Table of Contents

WLCG Operations Planning - 29nd November 2012 - minutes

Agenda

- https://indico.cern.ch/conferenceDisplay.py?confId=219317⧉

Attendance

- Local: Andrea Sciabà (secretary), Maria Girone (chair), Alessandro Di Girolamo, Nicolò Magini, Simone Campana, Ikuo Ueda, Marco Cattaneo, Dirk Duellmann, Borut Kersevan, Ian Fisk, Stefan Roiser, Domenico Giordano, Maarten Litmaath, Maria Dimou
- Remote: Jeremy Coles, Christoph Wissing, Rob Quick, Dave Dykstra, Alessandro Cavalli

Agenda items

Goals of the meeting (M. Girone)

Maria introduces the meeting pointing out that its main purpose is to review the needs and plans of experiments and sites, and in particular today the focus is on the experiment plans for 2013 in terms of computing operations.

ALICE plans (M. Litmaath)

Maarten briefly illustrates the foreseen processing, analysis and MC production activities for 2013. Even if the pressure will be less than during data taking, a lot of work is still planned, also in connection to conferences.

Then, he highlights some topics of particular relevance to ALICE: CERNVM & CVMFS, gLExec, the SHA-2 migration and the tracking tools evolution.

The main interest of CERNVM is to integrate the HLT farm with the offline resources, starting with small scale tests and then gradually expanding to the production system over a time scale of several months. There is a strong interest in doing this together with other experiments using common solutions.

The gLExec deployment is not easy for ALICE. They are also following the evolution of virtual machine technology and the introduction of cgroups in SL6 kernels, which would allow to lock up processes in their own "jails", which might be enough in the long run to separate users in worker nodes to avoid interference from a security perspective. This is a background activity, as the technology is still rather new. Even if ALICE will look into integrating gLExec in AliEN, a simpler solution might come later.

The SHA-2 migration affects everybody. The experiments, EGI and OSG will have to work together. A new CERN CA, capable of issuing SHA-2 proxies, will be available in January, though it will not be in IGTF, at least initially. The efforts are expected to significantly ramp up from January on, also in EGI. In the mean time, experiments can start testing with RFC proxies already from now. Even if some middleware components will not be ready for RFC proxies and SHA-2 before EMI-3, experiment-specific services can start to be tested for RFC readiness. Maarten would like to encourage the four LHC experiments to start testing with RFC proxies and the WLCG cooordination working group would monitor the progress.

Other topics, like the evolution of tracking tools and possibly even FTS 3, are potentially interesting for ALICE.

Maria G. suggests that Maarten could look into the tracking tools and CVMFS for ALICE, while Stefan and Fernando Barreiro could collaborate for what concerns testing CERNVM on the HLT for all the experiments. Maarten agrees that taking advantage of the work already done by others would be beneficial.

Concerning SHA-2, Marco says that Adrian Casajus could help on behalf of LHCb. Maarten replies that Adrian had already started looking into adapting DIRAC during the spring.

Maria D. asks if VOMRS will be included in the SHA-2 tests. Maarten replies affirmatively, although one has to take into account that VOMRS is at the end of its life and will be replaced with VOMS-Admin. Still Maarten hopes that it may be trivial to make VOMRS to work with SHA-2 certificates: a test he did recently was partially successful but for some issues that Steve has not yet had the time to investigate. If it is not the case, it will put additional pressure on the VOMRS-VOMS-Admin transition.

ATLAS plans (I. Ueda)

For ATLAS, 2013 will see some significant changes: the migration from DDM to Rucio; the introduction of a new production task definition system, JEDI (not so relevant for the sites); improvements in the ATLAS software (to introduce multi-core designs, in collaboration with Openlab, PH-SFT and CMS) and in the event data model.

The migration to Rucio will require to change the SURLs of all files at sites. ATLAS is studying how to do it by renaming the files, without moving any data; a possibility could be using WebDAV (or native functionality of the storage system, when appropriate). DDM and PanDA will soon start using the new name convention. More details will be available after the ATLAS T1/2/3 jamboree in December. The migration to Rucio will eventually remove the dependency on the LFC, by the end of LS1.

ATLAS is also testing the WAN data access and caching in the xrootd federation; when satisfied with the results, a global deployment will be requested.

The validation of the ATLAS software on SL6, the FTS 3 validation and the planning of database usage for LS1 and Run2 are also ongoing; the latter is expected to be finalized at the Jamboree.

The ATLAS needs for monitoring include Hammercloud and SAM. Hammercloud needs to keep running with the latest releases of the ATLAS software and the HC reports for the sites should be improved. The SAM framework needs to be extended to support external metrics and allow for more flexibility in testing services (e.g. services not registered in GOCDB, OIM or the BDII) and metaservices (e.g. test activities and sites rather than CEs, SEs, etc.). SAM/Nagios probes should be maintained as common tools and also experiment-specific probes should be reused by more experiments when possible; the WLCG operations coordination working group should take care of this.

ATLAS plans to reduce the computing shifts and not have on-call experts for the Tier-0 processing and export, nor ALARM tickets for the Tier-0 activities. Shifts for the grid activities will continue and ALARM tickets affecting critical ATLAS activities will be possible. The WLCG critical services list should be updated and available on the web. The AMOD (ATLAS distributed computing manager-on-duty) shifts may be remote and participation to the 3pm WLCG operations meeting be via phone. Moreover, ATLAS suggests to reduce the frequency of the 3pm meeting and use GGUS for the daily issues.

Maria G. proposes to have the daily meeting on Mondays (to review the weekend) and Thursdays (to review the week). All four experiments agree to it.

Maria G. comments on the monitoring for operations reminding that the WLCG TEG on operations and tools had a mandate also to understand monitoring needs and their evolution, but when the WLCG operations working group was formed, monitoring was not part of its mandate, as there is a wish to review the strategy on monitoring and create a dedicated task to include all monitoring activities. Still, she is in favour of

experiments and sites expressing their needs for operations-related monitoring already now in the operations working group, without interfering with the WLCG global strategy but allowing changes to be requested and made, for instance in Hammercloud and SAM. This would be acceptable, until the global strategy on monitoring is clarified.

Ian says that Hammercloud and SAM cannot be decoupled from operations; LHCb, ALICE and ATLAS support his statement.

Maria G. asks what is the timescale for the final migration from DDM to Rucio. Borut answers that it is not yet fully defined but it should start by the end of 2013. Simone adds that the file renaming will happen much sooner. Dirk asks when should it finish. Simone answers that he does not know yet how long it will take via WebDAV; for sites not supporting it (like CERN) there are dedicated discussions and for them quicker and cleaner ways to do the renaming are considered.

Maria G. asks LHCb what are the plans for the File Catalogue, given that by the end of LS1 LFC will not be used anymore. Marco answers that there is a lot of development ongoing but it is too early to say. Stefan adds that multiple strategies out of LFC are being considered.

Stefan asks how ATLAS thinks to adapt the site infrastructure to multi-core in Athena, for example by using special queues. Ueda and Borut answer that they are in fact using special queues, though it is probably not the final solution; on this they would like to collaborate with the other experiments and PH-SFT.

Gonzalo asks if proposals for "whole nodes" are still valid; Ian anticipates that he will discuss it in his talk.

CMS plans (I. Fisk)

Ian illustrates the foreseen evolution for the next 6-12 months and on a longer term.

Xrootd is planned to be rolled out on all sites, starting with enabling fallback (which is very easy to configure) and then enabling data access via xrootd. SAM is used to monitor the health of the functionality and the xrootd popularity monitoring to monitor the sustainability of the system.

Another planned change is to have a Tier-1 storage "cloud", that is allowing for transparent data access across Tier-1 sites using the OPN. The advantages will be to eliminate individual workflow problems and in the long term to enable a more dynamic usage of resources and the fallback to xrootd instead of tape, as discussed in the WLCG jamboree in Amsterdam. All experiments can potentially benefit from it and it would change the Tier-1 concept as an aggregated resource instead of several individual sites. Connecting the batch nodes to the OPN will require a fair deal of negotiation.

CMS plans also to ask Tier-1 sites to implement a disk/tape separation, via two PhEDEx endpoints, one for disk and one for tape. The disk endpoint should write only to disk and the tape endpoint should write to tape, but could also be hosted on disk. Suscribing to disk would cause a prestage and pin, while subscribing to tape would trigger a copy to the tape archive. This would provide better disk management and prevent unexpected staging. Sites can freely choose how to implement it (e.g. CERN did it by introducing a separate technology and RAL with a split of the existing technology).

In addition, CMS proposes a new split of the disk space at Tier-1/2 sites, whereby all T1 disk space and a fraction of T2 space is centrally managed and the rest of the T2 space is managed by users and groups. To save disk space, cached copies of data will be automatically cleaned up with Victor. This will reduce the couplings of groups and sites and slow down the increase of disk space needs.

CMS is now testing the HLT farm in "cloud-mode" for organized data processing using Openstack with EC2 interface to start resources and glideinWMS3 for resource control, with data accessed from EOS over xrootd. The goal is to run at scale during the Christmas stop.

ATLAS plans (I. Ueda)   3

Another goal is to eliminate the last dependencies on the gLite WMS. Next week CMS will develop a release plan for the next generation of analysis tools that will use pilots by default. Also the SAM tests will need to move away from the WMS.

Concerning tools for operations, it is important that Hammercloud and SAM are properly supported, as they proved very useful in achieving and maintaining a high site reliability level. Some new features are needed, like the addition of site-level metrics and of new service types in SAM, and full glidein support, better error reporting tools and cloud submission in Hammercloud. Development of probes should exploit the many commonalities with other experiments to save effort.

By the end of 2013 CMS will have a multi-threaded reconstruction software and will need to transition organized processing to multi-core using queues with up to 8-16 cores. Opportunistic computing using e.g. clouds and university clusters that can be used with short notice will also become interesting, and the goal is to demonstrate that CMS can spill work to O(30K) cores in clouds.

Coming back to Gonzalo's question, Ian explains that the plans to go multi-core that started last year by asking Tier-1 sites to provide a 10% of nodes in multi-core queues were postponed because the memory improvements in CMSSW 5 were very good, lessening the urgency of the transition.

Maria G. comments that the xrootd deployment could be handled by the WLCG operations working group and the use of the xrootd popularity expanded to monitor all federation resources. Domenico stresses that we have to work both on communication inside the WG and on the monitoring and the deployment. Ian suggests to join efforts with ATLAS to have a checklist for sites and have it pushed by WLCG; this should make it easier to convince sites. Maria G. proposes that ATLAS and CMS talk to Domenico to understand how to start this. Dirk suggests to involve F. Furano (whose working group on storage federations is ramping down).

Simone asks Ian if having multi-core queues requires to partition resources, also considering that most likely analysis jobs will not be multi-threaded. Ian says that the pilot should be smart enough to understand how many instances of the software to start, while the entire farm should be multi-core.

Maria G. asks the opinion of the other experiments about the Tier-1 "cloud". Simone answers that it was recently discussed and brought up with Michael Ernst, who discussed it with LHCONE but in a more general fashion. For ATLAS, the idea is interesting both for the OPN and LHCONE. Ian proposes to wait for the network workshop in two weeks and speak to Michael. He asks to LHCb if they have any objection, considering that it might be a disruptive change, if it requires long downtimes. Marco answers that actually the idea is interesting also for LHCb, and similar to what LHCb already does at Tier-2 sites, though so far there was no plan to do it at Tier-1 sites as well.

Gonzalo points out that very good monitoring is crucial and asks if it is available. Maria G. answers that it is, though it is being enhanced, with work ongoing in IT-ES. Domenico adds that for example xrootd access monitoring is in place for ATLAS and CMS at several sites and provides much more information than it did two months ago.

LHCb plans (S. Roiser)

Stefan describes the plans for data (re)processing for the rest of this year, 2013 and beyond. In particular, the incremental stripping, starting in autumn 2013, will be particularly challenging, as it will need to re-stage all reco output in a very short time. CERN cannot help, as all data are in Tier-1 sites. He shows a very useful table summarizing the expected load on the tape systems at each Tier-1 site, which shows that the incremental stripping will be the most demanding activity. Sites should give it a look.

He recalls the services recently retired: Oracle Conditions databases at Tier-1 sites, LFC instances at Tier-1 sites and Oracle Streams from the Tier-0 to the Tier-1 sites. Soon LHCb will also exclusively use direct submission to CREAM, thus bypassing the glite WMS, whose usage will be discontinued by summer 2013.

CMS plans (I. Fisk)   4

This will require to replace it also in SAM, for which a common solution is desired.

Maarten says that some details of the SAM job submission will not be the same for all the experiments. For instance, ALICE uses the gLite WMS only to send worker node tests to the site: the SAM test that directly submits to CREAM does not currently allow for that. Simone suggests CondorG submission as an alternative and Andrea says that is is an option also for CMS.

This topic: LCG > WLCGOpsMinutes121129
Topic revision: r8 - 2013-03-19 - AndreaSciaba