NATIONAL COMPUTATIONAL INFRASTRUCTURE

Lustre at the Australian National Computational Infrastructure (NCI)

Joseph Antony (NCI)
joseph.antony@anu.edu.au

Patrick Fitzhenry (DDN)
pfitzhenry@ddn.com

Australian Government
Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education

Australian National University

Australian Government
Bureau of Meteorology

CSIRO

Australian Government
Geoscience Australia

Australian Government
Australian Research Council

www.nci.org.au

- What is NCI ?
- Petascale Machine at NCI (Raijin)
- Root over Lustre
- Lustre Storage on the Petascale Machine
- Other Lustre Storage at NCI
- Future Plans & Collaboration Possibilities

WHAT IS NCI?

Mission:

- **to foster ambitious and aspirational research objectives and to enable their realisation, in the Australian context, through world-class, high-end computing services**

NCI is:

- **being driven by research objectives**
- **a comprehensive, vertically-integrated research service**
- **providing national access on priority and merit, and**
- **being built on, and sustained by, a collaboration of national organisations and research-intensive universities**

Australian Government
Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education

Australian National University

CSIRO

Australian Government
Bureau of Meteorology

Australian Government
Geoscience Australia

Australian Government
Australian Research Council

The University of Adelaide

The University of Sydney

The University Of Queensland

Monash University

UNSW

INTERSECT

qcif

Research Objectives

- Research Outcomes
- Communities and Institutions/ Access and Services
- Expertise Support and Development
- Digital Laboratories
- Data Centric Services
- Compute (HPC/Cloud) and Data Infrastructure

ARC Centre of Excellence for Climate System Science

CAASTRO ALL-SKY ASTROPHYSICS

space.gov.au

TERN Terrestrial Ecosystem Research Network

cudos An ARC Centre of Excellence

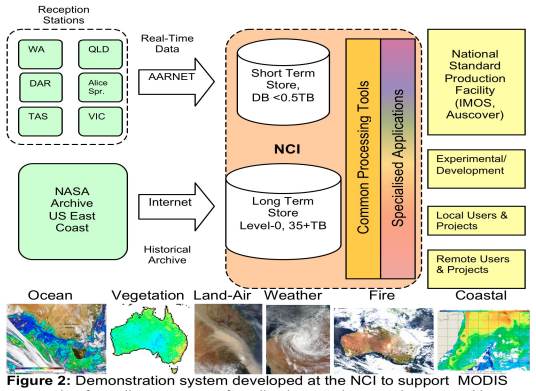
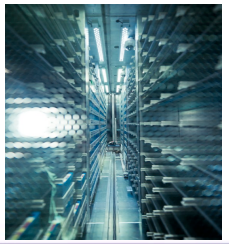


Figure 2: Demonstration system developed at the NCI to support MODIS processing for a diverse range of applications and research communities.



- In the nation's capital, at its national university …

Engagement: Research Communities

- **Specialised Support**
  - Climate Science and Earth System Science
  - Astronomy (optical and theoretical)
  - Geosciences: Geophysics, Earth Observation
  - Biosciences: Bioinformatics
  - Social Sciences

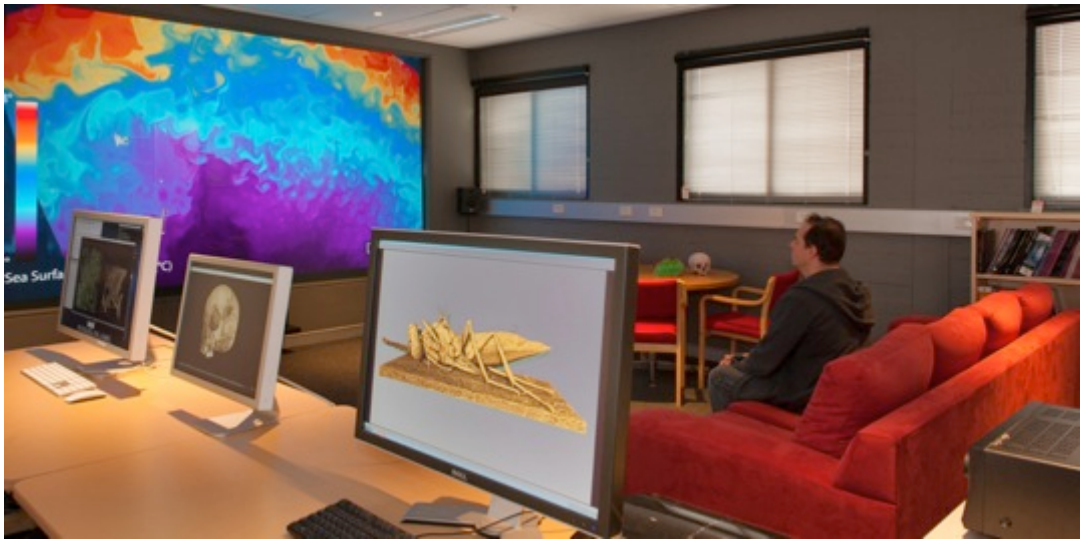
  - Growing emphasis on data-intensive computation
    - Cloud Services
    - Earth System Grid

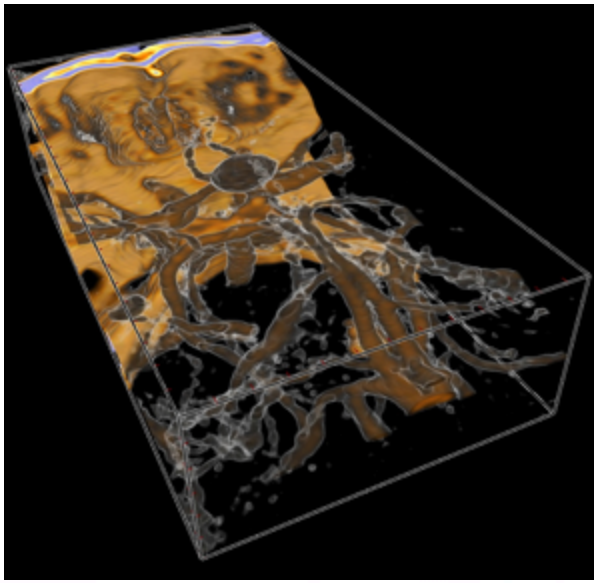
- **Engagement with RDSI and NeCTAR**
  - Approximately $100M in funding from the Australian Federal Government
  - RDSI – National Storage Initiative
    - NCI High-Performance Data Node
      - Hosting data collections of national importance, seeding storage initiatives across the country
  - NeCTAR – National Research Cloud Initiative
    - High-Performance node of NeCTAR Cloud
    - Major Participant in Virtual Labs (VLs)
      - Weather and Climate VL
      - All-Sky Virtual Observatory VL
      - Contributing to Characterisation VL, VEGL
    - Tools—volume visualisation in the cloud

- NCI VizLab in existence since early-1990s
- Innovative software development (Drishti and Voluminous)
- Skilled visualisation programmers who deal with multi-terabyte datasets
- Lustre use-case: access from visualization desktops, driving video walls, on-demand GPU clusters, on-demand volume visualization

http://nci.org.au/specialised-support/scientific-visualisation/vizlab-showcase/

http://youtu.be/1JxUYUKSnLs





PRIORITY SCIENCE AREAS

Engaging with Priority Research: Climate

Case Study: Building a National Climate Modelling Capability

Partners: CAWCR (Bureau of Met, CSIRO), ARC Centre for Climate Systems Science, NCI, Fujitsu

Goals:

- Enhance the value of investment in ACCESS model development
- Harness and develop Australia's international value in Climate Research (CAWCR + AU Universities)
- Build research infrastructure in harmony with operational environment

Requirements:

- High Performance Computing at NCI available at competitive level to support Climate
- Provide integrated environment for supporting:
  - Simulations
  - Data repository: Online and Deep Archive
  - Cloud capability for data processing, analysis and visualization

http://youtu.be/zUF2rsq7ej8

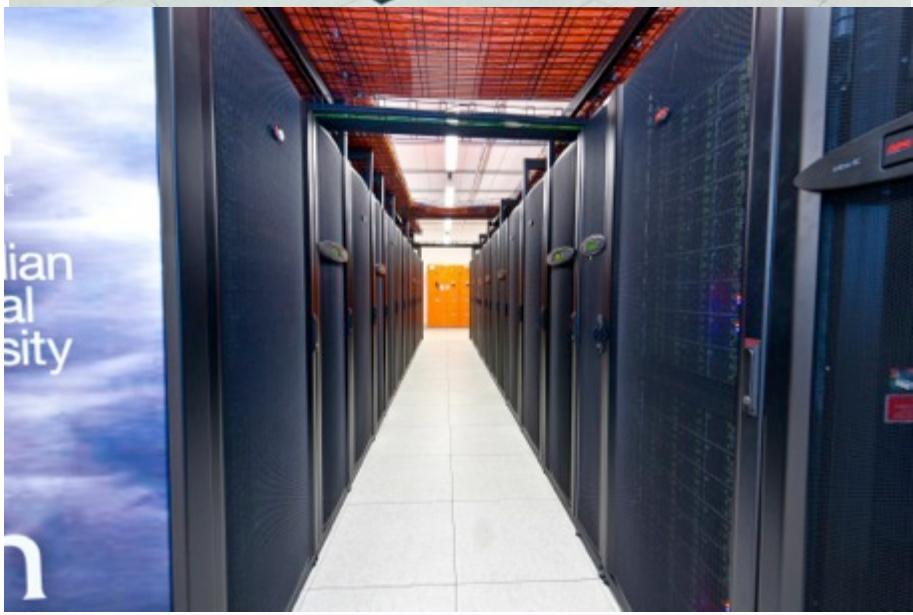
VIDEO: ANDY HOGG @ ANU

CURRENT INFRASTRUCTURE

Background: Evolution of Peak Facilities at NCI/APAC

System (Top500 rank)	Procs/ Cores	Memory	Disk	Peak Perf. (Tflops)	Sustained Perf. (SPEC)
2001–04 Compaq Alphaserver (31)	512	0.5 Tbyte	12 Tbytes	1 TFlop	2,000
2005–09 SGI Altix 3700 (26)	1920	5.5 Tbytes	30 (+70) Tbytes	14 Tflops	21,000
2008–12 SGI Altix XE (-)	1248	2.5 Tbytes	90 Tbytes	14 TFlops	12,000
2009–13 Sun Constellation (35)	11,936	37 Tbytes	800 Tbytes	140 TFlops	240,000
2013– Fujitsu Primergy (24)	57,500	160 Tbytes	12.5 Pbytes	1200 Tflops	1,400,000+

Fujitsu Primergy Petascale System (2013–)



Current Infrastructure - Compute

- Raijin—Fujitsu Primergy cluster—June 2013
- Approx. 57,500 Intel Sandy Bridge (2.6 GHz)
- 157 TBytes memory, 10 PBytes short term storage
- FDR Infiniband
- 150 GB/s bandwidth to filesystem
- Centos 6.4 Linux; PBS Pro scheduler
- Good Performance — well balanced, appreciated
  - 1195 Tflops, 1,400,000 SPECFPrate
- Significant growth in highly scaling application codes
  - Largest: 40,000 cores; many 1,000 core tasks

Current Infrastructure - Storage

Data Storage

- HSM (massdata) – DMF based: 8PB as at September 2012 [2 copies]

- /projects: SGI CXFS (Interactive f/s space) HSM (shared with massdata), achieves 2.5 GB/sec from tape

- Global Lustre Filesystem
  - 4.4 PB by end Sept 2012 and growing
  - Global bandwidth: 25 GB/sec
  - Migrating /projects off of CXFS

- Object Storage: Ceph
  - Initially object store for NeCTAR cloud
  - Considering use in long term on-disk copy when erasure coding backend is mature

Current Infrastructure - Cloud

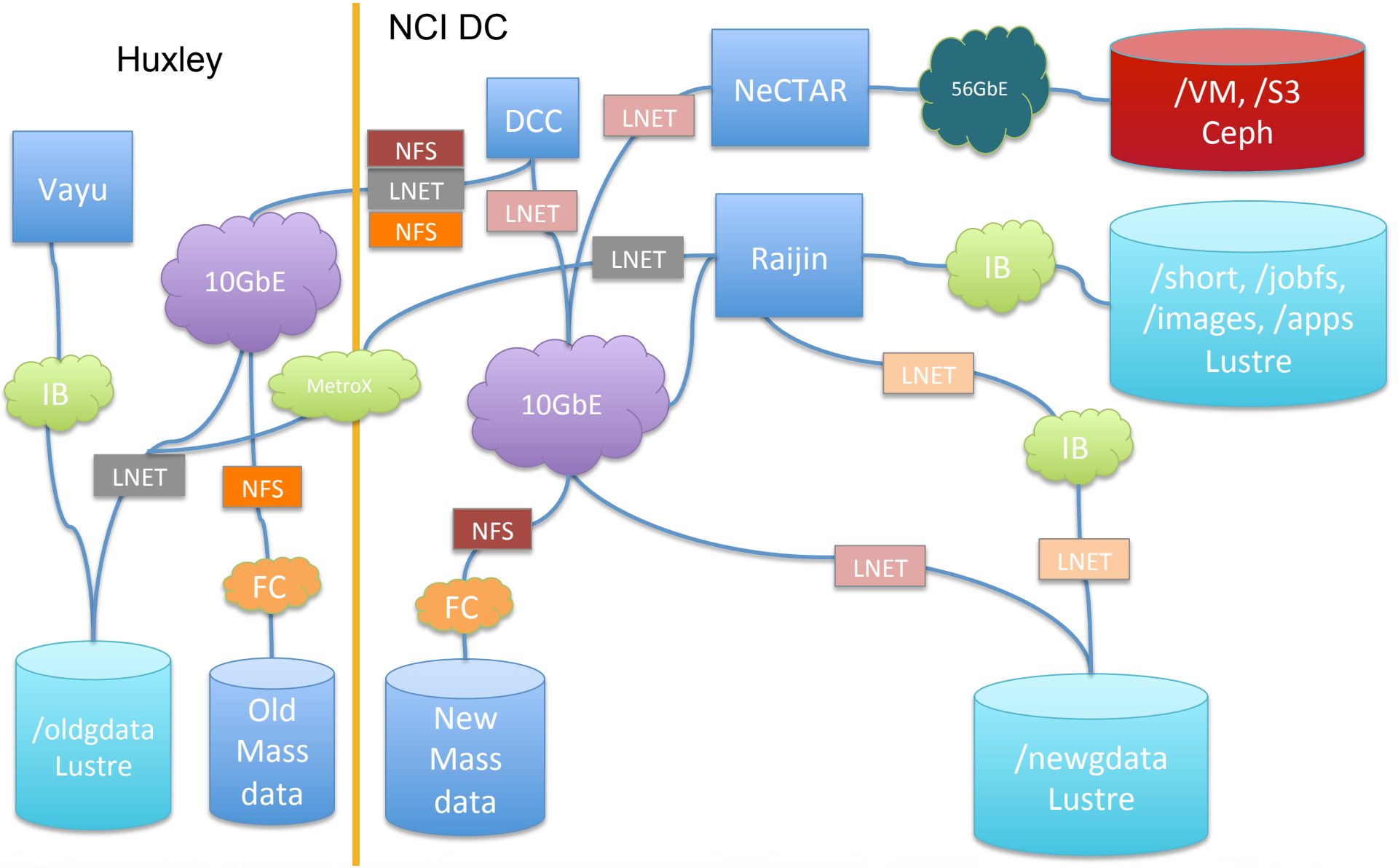
- VMware ESX cluster—providing mission-critical hosting of essential services in a high availability environment

- DCC : Specialised cluster for data-intensive applications
  - Climate, earth-system observation and bioinformatics
  - Part virtualized, part bare-metal

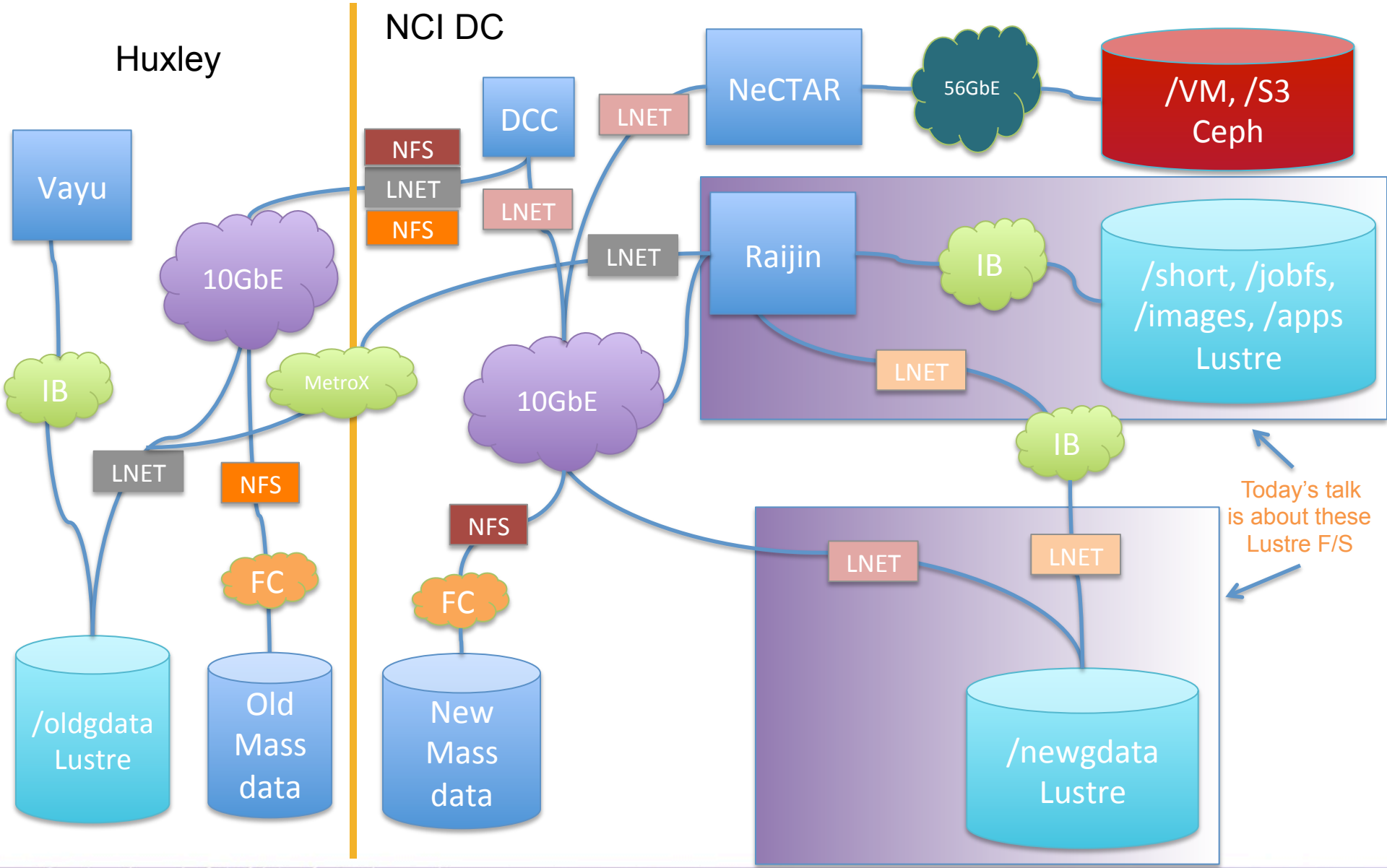
- Cloud computing
  - NeCTAR Research Cloud node at NCI
    - Australia's highest performance cloud
    - Architected for strong computational and I/O performance needed for "big data" research
    - Intel Sandy Bridge (3200 cores)
    - 160 TB of SSDs; 56GigE + RoCE for compute and I/O performance
      - RoCE for LNET

  - Private cloud: RedHat OpenStack
    - SLA centric, on-demand scientific computation

How does all of the pieces link together?



How does all of the pieces link together?



ROOT OVER LUSTRE

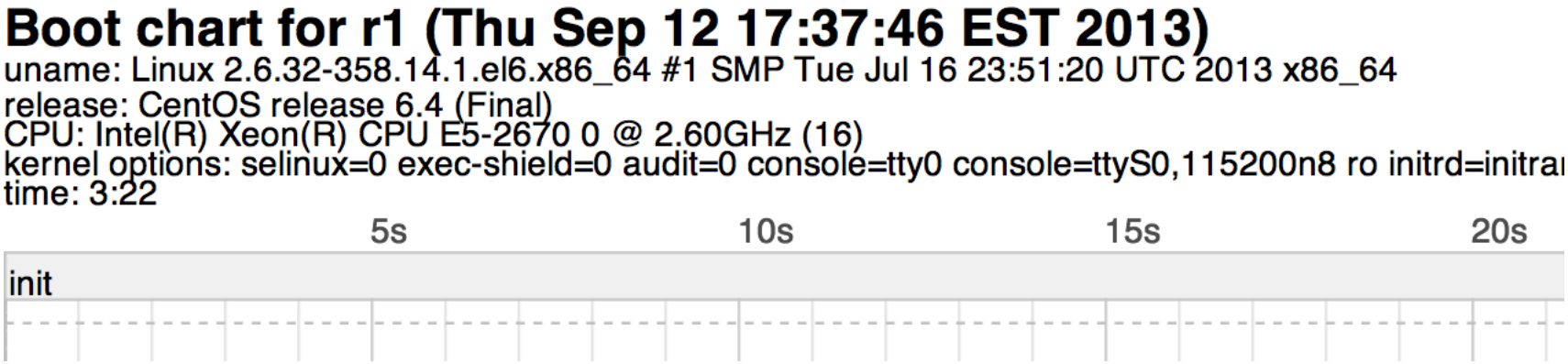
- What is root over Lustre?
  - The root filesystem is provided by Lustre
  - We use oneSIS for provisioning with minor patches
- Why?
  - **S**implicity: Ease of management
    - Diskless compute nodes
    - One golden image for multiple clusters
    - ‘yum update’ the entire cluster
  - **S**ynchronous: Rolling out updates
    - Once an update is made, all nodes see it
  - **S**ecurity: Better/Coherent patching
- We have been using root over Lustre since 2008

- Key feature: oneSIS loads Lustre kernel modules and parses the location of the root filesystem from its boot command line:

  lustreroot=10.9.103.1@o2ib3:10.9.103.2@o2ib3:/images/NCI/centos-6.4-compute-03

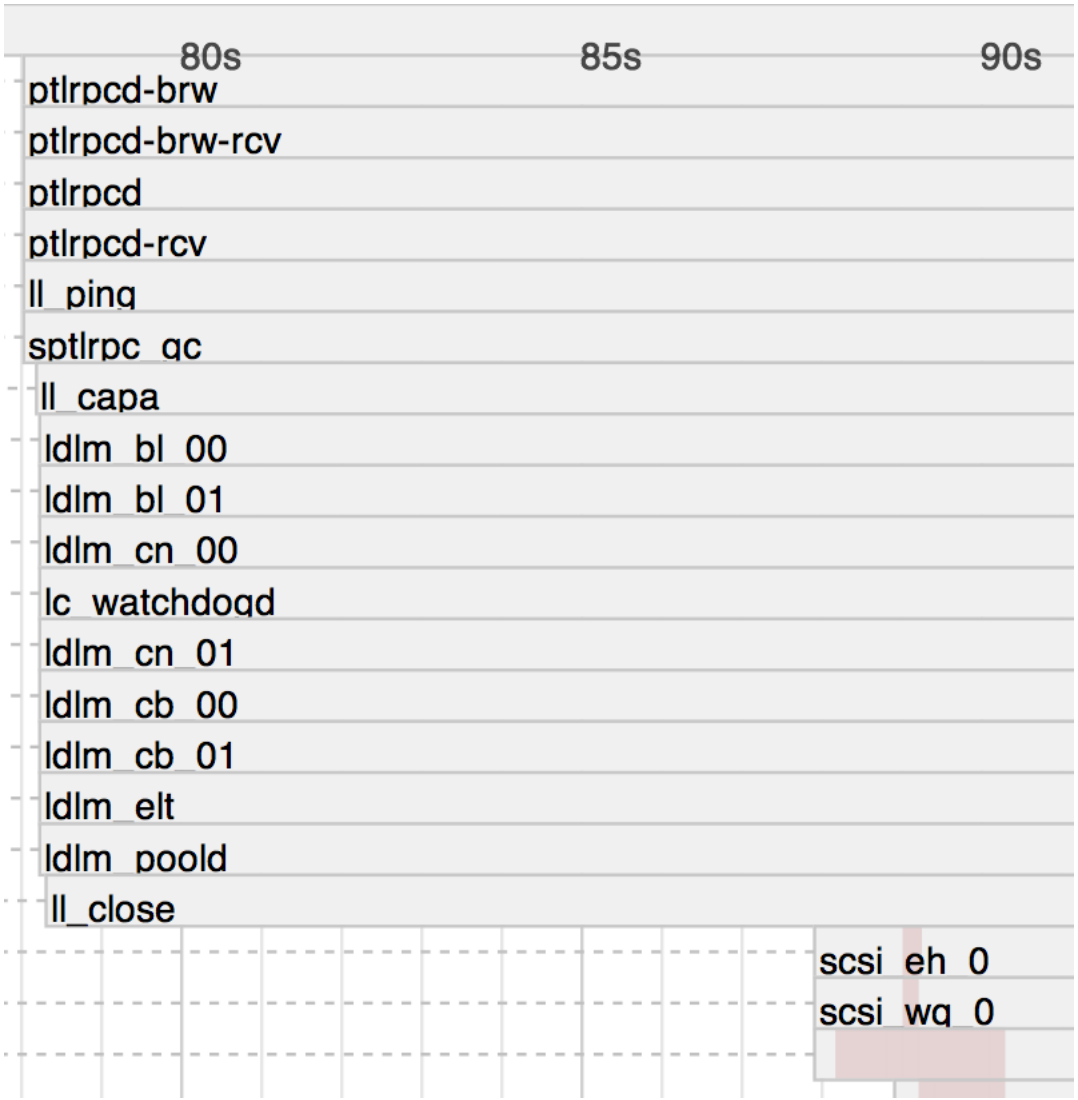
- NCI implements root-over-lustre by modifying oneSIS. Work done by Robin Humble

http://nf.nci.org.au/wiki/OneSIS/Root-on-Lustre

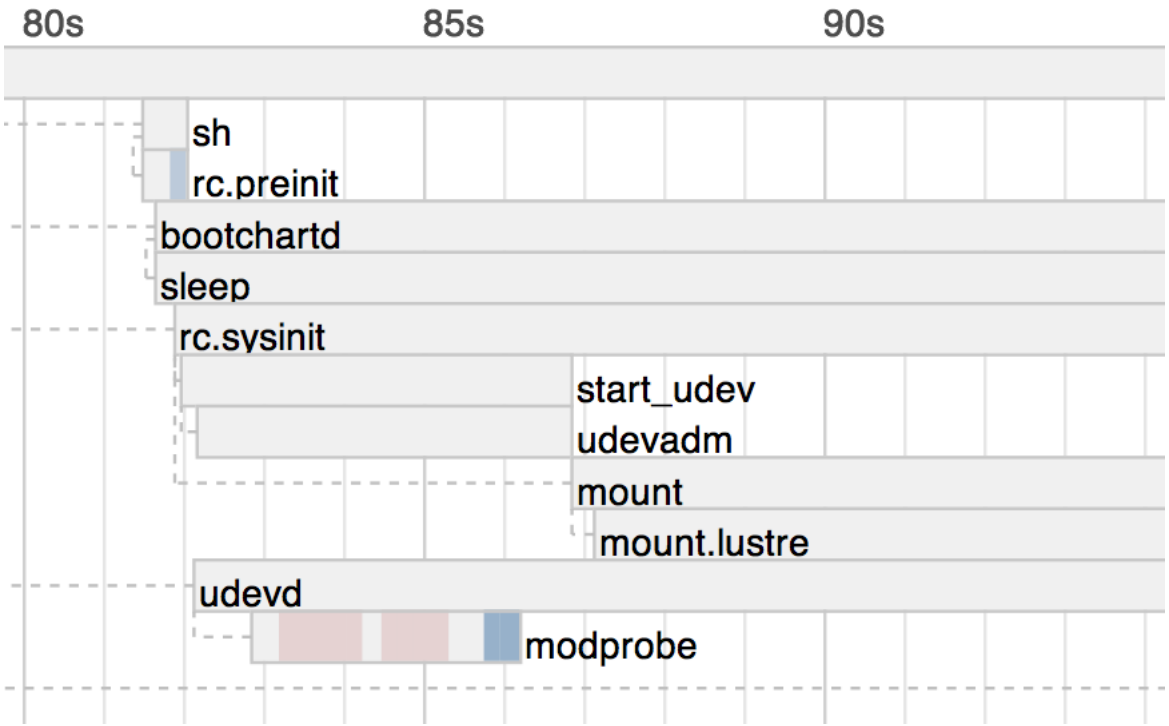

Boot chart for r1 (Thu Sep 12 17:37:46 EST 2013)
uname: Linux 2.6.32-358.14.1.el6.x86_64 #1 SMP Tue Jul 16 23:51:20 UTC 2013 x86_64
release: CentOS release 6.4 (Final)
CPU: Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz (16)
kernel options: selinux=0 exec-shield=0 audit=0 console=tty0 console=ttyS0,115200n8 ro initrd=initra
time: 3:22
5s
10s
15s
20s
init

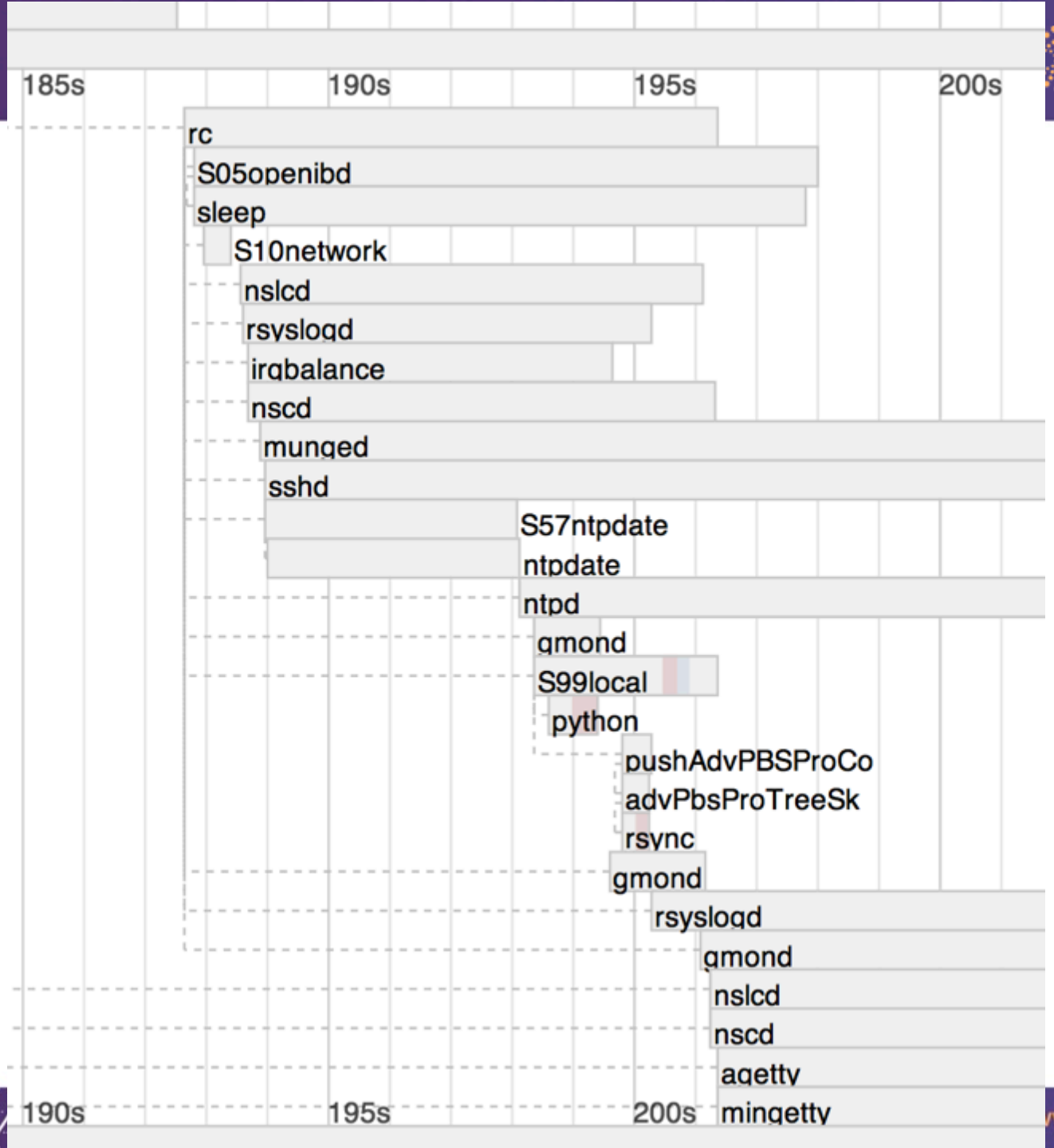
Boot Chart for Root over Lustre

Boot Chart for Root over Lustre



Boot Chart for Root over Lustre




185s
190s
195s
200s
rc
S05openibd
sleep
S10network
nslcd
rsyslogd
irqbalance
nscd
munged
sshd
S57ntpdate
ntpdate
ntpd
gmond
S99local
python
pushAdvPBSProCo
advPbsProTreeSk
rsync
gmond
rsyslogd
gmond
nslcd
nscd
agetty
190s
195s
200s
mingetty

- IB Flexboot provides boot over IB
- Initial bugs ironed out
- Planning to roll into next scheduled downtime window

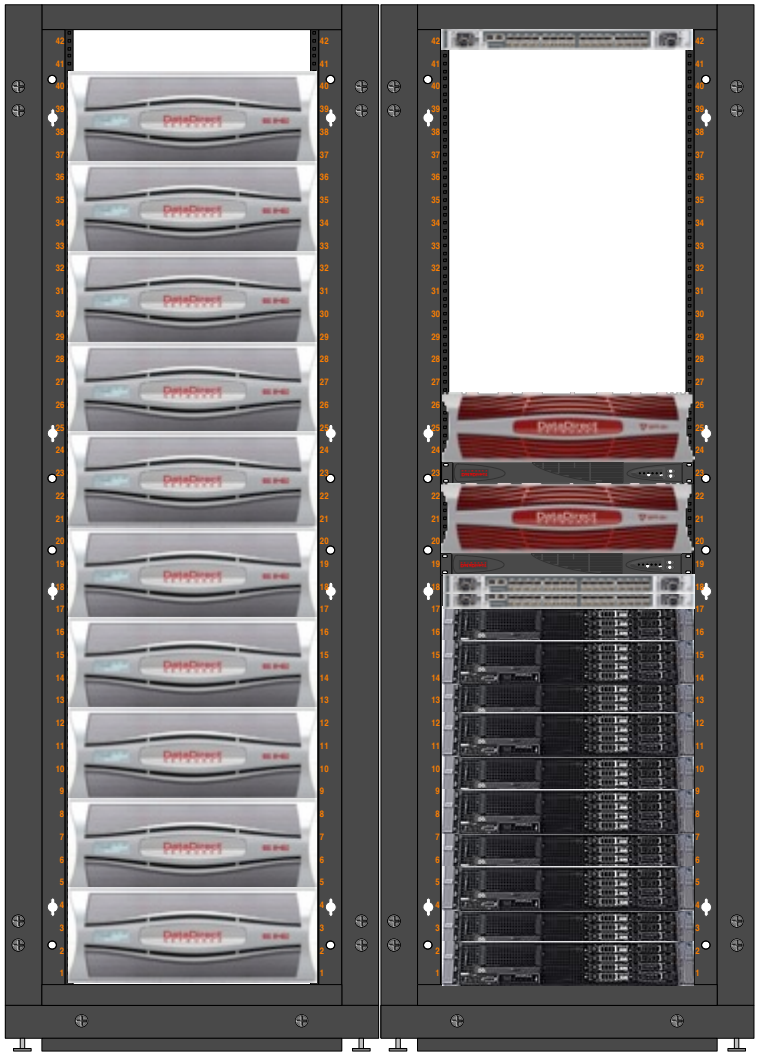
# of Nodes	Time to boot (minutes)
1 Node	6 min.
4 Nodes (1 chassis)	6 min.
72 Nodes (1 rack)	7 min. (±11 seconds)

LUSTRE ON RAIJIN

Storage Architecture of Raijin

- Storage for the Petascale machine provided by DDN SFA block appliances
- 5 storage building blocks of SFA12K40-IB with 10 x SS8460, 84 bay disk enclosures
- Each building block:
  70 x RAID6 (8+2) 3TB 7.2k SAS pools
  20 x RAID1 (1+1) 3TB 7.2k SAS pools
  40 x RAID1 (1+1) 900GB 10k SAS pools

  12 x 3TB 7.2k SAS hot spares
  8 x 900GB 10k SAS hot spares
- Building blocks scale diagonally with both capacity & performance



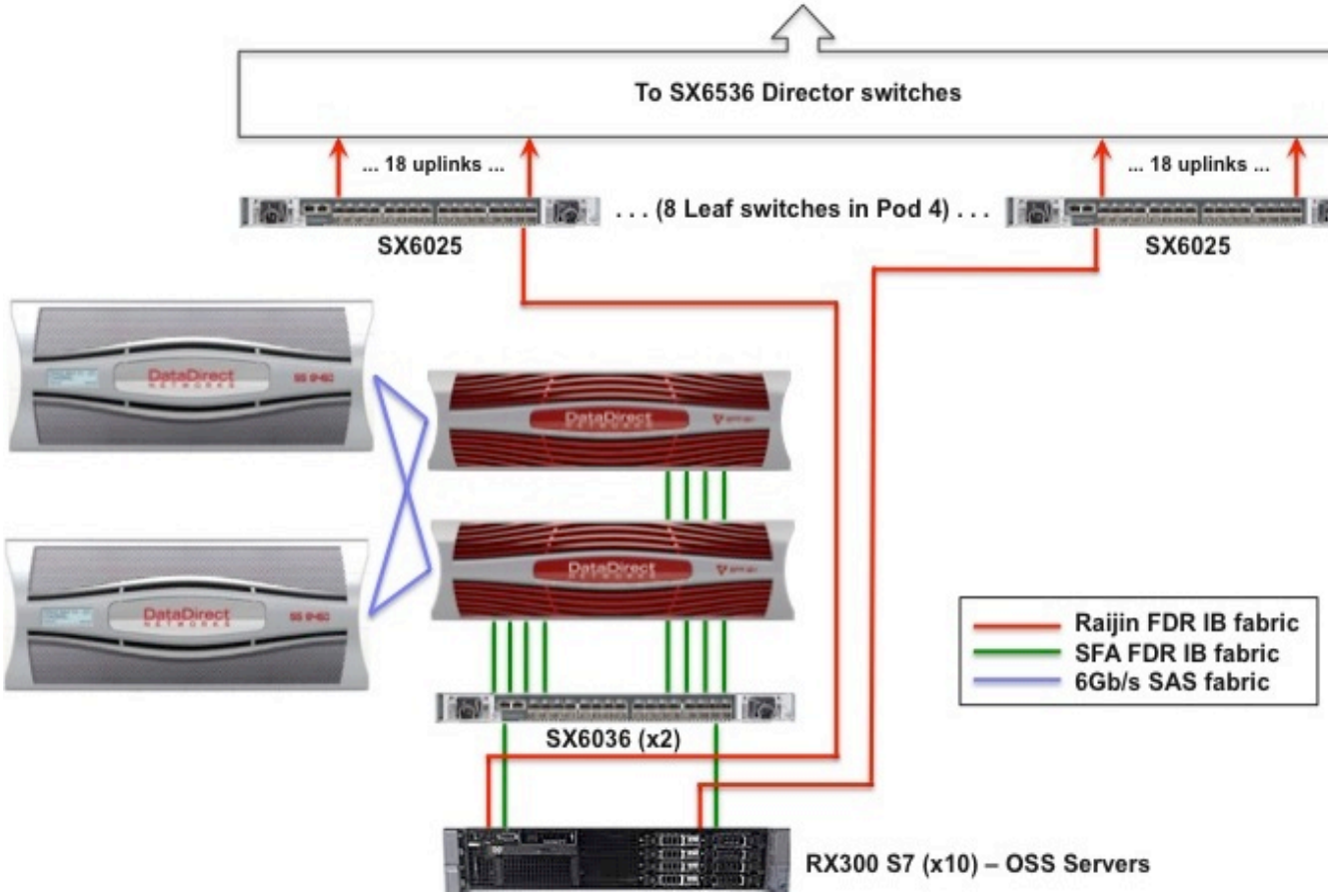
1 x SX6025

2 x SFA12K40-IB

2 x SX6036

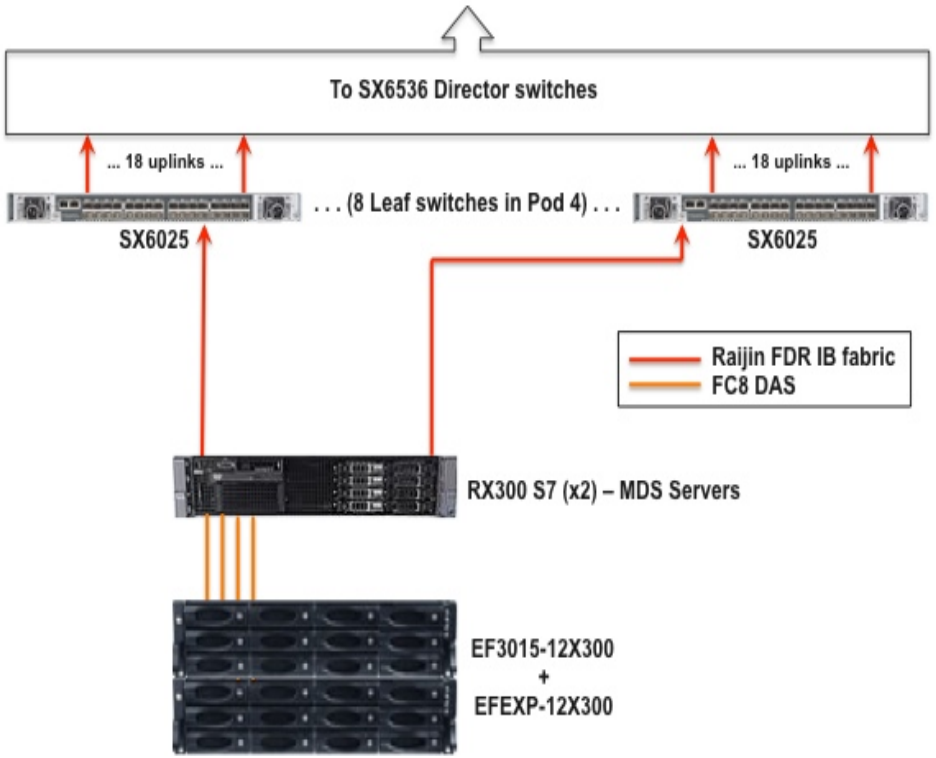
10 x OSS Servers

Lustre Network Fabrics for Storage Building Blocks



Fully redundant SAS enclosures, FDR IB between storage and OSSes and uplinks to Raijin

Lustre Network Fabrics for Metadata Building Blocks



- Metadata storage is based on the DDN EF3015 storage platform
- Each metadata storage block has 12 RAID1 (1+1) 300GB 15kSAS pools. There are 2/4 storage blocks for each MDS.
- Fully redundant Direct Attached FC8 fabric
- Fully redundant FDR IB uplinks to main cluster IB fabric

- Lustre servers are Fujitsu Primergy RX300 S7
  Dual 2.6GHz Xeon (*Sandy Bridge*) 8-core CPUs
  128/256GB DDR3 RAM

  6 MDS (3 HA pairs)
  50 OSS (25 HA pairs)

- **All Lustre servers are diskless**
  Current image is CentOS 6.3, Mellanox OFED 2.0, Lustre v 2.1.6, corosync/pacemaker
  (image was updated 8 September – simply required a reboot into new image)
  HA configuration needs to be regenerated whenever a HA pair is rebooted

- 5 Lustre file systems:
  /short – scratch file system (rw)
  /images – images for root over Lustre used by compute nodes (ro)
  /apps – user application software (ro)
  /home – home directories (rw)
  /system – critical backups, benchmarking, rw-templates (rw)

- NCI MDS requirements:

  MDT Storage on LVM on top of software RAID1 configuration of hardware RAID1 LUNs - 4-way mirror (1+1) + (1+1).

- NCI acceptance testing requirements for the scratch file system, **/short**

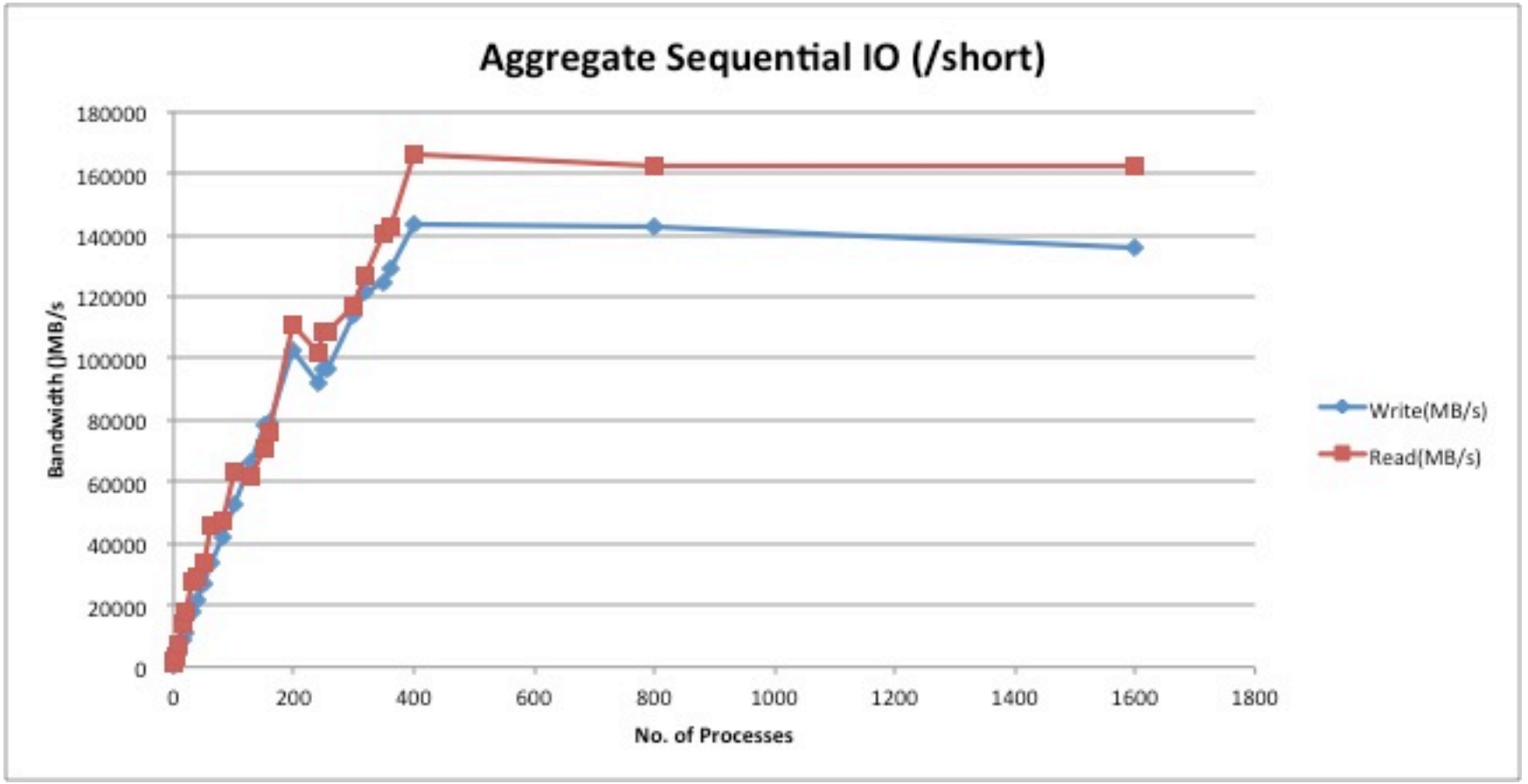
  Demonstrate IOR exceeds 120GB/s for sustained streaming write performance:
  Achieved 143 GB/s (Updated after reconfiguration 152 GB/s)

  Demonstrate IOR exceeds 7.5GB/s for random 1MB write performance:
  Achieved 75.5 GB/s

  Demonstrate mdtest test can create, stat and delete 65536 files in a shared directory within 53 seconds:
  Achieved

File Creation	3.57s
File Stat	2.88s
File Delete	6.20s
Total	**12.65s**

Lustre on Raijin: Sequential IO Performance



File System	RAID	OST/OSS	Total OST	Total Size	Performance*
/short	RAID6 (8+2) 7.2k SAS	7	350	7.5PB	152 GB/s
/images	RAID1 (1+1) 10k SAS	2	100	80TB	17.8 GB/s**
/apps	RAID1 (1+1) 10k SAS	2	100	80TB	17.9 GB/s**
/home	RAID1 (1+1) 7.2k SAS	1	50	135TB	6.9 GB/s**
/system	RAID1 (1+1) 7.2kSAS	1	50	135TB	8.1 GB/s

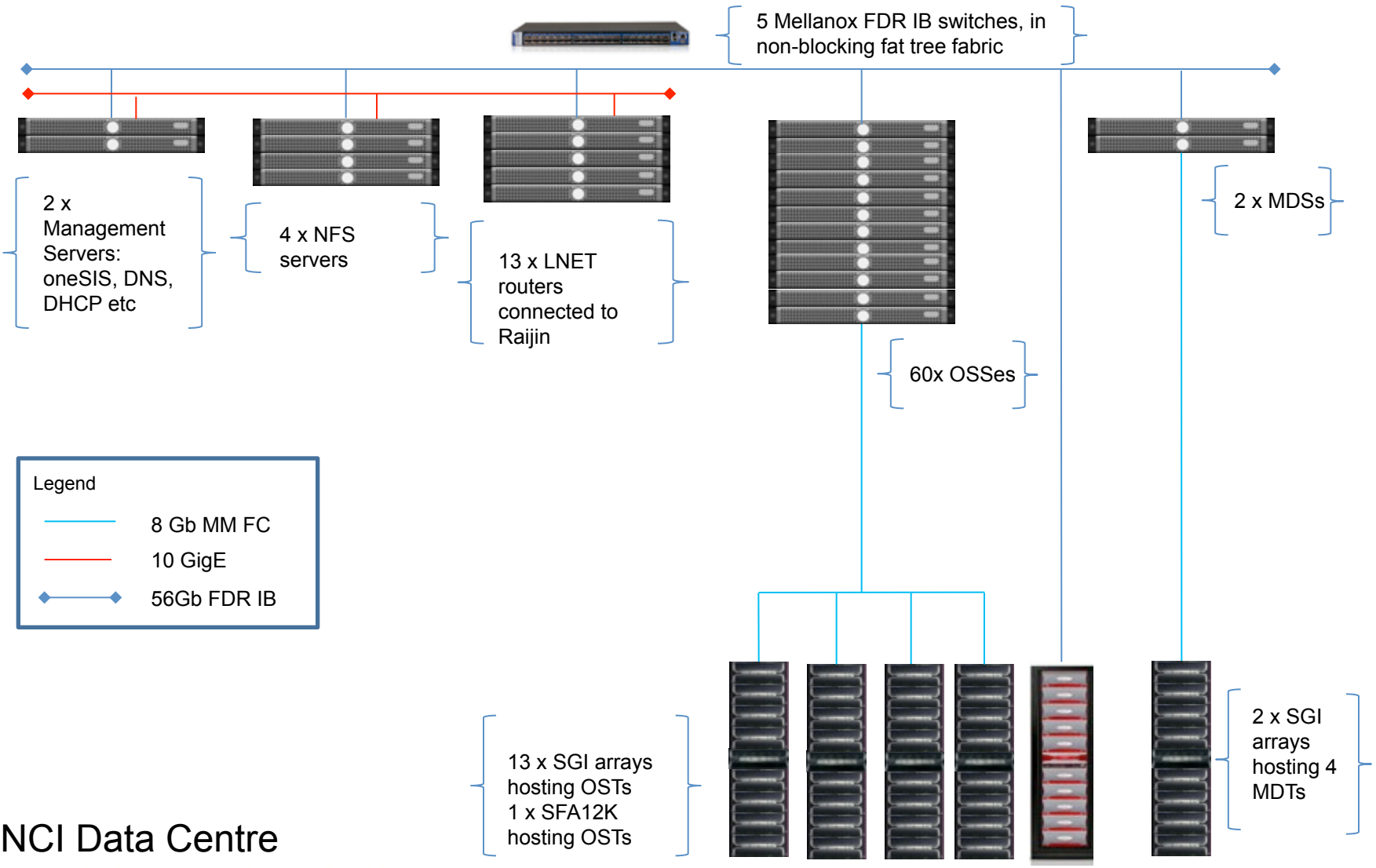
* Aggregate Sequential write bandwidth with IOR (Aug 2013)
** File system was not idle

- Currently investigating a Lustre read performance issue:
  During acceptance testing in Dec 2012 **/short** read performance was 160 GB/s.
  From later benchmarking (May 2013) **/short** read performance is 88 GB/s

SITE-WIDE LUSTRE

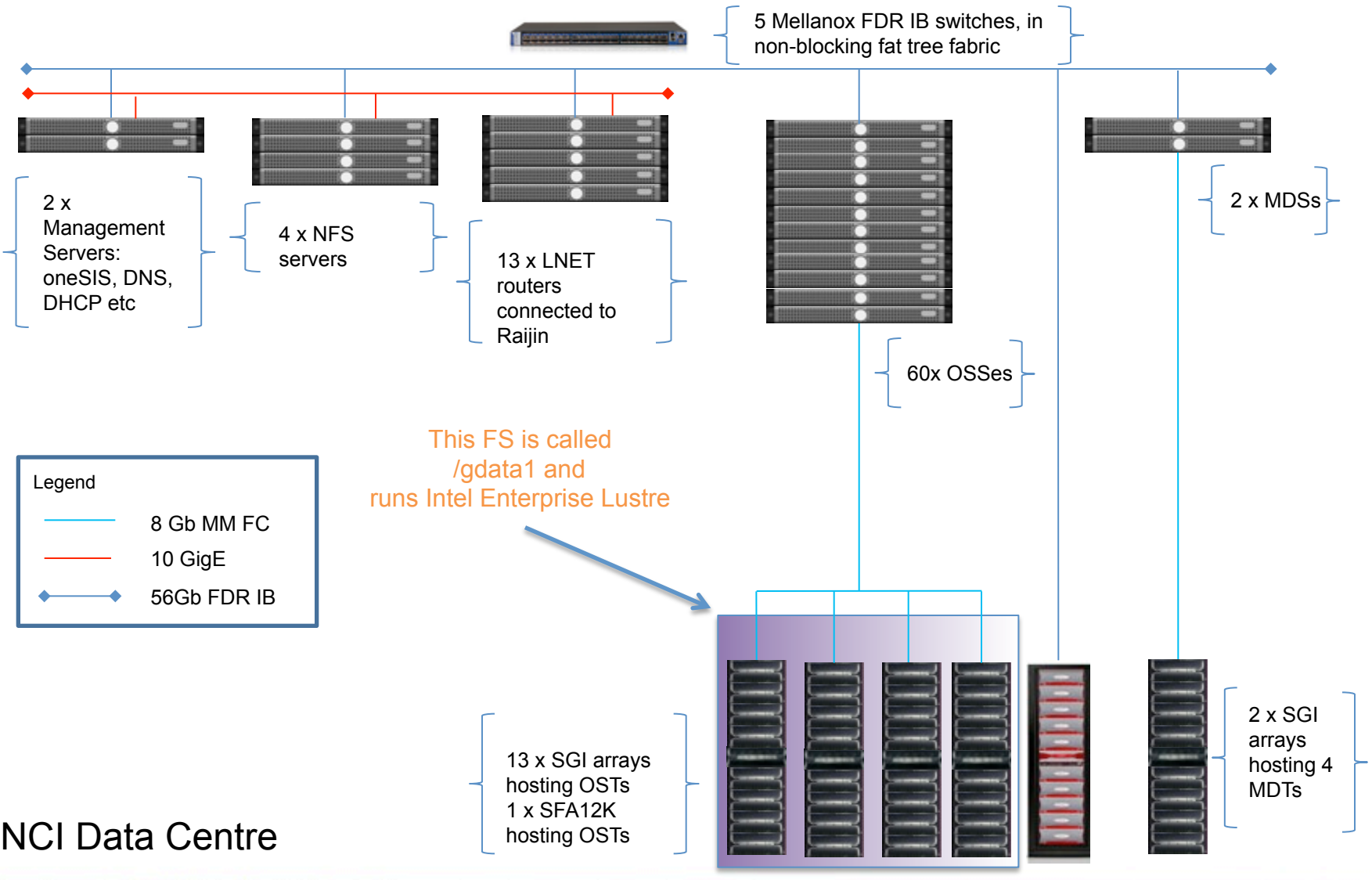
- In order to avoid moving data between clusters and storage, the NCI has implemented a site-wide Lustre F/S, visible both to compute clusters and virtual machine hosts
- We have decided to use islands of storage to create multiple Lustre F/S which are vendor/technology specific

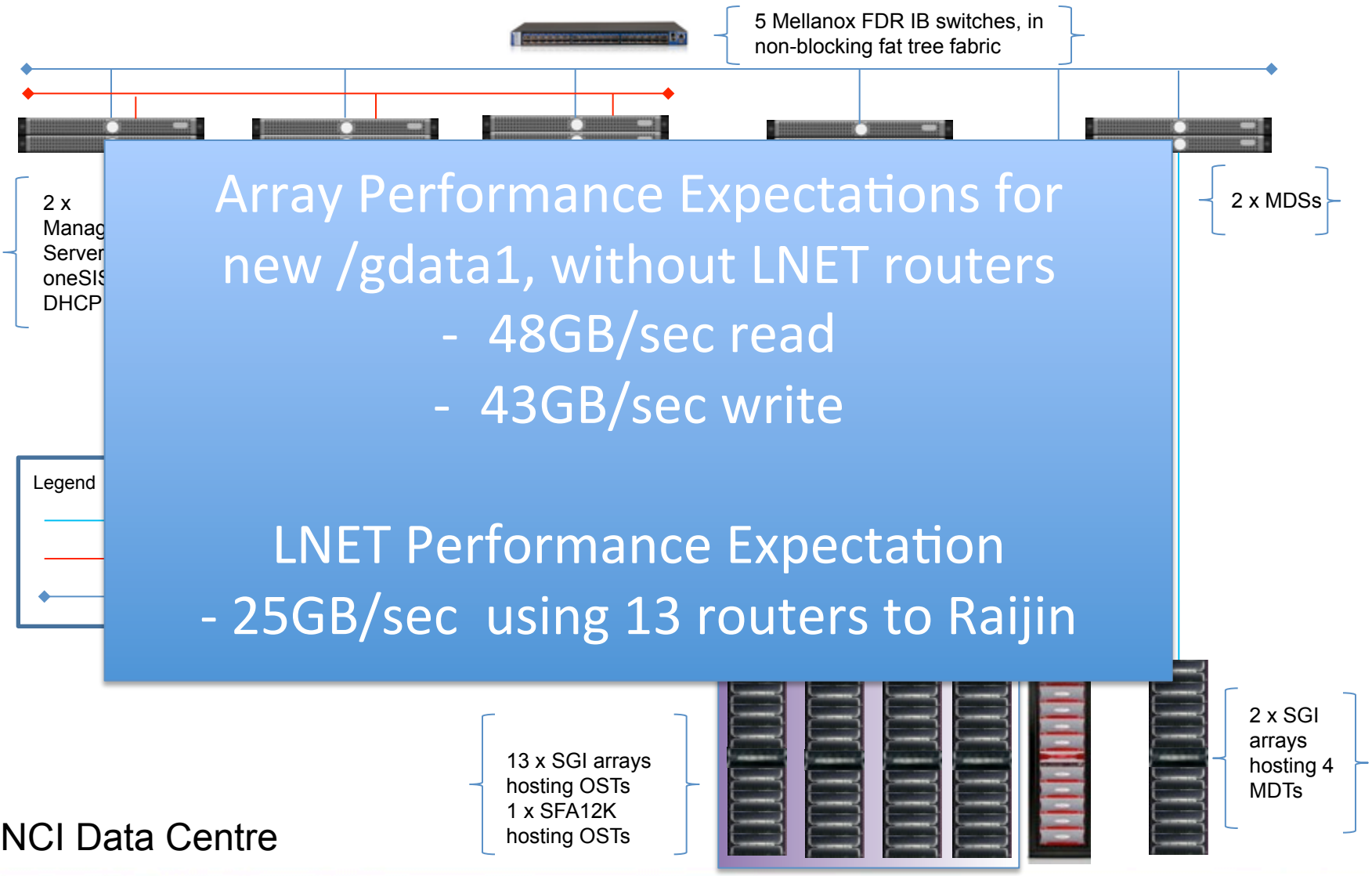
Site-wide Lustre – Functional Composition



NCI Data Centre

Site-wide Lustre – Functional Composition





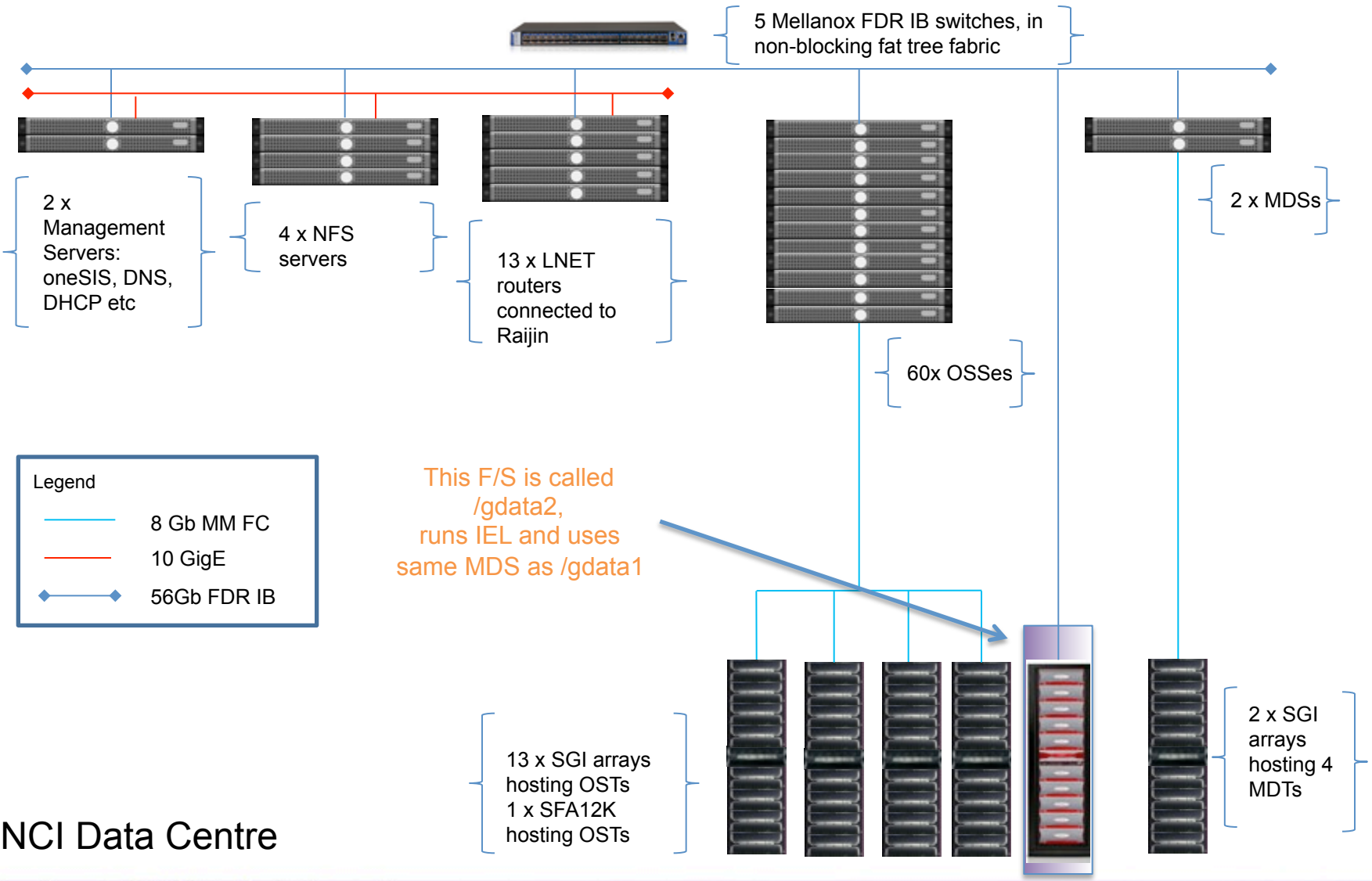
Array Performance Expectations for new /gdata1, without LNET routers

- 48GB/sec read
- 43GB/sec write

LNET Performance Expectation

- 25GB/sec using 13 routers to Raijin

Site-wide Lustre – Functional Composition



NCI Data Centre

- Site-wide Lustre to tie together HPC, Cloud and Visualization
- Complex workflows, post-simulation, will use the NCI's NeCTAR OpenStack node, and requires access to Lustre
- We are keen to implement Lustre HSM, WAN and Kerberos feature sets