Building a Prototype Data Analysis as a Service: the STFC experience

NOBUGS 2016 - Copenhagen

Frazer Barnsley

October 2016
Outline

• Motivations
• Architecture
• Technology overview
• Prototype
• Future developments
Motivations
The facilities

- ISIS, Diamond and CLF
- > 6000 scientists a year
- Users come to perform experiments and leave with their data
- Desktop analysis and simulation becoming more difficult
The target

• Provide users access to data and compute resources during and after experiment

• Provide users with a suitable software environment

• Provide appropriate compute resources with seamless access to data and software

• Provide suitable interfaces

• Expandable for different communities
Architecture
Technology Overview
Access: Remote Desktop

• NoVNC
  – Easy browser access

• SSVNC (secure VNC)
  – Fast access, requires user to install software

• RDP
  – Fast access, built into Windows
Access: Data

- Browser
  - Easy access
- WebDAV
  - Easy mount
- Globus / GridFTP
  - TB transfers
Resources: Cloud

- OpenNebula
  - 896 processing cores
  - 3.5TB memory

- CEPH
  - 750TB storage
  - Self healing capabilities
  - Decentralised design

- Managed by RAL Tier 1
Resources: SCARF

- General purpose batch system based on IBM Platform LSF
- 5808 cores
- 500 registered users
- >70 applications
- 300TB high speed storage (PaNaSaS)
Software: CernVM-FS

- Union file system
- Multi-tiered caching
- Clients installed on SCARF and the cloud
- Service managed by RAL Tier 1
Software: Configuration Mgmt

- Quattor
- Reuse of configuration across multiple systems
- Service managed by RAL Tier 1
Intelligence: ICAT

- Metadata catalogue
- Rule based permissions
- Links users to experimental data, instruments, parameters etc ...

- ISIS
  - 48TB data
  - >13 million data files

- Also used by DLS, CLF, ESRF, SNS ...
Prototype
The setup

• Web frontend to launch cloud VMs

• Graphical access to VMs via remote desktop

• VMs configured at contextualisation stage

• Software loaded from CernVM-FS via desktop icons

• Processing can be done locally on VMs or sent to HPC systems

• Persistent user data storage
The software - CCP4

“integrated suite of programs that allows researchers to determine macromolecular structures by X-ray crystallography, and other biophysical techniques” - www.ccp4.ac.uk

- Used on data non-exclusively from the MX beamlines at Diamond
- Project co-ordinated by STFC
User selects their environment

Create an analysis environment then launch a remote desktop session.

- CCP4 Analysis Machine
  - LAUNCH
  - DELETE
- CCP4 Analysis Machine

[CCP4] CREATE
Then select their preferred method for remote access.
Users can launch a full screen remote desktop session from their browser.
VMs also bring forward specific groups of software
Local processing on the VM
or launch jobs to HPC systems
Persistent user storage is accessible from each VM ... and can be mounted on the user's local machine
Future Developments

• Data transfer and management

• More integration with the ICAT family

• More engagement with user communities
frazer.barnsley@stfc.ac.uk