FermiGrid-HA Phase 2
(FermiGrid-HA2)

Keith Chadwick
21-Apr-2011

Abstract:

This document discusses the requirements, design and deployment for FermiGrid-HA (High Availability) Phase 2 (FermiGrid-HA2).

Document Revision History:

Version	Date	Author	Comments
0.1	16-Mar-2010	Keith Chadwick	Initial highly draft version.
1.0	18-Mar-2010	Keith Chadwick	Draft placed in CD-DocDB
1.1	24-Mar-2010	Keith Chadwick	Revise network diagram.
1.2	30-Sep-2010	Keith Chadwick	Update following meeting with facilities.
1.3	05-Jan-2011	Keith Chadwick	Update following work on 28&29-Dec-2010
1.4	03-Feb-2011	Keith Chadwick	Update following work on 25 to 27-Jan-2011 and 01&03-Feb-2011
1.5	17-Feb-2011	Keith Chadwick	Update following work on 08-Feb-2011 and 17-Feb-2011
1.6	01-Mar-2011	Keith Chadwick	Update following installation of Cisco Nexus 2248 Fabric Extender
1.7	02-Mar-2011	Keith Chadwick	Update following work on 01-Mar-2011
1.8	21-Apr-2011	Keith Chadwick	Add steps for the 152/153 subnet transition.

Table of Contents

Introduction 1
Core Services Rack Configuration 3
Gatekeeper Rack Configuration 4
OSG Services Rack Configuration 5
FermiGrid-HA2 Requirements 6
FermiGrid-HA2 Network Design 6
FermiGrid-HA2 Rack Configuration 8
FermiGrid-HA2 Transition Plan 9
Appendix 1 – FermiGrid-HA2 Rack #1 FCC2 Location 14
Appendix 2 – FermiGrid-HA2 Rack #2 GCC-B Location 15

Introduction

FermiGrid has previously architected the high availability FermiGrid-HA (reference CD-DocDB Document # 3579). As part of FermiGrid-HA planning, the potential requirements of a future enhancement to extend the FermiGrid-HA design and deployment to support Redundant and/or Resilient Services (FermiGrid-RS) were also considered. Due to the previously demonstrated reliability of the computer rooms in the Feynman Computing Center following the installation of the building wide UPS and generator, the FermiGrid-RS extensions were not viewed as being needed until the CY2011-2012 timeframe. The principle issues with the computer room had been the availability of cooling capacity, and due to the relatively low power required to operate the systems that provided the FermiGrid-HA services, FermiGrid personnel had been allowed to operate these systems during FCC cooling outages.

Unfortunately, there were two significant power interruptions in February 2010 and a second set of two significant power interruptions in November 2010 for the computer rooms in the Feynman Computing Center. The exact root cause is not pertinent to this document, but the incidents in February 2010 raised the need for continued operation of the set of FermiGrid-HA services, and thus the planning process was initiated for FermiGrid-RS. In addition, the name of FermiGrid-RS was changed to FermiGrid-HA Phase 2 (HA2).

The figure below shows the FermiGrid-HA "Service Communication Diagram":

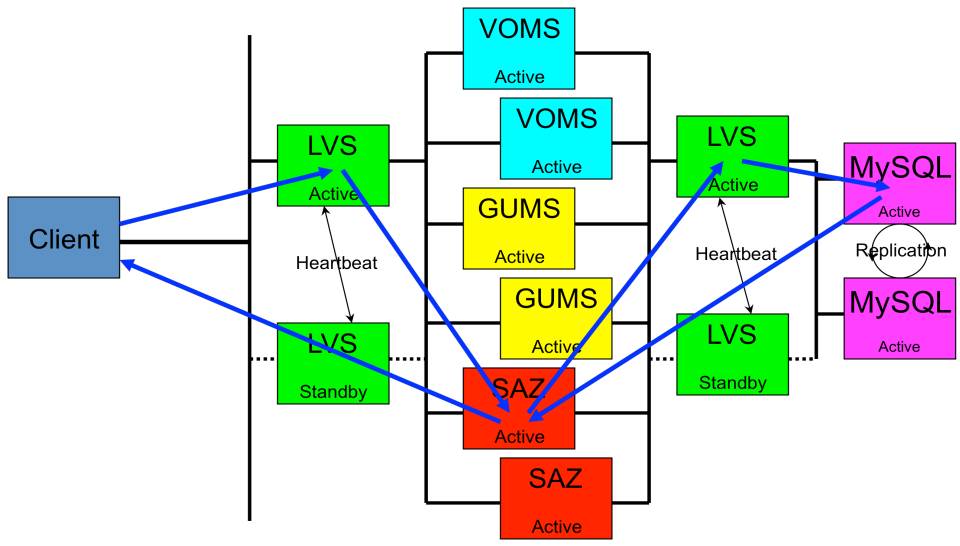


Figure – FermiGrid-HA Service Communication Diagram

Each of the services identified above were deployed in Xen virtual machines under Scientific Linux as show in the figure below:

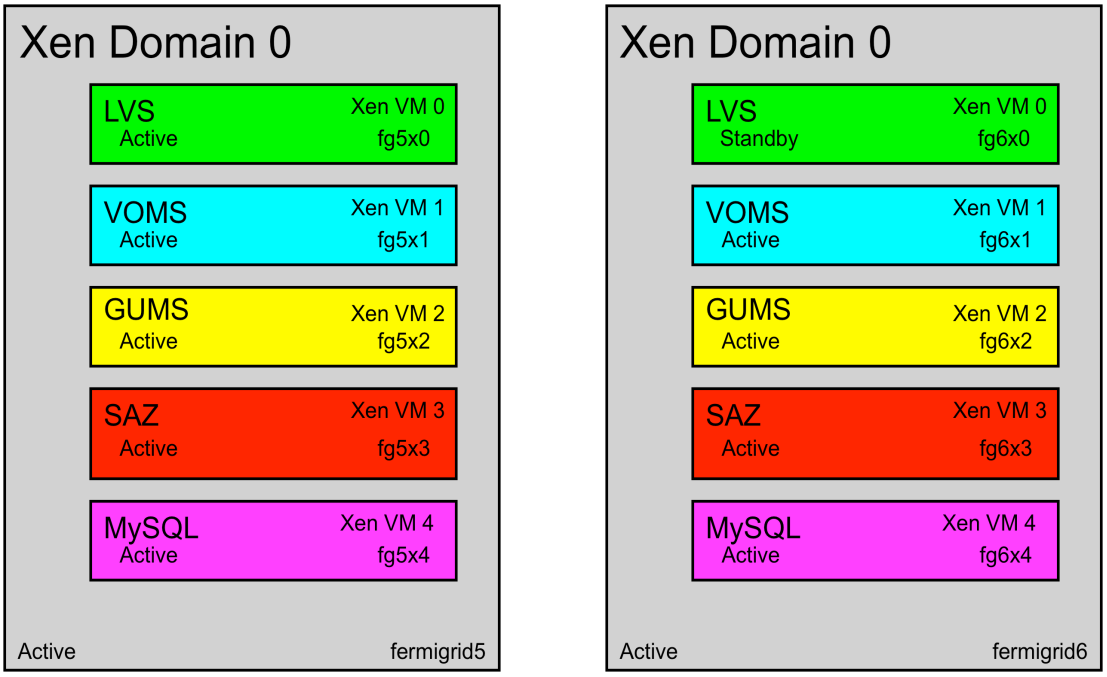


Figure – FermiGrid-HA Virtual Machines

Beyond the FermiGrid-HA / FermiGrid Core Services systems above, FermiGrid operates three classes of systems:

Core Services:
fermigrid[0-6] and associated virtual machines fg0x[0-7] and fg[1-6]x[0-5]

Gatekeeper Services:
fcdfsrv[0-5] and associated virtual machines fcdf[0-5]x[0-5]
d0osgsrv[1,2] and associated virtual machines d0osg[1,2]x[0-4]
fnpcsrv[3-9] and associated virtual machines

OSG Services:
Gratia[10-13] and associated virtual machines gr[10-13]x[0-5]
Ress[01,02] and associated virtual machines ress[1,2]x[0-3]

The following three sections show the current physical machine layout of these three classes of systems across the three FermiGrid racks:

Core Services Rack Configuration

Core Services Rack Front	Rack "U"	Core Services Rack Rear
	42	
	41	
	40	Cyclades AlterPath Console Server 16
	39	APC Transfer Switch
	38	
	37	
	36	
	35	
	34	
	33	
	32	
	31	
	30	
fermigrid0	29	fermigrid0
	28	
	27	
	26	Cyclades PM10-L30A
	25	Cyclades PM10-L30A
Display / Keyboard / Mouse	24	Display / Keyboard / Mouse
Omniview PS3 16 port KVM	23	Omniview PS3 16 port KVM
	22	Linksys SR2016 Private LAN Switch
	21	
	20	
fermigrid6	19	fermigrid6
	18	
fermigrid5	17	fermigrid5
	16	
fermigrid4	15	fermigrid4
	14	
fermigrid3	13	fermigrid3
	12	
fermigrid2	11	fermigrid2
	10	
fermigrid1	9	fermigrid1
	8	
	7	
	6	
	5	
	4	
	3	
	2	Cyclades PM10-L30A
	1	Cyclades PM10-L30A

Figure – FermiGrid Core Services Rack

Gatekeeper Rack Configuration

Gatekeeper Rack Front	Rack "U"	Gatekeeper Rack Rear
	42	
	41	Public LAN Switch Catalyst 2960G
	40	Cyclades AlterPath Console Server 32
	39	
d0osgsrv2	38	d0osgsrv2
	37	
	36	
	35	
d0osgsrv1	34	d0osgsrv1
	33	
fcdfsrv4	32	fcdfsrv4
	31	
fcdfsrv5	30	fcdfsrv5
	29	
fcdfsrv3	28	fcdfsrv3
	27	
	26	Cyclades PM10-L30A
	25	Cyclades PM10-L30A
Display / Keyboard / Mouse	24	Display / Keyboard / Mouse
Omniview PS3 16 port KVM	23	Omniview PS3 16 port KVM
	22	
	21	
	20	
	19	
	18	
	17	
fcdfsrv2	16	fcdfsrv2
	15	
fcdfsrv1	14	fcdfsrv1
	13	
fcdfsrv0	12	fcdfsrv0
fnpcsrv3	11	fnpcsrv3
fnpcsrv4	10	fnpcsrv4
fnpcsrv5	9	fnpcsrv5
	8	
	7	
fnpcsrv8	6	fnpcsrv8
fnpcsrv9	5	fnpcsrv9
	4	
	3	Cyclades PM10-L30A
	2	
	1	

Figure – FermiGrid Gatekeeper Rack Layout

OSG Services Rack Configuration

OSG Services Rack Front	Rack "U"	OSG Services Rack Rear
	42	
	41	Cyclades AlterPath Console Server 16
	40	
	39	
ress01	38	ress01
	37	
ress02	36	ress02
	35	
	34	
	33	
gratia12	32	gratia12
	31	
	30	
gratia10	29	gratia10
	28	
	27	
gratia11	26	gratia11
	25	
Display / Keyboard / Mouse	24	Display / Keyboard / Mouse
Omniview PS3 16 port KVM	23	Omniview PS3 16 port KVM
	22	
	21	
	20	
gratia13	19	gratia13
	18	
	17	
	16	
	15	
	14	
	13	
	12	
	11	
	10	
	9	
	8	
	7	
	6	
	5	
	4	
	3	
	2	
	1	

PM20-L30A

PM20-L30A

Figure – OSG Services Rack Layout

FermiGrid-HA2 Requirements

The requirements of FermiGrid-HA2 are very simple:

- Continued operation of the set of FermiGrid-HA services if either building hosting the FermiGrid-HA services is down within the limits of the service operation.

This simple requirement places significant constraints on the network design and service deployment.

FermiGrid-HA2 Network Design

The figure below shows the proposed FermiGrid-HA2 network design:

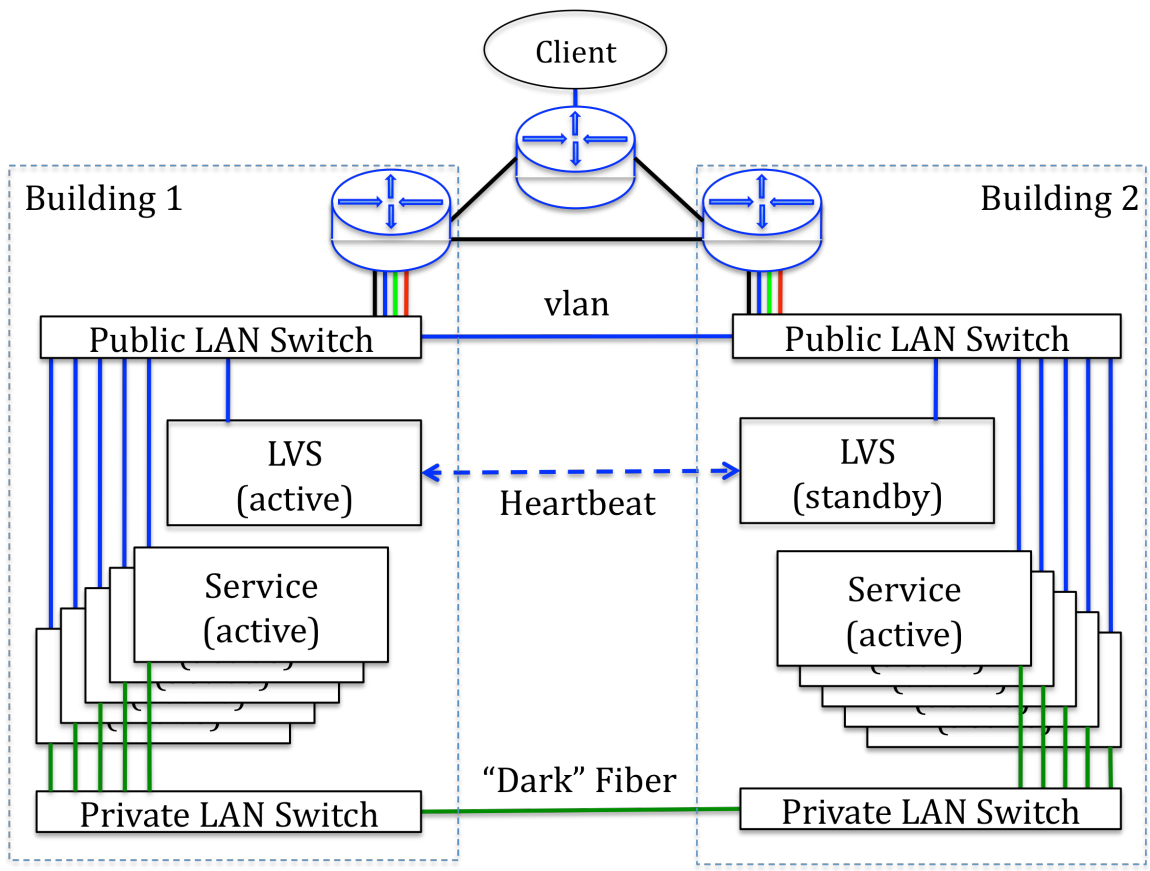


Figure – FermiGrid-HA2 Network Design

This design uses both the "public" Fermilab network (routers and switches attached to various workgroups) as well as an extended private LAN (to allow for secure replication of various "secrets" between the services). An example of such a "secret" is the replication of the drbd blocks that hold the private keys of the credentials that users have chosen to place in the FermiGrid MyProxy repository.

The "Public LAN" connections shall be the standard data communications switch and workgroup ports, and the services that require heartbeat (such as LVS) shall exchange "heartbeat" information across the public LAN interfaces.

The "Public LAN" switch shall offer multiple VLANs via VLAN trunking on the uplink and shall have a minimum of a 10 GB uplink to the Fermilab network infrastructure. The list of VLANs needed on this switch includes:

- The "Grid Services" 107 subnet.
- The CDF Grid clusters 240 subnet(s).
- The D0 Grid clusters 216 subnet(s).
- The GP Grid cluster 167 subnet(s).

Based on recommended by Data Communications, the options for the "Public LAN" switch include the Cisco Nexus 2248 or 2232 switches. While the racks remain in the FCC1 area, the existing individual network connections shall be run via a 24 port network patch panel.

The "Private LAN" connections shall be two Linksys SR2024 switches (unmanaged 24 port gigabit Ethernet switches with optional dual miniGBIC optical interfaces), connected via miniGBIC interfaces over "dark" fiber between building 1 (the Feynman Computing Center) and building 2 (Grid Computing Center).

FermiGrid-HA2 Rack Configuration

Based on the above FermiGrid-HA2 network design, the current FermiGrid systems can be allocated into two “identically” configured racks as shown in the Figure below:

FermiGrid-HA2 Rack Front	Rack "U"	FermiGrid-HA2 Rack Rear	
	42	Cisco Nexxus / 2960G Public LAN Switch	Vertical PDU - APC 7941 208VAC
blank - 2U	41	Cyclades AlterPath Console Server 16	
fnpcsrv8 / blank - 1U	40	fnpcsrv8 / blank - 1U	
fnpcsrv5 / fnpcsrv9	39	fnpcsrv5 / fnpcsrv9	
fnpcsrv3 / fnpcsrv4	38	fnpcsrv3 / fnpcsrv4	
blank - 1U	37	blank - 1U	
	36		
d0osgsrv1 / d0osgsrv2	35	d0osgsrv1 / d0osgsrv2	
	34		
blank - 2U	33	blank - 2U	
	32		
fcdfsrv3 / fcdfsrv4	31	fcdfsrv3 / fcdfsrv4	
	30		
fcdfsrv1 / fcdfsrv2	29	fcdfsrv1 / fcdfsrv2	
	28		
fcdfsrv0 / fcdfsrv5	27	fcdfsrv0 / fcdfsrv5	
	26	Cyclades PM10-L30A 120VAC	
blank - 2U	25	Cyclades PM10-L30A 120VAC	
Display / Keyboard / Mouse	24	Display / Keyboard / Mouse	
Raritan MasterConsole MCCAT116 KVM	23	Omniview PS3 16 port KVM	
blank - 1U	22	Linksys SR2024 Private LAN Switch	
Slave KDC - Sun Netra X1	21	APC Transfer Switch	
blank - 1U	20	blank - 1U	
	19		
	18		
gratia12 (FCC) / gratia13(GCC)	17	gratia12 (FCC) / gratia13(GCC)	
	16		
	15		
gratia10 (FCC) / gratia11 (GCC)	14	gratia10 (FCC) / gratia11 (GCC)	
blank - 1U	13	blank - 1U	
	12		
ress01 / ress02	11	ress01 / ress02	
	10		
fermigrid5 / fermigrid6	9	fermigrid5 / fermigrid6	
	8		
fermigrid2 / fermigrid3	7	fermigrid2 / fermigrid3	
	6		
fermigrid1 / fermigrid4	5	fermigrid1 / fermigrid4	
	4		
fermigrid0 / fermigrid7	3	fermigrid0 / fermigrid7	
	2	Cyclades PM10-L30A 120VAC	
blank - 2U	1	Cyclades PM10-L30A 120VAC	

Figure – FermiGrid-HA2 Rack Layout

FermiGrid-HA2 Transition Plan

The transition between the current FermiGrid Services deployment across the three current racks to the FermiGrid-HA2 deployment across the two racks shall consist of the following steps:

1. Identify and acquire the necessary equipment, power cables, network and serial console cables to implement the transition **(In process)**.

2. Identify the rack locations **(Done - 17-Dec-2010)**:

Rack	Proposed Location
FermiGrid-HA2 Rack 1	FCC 2 Computer Room – FCC-2-1464
FermiGrid-HA2 Rack 2	GCC Network Room B – GCC-B-3056

3. Identify power sources and circuits in the (proposed) rack locations **(Done)**:

FermiGrid-HA2 Rack 1	Power Source	Circuit
APC Transfer Switch	FCC UPS-1	UPS-PP2-23 CIR 41
APC Transfer Switch	FCC UPS-1	UPS-PP2-22 CIR 41
APC AP7941 208V	FCC UPS-1	UPS-PP2-23 CIR 21,23
APC AP7941 208V	FCC UPS-1	UPS-PP2-22 CIR 21,23

FermiGrid-HA2 Rack 2	Power Source	Circuit
Cyclades PM10-L30A #1	GCC UPS-2	PP-GCC-21A-4 Cir 12
Cyclades PM10-L30A #2	GCC UPS-2	PP-GCC-21A-4 Cir 14
Cyclades PM10-L30A #3	GCC UPS-2	PP-GCC-21A-4 Cir 16
Cyclades PM10-L30A #4	GCC UPS-2	PP-GCC-21A-4 Cir 18

Note that an "alternate" set of circuits exist in close proximity to relay rack GCC-B-3056, namely PP-GCC-21A-18, Cir 30, 28, 26, 24, 22. If possible, the even (or odd) numbered Cyclades PM10-L30A's in FermiGrid-HA2 Rack 2 might be plugged into two of these circuits to allow panel level redundancy for the FermiGrid-HA rack located in GCC-B.

4. Identify the racks that will be used to perform the migration – The FermiGrid Core Services rack and the OSG Services rack appear to be the best candidates. For the purposes of the remainder of this document, the following rack identification shall be used **(Done)**:

Current Rack Number	Current Rack Identification	Future Rack Identification	Future Rack Number
FCC-1-3465	FermiGrid Core Services	FermiGrid-HA2 Rack 2	FCC-2-1464

FCC-1-3467	FermiGrid Gatekeepers	none	
FCC-1-3470	OSG Services	FermiGrid-HA2 Rack 1	GCC-B-3056

5. Install the "base" infrastructure equipment in the FermiGrid Core Services and OSG Services racks (Cyclades ACS-16, Cyclades PM10-L30A, APC AP7941 208V PDU, Keyboard-Mouse-Monitor, KVM switches, APC power transfer switches, Linksys SR2024 "private LAN" network switches, public LAN network switches) **(Done - 28&29-Dec-2010)**.

6. Connect the Linksys SR2024 switches between the two racks using a Category 5e UTP cable **(Done – 28-Dec-2010)**.

7. Begin the system-by-system migration of the selected FermiGrid Core Services systems from the current FermiGrid Core Services rack to the current OSG Services Rack as shown in the FermiGrid-HA2 rack layout above. Verify full service functionality after the selected systems have been relocated **(Done 25-Jan-2011)**.

System	Initial Location	Target Location
fermigrid0	Core Services – 29/30	OSG Services – 3/4
fermigrid1	Core Services – 9/10	OSG Services – 5/6
fermigrid2	Core Services – 11/12	OSG Services – 7/8
fermigrid5	Core Services – 17/18	OSG Services – 9/10

8. Reorganize the remaining FermiGrid Core Service systems in the current FermiGrid Core Services rack as shown in the FermiGrid-HA2 rack layout above **(Done 27-Jan-2011)**.

System	Initial Location	Target Location
fermigrid8	WH8SE Fgtest Rack	Core Services – 3/4
fermigrid4	Core Services – 15/16	Core Services – 5/6
fermigrid3	Core Services – 13/14	Core Services – 7/8
fermigrid6	Core Services – 19/20	Core Services – 9/10

9. Reorganize the Fermilab and OSG Gratia collector systems in the current OSG Services rack as shown in the FermiGrid-HA2 rack layout above **(Done 01-Feb-2011)**.

System	Initial Location	Target Location
gratia10	OSG Services – 29/30/31	OSG Services – 12/13/14
gratia12	OSG Services – 32/33/34	OSG Services – 15/16/17

10. Reorganize the Fermilab and OSG Gratia reporter systems in the current FermiGrid Core Services rack as shown in the FermiGrid-HA2 rack layout above **(Done 01-Feb-2011)**.

System	Initial Location	Target Location
gratia11	OSG Services – 26/27/28	Core Services – 12/13/14
gratia13	OSG Services – 19/20/21	Core Services – 15/16/17

11. Reorganize the OSG Resource Selection Services (ReSS) systems across the OSG Services and FermiGrid Core Services racks as shown above **(Done 25&27-Jan-2011)**.

System	Initial Location	Target Location
ress01	OSG Services – 38/39	OSG Services – 19/20
ress02	OSG Services – 36/37	Core Services – 19/20

12. At this point, the two racks presently known as FermiGrid Core Services and OSG Services should be fully functioning copies of one another (within the service deployment limitations). A set of tests shall be performed to demonstrate the service failover between the two racks – power down individual systems, power down the one of the two racks, etc.. If any services fail to failover as expected, the cause shall be identified and rectified prior to proceeding to the next step in the FermiGrid-HA2 transition **(Done)**.

13. Once the service failover tests have been successfully completed, the racks formerly known as the FermiGrid Core Services and OSG Services racks shall be renamed to FermiGrid-HA2 rack 2 and FermiGrid-HA2 rack 1 respectively **(Done)**.

14. At this point, the remainder of the FermiGrid-HA2 repackaging shall be performed – the individual gatekeeper systems in the FermiGrid Gatekeepers rack shall be moved one-by-one to the allocated location in the FermiGrid-HA2 racks. Systems shall be moved one system at a time, and full system functionality shall be verified prior to the movement of any additional systems. Wherever possible, the relocation of the Gatekeeper systems shall occur during previously scheduled and agreed upon downtimes **(Done. D0 and CDF - 08-Feb-2011, GP - 17-Feb-2011)**.

System	Initial Location	Target Location
fnpcsrv3	Gatekeeper – 11	FermiGrid-HA2 Rack 1 –
fnpcsrv4	Gatekeeper – 10	FermiGrid-HA2 Rack 2 –
fnpcsrv5	Gatekeeper – 9	FermiGrid-HA2 Rack 1 –
???	???	FermiGrid-HA2 Rack 2 –
fnpcsrv8	Gatekeeper – 6	FermiGrid-HA2 Rack 1 –

fnpcsrv9	Gatekeeper – 5	FermiGrid-HA2 Rack 2 –
fcdfsrv0	Gatekeeper – 12/13	FermiGrid-HA2 Rack 1 – 27/28
fcdfsrv1	Gatekeeper – 14/15	FermiGrid-HA2 Rack 2 – 27/28
fcdfsrv2	Gatekeeper – 16/17	FermiGrid-HA2 Rack 1 – 29/30
fcdfsrv3	Gatekeeper – 28/29	FermiGrid-HA2 Rack 2 – 29/30
fcdfsrv4	Gatekeeper – 32/33	FermiGrid-HA2 Rack 1 – 31/32
fcdfsrv5	Gatekeeper – 30/31	FermiGrid-HA2 Rack 2 – 31/32
d0osgsrv1	Gatekeeper – 34/35	FermiGrid-HA2 Rack 1 – 35/36
d0osgsrv2	Gatekeeper – 38/39	FermiGrid-HA2 Rack 2 – 35/36

15. Once all systems have been removed from the FermiGrid Gatekeepers rack, FermiGrid personnel shall arrange for the removal of the FermiGrid Gatekeepers rack **(Done)**.

16. FermiGrid personnel shall install Raritan MasterConsole MCCAT 216 KVM Switches together with the necessary Cat5e cabling to replace the Belkin Omniview Pro3 KVM Switches and custom Omniview cables **(Done)**.

17. Site networking personnel shall install the Cisco Nexus 2224 Fabric Extenders in the top of the two FermiGrid-HA2 racks **("Loaner" Cisco 2248's were installed in the FermiGrid-HA2 racks on 28-Feb-2011, the production Cisco 2224's were installed on 05-Apr-2011 and 06-Apr-2011 – Done.)**.

18. Site networking personnel shall configure subnet 107 on the Cisco Nexus 2224 Fabric Extenders in the top of the two FermiGrid-HA2 racks **(Done – 01-Mar-2011)**.

19. FermiGrid personnel shall recable the subnet 107 systems in the two FermiGrid-HA2 racks to the Cisco Nexus 2224 (2248) Fabric Extenders **(Done – 01-Mar-2011)**.

20. Site networking personnel shall configure the /23 subnet [152/153] and configure on the Cisco Nexus 2224 Fabric Extenders in the top of the two FermiGrid-HA2 racks **(TBD)**.

21. FermiGrid personnel shall re-address the subnet 107 FermiGrid, ReSS and Gratia systems in the two FermiGrid-HA2 racks to addresses in the 152/153 subnet **(TBD)**.

22. FermiGrid personnel shall re-address and recable the subnet 240 (CDF) systems in the two FermiGrid-HA2 racks to the Cisco Nexus 2224 Fabric Extenders **(TBD)**.

23. FermiGrid personnel shall re-address and recable the subnet 161 (D0) systems in the two FermiGrid-HA2 racks to the Cisco Nexus 2224 Fabric Extenders **(TBD)**.

24. FermiGrid personnel shall re-address and recable the subnet 166/167 (GP) systems in the two FermiGrid-HA2 racks to the Cisco Nexus 2224 Fabric Extenders **(TBD)**.

25. The re-racked systems shall be operated for at least 7 days in the "final" configuration prior to proceeding to the next step in the transition plan **(TBD)**.

26. Relocate "FermiGrid-HA2 Rack 1" from the current location to the agreed FCC2 location identified in Transition Step 2 above. The height of the Dell racks (with the feet down) is 79 inches (the feet must be raised to move the racks – so the rolling height will be ~78.5 inches), the limiting doorway opening to exit the computer room is ~83 inches, and the FCC elevator opening is 84 inches. Due to weight restrictions, the systems in the upper half of the rack will need to be physically removed from the rack prior to movement and reinstalled in the rack when the rack has been moved to it's new location **(TBD)**.

27. Re-establish private LAN network connectivity between FermiGrid-HA2 Rack 1 and FermiGrid-HA2 Rack 2 via a "dark" (dedicated) single mode fiber pair with quantity LC connectors at each end **(TBD)**.

28. Re-establish public LAN network connectivity between FermiGrid-HA2 Rack 1 and the site network **(TBD)**.

29. The racks shall be operated for at least 7 days in these locations prior to proceeding to the next step in the transition plan **(TBD)**.

30. Relocate "FermiGrid-HA2 Rack 2" from the current location to the agreed GCC location identified in Transition Step 2 above. Due to weight restrictions, the systems in the upper half of the rack will need to be physically removed from the rack prior to movement and reinstalled in the rack when the rack has been moved to it's new location **(TBD)**.

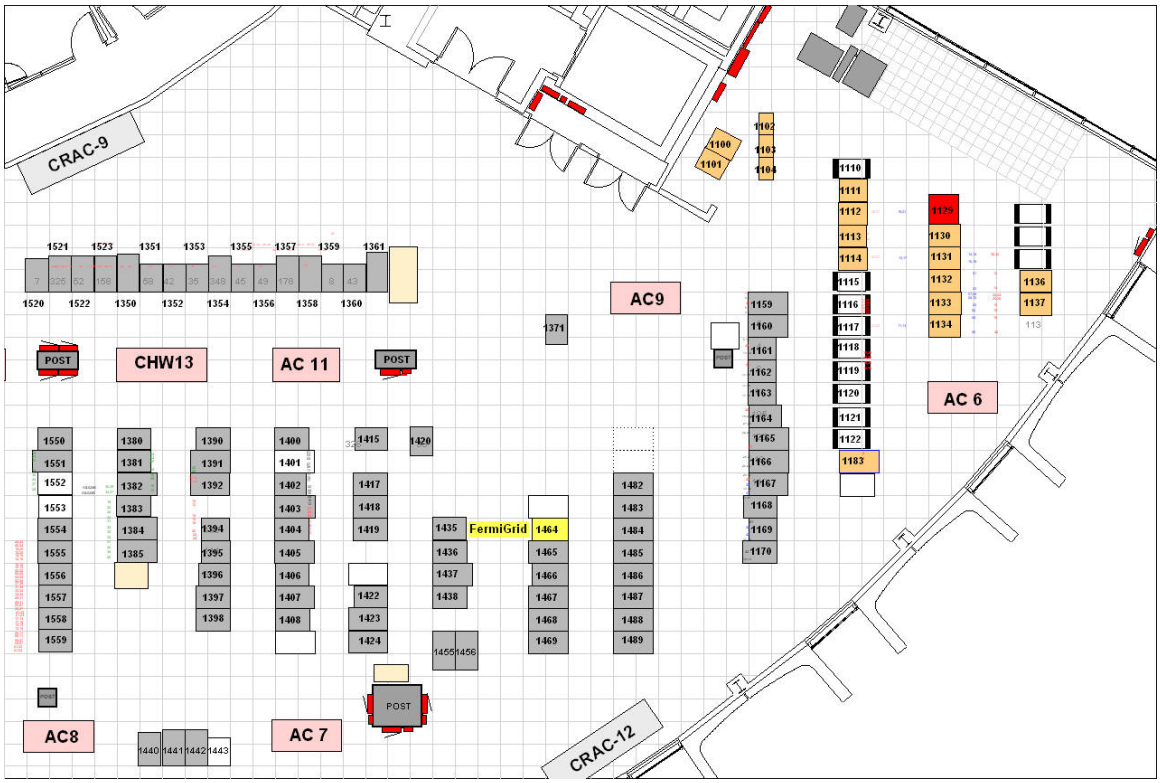
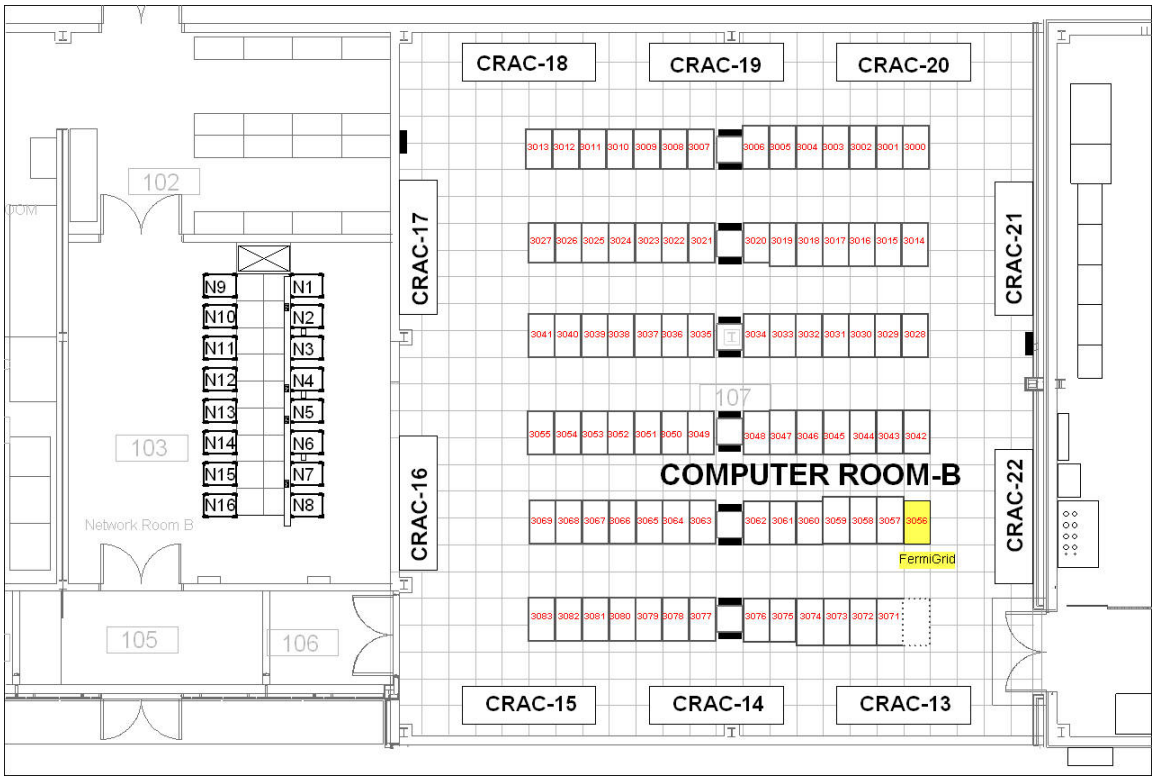
31. Re-establish private LAN network connectivity between FermiGrid-HA2 Rack 2 and FermiGrid-HA2 Rack 1 **(TBD)**.

32. Re-establish public LAN network connectivity between FermiGrid-HA2 Rack 2 and the site network **(TBD)**.

33. The racks shall be operated for at least 7 days in these locations prior to proceeding to the next step in the transition plan.

34. Transition Complete

Appendix 1 – FermiGrid-HA2 Rack #1 FCC2 Location



Appendix 2 – FermiGrid-HA2 Rack #2 GCC-B Location