WLCG Service Report

Jamie.Shiers@cern.ch

~ ~ ~

WLCG Management Board, 7th April 2009

Introduction

- This report is primarily about the service since last week's MB, but I would like to start with a brief summary of the quarter based on the draft contribution to the LCG QR
- Although this is not an exhaustive list of all serious problems in the quarter, we can still draw some conclusions from it:

Site	When	Issue
CERN	08/01	Many user jobs killed on lxbatch due to memory problems
CERN	17/01	FTS transfer problems for ATLAS
CERN	23/01	FTS / SRM / CASTOR transfer problems for ATLAS
CERN	26/01	Backwards incompatible change on SRM affected ATLAS / LHCb
CERN	27/02	Accidental deletion of RAID volumes in C2PUBLIC
CERN	04/03	General CASTOR outage for 3 hours
CERN	14/03	CASTOR ATLAS outage for 12 hours
CNAF	21/02	Network outage to Tier2s and some Tier1s
FZK	24/01	FTS & LFC down for 3 days
ASGC	**25/02**	**Fire affecting all site – services temporarily relocated**
RAL	24/03	Site down after power glitches. Knock-on effects for several days

QR – Conclusions (1/2)

- Not all sites are yet reporting problems consistently – some appear 'only' in broadcast messages which makes it very hard to track and (IMHO) impossible to learn

➢ **If you don't learn you are destined to repeat**

- e.g. from a single joint operations meeting (Indico 55792) - here

SARA: OUTAGE: From 02:00 4 April to 02:00 5 April. Service: dCache SE.
SARA: OUTAGE: From 09:30 30 March to 21:00 30 March. Service: srm.grid.sara.nl.
SARA: OUTAGE: From 15:13 27 March to 02:00 31 March. Service: celisa.grid.sara.nl. Fileserver malfunction.
CERN: At Risk: From 11:00 31 March to 12:00 31 March. Service: VOMS (lcg-voms.cern.ch).
FZK: OUTAGE: From 14:21 30 March to 20:00 30 March. Service: fts-fzk.gridka.de
INFN-CNAF: OUTAGE: From 02:00 28 March to 19:00 3 April. Service: ENTIRE SITE.
INFN-T1: OUTAGE: From 16:00 27 March to 17:00 3 April. Service: ENTIRE SITE.
NDGF-T1: At risk: From 12:31 27 March to 16:31 30 March. Service: srm.ndgf.org (ATLAS).
NDGF-T1: At risk: From 12:31 27 March to 13:27 31 March. Service: ce01.titan.uio.no.

- **As per previous estimates, one site outage per month (Tier0+Tiers1) due to power and cooling is to be expected**

☝ **It is very important to find some track of these through the daily operations meetings and weekly summaries**

- We must improve on this in the current (STEP'09) quarter – all significant service / site problems need to be reported and some minimal analysis – as discussed at the WLCG Collaboration workshop – provided spontaneously

➢ **I believe that there should be some SERVICE METRICS – as well as experiment metrics – for STEP'09 which should reflect the above**

- **See GDB tomorrow – they are not new by the way!**

QR – Conclusions (2/2)

- CASTOR and other data management related issues at CERN are **still too high** – we need to monitor this closely and (IMHO again) pay particular attention to CASTOR-related interventions
  - DM PK called about once per week; mainly CASTOR; sometimes CASTOR DB*; more statistics needed but frequency is painfully high...

- ASGC fire – what are the lessons and implications for other sites? Do we need a more explicit (and tested) disaster recovery strategy for other sites and CERN in particular? [Status later]

- ☺ Otherwise the service is running smoothly and improvements can clearly be seen on timescales of months / quarters

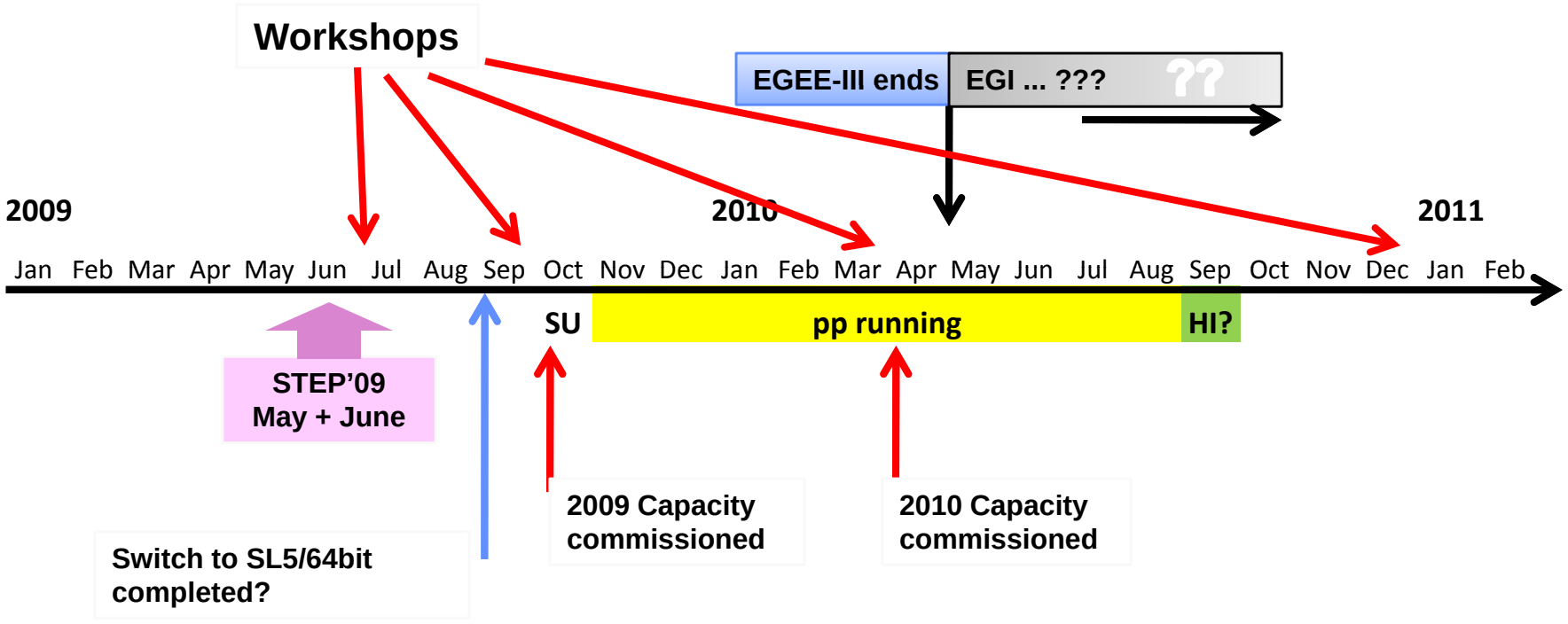
- The report from this quarter is significantly shorter than that for previous quarters – **this does not mean that all of the issues mentioned previously have gone away!**

HEPiX Questionnaire - Extract

4. Disaster recovery :
   a) How do you ensure your backup'ed files are recoverable?
      1. Do you have an off-site tape vault?
      2. Do you have a disaster recovery plan ?
      3. Did you verify it by a simulation?
   b) Have you got a total service hardware redundancy for main services?
      1. How many main services cannot be automatically switched?
      2. If possible, cite some of them
   c) How is staff availability addressed to cope with a problem:
      - Obligation? Money stimulus? Vacation stimulus? Willingness? Other?
5. Use Case 1 : fire problem in the computing room
   a) Is there any special automatic fire or smoke detection mechanism in the room?
   b) Is there any special automatic fire extinction mechanism? Which one?
   c) How much time before fire department arrives on scene?
   d) Do you have an evacuation plan at your disposal ?
   e) Did you simulate such a plan? which results ?

Proposal: all WLCG T0+T1 sites to complete!

WLCG timeline 2009-2010



A pre-CHEP 2010 (17-22 Oct) workshop probably does not make sense, nor does a more restricted A-P event (IMHOA)

GGUS Summaries

VO concerned	USER	TEAM	ALARM	TOTAL
ALICE	2	2	**7**	11
ATLAS	20	14	**11**	45
CMS	43	0	**8**	11
LHCb	10	1	**9**	20
Totals	35	17	**35**	87

Alarm testing was done last week: the goal was that alarms were issued and analysis was complete well in advance of these meetings!

- **From Daniele's summary of the CMS results:**
  - ***"In general, overall results are very satisfactory, and in [the] 2nd round the reaction times and the appropriateness of the replies were even more prompt than in 1st round"***
- **Links to detailed reports are available on the agenda page – or use GGUS ticket search for gory details!**
- **Discussions during daily meetings revealed mismatch: DM PK & expectations: there is no piquet coverage for services other than CASTOR, SRM, FTS and LFC.**
  (For all other services, the alarm tickets will be treated as best effort*)

RAL-ATLAS Example

- Public Diary:
  Dear ggusmail:

  The alarms email you sent was recognised as being an urgent request from an accredited VO. Your report has been escalated appropriately, but do please keep in mind that there is no guaranteed response time associated with your alarm. However, we will do our best to address the issue you reported as soon as possible.

  Your mail has been forwarded to the alarms list and escalated in our Nagios support systems.

  Regards,
  The RAL-LCG2 alarms response team

DE-T1 ATLAS Example

- **Description:** ALARM SYSTEM TEST: Fake ticket use case--> LFC down at your site

  Detailed description:
  Dear T1 site,

  this is another *TEST* GGUS Alarm ticket, as required by WLCG to all ATLAS T1s.
  This is specifically a test executed by ATLAS.

  For this test, ATLAS expert thinks that your LFC is down.

  Official instructions: sites please should "follow the procedure foreseen if this were a true alarm. The site will have to close the GGUS ticket confirming they understand what they would have done, had this been a true alarm".

  Thanks in advance

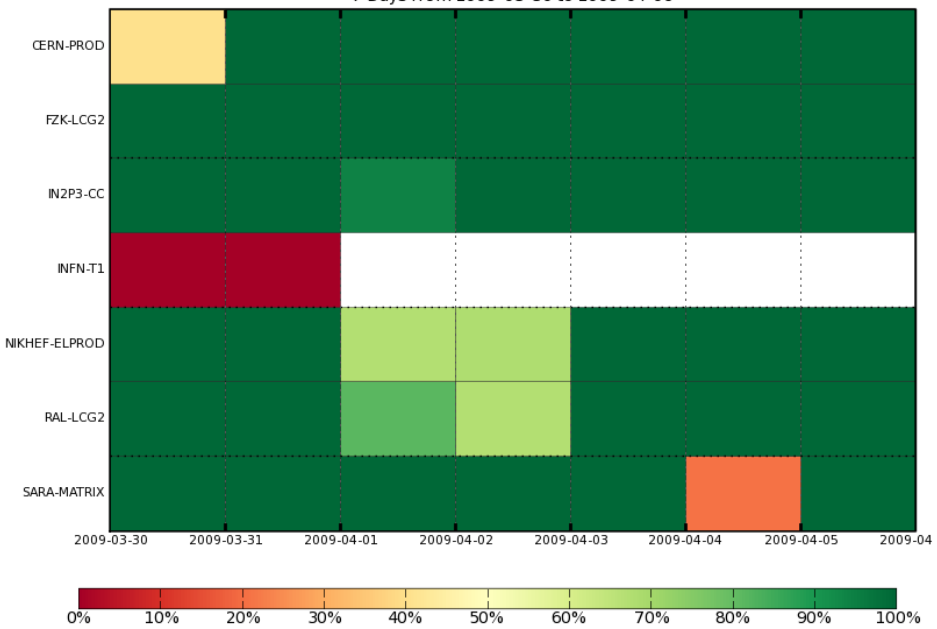
  Stephane **Solution:** Dear ATLAS-Team,

  the communication between the database backend and the LFC frontends have been tested and are working fine. All daemons are running and querying all frontends shows no errors.

  With kind regards,
  Angela
  This solution has been verified by the submitter.

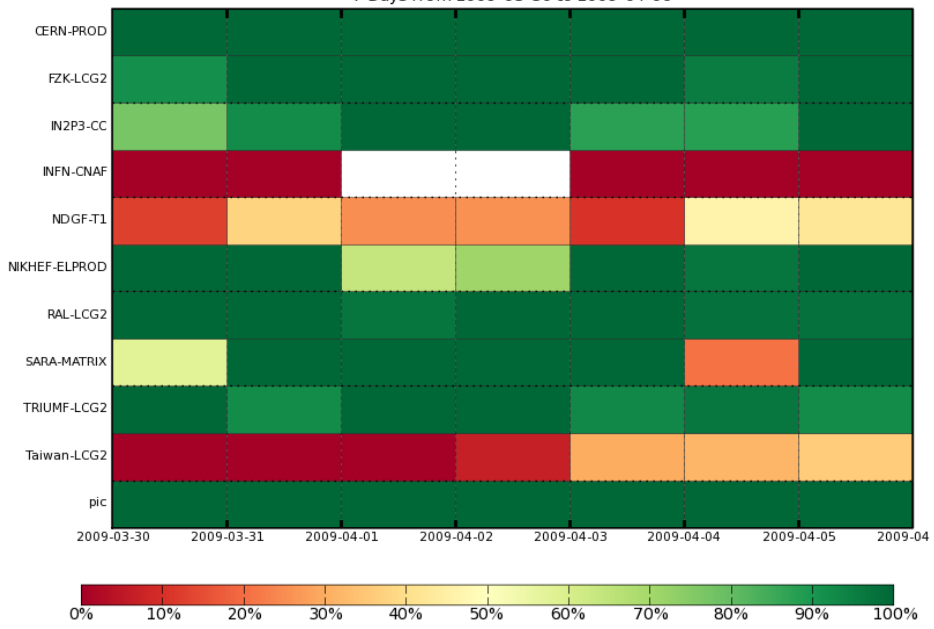
ite Availability using WLCG Availability (FCR critical

7 Days from 2009-03-30 to 2009-04-06



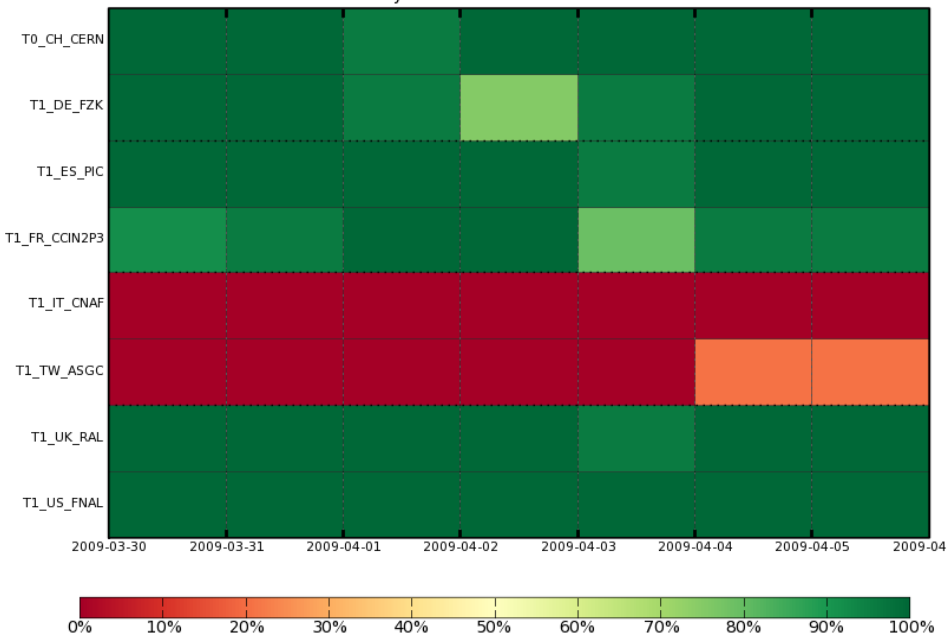
Site Availability using WLCG_SRM2

7 Days from 2009-03-30 to 2009-04-06



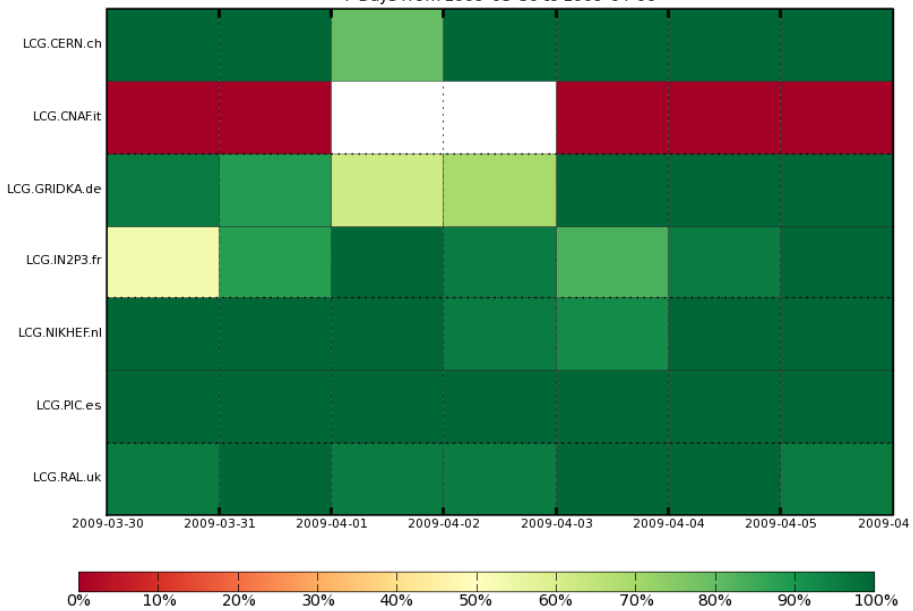
Site Availability

7 Days from 2009-03-30 to 2009-04-06



Site Availability using LHCb Critical Availability

7 Days from 2009-03-30 to 2009-04-06



ASGC Update

- [Jason] After relocating the facilities from ASGC DC to IDC, we took another week to resume the power on trial before entering the IDC. also, the complex local management policy have delay the whole progress for another week.
- Now, all T1 services should have been restored.

- Full details are in the following slides.

Service	Date	Details
ASGC Network	two days after the fire incident (Feb 27)	(relocate into main bld. of AS campus where we have main fiber connectivity from all other service provider). service relocated into computer room of IoP are: VOMS, CA, GSTAT, DNS, mail, and list.
ASGC BDII & core service	first week after the incident (Mar 7)	services consider in first relocation including also LFC, FTS (DB, and web frontend), VOMRS, UI, and T1 DPM.
ASGC core services	relocate from IoP/4F to IDC at Mar 18th	we took around two weeks to resume power on trial outside data center area due to the concerning dusting in the facilities that might trigger VESDA system in the data center. majority of the system have been relocate into rack space at Mar 25.
T2 core services	Mar 29	this including CE/SEs while the expired crl and out dated ca release have cause instability of the SAM probes the first few days. we later integrate the T1/T2 pool that all submission will turn to same batch scheduler to help utilizing the resources of new quad core, x86-64 computing nodes.
...		**[Full report in slide notes...]**

Summary

- ✓ **GGUS alarm tick**
- ☹ **ALL sites should (MoU targets) ar targets in this a**
- 👍 **The hard work th visible & recogni talk about the ab than to confirm t**
- **[Maybe we can successes in a si**
- ➢ ***"The priority to r only in May/Jun scale-testing at***