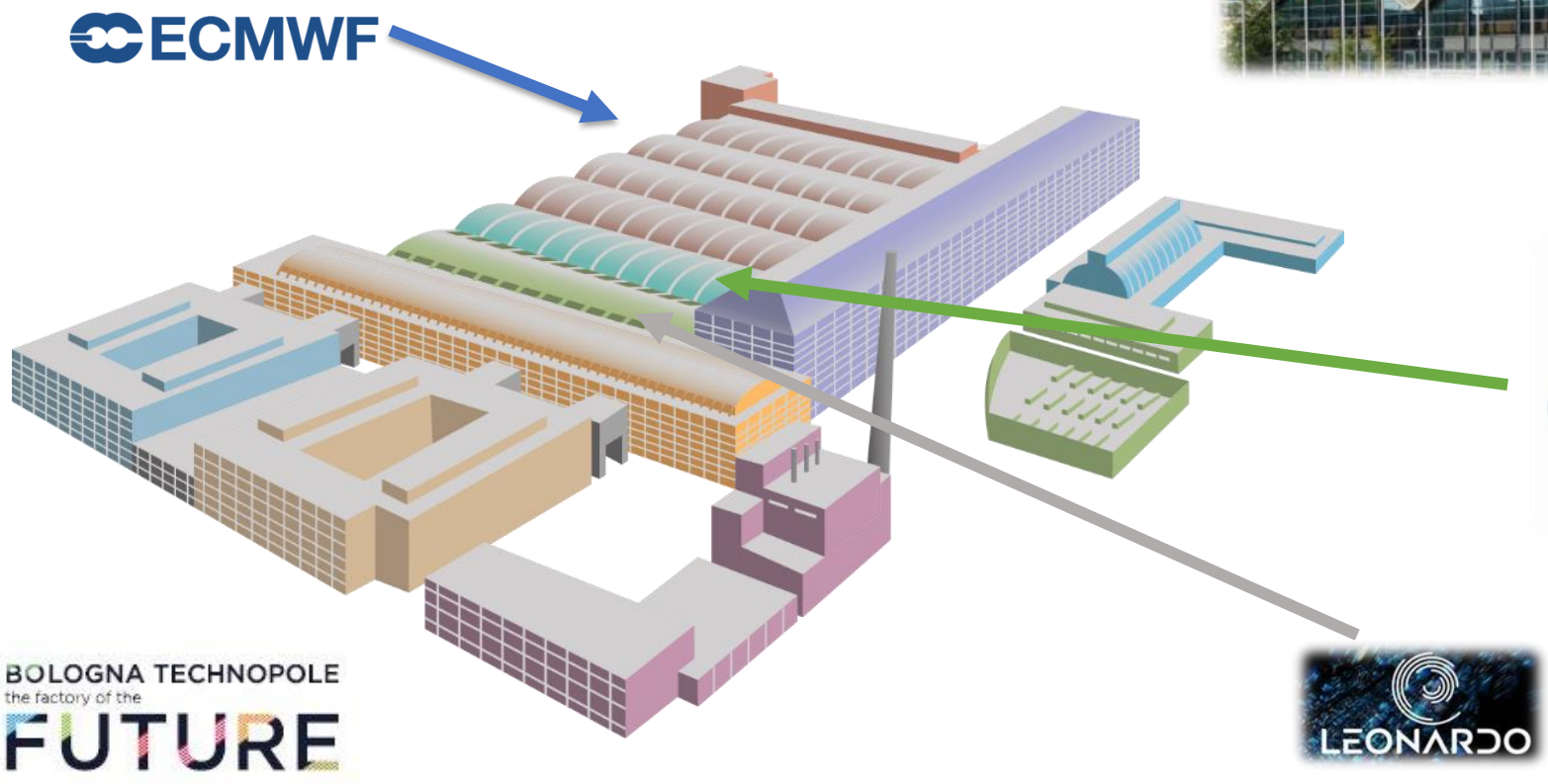
CNAF - report from integration efforts with the Leonardo supercomputer

HEP/HPC Strategy Meeting- 30-31 Jan 2025

D.Cesini – INFN-CNAF

INFN-T1 new Data Center

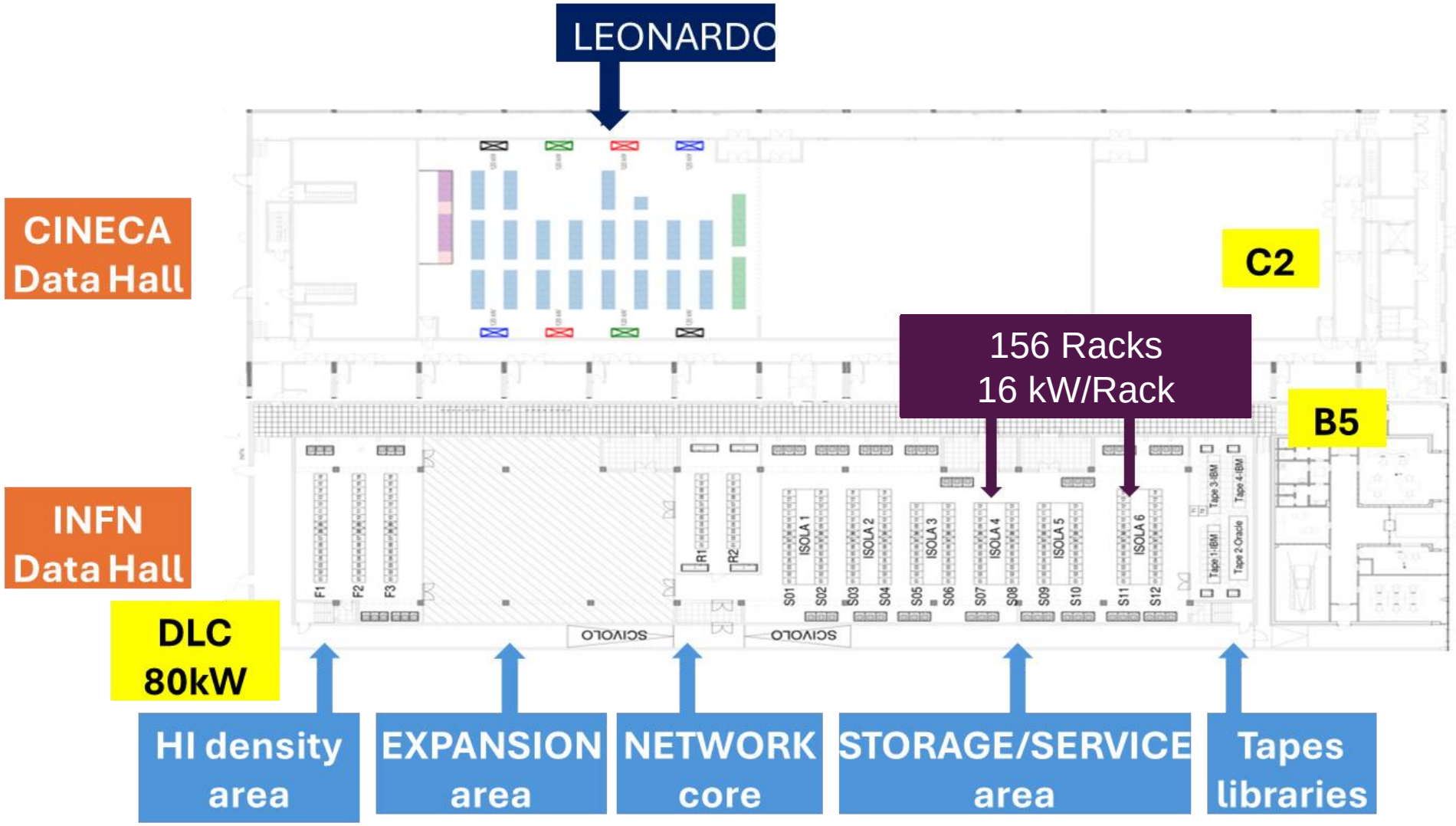


Our new DC building

Co-funded by
- EuroHPC Joint Undertaking
- Italian Ministry of University and Research
- INFN
- SISSA

Under the patronage of the Emilia Romagna Region in Italy

Layout of the new Data Center



CNAF 2025 Pledged Resources:
825k HS
101 PB_net Disk
233 PB Tape

20% year-over-year growth rate

- Air cooled Cold Corridor aisles
- DLC in High Density area
- 3+1 redundancy in all infrastructure facilities

Leonardo is hosted and managed by CINECA

- Booster Module
  - Features a custom BullSequana X2135 “Da Vinci” blade, composed of:
    - 1 x CPU Intel Xeon 8358 32 cores, 2,6 Ghz booster
    - 512 (8 x 64) GB RAM DDR4 3200 MHz
    - 4 X Nvidia custom Ampere GPU 64GB HBM2
    - 2 x NVidia HDR 2×100 Gb/s cards
  - Performance per node: 89,4 TFLOPs peak

Used by INFN theoretical physicists (mainly QCD) as any other CINECA user via SSH login

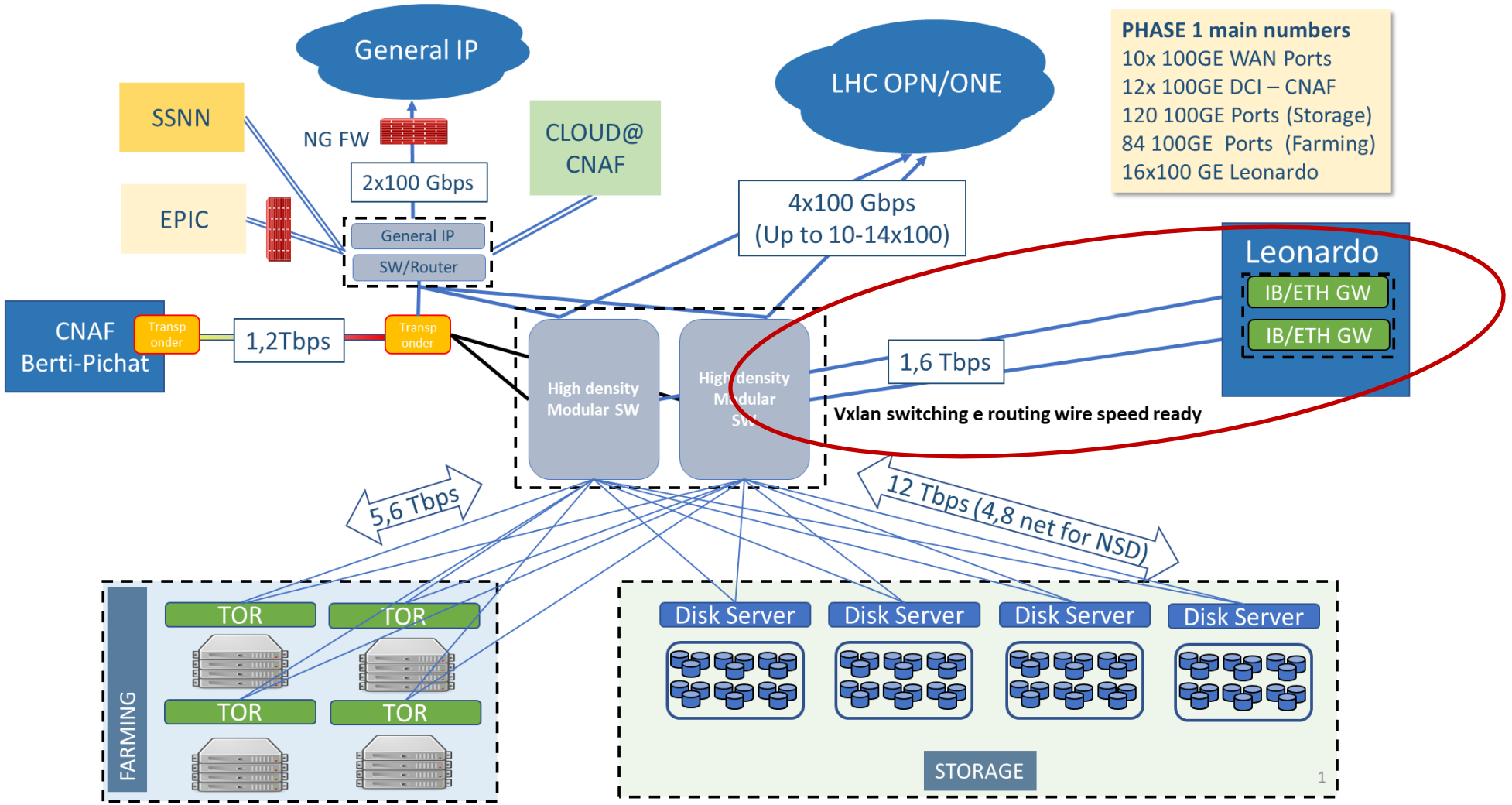
© https://leonardo-supercomputer.cineca.eu/hpc-system/

- Data Centric Module (aka General Purpose Partition)
  - Features BullSequana X2140 three-node CPU Blades
  - Each computing node is composed of:
    - 2x Intel Sapphire Rapids, 56 cores, 4.8 GHz
    - 512 (16 x 32) GB RAM DDR5 4800 MHz
    - 3xNvidia HDR cards 1x100Gb/s cards
    - 8 TB NVM

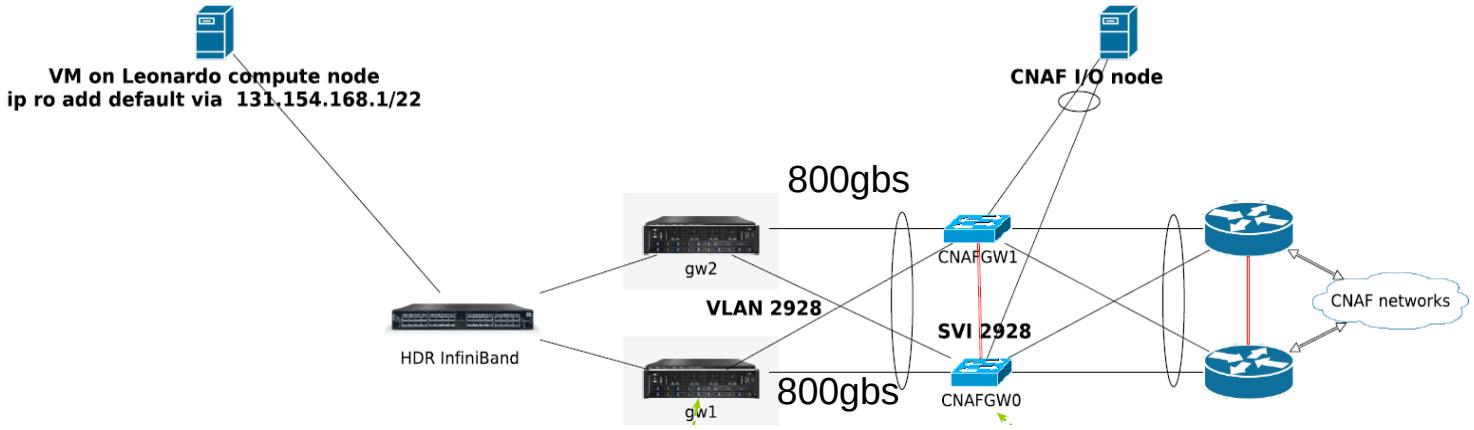
No Ethernet!!

Up to 200 nodes provisioned via SLURM, integrated as a transparent extension in the CNAF HTC farm to be accessed via Grid services.

CNAF Network Architecture



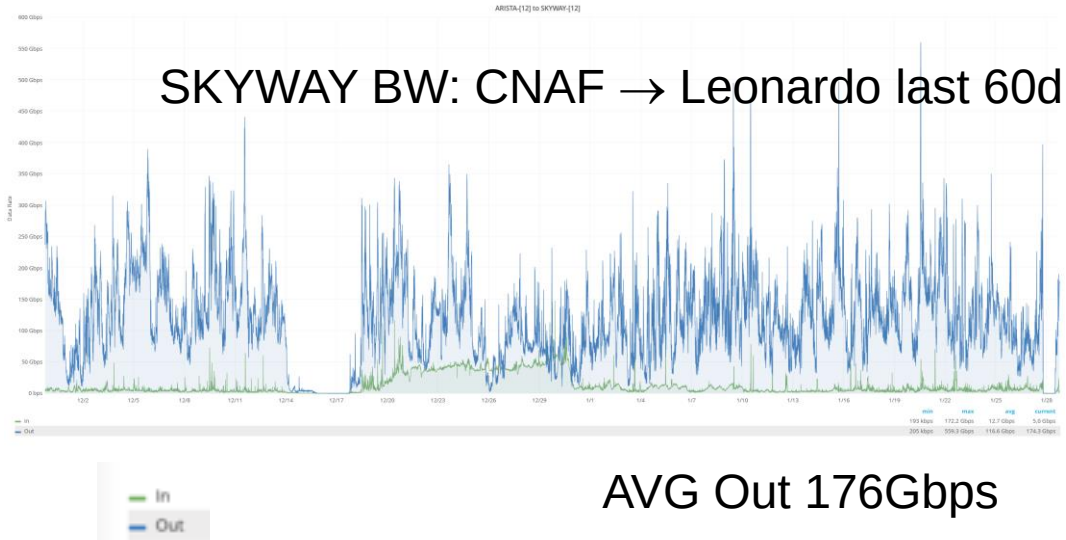
CNAF ⇔Leonardo set-up



- HTCondor WN created via CINECA SLURM whole-node jobs on the Leonardo General Partition (CPU-only)
- Public CNAF IP on IPoIB
- Inbound/Outbound connectivity via NVIDIA Skyway directly attached to our core switches → 1.6Tb/s

600Gbps

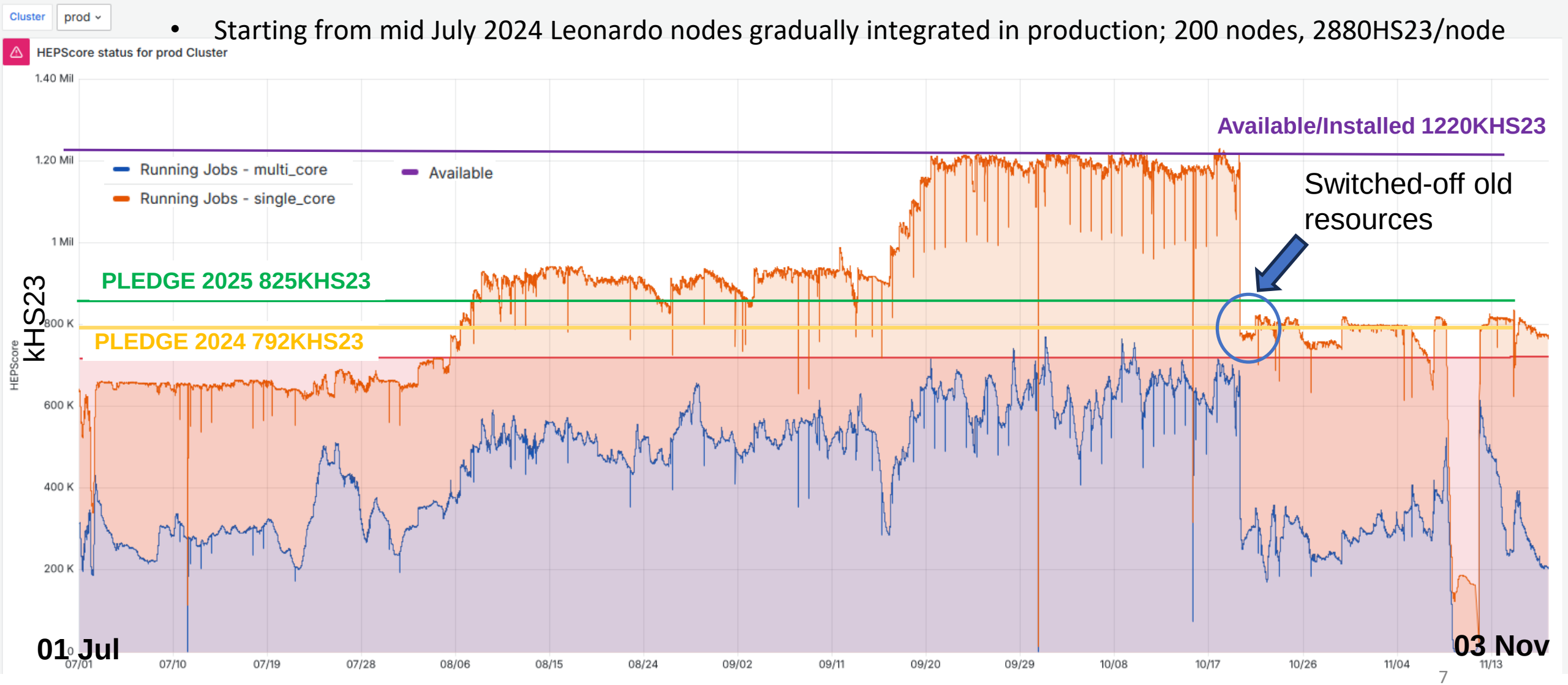
300Gbps



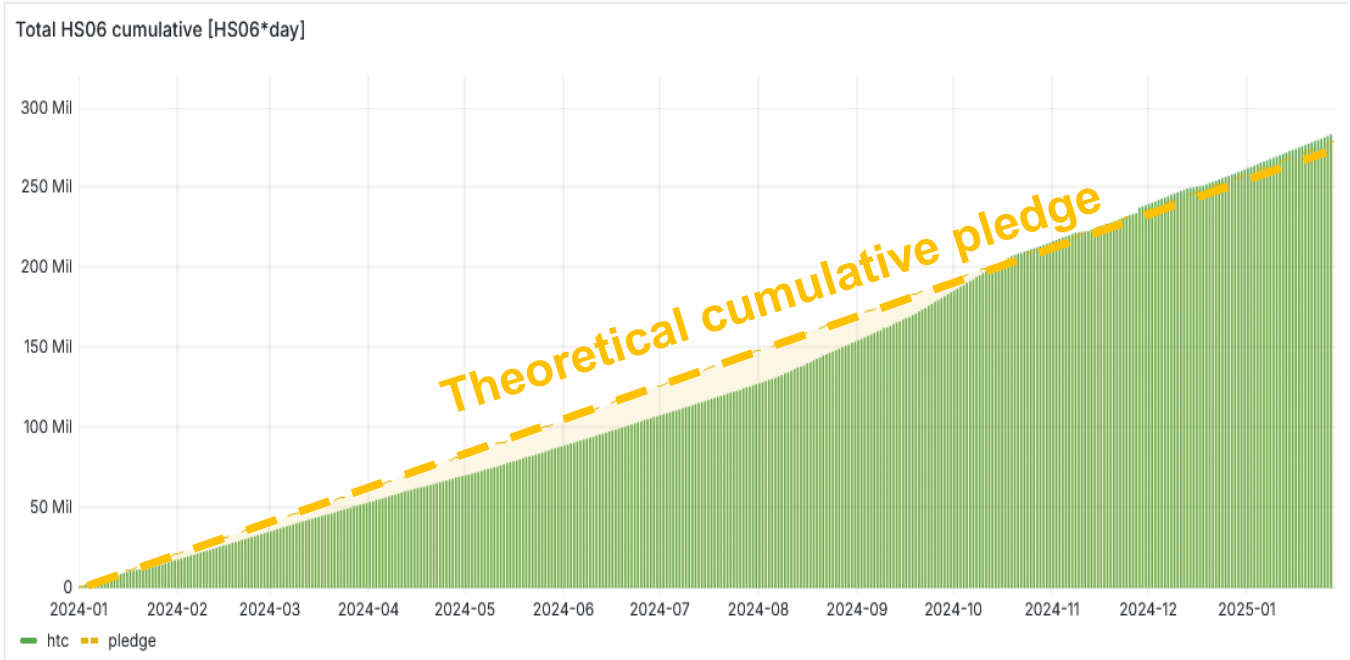
AVG Out 176Gbps

CPU - Farm

- **Pledge 2024: 792kHS23 – Pledge 2025: 825kHS23 → Total Installed: 1220KHS23**
- Starting from mid July 2024 Leonardo nodes gradually integrated in production; 200 nodes, 2880HS23/node



Advantages



- So far, so good…
  - Resources granted by ad-hoc agreements on the national quota
  - HTCondor/SLURM Integration worked smoothly
    - Transparent extension of the CNAF farm to Leonardo
    - We managed to execute “our WNs on un-modified Leonardo nodes using virtualization, including:
      - WLCG tools
      - CVMFS
      - direct mount of CNAF filesystems
  - Allowed us to recover from historical 2024 CPU under-pledge
  - Allowed us to grant over-pledge to several VOs that requested it

Drawbacks

- 576.000HS of **pledged** resources are now provided by Leonardo
  - 70% of the total CNAF HTC computing power
- ETH/IB gateways are working better than expected but…
  - …No IPv6 support
    - No SKYWAY fw updates can fix this
  - Having a reasonably-high bandwidth Ethernet NICs on Leonardo nodes would have saved a lot of pain
- HyperThread not enabled
  - Unbalance in the HS/core numbers
- We do not manage the HW
  - Very little control on resources availability
- Different downtimes schedule
  - Few possibilities to affect or agree on it
- Different sw/kernel updates policies
- Different infrastructure redundancy (electrical or cooling)
- In general, different requirements on A/R

Summary and Open Points

- «friendly» HPC center resources can be effectively and transparently integrated into HTC sites
- Being involved in the design and technical decisions of the HPC systems would be highly beneficial to our community
  - To provide handles for a smoother integration
    - i.e. ETH NICs, IPv6 support, networking ACLs, GPUs models, CPU/GPU ratio, processor architecture, etc…
- Different use cases ➔ different A/R requirements
  - Could be risky to pledge a too large fraction of the total resources
- HPC system in general has a different lifecycle schedule ➔ shorter lifetime compared to the HTC resources
  - Need to be prepared for the post-Leonardo