Table of Contents

Agenda

https://indico.cern.ch/event/319754/

Introduction - M. Jouvin

Next GDB and preGDB

- Ian Collier will take over chairmanship in January: GDB evolution will be decided after the discussions during workshop.
- February canceled: just one week after the workshop

Workshop in Lisbon 1-3 February: a very important event for WLCG

- **register and participate**: https://indico.cern.ch/e/WLCG-Workshop-Lisbon-2016
- Followed by DPHEP workshop: https://indico.cern.ch/event/444264/

HTCondor European workshop: Barcelona (ALBA synchrotron) 29 February 04 March

- Presentations, tutorials, individual F2F discussions. Covering Condor CE and ARC CE as well
- For sites running HTCondor and for those who are considering/planning to do it or interested to hear about it
    - Unique occasion to meet with HTCondor developers: proved to be very rich last year

ARGUS: collaboration working well, monthly meetings with summaries

- See Indico for details: https://indico.cern.ch/event/465818/
- EL7/Java 8 version ready for server and client, several bugs fixed
    - Pepd issue: bug in pool account allocation fixed
- Advanced prototype available for replacing gridmapdir by an in-memory database: a production version is expected by mid 2016

Actions in progress (https://twiki.cern.ch/twiki/bin/view/LCG/GDBActionInProgress)

- Still looking for more early adopters of Machine/job features: focus on LHCb sites
    - Sites which deployed it reported no difficulty to do it
    - Andrew McNab taking over from Stefan Roiser as the MJF TF leader
- Reminder that VOs are expected to fill information on:
    - Class 2 services: https://twiki.cern.ch/twiki/bin/view/LCG/Class2VOServices
    - storage protocol / implementation / experiment combinations: https://twiki.cern.ch/twiki/bin/view/LCG/WLCGDataAccessProtocolUse

Best wishes to Ian Collier and thanks to all the contributors and the note-takers during these almost 4 years

- Note takers by order of contribution (from 5 to 2 GDB): Maarten Litmaath, Jeremy Coles, Ulf Tigerstedt, Oliver Keeble, Stefan Roiser, Maria Alandes, Renaud Vernet, Catherine Biscarat, Andrew Samsum, Andrew McNab, Helge Meinhard

WLCG/DPHEP Workshop - A. Sciaba

WLCG workshop main goal: preparation of Run 3 and Run 4

- One day dedicated to long term: no detailed agenda yet, driven by experiments, "revolutionary ideas" welcome!
  - Ian Bird taking the responsibility of shaping the agenda, based on discussions with experiments
- Also address some medium-term issues and optimisations: compute, security and trust model, storage/data, monitoring and information system

Lisbon, February 1-3

- Indico event: see slides

DPHEP workshop: plans for certification, data management plans

- Mainly for T0 and T1s

Accounting Update - J. Gordon

Multicore usage: successful campaign to get the core information published, 99% of WLCG

- Missing resources are mainly due to an issue between ARC CE and PBS @DESY-HH
- Also a low-level of failed ARC CE jobs not reporting their core usage (the same fixed for CREAM CE)
- WLCG reports are now taking into account the core information published to report CPU efficiency

Accounting portal in the middle of a major rewrite

- Temporarily the new data (history limited to 18 months) is available as a new tree (EMI3)
- T1 and T2 views will also display soon the number of cores and wallclock*ncores in addition to raw wallclock

CPU accounting developments

- ARC parser: keeps local history, easy to republish
    - Still in test: VO info missing, to be added soon
- HTCondor CE: would like to hear a site who needs it to work on it before the CE is in prod...
    - CERN is running its own accounting tools so need for another site
    - Jeremy: a few UK sites looking at HTConcor CE, may be interested
- Would be good to get early alerts by batch system experts when something changes in the batch system accounting
    - Operations Coord may be one place to share problems before they reach production and document best practices (like cgroups configuration)

Cloud accounting: updated Usage Records to allow benchmark and structure within sites, infrastructure now in place

- Monthly reports for long running VMs too
- T2 view displays cloud usage but only 5 reporting cloud usage (1 for LHC VOs). No need to be part of EGI FedCloud to be in the WLCG views.
    - Also CERN publishing but no T1s
- Issue: cloud site name not always the same as the MoU site name: can REBUS be adapted to handle it. Probably some cleaning possible now that site structure is supported (no need for multiple sites for multiple clouds).
- Not all clouds report the CPU time used by the VM (or return it equal to WCT)
- Still the problem of matching job-based accounting (by the VO) with VM accounting (by the site): a good topic for the workshop!
- Core used vs. core allocated: currently the number of cores allocated is taken into account when computing efficiency (WC time is scaled by the number of cores)
    - When allocating more cores that effectively used (for memory size reasons for example), the job efficiency will be reported lower

Storage accounting: recently revived

- Recently got new data from a few DPM and dCache sites
    - Useful discussions about problems during DPM workshop: hopefully fixed soon
- Accounting storage view improved: EGI plans wider rollout in February
- Still need to fix how VO FQANs are counted against the VO (like for CPU): currently they are not

SL5 support for SL5 box will stop end of March 2016

- SSLv3 (Poodle) is not yet blocked at EGI Message Brokers but may be in the future: sites should upgrade their APEL box to SL6 asap
    - ♦ No easy way to identify SL5 sites

INDIGO Datacloud work: developing a RESTful interface to extract data

- Will be restricted to approved clients

Information System : Future Use Cases - M. Alandes

BDII news

- BDII development: no new release since Sept. 2014, CentOS 7 BDII ready for inclusion into UMD4
- BDII slapd issue: fix provided by RedHat and successfully tested in top BDII and ARC CE
    - Waiting for the RH release of the new version

Information System TF: lots of discussion, input from all the infrastructures and experiments

- 2 documents produced: current use cases, future use cases (see slides for URLs)

GLUE 2: time to simplify and to move away for GLUE 1.3 usage

- GLUE2 will allow to publish cloud resources, something impossible with 1.3
- All the resources published and validated in GLUE2 since a long time: time to consume it!
- EGI is planning to transition the operation tools during 2016
- OSG: ready to provide GLUE2 information in JSON format if there is a motivated need
- Experiments have expressed the feasibility of using GLUE2: no major issue identified. We need an official decision from WLCG to proceed further.
    - Not a problem for any of the WLCG VOs
    - Some help may be needed into migrating existing GLUE 1.3 queries: TF will be happy to help!

TF working on improving the definition of some attributes

- New proposed definitions currrently being checked with the community (LCG-ROLLOUT in particular)

Use cases document

- Information System: a central place owned by WLCG presenting information in a consistent/homogeneous way (service topology, installed capacity)
    - Information provided by infrastructures
    - Information cached and validated
- Proposal: develop a WLCG IS prototype based on AGIS
    - Allready has the ability to retrieve information from all the sources used by WLCG, supporting GLUE2
    - Provides caching and validation
    - See details in the presentation at TF

Discussion

- Ian B.: would WLCG IS be a layer on top of existing infrastructure, that experiments will connect to?
    - Maria: yes. It addresses the current situation where BDII is no longer universally supported and is not the only source of information (GOCDB...). WLCG would no longer need to maintain the BDII infrastructure for itself and would no longer depend on the top BDIIs. May be different for EGI and sites supporting non LHC VOs.
- where do the resources for developing WLCG IS come from?
    - Julia: still being discussed, mainly a validation work will be needed as AGIS is already functional. Effort much lower than if we started from scratch.

- Jeff: sounds all like a good idea, but need to avoid the need for sites to double-publish; also need to think who chases sites for providing correct BDII information.
    - Maria: No double publishing as WLCG IS will query the resource or site BDII. Validation of BDII information is already done by EGI, not WLCG.
- Oxana: proposal means that WLCG will take over an important additional responsibility.
    - Michel: no, we get rid of the top BDIIs.
    - Ian B: would expect that CERN runs it rather than the top BDIIs (partly run by sites).
- Stefan: LHCb is running a similar system; would need a system where the end points are announced (GOCDB?)
    - Ian B: need to ensure that experiments will adopt this WLCG IS, will query the experiment computing coordinators about their commitment to project in January MB

EGI AAI Future - P. Solagna

Currently AAI based on X509 only

- Can hide X509 from the user with a Science Gateway submission portal
  - Robot certificate used
  - Additional information added to the proxy (extensions) to pass information about the user and the VO

New work started as part of EGI-Engage: new layer of authn/authz to simplify management of user credentials

- Both web and non web applications
- X509 and SAML/OpenID
- Identify other identity vetting approaches
- Liaise with other similar activities in other projects, e.g. AARC
  - Trhough AARC, with PRACE, GEANT, EUDAT, WLCG...
- A set of services aggregating information from different identity (eduGAIN , EGI SSO...) and attribute providers and authenticating with the services through X509, SAML or OpenID
  - Federated access must be an enabler, not a barrier
  - Users should be identified uniquely and persistently
  - Flexible support of attribute retrieval, multiple Levels of Assurance (e.g. distingish LoA between self-provided attributes and home organisation provided ones)
  - Maintain access to services based on EGI/VO roles
- Make joining the new AAI platform easy for service provides without a direct agreement with all the IdPs: EGI AAI acting as a proxy
  - EGI AAI will also provide an integrated view of attribute authorities and IdPs to services: support both SAML-based and VOMS-based attributes
  - GOCDB defining site roles, it should also be considered as an attribute authority in the EGI context
- Token translation: hide certificate management from the user with a flexible online CA, in collaboration with AARC
  - Safe and cached storage of credentials, including private keys with MyProxy or SSH keys distribution
  - VOMS proxies
  - Build on CILogon service and other existing services (OA4PM, MyProxy, Shibboleth, SimpleCA)
- Timeline: integration of EGI tools by April, pilot services in July

Discussion

- Romain: fully agree with conclusions, but looks more like a wishlist. Collaboration with WLCG does not exist, same with other claims. Discussion and exchange is really needed. WLCG has been running a pilot for 2 years and had aimed for EGI compatibility, but now EGI seems to have diverged without letting people know.
  - Michel: support for latter point: WLCG communicated a lot about its pilot, including during GDBs and EGI conferences. Stranged that it seemed that EGI ignored this activity. Strange that all partners being involved in AARC, this was not identified: worrying about the project consistency.
  - Peter: accepted. Sometimes there is lack of communication between infrastructures, AARC was/is focussed on architectures. EUDAT has different requirements and goals, thus yet another different implementation (despite architectures quite close) but there will be users for both infrastructures, hence we need compatibility.
- Romain: WLCG IdP does not exist, the WLCG strategy is EDUgain.

    - Peter: EGI.eu joining EDUgain also but will cover EGI.eu employees only. Another issue is local username/password for users without an X.509 certificate.
- Romain: Would a monthly phone meeting between WLCG and EGI, including Hanna from AARC, help?
    - Peter: yes, let's start in January.
    - Michel: need a first discussion before the workshop in Lisbon...
    - Oliver: FTSweb is a demonstrator/pilot application, showing many similarities with Peter's details again calling for closer synchronisation.
- Jeff: discussions tend to focus on low-level technical details that few people understand, but the broad lines are not necessarily clear; a one-page description is required, should ask it from all players involved. Also some of the services branded EGI in Peter's presentation may have a broader meaning. NIKHEF people in AARC will get in touch with WLCG people.

UMD3 and UMD4 - V.Spinoso

SL5 support in UMD will stop by April 2016: all SL5 services must be decommissioned before

CentOS7: only in UMD4

- Main issue: need to create new validation procedures relying on Puppet rather than YAIM (no longer available)
- Need early adopters: first release foreseen early December
- Products currently in UMD4 (or about to be added): Frontier, FTS3, ARC, ARGUS, BDII, dCache
    - ♦ 1st release scheduled on December 17
- SL6 in UMD4: January

UMD preview repository: will replace EMI repo early 2016

- Products before verification/integration
- Monthly update of the repo with information provided by the product team

Discussion Michel: Is it sure that UMD Preview will start early 2016, already announced in the past?

- ♦ Vincenzo: Yes, additional effort has become available to make that happen

DPM Workshop - F. Furano

Well attended, very committed people.

- 60 PBs in 176 instances, largest instance with 3.3 PB
- Good country reports: several common concerns raised
    - In particular SRM issues
    - Several multi-PB instances
- BELLE2: 33% of the storage, interest in ACLs
    - Also using LFC as a file catalog
    - Currently using xrootd as the access protocol, interest in using http in the future

DPM 1.8.10 released mid-October and is the new baseline/recommended version

- Space reporting
- Drain/replication improvements: now using http
- gridftp redirection: a major step for SRM-less storage systems
    - Proved to be more difficult than anticipated due to the way DPM interfaces to Globus for this (using internal interfaces)
    - dpm-gsiftp needs to be rebuilt at each new version of Globus gridftp: will discuss this issue with Globus to find a better solution
- Multiple checksum supports, of different types

Release discussions

- Everything released in EPEL 5/6 currently
- Abandon metapackages: only causing delays
- Versionning: current independent versionning causing pain for sysadmins, looking at moving to a more unified versionning

New DPM web site, new build systems

- New web site based on Drupal
- New build system: Jenkins
    - Painful (unscheduled) transition... but now over

DPM shell (dmlite-shell): the new, feature rich, extendable interface to DPM

http/WebDav: good perfs, 3d party copy, scalable, stable support

- Interesting initiatives to use DPM/http with clouds, federations and S3 backends
- http is becoming mandatory in any DPM installation: required for some internal operations like drain/replication

Space reporting: 'du' like feature through a browser

- Not yet enabled by default: possible source of confusion as it may not stay in sync with SRM numbers

Configuration: Puppet is the recommended, supported way to configure DPM (and even for LFC)

- Standalone Puppet mode usable/recommended as an alternative to YAIM for sites not having a Puppet infrastructure

Related product: DynaFed/DataBridges

- plugin for dmlite
- Provides an aggregated view of http-based storage
- Can be used in cloud context with S3 systems

Future directions: support SRM-less storage, address historical issues linked with old stack

- Get rid of legacy DPM stack: still some dependencies (dpmd, rfio) between the old and new stack
    - DPMRest prototype: implemented as a fastCGI application
    - Implementation about to start... target: release in Q4/2016
- Provide file caching features (lightweight DPM)
- Directory quotas (with space reporting) to replace space tokens
- Checksum request queuing to avoid killing a site with these requests...

Performances: not much work on this topic during last year but CMS extensive stress ("chaos") tests reported that DPM was matching their requirements with proper tuning.

- Markus: Do we have guides for tuning DPM? Are they advertized and easy to found on the (new) web site?
    - Fabrizio: yes a comprehensive tuning guide is available on the old LCGDM web site. Will update slide with the link to guides and check the new web site.

CERN IT reorganization: DPM team joining DSS group but no significant impact expected on plans.

HS06 Scalability Analysis

Salability with LHCb Applications - P. Charpentier

Follow-up for September work, based on a comparison between DIRAC benchmark results and HS06 obtained through MJF

- In September, only 2 sites: GridKA, CERN
  - Rather consistent at GridKA, a bit less at CERN: a few reasons identified
- Today, update with 3 more sites: GRIF/LPNHE, GRIF/LAL and Imperial

Some known issues fixed, in particular number of slots at GridKA

- Minor issue at CERN: inconsistencies between LSF values and MJF
- Also an issue being fixed at CERN with HTCondor: $MACHINEFEATURES not pointing to the right location

Metrics

- DiracPower: result of DIRAC's fast benchmark
- JobPower: renormalized event rate (1/ CPUTimePerEvent)
- MJFPower: single processor power computed from MJF information

Main results

- JobPower/MJF
  - CERN: for most of the CPU types, ratio=1. Some with >1 due to boosting effect in some HW.
  - GridKA: similar results
  - LPNHE and LAL: showing the same lower correlation for some identical machine types
  - Site comparison for the same HW model showing a strong correlation of ratio/HW type
- JobPower/DiracPower: nice symmetric peaks but small tail
  - Larger than MJF when it scales
- See slides for detailed plots

Conclusions

- Still to few sites to draw definitive conclusions: statistics in this presentation based on ~10 % of the jobs executed by LHCb. **More sites providing MJF welcome**.
- LHCb simulation doesn't scale with HS06: big differences for some HW modes, up to 50%
  - Not an operational issue as the tail is always about sites providing more power than expected

Discussion

- Ian: does the application scale better with SpecInt?
  - Philippe: test not done, finding a "normalized" SpecInt information is not easy
- John: empty cores could give some extra performance.
- Alessandro dG: why do we need an accuracy better than 30% for job brokering?
  - Philippe: LHCb is running multiple payloads in one pilot job and is doing job masonry to fill the cores in multicore jobs. The better the precision, the highest the efficiency. But an efficiency better than expected is not a problem for LHCb: just a less optimal use of site resources and a site accounting underestimating the actual resources provided by the site. LHC uses a 30% margin when estimating the time left. The bottom line for LHCb is that all the work done by a non terminated payload is lost.

- Jeff: accuracy more important for LHCb as they want to use job slots for very long periods, longer than other VOs.
- Matthias: for grid site this margin is probably less problematic than for HPC sites
- Alessandro dG: you could checkpoint and upload events from time to time, to avoid loosing the work done by an uncompleted payload.

HS06 Analysis at - M. Alef

P. Charpentier's presentation in Sept. showed underestimated power at GridKA on some machines types when others were fine

- Affected HW types: 2 AMD and 2 Haswell

Investigations done with other VO jobs

- ATLAS: found also that Haswell and AMD had an unexpected HS06/application performance scaling
  - WN node stored in job logs, allowing correlation with machine types.
  - Pretty similar to LHCb simulation results
- Compared with different fast benchmarks with HS06: DIRAC, whetstone
  - Whetstone scaling well with HS06 on all the the processor types studied but having the same issue with DIRAC benchmark and application scaling

Haswell and AMD common features at GridKA: large RAM size per slot (3 to 4 GB per job slot)

- HS06 results analysis: slightly better (7.5%) HS06/jobslot/Ghz with Haswell compared to SandyBridge
  - major enhancement: AVX2 but requiring special compiler flags not used by LHC experiments

Evolution in GridKA config between Sept. and Nov. cgroups used to limit memory consumption (soft limit on RSS)

- In addition to the higer memory per core in last generation of machine
- Also high-memory jobs specifc scheduling

Possible other causes for differences observed by Philippe with Haswell

- faster memory (DDR4)
- Larger L3 cache
  - Wehtstone-double probably fits in L3 cache
- Different compiler flags used by different benchmarks

Discussion

- Jeff: could it be that HS06 fits into the cache of modern processors? More indications at NIKHEF that memory size, speed and usage plays a major role.
  - Agreeement that memory size/speed effect was probably underestimated until now and that we should look at it more closely

GDB Future - I. Collier

Entering a new phase of deployment: a good time to refresh GDB

- Continue to discuss and agree directions and details of implmentation
- Sponsor technical development where needed
- But avoid duplication with other meetings
- Main focus should be in-depth discussions

Exact evolution of GDB format and topic priorities will be decided after the workshop

- January will follow existing format
- February cancelled
- March probably in Amsterdam (clash with motor show)
    - ♦ Exact format not yet clear
- Same for pre-GDBs: what they will become will be decided after the workshop

Summary very valuable: help for note taking is important

Success of WLCG is not of any particular technology but in the ability to set ambitious goals and work together to achieve them without disrupting the infrastructure.

This topic: LCG > GDBMeetingNotes20151209
Topic revision: r1 - 2015-12-17 - MichelJouvin