Science & Technology Facilities Council

Cloud services at RAL, an Update

26th March 2015

Spring HEPiX, Oxford

George Ryall,

Frazer Barnsley, Ian Collier, Alex Dibbo, Andrew Lahiff

V2.1

Overview

- Background
- Our current set up
- Self-service test & development VMs
- Traceability, security, and logging
- Links with Quattor (Aquilon)
- Expansion of the batch farm into unused capacity
- What next

Background

- Began as small experiment 3 years ago
  - Initially using StratusLab & old worker nodes
  - Initially very quick and easy to get working
  - But fragile, and upgrades and customisations always harder
- Work until last spring was implemented by graduates on 6 month rotations
  - Disruptive & variable progress
- Worked well enough to prove usefulness
- Self service VMs proved very popular, though something of an exercise in managing expectations

Background (2)

- In March 2014, we secured funding for our current hardware, I became involved –setting up the Ceph cluster
- A fresh technology evaluation led us to move to OpenNebula.
- In September secured first dedicated effort to build on the previous 2 ½ years of experiments to produce a service with a defined service level. Two staff, me full time and another half time.

Where we are now

- Just launched with a defined, if limited, service level for users across STFC.
- Integrated in to the existing Tier 1 configuration & monitoring frameworks (yet to establish cloud specific monitoring).
- Now have an extra member of staff working on the project, bringing us close to two full time equivalents.

Our current set up

- OpenNebula based cloud with a Ceph storage backend
- This has 28 Hypervisors consisting of 892 cores and 3.4TB of RAM
- We have 750TB of raw storage in the supporting Ceph cluster (as seen in Alastair Dewhurst's presentation yesterday, performance testing described in Alex Dibbo's presentation this afternoon)
- This is all connected together at 10Gb/s
- Web Front End and headnode on VMs in our HyperV production virtulisation infrastructure
- Three node MariaDB/Galera cluster for DB (again on HyperV)

Science & Technology Facilities Council

Self-service test & development VMs

- First use case to be exposed to users in a pre-production way.
- Provide members of the Scientific Computing Department (~160 people) with access to VMs on demand for development work.
- Quickly provides VMs to speed up the development cycle of various services and offer a testing environment for our software developers (<1 minute to a useable machine) .
- Clear terms of service and defined level of service that is currently short of production..

Science & Technology Facilities Council

A simple web frontend

- To provide easy access to these VMs we have developed a simple web front-end running on a VM
- This talks to the OpenNebula head node through its XML RPC interface
- Provides a simpler, more customised interface for our users than is available through OpenNebula's sunstone interface

The web front end from a users perspective

https://cloud.stfc.ac.uk

SCD CLOUD | Machines | Login

Home

Terms of Service

Infrastructure

The cloud is based on OpenNebula for the virtualisation and CEPH for the storage. For the hardware, we use

- 28x Dell R420 for the hypervisors
- 30x Dell R520 for the storage nodes

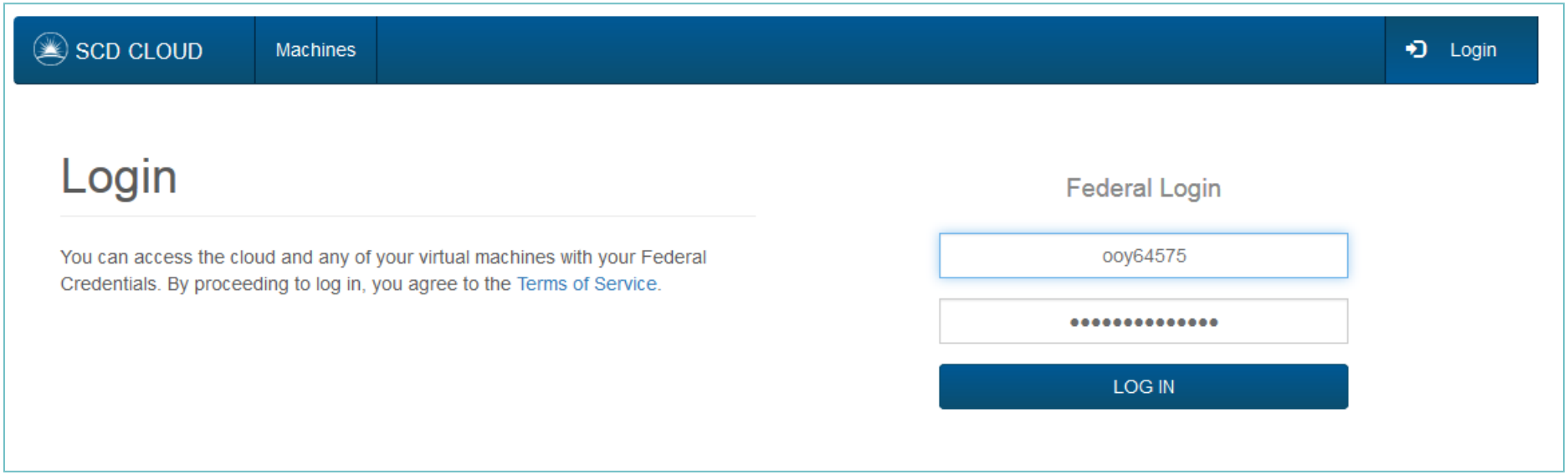
This gives us a total of **3.5TB** of memory, **896** processing cores and a storage capacity of approximately **750TB**.

OpenNebula

ceph

Science & Technology Facilities Council

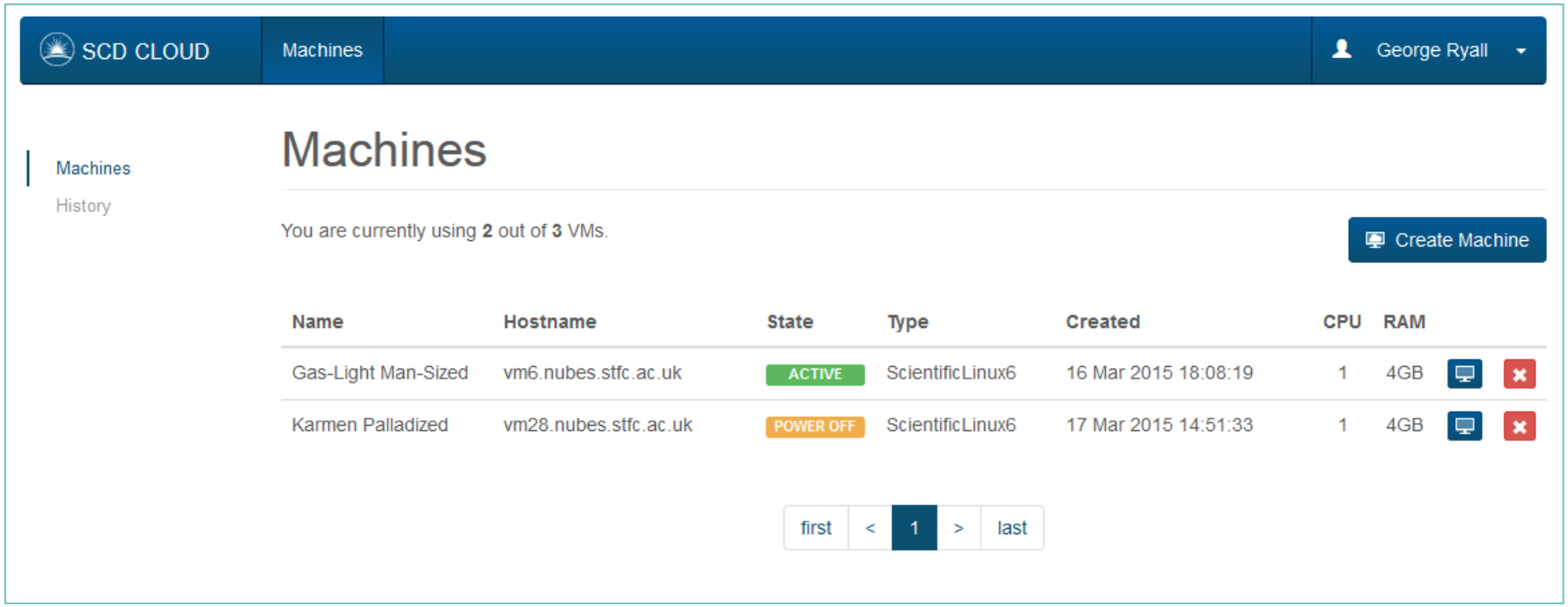
The web front end from a users perspective



User logs in with their organisation wide credentials (implemented using Kerberos)

Science & Technology Facilities Council

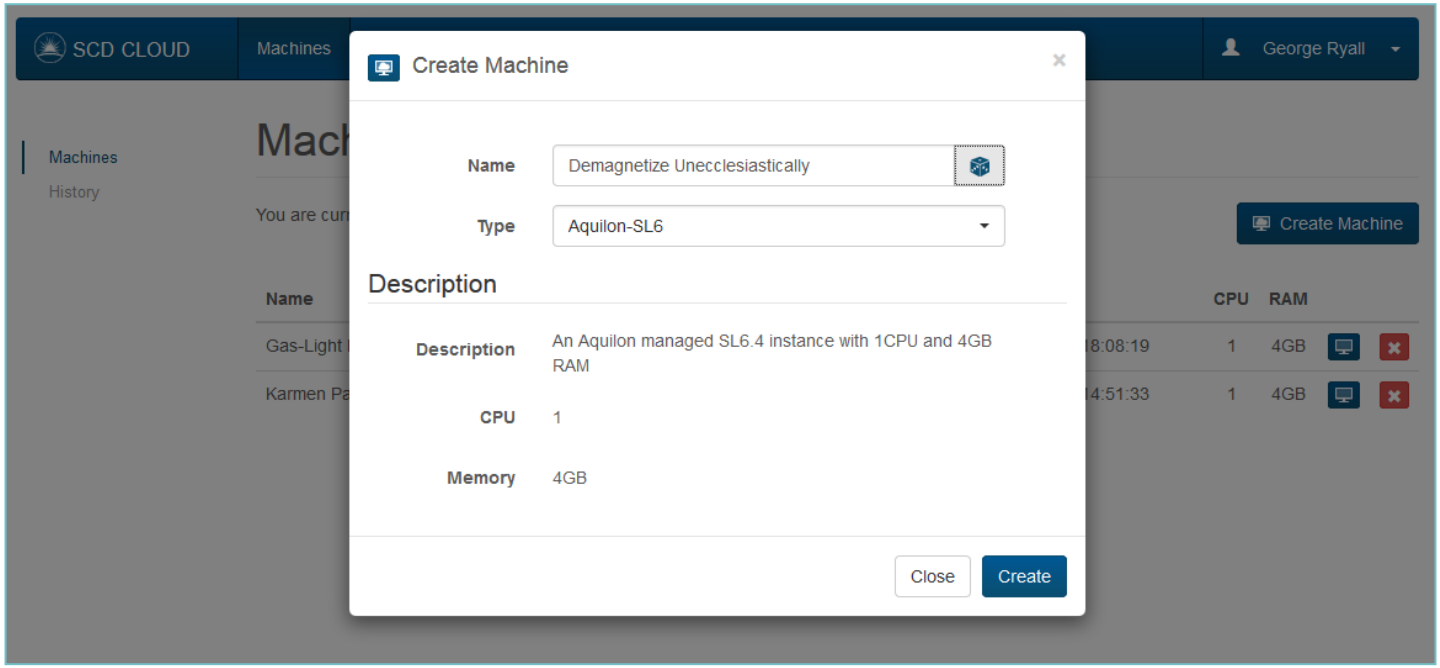
The web front end from a users perspective



The User is presented with a list of their current VMs, a button to launch more, and an option to view historical information

Science & Technology Facilities Council

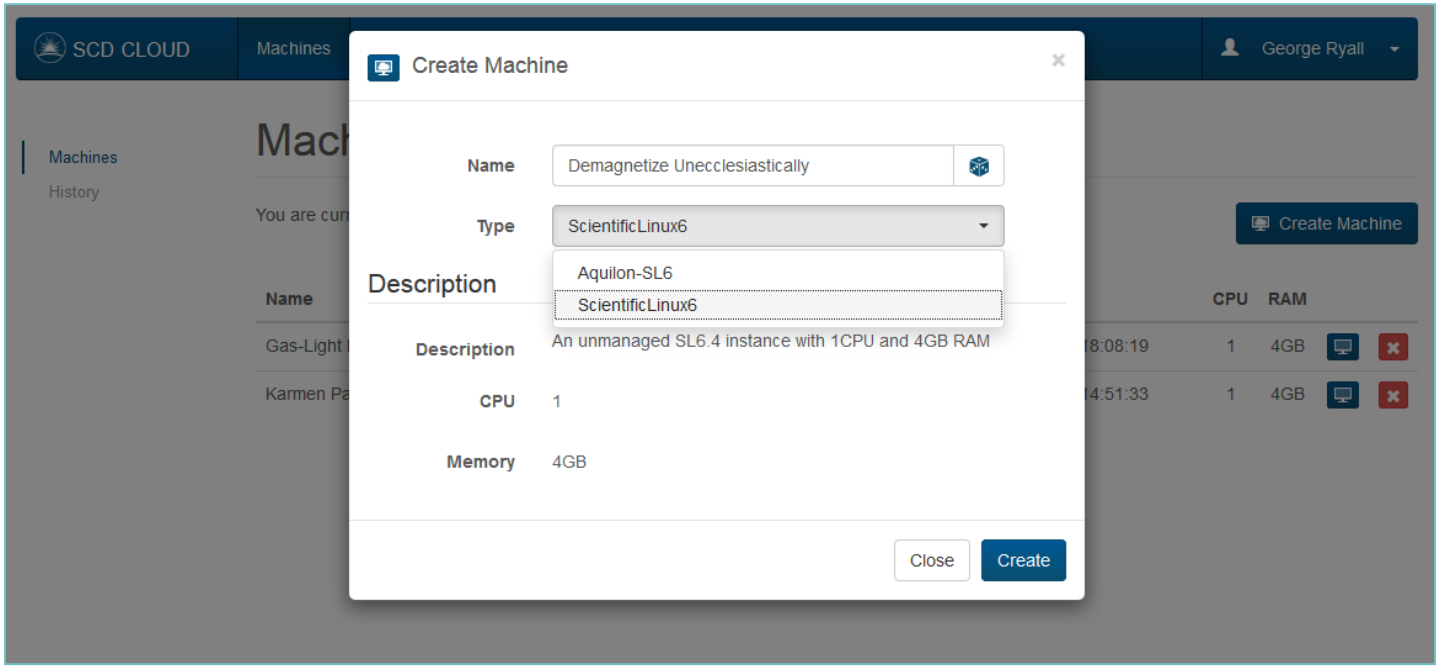
The web front end from a users perspective



The User clicks to “Create Machine”
(because they’re lazy they use our auto-generate name button)

Science & Technology Facilities Council

The web front end from a users perspective



The user is presented with a list of possible machine types to launch which is relevant to them
This is accomplished using OpenNebula groups and active directory user properties.

CPU and Memory are currently pre-set for each type, we can expand it later by request. We could offer a choice – but we suspect users, being users, will just select the most available with little thought.

Science & Technology Facilities Council

The web front end from a users perspective

SCD CLOUD | Machines | George Ryall

Machines
History

Machines

You are currently using **3** out of **3** VMs.

Create Machine

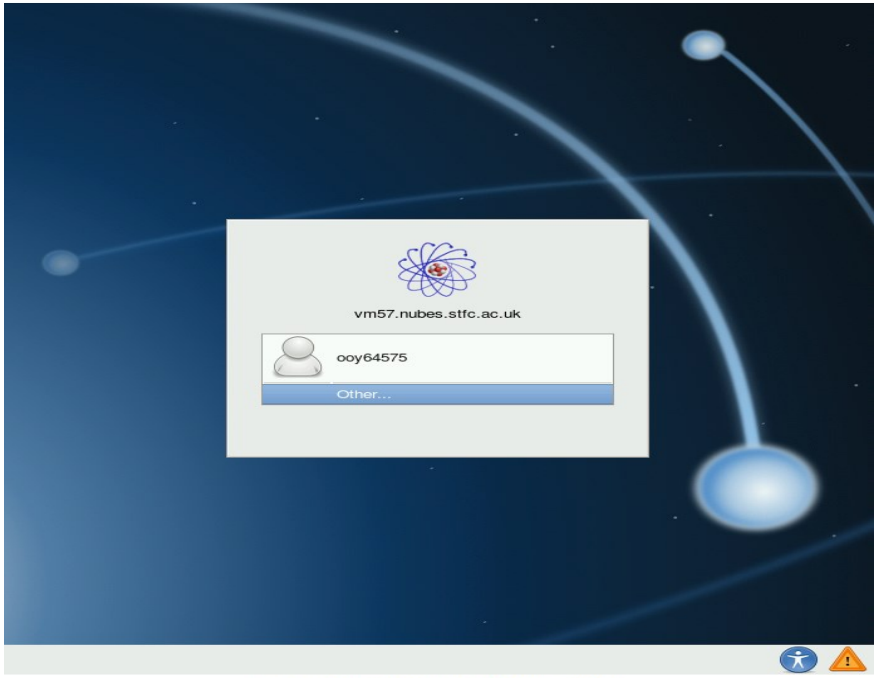
Name	Hostname	State	Type	Created	CPU	RAM
Gas-Light Man-Sized	vm6.nubes.stfc.ac.uk	ACTIVE	ScientificLinux6	16 Mar 2015 18:08:19	1	4GB
Karmen Palladized	vm28.nubes.stfc.ac.uk	POWER OFF	ScientificLinux6	17 Mar 2015 14:51:33	1	4GB
Demagnetize Unecclesiastically	vm57.nubes.stfc.ac.uk	PENDING	ScientificLinux6	25 Mar 2015 15:38:51	1	4GB

first < 1 > last

The VM is listed as pending for about 20 seconds, whilst OpenNebula deploys it on a hypervisor

Science & Technology Facilities Council

The web front end from a users perspective

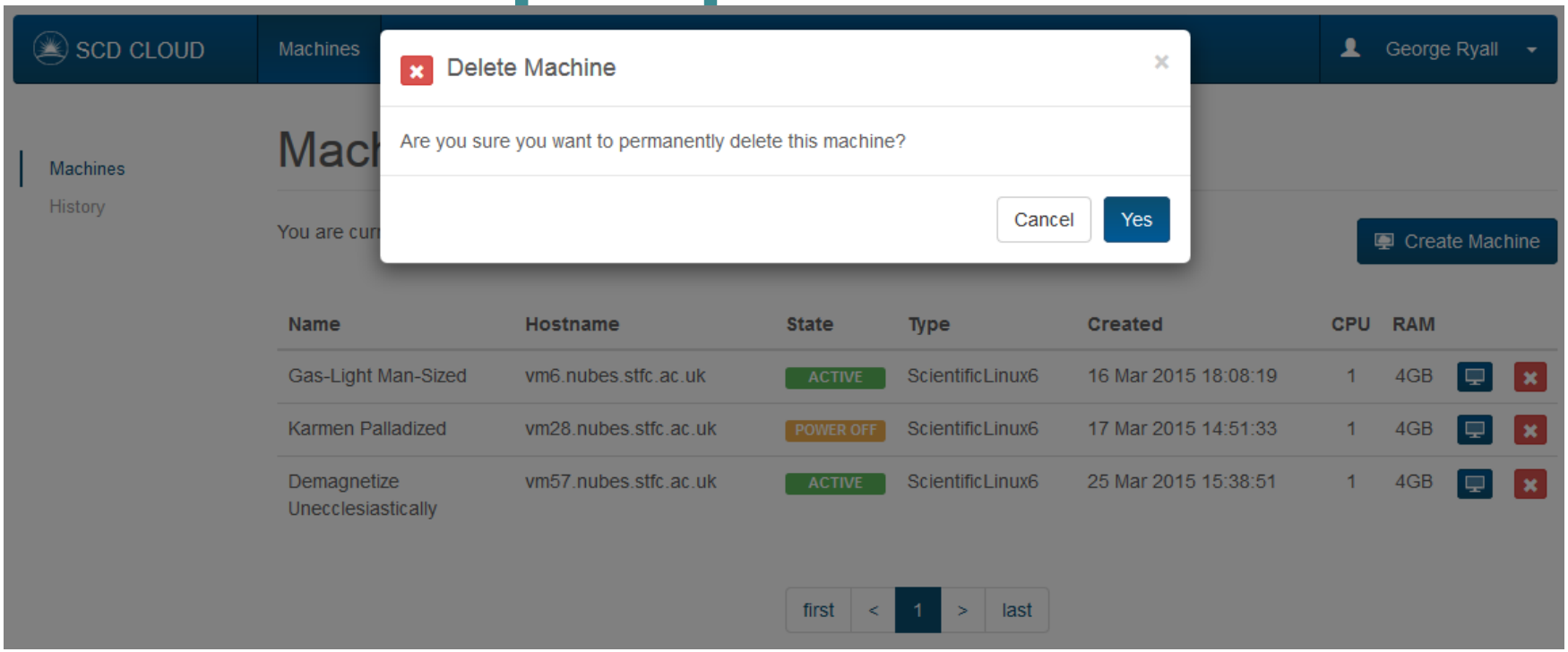


Connected (encrypted) to: QEMU (one-3587)

Once booted, the user can login with their credentials or they can SSH in with those same credentials

Science & Technology Facilities Council

The web front end from a users perspective



Once the user is done they click the delete button and from their perspective it goes away…

Science & Technology Facilities Council

Traceability

- …Actually for traceability reasons (as seen in Ian Collier's Tuesday afternoon presentation) we keep snapshots of the images for a short period of time.
- This allows us to allow us to investigate potential user abuse of short-lived VMs as well as being useful for debugging other issues.
- At VM instantiation, an OpenNebula hook creates a deferred snapshot to be executed when the machine is SHUTDOWN.
- A cron job runs daily to check all images are the right type and the age and deletes the relevant
- images.

Science & Technology Facilities Council

Security Patching

- Just like any other machine in our infrastructure, VMs need to have the latest security updates applied in a timely manner.
- For Aquilon managed machines, this will be done with the rest of our infrastructure.
- The unmanaged images come with Yum auto-update and local Pakiti reporting turned on .
- Our user policy expressly prohibits disabling this.
- The next step is to monitor this.

Logging

- We require all our VMs running on our cloud to log
- They are configured, like the rest of our infrastructure, to use syslog to do this
- Again, disabling this is specifically prohibited in our terms of service.
- Again, the next step is to implement monitoring of compliance with this.

Quattor (Aquilon) and Our Cloud

- All of our infrastructure is configured using Quattor (As seen in this mornings presentation by James Adams) – investigating UGent developed OpenNebula Quattor component, using UGent Ceph component.
- We build updated managed and unmanaged images for users using Quattor to first install them and then we strip them back for the unmanaged images.
- Managed VMs are available, but we re-use hostnames.
- Rather than dynamically creating the hosts when managed VMs are launched, we use a hook when they are removed to make a call to Aquilon's REST interface to reset their 'personality' and 'domain'.

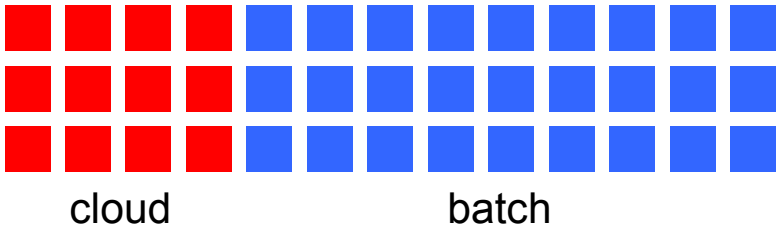
Expanding the farm into cloud

This work has already been presented by Andrew Lahiff and Ian Collier at ISGC – the content of the following slides has largely been provided by them.

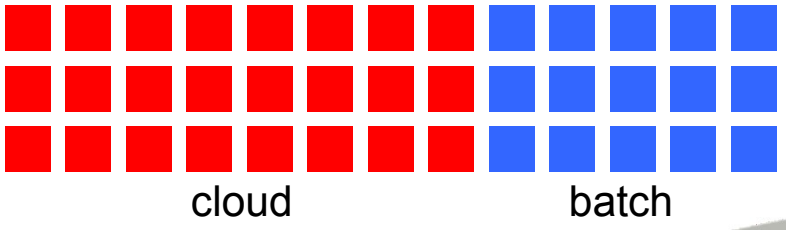
Much of it has also been presented at previous HEPiX's. So the following is a brief refresher and update.

Bursting the batch system into the cloud

- Initial situation: partitioned resources: Worker nodes (batch system) & Hypervisors (cloud)
- Ideal situation: completely dynamic
  - If batch system busy but cloud not busy
    - Expand batch system into the cloud



  - If cloud busy but batch system not busy
    - Expand size of cloud, reduce amount of batch system resources



Science & Technology Facilities Council

Bursting the batch system into the cloud

- This lead to an aspiration to Integrate cloud with batch system
  - First step: allow the batch system to expand into the cloud
    - Avoid running additional third-party and/or complex services
    - Leverage existing functionality in HTCondor as much as possible
- Proof-of-concept testing carried out with StratusLab in 2013
  - Successfully ran ~11000 jobs from the LHC VOs
- This will ensure our private cloud is always used
  - LHC VOs can be depended upon to provide work

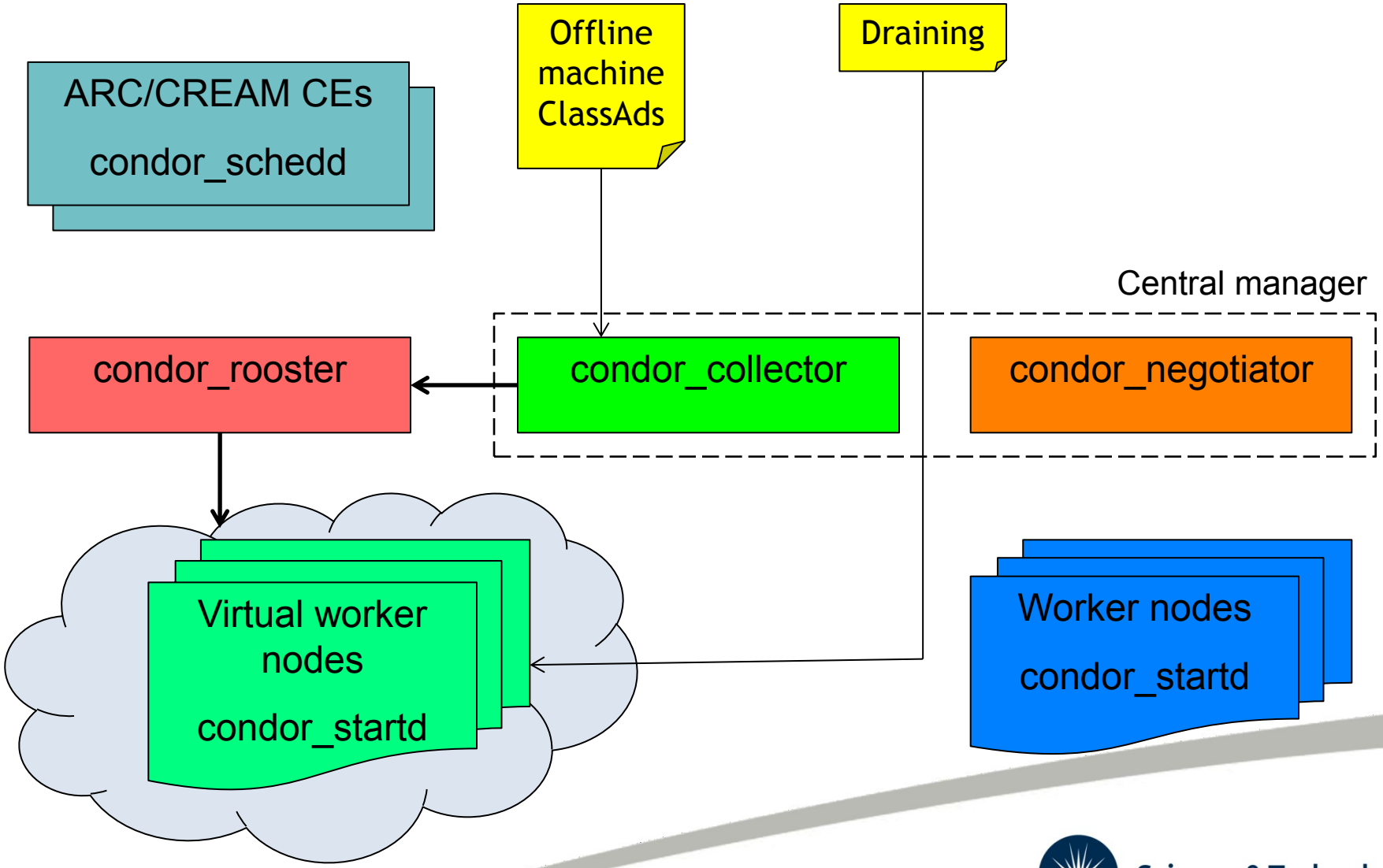
Science & Technology Facilities Council

Bursting the batch system into the cloud

- Based on existing power management features of HTCondor
- Virtual machine instantiation
  - ClassAds for offline machines are sent to the collector when there are free resources in the cloud
  - Negotiator can match idle jobs to the offline machines
  - HTCondor rooster daemon notices this match & triggers creation of VMs

Bursting the batch system into the cloud

- Virtual machine lifetime
  - Managed by HTCondor on the VM itself. Configured to:
    - Only start jobs when a health-check script is successful
    - Only start new jobs for a specified time period
    - Shuts down the machine after being idle for a specified period
  - Virtual worker nodes are drained when free resources on the cloud start to fall below a specified threshold



Science & Technology Facilities Council

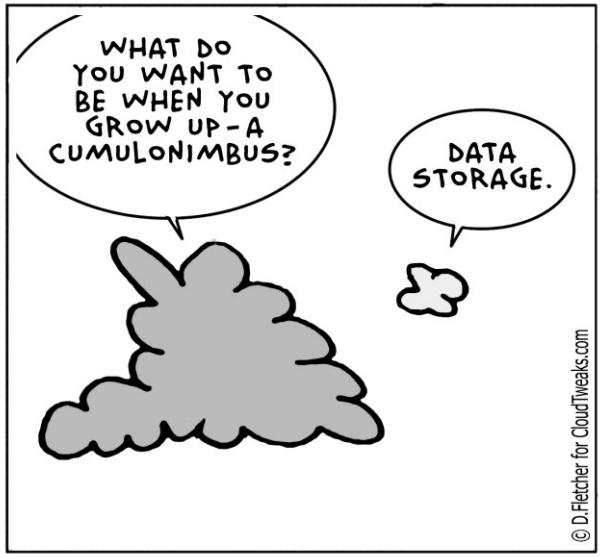
Bursting the batch system into the cloud

- Previously this was a short term experiment with StratusLab
- Ability to expand batch farm into our cloud is being integrated into our production batch system
- The challenge is to have a variable resource so closely bound to our batch service
- HTCondor makes it much easier – elegant support for dynamic resources
- But significant changes to monitoring
  - Moved to the condor health check – no Nagios on virtual WNs
  - This has in turn fed back in to the monitoring of bare metal WNs

What Next

- Consolidation of configuration, review of architecture and design decisions
- Development of new use cases for STFC Facilities (e.g. ISIS and CLF)
- Work as part of DataCloud H2020 project
- Work to host more Tier 1/WLCG services
- Continue work with members of the LOFAR project
- Engagement with non-HEP communities
- Start to engage with EGI Fed-cloud

Any Questions?



Science & Technology Facilities Council