**Centre de Calcul** de l'Institut National de Physique Nucléaire et de Physique des Particules

# Presentation of the CC-IN2P3

GDR Neutrino meeting – Bordeaux – October, 30 2019

# Plan

- The CC-IN2P3
  - Quick overview of the CC-IN2P3
- Services catalog
  - Computing
  - Storage
  - Collaborative tools
- Some (probably) more interesting tools
  - CVMFS, Dirac, GitLab, GPU, Singularity
- Useful links

**Centre de Calcul** de l'Institut National de Physique Nucléaire et de Physique des Particules

# CC-IN2P3

- Centre de Calcul de l'IN2P3 / CNRS

  Established in Villeurbanne since 1986

- Missions
  - Mass storage and computing infrastructure
  - Network and connectivity
  - Common and collaborative services (electronic mail, electronic document management, software versioning system, projects management, etc.)
- Staff
  - 84 people (engineers, technicians, administration and researchers)

# Buildings

- 2 computer rooms, 850 m$^2$ each

#VIL1 (1986)

#VIL2 (2011)

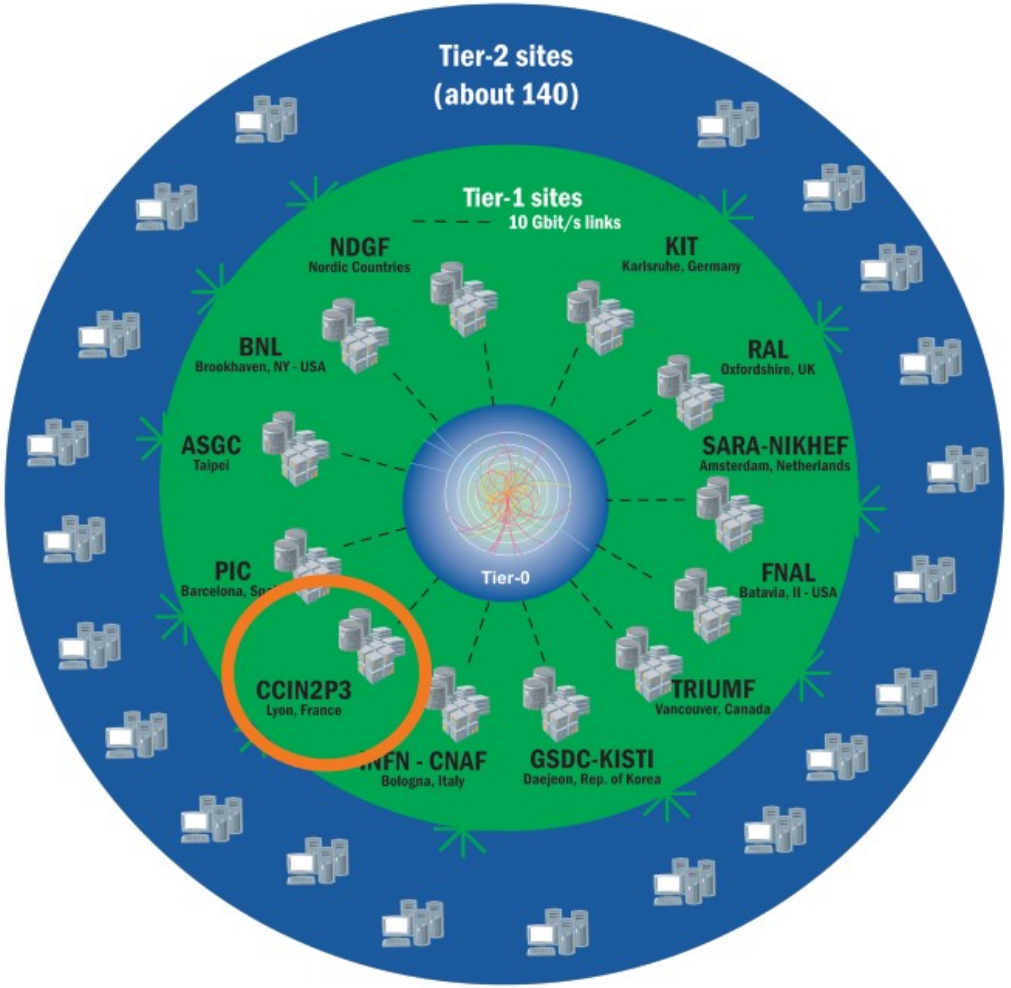
# Worldwide LHC Computing Grid - Tier 1

LCG-TDR-001
CERN-LHCC-2005-024

## LHC Computing Grid
### Technical Design Report

Editor: Jürgen Knobloch

www.cern.ch/lcg




Offering resources for the 4 LHC experiments
Alice, Atlas, CMS and LHCb.

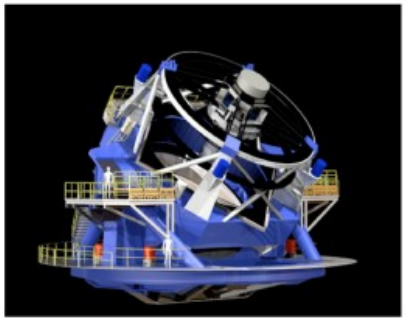
# Also working for…

## LSST

Whole dataset available at CC-IN2P3

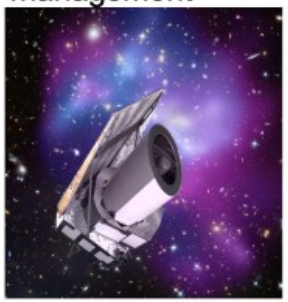
50% of the processing by CC-IN2P3

other 50% by NCSA



## EUCLID

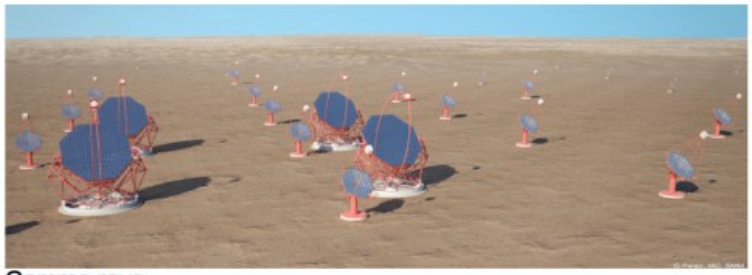
CC-IN2P3 is the French Data Center for processsing and data management



dark energy and dark matter

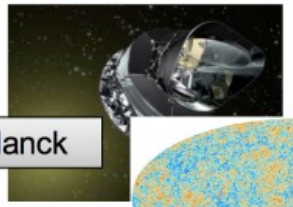
## CTA

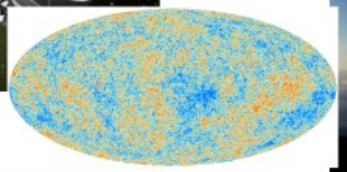
CC-IN2P3 should play a key role in the CTA data processing

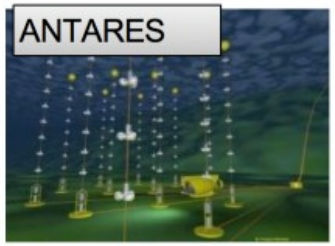


Gamma rays




**Centre de Calcul** de l'Institut National de Physique Nucléaire et de Physique des Particules

# Services Catalog

- Computing
- Storage
- collaborative tools

- High Throughput Computing farm
  - Single and multicore jobs (mostly 8 cores jobs)
  - ~40000 slots, with RAM 3 GB/slot
- High Performance Computing farm
  - Dedicated for OpenMP / MPI jobs
  - 512 cores, RAM 2048 GB
    - InfiniBand interconnect
- GPGPU
  - 40 GPU Nvidia Tesla K80 with 12 GB
    - InfiniBand interconnect
  - 24 GPU Nvidia Tesla V100 with 32 GB

- **Univa Grid Engine** as batch scheduler system
  - Submit jobs to all three clusters HTC, HPC and GPU
  - Worker nodes are running CentOS7 (RedHat7 like)
  - Singularity (container tool) allows to run a job in a different environment (OS, and/or softwares)
    - See later for more details

- Depending on what job you want to run
  - Select the most suitable cluster
  - Select the right batch queue depending on your job profile (CPU time required, memory, ...)

- Documentation :
  https://doc.cc.in2p3.fr/utiliser_le_systeme_batch_ge_depuis_le_centre_de_calcul

# Some more details about **UGE**

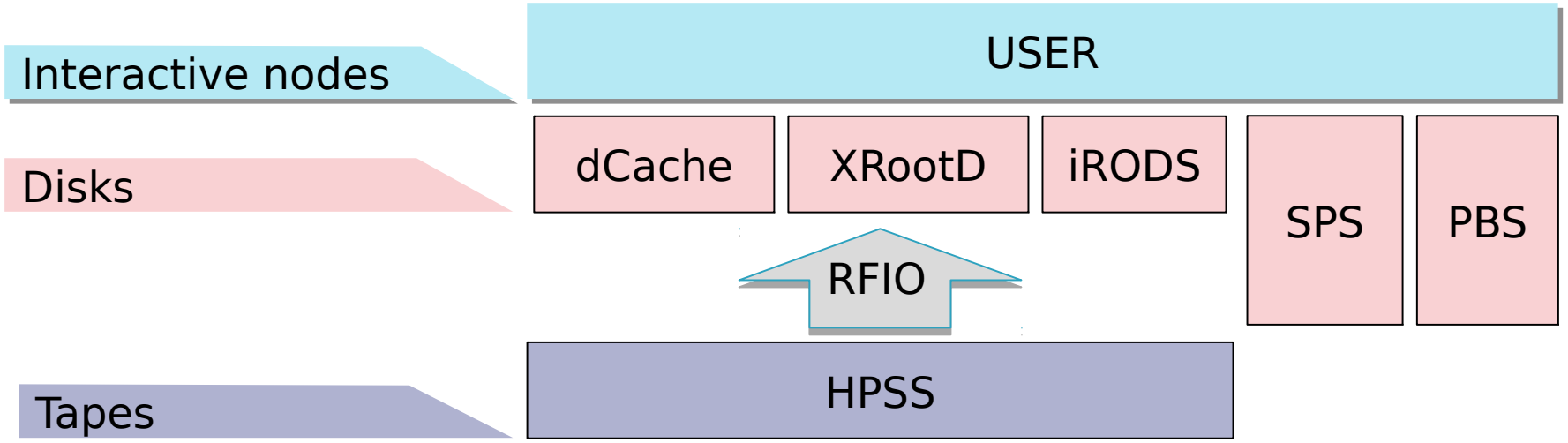
- UGE is a scheduler batch system
  - Orders by priority all tasks then submits them on the computing cluster
- What is a queue ?
  - A queue is a set of global properties and limits
    - Limits: CPU time, memory (virtual and/or resident), scratch disk, ...
  - Several queues are availables, with various limits
    - Check them to find the most appropriate one
- Priority mostly depends on the fairshare
  - Quota of computing resources reserved to you
    - Since the quota is not reached, UGE will submit your jobs asap
  - Quota are discussed every end of year between groups and CC-IN2P3

# Batch queues available

- http://cctools.in2p3.fr/mrtguser/info_sge_queue.php

| Queue name | Host list | Access list (if restricted) | Max CPU time (hh:mm:ss) | Max elapse time (hh:mm:ss) | Max virtual memory | Max resident memory | Max file size | Slots used | Available slots | Used/available |
|---|---|---|---|---|---|---|---|---|---|---|
| admin | @multicores @multiseqs | admins | 03:00:00 | 30:00:00 | 16G | 1G | 5G | 0 | 809 | 0.0 % |
| admin_test | @multicores @multiseqs | admins | 00:05:00 | 01:00:00 | 16G | 500M | 5G | 0 | 809 | 0.0 % |
| demon | @multiseqs | demonqueue | 29:00:00 | INFINITY | 16G | 2G | 2G | 8 | 1207 | 0.7 % |
| huge | @multiseqs | hugequeue | 72:00:00 | 86:00:00 | 32G | 10G | 110G | 1396 | 10525 | 13.3 % |
| interactive | @interacts | | 36:00:00 | 36:00:00 | 16G | 3G | 500G | 3 | 72 | 4.2 % |
| long | @multiseqs | | 48:00:00 | 58:00:00 | 16G | 4G | 30G | 10199 | 34491 | 29.6 % |
| longlasting | @multiseqs | longlastingqueue | 168:00:00 | 192:00:00 | 16G | 4G | 30G | 204 | 3947 | 5.2 % |
| mc_gpu_interactive | @interactsgpu | gpuqueue | 36:00:00 | 36:00:00 | INFINITY | 16G | 250G | 4 | 16 | 25.0 % |
| mc_gpu_long | @gpu | gpuqueue | 48:00:00 | 56:00:00 | INFINITY | 16G | 30G | 36 | 144 | 25.0 % |
| mc_gpu_longlasting | @gpu | longlastinggpuqueue | 202:00:00 | 226:00:00 | INFINITY | 16G | 30G | 8 | 72 | 11.1 % |
| mc_gpu_medium | @gpu | gpuqueue | 05:00:00 | 12:00:00 | INFINITY | 16G | 30G | 0 | 144 | 0.0 % |
| mc_highmem_huge | @highmem | mchighmemoryqueue | 144:00:00 | 150:00:00 | 2000G | 500G | 1000G | 0 | 3 | 0.0 % |
| mc_highmem_long | @highmem | mchighmemoryqueue | 72:00:00 | 72:00:00 | 2000G | 40G | 1000G | 0 | 40 | 0.0 % |
| mc_huge | @multicores @multiseqs | mchugequeue | 72:00:00 | 86:00:00 | 32G | 8G | 30G | 1480 | 9312 | 15.9 % |
| mc_interactive | @interacts | | 36:00:00 | 36:00:00 | 16G | 3G | 500G | 27 | 96 | 28.1 % |
| mc_long | @multicores @multiseqs | mcqueue | 48:00:00 | 58:00:00 | 16G | 3.6G | 30G | 18396 | 34584 | 53.2 % |
| mc_longlasting | @multicores @multiseqs | mclonglastingqueue | 202:00:00 | 226:00:00 | 16G | 3G | 30G | 576 | 20284 | 2.8 % |
| pa_gpu_long | @gpu | pagpuqueue | 48:00:00 | 56:00:00 | INFINITY | 16G | 30G | 0 | 144 | 0.0 % |
| pa_long | @parallels | paqueue | 48:00:00 | 58:00:00 | 16G | 3G | 30G | 32 | 512 | 6.3 % |
| pa_longlasting | @parallels | palonglastingqueue | 168:00:00 | 192:00:00 | 16G | 3G | 30G | 352 | 512 | 68.8 % |
| pa_medium | @parallels | paqueue | 05:00:00 | 12:00:00 | 16G | 3G | 30G | 0 | 512 | 0.0 % |

- Various storage available
  - Various use-cases (what do you want to do?)
  - different ‘Service Level Agreement’ (backed-up or fail-over)
  - Some allow to back up directly into tapes



- PBS : Permanent Backed up Storage
  - Snapshot every 12 hours, backup on tape every 24 hours
  - $HOME (20G personal space)
  - $THRONG (100G shared space within the group)

- SPS: Semi-Permanent Storage
  - High throughput file system for working data
  - NO backup, NO snapshot
  - Space quota varies depending on the group resources request

- iRods : integrated Rule-oriented data system
  - High level overview (user interface)
  - Management of metadata (search feature)
  - Rules for data life cycle (data management policy)
  - Sites federation
  - Migration to tape

- XRootD
  - Performant access to data
  - Large disk cache (5 PiB)
  - Local copy or remote access
  - High scalability
  - Migration to tape

- HPSS: High Performance Storage System
  - Magnetic tape storage
  - NOT an archive (single copy only, no life cycle management)
  - Scientific data only
    - Raw data
    - Long term usage
  - Suitable for files > 1GB

Storing ~71 PB
(full capacity 343 PB)

# Storage summary table

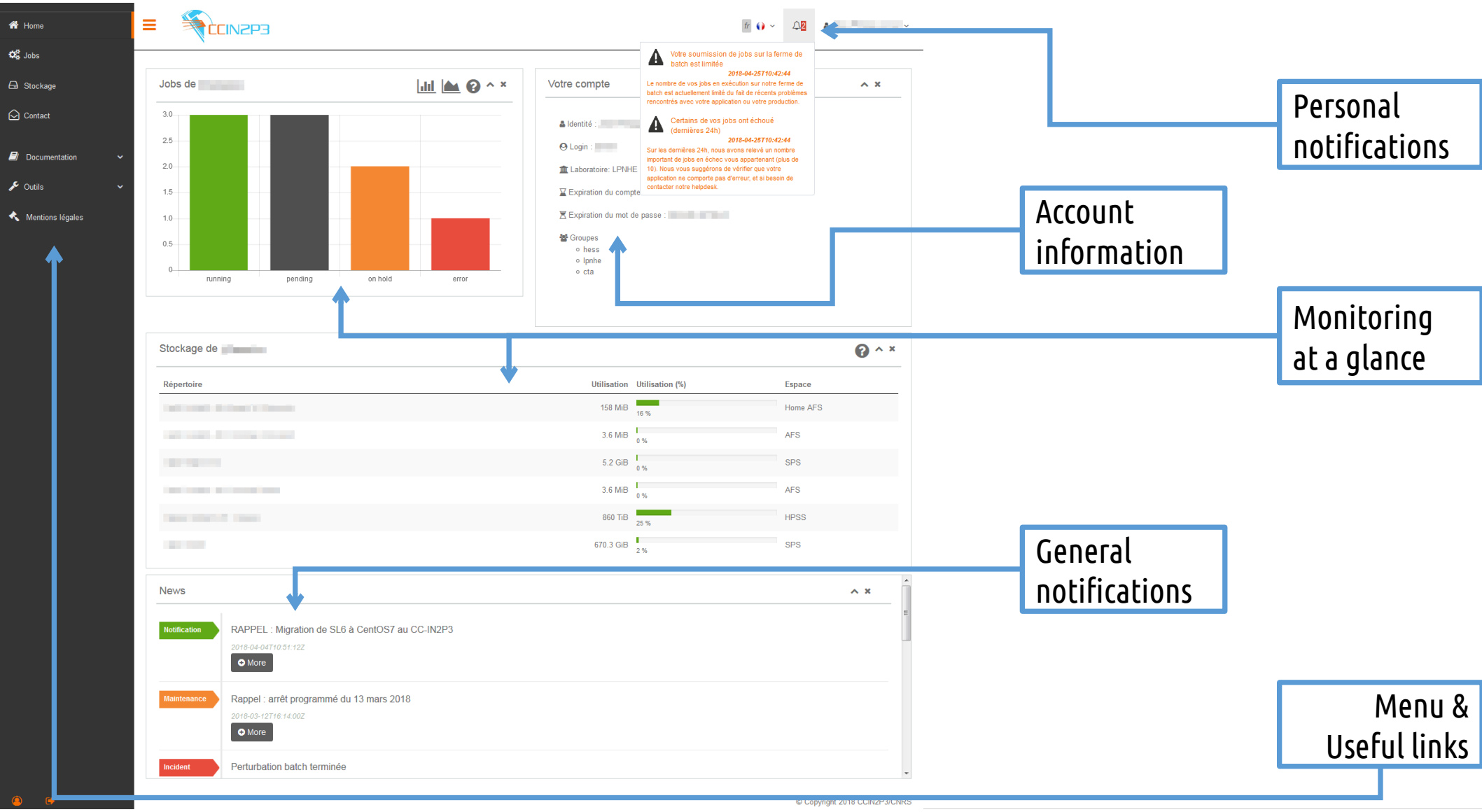
- https://doc.cc.in2p3.fr/en:stockage-et-transfert

| Where is the Data *before* the job? | File type / format | Access | Files / dataset types shared concurrently by jobs | Recommended dataset size |
|---|---|---|---|---|
| **dCache** | Any / ROOT (data) | read AND write<br>non posix (dCap, ROOT, local copy) | read | > 10MiB |
| **HPSS** | Any (data) | read OR write<br>non posix (local copy) | read | > 1GiB |
| **iRODS** | Any (data) | read OR write<br>non posix (local copy) | read | Any |
| **SPS** | Any + binaries + logs | read AND write<br>posix OR local copy | read AND write | ≤ 8GiB : direct access<br>> 8GiB : local copy |
| **XRootD** | ROOT (any) | read (non posix)<br>write (ALICE) (ROOT, local copy) | read | Any |

Ask us: https://cc-usersupport.in2p3.fr

# User portal

- Provides some batch accounting and storage monitoring
- Access with your CCA credential




# Some (more ?) interesting tools

# Collaborative tools

- Events management
  - Indico https://indico.in2p3.fr
- Documents management :
  - Atrium https://atrium.in2p3.fr
- Projects management
  - Forge https://forge.in2p3.fr
- Version control
  - GitLab https://gitlab.in2p3.fr

- Git / GitLab
  - One of the most powerful distributed control versions system
  - GitLab provides a user-friendly interface to do almost everything one can do with Git
  - Collaborative work thanks to the ‘issues’ which allow to track who needs to do what

- GitLab CI (CI : continous integration)
  - Can automate lots of things :
    - Building container
    - Unitary tests for code
    - Nightly builds for software…

https://gitlab.in2p3.fr/help

# Git / GitLab / CI

passed **Job #100156** triggered 3 weeks ago by **Sébastien GADRAT**

```
Running with gitlab-runner 12.3.0 (a8a019e0)
  on ccosvms0239@gitlab.in2p3.fr 96238d4c
Using Docker executor with image python:2.7-stretch ...
Pulling docker image python:2.7-stretch ...
Using docker image sha256:f764be8f15de134ad9abf1c8664da1b958e7b67aa3ab670bfe487b8e66d5b007 for python:2.7-stretch ...
Running on runner-96238d4c-project-5165-concurrent-0 via ccosvms0239...
Fetching changes...
Reinitialized existing Git repository in /builds/sgadrat/c3/.git/
Checking out b68ece72 as master...

Skipping Git submodules setup

Downloading artifacts for export_to_pdf (100155)...
Downloading artifacts from coordinator... ok        id=100155 responseStatus=200 OK token=k2yxFoWF
$ python -V
Python 2.7.16
$ apt-get update
Ign:1 http://deb.debian.org/debian stretch InRelease
Get:2 http://security.debian.org/debian-security stretch/updates InRelease [94.3 kB]
Get:3 http://deb.debian.org/debian stretch-updates InRelease [91.0 kB]
Get:4 http://deb.debian.org/debian stretch Release [118 kB]
Get:5 http://deb.debian.org/debian stretch Release.gpg [2365 B]
Get:6 http://security.debian.org/debian-security stretch/updates/main amd64 Packages [499 kB]
Get:7 http://deb.debian.org/debian stretch-updates/main amd64 Packages [27.4 kB]
Get:8 http://deb.debian.org/debian stretch/main amd64 Packages [7086 kB]
Fetched 7918 kB in 1s (4034 kB/s)
Reading package lists...
$ mkdir .public
$ cp -r * .public
$ mv .public public


Uploading artifacts...
public: found 112 matching files
Uploading artifacts to coordinator... ok            id=100156 responseStatus=201 Created token=c6wasX9c
Job succeeded
```

- Cloud OpenStack
  - https://doc.cc.in2p3.fr/le_cloud_openstack_du_ccin2p3
  - Allows to spawn dedicated virtual machines on demand
  - The user fully manages his VMs

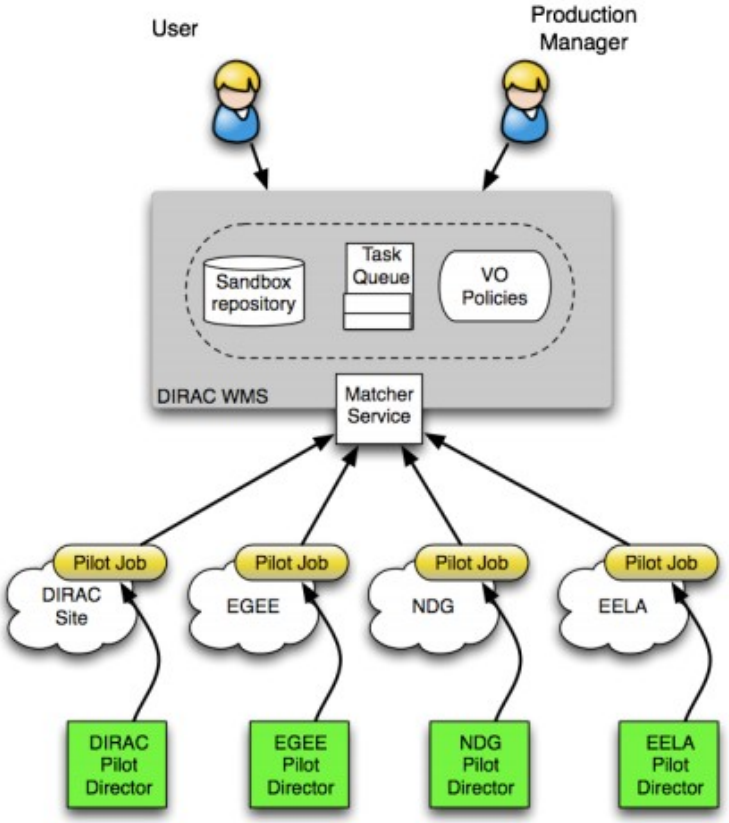
- Singularity (on the computing farm)
  - Allows to run a job in a specific and dedicated environment
    - Different from the default worker nodes environment

```
$ cat /etc/redhat-release
CentOS Linux release 7.7.1908 (Core)

$ singularity exec hello-world_latest.sif cat /etc/issue
Ubuntu 14.04.6 LTS

$ singularity exec hello-world_latest.sif ls /
anaconda-post.log  etc   lib64       mnt   root  singularity  tmp
bin        home  lost+found  opt   run   srv          usr
dev        lib   media       proc  sbin  sys          var
```

- provides high user jobs efficiency
  - hiding the heterogeneity of the the underlying computing resources (federate various sites).
- Uses pilot jobs to prepare the required job environment

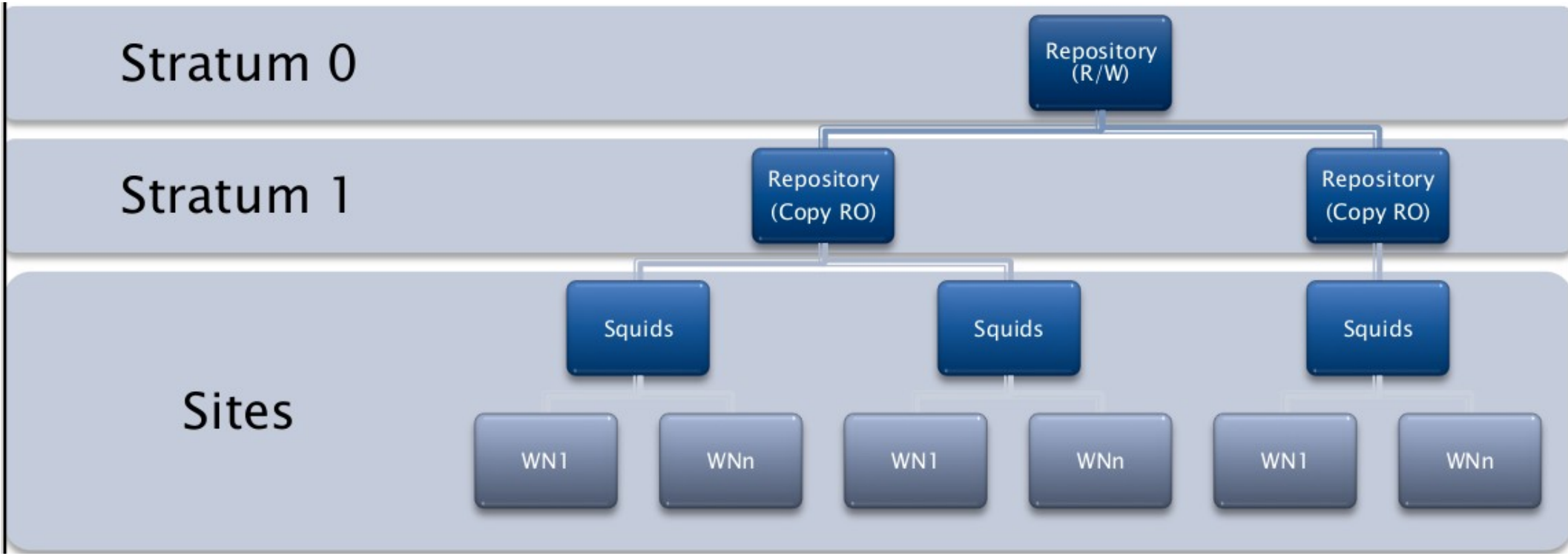



DIRAC
THE INTERWARE

https://dirac.readthedocs.io/en/latest/

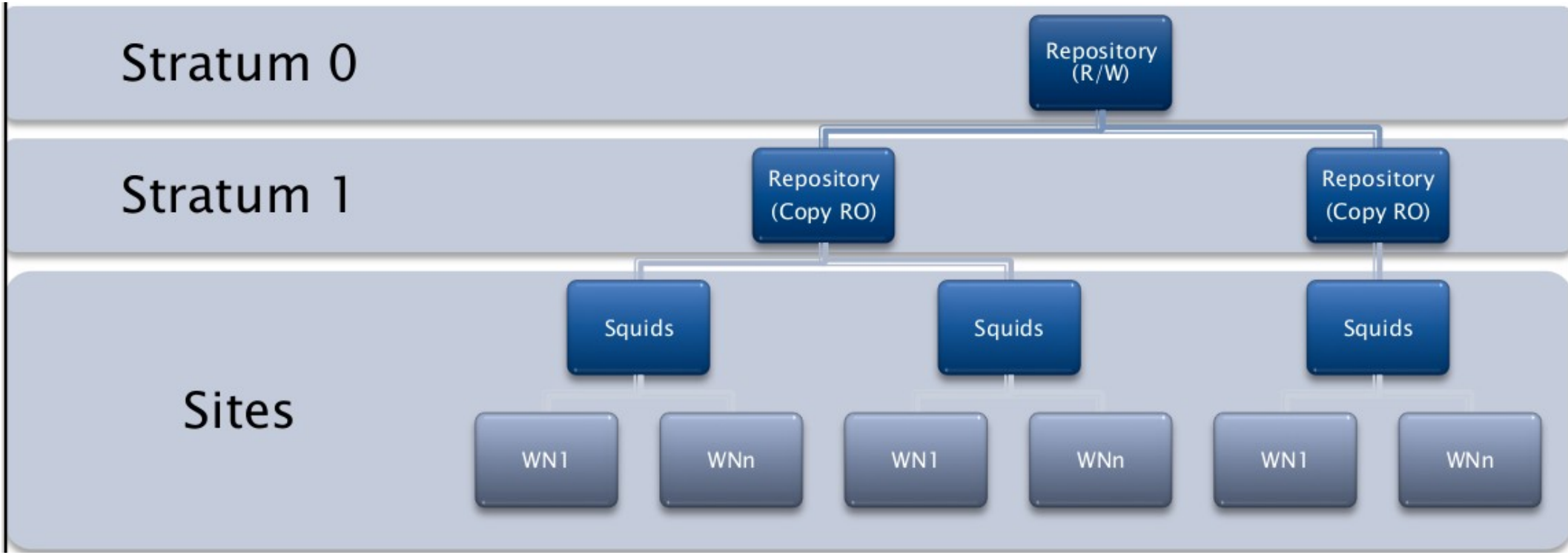
- Dirac scheduler can interface to
  - Cloud: OpenStack, OpenNebula, Amazon EC2
  - Most of the batch system: LSF, BQS, SGE, HTCondor, Slurm, …
  - Grid environment through gLite (grid middleware)
- Central File catalog (data management system)
  - To keep track of all the physical file replicas
- Storage client for the main storage protocol
  - SRM, XrootD, RFIO, …
  - gfal2 gives access to S3, WebDAV, ...

- CVMFS allows to distribute software among sites
  - Software always up-to-date on ALL sites
  - Few people to keep the software up-to-date

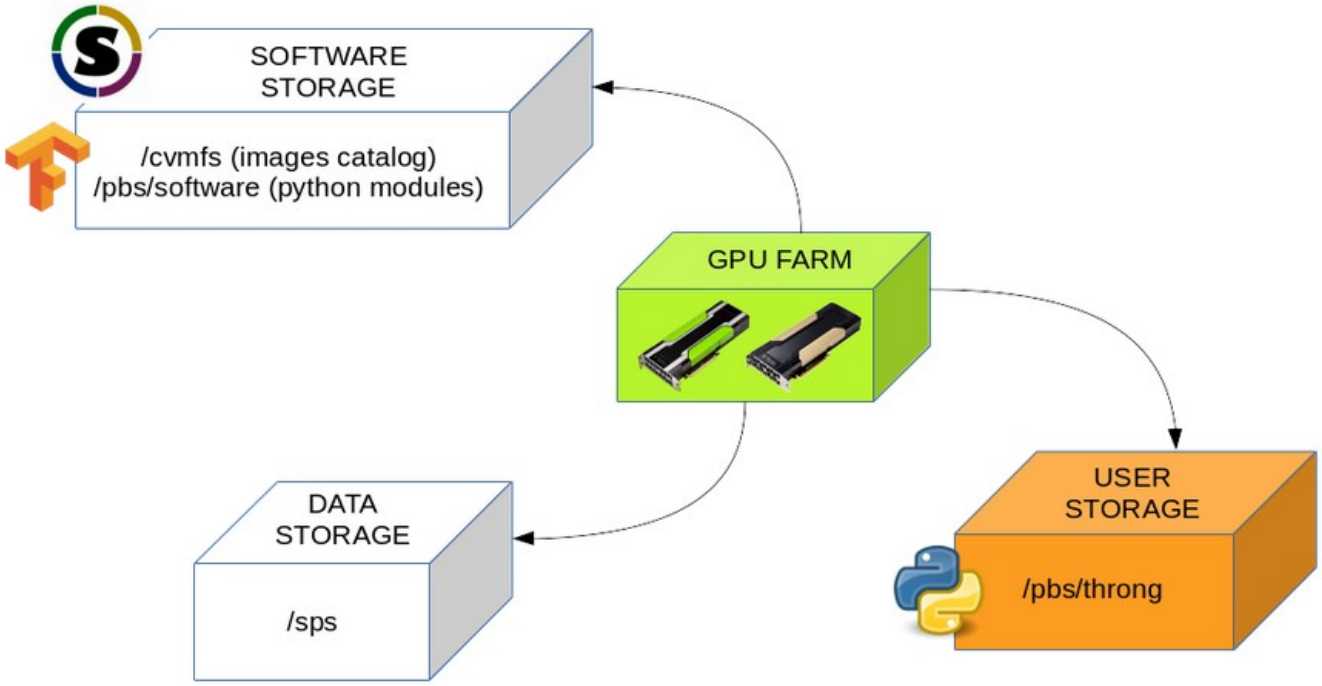


# How to distribute the collaboration software among sites : CVMFS

- Main repository is writable, used to deploy and update the software by the expert (‘stratum 0’)
- Software is then synchronised on the next layer (‘stratum1’)
- Sites provides local caches on the WNs, and Squids will feed these caches on jobs requests



- Two (small) Nvidia GPU clusters
  - K80 and V100
  - Softwares provided : Singularity images and TensorFlow
- Talk@GitLab (talk from a last GPU workshop @CC-IN2P3)



# Useful links

- User portal
  - https://portail.cc.in2p3.fr
- Ticketing system OTRS
  - https://cc-usersupport.in2p3.fr
- Documentation
  - https://doc.cc.in2p3.fr
  - Stay tuned, new documentation is on its way!


- Training @CC-IN2P3
  - https://indico.in2p3.fr/category/857/

# THANK YOU ! Any Question ?