Lustre Fraternal Twins

Daniel Rodwell
Manager, Data Storage Services

September 2018

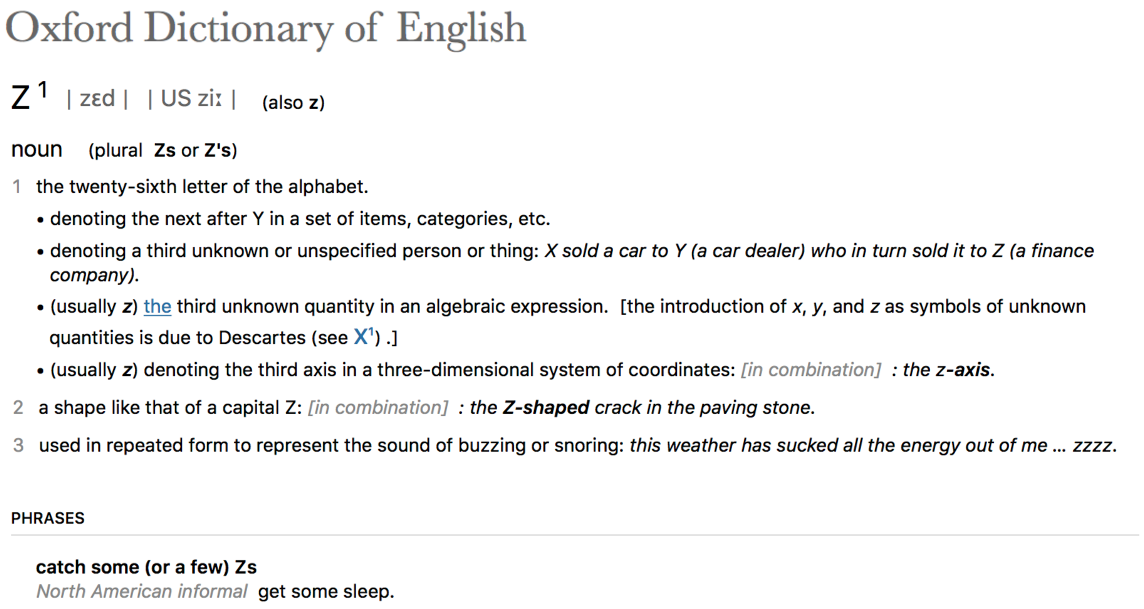
Agenda

- **Intro to NCI**
  - Systems & Growth
- **Gdata1 Replacement**
  - Requirements & Procurement
  - Design
- **Acceptance Testing & Performance**
  - Acceptance Testing
  - Performance
- **Migration**
  - Method & Tools
- **Reflection**
  - Good, Bad, Better?
  - Future

Zee or Zed

– Zee FS or Zed FS ?

- For this presentation, I may use both interchangeably.



Oxford Dictionary of English

Z^{1} | zɛd | | US ziː | (also z)

noun (plural **Zs** or **Z's**)

1 the twenty-sixth letter of the alphabet.

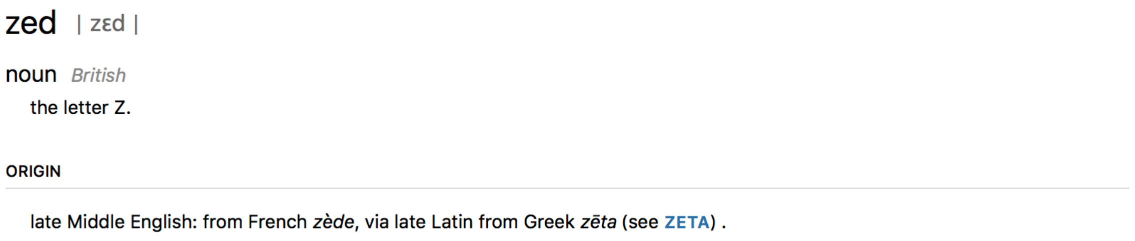
- denoting the next after Y in a set of items, categories, etc.
- denoting a third unknown or unspecified person or thing: *X sold a car to Y (a car dealer) who in turn sold it to Z (a finance company).*
- (usually ***z***) the third unknown quantity in an algebraic expression. [the introduction of *x*, *y*, and *z* as symbols of unknown quantities is due to Descartes (see X^{1}) .]
- (usually ***z***) denoting the third axis in a three-dimensional system of coordinates: *[in combination]* : *the **z-axis**.*

2 a shape like that of a capital Z: *[in combination]* : *the **Z-shaped** crack in the paving stone.*

3 used in repeated form to represent the sound of buzzing or snoring: *this weather has sucked all the energy out of me ... zzzz.*

PHRASES

catch some (or a few) Zs
North American informal get some sleep.

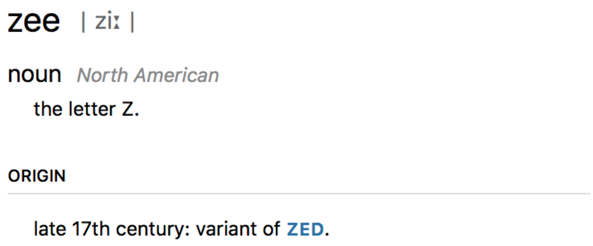


zed | zɛd |

noun *British*
the letter Z.

ORIGIN

late Middle English: from French *zède*, via late Latin from Greek *zēta* (see **ZETA**) .



zee | ziː |

noun *North American*
the letter Z.

ORIGIN

late 17th century: variant of **ZED**.

National High Performance Compute + Data Capability

Intro to NCI

NCI – an overview

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.

NCRIS
National Research
Infrastructure for Australia
An Australian Government Initiative

Australian National University

CSIRO

Australian Government
Bureau of Meteorology

Australian Government
Geoscience Australia

Australian Government
Australian Research Council

DEAKIN UNIVERSITY

INTERSECT

RMIT UNIVERSITY

qcif

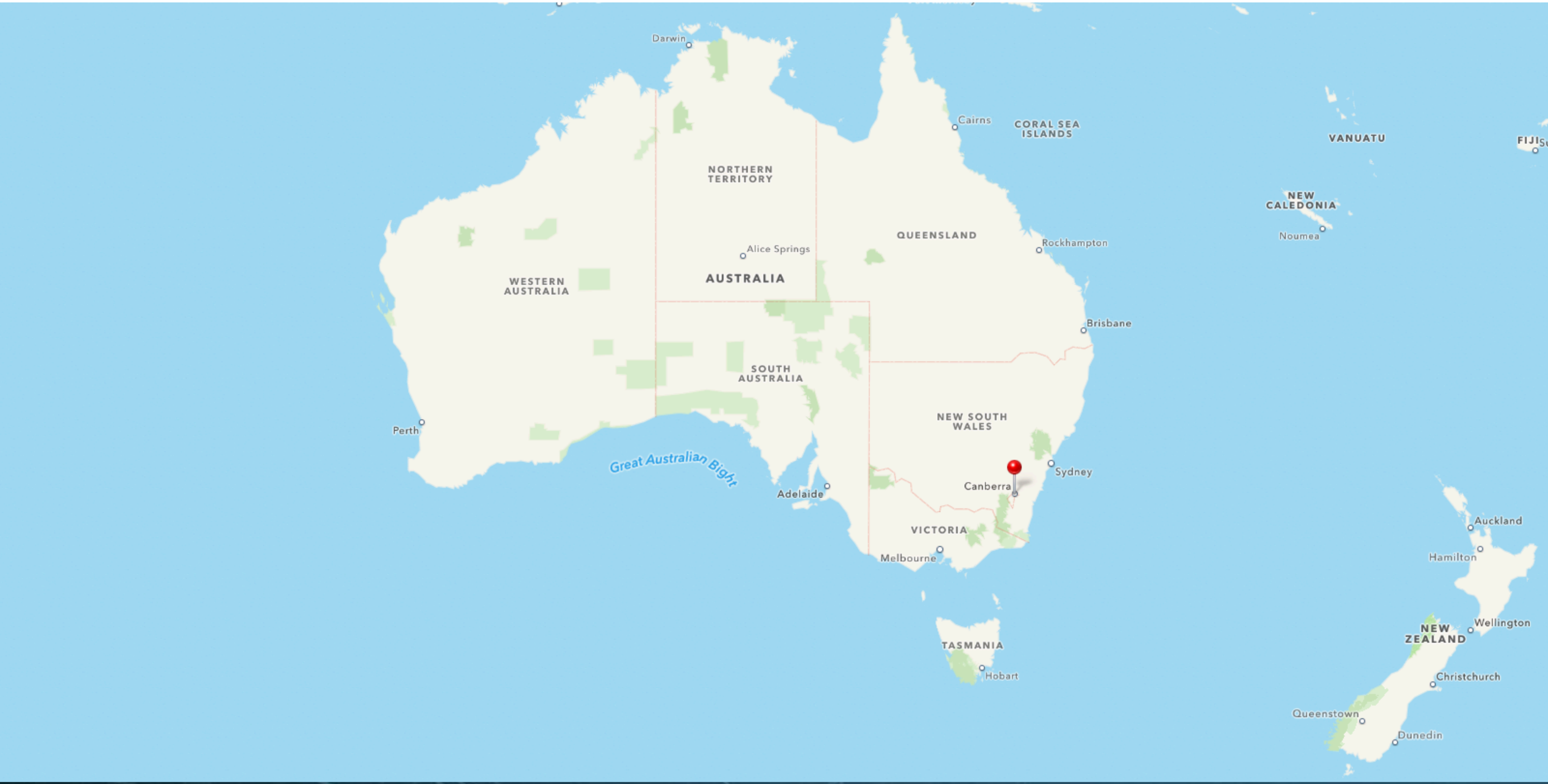
ACE CRC
ANTARCTIC CLIMATE & ECOSYSTEMS
COOPERATIVE RESEARCH CENTRE

UTS
UNIVERSITY OF TECHNOLOGY, SYDNEY

GARVAN INSTITUTE
OF MEDICAL RESEARCH

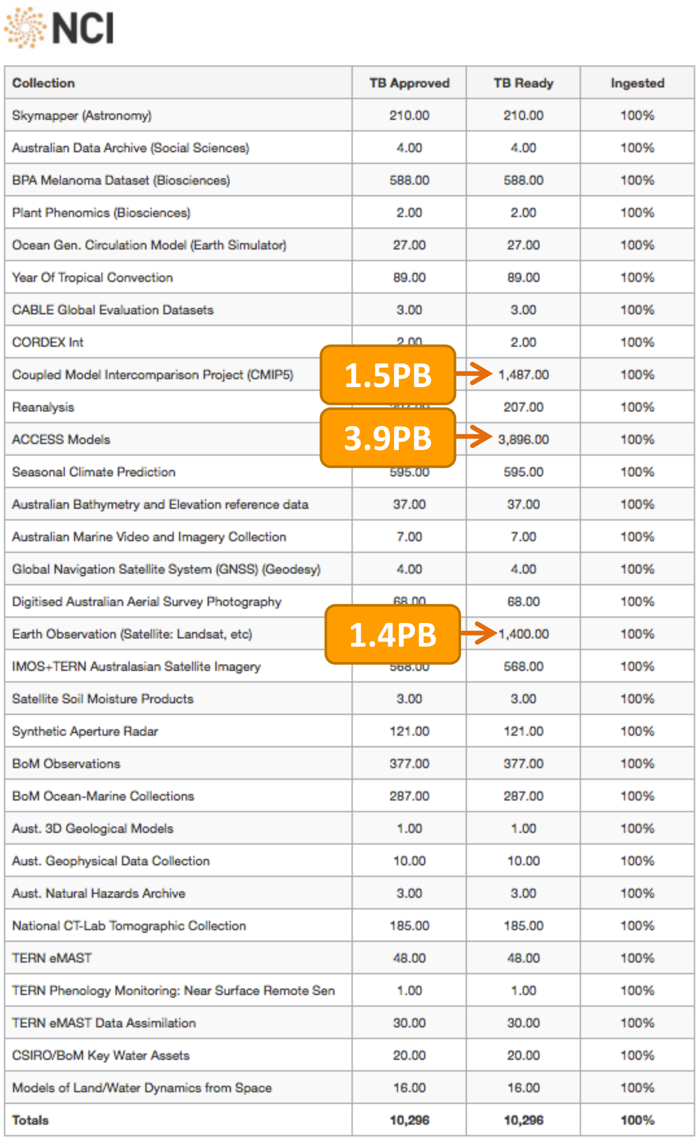
Where are we located?

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



What do we store?

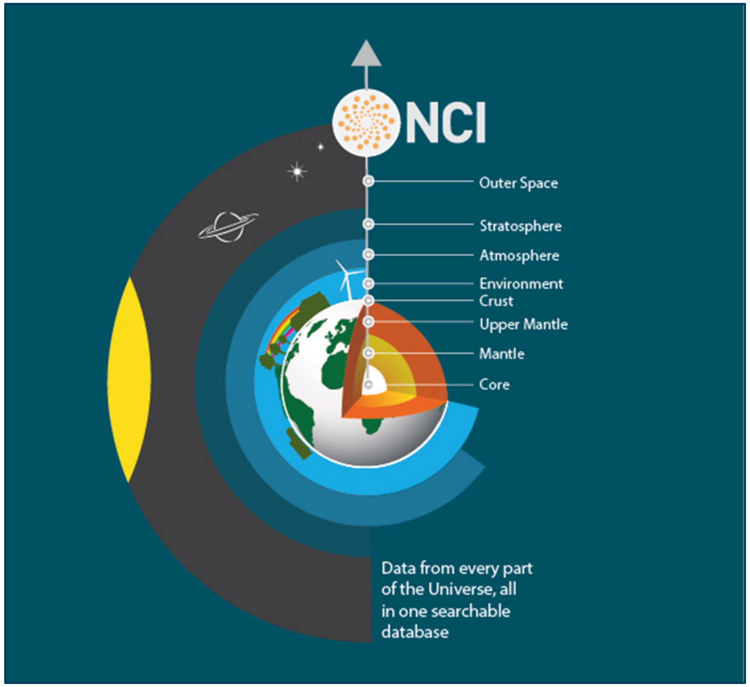
- **How big?**
  - Very.
  - Average data collection is 100+ Terabytes
  - Larger data collections are multi-Petabytes in size
  - Individual files can exceed 2TB or be as small as a few KB.
  - Individual datasets consist of tens of millions of files
  - Next Generation datasets are 6-10x larger.
  - Today:
    - 46.5PB Lustre, 30PB Tape
    - 1309 Million inodes stored on Lustre
    - 470GB/sec aggregate Lustre B/W across facility



NCI

Collection	TB Approved	TB Ready	Ingested
Skymapper (Astronomy)	210.00	210.00	100%
Australian Data Archive (Social Sciences)	4.00	4.00	100%
BPA Melanoma Dataset (Biosciences)	588.00	588.00	100%
Plant Phenomics (Biosciences)	2.00	2.00	100%
Ocean Gen. Circulation Model (Earth Simulator)	27.00	27.00	100%
Year Of Tropical Convection	89.00	89.00	100%
CABLE Global Evaluation Datasets	3.00	3.00	100%
CORDEX Int	2.00	2.00	100%
Coupled Model Intercomparison Project (CMIP5)	1.5PB →	1,487.00	100%
Reanalysis	[illegible]	207.00	100%
ACCESS Models	3.9PB →	3,896.00	100%
Seasonal Climate Prediction	595.00	595.00	100%
Australian Bathymetry and Elevation reference data	37.00	37.00	100%
Australian Marine Video and Imagery Collection	7.00	7.00	100%
Global Navigation Satellite System (GNSS) (Geodesy)	4.00	4.00	100%
Digitised Australian Aerial Survey Photography	68.00	68.00	100%
Earth Observation (Satellite: Landsat, etc)	1.4PB →	1,400.00	100%
IMOS+TERN Australasian Satellite Imagery	568.00	568.00	100%
Satellite Soil Moisture Products	3.00	3.00	100%
Synthetic Aperture Radar	121.00	121.00	100%
BoM Observations	377.00	377.00	100%
BoM Ocean-Marine Collections	287.00	287.00	100%
Aust. 3D Geological Models	1.00	1.00	100%
Aust. Geophysical Data Collection	10.00	10.00	100%
Aust. Natural Hazards Archive	3.00	3.00	100%
National CT-Lab Tomographic Collection	185.00	185.00	100%
TERN eMAST	48.00	48.00	100%
TERN Phenology Monitoring: Near Surface Remote Sen	1.00	1.00	100%
TERN eMAST Data Assimilation	30.00	30.00	100%
CSIRO/BoM Key Water Assets	20.00	20.00	100%
Models of Land/Water Dynamics from Space	16.00	16.00	100%
Totals	**10,296**	**10,296**	**100%**

- **What do we store?**
  - 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
      - (CMIP6 Planning in Progress)
    - Satellite imagery (Landsat, INSAR, ALOS)
    - Skymapper, Whole Sky Survey/ Pulsars
    - Australian Plant Phenomics Database
    - Australian Data Archive
    - EUMETSAT Copernicus Programme Sentinel Data (Sentinel 1,2 & 3)
  - Large Scale Genomics and Bioinformatics datasets



Storage Overview

Tier 1
HPC System, Job & Scratch Data

Highest Speed, Large Capacity Short Term Storage

/short – 7.6PB (Lustre)

Scratch and checkpointing storage for Active HPC jobs,
Typically stored for less than 90 days.

Tier 2
Active Project Data

High Speed , Bulk Capacity Long Term Storage

/g/data[1a,1b,2,3] – 39PB (Lustre)

Storage for Active Projects requiring online high speed
Storage for frequently used and reference data sets.

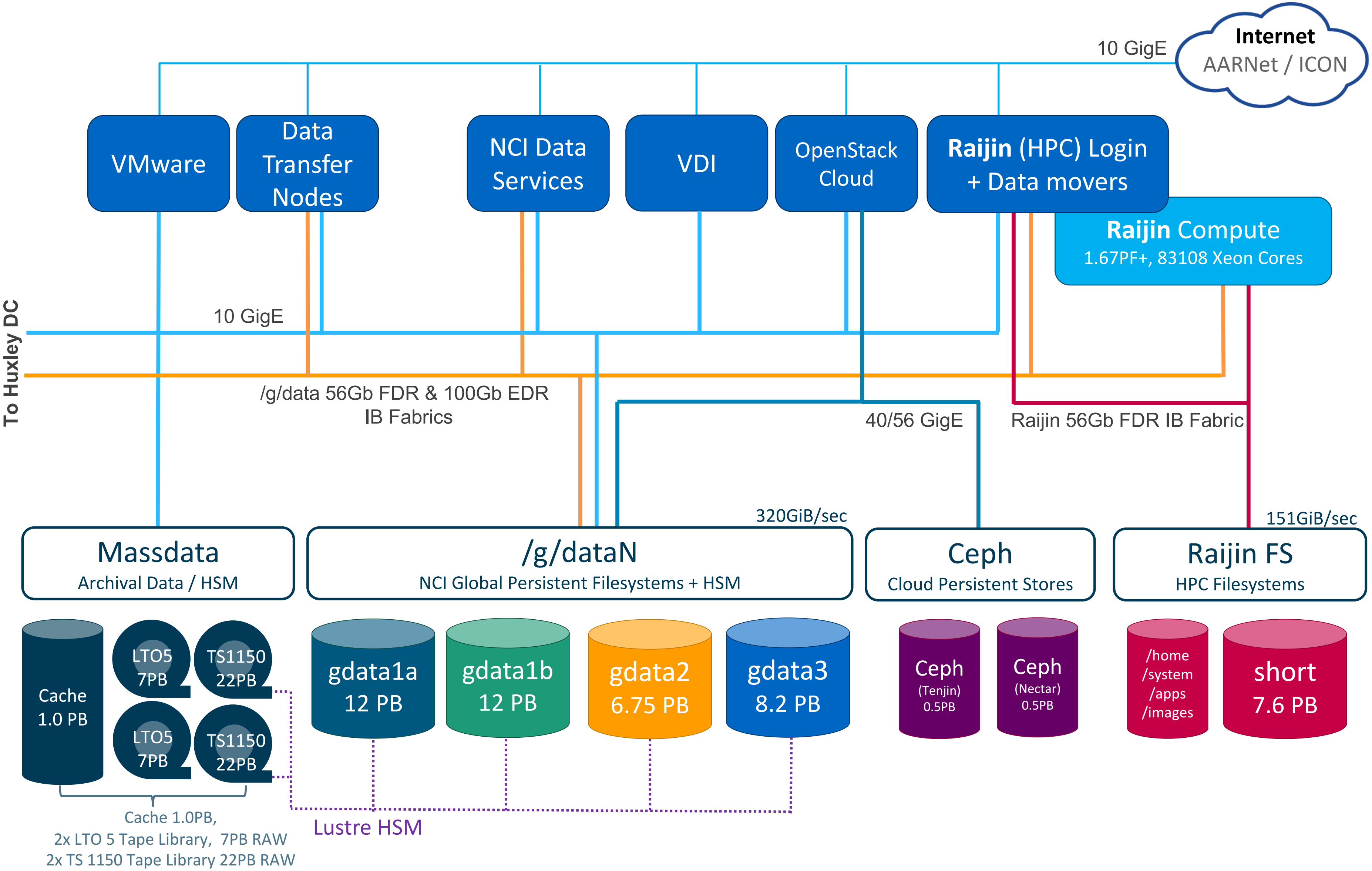
Tier 3
Archival Data

Lower Speed, Bulk Capacity Long Term Storage

/massdata – 30PB (DMF/CXFS)

Migrating Disk Cache <> Tape Archival Storage for
inactive projects or infrequently used data, dual site copy.

Systems Overview



Gdata1: 2013-2017

Gdata1 Replacement

– The Tale of Two Filesystems

- *It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way…*
  - *(The Tale of Two Cities, Charles Dickens)*

Alternative Presentation Names …

– Blue Pill, Red Pill.

- *You take the blue pill—the story ends, you wake up in your bed and believe whatever you want to believe. You take the red pill—you stay in Wonderland, and I show you how deep the rabbit hole goes. Remember: all I'm offering is the truth. Nothing more.*
  - *(The Matrix, 1999)*

Gdata1 Replacement

- **Gdata1: Nov 2013 – Sept 2017**
  - Ageing Infrastructure at End of Support Life – 30 Sept 2017
  - 7.4PB, 60GB/sec
  - 44 OSS, 2 MDS, 14 LNET Routers
  - Lustre 2.3.11 (IEEL 1)
  - 10x SGI IS4600 OST Arrays [Purchased 2011]
    - (LSI Engenio 7900 HPC) 480x 2TB SATA w/ FC interposer trays
    - RAID 6 (8+2) w/ hot spares
    - 6.4GB/sec Read, 5.4GB/sec Write
  - 2x SGI IS5500 OST Arrays (aka NetApp E5400)
    - 240x 3TB NL-SAS
  - 1x SGI IS5000 MDT Array
    - 40x 600G 15K SAS
  - 60x Dell R620 SandyBridge 1U Servers, FDR IB HCA + Switching
  - 14x 42U Racks of Equipment

New System Requirements

- Replacement System in operation by 30 September 2017
  - No further support renewal available on IS4600 storage hardware. Limited on site spares available
  - IS4600 Hardware originally purchased in Late 2010/Early 2011. Combined into one Lustre filesystem in 2013.

- At least 10-12PB Minimum Useable capacity, preferably 20PB+
  - Usage at gdata1 decomm: 92% used (7.4PB cap, 614TB free), 271M inodes

- Maintain or improve Performance & Stability
  - At least 60GB/sec performance
  - No regression from current 99.95+ % availability
  - Filesystem is home to 200+ projects
  - Modern Lustre release

- Minimal impact to user community
  - Least possible downtime
  - Consider impact to scripts and workflows

Nice-to-Have Features

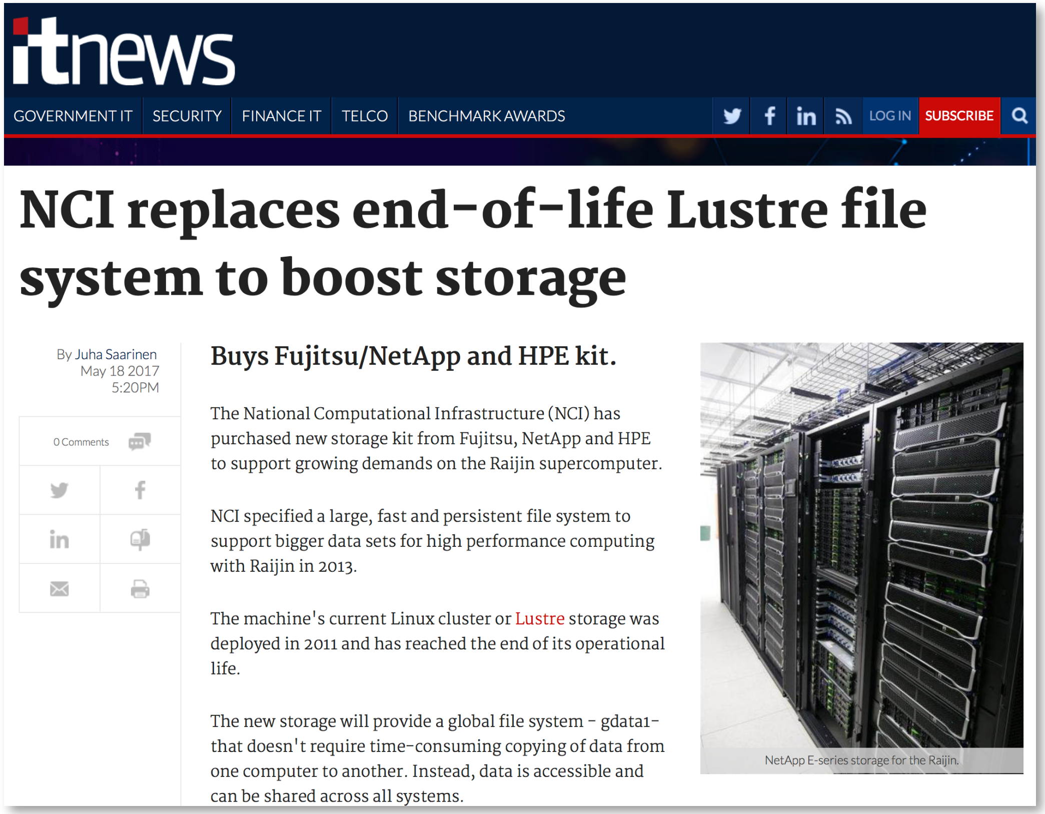
- Hardware
  - SSD based MDTs
  - No Fibre Channel Components – SAS or IB preferred

- Lustre
  - Metadata Performance improvement with DNE
  - Use new Lustre features where possible
  - Align with Lustre releases and ease future upgrade pain

- What about ZFS?
  - Snapshots, compression, consistency, scalability, lower cost JBOD hardware, write performance.
  - metadata performance? read performance?

Procurement

- University Procurement requirements. 18 Page RFP issued.
- Range of proposals received. Traditional RAID Controller + Server, Appliance Style Systems and JBOD Based storage solutions proposed.
- Two Systems selected within project budget.
- Selected on overall value, requirements alignment, capability and budget.



NCI replaces end-of-life Lustre file system to boost storage

By Juha Saarinen
May 18 2017
5:20PM

Buys Fujitsu/NetApp and HPE kit.

The National Computational Infrastructure (NCI) has purchased new storage kit from Fujitsu, NetApp and HPE to support growing demands on the Raijin supercomputer.

NCI specified a large, fast and persistent file system to support bigger data sets for high performance computing with Raijin in 2013.

The machine's current Linux cluster or Lustre storage was deployed in 2011 and has reached the end of its operational life.

The new storage will provide a global file system - gdata1- that doesn't require time-consuming copying of data from one computer to another. Instead, data is accessible and can be shared across all systems.

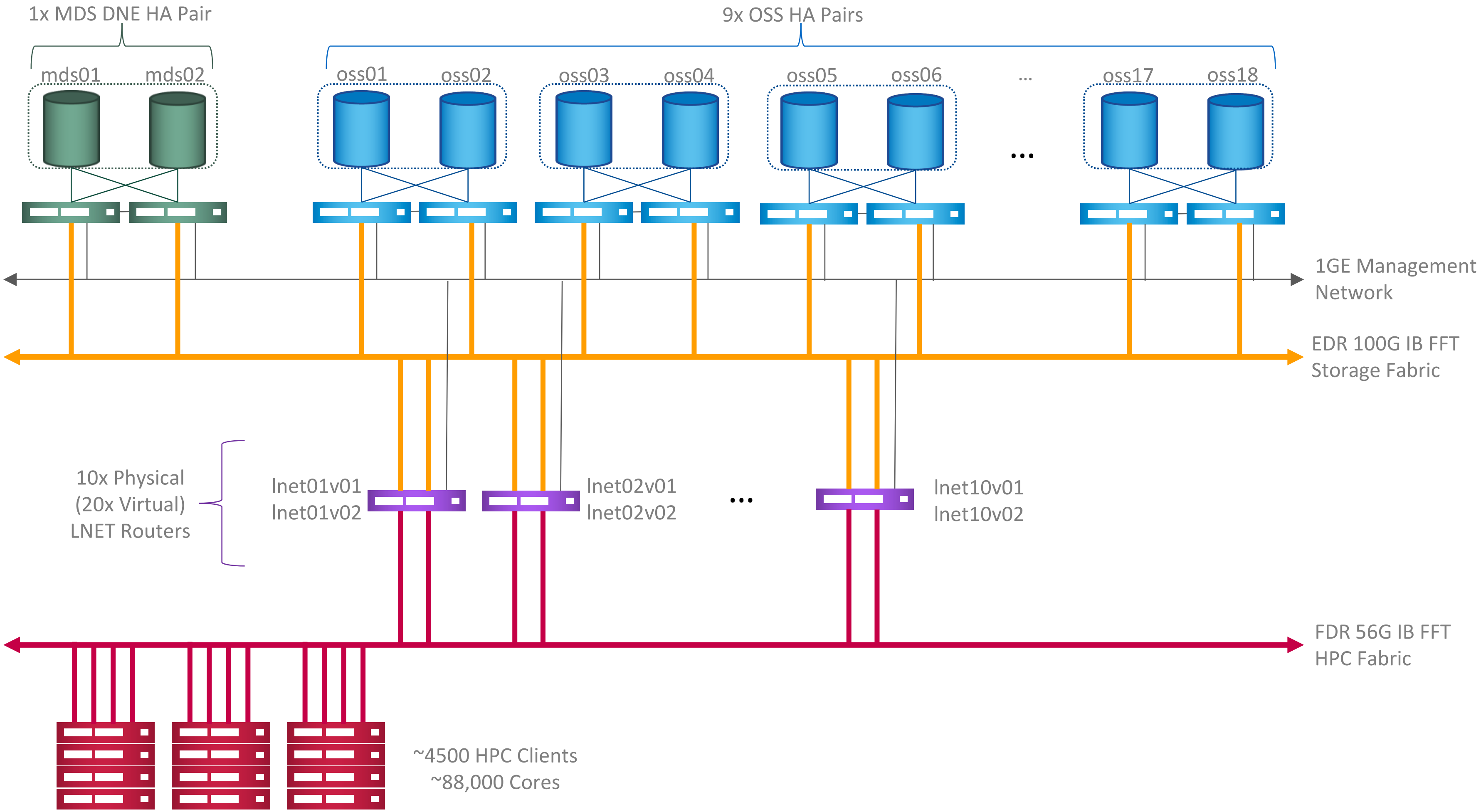
NetApp E-series storage for the Raijin.

Gdata1A

- **Gdata1A** (Fujitsu / NetApp)
  - Traditional LDISKFS, RAID Array Based Lustre Filesystem
  - 9x NetApp E5660 Disk Arrays (OST)
    - 180 Disk, 10TB NL-SAS configured as 8+2 RAID 6.
    - 18x OSTs per Array
    - 2x 4 Port 12G SAS Host Interface Card (SAS attached to OSS)
  - 18x Fujitsu RX2530-M2 (OSS)
    - 2x Xeon E5-2640v4 (Broadwell, 2.4Ghz, 10 core, 3.4Ghz TB)
    - 256GB DDR4
    - 1x ConnectX-4 VPI EDR 100Gbit
    - 2x LSI 9300-8e 12G SAS (4x 12G SAS connections to array)
  - 2x NetApp EF560 Disk Arrays (MDT)
    - 24 Disk, 800G SAS SSDs, 10 FDWPD
    - 8+2 RAID6 + LVM Mirror (20 Drives, 4 Hot spare)
    - 2x 2 Port FDR IB Host Interface Card (IB attached to MDS)
  - 2x Fujitsu RX2530-M2 (MDS)
    - 2x Xeon E5-2697v4 (Broadwell, 2.3Ghz, 18 core, 3.6Ghz TB)
    - 768GB DDR4
    - 2x ConnectX-4 VPI EDR 100Gbit Dual Port
    - 1x ConnectX-4 VPI FDR 56Gbit Single Port

Gdata1A - Overview

Gdata1A (Fujitsu / NetApp, RAID6 + LDISKFS)



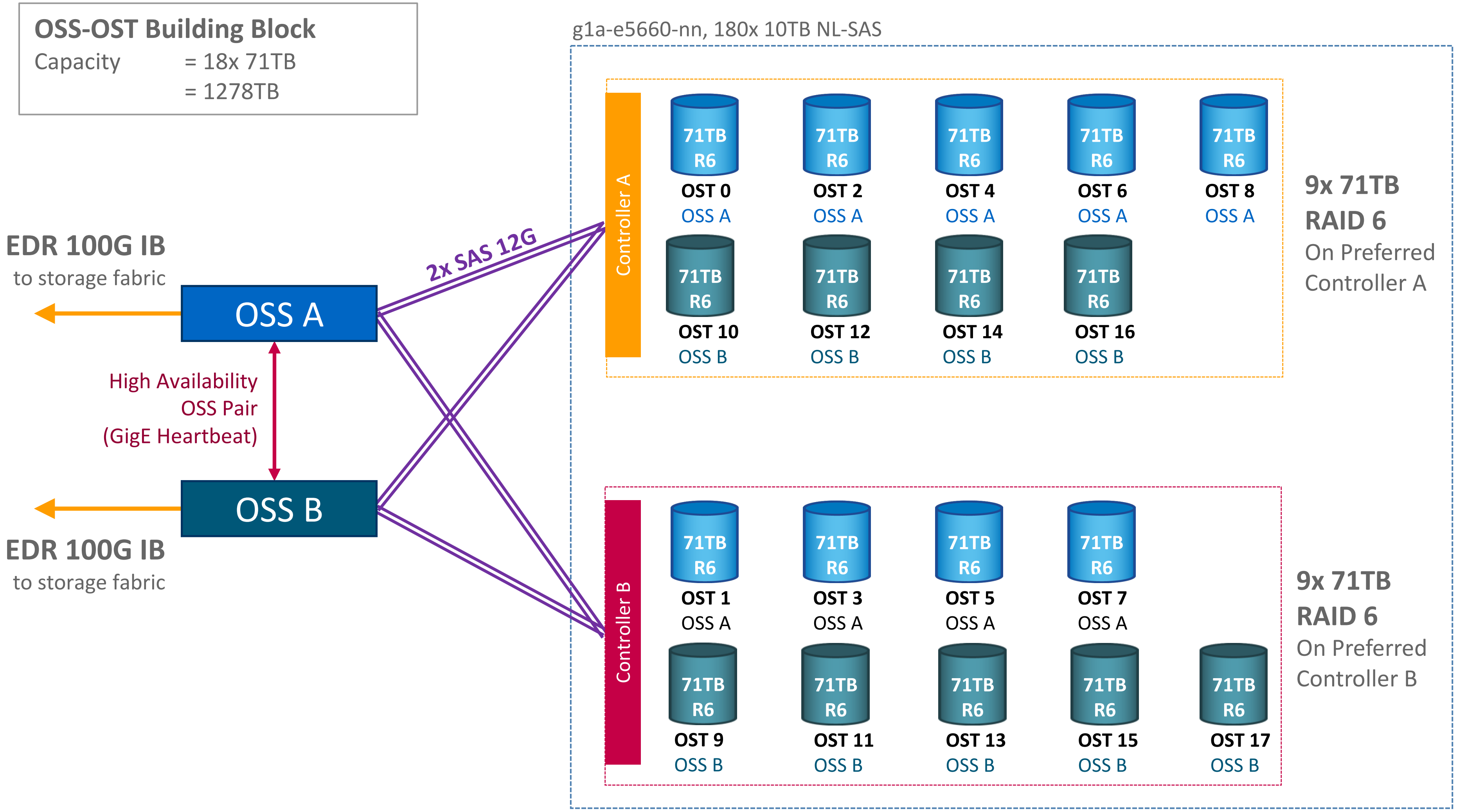
Gdata1A – OSS/OST Disk Layout

Gdata1A (Fujitsu / NetApp)

OSS-OST Building Block

Capacity = 18x 71TB

= 1278TB



g1a-e5660-nn, 180x 10TB NL-SAS

EDR 100G IB
to storage fabric

OSS A

High Availability
OSS Pair
(GigE Heartbeat)

OSS B

EDR 100G IB
to storage fabric

2x SAS 12G

Controller A

71TB R6 OST 0 OSS A

71TB R6 OST 2 OSS A

71TB R6 OST 4 OSS A

71TB R6 OST 6 OSS A

71TB R6 OST 8 OSS A

71TB R6 OST 10 OSS B

71TB R6 OST 12 OSS B

71TB R6 OST 14 OSS B

71TB R6 OST 16 OSS B

9x 71TB
RAID 6
On Preferred
Controller A

Controller B

71TB R6 OST 1 OSS A

71TB R6 OST 3 OSS A

71TB R6 OST 5 OSS A

71TB R6 OST 7 OSS A

71TB R6 OST 9 OSS B

71TB R6 OST 11 OSS B

71TB R6 OST 13 OSS B

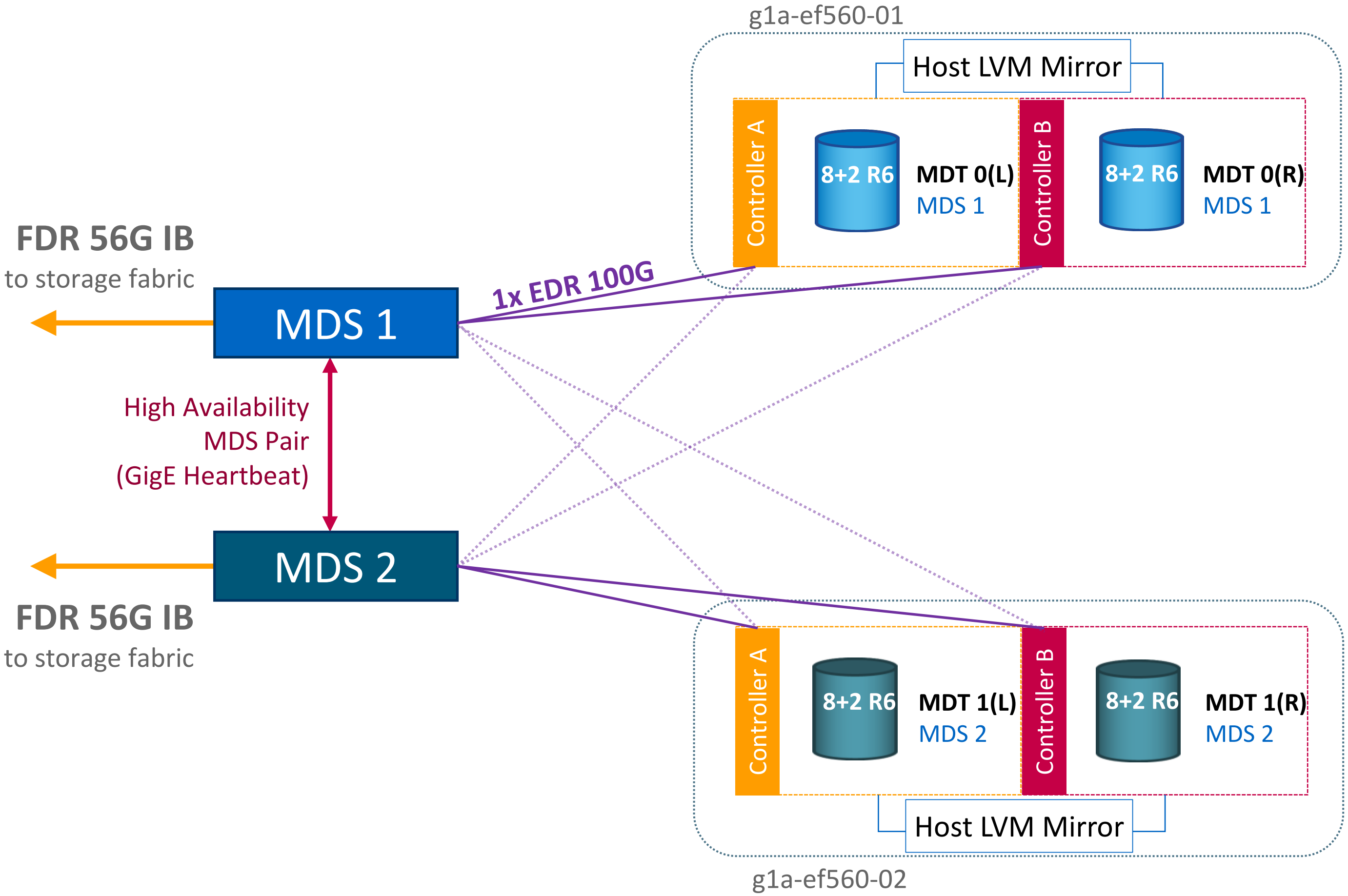
71TB R6 OST 15 OSS B

71TB R6 OST 17 OSS B

9x 71TB
RAID 6
On Preferred
Controller B

Gdata1A – MDS/MDT Disk Layout

Gdata1A (Fujitsu / NetApp)



g1a-ef560-01

Host LVM Mirror

Controller A

8+2 R6

MDT 0(L)
MDS 1

Controller B

8+2 R6

MDT 0(R)
MDS 1

FDR 56G IB
to storage fabric

MDS 1

1x EDR 100G

High Availability
MDS Pair
(GigE Heartbeat)

MDS 2

FDR 56G IB
to storage fabric

Controller A

8+2 R6

MDT 1(L)
MDS 2

Controller B

8+2 R6

MDT 1(R)
MDS 2

Host LVM Mirror

g1a-ef560-02

2x 5.9TB
RAID 6
1 Per Controller
LVM Host Mirror

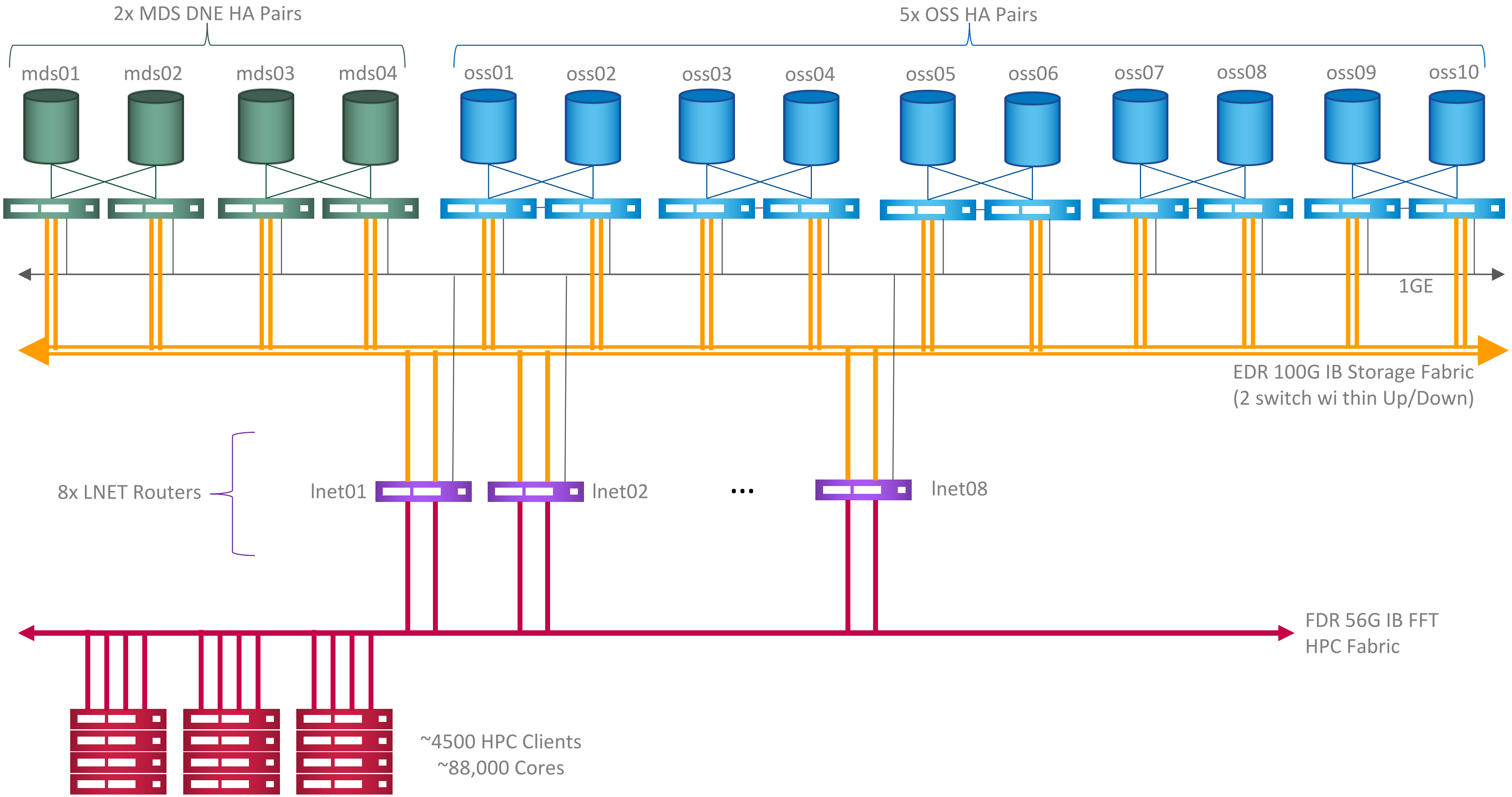
2x 5.9TB
RAID 6
1 Per Controller
LVM Host Mirror

Gdata1B

- **Gdata1B** (HPE)
  - ZFS, JBOD Based
  - 5x HPE Apollo 4520 Chassis, w/ 2x XL450 Nodes
    - 2x Xeon E5-2690v4 (Broadwell, 2.6Ghz, 14 core, 3.5Ghz TB)
    - 256GB DDR4
    - 1x ConnectX-4 VPI EDR 100Gbit Dual Port
    - 1x HP Smart Array H240/244br 12G SAS (internal)
    - 2x HP Smart Array P841 12G SAS (external)
    - 40x 8TB-NL SAS
  - 6x HPE D6020 JBODs per Apollo 4520 Chassis
    - 400x, 8TB NL-SAS
    - Raidz2 9d2p configured over 440 drives (with enclosure redundancy)
    - 440 drives total, 40x vdevs, 4x vdev per pool, 10x OSTs
  - 4x HPE MSA2040 (MDT)
    - 8x, 800G SAS SSDs, 10 FDWPD
    - 2x 4d RAID10 (2x RAID10 vdevs -> 1x zpool (mirror) -> 1x MDT
    - 2x 4 Port 12G SAS Controllers (SAS attached to MDS)
  - 4x HPE DL360G9 (MDS)
    - 2x Xeon E5-2697v4 (Broadwell, 2.3Ghz, 18 core, 3.6Ghz TB)
    - 768GB DDR4
    - 1x ConnectX-4 VPI EDR 100Gbit Dual Port
    - 2x HP Smart Array H241 12G SAS Adapters

Gdata1B - Overview

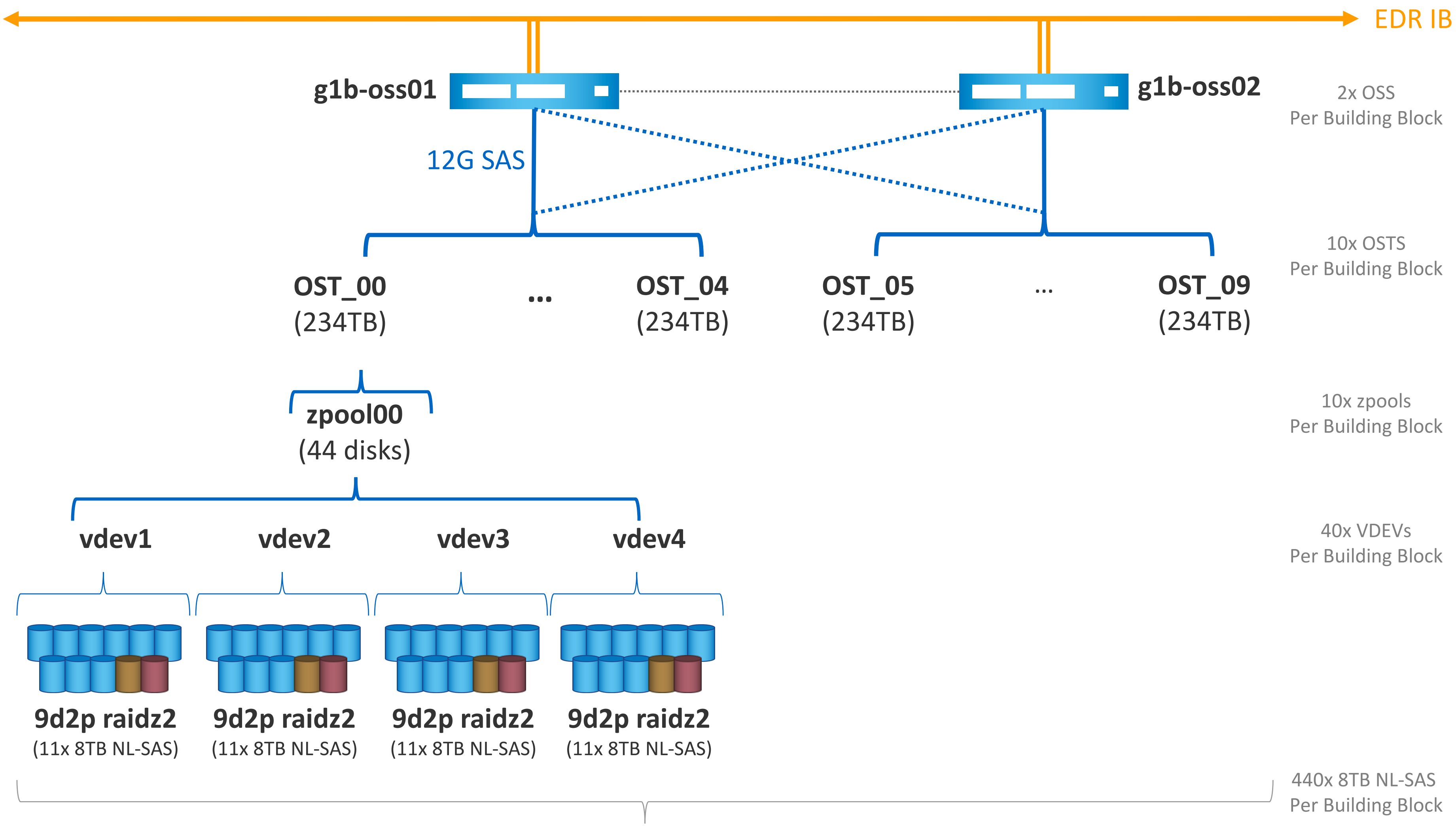
Gdata1B (HPE, JBOD+ZFS)



Gdata1B – OSS/OST Layout

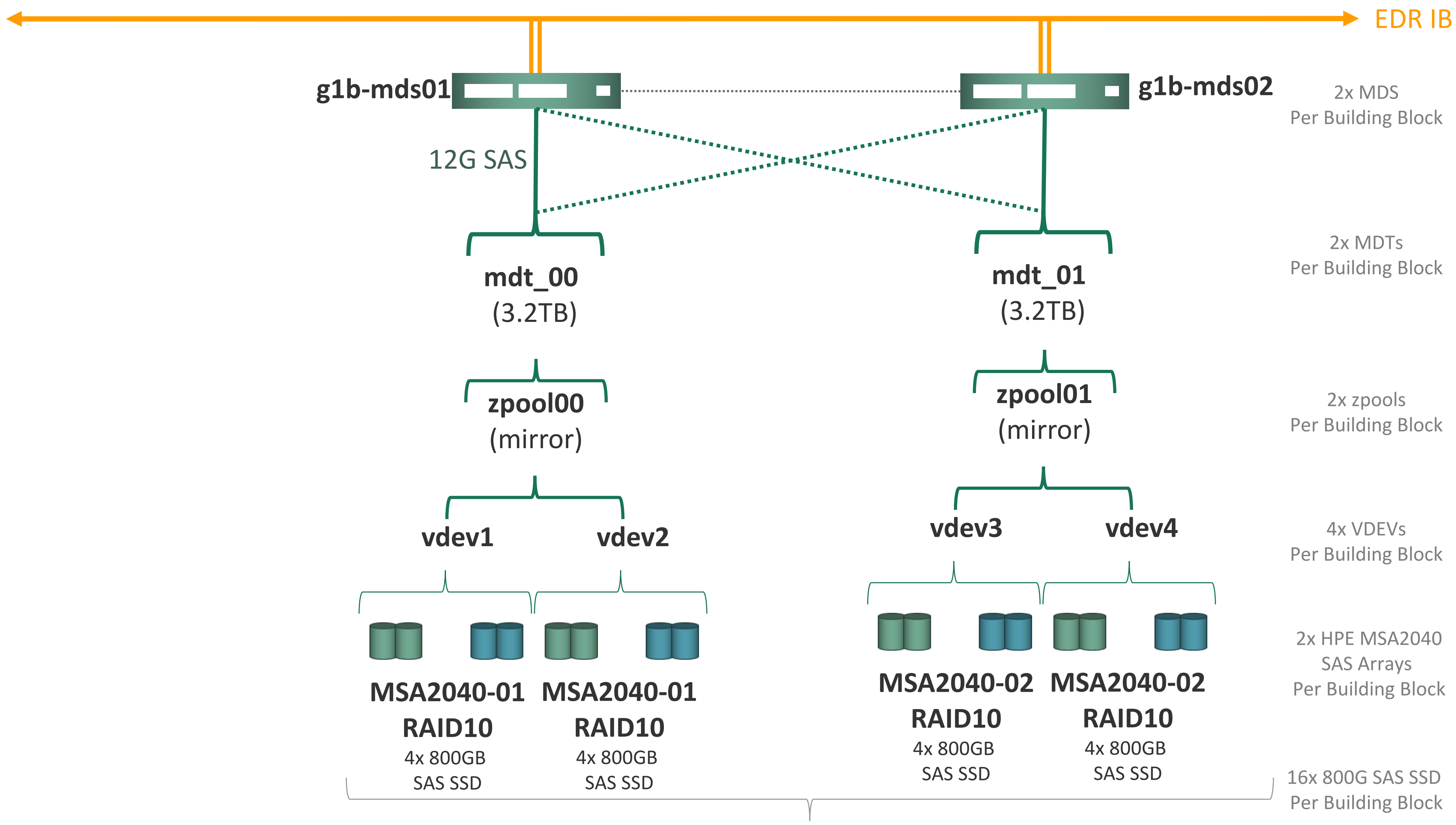
Gdata1B (HPE, JBOD+ZFS)

OSS-OST Building Block
Capacity = 10x 234TB
= 2340TB



Gdata1B – MDS/MDT Layout

Gdata1B (HPE, JBOD+ZFS)



Comparison Table

Property	Gdata1A	Gdata1B
BackingFS	LdiskFS, RAID6 (8+2)	ZFS, raidz2 (9d2p)
Lustre Useable Capacity	12PB	12PB
Lustre Version	IEEL 3.1.1 (Lustre 2.7.22/2.8)	2.10.4 LTS, ZFS 0.7.9
OSS	18	10
OST Count	162	50
OST size	70TB	234 TB
OST Drive, Count	10TB NL-SAS, 1620x	8TB NL-SAS, 2200x
MDS	2	4
MDT	2	4
RU, Rack Count	155RU, 4 Racks	200RU, 5 Racks
Multi-rail LNET	No	Yes
DNE	Yes, Phase 1	Yes, Phase 1
LNET Routers	10x Physical (20x Virtual)	8x Physical / Native with Multi-Rail
IB Topology	Full Fat Tree (1:1) EDR	2 Switch, Thin Up-Down, Bonded

Lustre Versions & Features

- Lustre Versions
  - Filesystems Built May-July 2017
  - Need to have at least 1 filesystem built and accepted by end of July to commence 8 week migration period, ahead of 30 Sept Deadline.
  - Lustre 2.10.0 LTS almost ready, but not yet released
  - Lustre 2.10.0 + ZFS 0.7.3 significant improvements, + CentOS 7.4 Compatibility

- gdata1A doesn't need ZFS improvements
  - gdata1A is "Plan A".
  - Stable, traditional ldiskfs + RAID6
  - Based on existing design at NCI, in use for 2+ years (gdata3)
  - Current vendor supported release is stable & mature
  - Gdata1A to be built on IEEL 3.1.1.0 (Lustre 2.7.21) [now DEEL? or WEEL?..]
  - Gdata1A best suited to hold majority of transferred projects based on access patterns
- gdata1B timeline slipped a few weeks to pickup 2.10.0 release
  - Upgraded to 2.10.2 before official production go-live

Lustre Features

- DNE
  - Both gdata1A and gdata1B use DNE Phase 1 (Remote Directory)
  - Projects are assigned a MDT on top level directory creation
  - # lfs mkdir –i <mdt_index> /g/data1a/projectid
  - It works!
    - Overall improvement in metadata performance
    - Isolates a badly behaving project to one MDT/MDS, other MDS/MDTs mostly unaffected
    - Failure of remote MDTs (index >0) filesystem still ok
    - Known and tested recovery mechanisms
    - Well worn code paths (DNE phase 1)

- Project Quotas
  - 2.7/2.8 – No support for Project Quotas
  - Upgrade to 2.10!
  - No – project quotas require params to be set at makefs. Cannot upgrade existing fs to get project quotas.

  - But what about gdata1B on 2.10?
    - Project quotas only for 2.10 + LDISKFS

Lustre Features

- Subdirectory Mounts, Nodelists
  - Want to use this for less trusted hosts
  - Attempted on gdata1B
  - See Tales of Woe slide
- Multi-Rail LNET
  - Partial implementation on gdata1B
  - Limitations hit in *LU-10153 - LNET route via two different networks not supported*

Gdata1A/1B

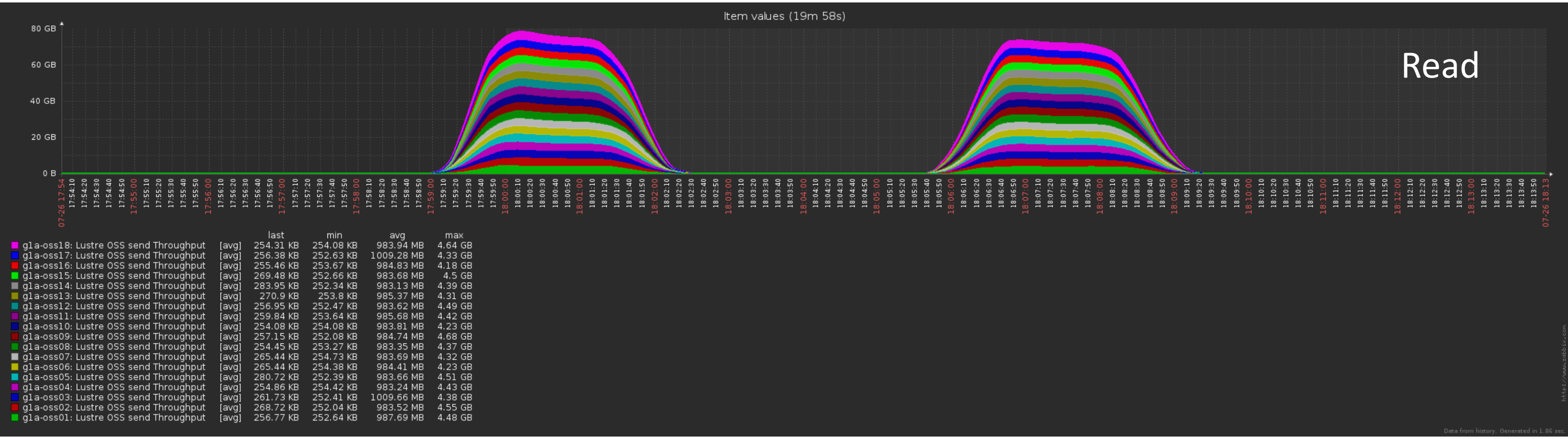
Acceptance Testing & Performance

Acceptance Tests

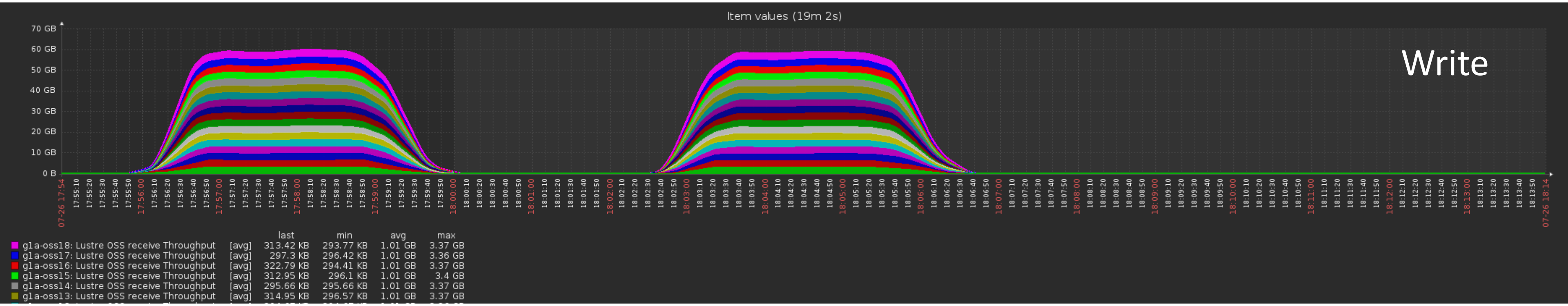
- **Acceptance Tests**
  - dd
    - Uniformity of devices / Volumes
  - IOR
    - 1MB Stream Read/Write, Multiple Pass, Pre-Allocated files
  - Mdtest
    - Directory Stat, File Stat
    - Directory Creation, File Creation
  - InfiniBand
    - ibdiagnet
    - ib_send_bw
    - Ib_send_lat
  - Power & Connectivity Loss
    - Partial Loss of Power (simulate UPS A or UPS B failure)
    - Total Loss of Power to All Components while under load
    - Recovery from Total Loss of Power
    - Storage Controller Failover
    - Loss of Host Link (storage to server)
    - Loss of Enclosure link (simulated SAS loop failure)
    - Drive Rebuild Deadline (4x failed drives within building block, separate volumes/zpools, less than 48hrs while under load)
  - 72 Hour Stability and Burn-in
    - 50% continual load on system, mix of Apps, Mdtest & IOR
    - Failure = start 72 hrs again

Performance

- **Gdata1A**
  - IOR Read: 69 GiB/sec (70,522 MB/sec)
  - IOR Write: 55 GiB/sec (56,552 MB/sec)
  - Likely bandwidth limited on Virtual LNET routers



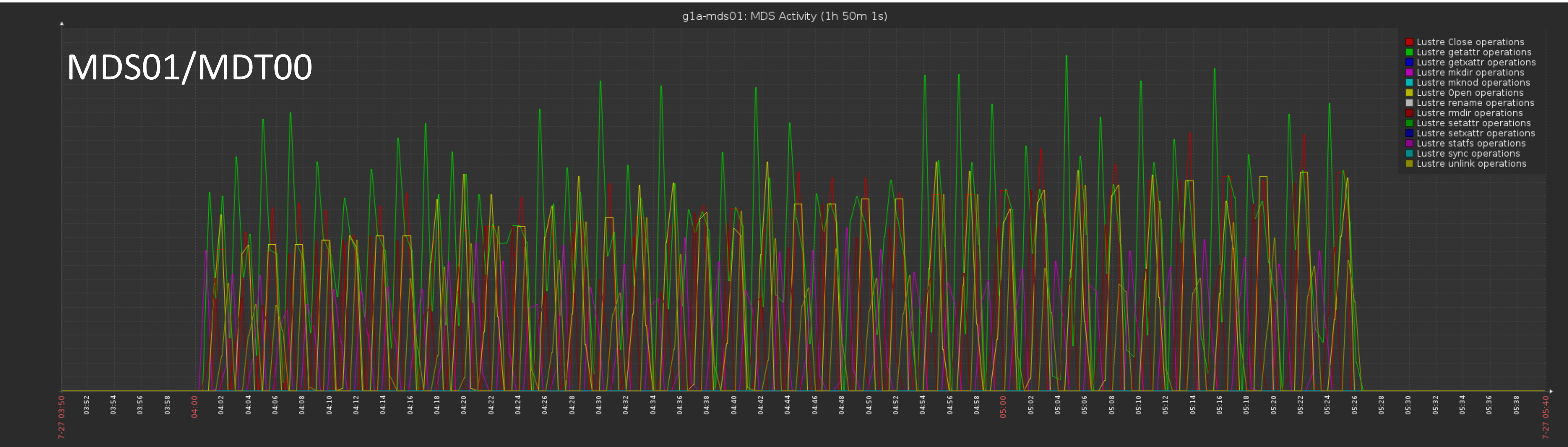
		last	min	avg	max
■ g1a-oss18: Lustre OSS send Throughput	[avg]	254.31 KB	254.08 KB	983.94 MB	4.64 GB
■ g1a-oss17: Lustre OSS send Throughput	[avg]	256.38 KB	252.63 KB	1009.28 MB	4.33 GB
■ g1a-oss16: Lustre OSS send Throughput	[avg]	255.46 KB	253.67 KB	984.83 MB	4.18 GB
■ g1a-oss15: Lustre OSS send Throughput	[avg]	269.48 KB	252.66 KB	983.68 MB	4.5 GB
■ g1a-oss14: Lustre OSS send Throughput	[avg]	283.95 KB	252.34 KB	983.13 MB	4.39 GB
■ g1a-oss13: Lustre OSS send Throughput	[avg]	270.9 KB	253.8 KB	985.37 MB	4.31 GB
■ g1a-oss12: Lustre OSS send Throughput	[avg]	256.95 KB	252.47 KB	983.62 MB	4.49 GB
■ g1a-oss11: Lustre OSS send Throughput	[avg]	259.84 KB	253.64 KB	985.68 MB	4.42 GB
■ g1a-oss10: Lustre OSS send Throughput	[avg]	254.08 KB	254.08 KB	983.81 MB	4.23 GB
■ g1a-oss09: Lustre OSS send Throughput	[avg]	257.15 KB	252.08 KB	984.74 MB	4.68 GB
■ g1a-oss08: Lustre OSS send Throughput	[avg]	254.45 KB	253.27 KB	983.35 MB	4.37 GB
■ g1a-oss07: Lustre OSS send Throughput	[avg]	265.44 KB	254.73 KB	983.69 MB	4.32 GB
■ g1a-oss06: Lustre OSS send Throughput	[avg]	265.44 KB	254.38 KB	984.41 MB	4.23 GB
■ g1a-oss05: Lustre OSS send Throughput	[avg]	280.72 KB	252.39 KB	983.66 MB	4.51 GB
■ g1a-oss04: Lustre OSS send Throughput	[avg]	254.86 KB	254.42 KB	983.24 MB	4.43 GB
■ g1a-oss03: Lustre OSS send Throughput	[avg]	261.73 KB	252.41 KB	1009.66 MB	4.38 GB
■ g1a-oss02: Lustre OSS send Throughput	[avg]	268.72 KB	252.04 KB	983.52 MB	4.55 GB
■ g1a-oss01: Lustre OSS send Throughput	[avg]	256.77 KB	252.64 KB	987.69 MB	4.48 GB

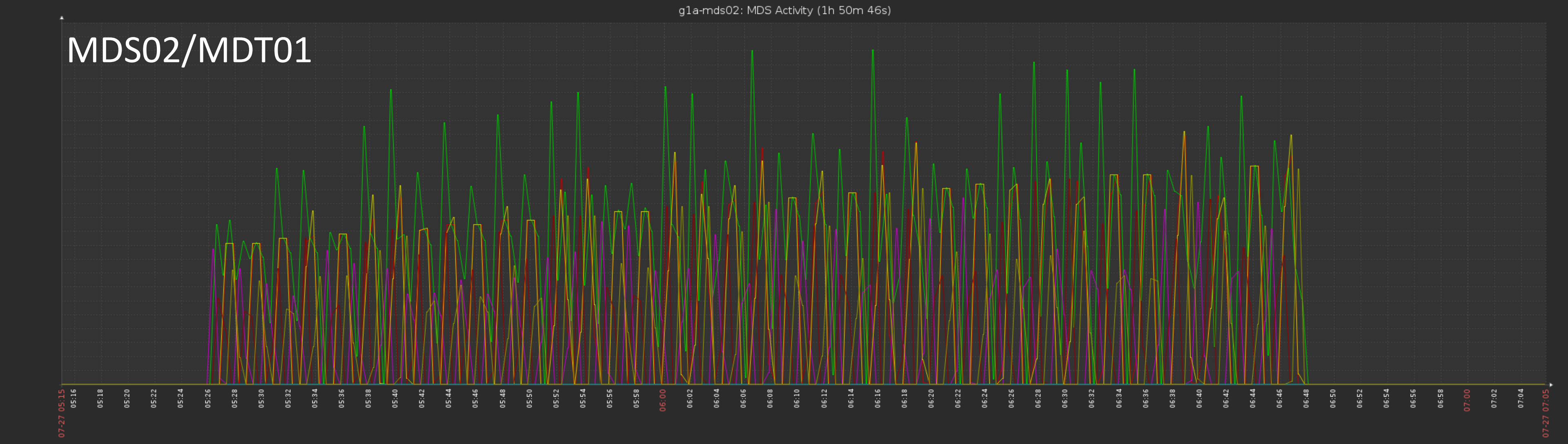


		last	min	avg	max
■ g1a-oss18: Lustre OSS receive Throughput	[avg]	313.42 KB	293.77 KB	1.01 GB	3.37 GB
■ g1a-oss17: Lustre OSS receive Throughput	[avg]	297.3 KB	296.42 KB	1.01 GB	3.36 GB
■ g1a-oss16: Lustre OSS receive Throughput	[avg]	322.79 KB	294.41 KB	1.01 GB	3.37 GB
■ g1a-oss15: Lustre OSS receive Throughput	[avg]	312.95 KB	296.1 KB	1.01 GB	3.4 GB
■ g1a-oss14: Lustre OSS receive Throughput	[avg]	295.66 KB	295.66 KB	1.01 GB	3.37 GB
■ g1a-oss13: Lustre OSS receive Throughput	[avg]	314.95 KB	296.57 KB	1.01 GB	3.37 GB

Performance

- **Gdata1A**
  - **MDtest, simultaneous run on MDT0 & 1. Scaling # of worker tasks.**





Performance

- **Gdata1B**
  - IOR Read: 63 GiB/sec (64,937 MB/sec)
  - IOR Write: 70 GiB/sec (71,586 MB/sec)

  - MDtest (4x MDT, DNE-1, lfs mkdir -i n /path/)

```
mdtest-1.9.4-rc1 was launched with 120 total task(s) on 30 node(s)

Command line used: /system/Benchmarks/Filesystem/mdtest/mdtest_ompi211 -u -d
/g/data1b/fu2/mdtest/mdt0_5@/g/data1b/fu2/mdtest/mdt1_5@/g/data1b/fu2/mdtest/mdt2_5@/g/data1b/f
u2/mdtest/mdt3_5 -n 32768 -i 2

Path: /g/data1b/fu2/mdtest
FS: 11714.4 TiB   Used FS: 4.5%   Inodes: 194.4 Mi   Used Inodes: 2.5%

120 tasks, 3932160 files/directories

SUMMARY: (of 2 iterations)
   Operation                      Max            Min           Mean        Std Dev
   ---------                      ---            ---           ----        -------
   Directory creation:      58523.642      41748.429      50136.036       8387.607
   Directory stat    :     378850.072     352877.543     365863.807      12986.264
   Directory removal :     156167.973     134823.142     145495.558      10672.416
   File creation     :     125104.777     122010.634     123557.705       1547.07
   File stat         :     304980.603     296501.890     300741.247       4239.356
   File read         :     314153.143     305879.803     310016.473       4136.670
   File removal      :     202218.173     197868.384     200043.278       2174.895
   Tree creation     :        114.098          4.485         59.292         54.806
   Tree removal      :         70.908         43.207         57.058         13.851
```

7PB Data, 271M inodes, 200+ Projects

Data Migration

Data Migration

- Need to move data off old gdata1 by 30 September 2017
  - ~7PB Data, 271 Million inodes, 200+ Projects
  - Maximum downtime allowable per project = 24 hours
  - 8 Weeks between Filesystem acceptance signoff and 30 Sept

- Tool
  - dcp2
  - LAD’16: Petascale Data Migration between Site-Wide Filesystems
    - https://www.eofs.eu/_media/events/lad16/05_petascale_data_migration_rodwell.pdf

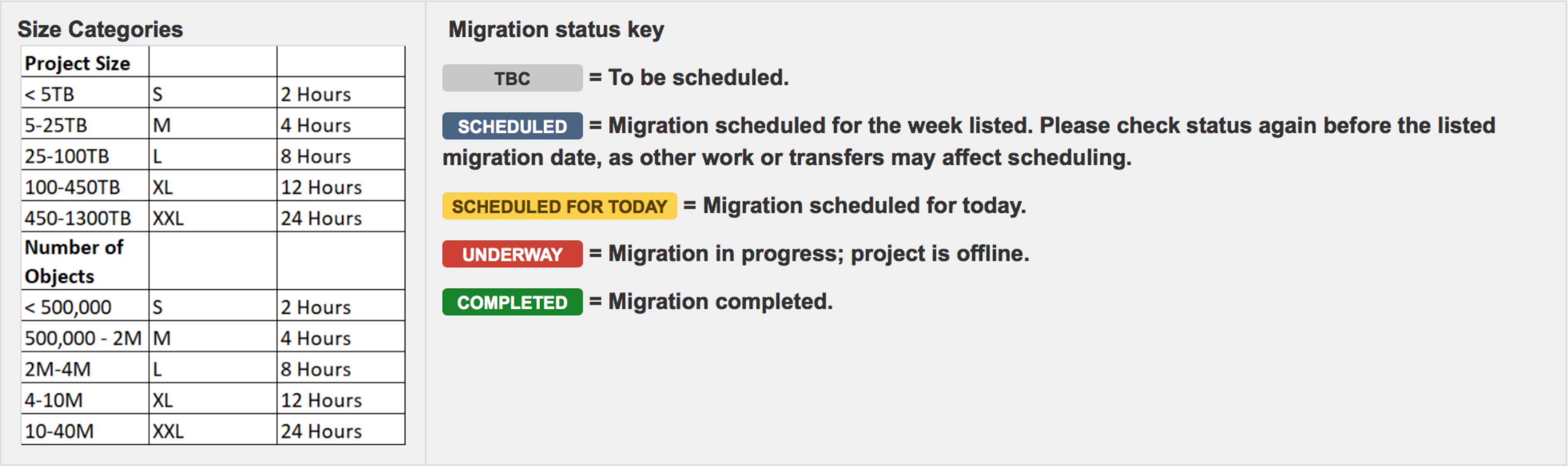
- Use Symlinks to avoid broken Job Submission Scripts
  - July 2017 Scheduled Maintenance: remount /g/data1 at /g/data1legacy
  - Mount new gdata1A filesystem: /g/data1a
  - Create symlink for each project after maintenance remount
  - Update symlink to repoint project to gdata1a

```
[dr@raijin data1]$ pwd; ls -l
/g/data1
lrwxrwxrwx 1 root root 18 Jul 14  2017 z80 -> /g/data1legacy/z80
lrwxrwxrwx 1 root root 13 Aug 24  2017 z81 -> /g/data1a/z81
```

Data Migration

– Method

- Use previous dcp2 migration logs to determine 4 classes of project size, based on inodes and size in bytes.
- Determine number of available nodes to run dcp on, and available “dcp hours” per day, while conscious of not stalling filesystem due to migration bandwidth load.
- Advertise migration schedule in advance, co-ordinate key dates with critical projects.
- Eg. Any calendar day may have 1x XXL migration or 10x S project
- All projects migrated in <6 Weeks, no weekends, minimal after hours
- No data corruption, user errors reported , or data lost
- Old copy of project held in root only directory on gdata1 until 2 January 2018.



Size Categories

Project Size		
< 5TB	S	2 Hours
5-25TB	M	4 Hours
25-100TB	L	8 Hours
100-450TB	XL	12 Hours
450-1300TB	XXL	24 Hours
Number of Objects		
< 500,000	S	2 Hours
500,000 - 2M	M	4 Hours
2M-4M	L	8 Hours
4-10M	XL	12 Hours
10-40M	XXL	24 Hours

Migration status key

TBC = To be scheduled.

SCHEDULED = Migration scheduled for the week listed. Please check status again before the listed migration date, as other work or transfers may affect scheduling.

SCHEDULED FOR TODAY = Migration scheduled for today.

UNDERWAY = Migration in progress; project is offline.

COMPLETED = Migration completed.

Gdata1A/1B

Reflection

Gdata1A Review

- ## 1A - A *few* things didn't go to plan
  - Hardware
    - Wrong LSI SAS Adapters shipped in hosts, maximum 16 targets visible (need up to 76).
    - 'lsscsi' command on Linux should see 72 NetApp standard targets (~70TB each), being unique 18 targets over 4 connections - 18x4=72). Additionally 4x NetApp Universal Export (Access LUN) target presentations should also be visible, for a total of 76 targets.
    - Easy fix, swap LSI/Avago MegaRAID SAS 9380-8e for LSI/Avago 9300-8e SAS HBAs.
    - Triggered a rare bug in Santricity 8.30.20 on E-series controllers - defect#LSIP200951958. Cause identified, solution in place.
  - Lustre
    - **LU-6602** ASSERTION(rec->lrh_len <= 8192) failed
    - **LU-6886** declare changelog store for POSIX ACLs in mdd_xattr_del
    - **LU-9740** Most of OSTs remounted read-only due to abort transaction in __ldiskfs_handle_dirty_metadata

- **1B- A *few more* things didn't go to plan**
  - Hardware
    - D6020 JBOD – Drive activity/ID lights inactive (v2.09 D6020 IOM FW)
    - D6020 JBOD - Drive activity/ID lights work, but OOM on boot (v.1.63 D6020 FW)
      - OOM 256GB RAM on Boot.
      - Each host typically home to 200x D6020 drives, 4x paths = 800 devices to enumerate
      - Prefer no OOM over no LEDs
    - HPSA 3.4.20-125 – HPSA driver can silently block IO with bad drive / SMART Errors
    - D6020 Firmware update speed – 30x mins per IOM
      - Each 4520+JBOD system has 6x D6020, 4x IOM per 6020. 24 IOMs, 30 Mins each, Sequential update process.
  - Linux SES
    - Long Boot & Device Discovery times due to repeated diagnostic page requests on systems with large device counts
    - https://patchwork.kernel.org/patch/10056895/ (credit: Dongyang Li)
  - ZFS
    - Lustre ZFS MDT softlockups under heavy Metadata Workload
    - ZFS (0.7.3, Fixed 0.7.6, included in 0.7.9/2.10.4):
    - https://github.com/zfsonlinux/zfs/pull/6986 (credit: Dongyang Li)

Gdata1B Review

- **1B- A *few more* things didn’t go to plan**
  - Lustre - Bugs hit and resolved in 2.10.4
    - **LU-10463** - Poor write performance periodically on repeated test runs
    - **LU-10460** - Poor fsync Performance (LLNL Patches)
    - **LU-11024** - Broken inode accounting of MDT on ZFS
    - **LU-10703** - All mds Nodemap filset will be cleared when do some nodemap operations (lustre 2.10.3)
  - Lustre – Bugs hit, unresolved on gdata1B, but fix available
    - **LU-10680** - MDT becoming unresponsive in 2.10.3 (Landed 2.12)
    - **LU-10124** - lnetctl: lnetctl import --add not importing peers correctly (Fix maybe 2.12)
  - Lustre – Unresolved
    - **LU-10635** - MGS kernel panic when configuring nodemaps and filesets
      - **LU-10390** MGS crashes in ldlm_reprocess_queue() when stopping (MGS unmount panic)
      - **LU-9838** registration mount fails with -108 but then succeeds if retried immediately (Client mount fails)
    - **LU-10153** - LNET route via two different networks not supported (use IB bonding workaround)
    - **LU-9704** - ofd_grant_check() claims GRANT, real grant 0 (dirty page discards on DTNs)

- **Final Thoughts**
  - Building 2x completely different 12PB filesystems simultaneously is hard.
  - Building a MajorVersion.0 System provides a source of daily surprise, entertainment & stress
  - Finding broken/failed/bad/pathologically-bad drives in a JBOD system still can be challenging
    - (NCI internal how-to guide is now at revision 4.)
  - What worked well?
    - Building 2 different systems simultaneously can provide risk management if one falters or is seriously delayed.
    - The gdata1A / gdata1B project has allowed us to maintain reliability, manage risk and hit project deadlines, while still being able to understand if ZFS based Lustre works for us in the longer term
    - Both systems are currently stable, and achieving ~99.98% uptime
    - Have a team of people who have prior experience with JBOD systems + ZFS
  - What would I do differently?
    - Use far fewer drives behind each OSS in ZFS
    - Don’t forget to order new 1GE management switches with your hardware!

Future

- **Future Activities**
  - Upgrade gdata1A to 2.10 LTS or 2.12 (LTS?)
  - Test ZFS compression and performance impact
  - Subdirectory mounts, UID Mapping & Kerberos revisit
  - Further testing on ZFS snapshots
  - Remove IB Bonds, full native multi-rail (dependent on LU-10153)
  - Convert Virtual LNET routers on gdata1A to native Multi-Rail
  - Upgrade / Configure Progressive File Layouts

NCI

Providing Australian researchers with
world-class computing services

NCI Contacts
General enquiries: +61 2 6125 9800
Media enquiries: +61 2 6125 4389
Support: help@nci.org.au

Address:
NCI, Building 143, Ward Road
The Australian National University Canberra ACT
2601
Australia

NCRIS
National Research
Infrastructure for Australia
An Australian Government Initiative

Australian Government
Bureau of Meteorology

Australian Government
Geoscience Australia

Australian Government
Australian Research Council

CSIRO

Australian
National
University

W nci.org.au

@NCInews