Security, AuthN/AuthZ, and Software Evaluation for FermiCloud

S. Timm
Fermilab Grid and Cloud Computing Dept.
For the FermiCloud Project
HPACG, Apr. 14 2011
Virtualization at Fermilab

- FermiGrid Services
  - Highly Available provisioned virtual services
  - SLF5+Xen
- General Physics Compute Facility
  - Deployment of experiment-specific virtual machines for Intensity Frontier experiments
  - Oracle VM (Commercialized Xen)
- Virtual Services Group
  - Virtualization of Fermilab business systems using VMWare
  - Windows
What FermiCloud is

- Infrastructure-as-a-service facility
  - Developers, integrators, and testers get access to virtual machines without sysadmin intervention.
  - Virtual machines are created by users and destroyed by users when no longer needed. (Idle VM detection coming in phase 2).
  - Testbed to let us try out new storage applications for grid and cloud.

- A Private cloud—on-site access only for registered Fermilab users.
- A Project to evaluate the technology, make the requirements, and deploy the facility
- Unique use case for cloud—on public production network, integrated with rest of infrastructure.
Drivers for FermiCloud

- Developer machines 8+ years old with limited memory and CPU, power outage killed off remainder.
- Developers need test machines on short turnaround
- Improved utilization of power, cooling, and employee time for managing small servers and integration machines.
- CERN IT + HEPiX Virtualisation Taskforce program to have uniformly-deployable virtual machines.
- Virtualization under extensive use by SNS, FEF, FGS, and CMS T1.
- 16+ core systems lend themselves to hosting multiple logical servers on same physical hardware.
FermiCloud project staff

- Steve Timm (FGS)—project lead
- Dan Yocum, Faarooq Lowe (FGS), hypervisor and cloud control software installation and evaluation, early user support.
- Keith Chadwick—management and security policy
- Gabriele Garzoglio, Doug Strain, storage evaluation
- Ted Hesselroth—authentication and authorization development
- Many other Grid dept. staff and stakeholders who come regularly to meetings and tried early versions of cloud.
Stakeholders and Early Adopters

- Joint Dark Energy Mission (WFIRST)
  - Distributed messaging system, testing fault tolerance, ideal application for cloud
- Grid Department Developers
  - Authentication/Authorization
  - Storage evaluation
  - Monitoring/MCAS
  - GlideinWMS
- Fermi Scientist Survey
- Running Experiments Dept, job forwarding server
- Extenci project
Phase 1 of FermiCloud Project

- Acquisition of hardware (complete)
- Development of requirements based on stakeholder requirements (complete)
- Review of how well open source technologies match our requirements (Eucalyptus, Nimbus, OpenNebula).
- Storage evaluation
- Deployed pilot service for use by early adopters.
- All 3 major cloud software packages made new major releases in summer of 2010, had to retest.
Hardware

- 2x Quad Core Intel Xeon E5640 CPU
- 2 SAS 15K RPM system disk 300GB
- 6x 2TB SATA disk
- LSI 1078 RAID controller
- Infiniband card
- 24GB RAM
- 23 machines total
- Arrived June 2010
- +25TB Bluearc NAS disk
FermiCloud Network Topology

Physical

- fcl001
- fcl002
- fcl003
- fcl004
- fcl005
- fcl023

Logical

- vm-dual-1
- vm-dual-2
- vm-public
- vm-pubpriv-hn
- vm-priv-wn1
- vm-priv-wn2
- vm-man-a1
- vm-man-b1
- vm-man-b2

Cluster Controller

VLAN

April 14, 2011

FermiCloud—S. Timm
http://www-fermicloud.fnal.gov
Requirements--Security

- New VM's subjected to network vulnerability and virus scan before allowed on Fermi Network, leveraging our laptop network jail.
- VM's must use standard site-wide patching mechanisms.
- Periodically wake up dormant virtual machines to be sure they get their patches.
- Must have either Kerberos or X509 credential to launch a virtual machine and to log into it once it's launched.
- Cloud daemons must communicate via secure protocols.
- If X509 used, must be possible to replace SimpleCA with IGTF-approved certs.
Hypervisor Evaluation: Xen

- Xen:
  - At Fermilab since 2004
  - Consists of hypervisor, paravirtualized kernel, user tools
  - Paravirtualization and full hardware virtualization
  - Open Source, Citrix/EMC distribute commercial version
  - FermiGrid uses paravirtualized Xen almost exclusively including on all production grid gatekeepers, auth servers, batch system masters, and databases.
  - Part of Sci. Linux since SL 5.2
  - Red Hat drops support for Xen hypervisor in RHEL6 but RHEL6 can still be a Xen guest.
  - If necessary, we could get Xen hypervisor rpm's from xen.org as we did before.
  - Some time instability seen in 32-bit guest OS from SLF5.4+
  - Paravirtualized performance very good, almost indistinguishable from bare metal.
Hypervisor Evaluation: KVM

- KVM:
  - Bought a few years ago by RedHat, fully virtualized, works on newer hardware.
  - Initially just gave 100 Mbit/s ethernet, IDE disk
  - Now “virtio” drivers give much better performance
  - Support is in stock RHEL kernel, no alternate kernel needed
  - Possible to overbook memory on a VM host
  - Possible to see real memory and cpu usage from “top” on a VM host.
  - Some performance issues particularly on complex I/O tasks like Root, Lustre server, etc.
Hypervisor Evaluation

- Commercial hypervisors
  - VMWare is cost-prohibitive for 50-slot cloud
  - Commercialized Xen products also available
    - Oracle VM, commercial HVM Xen-based solution, used by FEF
    - Citrix XenServer, and its open-source cousin XCP.
  - Commercial hypervisors have their place but features are gradually moving to their open-source cousins.
  - In a cloud environment, extra bells and whistles of commercial hypervisor aren't needed.

- KVM vs. Xen:
  - Past experience has told us it is difficult to work against RedHat when they pick a technology winner.
  - We will deploy most of FermiCloud on KVM
  - Keep capacity to run Xen for some I/O intensive applications.
Eucalyptus

• Philosophy
  – Produce a open-source emulation of Amazon EC2 cloud.
  – Cloud and cluster controllers for overall control
  – Node controller on each node that hosts VM's

• Strengths
  – Most complete implementation of Amazon API's
  – Cleanly packaged software (RPMS)
  – Easy to deploy a small installation
  – Emulates Amazon's S3 and EBS storage API's as well
  – Web GUI support via HybridFox 3rd party browser addon
Eucalyptus

• Weaknesses
  - Protocols are scalable in theory but not the way Eucalyptus implemented them.
  - Most network traffic and disk traffic goes through cluster controller—single bottleneck and single point of failure
  - When cluster controller reboots all VM’s are lost
  - Not flexible in the kind of VM’s you can create
  - Uses x509 authentication on SOAP API but with self-signed SimpleCA certs and passwordless keys.
  - Developers promise scalability improvements but only in enterprise version
  - Developers refuse to make any changes that break compatibility with EC2.
  - Takes manual operation to save state of running VM.
  - No notion of scheduling at all.
Nimbus

• Philosophy
  - Grows out of Globus Virtual Workspace project
  - Includes a Globus WSRF interface to take grid certificates
  - Project dedicated to enabling science users to use “science clouds” both at university and lab facilities and on EC2

• Strengths
  - Has Globus WSRF frontend that handles grid certificates
  - Has notions of user and group quotas
  - Has notion of machine reservations
  - Can launch virtual machines via pilot jobs into a batch system
  - Has context broker for easy coordination of cluster launches
  - Developers are local and eager to collaborate.
Nimbus

- Weaknesses
  - Documentation of early versions was exasperating, dozens of little gotchas. Most have been fixed in version 2.6 but examples still don't all work right.
  - Have to open up lots of permissions on libvirt sockets and in sudoers to get things to work right.
  - Default installation dependent on SimpleCA certificate authority and passwordless private keys, provides way to swap them out.
OpenNebula

• Philosophy
  - OpenNebula is part of EU Reservoir project,
  - Started as a virtual infrastructure manager and added cloud API's afterwards

• Strengths
  - Most flexibility in making the virtual machines we want.
  - Large developer and user base
  - Proven performance at HEP-lab scale at CERN
  - Good scheduling features
  - Least sysadmin time required to install it.
  - Fewest single points of failure and network bottlenecks
  - Most robust operations, daemons run well, recover after reboot.
OpenNebula

• Weaknesses
  - Ships with “simple” authentication—plain text passwords
  - Pluggable authentication allows for x.509, LDAP, SSL
  - No group quotas by default, can be added.
  - Limited Amazon ReST API functionality, no Amazon SOAP API, but it's on their roadmap.
# Authentication/Authorization comparison

<table>
<thead>
<tr>
<th>CLOUD SYSTEM</th>
<th>Upload Image</th>
<th>Launch VM CLI</th>
<th>Launch VM API</th>
<th>Login</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eucalyptus</td>
<td>X509</td>
<td>X509</td>
<td>X509, EC2_ACCESS_KEY</td>
<td>ssh-keypair</td>
</tr>
<tr>
<td>Nimbus</td>
<td>X509</td>
<td>X509</td>
<td>X509, EC2_ACCESS_KEY</td>
<td>ssh-keypair</td>
</tr>
<tr>
<td>OpenNebula</td>
<td>user/pass</td>
<td>user/pass</td>
<td>user/pass, EC2_ACCESS_KEY</td>
<td>ssh-keypair</td>
</tr>
</tbody>
</table>
Current FermiCloud Pilot Service Deployment

- 8 nodes deployed in OpenNebula
- 7 nodes deployed in Nimbus
- 3 nodes deployed in Eucalyptus
- 4 nodes dedicated to storage investigations
- Persistent virtual machines running include
  - GUMS servers,
  - MySQL servers,
  - MCAS servers
  - dCache servers
  - JDEM/WFIRST machines
- We supply a sample virtual machine OS install and a template to start a virtual machine.
Storage Virtual Machine

2 TB
6 Disks

Dom0:
- 8 CPU
- 24 GB RAM

Lustre Server VM

Lustre Client VM

6 x

FG ITB
Clients

mount

BA

mount

http://www-fermicloud.fnal.gov
Phase 2: Monitoring and backfill

- Goals:
  - Make sure the machines that are supposed to stay up all the time stay up
  - Make sure the appropriate cloud daemons are running and load of the head nodes is reasonable
  - Detect idle virtual machines and pause them based on policy
  - Fill in with worker node VM's and take jobs from the grid.
- Nimbus 2.7 already claims this feature
  - Currently checking it out
  - Evaluating other policy-based idle machine detection mechanisms as well.
Phase 2: Authentication and Authorization

- Fermilab developers have written x509 authentication plugin for OpenNebula
- Works with CLI and EC2 Query API
- Successfully tested with patched OpenNebula 2.2 prerelease, now making plans to deploy in production
- Have contributed our patches to OpenNebula trunk.
- Authorization based on Opennebula user database, user's DN is stored in place of where EC2 secret key would be.
FermiCloud Phase 2 Program of Work

- Complete storage evaluation
- Develop and deploy automated provisioning and patching
- Security Policy
- Infiniband and MPI
- Image repository
- Live Migration
- OSG collaboration
Conclusions

- FermiCloud has completed our technology evaluation and requirements gathering phase.

- No open source cloud software meets all our requirements—but for the retail VM's we are doing now OpenNebula is best. Most remaining requirements are in OpenNebula roadmap of future deployment.

- For private cloud, important to have more flexible API's than EC2 for max flexibility.

- We have deployed a pilot service which is gaining increased use by developers and integrators.

- We have a robust program of work defined to transform our pilot service into a production service.

- We welcome interest from new users and stakeholders.