NCI

NATIONAL COMPUTATIONAL INFRASTRUCTURE

Fine-grained File System Monitoring with Lustre Jobstat

Daniel Rodwell
daniel.rodwell@anu.edu.au

Patrick Fitzhenry
pfitzhenry@ddn.com

Australian Government Department of Education

Australian National University

Australian Government Bureau of Meteorology

CSIRO

Australian Government Geoscience Australia

Australian Government Australian Research Council

www.nci.org.au

Agenda

- **What is NCI**
- **Petascale HPC at NCI (Raijin)**
- **Lustre at NCI**
  - Lustre on Raijin
  - Site-wide Lustre
- **Job Aware Filesystem Monitoring**
  - Scheduler
  - Workloads
  - Challenges
  - Monitoring
  - Linking IO activity to specific jobs

WHAT IS NCI

In case you're wondering where are we located?

- In the Nation's capital, at its National University …

NCI – an overview

- 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

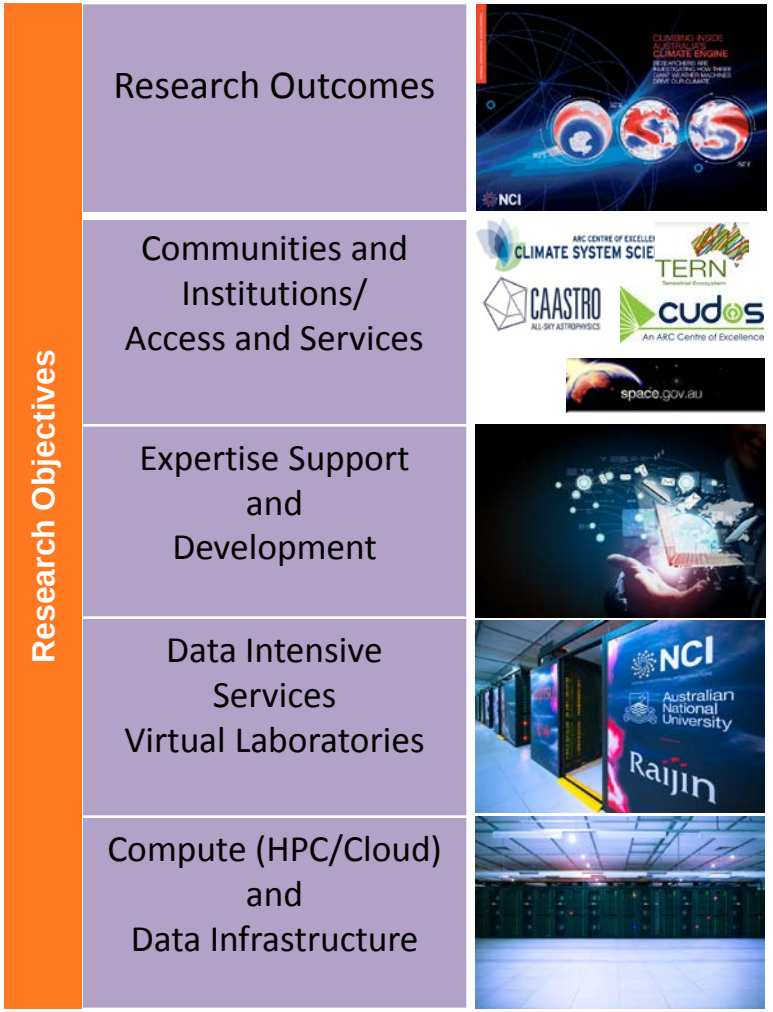
- As a partnership with a number of research-intensive universities, supported by the Australian Research Council.

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

NCI – an overview

Our mission is

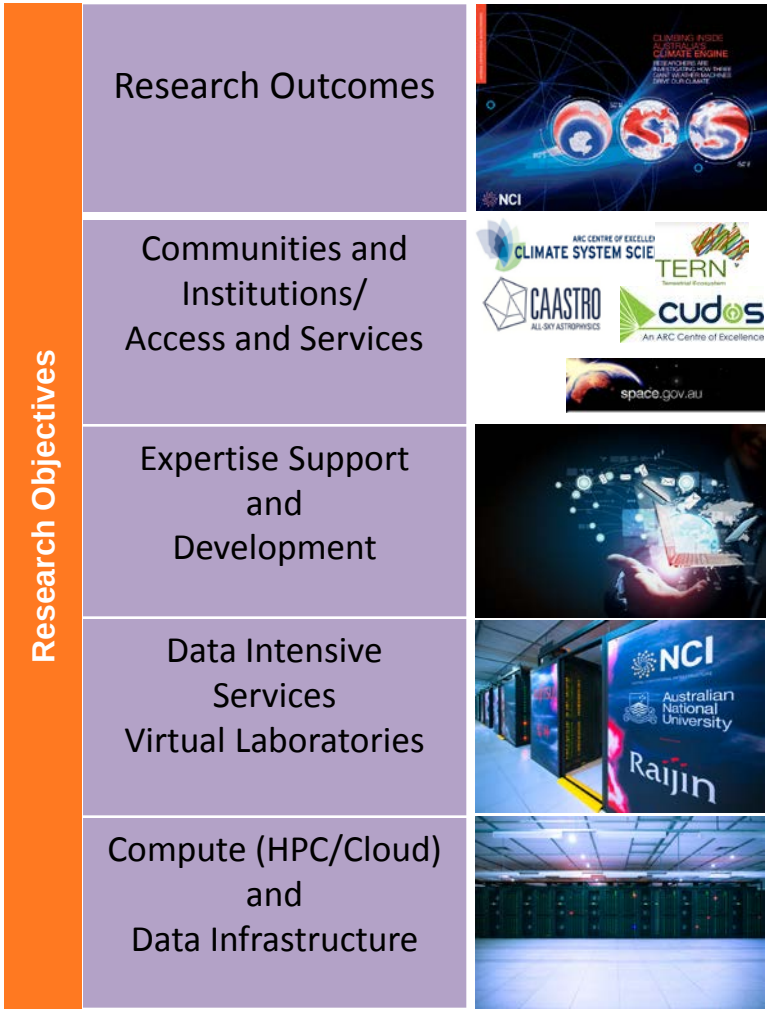
To foster ambitious and aspirational research objectives and to enable their realisation, in the Australian context, through world-class, high-end computing services.



NCI – an overview

NCI is:

- Being driven by research objectives,
- An integral component of the Commonwealth's research infrastructure program,
- A comprehensive, vertically-integrated research service,
- Engaging with, and is embedded in, research communities, high-impact centres, and institutions,
- Fostering aspiration in computational and data-intensive research, and increasing ambition in the use of HPC to enhance research impact,
- Delivering innovative "virtual laboratories",
- Providing national access on priority and merit, and
- Being built on, and sustained by, a collaboration of national organisations and research-intensive universities.



- **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

Raijin

PETASCALE HPC @ NCI

Raijin—Petascale Supercomputer

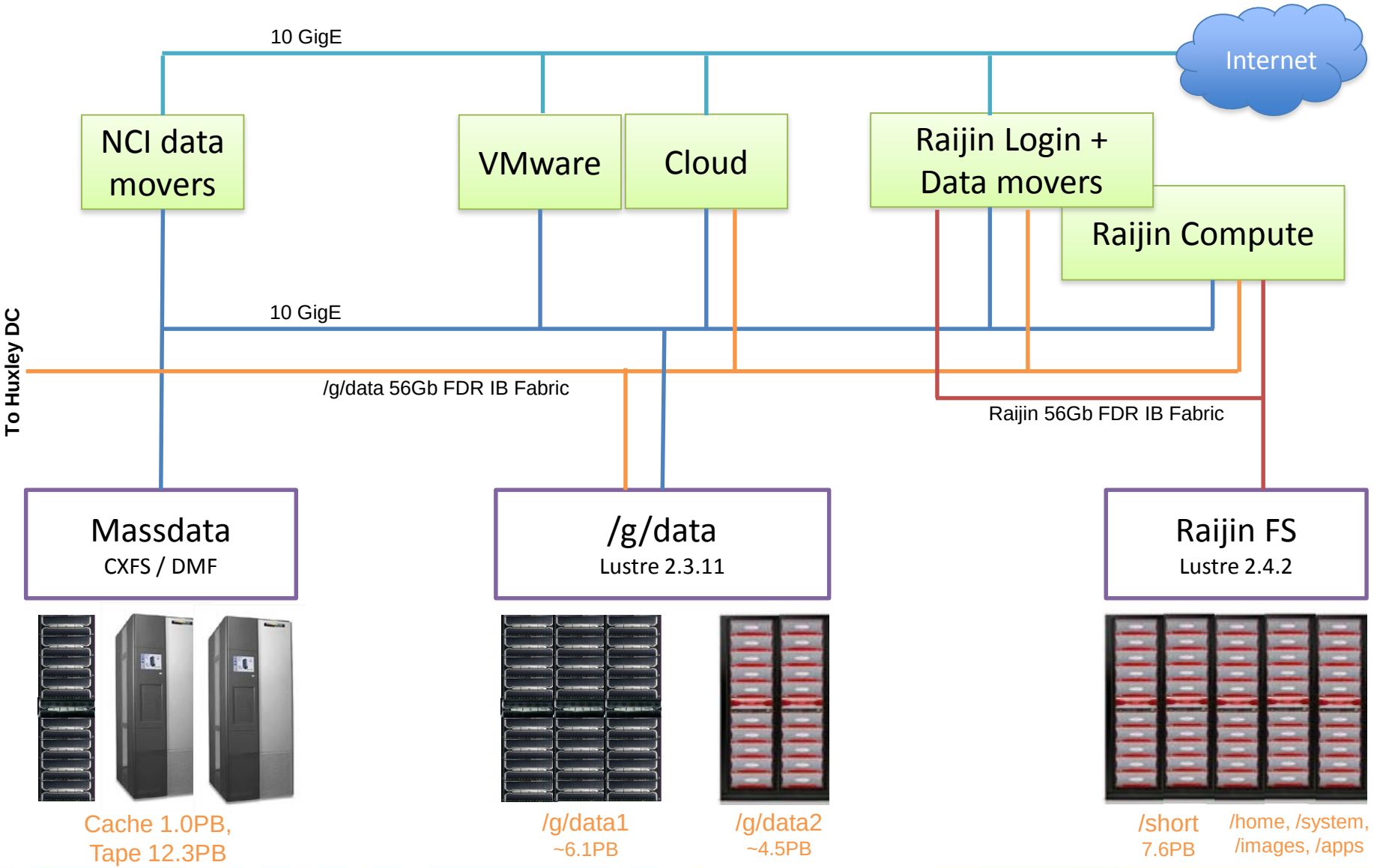
Raijin Fujitsu Primergy cluster, June 2013**:**

- 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 in debut (November 2012); first petaflop system in Australia
  - 1195 Tflops, 1,400,000 SPECFPrate
  - Custom monitoring and deployment
  - Custom Kernel, CentOS 6.4 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

LUSTRE AT NCI

Storage Overview



Raijin:/short

Filesystem:	raijin:/short (HPC short-term storage)
Type	Lustre v2.4.2 parallel distributed file system
Purpose	High Performance short term storage for peak HPC system
Capacity	7.6 PB
Throughput	Max 150GB/sec aggregate sequential write Avg 600MB/sec individual file
Connectivity	56 Gbit FDR Infiniband (Raijin compute nodes) 10 Gbit Ethernet (Raijin login nodes)
Access Protocols	Native Lustre mount (Raijin compute nodes) SFTP/SCP/Rsync-ssh (Raijin login nodes)
Backup & Recovery	No Backup or data recovery

Lustre on Raijin: Building Blocks and Composition

- 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)
40 OSS (20 HA pairs)

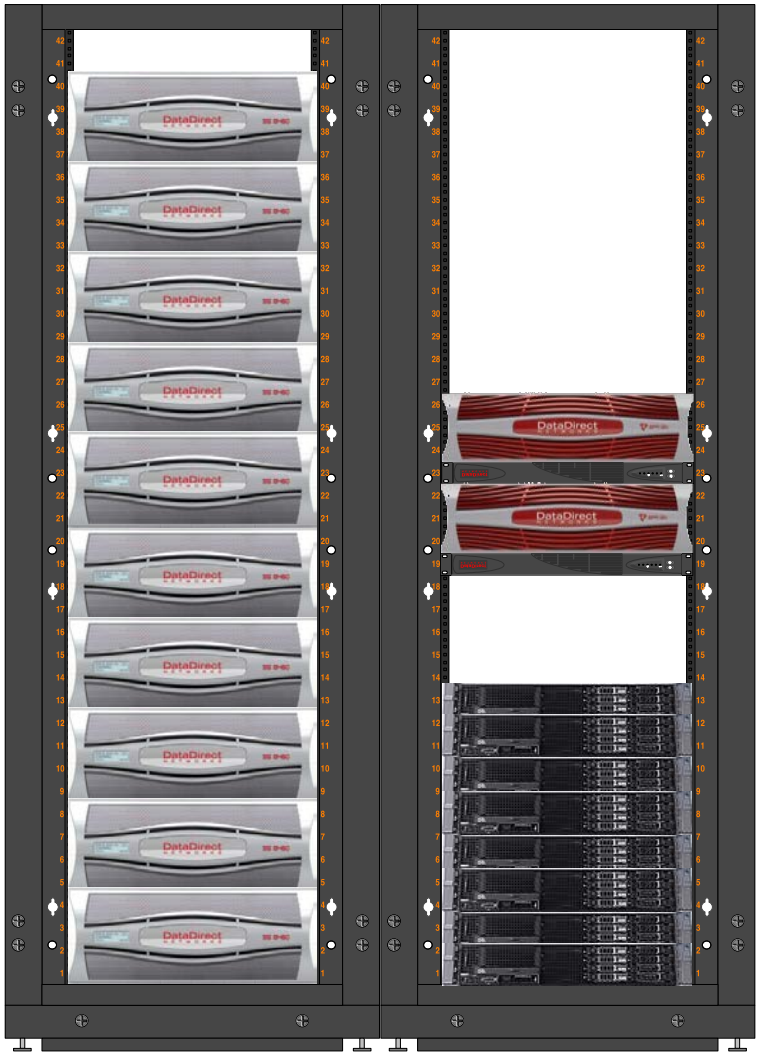
- **All Lustre servers are diskless**
Current image is CentOS 6.4, Mellanox OFED 2.0, Lustre v 2.4.2, corosync/pacemaker
(image was updated January 2014 – simply required a reboot into new image)
HA configuration needs to be regenerated whenever a HA pair is rebooted

- 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)

Storage Architecture of Raijin

- Storage for Raijin (HPC) provided by DDN SFA block appliances
- 5 storage building blocks of SFA12K40-IB with 10 x SS8460, 84 bay disk enclosures
- Each building block:
  72 x RAID6 (8+2) 3TB 7.2k SAS pools
  16 x RAID1 (1+1) 3TB 7.2k SAS pools
  32 x RAID1 (1+1) 900GB 10k SAS pools

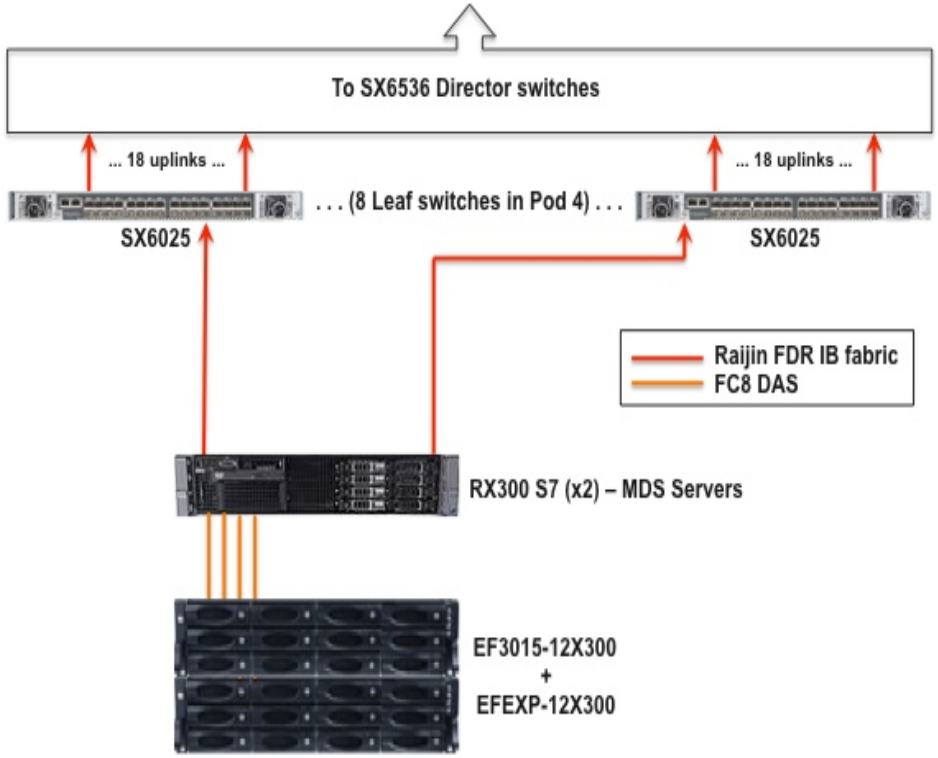
  12 x 3TB 7.2k SAS hot spares
  8 x 900GB 10k SAS hot spares
- Building blocks scale diagonally with both capacity & performance



2 x SFA12K40-IB

8 x OSS Servers

Lustre Network Fabrics for Metadata Building Blocks

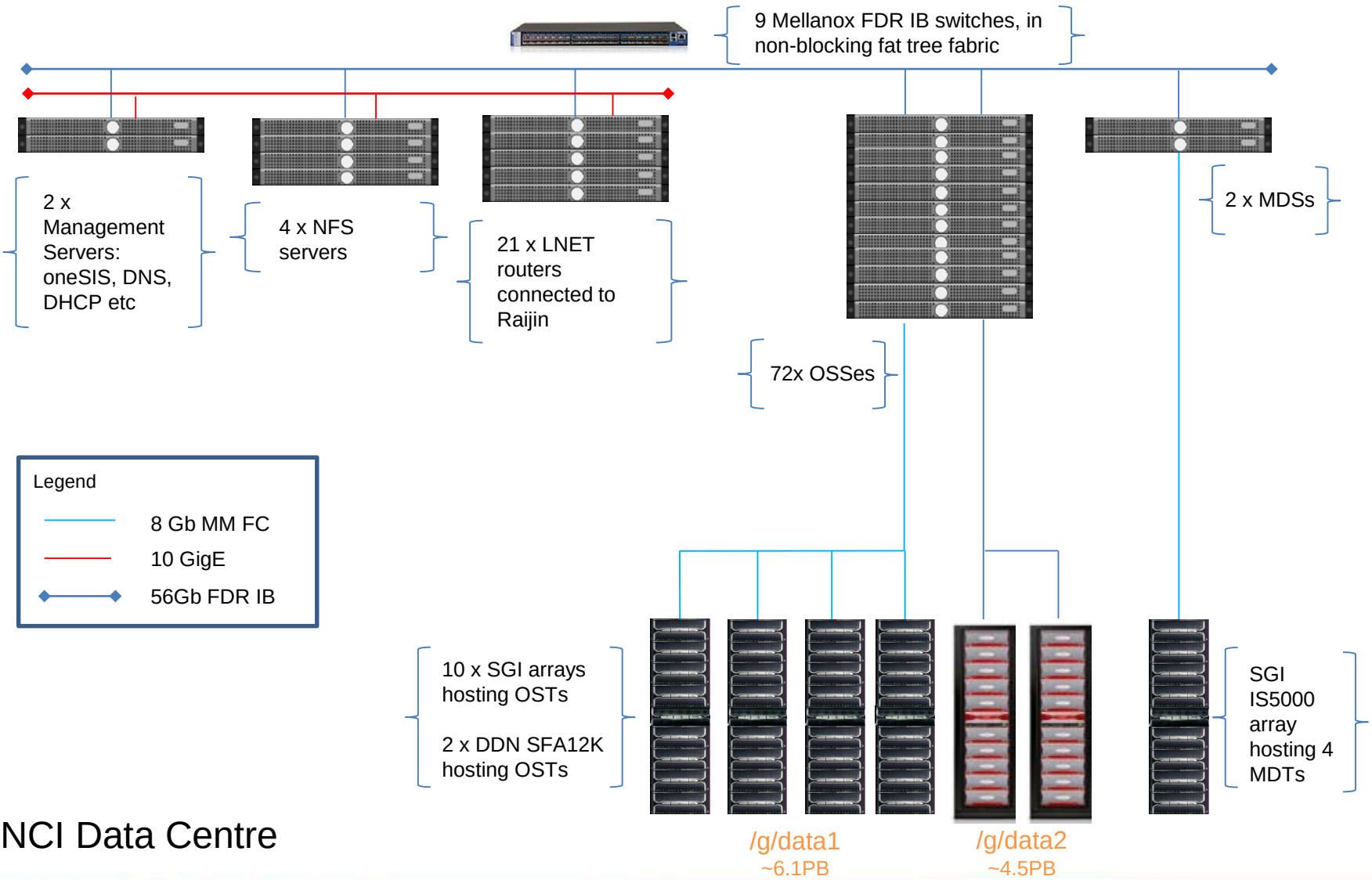


- Metadata storage is based on the DDN EF3015 storage platform
- Each metadata storage block has 12 RAID1 (1+1) 300GB 15kSAS pools. There are 2/4 storage blocks for each MDS.
- Fully redundant Direct Attached FC8 fabric
- Fully redundant FDR IB uplinks to main cluster IB fabric

/g/data

Filesystem:	/g/data (NCI Global Data)
Type	Lustre v2.3.11 parallel distributed filesystem
Purpose	High Performance Filesystem available across all NCI systems
Capacity	6.1 PB /g/data1 4.5 PB /g/data2
Throughput	/g/data1: 21GB/sec, /g/data2: 45GB/sec aggregate sequential write Avg 500MB/sec individual file
Connectivity	56 Gbit FDR Infiniband (Raijin compute nodes) 10 Gbit Ethernet (NCI datamovers, /g/data NFS servers)
Access Protocols	Native Lustre mount via LNET (Raijin compute nodes) SFTP/SCP/Rsync-ssh (NCI datamover nodes), NFSv3
Backup & Recovery	Lustre HSM, DMF backed with dual site tape copy (Q3 2014, Lustre 2.5.x)

Site-wide Lustre – Functional Composition



NCI Data Centre

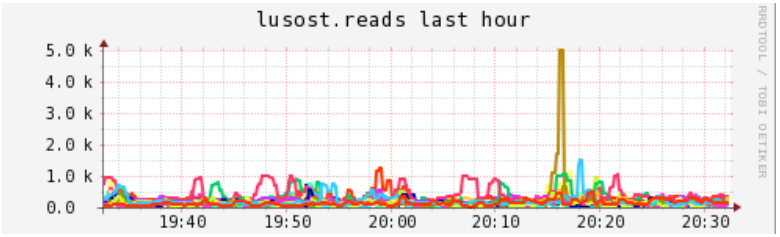
JOB AWARE LUSTRE MONITORING

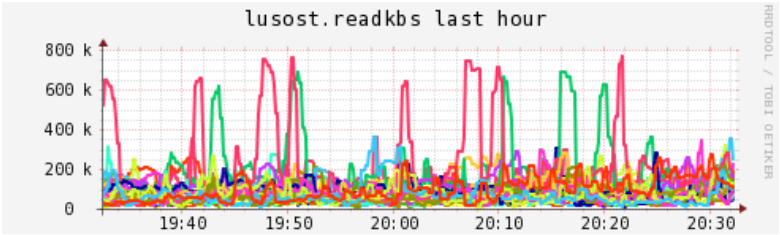
Job Aware Monitoring

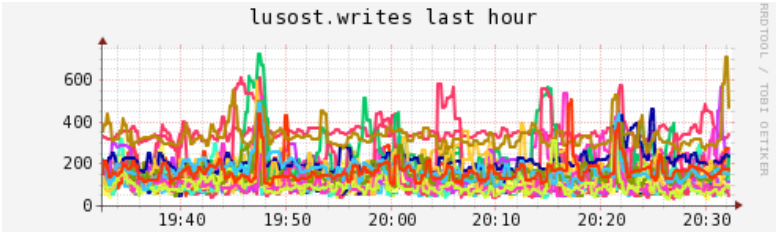
- Scheduler
  - Raijin uses a highly customised version of Altair's PBS Professional Scheduler
  - Typically 1000-2000 Concurrent jobs running, 2000-4000 queued
  - Walltime can be 100s of seconds to 100s of hours

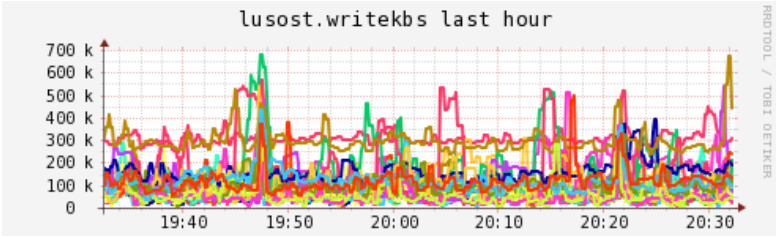
Job Aware Monitoring

- Workload
  - Significant growth in highly scaling application codes.
  - Largest: 40,000 cores; many 1,000 core tasks
  - Heterogeneous memory distribution
    - 2395 x 32GB nodes
    - 1125 x 64GB nodes
    - 72 x 128GB nodes
  - High variability in workload – applications & disciplines, cores, memory, code optimisation level.
  - Mix of optimised and suboptimal IO patterns









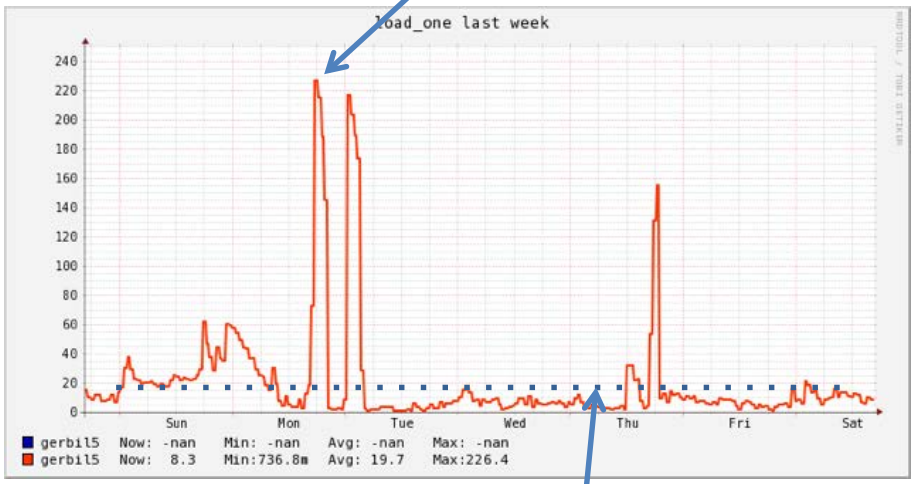
Legend: OSS activity

Job Aware Monitoring

Problem Jobs = Needle in a haystack

- Large number of concurrent jobs running
- Not all application codes are properly optimised
- A few 'badly behaved' jobs can impact on overall filesystem performance, degraded IO for all jobs
  - Compute allocation != IOPS consumption
  - Massive small IO is often detrimental to Lustre performance, particularly MDS load

/short active MDS
1m load avg = 226



/short active MDS
long term load avg ~20

Job Aware Monitoring

DDN has a current project for Job Aware Lustre Monitoring (Suichi Ihara, Xi Li, Patrick Fitzhenry) with NCI as an active collaborator (Daniel Rodwell, Javed Shaikh).

The project aims to develop a centralised Lustre statistics monitoring and analysis tool:

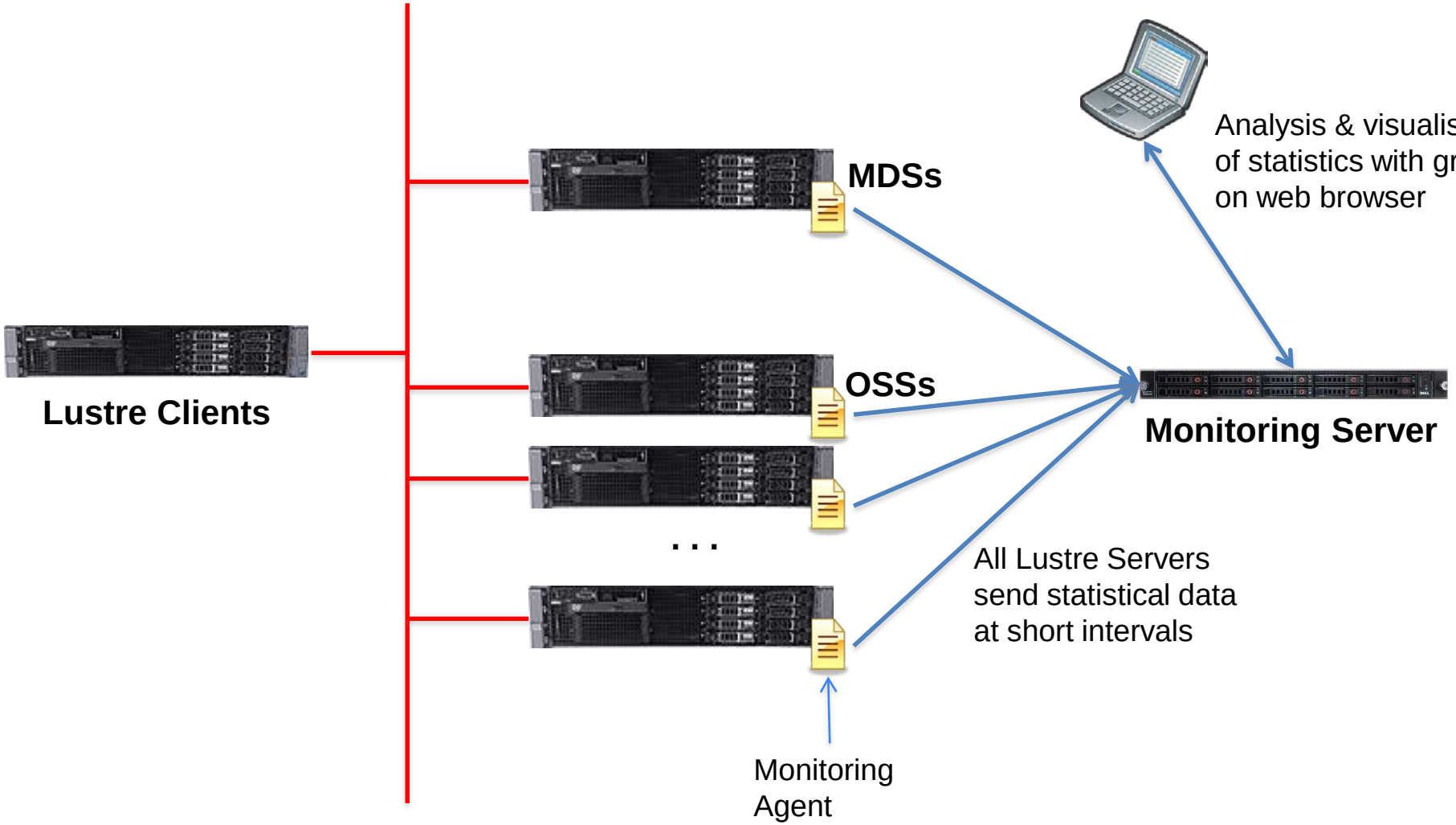
- Collect near real-time data (minimum of every 1 sec.) and visualise them
- All Lustre statistical information can be collected
- Support Lustre v1.8.x, v2.x and beyond
- Application Aware Monitoring (Job statistics)
- Administrators can make custom graphs within a web browser
- Easy to use dashboard
- Scalable, Lightweight and without performance impact

Job Aware Monitoring

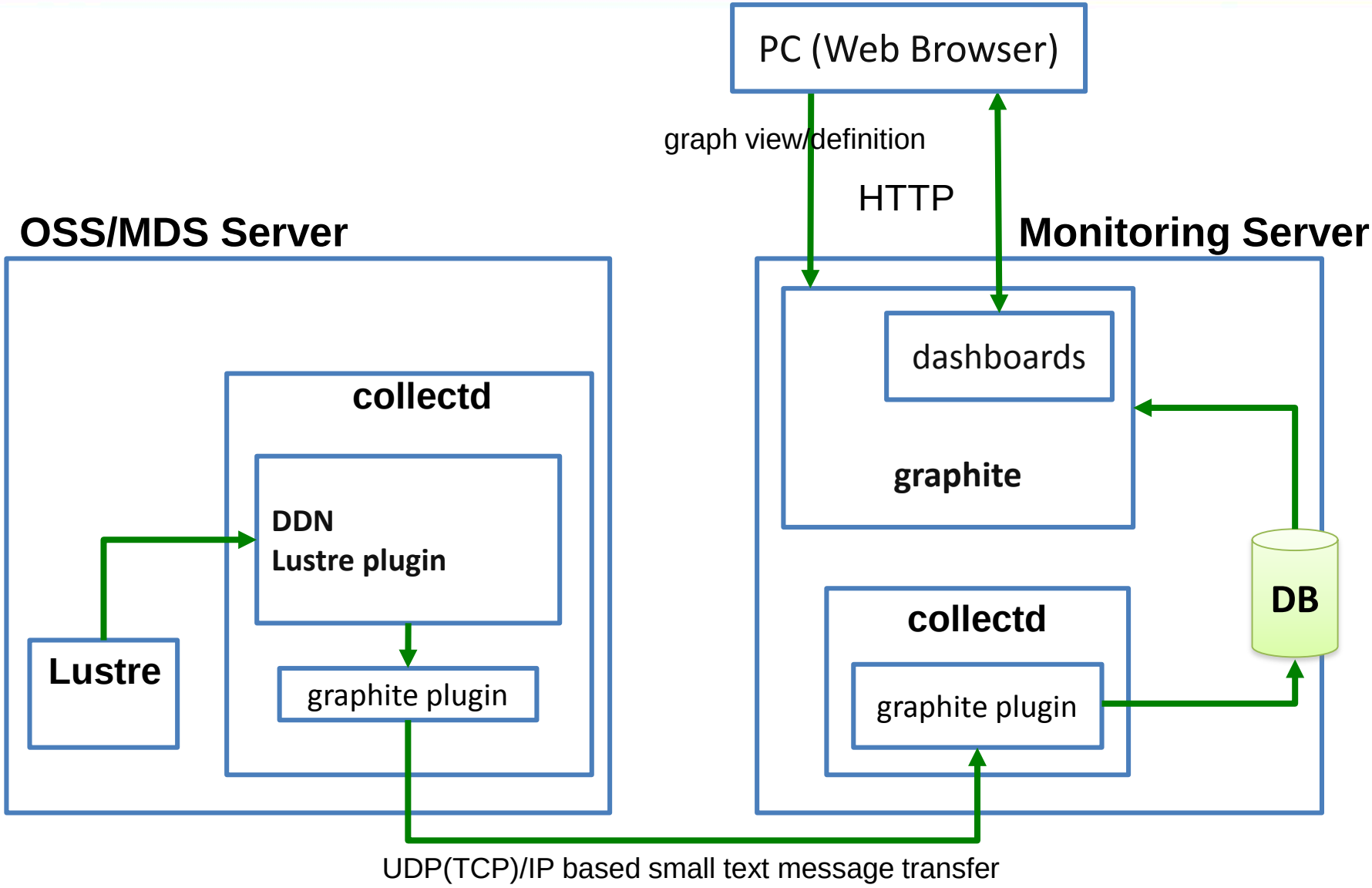
Lustre keeps a lot of statistical information in Linux's /proc and it can be quite helpful for debugging and I/O analysis.

Since Lustre v2.3, we can collect Lustre I/O statistics per JOBID. Administrators can see very detailed and application aware I/O statistics, as well as Cluster wide I/O statistics.

Monitoring System Architecture



Software Stack



Collectd

We chose collectd (http://collectd.org)

- Running on many Enterprise/HPC systems
- Written in C for performance and portability
- Includes optimisations and features to handle 10,000s data sets
- Comes with over 90 plugins which range from standard to very specialised and advanced topics
- Provides powerful networking features and is very extensible
- Actively developed and well documented

The Lustre Plugin extends collectd to collect Lustre statistics while inheriting its advantages

It is possible to port the Lustre Plugin to a different framework if necessary

XML Definition of Lustre's */proc* Information

Provides tree structured descriptions about how to collect statistics from Lustre */proc* entries

Modular

- A hierarchical framework comprised of a core logic layer (Lustre plugin) and statistics definition layer (XML files)
- Extendable without the need to update any source code of the Lustre plugin
- Easy to maintain the stability of the core logic

Centralised

- A single XML file for all definitions of Lustre data collection
- No need to maintain massive error-prone scripts
- Easy to verify correctness
- Easy to support multiple versions and update to new versions of Lustre

XML Definition of Lustre's */proc* Information

Precise

- Strict rules using regular expression can be configured to filter out all but what we exactly want
- Locations to save collected statistics are explicitly defined and configurable

Powerful

- Any statistic can be collected as long as there are proper regular expressions to match it

Extensible

- Any newly required statistics can be collected rapidly by adding definitions in the XML file

Efficient

- No matter how many definitions are predefined in the XML file, only in-use definitions will be traversed at run-time

Lustre Plugin Configuration

Structured configuration about which statistics to collect

```
LoadPlugin lustre
<Plugin "lustre">
        <Common>
                # The location of definition file
                DefinitionFile "/etc/lustre_definition.xml"
        </Common>
        <Item>
                # A unique item type defined in XML file which describes how to collect
                Type "ost_jobstats"
                # A Rule to determine whether to collect or not
                <Rule>
                        # One of the fields of collected data defined in XML file
                        Field "job_id"
                        # Any regular expression which determines whether the field matches
                        Match "dd_[[:digit:]][.][[:digit:]]+"
                </Rule>
                # Could configure multiple rules for logical conjunction
        </Item>
</Plugin>
```

Graphite - Analysis of Statistical Information

We chose Graphite/Carbon(http://graphite.wikidot.com/)

- An open source software
- Any graph can be made from collected statistics
- Each statistic is kept in its own binary file – file system(Cluster-wide), OST Pool, each OST/MDT statistics, etc…
- JOB ID, UID/GID, aggregation of application's statistics, etc…
- Archive of data by policy (e.g. 1 sec interval for one week, 10 sec for a month, 1 min for a year)
- It can export data to CSV, MySQL, etc…
- Hierarchical architecture is possible for large scale

User and Job Aware Monitoring

Where To From Here ?

We have seen that we can monitor the I/O of specific applications using the Lustre JobStats.

In the NCI context the next step is to integrate the PBSPro resource manager with this monitoring system. This will allow faster identification of user jobs with pathological I/O behaviour

NCI has a broader aim for developing a highly integrated monitoring system across the whole facility. This needs to encompass the high speed interconnects, the compute and the storage subsystems.

Other goals will be alerting and automated suspension & quarantine of problematic jobs.

Questions?

Thank You !