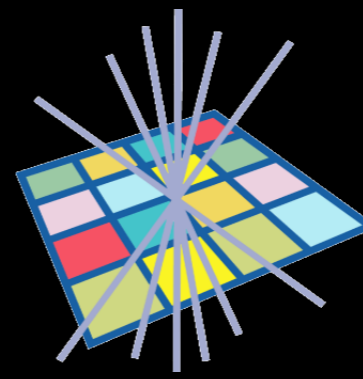


GridPP
UK Computing for Particle Physics



ScotGrid



**ScotGrid
vs.
Resilience**



**Graeme
(for the ScotGrid team) Disaster**

Resilience



**University
of Glasgow**



**Durham
University**

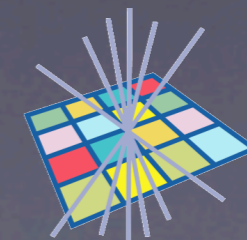
Overview

- Tier-2 Services
 - What do we have to do?
- Hardware, Fabric and Virtualisation
- Grid Services and Storage
- Monitoring, Documentation, Communications
- Results
- So, what about disasters?

WLCG/GridPP Tier-2 SLA

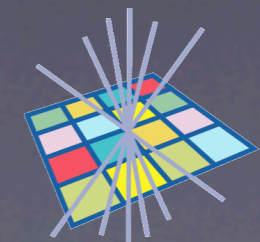
Service	Maximum delay in responding to operational problems		Average availability measured on an annual basis
	Prime time	Other periods	
End-user analysis facility	2 hours	72 hours	95%
Other services	12 hours	72 hours	95%

- What's an analysis facility? How does it differ from other services?
- What's a response?



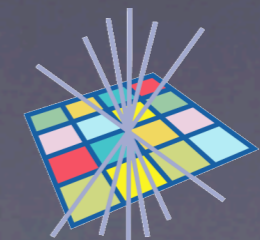
The ScotGrid View

- User analysis actually should be the core of what we do
- It means a successful end to end chain, starting with job definition and ending with a successful job run
- So it certainly covers the core site services of CE, batch system, storage and (probably) information systems
- So our target is usually a 2 hour response, not 12



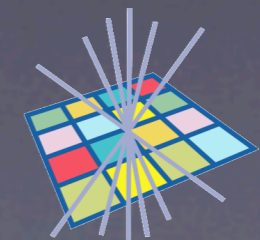
Hardware

- No site can run on poor or inappropriate hardware
 - We try to keep servers in maintenance
 - However, sometime budget constraints and procurement cycles force us to extend hardware beyond its natural life
- In this case hot spares and rapid re-installs are used
- Critical data must be backed up
- We probably do not exercise recovery from backup often enough



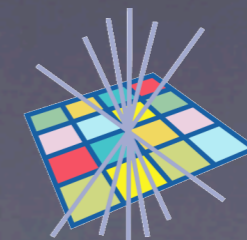
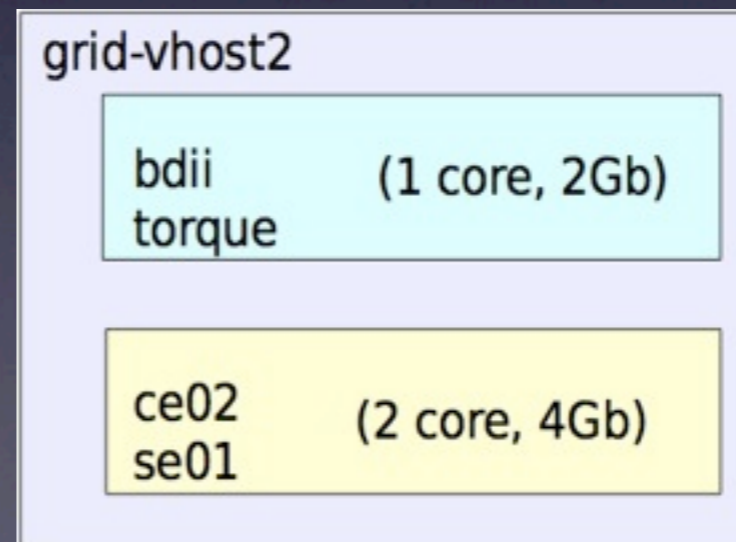
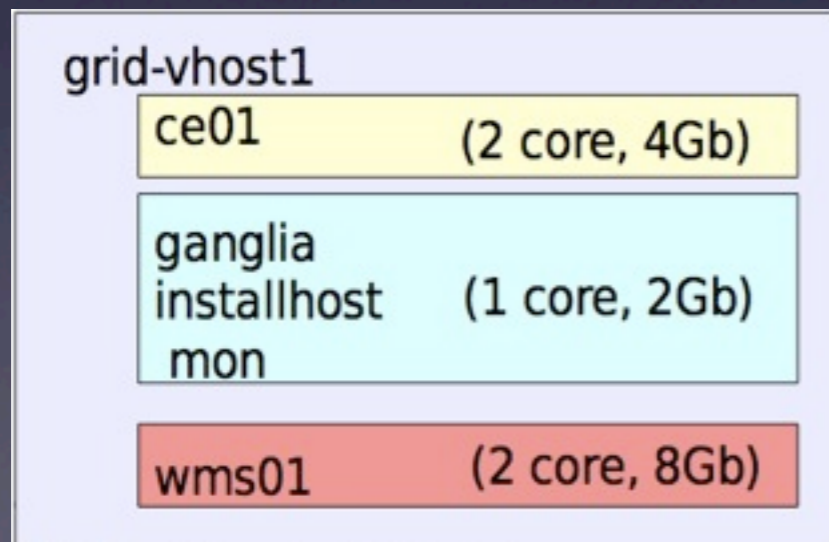
Fabric

- Both Durham and Glasgow use cfengine
 - This is configured in as much the same way as possible
 - This should provide a “known install” process
 - But it does not always work
 - Packages change underfoot, dependencies are usually in a mess, etc.
- Installation of pre-production machines has helped a lot here
- ECDF moving to System Imager, which will be able to snapshot grid front ends as well



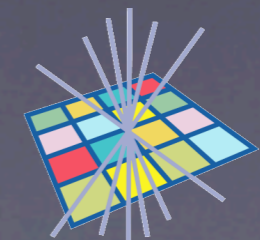
Virtual Machines

- Durham are our leading site here
 - Glasgow investigating this actively, but not yet deployed
- Needs integrated with rest of fabric system
- ECDF have their own virtualisation projects in conjunction with NGS
 - From which we should also profit

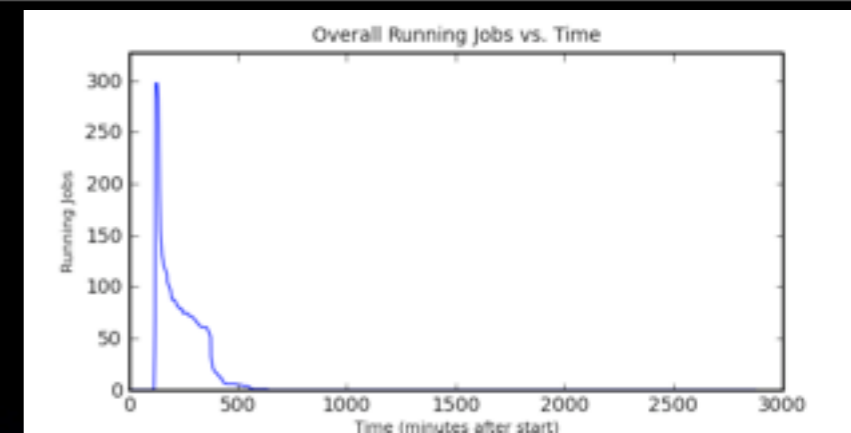


Redundancy, redundancy, ...

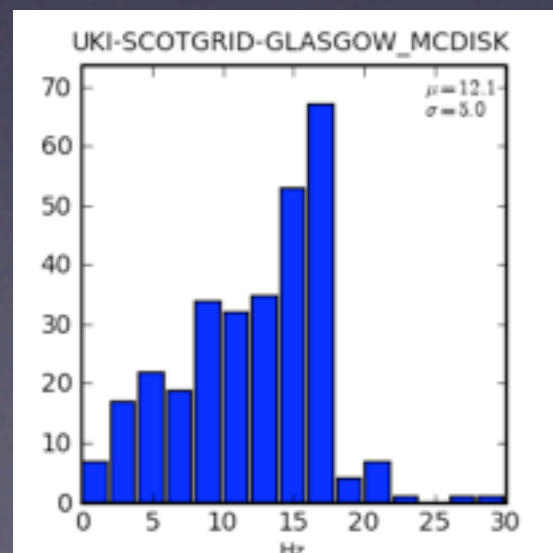
- Every site now has 2 CEs
 - This is probably more useful for SAM than anything else
 - I note in passing this is not well supported in Glue 1.3
- Glasgow even has two SEs
 - One used as an ATLAS PPS SE
 - Helped investigate ephemeral gSOAP errors
- The middleware is not well designed for high availability
 - however, we do know it well now and have workarounds for the problems we face
 - error messages are still very poor



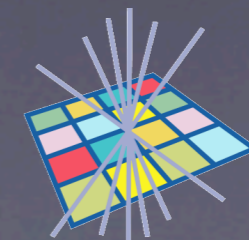
It's about data...



- Managing large SEs for production we largely understand
- We are still amateurs at providing user analysis services
 - 1500 missing replicas in DPM
 - Time consuming manual interventions
 - Lack-lustre performance of 300 jobs

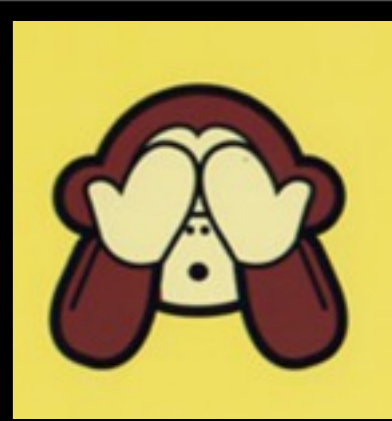


SITE	# PROCESSED	# EXPECTED
UKI-SCOTGRID-GLASGOW_MCDISK	85917	90504
TOTAL	85917	90504



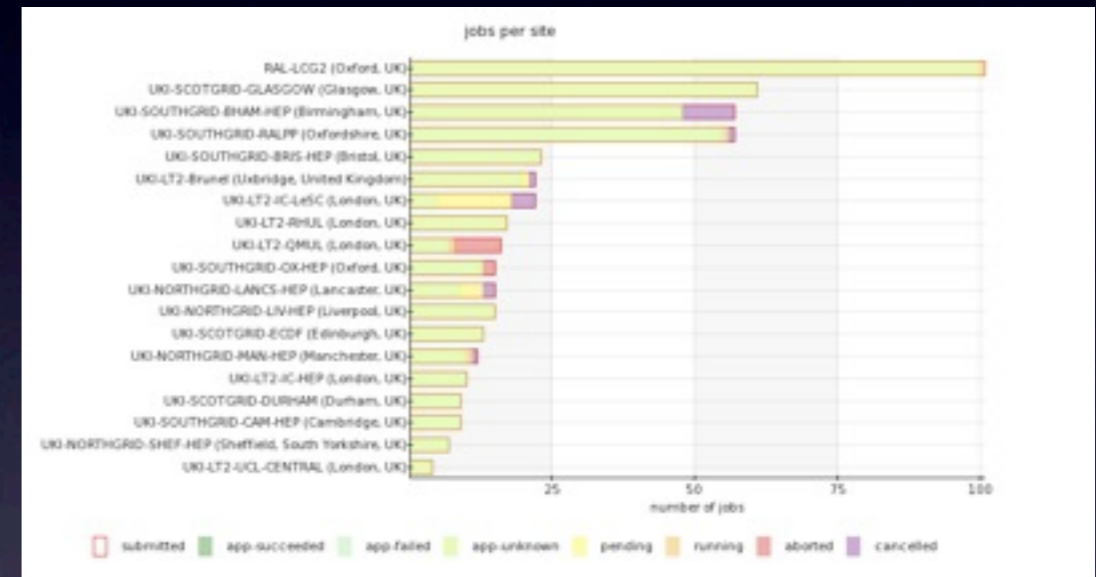
ScotGrid

Monitoring

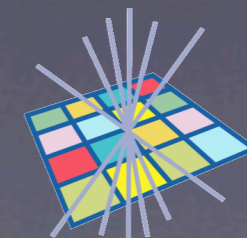


We're fine...

- Nagios infrastructure monitors SAM tests
- Ganglia useful for measuring trends and diagnosis
- We still lack specific tests and probes of performance
 - Need to share with other T2s
 - But is anyone co-ordinating?
- However, experiment relationships are strong and this helps a lot

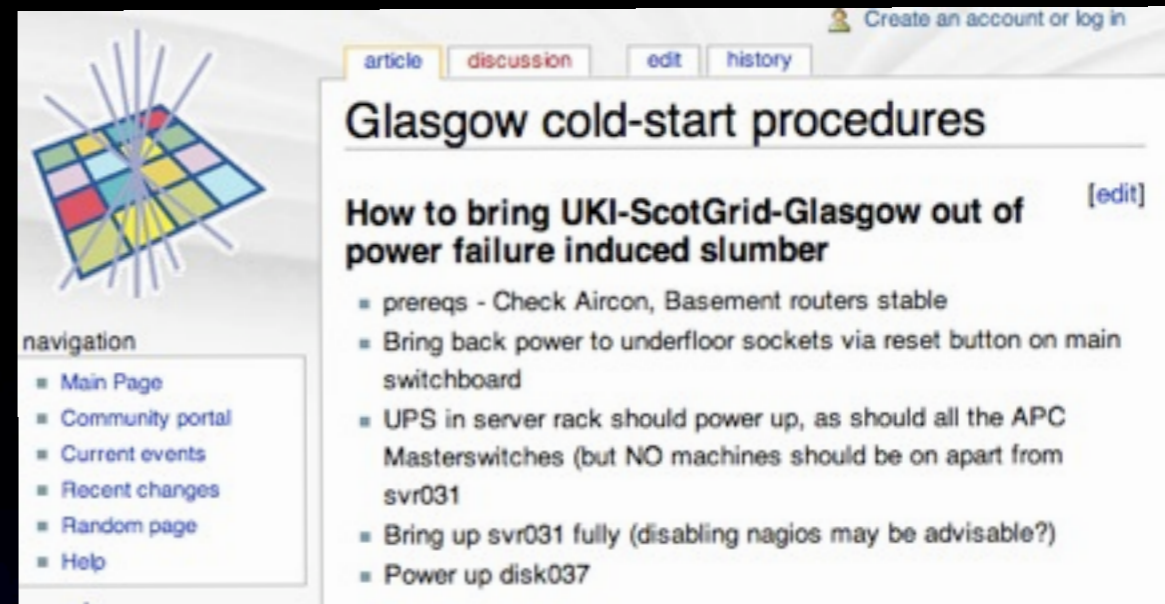


Host	Service	Status	Last Check	Duration	Attempt	Status Information
sv01a.scotgrid.ac.uk	sv.nagios.LockOwner_PerfCheck	CRITICAL	04-01-2009 12:07:40	78s 19m 31m 49s	64	CRITICAL - Socket timeout after 60 seconds
sv01a.scotgrid.ac.uk	sv.nagios.BCn-Check	CRITICAL	04-01-2009 12:04:59	110s 18m 0m 35s	64	Could not search/find objectclasses in GueStelUniqueID=UK-SCOTGRID-DURHAM-Msa-Via-Name=UK-SCOTGRID-DURHAM-D-one
sv01a.scotgrid.ac.uk	sv.nagios.DANCE-Check	CRITICAL	04-01-2009 12:04:14	207s 11m 30m 44s	64	CRITICAL - Socket timeout after 60 seconds
sv01a.scotgrid.ac.uk	sv.nagios.LockOwner_PerfCheck	CRITICAL	04-01-2009 12:12:05	187s 21m 49m 57s	64	CRITICAL - Socket timeout after 60 seconds
sv01a.scotgrid.ac.uk	sv.nagios.DnFTN-Check	CRITICAL	04-01-2009 12:10:54	0s 12h 4m 34s	64	Connection refused
sv021a.scotgrid.ac.uk	sv.nagios.LockOwner_PerfCheck	CRITICAL	04-01-2009 12:07:43	36s 0h 26m 17s	64	Connection refused
sv001a.scotgrid.ac.uk	sv.scs.GridProxy.Valid	CRITICAL	04-01-2009 12:04:59	131s 4h 44m 18s	33	Valid grid proxy doesn't exist.

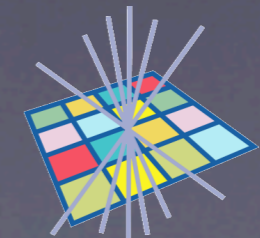


ScotGrid

Documentation

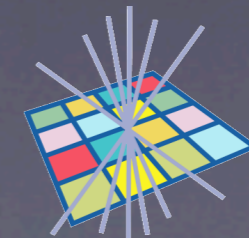
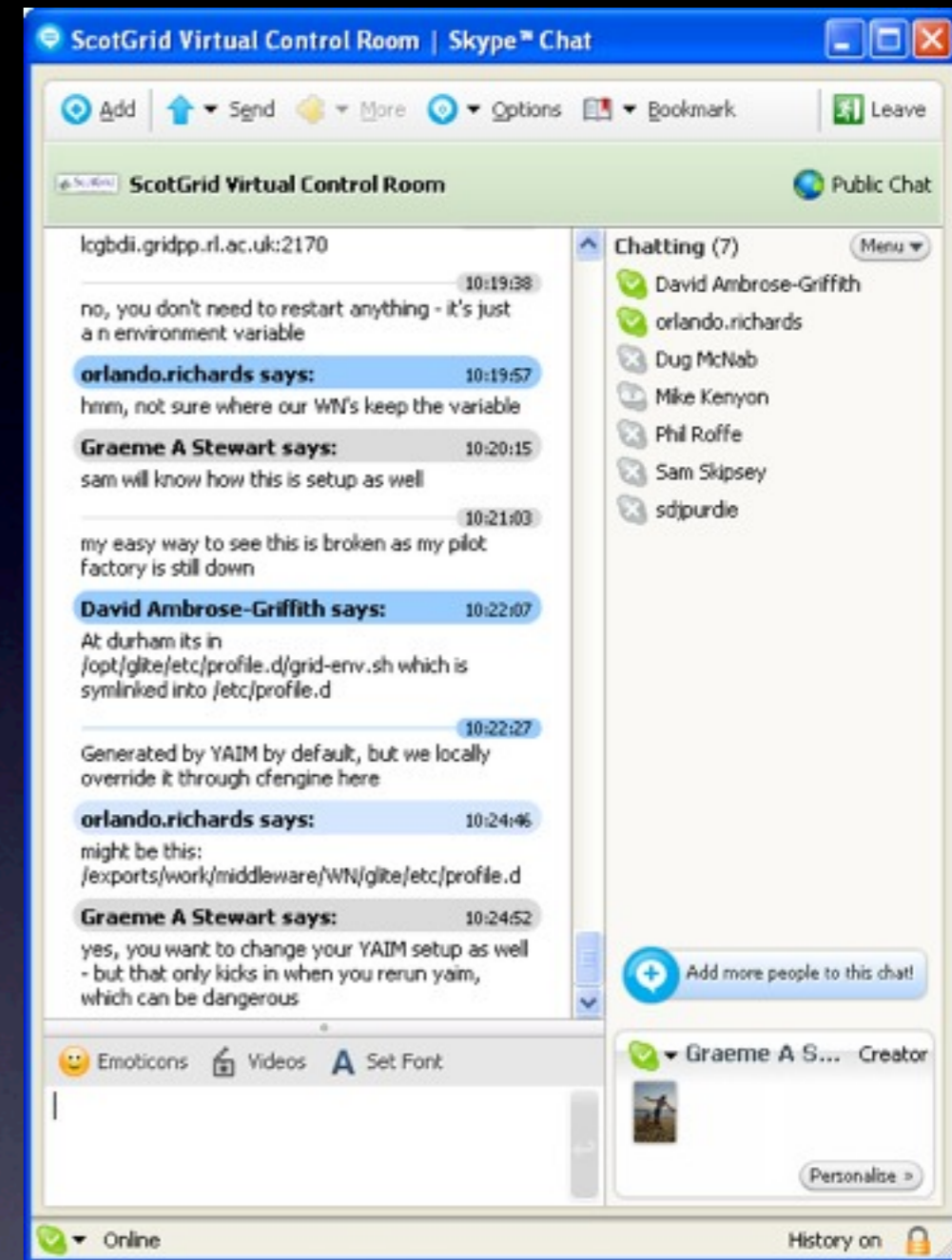


- We try to keep critical documentation up to date on the wiki
 - This is probably most advanced at Glasgow (more people)
- We learned the obvious stuff
 - Cold start procedure is printed out
- But again, documentation is inevitably imperfect so you need to ask for help...
 - Procedure to update documentation when flaws are discovered



Communications

- Multiple redundant communication channels help a great deal:
 - Jabber, Skype, Phone (Mobiles), Email
- Best new innovation is the ScotGrid Virtual Control Room
 - Group skype chat which is hugely helpful for live problem diagnosis and resolution
 - Within a couple of months used a lot



Results

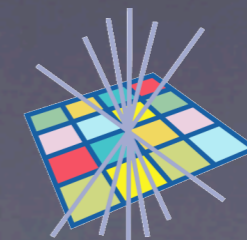
- We have had a few dry runs
 - e.g., two power cuts at Glasgow
- We are not doing too badly in the metrics

Institute	2Q08	3Q08	4Q08	1Q09
Durham	93%	92%	87%	1%
Edinburgh			85%	85%
Glasgow	98%	96%	97%	96%
Institute	2Q08	3Q08	4Q08	1Q09
LondonGrid	94%	89%	86%	87%
NorthGrid	96%	96%	85%	70%
ScotGrid	96%	95%	92%	91%
SouthGrid	96%	92%	96%	86%
Tier-1	93%	76%	70%	

Site	2Q07	3Q07	4Q07	1Q08	2Q08	3Q08	4Q08	1Q09
UKI-SCOTGRID-DURHAM	92%	91%	92%	94%	85%	95%	84%	97%
UKI-SCOTGRID-ECDF			13%	66%	72%	83%	99%	96%
UKI-SCOTGRID-GLASGOW	89%	93%	96%	85%	97%	96%	98%	97%
LondonGrid	79%	82%	80%	75%	72%	79%	84%	86%
NorthGrid	74%	80%	91%	89%	94%	96%	93%	96%
ScotGrid	91%	92%	73%	82%	85%	92%	94%	97%
SouthGrid	80%	86%	91%	94%	92%	93%	94%	95%

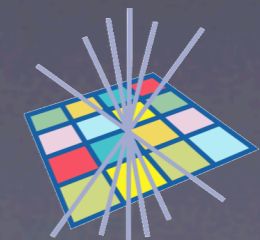
Brokered
ATLAS tests

SAM Tests



Our Disasters

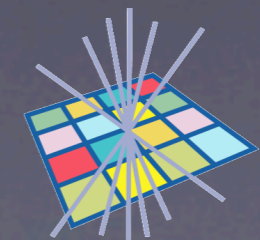
- Failure to deliver grid service on ClusterVision kit at Glasgow (Autumn 2006)
- Irretrievable loss of /etc on key cluster headnode at Glasgow (Autumn 2007)
- Failure to get grid jobs running properly on ECDF (2008)
- *In each case we probably did not deal with these situations optimally and we have to ensure that lessons are integrated into the team's experience and knowledge*



Disaster Planning



- Researchers find that disaster plans do not produce better responses to surprising crises, but that the processes of preparing them, does.
- ...it is more effective to, “create internal processes and organizational structures that build latent resilience ... so that they demonstrate positive adaptive behaviors when under stress.”



In other words...

- The plan itself is not so important
 - Disasters do not happen by the book
- The people who can make the plans are important
 - Have a good team of intelligent people who trust each other and work well together
- The Tier-2 needs to be agile to meet users' needs at all levels

