NCI

Providing Australian researchers with world-class computing services

LAD'14 - Lustre HSM

Daniel Rodwell
Manager, Data Storage Services

September 2014

- **What is NCI**
- **Petascale HPC at NCI (Raijin)**
- **Storage at NCI**
  - Lustre Filesystems
  - Other
- **Lustre HSM Project**
  - History & Requirements
  - Design
  - Sizing
  - Challenges

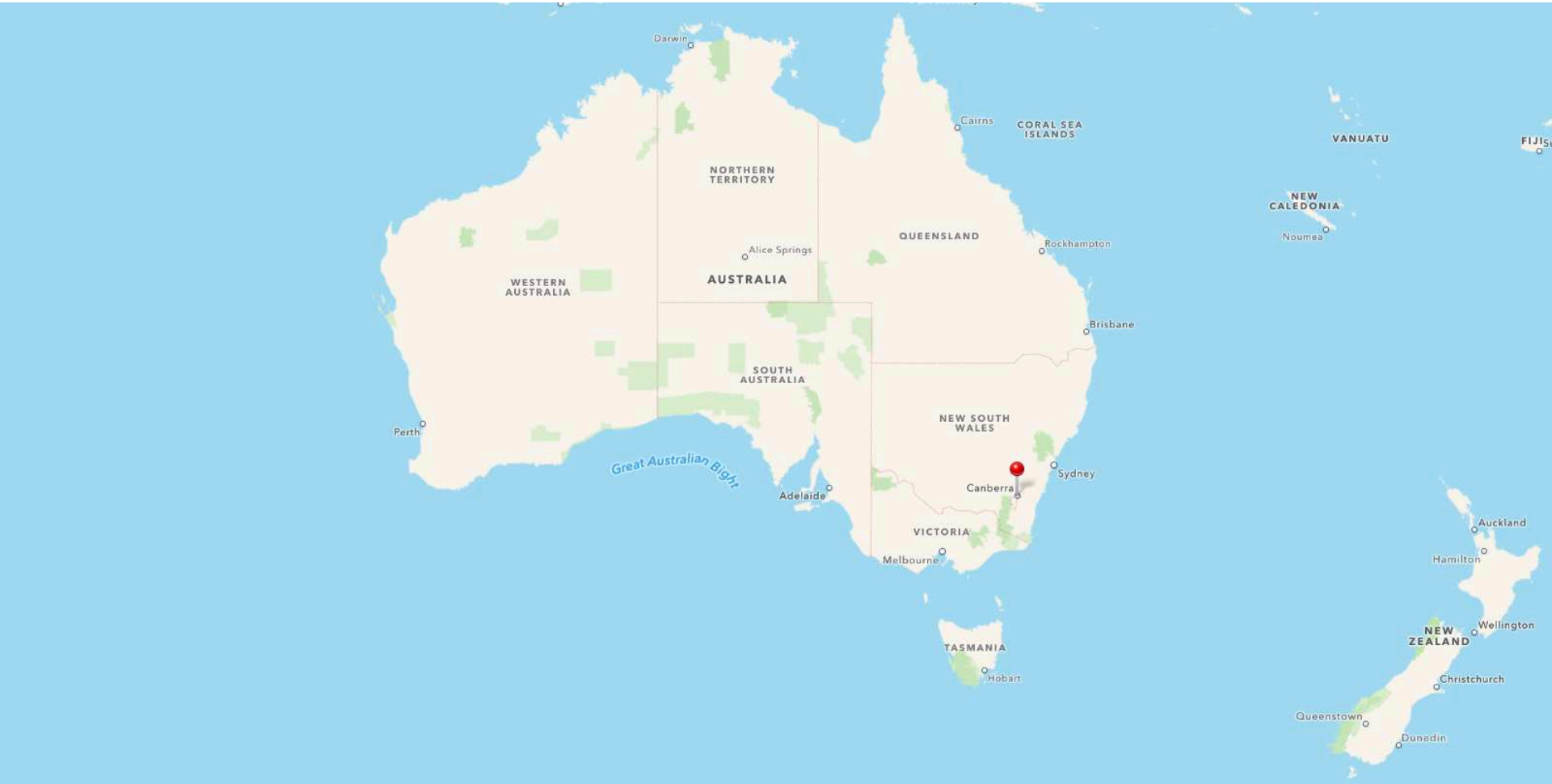
What is NCI?

- 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.

Australian Government Department of Education
Australian National University
CSIRO
Australian Government Bureau of Meteorology
Australian Government Geoscience Australia
Australian Government Australian Research Council
DEAKIN UNIVERSITY AUSTRALIA Worldly
qcif
INTERSECT

Where are we located?

- In the Nation's capital - Canberra, ACT
- at its National University – The Australian National University (ANU).



Research focus areas

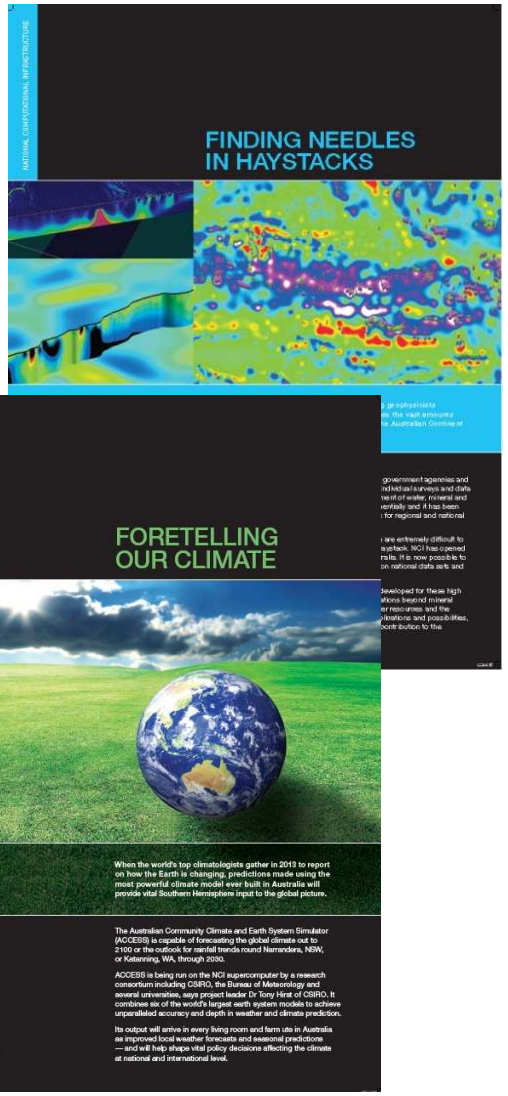
- Climate Science and Earth System Science
- Astronomy (optical and theoretical)
- Geosciences: Geophysics, Earth Observation
- Biosciences & Bioinformatics
- Computational Sciences
  - Engineering
  - Chemistry
  - Physics
- Social Sciences

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

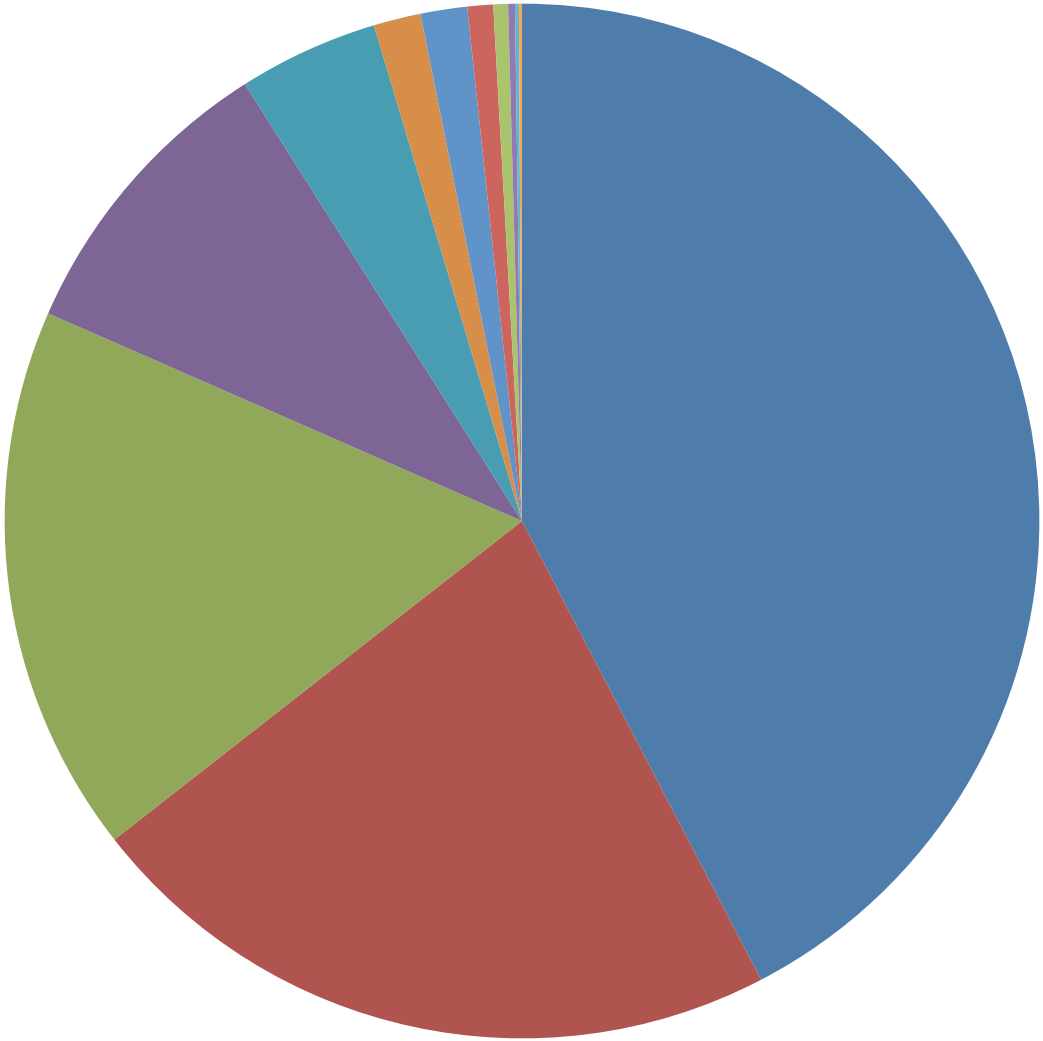
Who Uses NCI ?

- 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 ?



'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)

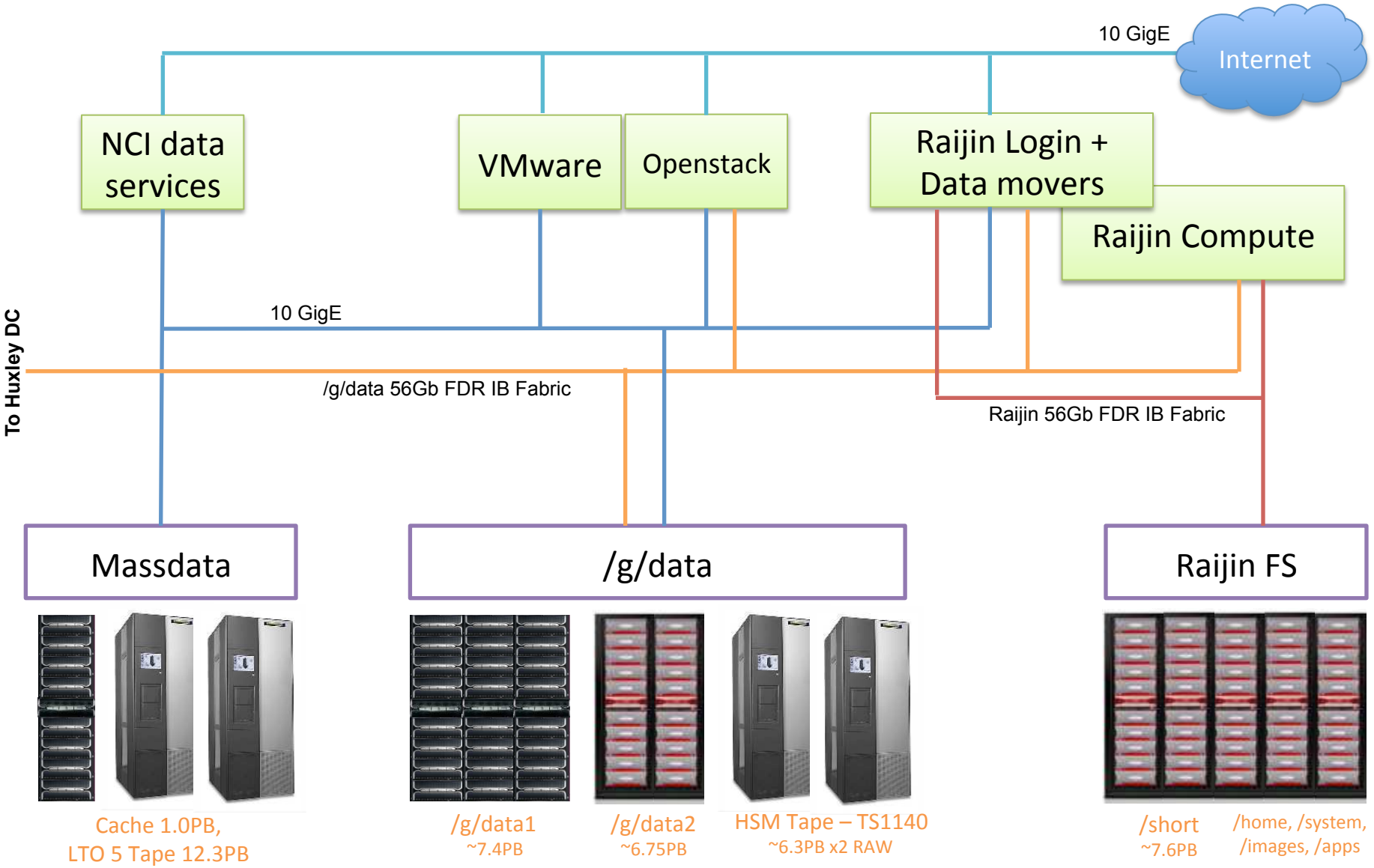
– 24th fastest in the world on debut (November 2012); first petaflop system in Australia

- 1195 Tflops, 1,400,000 SPECFPrate
- Custom monitoring and deployment
- Custom Kernel, CentOS 6.5 Linux
- Highly customised PBS Pro scheduler.
- FDR interconnects by Mellanox
  - ~52 KM of IB cabling.
- 1.5 MW power; 100 tonnes of water in cooling

22PB 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 2.4.2 + Custom patches (DDN). v2.5.3 scheduled for 1st October
  - Gdata1 – Persistent Data: /g/data1
    - 7.4PB @ 21GB/Sec (IOR Aggregate Sequential Write)
    - Lustre 2.3.11 (IEEL v1). IEEL 2 update scheduled for 1st half 2015
  - Gdata2 – Persistent Data: /g/data2
    - 6.75PB @ 46GB/Sec (IOR Aggregate Sequential Write)
    - Lustre 2.5.3 (IEEL v2.0.1)
- 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


10 GigE
Internet
NCI data services
VMware
Openstack
Raijin Login + Data movers
Raijin Compute
10 GigE
To Huxley DC
/g/data 56Gb FDR IB Fabric
Raijin 56Gb FDR IB Fabric
Massdata
/g/data
Raijin FS
Cache 1.0PB,
LTO 5 Tape 12.3PB
/g/data1
~7.4PB
/g/data2
~6.75PB
HSM Tape – TS1140
~6.3PB x2 RAW
/short
~7.6PB
/home, /system,
/images, /apps

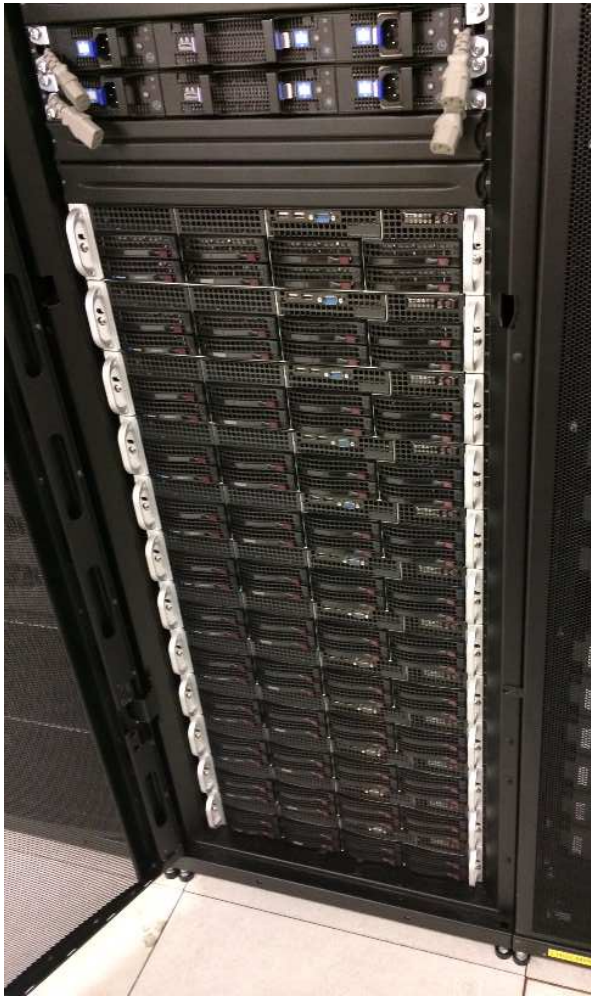
High Performance Persistent Data Stores

Lustre HSM Project

- Previous Systems
  - 900TB Lustre 1.8 Filesystem (/g/data)

  and

  - 1.4 PB CXFS Filesystem (/projects)
  - dual state HSM to DMF
  - Backed by 2x LTO5 Tape Libraries
  - Needed significant capacity growth to accommodate large reference data sets required by researchers
  - Scalability concerns for Petascale HPC and beyond workloads



Original 900TB gdata MDS/OSSes

- Data Storage Requirements
  - High Performance, High Capacity Storage capable of supporting HPC connected workload.
  - Persistent Storage for Active Projects and Reference Datasets, with ‘backup’ capability.
  - 14PB required by end of 2014.
  - Modular design that can be scaled out as required for future growth
  - 20+ GB/sec minimum performance, online
  - Available across all NCI systems (Cloud, VMWare, HPC) using native mounts and 10/40Gbit NFS.

SGI IS4600 Disk Enclosures ready to be installed

- Gdata Persistent Data Stores
  - /g/data 1 – 7.4PB capacity
    - 4.2PB used, 150M inodes
  - /g/data 2 – 6..75PB capacity
    - 0PB used, pre-production, go-live October 2014
  - Approx 300-400M inodes per /g/dataN
  - 14.1PB, 800M+ inodes (possibly 1B inodes?)
- Backups?
  - Traditional 'Backup' not viable – interval? Deep traversal of directory structures?
  - Data change between start and end of backup event?
  - Calculation of difference between backup events takes days/weeks
  - Backup impact on filesystem performance, particularly metadata load on MDS
- HSM as a backup - Lustre HSM & Changelogs
  - Lustre MDS knows which files are being accessed & altered
  - Activity logged in a 'changelog'
  - No need for deep traversal if you know what is being altered.
  - 'backup' is always occurring, light persistent load – not periodic intense loads

Design – gdata1 building blocks

- Gdata1 Lustre servers are Dell R620 v1

MDS

Dual 2.90GHz E5-2690 Xeon (*Sandy Bridge*) 8-core CPUs
768GB LRDIMM DDR3

OSS

Dual 2.00GHz E5-2620 Xeon (*Sandy Bridge*) 6-core CPUs
256GB RDIMM DDR3

2 MDS (1 HA pair)
44 OSS (22 HA pairs)

Current image

CentOS 6.4

In Kernel OFED

Lustre v2.3.11 (IEEL v1.0 + patches)

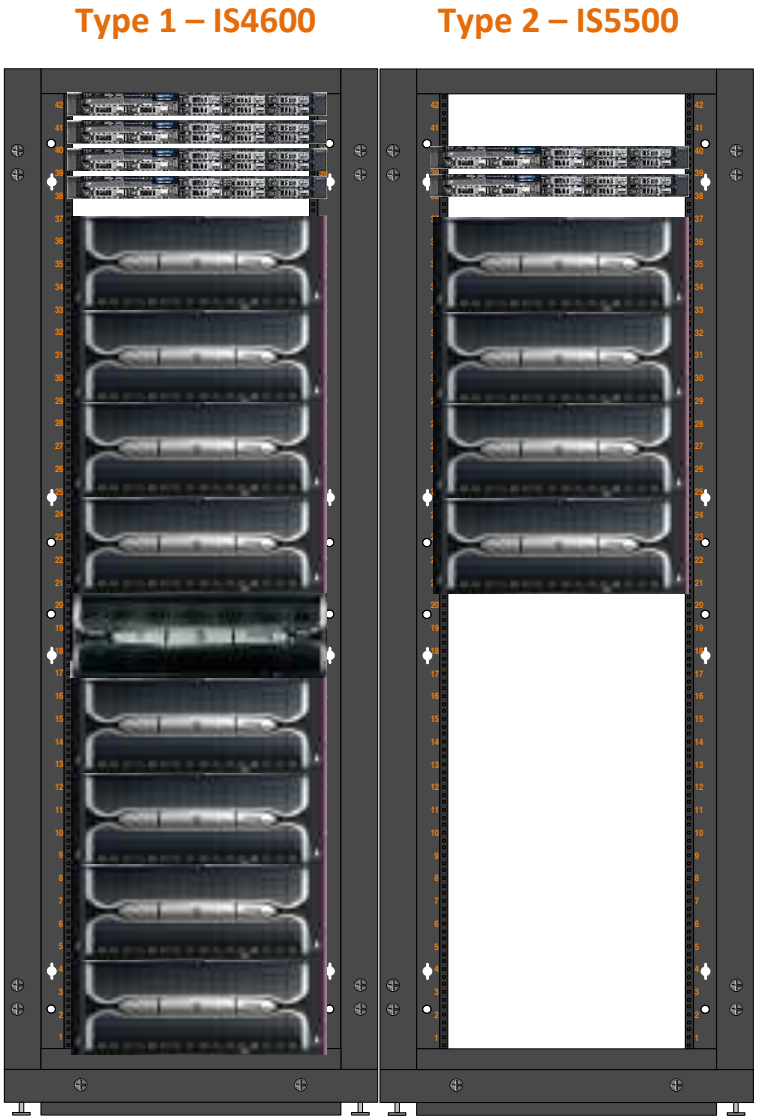
corosync/pacemaker

Gdata1 Object Store Building Blocks

- Storage for Gdata1 is built using SGI's Infinite Storage block storage arrays, with OSS-OST 8Gbit Fibre Channel interconnects

- Type 1 Building Block (x 10)
  - SGI IS 4600 Array
  - 480 x 2TB 7.2K SATA disk
  - 4 x OSSes per IS4600 (2x HA Pairs)
  - 46 x RAID 6 (8+2) 14.5TB pools (OSTs)
  - 11-12 OST per OSS, 22-24 per HA pair

- Type 2 Building Block (x2)
  - SGI IS 5500 Array
  - 240 x 3TB 7.2K NL-SAS disk
  - 2 x OSSes per IS5500 (1x HA Pair)
  - 30 x DDP 14.55TB pools (OSTs)
  - 15 OST per OSS, 30 per HA pair



- Gdata2 Lustre servers are Dell R620 v2

MDS

Dual 3.00GHz E5-2690v2 Xeon (*Ivy Bridge*) 10-core CPUs
768GB LRDIMM DDR3

OSS

Dual 2.60GHz E5-2630v2 Xeon (*Ivy Bridge*) 6-core CPUs
256GB RDIMM DDR3

2 MDS (1 HA pair)
24 OSS (12 HA pairs)

Current image

CentOS 6.5

Mellanox OFED 2.0.1-6

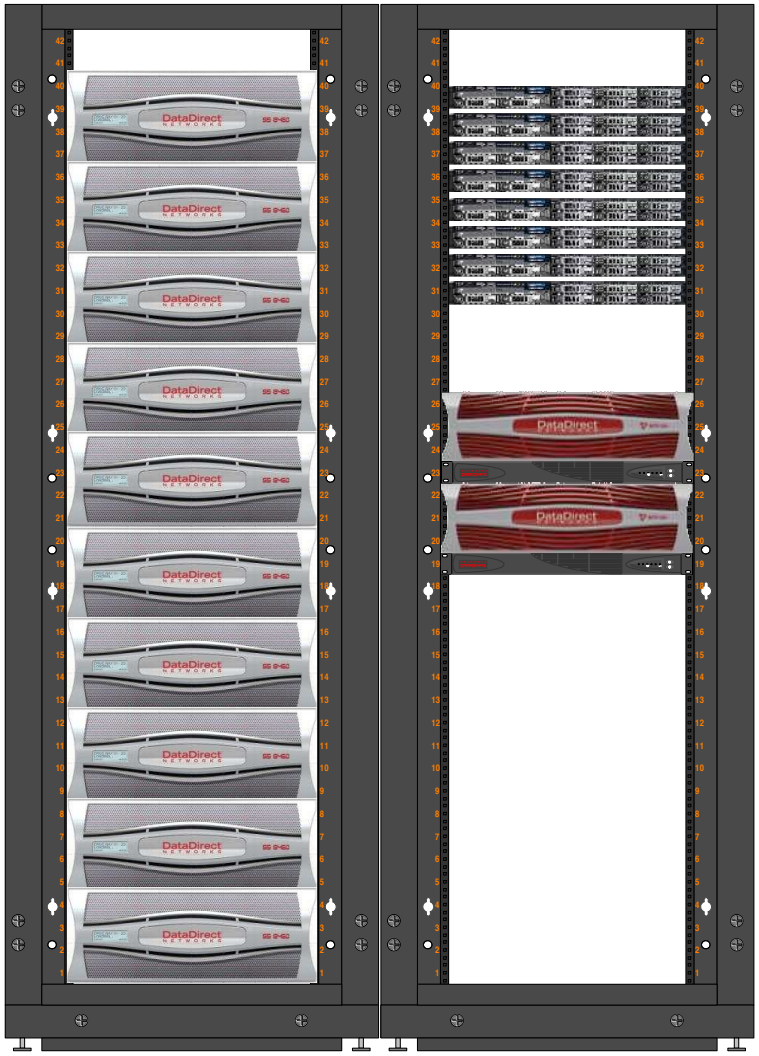
Lustre v2.5.2 (IEEL v2.0.2)

corosync/pacemaker

Gdata2 Object Store Building Blocks

- Storage for Gdata2 is built using DDN's SFA12K-40IB block storage appliances, with OSS-OST FDR IB interconnects.

- Building Block (x 3)
  - DDN SFA12K-40IB, with 10x SS860 84 bay enclosures
  - 800x 4TB 7.2K NL-SAS disk
  - 8 x OSSes per SFA-12K (4x HA Pairs)
  - 80 x RAID 6 (8+2) 32TB pools (OSTs)
  - 10 OSTs per OSS, 20 per HA pair



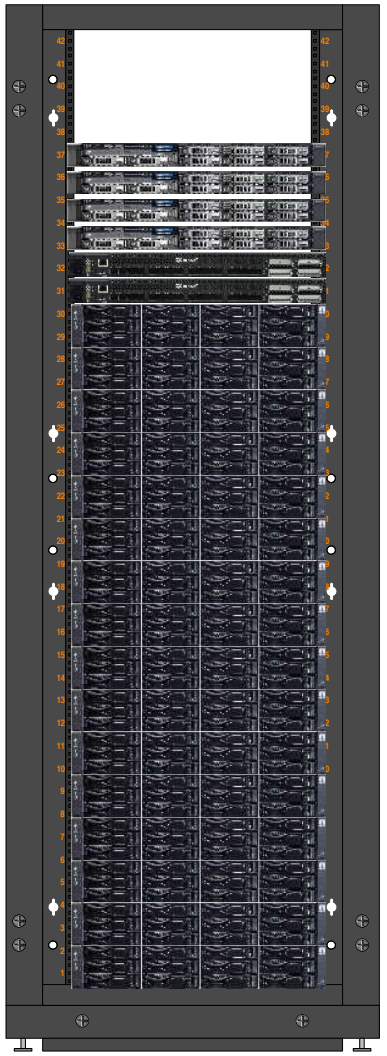
Gdata 1 & 2 Metadata Building Blocks

- MDT storage for both Gdata1 and Gdata2 is built using a shared SGI Infinite Storage 5000 block storage array, with MDS-MDT 8Gbit Fibre Channel interconnects

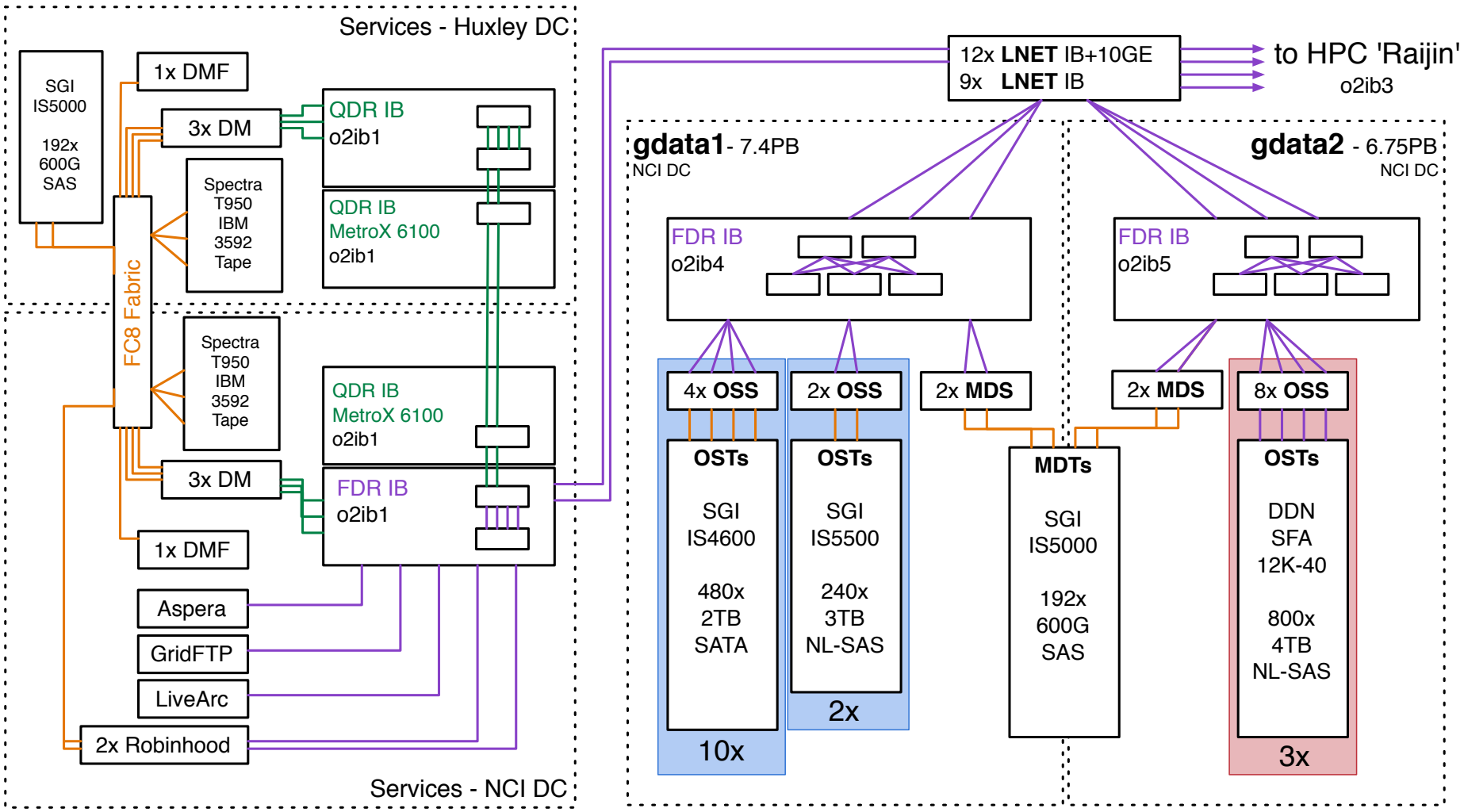
- Array:
  - 192 x 600G 15K SAS
  - Dual 8Gbit FC Controllers
  - 2 x Qlogic 8Gbit SanBox 5800 FC Switches

- Gdata1
  - 2x MDS
  - 40 disk 600G 15K SAS DDP pool

- Gdata2
  - 2x MDS
  - 2 x 48 Disk 600G 15K SAS pools, XVM together
  - 1 preferred pool per controller



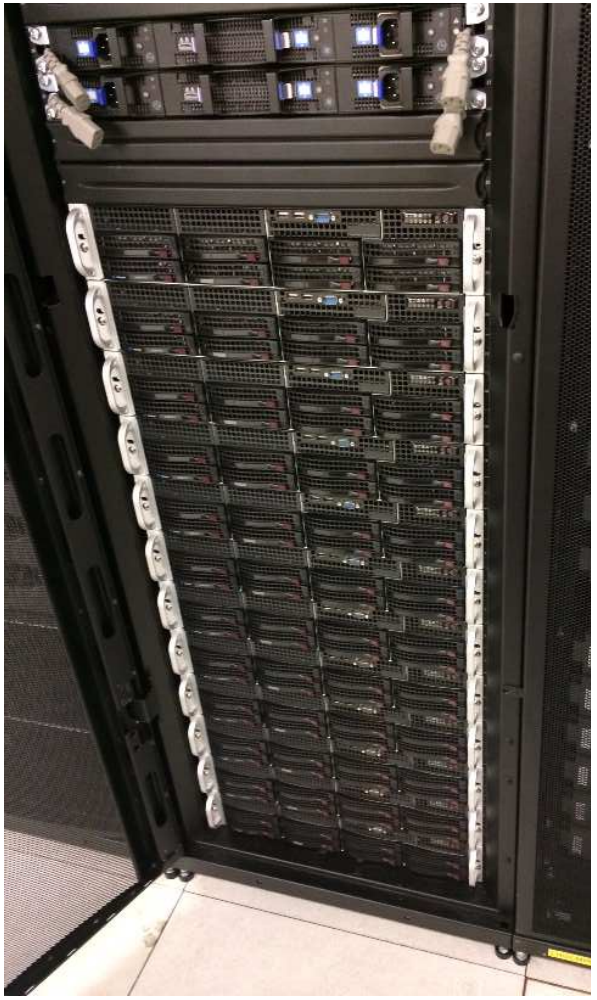
Fabric Layout



- HSM Configuration
  - Essentially create a backup, rather than migrating tiers
  - All Lustre objects to be Dual Stated – i.e. exist both on Lustre Disk, HSM Tape
  - Backend tape to be Dual Site – i.e. copied to primary and secondary tape library
    - for site level protection (Disaster Recovery) and
    - tape level protection (tape fault)

- HSM Stack
  - Lustre v2.5 Front End
  - Robinhood Policy Engine (2.5.3)
  - SGI DMF Copytool v1.0
  - SGI DMF 6.2 Tape Back-End (+ ISSP 3.2 / CXFS 7.2)
  - Spectra Logic T950 Tape Library
  - IBM 3592 Tape System, TS1140 Drives, JC Media

- Dedicated DMF instance to service workload
- 2x DMF Controllers
  - SGI C2108-RP2 Servers
  - Dual 2.6GHz E5-2670 8C Xeon *(Sandy Bridge)*
  - 256GB RDIMM DDR3
  - 2x Quad Port 8GBit FC HBAs
  - SLES 11 SP3
  - DMF 6.2, ISSP 3.2, CXFS
- 6x Lustre <> DMF Datamovers
  - SGI C2108 TY-11 Servers (c.2011)
  - Dual 2.4GHz E5620 8C Xeon *(Westmere)*
  - 48GB RDIMM DDR3
  - Dual Port QDR IB HCAs
  - 2x Dual Port 8GB FC HBAs
  - SLES 11 SP3 (3.0.10)



Lustre HSM Data movers
(previous generation OSSes)

Lustre HSM

Challenges & Experience so far

- Existing Tape System
  - Existing DMF deployment – Massdata Filesystem: 1PB cache, 6PB x2 LTO 5 Tape
  - Migrating filesystem - primarily offline, dual site copy tape
  - Mix of writes and recalls, many tape movements and load cycles
  - 2x Spectra Logic T950 libraries (reconfigured as 4 frame each)
  - 54x IBM Ultrium LTO 5 drives (140MB/sec ea) across 2 libraries
  - 4200 LTO5 Tapes each library. 1 Robotics unit per Library.

- Anticipated Lustre HSM Workload
  - Dual-state (online & Tape), Dual Site (2 libraries, 1 copy each)
  - Primarily write biased workload
  - Few recall events – required for recovery only (user, disaster)
  - Supporting high performance filesystems (21GB/sec + 46GB/sec)
    - Expected to grow to future gdata3, possibly 90+GB/sec
  - Large streaming writes (DMF will attempt to optimise tape layout)
  - Very large 'initial' copy on HSM enable (1-2PB?)
    - Likely to monopolise tape drives and library for weeks

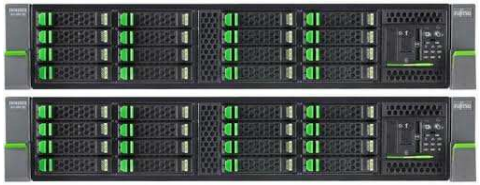
- Answer – Additional Tape Libraries
  - 2x new Spectra Logic T950 Universal Libraries
  - Configured for IBM 3592 'Jaguar' Tape
  - Fewer, but faster + larger capacity drives
  - Media up-format capability to newer generation
  - More Libraries = more robotics units
  - **Initially**:
  - 6x TS1140 drives per Library (350MB/sec ea)
  - 1584x 'JC' Media (4TB uncompressed ea)
  - **Early 2015:**
  - Upgrade to next generation drives (TS1150?)
  - Up-format existing media + higher capacity next generation media (xxTB?)
  - Well suited for write biased, streaming performance, fewer load cycles

New Spectra Logic T950 'Jaguar' TS Library build

- Robinhood – sizing resources
  - Server hardware – how big?
  - Database storage – performance?
  - Database tuning?
  - Changelog performance impact on MDS?

- Fact-finding
  - Configured & enabled lustre changelogs on /gdata1 during period of low filesystem utilisation
  - Observed performance over time (very little impact – likely ~ %5)
  - Analysed changelogs – entries per hour, entries per 24 hours
    - Rough guide of required database insertions/updates per sec
  - MDS/MDT performance – will be a limiting factor on initial scan
  - Varied DB storage types (SSD, 10K SAS, 15K SAS) and observed RBH changelog rate processing impact.

- Server Hardware – repurposed existing
  - 2x Fujitsu RX300 S7 (HA Pair), each with
  - Dual 2.6GHz E5-2670 8C Xeon *(Sandy Bridge)*
  - 128GB RDIMM DDR3
  - FDR InfiniBand HCA
  - Dual Port 8GBit FC HBAs
  - * 3x Dual Port Intel X520 10GE NICs for test below



- Storage Array for Robinhood MySQL Database
  - Evaluation system - NetApp EF-550 'All-flash' Array
  - 450,000 IOPS sustained
  - 24x 800GB SSDs in 2RU
  - 8x iSCSI 10GE host ports

  - Benchmarked up to 320,000 4K IOPS with single host, using 6 of 8 available 10GE ports (RX300 CPU limited)
  - Essentially uncapped IOPS for testing

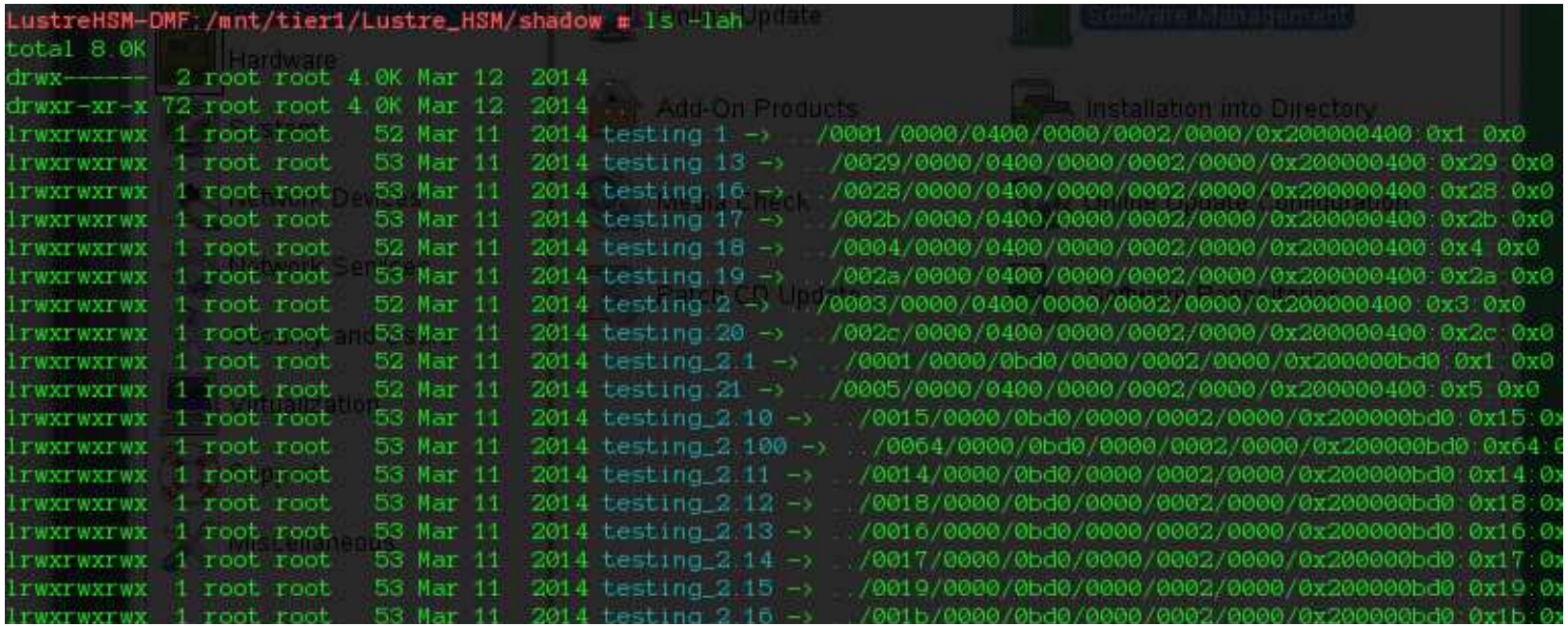
- Storage Array for Robinhood MySQL Database
  - For Production deployment
    - starting point, can grow later
  - SGI IS5000 #2 at 2nd datacentre, also hosting other applications
    - 192x 600G 15K SAS, 8Gbit Fibre Channel
    - Using 2x 10 Disk RAID 10 Pools, XVM together, 1 preferred pool per controller
    - ~ 4000 Read IOPS, 2000 Write IOPS
    - ~ 1.9TB size
- Database
  - Storage optimised for 4K IO, tune for transactional rather than streaming or capacity
  - 150M inodes used at time on /g/data1 = 99GB MySQL Database after initial scan
  - Can easily relocate MySQL database storage later
  - Aim for most of database / all indexes to be in memory (tune MySQL appropriately)
  - HA Pair of Robinhood/MySQL servers – 1 active per gdata filesystem

- Hit LU-5405: *'Performance Issue while using Robinhood in changelog mode'*
  - Very similar conditions to reported bug (filesystem - 150M inodes, 3600+ clients)
  - As the Lustre changelog grows, a condition can occur where an ever increasing amount of time is required to process the entries
  - MDS appears to level out at processing about 30 changelog records per second, irrespective of backend database server/storage
  - Compounding backlog of events to deal with (insert / clear)
  - Active MDS is eventually overwhelmed
    - High load average
    - MDS processing changelogs (slowly) rather than servicing OSS requests
    - Filesystem unresponsive (unscheduled downtime)
  - Currently disabled on /g/data1 (v2.3.11) until fix applied

Challenge #4 – DMF disk (home/cache/work filesystem)

- DMF home, cache, work filesystem
  - DMF uses an intermediate XFS filesystem in front of tape tier
  - Basically a pool of inodes acting as pointers
  - Allows presentation of browable ‘shadow’ filesystem

  - But…
  - 1 lustre object != 1 inode on DMF intermediate XFS filesystem (same for POSIX copytool)

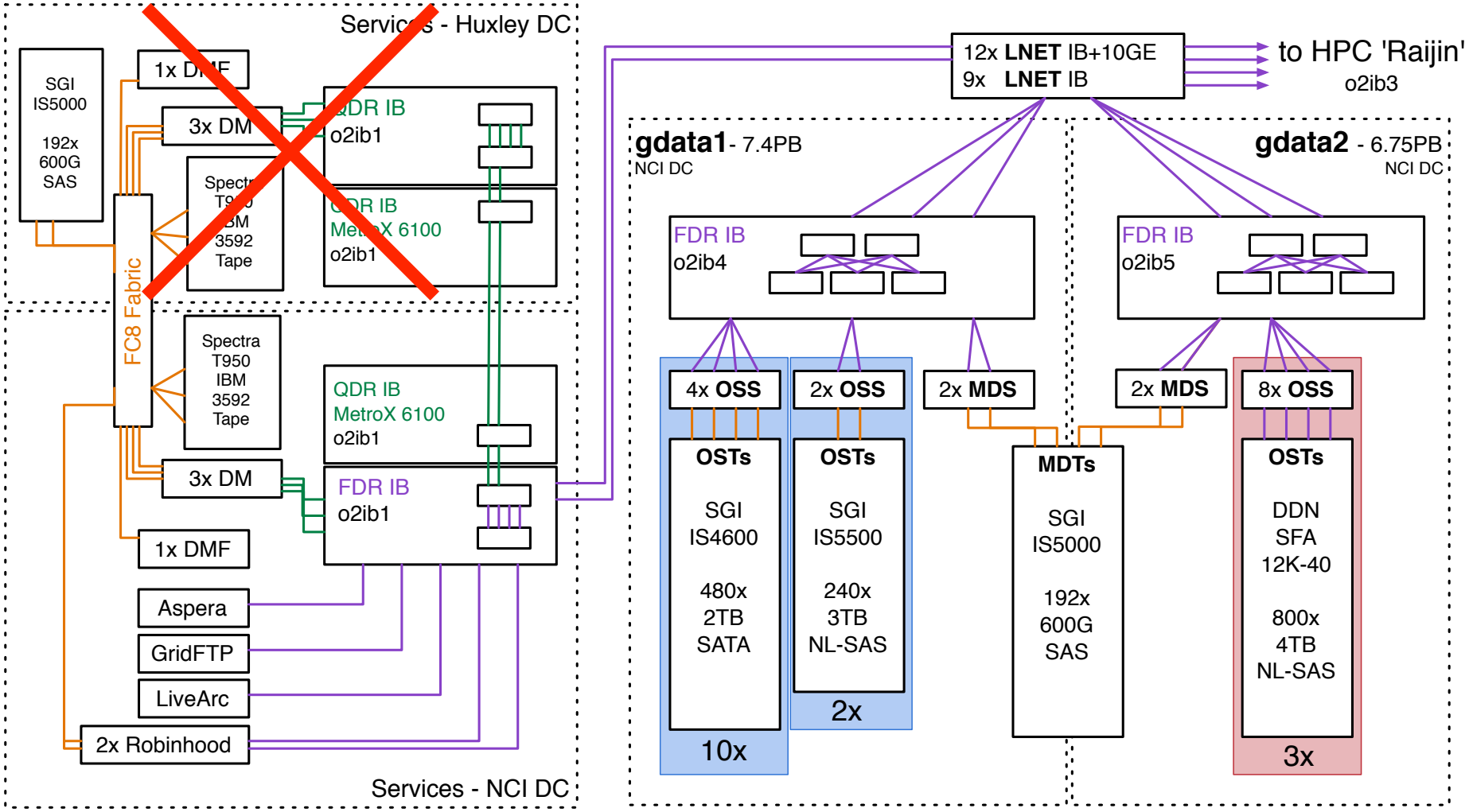


- Design of DMF disk
  - Normally DMF copytool will move data directly between Lustre Filesystem, and DMF tape
  - However, need to accommodate disaster recovery situation where Lustre may not be in a viable state
  - Need sufficient 'cache' available to recall data in abnormal conditions
    - at least as large as largest file
    - ability to perform specific tape maintenance if needed (i.e. recall tape on to disk other than Lustre)
  - High IOPS required to perform weekly XFS inode dump (backup of inode 'pointers')
  - Rough guide
    - IOPS – 2x 40 disk RAID 10, XVM (80 Disk R10)
      - Approx 16K IOPS read, 8K IOPS write, 23.4TB.
    - Capacity
      - TS 1140 'JC' Media is 4 TB Native, 8TB 2:1 Compressed
      - Guess at next generation: TS 1150 ? 'JD' media? Maybe 8-10TB, 2:1 @ 20TB??

- Data Recoverable after Primary Site incident (NCI DC)
  - Need to accommodate in design
  - Will have many other problems if total loss on primary site
    - No HPC system to use data
    - Where would you 'restore' 14PB to?
  - Still 'nice-to-have' capability
  - Configured so that data can be retrieved if necessary from only 2nd site (even if few TB at a time)
- Cross-site design
  - IS5000 with DMF (shared with Robinhood DB) at 2nd site
  - Existing cross site 8Gbit FC Fabric, between NCI DC <> Huxley DC
  - Mellanox Long Distance IB 'MetroX 6100' links both DCs, 6x QDR IB WAN capable
  - 3 Lustre HSM datamovers per site
  - 1 Spectra Logic T950 3592/TS1140 Library at each DC, dual site mirrored into each

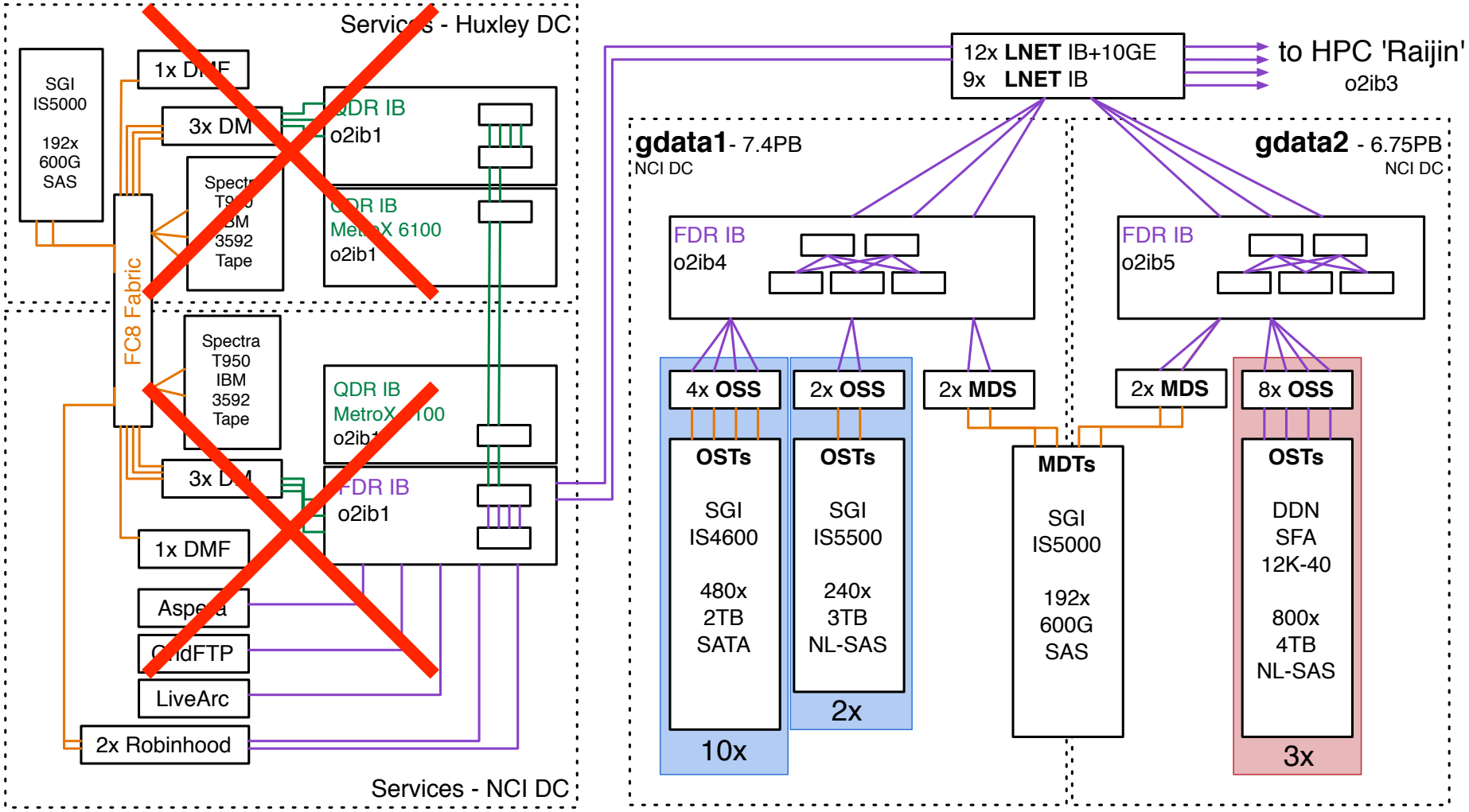
Loss of 2nd Site: OK

- Lustre not directly affected, DMF + Robinhood unavailable. 'HSM' capabilities offline.



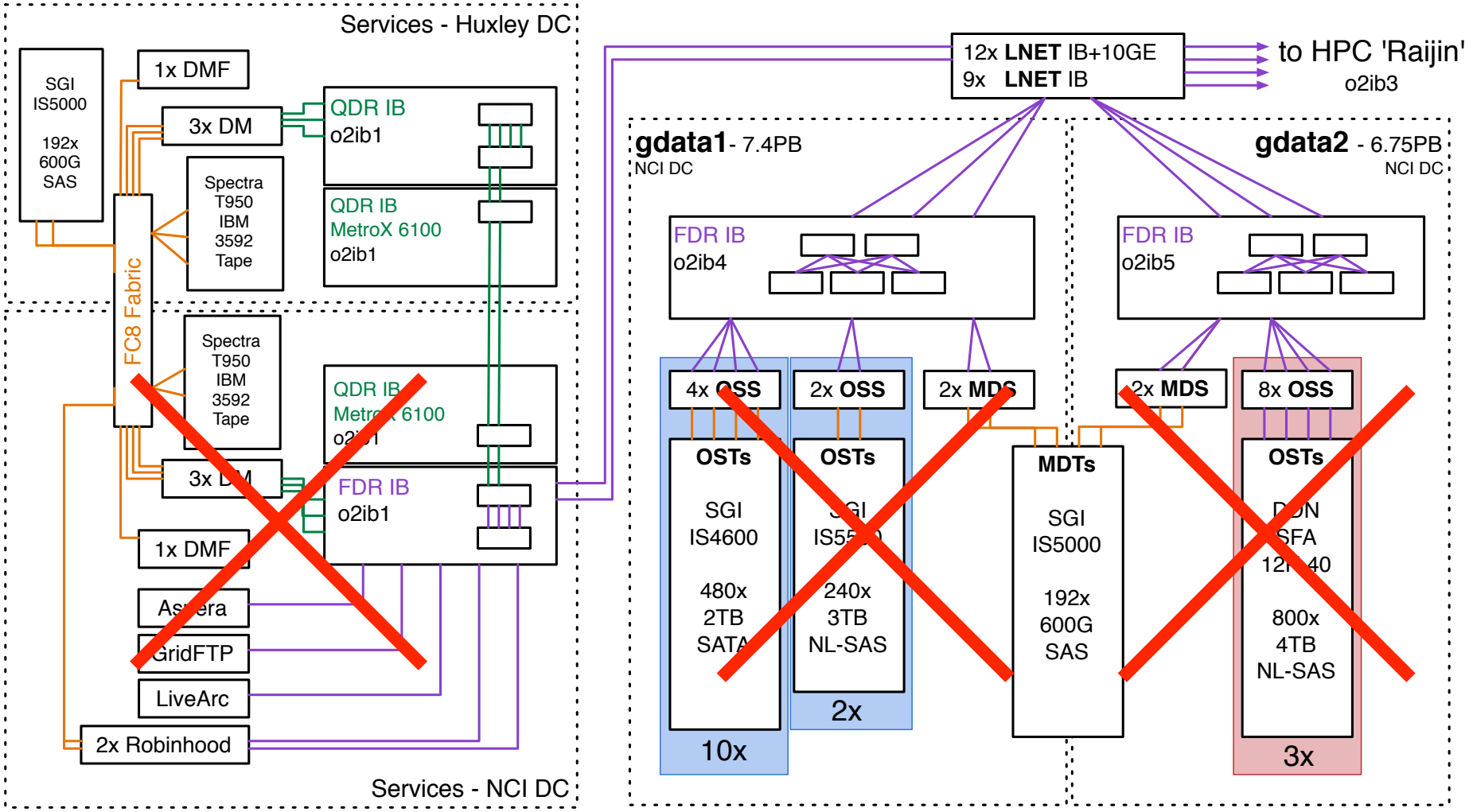
Loss of HSM components: OK

- Lustre not directly affected, DMF + Robinhood unavailable. ‘HSM’ capabilities offline.



Loss of Primary Site: ☹, but not fatal.

- Lustre offline. Data secure & recoverable at 2nd site (with low performance)



Lustre HSM

Roll-out plan

Roll-out plan

- Gdata2 is being built on IEEL v2.02 (Lustre 2.5.2)
- Existing Gdata1 (IEEL 1 + patches) will stay on 2.3.11 until gdata2 is successfully operating in full HSM Mode
  - End 2014 (Gdata2 in full HSM), Early 2015 (Gdata1 upgrade)
- Phased rollout of functionality
  1. Build /g/data2 filesystem as Lustre 2.5.2 / IEEL 2.0.2
  2. Enable + Tune Changelog config for RBH / HSM
  3. Build Robinhood, configure policies and types
  4. Build DMF components + Datamovers + Tape Library configuration
  5. Enable full HSM
- Phased approach allows for easier troubleshooting along the way

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