FermiGrid
Status and Plans

Keith Chadwick
chadwick@fnal.gov

Fermilab
Computing Division
Communications and Computing Fabric Department
Fabric Technology Projects Group Leader

People

FermiGrid Operations Team:

- Keith Chadwick (CD/CCF/FTP) – Project Leader
- Steve Timm (CD/CSS/FCS) – Linux OS Support
- Dan Yocum (CD/CCF/FTP) – Application Support

FermiGrid Stakeholder Representatives:

- Keith Chadwick – FermiGrid Common Services Representative
- Igor Sfiligoi – CDF Stakeholder Representative
- Ian Fisk – U.S.CMS Representative
- Amber Boehnlein & Alan Jonckheere – D0 Stakeholder Representatives
- Neha Sharma – DES & SDSS Stakeholder Representative
- Steve Timm – Fermilab General Purpose Farms Representative
- Ruth Pordes – OSG Stakeholder Representatives
- Eileen Berman & Rob Kennedy - Storage (enstore & dcache) Representatives

FermiGrid Web Site & Additional Documentation:

- http://fermigrid.fnal.gov/

What is it?

Like the old story about the blind men and the elephant,
the answer to the question “What is FermiGrid”
depends on who is asking about what and where they are.

VAW - Initial Strategy and Goals:

On November 10, 2004, Vicky White (Fermilab CD Head) wrote the following:

In order to better serve the entire program of the laboratory the Computing Division will place all of its production resources in a Grid infrastructure called FermiGrid. This strategy will continue to allow the large experiments who currently have dedicated resources to have first priority usage of certain resources that are purchased on their behalf. It will allow access to these dedicated resources, as well as other shared Farm and Analysis resources, for opportunistic use by various Virtual Organizations (VOs) that participate in FermiGrid (i.e. all of our lab programs) and by certain VOs that use the Open Science Grid. (Add something about prioritization and scheduling – lab/CD – new forums). The strategy will allow us:

- to optimize 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 (and eventually to BTeV) 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.

VAW - What FermiGrid Is:

FermiGrid is a meta-facility composed of a number of existing "resources", many of which are currently dedicated to the exclusive use of a particular stakeholder.

FermiGrid (the facility) provides a way for jobs of one VO to run either on shared facilities (such as the current General Purpose Farm or a new GridFarm?) or on the Farms primarily provided for other VOs.

FermiGrid will require some development and test facilities to be put in place in order to make it happen.

FermiGrid will provide access to storage elements and storage and data movement services for jobs running on any of the compute elements of FermiGrid

The resources that comprise FermiGrid will continue to be accessible in "local" mode as well as "Grid" mode

VAW - The FermiGrid Project:

This is a cooperative project across the Computing Division and its stakeholders to define and execute the steps necessary to achieve the goals of FermiGrid

Effort is expected to come from

- Providers of shared resources and services – CSS and CCF
- Stakeholders and providers of currently dedicated resources - Run II, CMS, MINOS, SDSS

The total program of work is not fully known at this time – but the WBS is being fleshed out. It will involve at least the following

- Adding services required by some stakeholders to other stakeholders dedicated resources
- Work on authorization and accounting
- Providing some common FermiGrid Services (e.g ….)
- Providing some head-nodes and Gateway machines
- Modifying some stakeholders scripts, codes, etc. to run in the FermiGrid environment
- Working with OSG technical activities to make sure FermiGrid and OSG (and thereby LCG) are well aligned and interoperable
- Working on monitoring and web pages and whatever else it takes to make this all work and happen
- Evolving and defining forums for prioritizing access to resources and scheduling

FCC - Feynman Computing Center

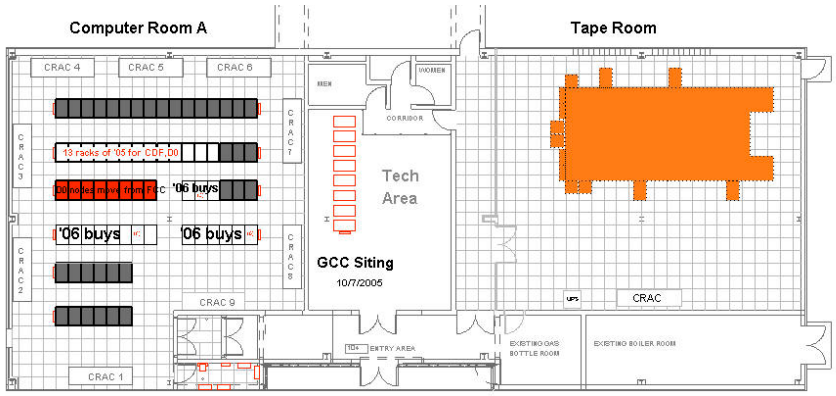
GCC – Computer Room A & Tape Robot Room

GCC – Grid Computing Center

Computer Room A – Specified for 2880 systems and being populated.

GCC reached 270 KVA in September 2005 which is 45% of FCC, GCC only ~30 percent occupied

Tape Robot Room – "Under Construction", expected completion April-May 2006.



GCC Computer Room A with rack allocations
Tape Room with ADIC robot

What FermiGrid Is Today:

Based on the initial FermiGrid project definition provided by Vicky White, the FermiGrid project team and stakeholder representatives developed the following 4 key components of FermiGrid:

FermiGrid Common Grid Services:

- Supporting common Grid services to aid in the development and deployment of Grid computing infrastructure by the supported experiments at FNAL

FermiGrid Stakeholder Bilateral Interoperability:

- Facilitating the shared use of central and experiment controlled computing facilities by supported experiments at FNAL
  - CDF, D0, CMS, GP Farms, SDSS, etc.

FermiGrid Development of OSG Interfaces for Fermilab:

- Enabling the opportunistic use of FNAL computing resources through Open Science Grid (OSG) interfaces.

FermiGrid Exposure of the Permanent Storage System:

- Enable the opportunistic use of FNAL storage resources (STKEN) through Open Science Grid (OSG) interfaces.

Common Grid Services:

The common grid services component of FermiGrid supports the deployment and operation of the following common grid services:

- Fermilab Site Wide Globus Gateway (fermigrid1.fnal.gov / fermigrid.fnal.gov).
  - The FermiGrid Site Wide Gateway accepts jobs from the Open Science Grid, and (following appropriate credential authorization), schedules these jobs for execution on Fermilab Grid resources via Condor-G.
- VOMS and VOMRS (fermigrid2.fnal.gov / voms.fnal.gov).
  - The FermiGrid VOMS server hosts the site wide VO management service.
- GUMS (fermigrid3.fnal.gov / gums.fnal.gov).
  - The FermiGrid GUMS server hosts site wide Grid User Mapping Services.
- Myproxy (fermigrid4.fnal.gov / myproxy.fnal.gov).
  - The FermiGrid myproxy server hosts the
- SAZ (fermigrid4.fnal.gov).
  - The Fermilab Site AuthoriZation service.
- Accounting and Monitoring (distributed).
  - A growing issue.

Hardware

Dell 2850 Servers with dual 3.6 GHz Xeons, 4Gbytes of memory, 1000TX, Hardware Raid, Scientific Linux 3.0.4, VDT 1.3.6

FermiGrid1:

Site Wide Globus Gateway

FermiGrid2:

Site Wide VOMS &
VOMRS Server

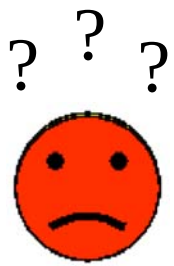
FermiGrid3:

Site Wide GUMS Server

FermiGrid4:

Myproxy server
Site AuthoriZation server

Site Wide Gateway – Why:



VOMS Server

Myproxy Server

GUMS Server

Site Wide Gateway

SAZ Server

CMS WC1 | CMS WC2 | CDF CAF | CDF | D0 CAB | D0 | SDSS TAM | GP Farm | LQCD | Desk tops

Site Wide Gateway Technique:

This technique is closely adapted from a technique first used at GridX1 in Canada to forward jobs from the LCG into their clusters.

We begin by creating a new Job Manager script in:

$VDT_LOCATION/globus/lib/perl/Globus/GRAM/JobManager/jobmanager-condorg

This script takes incoming jobs and resubmits them to Condor-G on fermigrid1

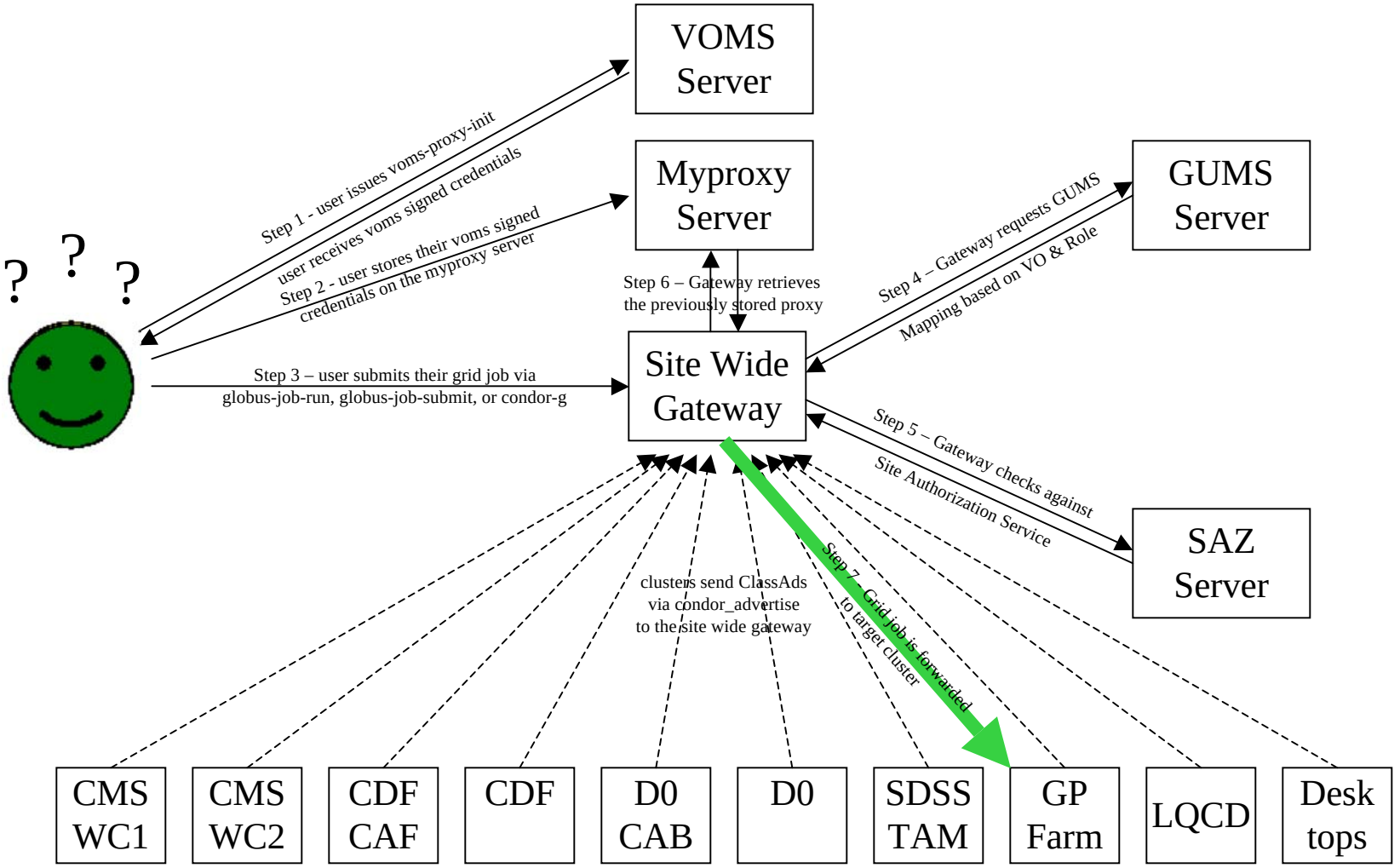
Condor matchmaking is used so that the jobs will be forwarded to the member cluster with the most open slots.

Each member cluster runs a cron job every five minutes to generate a ClassAD for their cluster. This is sent to fermigrid1 using condor_advertise.

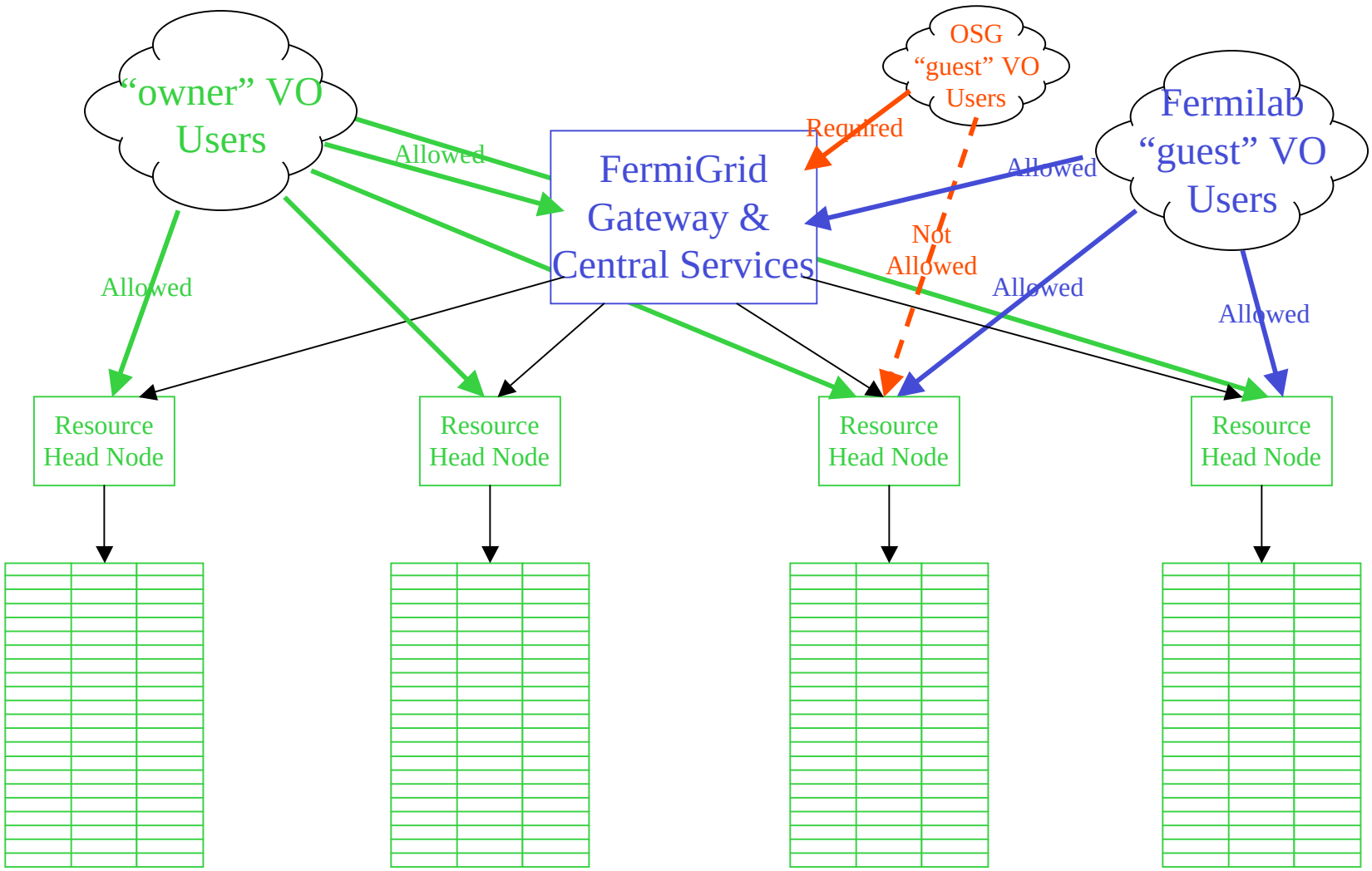
Credentials to successfully forward the job are obtained in the following manner:

1. User obtains a voms-qualified proxy in the normal fashion with voms-proxy-init
2. User sets X509_USER_CERT and X509_USER_KEY to point to the proxy instead of the usercert.pem and userkey.pem files
3. User uses myproxy-init to store the credentials, using myproxy, on the fermilab myproxy server myproxy.fnal.gov
4. jobmanager-condorg, which is running as the uid that the job will run on under fermigrid, executes a myproxy-get-delegation to get a proxy with full rights to resubmit the job.
5. Documentation of the steps to do this as a user is found in the Fermigrid User Guide: http://fermigrid.fnal.gov/user-guide.html

Site Wide Gateway Animation:



Guest vs. Owner VO Access:



VOMS & VOMRS:

FermiGrid deployed VOMS & VOMRS services on April 1st 2005.

Prior to this, VO management at Fermilab was performed via USCMS in a "back pocket" fashion.

- USCMS, SDSS, Lattice QCD, MINOS, etc.
- Realized that this was not a viable solution for the long term.
- And CMS would probably like to direct that effort towards their work.

Since April 1st 2005, the FermiGrid has provided the central infrastructure for the VO Management Server and Services.

- Existing VOs were migrated to the new VOMS / VOMRS on FermiGrid Common Grid Services (with agreement of the VO of course).
- New VOs are being created as requested/appropriate.
- VO administration roles are being delegated to appropriate members of the VOs.
- We periodically "sweep" Fermilab Kerberos users into the fermilab VO (4500+ user entries)

Virtual Organizations:

FermiGrid currently hosts the following Virtual Organizations:

- auger http://www.auger.org/
- cdf http://www-cdf.fnal.gov / (this is a replica of the cdf-grid VO at INFN)
- cms http://cmsinfo.cern.ch/ (this is a replica of the cms VO at Cern)
- des http://decam.fnal.gov/
- dzero http://www-d0.fnal.gov/
- fermilab http://www.fnal.gov/
- gadu http://www-wit.mcs.anl.gov/Alex/GADU/Index.cgi
- nanohub http://www.nanohub.org/
- sdss http://www.sdss.org/
- uscms http://www.uscms.org/
- ilc http://ilc.fnal.gov/
- lqcd http://lqcd.fnal.gov/
- I2u2 http://www-ed.fnal.gov/uueo/i2u2.html

GUMS:

FermiGrid uses the Grid User Management System (GUMS) developed at BNL to maintain GRIDMAP file identity mappings between certificate and the local user UID and the VO Privilege Project (PRIMA) to manage delegation of privilege roles within VOs.

We currently support the following two GUMS mapping:

- Many to 1 – (i.e. the entire VO mapped onto a single UID).
- Many to Some – (i.e. roles within a VO).

Additional customized mappings are made available as needed by VOs to perform grid related administration tasks and manage privileges, subject to negotiation between Fermilab and the VO.

We currently have over 13K entries in our GUMS mapping table.

Myproxy:

As integral part of the condor-g job forwarding mechanism, we are running myproxy.

Myproxy is also available for general fermilab grid use (members of uscms are using myproxy on FermiGrid4 to store their grid certificates to operate with LCG).

SAZ - Site AuthoriZation:

We are in the process of implementing a new Site AuthoriZation module for the Fermilab "Open Science Enclave".

Our current plan is to operate in a default accept mode for user credentials that are associated with known and "trusted" VOs and CAs:

Site authorization callout on globus gateway sends SAZ authorization request (example):

user: /DC=org/DC=doegrids/OU=People/CN=Keith Chadwick 800325
VO: fermilab
attribute:/fermilab/Role=NULL/Capability=NULL
CA: /DC=org/DC=DOEGrids/OU=Certificate Authorities/CN=DOEGrids CA 1

SAZ server on fermigrid4 receives SAZ authorization request:

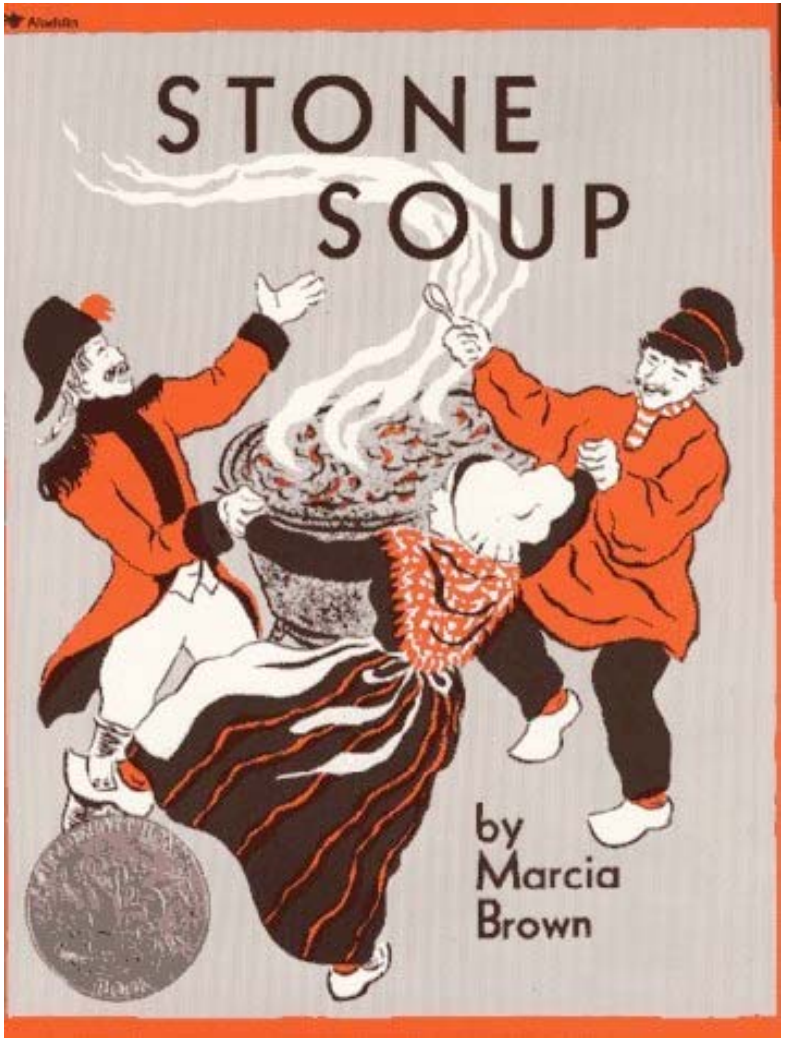
1. Verifies certificate and trust chain.
2. if the certificate does not verify or the trust chain is invalid; then
   SAZ returns "Not-Authorized"
   fi
3. Issues select on "user:" against the local SAZ user DB
4. if the select on "user:" fails; then
   a record corresponding to the "user:" is inserted into the local SAZ user DB with status = Authorized
   fi
5. Issues select on "VO:" against the local SAZ VO DB
6. if the select on "VO:" fails; then
   a record corresponding to the "VO:" is inserted into the local SAZ VO DB with status = Not-Authorized
   fi
7. Issues select "CA:" against the local SAZ CA DB
8. if the select on "CA:" fails; then
   a record corresponding to the "CA:" is inserted into the local SAZ CA DB with status = Not-Authorized
   fi
9. The SAZ server then returns the logical and of ("user:", "VO:", "CA:")

Stakeholder Interoperability:

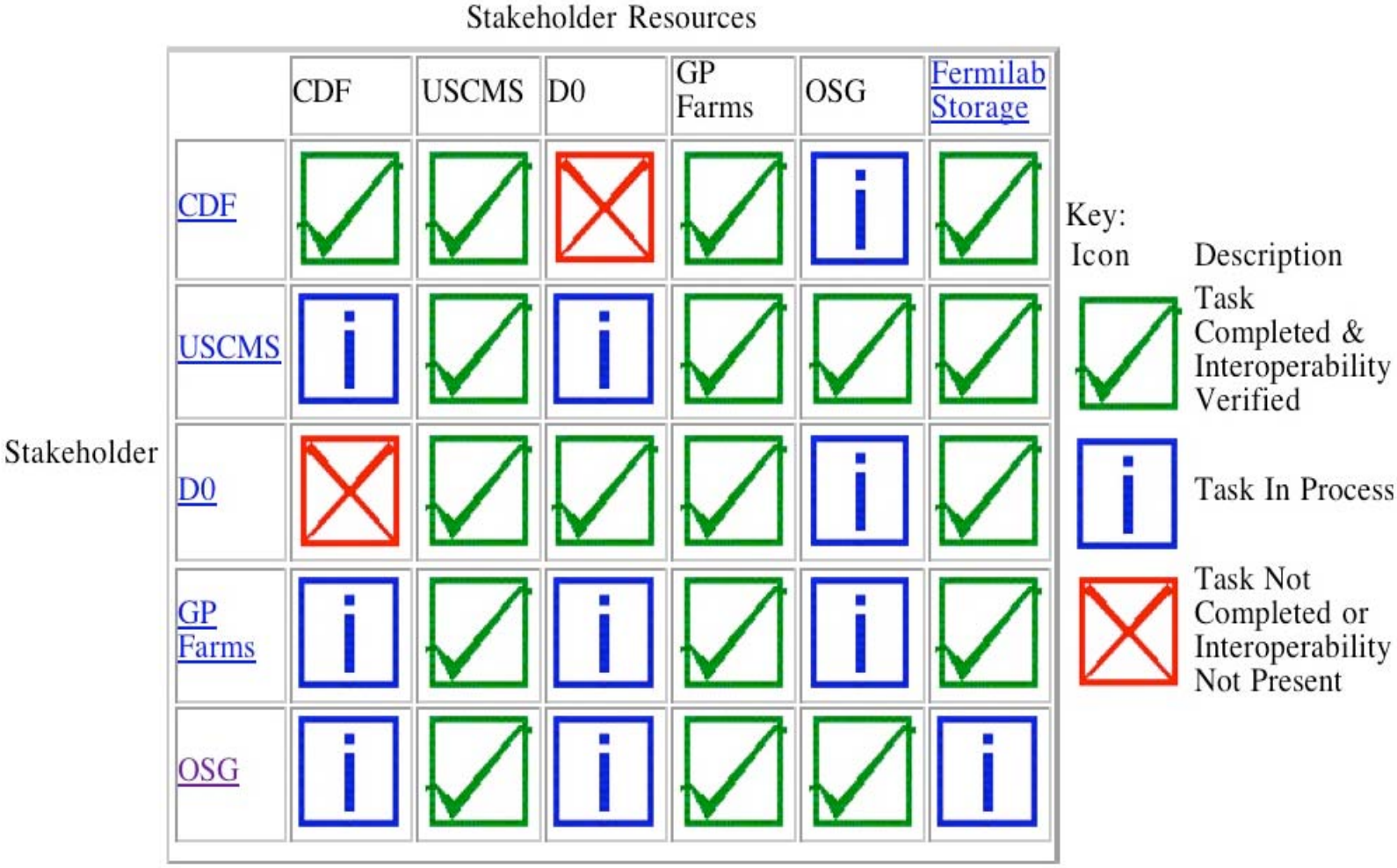
The second component of FermiGrid is the bilateral stakeholder interoperability across the stakeholders (CDF, D0, CMS, GP Farms, SDSS, etc.) computing resources.

Most of this work takes place in the various stakeholder organizations without direct FermiGrid Operations Team involvement.

But there are certainly vigorous discussions…



Stakeholder Progress:



OSG Interfaces for Fermilab:

The third component of FermiGrid is enabling the opportunistic use of FNAL computing resources through Open Science Grid (OSG) interfaces.

Most of this work to accomplish this happened in the context of the installation and configuration of the Fermilab Common Grid Services and deployment and integration on the GP Farms.

We are still wrestling with the problem of “sufficient” and “available” OSG storage.

Fermilab “job-forwarding” gateway has caused some problems for users that assume that a job submitted to the fork jobmanager will run on the same hardware that a job submitted to the condor-g jobmanager.

OSG Access to Permanent Storage:

The fourth component of FermiGrid is enabling the opportunistic use of our storage resources through Open Science Grid (OSG) interfaces.

The Fermilab storage resources are available in the context of the OSG Storage Element (SE). The storage elements are either buffering (used to cache and assemble data sets) or custodial (used for long-term retention of scientific data). The preferred Grid interface is via the Storage Resource Manager (SRM). The SRM client supports the following access methods:

- Direct GridFTP service (although it is not scalable).
- Http service for reads.
- Kerberized FTP service.
- A read-only service with weak passwords.

Fermilab has realized three separate storage system interfaces:

- CDFEN - Enstore Mass Storage Production Service for CDF Run II.
- DOEN - Enstore Mass Storage Production Service for D0 Run II.
- STKEN - Enstore Mass Storage Production Service for all other users.

The STKEN facility is exposed to the Open Science Grid and is available for use by Fermilab experiments. Fermilab is prototyping an opportunistic role with other experiments as well.

Timeline & Where Are We Today?

Feb 2005	The systems which host the FermiGrid Common Grid Services were delivered in the middle of February and installed by the end of February 2005.
Mar 2005	Configuration and shakedown testing of VOMS, VOMRS and GUMS.
1 Apr 2005	Production availability of the FermiGrid Common Grid Services occurred.
Apr – Jun 2005	Integration with USCMS and the Fermilab General Purpose Farms.
Jun – Jul 2005	Integration with the Open Science Grid 0.2 deployment.
	Participated in the OSG Consortium ribbon cutting on July 20, 2005.
Aug – Sep 2005	Forwarding gateway implemented.
	Metrics collection implemented.
Dec 2005	GUMS 1.1.0 and VOMS 1.6.7 upgrades.
Jan 2006	OSG 0.4 deployment.

Current Work

Public storage and storage element:

- Expose public dcache storage element (FNAL_FERMIGRID_SE).
- Add additional general availability storage.

Working with DZero to transition from static gridmap files to use FermiGrid Common Grid Services

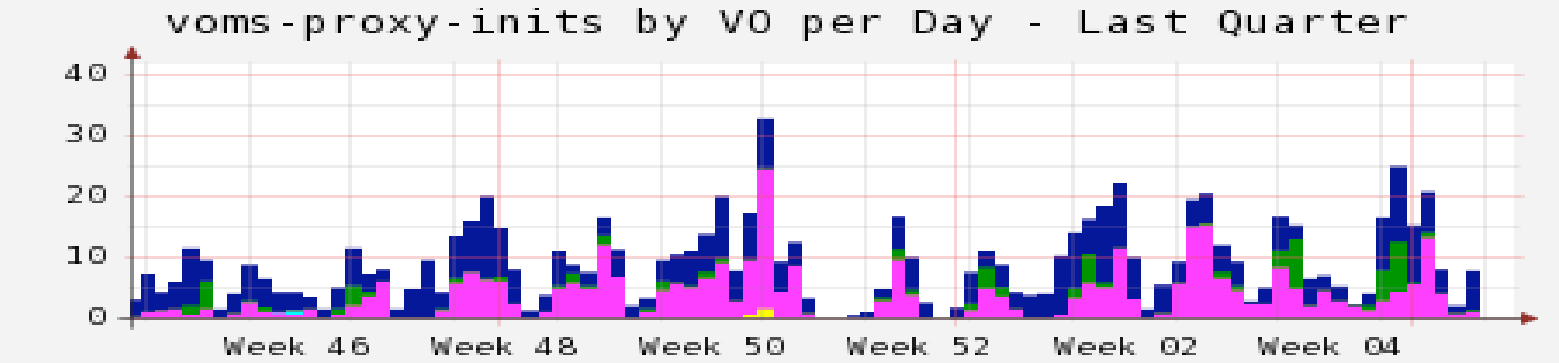
Working with the SAM & SAMgrid teams to utilize the FermiGrid Common Grid Services.

Working with D0 and CDF to implement OSG interfaces to their analysis clusters.

Service Failover:

- At the present time the services are running in non-redundant mode.
- We have plans to implement service failover via Linux-HA clustering.

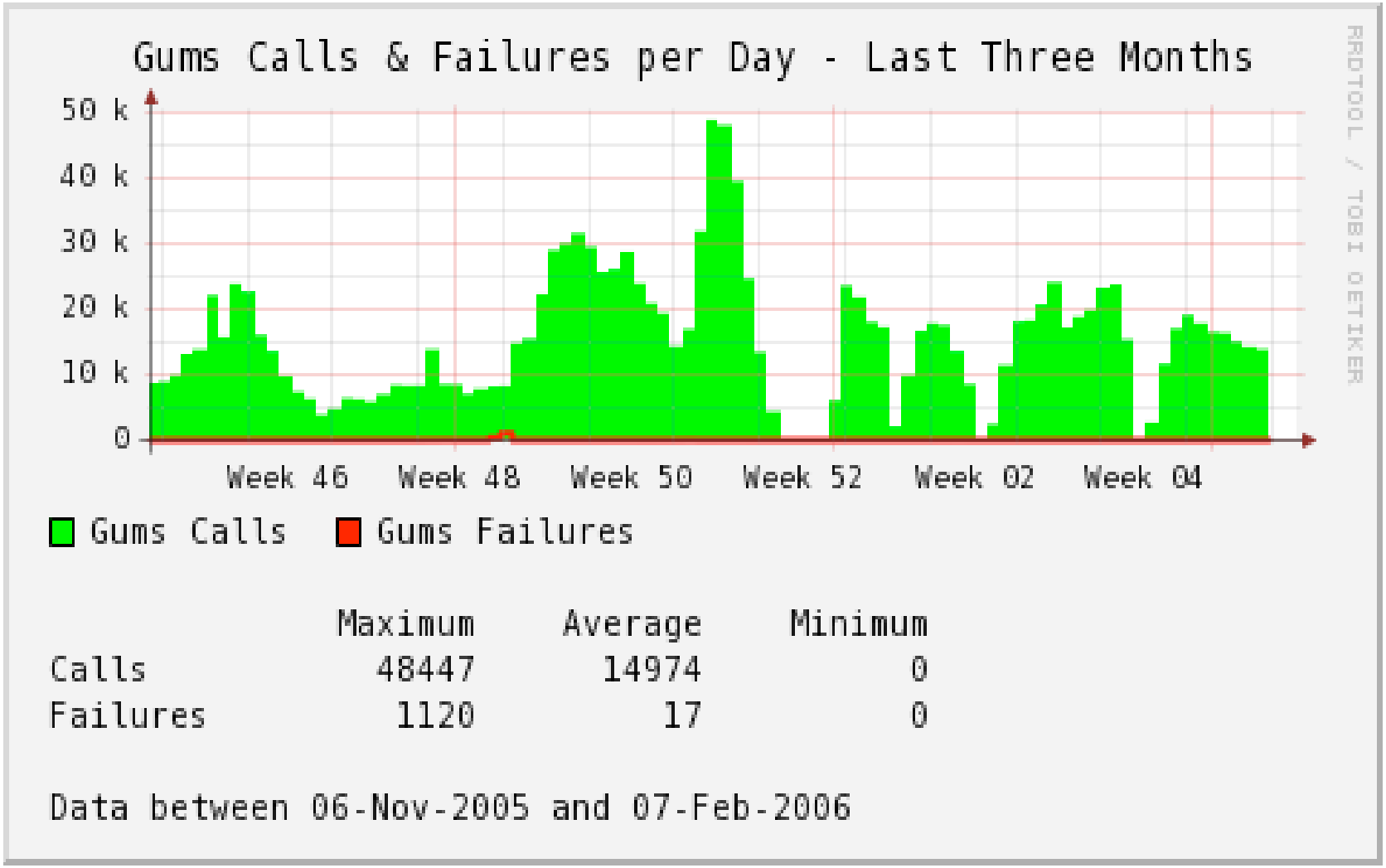
Monitoring and Accounting – VOMS:



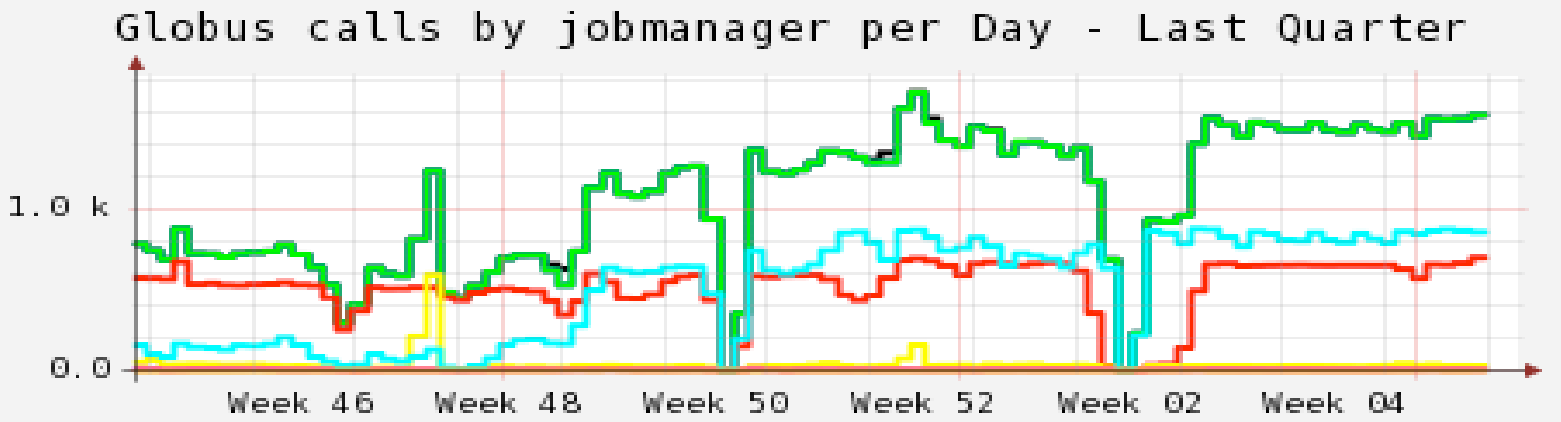
	Maximum	Average	Minimum
auger	0	0	0
cdf	0	0	0
cms	0	0	0
des	0	0	0
dzero	2	0	0
fermilab	23	4	0
gadu	1	0	0
ilc	0	0	0
lqcd	0	0	0
nanohub	0	0	0
sdss	8	1	0
uscms	13	4	0

Data between 06-Nov-2005 and 07-Feb-2006

Monitoring and Accounting – GUMS:



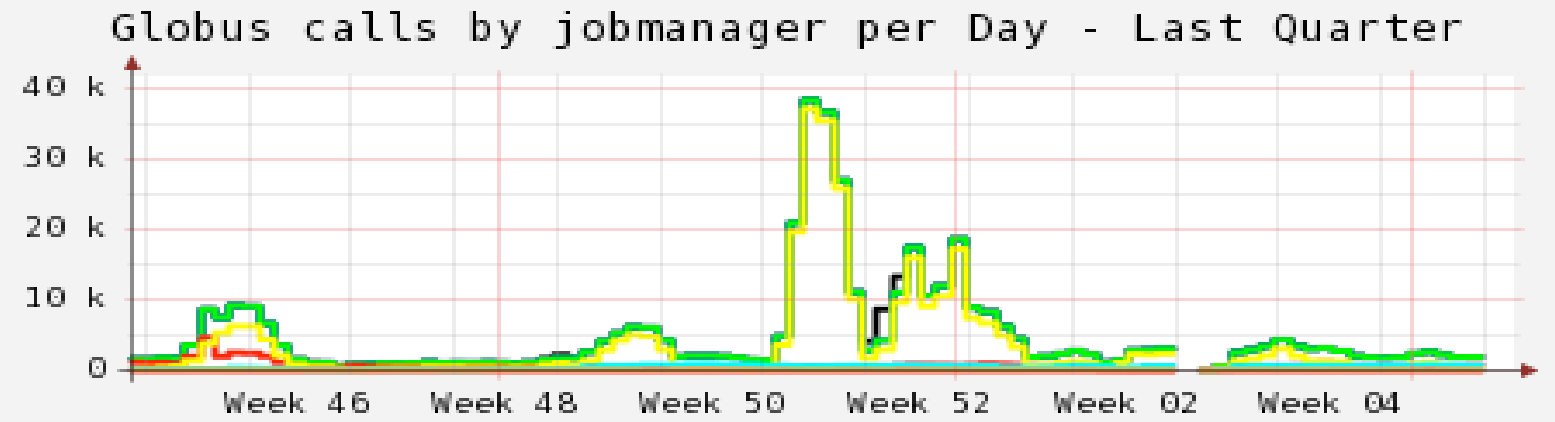
Monitoring and Accounting – fermigrid1:



	Maximum	Average	Minimum
authenticated	1726	1073	0
authorized	1715	1070	0
jobmanager	1719	1070	0
forks	696	517	0
condor	592	37	0
mis	12	6	0
other	876	511	0
condor-g	4	0	0

Data for between 06-Nov-2005 and 07-Feb-2006

Monitoring and Accounting – fngp-osg:



	Maximum	Average	Minimum
authenticated	38401	4875	0
authorized	38400	4773	0
jobmanager	38400	4773	0
forks	4702	737	0
condor	37154	3545	0
mis	21	7	0
other	891	485	0
condor-g	2	0	0

Data for between 06-Nov-2005 and 07-Feb-2006

Future Plans

Storage:

- More is better!

Desktop condor pool:

- Head node operated by the central FermiGrid operations team.
- Invite users to “donate” their unused Linux desktop cycles on an opportunistic basis.

Research & Development & Deployment of future ITBs and OSG releases…

- Ongoing work…
- Need to deploy testbed systems so we can perform the research, development & integration without impacting our production services.

Operational Experience:

User and Administrator Documentation:

- More is better:
- http://fermigrid.fnal.gov/user-guides.html
- http://fermigrid.fnal.gov/admin-guides.html
- http://fermigrid.fnal.gov/vo-guides.html (coming soon)

User Education:

- The <name deleted> VO were running jobs on the fngp-osg cluster which put a large (4.5GB) tarball in the $DATA area and then ran jobs on the worker nodes to untar it across nfs and write the output back to $DATA across nfs.
- This took several hours just to untar the tarball, then even more to read the data across NFS again.
- Steve Timm asked them to please use $WNTMP instead and they modified their application.

Service Level Monitoring:

- We have had two GUMS service outages.
- Analysis of these failures indicate that GUMS 1.0.1 was performing an inadvertent denial of service attack against the GUMS MySQL database due to the 4500+ members in the fermilab VO and the 13K+ entries in the GUMS mapping table.

Ongoing challenges:

- Learning how to operate a VO (fermilab) with 4500+ members.
- Learning how to operate a VOMS server with >12 VOs.
- Dealing with fallout caused by changes to the VOMS-GUMS interfaces.
- Security in the context of the Open Science Grid and DOE requirements.
- Progress limited by available personnel.
- Many others…

Fin

Any Questions?