FermiCloud - Enabling Scientific Computing with Integrated Private Cloud Infrastructures

Keith Chadwick for the
FermiCloud project

EGI Technical Forum 2011

Work supported by the U.S. Department of Energy under contract No. DE-AC02-07CH11359

Fermilab

What FermiCloud Is

- Infrastructure-as-a-service private cloud for the Fermilab Scientific Program.
- Integrated into the Fermilab site security structure.
- Virtual machines have full access to existing Fermilab network and mass storage devices.
- Scientific stakeholders get on-demand access to virtual machines without system administrator intervention.
- Virtual machines created by users and destroyed or suspended when no longer needed.
- Testbed for developers and integrators to evaluate new grid and storage applications on behalf of scientific stakeholders.
- Ongoing project to build and expand the facility:
  - Phase 1 - Technology evaluation, requirements, deployment,
  - Phase 2 - Scalability, monitoring, performance improvement,
  - Phase 3 - High availability and reliability.

Other Virtualization Projects 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, RHEL 5

Drivers For FermiCloud

- Continue program of virtualizing all scientific servers that can be virtualized.
  - Many experiment servers need minimal CPU, memory but want ports to themselves.
- Improve utilization of power, cooling and employee time (admins and developers) for managing small science servers.
- Had to replace 6 racks of legacy development machines with limited hardware budget and computer room space.
- CERN IT + HEPiX Virtualisation Taskforce program to have uniformly-deployable virtual machines. Expect LHC and future Fermilab experiments will eventually require cloud technology.

Science Stakeholders

- Fermilab Intensity Frontier
  - Monitoring Server (MCAS)
  - GridFTP endpoint server
  - Experiment-specific storage investigations
- Fermilab D0 Experiment
  - Job Forwarding Server
- Extenci project (Cloud activities, LHC)
  - Distributed storage on WAN.
- GEANT4
  - Validation server
- hosts for Scientific middleware development.
- Host developers and integrators of OSG middleware.
- Joint Dark Energy Mission->WFIRST->LSST
  - Distributed messaging system, testing fault tolerance.

FermiCloud 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
- Just delivered – 84 TB of Nexsan SAN disk

FermiCloud Software Technologies

- OS: Scientific Linux 5 & 6
- Hypervisor: Paravirtualized KVM
  - Fully virtualized KVM available as an option.
  - KVM allows sharing of read-only memory sections across multiple VMs with copy on write.
- Cloud Management: OpenNebula
- Modifications to OpenNebula CLI, Query API, GUI to use X.509 authentication to launch virtual machines.
- Secure credential store,
  - All security secrets loaded at boot time only.
- Site-wide patching and vulnerability scanning facilities.

X.509 Authentication

- We performed a requirements analysis and constructed a weighted decision matrix of the features of Eucalyptus, Nimbus & OpenNebula.
  - OpenNebula was chosen as the winner.
- Most 3rd-party tools use EC2 Query API so we have to make it work.
- Use pluggable authentication features of OpenNebula to use internal X.509 authentication.
  - "secret key" in user database -> X.509 DN.
- 3 components modified thus far:
  - command line, "econe" query daemon, and SunStone GUI.
- Patches contributed back to OpenNebula, released as part of OpenNebula 3.0 beta 2.
  - Clients: HybridFox works without modification. Condor-G modifying EC2_GAHP to support, first tests worked.
  - We have a beta of the ruby XACML callout and have tried it against GUMS (OSG Grid User Mapping Service), and it appears to be working.

X.509 Authentication Details

- For "econe" and "Sunstone":
  - X.509 certificate or proxy authenticated by "Gridsite",
  - If a certificate is present, DN is passed to the OpenNebula core for normal "password" check,
  - Server creates login token valid for subsequent operations,
  - Ruby plugin for X.509 authentication used.
- For CLI:
  - Present certificate via X509_USER_CERT,
  - Create authentication token via oneauth command,
  - Ruby plugin for X.509 used,
  - Also created certificate-based login for the admin user.
- For OCCI:
  - Haven't attempted yet but expect that the strategy used with "econe" would work.

Towards X.509 Cloud Authorization

- Authorization interoperability protocol currently used in OSG, EGI grids among others. Fermilab was part of effort. (GFD159),
- Clients make XACML-based callout to authorization servers,
- We want to implement VO and role-based authorization (and resource quotas) as we move towards a bigger cloud,
- Open-source cloud software needs to clean up its AuthZ code anyway, lots of little MySQL databases, all different.

High Availability & Service Levels

- We are currently working towards High Availability deployment:
  - Based on prior experience with FermiGrid-HA2 project,
  - Add SAN for live migration and large datablock capacity,
  - Split FermiCloud between two buildings,
  - Mirror storage between two buildings,
  - Set up high-availability procedures for failover of cloud controller and migration of virtual machines.
- Offer three service levels:
  - High availability 24x7,
  - Regular (9x5) virtual machine,
  - Opportunistic (spot market) can be pre-empted anytime,
  - Support overbooking (in conjunction with hyperthreading).
- Stakeholders will eventually be "billed" for usage according to an economic model, analogous to existing tape robot facility.
  - The draft economic model has a 24x7 non-hyperthreaded FermiCloud VM priced at slightly less than a 1 year pre-purchased "EC2 small" compute unit from Amazon,
  - FermiCloud does not charge for I/O and offers much better network connectivity to the storage at Fermilab,
  - Unbilled cycles will be used to "spin-up" additional worker node VMs to add to the existing Grid clusters.

Using Virtualization to Enable Science

- New interactive science applications that require ongoing interaction or unique network topologies and don't fit grid batch processing paradigm.
- Complicated software stacks where grid distribution has been difficult or impossible.
- Legacy experiments which require specific OS and library combinations.
- Extra compute capacity on demand for experiments that suddenly require it.
- Virtualization used on 32-core+ worker nodes to:
  - Pin applications to appropriate CPU-memory combinations for better performance
  - Sandbox applications to keep one rogue job from killing the other 31.
  - Memory segments can grow or be shared as needed.

Recent Success Stories

- OSG pacman to rpm refactoring:
  - From: Alain Roy – Thursday 04-Aug-2011
  - I want to give a big thanks for FermiCloud. The OSG software team has been using it to support our work with RPMs, and it's been invaluable. For the kind of work we are doing, we need to be able to work with pristine machines (or virtual machines) with root privileges so we can try things out that may break the computer. Being able to have a handful of VMs at my disposal to try things out is marvelous. The fact that they come setup with host certificate, access to my home directory, and a good default configuration is amazing--I can use them with very little effort. Thank you so much! Please share my gratitude with whoever is appropriate.

- Intensity Frontier Custom GridFTP Servers:
  - Specialized GridFTP servers with custom DN to Unix UID mappings to allow Intensity Frontier experiments to "push" files to the GridFTP endpoints and have the files be stored in the storage system with desired UID/GID ownership.
  - Multiple deployments, in order to handle cases where one DN is a member of multiple experiments (sub-VOs in the fermilab VO).
  - Additional custom GridFTP deployments can be "spun up" in a couple of hours.

Ongoing Software Development

- Accounting and billing
  - Cloud accounting add-ons to Gratia accounting project.
- Monitoring
  - How many machines are running, who is running them, is everything up that should be up?
- Authorization
  - Apply well-tested and interoperable grid authorization tools to cloud authorization as well.
- All of above in collaboration with other projects and standards bodies.

Interoperability Efforts

- We are in contact with HEPiX virtualisation taskforce, VM catalogs and repositories, image format standards,
- Making X.509 work with OCCI interface,
- We are in contact with Condor-G developers, they have agreed to support X.509-authenticated ReST protocol, early tests are good,
- We are working to assure Interoperability with GlideinWMS project in OSG.

Current Technology Investigations

- Testing storage services with real neutrino experiment codes, identify NFS alternatives.
- Using Infiniband interface to create sandbox for MPI applications.
- Batch queue look-ahead to create worker node VM's on demand.
- Submission of multiple worker node VM's, grid cluster in the cloud.
- Idle VM detection and suspension, backfill with worker node VM's.
- Leverage site "network jail" for new virtual machines.
- IPv6 support.
- Testing dCache NFS4.1 support with multiple clients in the cloud.
- Interest in OpenID/SAML assertion-based authentication.

Conclusions

- FermiCloud has successfully deployed a wide range of servers for the scientific program.
- FermiCloud has been a testbed for several evaluations of storage and middleware that benefit the scientific program.
- FermiCloud has already provided significant power and cooling savings, and significant convenience benefits to scientific stakeholders
- Now integrating our work with other internal Fermilab virtualization activities and external projects.
- We welcome interest from new users, stakeholders, and other cloud-based projects.