FermiCloud Update

Keith Chadwick
Fermilab

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

Outline

- Personnel
- Mission, Strategy, Project, Phases
- Hardware and VM Specifications
- Economic Model
- Authentication, Contextualization
- Fault Tolerance
- Monitoring, Accounting,
- Virtualized File Service,
- Grid and Cloud "Bursting"
- Virtualized MPI
- Stakeholders
- Summary

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
Douglas Strain

FermiGrid Services

Steven C. Timm (Leader)
Hyunwoo Kim (Visitor - KISTI)
~~Faarooq Lowe (departed 30 Mar)~~
~~Seo-Young Noh (Visitor - KISTI)~~
Neha Sharma
Daniel R. Yocum

FermiCloud – Mission, Strategy, Goals & History

- As part of the FY2010 activities, the (then) Grid Facilities Department established a project to implement an initial "FermiCloud" capability.

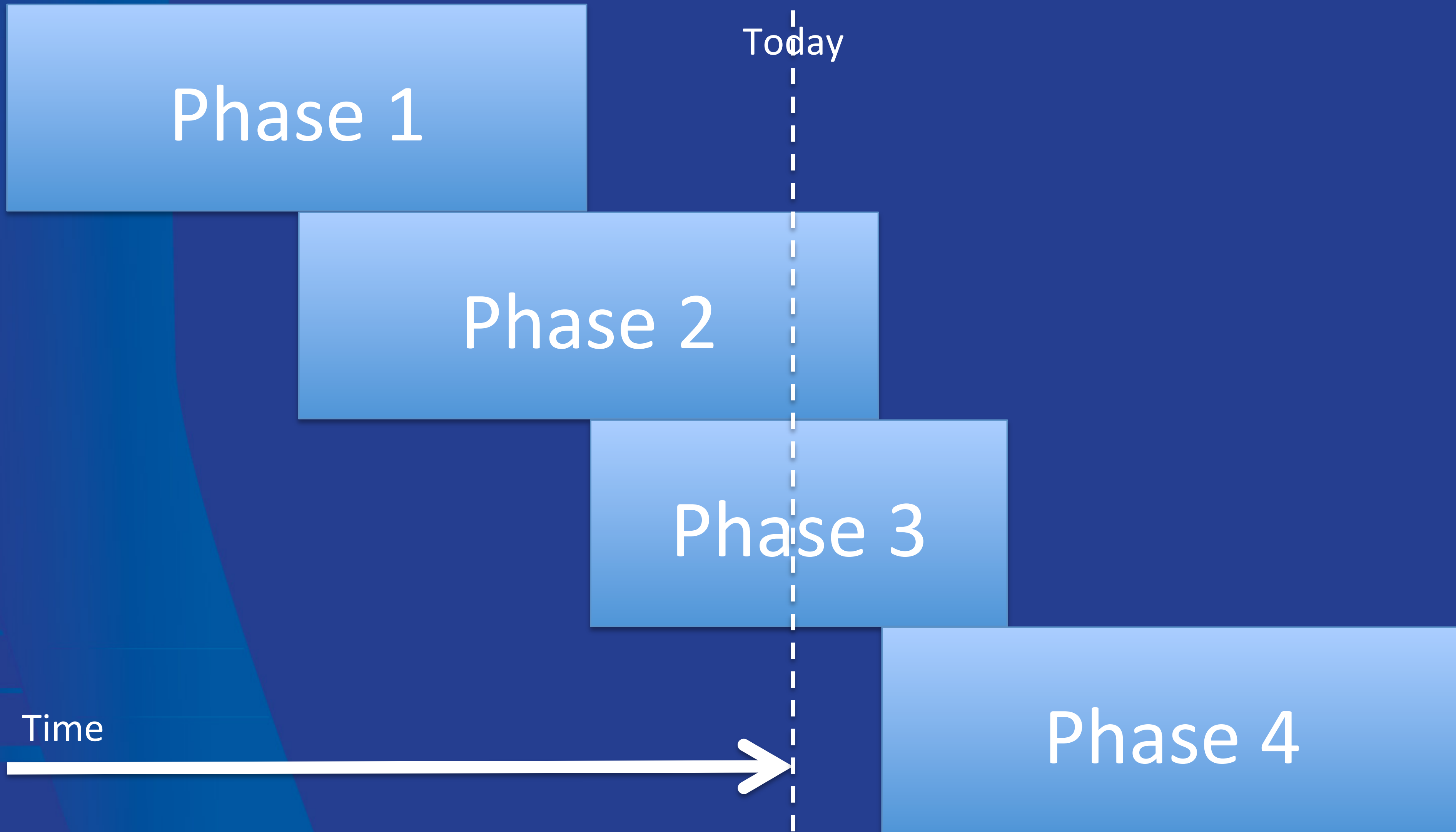
- In a (very) broad brush, the mission of FermiCloud is:

  - 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.

- This project is split over several overlapping phases.

Overlapping Phases



FermiCloud Phase 1

- Specify, acquire and deploy the FermiCloud hardware,
- Establish initial FermiCloud requirements and select the "best" open source cloud computing framework that best 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

FermiCloud Phase 2

- Implement x509 based authentication (patches contributed back to OpenNebula project and are generally available in OpenNebula V3.2), perform secure contexualization of virtual machines at launch.
- Implement monitoring and accounting,
- 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).

In Process

FermiCloud Phase 3

- Select and deploy a true multi-user filesystem on top of a distributed & replicated SAN,
- Deploy 24x7 production services,
- Deploy puppet & cobbler,
- Live migration becomes important for this phase.

In Process

FermiCloud – Hardware Specifications

Currently 23 systems split across FCC-3 and GCC-B:

- 2 x 2.67 GHz Intel "Westmere" 4 core CPU
  - Total 8 physical cores, potentially 16 cores with Hyper Threading (HT),
- 24 GBytes of memory (we deploying an upgrade to 48),
- 2 x 1GBit Ethernet interface (1 public, 1 private),
- 8 port Raid Controller,
- 2 x 300 GBytes of high speed local disk (15K RPM SAS),
- 6 x 2 TBytes = 12 TB raw of RAID SATA disk = ~10 TB formatted,
- InfiniBand SysConnect II DDR HBA,
- Brocade FibreChannel HBA (added in Fall 2011/Spring 2012),
- 2U SuperMicro chassis with redundant power supplies

FermiCloud Typical VM Specifications

- Unit:
  - 1 Virtual CPU [2.67 GHz "core" with Hyper Threading (HT)],
  - 2 GBytes of memory,
  - 10-20 GBytes of of SAN based "VM Image" storage,
  - Additional ~20-50 GBytes of "transient" local storage.

- Additional CPU "cores", memory and storage are available for "purchase":
  - Based on the (Draft) FermiCloud Economic Model,
  - Raw VM costs are competitive with Amazon EC2,
  - FermiCloud VMs can be custom configured per "client",
  - Access to Fermilab science datasets is much better than Amazon EC2.

FermiCloud – VM Format

- Virtual machine images are stored in a way that they can be exported as a device:
  - The OS partition contains full contents of the / partition plus a boot sector and a partition table.
  - Not compressed.
- Kernel and initrd are stored internally to the image,
  - Different from Amazon and Eucalyptus,
- Note that it is possible to have Xen and KVM kernels loaded in the same VM image and run it under either hypervisor.
- Secrets are not stored in the image,
  - See slides on Authentication/Contextualization.
- We are currently investigating the CERN method of launching multiple copies of same VM using LVM qcow2 (quick copy on write) but this is not our major use case at this time.
- We will likely invest in LanTorrent/LVM for booting multiple "worker node" virtual machines simultaneously at a later date.

FermiCloud Economic Model

- Calculate rack cost:
  - Rack, public Ethernet switch, private Ethernet switch, Infiniband switch,
  - $11,000 USD (one time).
- Calculate system cost:
  - Based on 4 year lifecycle,
  - $6,500 USD / 16 processors / 4 years = $250 USD / year
- Calculate storage cost:
  - 4 x FibreChannel switch, 2 x SATAbeast, 5 year lifecycle,
  - $130K USD / 60 Gbytes / 5 years = $430 USD / GB-year
- Calculate fully burdened system administrator cost:
  - Current estimate is 400 systems per administrator,
  - $250K USD / year / 400 systems = $1,250 USD / system-year

Service Level Agreements

24x7:

- Virtual machine will be deployed on the FermiCloud infrastructure 24x7.

9x5:

- Virtual machine will be deployed on the FermiCloud infrastructure 8-5, M-F, may be "suspended or shelved" at other times.

Opportunistic:

- Virtual machine may be deployed on the FermiCloud infrastructure providing that sufficient unallocated virtual machine "slots" are available, may be "suspended or shelved" at any time.

HyperThreading / No HyperThreading:

- Virtual machine will be deployed on FermiCloud infrastructure that [has / does not have] HyperThreading enabled.

Nights and Weekends:

- Make FermiCloud resources (other than 24x7 SLA) available for "Grid Bursting".

FermiCloud Draft Economic Model Results (USD)

SLA	24x7		9x5		Opportunistic
	No HT	HT	No HT	HT	--
"Unit" (CPU + 2 GB)	$250	$125	$90	$45	$24
Add'l memory per GB	$30	$30	$30	$30	$30
Add'l local disk per TB	$40	$40	$40	$40	$40
SAN disk per TB	$450	$450	$450	$450	$450
BlueArc per TB	$430	$430	$430	$430	$430
System Administrator	$1,250	$1,250	$1,250	$1,250	$1,250

FermiCloud / Amazon Cost Comparison (USD)

SLA (CPU Only)	FermiCloud	EC2 Small	EC2 Large	EC2 High CPU Medium
24x7 No HT	$250/yr	$220.50/yr	$910.00/yr	$455.00/yr
24x7 With HT	$125/yr	n/a	n/a	n/a
9x5 No HT	$90/yr	n/a	n/a	n/a
9x5 With HT	$45/yr	n/a	n/a	n/a
Opportunistic	$25/yr $0.00285/hr	$0.02/hr	$0.34/hr	$0.17/hr

Comments on Cost Comparison

- The FermiCloud "Unit" (CPU+2GB) without HyperThreading is approximately two Amazon EC2 compute units.
- Amazon can change their pricing model any time.
- The Amazon EC2 prices do not include the costs for data movement, FermiCloud does not charge for data movement. Since the typical HEP experiment moves substantial amounts of data, the Amazon data movement changes will be significant.
- The prices for FermiCloud do not include costs for the infrastructure (building/computer room/environmental/electricity) and the costs for operation (electricity).
- System administrator costs factor out of the comparison, since they apply equally to both sides of the comparison [FermiCloud / Amazon].
  - Our expectation/hope is that with the pupper & cobler deployment, the VM system administrator costs will decrease.

Amazon Data Movement Cost Range (USD)

Annual Data Movement (TB)	Data In (TB)	Data Out (TB)	Estimated Annual Cost
10	4	6	$1,331
25	10	15	$3,328
50	20	30	$6,656
100	40	60	$13,312
250	100	150	$24,064
500	200	300	$48,128
1,000	400	600	$96,256
2,000	800	1,200	$180,224
5,000	2,000	3,000	$450,560

FermiCloud – Authentication

- Reuse Grid x509 based authentication,
  - Patches to OpenNebula to support this were developed at Fermilab and submitted back to the OpenNebula project (generally available in OpenNebula V3.2).
- User authenticates to OpenNebula via graphical console via x509 authentication, EC2 query API with x509, or OCCI,
- VMs are launched with the users x509 proxy.

FermiCloud – Contextualization

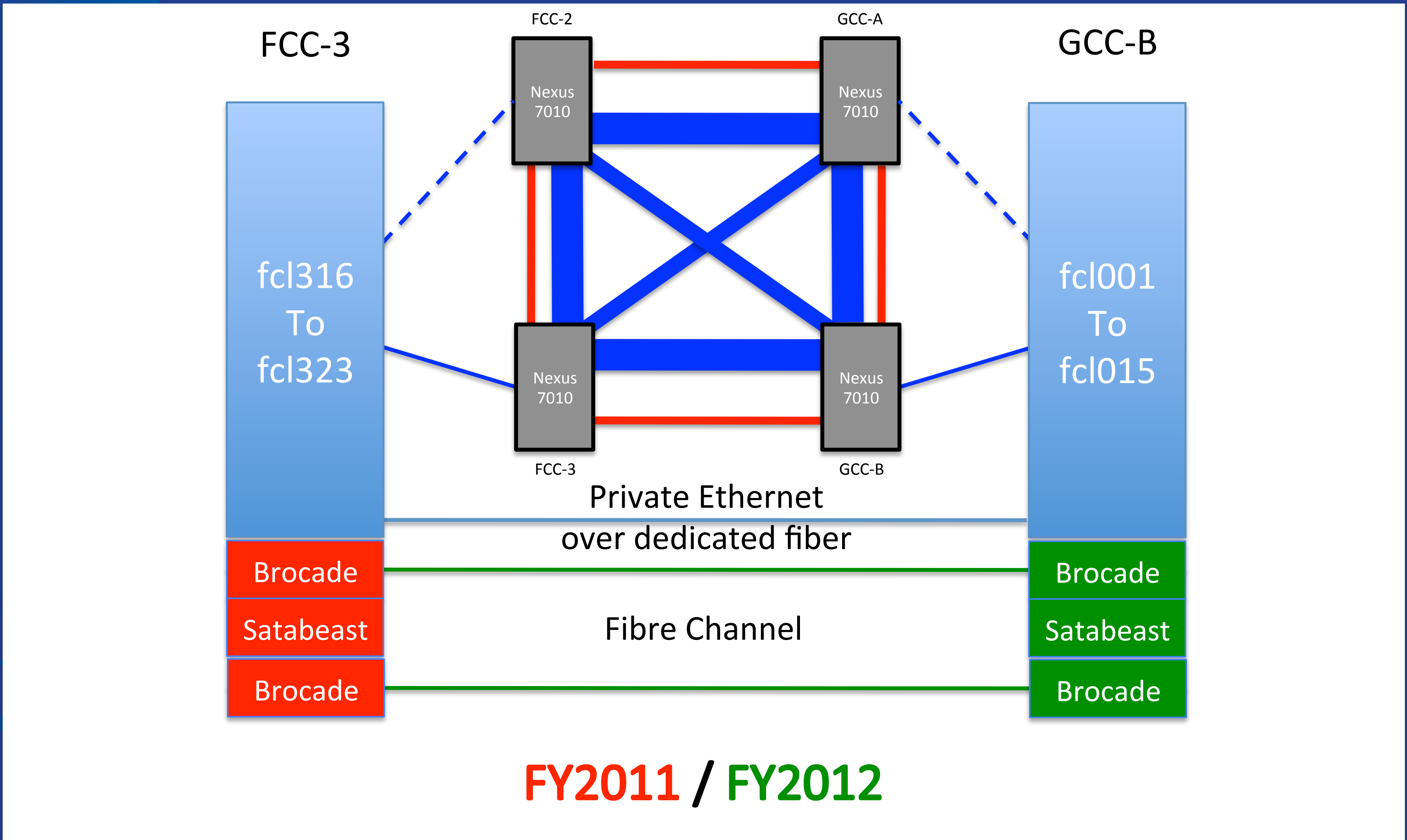
- VM contextualization accomplished via:
  - Use user x509 proxy credentials to perform secure access via openSSL to an external "secure secrets repository",
  - Credentials are copied into ramdisk within the VM and symlinks are made from the standard credential locations (/etc/grid-security/certificates) to the credentials in the ramdisk.
- On VM shutdown, the contents of the ramdisk disappear and the original credential remains in the "secure secrets repository".
- These mechanisms prevent the "misappropriation" of credentials if a VM is copied from the FermiCloud VM library,
  - No credentials are stored in a VM "at rest".
- This is not perfect – a determined user (VM administrator) could still copy the secure credentials off of their running VM, but this does not offer any additional risk beyond that posed by the administrator of a physical system.

FermiCloud – Fault Tolerance

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

- We are actively working on deploying the FermiCloud hardware resources in a fault tolerant infrastructure:
  - The physical systems are split across two buildings,
  - There is a fault tolerant network infrastructure in place that interconnects the two buildings,
  - We have deployed SAN hardware in both buildings,
  - We are evaluating GFS for our multi-user filesystem and distributed & replicated SAN.

FermiCloud – Network & SAN "Today"



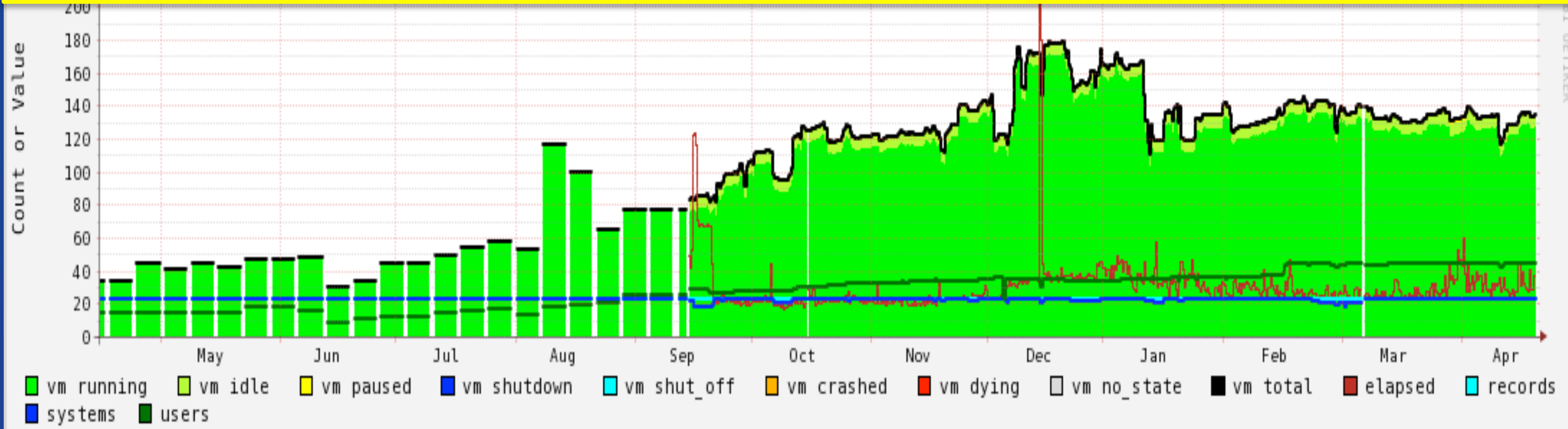
FermiCloud – Monitoring

- Temporary FermiCloud Usage Monitor:
  - http://www-fermicloud.fnal.gov/fermicloud-usage-data.html
  - Data collection dynamically "ping-pongs" across systems deployed in FCC and GCC to offer redundancy,
  - See plot on next page.

- FermiCloud Redundant Ganglia Servers:
  - http://fcl001k1.fnal.gov/ganglia/
  - http://fcl002k1.fnal.gov/ganglia/

- ***Preliminary*** RSV based monitoring pilot:
  - http://fermicloudrsv.fnal.gov/rsv

Note – ***FermiGrid*** Production Services are operated at 100% to 200% "oversubscription"



VM states as reported by "virsh list"

	Maximum	Average	Minimum	LastVal
records	23	23	23	23
systems	23	23	18	23
vm total	179	104	30	134
vm running	172	99	30	127
vm idle	7	7	6	7
vm paused	2	0	0	0
vm shutdown	0	0	0	0
vm shut off	0	0	0	0
vm crashed	0	0	0	0
vm dying	0	0	0	0
vm no state	0	0	0	0
users	45	29	9	45
elapsed	305	31	17	37

Note - vm states as reported by virsh list
Data for ferm ... 2012
Plot generate ... al.gov

FermiCloud Target

FermiCloud Capacity	# of Units
Nominal (1 physical core = 1 VM)	184
50% over subscription	276
100% over subscription (1 HT core = 1 VM)	368
200% over subscription	552

Description of Virtual Machine States Reported by "virsh list" Command

State	Description
running	The domain is currently running on a CPU. Note – KVM based VMs show up in this state even when they are "idle"
idle	The domain is idle, and not running or runnable. This can be caused because the domain is waiting on I/O (a traditional wait state) or has gone to sleep because there was nothing else for it to do. Note – Xen based VMs typically show up in this state even when they are "running"
paused	The domain has been paused, usually occurring through the administrator running virsh suspend. When in a paused state the domain will still consume allocated resources like memory, but will not be eligible for scheduling by the hypervisor.
shutdown	The domain is in the process of shutting down, i.e. the guest operating system has been notified and should be in the process of stopping its operations gracefully.
shut off	The domain has been shut down. When in a shut off state the domain does not consume resources.
crashed	The domain has crashed. Usually this state can only occur if the domain has been configured not to restart on crash.
dying	The domain is in process of dying, but hasn't completely shutdown or crashed.

FermiCloud – Monitoring Requirements & Goals

- Need to monitor to assure that:
  - All hardware is available (both in FCC3 and GCC-B),
  - All necessary and required OpenNebula services are running,
  - All "24x7" & "9x5" virtual machines (VMs) are running,
  - 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.

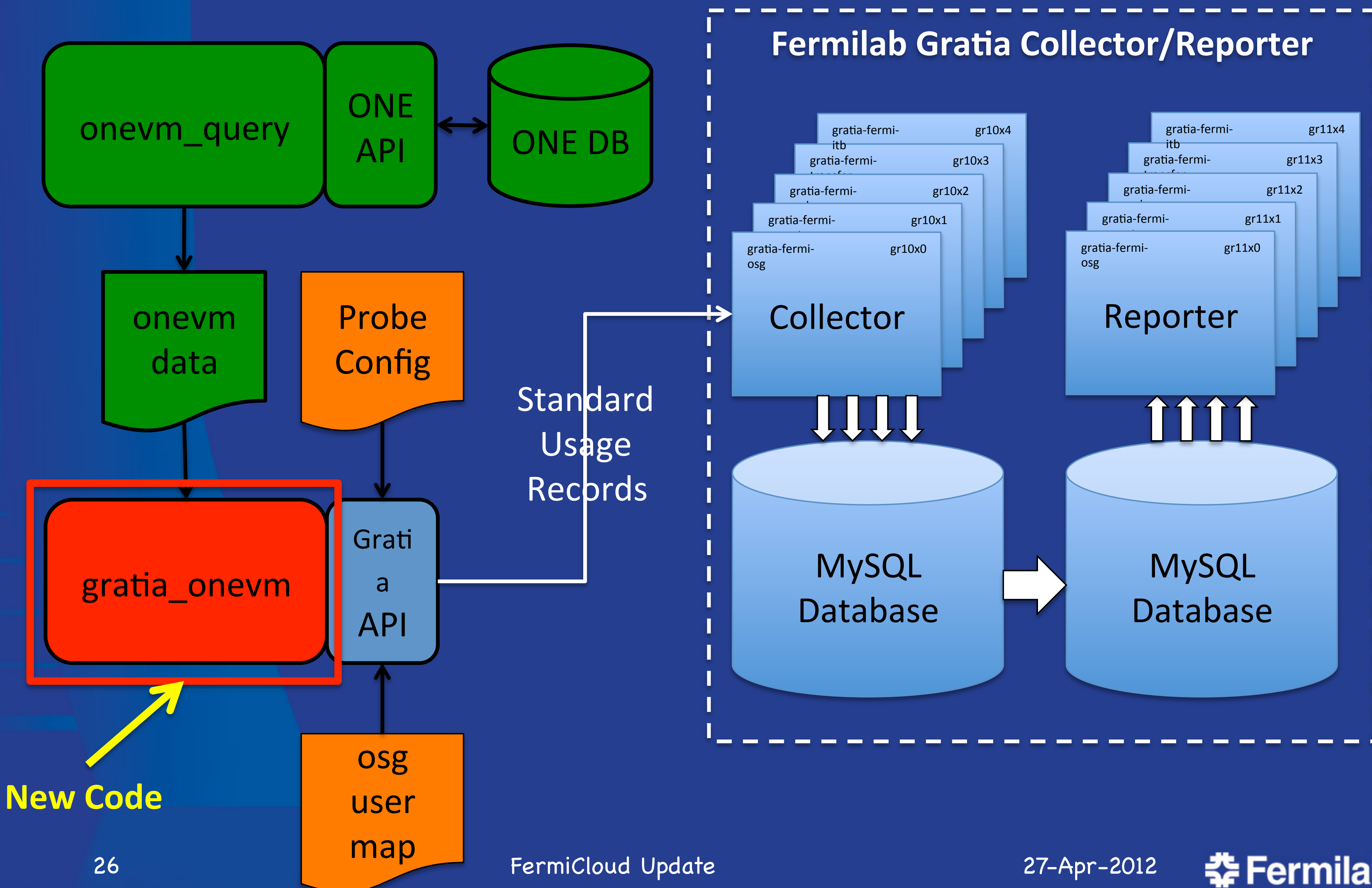
- We plan to replace the temporary monitoring with an infrastructure based on either Nagios or Zabbix during CY2012.
  - Possibly utilizing the OSG Resource Service Validation (RSV) scripts.
  - This work will likely be performed in collaboration with KISTI (and others).

- A "stretch" goal of the monitoring project is to figure out how to identify really idle virtual machines.
  - Unfortunately, at the present time we cannot use the "virsh list" output, since actively running Xen based VMs are incorrectly labeled as "idle" and idle KVM based VMs are incorrectly labeled as "running".
  - In times of resource need, we want the ability to suspend or "shelve" the really idle VMs in order to free up resources for higher priority usage.
  - Shelving of "9x5" and "opportunistic" VMs will allow us to use FermiCloud resources for Grid worker node VMs during nights and weekends (this is part of the draft economic model).

FermiCloud - Accounting

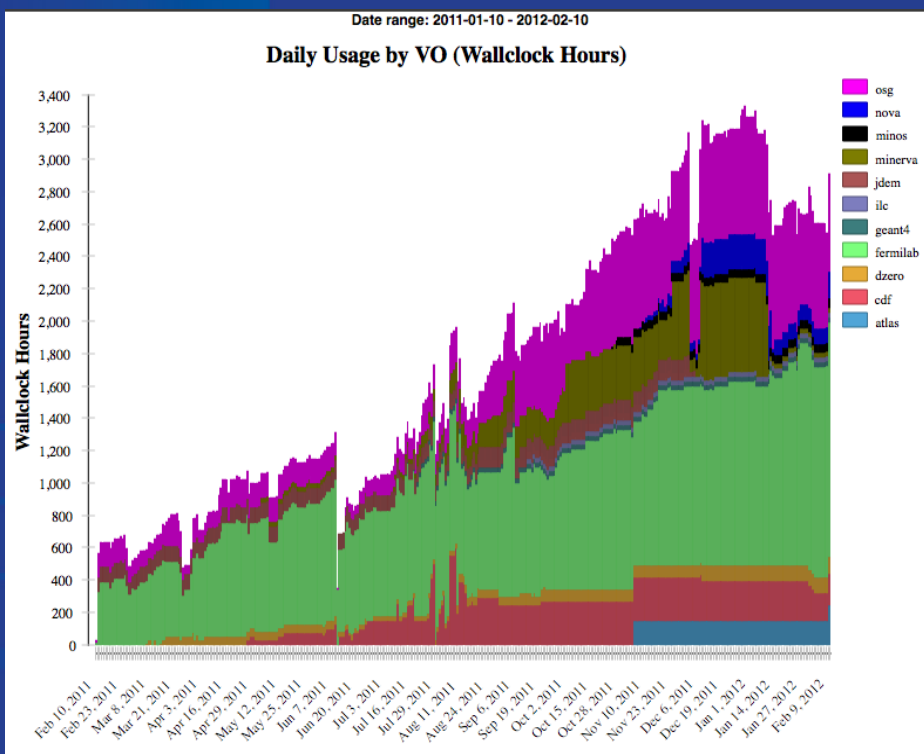
- Currently have two "probes" based on the Gratia accounting framework used by Fermilab and the Open Science Grid:
  - https://twiki.grid.iu.edu/bin/view/Accounting/WebHome
- Standard Process Accounting ("psacct") Probe:
  - Installed and runs within the virtual machine image,
  - Reports to standard gratia-fermi-psacct.fnal.gov.
- Open Nebula Gratia Accounting Probe:
  - Runs on the OpenNebula management node and collects data from ONE logs, emits standard Gratia usage records,
  - Reports to the "virtualization" Gratia collector,
  - The "virtualization" Gratia collector runs existing standard Gratia collector software (no development was required),
  - The development of the Open Nebula Gratia accounting probe was performed by Tanya Levshina and Parag Mhashilkar.
- Additional Gratia accounting probes could be developed:
  - Commercial – OracleVM, VMware, ---
  - Open Source – Nimbus, Eucalyptus, OpenStack, …

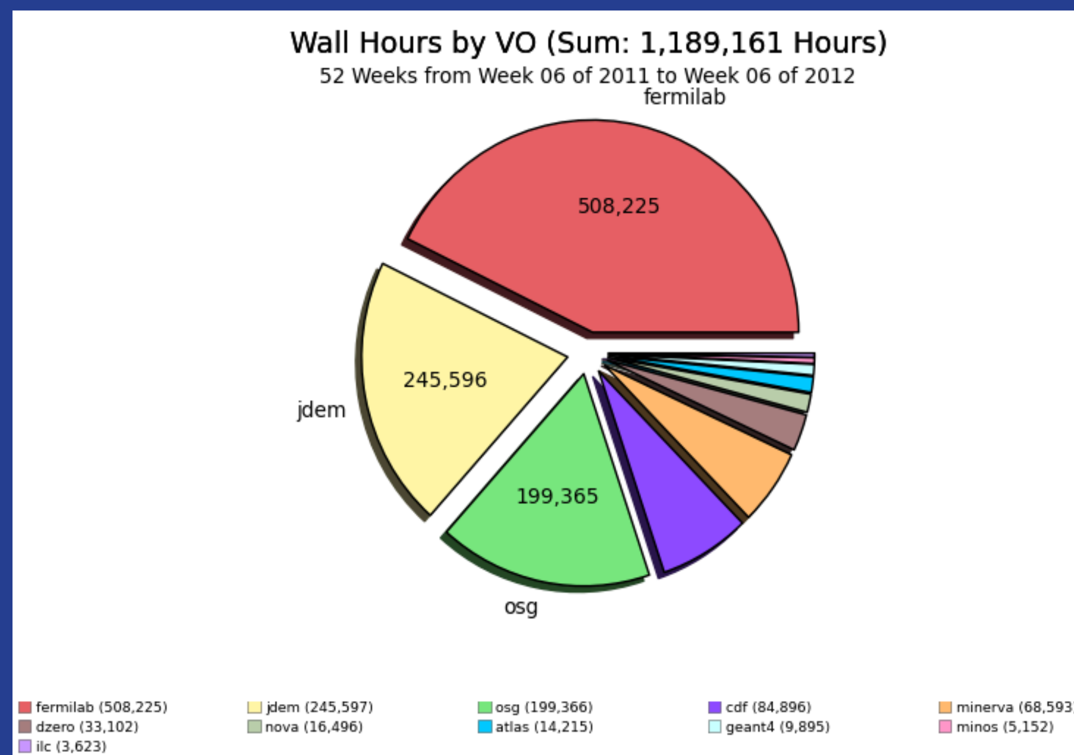
Open Nebula Gratia Accounting Probe



FermiCloud – Gratia Accounting Reports

Here are the results of "replaying" the previous year of the OpenNebula "OneVM" data into the new accounting probe:





Virtualized Storage Service Investigation

Motivation:

- General purpose systems from various vendors being used as file servers,
- Systems can have many more cores than needed to perform the file service,
  - Cores go unused => Inefficient power, space and cooling usage,
  - Custom configurations => Complicates sparing issues.

Question:

- Can virtualization help here?
- What (if any) is the virtualization penalty?

Virtualized Storage Server Test Procedure

Evaluation:

- Use IOzone and real physics root based analysis code.

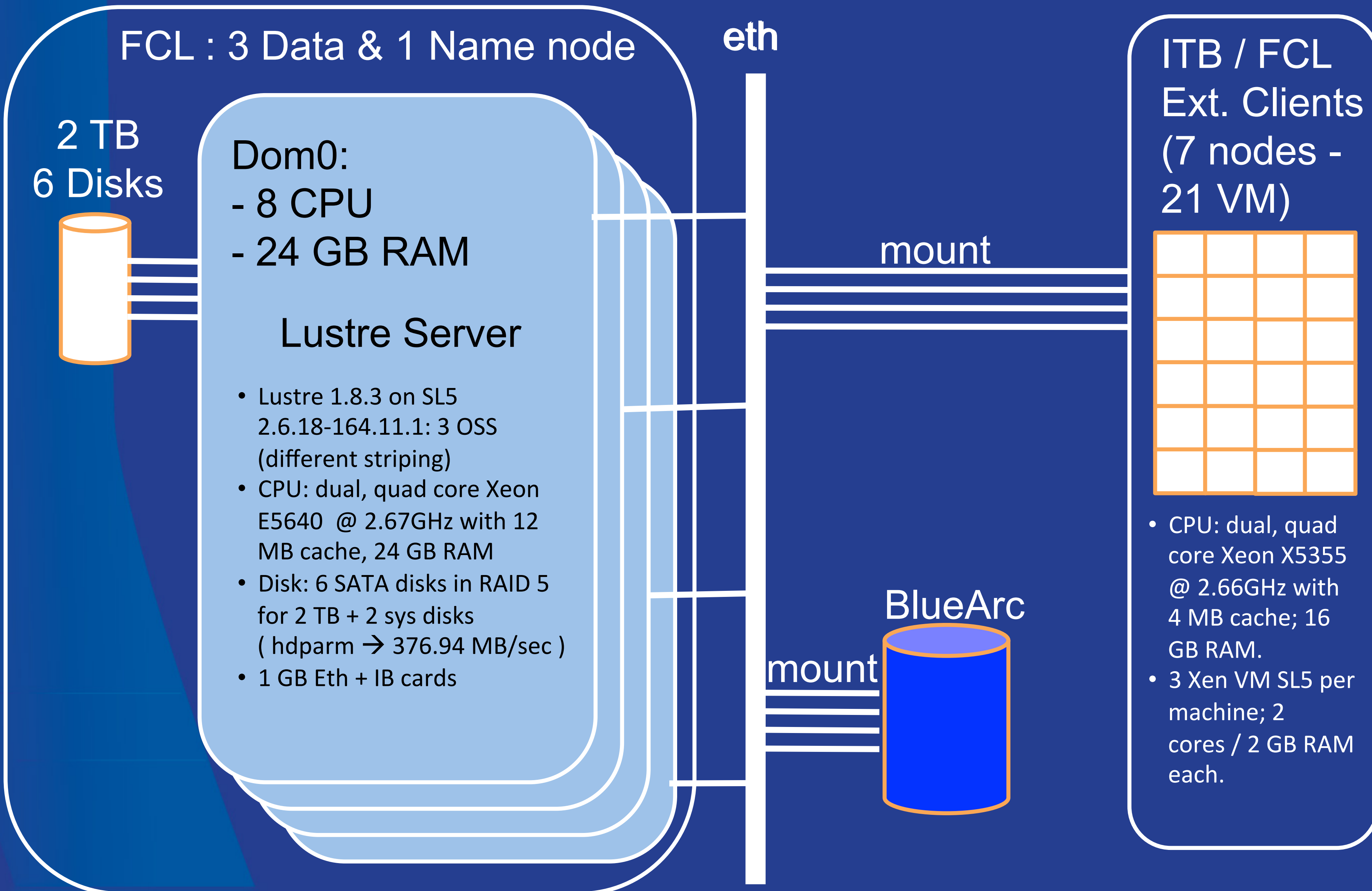
Phase 1:

- Install candidate filesystem on "bare metal" server,
- Evaluate performance using combination of bare metal and virtualized clients (varying the number),
- Also run client processes on the "bare metal" server,
- Determine "bare metal" filesystem performance.

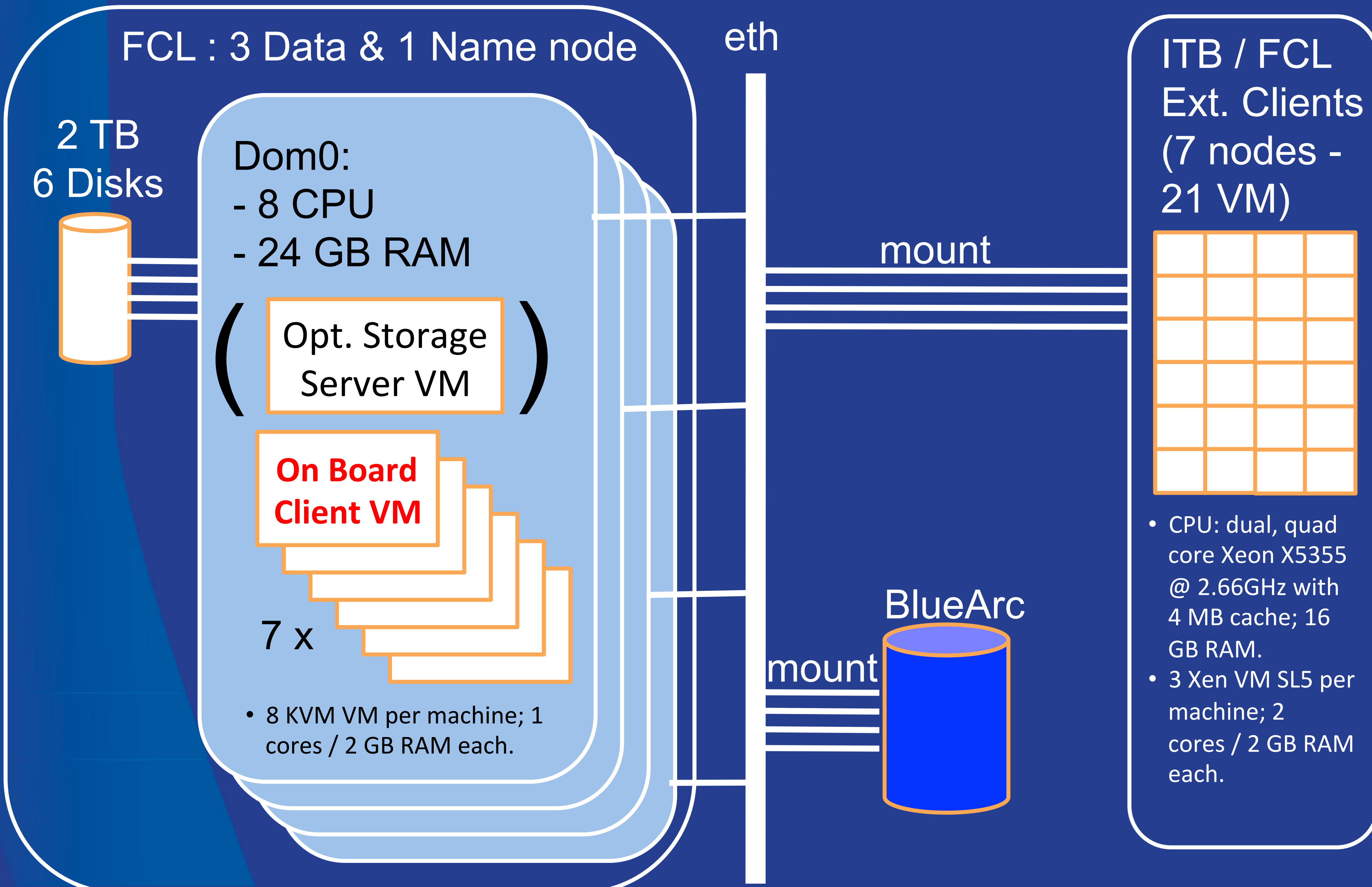
Phase 2:

- Install candidate filesystem on a virtual machine server,
- Evaluate performance using combination of bare metal and virtualized clients (varying the number),
- Also use client virtual machines hosted on same physical machine as the virtual machine server,
- Determine virtual machine filesystem performance.

FermiCloud Test Bed - "Bare Metal" Server



FermiCloud Test Bed - Virtualized Server



Virtualized File Service Results Summary

FileSystem	Benchmark	Read (MB/s)	"Bare Metal" Write (MB/s)	VM Write (MB/s)	Notes
Lustre	IOZone	350	250	70	Significant write penalty when FS on VM
	Root-based	12.6	-	-	
Hadoop	IOZone	50 - 240	80 - 300	80 - 300	Varies on number of replicas, fuse does not export a full posix fs.
	Root-based	7.9	-	-	
OrangeFS	IOZone	150 - 330	220 - 350	220 - 350	Varies on number of name nodes
	Root-based	8.1	-	-	
BlueArc	IOZone	300	330	n/a	Varies on system conditions
	Root-based	8.4	-	-	

See ISGC talk for the details - http://indico3.twgrid.org/indico/getFile.py/access?contribId=32&sessionId=36&resId=0&materialId=slides&confId=44

FermiCloud – Interoperability

- From the beginning, one of the goals of FermiCloud has been the ability to operate as a hybrid cloud:
  - Being able to join FermiCloud resources to FermiGrid resources to temporarily increase the Grid capacity or GlideinWMS with VMs (Grid Bursting),
  - Being able to join public cloud resources (such as Amazon EC2) to FermiCloud (Cloud Bursting via Public Clouds).
  - Participate in compatible community clouds (Cloud Bursting via other Private Clouds). Had the DOE Magellan project continued further we likely would have invested significant effort here (anyone looking for a collaboration?).

FermiCloud – Grid Bursting

- Join "excess" FermiCloud capacity to FermiGrid:
  - Identify "idle" VMs on FermiCloud,
  - Automatically "shelve" the "idle" VMs,
  - Automatically launch "worker node" VMs,
  - "worker node" VMs join existing Grid cluster and contribute their resources to the Grid.
  - "Shelve" the "worker node" VMs when appropriate.
- AKA – The "nights and weekend" plan for increased Grid computing capacity.
- At the moment, we are waiting for the results of the monitoring project later this year to (hopefully) allow us to correctly identify "idle" VMs.

FermiCloud – Cloud Bursting

- vCluster - Deployable on demand virtual cluster using hybrid cloud computing resources.
  - Head nodes launched on virtual machines within the FermiCloud private cloud.
  - Worker nodes launched on virtual machines within the Amazon EC2 public cloud.
- Work performed by Dr. Seo-Young Noh (KISTI).
  - Refer to his ISGC talk on Friday 2-Mar-2012 for more details:
  - http://indico3.twgrid.org/indico/contributionDisplay.py?contribId=1&confId=44

FermiCloud – Community Cloud

- There are efforts underway at Fermilab that may result in the deployment of additional cloud computing resources based on the FermiCloud model.
- If/When these efforts are successful, we will interoperate with them.
- We are also willing to collaborate on furthering interoperability with other cloud computing resources that use a compatible access model (x509).

FermiCloud – Running "External" VMs

- We participate in:
  - The HEPiX virtualization working group led by Tony Cass (CERN),
  - The "Security for Collaborating Infrastructures" (SCI) group led by Dave Kelsey (RAL).

- That being said…
  - It is our intention that FermiCloud (and likely FermiGrid at some future date) will support the submission of VMs for running on FermiCloud or FermiGrid resources via standard Cloud/Grid mechanisms. Such as Amazon S3, OCCI, globus-url-copy (GridFTP) or globus-job-run in addition to direct OpenNebula console access via x509 authentication.
  - As part of our security infrastructure we will likely reuse the existing "network jail" to treat new untrusted VMs similarly to how Fermilab treats new untrusted laptops on the site network.
  - If there is any issue with the VM, we will directly contact the people who:
    1. Transferred the VM to Fermilab (*yes, we do keep logs as well as monitor the actions of the VM on the network*).
    2. Launched the VM on FermiCloud and/or FermiGrid (*again – yes, we do keep logs as well as monitor the actions of the VM on the network*).
  - If we don't get satisfactory answers from both, then both of the DNs will likely wind up on the site "blacklist" infrastructure.

MPI on FermiCloud (Note 1)

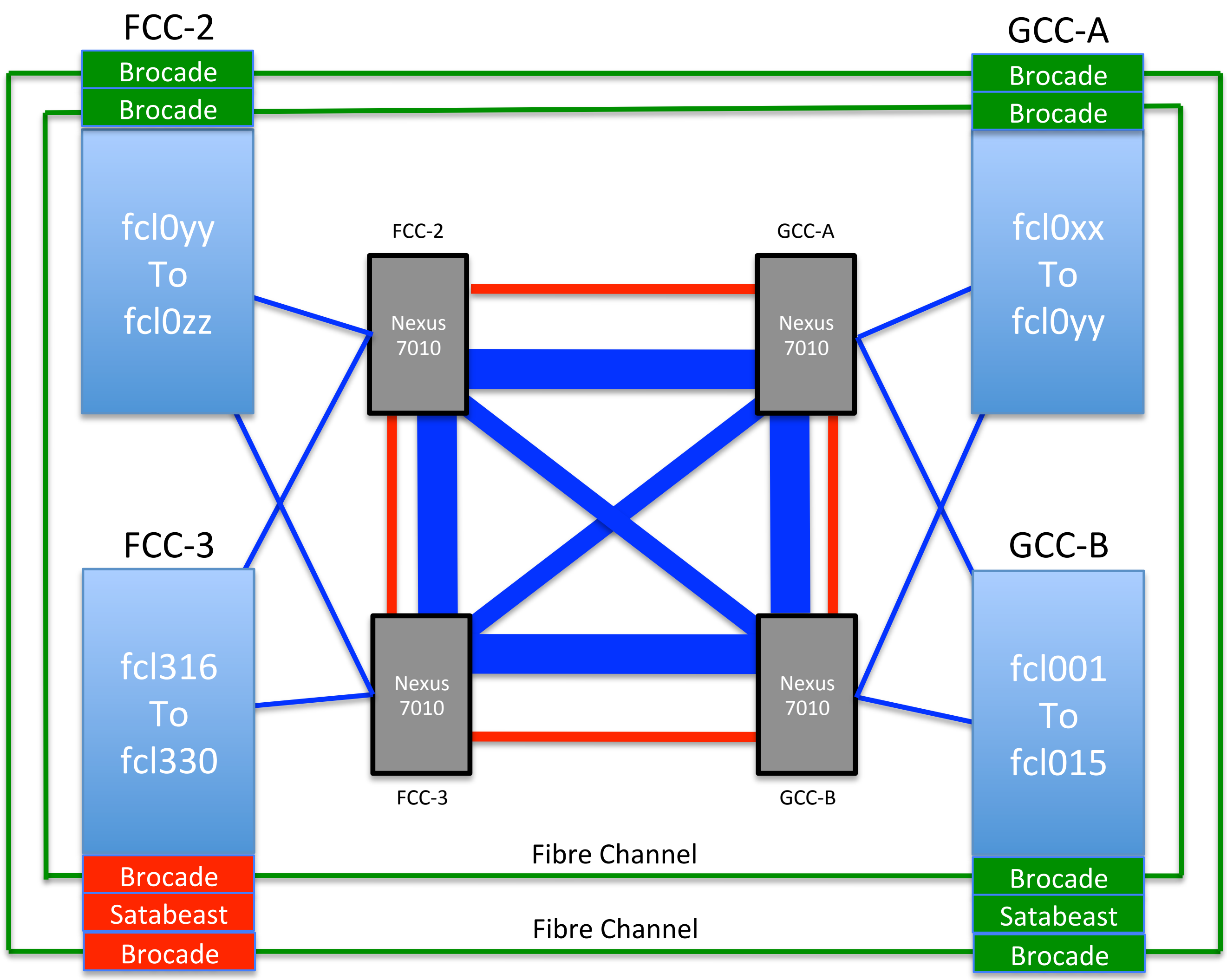
Configuration	#Host Systems	#VM/host	#CPU	Total Physical CPU	HPL Benchmark (Gflops)
Bare Metal without pinning	2	--	8	16	13.9
Bare Metal with pinning (Note 2)	2	--	8	16	24.5
VM without pinning (Notes 2,3)	2	8	1 vCPU	16	8.2
VM with pinning (Notes 2,3)	2	8	1 vCPU	16	17.5
VM+SRIOV with pinning (Notes 2,4)	2	7	2 vCPU	14	23.6

Notes:
(1) Work performed by Dr. Hyunwoo Kim of KISTI in collaboration with Dr. Steven Timm of Fermilab.
(2) Process/Virtual Machine "pinned" to CPU and associated NUMA memory via use of numactl.
(3) Software Bridged Virtual Network using IP over IB (seen by Virtual Machine as a virtual Ethernet).
(4) SRIOV driver presents native InfiniBand to virtual machine(s), 2nd virtual CPU is required to start SRIOV, but is only a virtual CPU, not an actual physical CPU.

Current Stakeholders

- Grid & Cloud Computing Personnel,
- Run II – CDF & D0,
- Intensity Frontier Experiments,
- Cosmic Frontier (JDEM/WFIRST),
- Korean Institute for Science & Technology Investigation (KISTI),
- Open Science Grid (OSG) software refactoring from pacman to RPM based distribution.

FermiCloud – Network & SAN (Possible Future – FY2013/2014)



FermiCloud Summary - 1

- The existing (temporary) FermiCloud usage monitoring shows that the peak FermiCloud usage is ~100% of the nominal capacity and ~50% of the expected oversubscription capacity.
- The FermiCloud collaboration with KISTI has leveraged the resources and expertise of both institutions to achieve significant benefits.
- FermiCloud has plans to implement both monitoring and accounting by extension of existing tools in CY2012.
- Using SRIOV drivers on FermiCloud virtual machines, MPI performance has been demonstrated to be >96% of the native "bare metal" performance.
  - Note that this HPL benchmark performance measurement was accomplished using **2 fewer** physical CPUs than the corresponding "bare metal" performance measurement!
- FermiCloud personnel are working to implement a SAN storage deployment that will offer a true multi-user filesystem on top of a distributed & replicated SAN.
- Science is directly and indirectly benefiting from FermiCloud:
  - CDF, D0, Intensity Frontier, Cosmit Frontier, CMS, ATLAS, Open Science Grid, …

FermiCloud Summary - 2

- FermiCloud operates at the forefront of delivering cloud computing capabilities to support physics 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 an Infrastructure as a Service cloud computing capability that supports science at Fermilab.
  - The Open Science Grid software team is using FermiCloud resources to support their RPM "refactoring".
- None of this 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 KISTI.
- We have a personnel opening to work with the FermiCloud project:
  - Cloud and Grid Administrator (Computer Services Specialist III)
  - https://fermi.hodesiq.com/job_detail.asp?JobID=2985317&user_id=

Thank You!

Any Questions?