



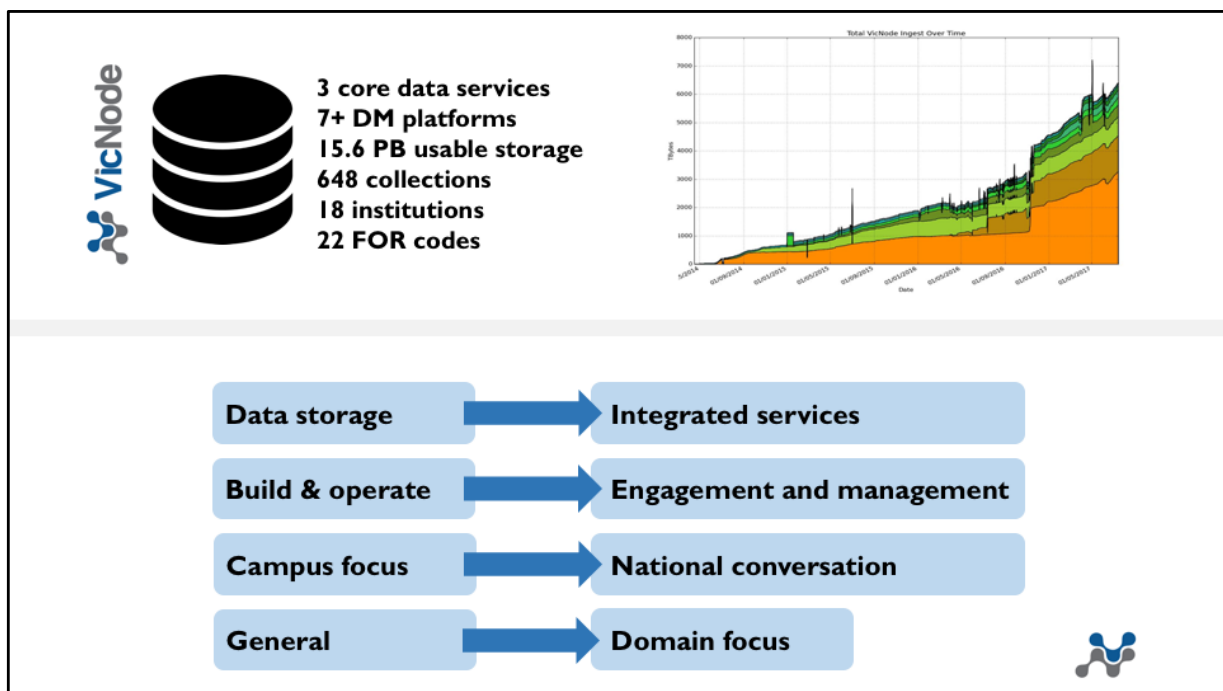
Enhancing research with VicNode: from campus instruments to international data consortia

Dr. Steven Manos –Vicnode, University of Melbourne
Stephen Dart –Vicnode, Monash University

CSIRO Conference on Computational and Data Intensive Science 2017



Aiming to show you today how Vicnode supports not only data storage but in concert with other investments and platforms supports large-scale integration of digital research platforms, at campus, state, national and international scales.



“Where Researchers can easily store and share research data (collections) through an affordable, secure and sustainable service”

- VicNode established as a primary node of RDSI as a joint partnership between the University of Melbourne, acting as the lead agent, and Monash University in late 2012, on behalf of all Victorian universities.
- Project teams established and initiated in December 2012 and hardware installed and available in May 2014 at both Universities. Storage expansions occurred in 2015 and 2016.
- Initial collections ingested were made available through a Merit Allocation process for collections of national and Victorian significance which ran from September 2013 through to June 2015.
- ✓ 101 research data collections of national significance (ReDS funded..)
- ✓ 6 research data collections of Victorian research significance

In total...


- 5.6 PB of usable storage at the University of Melbourne
- 10 PB of usable storage at Monash University
- 648 research data collections

- 18 institutions
- Representing all 22 FOR codes with a strength in Biological sciences, Engineering, Medical and Health Sciences, and Physical Sciences


The collection are one thing, but what Vicnode and RDS did in Victoria was fundamentally shift the conversations we were having and at the scale we were having them.

1. Increasingly Vicnode is a vehicle for conversations and needs beyond storage, as it is strongly tied to eResearch or Research ICT support teams at Melbourne and Monash, who are both also nodes of the Nectar research cloud. The conversations used to be about isolated storage allocations, but have now moved to integrated conversations around data + platforms, HPC, cloud, and so on.
2. Federal investment funded people to not only build, but the people to engage, manage collections , operate, etc. Value now is in on boarding and support.
3. Victorian benefit – there’s a coordinated approach – a Victorian view of the national landscape, ensuring that we’re involved in this initiatives.
4. The RDSI project has been a vehicle to have national conversations on (generic) data, and that continues to this day through domain activities. Exploring the data needs of domains – sharpening our understanding how to support these domains.


Communities




Culture and communities




Health and medical



Characterisation and imaging



Life sciences (genomics)



Omics

Platforms

	Aspera	figshare	DaRIS	Mediaflux	MyTardis	Omeka	OMERO
Presentation Layer	●	●	●	●	●	●	●
Visual representation of data		●	●	●	●	●	●
Publishing Capability		●	●	●	●	●	●
Suitable for big data	●		●	●	●		
Customisation / extendibility		●	●	●	●	●	

As VicNode and the broader national RDS program has evolved, there has been a focus on bridging the gap between digital infrastructure and research communities, through engagement and building platforms specific to those communities. Examples have included...

Culture and Communities – spent significant effort to understand HASS researcher needs and identify important cultural research datasets that would benefit from being digitised and made publically available. Regular training workshops are held to teach HASS researchers how to use Omeka as a means to display their data

Health and Medical research - identified security controls required for storing human and clinical datasets. Each node can now provide compliance levels to prospective data custodians of health and medical research datasets

Image and characterisation – Over 100 instruments in Victoria have been directly connected to cloud-based data management software to ensure that data generated by these instruments is automatically captured, managed and can be delivered to the cloud or downloaded to a desktop for analysis.

Life Sciences – Development of the GVL + run regular workshops to train researchers on how to use the GVL, mGVL

The Vicnode office also holds data custodian meet ups run every quarter to better understand our user base and their needs

There are various middleware platforms which have been established and are operated...

Aspera - data transfer technology that can significantly reduce the time taken to transfer data over long distances.

DaRIS - supply biomedical imaging data management and integration with instruments and research computing infrastructure.

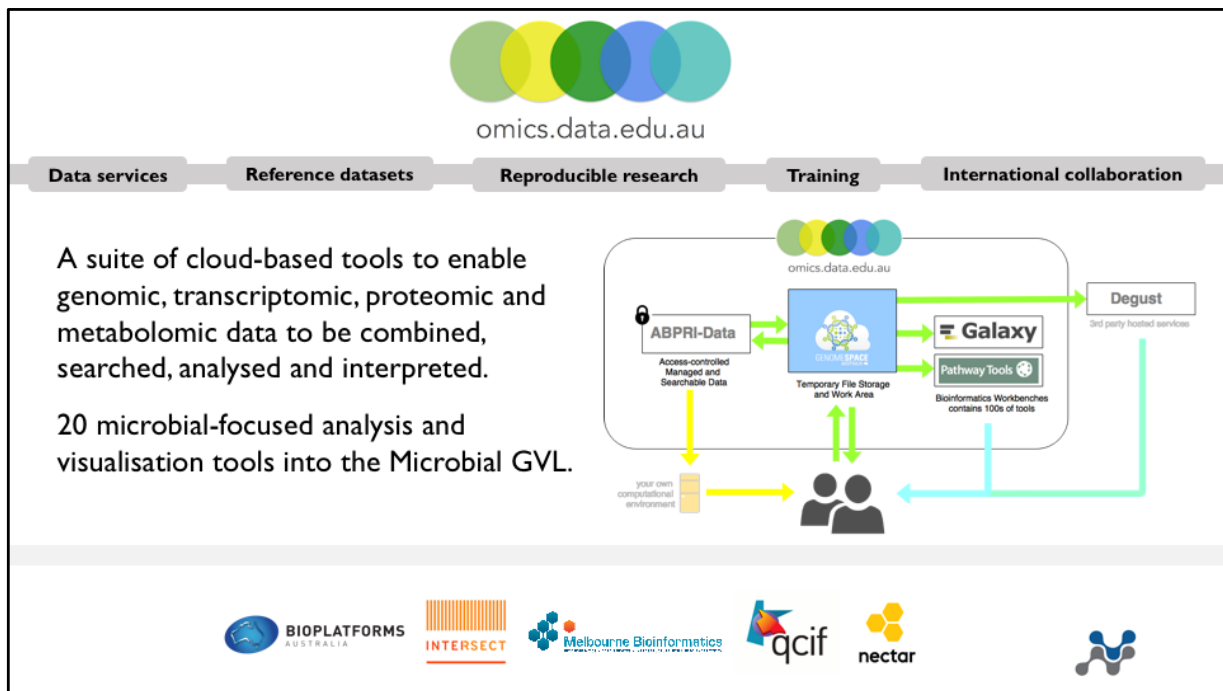
Figshare – tool to make research outputs shareable and/or publicly available.

Mediaflux – data management platform to manage any type of unstructured data, as well as structured data, and relationships between structured and unstructured data. It can be used for ingesting, storing, discovering and sharing any type of data.

MyTardis - open source solution built at Monash to allow to store large datasets and share them with collaborators online with a focus is on integration with scientific instruments, instrument facilities and research lab file storage.

Omeka - open source web-publishing platform for the display of archives, library, museum, and scholarly collections and exhibits.

Omero – tool that handles all your images in a secure central repository from microscope to publication



The RDS Omics project is an [RDS-funded flagship project](#) to provide cloud-based data services and tools for Australian Life Science Researchers to combine, analyse and interpret genomic, transcriptomic proteomic and metabolomic data.

Project is a partnership between RDS Nodes VicNode, QCIF, Intersect and Melbourne Bioinformatics. All aspects of the platform leverage previous NCRIS investments (RDSI, RDS, NeCTAR).

The data focus has been the BPA Australia Antibiotic Resistant Pathogens initiative framework dataset.

It is first Australian platform to allow 4 distinct 'omics' data types to be:

- co-analysed and stored in one system;
- managed through a common data management system;
- able to have bioinformatics analysis performed on these data via a common interface, as well as implement reproducible pipelines and protocols developed by omics experts, published within the data services platform.
- published to international repositories

And a component of **Training and education – producing materials and training**

programs which support the use of cloudbased eResearch infrastructure for multi-omics analysis.

The platform itself is made up of three main parts :

- Your temporary working data storage area: GenomeSpace
- A secure, searchable repository of omics data (Data Management Platform using RDS Storage)
- Bioinformatics workbenches with many tools:

To date...

Project has deployed 20 specialised microbial focused analysis and visualisation tools into the Microbial GVL ... along with self guided tutorials for each tool

These are across variant calling, genome annotation, protein and metabolite identification.. And the overlay of multi-omics data on metabolic pathways

The platform is currently available to the BPA consortium and going through the phase of user testing and training.

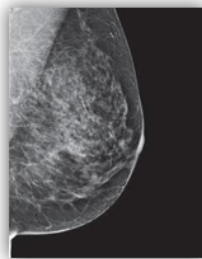
- An NBCF/Cancer Australia funded project (\$5M+)
- 54,000 + participants



- Consent, contract and correspondence tracking



- 500K+ DICOM image repository
- On-call for MD processing & analyses
- Future multimodal integration



An NBCF/Cancer Australia funded project - \$5 million+ investment

A national resource for women's health research, particularly breast cancer research

Recruiting women, nationally, who attend BreastScreen and to date, over 54,000 women are participating

For each participant, the project holds:

Consent and contact information

Comprehensive epidemiology questionnaire data

Several mammograms, captured by BreastScreen

There are also over 8K+blood biospecimens; 1200+ saliva biospecimens

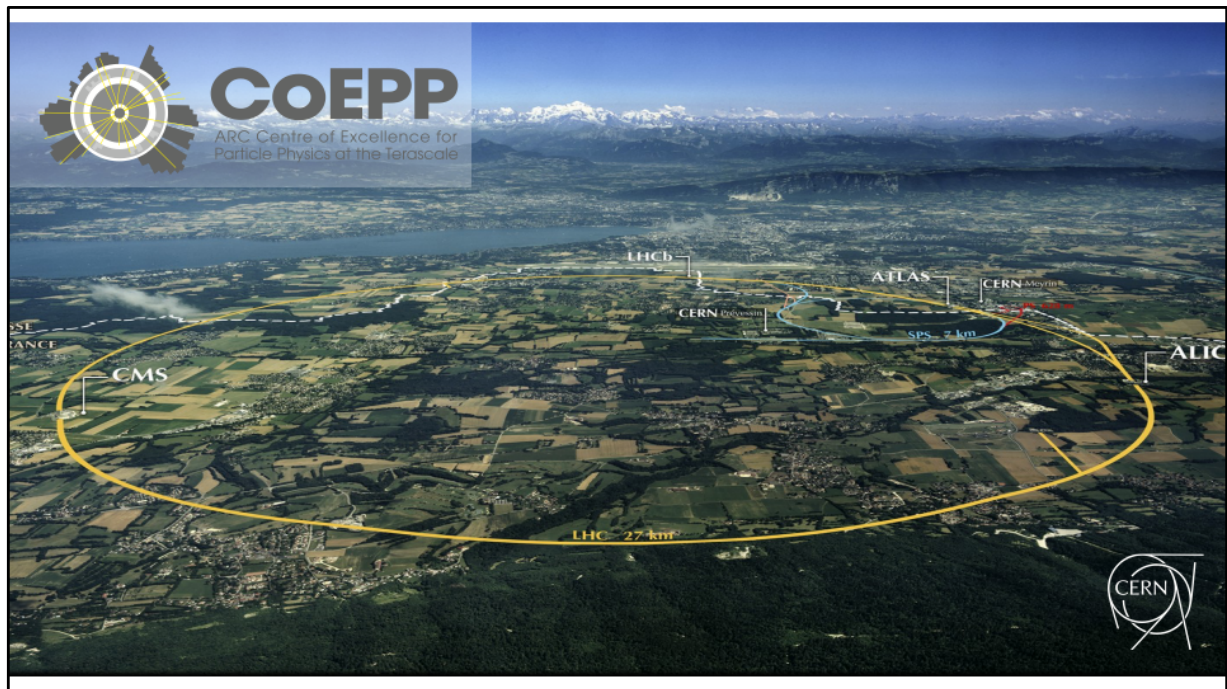
There are various elements to the underpinning digital infrastructure supporting lifepool

1. The Ark (<https://sphinx.org.au/the-ark>, NeCTAR research cloud hosted)

- An open-source web-based biomedical data management system
- For *lifepool*, primarily a consent, contact, and correspondence tracking database
- Questionnaire data - Microsoft SQL Server, Faculty hosted
- Scanned consent forms and paper questionnaires - gigabyte magnitude, stored on Faculty-provided network storage
- Mammograms
 - Approximately 500,000 image files in DICOM format
 - Totaling 17 terabytes of image data
 - Supporting Mammographic Density (MD) breast cancer research
 - VicNode hosted

Supports...

- Fine-grained access control, high-speed, secure transfer of image data t
- Potential to integrate mammogram storage w/ The Ark and MS SQL Server



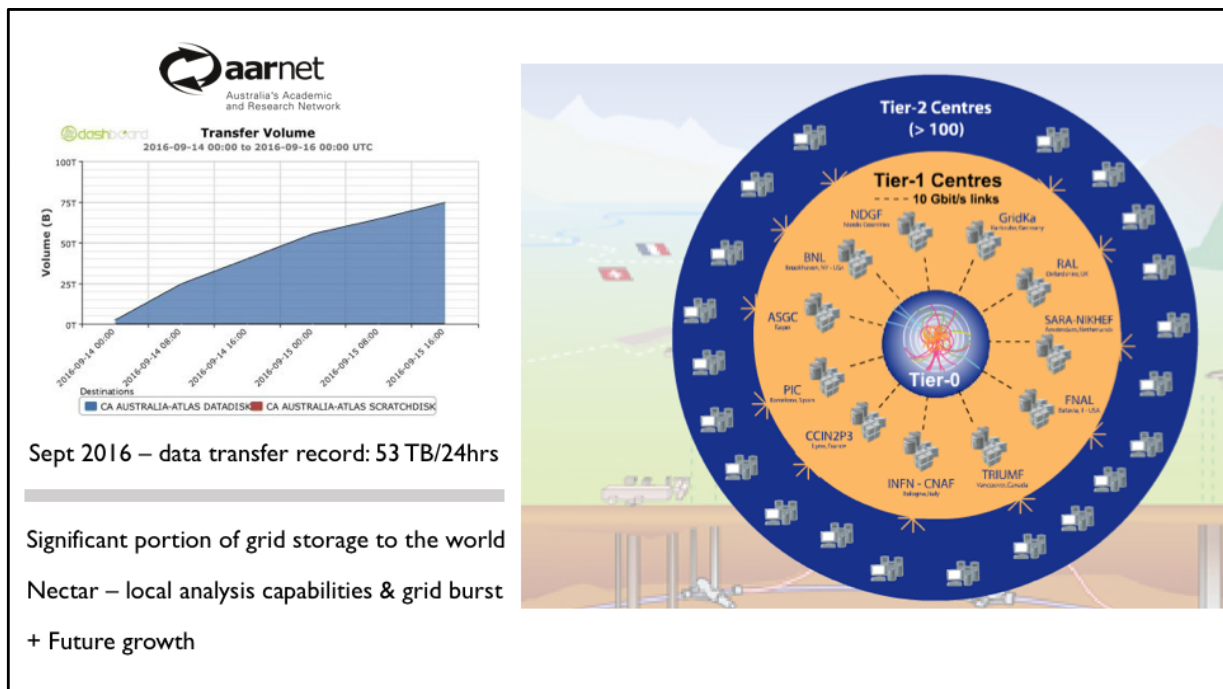
CoEPP is an Australian Research Council Centre of Excellence funded in 2011 for 7 years. \$35m plus \$7m in-kind funding.

Aerial shot of superimposed location of LHC particle accelerator tunnel near Lake Geneva – 27km circumference - and the 4 experiments at beam cross-over/collision points.

The CoEPP is one of the 170+ grid-connected computing centres in 42 countries worldwide that provide the linked-up computing and storage facilities required for analysing the ~30 Petabytes (30 million gigabytes) of data CERN's LHC produces annually.



ATLAS is one of two general-purpose detectors at LHC
investigates wide range of physics
 discovered Higgs boson
 searching for extra dimensions and particles that make up dark matter
particles collide at ~ 1 billion times per second
fast triggers keep data of only “interesting” events but still 17.4 PB of data for 2016



ATLAS-Australia a TIER 2 grid site

LHC chose grid computing to share the cost and expertise across all participating countries/institutes.

Tier 0

data recording, initial data reconstruction, data distribution

Tier 1

Permanent storage, re-processing, analysis

Tier 2

Simulation, researcher analysis

Tier 3

final analysis on local clusters/servers, non-grid resources

1.5PB grid storage (~350TB VicNode)

2k CPU (+1k CPU bursting into Nectar RC)

The network has proven to be incredibly robust. ... record last September in 2016... 53 terabytes transferred in 24hrs, at 100% efficiency, (nearly 5Gb/sec sustained over 24 hours)

University of Melbourne and a research network-connected site located in Germany.

Vicnode...

provides significant proportion of our grid storage to the world (RDS)

provides all our local end user analysis compute capabilities (Nectar RC)

provides grid burst capabilities during special production runs or busy machine periods (RC)

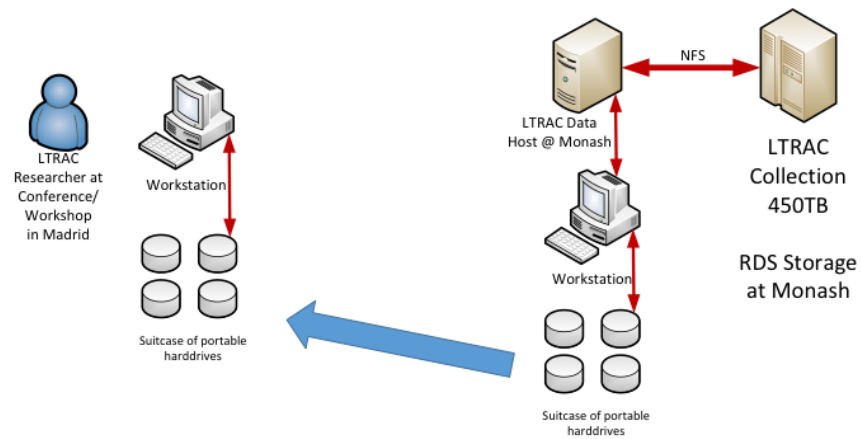
provides the platform for our future growth needs in both storage and compute

Assisting Research Collaboration using high speed data transfer

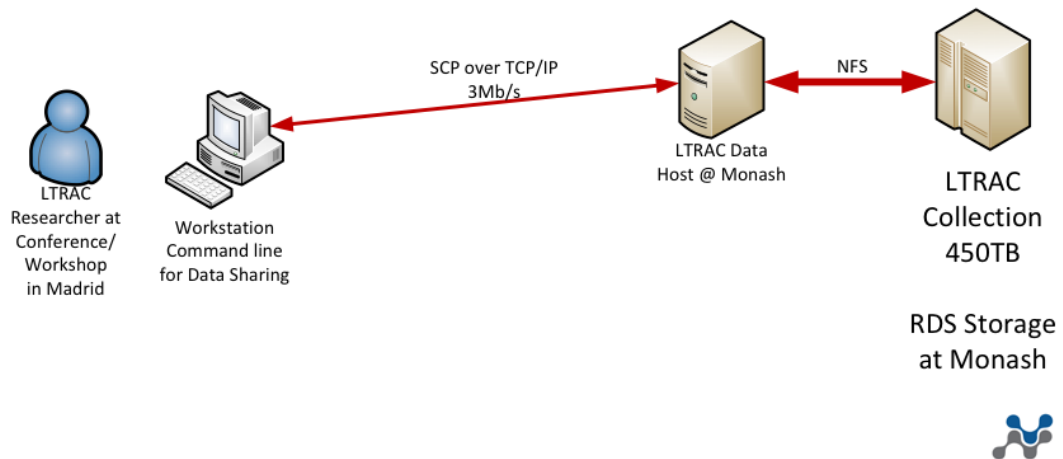
- 40 turbulence experts from around the world are attending a workshop to advance in the understanding of Coherent Structures in Wall-bounded Turbulence (COTURB Project).
- Conference and Workshop in Madrid
- Multiple Monash LTRAC Team members attending
- Large Collection ~450TB
- Not certain which parts are going to be useful
- Cannot take the entire collection



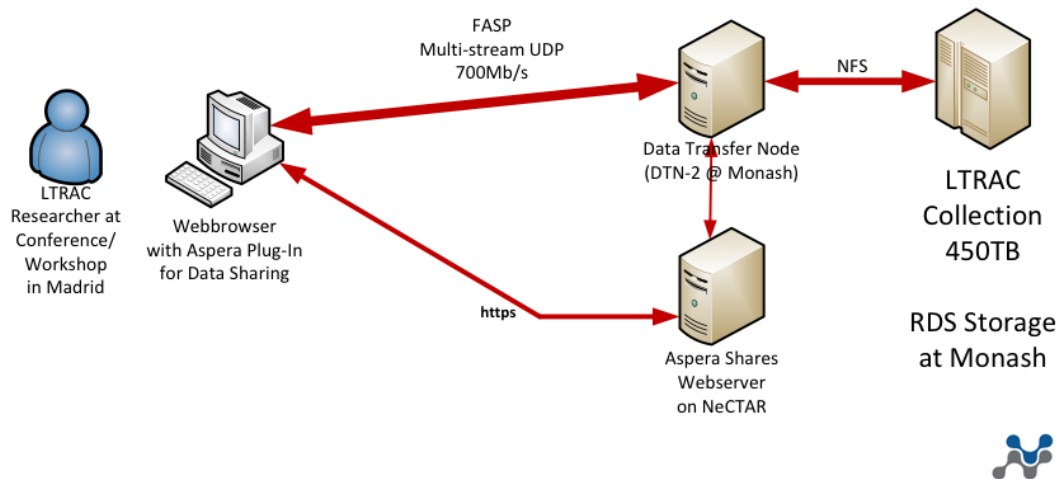
Carry the data there by hand luggage



Use SCP to the LTRAC host at Monash



Use Aspera with RDS/NeCTAR/DaShNet





Thanks to: Atsushi SEKIMOTO, Research Fellow
Laboratory for Turbulence Research in Aerospace & Combustion (LTRAC)
Department of Mechanical and Aerospace Engineering
Monash University



-finis-

