FermiGrid and FermiCloud Update

ISGC 2012

Keith Chadwick <chadwick@fnal.gov>
Fermilab Grid & Cloud Computing Dept.
28-Feb-2012

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

Feynman Computing Center (FCC):
- FCC2 Computer Room
- FCC3 Computer Rooms
- High availability services – e.g. core network, email, etc.
- Tape Robotic Storage (3 x 10000 slot libraries)
- UPS & Standby Power Generation
- ARRA project: upgrade cooling and add HA computing room - completed

Grid Computing Center (GCC):
- 3 Computer Rooms – GCC-[A,B,C]
- Tape Robot Room – GCC-TRR
- High Density Computational Computing
- CMS, RUNII, Grid Farm batch worker nodes
- Lattice HPC nodes
- Tape Robotic Storage (4 x 10000 slot libraries)
- UPS & taps for portable generators

Lattice Computing Center (LCC):
- High Performance Computing (HPC)
- Accelerator Simulation, Cosmology nodes
- Systems for Integration & Development
- No UPS

US EPA Energy Star award for 2010 & 2011
Grid & Cloud Computing Department

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

Distributed Offline Computing Services

Gabriele Garzoglio (Leader)
David Dykstra
Tanya Levshina
Parag Mhashilkar
Marko Slyz
Douglas Strain

FermiGrid Services

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

28-Feb-2012  FermiGrid and FermiCloud Update
The Fermilab Campus Grid (FermiGrid) consists of the high throughput computing clusters at Fermilab together with the underlying highly available service infrastructure that enables these resources for the Grid.
FermiGrid – Strategy & History

• In 2004, the Fermilab Computing Division undertook the strategy of placing all of its production resources in a Grid "meta-facility" infrastructure called FermiGrid.

• In April 2005, the first “core” FermiGrid services were deployed.

• In 2007, the FermiGrid-HA project was commissioned with the goal of delivering 99.999% core service availability (not including building or network outages).
  – In the first seven months of operation from November 2007 through June 2008, the measured core service availability was 99.997%.

• In 2011, the FermiGrid-HA2 project was commissioned with the goal of having the FermiGrid infrastructure being able to withstand significant building or network failures.
  – I’ll be presenting the results of this work later in this presentation.
FermiGrid – Goals of the Strategy

The FermiGrid strategy was designed to allow Fermilab:

• to insure that the large experiments who currently have dedicated resources to have first priority usage of those resources that are purchased on their behalf.
• to allow opportunistic use of these dedicated resources, as well as other shared Farm and Analysis resources, by various Virtual Organizations (VO's) that participate in the Fermilab experimental program and by certain VO's that use the Open Science Grid (OSG).
• to optimise use of resources at Fermilab.
• to make a coherent way of putting Fermilab on the Open Science Grid.
• to save some effort and resources by implementing certain shared services and approaches.
• to work together more coherently to move all of our applications and services to run on the Grid.
• to better handle a transition from Run II to LHC in a time of shrinking budgets and possibly shrinking resources for Run II worldwide.
• to fully support Open Science Grid and the LHC Computing Grid and gain positive benefit from this emerging infrastructure in the US and Europe.
## Current FermiGrid Statistics (as of February 2012)

<table>
<thead>
<tr>
<th>Cluster(s)</th>
<th>Batch System</th>
<th>Job Slots</th>
<th>Raw Occupancy</th>
<th>Effective Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDF (Merged)</td>
<td>Condor</td>
<td>6,260</td>
<td>94.3</td>
<td>72.8</td>
</tr>
<tr>
<td>CMS T1</td>
<td>Condor</td>
<td>7,784</td>
<td>98.3</td>
<td>88.2</td>
</tr>
<tr>
<td>D0 (Merged)</td>
<td>PBS</td>
<td>8,242</td>
<td>80.3</td>
<td>59.7</td>
</tr>
<tr>
<td>GP Grid</td>
<td>Condor</td>
<td>4,898</td>
<td>83.5</td>
<td>73.2</td>
</tr>
<tr>
<td>Overall-Today</td>
<td></td>
<td>27,184</td>
<td>89.3</td>
<td>73.3</td>
</tr>
<tr>
<td>Last Year</td>
<td></td>
<td>23,285</td>
<td>82.0</td>
<td>62.4</td>
</tr>
</tbody>
</table>

Slot = one CPU core that's assigned to a single user job (to first order),
Occupancy = % of Slots that are occupied by jobs,
Utilization = % CPU usage of occupied slots.
FermiGrid Overall Usage

Key:                
                    Busy     Idle     Waiting     Held

FermiGrid and FermiCloud Update
FermiGrid – Usage By Community

FermiGrid - Overall Major VO Usage - Last Year

Key:
- Red: FNAL
- Green: OSG
- Yellow: CMS
- Blue: Dzero

FermiGrid and FermiCloud Update
28-Feb-2012
FermiGrid-HA – Service Redundancy (deployed in 2007)

FermiGrid-HA uses three key technologies:
- Linux Virtual Server (LVS),
- Xen Hypervisor,
- MySQL Circular Replication.

The goal for FermiGrid-HA is greater than 99.999% service availability (Not including building or network outages).
## What Environment Do We Operate In?

<table>
<thead>
<tr>
<th>Risk</th>
<th>Mitigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service Failure</td>
<td>Redundant copies of Service (with service monitoring).</td>
</tr>
<tr>
<td>System Failure</td>
<td>Redundant copies of Service (on separate systems with service monitoring).</td>
</tr>
<tr>
<td>Network Failure</td>
<td>Redundant copies of Service (on separate systems with independent network connections and service monitoring).</td>
</tr>
<tr>
<td>Building Failure</td>
<td>Redundant copies of Service (on separate systems with independent network connections in separate buildings and service monitoring).</td>
</tr>
</tbody>
</table>

*FermiGrid-HA – Deployed in 2007*

*FermiGrid-HA2 – Deployed in 2011*
FermiGrid-HA2 – Network & Geographic Redundancy (2011)

Diagram showing the network and geographic redundancy setup with:
- FCC-2
- GCC-B
- Public LAN Switch
- LVS (active)
- LVS (standby)
- Service (active)
- Private LAN Switch
- Dedicated Fiber
- Heartbeat connection

28-Feb-2012
FermiGrid-HA2 – Geographic Locations

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

FCC has UPS and Generator.

GCC has UPS.
FermiGrid-HA2 Deployment

• The FermiGrid-HA2 software, hardware & physical reorganization was completed on Tuesday 7-Jun-2011 at ~1300.
  – Replicated critical services are hosted in two data centers (FCC-2 & GCC-B).
  – Non-critical services are split across the two data centers.

• The plan had been to utilize a scheduled power outage on 13-Aug-2011 for FCC2 to serve as the final acceptance test for the FermiGrid-HA2 project.
  – This power outage was later moved to 15-Oct-2011, and FemiGrid-HA2 functioned as designed for the scheduled power outage.

• Unfortunately, ~2 hours after the completion of the FermiGrid-HA2 reorganization, the GCC-B computer room suffered a major cooling outage.
  – All systems in GCC-B were shut down when the facilities personnel switched off the main electrical panel breakers.

• FermiGrid-HA2 functioned exactly as designed.
  – The critical services failed back to the remaining copy of the service on FCC-2.
  – The non-critical services went to reduced capacity.
  – When cooling and power were restored at ~1700, the second copy of the critical services in GCC-B transparently rejoined the service “pool”, and the non-critical services in GCC-B resumed operation.
# FermiGrid Service Availability (measured over the past year)

<table>
<thead>
<tr>
<th>Service</th>
<th>Raw Availability</th>
<th>HA Configuration</th>
<th>Measured HA Availability</th>
<th>Minutes of Downtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOMS – VO Management Service</td>
<td>99.667%</td>
<td>Active-Active</td>
<td>100.000%</td>
<td>0</td>
</tr>
<tr>
<td>GUMS – Grid User Mapping Service</td>
<td>99.663%</td>
<td>Active-Active</td>
<td>100.000%</td>
<td>0</td>
</tr>
<tr>
<td>SAZ – Site AuthoriZation Service</td>
<td>99.622%</td>
<td>Active-Active</td>
<td>100.000%</td>
<td>0</td>
</tr>
<tr>
<td>Squid – Web Cache</td>
<td>99.663%</td>
<td>Active-Active</td>
<td>100.000%</td>
<td>0</td>
</tr>
<tr>
<td>MyProxy – Grid Proxy Service</td>
<td>99.374%</td>
<td>Active-Standby</td>
<td>99.749%</td>
<td>1,320</td>
</tr>
<tr>
<td>ReSS – Resource Selection Service</td>
<td>99.779%</td>
<td>Active-Active</td>
<td>100.000%</td>
<td>0</td>
</tr>
<tr>
<td>Gratia – Fermilab and OSG Accounting</td>
<td>99.195%</td>
<td>Active-Standby</td>
<td>99.997%</td>
<td>60</td>
</tr>
<tr>
<td>MySQL Database</td>
<td>99.785%</td>
<td>Active-Active</td>
<td>100.000%</td>
<td>0</td>
</tr>
</tbody>
</table>
### FermiGrid “Core” Service Metrics

(measured over the past year)

<table>
<thead>
<tr>
<th>Service</th>
<th>Calls per Hour Average / Peak</th>
<th>Calls per Day Average / Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOMS – VO Management Service</td>
<td>84 / 175</td>
<td>984 / 2,969</td>
</tr>
<tr>
<td>GUMS – Grid User Mapping Service</td>
<td>17.3K / 114.1K</td>
<td>415K / 1.25M</td>
</tr>
<tr>
<td>SAZ – Site AuthoriZation Service</td>
<td>14.6K / 150.3K</td>
<td>350K / 1.23M</td>
</tr>
<tr>
<td>MyProxy – Grid Proxy Service</td>
<td>719 / 1,449</td>
<td>12.0K / 26.5K</td>
</tr>
</tbody>
</table>
FermiGrid Summary

• The FermiGrid-HA2 project to implement network & geographic redundancy proved itself two hours after completion!
  – And has been used several times since for both scheduled and unscheduled outages.

• The measured availability of the FermiGrid High Availability services is nothing less than extraordinary.
  – With the exception of MyProxy – Our MyProxy-HA service deployment using DRBD has proven to be extremely sensitive to Kernel version upgrades,
  – We have deployed additional policies, procedures and monitoring to work toward addressing the underlying issues on MyProxy service availability.

• FermiGrid (The Fermilab Campus Grid) is a world class infrastructure for scientific computing.
  – Across all three of the Fermilab research frontiers (Energy, Intensity and Cosmic),
  – FermiGrid stakeholders are routinely making opportunistic usage of other stakeholder resources across FermiGrid,
  – Together with opportunistic computing from the OSG and WLCG.
Network

Distributed Network "Core" Provides Redundant Connectivity
Distributed Network “Core” Provides Redundant Connectivity

Deployment completed in January 2012

- Disk Servers
- Robotic Tape Libraries (3)
- Grid Worker Nodes
- CCD-2
  - 20 Gigabit/s L3 Routed Network
- GCC-A
  - 80 Gigabit/s L2 Switched Network
- FCC-3
  - 40 Gigabit/s L3 Switched Networks
  - Note – Intermediate level switches and top of rack switches are not shown in the this diagram.
- FCC-2
  - Private Networks over dedicated fiber
- GCC-B
  - Nexus 7010
  - Disk Servers
  - Fermi Grid
  - Grid Worker Nodes
  - GCC-A
  - 80 Gigabit/s L2 Switched Network
  - Fermi Cloud
  - Grid Worker Nodes
  - Private Networks over dedicated fiber

Note – Intermediate level switches and top of rack switches are not shown in the this diagram.

Deployment completed in January 2012
FermiCloud

FermiCloud – Infrastructure as a Service (IaaS) Cloud Computing in support of the Fermilab Scientific Program
FermiCloud – 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.
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).
FermiCloud Phase 2

• 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.
• 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).
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.
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 are considering 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),
- 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 – Monitoring

• Temporary FermiCloud Usage Monitor:
  – Data collection dynamically “ping-pons” 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://fcl001k1.fnal.gov/ganglia/)

• Preliminary RSV-based monitoring pilot:
Note – **FermiGrid** Production Services are operated at 100% to 200% “oversubscription”

**FermiCloud Capacity**

<table>
<thead>
<tr>
<th># of Units</th>
<th>Nominal (1 physical core = 1 VM)</th>
<th>184</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50% over subscription</td>
<td>276</td>
</tr>
<tr>
<td></td>
<td>100% over subscription (1 HT core = 1 VM)</td>
<td>368</td>
</tr>
<tr>
<td></td>
<td>200% over subscription</td>
<td>552</td>
</tr>
</tbody>
</table>

---

**VM states as reported by “virsh list”**

<table>
<thead>
<tr>
<th></th>
<th>Maximum</th>
<th>Average</th>
<th>Minimum</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>records</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>systems</td>
<td>23</td>
<td>23</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>users</td>
<td>43</td>
<td>24</td>
<td>9</td>
<td>43</td>
</tr>
<tr>
<td>elapsed</td>
<td>305</td>
<td>31</td>
<td>17</td>
<td>32</td>
</tr>
</tbody>
</table>

---

Data for fermi states as reported by virsh list
Plot generated 28-Feb-2012
Note - 1 physical core = 1 VM
# Description of Virtual Machine States

Reported by "virsh list" Command

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>running</td>
<td>The domain is currently running on a CPU. Note – KVM based VMs show up in this state even when they are “idle”.</td>
</tr>
<tr>
<td>idle</td>
<td>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”.</td>
</tr>
<tr>
<td>paused</td>
<td>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.</td>
</tr>
<tr>
<td>shutdown</td>
<td>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.</td>
</tr>
<tr>
<td>shut off</td>
<td>The domain has been shut down. When in a shut off state the domain does not consume resources.</td>
</tr>
<tr>
<td>crashed</td>
<td>The domain has crashed. Usually this state can only occur if the domain has been configured not to restart on crash.</td>
</tr>
<tr>
<td>dying</td>
<td>The domain is in process of dying, but hasn't completely shutdown or crashed.</td>
</tr>
</tbody>
</table>
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.

• 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).
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](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

Fermilab Gratia Collector/Reporter

ONE DB

onevm_query

ONE API

Probe

Config

Gratia_onevm

MySQL Database

Collector

Report

MySQL Database

FermiGrid and FermiCloud Update

28-Feb-2012
Here are the **preliminary** results of “replaying” the previous year of the OpenNebula “OneVM” data into the new accounting probe:
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).
  – See his talk on Friday 2-Mar-2012 for more details.
### MPI on FermiCloud (Note 1)

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

Notes:
1. Work performed by Dr. Hyunwoo Kim of KISTI in collaboration with Dr. Steven Timm of Fermilab.
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.
FermiCloud – Network & SAN

“Today”

Private Ethernet
over dedicated fiber
Fibre Channel

FY2011 / FY2012

Note – Dashed Lines = Waiting for Equipment or Fiber
FermiCloud – Network & SAN (Possible Future – FY2013/2014)

Fibre Channel

FCC-2
Brocade
fcl0yy
To
fcl0zz

GCC-A
Brocade
fcl0xx
To
fcl0yy

GCC-B
Brocade
fcl001
To
fcl015

FCC-3
Brocade
Satabeast
fcl316
To
fcl330

Nexus 7010
FCC-2

Nexus 7010
GCC-A

Fibre Channel

Fibre Channel
FermiCloud – Support for Science (direct and indirect)

• FermiCloud resources are routinely being used by Grid middleware developers and integrators as well as CDF, D0 and the Cosmic and Intensity Frontiers.

• FermiCloud resources were used to perform a detailed analysis of filesystem performance for Grid applications (See talk on Thursday 1-Mar-2012).

• FermiCloud resources were used to determine the optimal service configurations for the XACML AuthZ profile (See talk on Wednesday 29-Feb-2012).

• FermiCloud resources are currently being used to examine NFS v4 performance and will shortly be used to explore the implications of IPv6 for Grid and Cloud applications and middleware.

• FermiCloud resources are being used by the OSG Software team to “re-factor” the OSG software distribution from a pacman based distribution to an RPM based distribution.

• Production services have been deployed on FermiCloud, including:
  – Specialized GridFTP virtual machines for the Intensity Frontier experiments,
  – NFS Administration virtual machines to allow Intensity Frontier experiments to perform file ownership management of their BlueArc volumes.
FermiCloud Summary

• 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.
Overall Summary

• FermiGrid continues to be a world class infrastructure for scientific computing:
  – Energy Frontier Experiments – CDF, D0 & CMS,
  – Intensity Frontier (neutrino) and Cosmic Frontier (astrophysics) experiments,
  – Opportunistic Computing (OSG and WLCG Virtual Organizations).

• 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.
Thank You!

Any Questions?
Extra Slides
# FermiCloud vs. Magellan

<table>
<thead>
<tr>
<th>FermiCloud</th>
<th>Magellan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom up requirements and design.</td>
<td>Top down requirements and mission.</td>
</tr>
<tr>
<td>Multi phase project, with each phase building on knowledge gained during previous phases.</td>
<td>Fixed term project without ongoing funding.</td>
</tr>
<tr>
<td>Evaluated available open source cloud computing frameworks (Eucalyptus, Nimbus, OpenNebula) against requirements and selected OpenNebula. We plan to “circle back” and evaluate OpenStack this year.</td>
<td>Spent a lot of time trying to get the open source version of Eucalyptus to work at scale, eventually switched to a combination of Nimbus and OpenStack late in the project.</td>
</tr>
<tr>
<td>Approached cloud computing from a Grid and high throughput computing (HTC) perspective.</td>
<td>Approached cloud computing from a high performance computing (HPC) perspective.</td>
</tr>
<tr>
<td>Significant prior experience delivering production Grid services via open source virtualization (Xen and KVM).</td>
<td>Unclear.</td>
</tr>
<tr>
<td>Have SRIOV drivers for InfiniBand.</td>
<td>Did not have SRIOV drivers for InfiniBand before the end of the project.</td>
</tr>
<tr>
<td>Actively sought collaboration (OpenNebula, KISTI).</td>
<td>Project was sited at NERSC and Argonne.</td>
</tr>
</tbody>
</table>