Computing Resources @IJCLab

Michel Jouvin, jouvin@lal.in2p3.fr

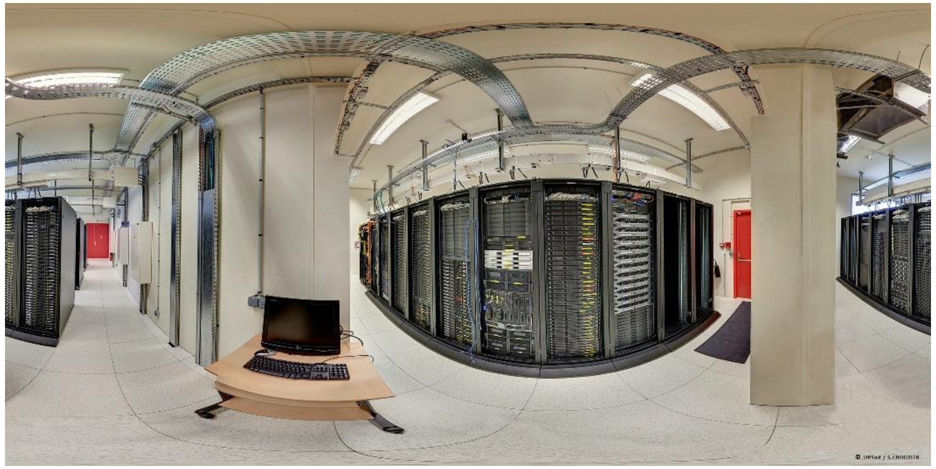
Séminaire A2C

April 1, 2020

Outline

- Physical resources @IJCLab
- Advanced computing resources
- Paris Saclay and national resources

VirtualData Datacenter


© UPSud / S.CROUZIER

- A joint initiative from the P2IO Orsay laboratories in 2011
  - 5 labs which are now IJCLab + IAS
  - Goal: build a scalable and energy-efficient datacentre for hosting our local computing resources in the next 15-20 years
  - Located at building 206: visits can be organised for those interested (after the confinement!)
- Modular infrastructure that can be expanded as needed up to 90 racks and 1.5 MW IT
  - Filling 1 rack requires 300 to 400 k€…
  - Currently : max capacity is 51 racks (22 empty), 600 kW IT, power redundancy for 300 kW
- Mainly used by IJCLab and IAS: most of their computing resources hosted here
  - In particular those presented today
- Also used by several University Paris Sud laboratories since 5 years
  - Growing interest in Paris Saclay (ENS, CentraleSupelec…) to use it after the recent extension

GRIF : the grid resource

- Probably not the most relevant resources for astrophysics and cosmology but was the first large-scale shared computing and storage resource operated by Orsay labs
  - Started in 2005: main driver and local user is the LHC experiments but also other significant users (including astro-particles, HESS/CTA)
  - IJCLab is one of the 4 GRIF subsites with LLR, Irfu and LPNHE
  - Currently 10000 cores (110 kHS06), 8 PB of disk storage: IJCLab hosting 1/3 of them
- Focused on High Throughput Computing (HTC)
  - Can be considered as a huge worldwide batch system: intended to run millions of independent jobs
  - No big shared memory and no low-latency interconnect (Infiniband): not suited for MPI jobs
  - Typical memory/core ratio : 2 GB
  - Good support for multi-core jobs inside a machine
- Very high usage ratio (> 95%)
- Support for containers provides an increased control on the job execution environment

VirtualData Cloud

- A more recent and pervasive computing infrastructure based on virtualisation technology
  - Give the user a full control of the execution environment
  - Dynamic provisioning of resources based on the needs: good basis for shared resources
  - Coupled with a storage infrastructure providing a dynamic provisioning of persistent volumes that can be attached to virtual machines (VMs)
  - Possible orchestration of services requiring several coordinated VMs
- This (OpenStack) cloud is the foundation to deploy the production services at IJCLab and most of the advanced services
  - Main part of the computing resources: 4000 cores with 2 GB/core (similar to a grid WN). A significant part funded by University Paris Sud and external users.
  - 4 servers with a large memory (40 GB/core) recently added, funded by the CLAC project: primarily dedicated to astrophysics and cosmology.
- VMs have a defined duration: users must confirm they still need the VM after this period
  - Allow to recycle unused resources

Cloud Short-Term Evolutions

- Complete the configuration of large-memory servers: started just before the confinement…
- Streamline the support of containers through the Magnum service
  - It’s up and running but deserves some documentation effort…
  - Several container orchestrators supported but we intend to support mainly Kubernetes
- Add a batch system (HTCondor) in front of the cloud
  - Remove from users the complexity of managing the VM system image and provisioning the VMs
  - Provide a queuing feature for accessing limited resources like large-memory machines
- Shared filesystem between VMs
  - Currently it is not straightforward to share data between VMs without relying on an external service
- Spot-instance type of service: VMs that can be killed on short notice

Storage: the P2IO Ceph Service

- Labex P2IO funded a couple of years ago a distributed storage infrastructure between all the P2IO labs (IJCLab, IAS, LLR, Irfu)
    - Storage spread over 3 sites for resilience
    - 1 usable PB initially
    - Use the open-source software Ceph
- Suffered huge delays because of network problems in Paris Saclay who failed to deliver the required 100G links between the 3 sites
    - Solutions/workarounds found recently and being deployed: confinement may not help (again)
- Once in production, will be usable to store physics data that could be processed with the grid or the cloud
    - Despite numbers, still a limited space per project (many projects interested)
    - A foundation infrastructure to add more storage as needed (and funded!) by projects
    - Different access protocols available, including S3

Spark

- Service for parallel processing of large volumes of data
  - Based on Apache Spark open-source framework, an extension of the map-reduce paradigm
  - Strong scalability as long as a large part of the processing can be done in // (map) on “data chunks” (subsets of the data) and that the result consolidation (reduce) phase can remain short
  - Very active ecosystem in the big data world
- R&D in the context of LSST started a while ago by Chris Arnault
- Now a major multi-tenant service in the VirtualData cloud
  - Dynamic provisioning of worker nodes implemented
  - Main user remains LSST with the FINK broker project but open to other uses
  - Particularly well adapted for processing a large amount of “events”
- Astrolab project: Spark-based developments for astrophysics
  - https://astrolabsoftware.github.io

JupyterHub

- A multi-tenant service to run and share Jupyter notebooks
  - Notebook: computational document allowing to mix code, text (e.g. documentation), images (ex: output from the notebooks like plots, images…)
  - Several Jupyter kernels (languages) implemented: Python, C++, possible to add more
  - Authentication/authorization integrated with University ADONIS
  - Run as a University service: used outside IJCLab, including Polytechnique CMAP
- Started 2 years ago and already used at large-scale by several teachers
  - From IJCLab: Jeremy Neveu, Jonathan Biteau, David Rousseau
- Plan: integrate this service with a Kubernetes cluster of containers
  - Allow the provisioning on demand of the compute resources needed to execute the notebooks
- Service evolution suffered from the too many things to do!
  - New use cases are still welcome, in particular on the research side

IJCLab Group Resources

- By definition depends on the group…
    - Also, in the previous labs, the policy used to be different: starting to harmonize…
    - Always small compared to the shared resources: large shared resources means that you always find significant available resources when you need them
- At LAL, started to migrate the group specific machines/clusters as resources in the clouds
    - No replacement of obsolete hardware for a couple of years
- With the CLAC project, funded by IdF region (DIM ACAV+), started to add to the shared computing platform some resources specifics to the needs of astrophysics/cosmology
    - Phase 1 (2019): 4 large-memory servers. Target: large single-core sky simulations.
    - Pahse 2 (2020): applied recently, waiting for the decision by the summer. Target: parallel (HPC/MPI) applications in a box with servers providing up to 250 cores.

IJCLab Computing Department

- Service Exploitation : system administrators with a strong expertise in distributed computing
  - Guillaume Philippon: 10+ years of experience in grid and cloud management, Service Exploitation leader
  - Adrien Ramparison: grid and cloud management with a focus on management of advanced services
  - Gérard Marchal Duval: cloud administrator
- Service Développement: software developers with a strong expertise in parallel data processing and application optimisation for parallel architectures
  - David Chamont and Hadrien Grasland: experts in performance optimisation and performance portability, involved in particular in several major HEP projects
  - Julien Peloton and Chris Arnault: experts in parallel, high throughput, data processing with a focus on Spark technology
- People are generally busy with current projects but are always happy to help!

HPC/GPU Resources

- No such resources owned by IJCLab, except a few machines owned by specific groups
  - CNRS basically forbids acquisition/installation of such clusters outside the HPC mésocentres and national centers (IDRIS, TGCC, CCIN2P3, CINES)
- IJCLab has access to several of these regional/national resources (in addition to GENCI resources)
  - CCIN2P3: cluster de 10 Dell 4130 (20 Nvidia K80) et 6 Dell 4140 (24 Nvidia Tesla V100)
    - Open to any CCIN2P3 user, meaning all the IJCLab members
    - Main GPU environments: CUDA for “traditional GPU apps”, TensorFlow for Machine/Deep Learning
    - A workshop 6 months ago to discuss the possible evolution, based on current and future needs
  - FUSION mésocentre run by CentraleSupelec: a University Paris Saclay HPC resource
    - Currently 6000 cores and 1000 GPUs, funded by CPER, extension requested in next CPER (as for VirtualData cloud)
    - Access possible to non CentraleSupelec users, with a best-effort support: a project needed to test it (discussion started for LiteBird)
  - IDRIS new Jean Zay machine (ML oriented) partly accessible outside GENCI calls

GridCL/ACP for Accelerators R&D

- Small GPU/FPGA cluster funded by Labex P2IO and hosted at LLR, started 6 years ago
    - A R&D and development platform: limited resources to test various approaches, algorithms, applications and assess the performances before going to production somewhere else
    - Periodic hardware refresh/additions: several generations and types of accelerators
    - Open to any P2IO user: no resource reservation, no call for projects
    - Batch or interactive access
- Current resources
    - 2 ser servers with 2 Nvidia K20 each and 64 Go RAM per server
    - 1 server with 2 AMD FirePro S9170 and 64Go RAM
    - 1 server with 6 NVidia GeForce GTX Titan and 128 Go RAM
    - 1 server with the last Xilinx FPGA
- Main asset: a community of experts around the resources that can help (best-effort) porting/developing your application

Paris Saclay: a Rich Context

- Informatique Scientifique @UPSaclay: a coordination around scientific computing, based on what was done at University Paris Sud
  - Several UPSaclay partners interested to share experience/expertise
  - VirtualData cloud and FUSION mésocentre as the 2 main, complementary, resources
  - First meeting planned just when the confinement started: bootstraping this coordination, even before the end of the confinement, currently discussed (2 convenors closely related to IJCLab, Marco Leoni and me)
- Center for Data Science (CDS): last phase of the project runs until the end of the year
  - Knowledge extraction from data: focus on machine/deep learning use for data processing
  - David Rousseau (PHE) is one of the CDS leader
  - https://www.datascience-paris-saclay.fr/contact-2/
  - Groupe mail : https://groups.google.com/forum/#!forum/cdsupsay
- 1 « Objet Transverse » proposed around HPC and performance portability challenges
  - Objet Transverse is “something” transverse to Graduate Schools: multi-disciplinary objects, not well defined yet
  - Maison de la Simulation (MdS) involved: Edouard Audit and me are the proposal leaders

Conclusions

- IJCLab has operated significant shared computing resources for scientific computing in the last 15 years
  - The VirtualData cloud is a corner-stone of scientific computing in Paris Saclay
  - The cloud has proven a very efficient technology to build a shared platform that can easily be tailored to different specific needs
  - CLAC project allowed to start funding resources matching the specific needs of astrophysics and cosmology
- IJCLab has access to several R&D and production HPC resources
  - No real sense to build our own: don't hesitate to request access to them
- IJCLab is part of a rich environment
  - Significant internal expertises: co-location of sysadmins and SW developers is a rather unique advantage
  - Paris Saclay has a lot of other communities with similar computing challenges and is building a framework to foster collaboration between them
  - CDS and MdS can provide help on several of our computing challenges

Useful links…

- https://www.mesocentre.u-psud.fr: main entry point for the scientific computing resources at (former) University Paris Sud
- https://openstack.lal.in2p3.fr: deserve a major refresh but some useful information to access the VirtualData cloud
- http://mesocentre.centralesupelec.fr: FUSION mésocentre
- https://calcul.docs.ipnl.in2p3.fr/documentation/ML/CC_GPU_Farm: CCIN2P3 GPU cluster
- https://www.datascience-paris-saclay.fr: Paris Saclay Center for Data Science
- https://www.maisondelasimulation.fr: Paris Saclay Maison de la Simulation