

FermiGrid and FermiCloud Updates

Keith Chadwick

Grid & Cloud Computing Department
Fermilab

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

Outline

FermiCloud Introduction:

- Acknowledgements, Staffing

FermiGrid Project:

- Activities over the past year,

FermiCloud Project:

- Background and Mission, Project Phases,

FermiCloud Facility and Software:

- Fault Tolerance, Distributed Replicated File System, FermiCloud-HA,
- Current Capabilities, Typical Use Cases, Proposed Economic Model & SLA,
- Cloud Computing Environment, Idle VM Detection, vCluster
- Usage, Personnel Effort, Review

Summary & Conclusions

Acknowledgements

None of this work could have been accomplished without:

- The excellent support from other departments of the Fermilab Computing Sector – including Computing Facilities, Site Networking, and Logistics.
- The excellent collaboration with the open source communities – especially Scientific Linux and OpenNebula,
- As well as the excellent collaboration and contributions from the Korea Institute of Science and Technology Information (KISTI).

Grid & Cloud Computing Department

Keith Chadwick (Department Head)
Gabriele Garzoglio (Associate Head)

Distributed Offline Computing Services

Gabriele Garzoglio (Leader)

David Dykstra
Hyunwoo Kim
Tanya Levshina
Parag Mhashilkar
Marko Slyz

2012 Summer students:

Siyuan Ma, Anupam Rajendran, Giovanni Franzini

FermiGrid Services (operations)

Steven C. Timm (Leader)

Gerard Bernabeu Altayo
Kevin Hill (OSG Security)
Neha Sharma
Karen Shepelak

FermiGrid

<http://fermigrid.fnal.gov>

FermiGrid Over the Past Year

The past year has been basically incremental progress and upgrades:

- “Routine” hardware upgrades and replacements,
- “Routine” software services upgrades,
- FermiGrid-HA/FermiGrid-HA2 infrastructure continues to deliver benefits.

We have run into a few issues along the way...

- The “three service Monte” that was used to maintain service availability during our Grid User Mapping Service (GUMS) upgrade mostly worked – a missed database connection parameter on the temporary GUMS server resulted in ~30 minutes of partial service function vs. the planned 10 seconds,
- MyProxy upgrade did not go well – appears to have been caused by a bad DNS caching configuration on the (new) MyProxy server VMs – we wound up having to back out the upgrade (having everything in VMs certainly helped here) and are working on further testing and will soon reschedule the upgrade,
- Coordination of the XACML AuthZ/AuthN deployment across the seven Grid clusters and associated worker nodes wound up taking much longer than initially anticipated,
- We performed the final (for Fermilab) VOMRS->VOMS migration yesterday.

FermiGrid-HA/FermiGrid-HA2

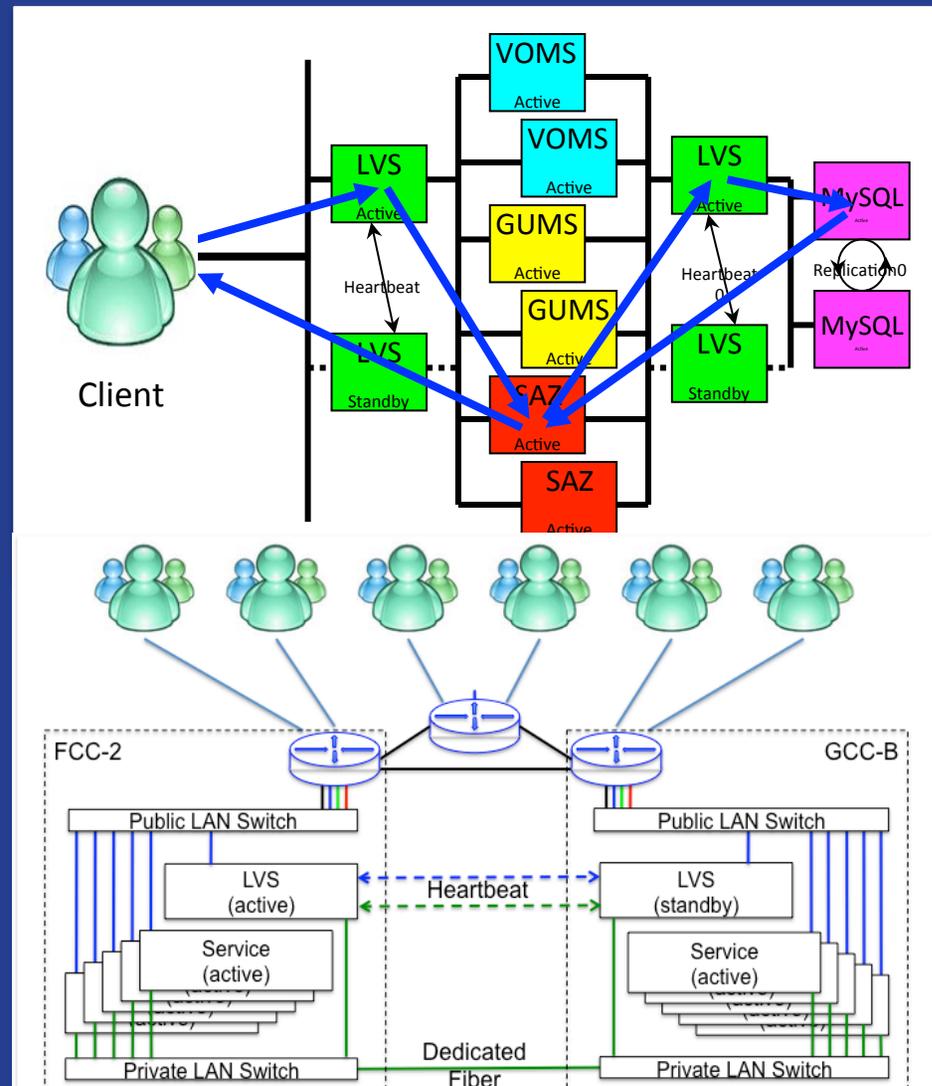
FermiGrid-HA uses three key technologies:

- Linux Virtual Server (LVS),
- Xen Hypervisor,
- MySQL Circular Replication.

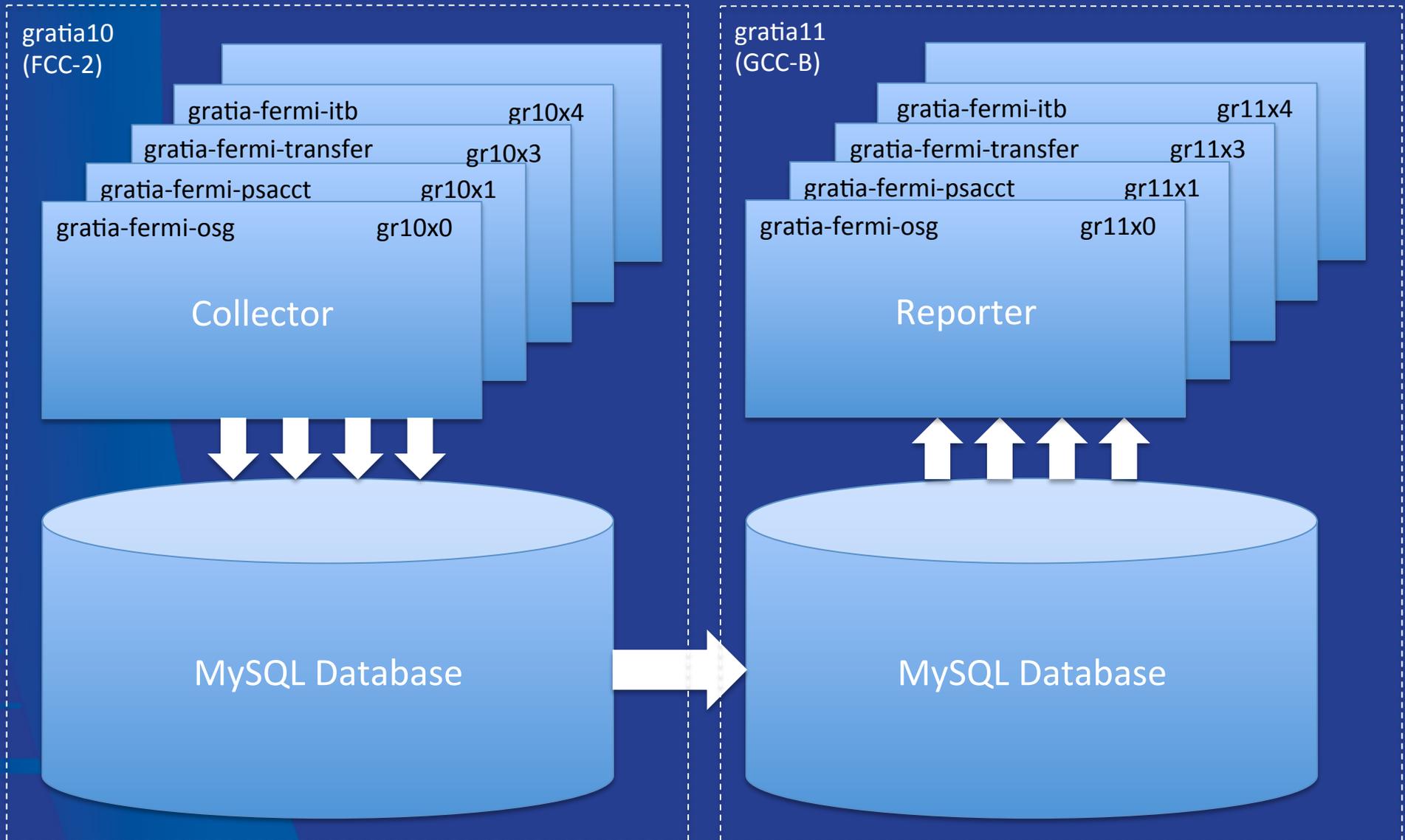
FermiGrid-HA2 added:

- Redundant services deployed across both FCC-2 and GCC-B computer rooms,
- Non-redundant services are split across both locations, and go to reduced capacity in the event of building or network outage.

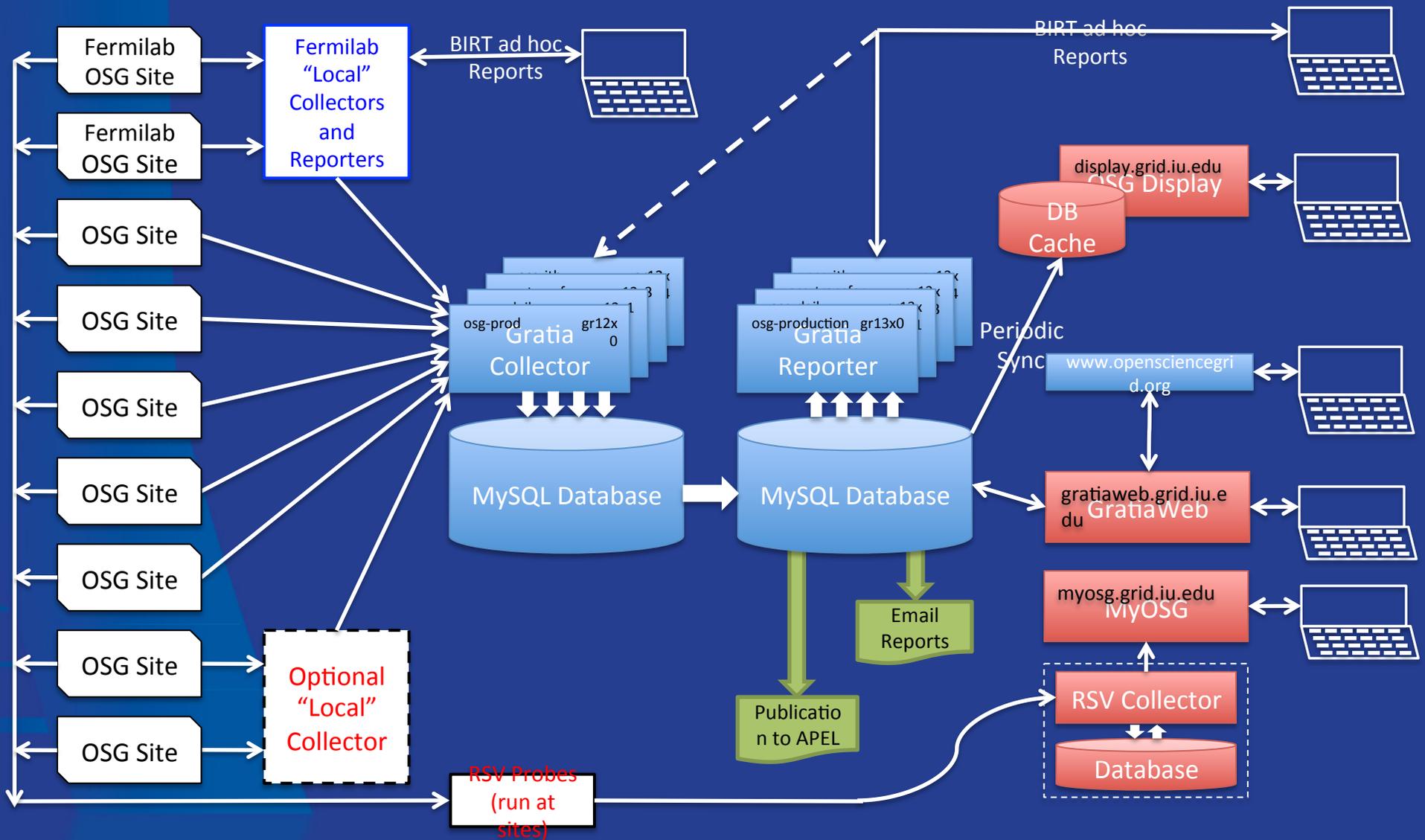
Deployment has been operated under real world conditions.



Fermilab Gratia Accounting



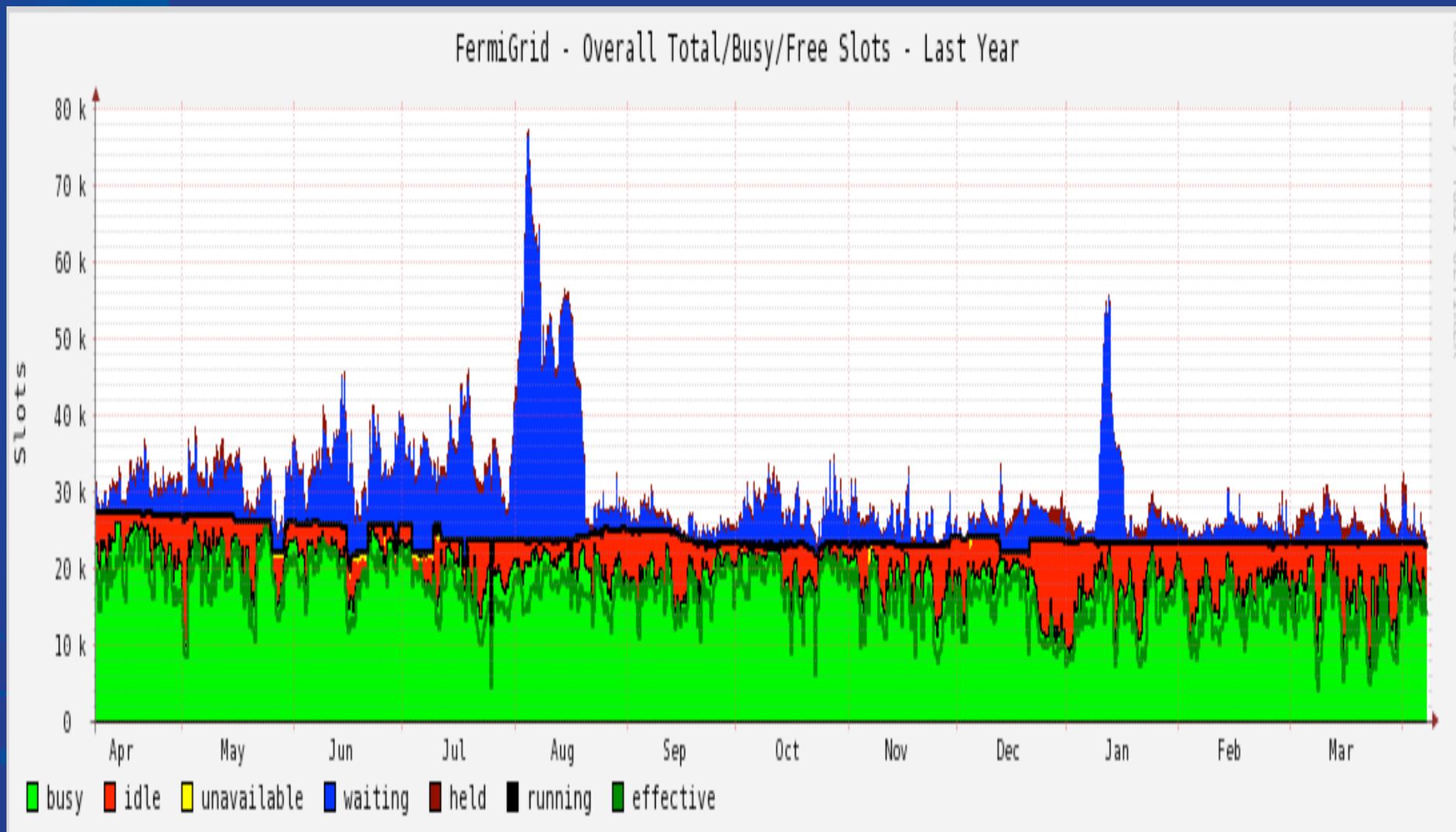
Gratia High Performance



Current Fermilab Campus Grid Statistics (as of April 2013)

Cluster(s)	Batch System	Job Slots	Raw Occupancy	Effective Utilization
CDF (Merged)	Condor	5268	81.8	70.0
CMS T1	Condor	6,272	90.3	85.5
D0 (Merged)	PBS	5,920	74.2	53.7
GP Grid	Condor	5,474	77.3	61.8
_____		_____	_____	_____
Overall-Today		22,934	80.4	68.3
Last Year		27,184	89.4	75.8

Fermilab Campus Grid



FermiGrid Service Availability (measured over the past year)

Service	Raw Availability	HA Configuration	Measured HA Availability	Minutes of Downtime
VOMS – VO Management Service	99.864	Active-Active	99.988%	60
GUMS – Grid User Mapping Service	99.863	Active-Active	100.000%	0
SAZ – Site AuthoriZation Service	99.863	Active-Active	100.000%	0
Squid – Web Cache	99.817	Active-Active	99.988%	60
MyProxy – Grid Proxy Service	99.781	Active-Standby	99.874%	660
ReSS – Resource Selection Service	99.915	Active-Active	99.988%	60
Gratia – Fermilab and OSG Accounting	99.229	Active-Standby	99.945%	300
MySQL Database	99.687	Active-Active	100.000%	0

Challenges for the next year

IPv6

- We hope to join the IPv6 testbed shortly...

SHA-1 => SHA-2

- Lots of middleware testing going on...

Merge FermiGrid Services into FermiCloud

- As well as other services...

FermiCloud

<http://fclweb.fnal.gov>

FermiCloud Background

Infrastructure-as-a-service facility for Fermilab employees, users, and collaborators,

- Project started in 2010,
- OpenNebula 2.0 cloud available to users since fall 2010,
- Condensed 7 racks of junk machines to 1.5 racks of good machines,
- Provider of integration and test machines to the OSG Software team,
- OpenNebula 3.2 cloud up since June 2012,

In a (very) broad brush, the mission of FermiCloud is:

- To deploy a production quality Infrastructure as a Service (IaaS) Cloud Computing capability in support of the Fermilab Scientific Program.
- To support additional IaaS, PaaS and SaaS Cloud Computing capabilities based on the FermiCloud infrastructure at Fermilab.

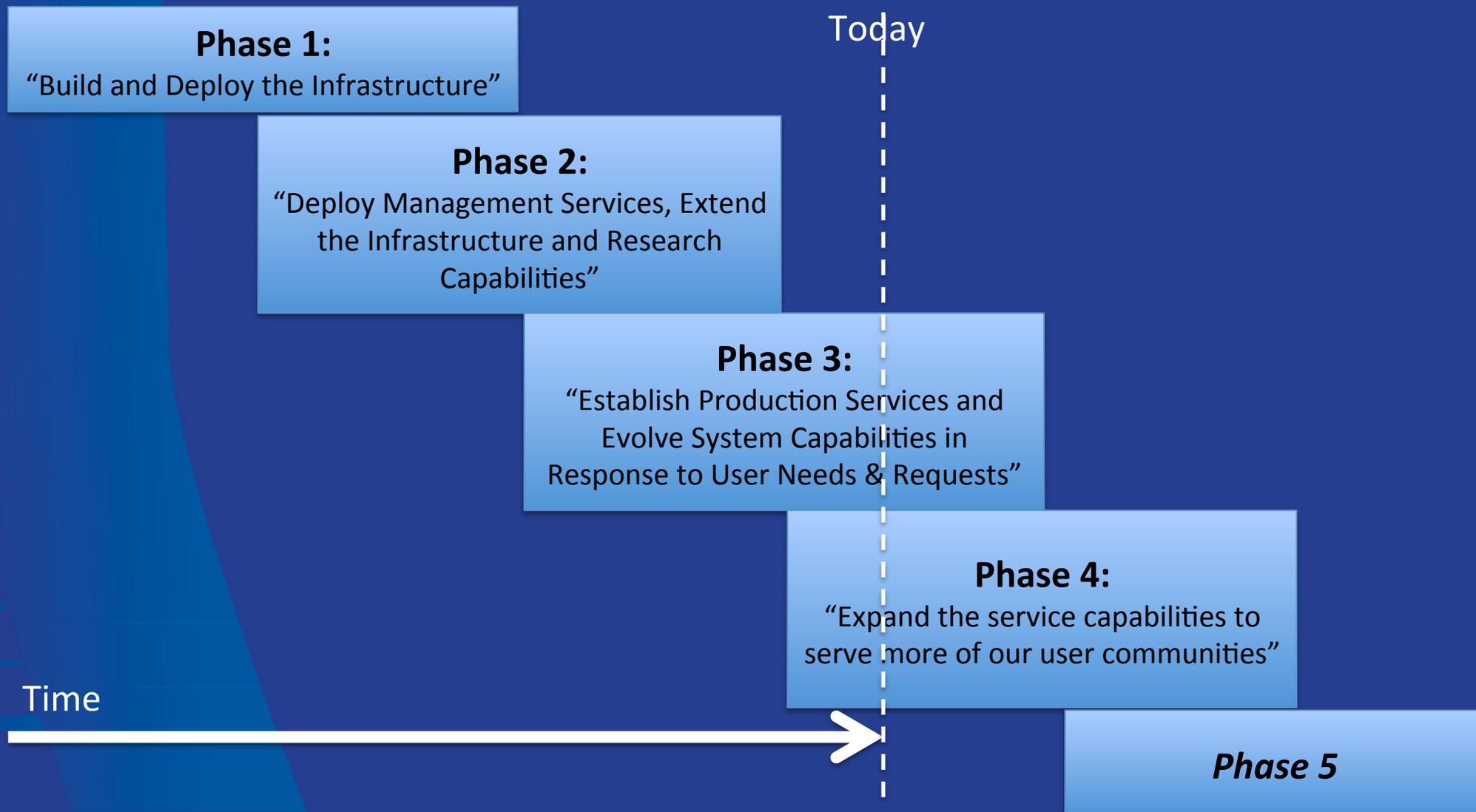
The FermiCloud project is split over several overlapping phases.

FermiCloud Development Goals

Goal: Make virtual machine-based workflows practical for scientific users:

- **Cloud bursting:** Send virtual machines from private cloud to commercial cloud if needed
- **Grid bursting:** Expand grid clusters to the cloud based on demand for batch jobs in the queue.
- **Federation:** Let a set of users operate between different clouds
- **Portability:** How to get virtual machines from desktop → FermiCloud → commercial cloud and back.
- **Fabric Studies:** enable access to hardware capabilities via virtualization (100G, Infiniband, ...)

FermiCloud Project Delivered Over Overlapping Phases



FermiCloud Phase 1:

“Build and Deploy the Infrastructure”

- Specify, acquire and deploy the FermiCloud hardware,
- Establish initial FermiCloud requirements and selected the “best” open source cloud computing framework that met these requirements (OpenNebula),
- Deploy capabilities to meet the needs of the stakeholders (JDEM analysis development, Grid Developers and Integration test stands, Storage/dCache Developers, LQCD testbed).

Completed late 2010

FermiCloud Phase 2: “Deploy Management Services, Extend the Infrastructure and Research Capabilities”

- Implement x509 based authentication (patches contributed back to OpenNebula project and are generally available in OpenNebula V3.2),
- Perform secure contextualization of virtual machines at launch,
- Perform virtualized filesystem I/O measurements,
- Develop (draft) economic model,
- Implement monitoring and accounting,
- Collaborate with KISTI personnel to demonstrate Grid and Cloud Bursting capabilities,
- Perform initial benchmarks of Virtualized MPI,
- Target “small” low-cpu-load servers such as Grid gatekeepers, forwarding nodes, small databases, monitoring, etc.,
- Begin the hardware deployment of a distributed SAN,
- Investigate automated provisioning mechanisms (puppet & cobbler).

completed late 2012

FermiCloud Phase 3:

“Establish Production Services and Evolve System Capabilities in Response to User Needs & Requests”

- Deploy highly available 24x7 production services,
 - Both infrastructure and user services.
- Deploy puppet & cobbler in production,
 - Done.
- Develop and deploy real idle machine detection,
 - Idle VM detection tool written by summer student.
- Research possibilities for a true multi-user filesystem on top of a distributed & replicated SAN,
 - CFS2 on FibreChannel SAN across FCC-A and FCC-B.
- Live migration becomes important for this phase.
 - Manual migration has been used, Live migration is currently in test, Automatically triggered live migration yet to come.
- Formal ITIL Change Management “Go-Live”,
 - Have been operating under “almost” ITIL Change Management for the past several months.

Approaching
Completion

FermiCloud Phase 4: “Expand the service capabilities to serve more of our user communities”

- Complete the deployment of the true multi-user filesystem on top of a distributed & replicated SAN,
- Demonstrate interoperability and federation:
 - Accepting VM's as batch jobs,
 - Interoperation with other Fermilab virtualization infrastructures (GPCF, VMware),
 - Interoperation with KISTI cloud, Nimbus, Amazon EC2, other community and commercial clouds,
- Participate in Fermilab 100 Gb/s network testbed.
 - Have just taken delivery of 10 Gbit/second cards
- Perform more “Virtualized MPI” benchmarks and run some real world scientific MPI codes,
 - The priority of this work will depend on finding a scientific stakeholder that is interested in this capability.
- Reevaluate available open source Cloud computing stacks,
 - Including OpenStack,
 - We will also reevaluate the latest versions of Eucalyptus, Nimbus and OpenNebula.

Starting Work ~ NOW

FermiCloud Phase 5

- Cover the program of work in the (draft) Fermilab-KISII FermiCloud Cooperative Research and Development Agreement (CRADA),
- We will likely start to move our production FermiGrid services deployment under the FermiCloud umbrella,
- This phase will likely involve work to enable the Fermilab CDF and D0 experiments Run II Data Preservation environment,
- This phase will also incorporate any work or changes that arise out of the Fermilab Scientific Computing Division strategy on Clouds and Virtualization,
- Additional requests and specifications are also being gathered!

Specifications
Under Development

Lessons from FermiGrid => FermiCloud must be Fault Tolerant

As we have learned from **FermiGrid**, having a distributed fault tolerant infrastructure is highly desirable for production operations.

Why:

- Users benefit from increased uptime,
- Service operators benefit from resiliency and gain flexibility to schedule routine maintenance.

How:

- Distribute physical systems across multiple buildings and computer rooms,
- Distributed and redundant "head nodes",
- Distributed "cloud services" across locations,
- Distributed network infrastructure,
- Heartbeat across public and private LAN to manage services,
- Distributed shared file system on SAN.

Goal:

- If a building is "lost", then automatically relaunch "24x7" VMs on surviving infrastructure, then relaunch "9x5" VMs if there is sufficient remaining capacity,
- Perform notification (via Service-Now) when exceptions are detected.

Some Recent Facility and/or Network Outages

FCC main breaker 4x (February–November 2010)

FCC-1 network outage 2x (Spring 2011)

GCC-B Load shed events (June–August 2011)

- This accelerated planned move of nodes to FCC-3.

GCC load shed events and maintenance (July 2012).

- FCC-3 cloud was ready just in time to keep server VM's up.

FCC-2 outage (October 2012)

- FermiCloud wasn't affected, our VM's stayed up.

Service Outages on Commercial Clouds

Amazon has had several significant service outages over the past few years:

- Outage in April 2011 of storage services resulted in actual data loss,
- An electrical storm that swept across the East Coast late Friday 29-Jun-20-12 knocked out power at a Virginia data center run by Amazon Web Services.
- An outage of one of Amazon's cloud computing data centers knocked out popular sites like Reddit, Foursquare Pinterest and TMZ on Monday 22-Oct-2012,
- Amazon outage affecting Netflix operations over Christmas 2012 and New Years 2013,
- Outage on Thursday 31-Jan-2013.

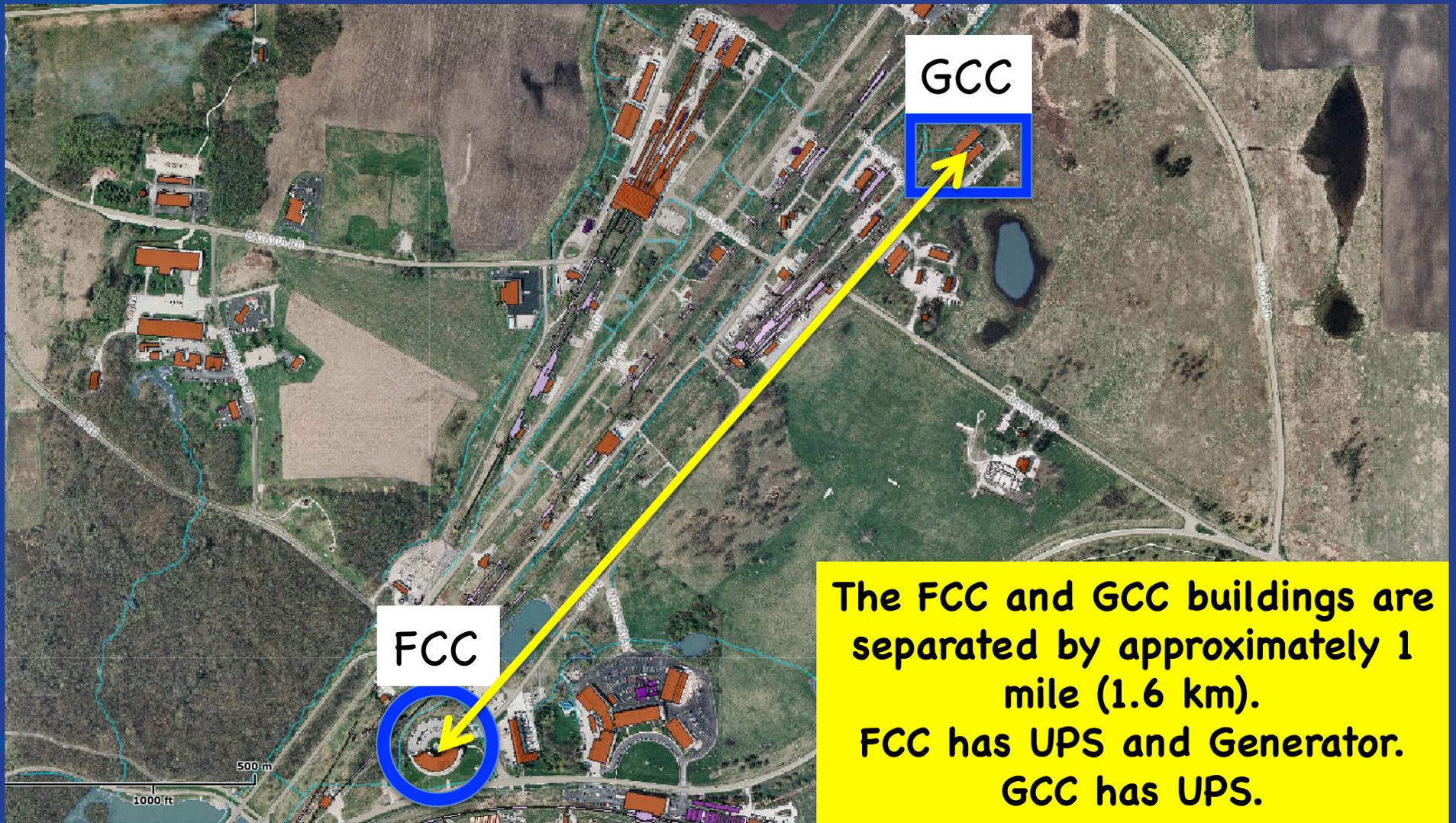
Microsoft Azure:

- Leap day bug on 29-Feb-2012.

Google:

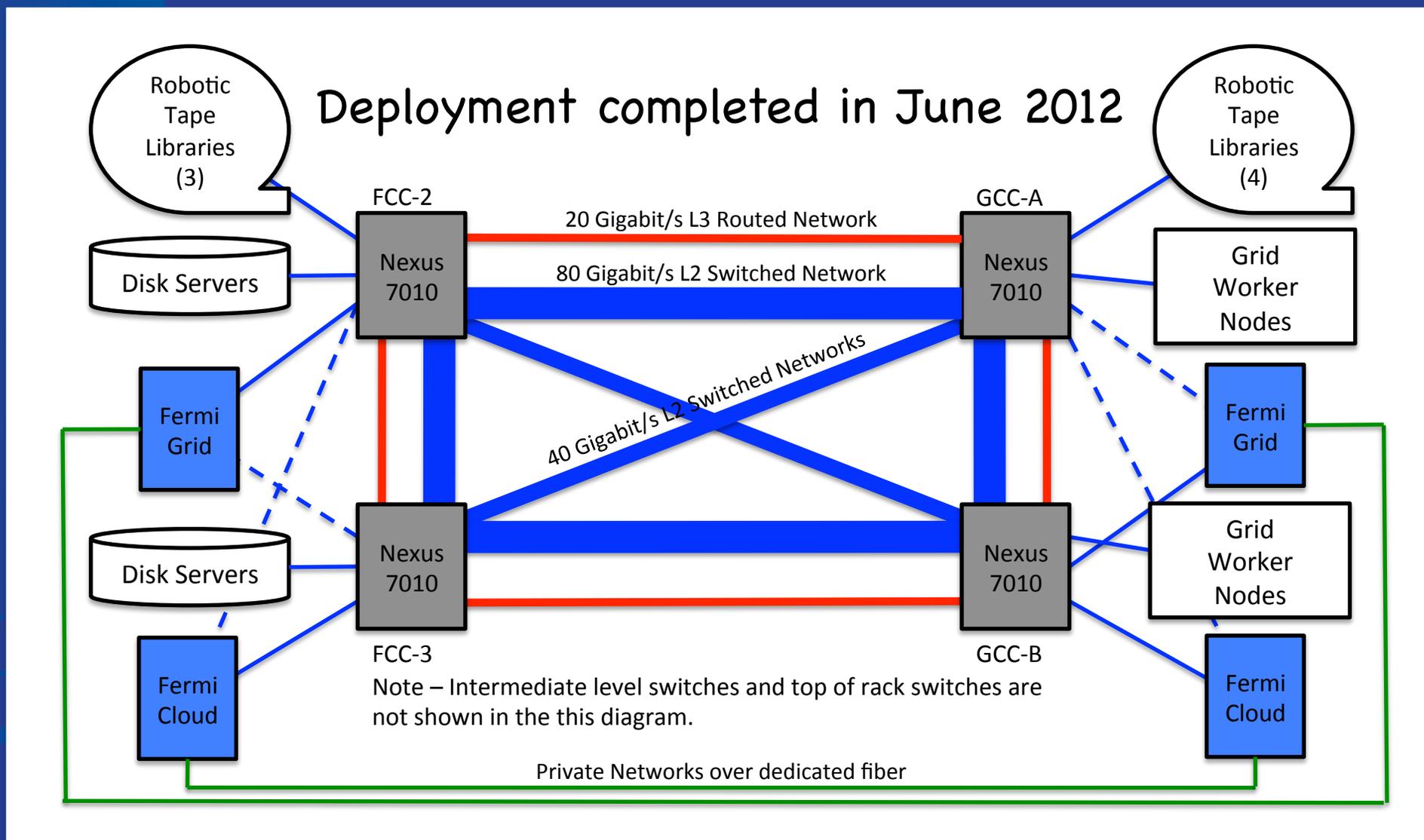
- Outage on 26-Oct-2012.

FCC and GCC



**The FCC and GCC buildings are separated by approximately 1 mile (1.6 km).
FCC has UPS and Generator.
GCC has UPS.**

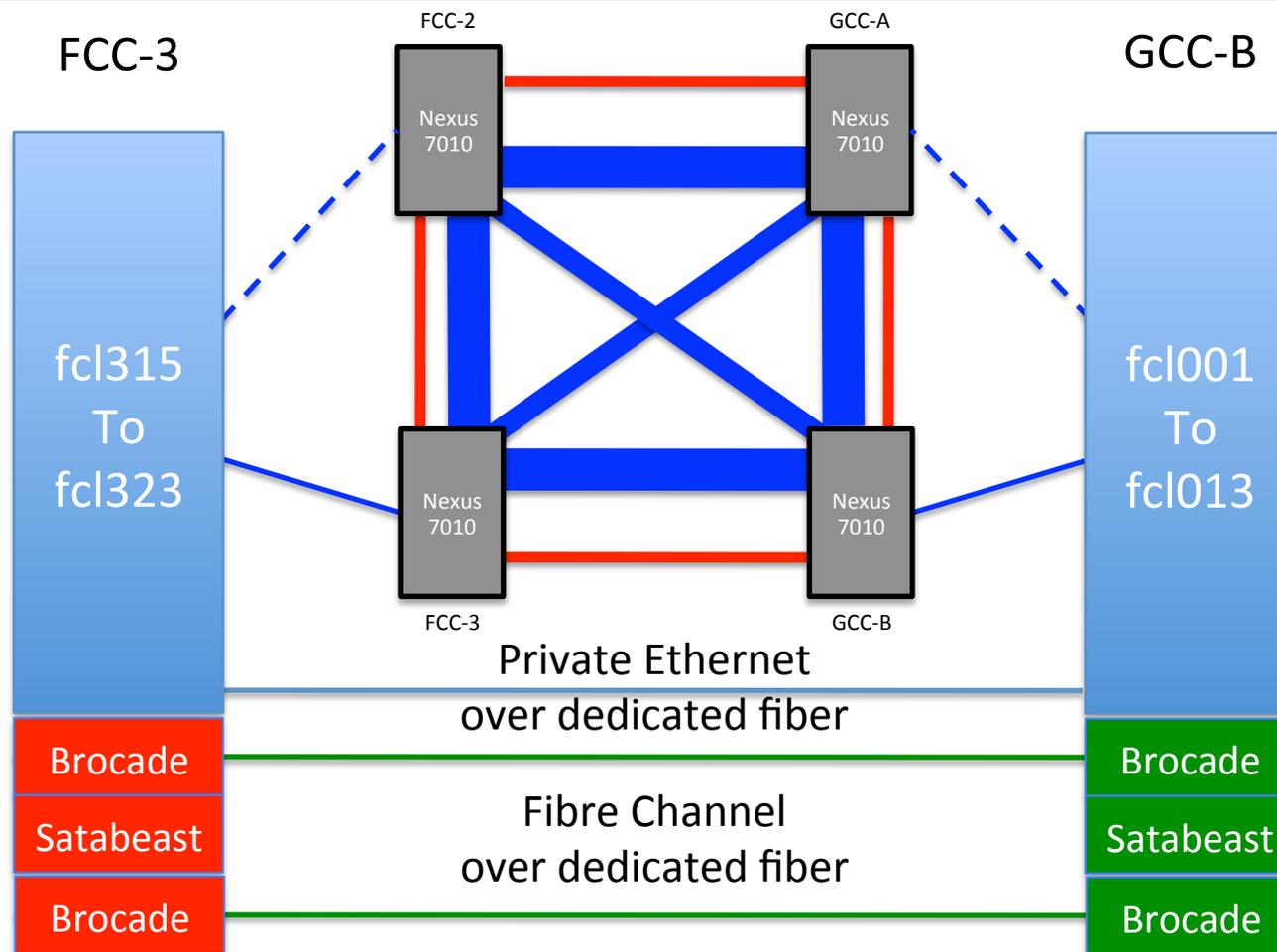
Distributed Network Core Provides Redundant Connectivity



Distributed Shared File System Design & Implementation

- Dual-port FibreChannel HBA in each node,
- Two Brocade SAN switches per rack,
- Brocades linked rack-to-rack with dedicated fiber,
- 60TB Nexsan Satabeast in FCC-3 and GCC-B,
- Redhat Clustering + CLVM + GFS2 used for file system,
- Each VM image is a file in the GFS2 file system,
- Next step—use LVM mirroring to do RAID 1 across buildings.
- Documented feature in manual, so this should be “easy” (for some definitions of “easy”).

FermiCloud – Network & SAN “Today”



FY2011 / FY2012

Distributed Shared File System Benefits

Fast Launch:

- Almost immediate as compared to 3-4 minutes with ssh/scp,

Live Migration:

- Can move virtual machines from one host to another for scheduled maintenance, transparent to users,

Persistent data volumes:

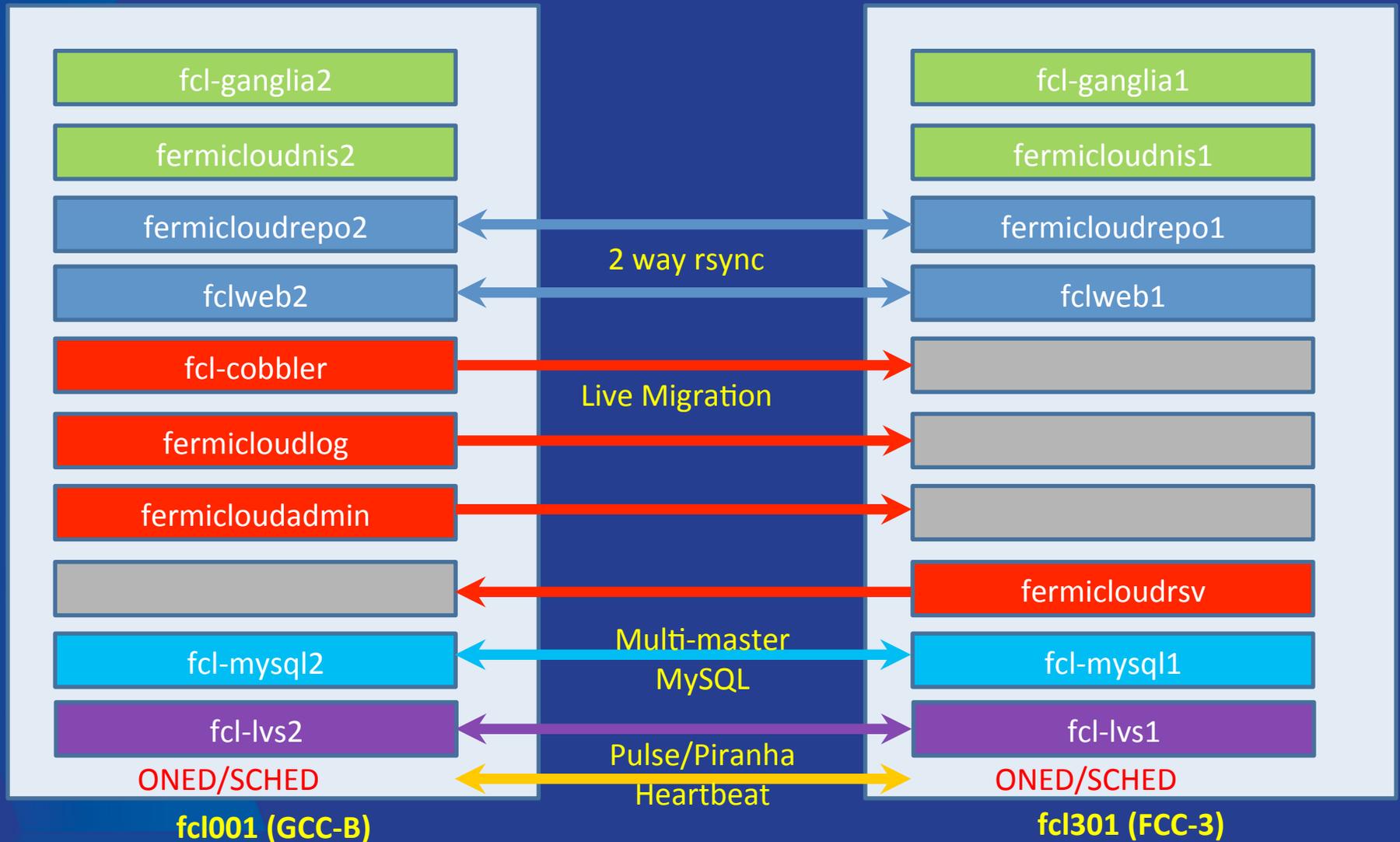
- Can move quickly with machines,

Once mirrored volume in place (soon):

- Can relaunch virtual machines in surviving building in case of building failure/outage,

Have recently moved the Head nodes from an independent shared file system based on GFS2/DRBD active/active to the SAN based GFS2.

FermiCloud-HA Head Node Configuration



Current FermiCloud Capabilities

Public network access via the high performance Fermilab network,

- This is a distributed, redundant network.

Private 1 Gb/sec network,

- This network is bridged across FCC and GCC on private fiber,

High performance Infiniband network to support HPC based calculations,

- Performance under virtualization is ~equivalent to "bare metal",
- Currently split into two segments,
- Segments could be bridged via Mellanox MetroX.

Access to a high performance FibreChannel based SAN,

- This SAN spans both buildings,

Access to the high performance BlueArc based filesystems,

- The BlueArc is located on FCC-2,

Access to the Fermilab dCache and enStore services,

- These services are split across FCC and GCC,

Access to 100 Gbit Ethernet test bed in LCC (via FermiCloud integration nodes),

- Intel 10 Gbit Ethernet converged network adapter X540-T1.

Who can use FermiCloud?

Any employee, user, or contractor of Fermilab with a current ID,

Most OSG staff have been able to get Fermilab "Offsite Only" ID's,

With Fermilab ID in hand, request FermiCloud login via Service Desk form,

Instructions on our new web page at:

- <http://fclweb.fnal.gov>

Note new web UI at:

- <https://fermicloud.fnal.gov:8443/>
- Doesn't work with Internet Explorer yet

Typical Use Cases

Public net virtual machine:

- On Fermilab Network open to Internet,
- Can access dCache and Bluearc Mass Storage,
- Shared Home directories between multiple VM's.

Public/Private Cluster:

- One gateway VM on public/private net,
- Cluster of many VM's on private net.

Storage VM:

- VM with large non-persistent storage,
- Use for large MySQL or Postgres databases, Lustre/Hadoop/Bestman/xRootd servers.

FermiCloud Draft Economic Model

SLA	24x7	9x5	Opportunistic
“Unit” (HT CPU + 2 GB)	\$125	\$45	\$25
Add'l HT core	\$125	\$125	\$125
Add'l memory per GB	\$30	\$30	\$30
Add'l local disk per TB	\$40	\$40	\$40
SAN disk per TB	\$475	\$475	\$475
BlueArc per TB	\$430	\$430	\$430
System Administrator	\$750	\$750	\$750
Specialized Service Support	“Market”	“Market”	“Market”

Note - Costs in the above chart are USD per year

Cloud Computing Environment

FermiCloud Security taskforce recommended to Fermilab Computer Security Board and the Computer Security Team that a new Cloud Computing Environment be established,

- This is currently under consideration.

Normal FermiCloud use is authenticated by Fermi MIT Kerberos credentials, or credentials (x.509) derived from Fermi MIT Kerberos.

Special concerns with Cloud:

- Users have root,
- Usage can be a combination of Grid usage (Open Science Environment) and Interactive usage (General Computing Environment),
- Planning for “secure cloud” to handle expected use cases: Archival systems at old patch levels or legacy OS, Data and code preservation systems, Non-baselined OS (Ubuntu, Centos, SUSE), Non-Kerberos services which can live only on private net,
- Cloud can incorporate non-Fermilab systems (Amazon, Microsoft Azure, etc.)

True Idle VM Detection

In times of resource need, we want the ability to suspend or “shelve” idle VMs in order to free up resources for higher priority usage.

- This is especially important in the event of constrained resources (e.g. during building or network failure).

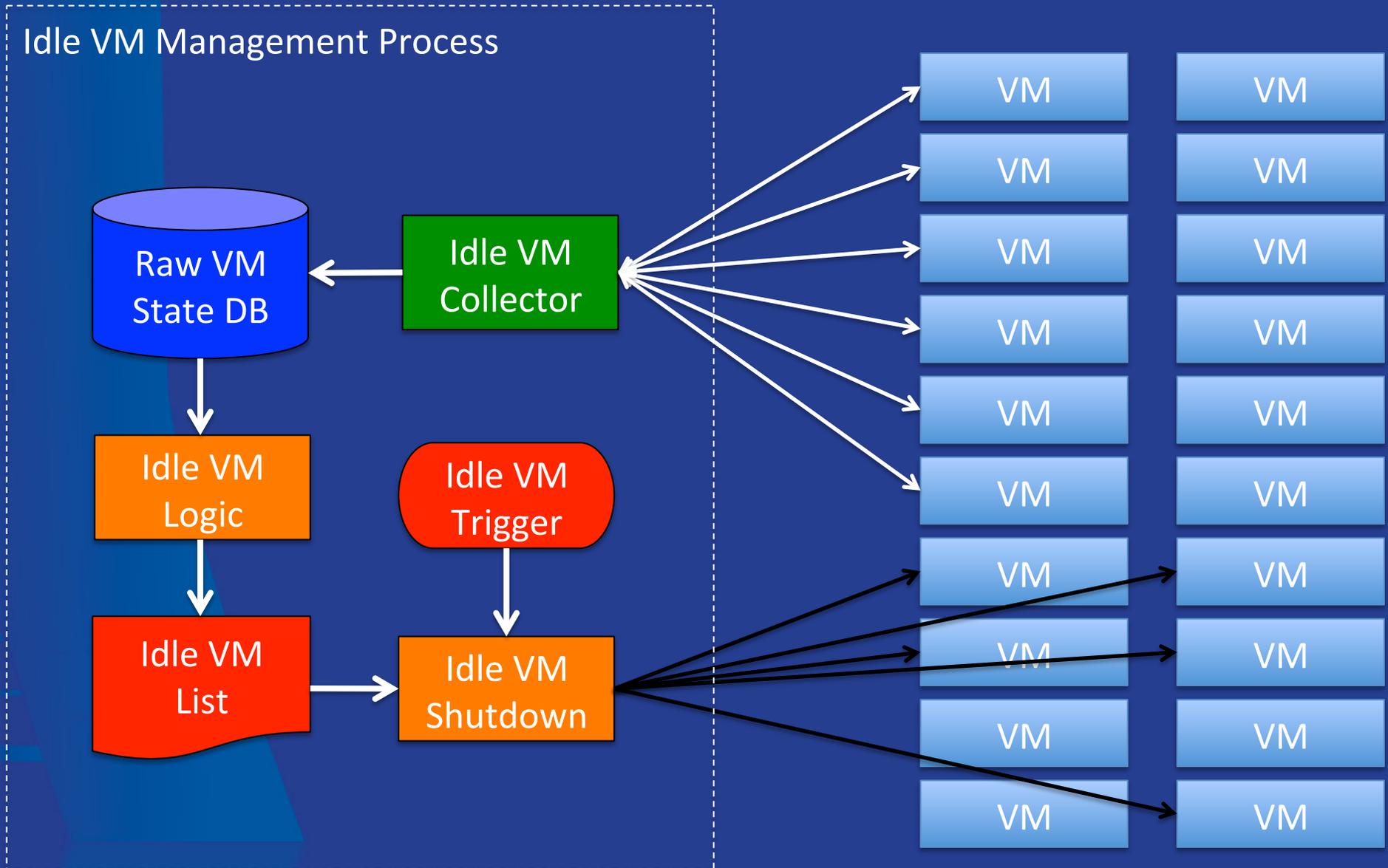
Shelving of “9x5” and “opportunistic” VMs allows us to use FermiCloud resources for Grid worker node VMs during nights and weekends

- This is part of the draft economic model.

Giovanni Franzini (an Italian co-op student) has written (extensible) code for an “Idle VM Probe” that can be used to detect idle virtual machines based on CPU, disk I/O and network I/O.

- This is the biggest pure coding task left in the FermiCloud project.

Idle VM Information Flow



VM Acceptance

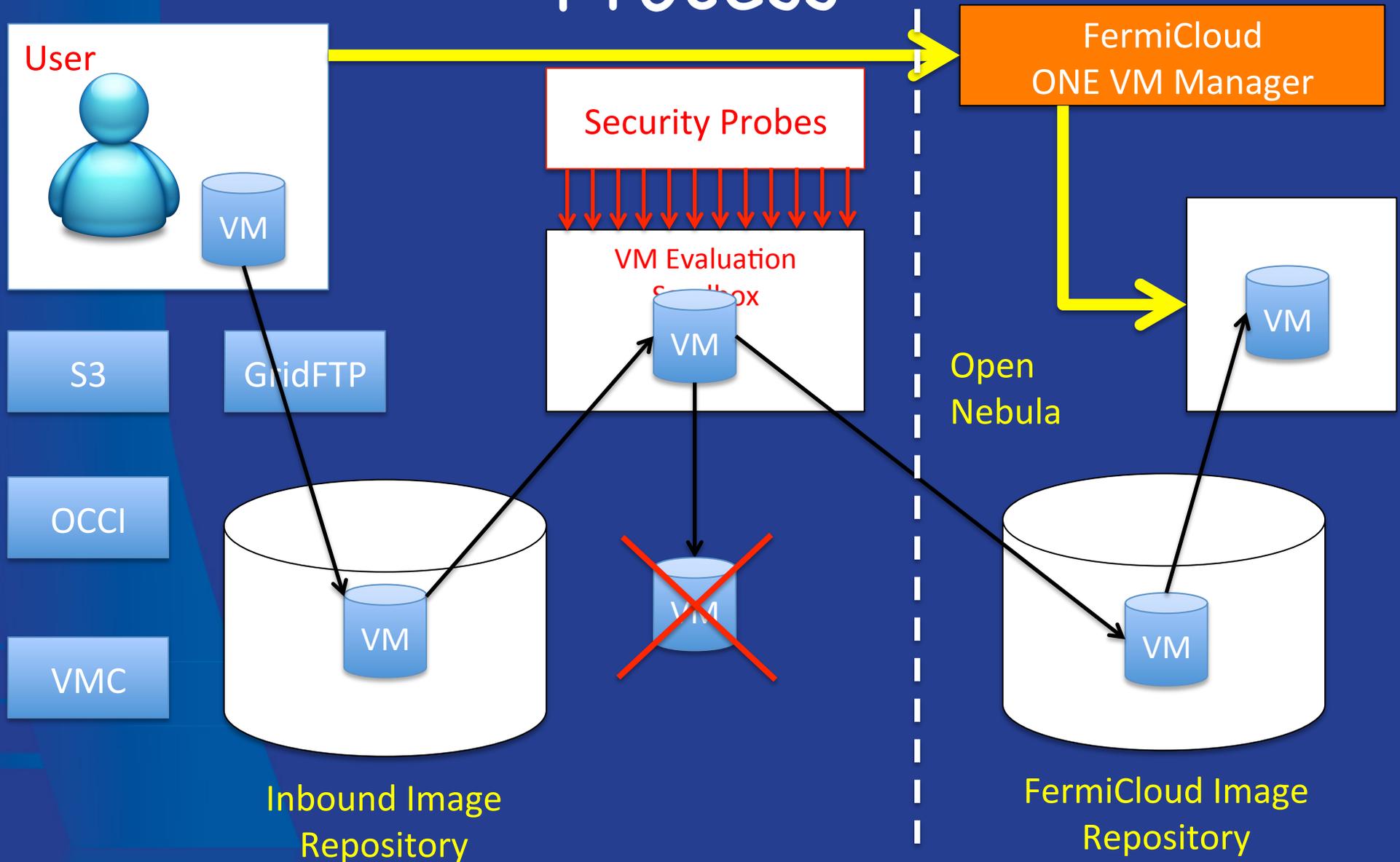
Fermilab does not accept that the VM signing/endorsing mechanisms recommended by the HEPiX Virtualization Working Group are sufficient.

Instead, we will:

- Keep logs of who transferred VM images to Fermilab,
- Keep logs of who requested VM images to be launched,
- Perform additional “pre-launch” security scanning (see next slide).

If any issues are identified, then we will contact the individuals who transferred the VM image or requested the VM launch.

Possible VM Image Acceptance Process



Virtual Infrastructure Automation and Provisioning

Demonstrate Interoperability and Federation:

- Accept Virtual Machines as Batch Jobs via cloud API's (such as native OpenNebula, Amazon EC-2 or OCCI),
- Test interoperability against other public (Amazon, Azure, etc.) and private (KISTI G-cloud, etc.) cloud computing stacks,
- Investigate interoperability with European efforts - EGI Cloud Federation Task Force (OpenNebula, OpenStack, StratusLab) - HelixNebula Federation of commercial cloud providers.

Grid and Cloud Bursting,

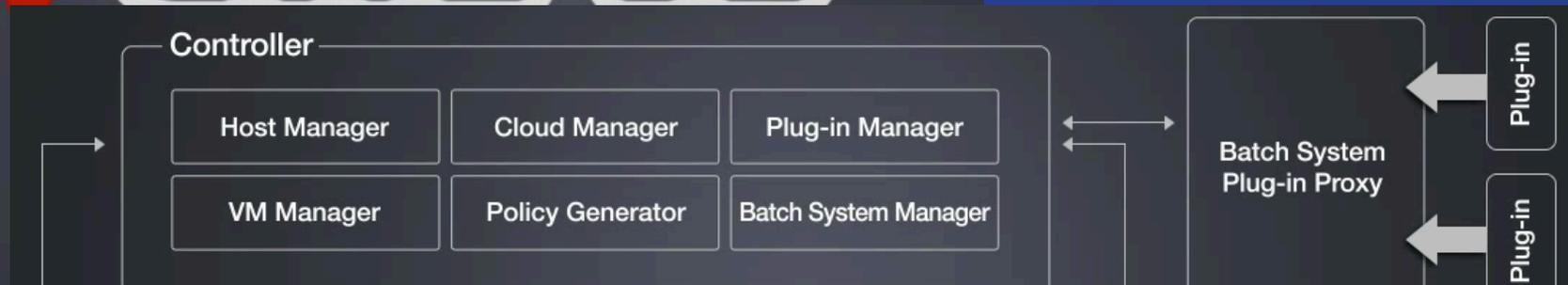
- Launch pre-defined user virtual machines based on workload demand,
- Use GlideinWMS to run Monte-Carlo production on public and private clouds.

Seo-Young Noh, KISTI visitor @ FNAL, showed proof-of-principle of "vCluster" in summer 2011:

- Look ahead at Condor batch queue,
- Submit worker node virtual machines of various VO's to FermiCloud or Amazon EC2 based on user demand,
- Machines join grid cluster and run grid jobs from the matching VO.

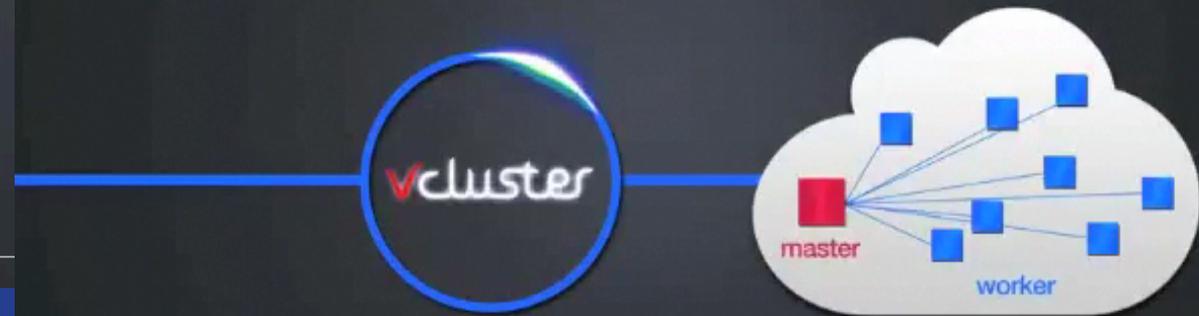
vCluster at SC2012

vcluster

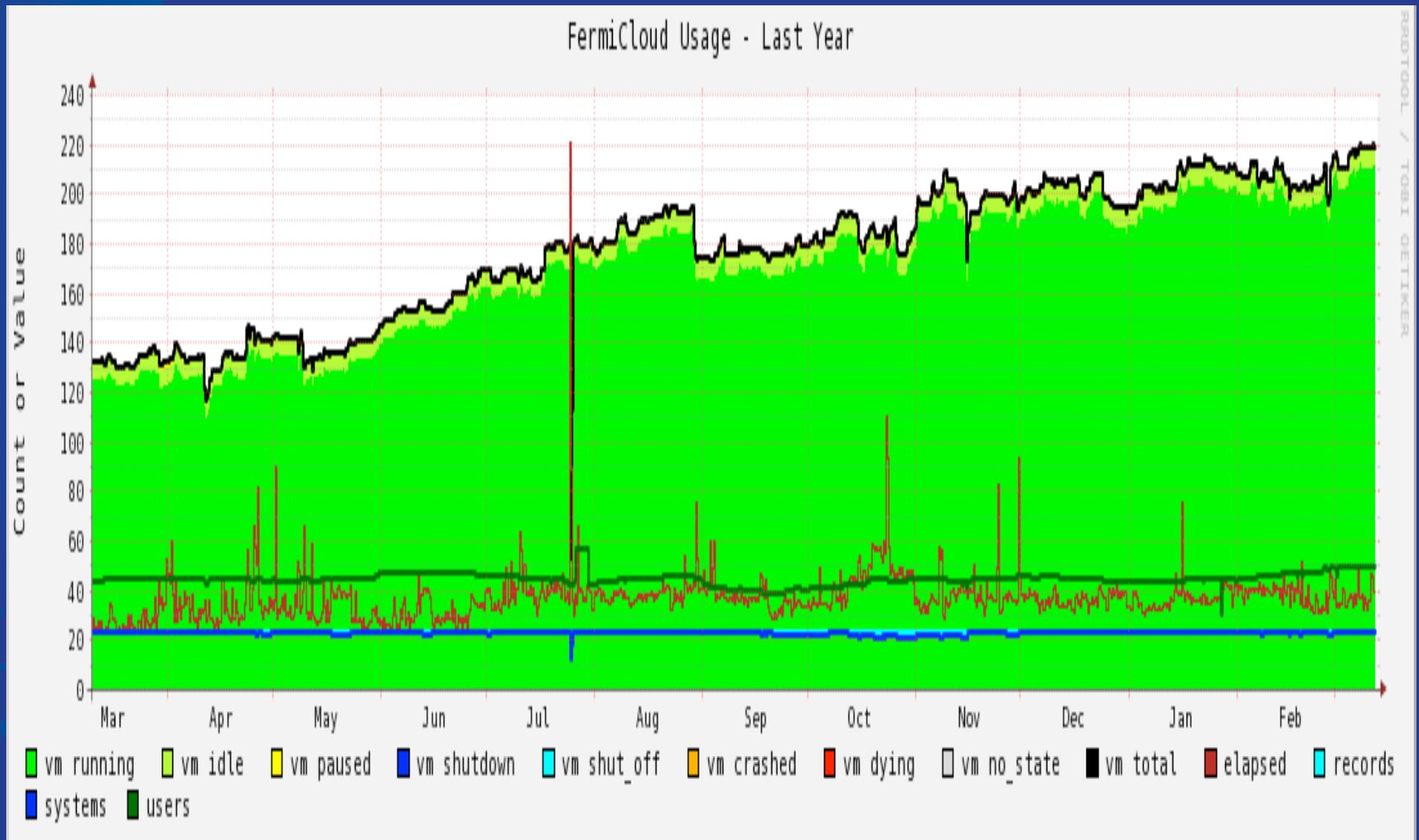


Auto Scalable on Demand

vcluster is automatically scalable. It regularly checks the queue and creates and removes virtual machines over multiple cloud systems.

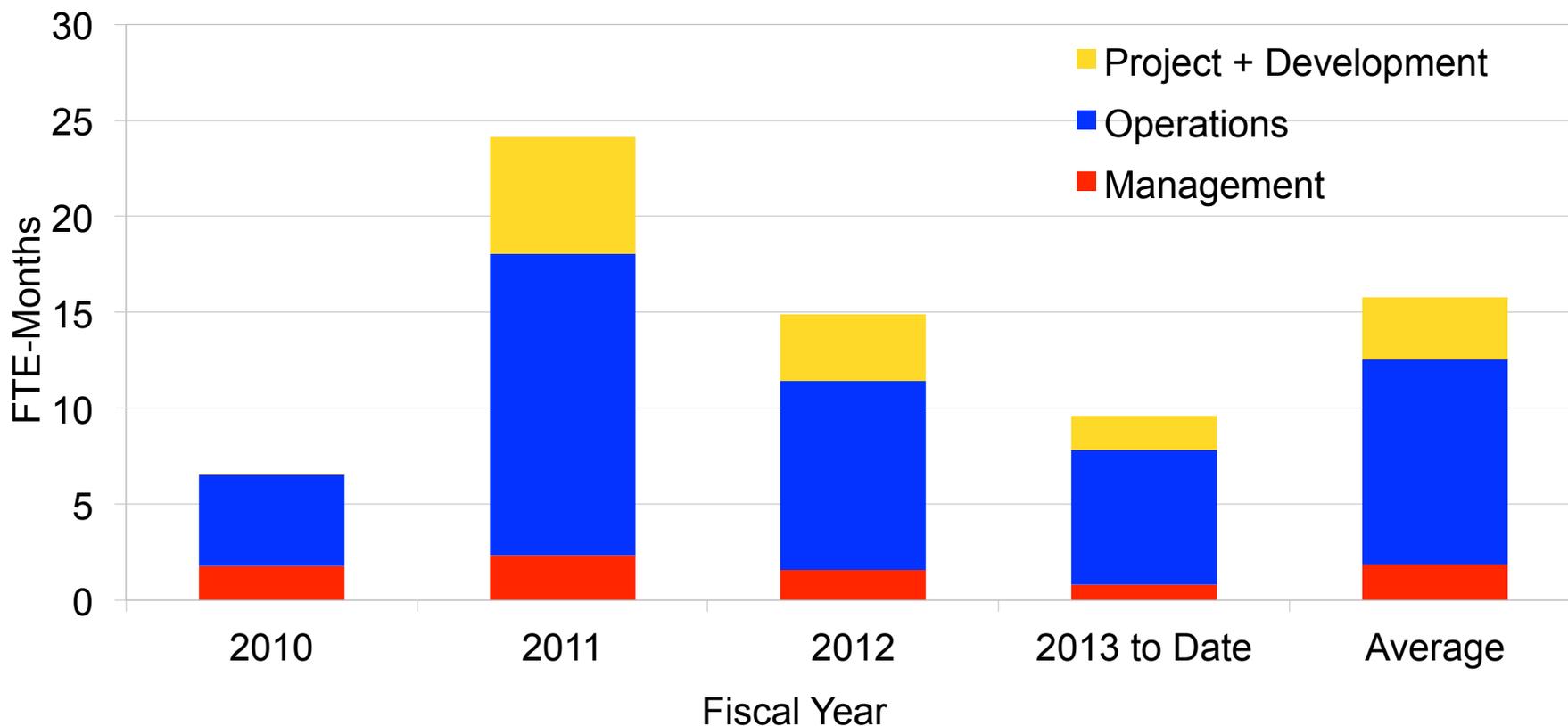


FermiCloud Usage



FermiCloud FTE Effort Plot (through Feb 2013)

FermiCloud Effort, FTE-Months per Year



FermiCloud Review

The FermiCloud Project recently underwent a management review that included both internal and external (to Fermilab) reviewers.

The review documents are available at:

- <https://indico.fnal.gov/conferenceDisplay.py?confId=6282>

Results of the review were very positive:

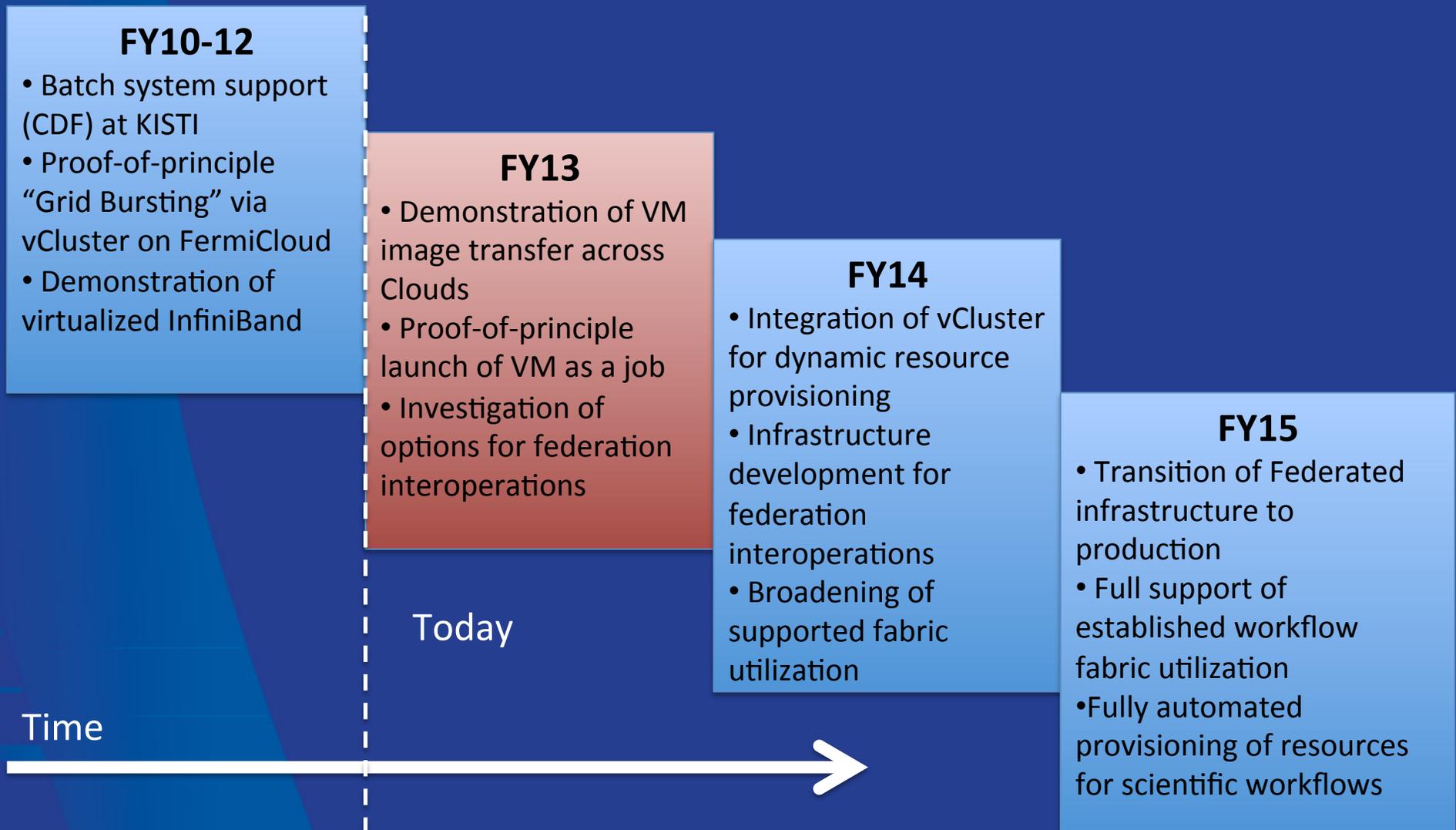
- Ready for production operations,
- Lots more work to do (see previous slides on project phases),
- Many items for senior line management to work on...

Current Stakeholders

Grid & Cloud Computing Personnel,
Run II – CDF & D0,
Intensity Frontier Experiments,
Cosmic Frontier (LSST),
Korea Institute of Science & Technology
Information (KISTI),
Open Science Grid (OSG).

Long Term Vision

[another look at some work in the FermiCloud Project Phases]



FermiCloud Summary - 1

Science is directly and indirectly benefiting from FermiCloud:

- CDF, D0, Intensity Frontier, Cosmic Frontier, CMS, ATLAS, Open Science Grid,...

FermiCloud operates at the forefront of delivering cloud computing capabilities to support scientific research:

- By starting small, developing a list of requirements, building on existing Grid knowledge and infrastructure to address those requirements, FermiCloud has managed to deliver a production class Infrastructure as a Service cloud computing capability that supports science at Fermilab.
- FermiCloud has provided FermiGrid with an infrastructure that has allowed us to test Grid middleware at production scale prior to deployment.
- The Open Science Grid software team used FermiCloud resources to support their RPM "refactoring" and is currently using it to support their ongoing middleware development/integration.

FermiCloud Summary – 2

The FermiCloud collaboration with KISTI has leveraged the resources and expertise of both institutions to achieve significant benefits.

vCluster has demonstrated proof of principal “Grid Bursting” using FermiCloud and Amazon EC2 resources.

Using SRIOV drivers on FermiCloud virtual machines, MPI performance has been demonstrated to be **>96%** of the native “bare metal” performance.

FermiCloud Project Summary - 3

FermiCloud personnel are currently working on:

- Finalizing a CRADA with KISTI for future collaboration,
- Completion of the Phase 3 deliverables, including a SAN storage deployment that will offer a true multi-user filesystem on top of a distributed & replicated SAN,
- Getting ready to start work on the Phase 4 deliverables,
- Collaborating on the development of the Phase 5 specifications.

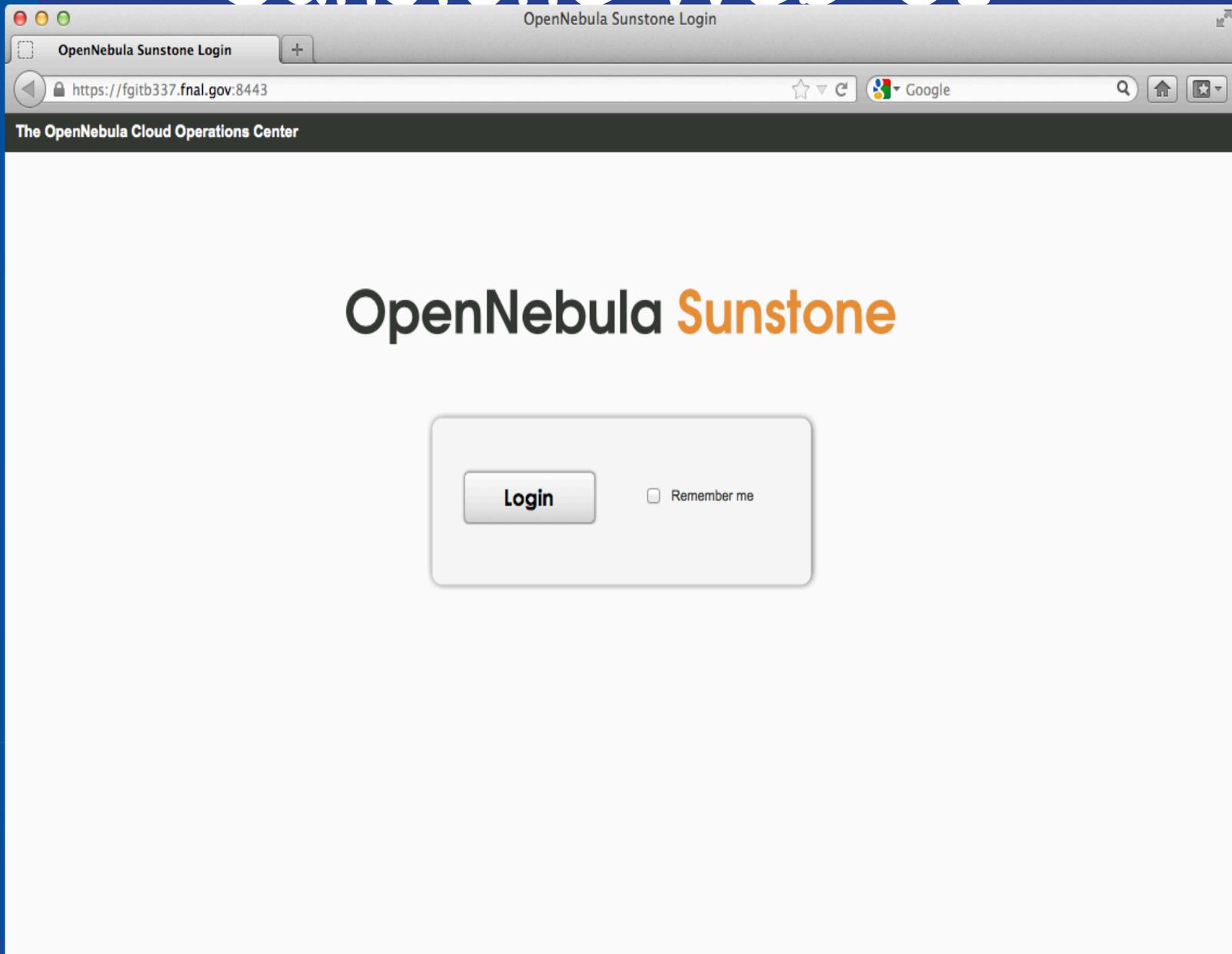
The future is mostly cloudy.

Thank You

Any Questions?

Extra Slides

Sunstone Web UI



Selecting a template

OpenNebula Sunstone Documentation | Support | Community Welcome timm2 | Sign out

+ New Update properties Instantiate Previous action Delete

Show 10 entries Search:

<input type="checkbox"/> All	ID	Owner	Group	Name	Registration time
<input type="checkbox"/>	50	oneadmin	oneadmin	SLF6 New Base	16:16:25 11/26/2012
<input type="checkbox"/>	53	oneadmin	oneadmin	SCT-slf58	22:31:08 11/28/2012

Showing 1 to 2 of 2 entries First Previous 1 Next Last

Copyright 2002-2012 © OpenNebula Project Leads (OpenNebula.org). All Rights Reserved. OpenNebula 3.2.1

Launching the Virtual Machine

The screenshot displays the OpenNebula Sunstone Cloud Operations Center interface. The browser address bar shows the URL <https://fgitb337.fnal.gov:8443>. The page title is "OpenNebula Sunstone: Cloud Operations Center". The navigation bar includes "OpenNebula Sunstone", "Documentation | Support | Community", and "Welcome hyunwoo | Sign out".

The left sidebar contains the following menu items: Dashboard, Virtual Machines (highlighted), Templates, Virtual Networks, Images, and Configuration.

The main content area shows a table of virtual machines. The table has the following columns: All, ID, Owner, Group, Name, Status, CPU, Memory, Hostname, Start Time, and VNC Access. A single entry is visible:

All	ID	Owner	Group	Name	Status	CPU	Memory	Hostname	Start Time	VNC Access
<input type="checkbox"/>	119	hyunwoo	users	one-119	PROLOG	0	0K	fgitb338	14:04:03 10/16/2012	

Below the table, it says "Showing 1 to 1 of 1 entries". Navigation controls include "First", "Previous", "1", "Next", and "Last".

At the top right of the main content area, there are buttons for "+ New", "Update properties", "Shutdown", "Previous action", and "Delete".

Copyright 2002-2012 © OpenNebula Project Leads (OpenNebula.org). All Rights Reserved. OpenNebula 3.2.1

Monitoring VM's

OpenNebula Sunstone Documentation | Support | Community **Welcome timm2 | Sign out**

Dashboard

- Virtual Machines
- Templates
- Virtual Networks
- Images
- Configuration

Summary of resources

VM Templates (total/public)	3 / 0
VM Instances (total/running/failed)	0 / 0 / 0
Virtual Networks (total/public)	1 / 0
Images (total/public)	4 / 0

Quickstart

New:

- VM Template
- VM Instance
- Virtual Network

Image

Historical monitoring information

Total VM count

total active error

Time	total	active	error
13:10:00	3	0	0
13:13:20	3	0	0
13:16:40	3	0	0
13:20:00	3	0	0
13:23:20	3	0	0
13:26:40	3	0	0

Total VM CPU

cpu_usage

Time	cpu_usage
13:10:00	273.000
13:13:20	273.000
13:16:40	273.000
13:20:00	273.000
13:23:20	273.000
13:26:40	273.000

Total VM Memory

mem_usage

Time	mem_usage
13:10:00	64.7G
13:13:20	64.7G
13:16:40	64.7G
13:20:00	64.8G
13:23:20	64.8G
13:26:40	64.8G

VM Network stats

net_tx net_rx

Time	net_tx	net_rx
13:10:00	23.3T	23.3T
13:13:20	23.3T	23.3T
13:16:40	23.3T	23.3T
13:20:00	23.3T	23.3T
13:23:20	23.3T	23.3T
13:26:40	23.3T	23.3T

Virtual Machines as Jobs

OpenNebula (and all other open-source IaaS stacks) provide an emulation of Amazon EC2.

Condor team has added code to their "Amazon EC2" universe to support the X.509-authenticated protocol.

Planned use case for GlideinWMS to run Monte Carlo on clouds public and private.

Feature already exists,

- this is a testing/integration task only.

Grid Bursting

Seo-Young Noh, KISTI visitor @ FNAL, showed proof-of-principle of "vCluster" in summer 2011:

- Look ahead at Condor batch queue,
- Submit worker node virtual machines of various VO's to FermiCloud or Amazon EC2 based on user demand,
- Machines join grid cluster and run grid jobs from the matching VO.

Need to strengthen proof-of-principle, then make cloud slots available to FermiGrid.

Several other institutions have expressed interest in extending vCluster to other batch systems such as Grid Engine.

KISTI staff has a program of work for the development of vCluster.

GlideinWMS project has significant experience submitting worker node virtual machines to cloud. In discussions to collaborate.

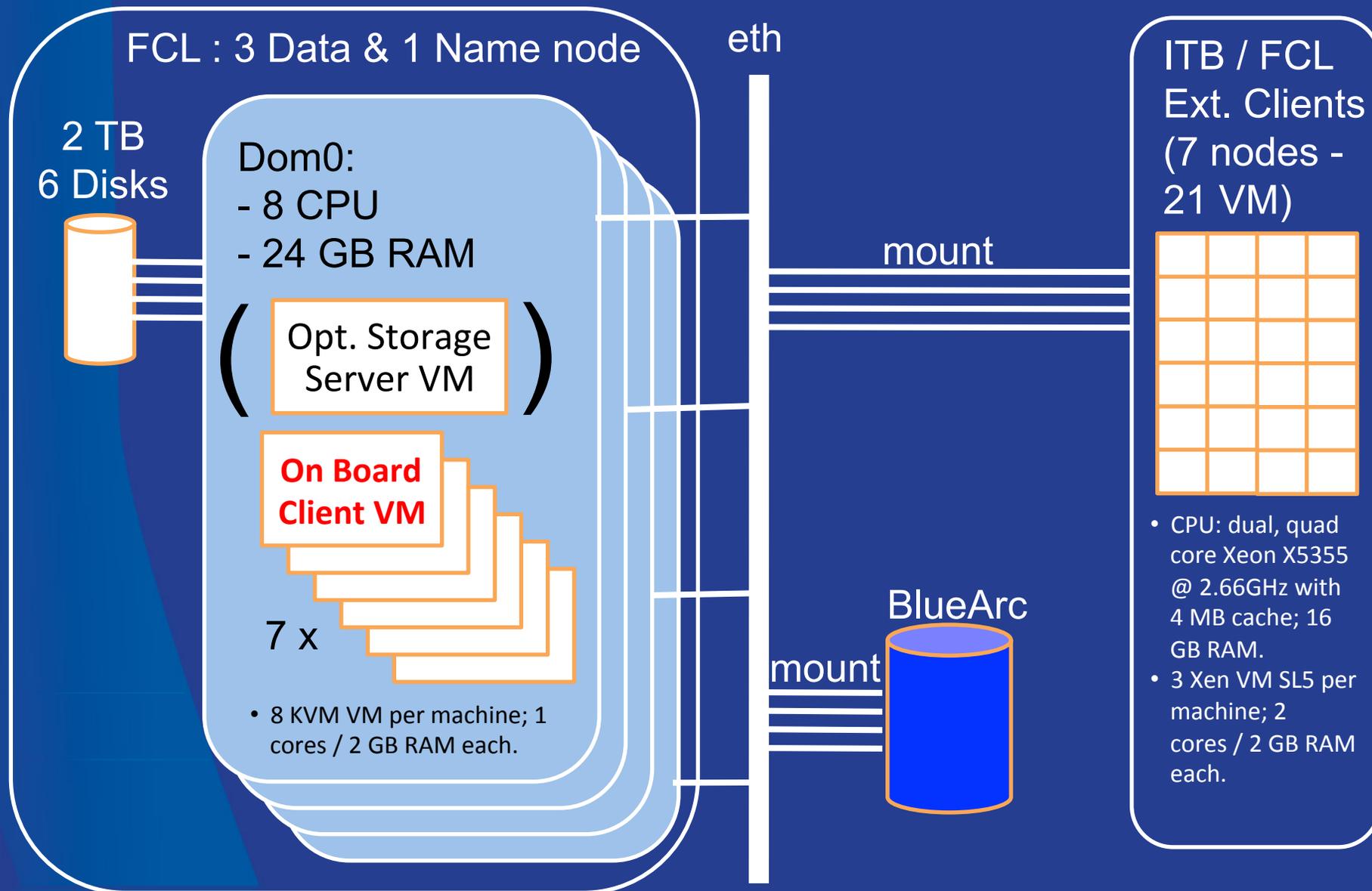
Cloud Bursting

OpenNebula already has built-in “Cloud Bursting” feature to send machines to Amazon EC2 if the OpenNebula private cloud is full.

Need to evaluate/test it, see if it meets our technical and business requirements, or if something else is necessary.

Need to test interoperability against other stacks.

FermiCloud Test Bed - Virtualized Server



Federation

Driver:

- Global scientific collaborations such as LHC experiments will have to interoperate across facilities with heterogeneous cloud infrastructure.

European efforts:

- EGI Cloud Federation Task Force – several institutional clouds (OpenNebula, OpenStack, StratusLab).
- HelixNebula—Federation of commercial cloud providers

Our goals:

- Show proof of principle—Federation including FermiCloud + KISTI “G Cloud” + one or more commercial cloud providers + other research institution community clouds if possible.
- Participate in existing federations if possible.

Core Competency:

- FermiCloud project can contribute to these cloud federations given our expertise in X.509 Authentication and Authorization, and our long experience in grid federation

Virtual Image Formats and Distribution

Clouds have different VM image formats:

- FS, Partition table, LVM , Kernel

Identify differences, find conversion tools

Investigate image marketplaces (Hepix, UVic)

Do we need S3 or Gridftp image upload facility?

- OpenNebula doesn't have one now.

Develop auto security scan for VM images

- Scan them like laptop coming onto site.

Interoperability/Compatibility of API's

Amazon EC2 API is not open source, it is a moving target that changes frequently.

Open-source emulations have various feature levels and accuracy of implementation:

- Compare and contrast OpenNebula, OpenStack, and commercial clouds,
- Identify lowest common denominator(s) that work on all.
- Contribute bug reports and fixes where possible.

High-Throughput Fabric Virtualization

Follow up earlier virtualized MPI work:

- Use it in real scientific workflows
- Example – simulation of data acquisition systems (the existing FermiCloud Infiniband fabric has already been used for such).

Will also use FermiCloud machines on 100Gbit Ethernet test bed

- Evaluate / optimize virtualization of 10G NIC for the use case of HEP data management applications
- Compare and contrast against Infiniband

FermiCloud X.509 Authentication

OpenNebula came with “pluggable” authentication, but few plugins initially available.

OpenNebula 2.0 Web services by default used “access key” / “secret key” mechanism similar to Amazon EC2. No https available.

Four ways to access OpenNebula:

- Command line tools,
- Sunstone Web GUI,
- “ECONE” web service emulation of Amazon Restful (Query) API,
- OCCI web service.

FermiCloud project wrote X.509-based authentication plugins:

- Available in OpenNebula 3.2 and onward.
- X.509 plugins available for command line and for web services authentication

FermiCloud X.509 Authorization

OpenNebula authorization plugins written in Ruby
As shipped: One DN → one userid—no VO support

- For now, must use Fermilab KCA short-lived cert.

Use existing Grid AuthZ routines:

- Use Ruby-Java bridge to call Java-based routines from VO Privilege project
- GUMS returns uid/gid, SAZ returns yes/no.
- Works with OpenNebula command line and non-interactive web services

Almost impossible to send VOMS proxy through web browser

- For OpenNebula's Sunstone Web GUI develop VO selection pull-down (similar to Gratia Admin Web UI)
- Get desired vo/role from user and poll GUMS/VOMS on server side

FermiCloud User Comments – 1

Bothered by no-root-squash NFS

- FermiCloud stores all host/http certs in volatile RAM disk for just that reason.

Clock problem with pause/resume

- FC admins solve that by launching a process to restart ntpd on a loop just before we pause. Will try to make this available to users.

Way for user to intervene on down VM

- We as admins can get serial console or access via virt-viewer

Can't save VM snapshot

- `Onevm saveas <vmid> 0 <name_of_image> ; onevm shutdown <vmid>`

IP finding tool needed

- `source /cloud/images/OpenNebula/scripts/one3.2/hostname.sh`
- Other OSG software team have written scripts too

FermiCloud User Comments – 2

- Change name of VM
 - Have asked OpenNebula for this feature.
- Community repo for users to share scripts
 - OSG SW team already added
- Better docs on best practices
 - Next version of OpenNebula expected to beat a lot of race conditions
 - Until then—we will modify onevm delete so it will not delete a VM still in “shut” or “epil” state
 - And add an “are you sure?” prompt for VM in “run” state.
- Make docs available as HTML pages, not word docs
 - See <http://fclweb.fnal.gov/fermicloud-dummies.html>
 - And <http://fclweb.fnal.gov/fermicloud-geeks.html>