



U.S. DEPARTMENT OF
ENERGY

Office of
Science

On-demand Services for the Scientific Program at Fermilab

Gabriele Garzoglio

Grid and Cloud Services Department, Scientific Computing Division,
Fermilab

ISGC – Thu, March 27, 2014

Overview

- Frontier Physics at Fermilab: Computing Challenges
- Strategy: Grid and Cloud Computing
- On-Demand Services for Scalability: Use Cases
- The Costs of Physics at Amazon

The Cloud Program Team

- **Fermilab Grid And Cloud Services Department**
 - Gabriele Garzoglio (Dept. Head)
 - Steven Timm (FermiCloud Project lead)
 - Gerard Bernabeu Altayo (lead FermiCloud operations)
 - Hyun Woo Kim (lead FermiCloud development)
 - Tanya Levshina, Parag Mhashilkar, Kevin Hill, Karen Shepelak (Technical Support)
 - Keith Chadwick (Scientific Computing Division Technical Architect)
- **KISTI-GSDC team**
 - Haeng-Jin Jang, Seo-Young Noh, Gyeongryoon Kim
- **Funded through Collaborative Agreement with KISTI**
 - Nick Palombo (consultant)
 - Hao Wu, Tiago Pais (2013 summer students, IIT)
 - Francesco Ceccarelli (2013 summer student, INFN Italy)
- **Professors at Illinois Institute of Technology (IIT)**
 - Ioan Raicu, Shangping Ren

Motivation: Computing needs for IF / CF experiments

- Requests for increasing slot allocations
- Expect stochastic load, rather than DC
- Focus on concurrent peak utilization
- Addressing the problem for IF / CF experiments as an ensemble under FIFE
- Collaborating with CMS on computing and infrastructure

Fermilab Scientific Computing Review*

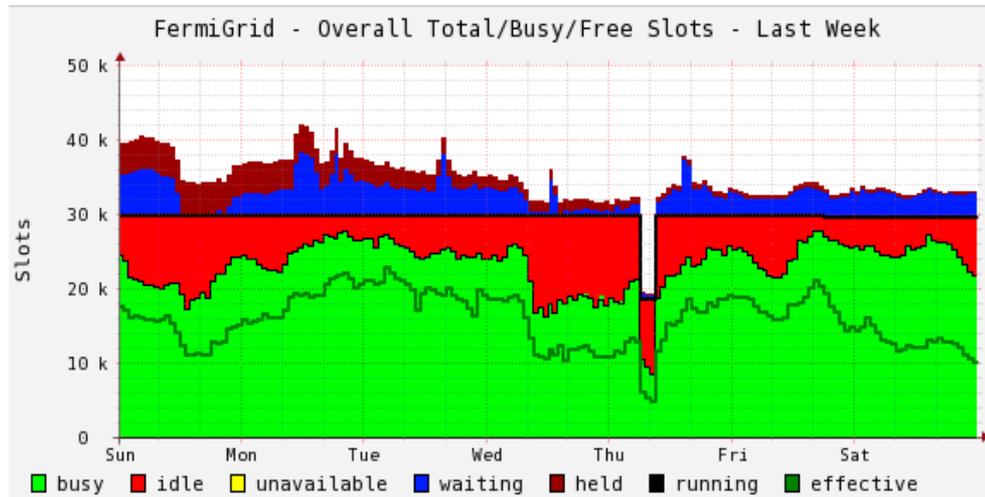
Experiment	Allocation (Slots)	Average Utilization (last quarter)		FY14 Request (Slots)
		Average Slots	Peak Slots	
ArgoNeuT	200	75	1712	0
Muon g-2	200	9	195	0
LBNE	500	30	1202	200
MARS	1225	212	1444	100
MicroBooNE	500	39	597	100
Minerva	1600	667	3580	500
Minos	1200	294	2427	0
Mu2e	500	597	2491	500
Nova	1300	410	1799	500
Total Requested	7225	2333	15447	1900

* Number of slots have been updated periodically since

“Fabric for Frontier Experiments at Fermilab” – Thu 3/27 14:40 – Rm2

Slots: Fermilab, OSG, & Clouds

- Current full capacity of FermiGrid \rightarrow \sim 30k slots
(Clusters: General Purpose, CMS, CDF, DZero)



30k Slots
FermiGrid

- Full capacity of OSG \rightarrow \sim 85k slots
- Additional OSG opportunistic slots \rightarrow 15k – 30k
- Additional per-pay slots at commercial Clouds
- **Porting experiments' computing to distributed resources is part of the strategy.**

Grid vs. Clouds (...in my world...)

Grid	Commercial Cloud	Community Cloud
Statically-allocated federated resources	Dynamically-allocated resources	
Diverse per-site computing environments	User-defined computing environment	
Mature workload management systems using Grid interfaces	Allocation and provision being integrated in workload management through diverse interfaces (EC2 emulation, Google, OpenStack, ...)	
Lot of capacity	Lot of capacity	Growing capacity
Dedicated slots: “free” at collaborating sites (after site onboarding)	Dedicated slots: ~\$0.12 CPU h ~\$0.09 – \$0.12 GB	Dedicated slots: “free” at (a few) collaborating sites. ~\$0.04 / CPU h @ BNL ~\$0.03 / CPU h @ FNAL
Opportunistic slots: “free” after onboarding; no guarantee	Spot pricing: bid your slot ~\$0.01 (better guarantee with higher bid)	Model not developed yet

On-demand Services for Intensity and Cosmic Frontiers

Goal: each experiment can access $O(10k)$ slots on distributed resources when needed

Problem: How many supporting services of type X do we deploy for experiment Y ?

Use cases for service X:

- ...**batch computation**
 - data transfer, submission nodes, coordinated clusters, DB proxies, WN for Grid-bursting, ...
- ...**interactive sessions**
 - build nodes, WN-like environment for debugging, DAQ emulation, ...
- ...**advanced use cases**
 - private batch systems, private-net clusters, Hadoop / MapReduce, generate customized VM, ...

Solution platform: dynamically instantiated virtualization

Dynamic provisioning drivers

- **Workloads become more heterogeneous**
 - Very large data sets (DES, Darkside)
 - Consecutive tasks on same large data set (DES)
 - Large memory requirements (LBNE)
 - Large shower libraries (Neutrino)
 - Order of minutes execution on O(30GB) data.
 - Whole node scheduling
- Need flexible ways to **reconfigure compute clusters and supporting services** for changing workload
- In FY16: 8 major experiments at Fermilab from 3 frontiers. Usage patterns: hard to plan over “slow” (yearly?) cycles. **Need to be more dynamic.**

Cloud Computing Environments

FermiCloud Services

- Production quality Infrastructure-as-a-Service based on OpenNebula. <http://www-fermicloud.fnal.gov>
- Made technical breakthroughs in authentication, authorization, fabric deployment, accounting and high availability cloud infrastructure
- Continues to evolve and develop cloud computing techniques for scientific workflows in collaboration with KISTI, S. Korea.
- Ref: ISGC 2011/12/13, ISGC12 Interop. Workshop, Keith Chadwick

Amazon Web Services

- Startup allocation to integrate commercial clouds

Community Clouds

- Resources available through collaborative agreements (KISTI, OSG, OpenStack Federation, etc.)

§1 - On-demand storage & data movement services

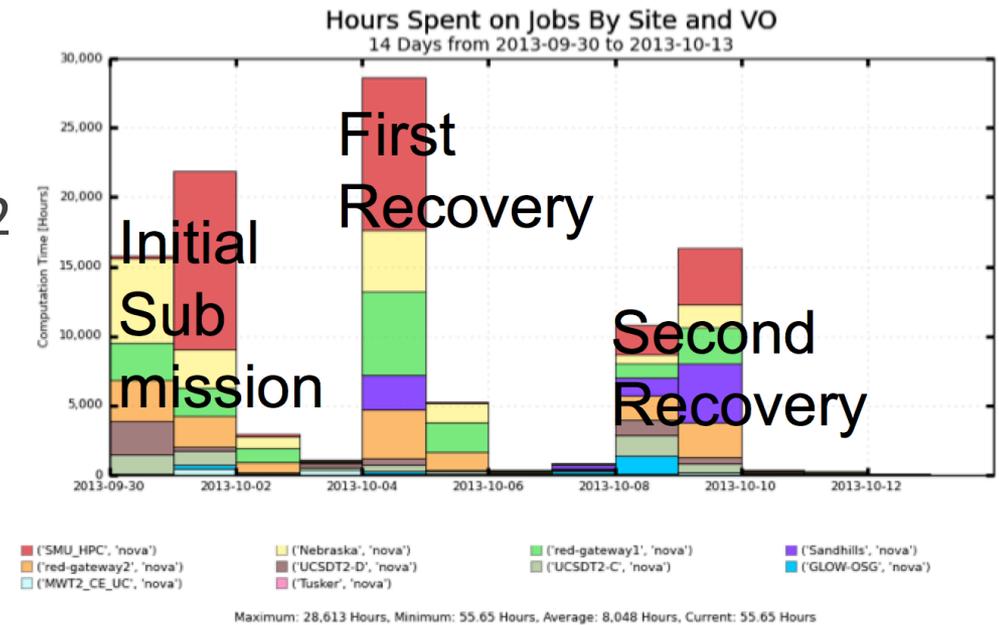
- Stochastic load of data movement
 - Different experiments running different workflows at different times
- Locally on FermiCloud...
 - scale experiment-specific data movement servers (GridFTP, xrootd, dCache doors, ...) with expected load
 - configure servers to work as an ensemble, gathering them up under a common entry gateway (SRM / BeStMan, cluster-level xrootd redirector, ...)
- At the Grid / Cloud site...
 - Helps to have a temporary storage element on remote grid/cloud site as well
 - Web caching and database proxies on the remote grid/cloud site

§2 - On-demand coordinated cluster

- Launch batch system head node + DB proxy + storage
 - Instantiate dynamically-provisioned WN depending on the characteristics of the workflow and current load
 - Supports legacy clusters, heterogeneous workflows, overlay networks
- Launch interactive computing environment with supporting services
 - CMS T3 “in-a-box”

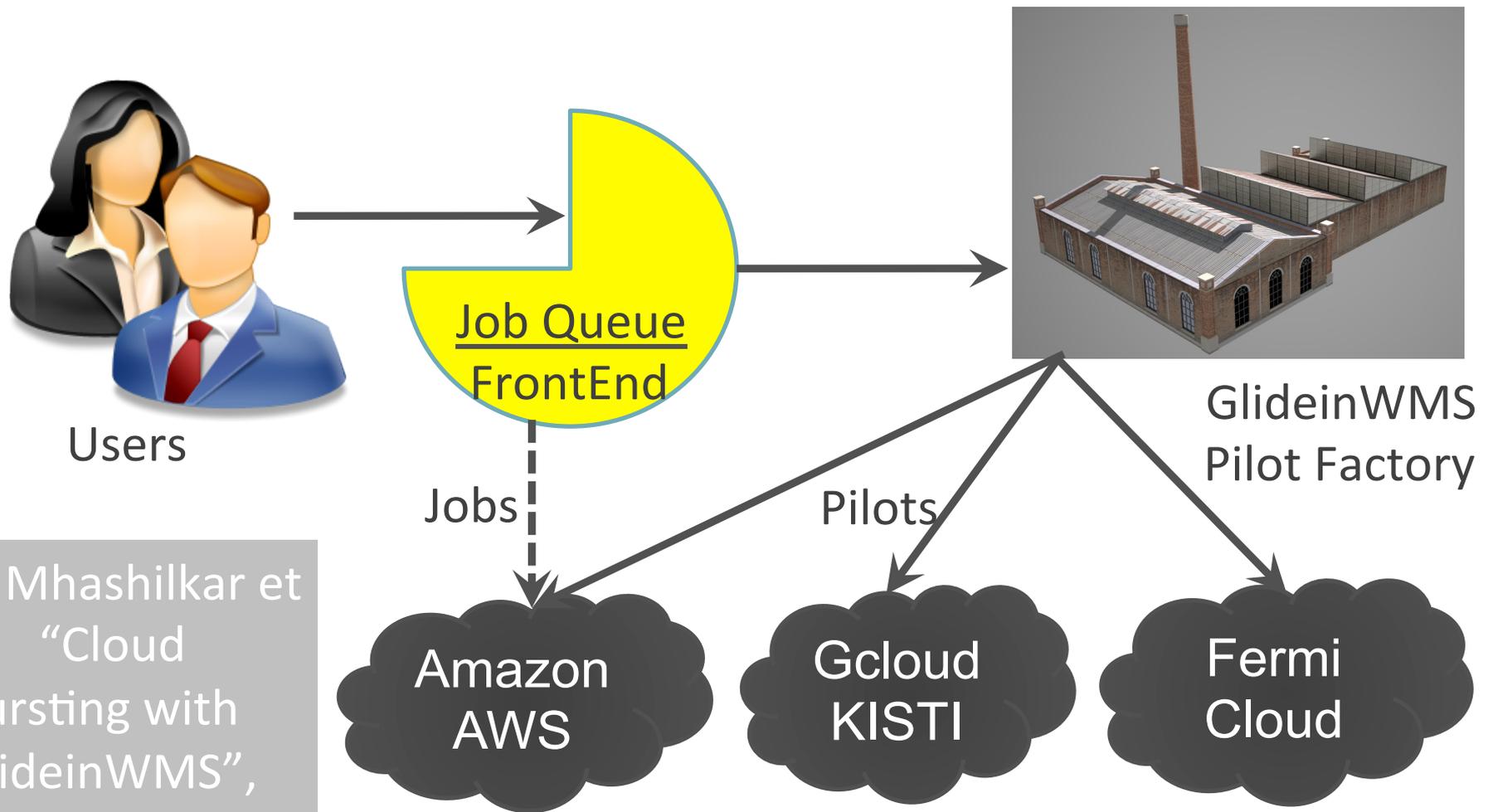
§3 - On-demand Grid-bursting

- Add WN from clouds to local batch systems
- 1,000,000 ev generated with ~10k jobs for 88,535 CPU h in 2 weeks of ops and 2 TB of data
- Run on OSG at SMU (dedicated), UNL, Uwis, UC, UCSD and O(100) jobs at FermiCloud
- Operations consisted of 1 submission + 2 recoveries
- Spent about 10% more resources than expected due to preemption



- We ran 400 jobs of this app on Amazon Web Services
- How much would it cost to run the full campaign on AWS ?

GlideinWMS – Grid Bursting



P. Mhashilkar et al, "Cloud Bursting with GlideinWMS", CHEP 2013

AWS prices

- Targeting use of AWS general purpose instances

Instance Family	Instance Type	Processor Arch	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	EBS-optimized Available	Network Performance
General purpose	m1.small	32-bit or 64-bit	1	1	1.7	1 x 160	-	Low
General purpose	m1.medium	32-bit or 64-bit	1	2	3.75	1 x 410	-	Moderate

- 1 ECU = 1.2 GHz Xeon in 2007
- Options: std instances and spot pricing
- FermiCloud economic model: \$0.03/h**
- HEPiX-Fall13 T. Wong: cost at RACF BNL: \$0.04 / h**
- ...watch out for data (network providers are negotiating with Clouds)

Standard Instances (\$/h)	
m1.small	0.06
m1.medium	0.12
Standard Spot Instances (\$/h)	
m1.small	0.007
m1.medium	0.013
Spot Instances (us-west-2c) Mar 12, 2014 (\$/h)	
m1.small	0.01
m1.medium	0.021
Data Transfer OUT From Amazon EC2 To Internet (\$ / GB)	
First 1 GB / month	0
Up to 10 TB / month	0.12
Next 40 TB / month	0.09
Next 100 TB / month	0.07
Next 350 TB / month	0.05

AWS / OSG Worker Node Equivalence

- AWS m1.medium \approx “Standard” OSG WN
 - Run 400 NOvA Monte Carlo jobs generating 100 events each on m1.medium (not enough memory on m1.small)

Site	Failed Jobs	RAM (GB)	Wall Avg (min)	Wall mn/mx (min)	Wall Tot (hr)	Trans (GB)	Cost
UNL via OSG	0	2.16	28.95	22/44	3.05	8.3	
FermiCloud	3	2.05	38.03	20/72	3.3	8.2	
AWS (m1.medium)	0	2.1	34.1	25/44	3.25	8.2	\$42

- HEPiX-Fall13 T. Wong (BNL): “EC2 jobs on m1.small took ~50% longer than on dedicated facility”

Cost range for NOvA event generation

- 1,000,000 events, 2TB data, 90,000 CPU h
 - Data transfer cost = ~ **\$240**
 - Std m1.medium Instance: **\$11k**
 - Spot price: **\$2k**
 - **FermiCloud-equivalent cost estimate: \$3k**
 - **RACF-equivalent cost estimate: \$4k**
- Next Test – MC generation of cosmics in the NOvA Near Detector
 - Jobs: 1000 jobs for ~1h / job & 250,000 events / job
 - Data: output per job: ~100MB. Total size 100GB
 - Estimate 1,000 CPU hours in a day with 100 VM
 - Std m1.medium: **\$130** (**\$12** data charges)
 - Spot price: **\$24** (**\$12** data charges)
 - **FermiCloud-equivalent cost estimate: \$30**
 - **RACF-equivalent cost estimate: \$40**

Recent publications

- Automatic Cloud Bursting Under FermiCloud
 - ICPADS CSS workshop, Dec. 2013.
- A Reference Model for Virtual Machine Launching Overhead
 - CCGrid conference, May 2014
- Grids, Virtualization, and Clouds at Fermilab
 - CHEP conference, Oct. 2013.
- Exploring Infiniband Hardware Virtualization in OpenNebula towards Efficient High-Performance Computing.
 - CCGrid Scale workshop, May 2014.

Summary

- Computing on distributed resources is strategic for the Intensity and Cosmic Frontiers experiments
- We will support these experiments with the Grid and Cloud paradigms
- We observe the emergence of strong drivers for on-demand service deployment
 - more heterogeneous workloads, cluster profile adaptation, ...
- Running jobs for the Intensity Frontier experiments on Cloud is successful at initial small scale