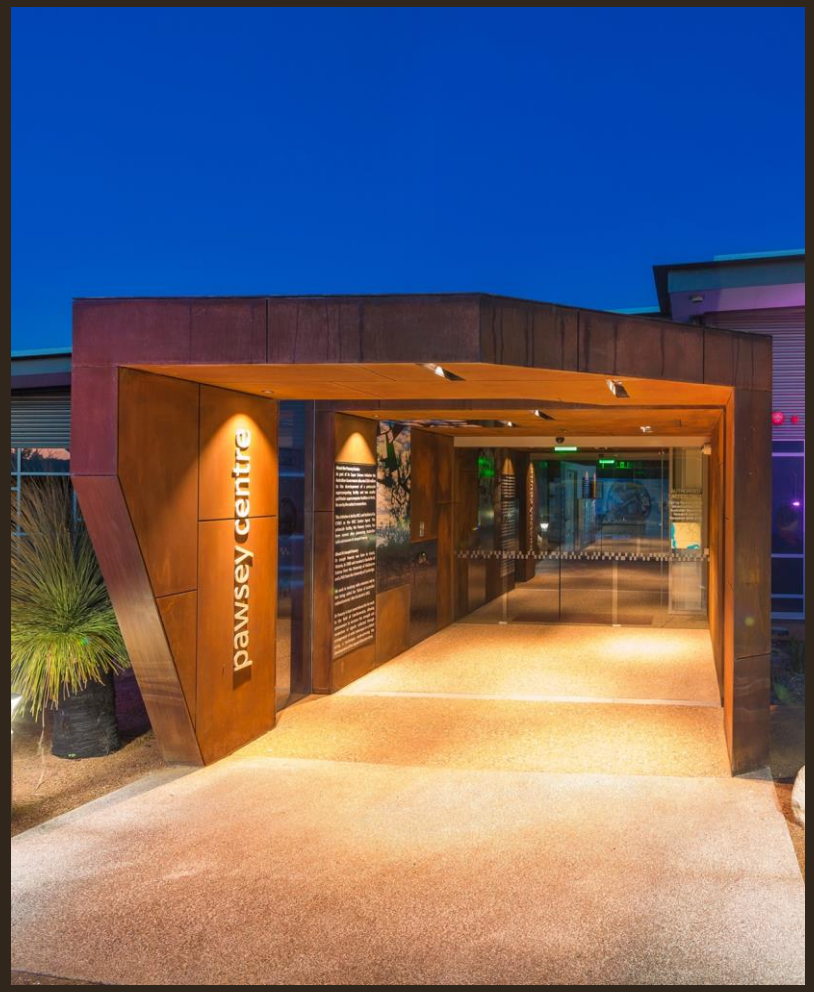


Innovative Long-Term Storage Solutions for HPC: The Pawsey Supercomputing Research Centre's Next Generation Architecture

Chris Schlipalius, Storage Manager, The Pawsey Supercomputing Centre
Bruce Gilpin, CEO, Versity Software

Introduction to Speakers

Chris Schlipalius
Storage Manager
Worked in University IT for 15 years - in both Enterprise IT and Research (storage, servers and data systems).
Pawsey Research Storage for 10 years

Bruce Gilpin
CEO and Co-founder of Versity Software

Describing the Pawsey Supercomputing Research Centre

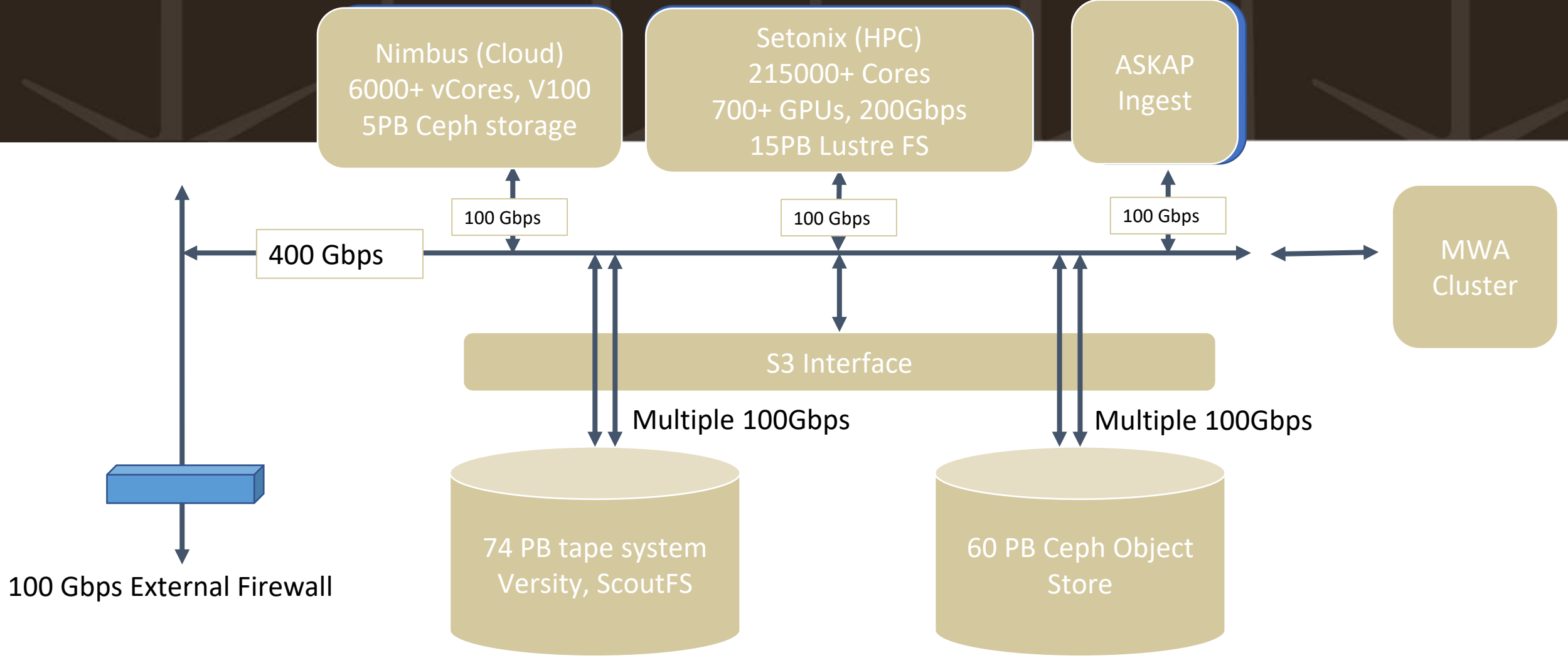
Australian Tier 1 Research Facility for HPC in Australia for all Australian researchers who have projects or scientific work of national merit.

- One of two national Tier 1 facilities in Australia (the other is NCI on the ANU campus in Canberra).
- We are physically located in Perth, Western Australia on Whadjuk country. The data Centre is operated by the CSIRO (CBIS).
- Our data centre was designed and built by the CSIRO in 2011 in order to support precursor SKA projects who operate in the remote radio quiet zone in Western Australia on Wajarri Yamatji country, 800km NNE of Perth – Murchison Radio Observatory (MRO).
- The Pawsey Centre (launched as such in 2014) is a non-incorporated joint venture between the four public universities of Western Australia and the CSIRO (National Science agency). Previously it was known as iVEC and started in the late 1990's (systems originally were distributed at University buildings).
- Funding for Infrastructure and Capital projects is from the Australian Commonwealth Government.
- ~60 Staff - operational costs are funded by the Western Australian Government.
- We support both General Science and Radio Astronomy Science, and we provide the majority of CPU hours to Australian researchers via NCMAS.

Setonix – HPE/CRAY Shasta HPC

≈50 Petaflops
200000+ AMD Milan (CPU) Cores
750+ AMD MI-250X GPUS (128GB HBM2/GPU)
548+ TB system memory
0.5 PB Near-node NVMoE storage
15 PB ClusterStor Lustre filesystem with 2.7PB SSD
Slingshot-11
Was 4th on Green500 - Water-cooled to component, waste heat is exchanged into groundwater (GWC). GWC pumps are PV-powered.
- https://www.top500.org/lists/green500/2023/11/

Pawsey's Architecture 2024



Banksia Project

Deliverables

- Affordable
- Reuse our existing investment in tape – two libraries, 64 Drives, ~15,000 tapes.
- Scalable/Expandable (server nodes/filesystem)
- S3 Interface for integration with HPC workflows
- REST API – modernized interface
- Open tape format
- Prefer Open Source and no lock-in
- Subscription licensing.
- Migration of 53PB Data (in DMF6 format).

More about Banksia

Ok, so let’s next dive into what the archive is,

how it's used,

and then go into detail about what it looks like,

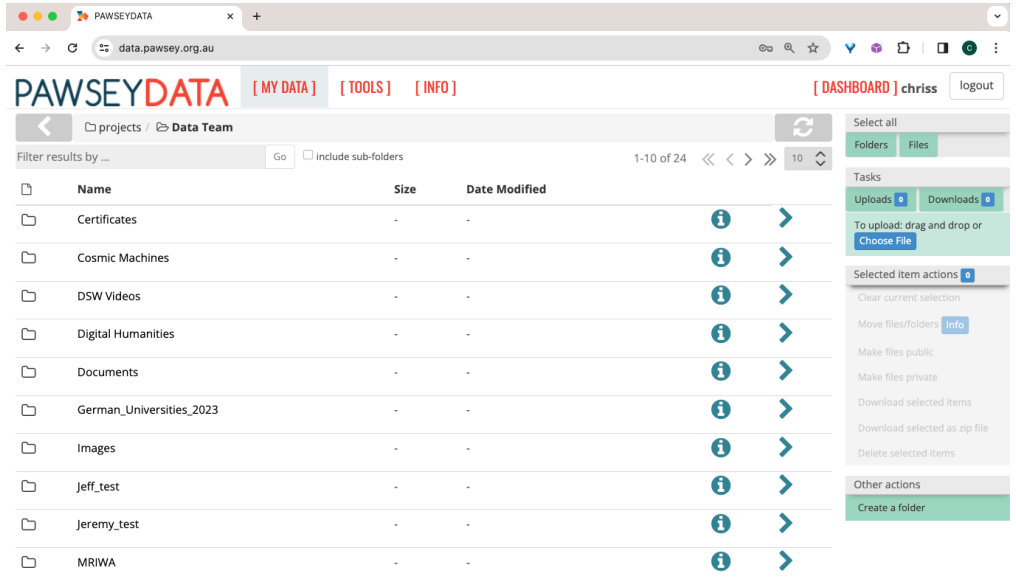
and I’ll talk about my experiences using and administrating it and why it is the best fit for the way we need to operate both now and in the future.

How is it used?

Three major Australian Scientific Research Data Services

- MWA ASVO
  - 38PB (was close to 45PB before some files were copied over to Acacia- Ceph S3).
  - Radio Astronomy observations going back over 20 years.
  - Largest publicly accessible research data set in Australia (to the best of my knowledge).
  - Frequently retrieved and used, reprocessed when necessary using new parameters (for e.g.).
- Pawsey Data Portal projects
  - Mediaflux software, custom interface made by Pawsey.
  - 13PB.
  - Over 60 data projects.
  - Able to support CLI integration for HPC jobs using pshell
- CASDA - CSIRO ASKAP Science Data Archive - testing 100TB – potentially 7.5PB.

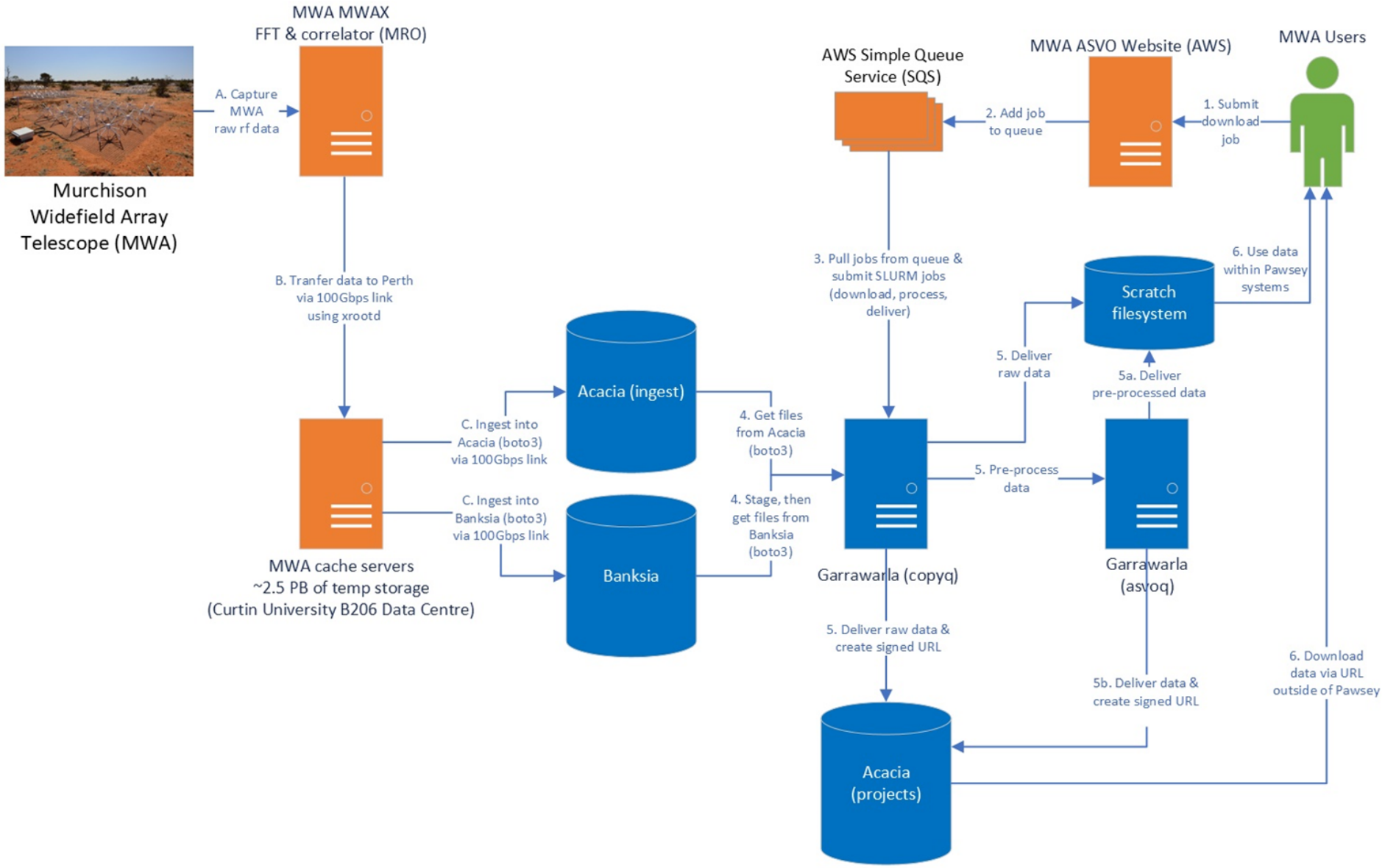
Pawsey Data Portal- Mediaflux



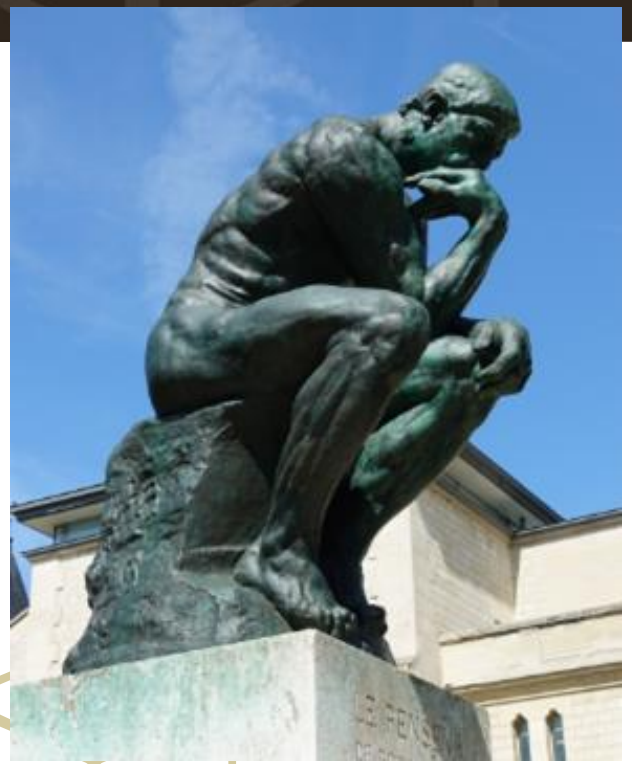
Our web GUI and CLI for Australian research data projects

- Data Allocations are merit based and have annual routine review and retention periods.
- It has a web front end and CLI – *pshell* and *pmount* utilities for use with pre and post batch run SLURM scripts for HPC workload to stage files ready for jobs and move results after runs.
- We customized the web front end over a number of years (new update due soon).
- We paid for the development work for Mediaflux to use Versity ScoutAM REST API

Banksia holds entire publicly-available data set for the MWA Virtual Observatory Archive



Administering Banksia



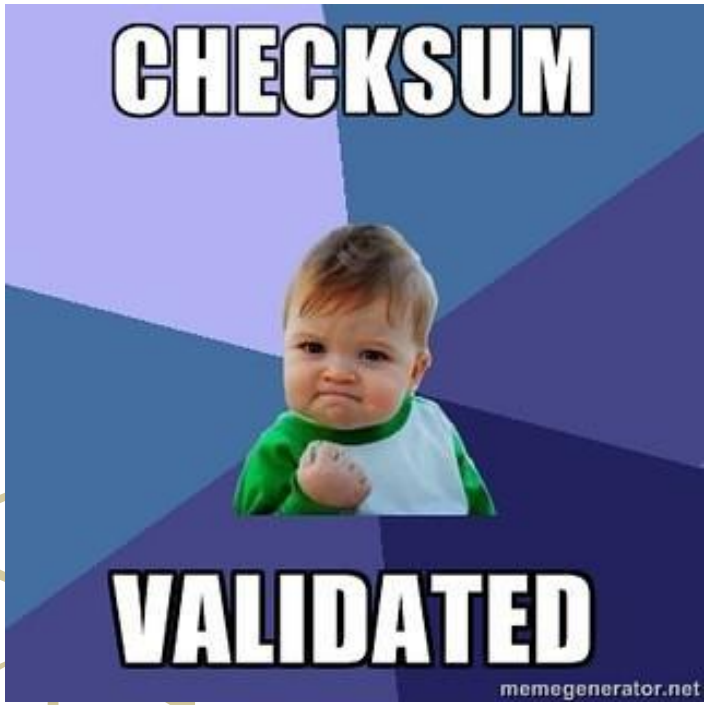
Questions:

Is it easy or hard to operate? How many people does it take?

Does it deliver what it promises? What other items did you need to use?

- Well, easy. It was designed to need less staff, the velocity was taken out of the libraries i.e. the churn and the many small files, less failing disks (than in CXFS), less failing drives (due to high use).
- Currently it takes 1.5 pax, soon to go to 2.5 pax, plus we have support contracts with DDN, Spectralogic and Xenon.
- Yes, it still needs some further enhancements and tooling - they are coming. Versity are very responsive and usually say yes.
- We need to get stuck into replicating damaged/failed tapes, plus mass migration of DMF vols - in order to sparse DMF read only tape vols.
- Other items we needed were Keycloak deployment, Kafka message bus for bulk stages for MWA ASVO, plus grafana, plus multiple iterations of the parser for DMF, oh and the S3 GW for boto3.

Banksia - Lessons learnt



Summary

1. The journey to RESTAPI, Object and migrating from DMF is ”not for the unprepared”. (Checksums are useful).
2. You need to support your users, help and retrain them, plus enter the “brave new world” for HSM System Administrators
3. Design and peer review and work with a good integrator (ie Xenon) and know precisely what you need.
4. Select and utilise a helpful, responsive and innovative software company to work together on functional enhancements and improvements.
5. Join the Versity Usergroup if you want to understand more about all this, useful if you work on backups or archive or HSM.


Versity

What We Do at Versity:

Manage large data collections at the lowest possible cost

- A Modern Software-Defined Storage Platform
- For Mass Storage & Large Archive Systems

Mission

- Improve the efficiency of long term data storage
- Make arrival storage easier to deploy and use
- Ensure that systems can scale to meet rapidly growing data collections
- Ensure vibrancy and innovation in the archival storage world

Trends

- More flash deployed in scratch tier (Vast, WEKA and others)
- Flash trend driving more demand for archival class storage
- Flash also changing the personality of the archive
- Much larger systems - Exabyte scale is becoming much more common
- More self contained and field serviceable rack modular tape systems
- More S3 or hybrid S3/Posix interfaces

Questions