LAL + GRIF Site Report

Michel Jouvin, jouvin@lal.in2p3.fr

HEPiX Barcelona, Octobre 9, 2018

Infrastructure

- Currently using the datacentre phase 1 built in 2012/2013
  - 30 water-cooled racks, 400 kW IT capacity (250 kW used)
  - Still working very smoothly: water-cooled racks proved to be a very resilient cooling solution
  - No power redundancy: turned out to be more a problem than anticipated due to the huge construction works in the area
  - Shared by several labs at Université Paris Sud (and a few others outside)
  - PUE < 1.3, low operation cost (~25 k€ including chiller contracts…)
- Extension project presented in Budapest still progressing very slowly…
  - Only due to “administrative problems”: managed by a structure not really agile…
  - Expecting the building work tender to be out in the coming days…
  - Expectation is now to have the extension in production at Fall 2019 (18 month delay!)
  - Extended capacity will be 51 water-cooled racks, 600 kW IT with power redundancy for 300 kW
    - Power redundancy using 2 different HV circuits + dual power supplies on critical machines

Hardware and OS Update

- Linux: SL5 finally over!
  - Last system powered off beginning of September
  - CentOS 7 on most of the service machines (web servers, DB servers, Indico, print servers…)
    - Also all cloud hypervisors (120)
  - SL6 is the dominating version currently
    - Still the main OS for our grid services and resources: upgrade planned in the coming months
  - HW: still focusing on Dell FX2 with integrated, stackable, 10 GbE in the chassis
    - Procurement in progress to phase out our remaining 100 IBM 3550 (8 years old)
- Windows: still a lot of XP boxes (~100) active… (long tail)
  - Mainly used for special cases like DAQ or electronics-related SW: no internet connection allowed
  - W10 is the main version used for user desktops (~100) but still W7 (~75) and W8.x (~35)
    - No active upgrade: done at HW renewal

SELinux

- Report on our early use of SELinux at HEPiX KEK (Fall 2017)
  - https://indico.cern.ch/event/637013/contributions/2749459/attachments/1541908/2422148/First_Experience_with_SELinux.pdf
- Attempt to activate it in enforcing mode on all machines hosting a service exposed to the internet: rather successful so far
  - Most of the services virtualized: SELinux in enforcing mode by default in VM images
  - Use a contextualization script to disable it when the VM starts for the few machines where it is a problem
- Main issue remains applications (e.g. httpd) requiring access to a NFS file system
  - SELinux context defined on the server not (yet) passed to the client
  - Most of our NFS file systems are hosted by a NetApp filer: no support for SELinux
  - If the NFS file system is used for read-only access, defining a context for the mount point is an option

Cloud

- OpenStack based: 120 hypervisors, 3500 cores, 1 PB (Ceph)
  - 10% of cores by old IBM 3550 (8 years old), with a significant perf impact (VM startup time)
  - Ceph: currently Infernalis, migration planned soon to Luminous (already in pre-prod)
  - OpenStack: currently Mitaka, migration to Pike in the coming month
    - Progressive migration planned: VM restart needed to upgrade Ceph client
  - Currently running the basic service (Nova, Neutron, Cinder)
    - Magnum on the todo list to provision Kubernetes clusters (JupyterHub, Spark…)
- User VM expiration: a local tool developed, LeaseIt
  - https://github.com/LAL/openstack-lease-it
  - 3 months by default with VM owner notification (email) 1 month before (with reminders)
  - Users can extend the “lease” with 1 click: VM stopped if lease is not renewed
  - Django + materialize (not based on OpenStack Mistral)

Service Virtualization

- Not a new activity but affects an increased number of services
- Service virtualization done in the cloud
  - Every virtualised service combines a “generic” system image and a persistent (Cinder with Ceph backend) volume that is used to store service configuration and live data
  - Goal: number of images to maintain << number of services (VMs)
  - A unique local tool to start all the VM instance from a short description (image, disk ID, network port ID, number of cores, contextualisation options)
- System image considered as an immutable object: no attempt to save anything in it. Ensure that we always restart from a known state
  - System images are built by our configuration system (Quattor): 1 image per service type (PHP-based web apps, Python/USWGI apps, Indico..)
  - Network addresses allocated by OpenStack (network port in a network pool) and then entered into DNS
- Next step: interactive systems and grid WNs

LAL Miscellaneous

- LAL Indico instance upgraded to v2 15 months ago
  - Moved from a service causing administration nightmares to a service running smoothly with a close to 0 support and administration
  - Thanks to the new backend (Postgres), version upgrades since then are very easy/smooth: no service downtime required, no risk (thanks to transactional DB upgrade)
  - Excellent support from CERN developer team
  - Considering using the room booking system: waiting v2.2 before final decision
- System configuration management still based on Quattor: no plan to change
  - A production instance of Quattor last generation configuration database (Aquilon) started
    - Aquilon presentation by RAL (HEPiX Annecy): https://slideplayer.com/slide/6104750/
  - Existing site templates can generally be migrated as is but a good occasion to review things
    - Most of the templates we rely on are from the template library which is identical in Aquilon
  - Hope to switch a significant number of machines in the coming year
    - Sysadmin life easier: no need to learn the Pan language for configuring a host with existing services

GRIF: Unified HTCondor Pool

- In production since 1 year between LAL ATLAS, LHCb) and LLR (CMS) subsites
  - 5 km distance, 10 Gb connection currently, 40 Gb soon
  - ~3500 cores in each site
  - No attempt to have a preferred site for each VO: really a global pool, with 1 CE per site giving access to the whole pool
  - No job efficiency impact seen
- Plan to extend the pool to another GRIF subsite in Paris (30 km)
  - May require a bit more coupling between data location and WN location: to be evaluated

Distributed Ceph Infrastructure

- Funding received end of 2017 to build a 1 PB (usable) distributed storage infrastructure based on Ceph
  - Resources spread over 3 sites, 5 km from each others: storage equally distributed
  - Assess usability and cost efficiency as the future basis for storage in P2IO labs (HEP, NP and astrophysics/cosmology)
  - Test various “replication” strategy: replication, erasure coding
  - Provide object storage (through Rados GW for S3, Xroot, may be NFS) and block devices
  - Involve 8 labs (wider than GRIF): human challenge of building a team out of it
- Procurement done a few months ago but deployment on hold because of delays in the 100G-based network infrastructure deployment at Univ. Paris Saclay
  - Hopefully by the end of this month…
- Goal: allow each lab to ask the disk servers needed by the projects it supports

GRIF Miscellaneous

- Working on a federated view of GRIF storage
  - Currently using DPM: not clear it can be done without developments or without creating SPOFs
  - Considering using the RAL ECHO approach: Rados gateways in front of Ceph
    - Inline with our investment in Ceph as the building block for storage services
  - Participation to the WLCG DOMA (and DOMA-FR) project
- IPv6: most of the storage dual-stacked
  - LAL is the late subsite: need to validate/improve DNS management tool (SLAM) to avoid manual, unsustainable management
    - https://github.com/LAL/SLAM
  - ½ CEs dual-stacked
  - No problem seen so far: minor issues don't affect production

Orsay Labs Refondation: LAL End of Life…

- Orsay Labs Refondation is a project to merge 4 IN2P3 labs and & 1 theoretical physics lab
  - Mentioned in my site report in Budapest: was the beginning of the process
  - Covering HEP, nuclear physics, astronomy, cosmology and accelerators
  - Potentially ~800 people (600 permanent staff): ~2/3 in technical divisions (computing: ~60)
  - All labs in a 500m x 100m rectangle
- 18 months of bottom-up discussion through phases of topical WGs
  - 300 people in phase 1, 200 people in phase 2
- Milestone this summer: IN2P3 and Université Paris Sud decided they wanted to see the project moving forward with the creation of a new lab early 2020
  - Still a lot of things to do, in particular decide the lab internal organization
  - A small group nominated to drive the elaboration process… but final decision by IN2P3 and University