

FermiGrid

D.R. Yocum, E. Berman, P. Canal, K. Chadwick, T. Hesselroth, G. Garzoglio, T. Levshina, V. Sergeev, I. Sfiligoi, N. Sharma, and S. Timm

Abstract— As one of the founding members of the Open Science Grid Consortium (OSG), Fermilab enables coherent access to its production resources through the Grid infrastructure system called FermiGrid. This system successfully provides for centrally managed grid services, opportunistic resource access, development of OSG Interfaces for Fermilab, and an interface to the Fermilab dCache system. FermiGrid supports virtual organizations (VOs) including high energy physics experiments (USCMS, MINOS, D0, CDF, ILC), astrophysics experiments (SDSS, Auger, DES), biology experiments (GADU, Nanohub) and educational activities.

Index Terms— Open Science Grid, Fermilab, High Energy Physics, infrastructure, authentication, authorization, certificates, VOMS, VOMRS, GUMS, glexec, gPlazma, SAZ.

1 INTRODUCTION

Fermilab is the premier high energy physics laboratory in the U.S. and supports a scientific program which includes experiments and scientists located across the globe. In order to better serve this community, Fermilab has placed its data analysis computer resources in a grid infrastructure called FermiGrid [1]. This model allows the large experiments at Fermilab to have priority access to their own resources and enables sharing of these resources in an opportunistic fashion across the various Virtual Organizations (VOs) that participate in FermiGrid. A virtual organization is a group of individuals which share grid resources to accomplish a common goal. As a member of the Open Science Grid (OSG) [2], Fermilab also allows usage of FermiGrid by several OSG virtual organizations. Through FermiGrid one may access our core grid services, the 3,000+ Fermilab resident CPUs, and the Fermilab Mass Storage System. The core grid services include a site wide Globus gateway [3], VO management services, a Fermilab site authorization service, grid user mapping service, job accounting, and monitoring. Access to these services is done via well supported grid interfaces.

2 USER REGISTRATION, AUTHENTICATION AND AUTHORIZATION

The OSG uses The Americas Grid Policy Management Authority (TAGPMA) federation of authorization providers as the base set of trusted certificate authorities

(CAs). Additional CAs have been added to FermiGrid systems on a case by case basis, but this is not encouraged for administrative reasons. Once a user obtains an X.509 certificate, they register their Distinguished Name (DN) with the VO of which they want to become a member using the Virtual Organization Membership Registration Service (VOMRS) [4] associated with that VO. Currently, Fermilab hosts a Virtual Organization Membership Service (VOMS) [5] for the following VOs:

- auger – Pierre Auger Cosmic Ray Observatory
- des – Dark Energy Survey
- dzero – D-Zero experiment at Fermilab
- fermilab – general Fermilab VO
- gadu – Genome Analysis and Database Update
- i2u2 – Fermilab Education and Outreach
- ilc – International Linear Collider
- lqcd – Lattice Quantum Chromodynamics
- nanohub – Network for Computational Nanotechnology
- sdss – Sloan Digital Sky Survey

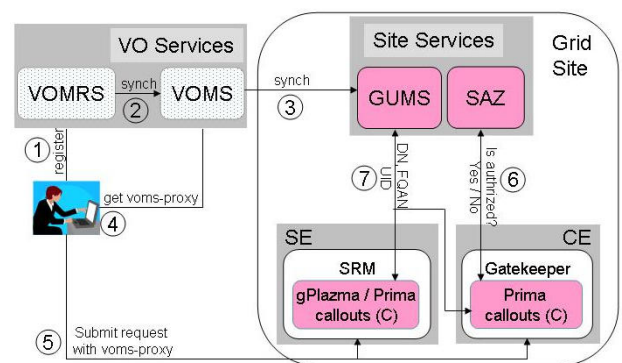


Fig. 1. Work flow of user registration, authentication, authorization.

- D.R. Yocum, Fermilab, Batavia, IL 60510 <yocum@fnal.gov>
- Eileen Berman, Fermilab, Batavia, IL 60510 <berman@fnal.gov>
- Philippe Canal, Fermilab, Batavia, IL 60510 <pcanal@fnal.gov>
- Keith Chadwick, Fermilab, Batavia, IL 60510 <chadwick@fnal.gov>
- Ted Hesselroth, Fermilab, Batavia, IL, 60510 <tdh@fnal.gov>
- Gabriele Garzoglio, Fermilab, Batavia, IL 60510 <garzogli@fnal.gov>
- Tanya Levshina, Fermilab, Batavia, IL 60510 <tlevshin@fnal.gov>
- Valery Sergeev, Fermilab, Batavia, IL 60510 <sergeev@fnal.gov>
- Igor Sfiligoi, Fermilab, Batavia, IL 60510 <sfiligoi@fnal.gov>
- Neha Sharma Batavia, IL 60510 <neha@fnal.gov>
- Steven Timm, Fermilab, Batavia, IL 60510 <timmm@fnal.gov>

VOMRS (Figure 1) maintains user registration attrib-

utes (contact information, home institution, etc.) and manages the registration workflow. A hierarchy of VO administrators approves the user registration to join groups and VO roles. Users often are members of multiple VOs and assume multiple roles within a VO. If plain grid certificates are used, that is, certificates without extended key attributes, a user would have to maintain individual certificates for each VO, group, and role combination. For users with many VO/group/role combinations, this can become an overly burdensome task. The infrastructure uses extended X.509 certificates to encode VO, groups and group role membership [6]. This allows a user to own a single certificate and belong to multiple VOs.

VOMRS pushes the user's DN and the VO extended attributes to the appropriate VOMS server, where they are stored. After this synchronization has completed, the user can generate a grid proxy with extended key attributes, also known as a VOMS proxy certificate, using `voms-proxy-init` (e.g., `voms-proxy-init -voms fermilab:/fermilab/usminos/Role=softadmin`).

Most, if not all, TAGPMA member CAs issue user grid certificates with a lifetime of one year, and a user can obtain a proxy certificate for up to the lifetime of the user certificate. This is not recommended for security reasons since the user's proxy certificate, by default, resides in `/tmp` for the lifetime of the certificate. As outlined above, a VOMS proxy certificate is composed of two parts: the user's grid proxy certificate and the VOMS extended key attributes proxy certificate. It is worth noting that the VOMS extended key attribute proxy certificate has a lifetime that is different from the grid proxy certificate; the default lifetime is 12 hours but can be easily modified to be longer (or shorter).

The Grid User Management System (GUMS) [7], maps grid credentials to site specific credentials. FermiGrid manages its own mapping as do other OSG sites. GUMS periodically polls the VOMS servers and updates its user DN plus extended attributes to local UID mappings database. This happens asynchronously on a 6 hour interval, introducing a possible delay before a grid user is able to execute a job on a batch system. The list of VOMS servers the GUMS server polls is manually configured by a system administrator.

Due to European privacy regulations, it has been discussed amongst the GUMS developers and the European grid security people to modify GUMS to only maintain identity information of users who run on a particular grid Site instead of maintaining a database of everyone in a particular VO, which is the current *modus operandi*. This enhancement would allow European grids to adopt this technology should they choose to do so.

There are three types of mapping supported by GUMS: many-to-one, one-to-one and one-to-self. In the many-to-one case, all members of the same VO having the same extended key attributes are mapped to a single local UID. In the OSG the many-to-one mapping is an acceptable usage case, however, in the European grids it

is not, due to the same privacy regulations mentioned above: with many-to-one mappings it is possible to obtain the proxy of another user running on the same batch system. Currently, the OSG prevents this through policies set in the Acceptable Usage Policy (AUP) document that all OSG users must electronically sign when they register with their VO. In the one-to-one usage case, a set of pool accounts is created on the batch system and GUMS maps a user DN with extended key attributes to a single, specific pool account. This mapping will be maintained in the GUMS database so that each consecutive request for mapping from that user using that specific set of extended key attributes will result in being mapped to the same pool account UID at that site (i.e., DN="Dan Yocum"+VO=cms/Role=cmsuser always maps to `uscms03` on FermiGrid worker nodes). The advantages of the one-to-one method enhance security by protecting the user's data and environment. One-to-one mapping also makes it easier to trace a rogue job in the queue to the originating user. One problem with one-to-one mapping is that it is the responsibility of the Site system administrator to make sure there are enough pool accounts available on each worker node (WN) in the batch system. The last type of mapping is one-to-self. As the name suggests, a user is mapped to her own local UID on the batch system. This mapping mechanism is employed at Brookhaven National Laboratory (BNL).

The mapping scheme that is used at a Site is a decision of the site administrators based on the request of the VO using the site. For instance, CMS has requested that users be mapped one-to-one on FermiGrid, while D-Zero is using the many-to-one mapping.

The scalability of the GUMS server has been remarkable. The hardware system, like all the FermiGrid critical systems, is comprised of a dual 3.06GHz Intel XEON with 4GB of PC2100 DDRAM, 2 mirrored 10K rpm SCSI system disks and a single gigabit ethernet connection to the LAN. GUMS is implemented as a Java Servlet and runs in Tomcat. The server is able to sustain a load of 450,000 mapping requests per day (5.2 mappings per second) with an average system load (1 minute) of about 2-3. It is anticipated that the system will easily be able to accommodate CMS production when the mapping requests will increase to >20Hz.

The final step in FermiGrid user authorization is the the Fermi Site Authorization Service (SAZ). SAZ allows security authorities of FermiGrid to impose a site-wide grid access policy based on user DN, VO membership or Certificate Authority. Registration with SAZ is automatic and transparent when a grid user first runs a job on FermiGrid and the user's DN and VO affiliation is entered into its database. The default policy is to allow access to users who present a valid and trusted VOMS proxy certificate. On a case-by-case basis, users can be allowed access to FermiGrid resources with valid non-extended grid credentials. Using SAZ's blacklist functionality, Site administrators have the ability, to prevent unauthorized access to compute and storage resources if

ing, the Collector attempts to find it from other sources (Gram, VOMRS, etc.)

OSG is currently deploying the Gratia Collector and Reporting Services on a single central server, located at Fermilab. In future releases, the collector and reporter will also be installed at each OSG Site.

To date, Gratia Probes have been developed for Condor, PBS, LSF, Sun Grid Engine, glexec, and dCache. These probes have been deployed both via RPMs, distributed from the OSG web site, and via the Virtual Data Toolkit (VDT) cache.

6 USER DISK SPACE AND MASS STORAGE

FermiGrid supplies shared data storage and access services to user jobs, providing two pools of data storage for general usage. The first is 14TB of disk space, served by an NFS server appliance for grid users to install software, stage input data, and write output data. This space is mounted on most of the worker nodes and uses quotas to restrict the amount of storage accessible to a user. In addition to this NFS disk, there is 6TB of disk spread across 5 storage servers. This disk is managed by the dCache system and does not use quotas but implements a least recently used cache management policy. The external interface to this storage resource is either direct via gridftp or by utilizing the Storage Resource Manager (SRM) [16] interface. SRM provides dynamic space allocation and file management functionality on local and remote data storage systems, using grid credentials for data access.

The FermiGrid dCache uses gPlasma [17] to interface to the GUMS service to map user DNs to local UIDs for file ownerships. The same DN to UID mapping schemes described earlier are available here: many-to-one, one-to-one and one-to-self. It should be noted again that in the many-to-one mapping, all members of the VO have access to all data written by other members of the same VO and data can be easily shared between members of the same VO. The same is true of the one-to-one mapping scheme if the owner of the file enables group read permissions.

Authorized users may have access from their jobs to the Fermilab tape storage system. In general, VOs not directly associated with Fermilab do not have access to the tape backed storage areas; however, upon special request this can be arranged.

FermiGrid provides a set of standard data locations referenced by environment variables so that users can access their data and applications in a consistent manner. The same is true for all OSG sites.

FermiGrid does not supply any tools for aiding the user in management of data collections.

7 WORKER NODES

As of February 2007, there are ~4450 CPUs at Fermilab that are available for use by OSG members. The breakdown by Fermilab experiment is shown in Figure 3.

| Experiment | Gatekeeper | CPUs | RAM | Disk |
|------------|----------------------|---------------------------|-----|-------|
| CMS | cmsosgce | 700 dual & dual dual core | 4GB | 250GB |
| CDF | fcdfosg1 fcdfosg2 | 520 dual core, dual | 4GB | 250GB |
| D-Zero | d0cabosg2 | 200 dual | 2GB | 250GB |
| GP Farm | fngp-osg | 220 various | 2GB | 250GB |

Figure 3: Fermilab CPUs Available to the OSG

VO Job Submissions for the Week of Feb 5-12, 2007

| Site Name | VO | sum |
|-----------------|-------------------|-------|
| FNAL_CDFOSG_1 | nanohub | 3 |
| | usecms | 3 |
| | gadu | 76 |
| | cdf | 5572 |
| | star | 101 |
| FNAL_CDFOSG_2 | nanohub | 12 |
| | star | 100 |
| | cdf | 6964 |
| | gadu | 38 |
| FNAL_DZEROOSG_2 | Unknown | 0 |
| FNAL_FERMIGRID | fermilab | 111 |
| | Unknown | 289 |
| | usecms | 2 |
| | star | 104 |
| | nanohub | 260 |
| | mis | 1551 |
| FNAL_GPFARM | mis | 1285 |
| | engage | 536 |
| | fermilab | 101 |
| | star | 8 |
| | ktev | 113 |
| | mipp | 4641 |
| | sdss | 317 |
| | cdms | 1316 |
| | nanohub | 561 |
| | ile | 57 |
| | LIGO | 1028 |
| | gadu | 11 |
| | USCMS-FNAL-WC1-CE | cem |
| uscms | | 45807 |
| gadu | | 25 |
| escience | | 51 |
| usatlas | | 76 |
| dteam | | 122 |
| fermilab | | 108 |
| nanohub | | 643 |
| dzero | | 977 |
| star | | 236 |
| cms | | 35804 |
| mis | | 984 |

Figure 4: Weekly Job Submissions to FermiGrid Gatekeepers (Unknown: submitted with vanilla grid proxies)

8 OPERATIONAL EXPERIENCE

Operational experience with FermiGrid has been very good. The FermiGrid services are being used by several experiment clusters and allow experiment support personnel to direct their effort to supporting their cluster rather than the middleware services. FermiGrid has also fostered the opportunistic use of idle cycles on the experiment clusters by multiple VOs (Figure 4).

The personnel required to integrate and operate the

FermiGrid hardware and middleware services have been approximately 3 FTEs over the course of calendar 2006. This includes several deployments of major enhancements to the underlying Grid middleware suite together with the development of an extensive operational metrics and service monitor infrastructure, which collects and publishes information for the underlying Grid middleware services.

The metrics collection occurs once a day and collects information for the previous day. The service monitors run multiple times per day (typically once per hour) and gather detailed information about the service that they are monitoring. The service monitors also verify the health of the service that they are monitoring (together with any dependent services), notify administrators if problems are detected and are instrumented to automatically restart the service(s) as necessary to insure continuous availability.

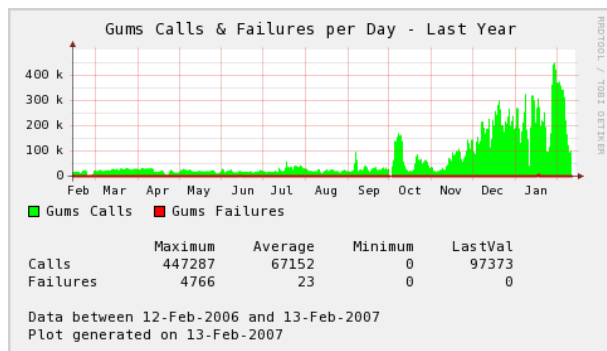


Figure 5: Increase in GUMS Calls

Use of FermiGrid services is expected to grow in the future. Increases in scale in the past year have been handled well by the deployed services. (See Figure 5).

9 FUTURE WORK

Fermilab is looking at improving the security of the Grid by introducing end-to-end integrity checks and encryption via a system called Epensys. With the proliferation of Workload Management Systems (WMS) and job portals, we have reached a situation, very much analogous to that of secure email transport, where the endpoints must be secure, but the intermediate services may not be.

However, unlike the secure email scenario, the final destination is not known at submission time, so Public Key Infrastructure (PKI) cannot be used for the encryption needed to protect the proxy used by the job at runtime. Epensys addresses the problem by requiring the use of a symmetric key and by storing the key in a Grid-central key store service. Moreover, since the intermediate services cannot be trusted, a very limited proxy, useful for user authentication and the running of the specific job only, is used in place of the full user proxy at submission time. After the payload reaches a trusted worker node, the symmetric key will be accessed by presenting

both the limited proxy and the host certificate.

Balancing the needs of our robust, shared job grid with those of a data grid will be necessary for the demands of both LHC experiments and the ever increasing number of OSG VOs. Work to support this will need to be addressed in the forthcoming year.

10 CONCLUSION

FermiGrid is a well designed and deployed, robust grid gateway system based on standard grid software and interfaces. On an average day our gatekeepers accept tens of thousands of job submissions and our GUMS server performs hundreds of thousands of user mapping authorizations, and authorizes the transfer of hundreds of terabytes of data in and out of the Fermilab Mass Storage System. Our uptime is over 99% and there are planning on making the system Highly Available to increase that reliability.

ACKNOWLEDGMENT

We would like to acknowledge all of the projects and people whose hard work went in to the construction of FermiGrid.

REFERENCES

- [1] FermiGrid <http://fermigrid.fnal.gov>
- [2] OSG <http://www.opensciencegrid.org>
- [3] Globus: <http://www.globus.org>
- [4] VOMRS: http://computing.fnal.gov/docs/products/vomrs/vomrs1_2/whhelp/whimpl/common/html/default.htm
- [5] VOMS: <http://hep-project-grid-scg.web.cern.ch/hep-project-grid-scg/voms.html>
- [6] Tuecke, S., Engert, D., Pearlman, L., Thompson, M., "Internet X.509 Public Key Infrastructure (PKI) Proxy Certificate Profile" RFC 3820
- [7] GUMS: <http://grid.racf.bnl.gov/GUMS/>
- [8] ReSS: <https://twiki.grid.iu.edu/twiki/bin/view/ResourceSelection>
- [9] CEMon: <http://grid.pd.infn.it/ceмон/field.php>
- [10] Raman, R., Livny, M., Solomon, M., "Matchmaking: Distributed Resource Management for High Throughput Computing", Proceedings of the Seventh IEEE International Symposium on High Performance Distributed Computing, July 28-31, 1998, Chicago, IL.
- [11] Condor: <http://www.cs.wisc.edu/condor/>
- [12] Andreozzi, S., Garzoglio, G., Reddy, S., Mambelli, M., Roy, A., Wang, S., Wenaus, T., "GLUE Schema v1.2 Mapping to Old ClassAd Format", Technical Document, July 24, 2006
- [13] MyProxy: <http://grid.ncsa.uiuc.edu/myproxy/>
- [14] glxec: <http://infn-ecgi.pi.infn.it/documentation/glexec.pdf>
- [15] Gratia: <http://gratia-fermi.fnal.gov:8882/gratia-reporting/>
- [16] SRM: <https://srm.fnal.gov/twiki/bin/view>
- [17] gPlazma: <http://www.dcache.org/manuals/Book/cf-gplazma.shtml>
- [18] Prima: <http://vdt.cs.wisc.edu/components/prima.html>