The New Innovation Game:
Rebooting Infrastructure for more productive Climate, Weather, Environmental Systems & Geophysics Research

Ben Evans
ICAS Workshop 2019
NCI’s national peak research system

• Collaboration with major partners - Aus Bureau of Met, CSIRO, Geoscience Australia and ANU
• As such, major priority is climate, weather, environmental and geoscience research
• NCI lead’s the national merit process: 10+% NCI, 10+% Pawsey, plus other specialized but minor systems
• Over 89,000 cores (Intel Xeon Sandy Bridge/Broadwell) in 4519 compute nodes
• 330 Terabytes of main memory
• Infiniband FDR/EDR interconnect
• 7 Petabytes of dedicated scratch high-performance storage (150 GB/sec)
• 54 Petabytes of high-performance Lustre collection and project storage
• Completed procurement
• Fujitsu is the primary contractor, mostly Cascade Lake, and increased in GPU, Centos 8
• https://opus.nci.org.au/display/Help/Gadi%3A+NCI%27s+New+Supercomputer

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gadi Phase 1 User Testing</td>
<td>Access for users to Gadi, Raijin running in parallel</td>
<td>11th November 2019</td>
</tr>
<tr>
<td>Raijin Decommission</td>
<td>Decommission of Sandy Bridge nodes, Removal of Raijin’s Broadwell, Skylake and GPU nodes</td>
<td>25 November 2019</td>
</tr>
<tr>
<td>Gadi Broadwell and Skylake</td>
<td>Broadwell, Skylake and GPU nodes installed</td>
<td>28 November 2019</td>
</tr>
<tr>
<td>Gadi GPU Phase 2 Install</td>
<td>More GPUs</td>
<td>11 December 2019</td>
</tr>
<tr>
<td>Gadi full production</td>
<td>User access to the full system</td>
<td>6 January 2020</td>
</tr>
</tbody>
</table>

Integrated HPC, internal cloud and storage

NCI persistent storage of over 50 PB high-speed (disk) filesystems, ~30 PB, duplicated on project (self-managed) archival (tape) storage.
Examples of priority research at NCI

- Modelling Extreme & High Impact events
- Monitoring the Environment & Ocean
- Natural Hazards
- Agriculture/food security
- Natural Hazard and Risk models: Tsunami, Ash-cloud
- 3D/4D Geophysics: Magnetotellurics, AEM, Forward/Inverse Seismic models and analysis
- Hydrology, Groundwater, Carbon Sequestration
For Climate, Weather, Environment and Geophysics, the recent activities have been

- CMIP6 – model and analysis prep at the tail-end of phase X, start of phase X+1
- Regional Copernicus Satellite Hub, and Australian region Landsat datasets and products
- Increasing geophysics datasets ready for HPC and interoperability
- Preparing for next generation needs for data, and data analysis
  1. Where/when do we internal cloud infrastructure – managed SaaS
  2. Data repository for trusted, managed, high-used dataset access
  3. Data analysis environments – eg., Jupyter, server-side processing
Challenge 1:
Technology changes within HPC centres
Most major compute centres are generally not holding back future compute capacity for legacy software:

- Componentry moves on, more attractive to purchase
- Models need to adjust
- Those that adapt go forward
- Legacy code that doesn’t adapt will get squeezed out
- This is the same as the previous vector -> micro change.

Eg 1. Oak Ridge - but most of HPC systems are in the same boat – with some exceptions for now

- Summit supercomputer system. Resource is GPUs and can’t access if not GPU-enabled
- Frontier – first Exascale (1.5 Eflops) in 2021. AMD EPYC CPU and AMD Radeon Instinct GPU technology

“As a second-generation AI system, Frontier will provide new capabilities for deep learning, machine learning and data analytics for applications ranging from manufacturing to human health.

eg 2. NERSC is the same -> (2020) GPU accelerators -> (2024) Exa -> (2028) post Moore/non-von Neumann
<table>
<thead>
<tr>
<th>Law</th>
<th>Description</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moore’s Law</td>
<td>Transistors density doubles every 2 years</td>
<td>Broken (2012)</td>
</tr>
<tr>
<td>Top500 HPL</td>
<td>Increase in performance of ~1.85x per year</td>
<td>Broken (2013)</td>
</tr>
<tr>
<td>Dennard Scaling</td>
<td>Transistors get smaller, circuitry gets faster, and their power density stays constant. So power use is proportional to area</td>
<td>Broken (2005)</td>
</tr>
<tr>
<td>Koomey’s Law</td>
<td>Performance per watt doubles every 1.57 years</td>
<td>Broken (2005)</td>
</tr>
<tr>
<td>Kryder’s Law</td>
<td>Areal density of disks doubles every thirteen months</td>
<td>Broken (2002)</td>
</tr>
<tr>
<td>Wirth’s Law</td>
<td>Existing software is getting slower more rapidly than hardware becomes faster.</td>
<td>Existing software not adapting easily to new hardware. New techniques &amp; algorithms are needed.</td>
</tr>
<tr>
<td>Parkinson’s Law</td>
<td><em>Analogue of the ideal gas law:</em> Work expands so as to fill time available for its completion.</td>
<td>Always true and unbeatable. So set a goal, then do it to highest quality</td>
</tr>
</tbody>
</table>
What are the options?

- DSLs – Domain Specific Languages to flexibly address architecture changes. PSyclone being used by UK Met Office. Others using gridtools.

- Use of lower precision arithmetic, and mixed precision, replacing some standard algorithms with elements such as low precision steepest gradient, and high precision final result.

- Replacing some algorithms in certain circumstance with known inference models based rather than brute-force for all calculations.

- Interprocesssor communication changes with MPI+X, where X is alternative communication protocols for better scalability.

- Computing on-the-fly rather than storing results – both for models and for dynamic data products through data services.
Challenge 2: Transparency in methods and outputs understanding scientific tolerances
How Trustworthy is your Digital Science?

• If someone gave you their digital science output, what is needed to convince you that it was correct?

• If this required a complex system that you couldn’t run yourself, what information do you need to trust it? Do you know what the assumptions and dependencies are?

• How much could and should you rely on results persisting over a longer time?
One set of Definitions: the Association for Computing Machinery (ACM):
https://www.acm.org/publications/policies/artifact-review-badging (based on International Vocab of Metrology - BIPM)

**Repeatability (Same team, same experimental setup)**
The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.

**Replicability (Different team, same experimental setup)**
The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author’s own artifacts.

**Reproducibility (Different team, different experimental setup)**
The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.
At large scale, it is not feasible/practical to repeat or completely verify the work at the large scale, then what do we need to gain trust?

For large computational digital science, the challenge is to be transparent. But how?

- Pretesting with visible unit tests, benchmark/scenario tests and procedures?
- (How) do we record and share the level of variability/precision?
- Systems maintainers to understand the underlying system state
  
  Does anyone know the “versions” of their big and unique systems? What about months ago?
- Transparency is supported by FAIR principles, but must apply more broadly than “output datasets”
- And we need a good scientific rationale of what is trying to be done and the level of precision required

“You must not fool yourself — and you are the easiest person to fool”, Richard Feynman.
To what extent should we be providing information about uncertainty/sensitivity analysis?
  • What needs to be put in place to Trust Deep Learning/ Machine Learning and AI?

How do we address next generation of workflow systems (the micro-services)
  • What is a PID on the chain?
  • Does a certified Blockchain play a role?

To what extent do we expect Provenance records to be available
  • Is PROV ready to handle the range of questions and return in a usable form?
  • Can these be available to provide information about the underlying platforms?
Challenge 3: Data for HPC, analysis & interoperability
Major Challenges for current users

- Datasets are too large/cumbersome to move, costly to download, hard to maintain coherency
- Number of files are very difficult to wrangle as a single dataset or to interoperate
- Data not generally fit for purpose - need to be converted/transformed
- Both HPC, data and coding skills required for manipulating the data are barriers to analyzing data
- Geospatial users would prefer to use standard tools and known protocols for important reference datasets
- Most users find it more convenient to analyse “on the desktop” – most compute and data not in central systems

Number of files growing rapidly:
3x every 2 years

For managed Datasets, number of files:
2018 - 330M
2020 – 1B
Preparing for the next generation of computing & data

Time Sharing

Data Centre/Cloud Computing
Preparing for the next generation of computing & data

Time Sharing

Data Centre/Cloud Computing

Fog Computing
NIST defines Fog computing:
“residing between smart end-devices and traditional cloud or data centers. This paradigm supports vertically-isolated, latency-sensitive applications by providing ubiquitous, scalable, layered, federated, and distributed computing, storage, and network connectivity.”


Fog = data+compute at edge interacts with Big Data central data center/cloud

In reality
- most data (80+) is at the edge
- most data requires authorization

- We need to plan for more intelligent interaction with better, high-speed data access

- Future of data and information doesn’t look like:
  - Login to this system
  - “show me your files, put in shopping cart, click for download”
Access patterns are increasing in complexity, and now part of BAU dependency

- Digital environments and analysis/visualisation portals
- Virtual Research Environments that span administrative domains and continents
- Programmable workflows, Jupyter/iPython notebooks, mobile apps, interconnected systems
- Other machine access types:
  - sensor-to-machine
  - device-to-machine
  - AI-to-machine
Challenge for Data Access

- Performant (real-time) interoperable access to large (PB+), highly curated and interoperable datasets
- Can’t afford to store the data in the way needed by special use-cases
- Can’t afford to recopy all data
- Can’t afford to coordinate to have in one location
- More demand for APIs and micro-services (e.g., data streams and data channels)
- Data that is needed might not even pre-exist. That is:
  - Dynamically generated datasets: not precomputed & stored
  - Intelligent services
    - Server-side algorithms
    - Deep Learning
High Performance for both Big Compute and Big Data Analysis

Potential Number of Users

Increasing Skills Level to Use

Vertical Slices (e.g., time, depth)

A Small Piece

2D Slices (e.g., XY spatial)

The whole cube

Increasing Capacity of Compute to Process
Make longtail data easily analyzable with the large reference data
A problem we all know so well ...

Cleaning and organizing data
12% preparing for analysis
Just 9% devoted to mining data for patterns

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Some of our progress at NCI to address
So, what is driving the current large data deluge?

Climate & Weather model data

Reanalysis data

Satellite Observations

Next domain with large growth is potentially genomics
Some focus area for these Datasets

**High Performance**
Ensure that the data is organized to allow effective access for High Performance applications

**FAIR**
Findable, Accessible, Interoperable, Re-usable

**Data Integrity**
Managed services to capture a workflow’s process as a comparable, traceable output. Repeatable science which can be conducted with less effort or an acceleration of outputs.

**Transdisciplinary**
To publish, catalogue and access self-documented data and software for enhancing transdisciplinary, big data science within interoperable data services and protocols.
Services-based approach with programmatic access

NCI NERDIP DATA SERVICES

- OGC Web Feature Service
- OGC Web Map Service
- OGC Web Coverage Service
- OPeNDAP
- OGC Web Processing Service

Technology

- GeoServer
- THREDDS
- GSKY

Services

- OGC Web Feature Service
- OGC Web Map Service
- OGC Web Coverage Service
- OPeNDAP
- OGC Web Processing Service

NCI Index Database

10 PB NCI NERDIP EARTH SYSTEMS, ENVIRONMENTAL AND SOLID EARTH DATA COLLECTIONS

- Landsat
- MODIS
- Himawari
- CMIP 5
- Numerical Weather Prediction
- Geophysics
- Hazards Models
- Bathymetry
- Elevation
- GPS
“Mischief Managed” – Ongoing improvements to managing data

- Managing the ongoing agreement with data providers and usage for others analysis/use
  - update frequency, licensing / access, provenance
  - Multi-domain use-cases, improving for data analysis, tracking success, training examples
- **Using technology-assisted process to improve ability to manage data at scale**
  - automated conformance checks, data quality and SLA
  - information exchanges of data / informatics
  - automated exchanges in DMP - tracking growth, usage, hot-spots
  - impact measures, citation
GSKY: A high performance, scalable, OGC Data-server

GSKY Userguide: https://gsky.readthedocs.io/en/latest/
GSKY service: http://gsky.nci.org.au/
Open Source Project: https://github.com/nci/gsky

Features
- OGC Standards (WMS, WCS, WPS)
- Scalable

© National Computational Infrastructure 2019
The Innovation Game    iCAS, Sept 2019    Ben.Evans@anu.edu.au
GSKY responds to Open Geospatial Consortium (OGC) API over http protocol:

- Web Map Service (for displaying the images on the map server)
- Web Coverage Service (for delivering the actual data as “coverages” - independent of the underlying storage format or files)
- Web Processing service

GSKY allows:

- Performant aggregations, subsetting, subsampling, polygon/pencil/pixel drills
- Execution of on-the-fly data transformations, re-projections and other algorithms

GSKY is implemented using

- Rich metadata server for data query e.g., spatial, temporal, other physical variables
- Clustered backend workers – high performance I/O and scale-out server-side compute
GSKY High Level Architecture

- Based on “flow-based programming” and “stream processing”.
- Data transformed by connected processes, forming a Directed Acyclic Graph (DAG).

GSKY + MAS + Data + ...
How do we find all that hidden data in millions of files?
2018: 300+M.  2020: 1B

data stores with poor latency -> geospatial-enabled databases

MAS: Metadata attribute service
Integrated with GSKY and now other tools and services
Provides database backend abstraction over the data

• It identifies individual data objects (datasets, variables, spatial and temporal extents).
• Can be sharded by different geospatial collections or by splitting collections into non-overlapping geographical extents.

As a good database, is able to process complex queries in milliseconds
Geospatially, temporally and variable-aware over X millions of files
POSIX is no longer in the right game, league, ...
GSKY-MAS file handling of common large Earth Observation datasets
Amount of data is scaling up with more modern satellite instruments
GSKY flexible to use other data formats (e.g., CoG and zarr)

MODIS
MODIS data from 2001 – present, with 4 timestamps per month and 1k files per timestamp.
GSKY is handling approx. 900k MODIS files as a coherent dataset.

Landsat-8 dataset
Landsat 8 dataset has data from 2013 – present, approximately 3k files per timestamp.
GSKY is handling approx. 7 million files.

Sentinel 2 ARD
Approx 1400 available days (2015-present) and about 12k files per day for this dataset.
GSKY is handling approx. 17 million files.
Managing Sparse very large satellite data via GSKY services
Comparing Observation and models

The Innovation Game

iCAS, Sept 2019

Ben.Evans@anu.edu.au
First demonstrations of CMIP6 data comparisons via GSKY
GSKY providing server-side analysis capability
ASTER geophysics data
Problem handling large variables
- 62 GBytes

Using:
• default TDS -> times out
• Tuned TDS -> 90s
• GSKY -> 1s
1. **Change detection / Feature extraction / Classification**

   Case study: Landsat 8 / OSM road detection

2. **Derive observed precipitation**

   *Case study: Geopotential height for three different levels of the atmosphere from*ERA-Interim climate model using Python Tensor objects that train neural networks*
Challenge 4: Tracking (research) Impact and use
Interoperability with following services
- ORCID
- DataCite
- CrossRef Event Data
- Scholix (Publishers)
- NCI
- GRID (Digital Science)
- SpringerNature
- GESIS
- Research Data Australia
- ARC, NHMRC, NIH
- Dryad,
- CERN,
- figshare
HPC innovation challenges:

• Technology Changes in HPC centres
• Transparency in methods and outputs, understanding scientific tolerances
• Data for HPC, analysis and interoperability
• Better tracking (research) impact