The CSIRO Galaxy Service
A collaboration between science and IT

Steve McMahon, CSIRO Scientific Computing
25 March 2014
Outline

• The Galaxy pilot project
• Galaxy Service status
• A better service
• DevOps
• Next steps
The Galaxy pilot project
The Galaxy pilot project

• A pilot because ...
  • We needed to learn the best way to deliver a production service
    – Didn’t know resources are required
  • The pilot required minimal extra resources – palatable to management
  • We needed to build and measure the demand of such a service to help plan for a full production

• A pilot service is a balancing act
  • between being big enough to be useful and gain support, and
  • small enough to get off the ground with minimal resources
Pilot infrastructure specifications

- Galaxy-compute: 32 cores, 192GB RAM
- Galaxy-db: 16 cores, 64GB RAM
- Galaxy-web: VM frontend – can scale out
- Galaxy-dev: VM frontend for development instance
- SAN storage: 4.5TB, used 1.9TB so far
- Remote job submission coming to bragg – a Top500 supercomputer
Pilot Project Outcomes

• A successful collaboration between science and IT
• Endorsement to plan and deliver a better, production, service
A better CSIRO Galaxy Service

Lessons learnt

• Need adequate infrastructure – storage and compute but mostly storage.
  • Decided on 100TB shared storage
  • Option to “bring your own”
• Remote submission to HPC too hard
  • Decided to collocate Galaxy with a major HPC installation
• Need to streamline way Galaxy is updated
• Need more development instances – one per developer ideally
• Collaboration between science and IT works really well
A better CSIRO Galaxy Service

Requirements

• Leverages existing HPC facilities – both hardware and software infrastructure
• Streamlined processes for ongoing development and implementation of new tools
• Ability to provide a service to bioinformatics as well as other science domains
• Integrates with data provided by the reference data project
New Galaxy Service infrastructure

CSIRO internal network

Production front end VMs
Test VMs
Development VMs

Database server
Shared Galaxy storage
Reference data
Bring your own data
HPC apps directory

HPC facility
Streamlining the process with DevOps

• DevOps is a software development movement that stresses a close relationship between software developers and system administrators
• The goal is to enhance and speed up the cycle of software production from creation to the delivery to users, with a special focus on quick resolution of user issues.
How to do DevOps

• Collaboration
• Infrastructure as code
  • Creating a new machine image from a script
• Continuous delivery
  • Build software in such a way that it can be released to production at any time
Steps towards DevOps

• Collaboration
• Infrastructure as code
  • Creating a new machine image from a script
• Continuous delivery
  • Build software in such a way that it can be released to production at any time
Infrastructure as code

• In touch with Olivier Inizan and Mikael Loaec from INRA, France
  • Developing a “Puppet” module for Galaxy
  • Also collaborating with Eric Rasche of Texas A&M University
• Considered GVL deployment scripts but much work to get them to work with our operating SLES
• HPC systems use SLES
• CSIRO IM&T already have Puppet expertise and infrastructure
Continuous delivery

• To be investigated ...
Next steps

- Continue to support existing Galaxy Service till it can be replaced
- Project proposal approved so compute and storage infrastructure can be requested
- Software development of “infrastructure as code” to be started