Table of Contents

WLCG Workshop, Barcelona, July 7-9, 2014

Agenda

https://indico.cern.ch/event/305362/other-view?view=standard

WLCG Status

WLCG Operations Report - A. Sciabà

See slides

Hot topic: reduce the efforts required for operations, in particular at sites

- "Virtual sites" concept: sites providing HW resources like a cloud for example, experiments managing them
- ATLAS+LHCb expressed a concerned about the impact of experiments taking the sysadmin role

Discussion

- Simone: "Virtual site" idea was launched one year ago as a medium/long term perspective. One year after, this is now short/medium term goal. Any progress
    - Michel: discussions in Copenhagen were about an evolution for the next 5 years, still at the beginning of a long process. But a serious risk of lacking short term milestones: need to make progress year after year. This is one of the main topic for this workshop: understand what are the real challenges, set short term milestones.

T0 Status - M. Barroso Lopez

Facilities

- Wigner is usual business now!
    - Additional capability being installed, mainly for OpenStack and EOS
    - Also some resources for business continuity
- Have not been able to reproduce the 100 Gb network problems: solved after cleaning the fibers?
- CERN decided to base SLC7 on CentOS rather than SL

Network

- LHCONE bandwith at CERN increased to 30 GB, working a AUP
- IPv6 deployed at CERN, several core services already dual-stacked
    - Several T1 with IPv6 connectivity over LHCOPN

Cloud

- Still growing: 2800 servers, 7000 VMs
- Work on SSO, Kerberos integration, accounting with ceilometer

Grid core services

- VOMRS to VOMS-Admin migration still on hold: waiting for a new release
- LFC: Atlas decommissionned, discussing with LHCb for the end date of their server
- FTS2: agreement to stop it August 1st
- New squid service requested by Atlas to cover all the squid usage, not only Frontier

Batch system

- SLC6 migration almost completed: no more submission to SLC5
    - But 20% of HW resources still running SLC5

- Migration from LSF to HT Condor: scalability, dynamism, dispatch rate, query scaling
  - Timeline not yet decided: starting a pilot that will be open to experiments
  - See https://indico.cern.ch/event/247864/session/5/contribution/22/material/slides/0.pdf

Discussion

- Marco C.: why still 20% of SLC5 resources almost one year after the SL6 migration deadline set by WLCG
  - Maite: machines still running SLC5 are old HW, not suitable for delivering SLC6 through VMs (OpenStack). But SLC6 migration done by adding a lot of new resources, so no lack of resources for anybody.
- Claudio: how are VM provisioned?
  - Maite: through static shares per experiment
  - Marco: we are losing the "elasticity" we had with a batch system
  - Maite: dynamic sharing of clouds is one of the big challenge ahead of us, still need to understand how to do it as this is not a standard feature in cloud MW (not a feature needed by most users/providers)

MW Readiness WG - M. Litmaath

Goal: assess that new packages/MC components are not only certified (per PTs) but also ready to be used in a production environment

- Integration with operations
- Integration with experiment workflows

YUM repositories: WLCG is fine with multiple repositories, PTs should the one they prefer

- EMI, Maven, WLCG, EPEL...
- Got commitment on long term maintenance of EMI repositories

New role: WLCG MW Officer

- Andrea Manzi
- Maintain and verify the baseline versions of MW components
- Orchestrates the readiness verification
- Decides when a new version can enter production
- Inform Ops Coord meeting
- Liaise with EGI and UMD for matters regarding WLCG MW

New too: WLCG MW Package Reporter

- New tool developed by L. Cons
- Collect MW packages running at sites, in production or testing
- Keeps collected data encrypted and with restricted access (MW Officer mainly)

Currently selected products for readiness verification: mainly storage products

- Also HTCondor, CVMFS, CEs on experiment request. HTCondor testing will be done in coordination of OSG (who does already most of the work).
- Sites found by Atlas and CMS, volunteering sites preparing a special setup dedicated to MW readiness verification
  - Availability/reliability of these resources will be reported separately: details are still being discussed

- 1st products to be verified to test the procedure/infrastructure: CREAM CE, then DPM

Client tool validation will use CVMFS grid.cern.ch repository

- Clone of AFS area, restructured to be a real mirror

For more information, see https://twiki.cern.ch/twiki/bin/view/LCG/MiddlewareReadinessArchive

Multicore Jobs - A. Perez-Calero

Scheduling multicore jobs require machines to be sufficiently drained

- Draining = idle CPUs
- If a single high priority job comes and take a slot drained, this result in wasted draining
- Minizing the waste of resources required the ability to do an efficient backfilling of drained slots with jobs shorter than the remaining time before the end of the drain
    - Require a good prediction of job duration
    - Entropy, i.e. large variety of job requirements, is needed
- Unfortunately, predicting WLCG job duration is difficult: depends on data acquisition characteristics (luminosity), data access time, pilots running multiple jobs...
- Another important topic to limit waste of resources in particular in absence of an efficient backfilling: keep a multicore slot occupied by multi-core jobs rather than being destroyed
- Would be better to have all experiments agree on a common multi-core slot size

Multicore submission models

- ATLAS: keep multi-core and single-core jobs separate, a pilot will run only one payload to maximise entropy, job duration will be passed to batch system (but still difficult to estimate!)
- CMS: partitionable slots. CMS will be able to use efficiently a multi-core slot, doing the internal optimal filling with multiple jobs
    - The price is that it reduces the entropy in the batch system...

First results

- ATLAS: sites exposed to multi-core jobs since last September
    - KIT (GE): no separate partition, longer wait time for shorter run time, bursty regime
    - RAL (HT Condor): no separate partition, tuning draining rate according to running/queue multicore jobs is critical to limit the waste of resources
    - NIKHEF: dynamic partitioning (Mcfloat), no draining required to support a constant multi-core load, multi-core jobs take longer to start if draining required. Also used at PIC.
- CMS: done at sites already exposed to ATLAS multi-core jobs since May
    - No further adjustments needed
    - CMS approach seems to take a longer time to ramp up but more stable
    - RAL noted lower CPU efficiency with multi-core jobs: 60% instead of 80%
- Combined multi-core jobs from ATLAS and CMS last week at PIC: promising results

WLCG Monitoring Consolidation - P. Saiz

Status

- Process of unification going well: working monitoring solutions
- Huge improvements in operation complexity by using CERN Agile Infrastructure
- Managed to decouple WLCG SAM from EGI SAM

New architecture: see slides

- New database backend allowing very diverse mining activities, including correlation of metrics
- Rich set of views already available
- APIs to build application on top of it
- Integration with site Nagios

Database Services @CERN - Eva Dafonte Perez

Oracle

- Moving to new storage backends (NetApp 62xx)
  - Simplified, consolidated setup: less controllers
  - 2-3x perf improvement, due to flash caching
  - Operation downside: less slots for intervention
- New server provisioned: 100, with 128+ GB of memory and 3x10 Gb connection
  - Managed/provisioned by Puppet
- Oracle 12c gradual migration underway
- New challenging database service set up for the machine, in particular the Quench Protection System
  - 150K rows/s written
- Replication at T1: plan to deploy Oracle Golden Gate and to phase out Streams
  - Everything managed at CERN

DBoD (DB on Demand) more and more popular

- MysQL, PostgresQL (since Sept. 2013), Oracle
  - MySQL being migrated to 5.6
- Main users are IT and PH

Hadoop

- Test setup in place for quite a while but direct use of Map/Reduce proved to be difficult in many cases
- Evaluating Cloudera Impala: SQL interface over Hadoop

WLCG Medium Term Evolution

Clouds - L. Field

At the first glance, a cloud is just an alternative way of provisioning compute resources in a environment using pilot jobs

Image management challenges

- Provides the software, the configuration and do the contextualization specific to the cloud used
- Must balance between pre and post instantiation operations: don't want to rebuild the image for every change
    - ♦ CVMFS is the key component to isolate the base image for the experiment requirements
    - ♦ Main feature required from an image is to be CVMFS-ready: (micro)CERNVM is the foundation
- Transient perspective : an instance is not update but rather destroyed/recreated
- Need for automated image build/management tools

(Fabric) Monitoring is now in the hand of the VOs

- Shift from resource to capacity management: ensure there are enough VMs running
- Require a componenet with some intelligence to decide whether a VM must be started/Stopped where
    - ♦ Standard tools are probably not enough
- Main fabric monitoring goal: check a VM health is ok or terminate it

Accounting and commercial providers

- Helix Nebula as a pathfind project: a great learning experience
- Need to understand how to cross check invoices with real usage
    - ♦ Need to run our own consumer-side accounting: course granularity acceptable
    - ♦ Ganglia is a good tool to base it on: already well adopted in the cloud world

Accounting and WLCG

- By default trust site accounting: no job information at the site level as there is no batch system
    - ♦ Unefficient use of a VM must be accounted to the VO
- Job information in the VO domain
- Dashboard to correlate them
    - ♦ Prototype developed at CERN

EGI Federated cloud: a community of providers sharing a set of required functions

- Common interface: OCCI
- Common accounting and monitoring
- pre-uploaded images: AppDB

State of adoption in WLCG experiments

- All experiments are using their HLT farm but LHCb is not running a cloud on it
- CernVM used by everybody except CMS
- Ganglia used by everybody except ALICE
- BOINC/VAC used by ATLAS and CMS
    - ♦ E.g. ATLAS@Home started last Spring

- WLCG entering production operations

Discussion

- John G.: why to keep the CERNVM name that tends to say that it is specific to CERN when in fact it is not?
- Simone: accounting is really the difficult challenge
    - HS06 is not very meaningful in the cloud context
    - Probably worth a specific workshop/pre-GDB

WLCG Network Outlook - T. Cass

Changes expected before Run2

- Responsibility of transatlantic links moving from Caltech to ESnet

Development expected at CERN before (LS2 during Run2)

- 40G for server connections
- Top of the rack switches with 100G uplinks
- 100G links to T1s and GEANT
    - GEANT connection no longer through switch
    - Major hub for fibres to Europe
- Distributed and layered firewalling
- NaaS over single network infrastructure
    - Not SDN-based at this term: most HW not SDN capable at CERN, standardization still not mature
    - Concentrating on decoupling the network configuration of dynamically provisioned resources (OpenStack) from core services

WLCG evolution before LS2

- LHCONE reaching more remote sites
- Major T1s/T2s adopting 100G
- Improved monitoring infrastructure: already good progress made

Beyond LS2

- 10GbaseT on laptops
- >1G WiFi everywhere: an alternative to copper cabling?
- Cheap 100G
- 400G to 1T interconnection links
- Wide IPv6 adoption
    - Will probably not happen during Run2: based on CERN experience, take a few years from early tests to production
- Possible use of SDN
    - Fine grained security policies
    - Best path for large flows and intelligent load balancing
- Computing models evolving from Tier-based to Storage/Processing roles
    - Storage: full data set per continent, local caches at every site, HSM across the network
    - IPv6 may allow a greater storage-network infrastructure: IPv6 address as dataset ID? Anycast routing for distribution?

IPv6 - D. Kelsey

Exponential growth of IPv6 addresses

- Google reports 20% of IPv6 connection from Belgium
- Microsoft has exhausted its IPv4 addresses for its Azur cloud: moving to IPv6
    - Similar to CERN motivation...

WLCG survey ran last May

- In fact a live table: sites who didn't answer yet can do it!
- 60% of WLCG sites answered: others have probably no plan yet
- Lack of IPv4 addresses: reported by CERN and a few T2s but not by T1s
    - Already 5 T2s reported IPv4 address exhaustion
- Most T1s have IPv6 plans in the short term (next 1-2 years) but only 11 T2s
    - Already 16 T2s with full IPv6 connectivity and 10 with a partial one

Application survey: still a good number of "unknown" status

- Most critical applications with external connectivity are now working

Experiments are now fully engaged

- ALICE move to dual-stack for all core services

Next steps

- All production services dual-stack: already working at several sites
- Join the perfSonar IPv6 testbed
- Test IPv6-only WNs

Network Monitoring and Metrics - M. Babic

perfSonar has now been commissioned in the whole WLCG

- Getting it installed at all sites is just the first step: commissioning NxN links is a squared effort!
- Tracking version changes and updates is a challenge
- Firewalls are a source of issues
    - Need both access to tests and metrics archive

New WG formed to ensure the proper metrics are collected and can be used by operations and applications

- Ensure that we share a common semantics of collected metrics
- Goal is to optimize experiment workflows based on network metrics
    - May include network bandwidth reservations for specific use case
    - NSF project ANSE involving both ATLAS and CMS
- Operations : implement alarming based on metrics collected and topology change monitoring
    - Topology monitoring could be used to influence data source selection

Kickoff meeting planned during Fall.

Data/Storage Management - W. Bhimji

Living without SRM: progress in all experiments

- ALICE never used it
- CMS claims to be able to use non SRM storage
- ATLAS will start validation after the end of the Rucio migration: need an alternative namespace-based usage reporting
- LHCb: work in progress for disk instances, want to keep SRM for tape access

Xrootd data federations in production

- All experiments using fallback for remote access
- Need to incorporate last sites
- Monitoring highly developed but not 100% coverage and underused
    - See May pre-GDB
- Remote read at scale: varying needs between experiments
    - Generally around 1-2 MB/s but ATLAS reported a use case linked to Higgs analysis with a need for 20 MB/s to be 100% CPU efficient
- XrdHTTP available in xrootd v4: allow xrootd sites to expose a http interface

http/WebDAV currently driven by ATLAS

- ATLAS: both user put/get, namespace operations and file deletion
- LHCb plans to use it if it is the best protocol available at site
- CMS has no plan

Commissioning of new standard protocols should trig the removal of old legacy, HEP-specific protocols to remove protocol zoo

Benchmarking activity ongoing

Important I/O developments in ROOT6

- Thread safety
- TTreeCache configurable by environment variables
- protocol redirection

Underlying storage technology evolution

- 6T disks available, new technologies for higher densities but performance not inline
- SSD and hybrid systems
- NVRAM based storage improving but quite expensive

Storage systems

- RAIN now common/standard
- New file systems: Ceph, HDFS
- Algorithmic placement

Protocols evolution

- http2: http with session reuse and smaller header
- Database: new SQL use model coming

Data science interest explosing in industry

- Help may come from outside HEP: see HiggsML challenge

Relaxing requirements: must demonstrate the benefits first

- Good example: data protection for read. Could probably move forward but storage providers/developers must demonstrate the performance benefit first...

Avoton Processor Evaluation - L. Dell'Agnello

See slides.

Avoton: 8 Atom cores in 1 chip based on the last microarchitecture (Silvermont)

- runs native x86_64 code, no recompilation needed
- Very low TDP: 12W
- Used in dense enclosure like HP moonshot: 45 nodes in 4 (4.3) U: reduced cabling

Experiment tests reported good scalability with traditional systems, despite being slower

- Generally reported to be ~50% slower
    - ATLAS reported 3x slower
- Still provides an advantage in term of energy consumption for the same computing power
    - 25% TCO saving estimated on the basis of a 4-year lifetime

Drawback: more nodes needed to achieve the same computing power

- More HW failures
- More manpower to keep all nodes running

Discussion

- Romain: according to CERN experience, exact power saving is very sensitive to kernel/library versions used
- Alessandro G.: discrepancies observed between HS06 and applications is scaring experiments
    - Michel: this is expected as SPEC2006 was not done for benchmarking this kind of architecture. This is one more reason for a new generation of benchmark and a requirements it must address. Expectation that SPEC2014 will cover all the currently existing infrastructures, including Avoton.

Experiment Session

ALICE - M. Litmaath

Run2 facts

- 25% larger event size
- Ressource compatible with flat budget, blessed by C-RSG
- New T2: Mexico (UNAM)
- New T1 at KISTI in production since January

Recent activities

- Improvement of analysis train: critical to improve efficiency
    - ♦ Seen a shift (6% of the ALICE load) from user analysis to analysis train
- SE stability: critical for analysis efficiency
    - ♦ Goal of 98% reliability
- Regular 'inactive data sets' cleanup
    - ♦ Popularity service being put into production
- Network: LHCONE to Asia/South America, IPv6

Short term plans

- Reprocessing of 20122-2013 data
- Cosmics rays data taking starting at the end of the summer for detector commissioning
- No grid data challenge plans: covered by daily operations

Data access model: central catalogue, all data access directly where they are, jobs located where data are

- Analysis train helps increase efficiency of I/O bound jobs: run multiple jobs on the same data
    - ♦ Require 2 MB/s/job to reach 100% CPU efficiency: far from being available at all sites
- SE monitoring relying on ApMon that collects monitoring information from Xrootd servers
    - ♦ A service on each VO box collecting the local data and doing the site aggregation
- SE functional tests: every 2h on redirectors + simplified tests on dynamically discovered disk servers
    - ♦ Monitor discrepancies between declared volume and total space observed
- Closest SE selection (for read and write) based on free space and reliability (+ a bit of randomness to avoid using always the same SE)

Site plans

- Will request upgrade to xrootd v4 as soon as it is stable
- EOS recommended for new sites
    - ♦ Already 5 sites having it

CVMFS monitoring based on MonALISA

- Already deployed at 35 sites

Opportunistic clouds: CERNVM Elastic Cluster

- HTCondor batch cluster started on demand with one command
- Grows and shrink according to load/queue
- Currently used for release validation

Working on installing an OpenStack cloud in the HLT farm to allow opportunistic use of these resources for MC jobs/CPU-intensive jobs

- I/O and uplinks from HLT seen as a potential bottleneck
- Prototype should be started in August

Towards Run3

- Running scenario: continous TPC readout, 50 khz PbPb interaction rate, 1.1 TB/s detector readout
- Data reduction strategy based on online reconstruction and compression: up to 20x
  - Store only the reconstructed data, discard raw data
  - 250K cores needed: HLT will have to provide half of it
  - Will require a very large disk buffer at P2: ~25 PB
- O2 (Offline and Online) envrionment based on ROOT6 and ALFA, the common framework developped by ALICE and GSI/FAIR
  - Capable of using GPU...
  - Improved I/O data model
- Virtually joining closed sites (by network latency) into Regional Data Clouds to deal only with a few clouds rather than individual sites
  - Clouds will be responsible to provide the appropriate reliable storage and the matching computing resources
- EOS the core piece of data management: no time to reinvent the wheel, already tested at scale!
  - Would need a scalable global name space to replace the central file catalog

Discussion

- Wahid/Michel: surprised by EOS being recommended for all sites. At last GDB, EOS developers expressed concerns with support but mainly with the fact that EOS is not targeted at small/medium size sites
  - Ian B.: really surprised by this ALICE reiterated statement after the previous discussions that already took place
- Markus: surprised by ALICE being so prescriptive about what must be deployed at site for achieving at the end results in term of efficiency very similar to other experiments

ATLAS - E. Lançon, S. Campana, A. Di Girolamo

Current situation

- 150K concurrent jobs
- Analysis: 50+% of jobs, 80% of access to data

Future challenges in the way to Run4 (10 years from now): deal with a pileup ~150

- Currently unable to simulate it
- Moore law will provide only a small part of the resource increase needed: have to gain 5x by SW improvements

Limitations of current model and tools

- Partitioning of resources: user analysis vs. production, T1 vs. T2
- Memory required for processing events with a higher pile-up
- Decreasing manpower

Run2 improvements

- Already achieve a 2x improvement in the reconstruction time
- Dataset lifetime introduced to optimize/reduce disk space needed: may be at the cost of more tape access/volumes
- Rebalance disk space between T1s and T2s: large T2s having a profile similar to T1s, may use them the same way
    - On the other hand, a lot of small sites requiring a lot of operation effort without contributing: would prefer less and larger sites
- Importance of computing opportunistic resources: HLT, cloud, HPC, ATLAS@Home, T3
    - At the price of more manpower in the experiment...
- HW resource increase should be ~inline with Moore law but have to check detailed replacement profiles at sites...

What to optimize

- Simulation: CPU
    - Integrated Simulation Framework (ISF) based on a mix of full GEANT4 simulation and fast simulation: only regions of interest are fully simulated, an order of magnitude in perf improvement
    - Either longer jobs with larger output files (grid) or shorter jobs for opportunistic resource use
- Reconstruction: CPU, memory
    - AthenaMP is the default in the last release: all production run on multicore
    - Optimizations in the code and algorithms
- Analysis: CPU, disk space
    - Common anlaysis data format (xAOD) to replace AOD and group ntuples: readable both by Athena and ROOT
    - Skimmed data produced with Data Reduction Framework: central analysis based on the analysis train concept

ProdSys2 based on DEFT + JJEDI + PanDA in production with some benefits

- Data loss: faster reproduction capability
- Automatic transient data handling
- Dynamic dimensionning of jobs
- Automatic rescheduling of failing jobs

Rucio migration in the last stage: DQ2 to Rucio migration

- No site impact
- Currently running the Rucio Full Chain Test: since mid-May, still a few more weeks
- When done, will migrate DQ2 to Rucio and deploy new DQ2 clients that are Rucio aware

WebDav tests in progress and comparison with FAX results

- Not yet production ready, instabilities at most sites, being troubleshooted

FAX: 56 sites, 90% of data

- In production: data access failover
- In testing: site overflow (relocation of payload to a less loaded site without copying the data)

Multicore: ATLAS will need 30-50% of the resources for multicore jobs

- All the MC simualtion, part of the reconstruction

Squid * Frontier: actively working with sites to get Squid 3 working * Known issues with a few large files in CVMFS: working to remove them

Clouds

- Currently using CloudScheduler to provision VMs
- Looking at Vac and Vcycle

3 day Jamboree planned December 3-5 at CERN

- http://indico.cern.ch/event/276502

Discussion

- Ian B.: doesn't understand the diminishing manpower, ATLAS collaboration not shrinking, a management responsibility to prioritize computing enough
  - Eric L.: cannot say why but it is very difficult to get any site commitment for the computing tasks nowadays
- Why Atlas is using much more resources than requested/validated by C-RSG
  - Eric L.: need in fact more simulation than "allowed" to put in our request and more full simulation that fast ones. Something we'll improve in the future.
- Simone: how to get request for WebDAV processed by each site?
  - Michel: as for any action like this (perfSonar, SHA2, EMI upgrades), after initial request through GDB and Ops Coord, probably need to open tickets against sites which have not done it yet

CMS - C. Wissing

Important SW performance improvements to stay within the achievable resources with flat budgets

- x3 gained in the SW

Multicore: dynamic partitioning of multi-core slots between single core and multicore jobs

- Transparently handled by pilots
- Assess good CPU efficiency up to 4 cores per jobs

Prompt reco during Run2

- Would require full T0 share + 50% of T1 capacity
- Will also use HLT farms in "inter-fill mode"
  - Investigating checkpointing mode for stopping VMs immediately without loosing the work in progress
  - Network to CERN CC has been upgraded to 60 Gb/s

Planning to use opportunistic computing resources, as everybody

- Required to do physics beyond what is allowed by the pledges
- Includes T3, academic and commercial clouds, HPC (e.g. NERSC, SDSC)
  - HPC quite complex as it is using specific Linux flavors and has no grid interface for data management

Data management

- AAA in full production for failover
    - 42/52 T2 sites already part of AAA
    - Test at scale during CSA14
- Create and remove dataset replica based on data popularity
- Disk/tape separation completed: T1 resources now open to end users without the risk of accidental recall from tape
- Changed from the disk space managed by physics group to a simpler model with 60% of disk space controlled centrally and 40% of unmanaged space * Unmanaged space is really unmanaged! Users responsible for what they put there

Discussing a new data format to replace AOD and ntuples: miniAOD

- Will test it during CSA14
- Decision taken afterward

CRAB3: a new end-user analysis tool

- Will be exercised during CSA14

Forward looking: real challenge will be Run4

- Not much changes expected during upgrade phase 1 (Run3)
- Just began to think about impact on computing: a workshop run recently
    - New analysis approaches based on MapReduce paradigm?
    - A few specialized Computing Centers (providing GPU or high-end computers for specific payloads requiring them) and more sites with power and cost effective HW
    - PRompt reconstruction will be very challenging in the pileup context of Run4: already challenging today...

Discussion

- Michel: 40% of unmanaged data is a lot... we know by experience that unmanaged space is always becoming a problem
    - Christoph: aware of this, no solution yet, would love to have more space under central control!
- Markus: surprised that CMS is thinking on his own about HL-LHC computing challenges, questions seem common to all experiments
    - Christoph: was just a first meeting, idea is really to have WLCG-wide meetings in the future

LHCb - M. Cattaneo, S. Roiser

HLT strategy changes will be the main changes in Run2 * Deferred HLT since 2012: take advantage of the time LHC is not delivering collisions (60-70%) to use the HLT farm to processed temporarily parked data * 2015: will implement a split HLT architecture that is a generalized deferred HLT strategy, will allow the prompt offline reconstruction to have the same quality as the most recent 2011-12 reprocessing * HLT1 will run all the data to the HLT farm disk buffer * HLT2 will run on the disk buffer with the calibration data available: closer to offline * Calibration will be done on the HLT farm instead of the grid: no need to wait 2-3 weeks before getting the reconstructed data and no more need for the 2-3 week data buffer * 12,5 KHz of data: 5 Khz processed as of today, 5 Khz parked for processing during LS2, 2.5 KHz for high priority processing on HLT farm

Reconstruction: no reprocessing until LS2 due to the improved prompt reconstruction strategy

- Not RAW tape access until LS2

- In practice, some iterations expected with this new approach with the early data in Run2

Stripping: due to previous changes, 2-3x times events/LHC-seconds to be written

- No choice but to reduce event size: move from 70 to 90% of microDST data
- Restripping planned twice a year: require accessing the full DST (FULL.DST), stored on tape, heavy load on tape
- Recalibration: LHCb DST format allows for recalibration during analysis. To reduce data size a new DST format, MDST. DST, that is of smaller size (all the info necessary for recalibration but less raw data)

Computing resources for Run2 inline with flat budgets except for tape

- Tape needs driven by 2 copies of RAW, 1 copy of FULL.DST and 2 copies of analysis datasets (including MC) for data preservation.
- Worried about this exploding tape requirements

Run3 is the real challenge for LHCb (one stage upgrade): x5 in luminosity

- TurboDST idea: RAW data + result of HLT reconstruction and selection, close to microDST from stripping
- During Run2, would like to explore the possibility to live without offline reconstruction (everything done in HLT) to reduce computing requirements

Data access: xrootd now available and will be put into production soon

- httpd will be evaluated then
- Significant work done to adapt DIRAC for FTS3 clients and GFAL2

New pilot written (v2): improved modularity through the use of plugins

- Same pilot used on all infrastructure
- Rely only on CVMFS for accessing the SW
- Possibility to run SAM probes whose results are fed into SAM

Storage federation: no real plan but a solution for data access failover

- Each job receives a list of all replica available for the files it needs
- Local replical is used first, other replica are used in case of a failure with the local replica

Multicore job support done by adding tags to jobs that are matched against queues

Cloud and similar resources

- Vac used at Manchester
- Vcycle (VAC concept for IaaS) on OpenStack @CERN
- VMDirac used to provision cloud resources
- Using CERNVM3 for virtual images
- Also using BOINC for "corridor resources"

Site News

- T2D commissioning in progress: already 1 PB
    - ♦ More sites welcome to join

- New T1 ramping up: Kurchatov Institute

Would like to thanks site for the excellent response time to GGUS tickets

- 50+% solved in less that 5 days

WLCG Future

EU-T0 - G. Lamanna

Project background

- Several European Funding Agencies (FA) contributing to WLCG have also responsibilities in Nuclear Physics, Astroparticles, Astrophysics, Astronomy and Ligh sources
- Some of the projects they support are very demanding in term of data management and computing
- Most project will run as observatories or facilities open to large communities
- FAs must plan for provisioning data processing infrastructure in a challenging funding context: want to build upon the WLCG heritage
- FAs agree to make their data centers accessible to all domain in a sustainable approach

EU-T0 launched Feb. 11, 2014

- Founders: CERN, CIEMAT, DESY IFAE, IN2P3, INFN, KIT, STFC
- To be extended to new partners and disciplines
- EU-TO "hub of knowledge" made of T0/T1 from WLCG who already have very significant computing infrastructures
- Consistent with the vision expressed last year by the e-IRG (e-Infrastructure Reflexion Group) and EIROForum
    - ♦ Capitalize on the investments in computing infrastructure made in the last decade
    - ♦ Address the current limitations and profit of the important advancements in cloud computing and new CPU architectures
- Bring research communities closer to each other and avoid repetition and fragmentation: share expertize, increase cross-fertilization
- Data and researcher centric: address the user needs and expectations from communities committed to major ESFRI projects

EU-T0 built upon WLCG but WLCG Run2 remains the priority for FAs. Want to promote:

- Integration of alternative authentication mechanism, like EDUGAIN, for communities that don't want to use certificates
- In collaboration with GEANT and NRENs, deploy LHCOPN-like networks for other communities
- Help projects with less expertise in data management and data preservation
- Coordination activities embracing the global e-Infrastructure: already done for several key topics by WLCG or others (EGI, EUDAT), leverage on this
- Convergence with HPC, collaboration with PRACE
- Increased relationships with industry
- Explore synergies with US and Asia

Involved in 2 H2020 projects due end of Sept.

- DataCloud: e-Infra-1-2014 subtopic 4&5
- Data backbone (ZEPHYR): e-Infra-1-2014 subtopic 7

OSG - L. Bauerdick

OSG providing computing resources to LHC (60%), other HEP (20%), other sciences (20%)

- Footprint of 120 campuses
- Staff: 26 people

- Funding: 5 year project, mid-term
    - Review planned next August

Main area of responsibility

- Distributed HTC infrastructure for (large/structured) communities who depend on it for running their workflows and for opportunistic harvesting of free resources by other communities/users
    - Size increasing due to the larger projects, in particular LHC
    - Opportunistic use at 25% level: delivering value to long tail science is one of the main criteria to evaluate OSG success
- User support
- Technology and SW development
- Campus Grids: a key concept for addressing long tail science needs
    - Do not rely on structured communities/VOs to get access to resources
    - From the Campus, get access to the whole OSG: BOSCO, OSG Connect Service

OSG Connect: OSG runs the services for campuses and their users

- Encapsulate all the main grid services
- Allow to use the identity management services of the users/communities
- A vehicle to extend OSG ecosystem to HPC centers and to use campus grids as ATLAS/CMS T3s
- Also bridging with industry commodity services: Amazon, Google....

Data is the real challenge

- Large VOs have sophisticated data management infrastructures
- OSG has not yet succeeded or event tried to make a full-fledged data service available to long tail science
    - A prototype being done based on iRods
    - OSG Connect has a rudimentary stash service
    - A simple data archiving service started at FNAL

Providing resources to long tail science doesn't mean give "free resources" to anybody: done through parternship between OSG and science project or campuses

- OSG rewarded for its contribution

EGI Future Plans - P. Solagna

Main activities last years

- Security : CSIRT and SVG
    - Incident response
    - Security training
    - Central emergency suspension
- Upgrade and deployment campaigns: probes, documentation, coordination/follow-up
- UMD: 16 updates in the last 12 months
- Service management: partnership with FedSM project * Lightweight SM framework for federated services

Federated cloud offered as a production service since May 2014

- Integration with EGI core services: accounting, monitoring, service registration, X509 authentication

EGI core services will continue to be supported by EGI partners as their contribution to EGI.eu

EGI-Engage: H2020 project (subtopic 6)

- Community engagement: strategy, policy and business development
  - Pay per use
  - IT Service Managemnent
- e-Infrastructure Commons
  - AAI
  - Permanent ID infrastructure
  - Marketplace: business model, pay per use
- e-Infrastructure Common Services
- Knowledge commons: competence center, user support, training (in collaboration with EUDAT and may be PRACE)
- Open Data commons: common initiative with EUDAT and GEANT to provide a federated open data infrastructure
  - Integrated with community IaaS clouds
  - EUDAT long term data preservation services
- Core Infrastruture: operation tools

Involved in several other H2020 projects, cloud related

- Integration with EGI infrastructure

Partnership with several other DCI projects

- GEANT: towards a European Marketplace of services for research and education
- PRACE and EUDAT: currently mostly on operation tools
- EU-T0: starting collaboration in the DataCloud project

Topics for the Future - I. Bird

Many ideas for evolution for Run2, based on Run1 experience, documented in the updated computing models

- Roles of tiers, networking, federated data, clouds...
- Huge effort by experiment to optimize SW
- Still much effort needed to live within likely resources, both HW and personnel

Only 10 years left before HL-LHC... HLC-LHC will bring new computing challenges

- Data rate: from 25 PB/year to 400 PB/year
  - Technology evolution should make it possible
- Compute growth needed: 50x
  - Max achievable with Moore law and flat budgets: 10x
  - Flat budgets are (very) optimistic

No choice but to reduce costs

- Use opportunistic resources as much as possible
- Reduce operation costs: require minimal staffing
  - Do not rely anymore on large EU or similar funding projects
- Move with evolving technology: clouds, IPv6, new proc architectures

Optimisation: the main keyword

- SW itself: SW collaboration/foundation * Meet the growing needs for simulation, reconstruction... * Promote maintenance and development of common projects * Enable the emergence of new projects addressing the challenges ahead of us * Enable new collaborators to become involved * 12 white papers received: a lot of commonality, broad agreement on goals and a lightweight structure, 2nd workshop planned later this year
- Global optimisation taking into account memory, storage, I/O, network...
    - Do not focus only on CPU efficiency: need to improve the overall performance per $
    - Need to come up with relevant metrics
- Some work started at the time of TEGs (in particular storage TEG) and some useful investigation at CERN-Wigner

Clouds

- Many sites are deploying cloud stacks
- Access to opportunistic resources likely to be through cloud interfaces
- Need to clarify our strategy: currently relying on the grid SW but it is not obvious we have the manpower to do it long-term maintenance
    - On the other hand, cloud stacks have large communities behind them
    - Move to cloud as the primary means to submit jobs to our sites? Focus on pilot factories as the site entry
    - We require some sort of scheduling: we are resource constrained

Operational cost reduction

- Reduce the diversity/amount of MW we use
    - Cloud may help in the long term but will make the situation worst in the short term...
    - We have MW we require (security) with missing support
- Simplify site management: reduce the service that are required
    - Explore BOINC as an alternative to access small site resources

Can EC help?

- EC intends to publish a funding call ICT-8 in Nov. 2014, with submission deadline in April 2015
- ICT-8 could offer an opportunity to procure state-of-the-art HW
    - Opportunity for EU-T0
    - EC will supplement by 20% the partner investments
    - Must prepare a cross-national procurement

Find our place in the Big Data world

- Funding for Big Data is focusing on improving the life of people: HEP has no real place here...
- We have most of the expertise required by other communities
- We need to engage with other initiatives, in particular in Europe and USA
    - We need to ensure that our developments can be useful to others
- We also need to be careful that our effort is used in directions that help us

WLCG scope

- Several requests from other HEP experiments (Belle2, ILC...) to benefit from WLCG infrastructure
- WLCG organisation currently scoped for LHC but would be crazy to think that the infrastructure is not common
- EU-T0 is a mechanism/framework that could help in Europe but doesn't address (yet) the global reach of WLCG and HEP
- We have the structure to collaborate: must take advantage of it

This topic: LCG > GDBMeetingNotes20140707
Topic revision: r7 - 2018-02-28 - MaartenLitmaath