CERN-CNAF Collaboration

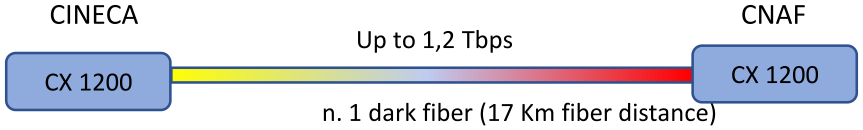
Topic-1.4: Network

08-10-2020

CNAF extension to CINECA HPC cluster

INFN TER1 Extension to HPC Resources @ CINECA

CINECA – CNAF DCI



CINECA

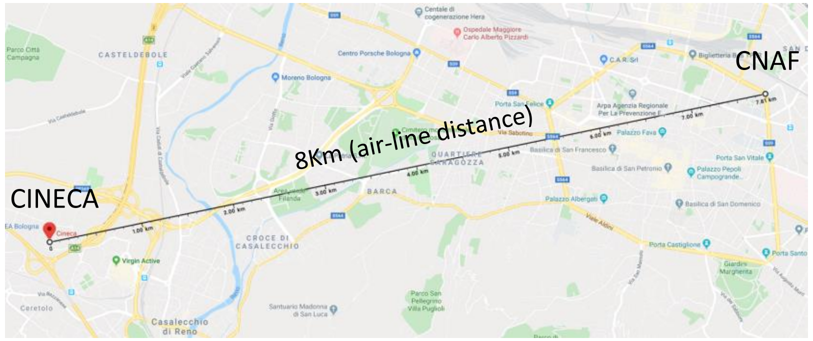
CX 1200

Up to 1,2 Tbps

CNAF

CX 1200

n. 1 dark fiber (17 Km fiber distance)



CINECA

Marconi A2 Partition

3600 nodes with 1 Xeon Phi 2750 (KNL) at 1.4 GHz and 96 GB of RAM

68 cores/node, 244800 cores

Peak Performance: ~11 Pflop/s

“THE GRANT”

The LHC Italy community successfully applied for a “PRACE Project Access” on the CINECA KNL partition

30McoreH allocated for running LHC Jobs

CNAF

Standard “GRID-like” HTC farm (30k cores, 400 kHS06)

•41 PB of disk

•90 PB of tapes on 2 libraries

HPC-HTC Integration (Challenges)

Matching LHC workloads with HPCs is not that easy because they derive from different ”User requirements”.

In the HPC centers there are usually **strict site policies not matching the LHC user requirements**

- Ad-hoc operating system, limited/absent external connectivity, user policies only for individuals, node hardware setup, ...

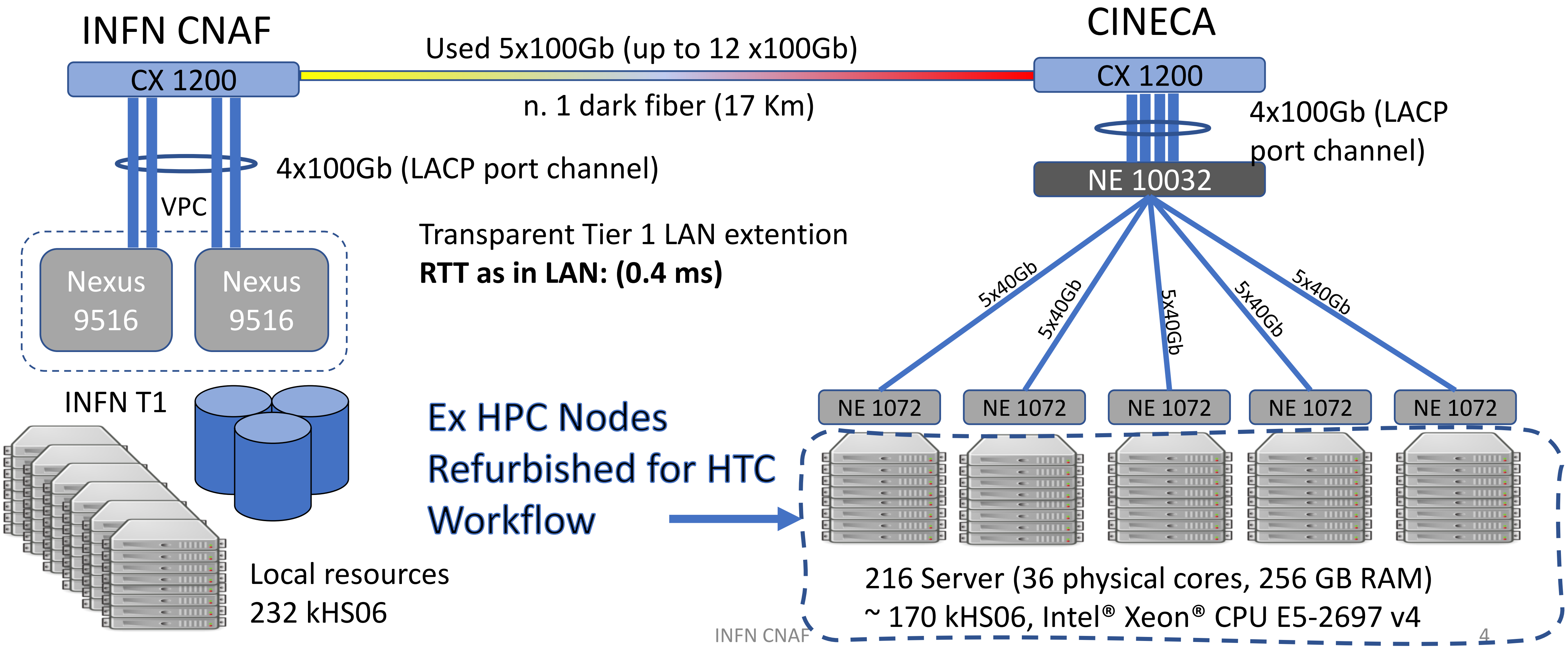
A lot of work needed to establish a “trust model” With CINECA and to understand the peculiarity of the HPC platforms running at CINECA and to adapt each workflows to run together.

Many people involved but very enriching experience for everyone.

A standard CINECA Marconi A2 is node configured with	A typical WLCG node has
An Intel(R) Xeon Phi(TM) CPU 7250 @ 1.40GHz: 68 or 272(HT4x) cores, x86_64, rated at ~¼ the HS06 of a typical Xeon per core	1-2 Xeon-level x86_64 CPUs: typically 32-128 cores, O(10 HS06/thread) with HT on
96 GB RAM, with ~10 to be reserved for the OS: 1.3-0.3 GB/thread	2GB/thread, even if setups with 3 or 4 are more and more typical (so a total 64-256 GB)
No outgoing connectivity from the node (only a very limited access to the external networks via NAT on bastion hosts) – CINECA overall WAN connectivity limited to 10Gbps (more than enough for its core users workflows)	Full outgoing external connectivity, with sw accessed via CVMFS mounts; additional experiment specific access needed (condition DBs, input files via remote Xrootd, ...)
No local disk (large scratch areas via GPFS/Omnipath)	O(20 GB/thread) local scratch space
Access to batch nodes via SLURM; Only Whole Nodes can be provisioned, with 24 h lease time	Access via a CE. Single thread and 8 thread slots are the most typical; 48+ hours lease time
Access granted to individuals (via passport / fiscal code identification)	Access via pilots and late binding; VOMS AAI for end-user access

CNAF – CINECA DCI role on HPC-HTC Integration

In order to address the problem on High throughput data rate access fron the HPC nodes at CINECA we decided to use the existing production DCI extention between CNAF and CINECA.



Transparent Tier 1 LAN extention

RTT as in LAN: (0.4 ms)

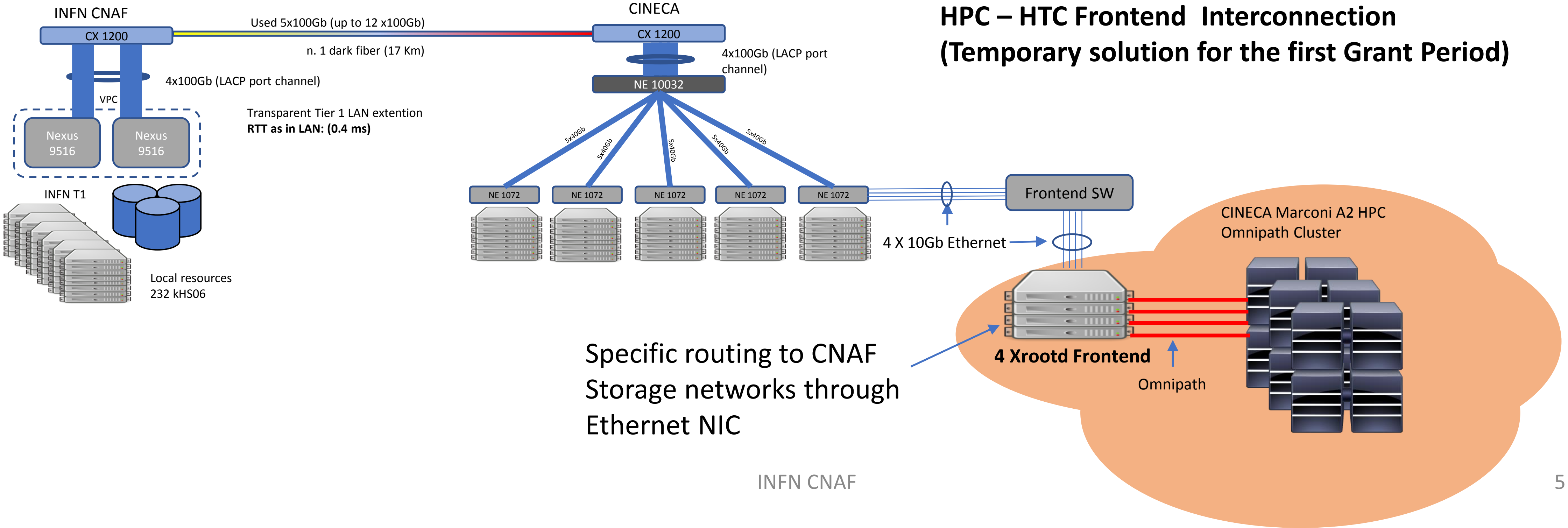
Ex HPC Nodes Refurbished for HTC Workflow

Local resources
232 kHS06

216 Server (36 physical cores, 256 GB RAM)
~ 170 kHS06, Intel® Xeon® CPU E5-2697 v4

Interconnecting the HPC cluster network

- *Not possible to directly intrconnect the Omnipath HPC Cluster internal network (due to security reasons)*
- *Use of dual homed (Omnipath and Ethernet) nodes as frontends in order to reach the Storage @CNAF without impacting on CINECA Ceneral IP link.*



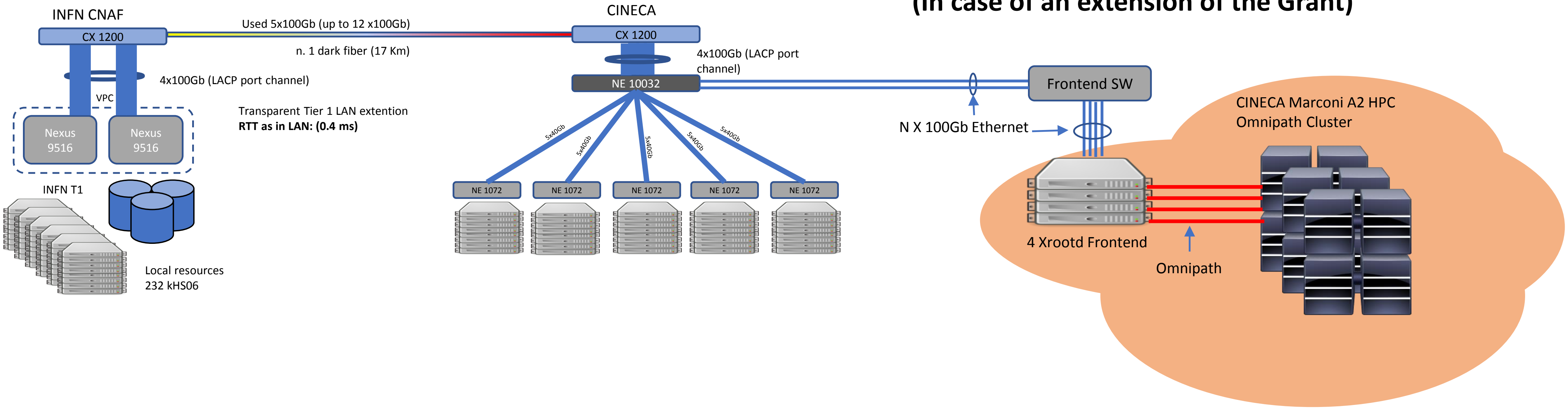
HPC – HTC Frontend Interconnection
(Temporary solution for the first Grant Period)

Specific routing to CNAF Storage networks through Ethernet NIC

Interconnecting the HPC cluster network

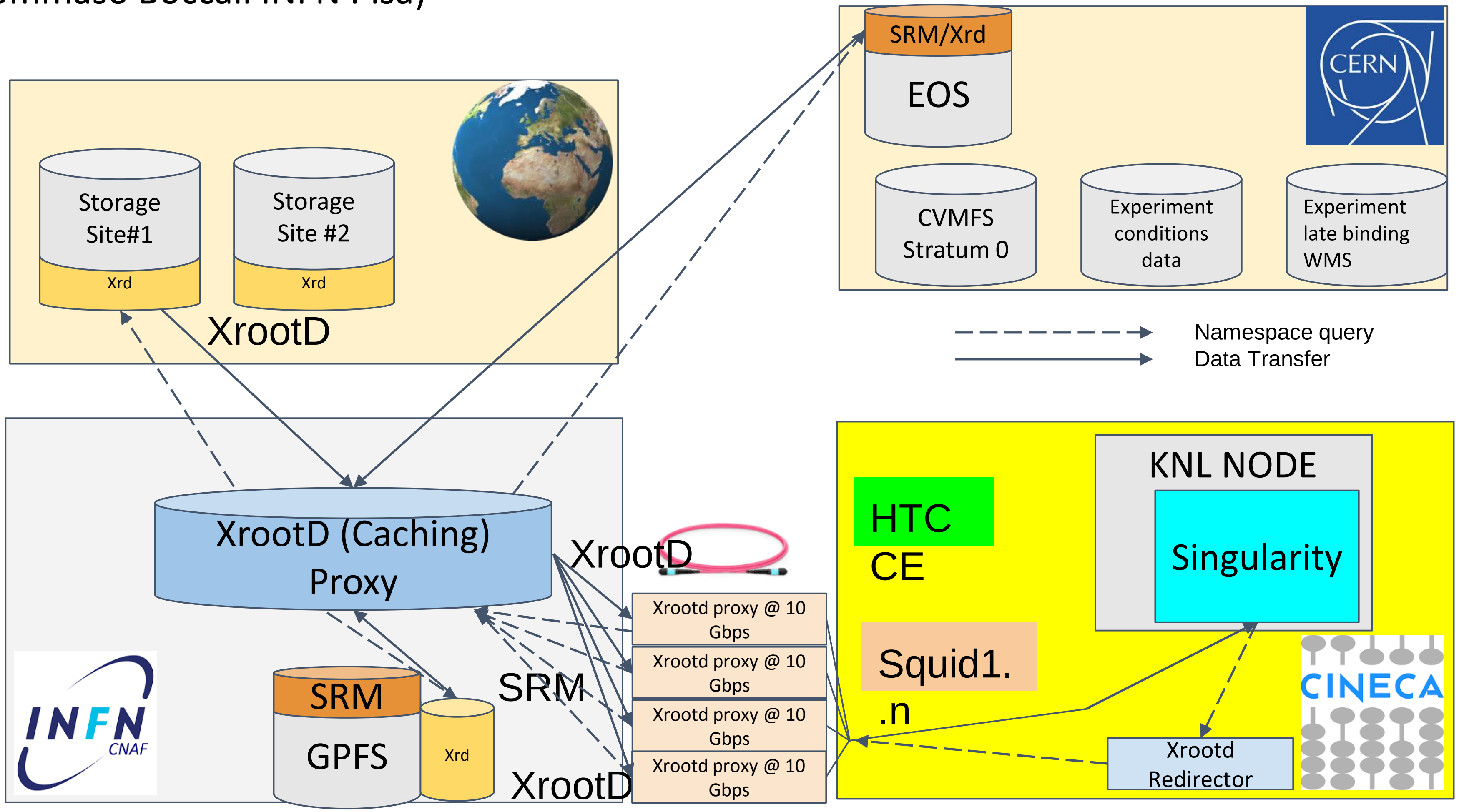
- In case of an extension of the “Grant” the interconnection to HPC cluster will be moved to CNAF aggregation switch in CINECA using 100G Ethernet connections, sizing properly the number of frontend.

HPC – HTC Frontend Interconnection
(In case of an extension of the Grant)



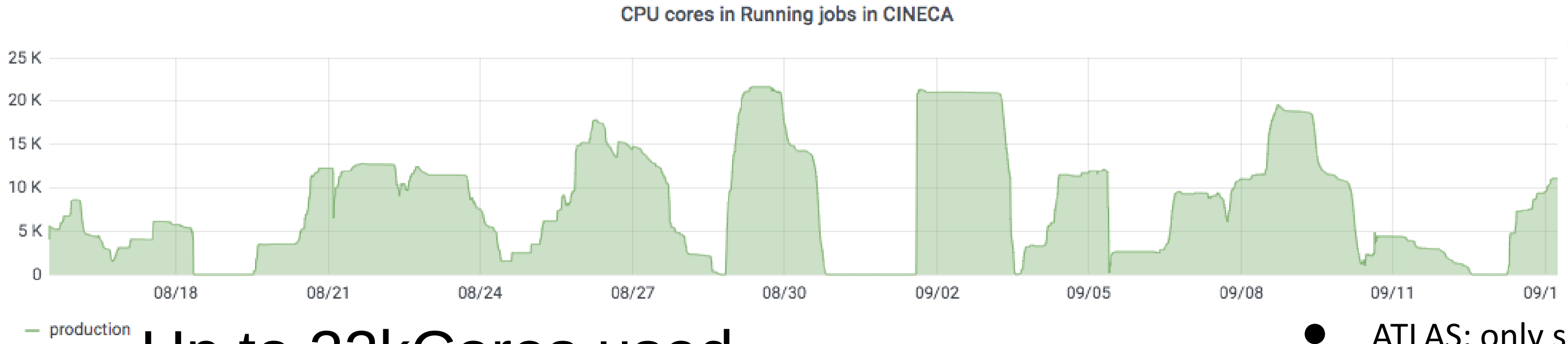
Data access architecture

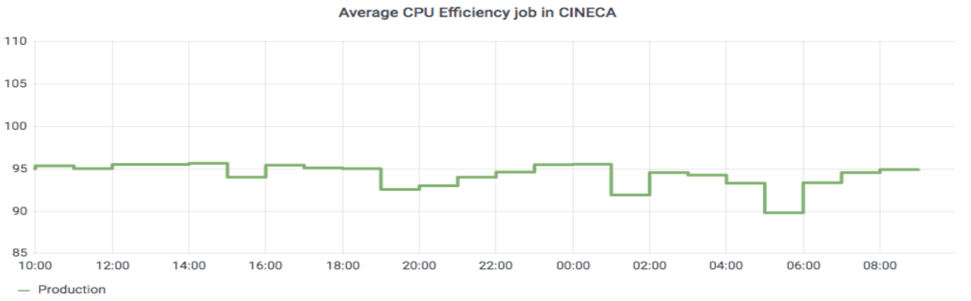
This is the chematic view of the Xrootd configuration implemented to use conveniently the given infrastructure (Courtesy Tommaso Boccali INFN Pisa)



Slide: Courtesy Tommaso Boccali (INFN Pisa)

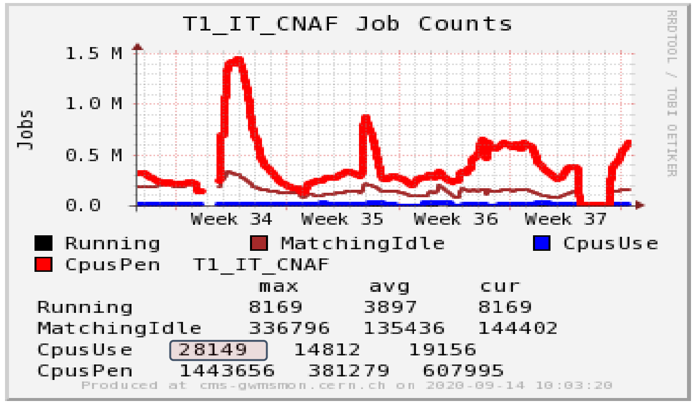
Encouraging results running WLCG workflows





- Up to 22kCores used
- Success rates 85-90% (mostly timeouts and application errors)
- CPU efficiency ~95% (with data intensive jobs!)
- Utilization of the link very low (only a small fraction of the CMS WFs require sizeable I/O in this period)
- CMS sees CNAF as able to provide 28kCores

- ATLAS: only simulation jobs, using 48 threads per node
  - Reached production level in May
- LHCb: simulation jobs, up to 136 threads per node
  - Reached production level in July (but still some software issues affecting physics)
- ALICE: do not use WMS, run local Run-3 O2 tests
  - Reached production level on these specific jobs (not the typical WLCG ones)
- CMS: all production jobs, up to 128 threads per node (we could go higher..)
  - Reached production level in Dec



	max	avg	cur
Running	8169	3897	8169
MatchingIdle	336796	135436	144402
CpusUse	28149	14812	19156
CpusPen	1443656	381279	607995

Produced at cms-gwmsmon.cern.ch on 2020-09-14 10:03:20

Slide: Courtesy Tommaso Boccali (INFN Pisa)

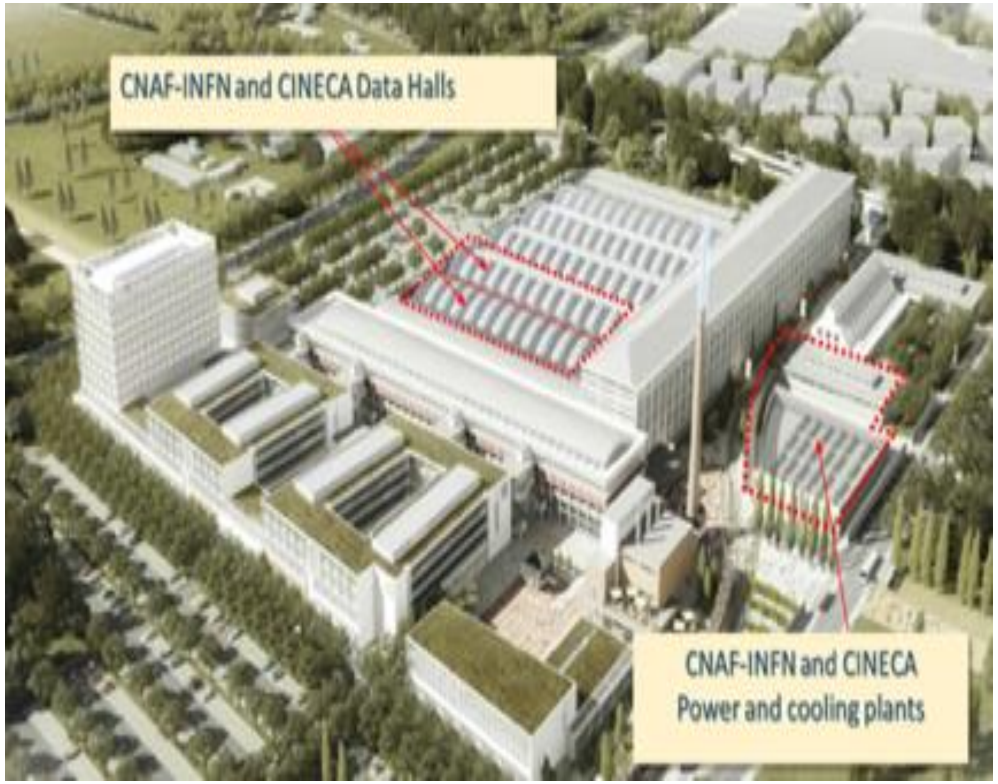
Extension to HPC clusters is challenging but very promising

CNAF and CINECA will have adjacent data halls (in 2021 - 2022) and "Leonardo" (CINECA-INFN-SISSA) PreExascale Supercomputer will be in place.

Leonardo (using Infiniband internally) will be interconnected to IP network at 2.5Tb/s using Mellanox SKYWAY gateways.

The interconnection between the two datacenters will be in Local Area Network with the minimum possible latency.

So theoretically the I/O rate from HPC nodes to WLCG data stored @CNAF TIER1 should be potentially extremely High.


CNAF-INFN and CINECA Data Halls
CNAF-INFN and CINECA Power and cooling plants