



LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

INCD Software Management

Joao Pina , Joao Martins

LIP Distributed Computing and Digital Infrastructures group



INCD - Infraestrutura Nacional de Computação Distribuída

INCD is a digital infrastructure:

- LIP Technical coordination
- Goals:
 - Provide **computing** and **data services** for the research community
 - Computing Services:
 - Cloud.
 - HTC and HPC (farm)





INCD operations centers in 2023



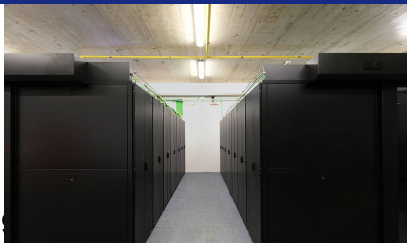
INCD-A @ LNEC in Lisbon
HPC / HTC / Cloud / Federation
6000 CPU cores
5 Petabytes online raw
100 Gbps
Includes the WLCG Tier-2



INCD-B @ REN in Riba-de-Ave
(DECOMMISSIONED in 2023)
HPC / HTC
2600 CPU cores
384 Terabytes raw
1 Gbps



INCD-L @ LIP in Lisbon
Tape storage
1 Petabyte backups
10 Gbps



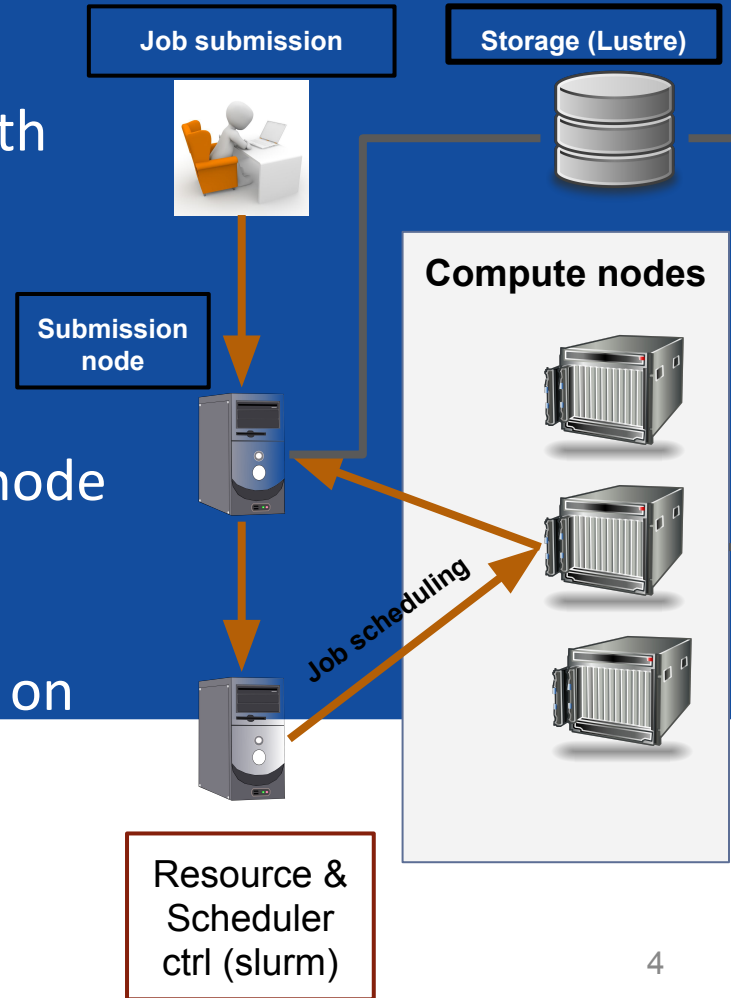
INCD-D @ UTAD in Vila Real
(BEING DEPLOYED)
HPC / HTC / Cloud / Federation
5000 CPU cores + IB HDR200
4 Petabytes online raw
100 Gbps



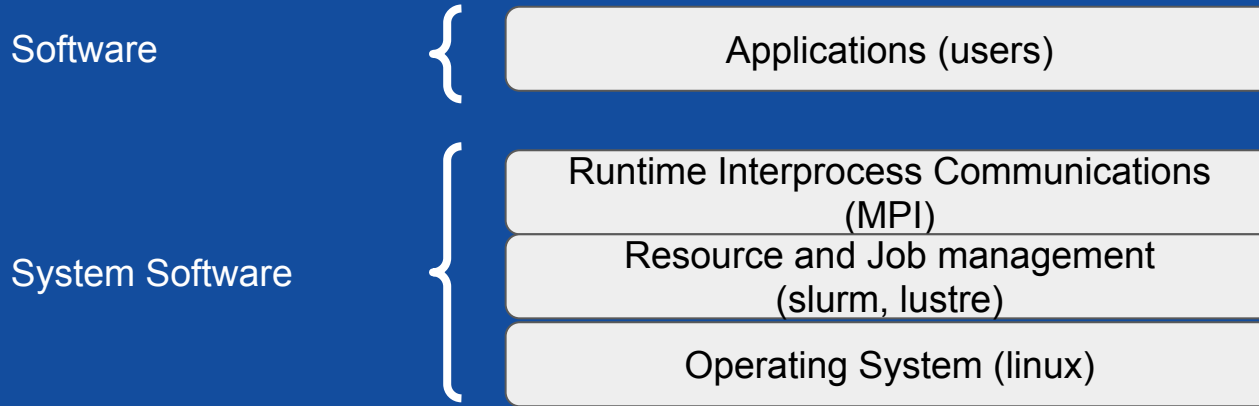
INCD-C @ UC in Coimbra
(BEING RENEWED)
Tape storage expansion
20 Petabytes
10 Gbps

Computing Farm:

- Access through a submission node with ssh keys authentication
- Applications run on compute nodes
- Compute nodes access through job scheduler
- Storage shared between submission node and compute nodes
- Hardware architecture common to submission node and compute nodes on new AlmaLinux 8



Software: Traditional Cluster Stack





INCD System Software

Software: System software

- Operating system:
 - Main distributions: Centos 7 and Alma Linux 8
 - Computing servers;
 - Storage (LUSTRE);
 - Cloud (OPENSTACK);
 - Kubernetes
 - Virtualization (KVM)
 - And many other services.
- Installation and service configuration;
 - Kick Start recipes for server common installations (WN, Virtualization, Storage)
 - Moving to ansible receipts for services (slow move)

Software: System software

- Runtime Interprocess Communications (MPI)
 - OpenMPI and MVAPICH2
 - INCD-A: 56 Gbps (infiniband)
 - INCD-D: 200 Gbps (infiniband)
 - GRID and local users: 1Gbps-10Gbps (copper)
- Storage provisioning through Lustre
 - INCD-D: 200 Gbps (infiniband)
 - INCD-A, GRID and local users: 1Gbps-10Gbps (copper)
 - 3.5 PB aggregate
- Resource and Job management:
 - Slurm
 - kubernetes (new)



INCD Software Applications

Software: Software Applications

- Multi user environment require a flexible setup:
 - Users may require different conflicting libraries and versions of the same application
 - Users may require multiple setups for the same application
 - Heterogeneous hardware architectures may require multiple builds
- We handle this issues with Environment Modules:
 - CentOS 7: package environment-modules
 - AlmaLinux 8: package lmod
- Module files customization for local usage over CVMFS repository mounts:
 - CentOS 7: /cvmfs/sw.el7/modules
 - AlmaLinux 8: /cvmfs/sw.el8/modules

Software: Software Applications

- CernVM File System (CernVM-FS) is a read-only file system on which files and file metadata are downloaded on demand and through standard HTTP.
 - Cache quota management;
 - Possibility to split a directory hierarchy into sub catalogs at user-defined levels
 - Capability to work in offline mode provided that all required files are cached
 - File system data versioning
 - Dynamic expansion of environment variables embedded in symbolic links
 - Support for extended attributes, such as file capabilities and SELinux attributes
 - Automatic mirror server selection based on geographic proximity
 - Automatic load-balancing of proxy servers
- Efficient replication of repositories
- Possibility to use S3 compatible storage instead of a file system as repository storage

Software Applications

- This strategy based on CVMFS allows to:
 - distribute the software and environment from a single central repository to multitude of clients spread locally and geographically
 - have good scalability, reliability and availability
 - easily maintenance of a complex environment
- CVMFS drawbacks:
 - low I/O performance
 - not suitable to share data sets, especially big
 - can not ensure privacy of restricted applications

Software Applications

- Customized per:
 - Operating System: CentOS 7, AlmaLinux 8
 - Community
 - Compiler: gcc, intel, aoc, cuda
 - Hardware architecture
- Over 300 different Software/Compilers builds
- Huge complexity and hard to maintain
- module examples:
 - module avail

module load openmpi/4.1.4

module list

module unload cuda/12.1

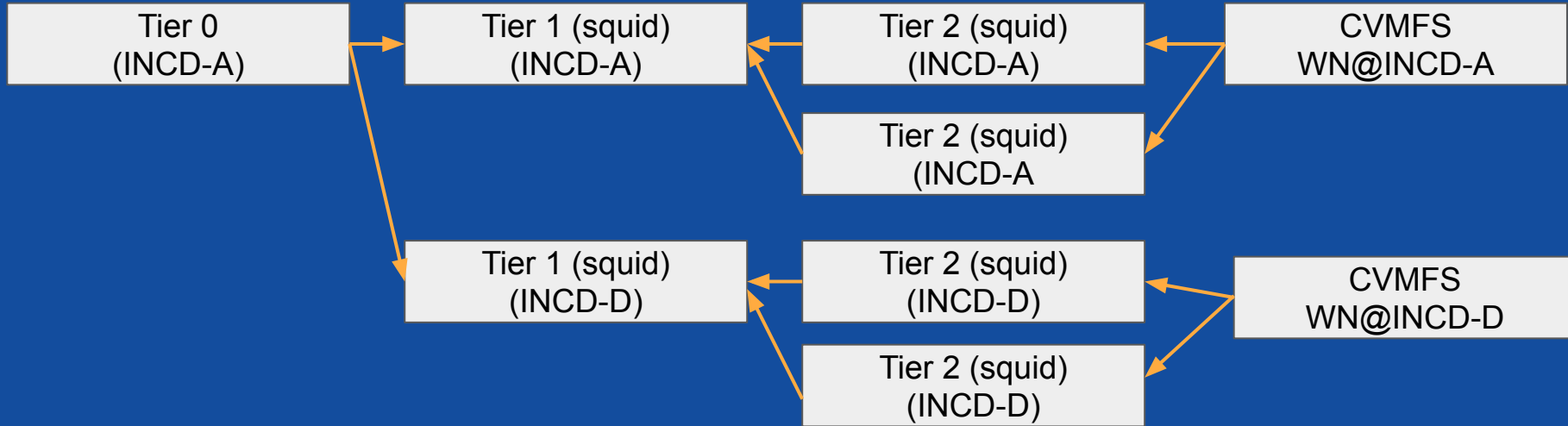
module purge

```
[jplna@cirrus08 ~]$ module avail
-----
aoc-4.0.0 gcc-11.3 gcc-8.5 intel/oneapi/2023 opam python/3.10.8 spack udocker/1.3.9 udocker/alphaFold/2.3.2 /cvmfs/sw.el8/modules/hpc/main
-----
aoc40/R/4.2.2 aoc40/lbbs/aocl/1.1.0 aoc40/lbbs/fftw/3.3.10 aoc40/lbbs/hdf5/1.14.0 aoc40/lbbs/lapack/3.11.0 aoc40/lbbs/nlopt/2.7.1 aoc40/mvapich2/2.3.7
aoc40/gromacs/2023 aoc40/lbbs/blas/3.11.0 aoc40/lbbs/gsl/2.7 aoc40/lbbs/jemalloc/5.3.0 aoc40/lbbs/libpng/1.6.39 aoc40/lbbs/openblas/0.3.21 aoc40/openmpi/4.1.4
-----
gcc85/R/4.2.2 gcc85/lbbs/fftw/3.3.10 gcc85/lbbs/hdf5/1.14.0 gcc85/lbbs/lapack/3.11.0 gcc85/lbbs/nlopt/2.7.1 gcc85/mvapich2/2.3.7 gcc85/netcdf-cxx/4.2 gcc85/netcdf-fortran/4.6.0 gcc85/plumed/2.9.0
gcc85/lbbs/blas/3.11.0 gcc85/lbbs/gsl/2.7 gcc85/lbbs/jemalloc/5.3.0 gcc85/lbbs/libpng/1.6.39 gcc85/lbbs/openblas/0.3.21 gcc85/netcdf-c/4.9.0 gcc85/netcdf-cxx4/4.3.1 gcc85/openmpi/4.1.4
-----
gcc11/R/4.2.2 gcc11/gromacs/2023.5 gcc11/lbbs/fftw/3.3.10 gcc11/lbbs/hdf5/1.14.0 gcc11/lbbs/lapack/3.11.0 gcc11/mvapich2/2.3.7 gcc11/netcdf-fortran/4.6.0 gcc11/openfoam/2012 gcc11/plumed/2.8.0
gcc11/cp2k/2023.1 gcc11/keras/2.10.0 gcc11/lbbs/gsl/2.7 gcc11/lbbs/jemalloc/5.3.0 gcc11/lbbs/libpng/1.6.39 gcc11/netcdf-c/4.9.0 gcc11/netcdf/2206 (D) gcc11/plumed/2.9.0 (D)
gcc11/gromacs/2021.4 gcc11/lbbs/blas/3.11.0 gcc11/lbbs/hdf5/1.14.0 gcc11/lbbs/nlopt/2.7.1 gcc11/netcdf-cxx/4.2 gcc11/netcdf-org/2.4.0 gcc11/openmpi/4.1.4 gcc11/tensorflow/2.10.0
gcc11/gromacs/2022.5 gcc11/lbbs/boost/1.80.0 gcc11/lbbs/jemalloc/5.3.0 gcc11/lbbs/openblas/0.3.21 gcc11/netcdf-cxx4/4.3.1 gcc11/openfoam-org/10 (D) gcc11/openmpi/4.1.4 gcc11/paraview/5.10.1
-----
intel/lbbs/blas/3.11.0 intel/lbbs/gsl/2.7 intel/lbbs/jemalloc/5.3.0 intel/lbbs/libpng/1.6.39 intel/lbbs/libpng/1.6.39 intel/mvapich2/2.3.7
intel/lbbs/fftw/3.3.10 intel/lbbs/hdf5/1.14.0 intel/lbbs/lapack/3.11.0 intel/lbbs/nlopt/2.7.1 intel/openmpi/4.1.4
-----
cuda/10.2 cuda/11.2 cuda/11.8 cuda/12.1 (D) nvhpc-byo-compiler/23.1 nvhpc-nonpt/23.1 nvhpc/23.1 /cvmfs/sw.el8/modules/gpu
-----
autodock-gpu-develop/11.3.0 bismark/0.23.0 fastqc/0.11.9 hmmer/3.3.2 jellyfish/2.2.7 /cvmfs/sw.el8/modules/bio samtools/3.0.0 transdecoder/5.5.0 vcf-tools/0.1.14
autodock-vlna/1.2.3 blast-plus/2.12.0 figtree/1.4.3 htslib/1.8 kallisto/0.48.0 raxml/8.2.12 sratoolkit/3.0.0 trimgalore/0.6.10
bcftools/1.8 bowtie2/2.4.2 freebayes/1.3.6 iqtree2/2.1.2 mrbayes/3.2.7a salmon/1.9.0 star/2.7.6a trinity/2.14.0
-----
gcc11/geant/4.10.7.3 gcc11/topas/3.9 gcc85/gdcm/2.8.9 /cvmfs/sw.el8/modules/LIP
```

Software Build

- Whenever possible we make a native build of applications on target hardware using:
 - configure, cmake and make utilities
 - spack package manager
 - opam package manager
 - dockers/udocker (<https://github.com/indigo-dc/udocker>) for applications demanding different operating systems
- Software installation and configuration base directory over the CVMFS repositories mount points
 - since this is a read-only directory we bind the path to a local directory with read-write permissions, for example:
 - `mount -bind /tmp/app /cvmfs/sw.el8/gcc85/app/<version>`
 - when ready we copy the installation tree to stratum 0 for publication

Software Management Topology



INCD in numbers:

- 2 x Tier 0 (1 TB)
- 4 x Tier 1 (2 INCD-A + 2 INCD-C)
- 8 x Tier 2 (5 INCA-A + 3 INCD-C)
- 150 x WN's (INCD-A + INCD-C)

Resume

- SQUID and CVFMS used for long time to deploy Software to the Worker nodes over several clusters (HTC + HPC)
 - Easy to maintain
 - Resilient
- Future
 - Use S3 compatible storage instead of a file system as repository storage
 - Cloud and Kubernetes

End

Questions?