FermiGrid

24x7 Service Support

Keith Chadwick
15-Oct-2008

Abstract:

This document describes the 24x7 support for FermiGrid Services

Document Revision History:

Version	Date	Author	Comments
0.1	09-Sep-2008	Keith Chadwick	Initial Version
0.2	17-Sep-2008	Keith Chadwick	New draft that incorporates initial feedback.
0.3	19-Sep-2008	Keith Chadwick	Incorporate feedback to V0.2 of this document.
0.4	22-Sep-2008	Keith Chadwick	More tweaks and enhancements.
0.5	06-Oct-2008	Keith Chadwick	Incorporate comments from Steve Wolbers
0.6	15-Oct-2008	Keith Chadwick	More document updates.

1. Introduction

This document assumes that the reader is familiar with the concepts and practices of Grid computing as deployed in the Open Science Grid (OSG) and the Fermilab Grid deployment (FermiGrid).

FermiGrid operates a large set of services in direct and indirect support of Grid computing at Fermilab. These services are categorized as either "critical" or "non-critical" (refer to Section 2 below for the definition of these categorizations). The current full FermiGrid Service Catalog is listed in Appendices A, B and C (the current set of "critical" services is also listed in Section 3 below).

As one component of assuring 24x7 service availability for "critical" services, FermiGrid has undertaken a strategic plan of deploying all services in a High Availability (HA) infrastructure. To accomplish this plan, FermiGrid has deployed several leading edge technologies (Xen, Linux Virtual Server and MySQL circular replication) in an infrastructure that is collectively referred to as "FermiGrid-HA".

The goal of the FermiGrid-HA deployment is to come as close as is reasonably achievable to Continuous Availability (CA) of those services in the FermiGrid Service Catalog. To accomplish this goal, the FermiGrid-HA infrastructure deploys services using several strategies and components that collectively offer continuous processing without skipping a "heartbeat".

Five 9's (99.999%) is a widely held industry standard for service availability. It is an extremely difficult standard to achieve since five 9's availability for a service corresponds to ~5 minutes total downtime for an entire year. The following table shows the total downtime for the entire year that is allowed to meet the two, three, four and five 9's availability standards:

Availability	**365 Days**	**8760 Hours**	**525600 Minutes**	**31536000 Seconds**
99%	3.65	87.6	5256	315360
99.9%	0.365	8.76	525.6	31536
99.99%	0.0365	0.876	52.6	3153.6
99.999%	0.00365	0.0876	5.3	315.36

For the period from 01-Dec-2008 through 30-Jun-2008, FermiGrid was able to achieve a service availability of 99.9969% (ten minutes of downtime in seven months of operation).

Ideally this measurement should include power and cooling. So far the FermiGrid service systems have been allowed to "ride out" the FCC cooling outages. For the purposes of FermiGrid availability calculations, since FermiGrid does not control the power and cooling availability, any outage of these services is factored out before calculating FermiGrid availability.

In an attempt to minimize the potential negative impact of power outages, the FermiGrid-HA systems (fermigrid0 through fermigrid6) are equipped with dual power supplies, each of the dual power supply cords are fed through independent power controllers plugged into independent power panels. Unfortunately, both power panels are connected to a single UPS. This is an item that should be considered when the FermiGrid systems are relocated from FCC1 to FCC2 as part of the ongoing FCC computer room consolidation.

In the event of catastrophic failure of either the FermiGrid-HA systems in FCC, FCC power and/or cooling, or FCC itself, the FermiGrid Contingency Plan calls for the utilization of the FermiGrid Test (fgtest) systems in the FermiGrid Dell relay rack located in "Falcon's Nest" area on Wilson Hall 8 East to provide service availability during the period that the production FermiGrid-HA systems are unavailable.

At the present time, the majority of the services listed in the FermiGrid Service Catalog are available as "HA" services. The current exceptions are the Globus Gatekeepers, the (old & new) MyProxy service, the OSG Resource Selection Service (ReSS) and the OSG Accounting Repository (Gratia). Efforts are underway to address these services. The eventual goal is to have all of the services listed in the FermiGrid Service Catalog available as "HA" services.

It should be noted that the "virtual machines" in the strategy list below are not actual physical systems; they are Xen Domain-User (Dom-U) systems Virtual Machines (VMs) running under the Xen Hypervisor (Domain-0 / Dom-0) on a small set of physical systems. When a service is deployed on multiple independent virtual machines, the virtual machines are hosted on distinctly separate physical systems.

FermiGrid employs several strategies to deploy "HA" services:

- "Trivial" monitoring or information services (such as Ganglia and Zabbix) are deployed on at least two independent virtual machines.

- Services that natively support "HA" operation (Condor Information Gatherer, FermiGrid internal ReSS deployment) are deployed in the standard service "HA" configuration on at least two independent virtual machines.

- Services that maintain intermediate routing information (Linux Virtual Server) are deployed in an active/passive configuration on at least two independent virtual machines. A periodic heartbeat process is used to perform any necessary service failover.

- Services that maintain intermediate context (Globus Gatekeepers, MyProxy), and do not natively support "HA" operation, are (will be) deployed in either an active/passive or active/active configuration on at least two independent virtual machines. These systems are configured to store all necessary state information in a Distributed Replicated Block Device (DRBD). If the service is deployed as an

active/passive service, then a periodic heartbeat process is used to perform any necessary service failover.

- Services that do not maintain intermediate context (i.e. are pure request/response services such as GUMS and SAZ) are deployed using a Linux Virtual Server (LVS) front end to active/active servers on at least two independent virtual machines.

- Services that support active-active database functions (circularly replicating MySQL servers) are deployed on at least two independent virtual machines.

The following services are not currently implemented as natively "HA" services:

- Globus gatekeeper services that are not currently implemented as "HA" services (such as the CDF and D0 experiment globus gatekeeper services) are deployed in "segmented" pools – loss of any single pool will reduce the available resources by approximately 50%.

- Certain Open Science Grid (OSG) services (such as the OSG accounting repository [Gratia] and the OSG Resource Selection Service [ReSS]) are not currently implemented as "HA" services. If the service fails, then the service will not be available until appropriate manual intervention is performed to restart the service. The FermiGrid FY2009 tactical plan lists activities, that if funded, would migrate these services onto a "HA" deployment.

2. Service Outage Impact Definitions

A critical service is defined as follows – any service that would prevent Grid jobs from being accepted and/or run on a Fermilab based Grid Computing Element (CE) or any service that would prevent data from being written to or read from a Fermilab based Storage Element (SE). Examples of such services are the Grid User Mapping Service (GUMS) and Site AuthoriZation (SAZ) service. At the start of each Grid job, the Globus gatekeeper must contact the FermiGrid GUMS and SAZ servers with the credentials of the users job to determine the account name that the Grid job should be run under and if the user that submitted the Grid job is authorized to run Grid jobs at Fermilab.

All other services are defined as non-critical services. An outage of a non-critical service may delay Grid job processing, reduce Grid job processing capacity, or impact other secondary services that a Grid job may optionally invoke. Examples of such services are the experiment Globus Gatekeepers and the Squid web cache. In the case of the experiment Globus Gatekeepers, multiple Globus Gatekeepers are deployed in front of the experiments worker nodes; loss of any single Globus Gatekeeper will only temporarily reduce the number of worker nodes available to the experiment. In the case of the Squid web cache, properly configured clients will first attempt to contact the web URL using the Squid web cache; if the Squid web cache is not available, then the web client will contact the web URL directly.

3. List of Critical Services

The current list of critical FermiGrid services is:

Service	Service Description	Service Hosts	Host Location
Xen Dom-0	Xen Hypervisor	fermigrid5 and fermigrid6	FCC1
LVS	Linux Virtual Server	fg5x0 and fg6x0	FCC1
GUMS	Grid User Mapping Service	fg5x2 and fg6x2	FCC1
SAZ	Site AuthoriZation Service	fg5x3 and fg6x3	FCC1
MySQL	Redundant circularly replicating MySQL	fg5x4 and fg6x4	FCC1

Additional services may be added to this list in the future.

4. Service Monitoring and Critical Service Verification

FermiGrid has implemented automated service monitoring that runs every hour for all services in the FermiGrid service catalog. The FermiGrid monitoring is currently hosted/managed from fermigrid0 and runs across all the FermiGrid systems. As part of the commissioning of the new fgtest systems, a second instance of the FermiGrid monitoring will be established on fgtest0 that will run at a staggered interval.

The results of the service monitoring are available at the following URL:

http://fermigrid.fnal.gov/fermigrid-metrics.html

The service monitoring verifies that:

- The system that the service is deployed on is running correctly.
- The service is running correctly.
- Collects a variety of "interesting" information about the services.

In the event that the service monitor detects a service outage, the service monitoring scripts will attempt an automatic restart of the failed service (if possible).

In addition, for the GUMS and SAZ services, FermiGrid has deployed automated service verification (every minute). The service verification verifies that:

- The service has successfully processed service requests within the previous 60 seconds.

- The required service processes exist on the system.

- The required service processes are "healthy" – the Virtual Memory Size (VMS) and Resident Set Size (RSS) are within predefined limits.

If the elapsed time between the current time and the last logged service request exceeds 60 seconds, or the service process parameters are above previously defined notification thresholds, the service verification script automatically notifies FermiGrid personnel (details of this notification process are in Section 5 below).

If the elapsed time between the current time and the last logged service request exceeds 300 seconds, or the service process parameters are above previously defined action thresholds, the service verification script automatically attempts to restart the service and notifies FermiGrid personnel.

The service process parameter limits (VMZ and RSS) together with the service response thresholds (60 and 300 seconds) were established based on a review of the FermiGrid service specific performance and diagnostic information available from the FermiGrid Service Monitor and Metrics web pages:

http://fermigrid.fnal.gov/fermigrid-metrics.html

5. Outages of Critical Services

As noted above, FermiGrid has implemented automated service monitoring for all services in the FermiGrid service catalog, and service verification for the GUMS and SAZ critical services. The service monitoring and service verification will notify FermiGrid personnel via email to the FermiGrid "MONITOR_MAILTO" service monitor list (fermigrid-monitoring@fnal.gov) and will automatically attempt to restart failed services. Notification of certain failures (such as restarts of critical services) are also emailed to the FermiGrid "MONITOR_PAGETO" service monitor email list.

Note that email to the MONITOR_MAILTO and MONITOR_PAGETO service monitor lists is **entirely separate** from any Computing Division HelpDesk notification or escalation procedures. This separation is deliberate – significant amounts of performance and diagnostic information is automatically sent to the MONITOR_MAILTO and MONITOR_PAGETO lists that is not appropriate to be included in the Computing Division HelpDesk notification or escalation procedures.

In the event of a critical service failure that is not resolved by the existing automated restart mechanisms, then the service client management or support personnel should escalate to the FermiGrid off-hours support rotation through the mechanisms detailed in Section 7 (below). In addition, they may elect to exercise the following options to assure service continuity while the personnel in the FermiGrid support rotation resolve the service outage incident:

5.1 **Temporarily** modify their Globus Gatekeeper and/or gLExec configurations to use the corresponding services hosted on the FermiGrid test/development systems identified in Appendix B. If the service client elects to perform this, then FermiGrid personnel **must** be notified of this action through the fermigrid-help@fnal.gov email list.

5.2 If the service outage is limited to SAZ, then as long as the Fermilab Computer Security Incident Response Team (FCIRT) is not actively investigating a Grid security incident, the service client management or support personnel may elect to **temporarily** remove the SAZ client callouts from their Globus Gatekeeper and/or gLExec configuration. If the service client elects to perform this action, then **both** FCIRT and FermiGrid personnel **must** be notified of this action through **both** the computer-security@fnal.gov and the fermigrid-security@fnal.gov email lists.

5.3 If the service outage is limited to GUMS, then the service client management or support personnel may elect to **temporarily** modify the PRIMA client callouts in their Globus Gatekeeper and/or gLExec configuration to utilize a GridMapFile. If the service client elects to perform this action, then FermiGrid personnel **must** be notified of this action through the fermigrid-help@fnal.gov email list.

6. Outages of Non Critical FermiGrid Services

Service client management or support personnel that identify outages of non-critical FermiGrid services should notify FermiGrid personnel by the appropriate CD HelpDesk ticketing system and/or by email to the fermigrid-help@fnal.gov email list.

FermiGrid personnel will offer 8x5 support plus "best effort" off-hour response for non-critical services.

7. Off Hours Contact of FermiGrid Personnel

In the event of a critical service failure that is not resolved by the existing automatic restart mechanisms, the service consumers should escalate to the FermiGrid off-hours support rotation through the standard Computing Division HelpDesk off hours paging procedures.

The Computing Division HelpDesk will contact the current FermiGrid Primary:

- Attempt 1 - Contact the current FermiGrid primary via pager or cell phone.
- Attempt 2 - Contact the current FermiGrid primary via pager or cell phone.
- Attempt 3 - Contact the current FermiGrid primary via home phone.

If the current FermiGrid Primary does not respond to the incident within the response interval, then the Computing Division HelpDesk will escalate to the current FermiGrid Secondary:

- Attempt 1 - Contact the current FermiGrid secondary via pager or cell phone.
- Attempt 2 - Contact the current FermiGrid secondary via pager or cell phone.
- Attempt 3 - Contact the current FermiGrid secondary via home phone.

It is expected that the person in the FermiGrid support rotation that receives the incident notification shall attempt to respond to the incident within 15 minutes if the notification occurs during standard business hours (Monday through Friday 8:00 through 17:00), and within 1 (one) hour for all other times, providing that this response interval does not create a hazard (i.e. don't drive 100 MPH in order to get home to log into FermiGrid in less than 1 hour).

If neither the current FermiGrid primary or secondary respond within the response interval, then the subsequent incident escalation shall be via the FermiGrid line management chain:

- FermiGrid Services Group Leader (or designee)
- Grid Department Head (or designee)
- Scientific Facilities Quadrant Head (or designee)
- Computing Division Head (or designee)

——————— Note ———————

It should be noted that there is a lot of documentation and procedures to be written, and cross training to be performed in order to make sure that the personnel in the FermiGrid off-hours support rotation (FermiGrid and/or otherwise) know what to do. This documentation and cross training will be an ongoing process that will require corresponding commitments from both the personnel in the FermiGrid off-hours support rotation and their line management.

8. Future FermiGrid Plans

The specific design and implementation of the FermiGrid High Availability (HA) infrastructure was selected in order to allow the eventual migration from a High Availability deployment to a Geographically Redundant Services (RS) deployment. The evolution from the existing FermiGrid-HA deployment to the FermiGrid-RS deployment is expected to require the acquisition of additional hardware that would be deployed at one or more locations other than FCC together with additional network infrastructure to interconnect these locations in an extended switched LAN, the existing software configuration already offers support for FermiGrid-RS.

FermiGrid project management is considering proposing the acquisition and deployment of the additional systems and network infrastructure that would be required to evolve the current FermiGrid-HA deployment into a FermiGrid-RS deployment during FY2010.

Appendix A – FermiGrid Production Service Catalog

Service	Critical (Y/N)	Service Host	Host Location	Service Description
Xen Dom-0	Y	fermigrid5 fermigrid6	FCC1	Xen Hypervisors. fermigrid5 hosts fg5x[0-4], fermigrid6 hosts fg6x[0-4]
LVS	Y	fg5x0 (active) fg6x0 (standby)	FCC1	Linux Virtual Server
VOMS	N	fg5x1 (active) fg6x1 (active)	FCC1	Virtual Organization Management Service
GUMS	Y	fg5x2 (active) fg6x2 (active)	FCC1	Grid User Mapping Service
SAZ	Y	fg5x3 (active) fg6x3 (active)	FCC1	Site AuthoriZation service
MySQL	Y	fg5x4 (active) fg6x4 (active)	FCC1	Redundant circularly replicating MySQL database service
Xen Dom-0, Monitoring	N	fermigrid0	FCC1	Xen Hypervisor. fermigrid0 hosts fg0x[0-7]. Collection of Service Metrics (daily), Service Monitoring (hourly), Service Verification (minute by minute).
Development	N	fg0x0	FCC1	
Development	N	fg0x1	FCC1	
Development	N	fg0x2	FCC1	
Development	N	fg0x3	FCC1	
Development	N	fg0x4	FCC1	
Development	N	fg0x5	FCC1	
Development	N	fg0x6	FCC1	
Development	N	fg0x7	FCC1	
Xen Dom-0	N	fermigrid1 fermigrid4	FCC1	Xen Hypervisors. fermigrid1 hosts fg1x[0-4], fermigrid4 hosts fg4x[0-4]
OSG ST&E	N	fg1x0	FCC1	Open Science Grid Security Test & Evaluation reporting service
Site Wide Gateway	N	fg1x1 (active) fg4x1 (standby)	FCC1	FermiGrid Site Wide Globus Gateway
Information Gatherer	N	fg1x2 (standby) fg4x2 (active)	FCC1	Information Gatherer used by FermiGrid Site Wide Globus Gateway
Condor Master	N	fg1x3 (active) fg4x3 (future)	FCC1	Condor Master for FermiGrid Site Wide Globus Gateway
OSG-TG Gateway	N	fg1x4 (standby) fg4x4 (active)	FCC1	Open Science Grid to TeraGrid Gateway
Xen Dom-0	N	fermigrid2 fermigrid3	FCC1	Xen Hypervisors. fermigrid2 hosts fg2x[0-4], fermigrid3 hosts fg3x[0-4]
Syslog-Ng	N	fg2x1 (future) fg3x1 (active)	FCC1	FermiGrid Syslog-Ng collectors / forwarders to CST clogger
Ganglia Zabbix	N	fg3x2 (active) fg3x3 (active)	FCC1	FermiGrid Ganglia and Zabbix Monitoring Hosts
Squid	N	fg2x3 (active) fg3x3 (active)	FCC1	FermiGrid Squid Web Caching Service
MyProxy	N	fg2x4 (active) fg3x4 (active)	FCC1	(new) FermiGrid MyProxy Service
MyProxy	N	fg4x1 (active)	FCC1	(old) FermiGrid MyProxy Service
Web Server	N	fermigrid1	FCC1	FermiGrid Web Server
Vomrs	N	fg2x1	FCC1	VOMRS

Appendix B – FermiGrid Testbed Service Catalog

Service	Critical (Y/N)	Service Host	Host Location	Service Description
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	
	N		WH8SE	

Appendix C – Other FermiGrid Services

Service	Critical (Y/N)	Service Host	Host Location	Service Description
Globus Gatekeepers	N	D0cabosg1 D0cabosg2	FCC1	
Pbs Masters	N	D0cabsrv1 D0cabsrv2	FCC1	
Globus Gatekeepers	N	Fcdfosg1 Fcdfosg2	FCC1	
Condor Master	N	Fcdfcm1/2	FCC1	
Globus Gatekeepers	N	Fcdfosg3 Fcdfosg4	FCC1	
Condor Master	N	Fcdfcm3	FCC1	
Xen	N	Fnpcsrv3	FCC1	
Globus Gatekeeper	N	Fnpc3x1	FCC1	
Xen	N	Fnpcsrc4	FCC1	
Globus Gatekeeper	N	Fnpc4x1	FCC1	
Xen	N	Fnpcsrv5	FCC1	
Condor Master	N	Fnpccm1 Fnpc5x1	FCC1	
Gratia	N	Gratia01	FCC1	
Gratia	N	Gratia02	FCC1	
Gratia	N	Gratia03	FCC1	
Gratia	N	Gratia05	FCC1	
Gratia	N	Gratia06	FCC1	
Gratia	N	Gratia07	FCC1	
Gratia	N	Gratia08	FCC1	
Gratia	N	Gratia09	FCC1	
ReSS (OSG production)	N	Osg-ress-1	FCC2	
ReSS	N	Osg-ress-2	FCC2	
ReSS	N	Osg-ress-3	FCC2	
ReSS (OSG itb)	N	Osg-ress-4	FCC2	
SamGrid	N	samgsrv2	FCC2	
Master YP Server for GP Grid cluster	N	fnpc5x4	FCC1	

Appendix D –Diagnosis of Service Failures

Service failures are observed in one or more of the following locations:

1. The Globus Gatekeeper and corresponding log.
2. The gLExec logs on the worker nodes.
3. The users log.
4. The FermiGrid Service Monitor Web Page (http://fermigrid.fnal.gov/fermigrid-metrics.html)

Physical Systems, Xen Hypervisor (Dom-0) and Xen Dom-U system Diagnostics:

Here is the process to diagnose failures of the systems:

1. Verify that the required physical environment is nominal:

   FCC1 has power.
   FCC1 environment (HVAC) is acceptable.
   FCC1 network subnet 107 is functioning.
   Primary and secondary name servers are functional.

2. Verify that the physical system and/or Xen Hypervisor (Dom-0) is nominal:

   ping –c 3 –t 10 –v <physical system IP name>
   ssh –v root@<physical system IP name>
   xm list

3. Verify that both the Xen VM (Dom-U) systems are nominal:

   ping –c 3 –t 10 –v <Dom-U system IP name 1>
   ssh –v root@<Dom-U IP name 1>

   ping –c 3 –t 10 –v <Dom-U system IP name 2>
   ssh –v root@<Dom-U IP name 2>

Globus_Gatekeeper Diagnostics:

External diagnostics:

The utilities for external diagnosis of the globus gatekeeper are the globusrun, globus-job-run and condor_submit utilities:

globusrun –a –r <gatekeeper-name>/jobmanager-fork

globus-job-run <gatekeeper-name>/jobmanager-fork /bin/hostname

condor_submit <???>

Note that a VOMS qualified proxy must be generated prior to calling these utilities.

Internal diagnostics:

The globus gatekeeper log file locations are:

Globus Toolkit Version	Gatekeeper Log Location
Pre Web Services (GT 3.2)	/usr/local/vdt/globus/var/globus-gatekeeper.log
Web Services (GT 4.0)	/usr/local/vdt/globus/var/container.log /usr/local/vdt/globus/var/container-real.log

tail –f <gatekeeper_log_location>

Linux Virtual Server (LVS) Diagnostics:

External diagnostic:

There are no external utilities to directly verify the LVS service function.

Internal diagnostic:

Log onto the LVS service Dom-U systems and ???

Grid User Mapping Service (GUMS) Diagnostics:

External diagnostic:

<?need new tool?>

Internal diagnostic:

Log into the GUMS server systems (currently fg5x2 and fg6x2) and tail the GUMS server log:

tail –f /usr/local/vdt/tomcat/v55/logs/gums-service-admin.log

Site AuthoriZation (SAZ) Service Diagnostics:

External diagnostic:

Generate a voms proxy and invoke:

sazclient –s saz.fnal.gov –p 8888 –x <proxy>

Internal diagnostic:

Log into the SAZ server systems (currently fg5x3 and fg6x4) and tail the SAZ server log:

tail –f /usr/local/vdt/saz/server/log/sazserver.log

MySQL Diagnostics:

External diagnostic:

The MySQL systems are protected via IPtables, so a general external diagnostic will not show meaningful data.

Internal diagnostic:

Log into the MySQL server systems (currently fg5x4 and fg6x4) and issue the following MySQL queries:

mysql –u root SAZDB –e "select * from user"

mysql –u root GUMS_1_1 –e "select * from MAPPING"

Squid Diagnostics:

External diagnostic:

Export the http_proxy environment variable:

export http_proxy=http://squid.fnal.gov:3128
export http_proxy=http://fg2x3.fnal.gov:3128
export http_proxy=http://fg3x3.fnal.gov:3128

After each of the above exports, issue the following command:

wget –O /dev/null http://fermigrid.fnal.gov/index.html

Internal diagnostic:

Log into the squid servers and tail the log file:

tail –f /usr/local/vdt/squid/var/log/access.log