Lustre Admins & Developers Workshop 2018

ZFS or Idiskfs?
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 FS or Zed FS?

• For this presentation, I may use both interchangeably.

Oxford Dictionary of English

Z ¹ | zɛd | /zɛd/ | US zɛː (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 X¹).]
   • (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 zêta (see ZETA).

zee | zɛː |

noun North American
the letter Z.

Origin
late 17th century: variant of ZED.
Intro to NCI

National High Performance Compute + Data Capability
• 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)
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
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

To Huxley DC

- VMware
- Data Transfer Nodes
- NCI Data Services
- VDI
- OpenStack Cloud
- Raijin (HPC) Login + Data movers

/g/data 56Gb FDR & 100Gb EDR IB Fabrics

- Internet AARNet / ICON

10 GigE

Raijin Compute
1.67PF+, 83108 Xeon Cores

Cache 1.0PB, 2x LTO 5 Tape Library, 7PB RAW
2x TS 1150 Tape Library 22PB RAW

Massdata
Archival Data / HSM

/g/dataN
NCI Global Persistent Filesystems + HSM

Ceph
Cloud Persistent Stores

Ceph (Tenjin) 0.5PB
Ceph (Nectar) 0.5PB
/home/system/apps/images

Raijin FS
HPC Filesystems

short
7.6 PB

Raijin FS
HPC Filesystems

Cache 1.0PB
TS1150 22PB

/LTO 5 7PB

/g/data1a
12 PB

/g/data1b
12 PB

/g/data2
6.75 PB

/g/data3
8.2 PB

Lustre HSM

320GiB/sec

151GiB/sec

300GiB/sec

40/56 GigE

Raijin 56Gb FDR IB Fabric

10 GigE

To Huxley DC

Internet AARNet / ICON
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)*
— 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: 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?
– 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.
• **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 (Fujitsu / NetApp, RAID6 + LDISKFS)**

- **1x MDS DNE HA Pair**: mds01, mds02
- **9x OSS HA Pairs**: oss01, oss02, oss03, oss04, oss05, oss06, ..., oss17, oss18
- **10x Physical (20x Virtual) LNET Routers**: Inet01v01, Inet01v02, Inet02v01, Inet02v02, ..., Inet10v01, Inet10v02
- **~4500 HPC Clients ~88,000 Cores**
Gdata1A – OSS/OST Disk Layout

**Gdata1A (Fujitsu / NetApp)**

**OSS-OST Building Block**
- Capacity = 18x 71TB
- = 1278TB

**High Availability OSS Pair**
- (GigE Heartbeat)

**EDR 100G IB**
- to storage fabric

**EDR 100G IB**
- to storage fabric

**Controller A**
- 2x SAS 12G
- 9x 71TB RAID 6
- On Preferred Controller A

**Controller B**
- 9x 71TB RAID 6
- On Preferred Controller B

**g1a-e5660-nn, 180x 10TB NL-SAS**
**Gdata1A** (Fujitsu / NetApp)

- **FDR 56G IB** to storage fabric
- **High Availability MDS Pair** (GigE Heartbeat)
- **2x 5.9TB RAID 6**
  - 1 Per Controller
  - LVM Host Mirror

### MDS Layout

- **MDS 1**
  - 8+2 R6 MDT 0(L) MDS 1
  - Controller A
  - 1x EDR 100G
- **MDS 2**
  - 8+2 R6 MDT 0(R) MDS 1
  - Controller B

- **Host LVM Mirror**
- **g1a-ef560-01**
- **g1a-ef560-02**

- **2x 5.9TB RAID 6**
  - 1 Per Controller
  - LVM Host Mirror
• **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)

2x MDS DNE HA Pairs

5x OSS HA Pairs

8x LNET Routers

~4500 HPC Clients
~88,000 Cores
Gdata1B – OSS/OST Layout

Gdata1B (HPE, JBOD+ZFS)

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

EDR IB

2x OSS Per Building Block
10x OSTs Per Building Block
10x zpools Per Building Block
40x VDEVs Per Building Block
440x 8TB NL-SAS Per Building Block

OST_00 (234TB)
OST_01 (234TB)
OST_02 (234TB)
OST_03 (234TB)
OST_04 (234TB)
OST_05 (234TB)
OST_06 (234TB)
OST_07 (234TB)
OST_08 (234TB)
OST_09 (234TB)

vdev1
vdev2
vdev3
vdev4

9d2p raidz2 (11x 8TB NL-SAS)
9d2p raidz2 (11x 8TB NL-SAS)
9d2p raidz2 (11x 8TB NL-SAS)
9d2p raidz2 (11x 8TB NL-SAS)

zpool00 (44 disks)

12G SAS
Gdata1B – MDS/MDT Layout

Gdata1B (HPE, JBOD+ZFS)

EDR IB

12G SAS

msa2040-01
RAID10
4x 800GB SAS SSD

MSA2040-02
RAID10
4x 800GB SAS SSD

2x MDS
Per Building Block

2x MDTs
Per Building Block

2x zpools
Per Building Block

4x VDEVs
Per Building Block

2x HPE MSA2040 SAS Arrays
Per Building Block

16x 800G SAS SSD
Per Building Block
<table>
<thead>
<tr>
<th>Property</th>
<th>Gdata1A</th>
<th>Gdata1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>BackingFS</td>
<td>LdiskFS, RAID6 (8+2)</td>
<td>ZFS, raidz2 (9d2p)</td>
</tr>
<tr>
<td>Lustre Useable Capacity</td>
<td>12PB</td>
<td>12PB</td>
</tr>
<tr>
<td>Lustre Version</td>
<td>IEEL 3.1.1 (Lustre 2.7.22/2.8)</td>
<td>2.10.4 LTS, ZFS 0.7.9</td>
</tr>
<tr>
<td>OSS</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>OST Count</td>
<td>162</td>
<td>50</td>
</tr>
<tr>
<td>OST size</td>
<td>70TB</td>
<td>234 TB</td>
</tr>
<tr>
<td>OST Drive, Count</td>
<td>10TB NL-SAS, 1620x</td>
<td>8TB NL-SAS, 2200x</td>
</tr>
<tr>
<td>MDS</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>MDT</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>RU, Rack Count</td>
<td>155RU, 4 Racks</td>
<td>200RU, 5 Racks</td>
</tr>
<tr>
<td>Multi-rail LNET</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>DNE</td>
<td>Yes, Phase 1</td>
<td>Yes, Phase 1</td>
</tr>
<tr>
<td>LNET Routers</td>
<td>10x Physical (20x Virtual)</td>
<td>8x Physical / Native with Multi-Rail</td>
</tr>
<tr>
<td>IB Topology</td>
<td>Full Fat Tree (1:1) EDR</td>
<td>2 Switch, Thin Up-Down, Bonded</td>
</tr>
</tbody>
</table>
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
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
- **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*
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
- **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
• **Gdata1A**
  – MDtest, simultaneous run on MDT0 & 1. Scaling # of worker tasks.
**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/fu2/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)**

<table>
<thead>
<tr>
<th>Operation</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directory creation</td>
<td>58523.642</td>
<td>41748.429</td>
<td>50136.036</td>
<td>8387.607</td>
</tr>
<tr>
<td>Directory stat</td>
<td>378850.072</td>
<td>352877.543</td>
<td>365863.807</td>
<td>12986.264</td>
</tr>
<tr>
<td>Directory removal</td>
<td>156167.973</td>
<td>134823.142</td>
<td>145495.558</td>
<td>10672.416</td>
</tr>
<tr>
<td>File creation</td>
<td>125104.777</td>
<td>122010.634</td>
<td>123557.705</td>
<td>1547.07</td>
</tr>
<tr>
<td>File stat</td>
<td>304980.603</td>
<td>296501.890</td>
<td>300741.247</td>
<td>4239.356</td>
</tr>
<tr>
<td>File read</td>
<td>314153.143</td>
<td>305879.803</td>
<td>310016.473</td>
<td>4136.670</td>
</tr>
<tr>
<td>File removal</td>
<td>202218.173</td>
<td>197868.384</td>
<td>200043.278</td>
<td>2174.895</td>
</tr>
<tr>
<td>Tree creation</td>
<td>114.098</td>
<td>4.485</td>
<td>59.292</td>
<td>54.806</td>
</tr>
<tr>
<td>Tree removal</td>
<td>70.908</td>
<td>43.207</td>
<td>57.058</td>
<td>13.851</td>
</tr>
</tbody>
</table>
7PB Data, 271M inodes, 200+ Projects

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](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

```bash
[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
```
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

<table>
<thead>
<tr>
<th>Project Size</th>
<th>S</th>
<th>M</th>
<th>L</th>
<th>XL</th>
<th>XXL</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 5TB</td>
<td>2 Hours</td>
<td>4 Hours</td>
<td>8 Hours</td>
<td>12 Hours</td>
<td>24 Hours</td>
</tr>
<tr>
<td>5-25TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-100TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100-450TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>450-1300TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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.
Reflection

Gdata1A/1B
• **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).
  • 'Iscsich' 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 __Idiskfs_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)
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 - Inetctl: Inetctl 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 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