Science & Technology Facilities Council

www.stfc.ac.uk

Maintaining Traceability in an Evolving Distributed Computing Environment

Ian Collier, STFC, Romain Wartel, CERN

ian.collier@stfc.ac.uk

Introduction

- Security incidents are an operational reality on our grid based distributed computing infrastructure.
- We are used to managing the risks inherent in running distributed computing infrastructures.
- *When* things go wrong we need, at a minimum, to know:
  - *WHO* did *WHAT*, *WHEN* they did it, and *WHERE* they did it.
- This allows us to contain the impact of incidents, preserve reputations and ensure that resources are available for their intended purposes.
- As our infrastructure evolves to include cloud resources we must ensure we maintain the traceability we depend upon.

Current traceability model

- We have well developed incident response procedures.
- Assumes sites have sight, and control, of the execution environments.
- Sites log, in detail, from the execution environment (worker nodes) and from CEs, batch systems etc. to central loggers.
  - More or less sophisticated tools (often variants of grep) to search
- Granular authorisation & traceability in multi user pilot jobs with glexec (although this is not implemented universally).
- We have developed and agreed incident response procedures
  - with clearly identify contact points
  - and established trust relationships
- Facilitates both the analysis of and response to problematic activities.

Emerging clouds

•Private, public and federated cloud resources bring changes to many aspects of distributed computing.

•Changes to workflows:

- Sometimes removing complexity for users (that is the aim).
- Changing things for providers - some things are easier.

•Clouds also introduce new software components (cloud management frameworks of varying complexity) and (complex) new frameworks for federating cloud resources are being actively developed.

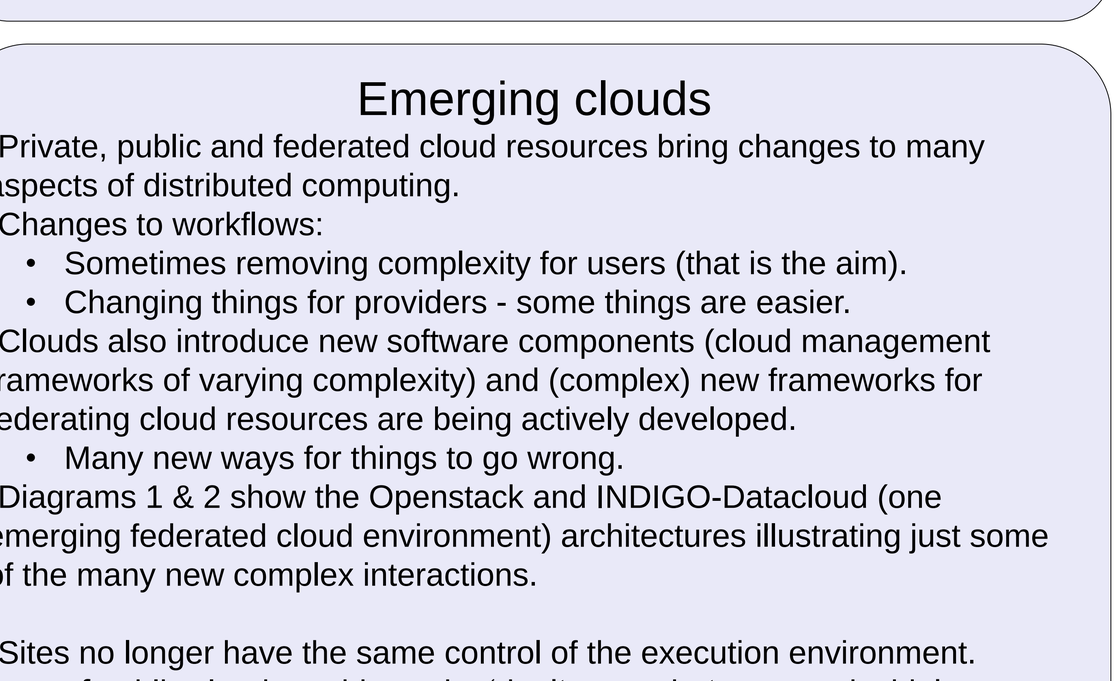
- Many new ways for things to go wrong.

•Diagrams 1 & 2 show the Openstack and INDIGO-Datacloud (one emerging federated cloud environment) architectures illustrating just some of the many new complex interactions.

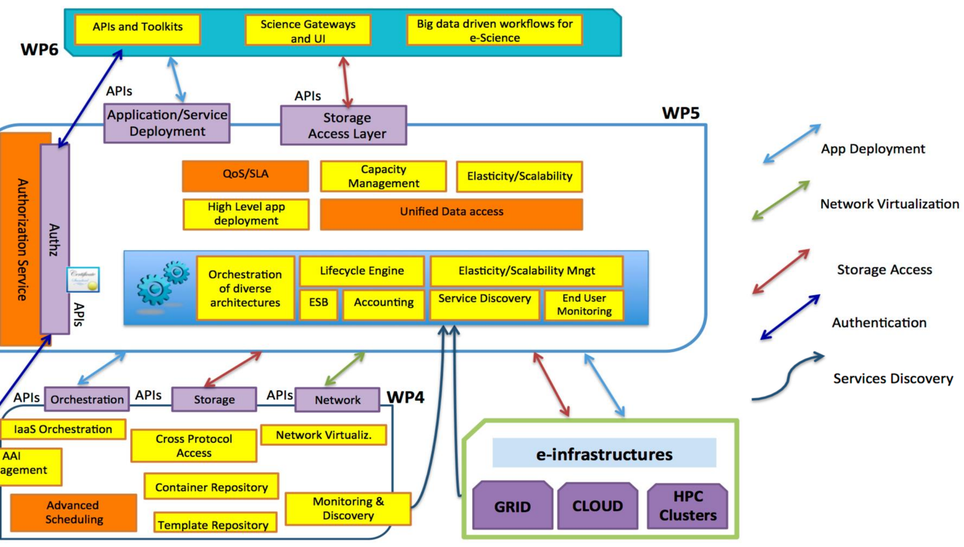
•Sites no longer have the same control of the execution environment.

- cf public cloud providers who ‘don’t care what goes on inside’.

•VMs launched by VOs – or by their workload management systems.



1. OpenStack Folsom architecture diagram



2.INDIGO-DataCloud architecture diagram

Solutions - Logging

- Increase focus on externally observable behaviour
  - Hypervisor & Cloud management framework
  - Federated resource frameworks
  - Network activity & flows (neglected until now)
- Also connect VMs to central loggers at sites - requires standardised hooks in VM images
- Aggregation of and cross checking between multiple sources is vital
- Improved tools for storing, aggregating & searching increasingly important

Solutions - Quarantining Images

•Virtualisation allows us to capture VM images for forensic examination – a big advantage.

- What if a hypothetical attacker deliberately uses short lived VMs?

•Need images to be retained for a, tunable, period after shutdown

- Some cloud platforms already do this
  - Implemention required for others - OpenNebula & OpenStack using Ceph are common combinations which need this.

Solutions - VOs as partners in Incident Response

- For some grid jobs we already need to go to VOs to find out what user ran specific jobs. (So that we can suspend just that user.)
- Within WLCG VOs already log workflows to support debugging & workload management.
  - Could changes to VO logging better support traceability?
- Rather than attempt an up front gap analysis, traceability service challenges can be used to identify limits of current logging and suggest enhancements.
  - payloads and challenge methodologies are currently in development
- This is an opportunity to formally recognise the existing reality that we need the active participation of VOs in order to maintain traceability.

Conclusions & Future work

- WLCG Cloud Traceability Working Group established to carry out practical development of possible solutions.
- Use the results from all these areas of investigation to develop:
  - Updated policies setting out requirements for running these new forms of distributed computing infrastructures without compromising traceability - perhaps even improving it.
  - Best practise recommendations for how to gather additional logging information and how to configure management frameworks and VM images.
- While this work is focussed in the already well developed WLCG collaboration the policy and best practise we produce can provide a model for emerging cloud & virtualisation based distributed computing infrastructures.
- Open question how to develop trust frameworks that will allow sites to accept VOs as full partners in incident response. Perhaps not unlike work that allowed workloads to be distributed across grids.

CERN

WLCG Worldwide LHC Computing Grid

GridPP UK Computing for Particle Physics