INFN Tier-1: a distributed site

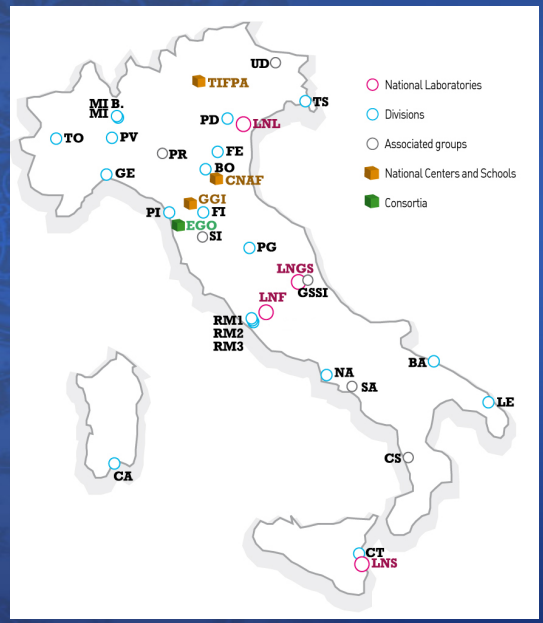
Luca dell'Agnello & Gaetano Maron (INFN-CNAF),
Tommaso Boccali (INFN Pisa)

CHEP 2018 conference

Sofia, July 12 2018

INFN

- **National Institute for Nuclear Physics** (INFN) is funded by Italian government
- Main mission is the research and the study of elementary particles and physics laws of the Universe
- Composed by several units
  - ~ 20 units dislocated at the main Italian University Physics Departments
  - 4 Laboratories
  - 2 National Centers dedicated to specific tasks
- CNAF, located in Bologna, is the National Center dedicated to computing applications



The INFN Tier-1

- First incarnation dates back to 2003 as computing center for BaBar, CDF, Virgo and prototypical for LHC experiments (Alice, ATLAS, CMS, LHCb)
- After a complete infrastructural refurbishing in 2008, it nowadays provides services and resources to more than 30 scientific collaborations
  - Farm power: 360 kHS06
  - ~34 PB of disk, ~50 PB of tape
- Planning now a new data center to cope with the high demanding computing requirements of HL-LHC and newly coming experiments
  - CNAF data center can host resources up to the end of LHC Run 3

INFN Tier-1 beyond CNAF data center

- Since 2015, we have started testing use of remote CPU resources to extend our data center beyond CNAF site
- Several reasons
  - Long-term lease of CPU resources available in other sites
  - Possibly accommodate unexpected CPU requests (cloud bursting)
  - Evaluate new operational models
- Internal requirements
  - Transparent extension for users
    - INFN Tier-1 CEs as unique access points also to these resources
  - Key issue is data access (i.e. remote or cache)

Various types of extensions…

- Functional tests on commercial clouds
  - Aruba (2015-2016)
  - Azure (2017)
- Participation to scalability tests with hybrid cloud in the context of EU project HNSciCloud
- Static allocation of remote resources
  - Bari-RECAS (since 2017)
  - CINECA (since 2018)
- Opportunistic CPU on HPC
  - Scheduled test on CINECA HPC resources

Opportunistic computing on commercial clouds

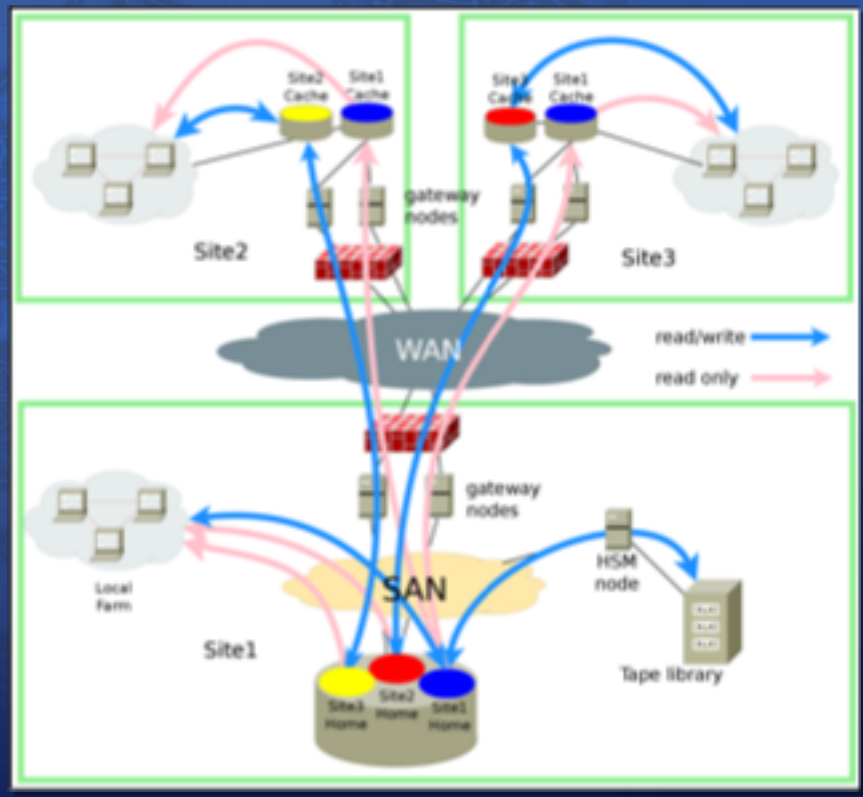
- Tested by CMS on 2 different clouds:
  - Aruba cloud (2015-2016): small scale test (up to 150 cores max) on idle resources
  - Azure cloud (2017): grant to CMS to use with MS Azure Cloud Infrastructure (272 cores)
- Based on an in-house developed application, dynfarm
  - Authenticates connection requests coming from remote hosts and delivers the information needed to create a VPN tunnel
    - Communication enabled only through remote WNs and local CEs, LSF and Argus
    - LSF binaries etc.. accessible via a cache of the shared file system (or synced at boot on VMs)
    - All other traffic goes through its default route
- Remote data access via XRootD on GPN
  - XRootD redirector can be added in the remote cloud
- Job efficiency (CPT/WCT) depends obviously on type of job
  - Very good for MC
  - Low on average (~0.4 vs. 0.80) due to GPN usage

Farm extension on HNSciCloud resources

- Besides testing based on full site instantiation with DODAS (CMS) or VCYCLE (Belle2)
- Production jobs submitted to HNSciCloud resources via INFN Tier-1
  - Dynfarm application
  - Limited storage resources required (only MC jobs)
  - VMs only access cvmfs and frontier-squid
  - Currently supporting CMS and ATLAS multicore jobs, other LHC experiments to come

Farm extension to Bari-ReCaS

- Since 2017 a fraction (~22 kHS06) of pledged CPU for WLCG experiments leased from Bari-RECAS data center (600 km far, RTT ~10 ms)
  - Transparent access for WLCG experiments
    - CNAF CEs and LSF as entry-point
    - Auxiliary services (i.e. squids) replicated in Bari
  - 20 Gbps L3 VPN provided by GARR
    - All traffic with farm in Bari routed via CNAF
  - Direct Posix access to data preserved via local cache (GPFS/AFM)
    - "Transparent" extension of CNAF GPFS
    - XRootD used by Alice (standard) and CMS (fallback)
  - Low efficiency for I/O intensive jobs ☹

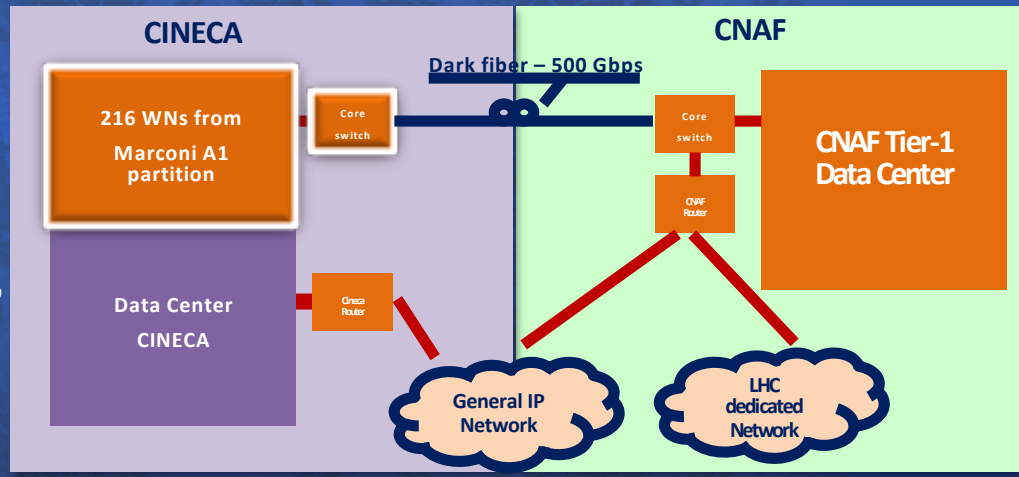


Farm extension to CINECA (1)

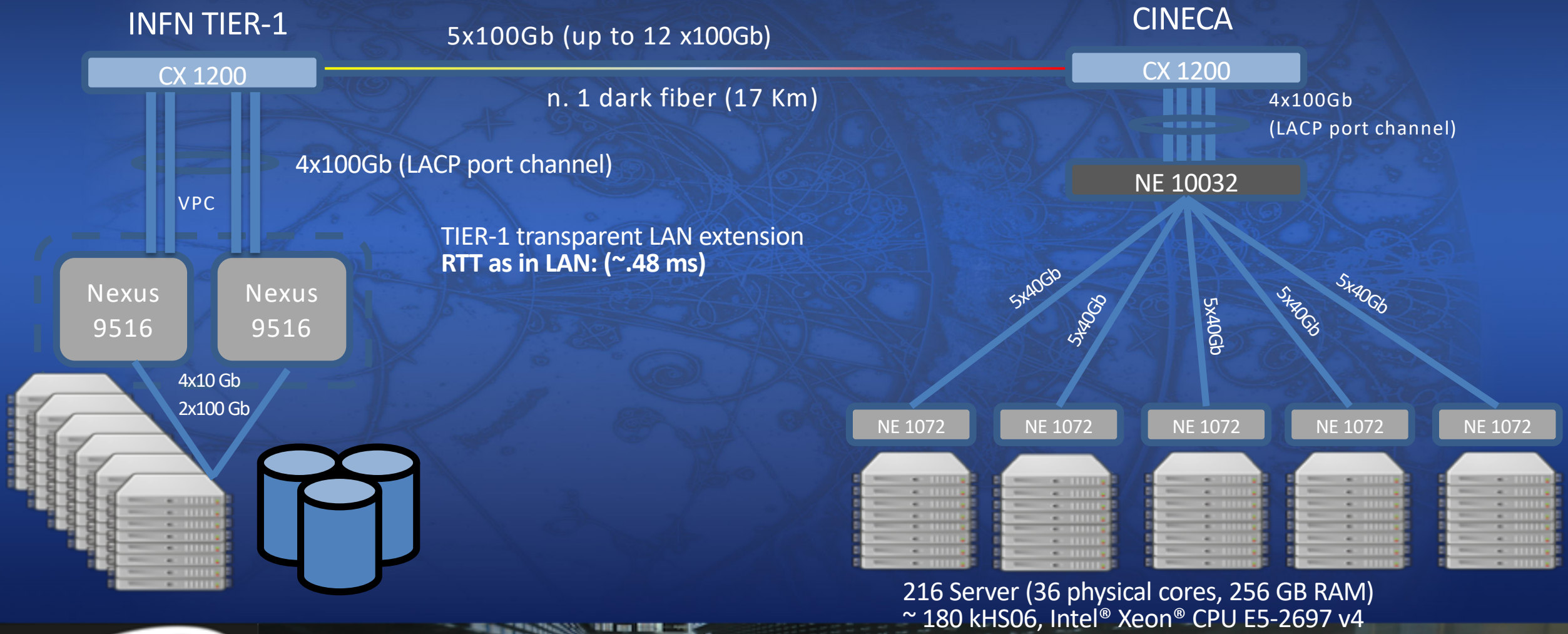
- CINECA, located in Bologna too, is the Italian supercomputing center and Tier-0 for PRACE
  - Close to CNAF (17 Km far away)
- Dedicated fiber directly connecting Tier-1 core switches to our aggregation router at CINECA
  - 500 Gbps (upgradable to 1.2 Tbps) on a single fiber couple via Infinera DCI
  - Quasi-LAN situation (RTT: 0.48 ms vs. 0.28 ms on T1 LAN)

Farm extension to CINECA (2)

- In 2018-2020 ~180 kHS06 leased from CINECA (from dismissed Marconi A1 partition)
  - 216 WNs (72 cores HT each), 3.5 GB/core
  - 10 Gbit connection to rack switch
  - 4x40 from rack switch to router aggregator
- Managed by LSF@Tier-1, access through Tier-1 CEs
- No disk cache, direct access to CNAF storage
- In production since March
  - Slowly opening to non-WLCG experiments (CentOS7)
  - Efficiency comparable to partition @CNAF



CNAF - CINECA Data Center Interconnection



The INFN Tier-1 farm: some figures

- Before the flood (2017 pledges)
  - ~22.800 computing slots, ~236 kHS06
    - Small part of the WNs @Bari-ReCaS
- After the flood (2018 pledges)
  - ~38.300 computing slots, ~340 kHS06
  - More than half of computing power outside CNAF

	WNs	Computing slots	kHS06
CNAF	1000	20288	215
Bari-ReCaS	40	2496	21
TOTAL	1040	22784	241

	WNs	Computing slots	kHS06
CNAF	611	20288	160
Bari-ReCaS	40	2496	21
CINECA	216	15552	179
TOTAL	867	38336	360

(No) Operational issues

- Similarities between the 2 cases (Bari-ReCaS and CINECA)…
  - WNs installation/configuration/operativity controlled by INFN Tier-1 staff
  - Physical layer (i.e. network configuration, hw support) managed by remote staff
- .. but with some formal differences
  - Bari-ReCaS site managed by INFN personnel
    - “Informal” procedure for the support
  - CINECA separate entity from INFN
    - MoU signed
    - Escalation procedure defined (after a few iterations i.e. problems ☺)
- In general very good collaboration with both sites

Plan for tests on CINECA-HPC

- Try and use for LHC the CINECA Marconi A2 partition (KNL based)
  - 64x4 cores per node, 96 GB
  - Currently no virtualization, no external network access
- Plan:
  - Use parasitic + grant based allocation
  - Install CVMFS + Singularity
  - Submit from CNAF via CREAM tunneling to CINECA's Slurm
  - Tunnel all external access via the Infinera DCI link

Marconi- A2

Model: Lenovo Adam Pass
Architecture: Intel OmniPath Cluster

Nodes: 3.600
Processors: 1 x 68-cores Intel Xeon Phi 7250 CPU (Knights Landing), 1.40 GHz
Cores: 68 cores/node (272 with HyperThreading), 244.800 cores in total
RAM: 16 GB/node MCDRAM + 96 GB/node DDR4
Internal Network: Intel OmniPath Architecture 2:1

Peak Performance: 11 PFlop/s

Conclusions

- The increasing demand of computing resources led to the investigation of several techniques to extend the existing farm of INFN Tier1
  - On commercial clouds (with low efficiency for average jobs out of the box: ~0.4) to accommodate temporary peaks
  - With statically allocated resources on remote sites for pledges
- Currently more than 50% of the CPU is hosted outside CNAF
  - Key point is (obviously) data access and cache technology to be carefully evaluated
- First experience with 'quasi local' resources positive
- Probable new leases of CPU for next years
- 2026+ configuration still unclear, but efforts are in the direction to have experience with all the possible scenarios: data lake, commercial clouds, HPC co-utilization