GridPP

UK Computing for Particle Physics

Cloud & Virtualisation Update at the RAL Tier 1

Ian Collier

Andrew Lahiff

STFC RAL Tier 1

HEPiX, Lincoln, NEBRASKA, 17th October 2014

Science & Technology Facilities Council

e-Science

GridPP
UK Computing for Particle Physics

Virtualisation @ RAL Tier 1

- Recap
- Hyper-V Across STFC
- Scientific Computing Department Cloud
- Batch farm related virtualisation

Science & Technology Facilities Council
e-Science

Recap - HyperV at RAL Tier 1

- 2010 began investigating virtualisation platform
- Settled on Hyper-V
  - A challenge for Linux admins
- Production local storage Hypervisors in 2011
- Shared EqualLogic iSCSI storage a year or so later
  - As Martin mentioned it has recently been problematic
- Even with some problems, life – and service provisioning - much easier
  - Matter of minutes to hours to provision new host
- See previous talks for more detail

Hyper-V Issues

- Challenge for admins without professional Windows experience to manage
  - Especially when there are difficult issues
  - Support from overstretched Windows experts has been limited
- Recent shared storage issues saw hypervisors losing contact with shared volumes
  - Our attempts to recover made it worse not better (see above re Windows expertise)
  - Underlying cause may have been subtle firmware clashes
  - Rebuilding very carefully
  - But have just hired someone with specific expertise in this area

GridPP
UK Computing for Particle Physics

Virtualisation & Cloud @ RAL

- Context at RAL
- Hyper-V Across STFC
- Scientific Computing Department Cloud
- Batch farm related virtualisation

Science & Technology Facilities Council
e-Science

Context at RAL Tier 1

- 2014 STFC has run a review of corporate IT services
- Working group on virtualisation
- Several separate Hyper-V deployments across organisation
  - Tier 1 has by far the largest
- Also some VMWare
- Little shared effort or experience
- An intention emerging for common configurations and experience sharing

GridPP
UK Computing for Particle Physics

Virtualisation & Cloud @ RAL

- Context at RAL
- Hyper-V Across STFC
- Scientific Computing Department Cloud
- Dynamically provisioned worker nodes

Science & Technology Facilities Council
e-Science

Recap - SCD Cloud

- Began as small experiment 2 1/2 years ago
  - Initially using StratusLab & old worker nodes
  - Very quick and easy to get working
- Deployment until now implemented by graduates on 6 month rotation
  - Disruptive & variable progress
- Worked well enough to prove usefulness
- Something of an exercise in managing expectations

SCD Cloud Future

- Develop to full supported service to users across STFC
- IaaS upon which we can offer PaaS
  - One platform could ultimately be the Tier 1 itself
  - Integrating cloud resources in to Tier 1 grid work
- Participation in cloud federations
- Carrying out fresh technology evaluation.
  - Things have moved on since we started with StratusLab
  - Currently favour Opennebula

SCD Cloud

- March this year secured funding for 30 well specified hypervisors & 30 storage servers
  - ~1000 cores (plus ~800 in old WNs)
  - ~1PB raw storage
- Using OpenNebula
  - Collaborating with UGent on configuration tools
  - Ceph for image store and running images
  - AD as initial primary authentication
- 2 staff began work last month one full time, one half time

SCD Cloud

- Initially private cloud
- Closely coupled to Tier 1 batch – see later
- Provide support for Scientific Computing Department projects
  - Eg Horizon 2020 projects coming on line mid next year
- Also potential use cases with other STFC departments
  - Parallel to data services already provided to Diamond & ISIS

SCD Cloud

- Ceph backend
  - Been investigating for a couple of years
  - Have separate test bed (half a generation of standard disk servers (1.7PB raw) being tested for grid data storage
  - Will use for image store, running images, and possibly a data service
- Summer student worked on benchmarking & tuning
  - Investigated moving journals to SSDs
  - Although OSD benchmarks were much faster, overall performance no different
  - Turned out it was caches in the raid controllers
  - Will probably run on those systems with out the journal

GridPP
UK Computing for Particle Physics

Virtualisation & Cloud @ RAL

- Context at RAL
- Hyper-V Across STFC
- Scientific Computing Department Cloud
- Batch farm related virtualisation

Science & Technology Facilities Council
e-Science

Bursting the batch system into the cloud

- In Annecy described leveraging HTCondor features designed for power saving to dynamically burst batch work in to our private cloud
- Aims
  - 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 cloud
  - Successfully ran ~11000 jobs from the LHC VOs
- This allows us to be sure good utilization of our private cloud as LHC VOs can always provide work

The Vacuum Model

- "Vacuum" model is becoming popular in the UK
  - Alternative to CE + batch system or clouds
  - No centrally-submitted pilot job or requests for VMs
  - VMs appear by "spontaneous production in the vacuum"
  - VMs run the appropriate pilot framework to pull down jobs
  - See Andrew McNab's talk in Annecy
- Can we incorporate the vacuum model into our existing batch system?
  - HTCondor has a "VM universe" for managing VMs

The Vacuum Model & HTCondor

- Considered using fetch work hooks on the worker nodes
  - Unfortunately fairshares would not be respected
- Now using a scheduler universe job which runs permanently & submits jobs (VMs) as necessary
  - One for each VO
  - If no work or an error, only submit 1-2 VMs every hour
  - If VMs are running real work, submit more VMs
  - Fairshares are maintained since the negotiator decides what VMs to run

The Vacuum Model & HTCondor

- Running VMs using HTCondor
  - Using the CernVM 3 image
    - ISO image only 15MB in size, CVMFS provides root partition
  - Using a job hook for:
    - Creating a 20 GB sparse hard disk image file for the CVMFS cache
    - Creating a 40 GB sparse hard disk image file to provide additional disk space mounted as /scratch on the VM
    - Contextualization
    - Setting up NFS so that the pilot log files can be written on the worker node's disk
  - Using condor_chirp for:
    - Adding information about the status of the pilot to the job's ClassAd

The Vacuum Model & HTCondor

- Testing so far includes
  - Successfully running regular SAM tests from the GridPP DIRAC instance
  - ATLAS validation jobs

Job attribute summary	Sort by count, alpha
cloud (1)	UK (1798)
computingelement (1)	RAL-LCG2_VAC (1798)
computingsite (1)	RAL-LCG2_VAC (1798)
destinationse (1)	RAL-LCG2_VAC (1798)
jobstatus (2)	failed (326) finished (1472)
priorityrange (1)	4100:4199 (1798)
prodsourcelabel (1)	install (1798)
produsername (1)	Alessandro De Salvo (1798)
transformation (1)	sw-mgr (1798)

- First attempt at running jobs from ATLAS. Need to investigate the cause of the failures:
  - Payload ran out of memory
  - Error in copying the file from job workdir to local SE

Integrating Batch & Cloud

- All these developments allow us more flexibility in managing & delivering Tier 1 computing services

GridPP
UK Computing for Particle Physics

Virtualisation & Cloud @ RAL

- Context at RAL
- Hyper-V Services Platform
- Scientific Computing Department Cloud
- Batch farm related virtualisation

Science & Technology Facilities Council
e-Science

Summary

- Provisioning already much more responsive.
- Private cloud has developed from a small experiment to the beginning of a real service
  - With constrained effort - Slower than we would have liked
  - The prototype platform has been well used
- Demonstrated transparentl expansion of batch farm into cloud.
- Proof of concept vacuum model implementation using HTCondor
- Whole Tier 1 service becoming more flexible

Questions