Galaxy AUSTRALIA

genomics virtual laboratory

Galaxy Australia

A public bioinformatics compute resource distributed nationally.

Simon Gladman (presented by Andrew Lonie)

Development partners

qcif

The University Of Queensland AUSTRALIA | RCC Research Computing Centre

Melbourne Bioinformatics
Bioinformatics • Data Services • Infrastructure, for life sciences today

Supported by

BIOPLATFORMS AUSTRALIA

EMBL Australia Bioinformatics Resource

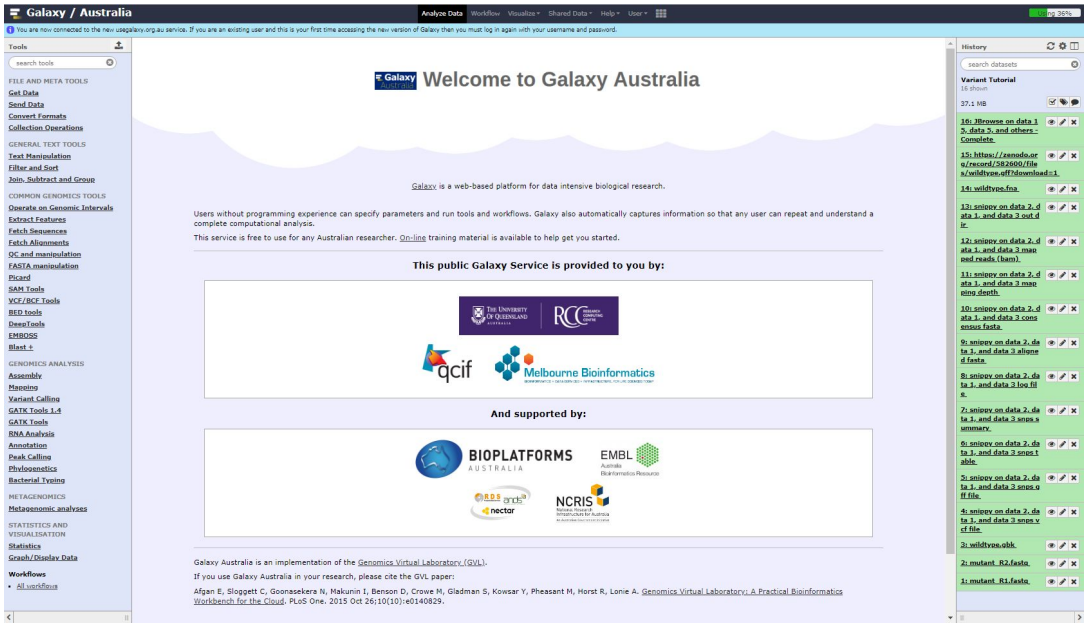
RDS Research Data Services | ands | nectar

Outline

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

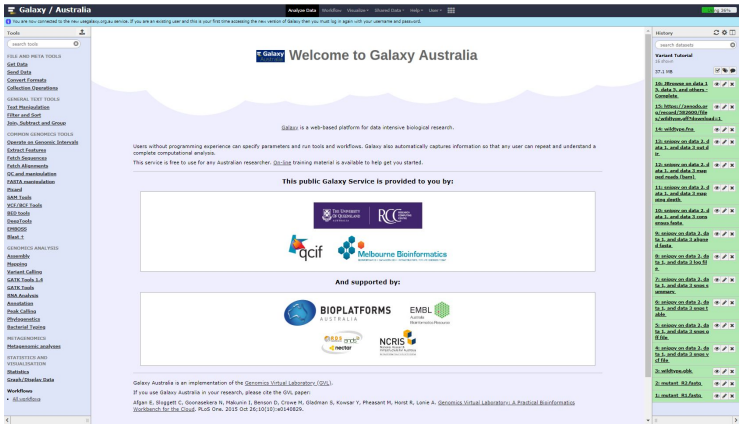
What is Galaxy?

Galaxy AUSTRALIA



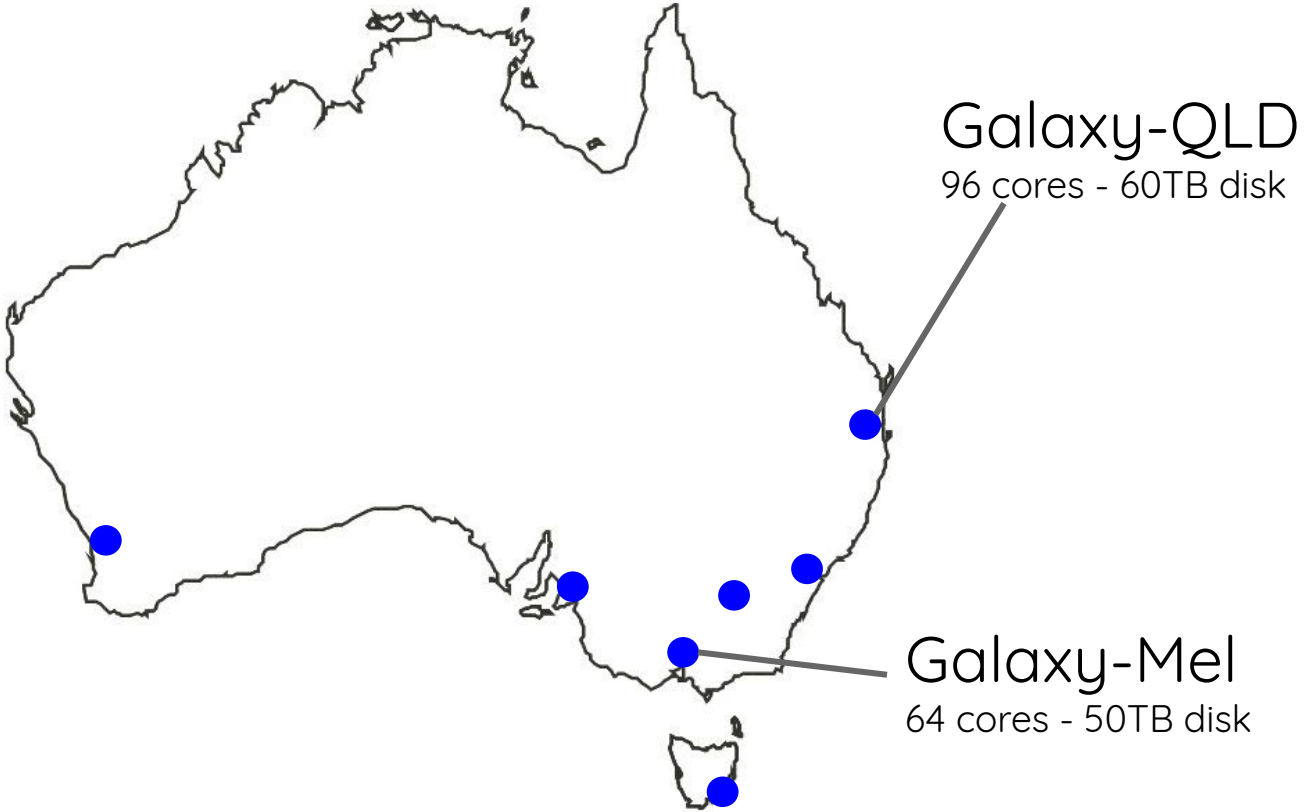
What is Galaxy?

- 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



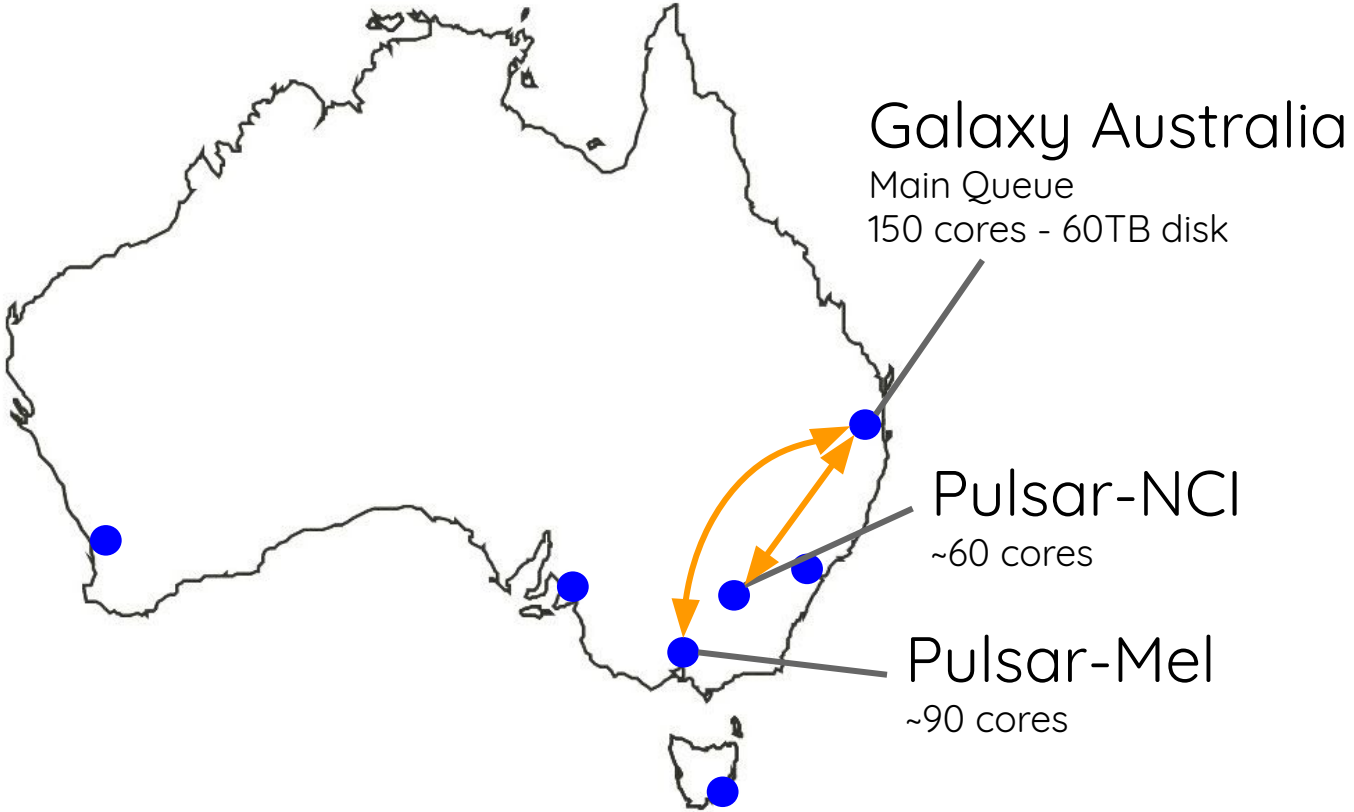
Galaxy Australia Infrastructure Locations

2015-2017

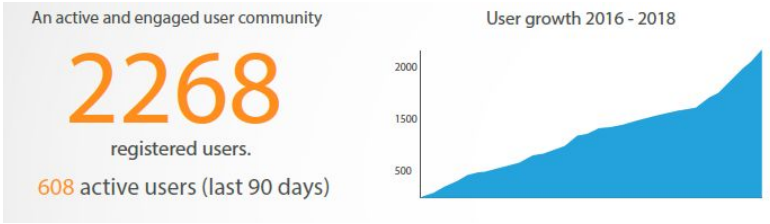


Galaxy Australia Infrastructure Locations

2018-now

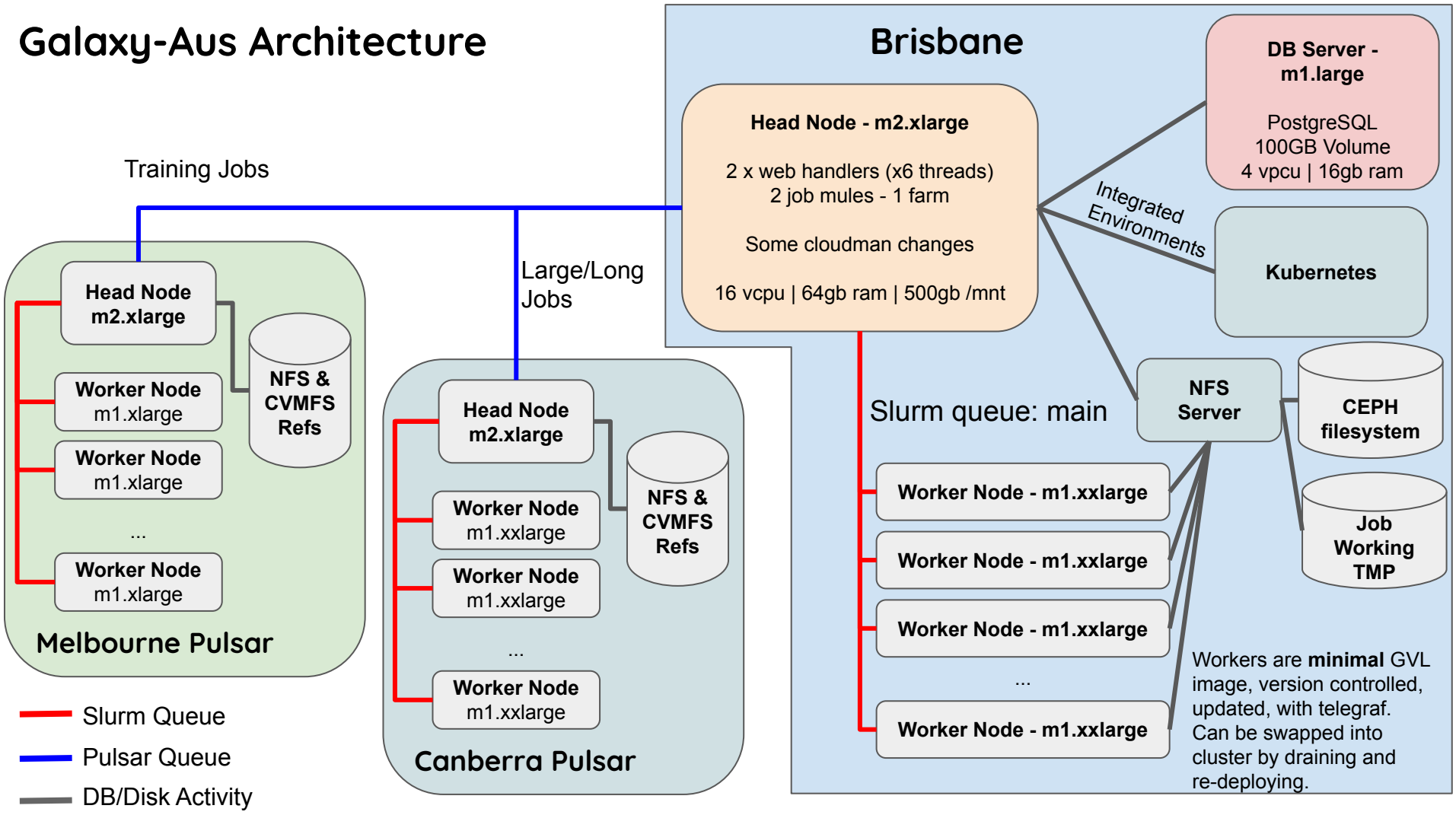


Galaxy Australia - Usage

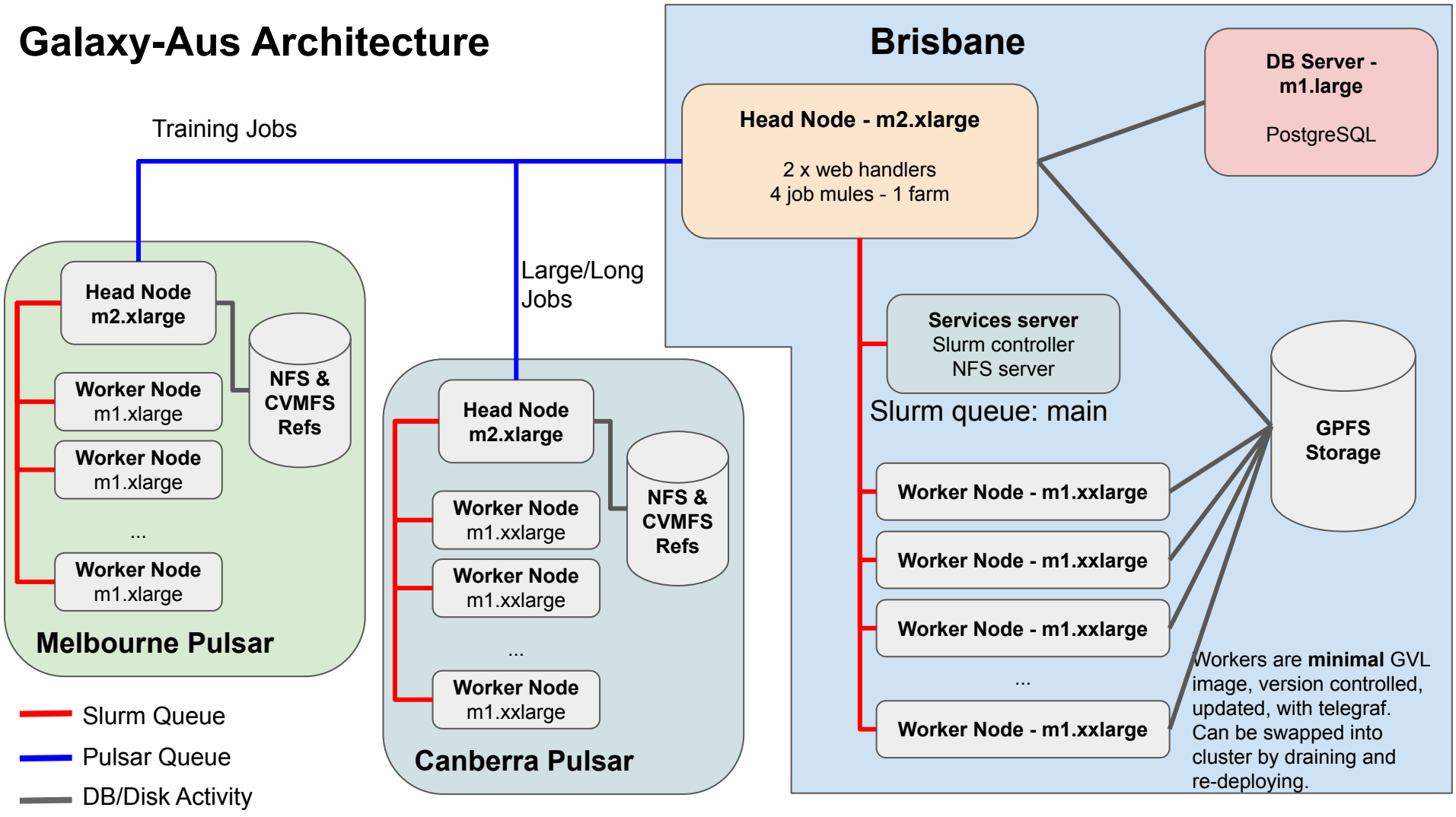


- Public and free Galaxy server in Australia
- Currently running > 50,000 jobs month.

Galaxy-Aus Architecture



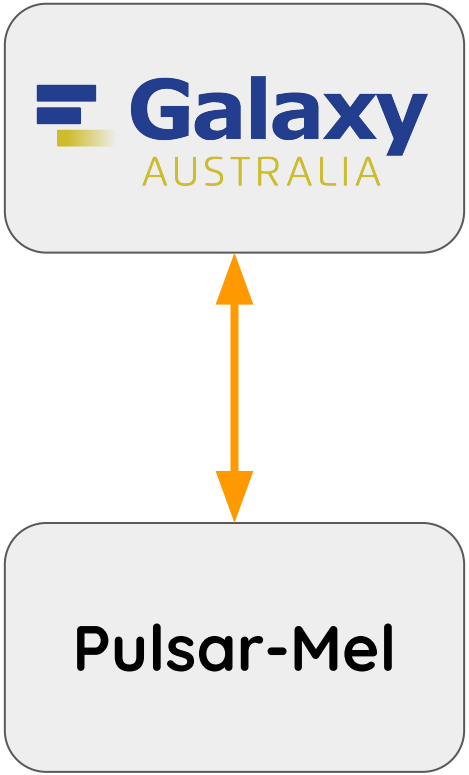
Galaxy-Aus Architecture



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

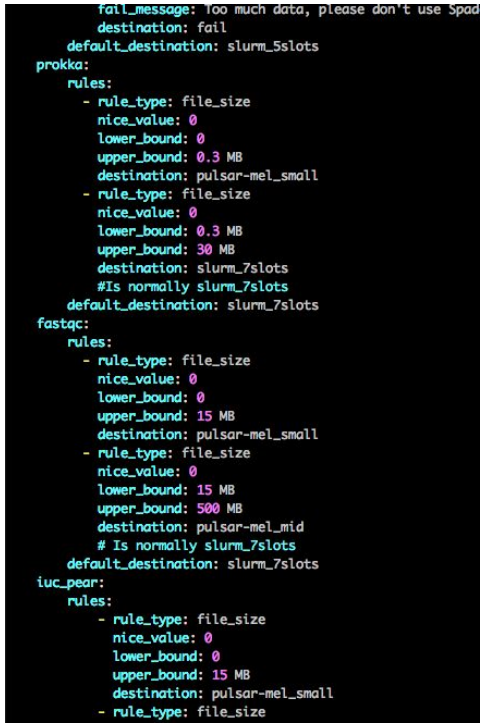
- Training jobs / Workshops
- Runs lots of small jobs
- Usually 2-4 cores
- Allows a tutorial or workshop to run without interfering with main queue

Pulsar-NCI

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

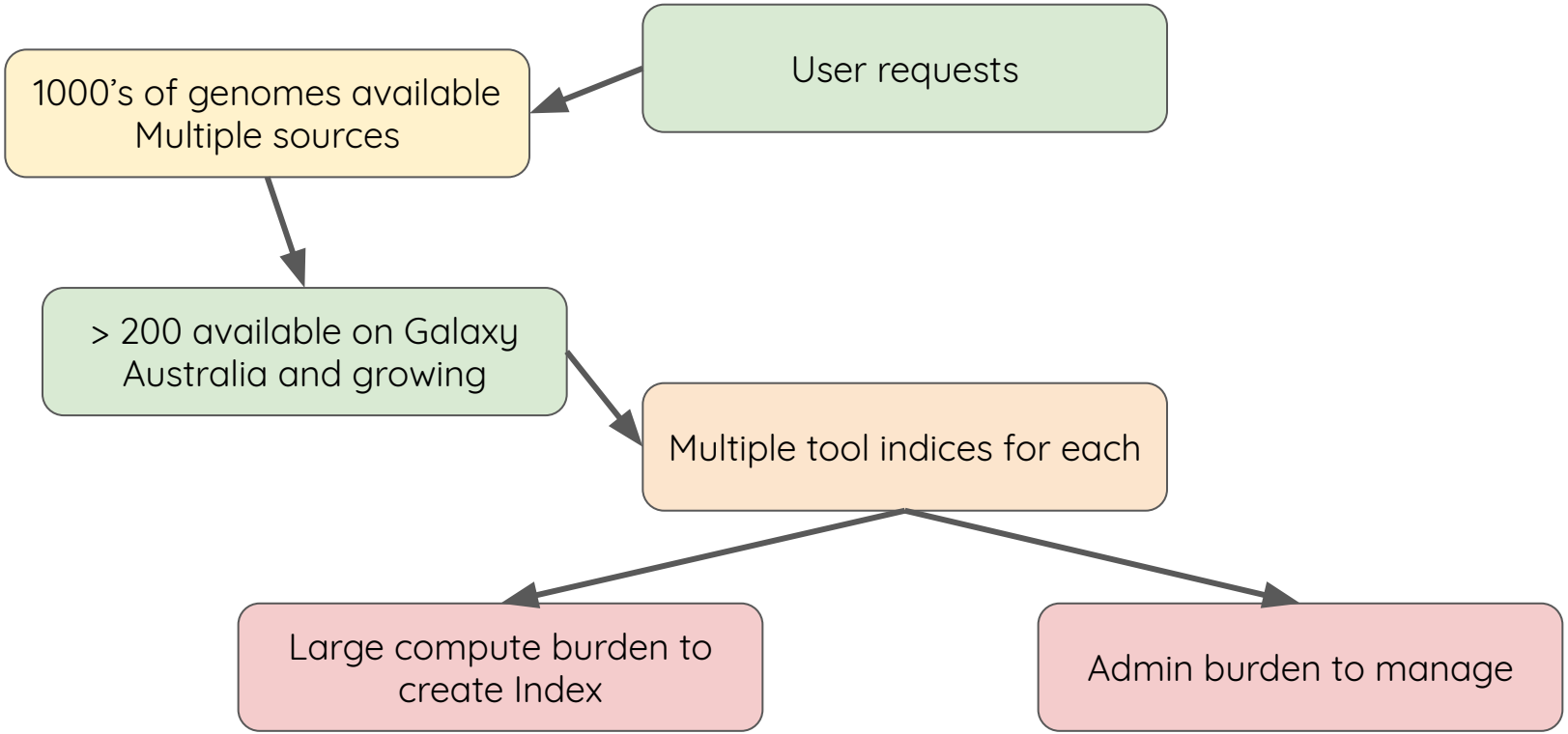
Dynamic Tool Destinations

- 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



The problem of reference data:

There's a lot and it's hard to deal with



So we started talking with other large Galaxies!

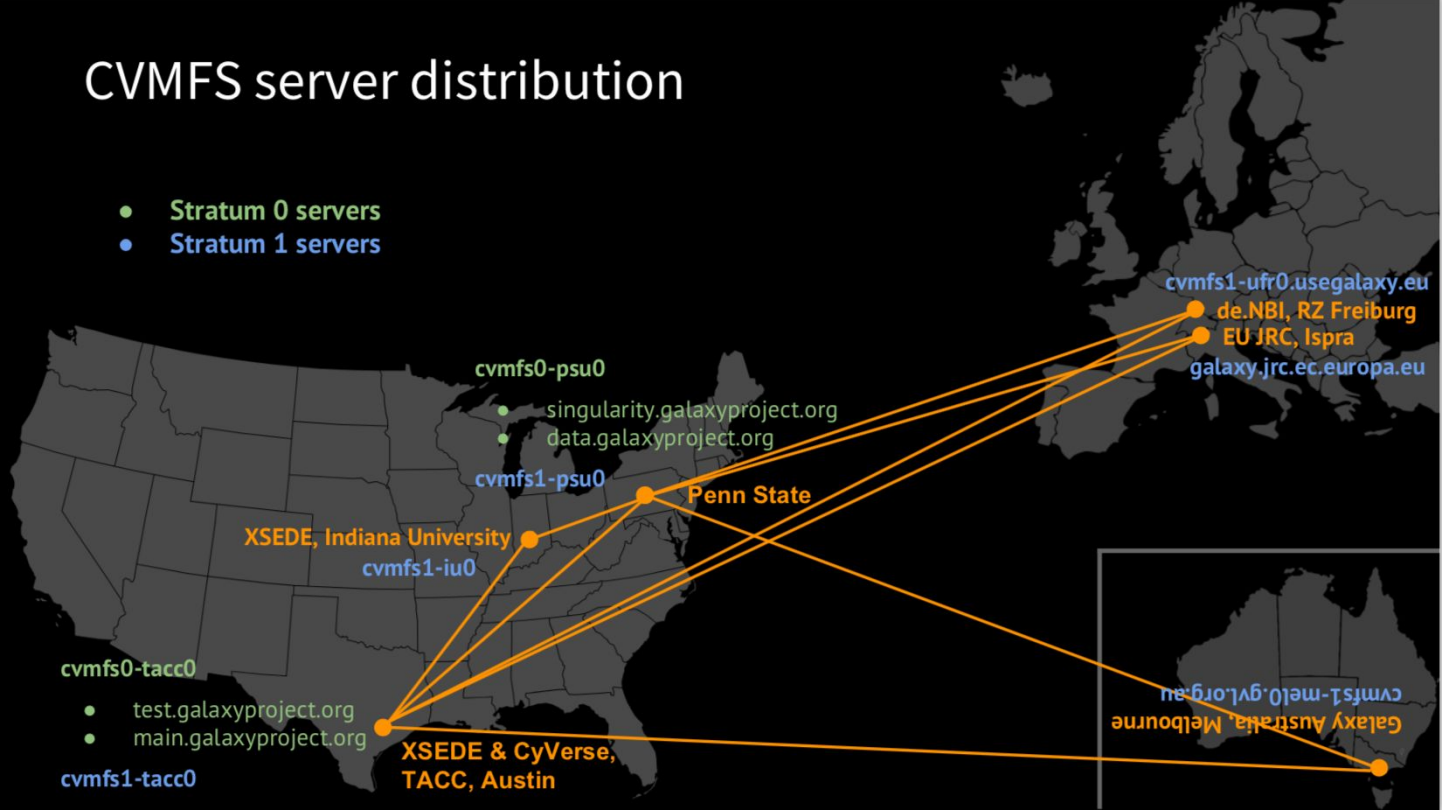
Share the burden

- 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

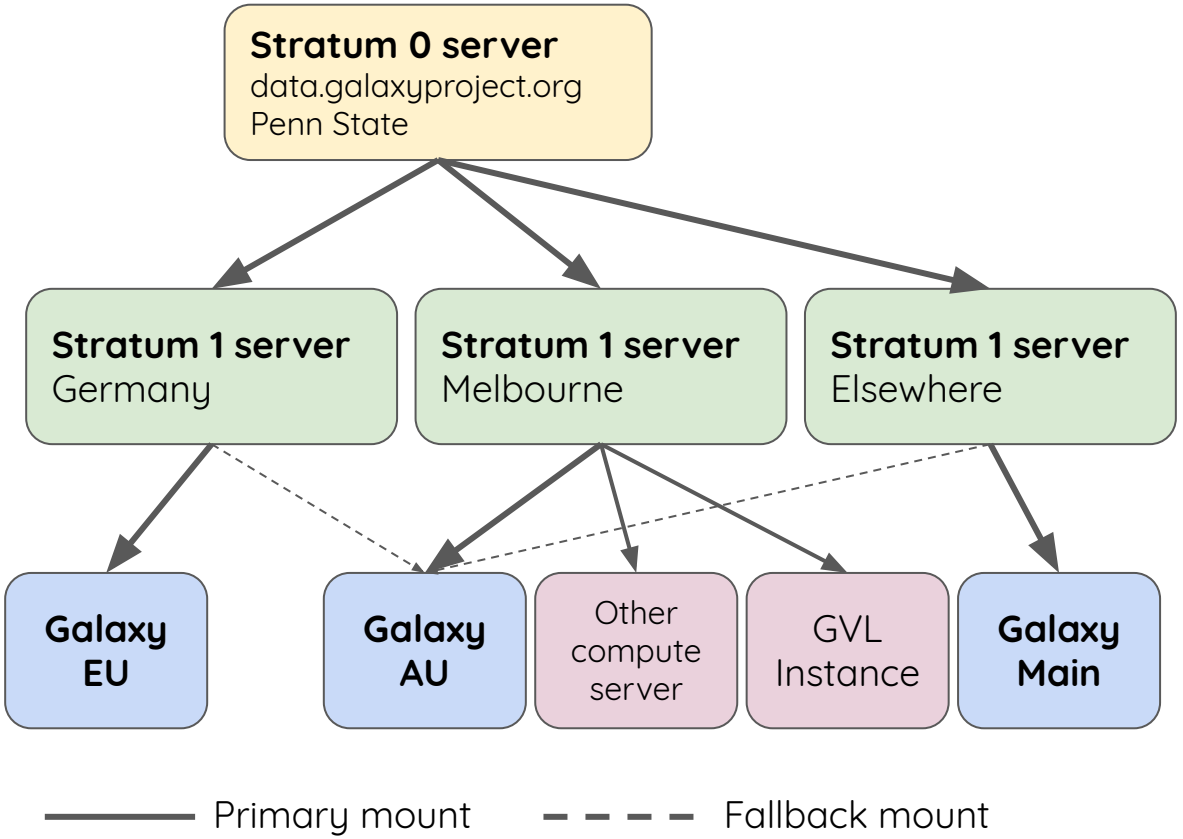


CVMFS global structure

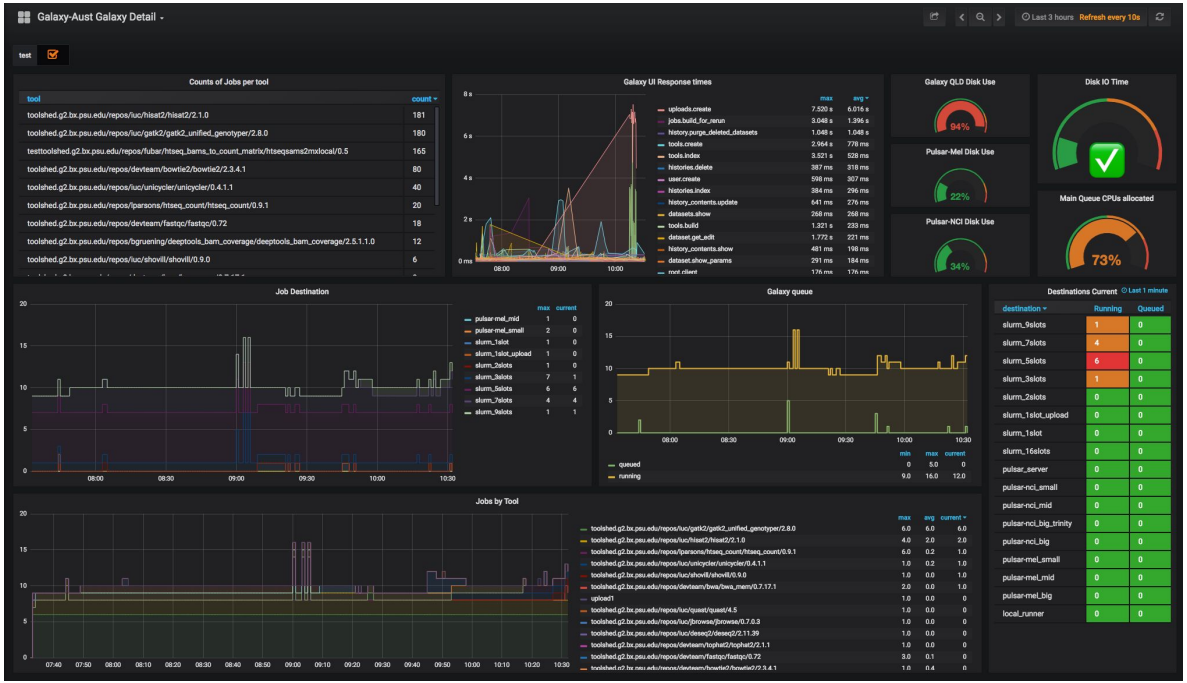
Stratum 0: The canonical source
Transactional updates

Stratum 1: Multiple servers
Mirrors Stratum 0 server
Continuous updates

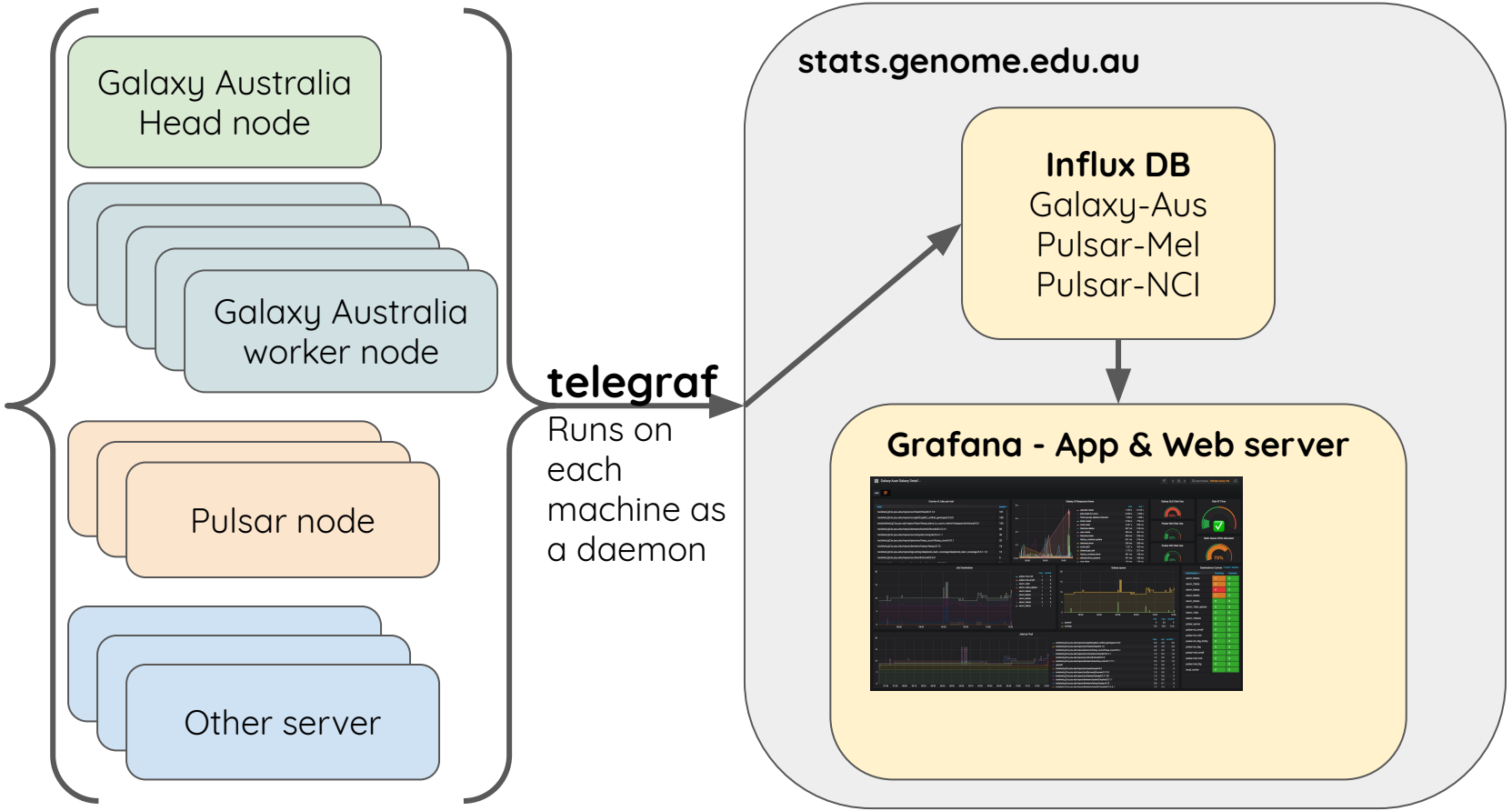
User servers: Many multiple servers
Mounts repo from stratum 1
Based on GEO-API
With fallback to other stratum 1s



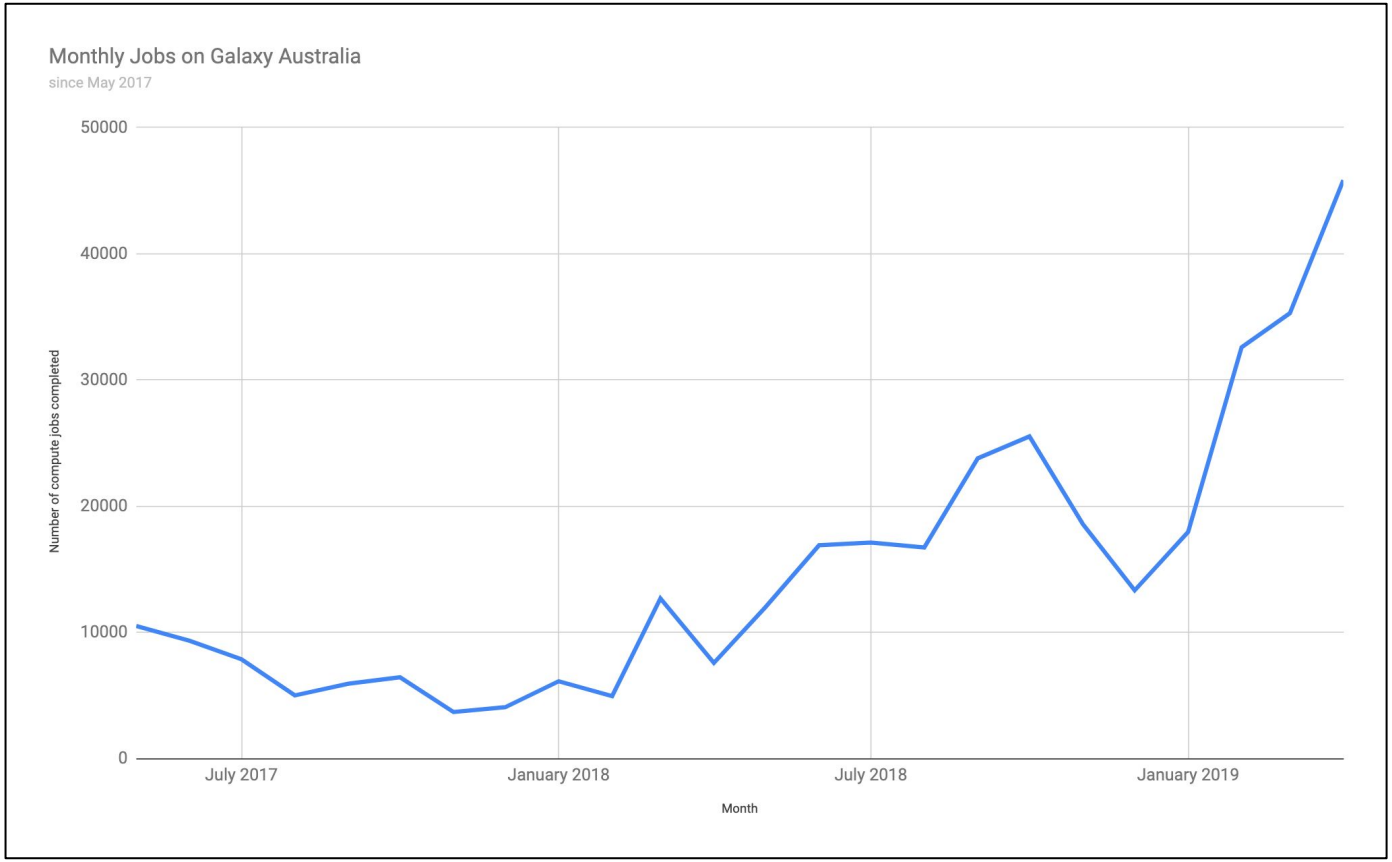
stats.genome.edu.au



Structure


Galaxy Australia
Head node
Galaxy Australia
worker node
Pulsar node
Other server
telegraf
Runs on
each
machine as
a daemon
stats.genome.edu.au
Influx DB
Galaxy-Aus
Pulsar-Mel
Pulsar-NCI
Grafana - App & Web server

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**