WLCG Service Report

Jamie.Shiers@cern.ch

~ ~ ~

WLCG Management Board, 18th November 2008

Overview – from last week

- **Overall goal is to move rapidly to a situation where (weekly) reporting is largely automatic**
- ➢ **And then focus on (i.e. talk about) the exceptions…**
- Have recently added a table of Service Incident Reports
  - Need to follow-up on each case – cannot be automatic!

Propose adding:

- Table of alarm / team tickets & timeline
  - Maybe also non-team tickets – also some indicator of activity / issues
- Summary of scheduled / unscheduled interventions, including cross-check of advance warning with the WLCG / EGEE targets
  - ➢ **e.g. you can't "schedule" a 5h downtime 5' beforehand…**
- Some "high-level" view of experiment / site activities also being considered
  - How to define views that are representative & comprehensible?

Summary of the week

- Relatively smooth week – perhaps because many people were busy with numerous overlapping workshops and other events!
- **ASGC – more problems started on Friday and by Monday only 70% efficiency seen by ATLAS (CMS also sees degradation) – expectation is that this will degrade further. In contact with ASGC – possible con-call tomorrow**
- FZK – some confusion regarding LFC r/o replica for LHCb. A simple test shows that entries do not appear in the FZK replica whereas they do in others (e.g. CNAF). This should be relatively low-impact but needs to be followed-up and resolved.
  - Not entirely clear how LHCb distinguish between an LFC with stale data that is otherwise functioning normally and an LFC with current data...
- Weekly summaries: GGUS tickets, GOCDB intervention summary & **Baseline versions** now all part of meeting template.

➢ **A relatively smooth week!**

WLCG “Data Taking Readiness” Workshop

Some Brief Comments...

Overview

- Over 90 people registered and at (many) times “standing room only” in IT amphi (100 seats)
- Attendance pretty even throughout the entire two days – even early morning – with slight dip after lunch
- Not really an event where major decisions were – or even could be – made; more an on-going operations event
- **Probably the main point: on-going, essentially non-stop production usage from the experiments → on-going production service!**
- Overlapping of specific activities between (and within) VOs should be scheduled where possible... To be followed...
- Personal concern: we are still seeing too frequent and too long service / site degradations. Maybe can “tolerate” this for the main production activities – what will happen when the full analysis workload is added?
- Matching opportunity: compared to this time last year – when we were still “arguing” about SRM v2.2 service roll-out – we have (again) made huge advances. Can expect significant service improvement in coming months / year. But remember: late testing (so far) has meant late surprises. (aka the “Fisk phenomenon”)

Actions

- A dCache workshop – most likely hosted by FZK – is being organised for January 2009
- Discussions on similar workshops for other main storage implementations with overlapping agendas:

1. Summary of experiment problems with the storage
   - what does not work that should work because it is needed
   - what does not work but you just have to deal with it
2. Instabilities of the setup
   - timeouts, single points of failure, VO reporting, site monitoring known (site specific) bottlenecks
3. Instabilities inherent with the use of the storage MW at hand
   - bugs/problems with SRM, FTS, gridFTP, used clients

Summary on Data-taking readiness WS

Patricia Mendez, R. Santinelli on behalf of the "*Rapporteurs*"

Alberto Aimar, Jeremy Coles

LHC status

- Interesting (technical) insights into the 19th Sep incident and description of the operation to get it back to life
- Question: when it will restart operating? Simply don't know. A more clear picture expected by the end of December.
- Enormous temptation to relax a bit: never mind.

Ed: I think that this means "don't"

- CRSG report to C-RRB
  - Harry reported the report to C-RRB about the scrutiny the subgroup C-RSG.
  - First time public assessments on the LHC exps requirements scrutinized
  - Some discrepancies between 2008/09 reqs and advocated resources per each exp
    - because of LHC shutdown.
    - Fairly old TDRs (2005)
  - This is just an advisory, the first step of a resources allocation process that has the final decision from the LHCC body
  - Burinng question from the C-RRB:
    - 2009 envisaged like 2008. Why more?
    - Usually site usage shown only 2/3 of resources used. Why more?

Experiments plans: re-validation of the service

- Common points:
  - No relax but keep the production system as it was in data taking
  - Cosmic data is available (at least ATLAS and CMS, LHCb enjoys muons from dump of SPS)
  - Plans to be tuned with LHC schedule. Assumption is that in May LHC is operative and then 2/3 real data taking (cosmic/beam collisions) and 1/3 to final dress rehearsal tests.
  - No need to have a CCRC'09 like on 2008 being the experiments more or less continuously running activity.
  - Some activities should be at least advertized (throughput)
  - Some other need to be carried out (staging at T1 for all exps)

CMS

- recent global runs CRUZET CRAFT (with 3.8 T magnetic field)
  - Analysis of these data now. Reprocessing these data with new releases of CMSSW in few weeks and on January.
- Analysis end-to-end. It is not a challenge but we need to go through to convince the quality of the process by the raw till the histogram (validation of the process).
- Another CRUZET global run with all subdetectors on will take place before LHC start (April/May)
- 64 bit: CMS sw not ready yet not time line for move to. WLCG should provide 64 bit resources with 32 bit compatibility
- For what concerns facilities (monitoring tools, procedures) CMS feels ready.

ATLAS

- September + December 2008: cosmic ray data taking and preparations re-processing and analysis
- Early 2009: reprocessing 2008 cosmic ray data (reduction to ~20%) and data analysis
- Presented resources requirements for these activities (pre-staging might be a problem, not now having already prestaged cosmic data at T1 →1000CPUs site ~1 TB/hour on-line
- New DDM (10 times more performant, functional tests)
- Distributed Analysis Test (done in Italy) should be carried on all clouds (at T2 1.2 GB/sec (10 Gb) per 400 CPU's)

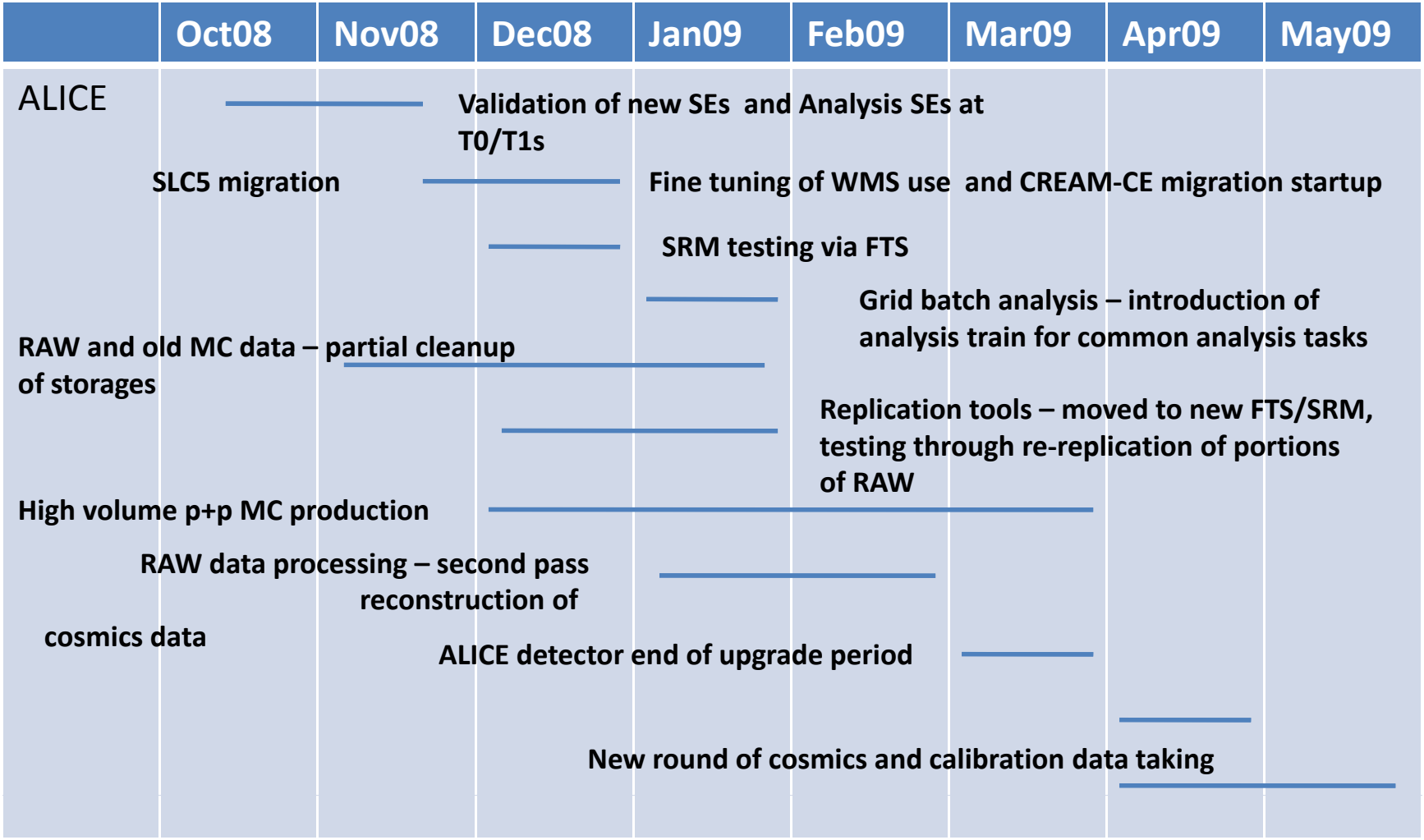
LHCb

- DIRAC3 is now in production since July and DIRAC2 dismissed fall 2008.
- Some Montecarlo for physics study of the detector and benchmark physics channels
- Some analysis of DC06 data via DIRAC3
- FEST09 started already (for its MC production now going to merge).
  - Full Experiment System Test (from the HLT farm to the user desktop exercising all ops procedures) in January/February tests and then March/April the exercise. A full scale FEST expected if LHC delay
- Generic pilot sentences
- CPU normalization sentences
- Concept of CPU peak requirements: site are asked to provide certain amount of CPU per exps but from time to time the needs fluctuate (from half or 0 to twice this pledge. Opportunistic usage of other VO's allocated resources at the site not always work

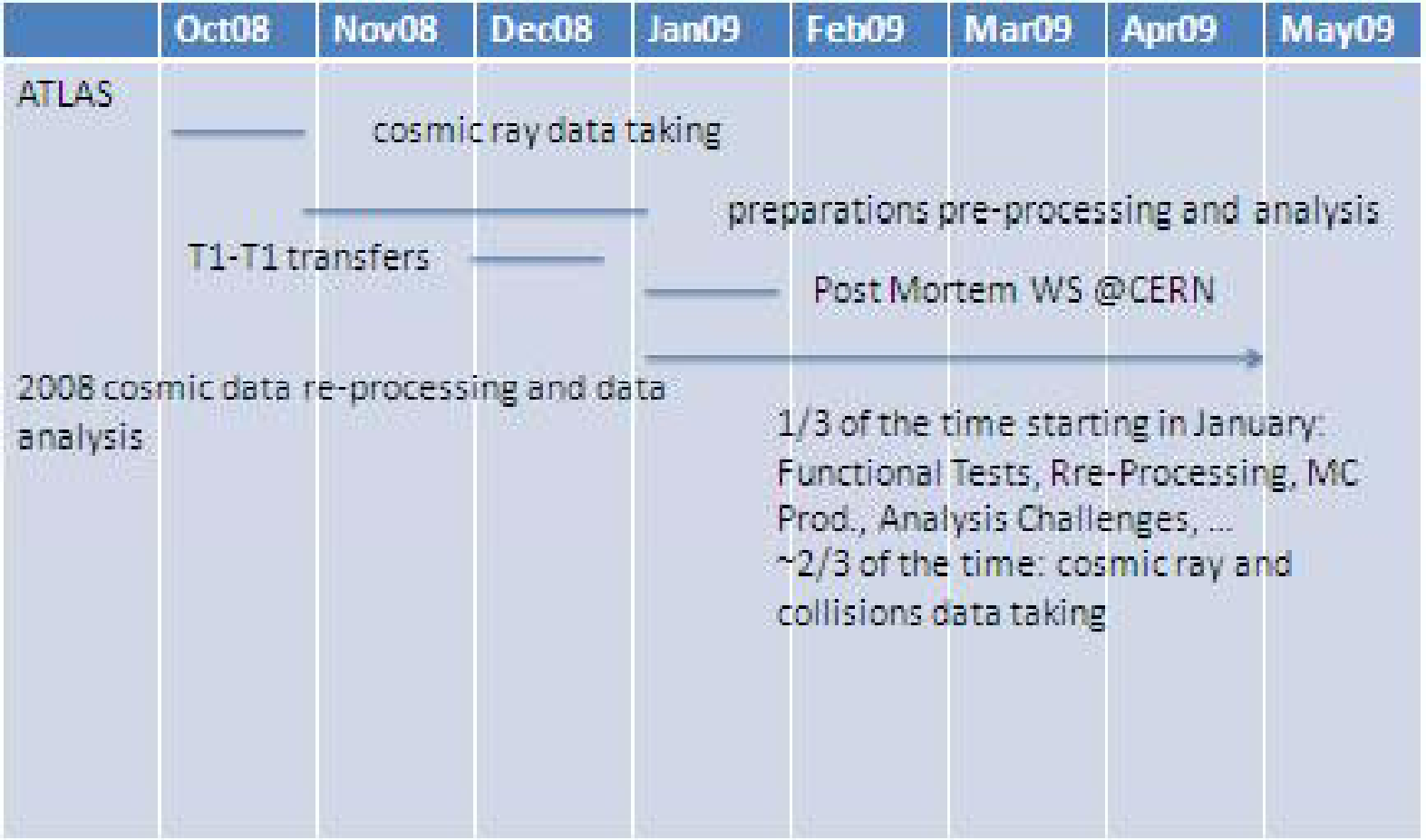
ALICE

- Data in 2008
  - 310 TB@CERN and 200TB replicated to T1 (FTS) in including cosmics and calibration
  - Discussion about data cleanup of old MC productions ongoing with the detectors groups and to be coordinated with IT
- Offline Reconstruction
  - Significant improvements of the processing configurations for all detectors
  - All reconstructible runs from 2008 cosmics data taking are processed
  - Development of quasi-online processing framework ongoing with new AliRoot releases
- Grid batch user analysis: High importance task
- Specific WMS configuration ongoing
- High interest on CREAM-CE deployment
  - Already tested by the experiment summer 2008 with very impressive results
- Tests of SLC5 going on at CERN
  - System already tested in 64b nodes
- New run of MC production from Dec08
- Storage
  - Gradually increasing the number of sites with xrootd-enabled SEs
  - Emphasis on disk-based SEs for analysis
    - Including a capacity at T0/T1s
    - Storage types remain unchangde as defined in previous meetings

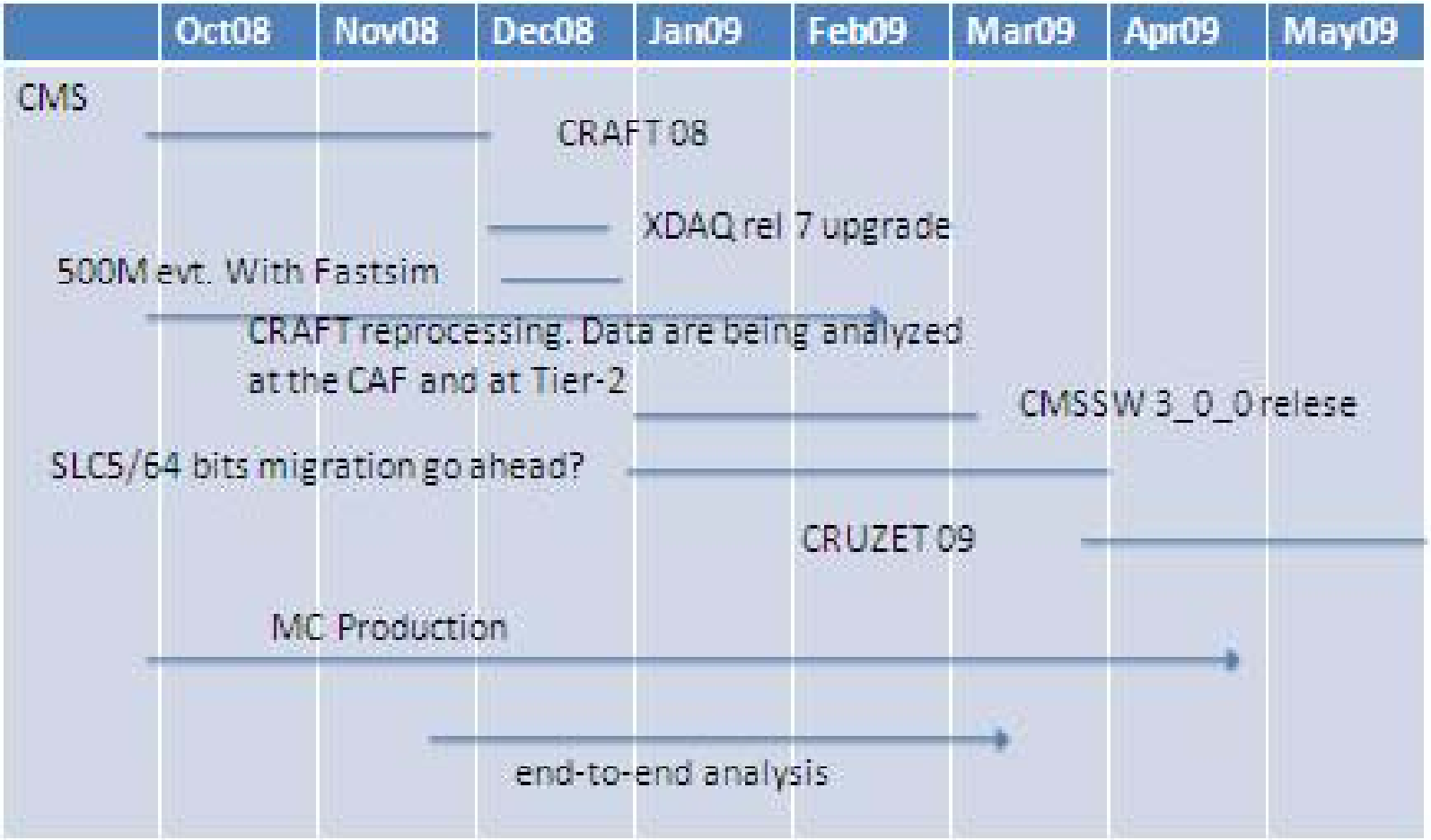
Experiments Plans



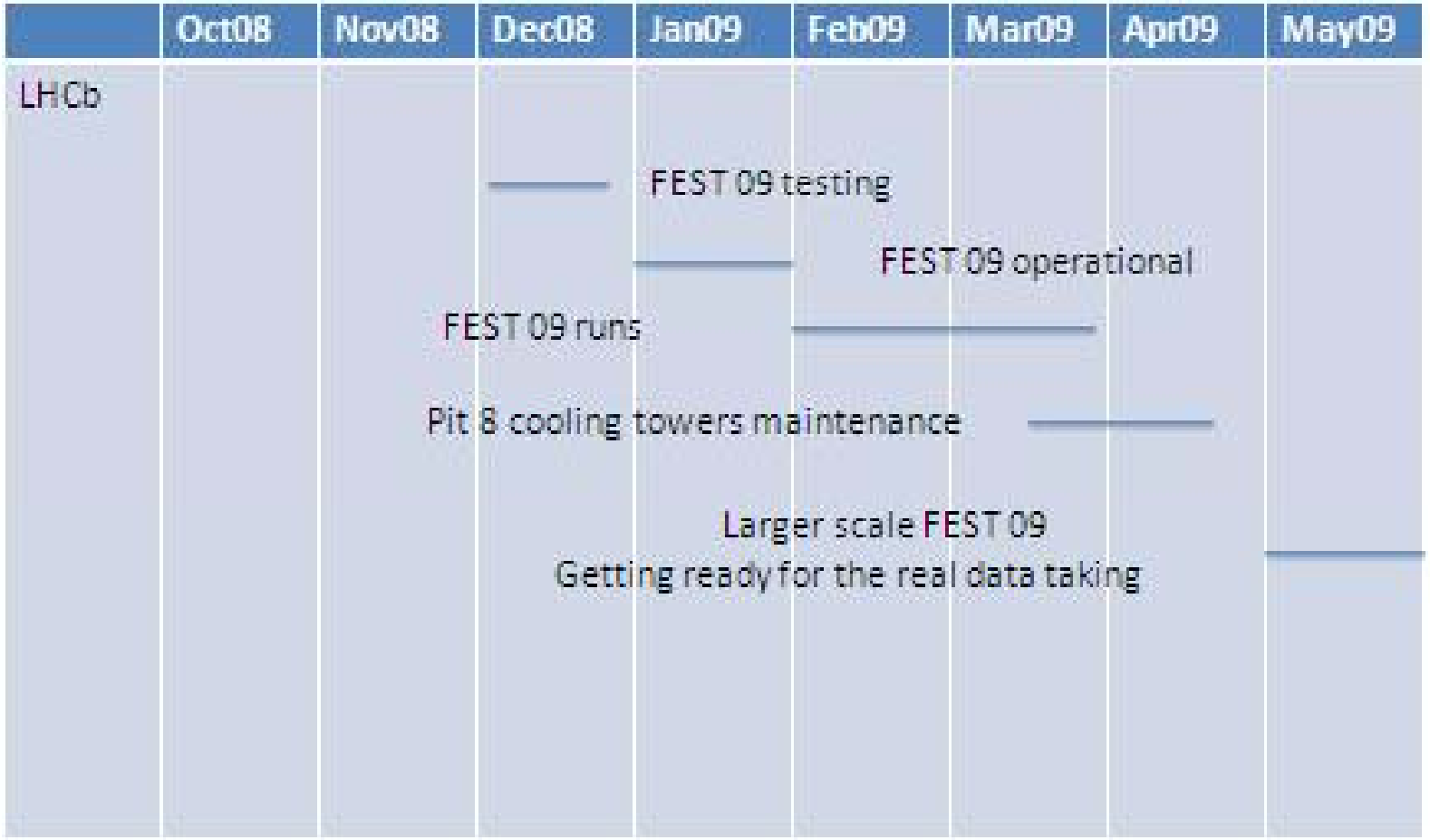
Experiments Plans



Experiments Plans



Experiments Plans



DM (I)

- Presented the status of various DM services, versions and experiences .

☺ **LFC (<1 incident/month in WLCG) and FTS (1.5/month)**

☺ **fairly good reliability observed.**

- Presented general storage issues: Main problems **still** on the Storage area (robustness and stability).
  - dCache Pnfs performances (mainly affected when expensive srmls are run with stat (FTS staging files hammering does not help)
  - Storage DB issues (RAL ASGC)
  - Files temporary lost should be marked as UNAVAILABLE so that experiments know it is not accessible.
  - Issues in the pre-staging exercise of ATLAS at various T1. Not problem for CASTOR but for dCache having different MSS system.
  - Various outages due to disk hotspots. Difficult disk balancing.

DM (II)

- Pictorial view of DM layers also given. Error are frequent to happen at any of each of these layers or between them (sub-optimized)
- The error often obscure. Must invest on a more exhaustive logging from DM
- Operations cost is still quite high

➢ **Sharing experiences across experiments but also sites would save time**

- We can survive already with the current system!

Ed: watch out for analysis Use Cases!

Distributed DB

- Smooth running in 2008 at CERN, profiting from larger headroom from h/w upgrade
  - Comfortable for operating the services
  - Adding more resources can only be done with planning
    - Use of data-guard also requires additional resources
- New hardware for RAC2 replacement expected early 2009
  - 20 dual-cpu quad-core (possibly blade) servers, 32 disk arrays
- Policies concerning integration and production database services remain unchanged in 2009
  - s/w and security patch upgrades and backups
- Hope that funding discussions for online services with ATLAS and Alice will converge soon

Ed: ATLAS conditions strategy to be clarified & tested, in particular for Tier2s. Gremlin over LHCb LFC/FZK ?

T2

- No major problems in 2008→Only CMS seems to have tried out analysis at T2's.
- Worries for 2009 :
  - Adequate the storage for analysis,: will real data and real activity come to T2?
  - Monitoring of sites...too many web portals and lack of a homogeneous source (site view).
- Hot topics for T2.
  - How to optimize CPU/IO bound? The LRMS should be smart enough to send high CPU bound jobs to machine with "poor" LAN connection and viceversa high bandwidth boxes serving high IO application.
  - How much the shared area size?

T1

- Stable CE/Information System Pilot
  - Pilot jobs would further alleviate
- Major feedbacks received about SEs
- Scalability of current dCache system
  - CHIMERA and new PostGres would alleviate
- Currently SRM system can handle 1Hz SRM requests.
- Sites would like to know how far from the experiment's targets they are
- Site view of experiment activity is also another important issue

T1 reliability

- Too often T1 are seemed to break the MoU
- 2 key questions risen to T1 (only few answered)
- Best practice now known to sysadmins but:
  - Are really these recommendation followed by T1's
- Proposal to have T1s reviewing each others seems interesting which is a good way to share expertise.

T0

- Services running smoothly now.
  - hardware issues and hotspots problems mainly.
- Procedure documentations should be improved
  - now the “Service on Duty” covers 24X7 and more people have to be in the game
- One main area of failure
- Interesting rules of thumb for upgrading a service adopted at CERN
  - PPS seems to help
- Illustrated the Post-Mortem concept:
  - *“According to WLCG a Post Mortem is triggered when a site’s MoU commitment has been breached”.*
- Time spent by sysadmin to catch and understand problems could be dramatically alleviated by a better logging system.
- Often we know a problem already and may be effort addressed to make it not happening again

Middleware

- meaning of the baseline and targets approved by MB for data taking
  - Defined the minimum required version of packages.
- intended baseline
  - WN SLC5 and new compiler versions (UI later)
  - glexec/SCAS: target: deployment of SCAS (now under certification)
  - GLUE2:no really target
  - Rationalize the publication of heterogeneous clusters.
  - WMS a patch fixing many problems is in certification.
    - ICE for submitting to CREAM there.

Storageware

- CASTOR
  - CASTOR Core: 2.1.7-21
  - CASTOR SRM v2.2: 1.3-28 on SLC3

  The recommended release is:
  - CASTOR Core: 2.1.7-22 (released this week)
  - CASTOR SRM v2.2: 2.7-8 on SLC4

  (srmCopy, srmPurgeFromSpace, more robust)

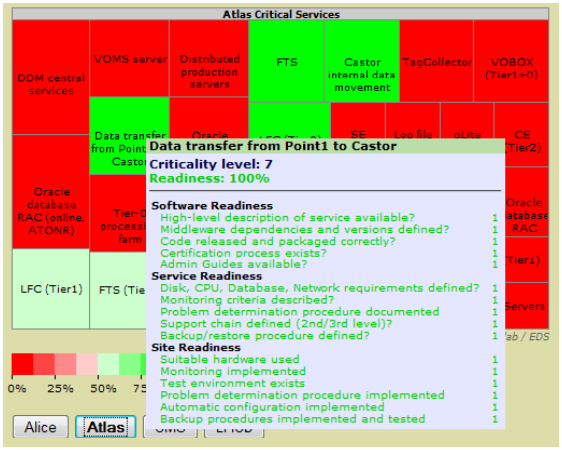
  The next version of CASTOR, 2.1.8, is being considered for deployment at CERN for the experiment production instances.
- dCache (1.9.0 recommended also fast PNFS)
  - 1.9.X (>1) comes with new features (New pool code (November) , Copy and Draining Manager PinManager (December, January), gPlazma,Tape protection (MoU) srmReleaseFiles based on FQAN (MoU), Space Token protection (MoU), New information providers, ACLs (January), NFS v4.1, Bug Fixes (gsiDcap, UNAVAILABLE Files, etc.)
- DPM V.1.6.11-3 is the last stable release
  - V. 1.7.0 currently in certification (srmCopy, Write permission to spaces limited to multiple groups (MoU), srmReleaseFiles based on FQAN (MoU), srmLs can return a list of spaces a given file is in, New dpm-listspace command needed for information providers (installed capacity), DPM retains space token when the admin drains a file from a file system)
- StoRM 1.3.20-04 currently in production
  - Major release 1.4 dec. 2008

Definition of readiness

Questions chosen to determine the level of "readiness"

Software readiness
High-level description of service available?
Middleware dependencies and versions defined?
Code released and packaged correctly?
Certification process exists?

ALICE ok. ATLAS ok – but relying on experts. CMS services basically ready (but still debugging). LHCb lack procedures/documentation in many areas



Support chain defined (2nd/3rd level)?
Backup/restore procedure defined?
Site readiness
Suitable hardware used?
Monitoring implemented?
Test environment exists?
Problem determination procedure implemented?
Automatic configuration implemented?
Backup procedures implemented and tested?

Reliability, redundancy etc. not part of this definition!

Analysis WG

- Presented a WG setup for distributed analysis
  - what is the minimum amount of services requirement.
  - What the models from the experiments.
- deliverable:
  - documented analysis models
  - report on reqs for services and devs with priorities and timescales.
- Understand how the experiments want to run analysis and advertize sites how best they should configure the resources.

New metrics to be defined?

- “Simulated” downtime of 1-3 Tier1s for up to – or exceeding – 5 days (window?) to understand how system handles export including recall from tape
- Scheduled(?) recovery from power-outages
- Extensive concurrent batch load – do shares match expectations?
- Extensive overlapping “functional blocks” – concurrent production & analysis activities
- Reprocessing and analysis use cases (Tier1 & Tier2) and conditions "DB" load - validation of current deployment(s)
- Tape rates - can sites support concurrent tape rates needed by all VOs it supports concurrently?

Goals and metrics for data taking readiness "challenges" – Ian/Jamie

- Summary of C-RRB/MB discussion. Should there be a CCRC'09? No but do need specific tests/validations (e.g. combined running, reprocessing …).
- Experiment plans shown yesterday, no time coordination between activities.
- Areas to be considered listed on the agenda (based on list from operations meeting)
- What tests? How will buffers deal with multiple site outages. Real power cuts break discs, simulated outages have less problems!
- Many things happen anyway – need to document the recovery better. Follow up on things that went wrong. Expts. have to expect problems and be able to deal with them in a smooth way.
- Sites still waiting for data flow targets from the expts. CCRC08 rates are still valid. ATLAS plan on continuous load but DDM exercise in 2 weeks will provide peak load. Worry about tape rates – limits in number of available drives. CMS – alignment of smaller activities would be useful. ALICE – live or die by disk based resources.
- Reprocessing and analysis cases of particular concern. Conditions DB in ATLAS (T2 frontier/SQUID) needs testing. Other than concurrent tape rates CMS has own programme of work. When a more complete schedule exists looking at overlaps would be interesting. ATLAS hope all clouds involved with analysis by end of year. LHCb- can plan modest work in parallel – sometimes need to solve problems in production when they occur.
- Allow expts. To continue with programmes of work. When ATLAS read -> common tape recall tests. Try time when can push lots of analysis. Things should be running and then focus on resolving issues in post-mortems. Need to test things that change (e.g. FTS).

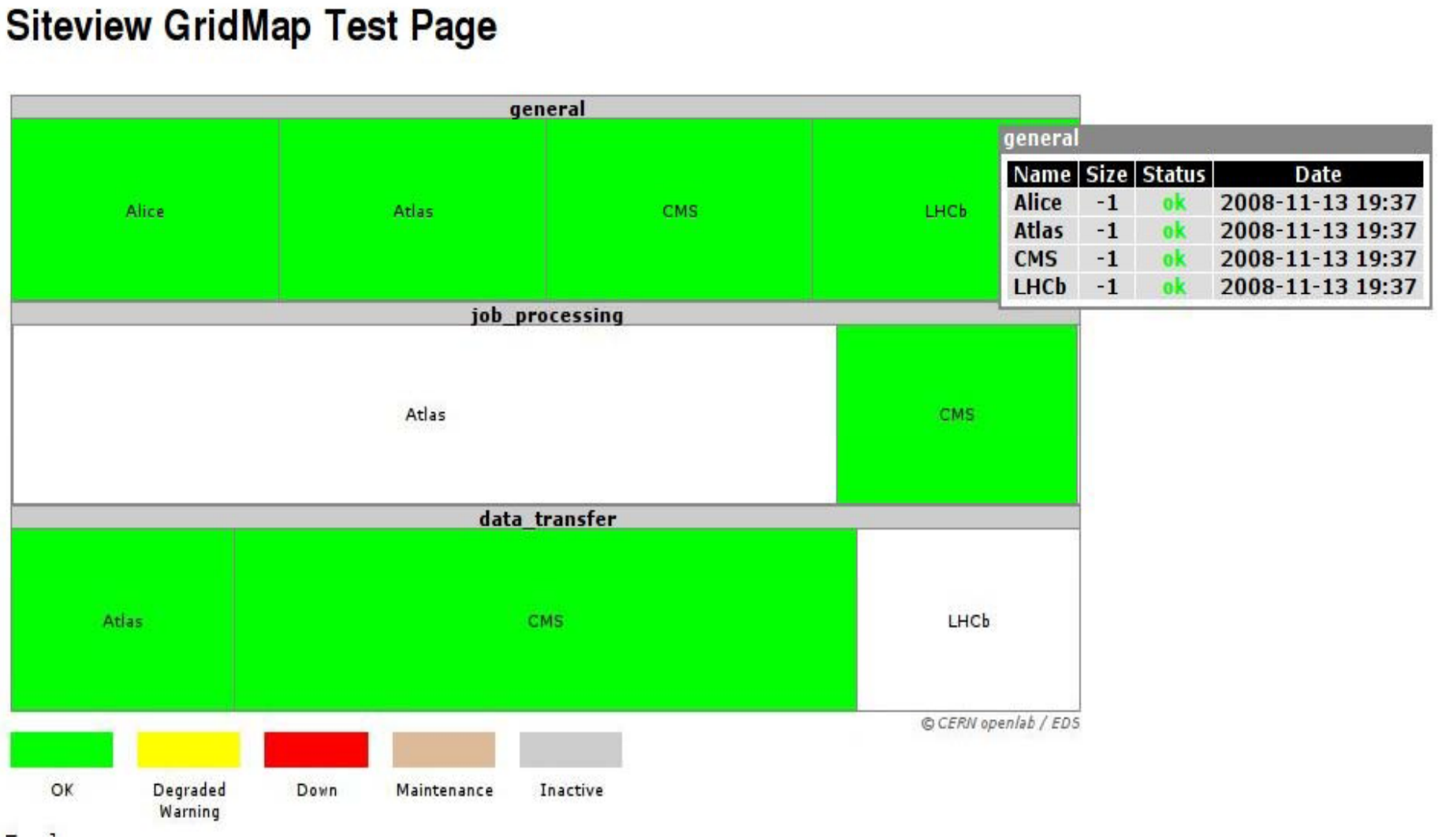
DM software

- Presentation of the DM s/w status and ongoing activities at the T0
- Status of Castor@T0
  - Used for migration of LHC data to tape and T1 replication
  - System in fully production state
  - Stressed during the CCRC'08 but few improvements neccessary from the point of view of the analysis

Calendar of major improvements

	Today	Spring09	Mid 09	End 09	End 10
Security	End user access to CASTOR has been secured	Support both PKI (Grid certificates) and Kerberos	Castor Monitoring and SRM interface being improved		
Tape efficiency	New tape queue management supporting recall / migration policies, access control lists, and user / group based priorities		New tape format with no performance drop when handling small files	Data aggregation to allow managed tape migration / recall to try to match the drive speed (increased efficiency)	
Access latency	a) Removed the LSF job scheduling for reading disk files b) Direct access offered with the XROOT interface which is already part of the current service definition c) Mountable file system possible		Plan to remove job scheduling also for write operation		Additional access protocols can be added (eg. NFS 4.1)

One desired answers to many sysadmins requests in the path…the site view gridmap.



Exp. Operations

- Experiment Operations rely on multilevel operation mode
  - First line shift crew
  - Second line Experts On-Call
  - Developers as third line support
    - not necessarily on-call
- Experiments Operations strongly integrated with WLCG operations and Grid Service Support
  - Expert support
  - Escalation procedures
    - Especially for critical issues or long standing issues
    - Incidents Post Mortems
  - Communications and Notifications
    - EIS personally like the daily 15:00h meeting
- **Problem with a VO in a site should be notified to the local contact person**

  That is known centrally by the VO and will transmit to the local responsible
    - Especially for US and Asia regions
- All experiments recognize the importance of experiment dedicated support at sites
  - CMS can rely on contacts at every T1 and T2
  - ATLAS and ALICE can rely on contacts per region/cloud
    - Contact at all T1s, usually dedicated
    - Some dedicated contact also at some T2
  - LHCb can rely on contacts at some T1

MoU Targets

- Do they well describe the real problems we face every day?
- GGUS is the system to log problem resolution
  - MoU categories introduced in some critical tickets (TEAM and ALARM)
  - Are these categories really comprehensive?
  - Should they provide a view by service or by functional block (i.e raw distribution)? (the former implies the submitter knows the source of the pb)
  - Time resolution: are they realistic?
  - No conflict with the experiment monitoring systems GGUS is not for measuring the efficiency of the site.
  - Response time to a GGUS ticket is in the MoU not solving time.
  - The sysadmin must any way to close as soon as he's confident the problem is fixed.
- In GGUS we have all data to build up complete and reliable report for assessing the intervention has been carried properly and accordingly MoU
- Use SAM test for each of the MoU categories and with GridView computing the site availability for automatic monitoring.