

The FermiCloud Project: Pilot Service Deployment and Future Plans:

Steven C. Timm
Fermilab Grid [and Cloud] Facilities Dept.

Computing Techniques Seminar
December 9, 2010

Outline

- Introduction to Cloud Computing and FermiCloud Project
- Requirements
- Technology Evaluation
- Pilot Service Deployment
- Program of work for FermiCloud Phase 2

Cloud Computing Introduction

- 3 basic types of Cloud Computing Services
 - Software-as-a-service (salesforce.com, Kronos)
 - Platform-as-a-service (Windows Azure, Google App Engine)
 - Infrastructure-as-a-service (FermiCloud, Magellan, Amazon Web Services)
- 3 types of Cloud:
 - Public cloud—Web API allows all authorized users to launch virtual machines remotely on your cloud. (Amazon)
 - Private cloud—Only users from your facility can use your cloud (FermiCloud)
 - Hybrid cloud—Infrastructure built from mix of public and private.
- Object-oriented storage (Hadoop, etc.) closely linked to cloud paradigm.

Common Cloud concepts

- Overall User Interface for requesting a VM (cloud controller + API)
- One or more Cluster Controllers which control a group of nodes
- A Node Controller on each node which can activate virtual machines
- A repository of virtual image files
- “Ecosystem”--the group of developers and users who make 3rd-party tools for cloud computing
- Hypervisor—the part of operating system which manages virtual machines.

Related Fermilab Virtualization Projects

- 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.

Why Developers and Integrators need FermiCloud

- Old developer machines in FAPL/Gridworks were old with limited memory and CPU
- Most were either burned out in Feb. power outage or forced to be turned off after that.
- Developer's order for standalone test machines can be time-consuming, many months.
- In FY10 CET and DOCS pooled their hardware funds into FermiCloud project, along with FGS funds, for a share of FermiCloud.

Fermilab Drivers for Virtualization and Cloud Computing

- 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.

What FermiCloud is not [yet?]

- Won't allow keeping vintage OS around past support deadlines.
- Won't allow you to do anything security-wise that you aren't allowed to do on a real machine.
- It's not a plan to virtualize all worker nodes in FermiGrid.
- It's not a plan to take virtual machines as jobs from the grid.
- The web services are not open for off-site users to launch untrusted VM's, only Fermi users can launch VM's with supported Fermi OS.

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 (Kowalkowski/Paterno)
 - Distributed messaging system, testing fault tolerance, ideal application for cloud
- Grid Department Developers
 - Authentication/Authorization
 - Storage evaluation
 - Monitoring/MCAS
 - GlideinWMS
- dCache Developers
- LQCD testbed
- OSG

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

December 9, 2010

Phase 1 of FermiCloud Project

- In the first year of the project we have experimented with the leading open source technologies for hypervisors and cloud computing software
- We wrote final requirements based on our experience of putting real users on the existing software
- We then made a weighted decision matrix based on how well the major software packages meet our requirements
- All 3 major cloud software packages made new major releases this summer, had to redo some testing.
- The following slides are a summary of requirements

Requirements—OS and Hypervisor

- Head nodes can run SL as host OS.
- Guest OS—Sci. Linux Fermi, Fermi STS (Fedora), Windows
- Support KVM and Xen Hypervisors, VMWare optional
- Userspace tools for Linux, Windows, Mac
- Documentation of how to move VM's from desktop to cloud and back
- Flexible creation of virtual machines—IB, multiple network, multiple disks.

Requirements—Provisioning + Contextualization

- Option to boot Linux install image and install OS from kickstart
- Leverage CFEngine or Puppet
- Automated and supportable by groups that manage large numbers of machines
- Creation of clusters of virtual machines so they all know where the others are.
- Allow for installation or creation on the fly of machine-specific files like kerberos keytabs, host certs, etc.

Requirements—Object Store

- Hold at least 300 images at average of 10GB each
- Clear instructions on image format, how to make and convert. Encryption and compression a plus.
- Launch up to 400 copies of same image simultaneously, or 100 different images, at speed consistent with network wire speed.
- User allowed to keep images private
- Amazon S3 or EBS emulation a plus if scalable.
- No machine-dependent secrets stored on machine, (kerberos keytab, x509 host cert and key).

Requirements--Interoperability

- Support Amazon EC2 ReST (Query) API and SOAP API
- Accept submissions from Condor-G's ec2 adapter
- Able to run CERNVM formatted images
- Able to use CVMFS, http-based read only file system of CERNVM
- Able to run standardized VM's being designed by Hepix virtualisation taskforce
- Able to run sample LHC Tier-3 machines
- Have configurable capability to “cloudburst” to external providers such as Amazon EC2 if desired
- Extra API's such as Open Cloud Computing Initiative a plus
- Users must be able to launch and monitor machines from web GUI.

Requirements--Functionality

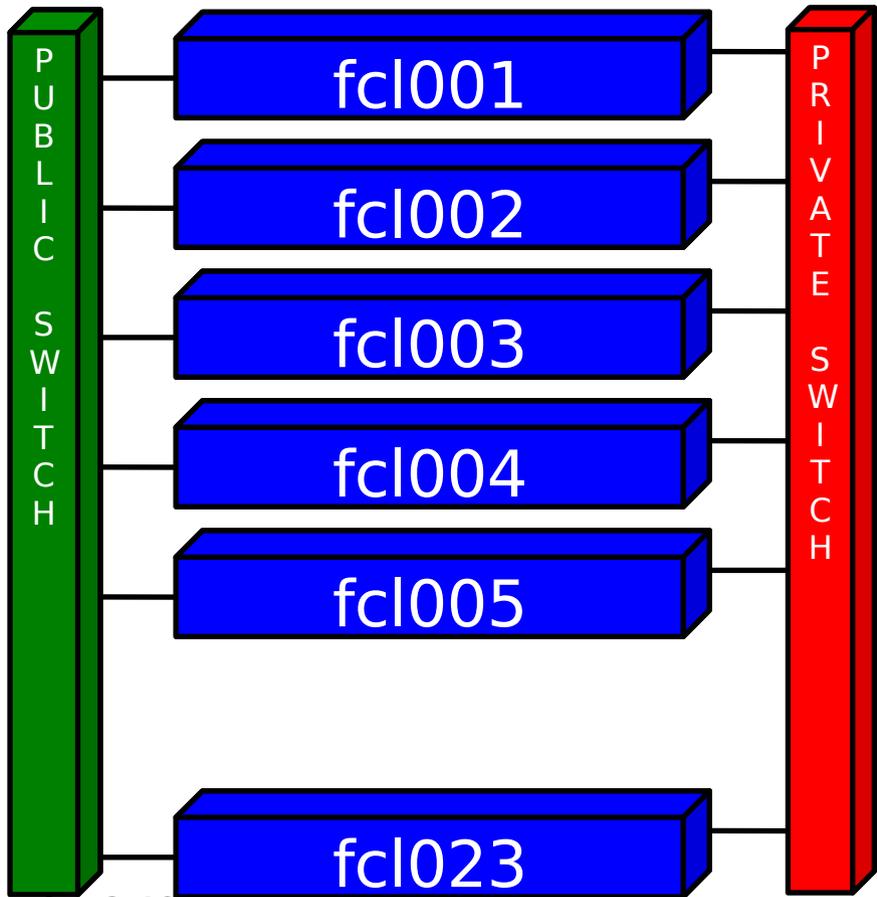
- Pause and resume virtual machine, and save state
- Detect idle virtual machines, pause them, and backfill with worker node virtual machines
- Launch VM's at fixed time or reserve a block of VM's in advance.
- Recover from reboot of single virtual machine host without confusion
- Recover from reboot of cloud controller without losing all VM's
- Live migration of VM's from one node to another
- Architected so that it can be made highly available
- User and group quotas, fair-share scheduling
- Stable running of daemons for long periods of time

Requirements--Network Topology

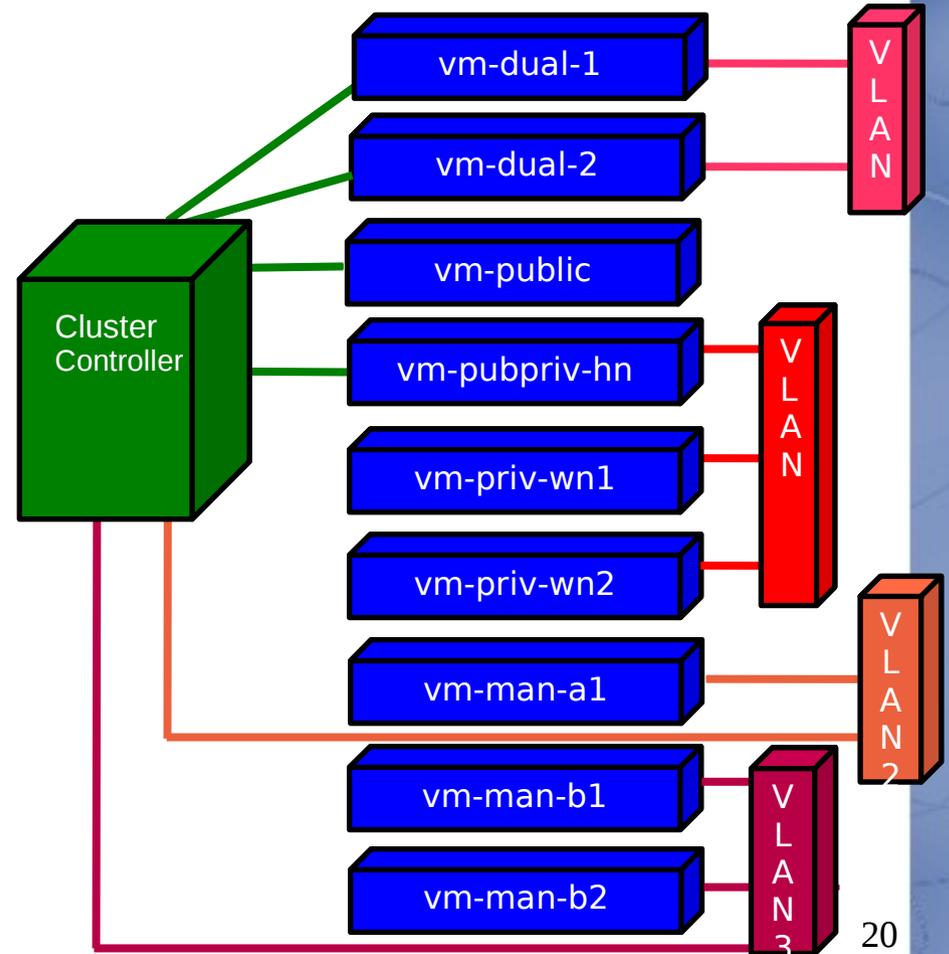
- IPv6 support
- Launch VM with same public IP address every time
- Launch VM with one of a pool of public IP addresses and load machine-specific data on the fly
- Launch cluster of VM's with one head node that can see public and private net and workers that can only see private net.
- Launch machines with public and private network, with public net only to be activated after patching and virus scan.
- Launch machine on private network and attach public IP address later (Amazon “Elastic IP”).

FermiCloud Network Topology

Physical



Logical



December 9, 2010

Requirements—Billing and Accounting, Clustered File Systems

- Billing and Accounting:
 - How many VM's run by user and group
 - How long they ran
 - How active they were while they ran
 - Cloud accounting/billing will be a Gratia extension.
- Must work with following clustered file systems:
 - Lustre, Hadoop, dCache, BestMan (xrootd and gateway) and NFS
 - Support for OCFS2 is a plus.

Requirements--Security

- New VM's subjected to network vulnerability and virus scan before allowed on Fermi Network, leverage laptop network jail if possible.
- 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.

Xen network performance

Xen iperf tests (2 min. test)	Bandwidth	Host cpu %	VM cpu %
Bare metal to bare metal	947 Mbps	11	
Xen to bare metal, 1 stream	949 Mbps	0	11
2 Xen to bare metal stream A	474 Mbps	0	3.5
2 Xen to bare metal stream B	478 Mbps	0	3.5
2 Xen to 2 bare metal stream A	265 Mbps	0	2
2 Xen to 2 bare metal stream B	681 Mbps	0	5
Xen to Xen within same host	1860 Mbps	0	15

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.

KVM Network performance

KVM iperf tests (2 min. test)	Bandwidth	Host cpu %	VM cpu %
Bare metal to bare metal	950Mbps	14	
KVM to bare metal 1 stream	949 Mbps	110	56
2 KVM to bare metal stream A	474 Mbps	56	11
2 KVM to bare metal stream B	483 Mbps	56	11
2 KVM to 2 bare metal stream A	463 Mbps	56	11
2 KVM to 2 bare metal stream B	490 Mbps	56	11
KVM to KVM within same host	2910Mbps	120	56

KVM vs. Xen Disk performance

KVM Disk test (Bonnie)	Size (MB)	Block read KB/sec	Block read CPU	Block write KB/sec	Block write CPU
Bare metal	30000	67908	15.4	51341	3.8
KVM VM	30000	23073	5.5	58959	7.4

Xen Disk test (Bonnie)	Size (MB)	Block read KB/sec	Block read CPU	Block write KB/sec	Block write CPU
Bare metal	30000	54117	10.6	62093	0.8
Xen VM	30000	59754	7.0	69929	0.1

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
 - Default security is wide-open.
 - Has “pluggable authentication module.” You bring the plug.
 - Limited Amazon ReST API functionality, no Amazon SOAP API.

Authentication/Authorization comparison

CLOUD SYSTEM	Upload Image	Launch VM CLI	Launch VM API	Login
Eucalyptus	X509	X509	X509, EC2_ACCESS_KEY	ssh-keypair
Nimbus	X509	X509	X509, EC2_ACCESS_KEY	ssh-keypair
OpenNebula	user/pass	user/pass	user/pass, EC2_ACCESS_KEY	ssh-keypair

Current FermiCloud Pilot Service Deployment

- 8 nodes deployed in OpenNebula
- 7 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.

Sample Management Console Displays

OpenNebula Management Console

Logged in as oneadmin - logout | version: 1.0.1

vm overview vm manager hosts networks users

Cloud vm's:

Id	User	Name	VM State	LCM State	Cpu	Memory	Host	VNC Port	Time		
70	oneadmin	cloudmysql.fnal.gov	active	running	0	2146304	fcl004	n/a	28d 2:47:26	[details] [log]	<input type="checkbox"/>
72	oneadmin	cloudlvs.fnal.gov	active	running	0	2146304	fcl004	n/a	25d 4:16:29	[details] [log]	<input type="checkbox"/>
110	timm	cloudgums1.fnal.gov	active	running	0	4194304	fcl007	n/a	7d 3:53:38	[details] [log]	<input type="checkbox"/>
111	timm	cloudgums2.fnal.gov	active	running	0	4194304	fcl007	n/a	7d 3:43:54	[details] [log]	<input type="checkbox"/>
121	parag	mcas	active	running	0	4194304	fcl007	n/a	5d 1:23:14	[details] [log]	<input type="checkbox"/>
122	dwd	dwd	active	running	0	4194304	fcl007	n/a	4d 3:9:41	[details] [log]	<input type="checkbox"/>
123	fmoscato	fmoscato	active	prolog	0	0	fcl005	n/a	0d 0:5:38	[details] [log]	<input type="checkbox"/>
124	fmoscato	fmoscato	active	prolog	0	0	fcl004	n/a	0d 0:5:34	[details] [log]	<input type="checkbox"/>
125	fmoscato	fmoscato	active	prolog	0	0	fcl003	n/a	0d 0:5:33	[details] [log]	<input type="checkbox"/>
126	fmoscato	fmoscato	active	prolog	0	0	fcl008	n/a	0d 0:5:32	[details] [log]	<input type="checkbox"/>

hold ok

Deploy vm:

VM Template: fnpcsrvtb

Amount:

deploy

OpenNebula Management Console

Logged in as oneadmin - logout | version: 1.0.1

vm overview vm manager hosts networks users

Cloud hosts:

Id	Name	Running VM	Total cpu	Free cpu	Assigned cpu	Total memory	Free memory	Status		
13	fcl003	0	1600	1504	96	24676608	24190724	on	[details]	<input type="checkbox"/>
14	fcl004	2	1600	1406	193	24676608	19375944	on	[details]	<input type="checkbox"/>
16	fcl005	0	1600	1499	100	24676608	23888192	on	[details]	<input type="checkbox"/>
8	fcl006	0	1600	1600	0	24676608	24392036	on	[details]	<input type="checkbox"/>
17	fcl007	4	1600	1416	184	24676608	15691544	on	[details]	<input type="checkbox"/>
12	fcl008	0	1600	1427	172	24676608	24203596	on	[details]	<input type="checkbox"/>
11	fcl009	0	1600	1600	0	24676608	24295652	on	[details]	<input type="checkbox"/>

enable ok

Create host:

Hostname:

IM:

VMM:

TM:

create

FermiCloud Phase 2 Program of Work

- Authentication/Authorization
- Provisioning and Patching
- Security Policy
- Infiniband and MPI
- Storage evaluation
- Image repository
- Live Migration
- Monitoring and Backfill
- OSG collaboration

Authentication and Authorization

- EC2 ReST API—very handy, all open source clouds support it, all web GUI's depend on it
 - Is it secure enough?
 - Can we make it secure enough?
 - Is it worth our trouble to make it secure enough?
 - Ted Hesselroth working on these questions, has a proof-of-principle demonstration.
- EC2 SOAP API—x509 access—vital for Condor access
 - By default uses self-signed certs with no-password private keys, can we fix?
 - OpenNebula doesn't have it at all, can we or they add it?
- Command line—which authentication is used, what security is appropriate?
- Can x509 systems be made to leverage grid tools like GUMS, VOMS, SAZ?
- As far as we know, we are first to do something about these questions, can be unique contribution to larger cloud community.

Provisioning and patching

- We want a “kickstart-me-now” feature in which a user can request a PXE boot and a clean kickstart install of Sci. Linux with a user-specified KS file.
- Leverage Fermilab's scanning and node registration service and treat the new VM like an incoming hostile laptop, don't let it on the net until it is scanned, registered and patched.
- Wake dormant machines up from time to time and give them their patches
- Both Eucalyptus and Nimbus give only the system admins the power to upload kernels and ramdisk—gives control over which OS you will support and keeping the kernels updated.

Security Policy

- FermiCloud currently operates under Open Science Enclave policy
- Stakeholder interest in private-only networks, passwordless rsh and ssh
- Stakeholder interest in OS's of alternate patch level or possibly not baselined at all.
- Users could request variance from CSExec just like on a real machine
- Policy work needs to be done to understand the effect on the cloud service that is hosting security-variant VM's.

Infiniband and MPI

- Lattice QCD cluster wants virtual mini-MPI testbed of 4 virtual nodes
- Let users test application with actual infiniband drivers and hardware without disrupting large LQCD clusters.
- For this we want actual Infiniband drivers and hardware present and visible in the VM's.
- 3rd generation “Infiniscale” cards can be passed through to 1 VM
- 4th generation “Connect-X” cards claim that they can be shared with several VM on same machine. Investigation continues.
- Can use Infiniband for private IP network applications if faster private net is necessary
- Infiniband can also be important for connection to storage and SAN.

Storage Evaluation

- Closely related Storage Evaluation project led by G. Garzoglio
- Trying to find new many-to-many storage technology for Intensity Frontier
- On FermiGrid these experiments currently use Bluearc NAS device.
- Trying Lustre and Hadoop both on bare iron and under KVM
- 8-core machine probably unnecessary to serve 10TB of disk. Allocate 2 cores to serve storage and let the rest of cores be compute virtual machines.
- Testing with actual neutrino experiment “root” application, benchmarking different solutions.

Storage Virtual Machine

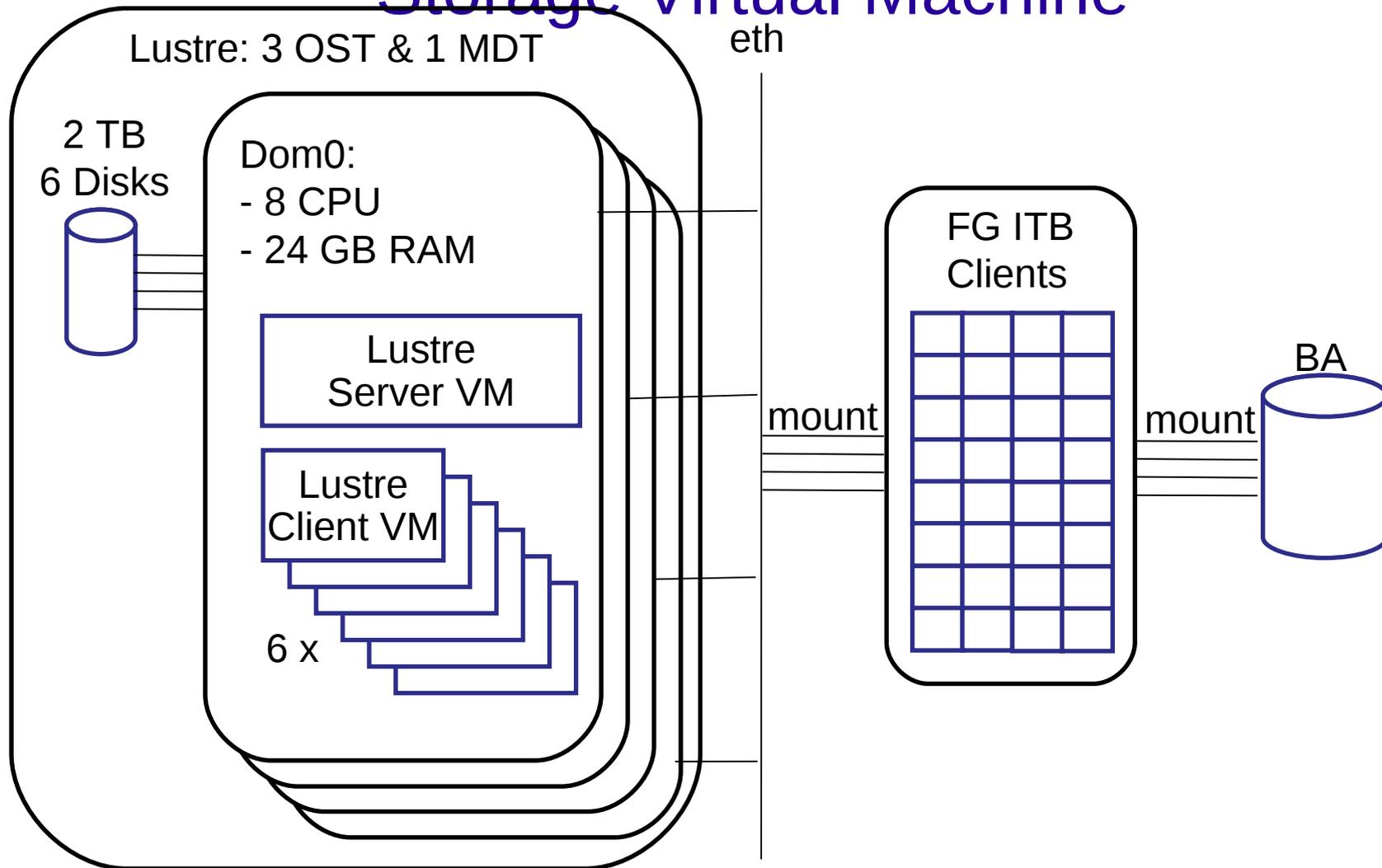


Image Repository

- Currently demo clouds are storing VM images on single node and copying with scp to final destination.
- NFS could also be used for OpenNebula—faster launch time but performance hit once you get it there.
- LVM has been used by CERN—good for putting 1000 of the same image across a cluster of worker nodes.
- We have ~350 cores available and will be running lots of different virtual machines, local caching of limited use.
- 3rd party image repositories also becoming available.

Live Migration

- Support of servers means we need to do live migration
- A SAN is usually used as back-end storage for virtualized systems that are doing live migration
- It may also be possible to use NFS for this
- We will make evaluation and recommend purchase of small SAN if necessary.

OSG Collaboration

- Discussions with OSG at early stage
- OSG has a cloud allocation on Amazon EC2
- Investigating constructing a model cloud-based computing facility for various virtual organizations to use.

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.
- A potential collaboration with Condor team here
 - Condor has most of pieces to determine if node is idle, suspend, and resume.
 - Using them for cloud would be new application.

Conclusions

- FermiCloud has completed our technology evaluation and requirements gathering phase.
- No open source cloud software meets all our requirements but we can meet them with a combination.
- 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
- We welcome other departments of CD to join the effort and collaborate.