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**:

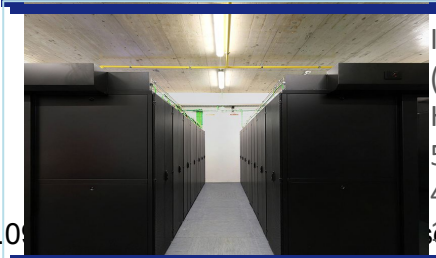
- LIP Technical coordination
- Goals:
  - Provide **computing** and **data services** for the research comm
  - Computing Services:
    - Cloud.
    - **HTC and HPC (farm)**

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

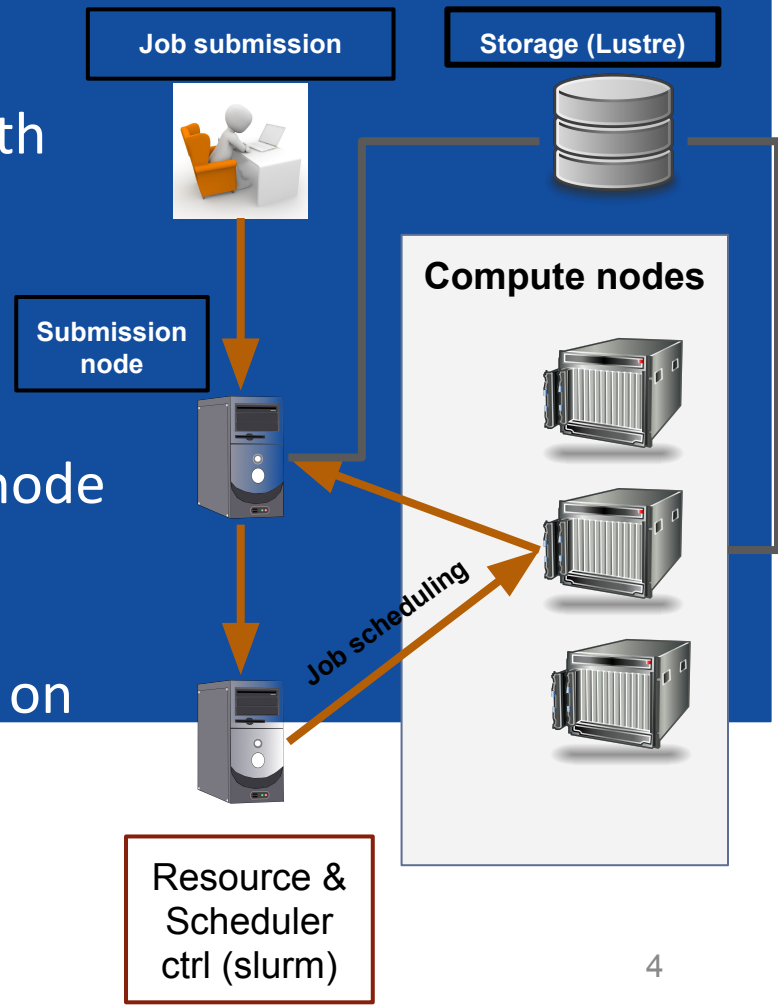


INCD-D @ UTAD in Vila Real
(BEING DEPLOYED)
HPC / HTC / Cloud / Federation
5000 CPU cores + IB HDR200
4 Petabytes online raw
10 Gbps

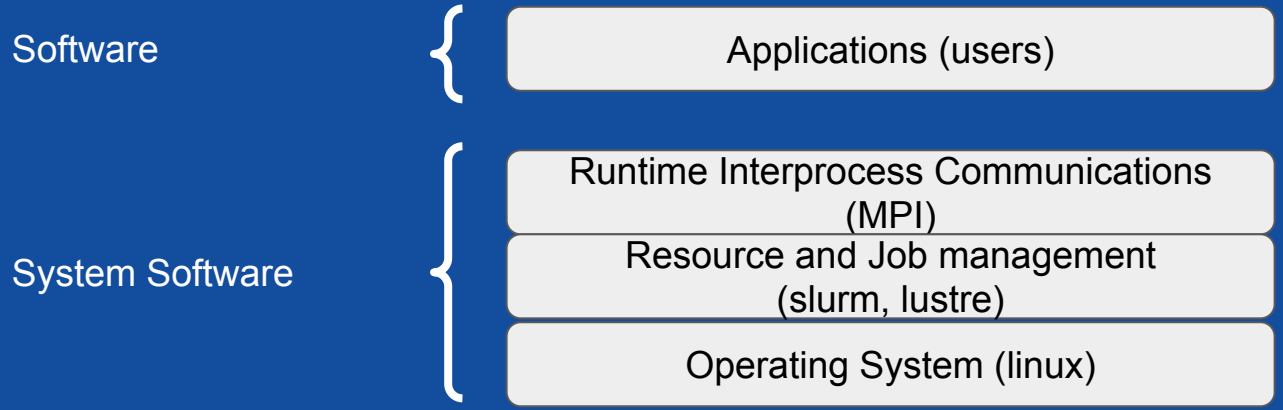
INCD-C @ UC in Coimbra
(BEING RENEWED)
Tape storage expansion
20 Petabytes
10 Gbps

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

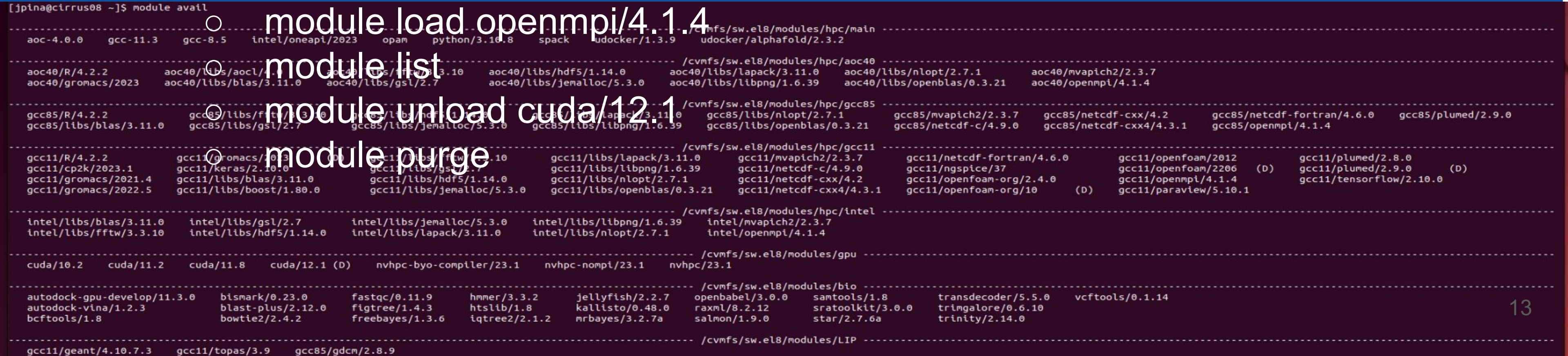
- CernVM File System (CernVM-FS) is a read-only file system on which files and file metadata are downloaded on demand and trough standard HTTP.
  - Cache quota management;
  - Possibility to split a directory hierarchy into sub catalogs at user-defined levels
  - Capability to work in offline mode provided 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

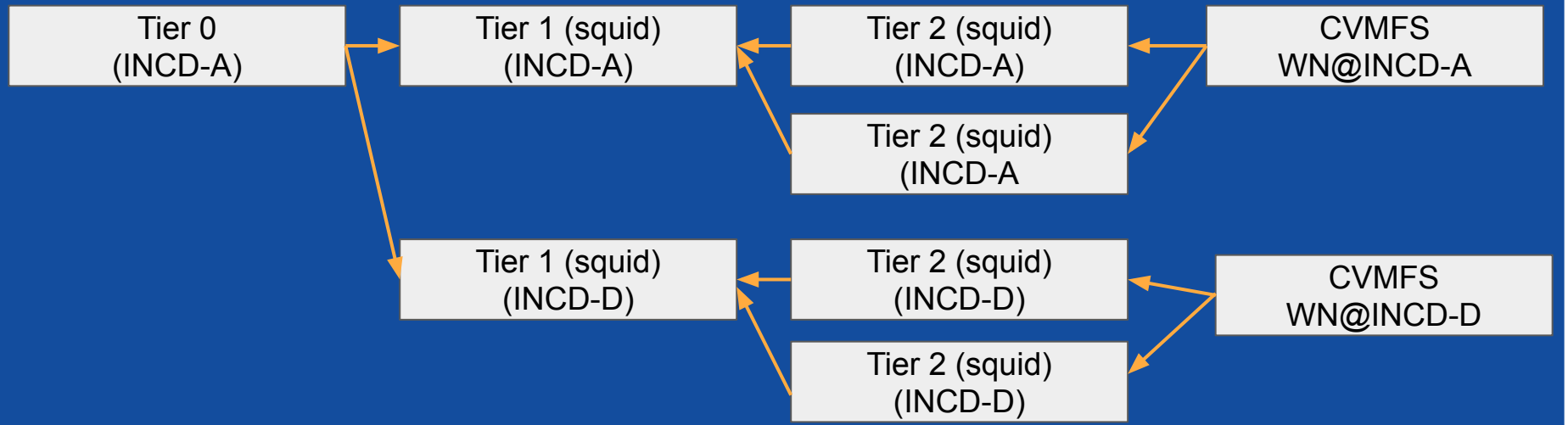
- Customized per:
  - Operating System: CentOS 7, AlmaLinux 8
  - Community
  - Compiler: gcc, intel, aoc, cuda
  - Hardware architecture
- Over 300 different Software/Compilers builds
- Huge complexity and hard to maintain
- module examples:
  - module avail
  - module load openmpi/4.1.4
  - module list
  - module unload cuda/12.1
  - module purge



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 stractum 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?