

Galaxy Australia

A public bioinformatics compute resource
distributed nationally.

Simon Gladman (presented by Andrew Lonie)

Development partners



Supported by



Outline

- A bit about Galaxy Australia
- Nationally distributed compute & Pulsar
- Dynamic job handling & DTD
- Reference data management & CVMFS
- Monitoring and stats collection
- Next?

What is Galaxy?

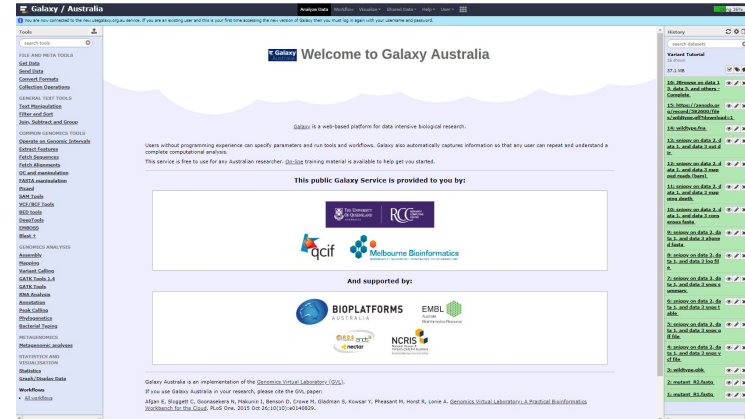


A screenshot of the Galaxy Australia web interface. The page has a dark header with navigation links like 'Home', 'Workflows', 'Databases', 'Shared Data', 'Help', and 'User'. Below the header is a light blue banner with the text 'Welcome to Galaxy Australia'. The main content area is white and contains a paragraph about Galaxy being a web-based platform for data-intensive biological research. It mentions that users without programming experience can specify parameters and run tools and workflows. Below this, there is a section titled 'This public Galaxy Service is provided to you by:' which lists logos for The University of Queensland, RCC, qcif, and Melbourne Bioinformatics. Another section titled 'And supported by:' lists logos for BIOPLATFORMS AUSTRALIA, EMBL Australia, NCRIS, and nector. At the bottom, there is a citation for the Genomics Virtual Laboratory (GVL) and a reference to a paper by Algan et al. (2015). On the left side, there is a sidebar with a search bar and a list of tool categories such as 'FILE AND META TOOLS', 'GENERAL TEXT TOOLS', 'COMMON GENOMICS TOOLS', 'VARIANT CALLING', 'ASSEMBLY', 'METAGENOMICS', 'STATISTICS AND VISUALISATION', and 'Workflows'. On the right side, there is a 'History' sidebar showing a list of recent datasets with their names and sizes.

What is Galaxy?

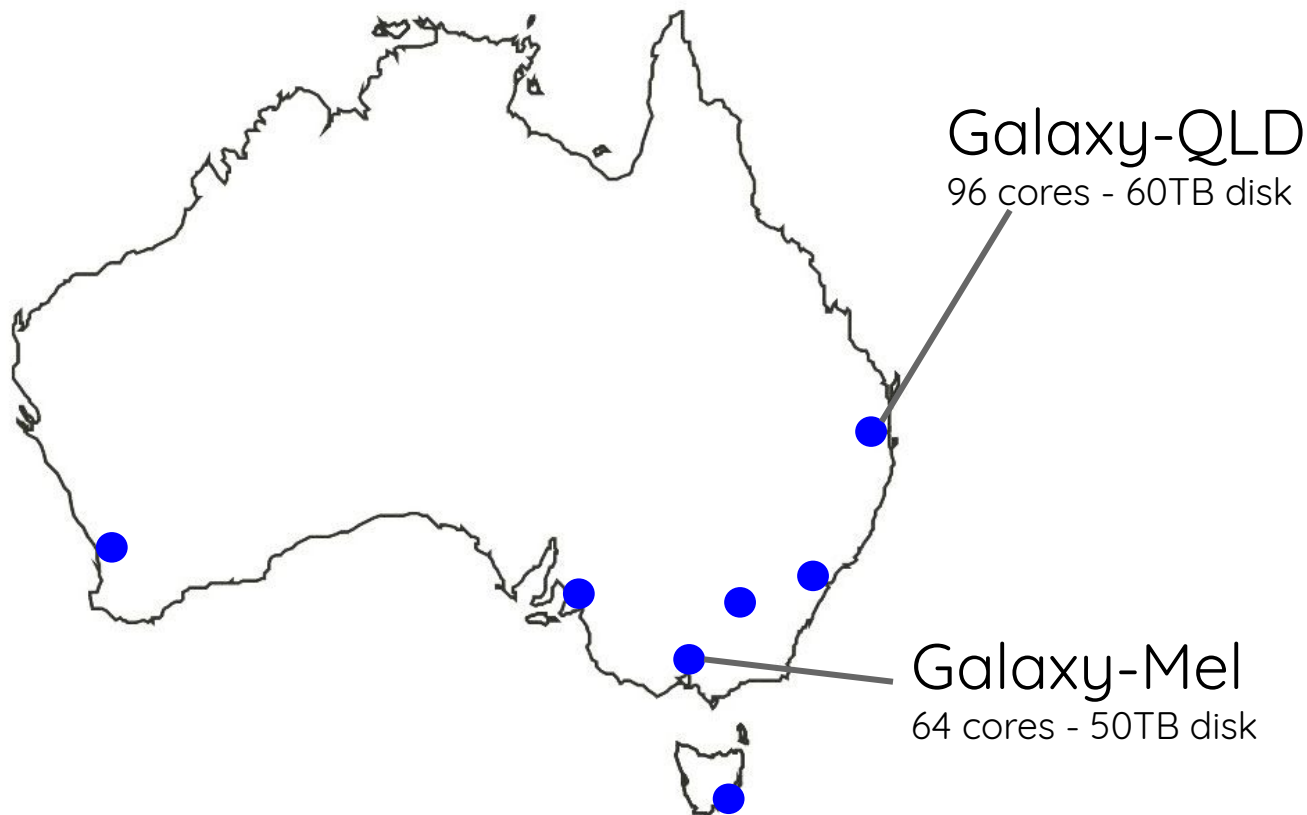


- Web based bioinformatics platform
- Retains “histories” of analyses
- Has large number of cutting edge tools
- Conda environments/Singularity containers for dependency management
- Has an app store
- Has a large support community



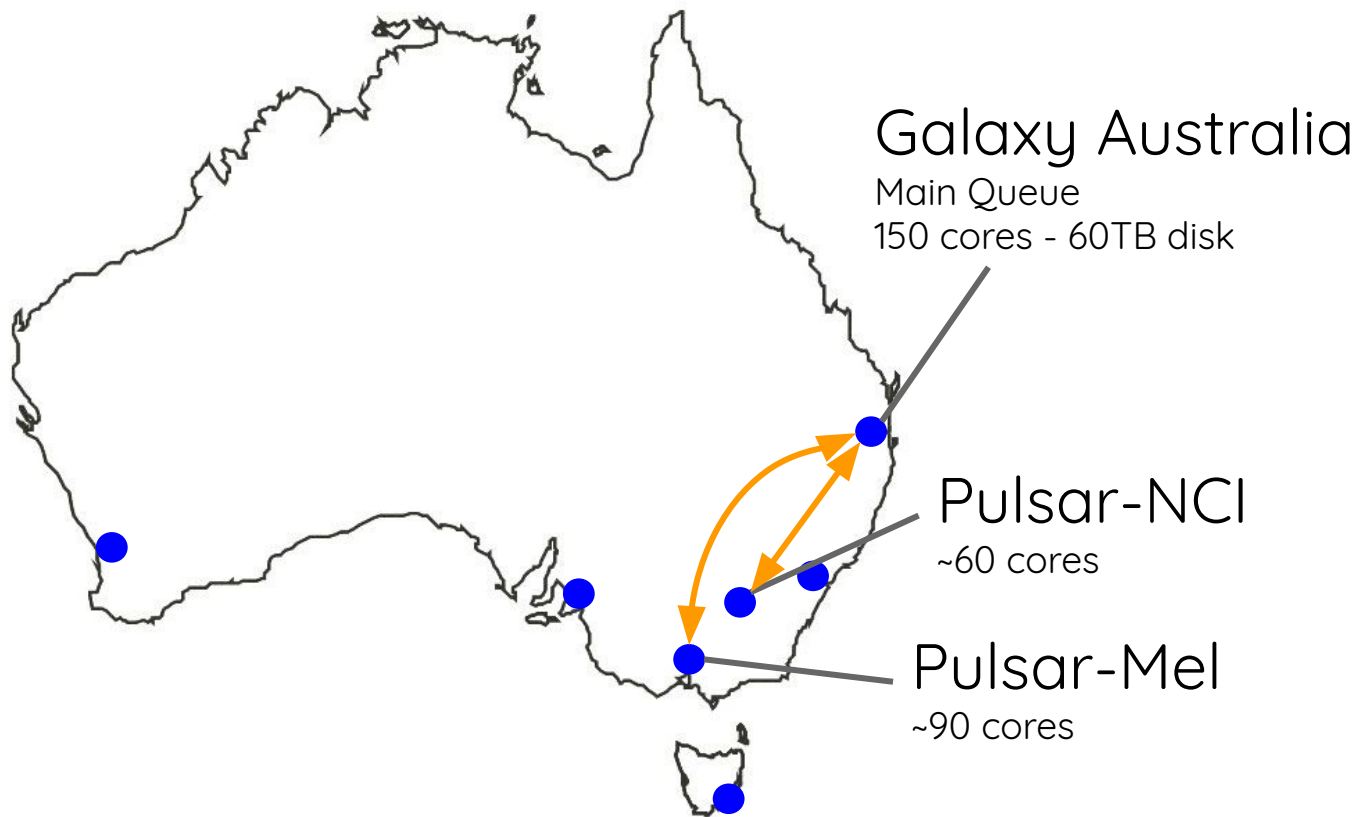
Galaxy Australia Infrastructure Locations

2015-2017



Galaxy Australia Infrastructure Locations

2018-now



Galaxy Australia - Usage

An active and engaged user community

2268

registered users.

608 active users (last 90 days)

User growth 2016 - 2018



Registered users in Australia from:



Australian Universities



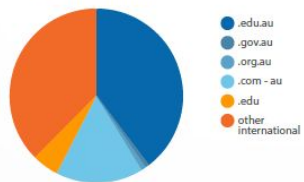
Medical Research Institutes or Organisations



Other Research Organisations

- Public and free Galaxy server in Australia
- Currently running > 50,000 jobs month.

Users per domain



Users represented across

338

organisations

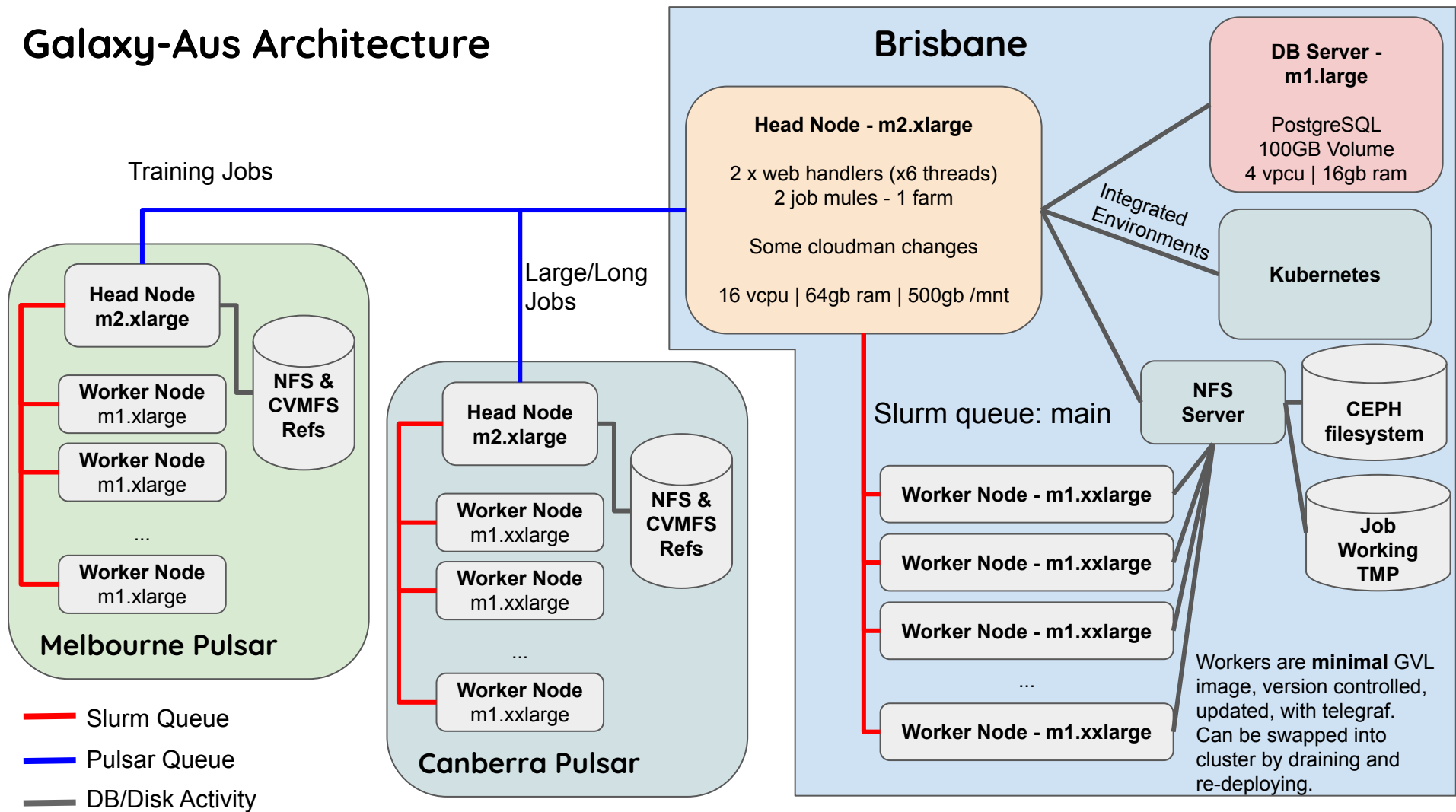
Users represented across

50

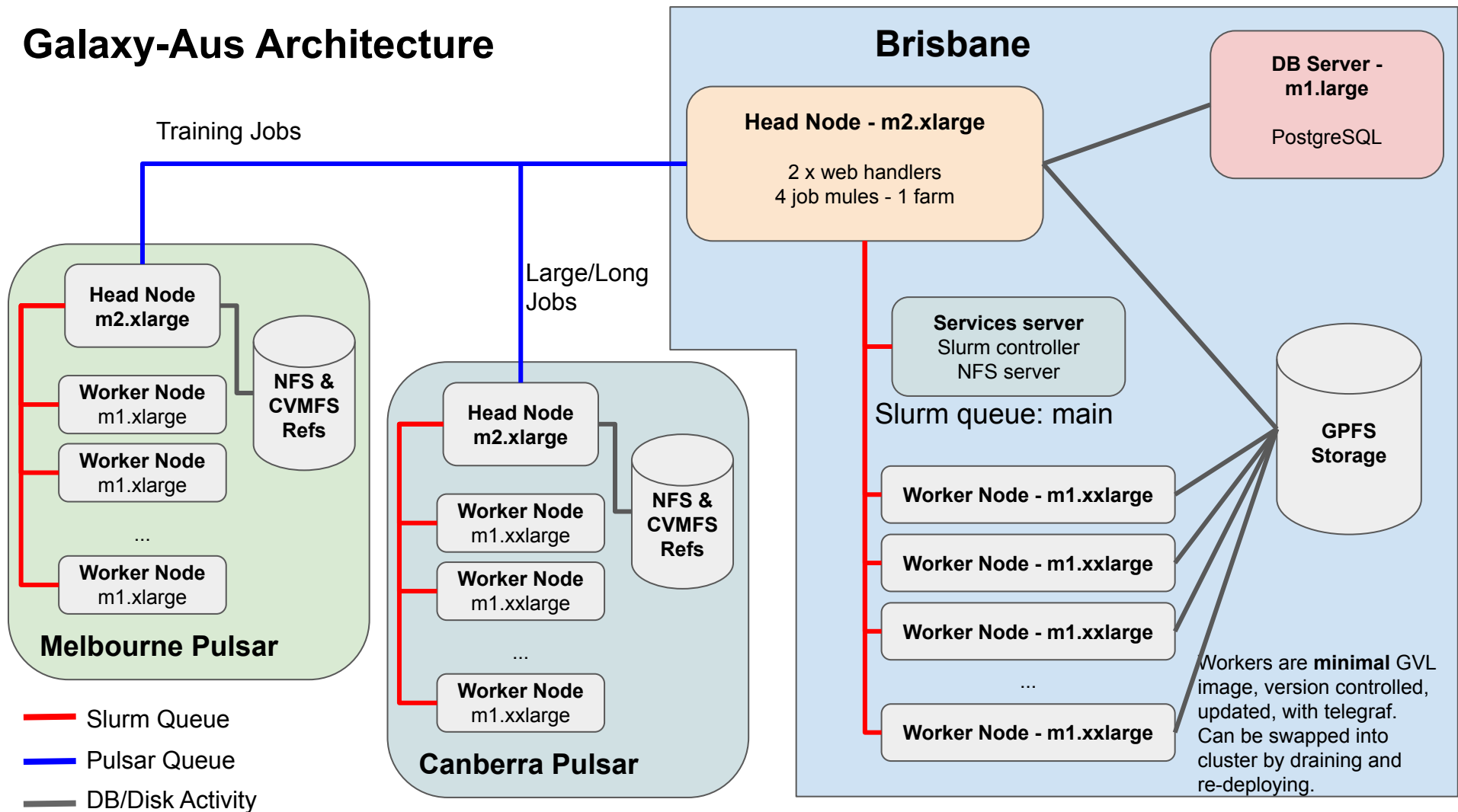
countries



Galaxy-Aus Architecture



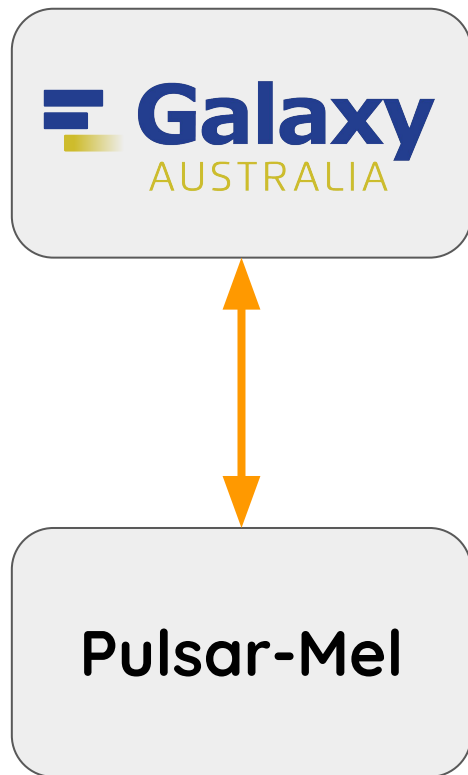
Galaxy-Aus Architecture



What is Pulsar?

- Python server application
- Allows a Galaxy server to run jobs on a remote system
- No shared file system required
- Configurable
- Securable
- Can submit jobs to HPC queueing system
- **Automatically handles dependency management**

How Pulsar works



1. User clicks “Execute”
2. Galaxy packs up and sends:
 - Data
 - Config files
 - Tool name & version
 - Parameters and other job metadata
3. Pulsar accepts the job
4. Pulsar checks if tool is installed locally
 - If not - Installs tool with Conda or Docker
5. Pulsar submits job to local queue
6. Pulsar waits until job complete
7. Pulsar packs up result and sends it back to Galaxy

What we use it for

Pulsar-Mel

- Training jobs / Workshops
- Runs lots of small jobs
- Usually 2-4 cores
- Allows a tutorial or workshop to run without interfering with main queue

Pulsar-NCI

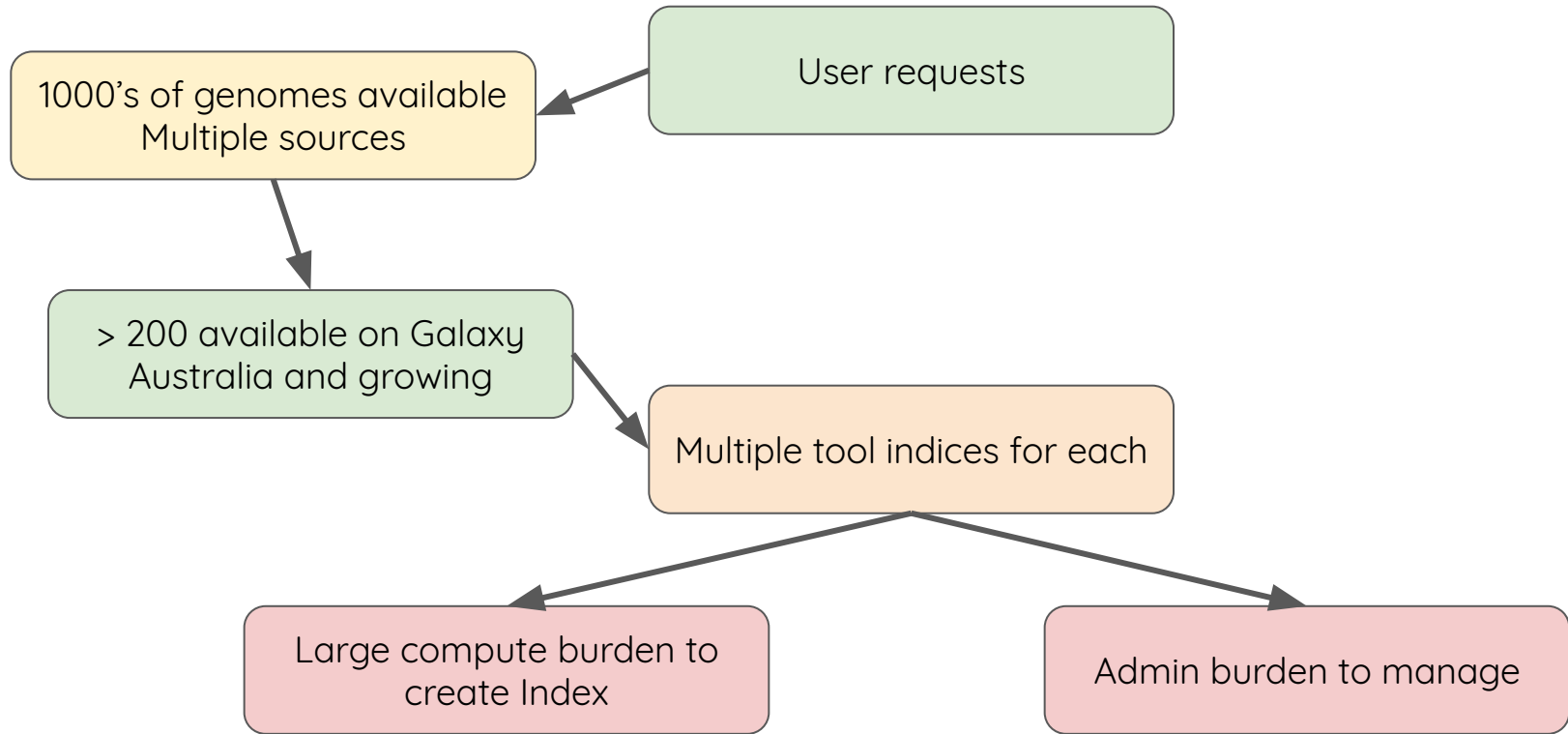
- Large long running jobs
- Jobs that could run for days
- Usually 16+ core jobs
- Lots of Disk-IO
- Genome assemblies / Transcriptome assemblies

Dynamic Tool Destinations

- Tool vs Destination rules
- YAML
- Can have rules for
 - File size
 - User
 - Tool parameters
- More complex rules in python functions
- Can alter without Galaxy restart

```
fail_message: too much data, please don't use Spad
destination: fail
default_destination: slurm_5slots
prokka:
  rules:
    - rule_type: file_size
      nice_value: 0
      lower_bound: 0
      upper_bound: 0.3 MB
      destination: pulsar-mel_small
    - rule_type: file_size
      nice_value: 0
      lower_bound: 0.3 MB
      upper_bound: 30 MB
      destination: slurm_7slots
      #Is normally slurm_7slots
      default_destination: slurm_7slots
fastqc:
  rules:
    - rule_type: file_size
      nice_value: 0
      lower_bound: 0
      upper_bound: 15 MB
      destination: pulsar-mel_small
    - rule_type: file_size
      nice_value: 0
      lower_bound: 15 MB
      upper_bound: 500 MB
      destination: pulsar-mel_mid
      # Is normally slurm_7slots
      default_destination: slurm_7slots
iuc_pear:
  rules:
    - rule_type: file_size
      nice_value: 0
      lower_bound: 0
      upper_bound: 15 MB
      destination: pulsar-mel_small
    - rule_type: file_size
```

The problem of reference data: There's a lot and it's hard to deal with



So we started talking with other large Galaxies!

Share the burden

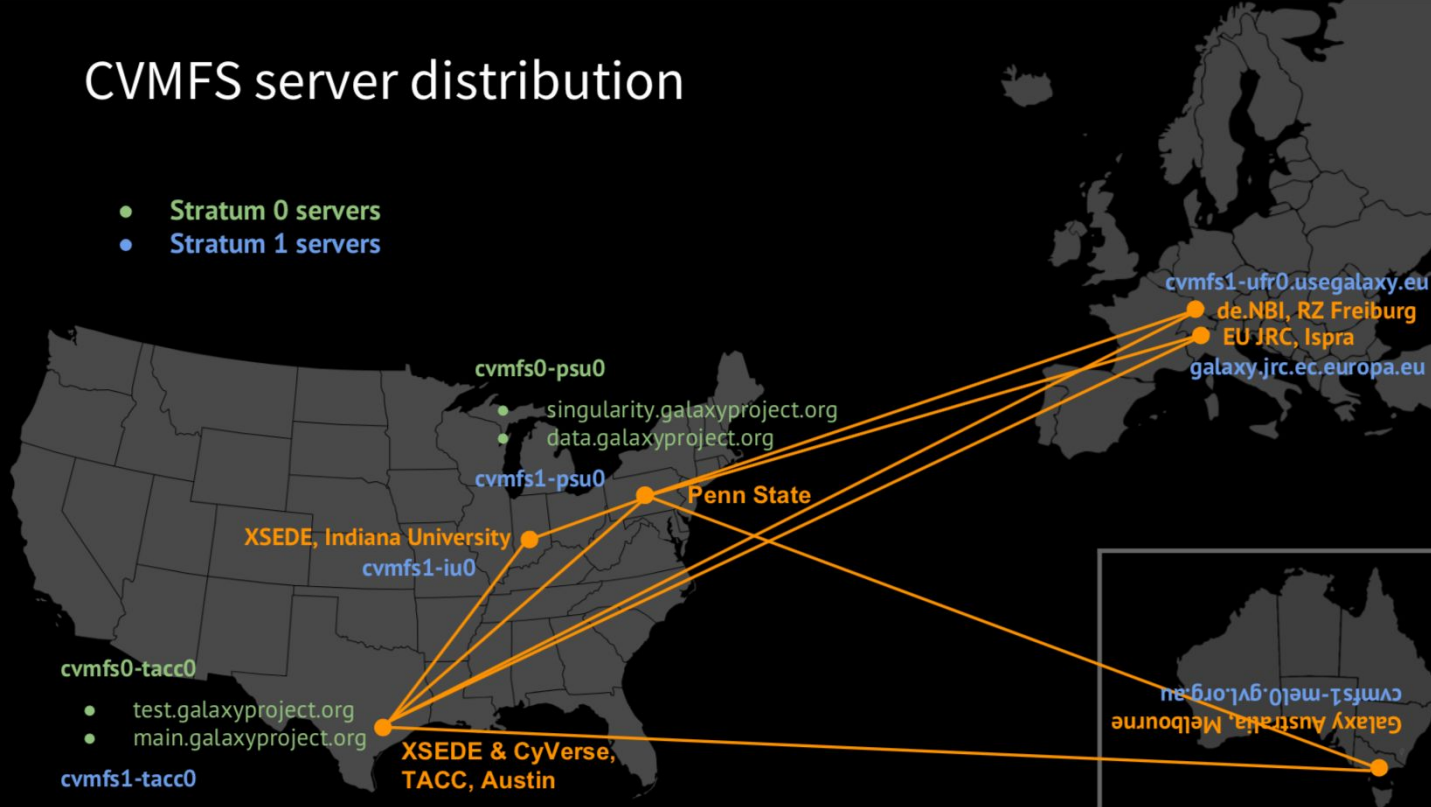
- Admin of references
- Compute requirements

We decided to share it all via CernVM-fs

- Currently controlled by the Galaxy Project
- Moving to a community model soon

CVMFS server distribution

- Stratum 0 servers
- Stratum 1 servers

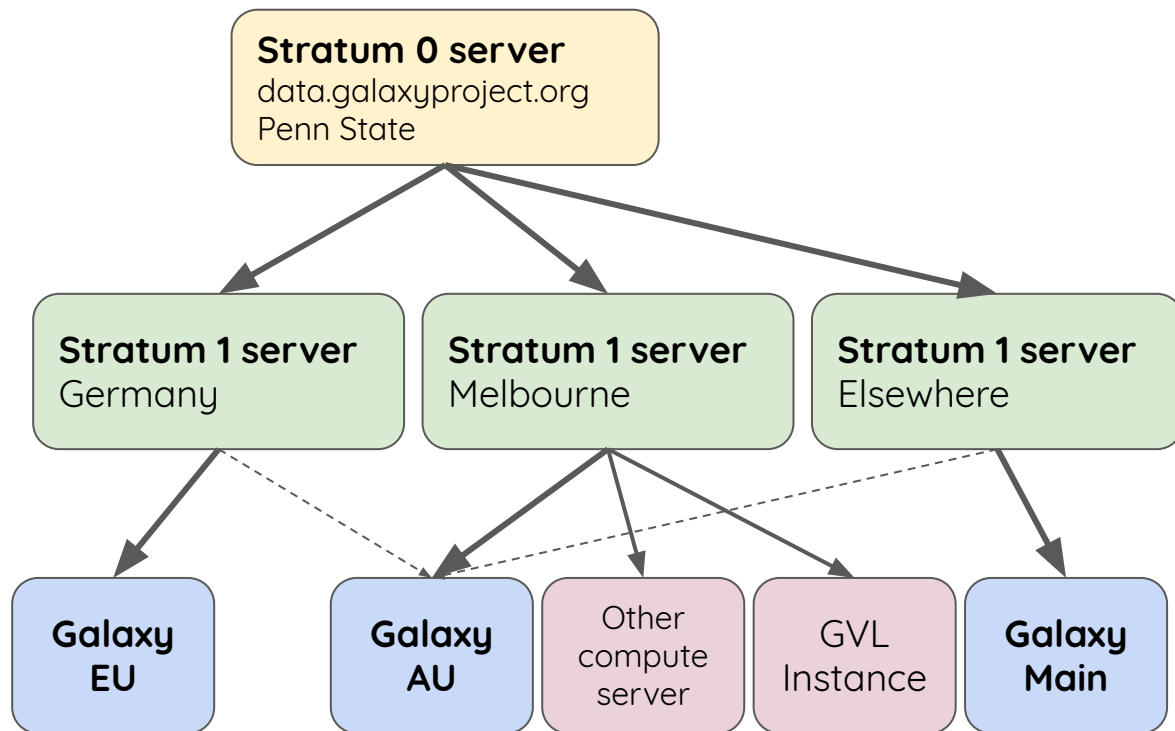


CVMFS global structure

Stratum 0: The canonical source
Transactional updates

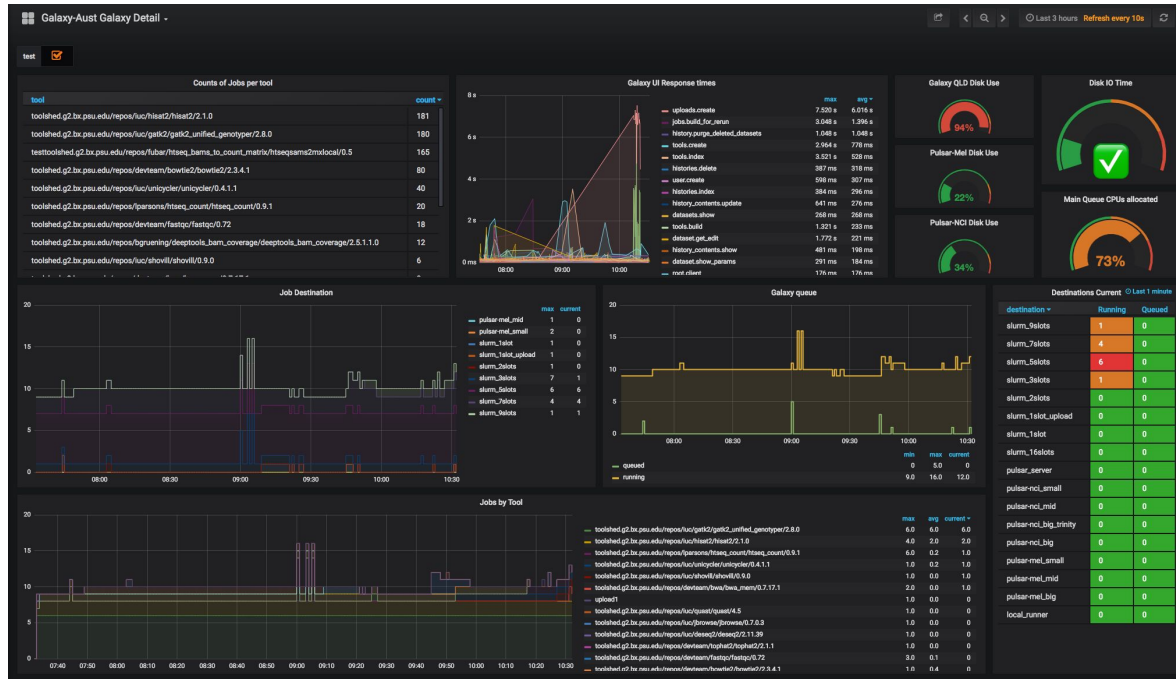
Stratum 1: Multiple servers
Mirrors Stratum 0 server
Continuous updates

User servers: Many multiple servers
Mounts repo from stratum 1
Based on GEO-API
With fallback to other stratum 1s

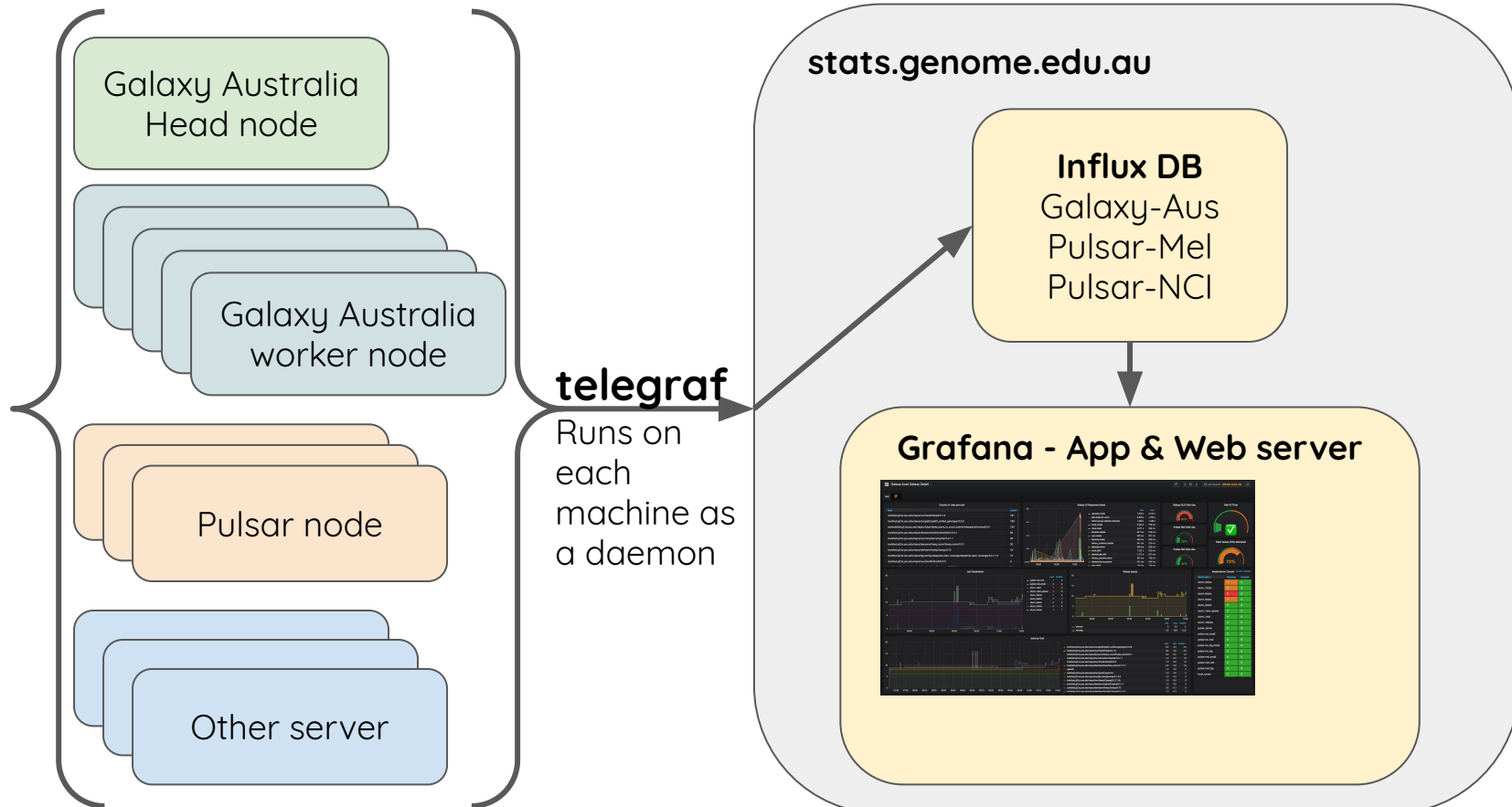


———— Primary mount - - - - Fallback mount

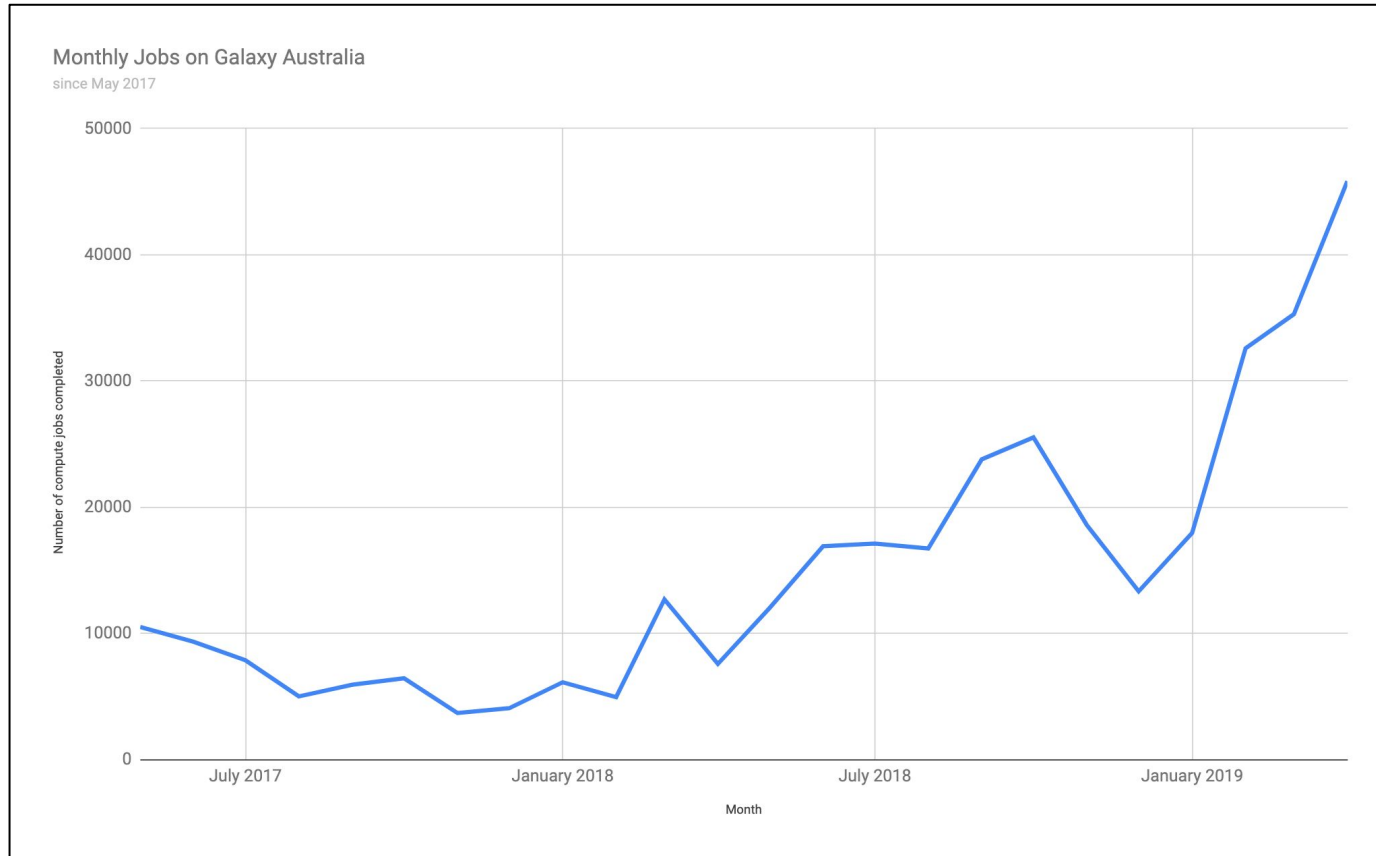
stats.genome.edu.au



Structure



All very nice tech, but is it used??



What's next?

- More compute nodes around the country
- Cloudstor?
- BYO resources
- More internationalisation of resources

Stay in Touch

- Galaxy Australia: Twitter
<https://twitter.com/galaxyaustralia>
- Galaxy Australia Community
<https://www.embl-abr.org.au/galaxyaustralia>
- And of course a final reminder:
 - Galaxy Australia <https://usegalaxy.org.au>