CNAF-T1 computing

Andrea Chierici

On behalf of CNAF-T1

computing group

Computing staff

- Giusy → EPIC and cloud in general
- Diego → Cloud and batch
- Alessandro 50%, Daniele 50% → batch, HPC and general farm
- Andrea → head of the group

The computing farm

- Originated as LHC and (later) WLCG
- Managed as a single batch system
- No direct access by users
- All nodes share the same configuration, even if hardware is different

- We currently run on free software only

- We manage several different systems
  - HTC, HPC, Cloud

The hardware

- Hardware procured with public tenders
  - Different vendors
  - Our standard rack node is the so called twin2

- BMC configured generally via shared access
  - Static IP (to evolve soon)
- Nodes have public IP (filtered via firewall on local nodes)
- Different hardware complicates management
  - In the future, with the possible adoption of liquid cooling, things may change

OS and middleware

- All nodes share same software configuration
  - Selinux disabled
- Agreement with LHC experiments and WLCG in general
- Cvmfs is the major driver for software distribution
  - Container images are available too
- We are facing an important update due to CentOS/RedHat policy change

Users' facilities

- Computing nodes share a single user domain
  - Fundamental for accessing files on shared storage
- User Interfaces provided to test software and as general gateway to the DC
  - Mostly used by non-WLCG experiments
  - Home dirs on shared FS

Batch system

- During life of our data center we changed several times
- PBS + moab → initial setup
- LSF → first production setup
- HTCondor → current solution for HTC
  - Most of the WLCG data centers chose this solution
- SLURM → for HPC cluster

What we provide to users (batch)

- Most of the computing power is served via batch system
  - 933 computing nodes
  - 47.900 cores
  - 662k HS06
- Small HPC farm to deal with specific use cases
  - Users requiring GPUs
- Cloud (IAAS)
  - Provided via OpenStack
    - Both ISO27001 certified and “standard” one
  - Significant increase in resources provided expected within the end of 2024, due to post-covid19 funding (both CPU and GPU)
- Heterogenous computing
  - Some Aarch64 nodes used by LHC collaborations
  - RISC-V systems to be tested soon

What we provide to users (Cloud Computing)

- Cloud@CNAF local IaaS infrastructure
  - 82 HV, 6000 core, 38TB RAM, 3.8PB storage,
  - 34 GPU (Nvidia A100, V100, T4, RTX5000, AMD MI210), 8 FPGA (AMD Xilinx U250, U50)
  - 700 VM managed per day
- EPIC Cloud ISO 27001/17/18 certified infrastructure for biomedical disciplines
  - 22 HV, ~1400 core, ~10TB RAM, 2.4PB storage
  - 6 GPU (Nvidia A100)
  - 140 VM managed per day
- One of the two INFN Cloud geographically distributed IaaS infrastructure regions
  - Federation point of all INFN cloud infrastructures
  - Providing PaaS and SaaS layer (many software developed by INFN)
  - 1400 VM, 4000 CPU, 16TB RAM, 380TB disc of allocate resources between all federated sites
- Based on
  - Openstack open-source IaaS software
  - Ceph open-source software defined storage (protocols used: RBD, S3, CephFS)

Provisioning and Configuration

- Evolved during the time, like batch system

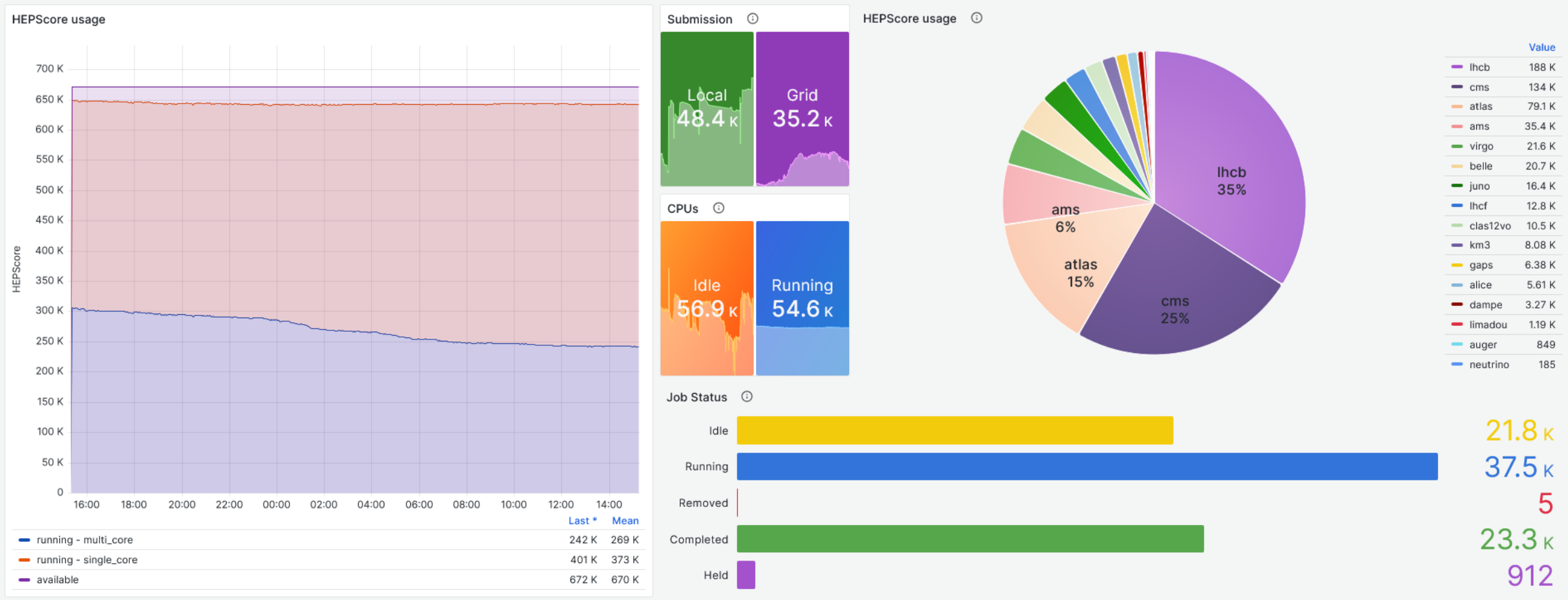
- Provisioning based on ”The Foreman”:
  - Centrally managed infrastructure
  - Host inventory and classification
  - Automatic configuration of DHCP, PXE and TFTP for unattended network installation

- Configuration based on Puppet
  - One puppet environment for each CNAF group (storage, farming, network, …)
    - Environment: collection of puppet modules coming from upstream or self-developed
    - Share of knowledge among CNAF groups

Monitoring and accounting

- Evolved during the time, like batch system, started with Nagios and Graphite
- Current solution based on SensuGo
  - Check health status of machines and their services
  - Collect monitoring metrics in time series databases InflixDB
  - Multiple notifications devices
    - Web page
    - Slack
    - MS teams
    - e-Mail
- Accounting built on local solutions
- Monitoring and accounting data displayed with Grafana

Grafana HTCondor job overview



Farm extensions

- We tested several different solutions, and all proved to be viable to solve specific requirements
  - Commercial cloud → main issue the expense prediction
  - External data centers

- Leonardo
  - Will be the main provider of computing resources for the near future
  - Sets a completely new challenge due to the way Leonardo is managed

How we provide services

- Two virtualization infrastructures used to run “background” services
  - Vmware and ovirt (to be replaced)
  - Roughly 100 VMs
  - High availability both at hardware and at VM level
    - Automatic restart in case of failure

- To be fully redundant we run some services also on physical nodes

Future collaborations

- Submitted project proposal to Call 2024-26 of MoU R&I IT-SRB
  - «Low Power Platforms for Scientific Computing»
  - Participants:
    - INFN, Italy
    - Vinca Institute, University of Belgrade, Serbia
  - Awaiting approval