Worldwide Trends in Computer Architectures for Data Science

NeSI
New Zealand eScience Infrastructure

13 February, 2020

Jeff Zais

NIWA
Taihoro Nukurangi

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

1:30 - 1:50 pm - **Jun Huh -** User journey-driven product management

1:30 - 5:30 pm - **Blair Bethwaite** - Containers in HPC tutorial

1:50 - 2:10 pm - **Brian Flaherty** - Where Data Lives: NeSI, taonga and growing repository services

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

LRZ
Garching (Munich area), Germany

- Main focus on energy efficiency – direct water cooling design
- Storage

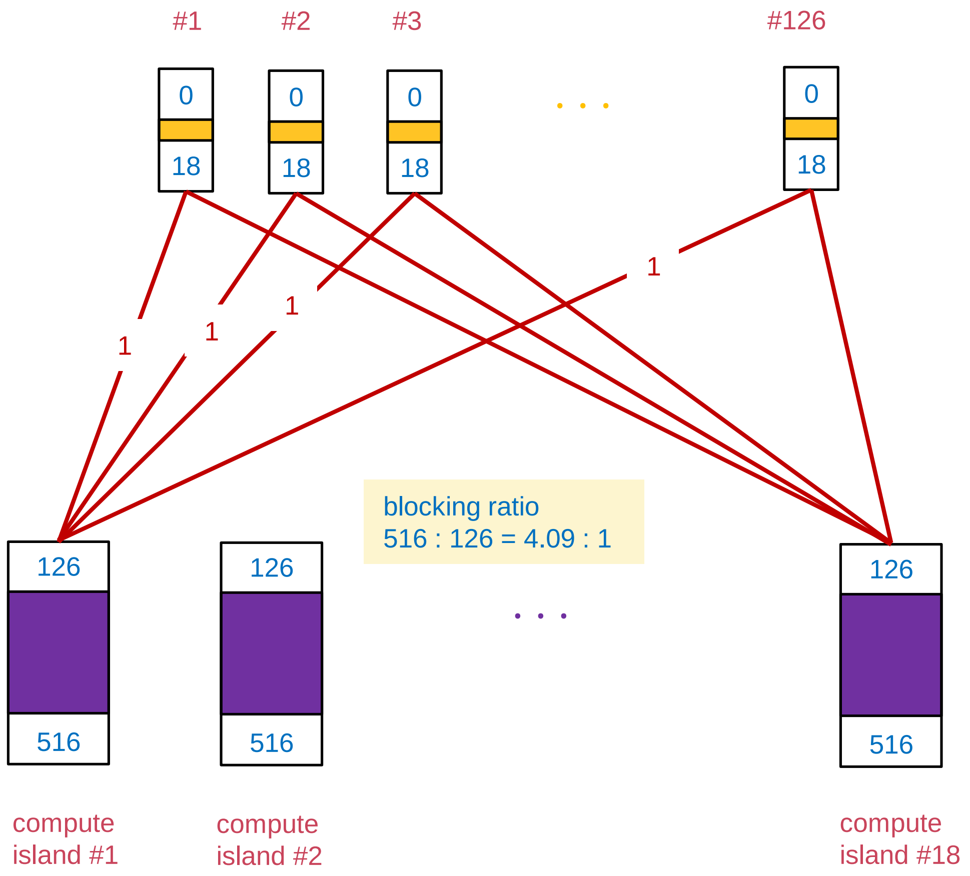
lrz Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities

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



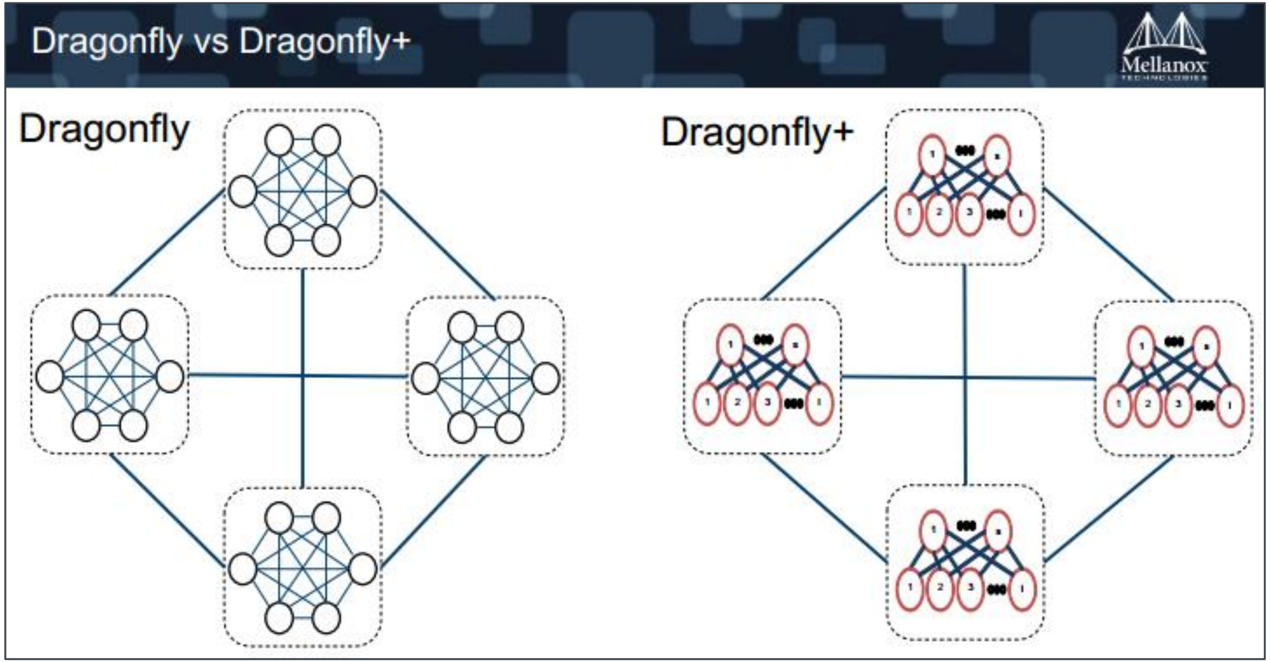
SciNet
Toronto, Canada

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

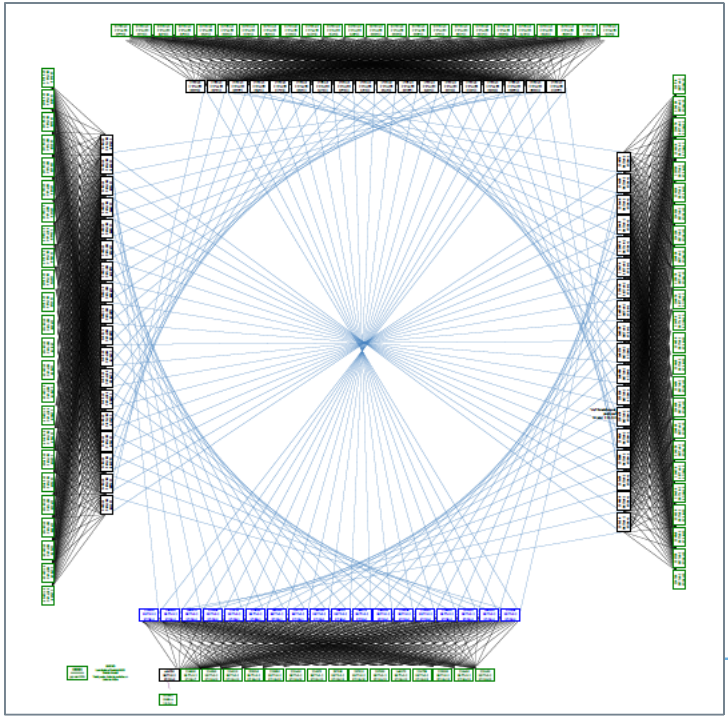


NCI, at Australia National University

Canberra, Australia

Lustre storage:
54 PB, 150 GB/s

- 2019 refresh of the compute cluster in Canberra
- Main focus on serving the broad academic community – with a bonus on energy efficiency



Argonne

Lemont (Chicago area), Illinois

- 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

NeSI
New Zealand eScience Infrastructure

NIWA
Taihoro Nukurangi

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

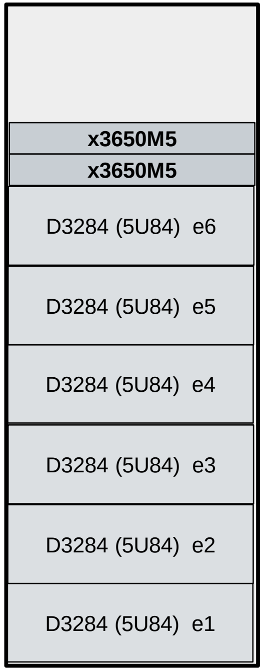
Storage essentials from vi4io.org site

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

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

Lenovo DSS G260



502 x NL-SAS
2x SSD

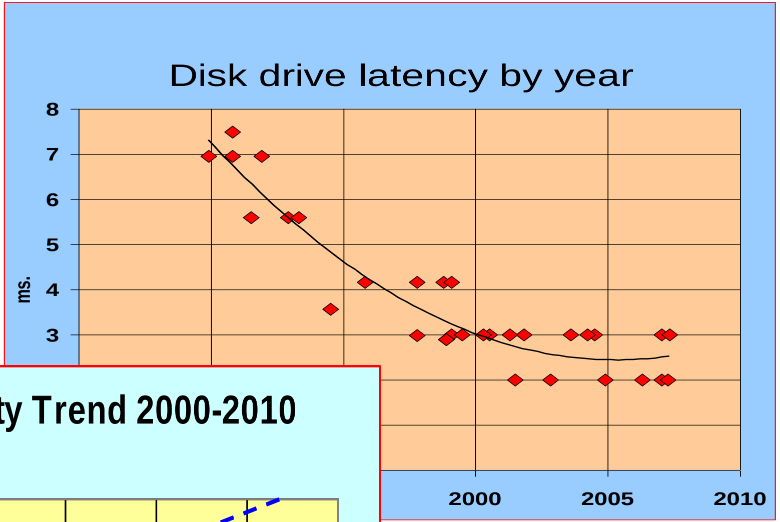
Basic connections into storage are straightforward

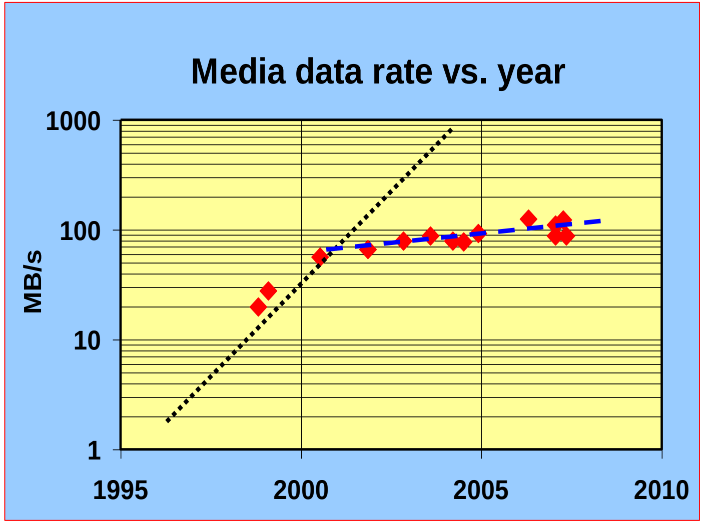
How do we connect via InfiniBand?

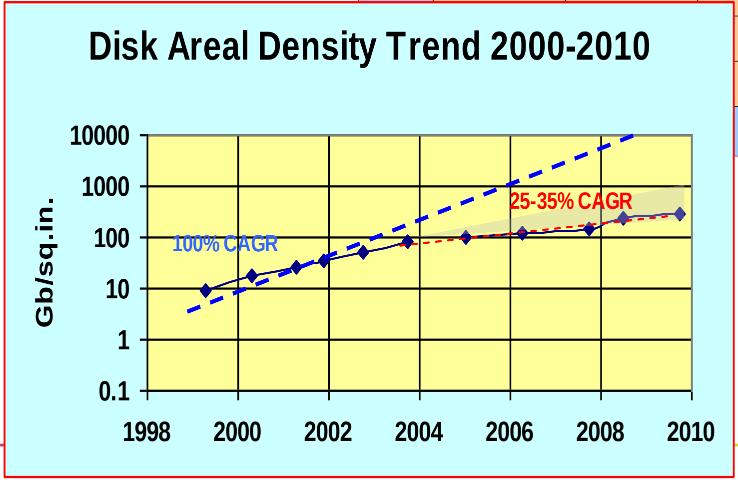
- Current generation is EDR, based on 200 Gb/s = 25 GB/s peak performance
- Typical actual performance (per wire) is 17 GB/s, so we just need a dozen or so wires to achieve 200 GB/s
- Latency will be excellent (1 ns typical)

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

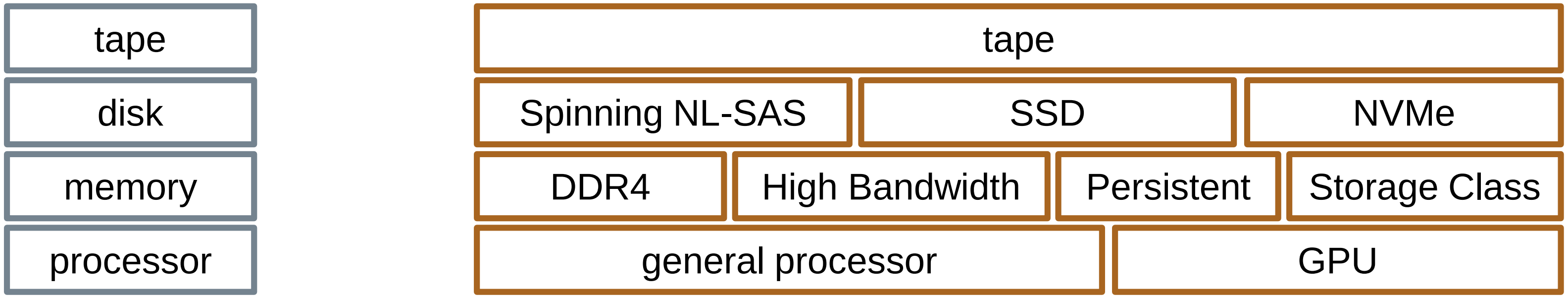






Well, then what is *not* so straightforward?

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



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

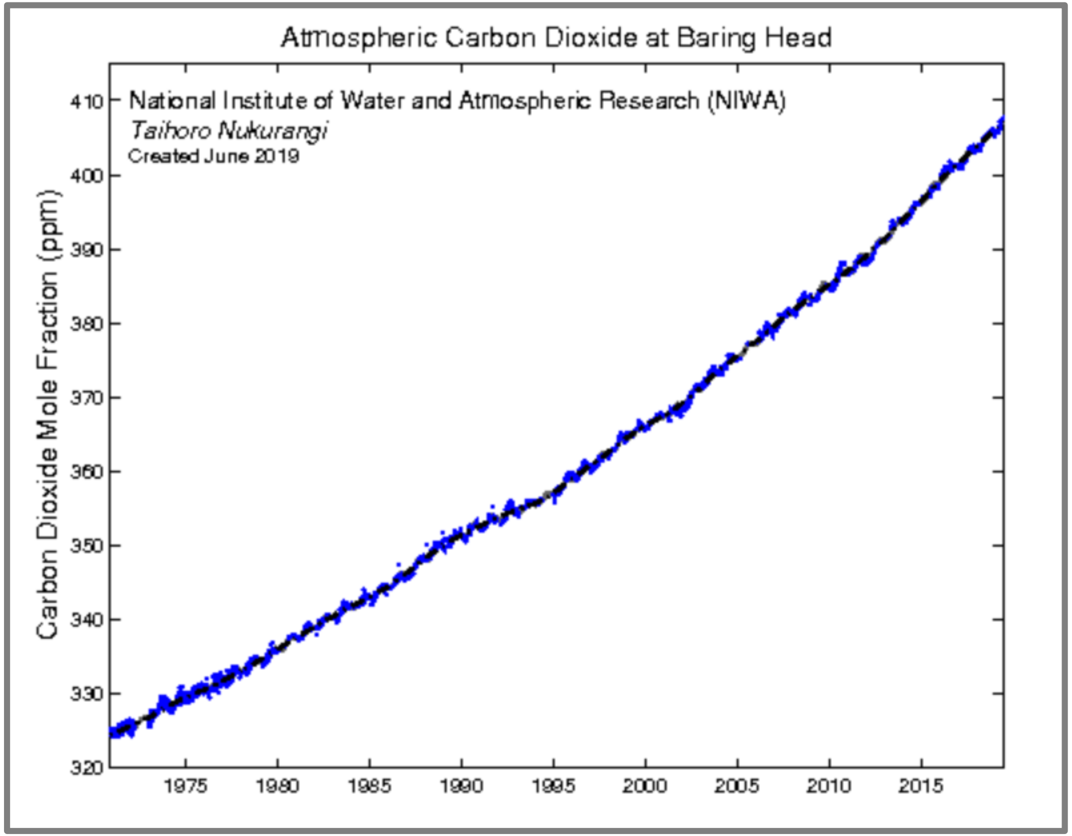
NeSI
New Zealand eScience Infrastructure

Bonus material !!

NIWA
Taihoro Nukurangi

Getting close to 50 years of data

Measurements in the Southern Hemisphere



Gathering basic data in space and 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

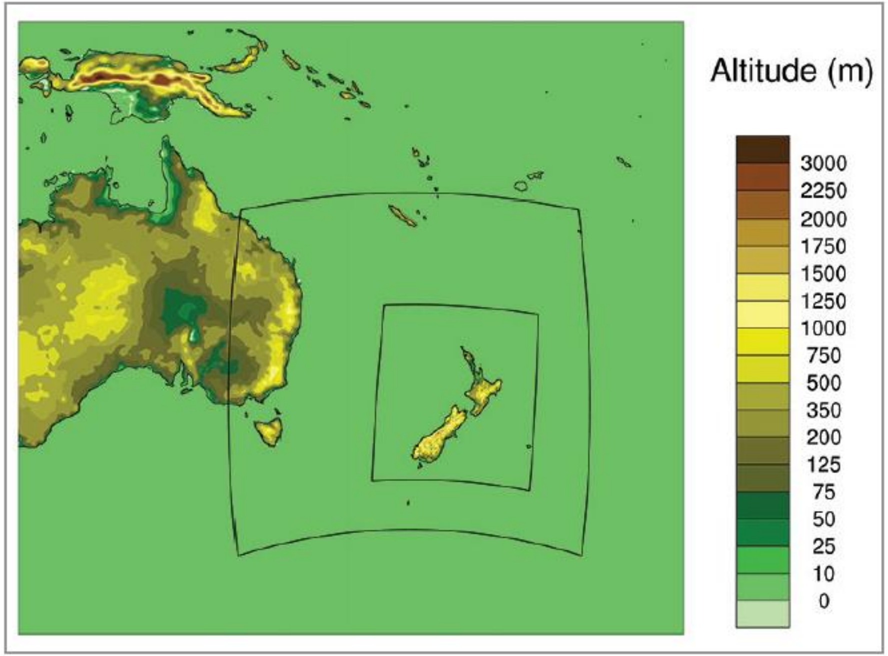


Figure 1 – Global model zoomed in over Australia and New Zealand showing the NZLAM (outermost) and NZCSM (innermost) domain boundaries.

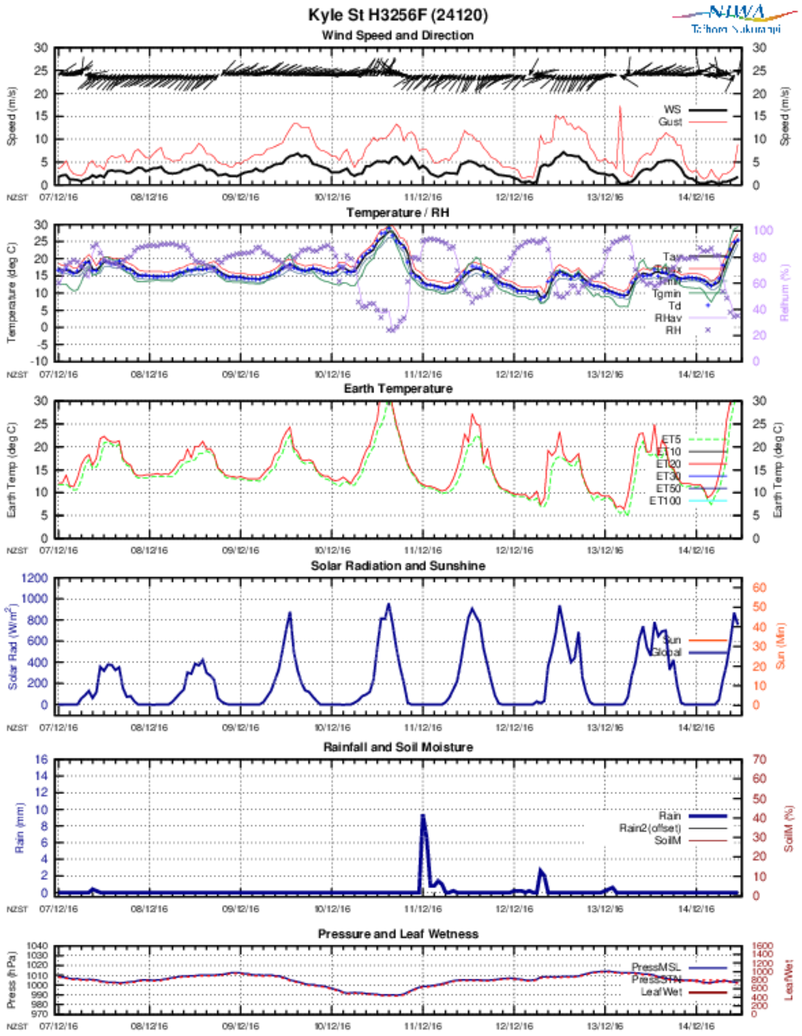
Gathering additional data in space and time

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



Preserving data to evaluate accuracy of weather forecasts

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

SC19 visit

Sony Optical Disc Archive Technology Version 3 – storage variant of Blu-ray

Joint project between Microsoft and NIWA

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.