Table of Contents

Summary of GDB meeting, January 16, 2013

Agenda

https://indico.cern.ch/conferenceDisplay.py?confId=197796

Welcome - M. Jouvin

Meeting organisation

- Looking for more volunteers for note-taking, preferably one per major country. MJ will do the final editing
- Video problems reported to CERN team, not fixed yet
    - News since GDB: hopefully fixed with next version of H323 gateway expected these days

Next GDBs

- March GDB will be external in KIT
    - Registration mandatory
    - Look at Indico for details
- April GDB moved to week before b/c of clash with EGI forum
    - May be canceled later as it is very close to March GDB (3 weeks only, including Easter)
- preGDB in Feb on "AAI on storage systems", pls fill in doodle

Actions in progress: see slides

- multi-core job support : Ian Fisk (IF) CMS plans for scheduling multi-core jobs by end of LS1, hoping to have a multi-threaded framework by Aug/Sept.
- MW client in CVMFS: ready for testing

Clarification of preGDB, GDB and Operations Coordination

- GDB to be discussing on work in progress
- preGDB should be focusing on a specific topic.
- WLCG operations coord is following up actions, in particular those that come out of general discussions in GDB.
- Ian Bird (IB) - we do not have in depth discussions in GDB, in depth shall be in preGDB
- Philippe Charpentier (PC), agendas shall be consistent and no overlap with other meetings, good coordination of topics in agendas of meetings needed, no duplication.

Experiment Support after EGI-Inspire SA3

End of EGI-Inspire SA3 has a big impact on CERN, in particular Experiment Support and Grid Technology groups.

- Respectively 33% and 45% reductions in CERN ES and GT groups

Next framework program at EU: Horizon 2020 with 3 priorities

- Excellent science
- Industrial leadership
- Societal challenges

Final funding of Horizon 2020 is stil under negociation and unclear

- Adoption of legislative acts by Parliament and Council: mid-2013
- Start Jan. 2014: there will be a 1 year gap

CERN has ongoing commitments to EGI.eu and NGIs, as we expect EGI to support WLCG

- Inline with what we need

EMI

- Complex management/coordination structure no longer needed
- IT needs to maintain its effort in DM

Effort available left should be refocused on activities benefiting to more than one experiment

- ~24 people left compared to ~40 today
- Probably less people labeled for one experiment: an issue for LHCb and ALICE in particular
- WLCG Operatons Coordination effort
- Dashboard: more and more commonalities between experiments
- Common activities: either coordination by IT or IT-driven
    - Seen as strategic by experiments for the future
- Details of IT reorganization after April has not yet been agreed

What can no longer be supported

- Ganga: hand over to Ganga collaboration
- POOL: now ATLAS specific
- No interest anymore for Gridview, SE-catalog sync: will be stopped

Potential for future projects: only in 2014, must be broader than just HEP

- Must involve industrial partners: HELIX/Nebula example
- Must not be perceived as grid development
- Can no longer be IT alone: must have commitments across the board
    - In particular PH/SFT must be involved
- Should target large project (EGEE-like scale) as smaller project have a too large overhead

I. Fisk: is the activities proposed to be maintained sustainable with the effort left?

- I. Bird: I think that yes but must foster common solutions

Future Computing Strategy - I. Bird

Summary of input to European Strategy for PP

- Contributed by IT and PH

HEP computing needs after 2020 (post LS2) are much larger than today: existing model cannot scale

- We are only at the beginning of the life of experiments
- Vital that funding for T1s and T2 is maintained

Data management is a strong selling point for HEP: must work with other communities to either buid community tools for more communities or move to more standard solutions

- Must include collaboration with industry

Data preservation: some other communities more advanced, must collaborate with them

- HEP has another scale...

e-Infrastructures

- Grid computing has been very successful for us but not for other sciences
- Should generalise this as a more general HEP-wide infrastructure: do not duplicate WLCG for each big facility/experiments (eg. ILC)
- Must understand how new technlogies will help/fit: terabit networks, clouds (commercial vs. private)...

Must also invest into SW to make an effective use of new HW architectures

- Several big issues and some initiatves started, eg. concurrency forum
- Becomes well-known we have an efficiency problem with new architectures

HEP needs a forum where these strategic issues can be discussed

- CERN is planning a workshop on these to kickstart these activities

Recently, LHCC asked for an update of computing models for the the period after 2015

- Explain use of new technologies: improved date management strategies, better use of current and future CPU architecture....
- Timescale: draft report at C-RRB in Oct. 2013
    - Need to have a good draft for discussion by Sept.
    - Will start soon by a discussion with experiment computing coordinators and then probably working groups
- Must coordinate with architects forum: discussions in progress

Discussion

- P. Charpentier (PC): agree with review of computing model but budget seems to be seen as fixed budget for replacing hardware. While LHC is running data increases.
    - I. Bird (IB): I think funding agencies are aware of the large increasing data set.
    - PC heard saying that not only budget but resources will be flat. IB yes one country said that not all countries.

    - I. Fisk (IF): if we continue with current thresholds in 2015 its a very different problem.
    - IB: by 2014 we need to say that we need increasing resources, need to careful how to increase.
- M. Jouvin (MJ) : look also into astro-physics which will take large datasets. They are very far from our distributed computing approach.
    - IB: astro-physics is too close, need to look into biologists, e-health. If we cannot convince astro-physics, how can we convince biologists, etc. In the future this needs to be driven by scientists not infrastructure providers.
    - MJ: experience in France is the opposite. Biologists are easier to convince and more ready to work with us.
- PC what is the commonality with these other sciences.
    - IB they have large data.
    - Jeff Templon (JT) we have to find corners where we can have large impact.
    - IB we shall start now contacting those communities before they have solved the problem for them.

EGI Report - P. Solagna

MW upgrade

- Huge progress in upgrade of unsupported MW (except DPM/LFC/WN) : ~15 sites remaining
- DPM: 32 to be upgraded
- LFC: 2 remaining
- WN: 92 CEs affected but many shared clusters
- EMI-1 probe not yet deployed but ready
- Tarball worker node was made available few days ago: sites providing reasonable schedule will be given a (short) extension if they need it.

Central user banning plans: change proposed an extension to the current "Service operation security policy" but not yet endorsed by OMB as implications for sites were not clear enough

- ARGUS is currently the only available solution: are they possible alternatives for smaller non WLCG sites
    - WLCG sites required to deploy glexec with generally implies ARGUS
    - NGI-level instance that could be used directly by smaller sites?
    - Publish list of suspended users are plain text?

Configuration management tools survey

- Support of YAIM core after EMI under discussion: may b dropped
- Several sites are already using various configuration management tools: site-info.def may disappear
- A survey will be sent soon
    - Expected outcome: collect and sharing of best practices by NGIs/sites

Discussion

- Helge Meinhard (HM) on configuration management: not a single solution for all sites, but Puppet attracting more and more sites, shall we have a common framework? e.g. with DESY, proposal will be communicated on sharing Puppet configs. Also planning a WG in HEPiX dedicating to sharing Puppet knowlegde and config.
    - Tiziana Ferrari (TF): Puppet only machine configuration model but also middleware?
    - HM: not defined yet, but probably yes as most sites are also WLCG sites

T0 Update - W. Salter

CC Upgrade Project: solve cooling issue for critical UPS rooms, increase capacity, decouple A/C for CC from adjacent office building

- Includes mvoing AHUs to UPS: 10 to 15mn, cover most of the power cuts experienced
- Will provide a physical separation of critical systems from non criticals: will imply moving some systems

Wigner Data Center: T0 extension in Budapest

- Should be transparent to experiments/users despite the increased network latency
  - Tests already done by introducing a delay between lxbatch and storage with no impact noticed
- All operations will be done remotely from CERN, except those requiring physical intervention
- Made of 3 rooms that will be used one after the other for a total of 2.7 MW
  - A 4th one available but will not be used
  - 1st room already available, other ones should be completed by June 2013
- 2 100 Gb links ordered: 1 commercial (T-System) and 1 DANTE (very different physical path)
  - Expected by end of Feb.
- Many dicussions on rack and network layout: 5 racks per row
  - Intelligent PDUs in each rack to allow remote control
- Network routers and siwtches are being delivered to Wigner
- CPU and disk servers currently ordered, delivery to Wigner in March
- Draft SLA
- Major opening ceremony in May/June

Business continuity plans

- First classification of services in 3 categories: backup, load balancing, reinstallatioin
- Internal study to see what would required to implement BC at the network level: currently only one network hub
  - Not before LS1 is finished due to the amount of work (eg. new fibres) needed (probably 2015)
- Plan to start with second delivery in 2013 but full BC not before 2015

Summary of Operations Coordination pre-GDB - M. Girone

Meeting longer that usual fortnightly meetings (typically 1.5h)

- Focused on areas in active deployment and integration
  - ♦ Many experiment improvements planned for LS1
- Willingness of sites and experiments to do more things in common
  - ♦ Ops Coordination has a strong experience in helping with this
  - ♦ Need to foster links between OSG and EGI

Clear progress on tasks with people who accepted to take ownership of the activities

- Eg. CVMFS but other TF using the same approach
- Coordination needs to be reinforced by sites and experiments

Security

- SHA-2: main milestone is to have all the SW SHA-2 ready by early summer
  - ♦ Recent good news from dCache which found a solution to provide SHA-2 support without RFC proxies: no need anymore to upgrade both to SHA2 and RFC proxies.
    - ◊ May have to move to RFC proxy in the future anyway but no pressure to do it. Will be done after the SHA-2 migration which is basically the same thing as the EMI-3 migration for services not yet SHA-2 compliant in EMI-2
  - ♦ One last issue with CAs having email addresses in their certificate names.
- glexec : have a fully validated system at scale bu the end of LS1
  - ♦ First get it deployed everywhere by end of 2013

Information System: proposal for a central Discovery Service for WLCG

- Idea: aggregate information from different places into one place
- WG needs to come back with more details about implementation and timeline for a decision to be taken

Data management

- Experiments agreed that FTS3 deployment must be finished: interested by new features
- Catalog improvements planned by several experiments
- Remote data access/federation in progress for all experiments
  - ♦ Currently based on xrootd, some tests with httpd
- Consensus on moving forward with disk and tape separation
- CMS proposal to enable the use of OPN for remote data access from WNs at T1s: first feedback rather positive, discussions in progress

Clouds and HLT farms: see cloud discussion introduction

- HLT resources in the order of 10-15% of ATLAS/CMS grid resources
  - ♦ ALICE HLT farm could be 250k core in 2018
- Lot of commonalities in the strategies for using the agile infrastructure and the HLT infrastructure

Future Work on Virtualization and Clouds

Intro - Michel Jouvin

Today is a follow up for the discussion we didn't have the time to have in December after Tony's proposal for a future work in WLCG.

pre-GDB yesderday has the whole afternoon dedicated to work of experiments about clouds

- Issues with accounting: APEL should be able to cope with cloud accounting for private clouds operated by our community
    - Not possible for public clouds where we don't have access to the cloud accounting service but do we really want/need this into WLCG accounting?
- ATLAS and CMS tests of CERN Agile infrastructure: a common work driven by CERN IT/ES, essentially successful
    - Test jobs and real production jobs, ~200 VMs of 4 cores/8 GB for each experiment, Condor used to instantiate VMs
    - A CPU efficiency problem identified recently by ATLAS but not yet analyzed: almost convinced this is due to some misusage of the resource as it doesn't match any number observed so far
- Plans to use HLT farms as standard computing resources during LS1 by **all** experiments
    - All experiments except LHCb plan to run a cloud to achieve this

Ian Bird's questions that he would like to see discussed:

- Can we use cloud sites instead of grid sites, what would be the advantage?
- Shall we use opportunistic clouds, scientific clouds?
- Who does the integration with the different technologies?
- Who will pay for it?

T0 View - M. Guijarro

Clouds are a fact and experiments began to deal with it

- HELIX/Nebula
- Site clouds at several places: BNL, CNAF, IN2P3, UVIC; HLT farms, CERN IT OpenStack

Must review the assumptions the previous work done (HEPiX) was based on

- Keep as much as possible inline with industry direction

Main topics to address

- Accounting: wall clock vs. CPU, mutiple tenants, integration with APEL
- Scheduling: how to deal with limited resources? How to claim back resources? Fairshare?
- Federated identity management
- Do we need to restrict the variety of clouds we support

Definitely in favour of a WG

LHCB View - P. Charpentier

Use pilot jobs in charge of pulling workload from a central queue

- A pilot job is in charge of running "job agents" that will start actual payload
  - A pilot job can start several job agents
  - A job can be multithreaded

Role of batch system in this model is very limited

- Place a pilot job on WN
- Ensure fair shares: not clear where they are coming from and how they are enforced
  - Some sites enforces it as a max limit, other as an average value
  - Fairshare should ideally give free resources when there is no competition without impact on the future...
- Job/resource monitoring and limiting: not necessarily a good thing...

VMs are interesting if they are allow to run for a long time with several cores

- Not making sense to run a VM per core
- Under the responsibility of the VO to make an efficient use of the resource: mix of application profiles, run // jobs, optimise memory footprints
- Accounting on wall-clock time as with commercial clouds
  - CPU time must also be accounted but not used as the main metrics for fairshare
  - Risk is paying for low efficiency if site is badly configured

Ideal scenario

- Start VM if there are resources available, contextualized for the VO and starting a pilot for the VO
- Start pulling jobs from CTQ based on WN configuration and jobs in queue: requires ability to get information on the WN
- Communication with sites
  - Max time for a VM to ensure fairness when running over pledges
  - VO should commit not to match new jobs if requested to stop with a reasonable grace period (~1 day): could be used to claim resources when there is competition between VOs
  - VM is shutdown if there is no more jobs to match

DIRAC has the ability to instantiate VMs.

ATLAS

- In 2011, tried building Condor pool in cloud: ostly following this solution.
  - VMs connecting to Condor pool.
  - Machines running indefinitely. Could use cloud scheduler.
- Also tried batchless configuration.
  - Pilot running in infinite loop.
  - Killing machine if needed. Part of lxcloud project.
  - No need for pilot factory, or Condor pool.

CMS - D. Collins

CMS is happy with the current resource allocatin system: any change acceptable if it is not broken or made inefficient

CMS active adapting job submission to clouds

- Mostly for peak offload
- Testing done on StratusLab, adapting to OpenStack
- EC2 as the only interface supported

ALICE - P. Buncic

No official strategy for use of clouds in ALICE: more a personal view

- Some initial work started as part of the attempt to use HLT farm for offline computing
  - Based on CernVM family of products

ALICE computing model relatively flat and thus cloud friendly

- Uniform data access though xrootd
- No real distinction between T1 and T2
- Single task queue
- In the future would welcome pure cloud sites if they offer an API compatible with public clouds (EC2)
  - No need for a batch system: job agent (cloud agent) can interact directly with task queue

Vision: clusters on demand instantiated on various clouds

ALICE expects to live in a mix grid/cloud world for a while.

Discussion

There was a long and rich discussion on many different aspects, below is a summary of the main points.

Two parallel infrastructures or a transition from grid to cloud for accessing computing resources: what's the vision?

- LHCb: currently a batch system is underused, would prefer to have a simpler interface. VMs are more flexible.
  - Just need an interface to start/stop VMs: preferrably a common one for all clouds. Currently using OpenStack interface at CERN but DIRAC also has (or is developping) backends for other cloud MW
  - I. Fisk (IF): we shall agree on one standard which should be EC2
- CMS: pretty good situation to submit to EC2 interface, but profiting from work on Stratuslab, Helix Nebula, open science grid
  - IF: we shall not put grid infrastructure on top of clouds
- Predrag Buncic (PB): we shall use the cloud infrastructure as an opportunity to consolidate our current model, removing software layers
- ATLAS: we need some time to test and integrate the system
- Helge Meinhard (HM): will simplify world if LHC VOs would go to virtualised infrastructure.
- Too early to say if a cloud can be an alternative to a grid CE for accessing computing resources but all the experiments doing some work in this direction
  - The recent work is more than using opportunistic resources
  - This is something quite different from using public/commercial clouds for bursting out load

Fair shares in clouds

- J. Templon (JT): no need for fair share when you are on a resource you own entirely (a private cloud like HLT farm or a commercial cloud where you buy the capacity you use). But in a shared cloud without a fair share mechanism, VO would loss a significant percentage of their resources. Need to keep a batch scheduler in this context.
    - I. Bird (IB): there is a scheduler for provisionning VMs in a cloud MW. Currently it is at a simpler level than a batch system but a lot of work in progress to get more sophisticated schedulers for VM provisionning.
- HM: we shall not have fixed partitioning of infrastructure to VOs.
    - PC: no, we need to re-establish the "fair share" but not too often, e.g. every week.

VM lifetime

- P. Charpentier (PC): VMs should not live forever but for long, longer than current queues, why need to restart a VM/pilot if work needs to be done. But for this to be acceptable for sites, a mechanism must exist that allow a site to claim back allocated resources whatever the reasons (site maintenance, "fair share" implementation between VOs...)
- IF: what shall be the rebalancing time?
- Several sites pointing out the risk of inefficiency of too long life time
- M. Jouvin (MJ) shortened the discussion on this topic: not the main objective for today, many other important issues to address first, avoid to restart the long standing debate about fair share in batch systems
    - Try to define some concrete experimental work to assess if the proposed approach (like Philippe's one) may work

Whole node scheduling

- PC: this is an opportunity to go to whole node scheduling
    - For making it acceptable for sites and making the VO responsible for the efficiency, should move to wall clock accounting rather than CPU accounting for pledges
    - IB: can be done by applying a conversion factor rather than changing the rule in the short term
- IF: pilot we could use it as transitioning to cloud infrastructure, pilot framework will decide what to do with a number of cores available in a VM.

What about cloud storage strategy?

- IB commercially available on Amazon, others not, use it as cache but probably not further at the moment.

Miscellaneous

- What will be the effect of moving responsibility for nodes from sites to VOs that are responsible for what happens inside their VMs?

Conclusions

- IB: we converged on a plan on what to do, even though we are still quite far away from a precise work plan
    - Proposing future project using Cloud technology will attract EU money where grid middleware won't
- Gonzalo Marino (GM): we are at the stage to make technical proposals e.g. for sites supporting multiple VOs e.g. for accounting, sharing resources.
- MJ: will try to summarize the rich discussion and organize a follow-up discussion with people interested in trying to establish a concrete work plan to progress on these issues

- Need both experiments (now all working on this topic) and sites (volunteers welcome in addition to T0)

This topic: LCG > GDBMeetingNotes20130116
Topic revision: r4 - 2013-01-22 - MichelJouvin