

# FermiGrid Scalability and Reliability Improvements

K. Chadwick, F. Lowe, N. Sharma, S. Timm, D. R. Yocum  
Grid And Cloud Computing Department  
Fermilab  
Spring 2011 HEPiX

Work supported by the U.S. Department of Energy under contract No. DE-AC02-07CH11359

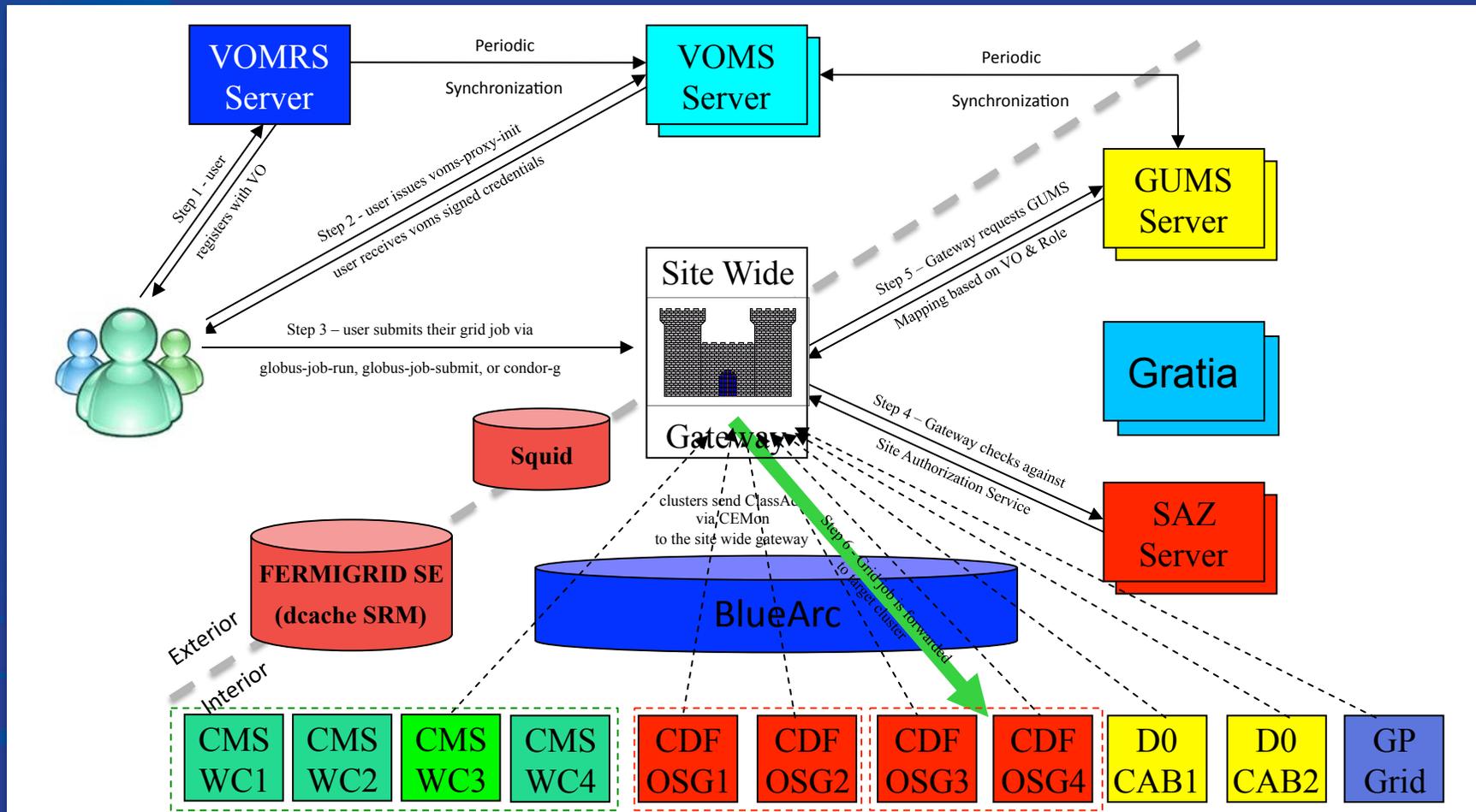
# What is FermiGrid?

- A Meta-facility that provides grid infrastructure for scientific computing at Fermilab,
- Provides highly-available centralized authorization and authentication services,
- Provides site gateway for Globus job submission,
- Coordinates interoperability among stakeholders,
- Provides grid-enabled mass storage services,
- More than 210,000,000 CPU-hours delivered since Gratia accounting started, most in OSG,
- Currently have over 22,500 batch slots, 10 compute elements.

# Key Policies

- Early management directions:
  - Special security enclave for nodes where grid jobs run,
  - There WILL be a single site gateway for OSG jobs,
  - All major users and clusters MUST interoperate,
  - There WILL be a unified DN/VO/FQAN to Unix UID mapping,
  - There WILL be a central banning server,
  - Pilot jobs MUST use gLexec to authenticate 3rd-party jobs,
  - We WILL have our own Certificate Authority (the Fermilab Kerberos Certificate Authority),
  - We WILL keep more than one batch system in production.

# FermiGrid Architecture



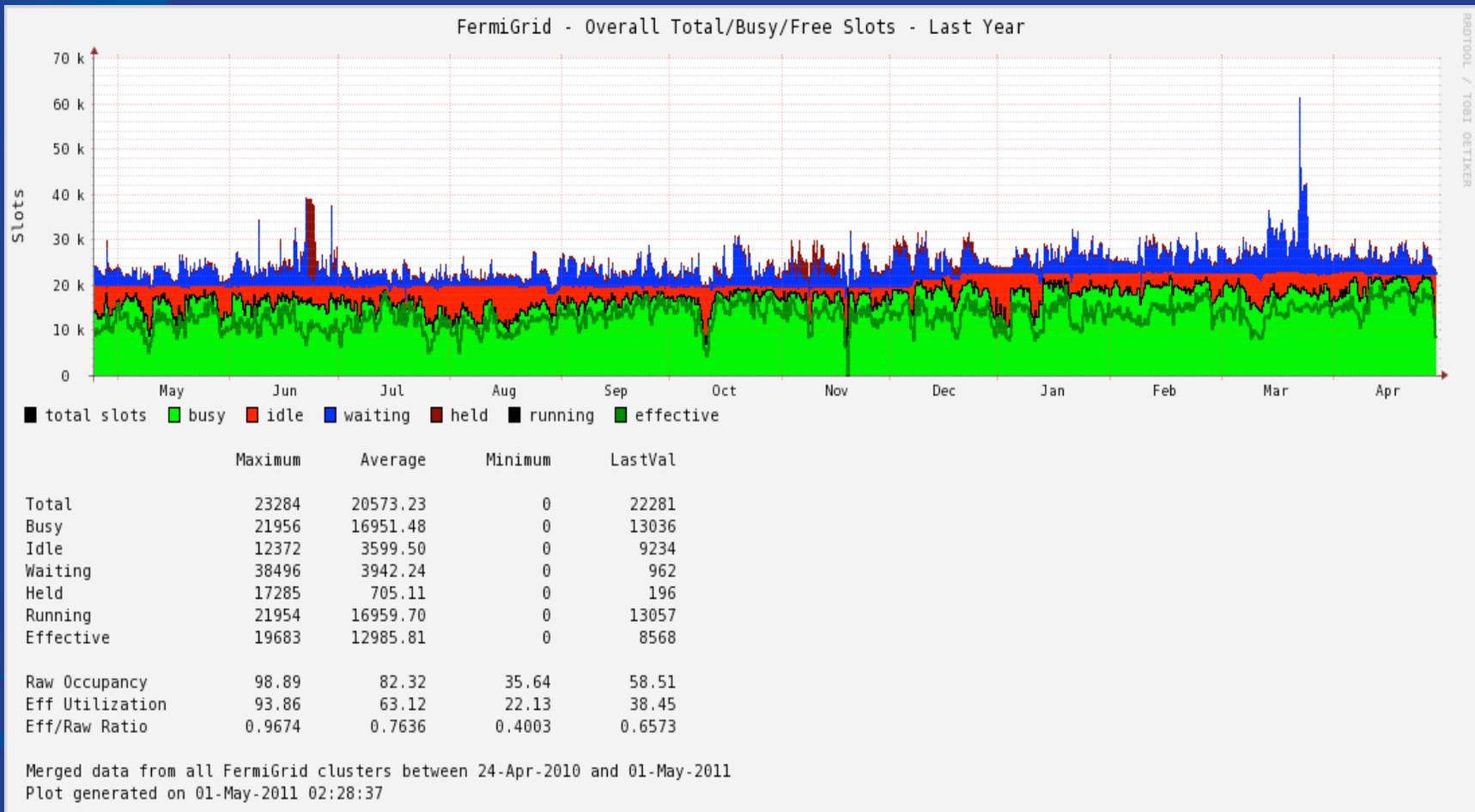
# Condor Classads

- Start with Glue Schema 1.3 LDIF information for each Computing Element and associated Storage Elements,
- Clusters divided up into subclusters based on hardware type,
- gLite CEMon plugin makes one classad for each unique combination of {VO, subcluster, Storage Area},
- Example: 60 VO's x 4 subclusters x 2 Storage Areas = 480 classads in just 1 cluster,
- FermiGrid presently has 7,174 classads,
- OSG presently has 21,981 classads,
- Classads transmitted by CEMon to Resource Selection System,
- Raw LDIF transmitted by CEMon to OSG BDII.

# Recent use case - OS Version

- Needed to upgrade from SLF4 to SLF5, one subcluster at a time,
- Each subcluster advertises:
  - Current OS Version,
  - Extra RSL fields to make sure you run on those nodes when your job matches to that subcluster.
- Users submit job requesting OS (4.7 or 5.3),
- Site gateway finds a matching subcluster and automatically appends right RSL, and forwards job.
- Can also select on memory, disk, batch system.

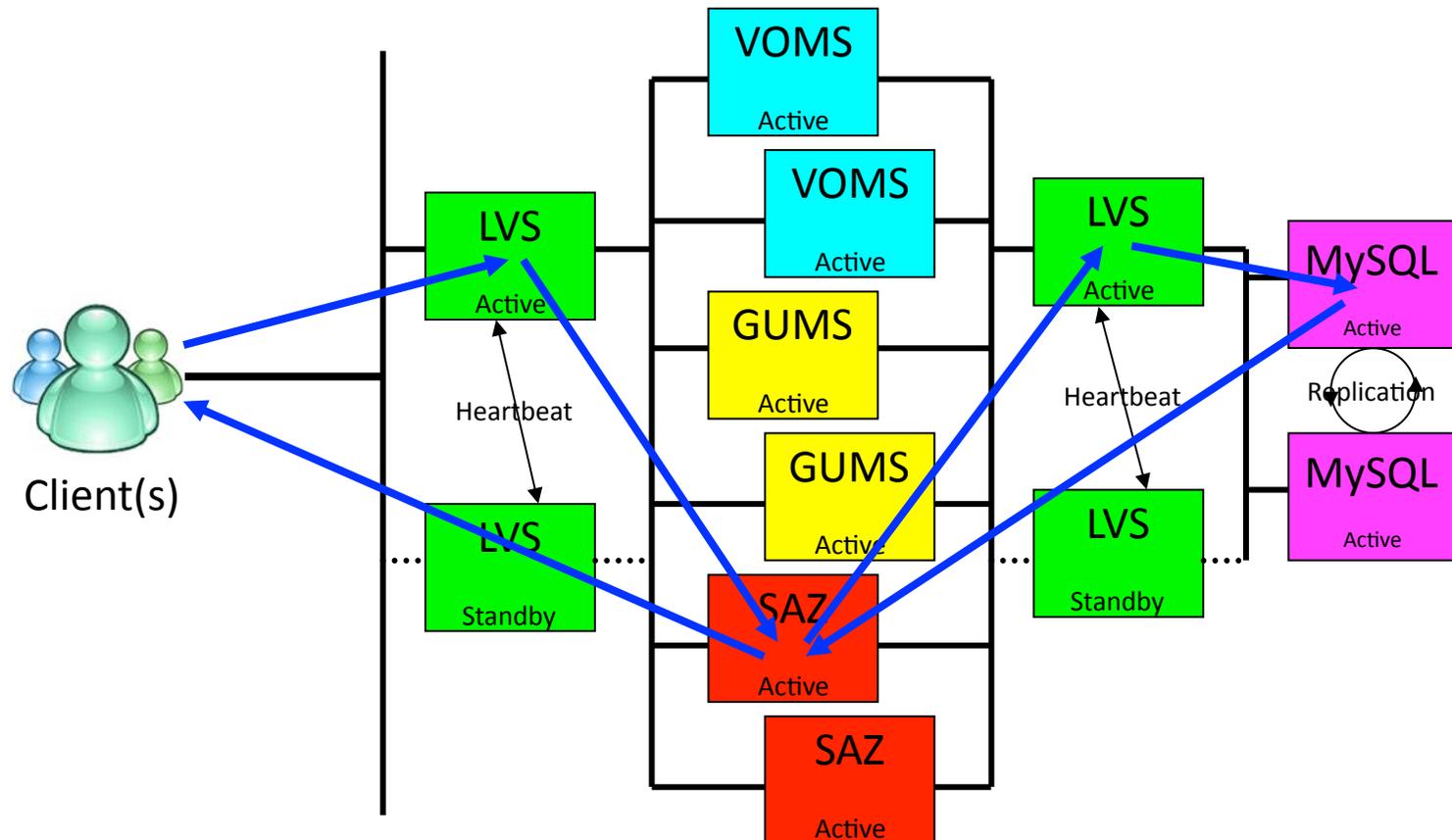
# Overall Occupancy & Utilization



# Batch Systems, Occupancy & Utilization

Cluster	Cluster Batch System	Current Cluster Size (Slots)	Average Cluster Occupancy (%)	Average Cluster Utilization (%)
CDF	Condor	5,600	89.3	63.4
CMS	Condor	7,485	88.8	77.3
D0	PBS	6,916	73.4	47.1
GP	Condor	3,284	68.7	62.3
Overall	----	23,285	82.3	63.1

# FermiGrid HA Services - 1



# FermiGrid-HA Services - 2

## Xen Domain 0

LVS                      Xen VM 0  
Active                      fg5x0

VOMS                      Xen VM 1  
Active                      fg5x1

GUMS                      Xen VM 2  
Active                      fg5x2

SAZ                      Xen VM 3  
Active                      fg5x3

MySQL                      Xen VM 4  
Active                      fg5x4

Active

fermigrid5

## Xen Domain 0

LVS                      Xen VM 0  
Standby                      fg6x0

VOMS                      Xen VM 1  
Active                      fg6x1

GUMS                      Xen VM 2  
Active                      fg6x2

SAZ                      Xen VM 3  
Active                      fg6x3

MySQL                      Xen VM 4  
Active                      fg6x4

Active

fermigrid6

# Measured FermiGrid Service Availability for the Past Year\*

Service	Availability	Downtime
VOMS-HA	100%	0m
GUMS-HA	100%	0m
SAZ-HA (gatekeeper)	100%	0m
Squid-HA	100%	0m
MyProxy-HA	99.943%	299.0m
ReSS-HA	99.959%	215.3m
Gratia-HP	99.955%	233.3m
Database-HA	99.963%	192.6m

\* = Excluding building or network failures

# SAZ – Central Banning Service

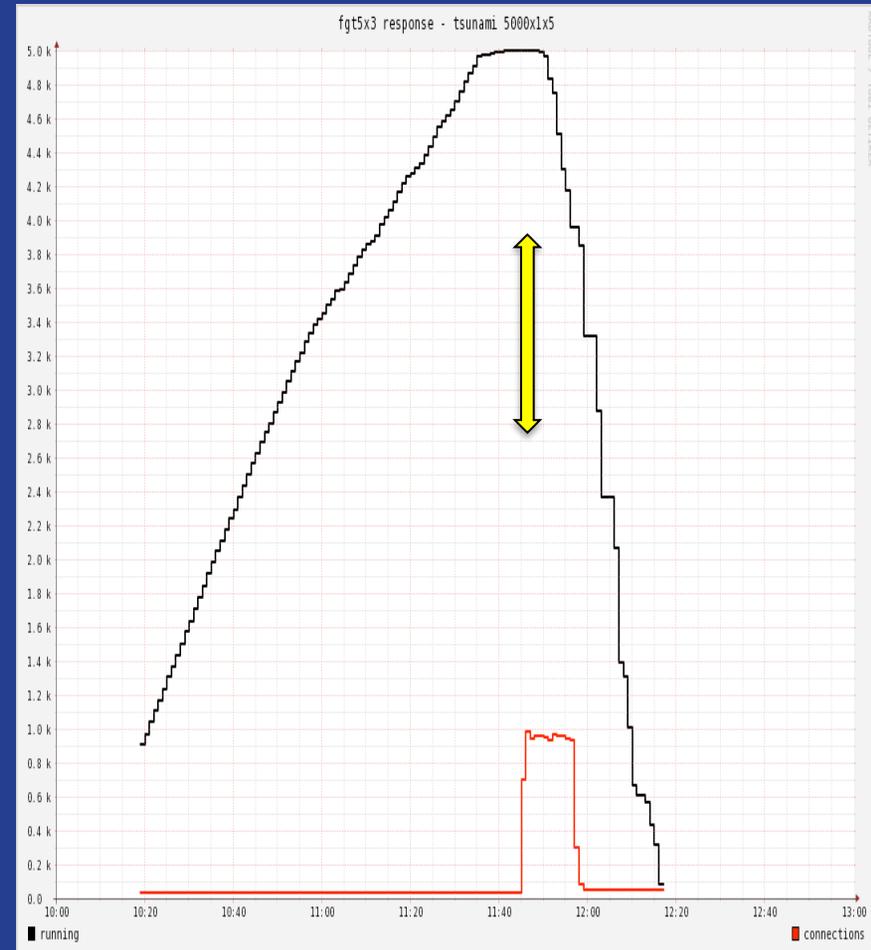
- Site AuthoriZation service, developed at Fermilab,
- Allows us to ban any DN, VO, FQAN, or CA,
- Don't have to wait for CA to revoke the certificate or VO to remove from membership,
- Available as Pacman package,
- Details at <http://saz.fnal.gov>
- With glideins, every worker node can be a client simultaneously (have seen 5000+ active clients),
- Significant work has been done to make it more resilient and scalable,
- Code also contains support for new XACML-based authorization protocol.

# Why a new SAZ Server?

- Previous SAZ server (V2.0.1b) had shown itself extremely vulnerable to user generated authorization "tsunamis":
  - Very short duration jobs
  - User issues condor\_rm on a large (>1000) glidein.
- This was fixed in the new SAZ Server (V2.7.2) using tomcat and pools of execution and hibernate threads.
- Various other bugs were found and fixed in the current SAZ server and sazclient.
- Added support for the XACML protocol (used by Globus).
  - NOT transitioning to using XACML (yet).
- New SAZ Servers (V2.7.2) have been in operation since May 2010.
  - Have repeatedly demonstrated in production that they are robust against authorization "tsunamis".

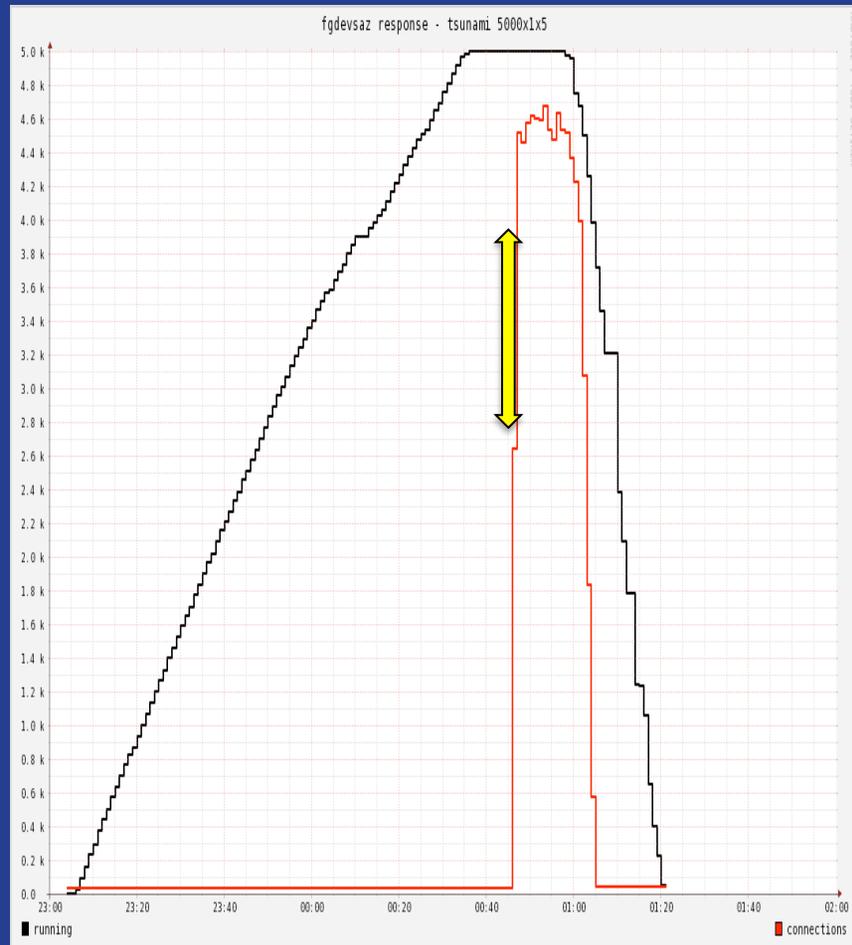
# SAZ Server V2.0.1b Performance

- Using development server,
- Black = number of condor jobs,
- Red = number of SAZ network connections.
- Trigger @ 11:45:12
- Failures start @ 11:45:19
- 25,000 Authorizations
- 14,183 Success
- 10,817 Failures
- Complete @ 11:58:28
- Elapsed time 13m 16s



# SAZ Server V2.7.2 Performance

- Using development server
- Black = number of condor jobs,
- Red = number of SAZ network connections.
- Trigger @ 00:46:20,
- 25,000 Authorizations,
- 25,000 Success,
- 0 Failures,
- Complete @ 01:05:03,
- Elapsed time = 18m 43s,
- 22.26 Authorizations/sec.



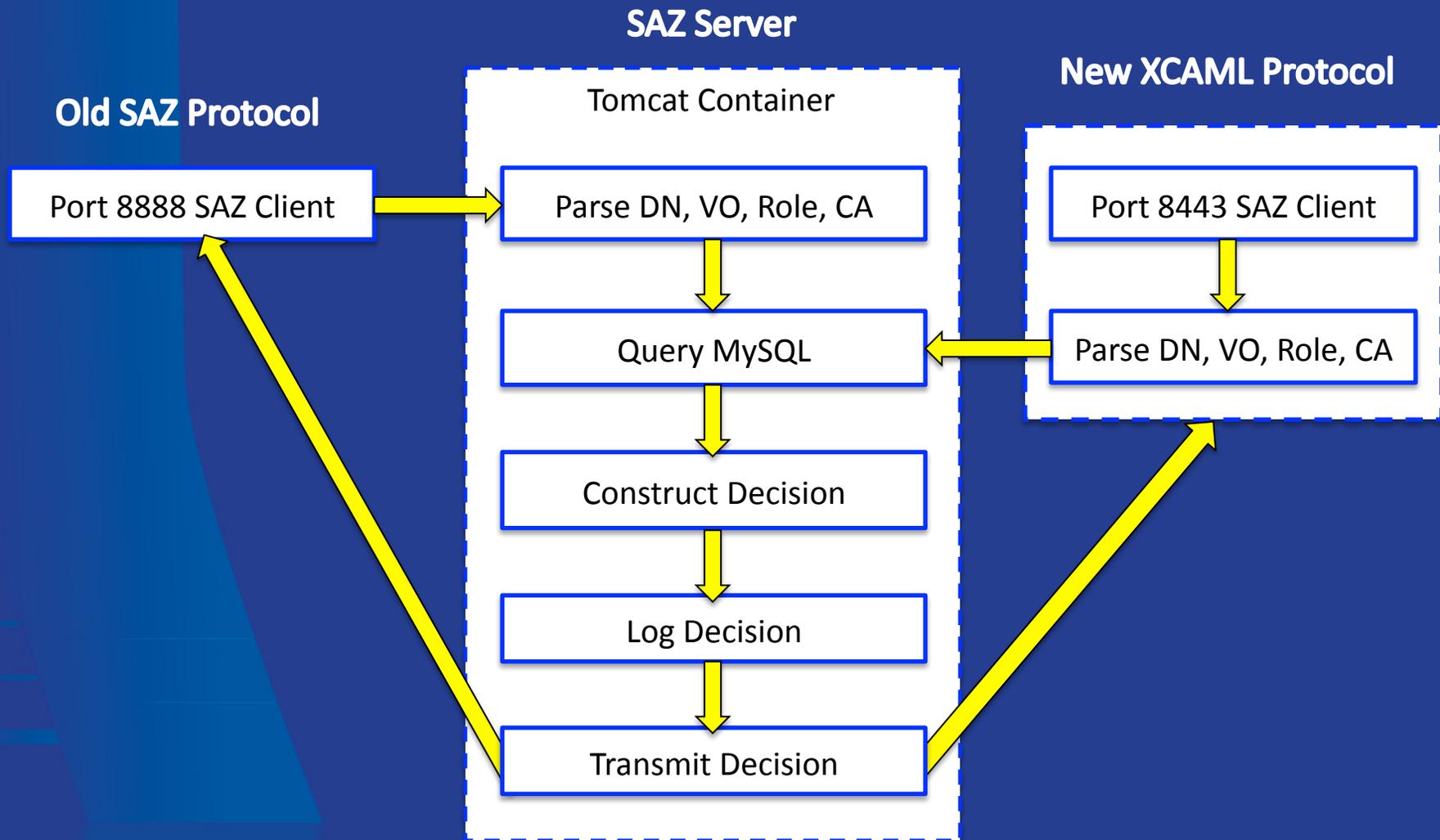
# Old (“current”) SAZ Protocol – Port 8888

- Client sends the public certificates in the users proxy to the SAZ server via port 8888.
- Server parses out DN, VO, FQAN, CA.
  - In the old SAZ V2.0.0b, the parsing logic frequently did not work well, and as a workaround, the SAZ server had to invoke a shell script voms-proxy-info to parse the proxy.
  - In the new SAZ V2.7.2, the parsing logic has been completely rewritten, and it no longer has to invoke the shell script voms-proxy-info to parse the proxy.
- Server performs MySQL queries.
- Server constructs the answer and sends it to the client.

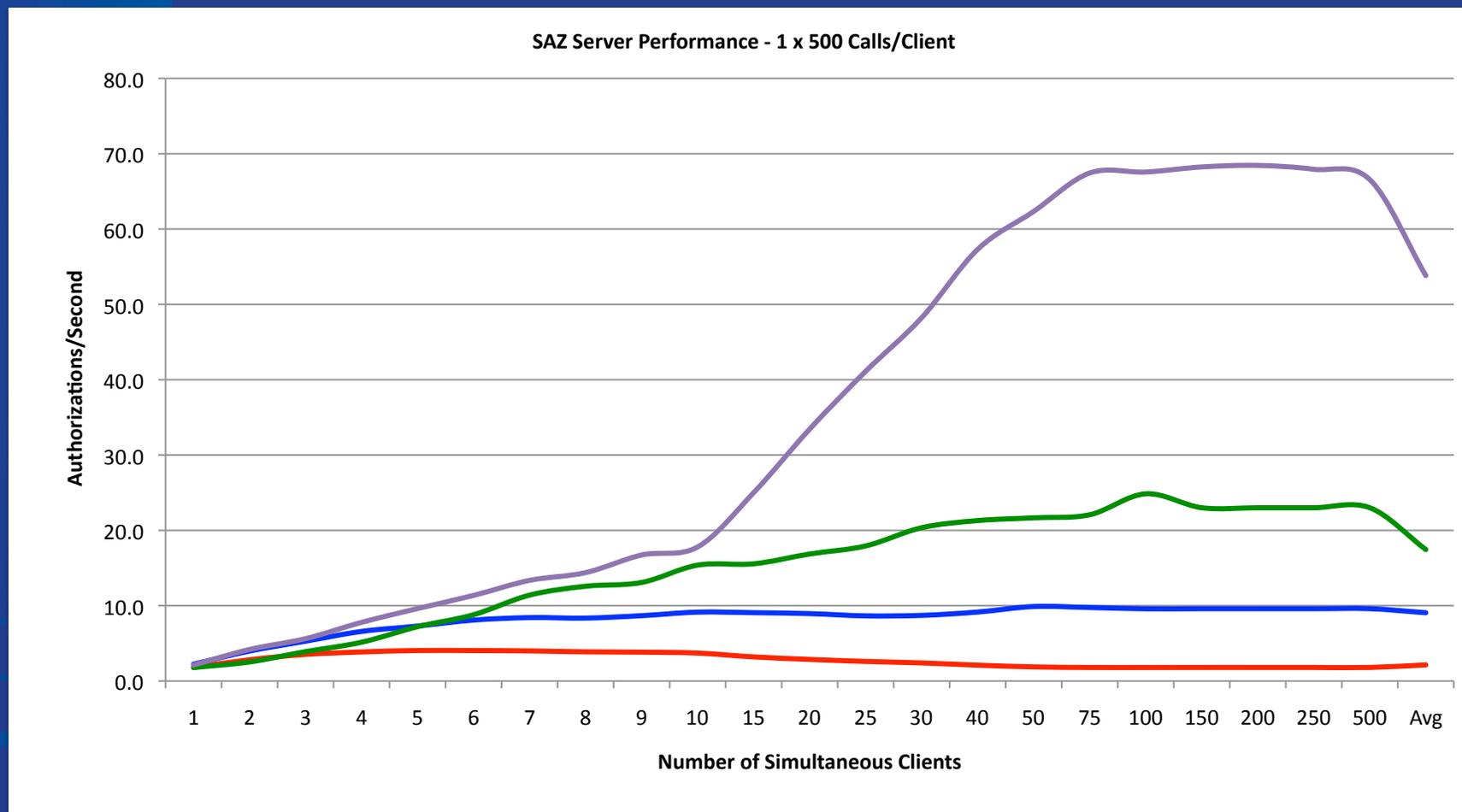
## New SAZ (XACML) Protocol – Port 8443

- Client parses out DN, VO, FQAN, CA and sends the information via XACML to the SAZ server via port 8443.
- Server performs MySQL queries.
- Server constructs the answer and sends it to the client.
- SAZ server V2.7.2 supports both 8888 and 8443 protocols simultaneously.

# Comparison of Old & New Protocol



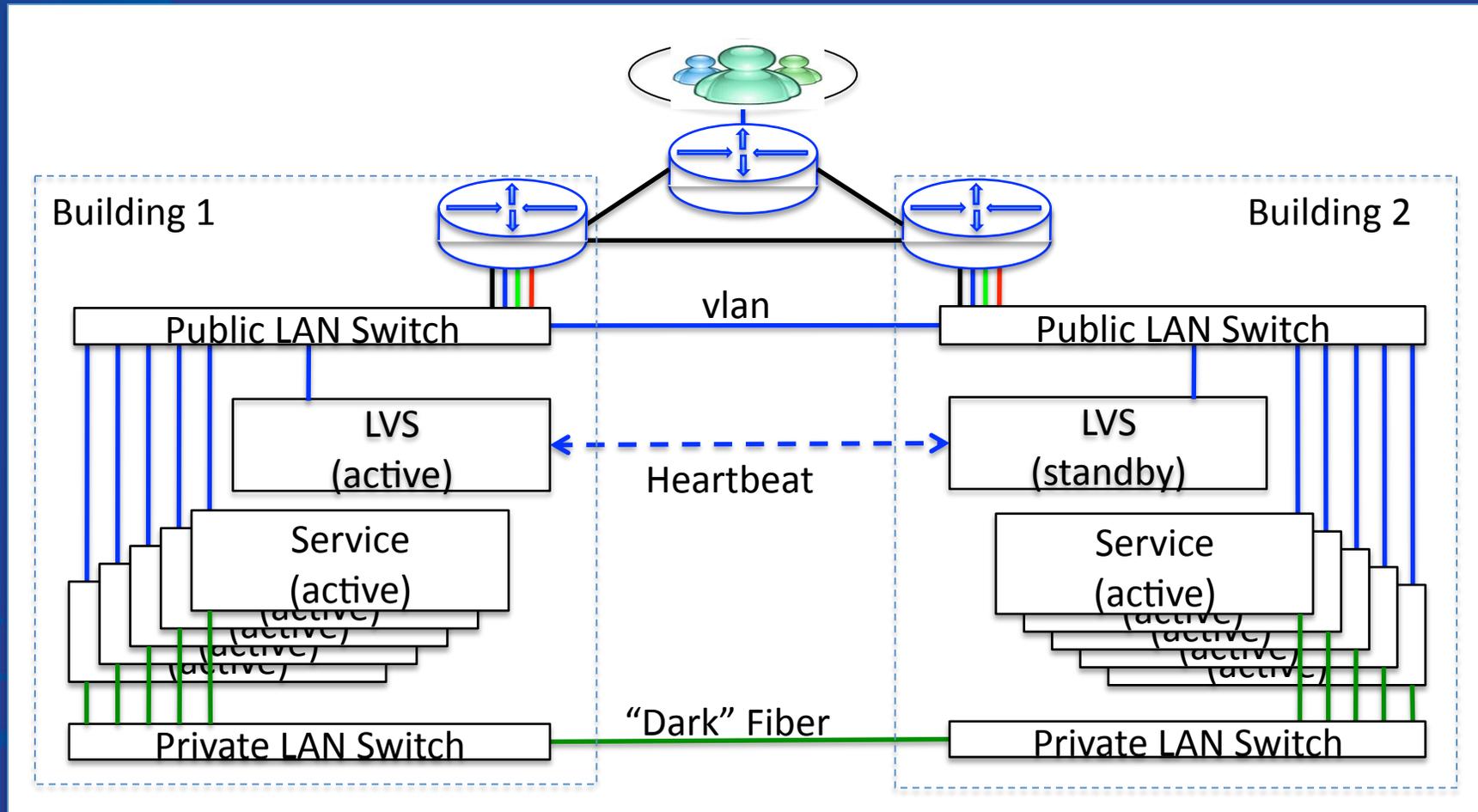
# Multiple Client SAZ production performance V2.0.1b (red) vs. V2.7.2 (purple)



# FermiGrid-HA2 Project

- At present, the FermiGrid machines are all in the FCC1 computer room – a single building with the corresponding power and network infrastructure,
  - Vulnerable to building issues (power issues 4X in past 16 months),
  - Vulnerable to network issues (6X in late February 2011).
- The existing FermiGrid-HA high availability infrastructure:
  - Web services: LVS active/active with MySQL back end,
  - File systems: DRBD + Heartbeat, (MyProxy and Gratia now, Compute Element soon),
  - Native HA – Condor collector/negotiator.
- The goal of the FermiGrid-HA2 project (started in March 2010) is to spread the systems and services between two buildings to lessen the chance of network cut or power outage disrupting all service.

# FermiGrid-HA2 Network



# FermiGrid-HA2 Rack Layouts

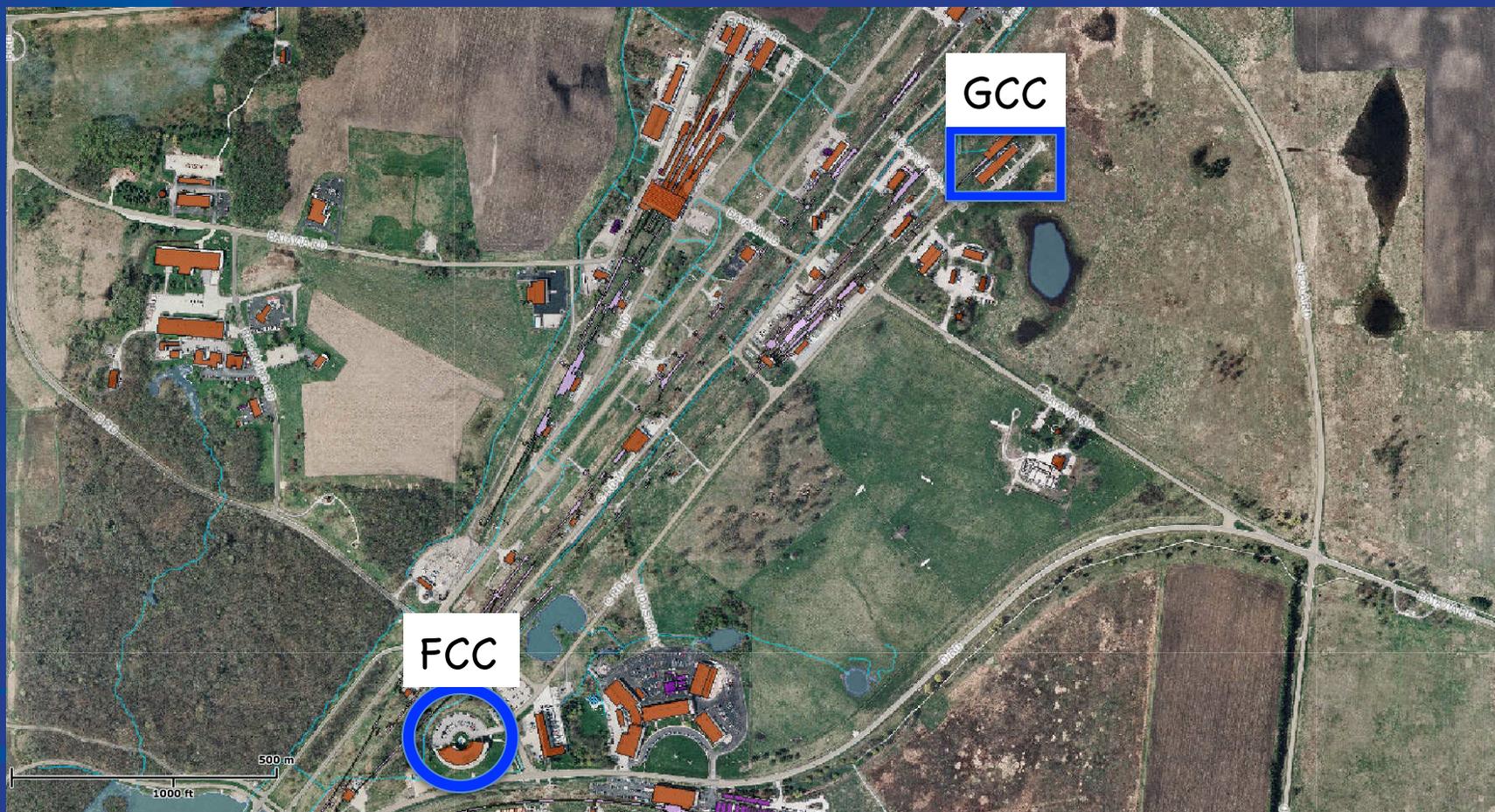
- Two (almost) identically configured racks:
  - One with 120VAC power distribution,
  - One with 208VAC power distribution,
- Both currently located in FCC1 computer room,
- Currently waiting on network configuration changes,
- 1<sup>st</sup> rack will be moved to FCC2 computer room,
- 2<sup>nd</sup> rack will be moved to GCCB computer room.

FermiGrid-HA2 Rack Front	Rack "U"	FermiGrid-HA2 Rack Rear
	42	Cisco Nexxus / 2960G Public LAN Switch
	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
blank - 2U	34	blank - 2U
	33	
	32	
fcdfsrv3 / fcdfsrv4	31	fcdfsrv3 / fcdfsrv4
	30	
fcdfsrv1 / fcdfsrv2	29	fcdfsrv1 / fcdfsrv2
	28	
fcdfsrv0 / fcdfsrv5	27	fcdfsrv0 / fcdfsrv5
	26	
blank - 2U	25	Cyclades PM10-L30A 120VAC
Display / Keyboard / Mouse	24	Cyclades PM10-L30A 120VAC
Raritan MasterConsole MCCAT116 KVM	23	Display / Keyboard / Mouse
blank - 1U	22	Omniview PS3 16 port KVM
Slave KDC - Sun Netra X1	21	Linksys SR2024 Private LAN Switch
blank - 1U	20	APC Transfer Switch
	19	blank - 1U
	18	
gratia12 (FCC) / gratia13(GCC)	17	gratia12 (FCC) / gratia13(GCC)
	16	
gratia10 (FCC) / gratia11 (GCC)	15	gratia10 (FCC) / gratia11 (GCC)
blank - 1U	14	blank - 1U
	13	
ress01 / ress02	12	ress01 / ress02
	11	
fermigrd5 / fermigrd6	10	fermigrd5 / fermigrd6
	9	
fermigrd2 / fermigrd3	8	fermigrd2 / fermigrd3
	7	
fermigrd1 / fermigrd4	6	fermigrd1 / fermigrd4
	5	
fermigrd0 / fermigrd7	4	fermigrd0 / fermigrd7
	3	
	2	Cyclades PM10-L30A 120VAC
blank - 2U	1	Cyclades PM10-L30A 120VAC

# FermiGrid-HA2 Pictures



# Geographical Redundancy



# XACML Performance Testing

- Recently, we have been utilizing the same techniques that were used to stress test the new SAZ server in early 2010 to stress test the XACML interfaces for the OSG Grid User Mapping Service (GUMS) and the FermiGrid banning service (SAZ).
- Initial tests showed adequate performance of the XACML interface under a “base” load of 50 clients continuously requesting authentications/authorizations.
  - The measured success rate was 100% for both SAZ and GUMS.
- Running a authorization “tsunami” of 5,000 jobs suddenly requesting near simultaneous authentications/authorizations showed poor performance.
  - The measured success rate was 39.8% and 50.7% for SAZ and GUMS respectively,
  - This is a real world example – we had 4,000 job user triggered authorization tsunami at 0308 on Saturday 23-Apr-2011.
- Investigation uncovered a coding flaw in the XACML client:
  - It had a hardcoded “TCP connection timeout” of only 170 milliseconds.
  - We have a test client with a more reasonable timeout of 300 seconds.
- Tests with the 300 second timeout client:
  - Measured success rates of 98.7% and ??.% for SAZ and GUMS respectively.
  - We are currently exploring the “TCP connection timeout” value vs. success rate curve and hope to have a recommended default value soon (maybe by the end of this week).

# Conclusions

- Good initial policy and planning gave FermiGrid an extensible architecture that we have been able to scale,
- More than 210,000,000 CPU-Hours delivered thus far, most in Open Science Grid,
- Testing at very large scale is the key to successful operation at very large scale,
- Expect completion of the FermiGrid-HA2 project by early June 2011:
  - Tuesday 24-May-2011 "Build & Test" - Move FermiGrid-HA2 Rack #1 to FCC2.
  - Tuesday 07-Jun-2011 "Go Live" - Move FermiGrid-HA2 Rack #2 to GCC-B.
  - Will give us resilience against single building failure,
  - We have to be out of the FCC1 computer room no later than 01-Jul-2011,
  - FCC power outage currently scheduled for 30-Jul-2011 or 06-Aug-2011 will be an important test.
- Ongoing planning and testing is key to robust and reliable service delivery.

# Fin

- Any Questions?