CCRC'08

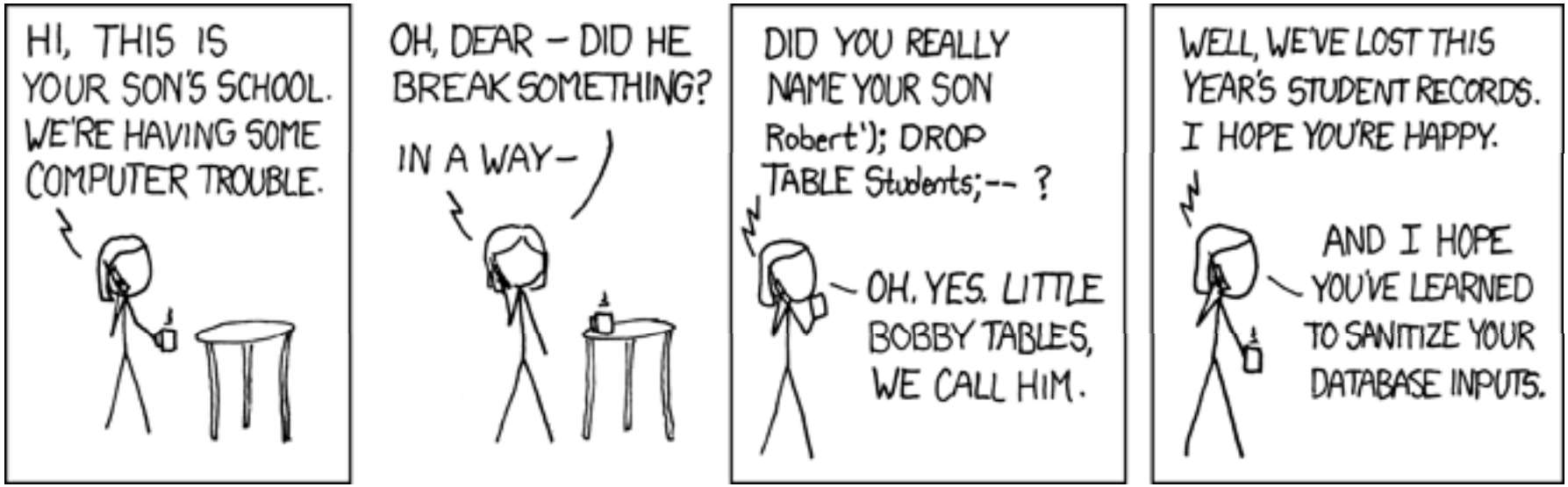
Jeff Templon

NIKHEF

JRA1 All-Hands Meeting

Amsterdam, 20 feb 2008

At least our users aren't malicious



What happens when

- Each experiment streams data into the T0
- The experiments' data model is followed T0-T1-T2
- Necessary computing (reconstruction, calibration) is done
- Sites try to reach the MoU targets for uptimes
- Graduate students try to analyze the data as it comes in
- All four experiments try this at the same time, at scale

CCRC

CCRC Goals

- For February, we have the following three sets of metrics:

1. The **scaling factors** published by the experiments for the various **functional blocks** that will be tested. These are monitored continuously by the experiments and reported on at least weekly;

2. The lists of **Critical Services**, also defined by the experiments. These are complementary to the above and provide additional detail as well as service targets. It is a goal that all such services are handled in a standard fashion – i.e. as for other IT-supported services – with appropriate monitoring, procedures, alarms and so forth. Whilst there is no commitment to the problem-resolution targets – as short as 30 minutes in some cases – the follow-up on these services will be through the daily and weekly operations meetings;

3. The services that a site must offer and the corresponding availability targets based on the **WLCG MoU**. These will also be tracked by the operations meetings.

- Phase 2 of CCRC in May

Scaling Factors

- We won't make it in Feb ... Functional problems
  - Castor can't handle load
  - At least on exp't framework can't handle load
  - FTS corrupt proxy problems (race condition)
- Note how "data driven" HEP is : without functional data flow, test at scale is not possible!

FTS "corrupted proxies" issue

- The proxy is only delegated if required
  - The condition is lifetime < 4 hours.
- The delegation is performed by the glite-transfer-submit CLI. The first submit client that sees that the proxy needs to be redelegated is the one that does it - the proxy then stays on the server for ~8 hours or so
  - Default lifetime is 12 hours.
- **We found a race condition in the delegation - if two clients (as is likely) detect at the same time that the proxy needs to be renewed, they both try to do it and this can result in the delegation requests being mixed up - so that that what finally ends up in the DB is the certificate from one request and the key from the other.**
- We don't detect this and the proxy remains invalid for the next ~8 hours.
- The real fix requires a server side update (ongoing).

BDII Scaling Problem

- BDII/SRM problem @ NIKHEF / SARA
- Discovery : only possible via monitoring of
  - Jobs success by exp'ts (not always optimum)
  - Site services by site
  - Coupled phenomenon
- BDII developer hears via 'vocal site person' about situation
- Active support
  - Checking deployment scenario
  - Asking for log files
  - Making recommendations

Post mortem by developer

Summary:

After fixing the initial problem with the missing index., the SRM failed at SARA and the lhcb jobs went into a lcg-gt loop which put a high query load onto the BDII.

Here are a number of recommendations following the incident.

Recommendation 1: Update the deployment and trouble shooting documentation explaining clearly the dangers of co-hosting the BDII with other services which could generate a high load.
Recommendation 2: Produce a new release of the BDII in which the new index is set.
Recommendation 3: Improve the efficiency of the lcg-utils commands, in particular has some kind of cache is need to avoid repeated queries to to BDII.
Recommendation 4: Use the logs gathered from the incident to test the performance of the BDII in such situation and address and performance bottlenecks found.
Recommendation 5: Monitor the load on the BDII at NIKHEF. If the load is consistently high, consider adding and additional machine for load balancing.
Recommendation 6: Advise the VOs the dangers of fail over methods that can do DOS loops. All fail over should contain some exponential backup.
Recommendation 7: Implement, as planned service discovery APIs. These interfaces to the information systems should contain limiter that prevent single threads (better processes) from issuing rapid fire queries. This limiter is needed to prevent accidental DOS attacks that make the whole resource unusable

CCRC Goals

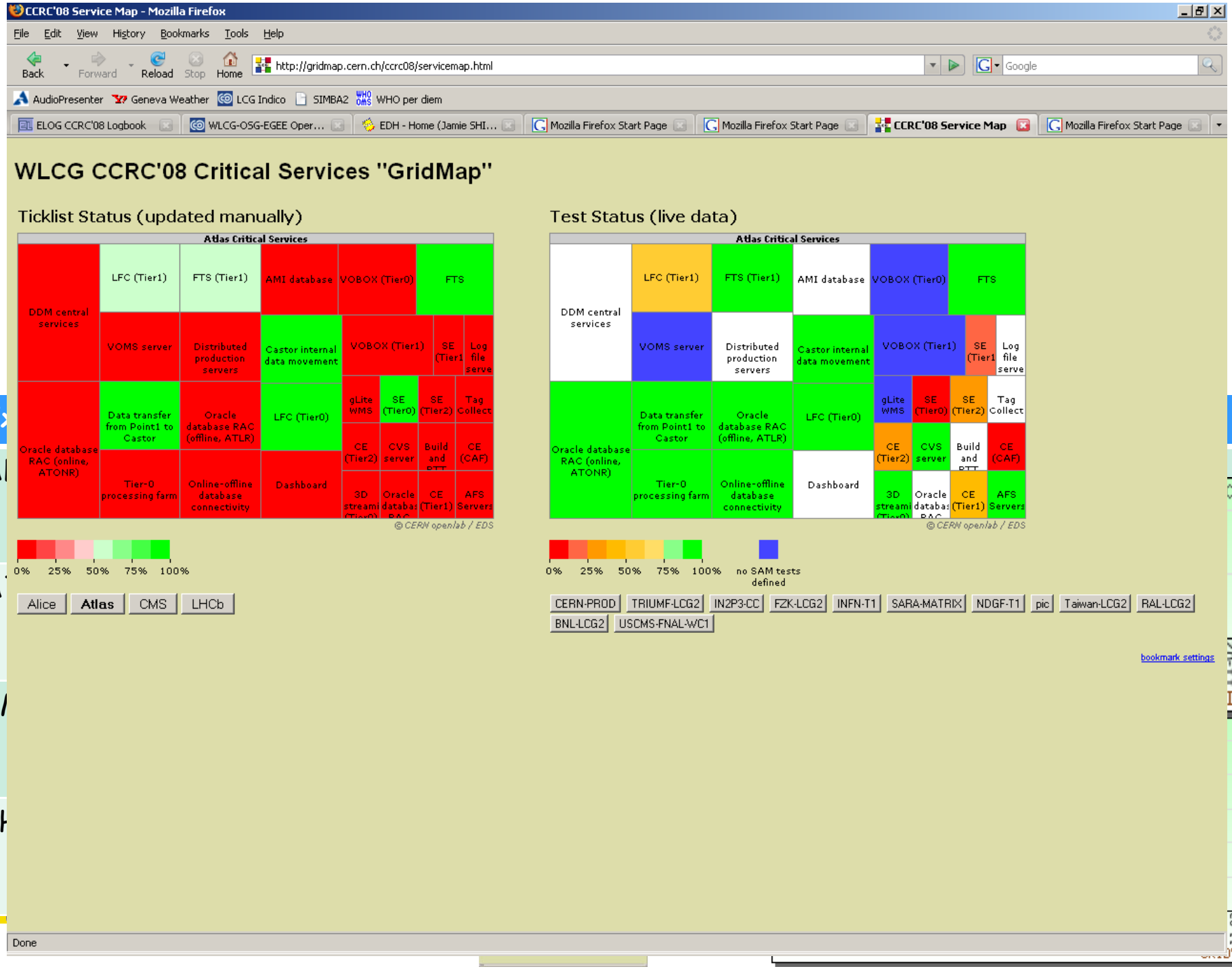
- For February, we have the following three sets of metrics:

1. The **scaling factors** published by the experiments for the various **functional blocks** that will be tested. These are monitored continuously by the experiments and reported on at least weekly;

2. The lists of **Critical Services**, also defined by the experiments. These are complementary to the above and provide additional detail as well as service targets. It is a goal that all such services are handled in a standard fashion – i.e. as for other IT-supported services – with appropriate monitoring, procedures, alarms and so forth. Whilst there is no commitment to the problem-resolution targets – as short as 30 minutes in some cases – the follow-up on these services will be through the daily and weekly operations meetings;

3. The services that a site must offer and the corresponding availability targets based on the **WLCG MoU**. These will also be tracked by the operations meetings.

- Phase 2 of CCRC in May


CCRC'08 Service Map - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Back Forward Reload Stop Home
http://gridmap.cern.ch/ccrc08/servicemap.html
Google
AudioPresenter Geneva Weather LCG Indico SIMBA2 WHO per diem
ELOG CCRC'08 Logbook
WLCG-OSG-EGEE Oper...
EDH - Home (Jamie SHI...
Mozilla Firefox Start Page
Mozilla Firefox Start Page
CCRC'08 Service Map
Mozilla Firefox Start Page
WLCG CCRC'08 Critical Services "GridMap"
Ticklist Status (updated manually)
Atlas Critical Services
DDM central services
LFC (Tier1)
FTS (Tier1)
AMI database
VOBOX (Tier0)
FTS
VOMS server
Distributed production servers
Castor internal data movement
VOBOX (Tier1)
SE (Tier1
Log file serve
Oracle database RAC (online, ATONR)
Data transfer from Point1 to Castor
Oracle database RAC (offline, ATLR)
LFC (Tier0)
gLite WMS
SE (Tier0)
SE (Tier2)
Tag Collect
CE (Tier2)
CVS server
Build and
CE (CAF)
Tier-0 processing farm
Online-offline database connectivity
Dashboard
3D streami
Oracle databas
CE (Tier1)
AFS Servers
© CERN openlab / EDS
0% 25% 50% 75% 100%
Alice Atlas CMS LHCb
Test Status (live data)
Atlas Critical Services
DDM central services
LFC (Tier1)
FTS (Tier1)
AMI database
VOBOX (Tier0)
FTS
VOMS server
Distributed production servers
Castor internal data movement
VOBOX (Tier1)
SE (Tier1
Log file serve
Oracle database RAC (online, ATONR)
Data transfer from Point1 to Castor
Oracle database RAC (offline, ATLR)
LFC (Tier0)
gLite WMS
SE (Tier0)
SE (Tier2)
Tag Collect
CE (Tier2)
CVS server
Build and
CE (CAF)
Tier-0 processing farm
Online-offline database connectivity
Dashboard
3D streami
Oracle databas
CE (Tier1)
AFS Servers
© CERN openlab / EDS
0% 25% 50% 75% 100%
no SAM tests defined
CERN-PROD TRIUMF-LCG2 IN2P3-CC FZK-LCG2 INFN-T1 SARA-MATRIX NDGF-T1 pic Taiwan-LCG2 RAL-LCG2
BNL-LCG2 USCMS-FNAL-WC1
bookmark settings
Done

Service Readiness

CERN IT Department

#	Question	Comments
1	High-level description of service available?	with architecture diagam
2	Middleware dependencies and versions defined?	OS deps, M/W deps, platforms supported
3	Code released and packaged correctly?	Repository + Tagging process, rpms/tarballs
4	Certification process exists?	
5	Automatic Configuration code exists?	e.g. Yaim, NCM, ...
6	Admin Guides available?	Installation, monitoring, problem determination
7	Disk, CPU, Database, Network requirements defined?	
8	Monitoring criteria described?	
9	Problem determination procedure documented	
10	Support chain defined (2nd/3rd level)?	
11	Backup/restore procedure defined?	
12	Suitable hardware used	
13	Monitoring implemented	
14	Test environment exists	
15	Problem determination procedure implemented	
16	Automatic configuration implemented	
17	Backup procedures implemented and tested	

Key:
Software Readiness
Service Readiness
Site Readiness

- Measure of how 'production-ready' a service :
  - In terms of software, service and deployment
- Manually edited (under SVN control) by responsibles
  - EIS team, service managers, deployment team

CCRC Goals

- For February, we have the following three sets of metrics:

1. The **scaling factors** published by the experiments for the various **functional blocks** that will be tested. These are monitored continuously by the experiments and reported on at least weekly;

2. The lists of **Critical Services**, also defined by the experiments. These are complementary to the above and provide additional detail as well as service targets. It is a goal that all such services are handled in a standard fashion – i.e. as for other IT-supported services – with appropriate monitoring, procedures, alarms and so forth. Whilst there is no commitment to the problem-resolution targets – as short as 30 minutes in some cases – the follow-up on these services will be through the daily and weekly operations meetings;

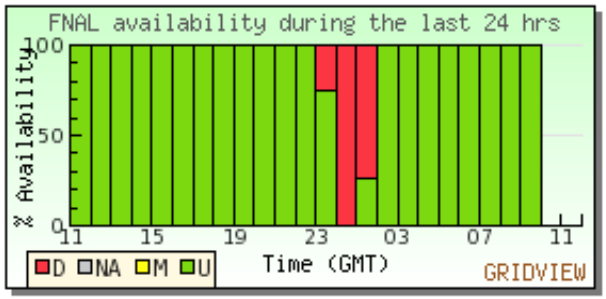
3. The services that a site must offer and the corresponding availability targets based on the **WLCG MoU**. These will also be tracked by the operations meetings.

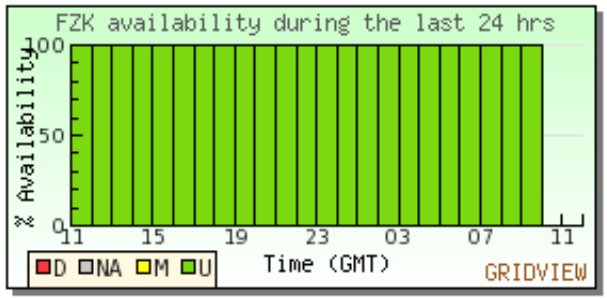
- Phase 2 of CCRC in May

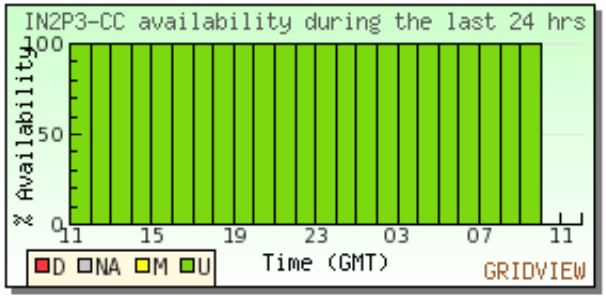
Site performance

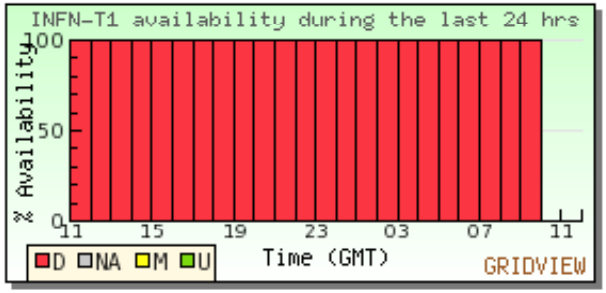
Service	*Maximum delay in responding to operational problems*			*Average availability measured on an annual basis*	
	Service interruption	Degradation of the capacity of the service by more than 50%	Degradation of the capacity of the service by more than 20%	During accelerator operation	At all other times
Acceptance of data from the Tier-0 Centre during accelerator operation	12 hours	12 hours	24 hours	99%	n/a
Networking service to the Tier-0 Centre during accelerator operation	12 hours	24 hours	48 hours	98%	n/a
Data-intensive analysis services, including networking to Tier-0, Tier-1 Centres outwith accelerator operation	24 hours	48 hours	48 hours	n/a	98%
All other services – prime service hours[6]	2 hour	2 hour	4 hours	98%	98%
All other services – outwith prime service hours	24 hours	48 hours	48 hours	97%	97%

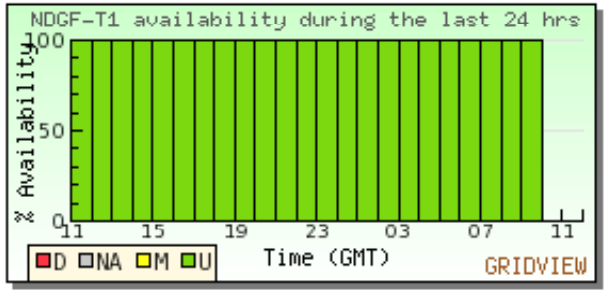
99% availability means < 100 minutes per week!

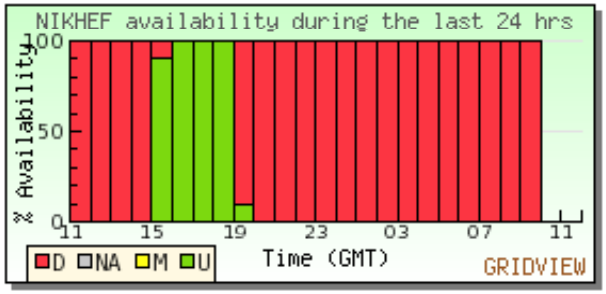

FNAL availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW

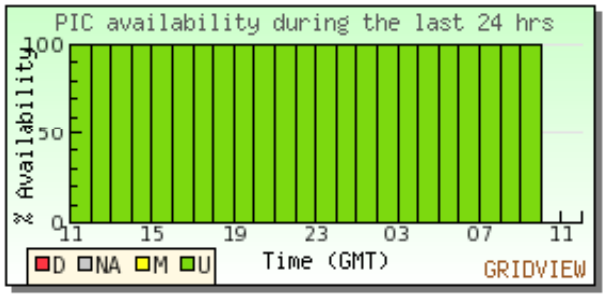

FZK availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW

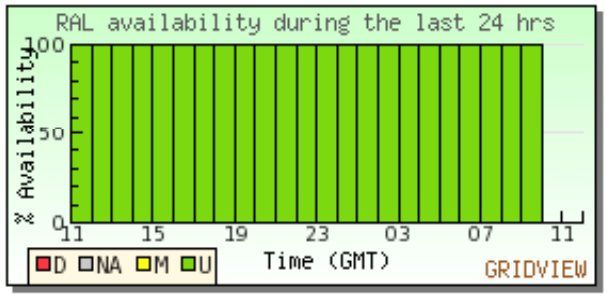

IN2P3-CC availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW

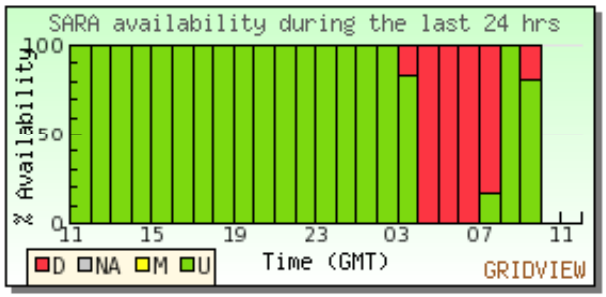

INFN-T1 availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW

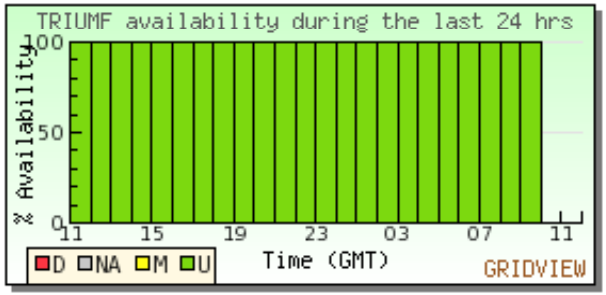

NDGF-T1 availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW


NIKHEF availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW

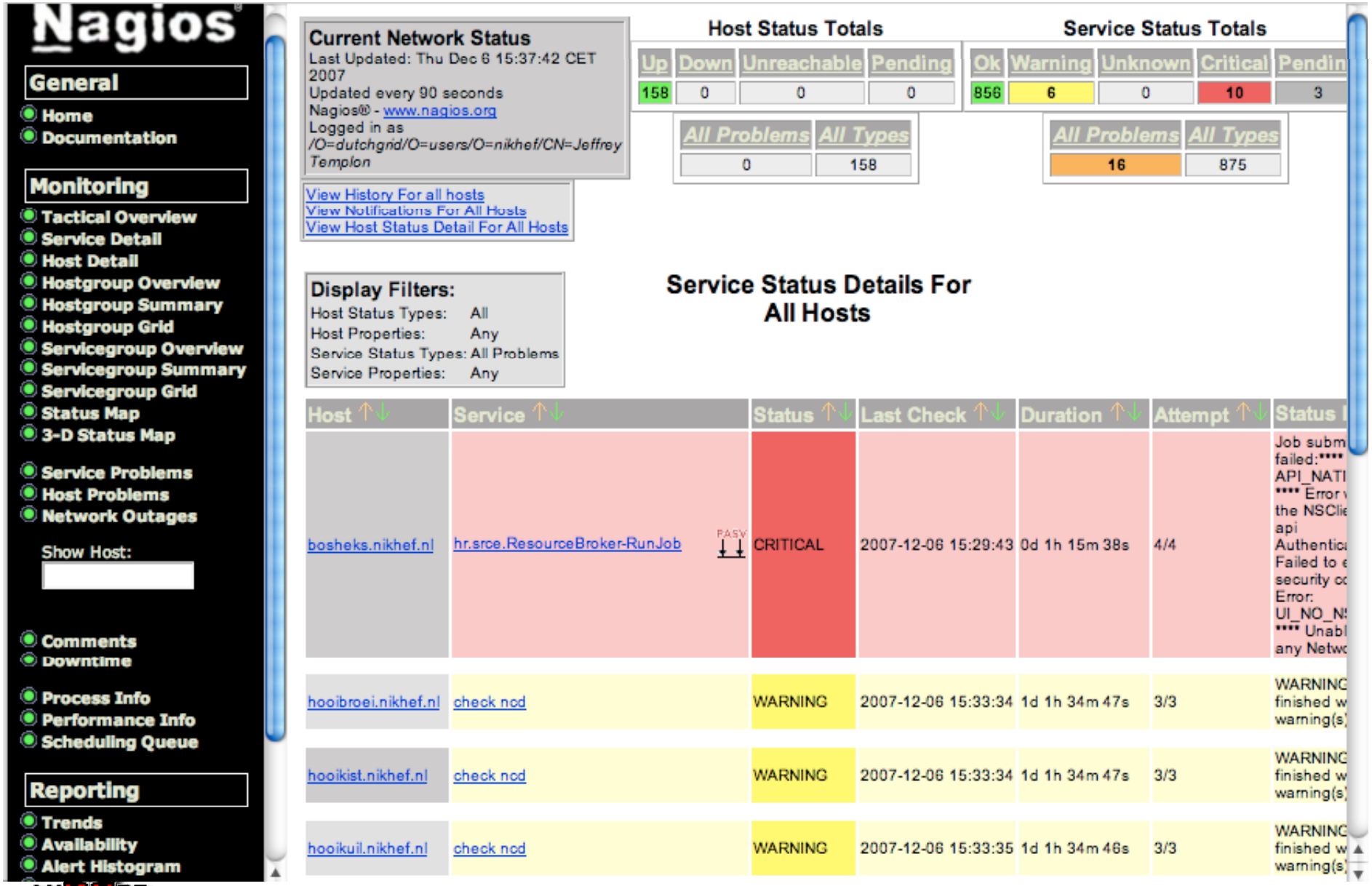

PIC availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW


RAL availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW


SARA availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW


TRIUMF availability during the last 24 hrs
% Availability
100
50
0
11 15 19 23 03 07 11
D NA M U
Time (GMT)
GRIDVIEW

GGUS is not fast enough ...



Middleware Coverage

- AAA : already reasonably well stressed
- FTS : broader range of usage, target SRMs, new SRM interface, higher rate (race condition ...)
- SRMs : stressed to max
- WMS : unknown to what extent. LHCb apparently using RB.
- CREAM : big miss. Should push extremely hard to get this ready for phase 2.
- Glexec / LCMAPS-server : another big miss
- All products could use improvement in logging / diagnostics / monitoring!!!!!

Handling Problems…

- Need to clarify current procedures for handling problems – some mismatch of expectations with reality
  - e.g. no GGUS TPMs on weekends / holidays / nights…
    - **c.f. problem submitted with max. priority at 18:34 on Friday…**
  - Use of on-call services & expert call out as appropriate
    - {alice-,atlas-}grid-alarm; {cms-,lhcb-}operator-alarm;
  - Contacts are needed on all sides – sites, services & experiments
    - e.g. who do we call in case of problems?
- Complete & open reporting in case of problems is essential!
  - Only this way can we learn and improve!
  - **It should not require Columbo to figure out what happened…**
- Trigger post-mortems when MoU targets not met
  - This should be a light-weight operation that clarifies what happened and identifies what needs to be improved for the future
  - Once again, the problem is at least partly about communication!

Don't panic

- Many EGEE / JRA1 services are in considerably better shape than exp't middleware
- BUT this is no license to slow down or slack off :
  - exp't efforts are often much more focused, they can catch up quickly
  - EGEE services are more critical : problems here affect all VOs / entire site. You *must* do better!
  - If exp'ts catch up and pass us, they will be merciless