The NIWA/NeSI HPC Replacement Project:
A voyage in complexity integrating (multi-site) XC, CS, ESS and OpenStack systems

- M. Uddstrom\(^1\), B. Corrie\(^1\), N. Jones\(^1\), F. Cantos\(^1\), A. Hicks\(^1\), W. Hayek\(^1\), D. Kelly\(^2\), P. Balle\(^2\), A. Sachitano\(^2\), B. Gilmer\(^2\), J. Altorf\(^2\), D. McCurdy\(^3\), A. Beattie\(^3\).

\(^1\)NIWA/NeSI, \(^2\)Cray, \(^3\)IBM

New Zealand eScience Infrastructure
About NIWA and NeSI

• **NIWA**: National Institute of Water and Atmospheric Research (a Crown Research Institute - CRI):
  • NIWA’S purpose is to enhance the economic value and sustainable management of New Zealand’s aquatic resources and environments, *to provide understanding of climate and the atmosphere and increase resilience to weather and climate hazards to improve safety and wellbeing of New Zealanders.*

• **NeSI**: New Zealand eScience Infrastructure (collaboration: University of Auckland, University of Otago (Dunedin), Manaaki Whenua (CRI, Lincoln) and NIWA (CRI, Wellington))
  • NeSI’s purpose is to *grow the computing capability of researchers* to ensure New Zealand’s future prosperity
  • ~50% Government funded, provides National HPC services
Background and Context

- **NeSI**: HPC legacy:
  - Established in 2011;
  - Share of NIWA’s IBM P575/P6: 2,208 cores;
  - IBM iDataPlex (2012/13/14): 5,784 cores (WSM, SAB, IVB), GPFS

- **NIWA**: HPC legacy:
  - Cray T3E 1200e (1999/04): 544 cores
  - IBM P575/P6 (2010/13): 3,392 cores, GPFS

- **The Challenge**:
  - Four investing institutions (2×CRIs, 2×Uni’s);
  - Design - coming to agreement;
  - RFP…
Design Decisions

• National data-centric research and operational computing environment:
  • Single site: NIWA, Wellington;
  • Capacity Cluster: High Throughput and Private Cloud;
  • Capability: Large simulations & NIWA forecasting;
  • Capability (DR): NIWA forecasting (Auckland);
  • High performance filesystems.

• New user services:
  • Virtual Labs;
  • Remote visualisation;
  • OpenStack private cloud;
  • Advanced data analytics;
  • Hierarchical storage management services.

• Back-end services that mitigate the risk of data loss.
Procurement Strategy

- **Maximise return on investment:**
  - One RFP (3 HPCS);
  - Single Site, shared storage;
  - NIWA ($15.2M), UoA, UoO and MW ($4.8M);
  - Separate NIWA DR site (Auckland).

- **Benchmark driven:**
  - Capacity: NAMD, ANSYS, GROMACS, GATK, NWCHEM;
  - Capability: Unified Model, NEMO, EMOD3D;
  - I/O: IOR, MDTEST, IOZONE;
  - Workflow:
    - Real use case ( cylc NWP cycle including post processing)
    - Workloads.

---

Request for Proposals:
NeSI/NIWA Platforms Refresh

RFP NeSI-002

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

Notices: Commercial & Confidential
Version 4.0 Release
Date: 31st January 2017

Authors: NeSI Platforms Manager
NeSI Solutions Manager
Solution (Wellington)

• **NIWA/NeSI:**
  • CS400: 9,604 cores (BRW)
  • CS400: Test Cluster
  • **ESS: 10.1PB** (GS4, GS4S, GL6S)
  • XC50-LC: 18,560 cores (6148)
  • CS500: 1,120 cores (6148)
  • CS400: 180 cores (BRW)
  • Nvidia P100: 16
  • S822L: 30TB

• **Ancillary Nodes:**
  • OpenStack VMs or Baremetal;
  • NICE DCV Remote Visualisation;
  • Virtual Labs;
  • Pre/Post Processing.

---

New Zealand eScience Infrastructure

CUG Stockholm (23-May-2018)
Storage Architecture (Wellington)

• **User Storage (ESS):**
  - GPFS (aka Scale) clients:
    - CS400;
    - CS500.
  - DVS GPFS nodes:
    - XC50.
  - BOS VM via SR-IOV;
  - Other systems:
    - Protocol nodes.

• **Flexible Storage (S822L):**
  - Provisioning OSs;
  - Databases;
  - ELK, etc.
### Some Performance Measures

<table>
<thead>
<tr>
<th>XC core performance relative to P575/P6 at P6 core counts</th>
<th>Large Simulation Codes: (UM, NEMO, EMOD3D)</th>
<th>Large Simulation Codes + NAMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRW E5-2695v4 (2.1 GHz, 18 cores/socket) RFP reference</td>
<td>1.60</td>
<td>1.56</td>
</tr>
<tr>
<td>SKL 6148 (2.4GHz, 20 cores/socket) BAFO estimate</td>
<td>2.08</td>
<td>1.97</td>
</tr>
<tr>
<td>SKL 6148 (2.4 GHz, 20 cores/socket) Acceptance Tests</td>
<td>2.14 (33.8%)</td>
<td>2.07 (32.7%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ESS (4×GL6S, 1×GS4S) GPFS v5.0</th>
<th>CS400 (Native GPFS Client)</th>
<th>XC 50 (via DVS – 36 nodes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDTEST (4KB creates, unique dirs.)/s</td>
<td>156,900</td>
<td>35,766 (23%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Read (GB/s)</th>
<th>Write (GB/s)</th>
<th>Total (GB/s)</th>
<th>Read (GB/s)</th>
<th>Write (GB/s)</th>
<th>Total (GB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOR (Single Stream) 4KB</td>
<td>2.1</td>
<td>1.3</td>
<td>3.2 (w/IOBUF!)</td>
<td>3.2 (w/IOBUF!)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOR (Single Stream) 8MB</td>
<td>5.1</td>
<td>3.3</td>
<td>2.5</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOR (total bandwidth)</td>
<td>59.5</td>
<td>86.7</td>
<td>146.3</td>
<td>63.0</td>
<td>64.0</td>
<td>126.9</td>
</tr>
</tbody>
</table>
Status

• **NIWA DR** (XC50, CS500, CS400, OpenStack) installed and operational since 01/2018:
  • P575/P6 (Wellington) users/data transferred (~6PB) 03-Dec-2017;
  • NIWA operational forecasting service 09-Jan-2018.

• **NIWA/NeSI** (CS400 Cluster, XC50, CS500, CS400, OpenStack) installation started: 26-Feb-2018:
  • Upgrade GPFS 4.3 to GPFS 5.0 (s/w and filesystems);
  • Passed acceptance tests: 30 Apr 2018;
  • Bright OpenStack completed 18-May-2018;
  • Expect to put first users on the system – Mid June.

• **Wellington Plant Upgrade**: 
  • Increased data centre efficiency, power and cooling redundancy.

• **Issues**: DVS (need a better solution for GPFS sites).
Additional Slides
NIWA/NeSI - Wellington

CUG Stockholm (23-May-2018)
Solution (Auckland)

- **NIWA:**
  - XC50-AC: 4160 cores (6148)
  - CS500: 440 cores (6148)
  - CS400: 144 cores (E5-2695v4)
  - Nvidia P100: 2
  - ESS: 4.5PB (GS4S, GL6S)
  - S822L: 30TB

- **Ancillary Nodes:**
  - OpenStack VMs or Baremetal;
  - Virtual Labs;
  - Remote Visualisation;
  - Pre/Post processing...

- **IBM ESS Storage**
  - Persistent
  - Scratch
  - Operational

- **IBM Spectrum Protect**
  - Backup
  - HSM

- **IBM TS4500 (Offline Storage)**
  - Tape Library
  - LTO

- **IBM S822L Storage**
  - Bright OpenStack
  - Logs
  - Databases

- **CS400 (EDR) (Shared Services)**
  - SLURM
  - DTN
  - Protocol Nodes
  - ELK
  - FreeIPA
  - Ansible Tower
  - Web servers
  - Librarian... etc.

- **XC50 Capability (Kupe)**
  - Compute
  - NIWA/NeSI shared

- **CS500 (EDR) Ancillary Nodes**
  - Large memory
  - GPGPUs

- **CUG Stockholm (23-May-2018)**
NIWA DR - Auckland

Tamaki: System Overview with IBM storage

HPC3/XC50
(104 compute nodes
19 DVS links)

HPC3 Multi-Purpose
(11 x C5500 nodes)

HPC3 Shared SVC
(4 x C5400 nodes)

All links in the network are EDR Infiniband

- IB-Eth
  - 12-port switch for ib-to-Ethernet
  - Gateway; 2 x 16 links to core switches

- 10/40Gb link(s) to Tamaki site

Research & Persistent & Operational & Nearline Storage
1 x GS4s + 2 x GL6s
(4 x EDR IB links to each GS4 and GL6s)

Flexible Storage
2 x Power Servers with internal drives, running Spectrum Scale
Data Management Edition
(2 x EDR IB links to each server)

NIWA DR - Auckland

New Zealand eScience Infrastructure