

Fermilab Grid Facility Department & KISTI: Opportunities for Collaborations

Overview

- Fermilab, the Computing Division, and the Grid Facility Department
- Opportunities for Collaboration

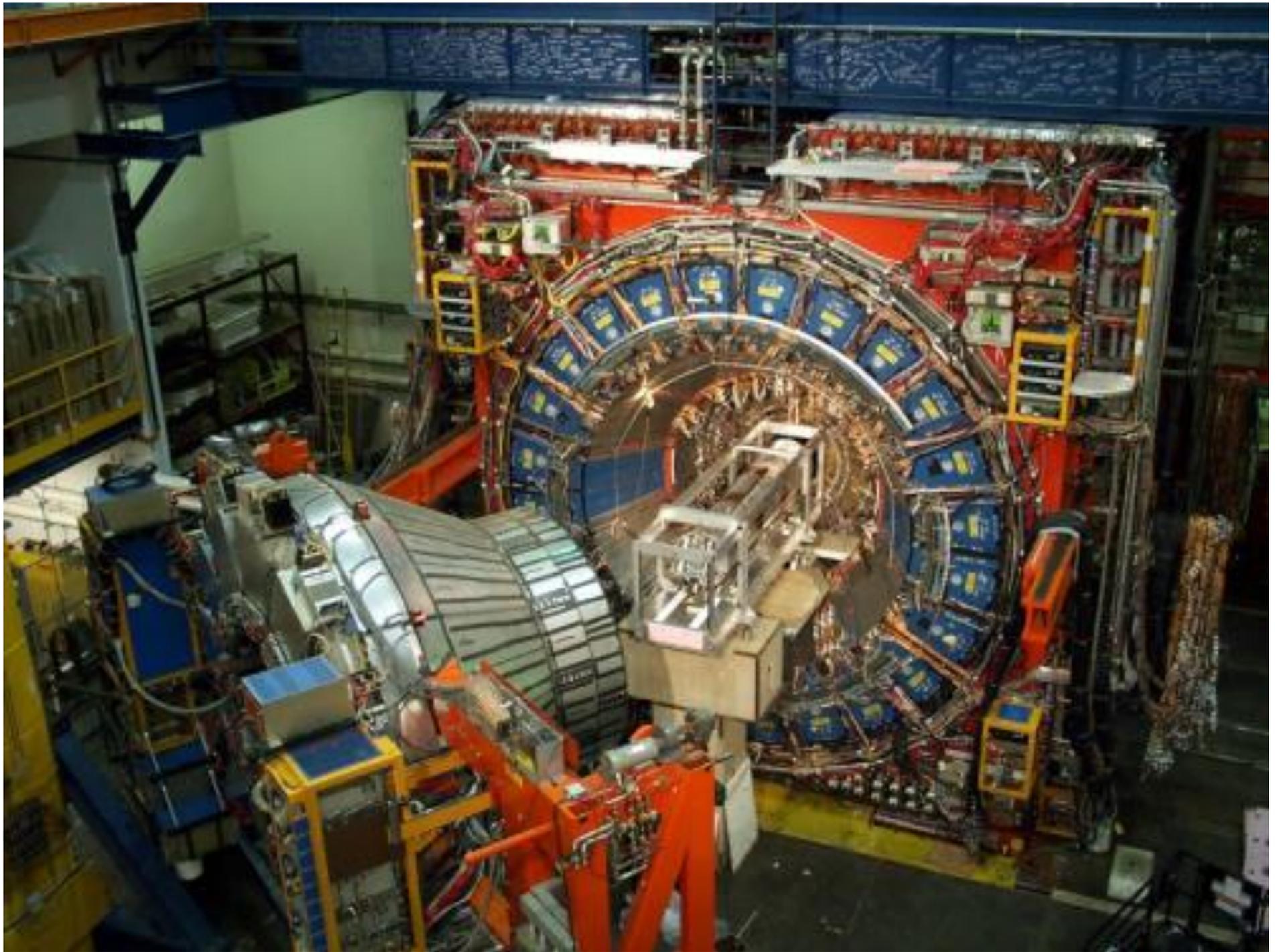
Meeting with KISTI
Oct 26, 2010

Gabriele Garzoglio
Grid Department, Associate Head
Computing Division, Fermilab

Fermi National Accelerator Laboratory







The Energy Frontier

Origin of Mass

Matter/Anti-matter
Asymmetry

Dark Matter

Origin of Universe

Unification of Forces

New Physics
Beyond the Standard Model

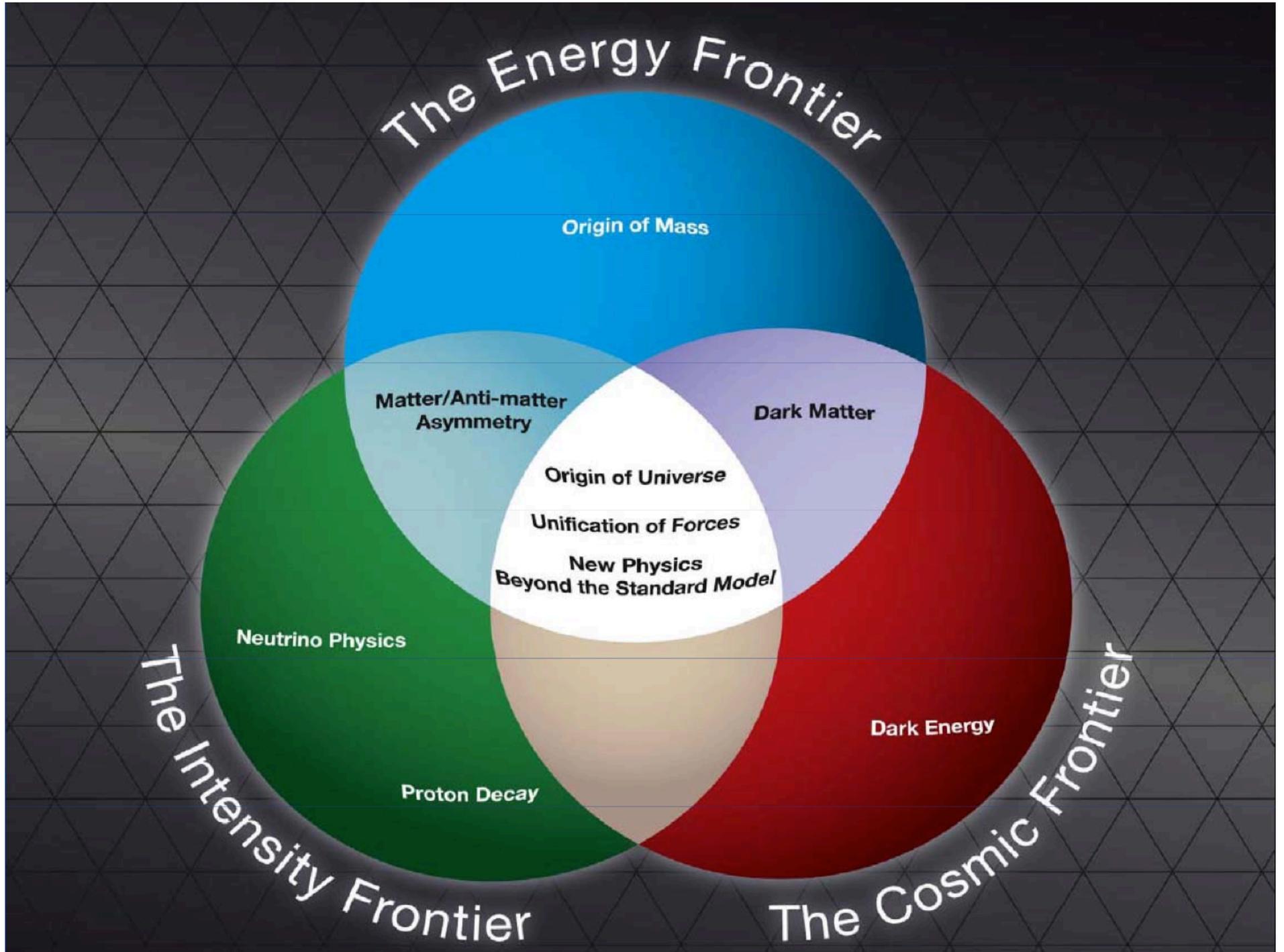
Neutrino Physics

Dark Energy

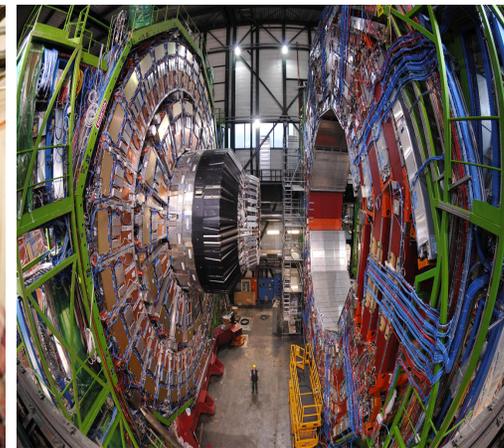
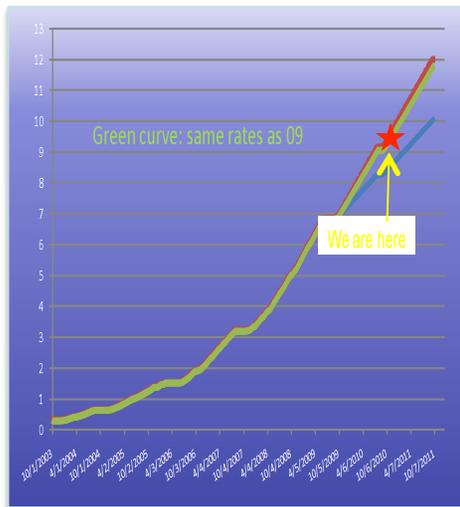
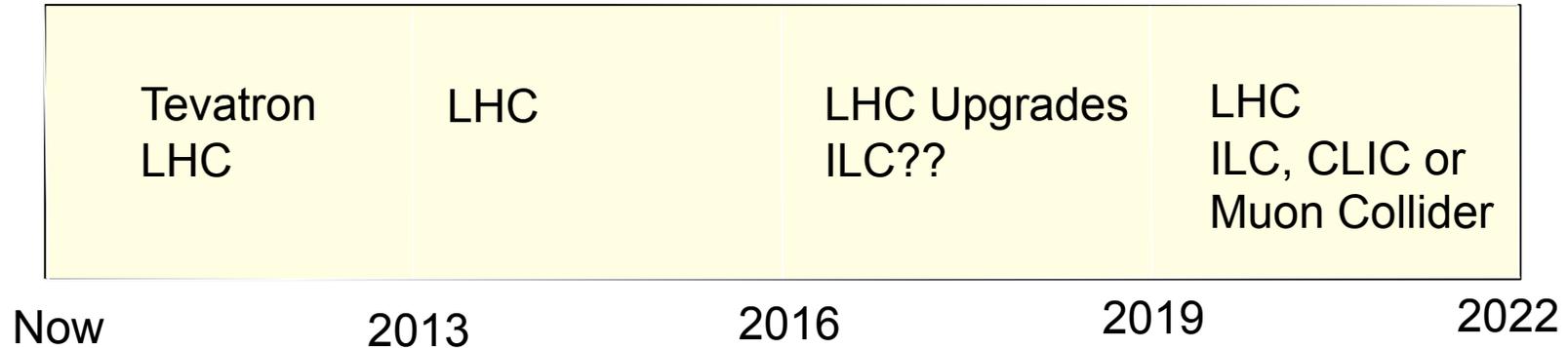
Proton Decay

The Intensity Frontier

The Cosmic Frontier



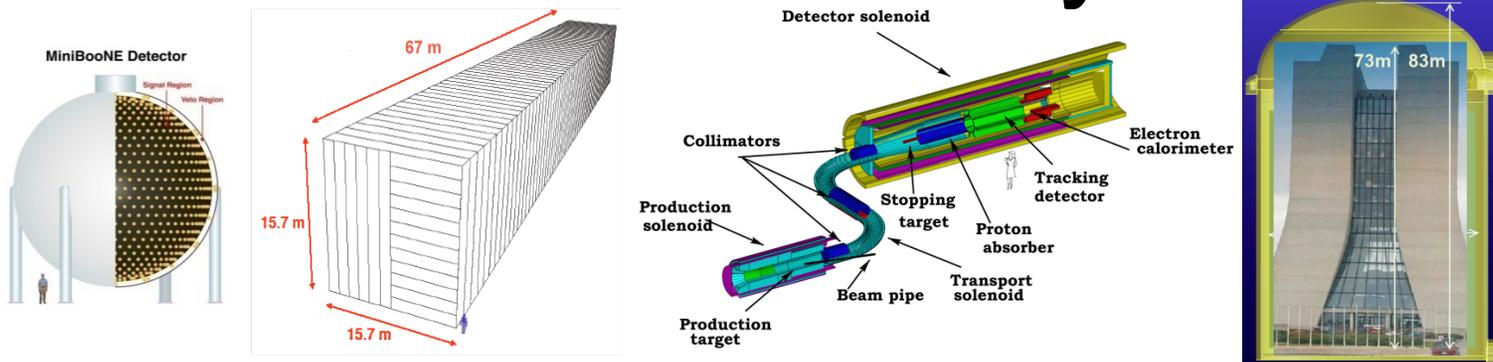
Present Plan: Energy Frontier



P. Oddone, DRAFT P5 Meeting, October 15, 2010

Oct 26, 2010

Present Plan: Intensity Frontier



MINOS MiniBooNE MINERvA SeaQuest	NOvA MicroBooNE g-2? SeaQuest	LBNE Mu2e	Project X+LBNE μ , K, nuclear, ... ν Factory ??
-------------------------------------------	----------------------------------------	--------------	---------------------------------------------------------------

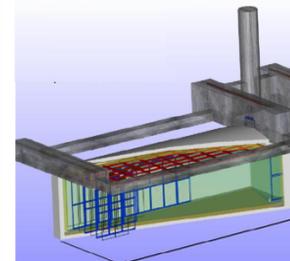
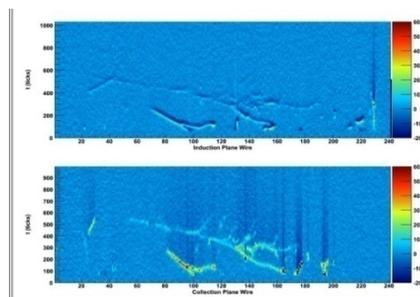
Now

2013

2016

2019

2022



P. Oddone, DRAFT P5 Meeting, October 15, 2010

7

The Fermilab Computing Division

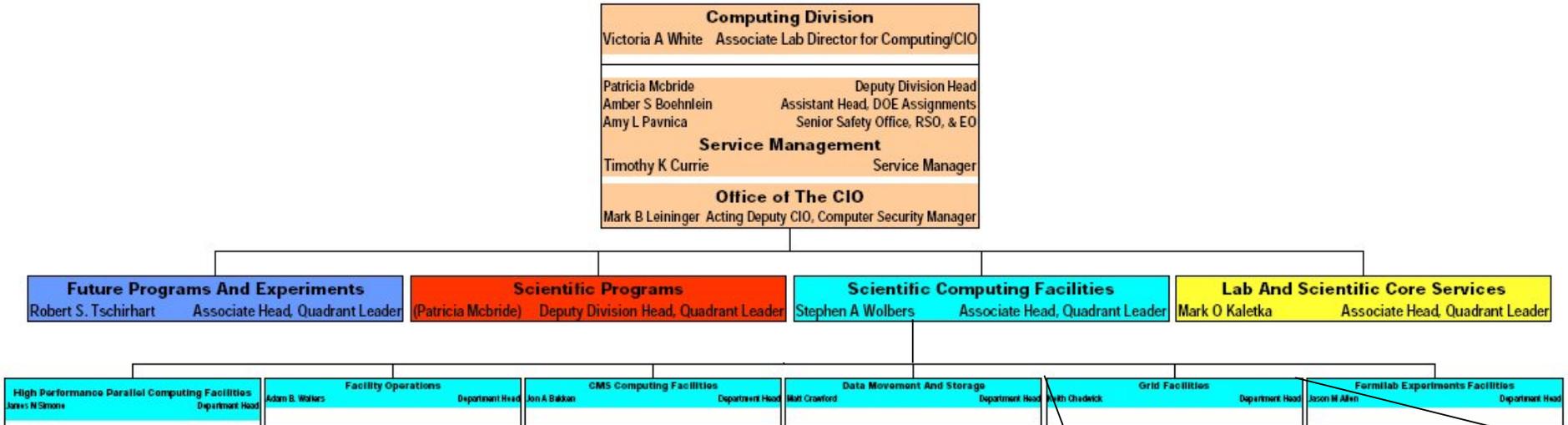


Computing Division Direct support for:	Energy Frontier	Intensity Frontier	Cosmic Frontier
	CDF, D0, LHC	MINOS, MiniBOONE, MINERvA, SciBooNE	SDSS, P. Auger, CDMS, COUPP
	Theory (Lattice QCD)	Theory (Lattice QCD)	DES, Auger N?
	LHC upgrades	MicroBooNE, NOVA, Argoneut, g-2?	Holometer, Other initiatives?
	ILC or Muon Collider ?	Mu2e, LBNE, Project X	JDEM?
Scientific Software	Scientific Experiment Computing	High Performance Computing	DAQ and Trigger
Geant4 Generators Frameworks Accel.modeling HPC tools	FermiGrid, FermiCloud Open Science Grid CMS T1 center, LPC CAF & ROC Data Storage & access	USQCD Facility Accelerator Modeling cluster Computational Cosmology cluster	NOvA DAQ Optical Links Pixel Telescope Future DAQ CCD readout
Physical Computing Facilities	IT Services	Document and Information Systems	Networks
Space Power Cooling	Email, Web, Service Desk, CyberSecurity, Windows, Mac, Linux infrastructure	For lab business For Projects For managing the lab	Campus Network Scientific Data movement, Advanced WAN

CD by the Numbers...

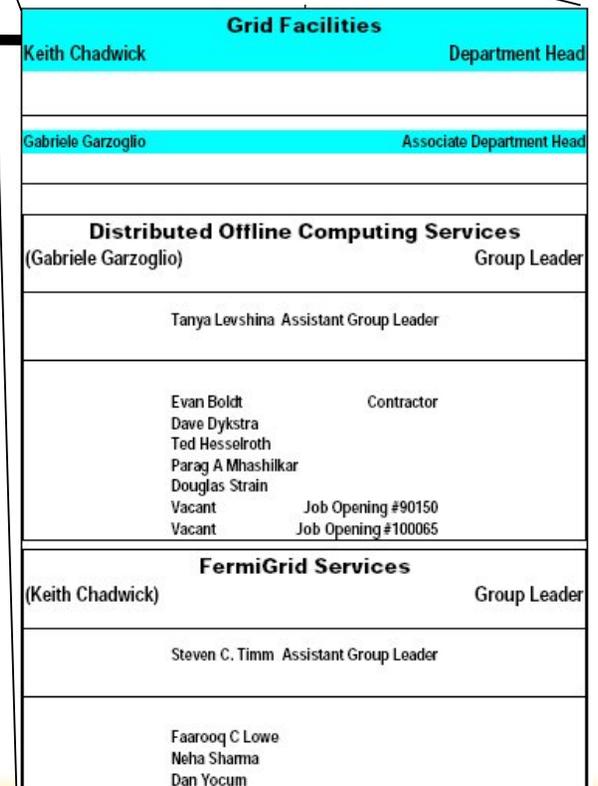
- The Computing Division (CD) operates...
 - 9 computer rooms (FCC 1, GCC A, LCC, ...)
 - 3 communications rooms
 - 3,000 m² of space for computing equipment.
- CD manages...
 - > 6500 computers (multi-CPU, multi-Core)
 - > 1 PetaByte of disk storage (50% for CMS)
 - 12 tape robots
 - > 3.5 megawatts (MW) of power
- CD has 20 PB of data as of Jan 2010 (in 2 buildings: FCC and GCC)
- 2 WAN connections (for redundancy):
 - One from GCC to LCC, through other points onsite, then out near Giese and Kirk Rds to Argonne
 - One from FCC to Northwestern U's Starlight hub, a 1GigE and 10GigE switch/router facility for high-performance access to participating networks

Fermilab Grid Facilities Department & KISTI: Opportunities for Collaborations



To establish and maintain Fermilab as a high performance, robust, highly available, expertly supported and documented Premier Grid Facility which supports the scientific program based on Computing Division and Laboratory priorities.

The department provides production infrastructure and software, together with leading edge and innovative computing solutions, to meet the needs of the Fermilab Grid community and takes a strong leadership role in the development and operation of the global computing infrastructure of the Open Science Grid.



Opportunities for Collaboration

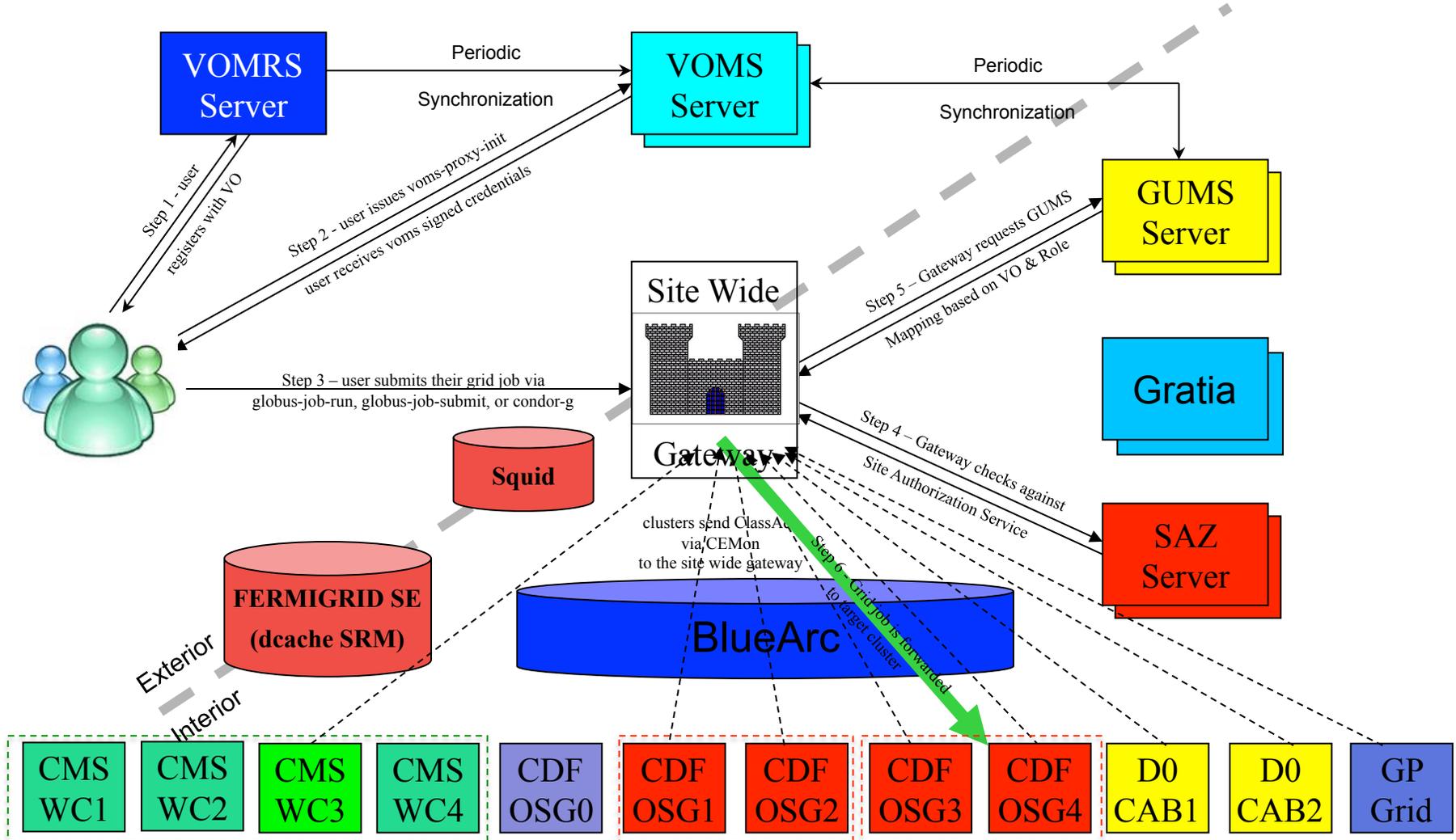
- **FermiGrid**
 - The Fermilab Campus Grid
- **FermiCloud**
 - The Fermilab Infrastructure as a Service
- **GlideinWMS**
 - Grid-wide overlay batch system
- **Center for Enabling Distributed Petascale Science**
 - CEDPS at Fermilab: data stage-out on overlay batch systems
- **OSG Storage**
 - Grid-wide data management
- **Gratia**
 - Grid accounting
- **Metric Correlation & Analysis Services**
 - Knowledge representation on a dashboard

FermiGrid

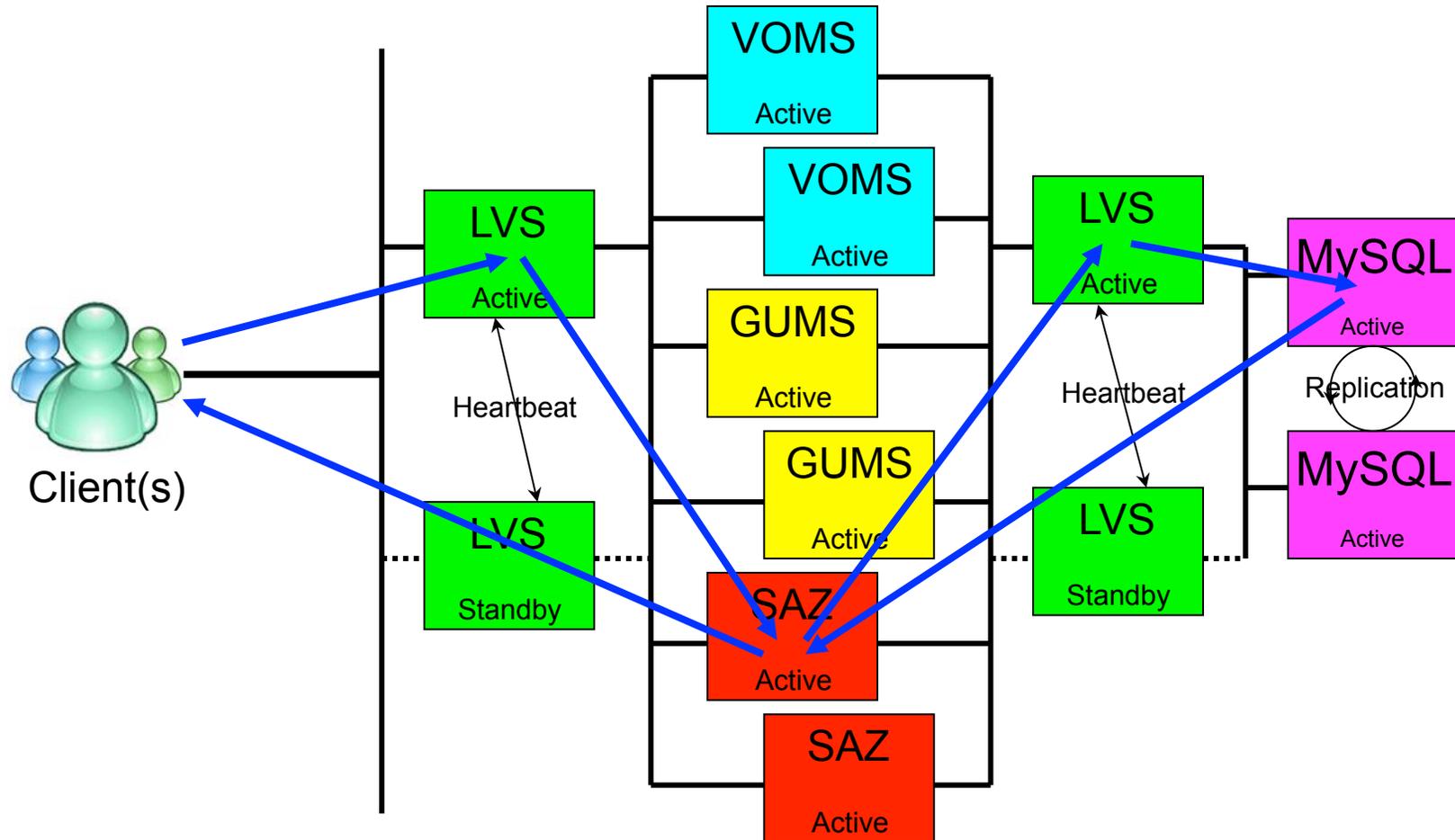
- Site Wide Globus Gatekeeper (FNAL_FERMIGRID).
- Centrally Managed Services (VOMS, GUMS, SAZ, MySQL, MyProxy, Squid, Accounting, etc.)

Compute Resources	# Clusters	# Gatekeepers	Batch System	# Batch Slots
CDF	3	5	Condor	5685
D0	2	2	PBS	5305
CMS	1	4	Condor	6904
GP	1	3	Condor	1901
Total	7	15	n/a	~19,000
Sleeper Pool	1	2	Condor	~14,200

FermiGrid - Architecture

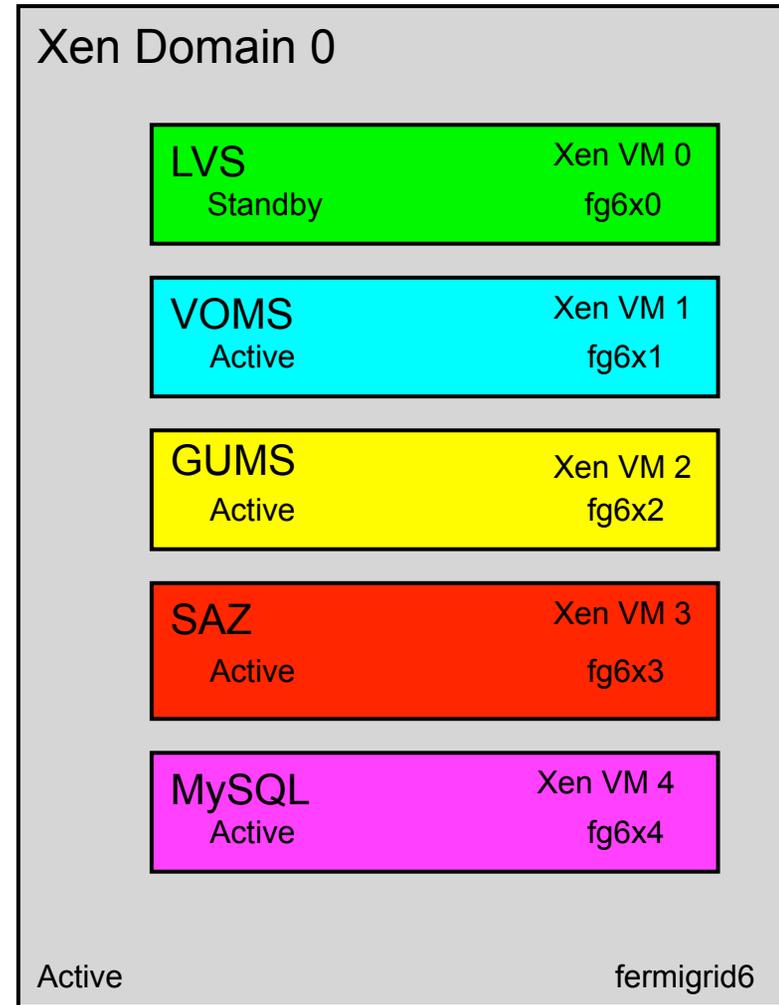
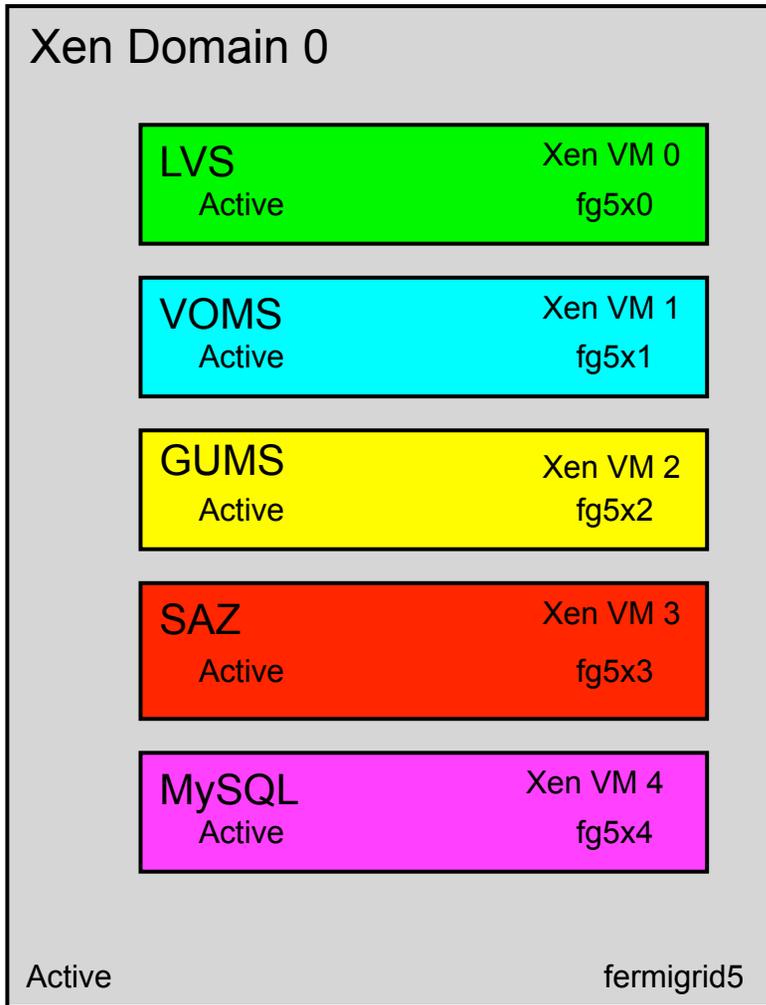


FermiGrid HA Services - 1

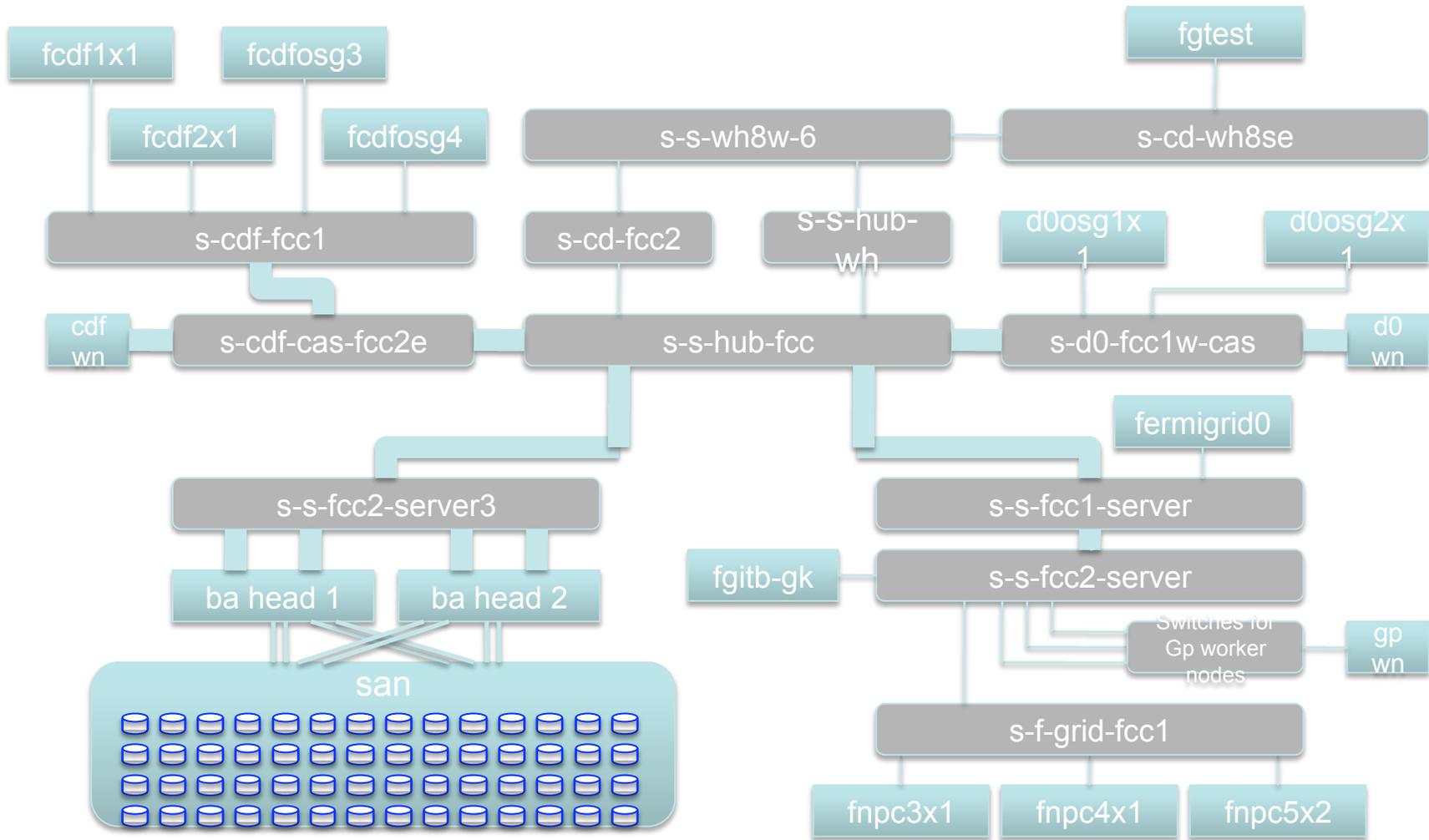


Campus Grids

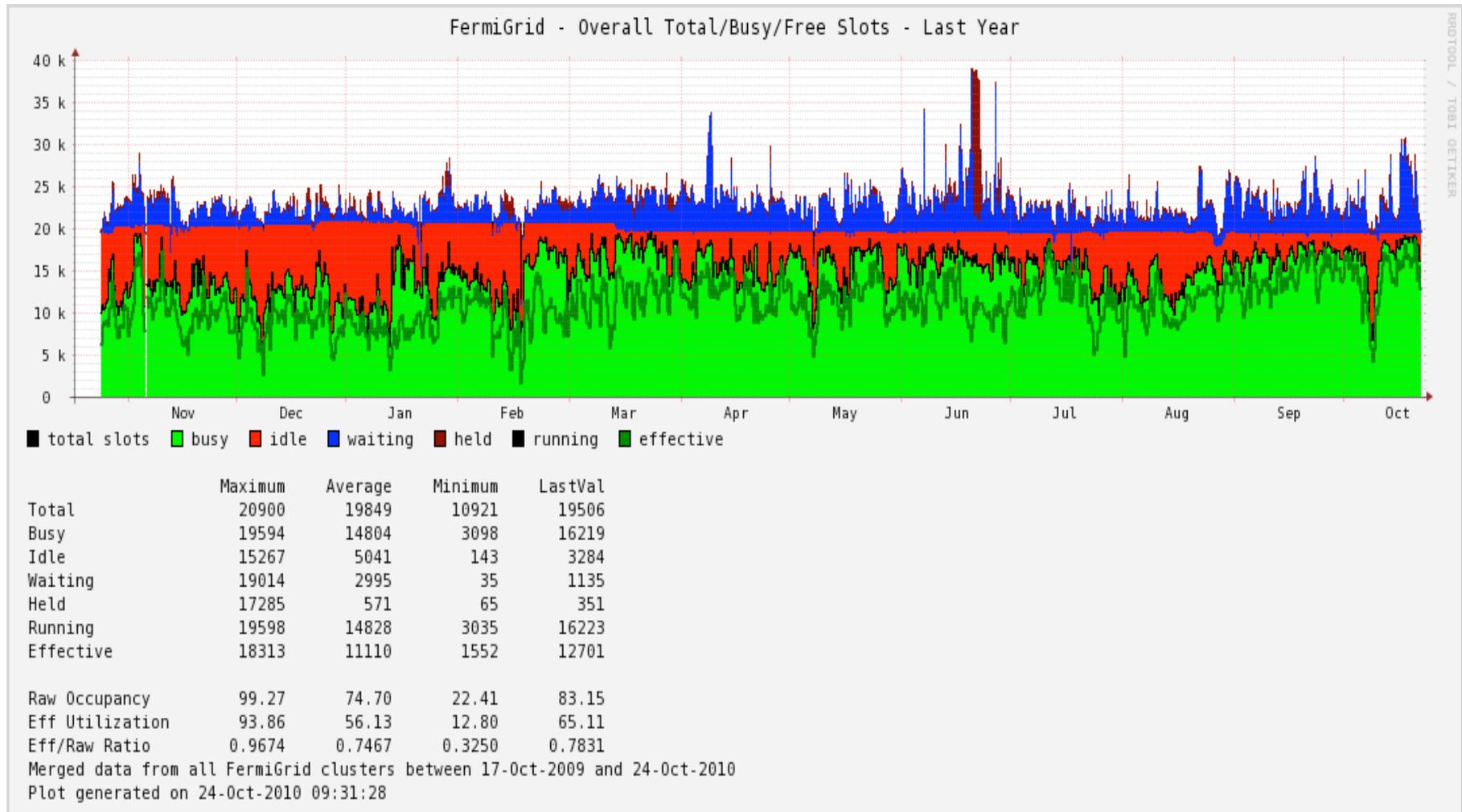
FermiGrid HA Services - 2



(Simplified) FermiGrid Network

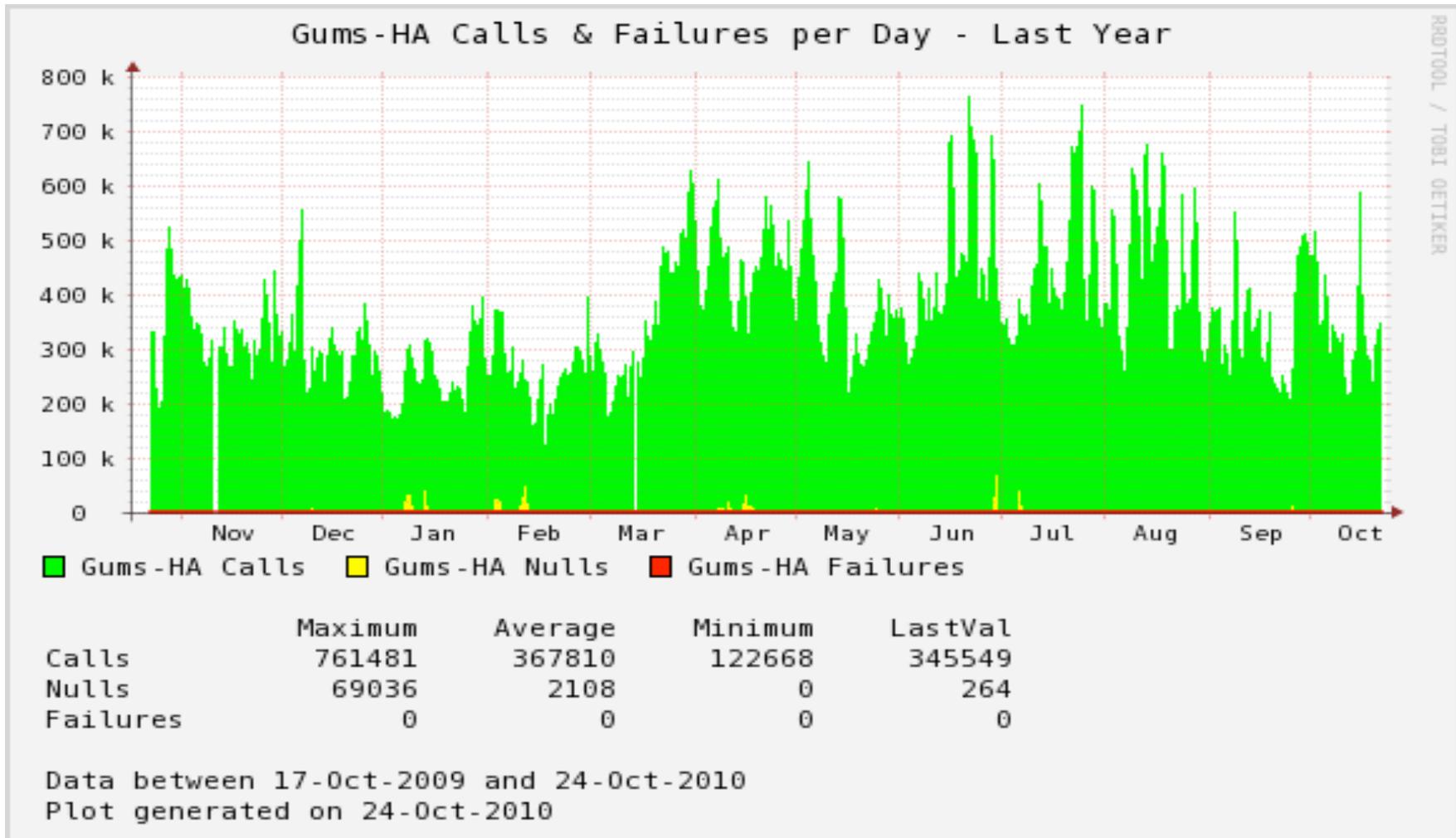


FermiGrid Utilization

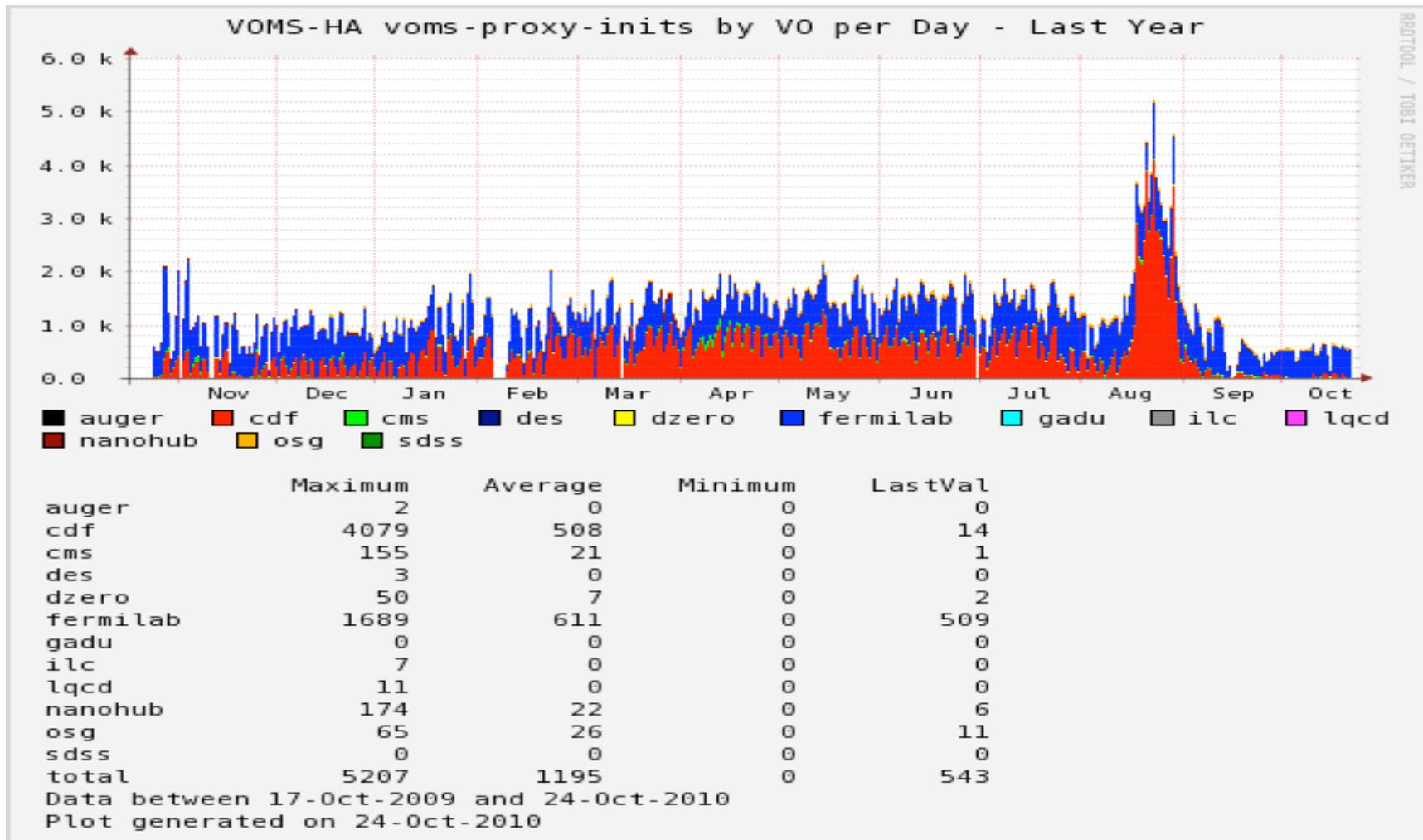


ROOT/TOOL / TOBI OETIKER

GUMS calls



VOMS-PROXY-INIT calls



Opportunities for Collaboration

- **FermiGrid**
 - The Fermilab Campus Grid
- **FermiCloud**
 - The Fermilab Infrastructure as a Service
- **GlideinWMS**
 - Grid-wide overlay batch system
- **Center for Enabling Distributed Petascale Science**
 - CEDPS at Fermilab: data stage-out on overlay batch systems
- **OSG Storage**
 - Grid-wide data management
- **Gratia**
 - Grid accounting
- **Metric Correlation & Analysis Services**
 - Knowledge representation on a dashboard

FermiCloud

- A private Infrastructure-as-a-service for...
 - ...grid and storage developers, integrators, testers
 - ...production services.
- VM are created on demand with no expiration time
- Excess is used by opportunistic scientific computing
- Machines have to meet all Fermilab security requirements
- The pilot service runs OpenNebula and Eucalyptus with KVM from SL5.5

FermiCloud Project

- Technology Evaluation and Evolution
 - Hypervisors
 - Open Source and Commercial Cloud Control
 - **Authentication** and **Authorization**
 - Software **Provisioning** and **Contextualization**
 - **Scheduling**
 - File Systems
- Requirements Gathering
 - HW and SW technical
 - Security
 - Stakeholder needs
- Customization and Deployment

Hardware

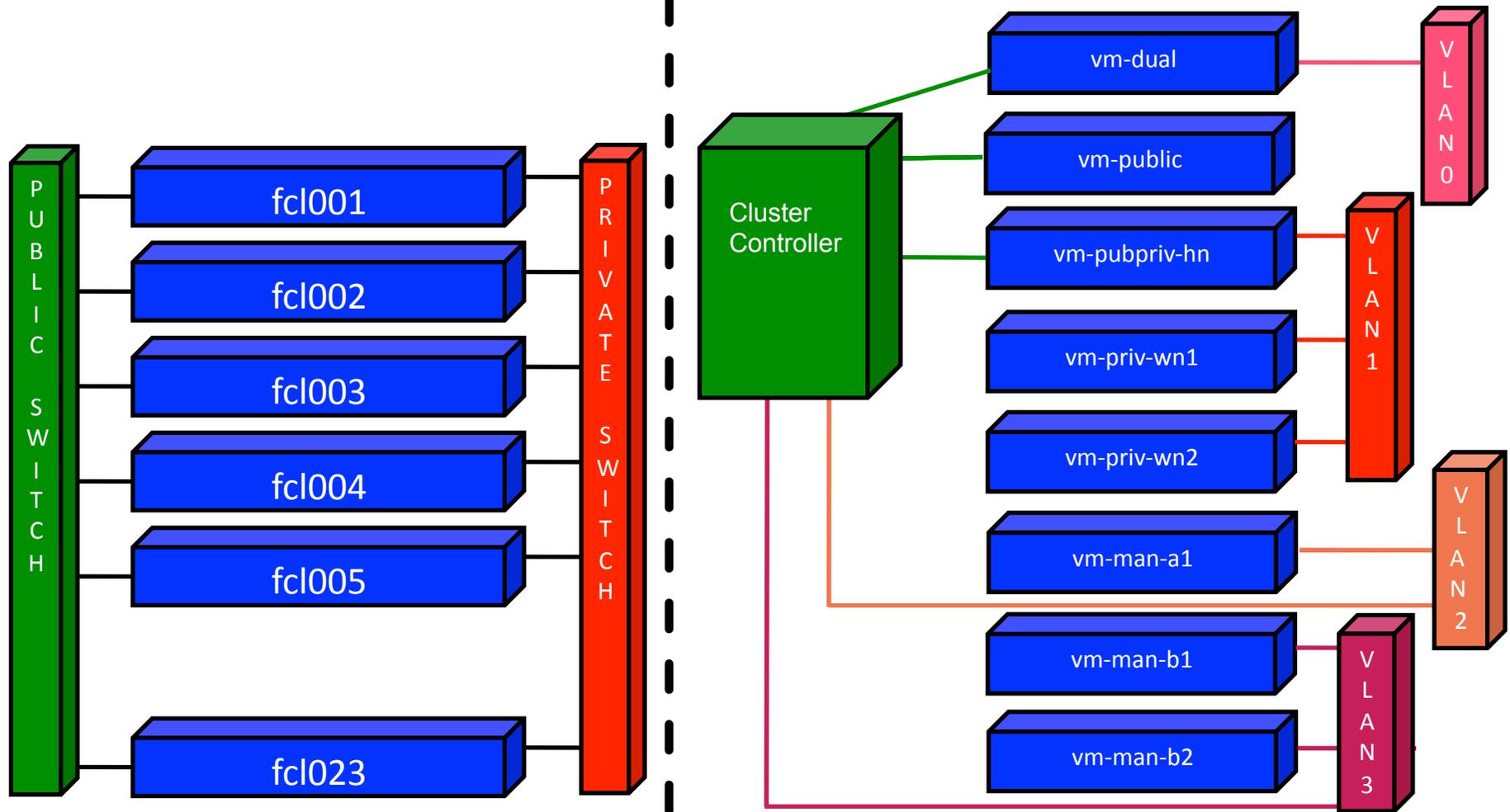


- 2x Quad Core Intel Xeon E5640 CPU
- 2 SAS 15K rpm system disk
- 300GB6x 2TB SATA disk
- LSI 1078 RAID controller
- Infiniband card
- 24GB RAM
- 368 logical cores on 23 machines total

Network Topology

Physical

Logical



Opportunities for Collaboration

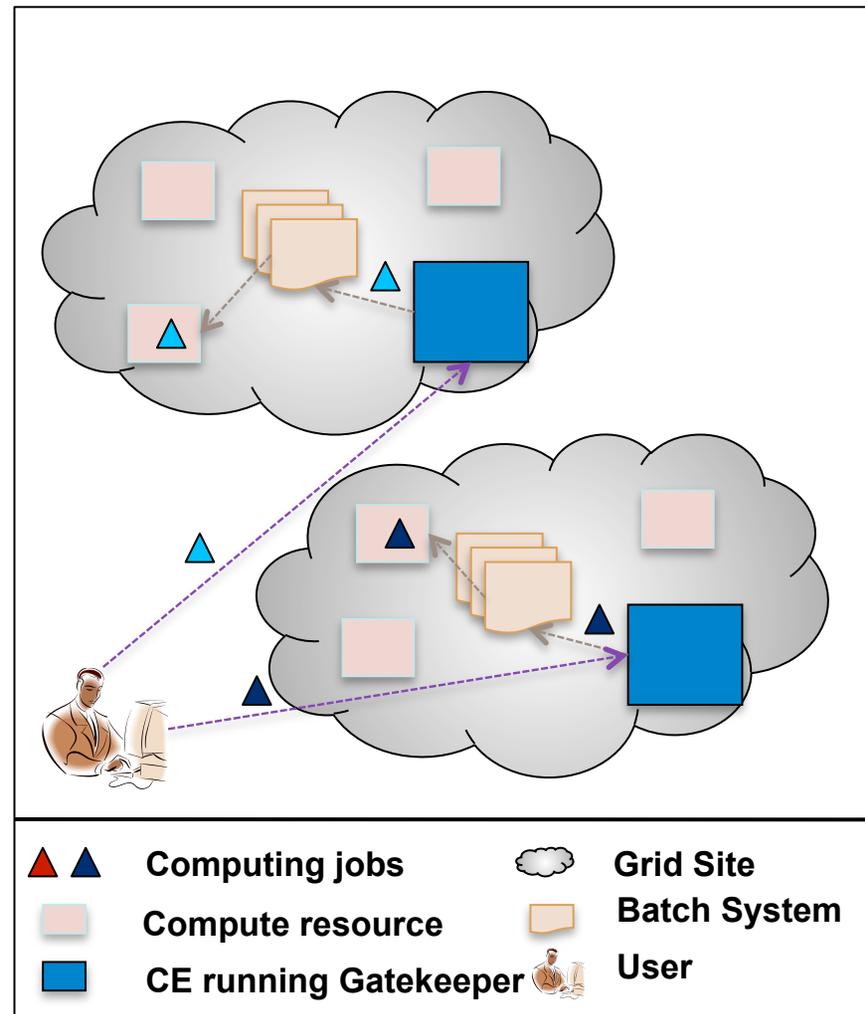
- **FermiGrid**
 - The Fermilab Campus Grid
- **FermiCloud**
 - The Fermilab Infrastructure as a Service
- **GlideinWMS**
 - Grid-wide overlay batch system
- **Center for Enabling Distributed Petascale Science**
 - CEDPS at Fermilab: data stage-out on overlay batch systems
- **OSG Storage**
 - Grid-wide data management
- **Gratia**
 - Grid accounting
- **Metric Correlation & Analysis Services**
 - Knowledge representation on a dashboard

GlideinWMS & Grid Computing

- Distributed computing paradigm spanning many administrative domains.
- Widely deployed for scientific communities with high computing demands
 - High Energy Physics (HEP)
 - Astro Physics Communities
 - Weather Surveys
 - Biology
 - [...]
- General purpose Grids used by the scientific communities
 - Open Science Grid (OSG)
 - European Grid for E-SciEnce (EGEE)
 - [...]

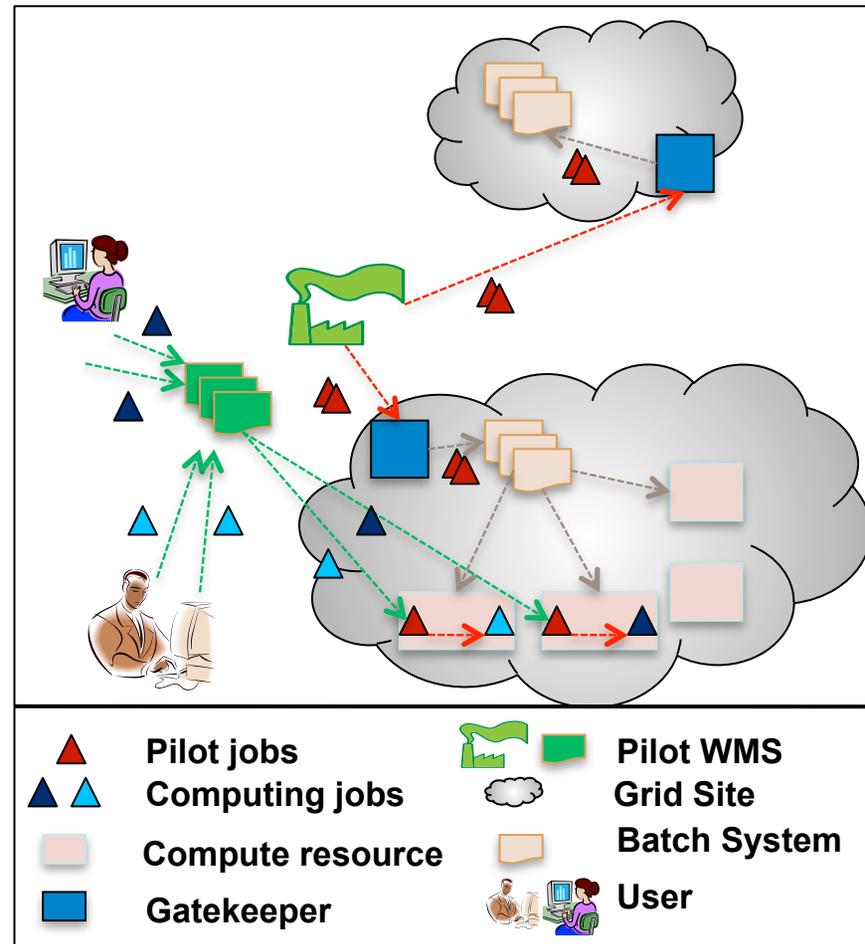
Typical Grid Use Case

- Grid Site: an admin. domain
 - Admins deploy Grid middleware
 - Compute Element (CE): Gateway for user jobs to the local batch system
 - Client tools on compute resources to access common Grid services
 - Local Batch System (BS)
- From user's perspective
 - Pros
 - Large pool of resources
 - Cons
 - Middleware problems in managing the job
 - Job progress is hidden and monitoring is complicated
 - Heterogeneity of the resources over the Grid
 - Need a meta WMS to manage Grid jobs



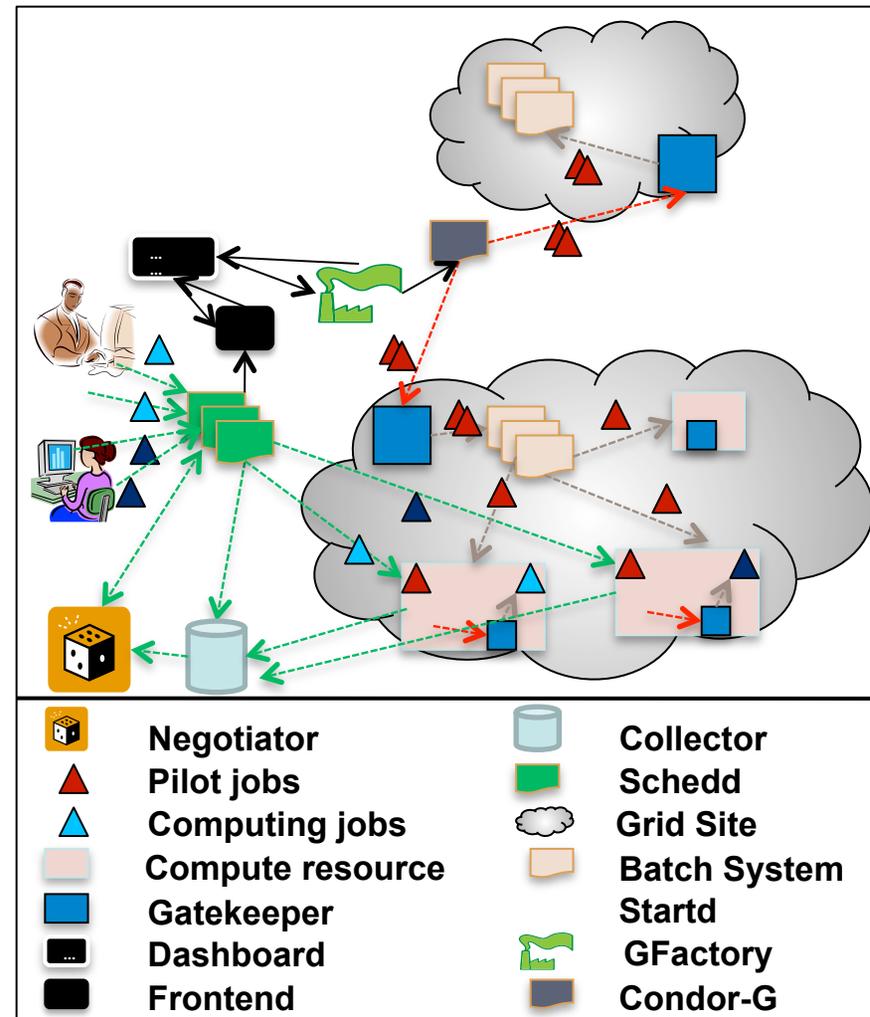
Pilot WMS Paradigm

- Pilot paradigm
 - Pilot factory submits pilot jobs to grid sites
 - Pilots start running on the computing resources and fetch user jobs from the user job queue
- Advantages of pilot based WMS
 - Forms an overlay pool of computing resources
 - Partially hides heterogeneity of grid sites from the user
 - Pilot jobs
 - If the environment is bad, pilot exits, preventing the user job to start and thus fail
 - Act as a wrapper and makes sure that the environment is right for the user job to execute.



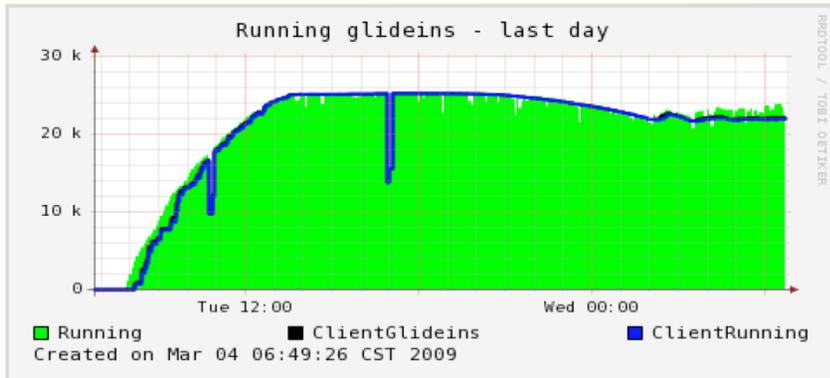
GlideinWMS Implementation

- GlideinWMS uses Condor as overlay batch system
- Factory
 - Manage jobs via CondorG
- VO Frontend
 - Allows Factory to regulate the flow of pilot jobs
- All network traffic is authenticated and integrity-checked
- Support pseudo-interactive monitoring

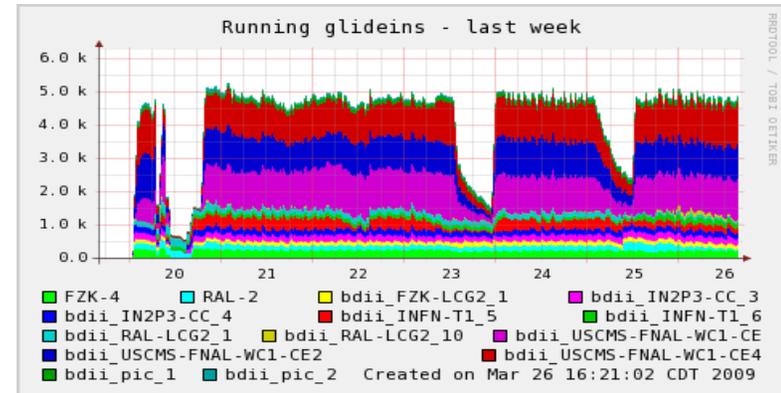


GlideinWMS in CMS Operations

32

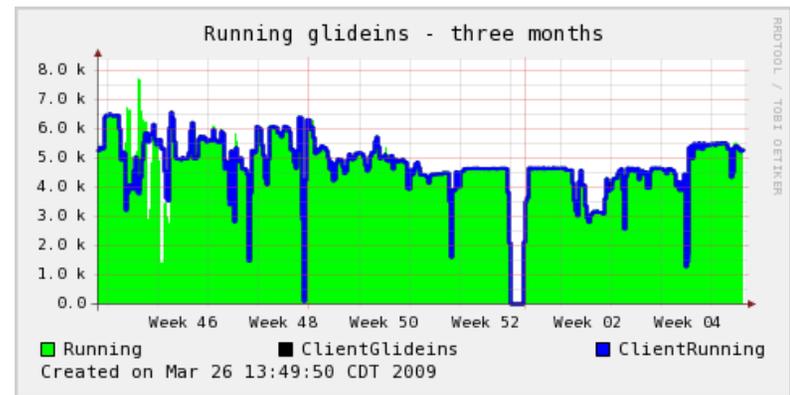


Running more than 20k glideins at any given time



CMS operations using glideinWMS at its seven archival storage sites

Criteria	Design goal	Measured
Total number of queued user jobs	100k	200k
Number of glideins in the system	10k	~26k
Number of running jobs per user queue	10k	~23k
Grid sites handled	~100	~100



CMS operations at Tier1 site at Fermilab

Opportunities for Collaboration

- **FermiGrid**
 - The Fermilab Campus Grid
- **FermiCloud**
 - The Fermilab Infrastructure as a Service
- **GlideinWMS**
 - Grid-wide overlay batch system
- **Center for Enabling Distributed Petascale Science**
 - CEDPS at Fermilab: data stage-out on overlay batch systems
- **OSG Storage**
 - Grid-wide data management
- **Gratia**
 - Grid accounting
- **Metric Correlation & Analysis Services**
 - Knowledge representation on a dashboard

The Center for Enabling Distributed Petascale Science

- The five year project started in 2006, funded by DOE
- Goals: Produce innovations for...
 - rapid and dependable data placement within a distributed high-performance environment
 - reliable and high-performance processing of computation and data analysis from many remote clients
 - troubleshooting and other ultra-high-performance distributed activities
- Collaborative Research
 - Math & Computer Science Div., Argonne National Lab
 - Lawrence Berkeley National Lab
 - Information Sciences Institute, U. of S. California
 - Dept of Computer Science, U. of Wisconsin Madison
 - Computing Division, Fermilab

CEDPS at Fermilab

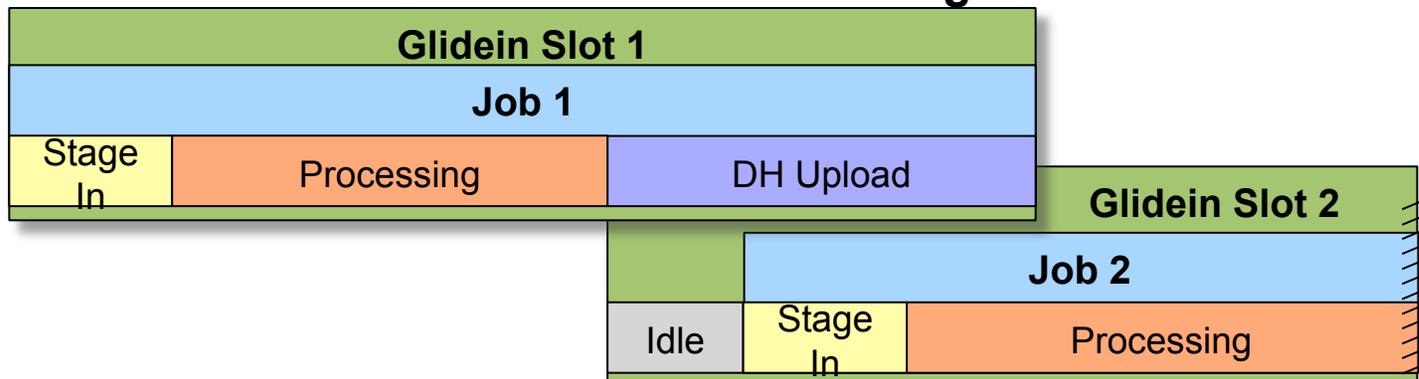
- Enhance the glideinWMS by
 - Increasing the CPU utilization of Condor resources on the Grid through CPU and network I/O overlap via asynchronous transfers of sandboxes
- What does this mean?
 - Pipeline the transfer of asynchronous sandboxes in Condor using globus.org i.e. a cloud-hosted high-performance, reliable, secure data movement service
 - A new job can start running if the previous job has entered stage-out state
 - Multiple transfers can take place via “transfer” slots
 - Reattempt failed transfers and use alternative routes as needed
 - Support multiple transfer protocols using transfer Plug-ins
- Work in progress
 - Implement support for transfer of asynchronous sandboxes in Condor

Slot Management

Condor Job Management Alone



DH Slot Management



Increase in throughput by creating new slots when stage out begins

Opportunities for Collaboration

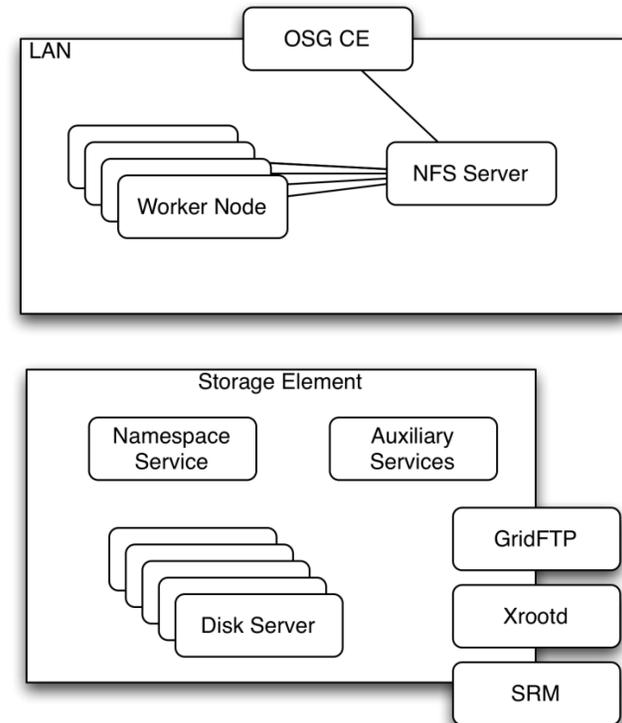
- **FermiGrid**
 - The Fermilab Campus Grid
- **FermiCloud**
 - The Fermilab Infrastructure as a Service
- **GlideinWMS**
 - Grid-wide overlay batch system
- **Center for Enabling Distributed Petascale Science**
 - CEDPS at Fermilab: data stage-out on overlay batch systems
- **OSG Storage**
 - Grid-wide data management
- **Gratia**
 - Grid accounting
- **Metric Correlation & Analysis Services**
 - Knowledge representation on a dashboard

OSG Storage Group

- Group members (all part time):
 - Tanya Levshina (OSG Storage coordinator)
 - Neha Sharma (support, dcache, probes, test suites)
 - Alex Sim (bestman developer and support)
 - Douglas Strain (rsv probes, xrootd, pigeon tools)
- Packages certification
- Test suites development, test stands
- Auxiliary software development
 - Gratia and RSV probes
 - Discovery tools
 - Pigeon tools
- Documentation
- Support for site administrators
 - GOC Tickets creation/monitoring
 - Liaison to developers groups
- Active mailing list: osg-storage@opensciencegrid.org

OSG Storage Architecture

- **SE Topology**
 - POSIX-mounted storage
 - Mounted and writable on the CE
 - Readable from the worker nodes
 - Not-scalable under heavy load
 - High-performance FS is not cheap
 - Space management is not trivial
- **SE Services**
 - SRM endpoint
 - Provides GridFTP Load balancing
 - Transfers via GridFTP servers
 - May provide internal access protocols (xroot, Posix)



Pictures from B. Bockelman's presentation at OSS2010

OSG SE Statistics

- Unofficial statistics based on BDII:
 - Number of sites providing Storage Elements: 49
 - Number of sites running dCache: 12
 - Number of sites running BeStMan-gateway: 37
 - HDFS 6
 - Xrootd 3
 - Lustre 3
 - REDDNet 1
 - All other sites: Local disk, NFS?
 - Number of sites reporting Gratia GridFTP Transfer Probes: 15 (daily transfer ~170,000 files, 800 TB)

Virtual Data Toolkit

VDT provides:

- A standard procedure for installation, configuration, services enabling, startup and shutdown
- Simplified configuration scripts
- All packages in one cache:
 - BeStMan
 - GridFTP
 - CA certificates, CRL installation, update
 - Log rotation scripts
 - Probes
- Straightforward upgrade procedure

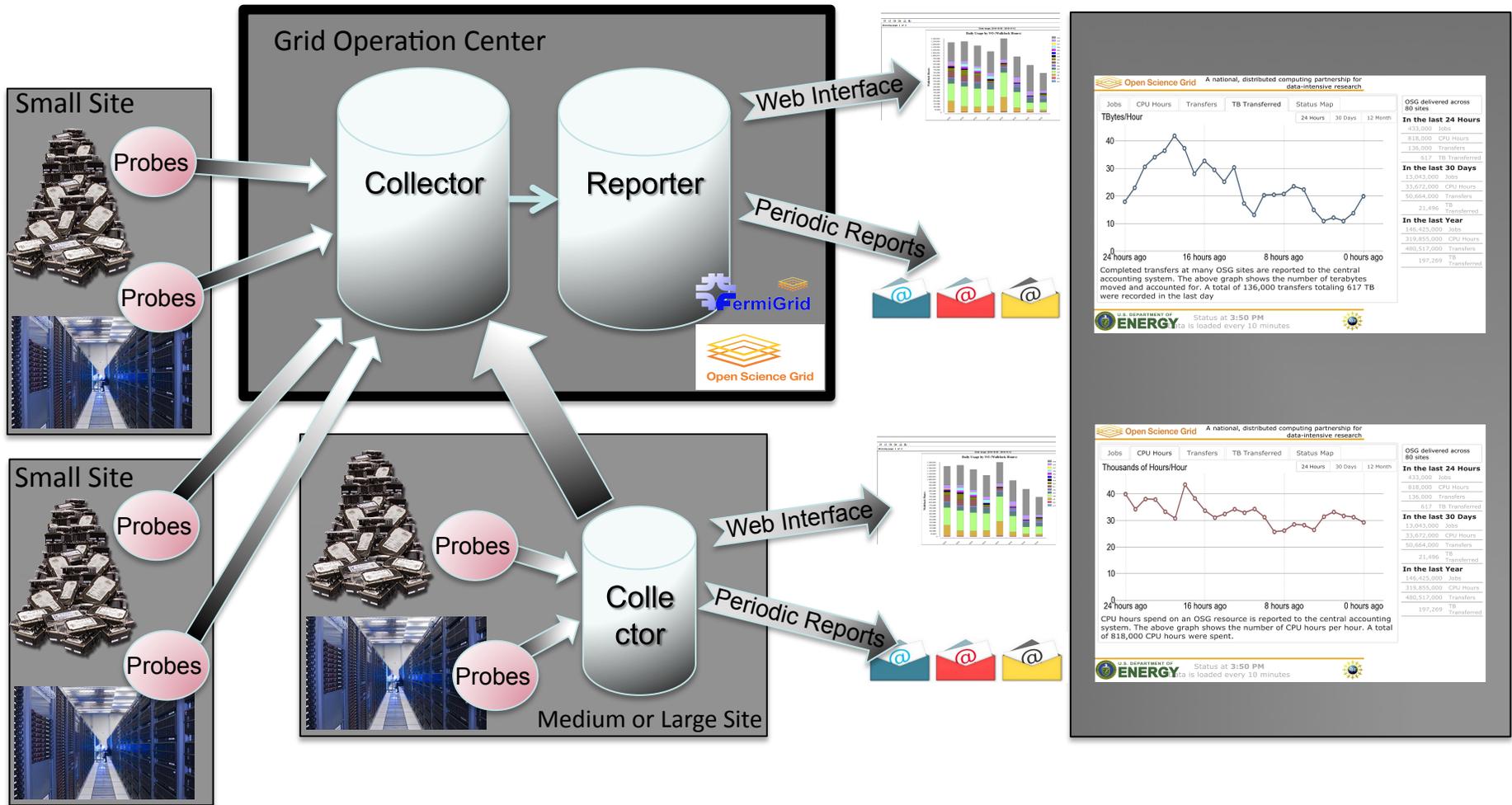
Opportunities for Collaboration

- **FermiGrid**
 - The Fermilab Campus Grid
- **FermiCloud**
 - The Fermilab Infrastructure as a Service
- **GlideinWMS**
 - Grid-wide overlay batch system
- **Center for Enabling Distributed Petascale Science**
 - CEDPS at Fermilab: data stage-out on overlay batch systems
- **OSG Storage**
 - Grid-wide data management
- **Gratia**
 - Grid accounting
- **Metric Correlation & Analysis Services**
 - Knowledge representation on a dashboard

Gratia Accounting Service

- Robust, scalable, trustable, dependable grid accounting service.
- The Gratia system consists of many probes operating on and uploading data from remote locations to a network of one or more collector-reporting systems.
- Data could be about batch jobs, grid transfers, storage allocation, site availability tests or process accounting.
- The primary focus of the Gratia system is to provide an accounting of jobs executed on the Open Science Grid.

Gratia Architecture



OSG Job Accounting Information By Site

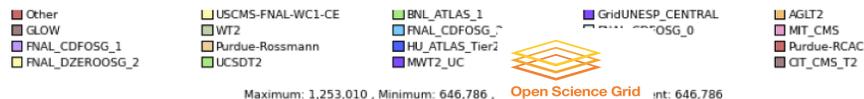
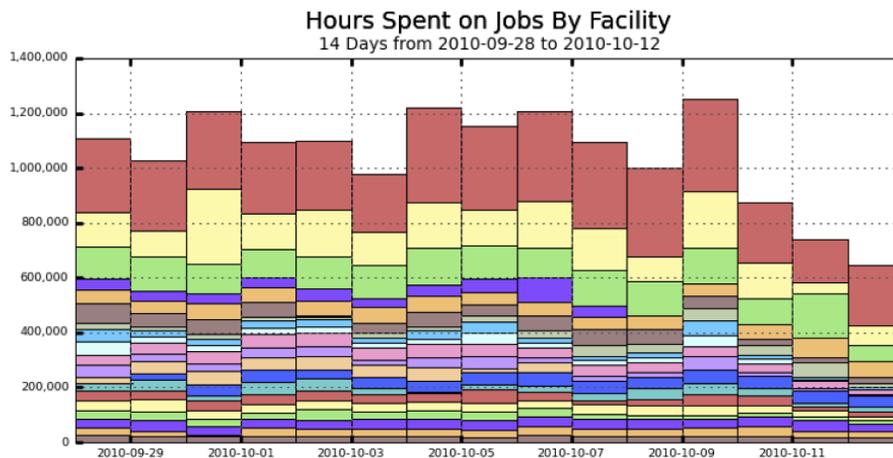
You are not authenticated! In order to be able to access a broader set of information, [click here](#).

Computation Hours

Navigation

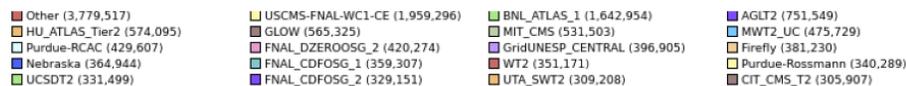
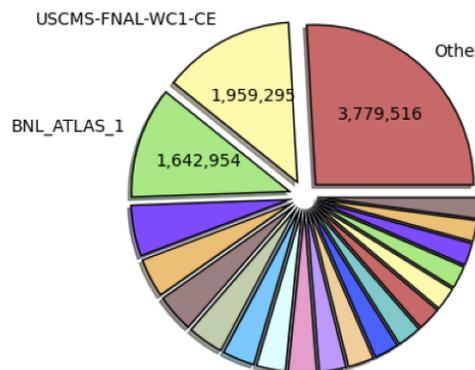


- Home
- Grid-wide View
- Monitoring By Site
- Accounting By Site
- Accounting By VO
- Opportunistic Usage
- Monitoring By VO
- Other View
- Non-HEP VOs by vo
- CMS T2 Sites by site
- CMS T2 Sites by vo
- ATLAS T2 Sites by vo
- ATLAS T2 Sites by site
- Non-HEP VOs by site
- WLCG Reporting (overview)
- WLCG Reporting (detailed)
- Developer Pages
- Custom View**



Wall Hours by Facility (Sum: 14,599,458 Hours)

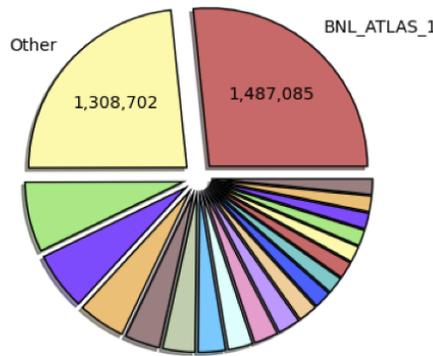
14 Days from 2010-09-28 to 2010-10-12



Fermilab Grid Facilities Department & KISTI: Opportunities for Collaborations



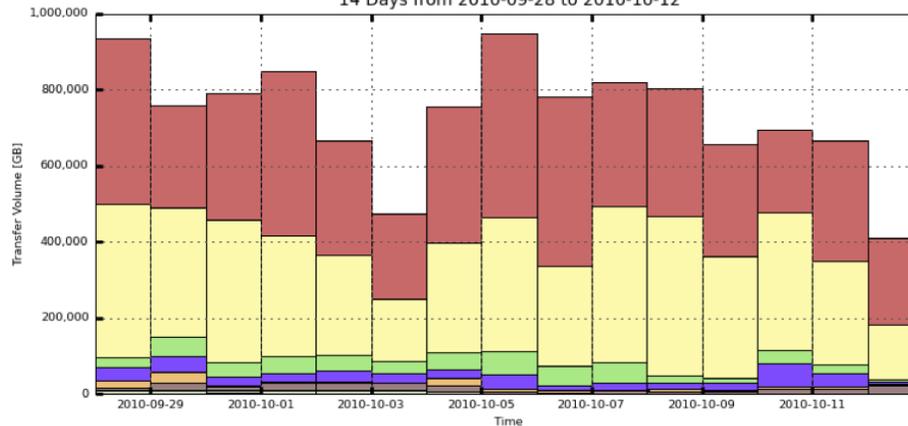
Job Count by Facility (Sum: 5,587,071)
14 Days from 2010-09-28 to 2010-10-12



- BNL_ATLAS_1 (1,487,085)
- Other (1,308,702)
- MWT2_UC (399,835)
- AGLT2 (331,414)
- USCMS-FNAL-WC1-CE (269,374)
- WT2 (200,304)
- Nebraska (183,855)
- GLOW (155,335)
- BU_ATLAS_Tier2 (141,002)
- UCSDT2 (137,007)
- HU_ATLAS_Tier2 (126,419)
- MIT_CMS (108,513)
- FNAL_CDFOSG_1 (99,255)
- FNAL_CDFOSG_2 (95,202)
- UFlorida-HPC (95,022)
- Purdue-RCAC (91,157)
- OU_OCHEP_SWT2 (90,443)
- Tufts_ATLAS_Tier3 (90,033)
- FNAL_CDFOSG_4 (88,926)
- FNAL_CDFOSG_3 (88,188)

Transfer Volumes

Volume of Gigabytes Transferred By Facility
14 Days from 2010-09-28 to 2010-10-12



- USCMS-FNAL-WC1-SE
- BNL_ATLAS_SE
- QT_CMS_SE2
- UCSDT2
- UFlorida-SE
- HEPGRID_UERJ
- NERSC_SE
- UCR-HEP-SE
- IJHEP
- UColorado_SE
- GridUNESP_CENTRAL_SE
- SGrid-Harvard-East
- brown-cms
- IllinoisHEP-SE
- umd-cms
- dinvestav-SE

Maximum: 949,328 GB, Minimum: 411,689 GB, Average: 734,962 GB, Current: 411,689 GB

Opportunities for Collaboration

- **FermiGrid**
 - The Fermilab Campus Grid
- **FermiCloud**
 - The Fermilab Infrastructure as a Service
- **GlideinWMS**
 - Grid-wide overlay batch system
- **Center for Enabling Distributed Petascale Science**
 - CEDPS at Fermilab: data stage-out on overlay batch systems
- **OSG Storage**
 - Grid-wide data management
- **Gratia**
 - Grid accounting
- **Metric Correlation & Analysis Services**
 - Knowledge representation on a dashboard

MCAS

Connect the dots ...

dCache service

Quick Finder

- Cell Services
- Pool Usage
- Tape Transfer Queue
- Detailed Tape Transfer Queue
- Pool Transfer Queues
- Action Log
- Pool Selection Configuration

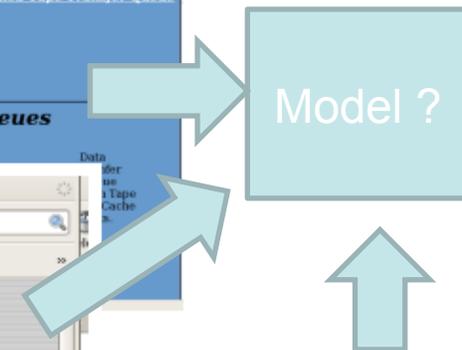
Status **Various Queues**

Availability and response

SRM Monitoring

- [SRM Requests \(Copy, Put, Get\)](#)
- [Browse SRM Requests](#)
- [Breakdown of errors for failed requests](#)
- [Number of Requests vs time](#)
- [Average Time Spent in each state for requests](#)
- [Remote Transfers](#)
- [Number of Remote Transfers vs time](#)
- [Average Time Spent in each state for Remote Transfers](#)
- [Requests for Space Reservation](#)
- [Get filerequests in state "Ready" created in last 24 hours](#)
- [Put filerequests in state "Ready" created in last 24 hours](#)

http://cmsdcam3.fnal.gov:8081/srmwatch/servlet/StartPage



PhEDEx - CMS Data Transfers

DB Instance: Production

Info Activity Data Requests Components

Rate Rate Plots Queue Plots Quality Plots Routing Transfer Details Deletions Recent Errors

Graph: Transfer Rate by Destination filter source destination

Period: Last 96 Hours up to

CMS PhEDEx - Transfer Rate
96 Hours from 2010-03-13 16:00 to 2010-03-17 16:00 UTC

Transfer Rate [MB/s]

Home System Servers Encp Help Mass Storage

STKEN: Enstore for the Masses

Ganglia FNAL TIER-1 RESOURCES Grid Report for Wed, 17 Mar 2010 10:55:32 -0500

FNAL TIER-1 RESOURCES Grid (13 sources)

CPU's Total: 12537
Hosts up: 1766
Hosts down: 20

unspecified (physical view)
CPU's Total: 36
Hosts up: 5
Hosts down: 1

CMSLPC (physical view)
CPU's Total: 96
Hosts up: 13
Hosts down: 0

Overall Status

- CD Mezzanine Powderhorn
- CMS Mezzanine Powderhorn
- SL8500s at GCC
- CMS dCache

Enstore Individual Servers

Servers

- Accounting Server
- Configuration Server
- Event Relay
- Inquisitor
- Volume Clerk

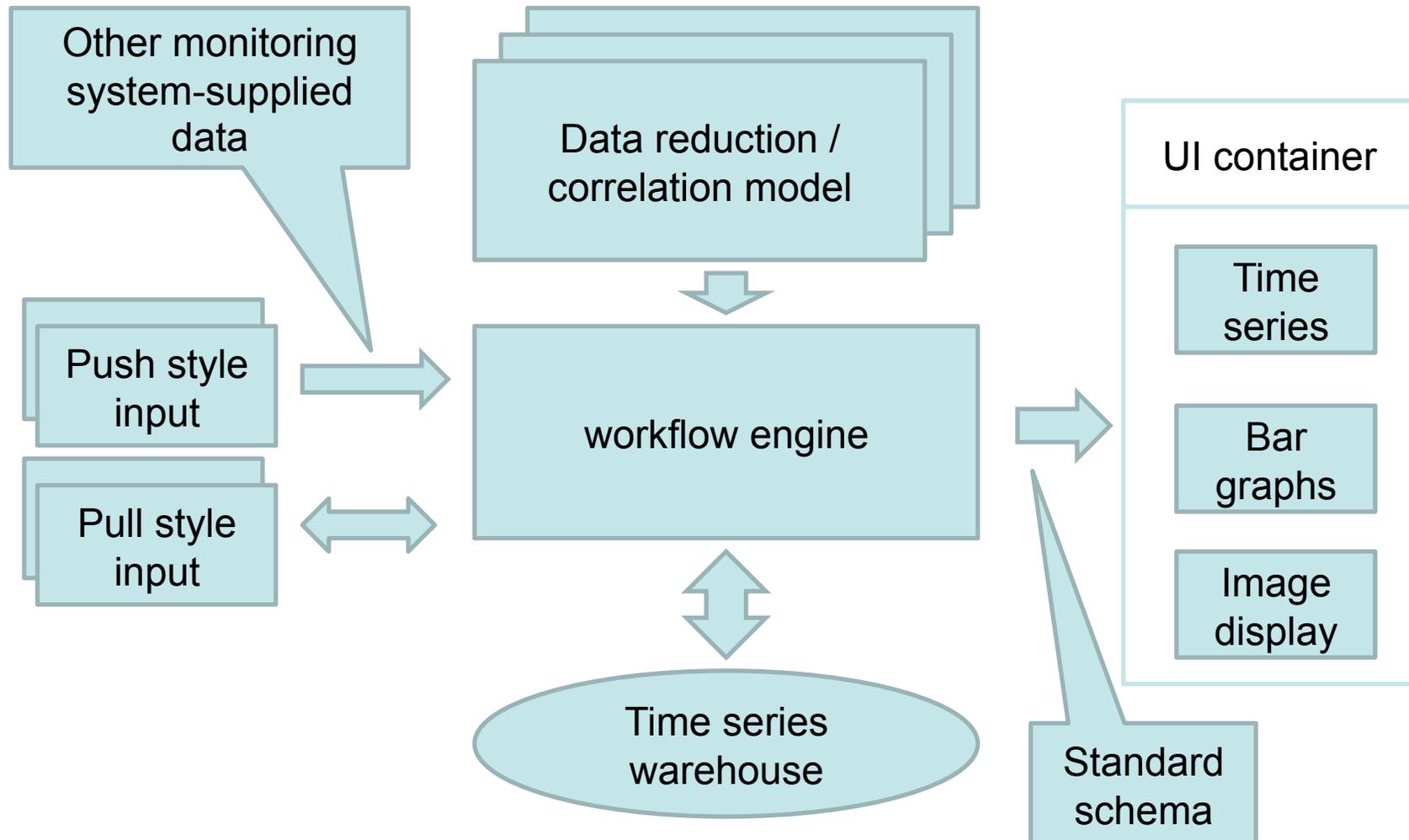
Library Managers

- 0940.library_manager (down/power)
- CD-9940B.library_manage
- CD-LTO3.library_manage
- CD-LTO4F1.library_manage
- CD-LTO4G1.library_manage

Metric Correlation and Analysis: Use cases

- Access, reduce and represent disjoint metrics data
- Infrastructure to bridge providers of monitoring/status information
- Summarization of metrics on at-a-glance view
- Toolkit for building and hosting time series, health bar graphs, and image displays

Project architecture



MCAS Workflow: cms

doors

Input data

poolgroups

pools

CMSDcacheDoors
name=CMSDcacheDoors
type=XML
mode=PULL

CMSpools
name=CMSpools
type=XML
mode=PULL

CMSQueues
name=CMSQueues
type=XML
mode=PULL

CMS5rm
name=CMS5rm
type=XML
mode=PULL

CMSDcap
name=CMSDcap
type=XML
mode=PULL

Merge and summarize

LPCCondorSched
name=LPCCondorSched
type=CSV
mode=PUSH

T1CondorSched
name=T1CondorSched
type=CSV
mode=PUSH

dCacheHealth
name=dCacheHealth
type=XML
mode=PULL

LPCCondorFarmStats
name=LPCCondorFarmStats
type=XML
mode=PULL

T1CondorFarmStats
name=T1CondorFarmStats
type=XML
mode=PULL

dCacheHealthTs
name=dCacheHealthTs
type=XML
mode=PULL

CmsFacilityHealthTs
name=CmsFacilityHealthTs
type=XML
mode=PULL

Plot

LPCCondorFarmHealth
name=LPCCondorFarmHealth
type=XML
mode=PULL

T1CondorFarmHealth
name=T1CondorFarmHealth
type=XML
mode=PULL

dCacheHealthTsPlot
name=dCacheHealthTsPlot
type=XML
mode=PULL

CondorFarmHealth
name=CondorFarmHealth
type=XML
mode=PULL

dCacheHealthTsAgg
name=dCacheHealthTsAgg
type=XML
mode=PULL

claimed_html

TimeCountHistogram

lpc_claimed_html

CmsFacilityHealth
name=CmsFacilityHealth
type=XML
mode=PULL

T1DetailImages
name=T1DetailImages
type=HTML
mode=PULL

SrmDetailImages
name=SrmDetailImages
type=HTML
mode=PULL

LPCDetailImages
name=LPCDetailImages
type=HTML
mode=PULL



http://mcas-int.fnal.gov:8080/portal/portal/cmsDcache



Google

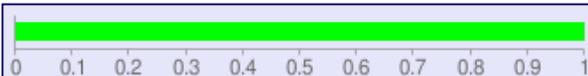
Most Visited | Fermi Linux | Fermilab | FermiLinux | Linux Distros | FINM 33100 (Winter ...

JBoss Portal

Log

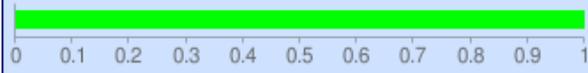
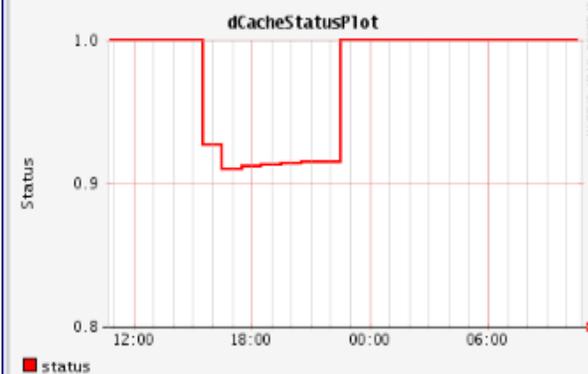
default

Cms Facility Health

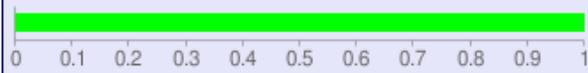


dCache

Total Value: 1
 Green Value: 1
 Graph URL: <http://chart.apis.google.com/chart?cht=bhs&chs=328x40&chd=t:1|0&chf=bg,s,E6E6FA|c,s,E6E6FA&chco=00FF00,FF0000,E6E6FA&chbh=10&chts=000000,11&chds=0,1&chxt=x&chxr=0,0,1>
 1. name: dCache
 2. status: 1
 3. statusMax: 1



SRM

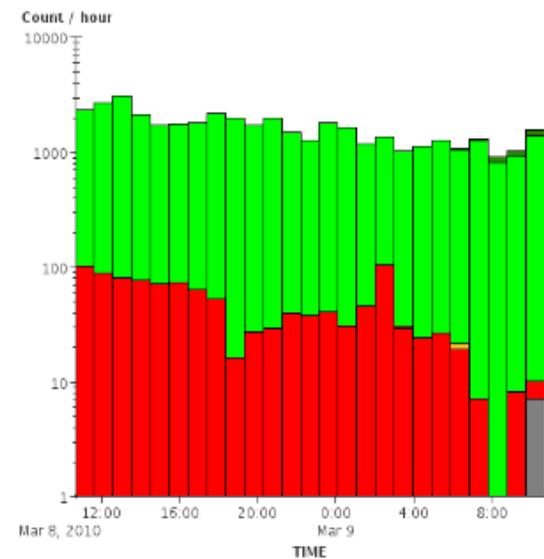
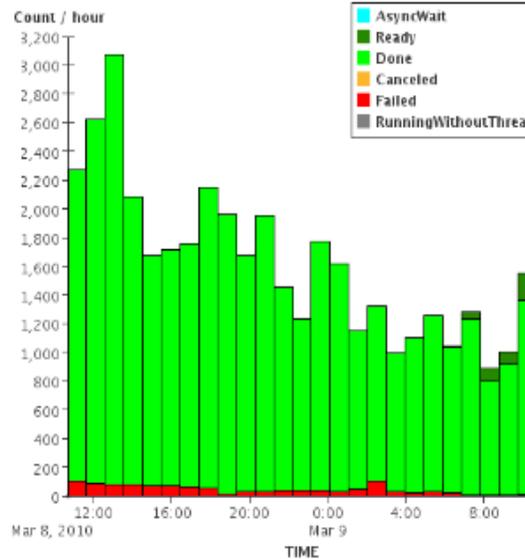


Pools



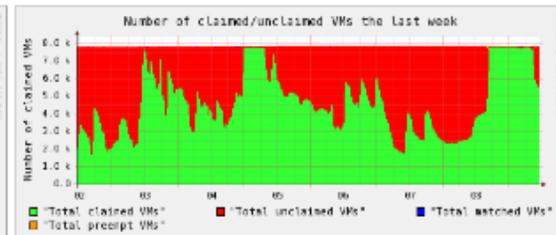
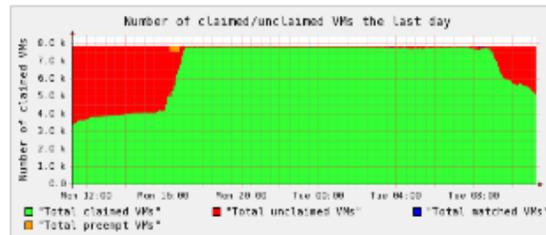
CMSSRMDetailImages

Click the image to show it in original size



CMST1FarmDetailImages

Click the image to show it in original size



Conclusions

- The Grid Department is thrilled of establishing a collaboration with the KISTI Institute
- We want to build on and strengthen our current relationship
 - KISTI as an OSG Site
 - KISTI as a remote center of CDF computing
- Early this year, we discussed the option of hosting KISTI personnel at Fermilab: it's a win-win opportunity