Service Delivery & Operations

~~~

Tier0, Tier1 and Tier2 experiences from Operations, Site and Experiment viewpoints

Jamie.Shiers@cern.ch

~~~

LCG-LHCC Referees Meeting, 16th February 2010

Structure

- Recap of situation at the end of STEP'09
  - Referees meeting of July 6th 2009 + workshop
- Status at the time of EGEE'09 / September review
- Issues from first data taking: experiment reports at January 2010 GDB
- Priorities and targets for the next 6 months
- Documents & pointers attached to agenda – see also experiment reports this afternoon

The Bottom Line…

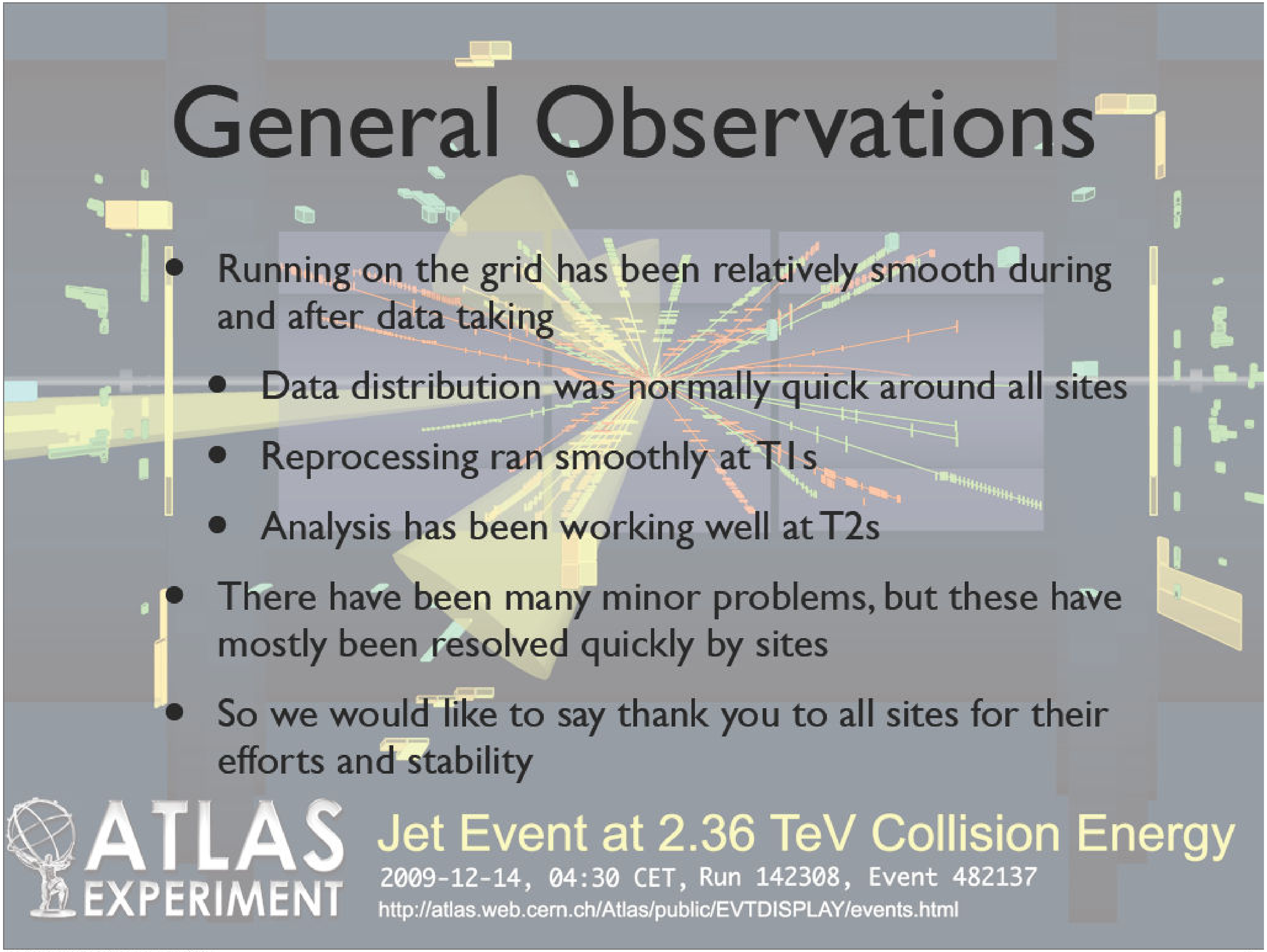
- From ATLAS’ presentation to January GDB

➢ **“The Grid worked… BUT”**

- There are a number of large “BUTs” and several / many smaller ones…

- Focus on the large ones here: smaller ones followed up on via WLCG Daily Operations meetings etc.

❑ **The first part of the message is important!**



General Observations

- Running on the grid has been relatively smooth during and after data taking
  - Data distribution was normally quick around all sites
  - Reprocessing ran smoothly at T1s
  - Analysis has been working well at T2s
- There have been many minor problems, but these have mostly been resolved quickly by sites
- So we would like to say thank you to all sites for their efforts and stability

ATLAS EXPERIMENT

Jet Event at 2.36 TeV Collision Energy

2009-12-14, 04:30 CET, Run 142308, Event 482137

http://atlas.web.cern.ch/Atlas/public/EVTDISPLAY/events.html

Wednesday, 13 January 2010

2

The Big Buts…

- Will be covered in more detail later, including major improvements in the associated areas in the past 6 months

Tier	Issue
0	Many critical – and sometimes unique – services run at the Tier0. Improvements in transparency in scheduling interventions is required. This is on-going – recently agreed pre-intervention "Risk Analysis" being put in place: hope to see measurable improvement by July.
1	There are concerns with the services at two Tier1s – one already flagged at the July review – that need further investigation and action. [But 2 of the 3 sites discussed at that time have since resolved their problems & re-testing has confirmed that these sites perform ok]
2	On-going concerns about data access as well as support models for end user analysis. Also issues around internal and external networking for these sites. [Good progress on Analysis stress tests in Q3/Q4]

What Were The Metrics?

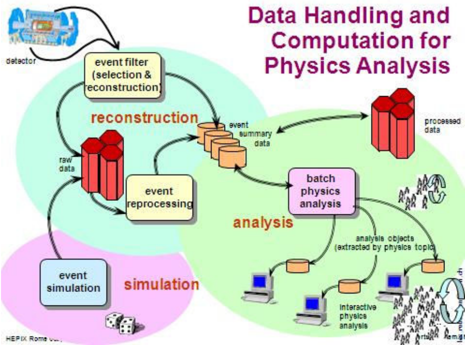
- Those set by the experiments: based on the main "functional blocks" that Tier1s and Tier2s support

- Primary (additional) Use Cases in STEP'09:

  1. (Concurrent) reprocessing at Tier1s – including recall from tape
  2. Analysis – primarily at Tier2s (except LHCb)

- In addition, we set a **single** service / operations site metric, primarily aimed at the Tier1s (and Tier0)

- Details:
  - ATLAS (logbook, p-m w/s), CMS (p-m), blogs
  - Daily minutes: week1, week2
  - WLCG Post-mortem workshop


Data Handling and Computation for Physics Analysis
detector
event filter (selection & reconstruction)
reconstruction
event summary data
processed data
raw data
event reprocessing
batch physics analysis
analysis
analysis objects (extracted by physics topic)
event simulation
simulation
interactive physics analysis
HEPIX Rome

STEP'09: What Were The Results?

☺ **The good news first:**

- ✓ **Most** Tier1s and **many** of the Tier2s met – and in some cases exceeded by a significant margin – the targets that were set

- In addition, this was done with **reasonable** operational load at the site level **and** with quite a high background of scheduled and unscheduled interventions and other problems – including 5 simultaneous LHC OPN fibre cuts!

➢ **Operationally, things went really rather well**

- Experiment operations – particularly ATLAS – overloaded
- ☺ **This has since been corrected – ATLAS now have a rota for this activity**

☹ **The not-so-good news:**

- Some **Tier1s** and Tier2s did not meet one or more of the targets

STEP '09: Tier1s: "not-so-good"

- Of the Tier1s that did not meet the metrics, need to consider (alphabetically) ASGC, DE-KIT and NL-T1
- In terms of priority (i.e. what these sites deliver to the experiments), the order is probably DE-KIT, NL-T1, ASGC

➢ **Discussions were held with KIT, formal reviews with NL-T1 and ASGC**

☺ **The situation with both KIT and NL-T1 has improved significantly: the issues with these sites can now be considered resolved.**

➢ **RAL suffered a period of major instability – much of which can be attributed to the machine room move – and a formal review, organized by GridPP, was held in December 2009 [important lessons here.]**

➢ **The situation with ASGC continues to be critical: here too a major fire had significant consequences but staffing and communication remain**

☛ **In depth independent analysis of these two site issues is required: review material and SIRs important input but not sufficient**

ASGC

- ASGC suffered a fire in Q1 which had a major impact on the site
- They made very impressive efforts to recover as quickly as possible, including relocating to a temporary centre

➢ **They did not pass the metric(s) for a number of reasons**

- It is clearly important to understand these in detail and retest once they have relocated back (on-going)

💣 **But there have been and continue to be major concerns and problems with this site which pre-date the fire by many months**

- The man-power situation appears to be sub-critical
- Communication has been and continues to be a major problem – despite improvements including local participation in the daily operations meeting

☟ **Other sites that are roughly equidistant from CERN (TRIUMF, Tokyo) do not suffer from these problems**

Site Problems: Follow-up

- Site reviews were proposed as a mechanism for following up on major issues at a previous LHCC review

➢ **These should be triggered (by the MB?) when there is a major problem lasting weeks or more**

- **As an addition to the previous proposal, the review “panel” could / should be responsible for follow-up on the recommendations for a period of 1-2 quarters**

☝ **Some major site problems have been triggered by major machine room moves: we should be aware of this in the case of future upgrades / moves which are inevitable over the lifetime of the LHC**

Site Problems: Root Causes?

- It is not clear that the real root causes behind e.g. the site problems at ASGC and RAL have been fully identified
- There may well be a number of contributing factors – one of which is likely related to **service complexity**
- The news from CNAF regarding their migration away from CASTOR as well as their experience in the coming 6 months will be extremely valuable input into a potential "Site Storage Review" that could be a major theme of the July 2010 WLCG Workshop
  - Commercial solutions (DMF, Lachman(?), HPSS, TSM) are used for the "tape layer" at many Tier1/2 sites
  - Simplification / lower cost of ownership is an important factor for all!
- **"The 5 whys" – we must drill down until we fully understand the root causes…**

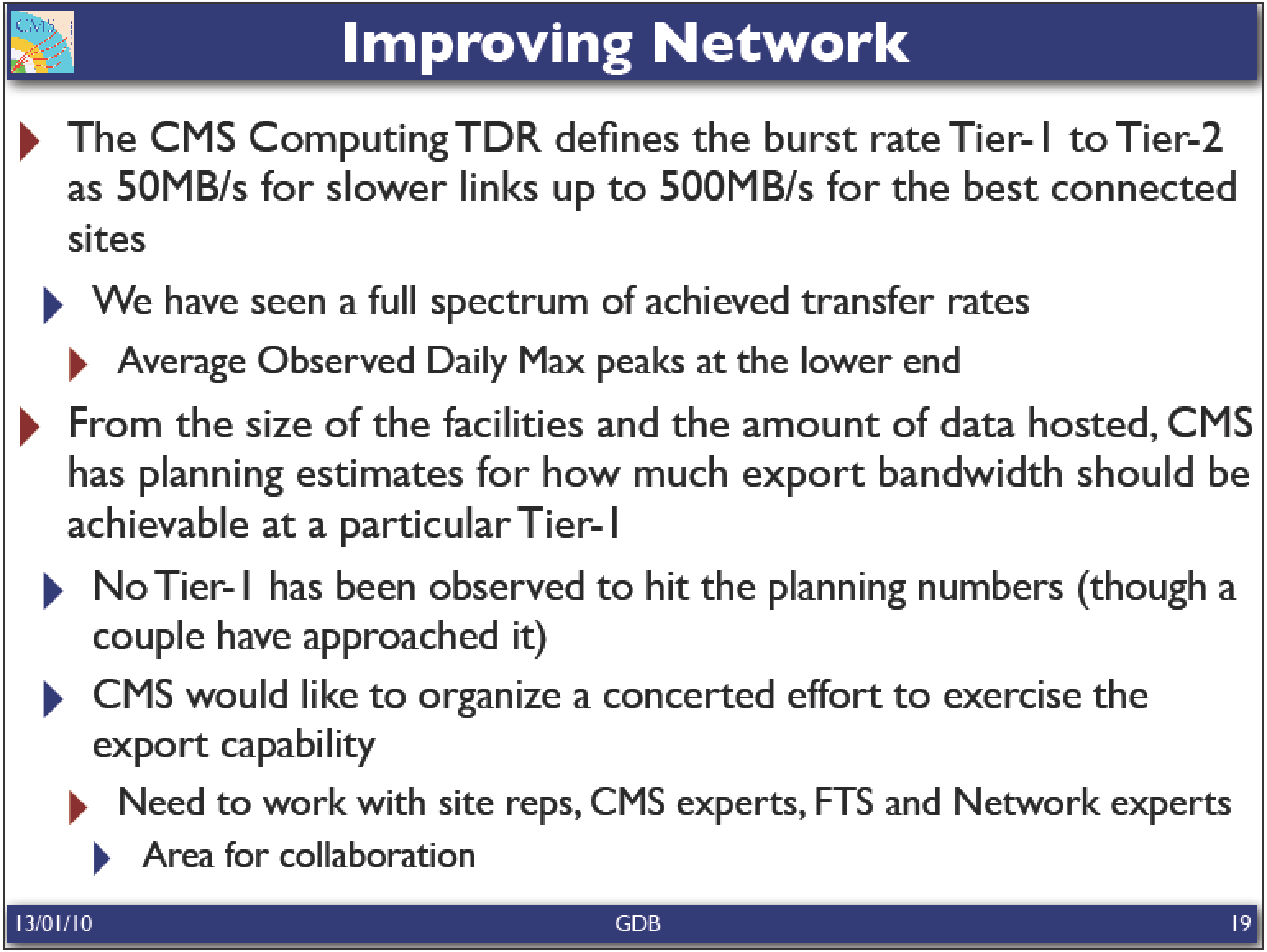
Outstanding Issues & Concerns @ EGEE '09

Issue	Concern
Network	T0 – T1 well able to handle traffic that can be expected from normal data taking with plenty of headroom for recovery. Redundancy?? T1 – T1 traffic – less predictable (driven by re-processing) – actually dominates. Concerns about use of largely star network for this purpose. Tn – T2 traffic – likely to become a problem, as well internal T2 bandwidth
Storage	We still do not have our storage systems under control. Significant updates to both CASTOR and dCache have been recommended by providers post-STEP'09. Upgrade paths unclear, untested or both.
Data	Data access – particularly "chaotic" access patterns typical of analysis can be expected to cause problems – many sites configured for capacity, not optimized for many concurrent streams, random access etc.
Users	Are we really ready to handle a significant increase in the number of (blissfully) grid-unaware users?

These statements were to stimulate discussion (which they did…)

Outstanding Issues - Progress

- Network: work going on in the LHC OPN community to address topology, backup links, T1-T1 and T1-T2 connections; strong interest from CMS in particular (ATLAS too?) in addressing network issues (next)
- Storage: **significant progress** in addressing stability issues in recent months seen in dCache – migrations to Chimera have been performed successfully: **this is a major improvement and should be acknowledged!**
- Improvements in the scheduling and execution of CASTOR+SRM have been **requested** – e.g. "Risk Analyses" to help in scheduling of interventions were discussed at the January GDB: we should review this in 3 – 6 months
- Data access: **still an issue** – a "Technical Forum" working group has been proposed in this area



Improving Network

- The CMS Computing TDR defines the burst rate Tier-1 to Tier-2 as 50MB/s for slower links up to 500MB/s for the best connected sites
  - We have seen a full spectrum of achieved transfer rates
    - Average Observed Daily Max peaks at the lower end
- From the size of the facilities and the amount of data hosted, CMS has planning estimates for how much export bandwidth should be achievable at a particular Tier-1
  - No Tier-1 has been observed to hit the planning numbers (though a couple have approached it)
  - CMS would like to organize a concerted effort to exercise the export capability
    - Need to work with site reps, CMS experts, FTS and Network experts
      - Area for collaboration

13/01/10 GDB 19

Tier2s

- Tier2 issues are now covered regularly at WLCG Daily Operations meetings: the main issues and tickets are reported by the experiments in their pre-meeting reports: the number of tickets is low & their resolution usually sufficiently prompt (or escalated…)
  - The calls are open but it is not expected that Tier2s routinely participate **[although Tier0 + Tier1s should and largely do!]**
  - The current activity is low – the number of issues will no doubt increase during data taking

Some of the key issues seen by the experiments are covered in the next slides

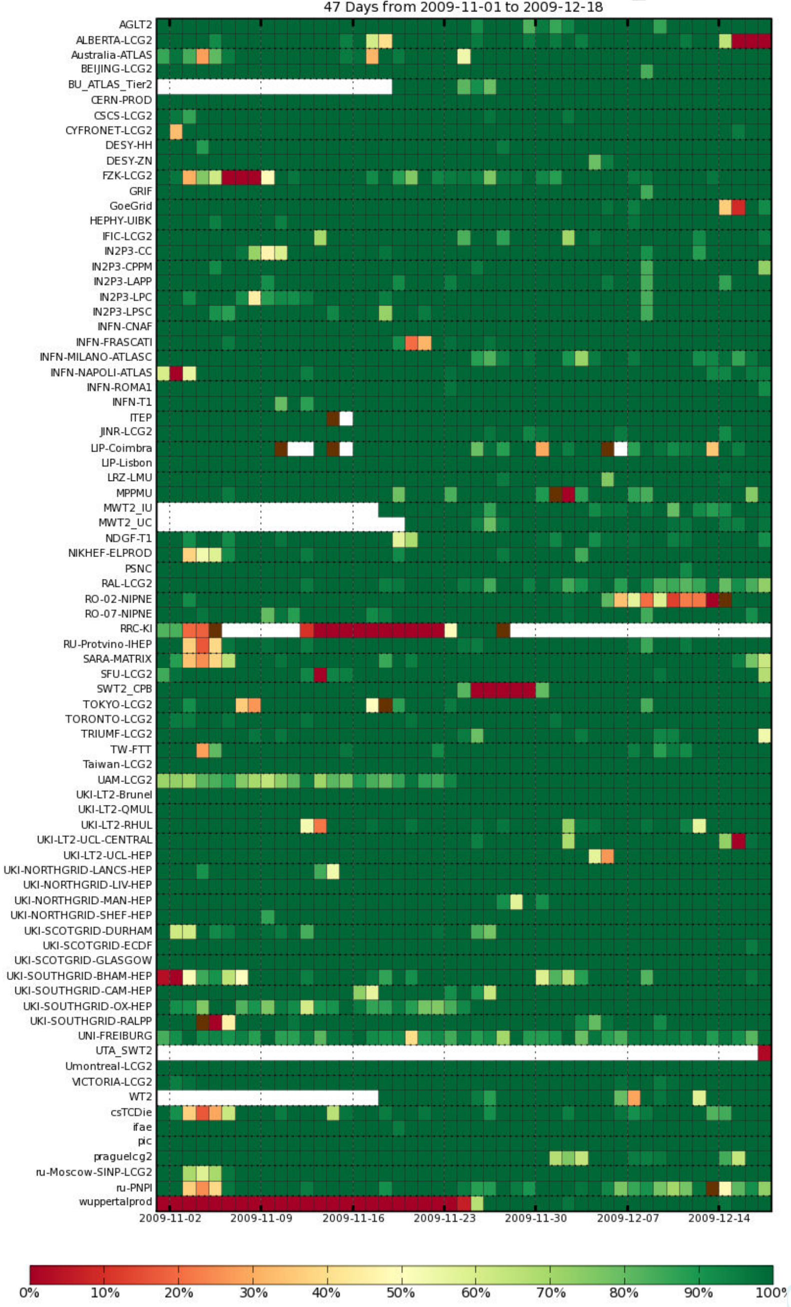
- The WLCG Collaboration workshop in July is foreseen to be held at the Tier2 at Imperial College in London: Tier2 involvement & issues will be a key element of this and indeed all such workshops

Tier2 Status

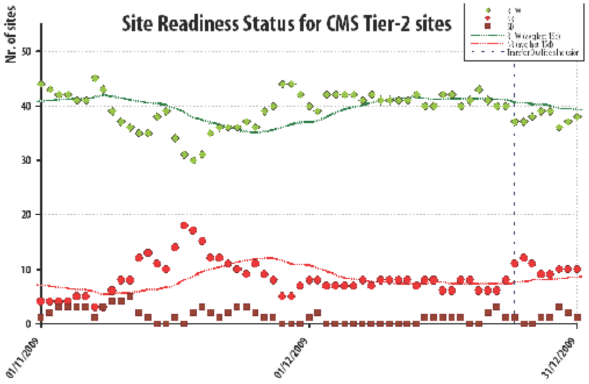
Experiment	Issue
ALICE	Just 2 Tier2s blacklisted as not running SL5 WNs & 2 Tier2 sites not yet running gLite 3.2 VO boxes
ATLAS	“Analysis has been working well at T2s”; storage reliability an on-going problem
CMS	1 Tier-1 and 10 Tier-2s that had to update to the latest release FroNTier/Squid release at time of January GDB; site availability has stabilized a lot since October
LHCb	Shared area issue: just looking at the last 3 months GGUS tickets, out of 170 tickets, ~70 were open against sites with problems with shared area (permission, accessibility, instability)

Site Reliability using Storages_SRMv2

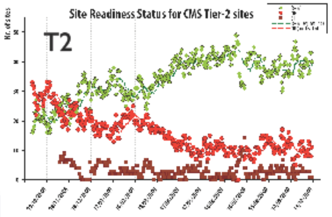
47 Days from 2009-11-01 to 2009-12-18



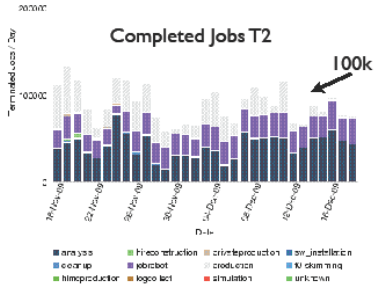
Tier-2 Readiness

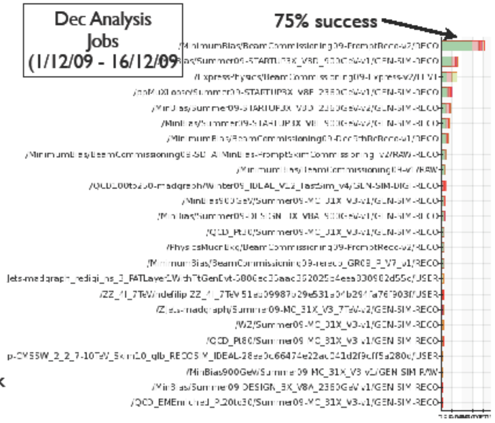


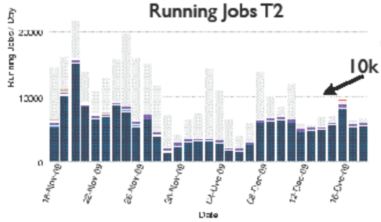
Looking back to Oct., Tier-2s have stabilized



Access at Tier-2s







Recommendations

1. Introduce **Risk Analyses** as part of decision making process / scheduling of interventions (Tier0 and Tier1s): monitor progress in next 6 months

2. Site visits by review panel **with follow-up** and further reviews 3-6 months later

3. Prepare for in-depth site **storage review**: understand motivation for migrations (e.g. CNAF, PIC) and lessons

4. Data access & User support: we need clear **targets** and **metrics** in these areas

Overall Conclusions

- The main issues outstanding at the end of STEP '09 have been successfully addressed
- Some site problems still exist: need to fully understand root causes and address at WLCG level
- Quarterly experiment operations reports to the GDB are a good way of setting targets and priorities for the coming 3 – 6 months

- "The Grid worked" AND we have a clear list of prioritized actions for addressing outstanding concerns