An ESS implementation in a Tier 1 HPC Centre

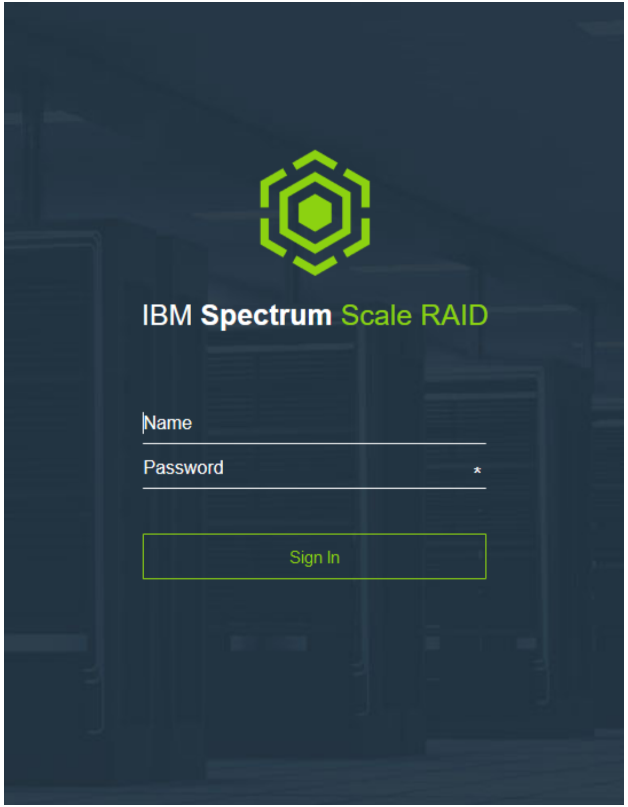
Maximising Performance - the NeSI Experience

José Higino (NeSI Platforms and NIWA, HPC Systems Engineer)

New Zealand eScience Infrastructure

Outline

- What is NeSI?
- The National Platforms Framework
- Our Multicluster
- I/O Performance Upgrade
- Dual Cluster Structure
- Single point of management (EMS)
- Finally a good Web Interface (GUI)
- ILM Policies and REST API
- Integrating Spectrum Scale with SR-IOV / OpenStack
- Protocol Nodes using OpenStack VMs
- ESS flash rebuilds using GNR
- Benchmarks



What is NeSI?

- Infrastructure and Services for Advanced Research Computing
  - High performance computing and data analytics services
  - Data, scientific consultancy, and training services
- Funding Institutions
  - NIWA, UoA, UoO, LR
- Available to the NZ Research Sector

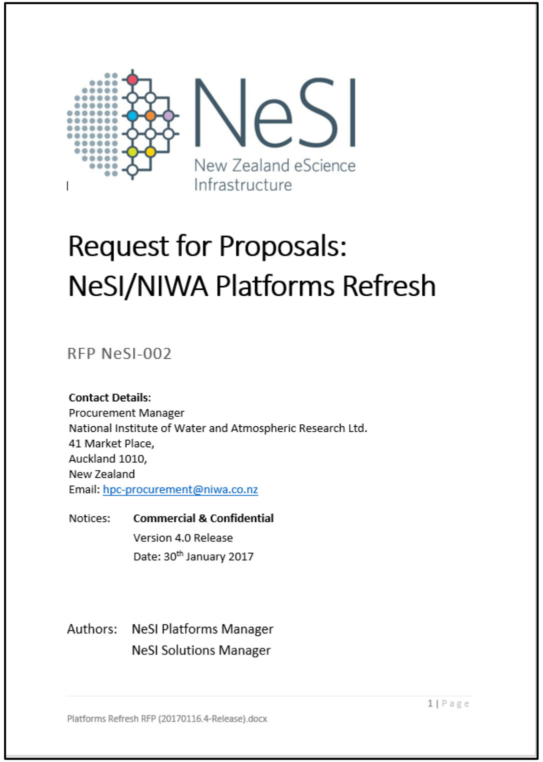
NeSI
New Zealand eScience Infrastructure

The Power Behind Researchers

Growing the computing capability of New Zealand researchers to ensure our future prosperity

Implementing the National Platforms Framework

- Maximise value through combined investment:
  - One RFP for 3 HPC Systems;
  - Single Site (Greta Point);
  - Capacity & Capability HPCs share same Storage.
- Minimise data movement:
  - Pre and Post processing services;
  - Virtual Labs & Remote Visualisation;
  - HPC Data Analytics software stack;
  - Offline storage.
- The “Data Centre” becomes a “Centre of Data”.



NeSI
New Zealand eScience
Infrastructure

Request for Proposals:
NeSI/NIWA Platforms Refresh

RFP NeSI-002

Contact Details:
Procurement Manager
National Institute of Water and Atmospheric Research Ltd.
41 Market Place,
Auckland 1010,
New Zealand
Email: hpc-procurement@niwa.co.nz

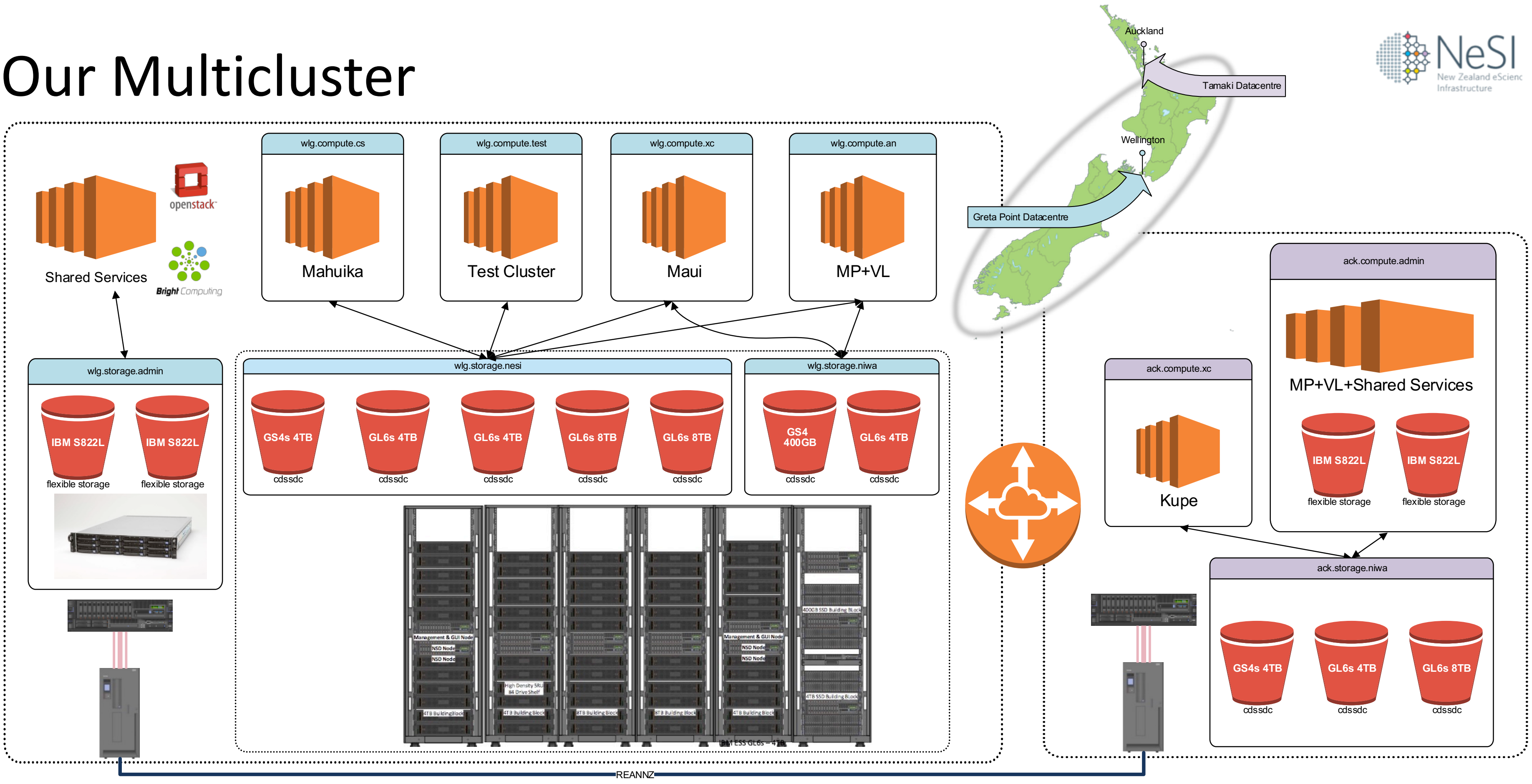
Notices: **Commercial & Confidential**
Version 4.0 Release
Date: 30th January 2017

Authors: NeSI Platforms Manager
NeSI Solutions Manager

1 | Page

Platforms Refresh RFP (20170116.4-Release).docx

Our Multicluster

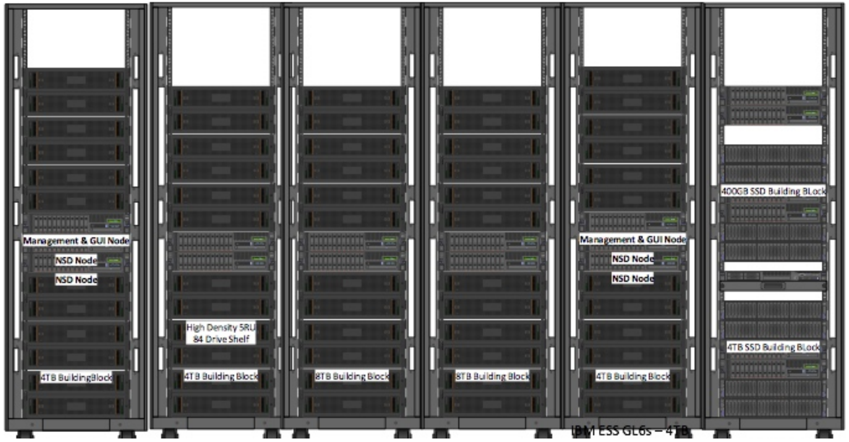


NeSI New Zealand eScience Infrastructure

NIWA Taihoro Nukurangi

I/O Performance Upgrade

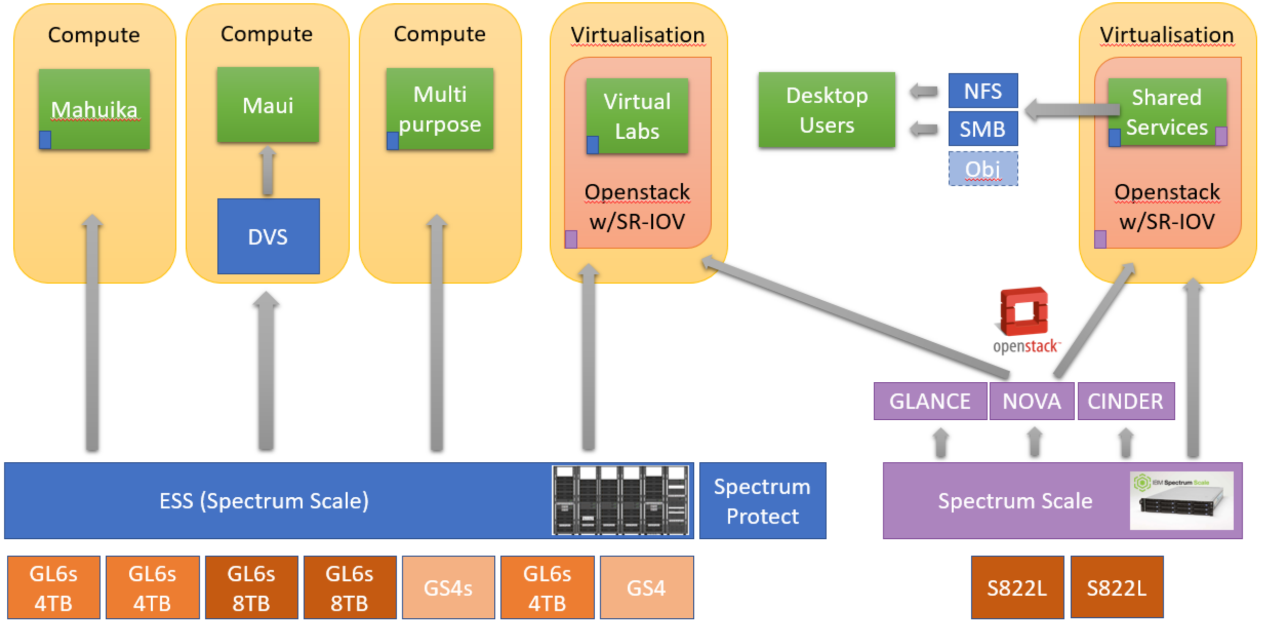
- Old DCS9900 (1500 disks), New ESS (2500 disks), excluding SSDs
- Bandwidth to Disk
  - Previous storage systems:
    - (DSxxxx models) Pan: ~1 GB/s;
    - (DCS9900) FitzRoy: ~8 GB/s;
  - **New shared storage: >165 GB/s.**



- Metadata performance (4KB)
  - Previous storage systems:
    - (DSxxxx SSDs) Pan: ~3K file creates/s;
    - (V7000 SSDs) FitzRoy: ~2.5K file creates/s;
  - **New shared storage >200K file creates/s.**
- New SSD storage pools (>132TB) – Multipurpose/Services
- 8MB (16MB) Filesystem Block Size (previous systems had 1MB and 4MB)

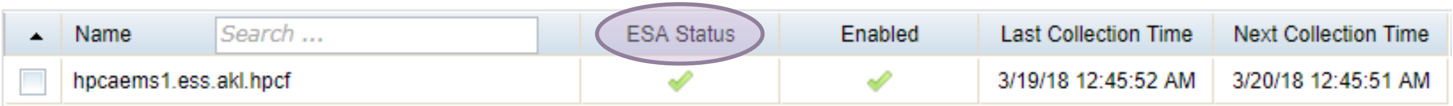
Dual Cluster Structure

- Flexible Storage (S822L)
  - ✓ Provisioning OSes (OpenStack VMs)
  - ✓ Databases (persistent)
- Main Data Storage (ESS)
  - ✓ VM access via SR-IOV
  - ✓ Direct access
  - ✓ HSM Filesystem
  - ✓ Cray DVS nodes
  - ✓ Other Clusters (Protocol Nodes)

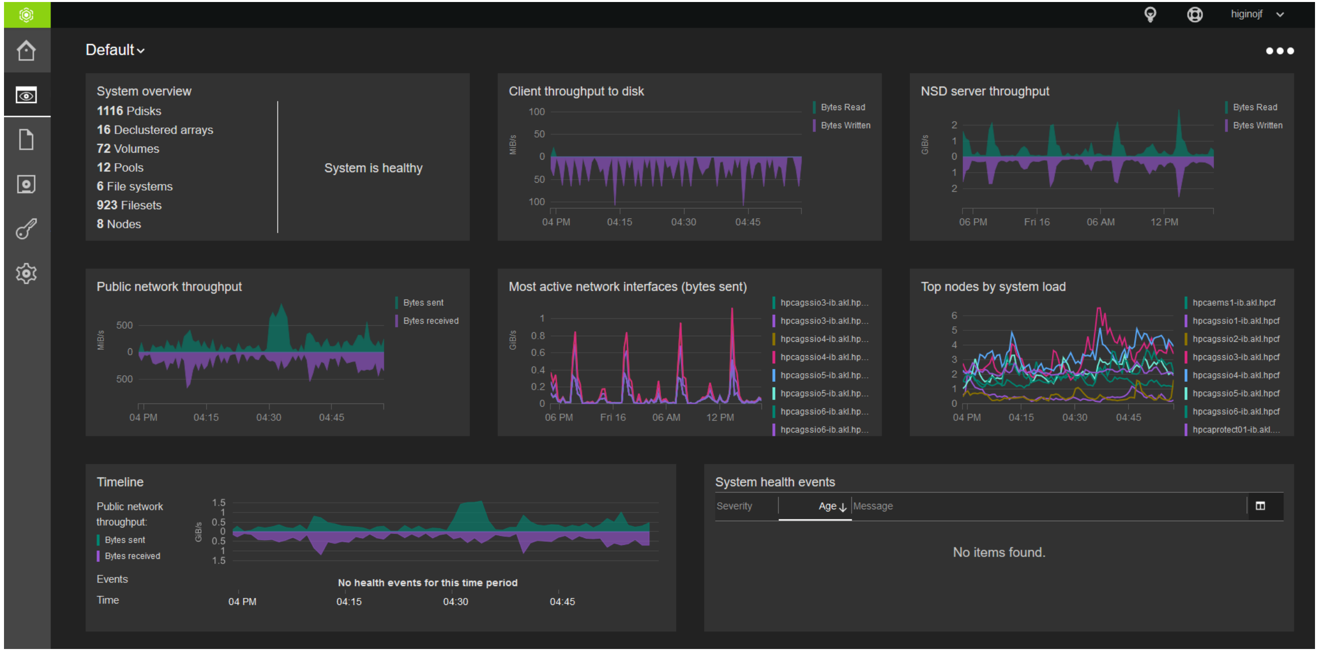


EMS (Single point of management)

- GUI Server
- xCAT based
- ESS Deployement
- Monitoring
  - ✓ Performance
  - ✓ Events/Faults
  - ✓ Advices
- Call Home (ESA)
- Upgrades



▲	Name	ESA Status	Enabled	Last Collection Time	Next Collection Time
□	hpcaems1.ess.akl.hpcf	✓	✓	3/19/18 12:45:52 AM	3/20/18 12:45:51 AM



System overview
1116 Pdisks
16 Declustered arrays
72 Volumes
12 Pools
6 File systems
923 Filesets
8 Nodes

System is healthy

System health events

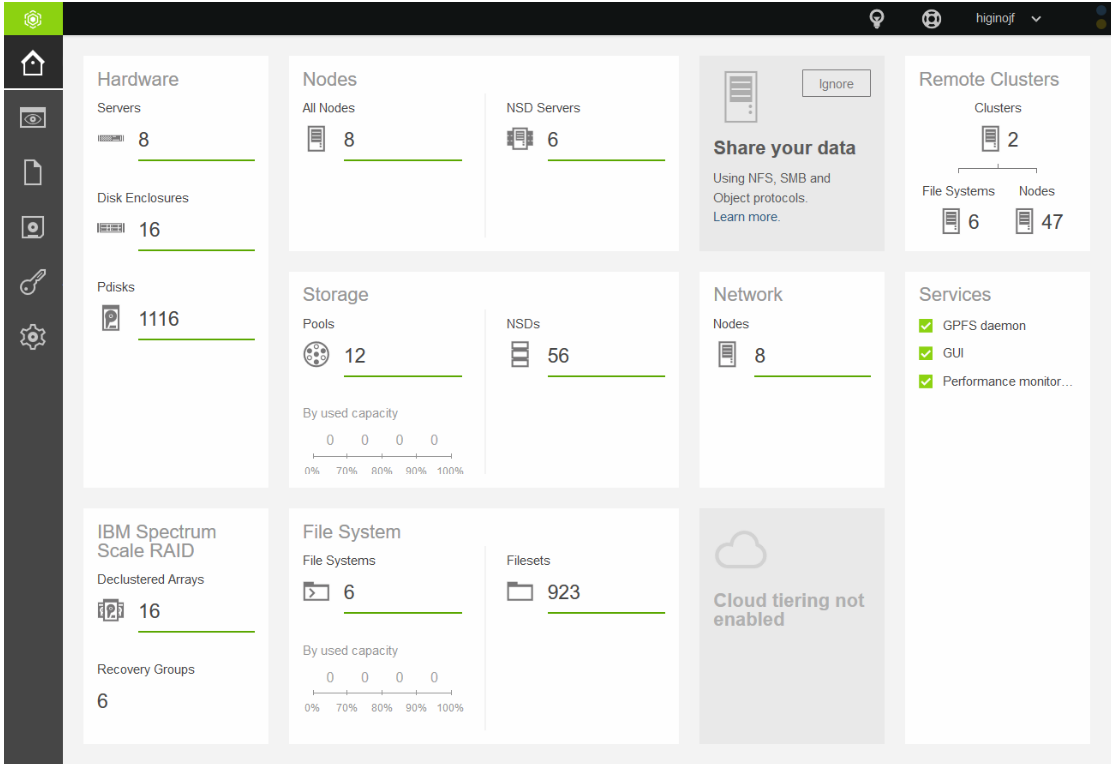
Severity	Age ↓	Message

No items found.

No health events for this time period

Spectrum Scale Web Management Interface (GUI)

- Hardware maintenance
- Statistics and Events
- Create and manage:
  - ✓ Filesystems
  - ✓ Filesets
  - ✓ Snapshots
  - ✓ Quotas
  - ✓ ILM Policies
- User access
- Notifications



higinojf

Hardware

Servers 8

Disk Enclosures 16

Pdisks 1116

Nodes

All Nodes 8

NSD Servers 6

Ignore

Share your data

Using NFS, SMB and Object protocols. Learn more.

Remote Clusters

Clusters 2

File Systems 6

Nodes 47

Storage

Pools 12

NSDs 56

By used capacity

0 0 0 0

0% 70% 80% 90% 100%

Network

Nodes 8

Services

☑ GPFS daemon

☑ GUI

☑ Performance monitor...

IBM Spectrum Scale RAID

Declustered Arrays 16

Recovery Groups 6

File System

File Systems 6

Filesets 923

By used capacity

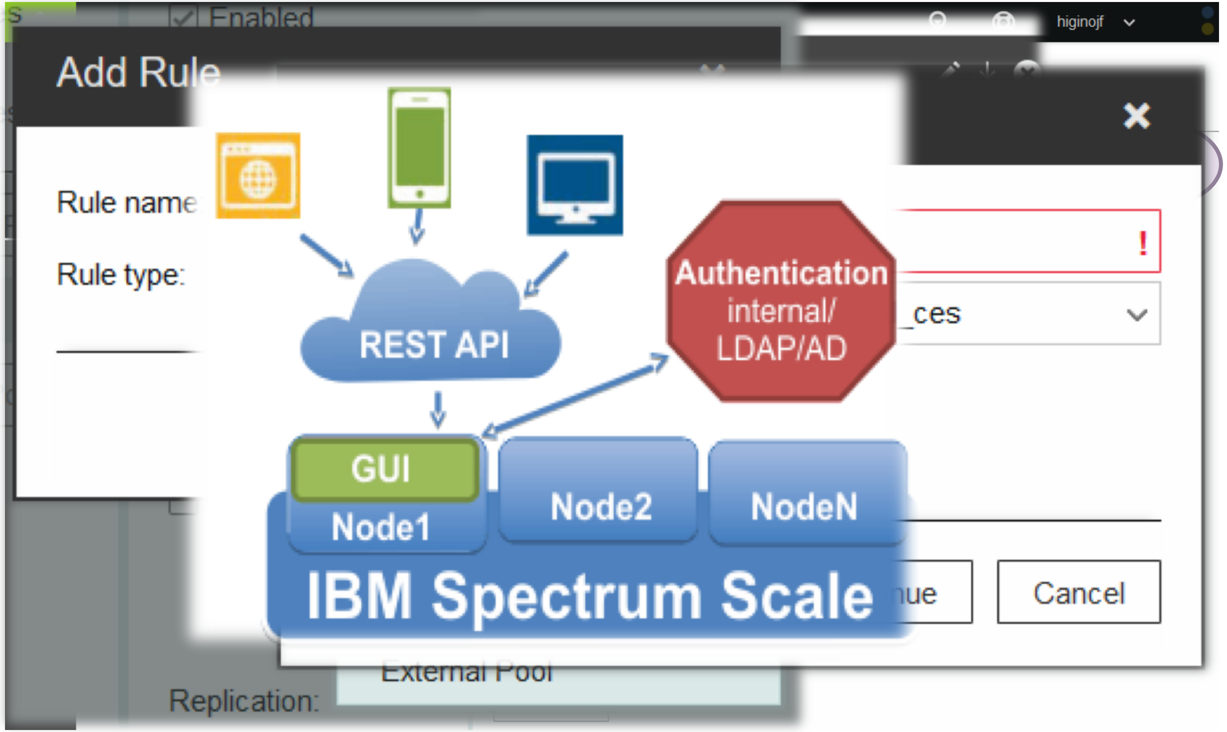
0 0 0 0

0% 70% 80% 90% 100%

Cloud tiering not enabled

Managing is now very simple!

- Policy Management:
  - ✓ Create/Delete
  - ✓ Enable/Disable
  - ✓ Predefined rules
  - ✓ Order rules
- See the text code
- Import policy files
- REST API (via GUI)



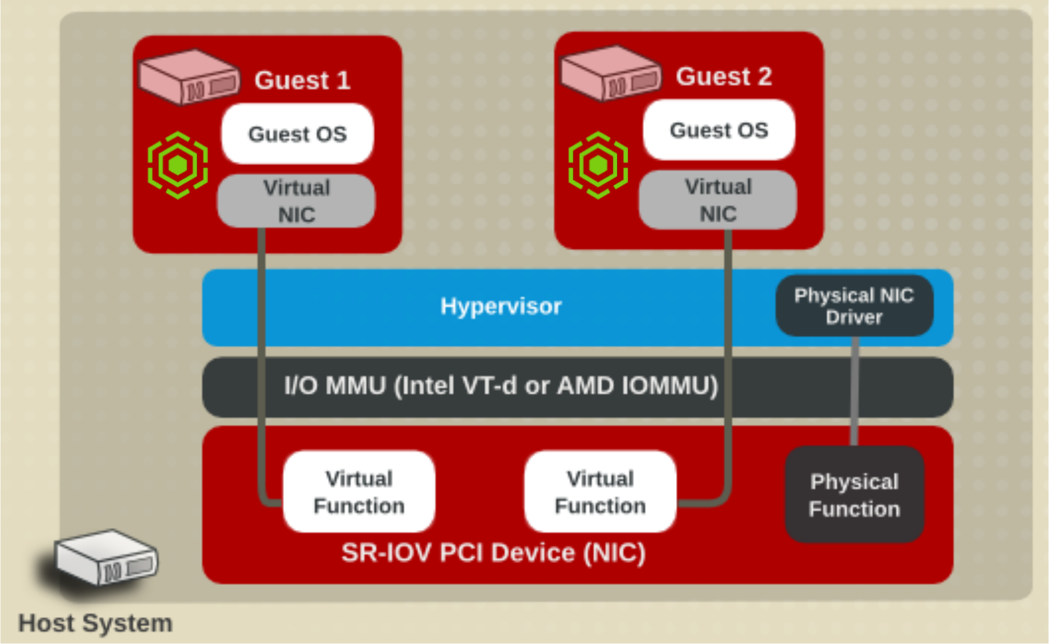
Spectrum Scale with SR-IOV / OpenStack

openstack™

- Virtual Interfaces on VMs
  - Infiniband (with RDMA)
  - Ethernet (1/10 Gbps)
- Orchestration via Bright OpenStack
- Heterogeneous Clusters (VM+BareMetal)

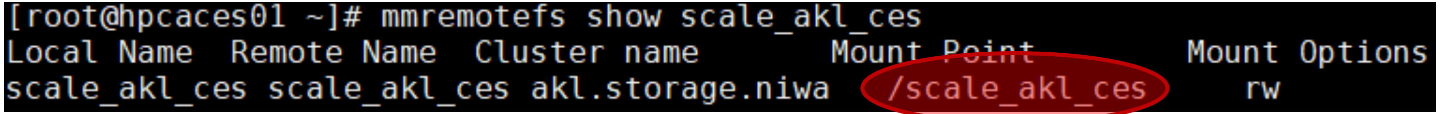
IBM Spectrum Scale

Bright Computing

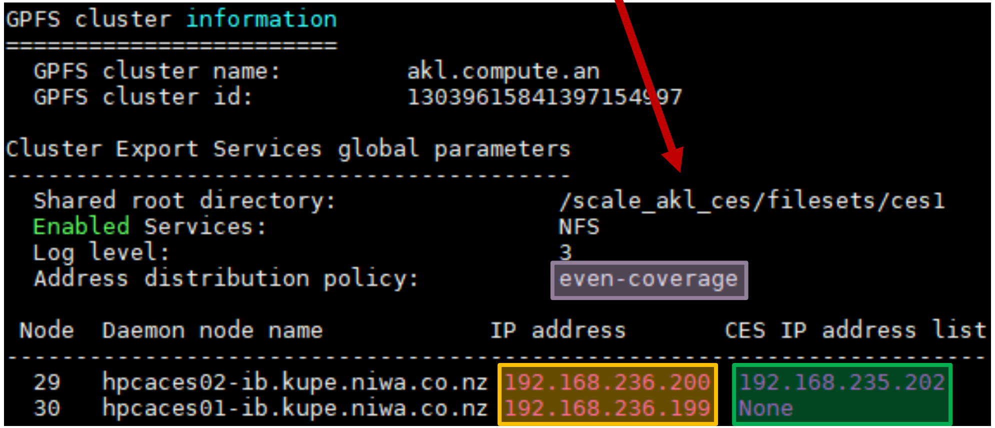


Protocol Nodes using OpenStack VMs

- High Available Services
  - ✓ Samba
  - ✓ NFS
  - ✓ (Planned) Object (Swift/S3 API)
  - ✓ (Exploring) File Auditing*1
- IP Distribution/Failback Policy
- Spectrum Scale Scalability
- Infiniband/10Gbps Ethernet
- Multicluster Support *1



```
[root@hpcaces01 ~]# mmremotefs show scale_akl_ces
Local Name   Remote Name   Cluster name       Mount Point        Mount Options
scale_akl_ces scale_akl_ces akl.storage.niwa  /scale_akl_ces       rw
```

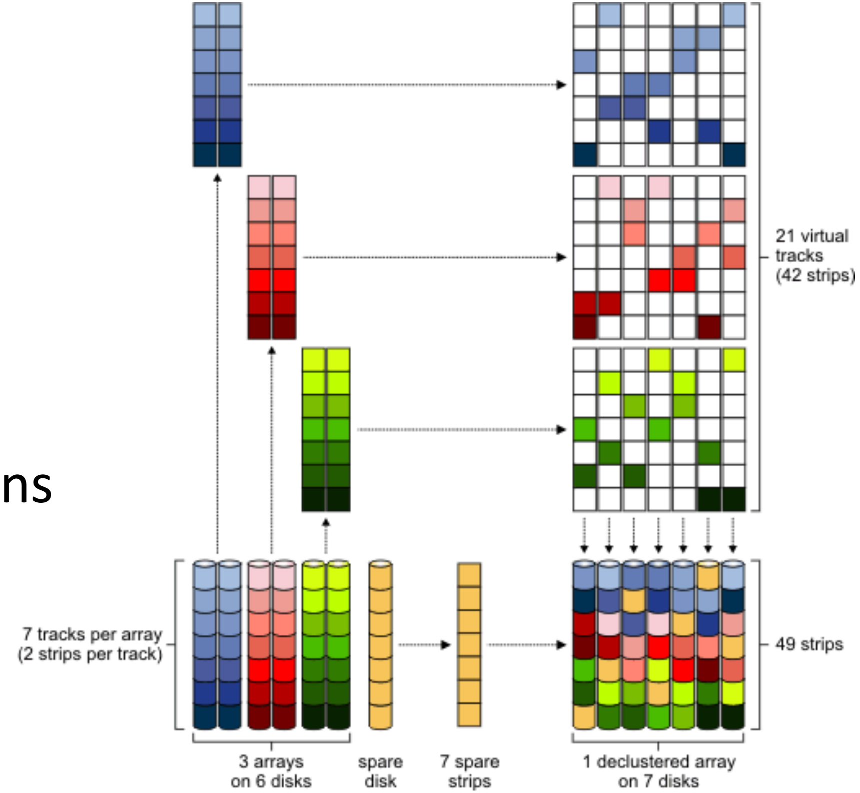


```
GPFS cluster information
========================
  GPFS cluster name:         akl.compute.an
  GPFS cluster id:           13039615841397154997

Cluster Export Services global parameters
-----------------------------------------
  Shared root directory:                /scale_akl_ces/filesets/ces1
  Enabled Services:                     NFS
  Log level:                            3
  Address distribution policy:          even-coverage

 Node  Daemon node name            IP address       CES IP address list
-----------------------------------------------------------------------
  29   hpcaces02-ib.kupe.niwa.co.nz 192.168.236.200  192.168.235.202
  30   hpcaces01-ib.kupe.niwa.co.nz 192.168.236.199  None
```

*1 – File Auditing not yet available across Multicluster configurations (2018-03-24)

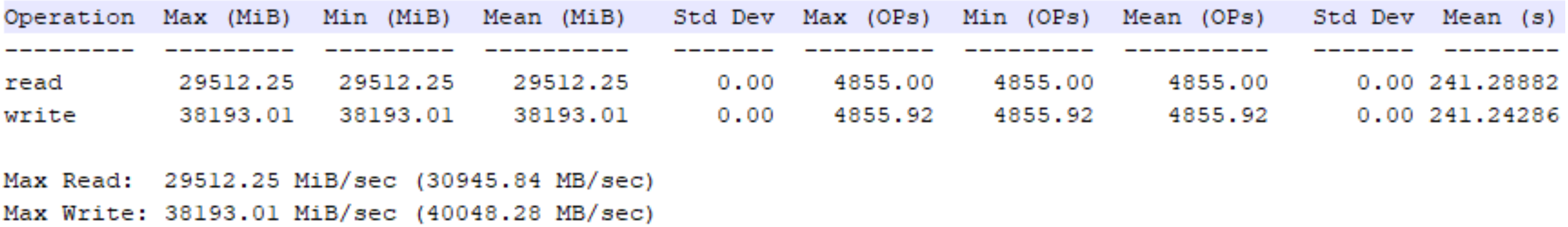
ESS flash rebuilds (GPFS Native RAID)

- Declustered arrays
  - ✓ Distributed Parity (less localized)
  - ✓ Software (uses system memory)
  - ✓ Increased IO Distribution
  - ✓ Higher Capacity available
- Powerful Resilience for Large Installations
- Dual path (hardware) Recovery Groups
- Tolerant to multiple disk failures


21 virtual tracks (42 strips)
7 tracks per array (2 strips per track)
49 strips
3 arrays on 6 disks
spare disk
7 spare strips
1 declustered array on 7 disks

Benchmarks (2x GL6s + GS4s)

- MDTEST (Kupe, Auckland)
  - ✓ From Cray XC50 (32 nodes with 2 tasks/node), in:
    - Unique directory file creation: 28.527 sec, 36757.358 ops/sec
    - Single directory file creation: 39.952 sec, 26245.610 ops/sec
- IOR (Kupe, Auckland)
  - ✓ From Cray XC50 (2x 52 nodes writing/reading with 2 tasks/node), 8MB Block Size:



```
Operation  Max (MiB)  Min (MiB)  Mean (MiB)   Std Dev  Max (OPs)  Min (OPs)  Mean (OPs)   Std Dev  Mean (s)
---------  ---------  ---------  ----------   -------  ---------  ---------  ----------   -------  --------
read        29512.25   29512.25    29512.25      0.00    4855.00    4855.00     4855.00      0.00 241.28882
write       38193.01   38193.01    38193.01      0.00    4855.92    4855.92     4855.92      0.00 241.24286

Max Read:  29512.25 MiB/sec (30945.84 MB/sec)
Max Write: 38193.01 MiB/sec (40048.28 MB/sec)
```

Summary 1/2

- Our facts and what we value most:
  - Continuously running GPFS filesystems since 2010 while:
    - Rolling software upgrades with no filesystem downtime;
    - Never losing data;
    - Shifting disk resources between live filesystems to meet new requirements in space and performance;
  - Continuous performance improvements and bug-fixing;
  - Flexibility of Spectrum Scale Features/Multicluster environments;
  - Provide SMB and NFS services via Spectrum Scale Protocol Nodes;
  - Integration with Spectrum Protect, providing Hierarchical Storage Management (Tape Storage).

Summary 2/2

- Where we are going next:
  - Upgrade ESS to support bigger sub-block division (change to 16MB Block Size) and reduced IO latency;
  - Fine tune Spectrum Scale and Spectrum Protect clusters for replication of backups between sites (Auckland and Wellington);
  - ESS LDAP integration (and GUI);
  - Enhancing Automation using REST API;
  - Implement Samba and Object Services (Protocol Nodes);
  - Benchmark performance of Spectrum Scale over SR-IOV.

Mahuika: HPC Cluster

- 234 compute node Cray CS400 cluster (8,424 x 2.1 GHz Broadwell cores, CentOS)
- FDR Infiniband network on compute nodes
- CS400 Virtual Labs, pre and post processing nodes (640 x 2.1GHz Broadwell cores, CentOS)
- Huge Memory node (4TB)
- Remote visualization
- GPGPUs (8 x P100)
- 100% NeSI access

Maui: HPC Supercomputer

- 464 compute node Cray XC50 supercomputer (18,560 x 2.4GHz Skylake cores, SLES)
- Cray Aries network
- CS500 Virtual Labs, pre and post processing nodes (1,120 x 2.4GHz Skylake cores, CentOS)
- Urika-XC Advanced Data Analytics
- Remote visualization
- GPGPUs (8 x P100)
- 57% NeSI access

Shared Storage

IBM ESS GL4S and GL6S disk storage (>10PB, >165 GB/s), Spectrum Scale (aka GPFS)

EDR Infiniband network

Spectrum Protect Hierarchical Storage Management system (storing >150PB in tape)