WLCG Service Report

Harry.Renshall@cern.ch

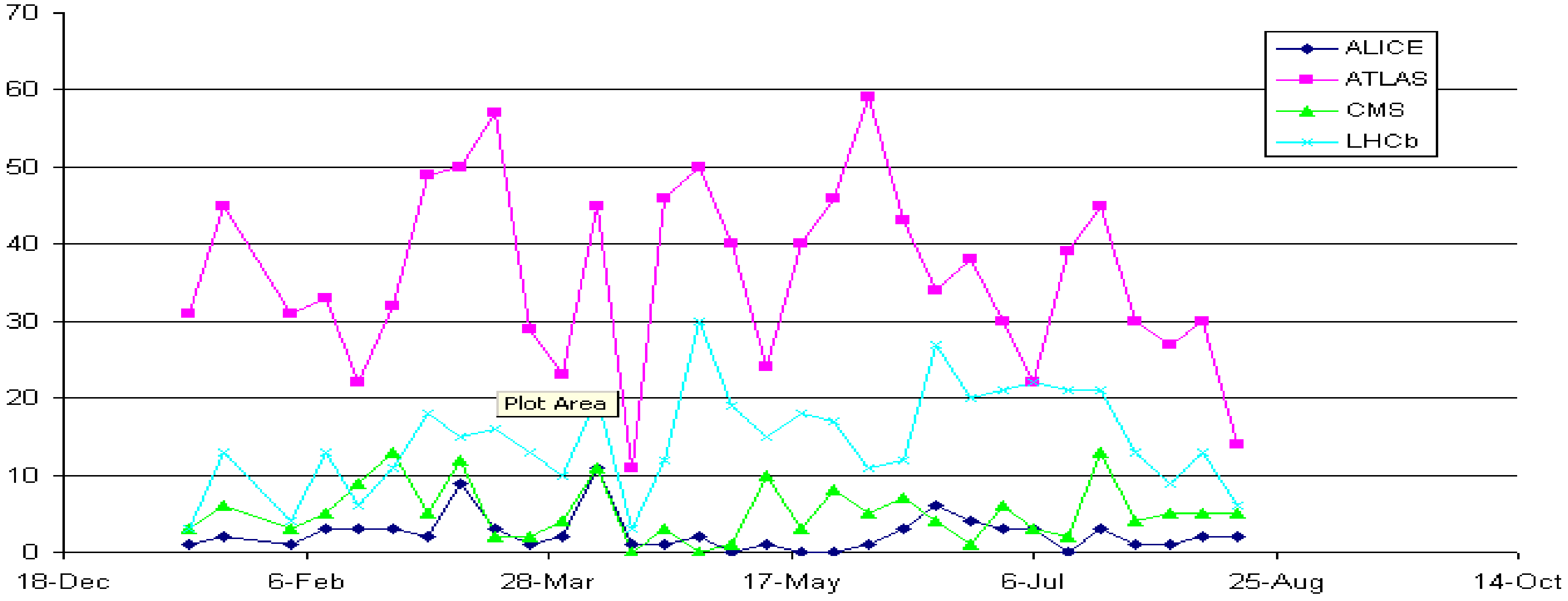
~~~

WLCG Management Board, 18th August 2009
~~~

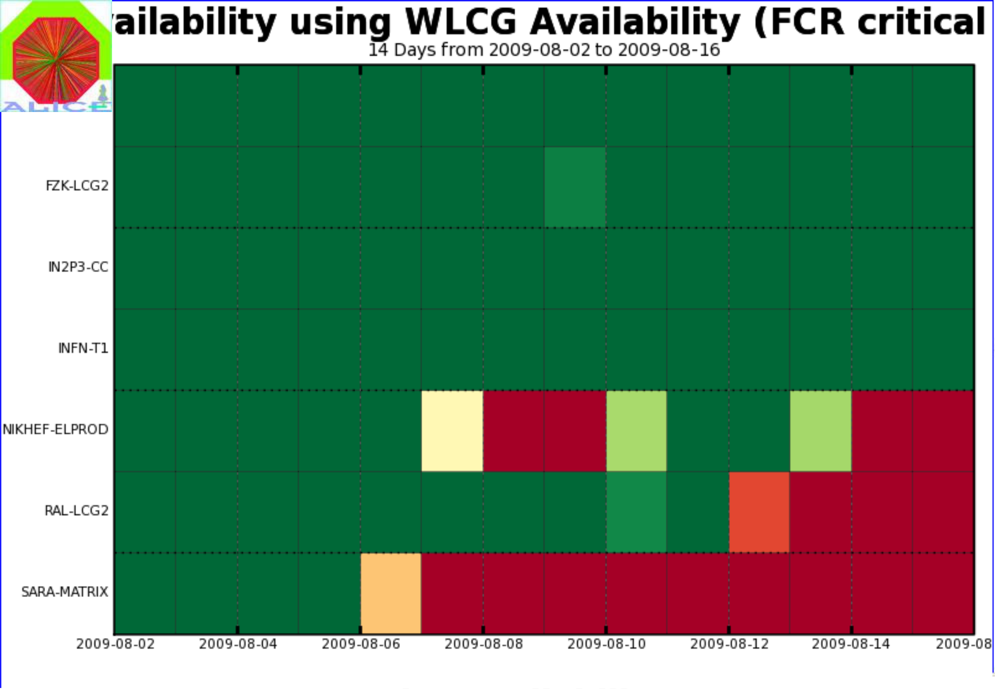
Introduction

- Covers the two weeks 2nd to 15th August
- Mixture of problems – mostly site related
- Three alarm tickets:
  - from CMS to CERN on 12 August when an ETICS virtual machine flooded network with DHCP requests (not reproducible) overloading a network switch then a router in front of castorcms and castoralice. Switch was stopped overnight.
  - ATLAS to DE-KIT on 9 August (when MC disk space filled up) exceptionally to keep MC Production running. 2TB added (with thanks from ATLAS) but KIT think this did not warrant an alarm ticket – experiment production space should be well planned – ATLAS agree but at the time did not know status of some obsolete data deletion services at FZK.
  - ATLAS to RAL on 7 August for hanging LFC connections. Front end servers were rapidly restarted fixing problem.
- Incidents leading to service incident reports
  - RAL air conditioning (chiller) failure from 12-17 August. Draft SIR available (see later report).
  - Fibre cut between Madrid and Geneva – LHCOPN primary and secondary down. SIR requested.

GGUS summary (2 weeks)

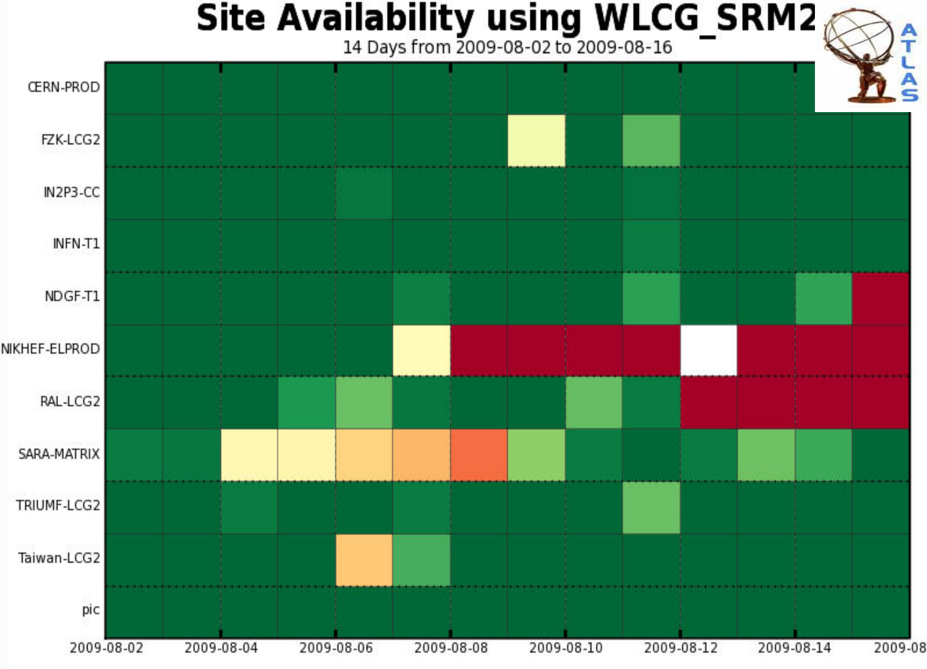
VO	User	Team	Alarm	Total
ALICE	4	0	0	4
ATLAS	10	32	2	44
CMS	8	1	1	10
LHCb	0	19	0	19
Totals	22	52	3	77





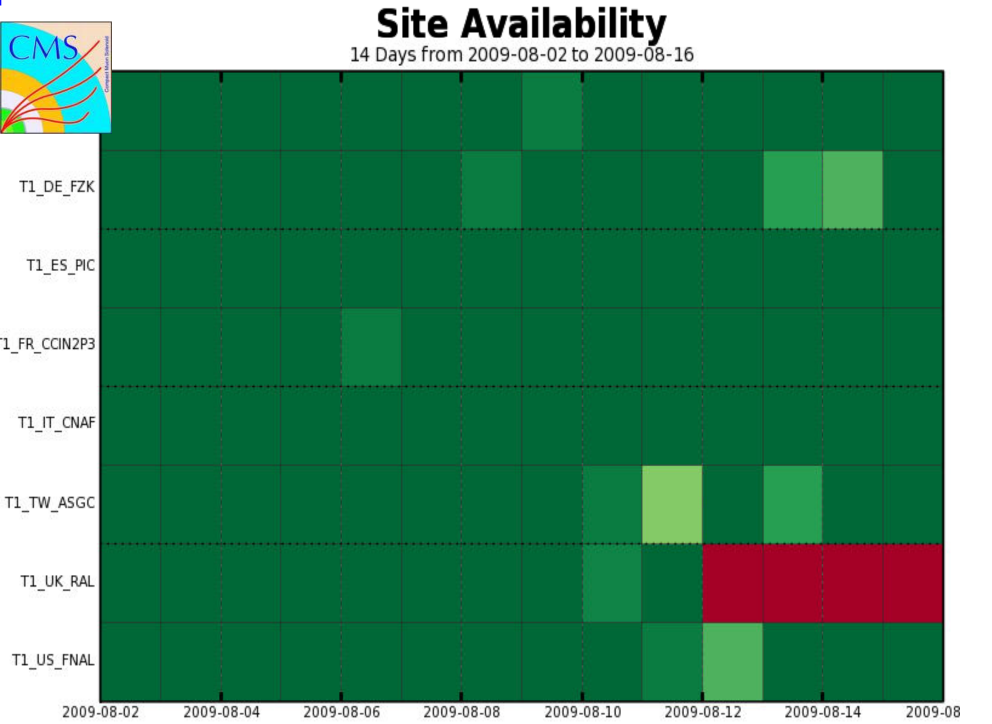
ailability using WLCG Availability (FCR critical

14 Days from 2009-08-02 to 2009-08-16



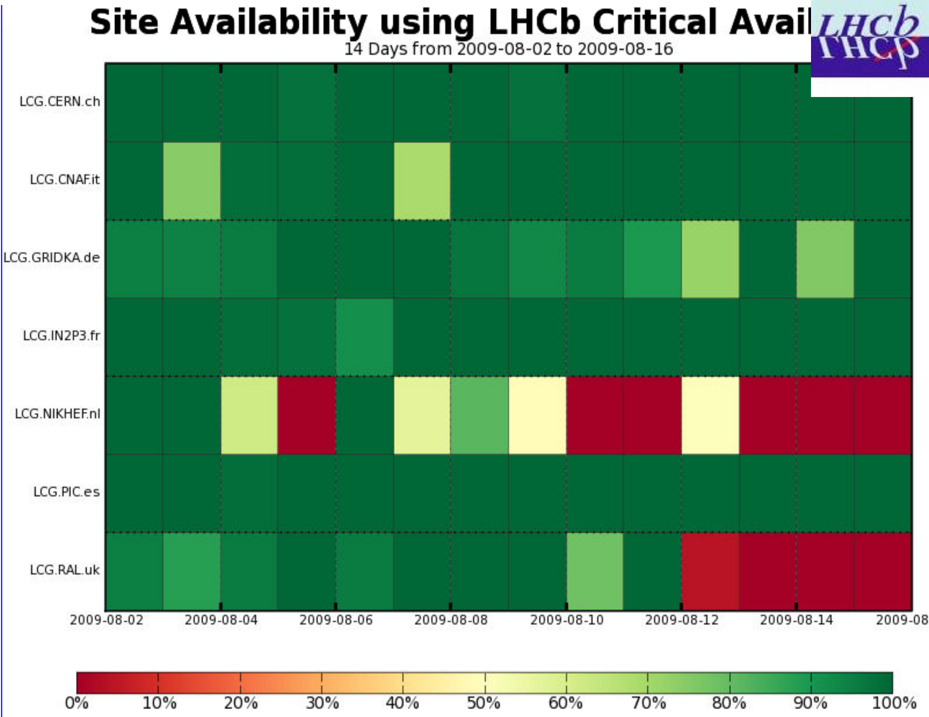
Site Availability using WLCG_SRM2

14 Days from 2009-08-02 to 2009-08-16



Site Availability

14 Days from 2009-08-02 to 2009-08-16



Site Availability using LHCb Critical Avail

14 Days from 2009-08-02 to 2009-08-16

Experiment Availabilities Reports

NIKHEF scheduled move to new computer centre performed from 10 to 14 for normal servers back up on 14th. This week moving disk servers and worker nodes.

The ALICE earlier but ongoing downtime at SARA was due to a failed ALICE VObox there. There is a separate ALICE VObox at NIKHEF – not clear if/how the two are coupled.

RAL had air conditioning stoppages due to water chiller failures starting on 12th continuing till Monday 17th.

RAL Air Conditioning stoppage (1/2)

The RAL Tier1 (RAL-LCG2) carried out an emergency power down following air conditioning failure during the night Tuesday-Wednesday 11-12 August, this was the second event in 2 days. All batch and CASTOR services had to be halted (and remain down) other critical services such as RGMA, the LFC and FTS have remained up the whole time. On Thursday 13th we also suffered a water leak (condensation) onto our main CASTOR tape robot.

Status as of 14.00 today is that All services are now back up and the downtime was ended in the GOCDB by 10am this morning. The over-pressure sensor that took down the Tier-1 has been re-configured to provide an alarm only but it is not yet completely clear if there actually was an over pressure (in the chilled water system) and if so, what caused it. We are actively seeking answers to these questions but have to work with the contractors and within the warranty constraints. Our best assessment is that there is a 5-10% chance of a recurrence.

RAL Air Conditioning stoppage (2/2)

We lost one D1T0 disk array from an ATLAS MCDISK pool on restart containing 99000 files. List being passed to ATLAS.

A rough draft of an SIR is on the GridPP website at http://www.gridpp.ac.uk/wiki/RAL_Tier1_Incident_20090810

The water leak onto the tape robot was traced to an overflowing drip tray on an aircon unit in the upper floors of the building. The water had made its way some distance and had leaked through a crack in the ceiling into the machine room. The drip tray has been replaced with a larger one and the reasons for the overflow are being investigated. An inventory of all water sources is being made. Some damaged electronics has been replaced, and various tapes and driveheads have been examined and we believe only suffered superficial splash marking. We will monitor the drive error rate for any increase; we believe the leak had been ongoing for some time.

Miscellaneous Reports (1/3)

- CERN responded to the Redhat 4 and 5 zero-day exploit of loaded modules that was exposed on the morning of Friday 14th August.
- The workaround of modifying /etc/modprobe.conf was validated as good enough for the weekend for protecting lxplus/lxbatch. Was rapidly propogated but nodes with a suspect module loaded (mostly bluetooth was found) needed reboot – one lxplus (of 50) and 150 lxbatch (of 1000).
- Experiment-info list was informed of reboots with apologies for any losses and also of later decision to reboot lxbuild machines.
- Linux for controls list was also advised to include WAN accessible DAQ and accelerator services.
- CERN actions were minuted in the daily report.

Miscellaneous Reports (2/3)

Fibre Madrid to Geneva was cut about 12.30 on Wednesday 5th August due to public construction work and taking out both primary and secondary OPN links. First indication was a GGUS ticket from an ATLAS user at 17:04 on the 4 August reporting file transfer errors showing in the dashboard ddm web page for PIC_DATADISK where PIC thought problems lay with FZK and INFN. Reply from OPN was then confusing as it referred to previous non-OPN routing problems between PIC and FZK and CNAF. Did PIC know the OPN was down ?

This ticket was superseded by an ATLAS team ticket at 20.00. Reply on 5 August did mention that the OPN was down and referred to a GGUS LHCOPN ticket. These cross references are part of the agreed procedure. I think the T1 were confused between the issue of the routing when the OPN is down and the OPN failure itself and the tickets reflect this.

Main concern is why there was no GGUS LHCOPN ticket earlier if the break happened at 12.30 on 4th as this would have saved experiment and sites time. We cannot see any dashboard style monitoring available to us of the OPN – some monitoring is password protected.

Total downtime was 26 hours so a SIR has been requested.

Miscellaneous Reports (3/3)

- Several FZK disk servers (ALICE and LHCb) were hit over several days by a bug in the bios that falsely detected overheating and shut down the machine.

- One of three FZK tape robots broke down several times from 8 August leaving the tape service intermittently degraded. Finally fixed on 13 August.

- WLCG now formally requesting all sites to upgrade worker nodes to SL(C5) following ATLAS confirmation.

Summary

- Should space exhaustion warrant an alarm ticket to save production time.
- This months computer centre infrastructure failure was air conditioning at RAL.
- LHCOPN monitoring/alarms/information flow for the sites and experiments needs improvement.
- Serious zero-day exploit on SL4 and 5 diverted a large amount of expertise.
- Universal migration to SL(C)5 worker nodes requested.