Worldwide Trends in Computer Architectures for Data Science

13 February, 2020

Jeff Zais
### NeSI @ eResearch NZ - Talks & Workshops:

**Wednesday 12 Feb**
- **1:30 - 1:50 pm** - Megan Guidry - Training: It's better together
- **1:30 - 5:30 pm** - Chris Scott - First steps in machine learning with NeSI
- **1:50 - 2:10 pm** - Callum Walley - Engineering HPC: What’s going on?
- **2:10 - 2:30 pm** - Marko Laban - Cloud-native technologies in eResearch: Benefits & challenges
- **2:50 - 3:00 pm** - Jun Huh - Learning how to learn
- **3:30 - 4:30 pm** - Megan Guidry - Building and supporting a NZ digital literacy training community
- **3:30 - 4:30 pm** - Blair Bethwaite - Research Cloud NZ

**Thursday 13 Feb**
- **11:00 - 11:20 am** - Wolfgang Hayek - Singularity containers on HPC
- **11:00 am - 12:20 pm** - Brian Flaherty - Building a national/regional data transfer platform: Globus BoF
- **1:30 - 1:50 pm** - Nick Jones - Advancing New Zealand’s computational research capabilities and skills
- **1:50 - 2:10 pm** - Brian Flaherty - Where Data Lives: NeSI, taonga and growing repository services
- **2:10 - 2:30 pm** - Dinindu Senanayake - HPC for life sciences: Handling the challenges posed by a domain that relies on big data
- **3:30 - 5:30 pm** - Jana Makar - Growing the eResearch workforce in an inclusive way

**Thursday 13 Feb (cont.)**
- **1:50 - 2:10 pm** - Jeff Zais - Worldwide trends in computer architectures for data science
- **2:10 - 2:30 pm** - Dinindu Senanayake - HPC for life sciences: Handling the challenges posed by a domain that relies on big data
- **3:30 - 5:30 pm** - Jana Makar - Growing the eResearch workforce in an inclusive way

**Friday 14 Feb**
- **11:20 - 11:40 am** - Alexander Pletzer - Enhancing eResearch productivity with NeSI's consultancy service
- **1:30 - 3:40 pm** - Nooriyah Lohani - Research Software Engineering (RSE) community update and next steps in New Zealand
Worldwide Trends in Computer Architecture for Data Science

A - Survey of large academic research centres

• NCI (Australia)
• LRZ (Germany)
• SciNet (Canada)

B - Trends & implications

• Processors
• Memory
• Networking
• Storage
Some architectural examples
Main focus on energy efficiency – direct water cooling design
Storage
LRZ Phase 1 fabric (9000+ nodes) Islands, with director switches at lowest level. Reduced bandwidth, independent islands ease bring-up.

For LRZ NG, storage island is a mix of DSS-G units.

50 PB @ 500 GB/s
20 PB @ 70 GB/s
SciNet
*Toronto, Canada*

**Burst Buffer**
- 80 NVMe drives in 10 servers
- 20M random read 4k IOPS
- 148 GB/s write
- 230 GB/s read

- Leading research compute centre in Canada
- Storage 12 PB capacity, with **NVMe burst buffer** based on Excelero technology
➢ 2019 refresh of the compute cluster in Canberra
➢ Main focus on serving the broad academic community – with a bonus on energy efficiency
➢ ALCF (Argonne Leadership Computing Facility) will deploy a new Cray ClusterStor E1000

➢ Computational capacity: “Grand” provides 150 PB, at 1000 GB/s

➢ Simplified data-sharing: “Eagle” provides 50 PB
Data movement –
Trends and connections
Beneficiaries of an arms race

- Increase in core count and core performance drives a requirement for increased memory bandwidth
- Best solution to that has been adding more (and more) memory channels
- 4 to 6 to 8 channels per socket
- At the same time DDR4 DIMM size continues to increase
- **Straightforward to configure nodes with 6+ terabytes** for simulations suited to large/huge memory
- Also, persistent memory modules with high capacities (128/256/512 GB) offer unique capabilities, evaluation is needed
## Storage essentials from vi4io.org site

<table>
<thead>
<tr>
<th>#</th>
<th>site.institution</th>
<th>site.storage_system.net capacity</th>
<th>site.supercomputer.compute peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>in PiB</td>
<td>in PFLOPS</td>
</tr>
<tr>
<td>1</td>
<td>National Energy Research Scientific Computing Center</td>
<td>580.72</td>
<td>35.14</td>
</tr>
<tr>
<td>2</td>
<td>Oak Ridge National Laboratory</td>
<td>278.00</td>
<td>220.64</td>
</tr>
<tr>
<td>3</td>
<td>Los Alamos National Laboratory</td>
<td>72.83</td>
<td>11.08</td>
</tr>
<tr>
<td>4</td>
<td>German Climate Computing Center</td>
<td>52.00</td>
<td>3.69</td>
</tr>
<tr>
<td>5</td>
<td>Lawrence Livermore National Laboratory</td>
<td>48.85</td>
<td>20.10</td>
</tr>
<tr>
<td>6</td>
<td>RIKEN Advanced Institute for Computational Science</td>
<td>39.77</td>
<td>10.62</td>
</tr>
<tr>
<td>7</td>
<td>National Center for Atmospheric Research</td>
<td>37.00</td>
<td>5.33</td>
</tr>
<tr>
<td>8</td>
<td>National Center for Supercomputing Applications</td>
<td>27.60</td>
<td>13.40</td>
</tr>
<tr>
<td>9</td>
<td>Global Scientific Information and Computing Center</td>
<td>25.84</td>
<td>17.89</td>
</tr>
<tr>
<td>10</td>
<td>Joint Center for Advanced HPC</td>
<td>24.10</td>
<td>24.91</td>
</tr>
<tr>
<td>11</td>
<td>Cineca</td>
<td>23.71</td>
<td>12.93</td>
</tr>
<tr>
<td>12</td>
<td>Argonne National Laboratory</td>
<td>21.32</td>
<td>10.00</td>
</tr>
</tbody>
</table>
Storage connections - driven by capacity and bandwidth

- Capacity ranges up to 40 or 60 petabytes (ignoring extreme sites up to hundreds of petabytes)
- Bandwidth (in GB/s) range up to several hundred (ignoring extreme sites up to 1000+ GB/s)
- Typical building block (DDN / Lustre, Lenovo/IBM Spectrum Scale) based on a rack with 500+ drives, 5+ petabytes capacity, 35 GB/s bandwidth
- These rack building blocks scale nicely to large sizes
Basic connections into storage are straightforward

How do we connect via InfiniBand?

- Current generation is EDR, based on 200 Gb/s = 25 GB/s peak performance
- Typical actual performance (per wire) is 17 GB/s, so we just need a dozen or so wires to achieve 200 GB/s
- Latency will be excellent (1 ns typical)
Handling big disc clusters is also straightforward

- (Lots of disks) + (technology trends) = disk failure every 4 days
- Combination of RAID6 and Declustered RAID allows for efficient and barely noticeable rebuilds
Well, then what is *not* so straightforward?

- Our data can follow more paths than ever before
- How can we
  - Maximize performance
  - Minimize hassle  *(for the scientist, the end user)*

**Notes:**

- tape
- disk
- memory
- processor

**Diagram:**

- tape
- Spinning NL-SAS
- SSD
- NVMe
- DDR4
- High Bandwidth
- Persistent
- Storage Class
- general processor
- GPU
NIWA data
NIWA pursues data
NIWA pursues data through space and time
NIWA pursues and preserves data through space and time
NIWA pursues, processes, and preserves data through space and time

Bonus material!!
Getting close to 50 years of data

Measurements in the Southern Hemisphere

[Graph showing atmospheric carbon dioxide levels over time]
Data used for weather models:
Output from the UK Met global forecast serves as input into the more detailed New Zealand regional forecast – four times a day.

Figure 1 – Global model zoomed in over Australia and New Zealand showing the NZLAM (outermost) and NZCSM (innermost) domain boundaries.
Gathering additional data in space and time

Additional sources of input into the detailed New Zealand regional forecast

- 28 weather stations located throughout New Zealand
- Data collected automatically from ships
- Data collected automatically from aircraft
- Data collected from sounding balloons
- Images from satellites
Processing and filtering data – research underway in this area

Applying imaging techniques to identify and remove bad data points
Continuous process improvement

- Take yesterday’s forecast
- And the forecast from some days before
- Compare to the measurement, from both fixed and mobile stations
- Close the comparison loop to see if there should be some changes to the forecast methodology
- *Repeat daily*
Challenge: A century of data
Sony Optical Disc Archive Technology Version 3 – storage variant of Blu-ray
Q: how can NIWA take advantage of handwritten data which spans back over many decades?

A: apply Artificial Intelligence techniques and training to learn the handwriting of the day, process the records digitally.
Thank you.

Kia ora koutou.