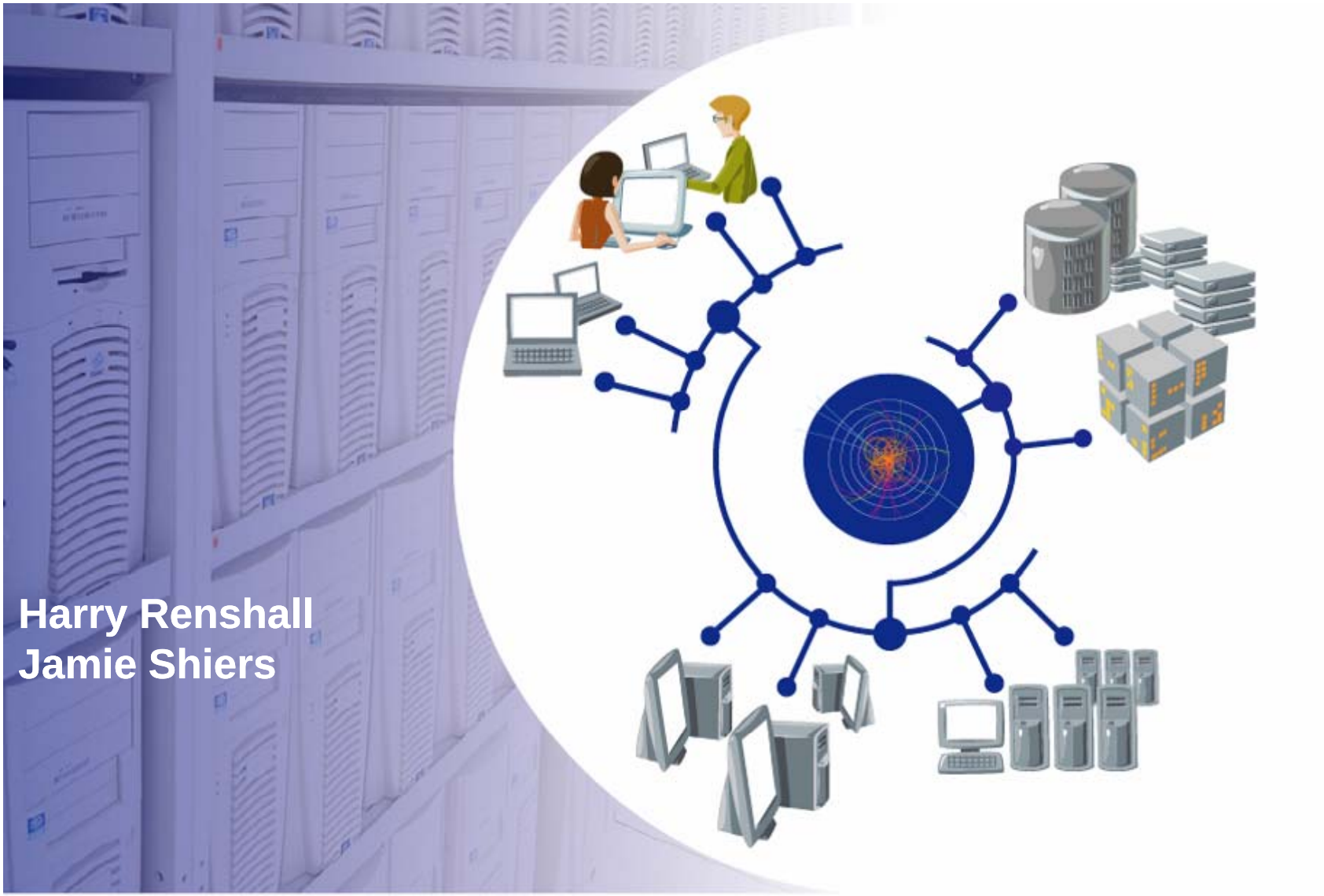
Summary of the WLCG Collaboration workshop Prague 21-22 March 2009



Harry Renshall
Jamie Shiers

WLCG Collaboration Workshop (Tier0/Tier1/Tier2)
21-22 March Prague

Third workshop in the pre-Chep series (after Mumbai and Victoria) and had a strong WLCG operations focus. A lot of ground was covered.

The opening remarks by Jamie Shiers were that the most important point is that it was a **WORKSHOP** and not a conference. Speakers were asked to please leave time for questions and discussions as being essential for the event's **success** and this summary reflects those discussions (with my personal judgement of what to bring out).

Another opening remark was that the workshop should: Basically discuss, clarify, agree how we will perform "WLCG Operations" for the 2009 – 2010 data taking run

□□ **Hint: it will probably look quite a bit like what we do now!**

228 participants: 44 Tier 0, 49 Tier 1, 124 Tier 2, 10 Tier 3, 1 Other (taken from registration Institute affiliation - some guesswork).
This mixture worked well and increased the understanding of operational issues between the Tier levels and with the experiments.

Main Session Themes

WLCG and Experiment roadmaps 2009/2010

WLCG reviews:

- Sites
- Database Services
- Data Management Services and Outlook

Analysis Services:

- User Analysis working group
- Supporting Analysis Services and Users

WLCG Operations

- Procedures and Tools, Incident Response
- Monitoring
- Support issues, ticket handling

WLCG and Experiment Roadmaps 2009/2010 (1/6)

Session started with WLCG roadmap from Ian Bird, LCG Project leader then one from each experiment. Summary of discussions at end.

WLCG Roadmap: Simple – we have to take data sometime but there are a few rocks on the way.

Current LHC schedule is a 44 week run with a 2 week Xmas stoppage and, with a low machine efficiency, integrating to 6.1*10**6 seconds of effective beam time. Planning heavy ion run in last few weeks.

Issues include the effect on the experiment models of no long winter shutdown, middleware upgrade plans (CREAM-CE etc.), a need for a CCRC'09 as we have not seen all experiments testing together tape recall and processing (and while writing raw data to tape), and the EGEE3 to EGI transition.

WLCG and Experiment Roadmaps 2009/2010 (2/6)

ALICE:

No major issues in registration/replication of raw data
Confidence in the storage and Grid tools is high
- Middleware and computing centres storage (T0 and T1s) have been fully certified

The RAW registration at T0 will be tested at full p+p and Pb+Pb rates (up to 1.5GB/sec) in May-June 2009
Cosmics: Resume data taking in July 2009, ~300TB of RAW p+p runs
Running a few days @ 0.9 GeV (October 2009), 11 months @ 10 TeV
Address the ALICE p+p physics program and provide baseline measurements for AA
A+A run: Fall 2010 - a standard period of Pb+Pb running
- Computing resources **must be sufficient** to process these data within 4 months after data taking as per the computing model

Monte Carlo: 2009-2010 are standard years for Monte Carlo production

WLCG and Experiment Roadmaps 2009/2010 (3/6)

ATLAS:

- Cosmics data taken in August-November 2008
- Few runs of single-beam events the week of Sept 10th

All RAW data were processed at Tier-0 → ESD, AOD, DPD, TAG
Distributed according to the Computing Model
Good data selected (about 35%) and re-processed except at 2 Tier 1
To be repeated in the summer with all Tier 1 participating and will physically remove input data from disk.

MC09 will soon start (wait for final release, final geometry)
Using new Geant4 release, produce samples for 10 TeV and cosmics running
Now at 1000 KSi2K seconds/event – more than expected
Nearly 20% of wall time is wasted on failed jobs
All clouds will be served from Oracle based Panda servers at CERN

End May/beginning June there will be a timing window when detector commissioning need for Tier 0 facilities stops. Opportunity for a CCRC.

WLCG and Experiment Roadmaps 2009/2010 (4/6)

CMS:

Data processing of CMS Global Runs, CRAFT is working well.
– Data were re-reconstructed twice with latest software and calibrations.
• Monte Carlo production at Tier2 sites is well established.
• Improvements of the Tier-1/Tier2 infrastructure are urgently required to reach production availability and reliability
– Better monitoring tools and commissioning policies are available,
– Monitoring will also be done by computing shifts…
– Stress testing Tier-1 sites has started
• Combined tests with ATLAS need to be defined and scheduled.
• Resource requirements for 2009/10 are being re-evaluated based on LHC schedule

WLCG and Experiment Roadmaps 2009/2010 (5/6)

LHCb:

2008 CCRC very useful for LHCb (although irrelevant to be simultaneous due to their low throughput)

DIRAC3 fully commissioned

- Production in July
- Analysis in November
- As of now, called DIRAC

Last processing on DC06

- Analysis will continue in 2009

Commission simulation and reconstruction for real data 2009-10

Large simulation requests for replacing DC06, preparing 2009-10

Full Experiment Services Test'09: ~1 week a month and 1 day a week

Resource requirements being prepared for C-RRB in April

Services are not stable enough yet!

WLCG and Experiment Roadmaps 2009/2010 (6/6)

Questions/Discussion points:

EGEE services expected to be covered by EGI include GGUS and ROCs. Most of the needed middleware is already supported by HEP sites.

The Xmas 2 week LHC break will not reduce computing activities.

Changes are inevitable during the 44 week run so we will have to learn how to test and make them while in production mode. Experiments will have to cope with some long Tier 1 downtimes.

With only one third of previous effective beam time 2009/10 what computing resources will be needed ? All experiments are recalculating.

Number of cores increasing faster than memory. Intelligent configuration of virtual/physical memory limits on worker nodes is needed.

WLCG Site Reviews (1/3)

Site Reviews – Status and Outlook (1/3)

Problems include:

- Site/service unreliability and unavailabilty
- Frequency of major incidents and consistent follow up
- Problems in procuring/commissioning resources to the pledge level

Serious incidents in the last 6 months:

- Castor – ASGC, CERN, CNAF, RAL
- dCache – FZK, IN2P3, NL-T1
- Oracle – ASGC, RAL
- Power – ASGC, PIC, NL-T1, CNAF
- Cooling- CERN, IN2P3
- Network- CNAF, PIC, BNL, CERN
- Other – CNAF, RAL, NL-T1,
- Fire – ASGC

Tier1s **will** be down. Experiment models should cope.

WLCG Site Reviews (2/3)

Site Reviews – Status and Outlook (2/3)

Simple actions to take

- Ensure sites have sufficient local monitoring; including now the grid service tests/results from SAM and experiments
- Ensure the response to alarms/tickets works and is appropriate – and test it
- Follow up on Site Incident Reports – does your site potentially have the same problem ?
- If you have a problem be honest about what went wrong – so everyone can learn
- Run workshops to share experience and knowledge on how to run reliable/fault tolerant services

Open Questions: Is communication adequate, is staffing of services adequate (e.g. full time DBA and MSS support where needed).

WLCG Site Reviews (3/3)

Site Reviews – Status and Outlook (3/3)

Discussion:
SAM framework is changing (Nagios based) – we should exploit this.
Tier 1 have very different availabilities.
LHCb SAM tests show strong correlation between their reports and real problems
Sites should detect problems before the experiments do – construction of the middleware often does not help here.
Experiments should tell sites when they are not of production quality.
Many of the VO tests are only understood by the experiments and experiment site reps usually have no backup.
CMS not seeing problems where sites are not paying attention but more systematic issues that crop up from time to time.
WLCG will make site visits including some Tier 2.

WLCG Database Services Reviews (1/2)

Examples and Current High-Priority Issues

Atlas, Streams and LogMiner crash 12-12-2008: Workaround no good
PVSS replication down 4 days, Conditions replication down 4 days
CMS Frontier, change notification and Streams incompatibility
Intermittent streams capture process aborts – patch did not solve
Castor, BigID issue – Wrong bind value being inserted (hit RAL and ASGC badly)
Castor, ORA-600 crash – ATLAS stager down 10 hours.
Solved with patch but needs merge of two existing Oracle patches.
Castor, crosstalk and wrong SQL executed
RAL reported an incident causing loss of 14K files – fixed
Oracle clients on SLC5, SELINUX connectivity problem
Cannot use multiple nodes to restore a VLDB – 1 Day for 10TB

WLCG Database Services Reviews (2/2)

Follow up Actions:

Following proposals to the WLCG MB and elsewhere WLCG Oracle Operational review meetings are being established.

Such meetings will initially take place quarterly.

Input from regular WLCG Distributed Database Operations (formerly-3D) conference calls and workshops will be essential in preparing and prioritizing lists of issues, as well as to exchange knowledge on workarounds and solutions.

WLCG Data Management Reviews and Outlook (1/5)

Summary of FZK Dcache Workshop 14-15 January 2009 (1/2)

Goals were to: Discuss mutual issues, practices, procedures
Improve communication
Meet/Unite administrators
Absence of experiments reps. at the workshop allowed more depth of detailed discussions (though they were invited).

The presentation reviewed sources of storage (dcache) instability:

Complexity: not solved
Increasing resource footprint: not solved
Databases in the line of fire: not solved
Asynchronous operations: that's gridlife
and drew the analogy: the Doctor is not in, dCache.org has pills & bandages, Call the nurse

WLCG Data Management Reviews and Outlook (2/5)

Summary of FZK Dcache Workshop 14-15 January 2009 (2/2)

The final points were on sustainability pointing out that:

dCache is developed by a small group

Documentation is lacking for problems T1 sites are confronted with.

Do T1s need/have a plan B?

This lead to a lively discussion with the speaker being challenged as to how the user community was going to solve their difficulties with dcache support. This brought out several actions/recommendations that had been decided at the workshop as can be seen from their minutes.

Sites can look at the source code (Java) and T1 Admins should be able to understand it at entry level. Dcache.org will organise master classes to help (e.g. on Chimera setup) and setup a Twiki where T1 sites can enter Setup files, How-to's, FAQ etc. to improve knowledge consolidation and exchange.
Sites would do whatever they could to reinforce the dcache team and the community understood it had to be more self supporting.
There will be a follow-up workshop.

WLCG Data Management Reviews and Outlook (3/5)

Summary of CASTOR External operations meeting: RAL 18-19 Feb 2009 (1/2)

Four site reports – all wanting to improve resilience and monitoring. RAL and ASGC hit by the same fatal “bigID” ORACLE bug. ASGC suffered 8 weeks of instability with corruptions difficult to recover in their uncommon configuration. CNAF only use CASTOR for T1D0 data. CERN going from 3 to 2 FTE in operations support.

Developers reported rich set of enhancements in 5 project areas:

- Tape efficiency – file aggregation on tape, less tape marks, user priorities
- File access protocols and latency – support analysis with small files and many concurrent streams. For low open latency use xrootd protocol and i/o server with CASTOR specific extensions.
- Security – every user authenticated, every action logged.
- SRM and database schema – merge the SRM and stager software & DBs
- Monitoring – key indicators, alarms before performance limits reached

WLCG Data Management Reviews and Outlook (4/5)

Summary of CASTOR External operations meeting: RAL 18-19 Feb 2009 (2/2)

Developers also looking further ahead to be ready for new challenges :

New mountable storage, safety of disk only data, 10 GigE interfaces, iscsi storage, solid state disks, distributed file systems

Developers recommend CASTOR 2.1.8 for LHC startup. Preparing 2.1.9 with an improved nameserver, further xrootd integration and new build infrastructure.

Significant effort is being put on development of monitoring integration tools, tape efficiency and core software.

Discussions/comments:

Would be interesting for experiments to attend these meetings to learn how to better use CASTOR and developers would learn experiment use cases.

Is T0D2 support planned (double disk copy) – Yes, in 2.1.8

ATLAS did not see much for them in 2.1.8 and pointed out the upgrade to 2.1.7 was painful. LHCb want 2.1.8 for the xrootd enhancements for T1 analysis.

WLCG Data Management Reviews and Outlook (5/5)

Experiment Data Management Requests for 2009

Overall Conclusions were:

- Focus must be on stability
- Improve SRM scalability
- More support for xrootd
- More coordination among data management client and server developers
- Prestaging tests a must and must be concurrently with more VOs
- Still unclear if sites and computing systems are ready for the user data analysis - extensive tests are foreseen
- Authorization on storage systems is still too crude

Analysis Services (1/4)

User Analysis Working Group Update – Markus Schultz (chairman) (1/3)

Understand and document analysis models - ideally in a way that allows to compare them easily
Answer the question how unpredictable Chaotic User Analysis will be
Identify additional requirements coming from the analysis use cases

Received several summary write ups and comments to questions
And roughly 100 reference documents and presentations

All experiments have well established analysis frameworks
Communication channels between experiments and T2s are well organized
Storage at T2/T3 resembles a Zoo
Large experiments each have > 1000 users who used the grid during 2008
Users per week on the order of 100-200/experiment
Expected increase during summer: Factor 2-3
Current resource usage: 15 – 30 % of the production use (30k jobs/day)

Analysis Services (2/4)

User Analysis Working Group Update (2/3)

Issues in access to storage resources have three main problem domains
- I/O performance (including network)
- Access Control
- SRM commands

Documentation of analysis models seems to be not an urgent task

Detailed resource allocation at T2s are best coordinated by experiments directly

We should concentrate on the following issues:
- I/O for analysis, Understanding the different access methods
- How can we help the T2s to optimize their systems?
- Can the ALICE analysis train model help?
- Role of xrootd as an access protocol?
- Clarification of requirements concerning storage ACLs and quotas

Analysis Services (3/4)

User Analysis Working Group Update: Discussion (3/3)

For very large analysis centres (6000 cores) would have to partition storage
Clear requirements on fabric bandwidth to worker nodes are needed
Batch system priorities should be dynamically adjusted
Number of analysis jobs per day is not the good metric
ALICE emphasising need for site quality of service (performance and stability) to support a high level of analysis
No mention was made of the danger of overloads leading to abrupt degradation – a message to all middleware developers.
Sites need to cope with both pilot jobs where experiments control the fair shares and local jobs needing local fair shares.
All experiments use root – important to structure root trees properly. Involve the root team and experiment data modellers in the working group.
Do we have to optimise all remote access methods – can also copy to local disk
Sites would prefer to support a single access protocol

This was probably the hottest topic.

Analysis Services (4/4)

Supporting Analysis Services and Users

Review of experiment site services – the VO-box world
Analysis support on the grid(s) is effectively a catchall for many users problems:

- Is that site down? Is there a problem with my code? Are my experiment analysis services down?

So far steady growth in analysis usage - problem resolution time will be critical.
Use of Ganga within ATLAS and LHCb was reviewed – backend independent analysis job submission. Support through tutorials, help forum and validation (Ganga job robot will disable failing sites).
ATLAS find shared gmail account useful for users.
ATLAS use Ganga to run Hammercloud for large automated stress tests – it does what its name says. Proving a very valuable tool.

WLCG Operations (1/9)

Procedures and Tools – reporting and escalation of problems

SAM central monitoring moving to Nagios

Problem reporting should be in Global Grid User Support. Now includes:

- Alarm & team tickets
- Direct routing to sites
- LHC Optical Private Network now covered

Can now escalate problems yourself in GGUS and use daily ops meeting to decide if escalation needed to GDB or MB levels.

Downtimes and at-risk must be entered in GocDB- now has user defined notifications and automatic classification into un/scheduled

There are many issues from site configurations – keep VOid cards up to date (supported in the discussions that followed)

WLCG Operations (2/9)

Service Incident Reports (1/2)

Gave practical advice on handling incidents based on IT/FIO experiences

An 'Incident' is any event which is not part of the standard operation of the service and which causes, or may cause, an interruption or a reduction of the quality of the service.

Restore operations as quickly as possible and with minimal impact

-but tell everyone first

Users can help – look for announcements before complaining, propagate to internal news, do not contact service managers directly

Industry finds 80-90% of incidents result from changes

Must manage changes rather than just avoid them

WLCG Operations (3/9)

Service Incident Reports (2/2)

IT/FIO prefers change management to be periodically scheduled
- Aggregation of changes
- Freeze, test and certify
- Plan ahead with clear responsibilities and timelines and exercise it

Should sites implement a Change Management Board ?

Discussion: Need to balance depth of a report, too many details are not useful for the users

Cannot use a Change Management Board in emergency situations

A realistically sized pre-production system is essential

Please encourage all sites to produce Incident reports

WLCG Operations (4/9)

Site Actions on Major or Prolonged Problems

Currently EGEE central operations suspend sites (remove from the information system) :

- When a site is unresponsive to a ticket
- When a site is in never ending down time (>1 month)
- When a site poses a security threat

Aim is to increase Grid operational efficiency and reduce the experiment job failure rates.

However, ATLAS (and others) routinely set sites offline as part of their operations/site commissioning and ticket the site concerned.

In discussion it was agreed the suspension procedure needs a rethink and is anyway a regional grid issue rather than for WLCG.

WLCG Operations (5/9)

Downtime Handling

Downtimes can be scheduled, unscheduled or at risk.

EGEE site admins register in GOCDB, OSG use OIM and the CIC operations portal sends mail feeds.

Experiment actions:
ALICE: No specific action
ATLAS: There are defined shifters manual procedures
CMS: Site removed from CRAB. Site commissioning updated.
LHCb: Site blacklisted in DIRAC. To put it back in production requires manual intervention. To see scheduled downtimes use Google calendar.

The discussions thought the Google calendar a good idea and would like the possibility of sites to be flagged as ‘partially working’

WLCG Operations (6/9)

Monitoring: VO-views, Site views, Availability

Experiment dashboards are now quite mature but from the WLCG CCRC'08 postmortem workshop came a clear requirement for site views.

- Compare the experiment's view of the site contribution to the information the sites get from their own monitoring systems.
- Understand if their site is contributing to the VO activity as expected.
- A new tool based on GridMap for visualisation has been developed to provide an overall and detailed view of all the activities going on at a site, easy to use and requiring no knowledge of a particular experiment.
- For each activity and VO it provides drill down links to the VO information sources to facilitate problem solving.
- There is a GridMap for each site with a first level view of Activities for all the supported VO's and a second level of the VO's information.

WLCG Operations (7/9)

Automatic Resource and Usage Monitoring

The WLCG resource pledges must be validated.

- Are resources available comparable at T0, 1 and 2?
- To what level are these resources being utilized?
- The WLCG MB needs to know these commitments are being met on a month by month basis.

There is now a view in GridMap allowing regions and sites to get a good view of their measured capacity.

Compute resources are well defined (GLUE schema), information providers are done, validation well under way, reporting understood.

Storage is well defined, most providers are done, validation started, reporting is being developed.

WLCG Operations (8/9)

Support Issues : Ticket Usage, Incident Response

This was a general discussion session:

Sites need to understand activities and would like to have experiment links which reveal why a site is down for them.

Many displays rely on colour – support is needed for the colour blind.

FZK alarm systems are service related and send SMS directly to service managers. Most other sites use a first level filter. Ideas to put more intelligence in alarm tickets to allow parsing.

Sites cannot have a piquet (24 by 7 on call) for all services.

Important for an experiment is to have ticket reception acknowledged.

Working on a report generator analysing ticket response times.

Idea for a lightweight alarm ticket for next business day response.

WLCG Operations (9/9)

Service Coordinator on Duty (SCOD)

There has been a SCOD mailing list for the last year – used as a feed into the daily WLCG operations meeting.

Now is the time to step up WLCG operations a gear.

Encourage more attendance at the daily meeting – participants can upload their contributions or use the SCOD mailing list. Add some regular (weekly) specialised segments.

Extend the SCOD role to a small team on a rotational basis with 52 weeks per year coverage.

Start now and slowly expand the team – initially 1-2 people from the 4 CERN WLCG involved groups. Typically on duty for weeks not months.

Common reprocessing session – CCRC'09 ?

The recent LHCC review recommended a joint testing of tape recall and event processing at nominal rates and first ideas were explored.

ATLAS have a full program of exercising their computing model and suggested a 3 week period:
Week 1 June 1-5: Setting up all tests at lower rates
Week 2 June 8-12: Real CCRC09, run all tests at full rate
Week 3 June15-19: Contingency and reporting

CMS position is that another CCRC does not fit their program.
– We have to fulfill many CMS schedule constraints.
– CMS production has to have priority over any scale testing exercise.
– We cannot use all resources (human, hardware & services) for a "Challenge"
– We are working with sites to improve "Site-Readiness" with high priority NOW.

Common reprocessing session – CCRC'09 ?

For any combined VO activity CMS proposes STEP-09:

- Scale Testing for the Experiment Program at the WLCG 2009
- Possible Schedule: May/June, to finish by end June!

Discussion points:

IN2P3 already have a scheduled downtime for June week 1

A separate recall from tape exercise would exercise many components

We must try a full rate Tier 0 Raw data to Castor with all 4 experiments

Both CMS and ATLAS start detector runs in July

Approach as a series of stress tests – those ready could be run earlier giving time to fix something e.g. ATLAS packaged pre-staging

Networks are a common resource and we lack metrics. More than we have ever tested will be re-reconstruction during data taking.

Squeezing into 1 week increases likelihood of failure – maybe 2 weeks?

Conclusion: June is the most realistic period – to be continued…..

Major Conclusions

Strong requirements for **STABILITY**

Prepare for 44 week 5 TEV run with 6.1*10**6 seconds beam live time.

- New experiment requirements soon (expecting mostly less tape).
- Sites to complete 2009 resources by September 2009
- Maintain 2010 site resources schedule for April 2010

Staff shortages are hurting operations of some complex systems

Analysis is the least well understood area but good work has started.

CCRC'09 No – STEP'09 Yes: May/June. End June is hard limit.

The WLCG operations model is agreed and operations have considerably matured and hardened since CHEP07.

Finally

Workshop participants strongly motivated and enthusiastic

For WLCG the ‘fun’ has already begun

Looking forward to the challenges of real data