Dalla rete al Tecnopolo: 25 anni di esperienza al CNAF

Luca dell’Agnello INFN-CNAF

May 7, 2021

Abstract

Con il progetto di migrazione del data center al Tecnopolo, siamo alla vigilia di un’importante trasformazione del CNAF. Non è la prima volta nella nostra storia.

In questo seminario ripercorrerò l'evoluzione tecnologica a cui ho assistito durante i miei 25 anni di attività al CNAF, dal ruolo rivestito dal Centro nella nascita del GARR, al passaggio al calcolo distribuito con i primi progetti sulla Grid, alla realizzazione del prototipo del Tier1, fino al completamento e alla gestione di uno dei maggiori data center di WLCG.

Concluderò con una panoramica su una nuova attività di mio interesse relativa all’applicazione delle tecnologie digitali alla conservazione dei Beni Culturali.

Disclaimer

These slides reflect my personal view, based on my own experience: they are not meant as an assessment of the relevance of what has been accomplished at CNAF over the last 25 years, thanks to the great work of many people!

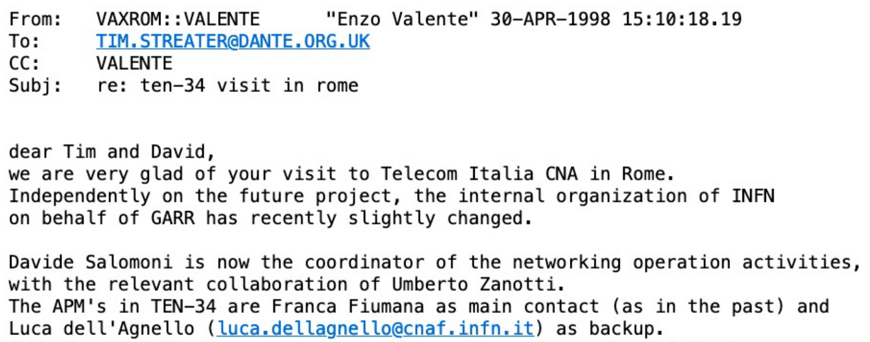
My first years at INFN

- I have devoted almost my entire career to scientific computing
  - Even before joining CNAF in 1996
  - Managing computing resources, doing MC simulations (LVD, Nestor experiments)
- In 1996, CNAF was the center of network development
  - GARR consortium was not yet established
- My activity in the first period at CNAF (1996-2001) was focused on network and related services
  - My first accomplishment was the migration of INFN mailing (from X.400 to RFC822)
  - Study and tests on IPv6
  - Participation to GARR-B project to set-up the new network
- More recently (2009-2015): member of GARR CTS

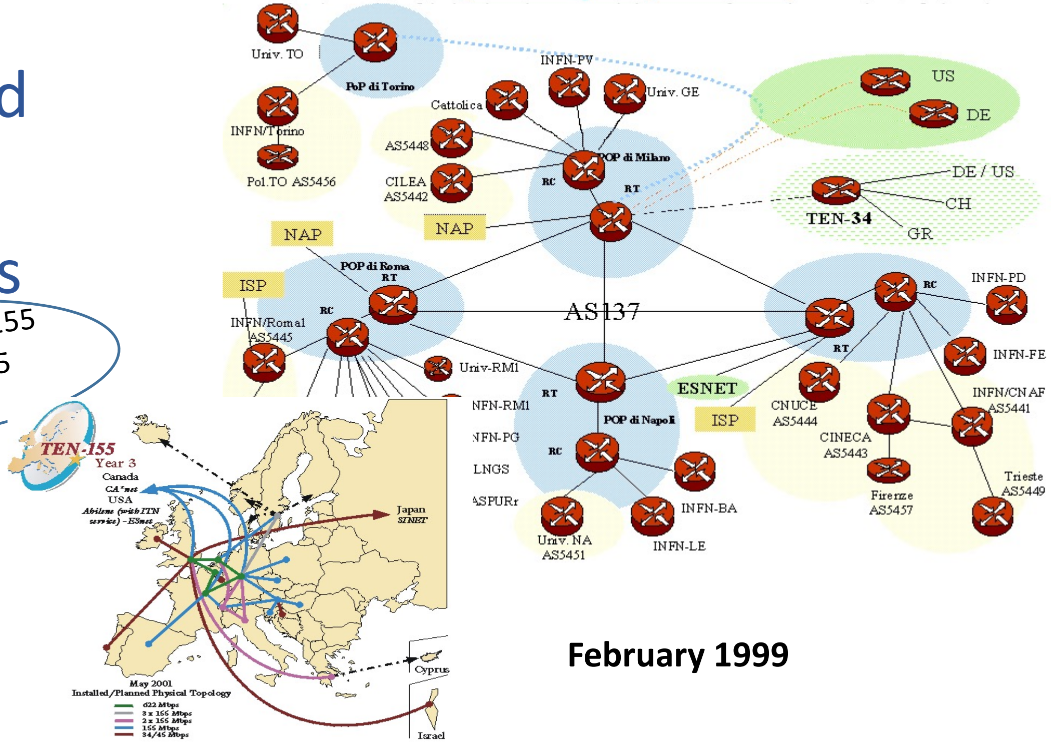
GARR-B

- In 1998 the NOC of GARR-B network was established at CNAF
  - Together with various services (e.g., LIR)
  - Several people hired
- I contributed also to the planning and set-up of GARR-CERT service (1999)
- GARR interconnected to other NRENs via TEN-34
- From 2000 via TEN-155
  - Appointed as GARR representative in TEN-155 (2000-2002)
- Participation to GARR-G pilot (2000)

The suffixes 34,155 are for 34, 155 Mbps!



```
From:   VAXROM::VALENTE      "Enzo Valente" 30-APR-1998 15:10:18.19
To:     TIM.STREATER@DANTE.ORG.UK
CC:     VALENTE
Subj:   re: ten-34 visit in rome


dear Tim and David,
we are very glad of your visit to Telecom Italia CNA in Rome.
Independently on the future project, the internal organization of INFN
on behalf of GARR has recently slightly changed.

Davide Salomoni is now the coordinator of the networking operation activities,
with the relevant collaboration of Umberto Zanotti.
The APM's in TEN-34 are Franca Fiumana as main contact (as in the past) and
Luca dell'Agnello (luca.dellagnello@cnaf.infn.it) as backup.
```



February 1999

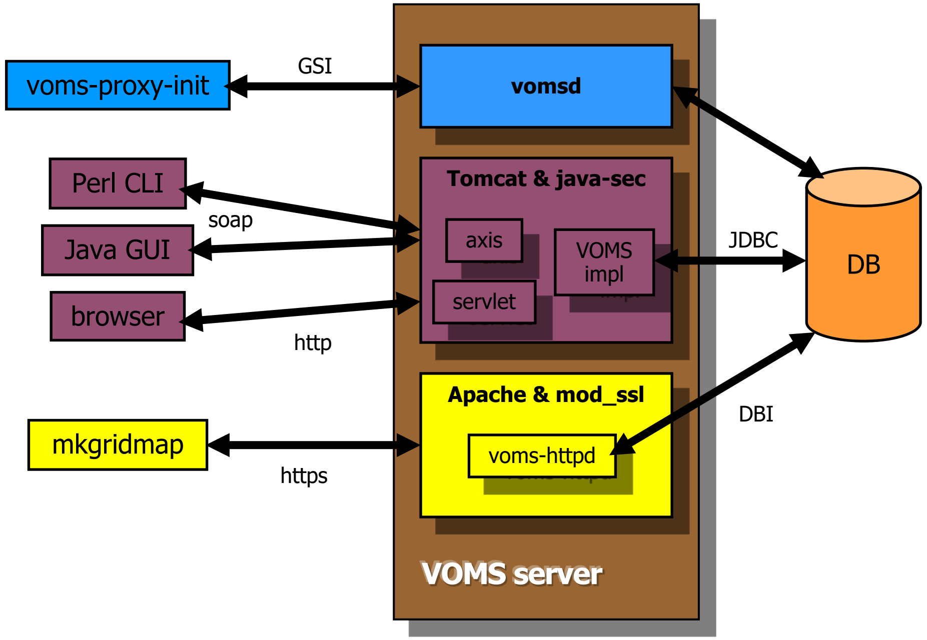
A new beginning: the Grid

- INFN (2000) joins the common effort on distributed computing in view of LHC
  - *CNAF has been one the main players in the building of the Grid with DataGrid and DataTag projects (2000-2003) and in the management of INFNGrid*
  - Decision to realize a Tier1 for experiments at LHC
- Most of effort on WP1 of DataGrid (e.g., WMS, ...)
- Some of us joined DataGrid Fabric WP (WP4) propaedeutic to the activity for the Tier1
  - Gaining experience in management of “large-scale” computing infrastructures
  - Installation, configuration, monitoring , gridification

In the meanwhile (2002):
The GARR consortium is born
GARR NOC and the other services move to Rome

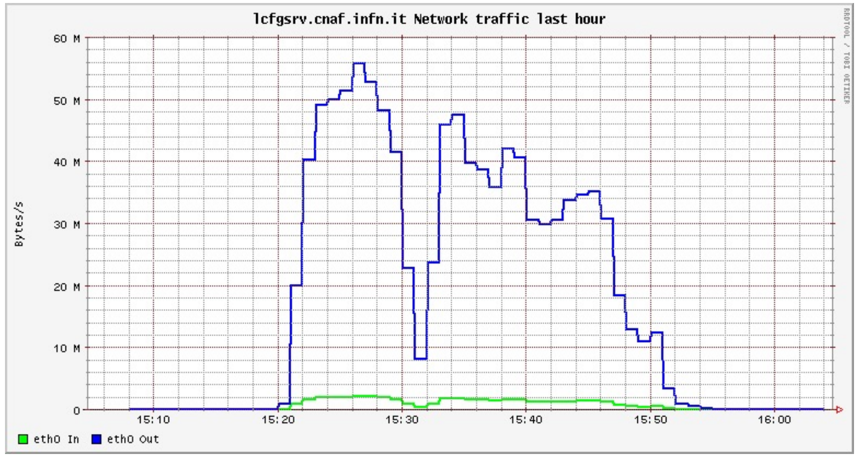
VOMS

- Being interested in Security and AAI issues, with other INFN collogues we formed the Authorization group of DataGrid project
  - AAI initially not really considered an issue ☺
    - Relying on standard Globus mechanism
      - Not suitable, not scalable
- Work on design of VOMS (Authorization system on the grid) and related tools
  - (e.g,. *mkgridmap*)
- Thanks to the effort of CNAF developers, this is one of the most successful products within Grid middleware
  - Still used in WLCG
- *Since then, strong expertise on AAI at CNAF*
  - *Indigo-IAM (it will substitute VOMS in WLCG)*



The Tier1: the prototype

- First prototype (2001-2002) in rooms 44-45
  - ~60 dual processors (800 MHz-1 GHz) with FastEthernet
    - First batch of 1U servers (aka “pizza boxes”)
  - NAS (6 TB raw)
  - (Small) Robot L180 StorageTek (1.8 TB) for backup (no HSM)
- Preparing for the new data center
- Testing management systems developed in the framework of DataGrid project
  - Es. Scalability test of LCFG (first provisioning tool)



Installation of 50 nodes

ISTITUTO NAZIONALE DI FISICA NUCLEARE

Progetto Speciale Centro Regionale di Calcolo Tier1

Piano della Realizzazione delle Infrastrutture di Base

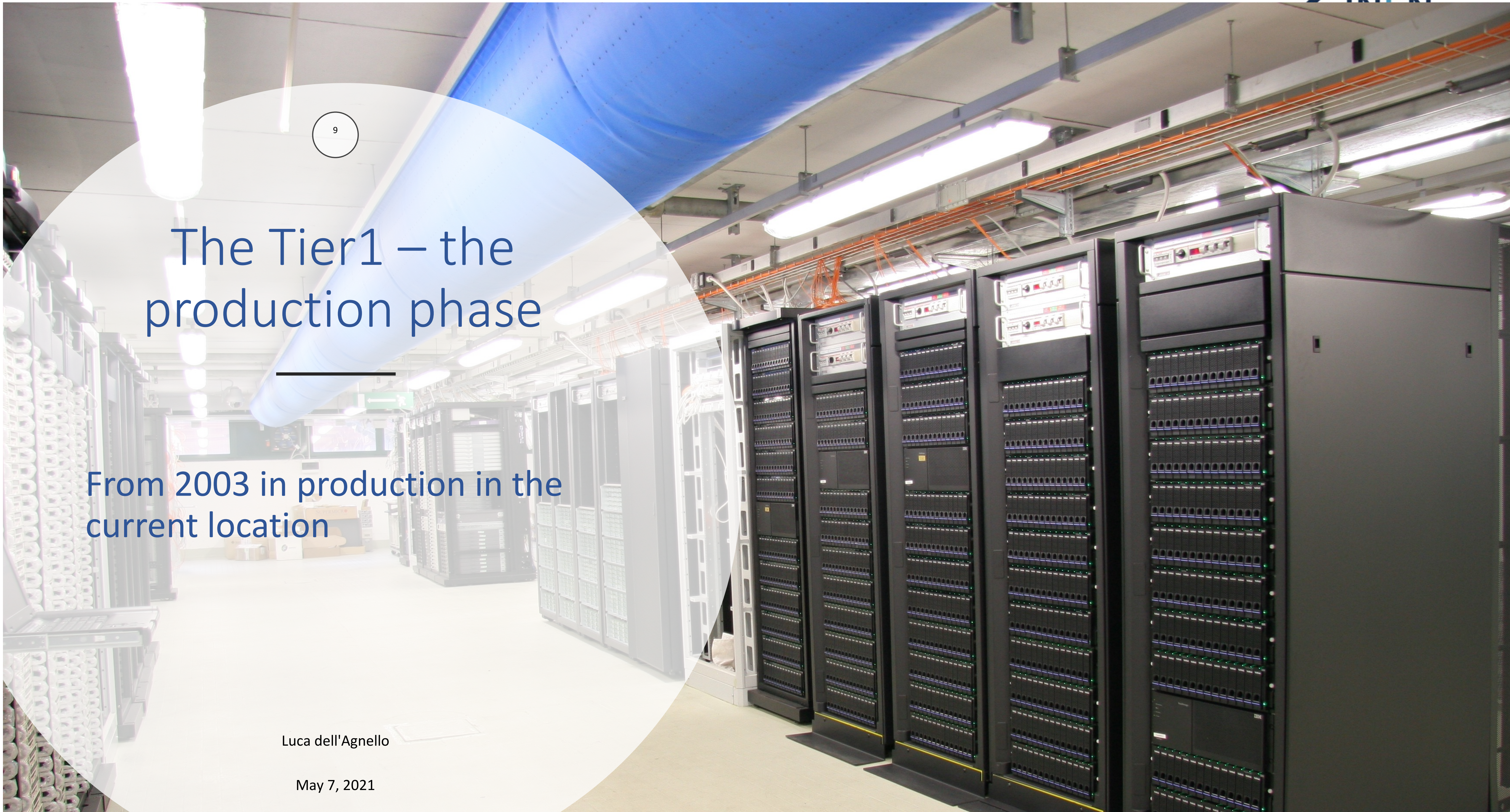
ISTITUTO NAZIONALE DI FISICA NUCLEARE

Progetto Speciale Centro Regionale di Calcolo Tier1
D1.3

Piano dettagliato degli acquisti di materiale

Autori
Luca Dell'Agnello, Pietro Matteuzzi, Federico Ruggieri, Stefano Zani
INFN – CNAF, Bologna

DRAFT V0.3 del 21 Luglio 2001



The Tier1 – the production phase

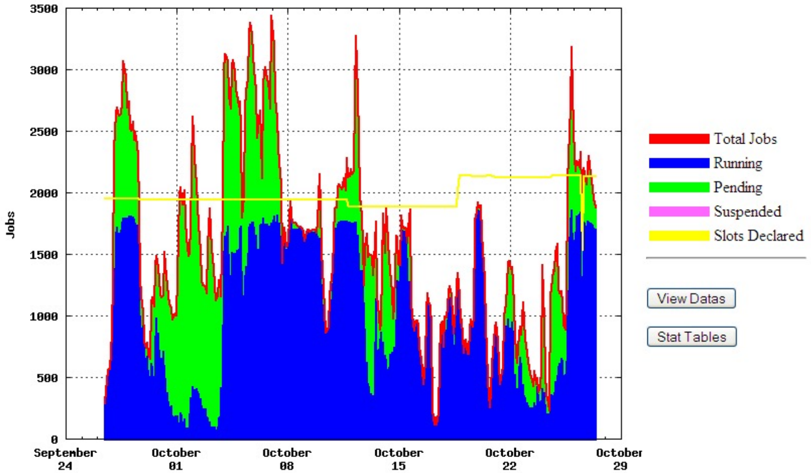
From 2003 in production in the current location

Luca dell'Agnello

May 7, 2021

The Tier1 – “first version” (2003-2007)

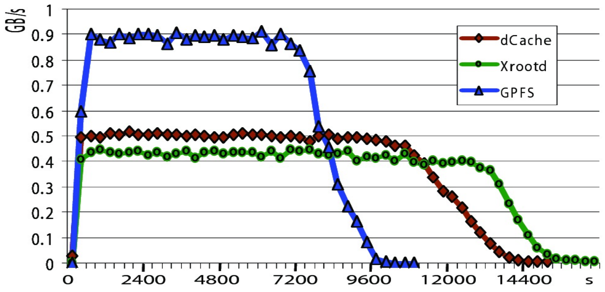
- Design, start-up and management coordination of the first farm (2003)
  - 160 1U bi-processor nodes (single core!), FastEthernet network
  - Managed by Torque/Maui
  - Static allocation of resources for each collaboration (Virgo, LHC)
    - For CDF and BaBaR legacy (dedicated) farms
    - Cluster ”segregation” with dedicated VLANs
- Since 2004 legacy farms merged in a common farm
  - ~1000 bi-processor nodes
  - Fair share mechanism implemented
  - Since 2005 farm managed by LSF
- Storage system still in evolution (2005)
  - GPFS, NFS and XRootD for disk-only storage (non grid access)
  - CASTOR as HSM (mainly for LHC experiments)
    - CASTOR SRM for Service Challenges from 2006
  - New tape library installed (StorageTek with 9840 and LTO drives)
- Design and set-up of AAI for CNAF
- WAN access: 1 Gbps
- LHCOPN: 10 Gbps (from 2006)



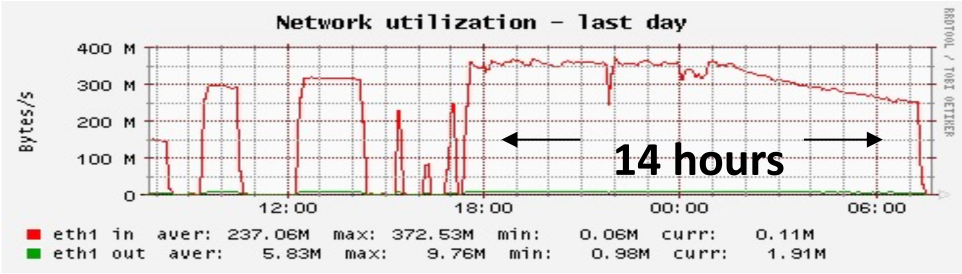
October 2005

The Tier1 – the “roaring years” (1/2)

- 2005-2011: coordinator of Tier1 Data Management group
  - Management of disk and tape storage, databases, FTS, etc.
  - Coordination with WLCG for Data Challenges, etc.
  - INFN representative in WLCG MB (2006-)
- First goal: consolidation of storage configuration
- Comparative test campaigns (2006) to validate storage model based on GPFS
  - CASTOR showed too demanding for support effort and not fitting with disk-only usage
    - Other issues with CASTOR acknowledged in 2007 from CERN after ATLAS (K. Boss) complaints and tests by B. Panzer-Steindel
  - Evaluated also Lustre and HEP specific solutions (dCache, XRootD)
- Measurements both with standard benchmarks and experiments real use cases
  - Stress test with real analysis and reprocessing jobs
- Also, StoRM tested in production environment



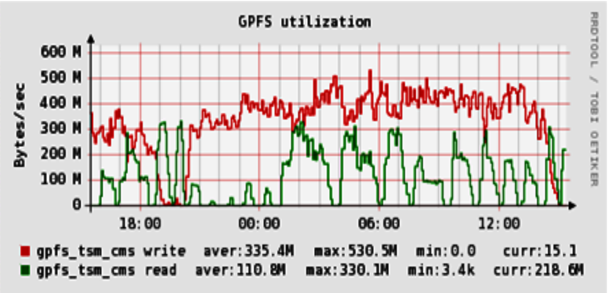
~ 1000 LHCb concurrent analysis jobs



15 parallel streams/transfer to GPFS
120 concurrent transfers (4 GridFTP servers)
~ 100K files (17 TB)

The Tier1 – the “roaring years” (2/2)

- 2006: All legacy storage systems (XRootD, NFS) moved to GPFS
- Disk-only system for LHC experiments (Atlas, LHCb) moved from CASTOR to GPFS (2007)
  - StoRM in production
- 2008: new library installed (SL 8500) with 1 TB tapes
- Sept 19, 2008: LHC RUN1 interrupted due to damage to some magnets ”following a large helium leak”
- 2008-2009: R&D to extend GPFS usage also to tape back-end storage (HSM) before the start of RUN1
  - GEMSS: integration of GPFS with TSM and interfaced to StoRM
    - Collaboration with IBM
  - Solution validated with a WLCG data challenge (STEP09)
- October 2009-March 2010: completed migration of all storage systems to GEMSS
  - LHC RUN1 just started….
- Q3 2010: transition to 10 Gbit technology on servers


GPFS utilization
Bytes/sec
600 M
500 M
400 M
300 M
200 M
100 M
0
18:00
00:00
06:00
12:00
RRDTOOL / TOBI OETIKER
🟥 gpfs_tsm_cms write aver:335.4M max:530.5M min:0.0 curr:15.1
🟩 gpfs_tsm_cms read aver:110.8M max:330.1M min:3.4k curr:218.6M

Throughput from tape 🟥

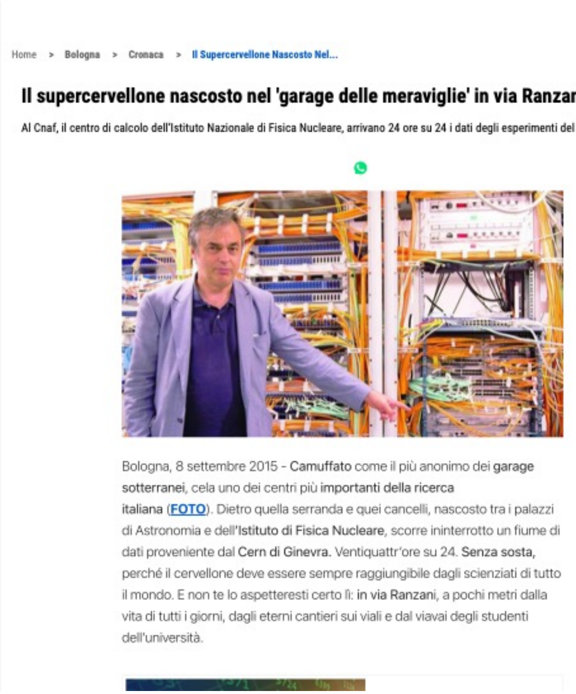
Throughput to tape. 🟩

In the meanwhile (2008):
Data center completely renewed
Study of virtualization – WNoDeS
Strong activity in the grid group (middleware development and infrastructure management)

The consolidation decade: 2011-present

- 2011—coordinator of Tier1
  - 5 groups (IT services, technological plants and User Support)
- Growing number of supported experiments (currently > 40)
- Optimization of performances of Tier1
  - LAN reconfigured to use JF
  - LHCOPN/ONE upgrade to 2x100 Gbps
  - Disk-servers” 10 → 20 → 40 → 2x100 Gbps
  - Farm always used 100%
  - Second library installed (2020) with 20 TB drives
- Organizational effort
  - Study and standardization of services offered to experiments
    - Crucial for Data Management
    - WLCG as term of reference
  - Reorganization of the support
  - Enforcing a more efficient collaboration among the groups
  - Bureaucratic work (tenders etc...)

In the meanwhile:
cloud@CNAF, Cloud@INFN, EPIC
A lot of projects (Indigo, XDC, DEEP, etc….)
SSNN and Information system

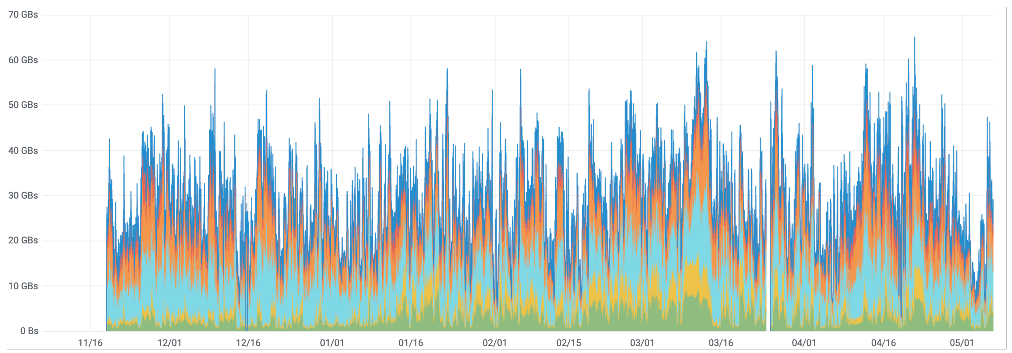


Home > Bologna > Cronaca > Il Supercervellone Nascosto Nel...

Il supercervellone nascosto nel 'garage delle meraviglie' in via Ranzani

Al Cnaf, il centro di calcolo dell'Istituto Nazionale di Fisica Nucleare, arrivano 24 ore su 24 i dati degli esperimenti del Cern

Bologna, 8 settembre 2015 - **Camuffato** come il più anonimo dei **garage sotterranei**, cela uno dei centri più **importanti della ricerca italiana** (FOTO). Dietro quella serranda e quei cancelli, nascosto tra i palazzi di Astronomia e dell'**Istituto di Fisica Nucleare**, scorre ininterrotto un fiume di dati proveniente dal **Cern di Ginevra.** Ventiquattr'ore su 24. **Senza sosta,** perché il cervellone deve essere sempre raggiungibile dagli scienziati di tutto il mondo. E non te lo aspetteresti certo lì: **in via Ranzani**, a pochi metri dalla vita di tutti i giorni, dagli eterni cantieri sui viali e dal viavai degli studenti dell'università.

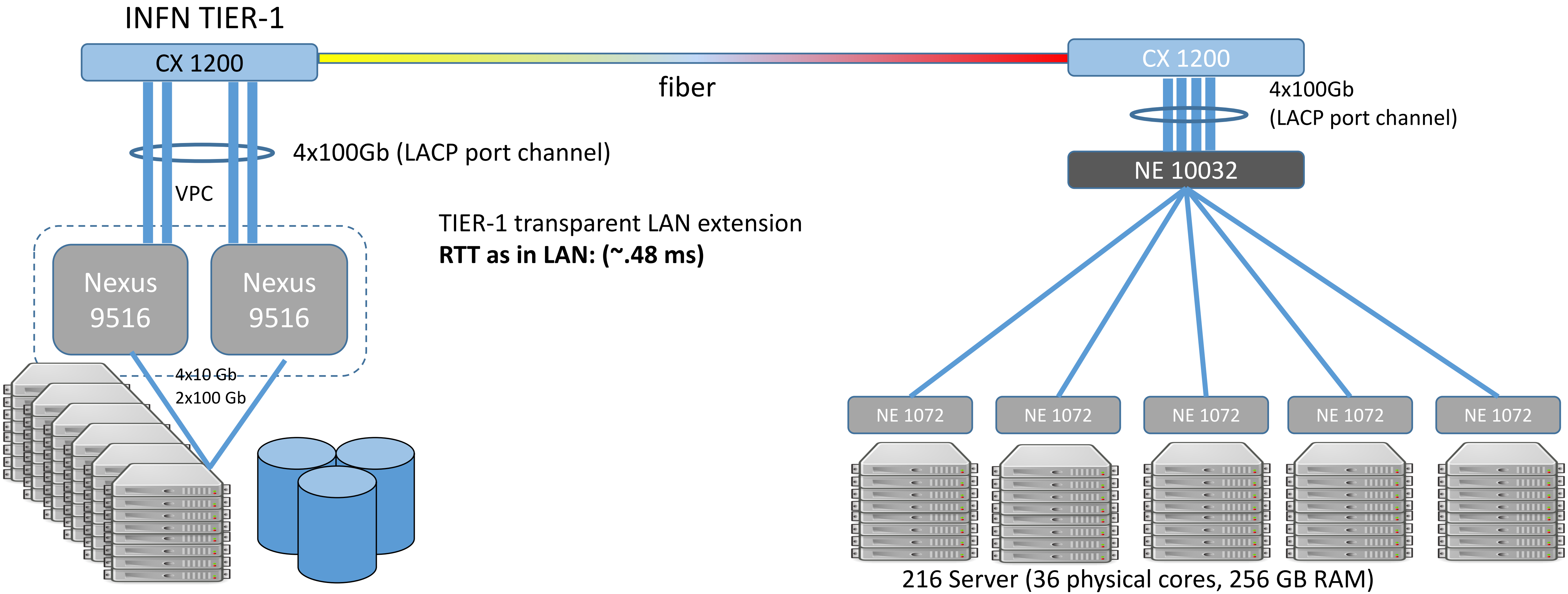


Reading access to servers

INFN Tier1 beyond CNAF data center

- Since 2015, we have started testing use of remote CPU resources to extend our data center beyond CNAF site in a transparent way for our users
  - INFN Tier1 CEs as unique access points also to these resources
  - Key issue is data access (i.e., remote or cache)
- 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
  - Grant to WLCG experiments (ended Q1 2021)

CNAF - CINECA Data Center Interconnection

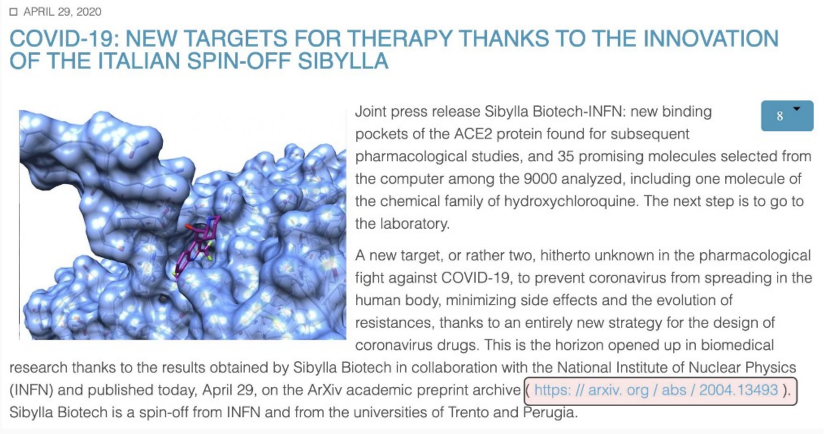


TIER-1 transparent LAN extension
RTT as in LAN: (~.48 ms)

216 Server (36 physical cores, 256 GB RAM)
~ 180 kHS06, Intel® Xeon® CPU E5-2697 v4

COVID-19

- Participation to EU project Exscalate4Covid
  - We provide data archiving service
- March 2020: SIBYLLA BIOTECH (spin-off from INFN) needed computing power to simulate the 3D folding of ACE2 protein
  - ~half of Tier1 farm and many resources from Tier2s used for this goal
  - Quite straightforward reconfiguration of our resources for this simulation
    - Our configuration does not depend on the particularity of any experiment

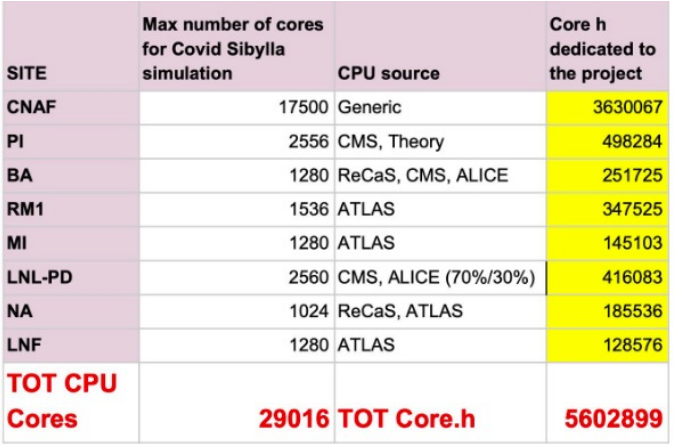


□ APRIL 29, 2020

COVID-19: NEW TARGETS FOR THERAPY THANKS TO THE INNOVATION OF THE ITALIAN SPIN-OFF SIBYLLA

Joint press release Sibylla Biotech-INFN: new binding pockets of the ACE2 protein found for subsequent pharmacological studies, and 35 promising molecules selected from the computer among the 9000 analyzed, including one molecule of the chemical family of hydroxychloroquine. The next step is to go to the laboratory.

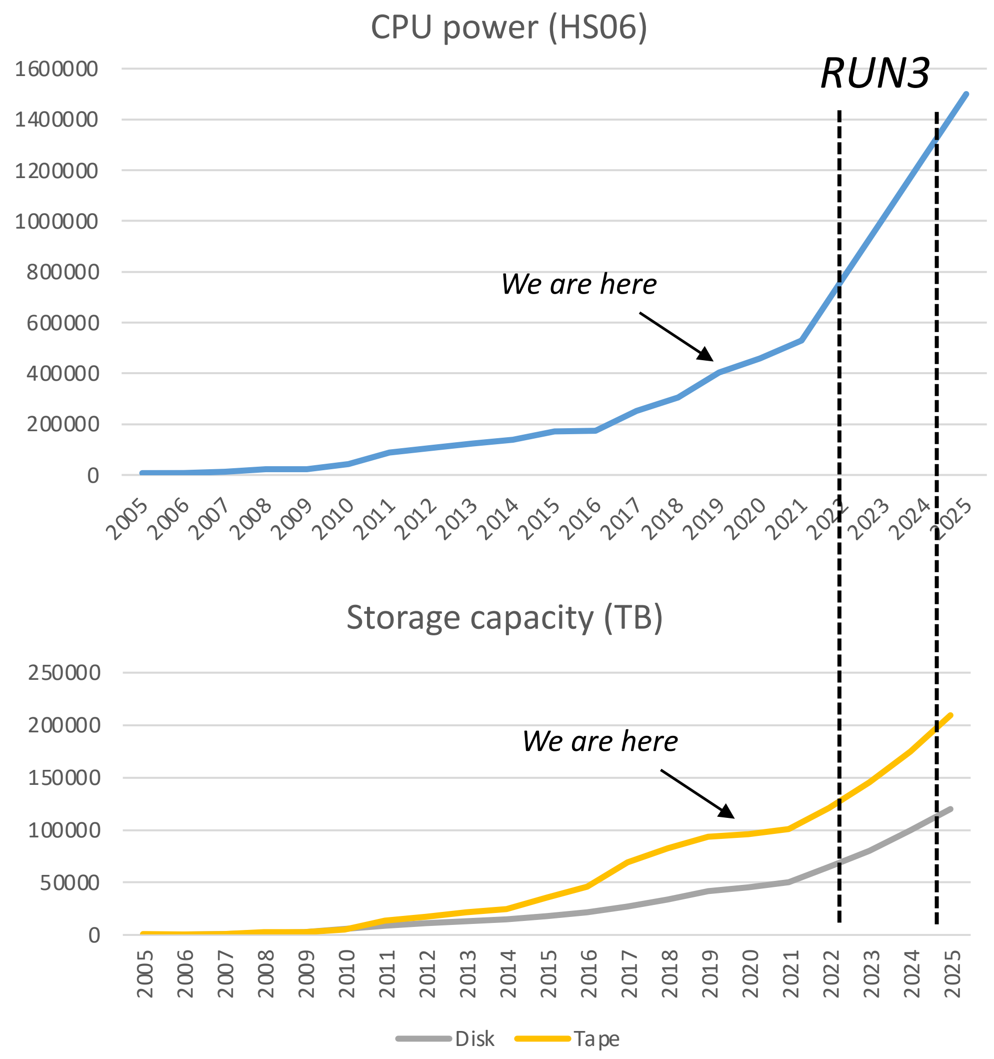
A new target, or rather two, hitherto unknown in the pharmacological fight against COVID-19, to prevent coronavirus from spreading in the human body, minimizing side effects and the evolution of resistances, thanks to an entirely new strategy for the design of coronavirus drugs. This is the horizon opened up in biomedical research thanks to the results obtained by Sibylla Biotech in collaboration with the National Institute of Nuclear Physics (INFN) and published today, April 29, on the ArXiv academic preprint archive (https: // arxiv. org / abs / 2004.13493). Sibylla Biotech is a spin-off from INFN and from the universities of Trento and Perugia.

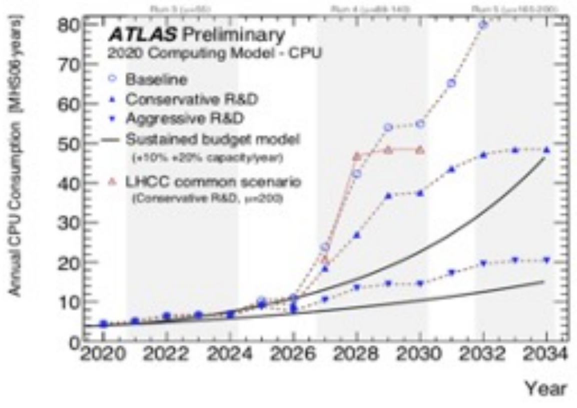


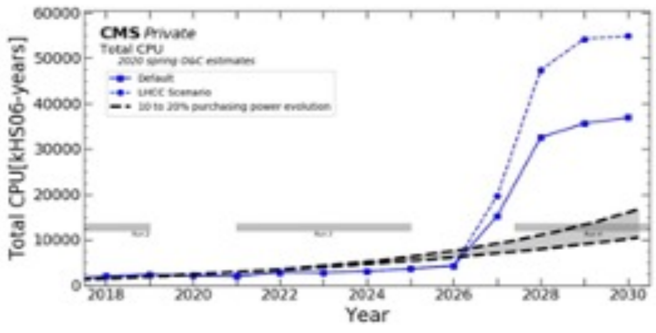
SITE	Max number of cores for Covid Sibylla simulation	CPU source	Core h dedicated to the project
CNAF	17500	Generic	3630067
PI	2556	CMS, Theory	498284
BA	1280	ReCaS, CMS, ALICE	251725
RM1	1536	ATLAS	347525
MI	1280	ATLAS	145103
LNL-PD	2560	CMS, ALICE (70%/30%)	416083
NA	1024	ReCaS, ATLAS	185536
LNF	1280	ATLAS	128576
TOT CPU Cores	**29016**	**TOT Core.h**	**5602899**

The Tier1: some figures

- More than 40 scientific collaborations supported
  - Farm power: 420 kHS06 (~42k cores)
  - ~34 PB of disk, ~90 PB of tape
- Huge increase of resources foreseen in the coming years
  - By 2025:
    - 1500 kHS06 (~120-140k cores) for the farm
    - 120 PB (net) of disk
    - 200 PB of tapes
  - And even more (x3-4 wrt 2025) from 2027 (HL-LHC)
- Present data center could host resources up to the end of Run3 but not for HL-LHC (2017 survey)

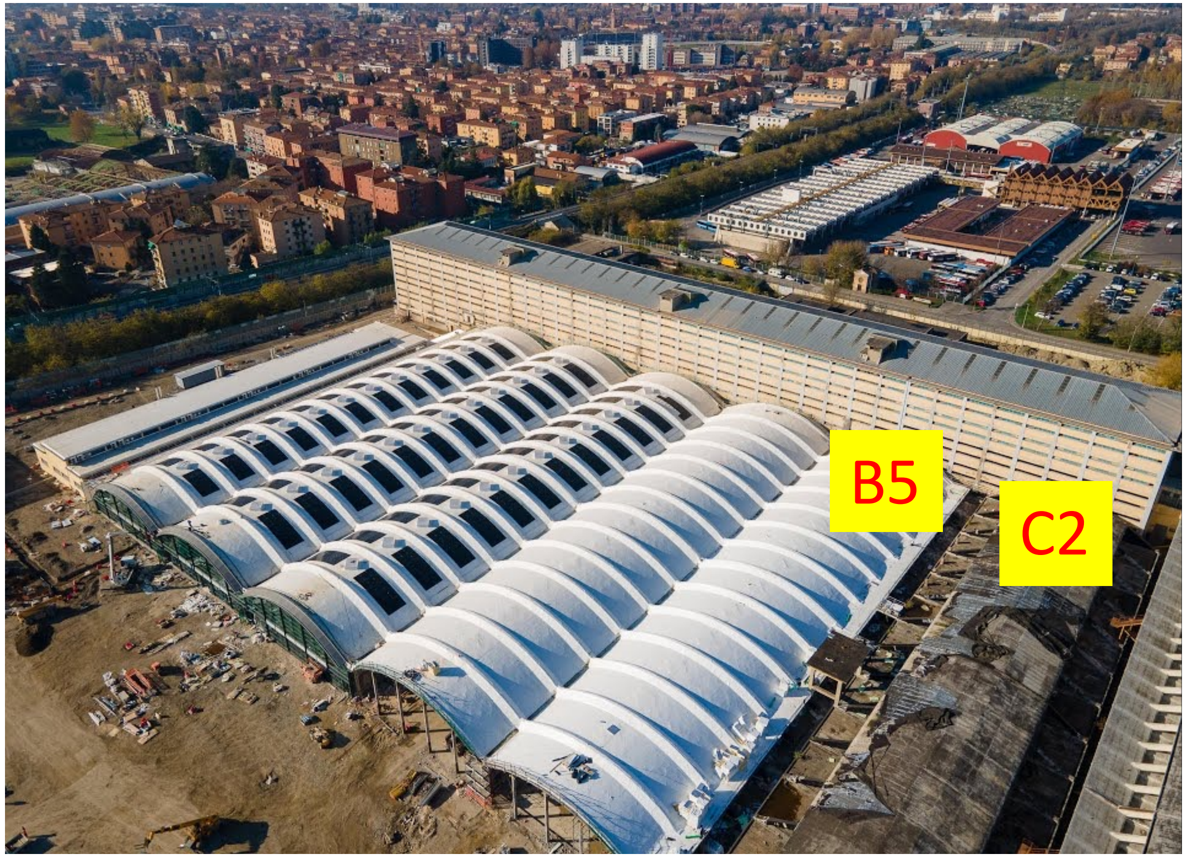






The new challenge

- Long term project to satisfy computing needs of INFN for several years
  - 1st phase driven by LHC Run 3 (2022-2025)
  - 2nd phase driven by HL-LHC (2027-)
- INFN Tier1 will be migrated to the hall B5
- The pre-exascale machine ***Leonardo*** will be installed in the adjacent hall C2
- A direct network interconnection will be available between the two data centers
- INFN and CINECA data centers will share
  - Technological plants (power and cooling)
  - Security and surveillance services

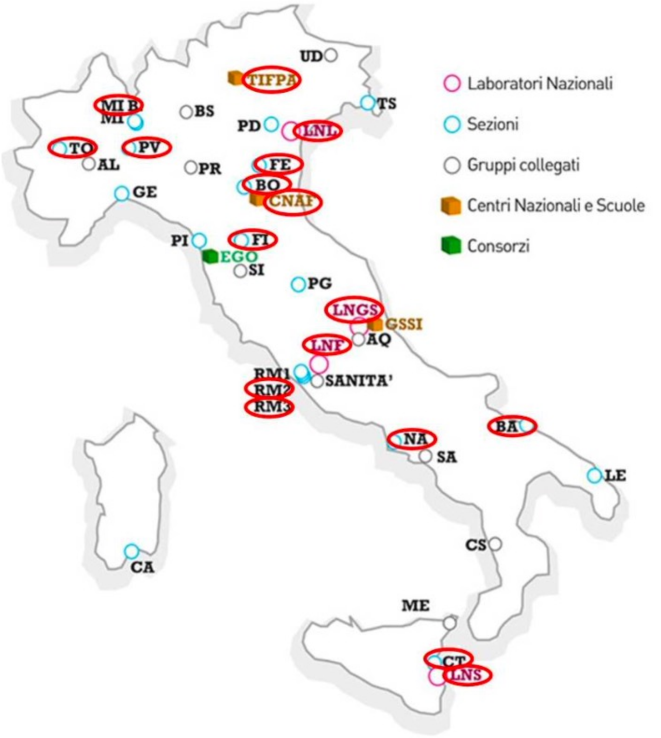


Summary

- The next 4 years will be crucial for CNAF
  - Migration (w/o service interruption) of the data center to the Tecnopolo area
  - Evolution of services for the HL-LHC era
  - Reshape the CNAF organization to better fit the new requirements
- In preparation for the HL-LHC phase, a major evolution of the INFN computing infrastructure is foreseen
  - Shift to data lake and cloud paradigms
- CNAF will play a central role having the capabilities and the staff to manage this evolution
- With the migration at Tecnopolo, CNAF will be able to consolidate its strategic role in the framework of INFN computing

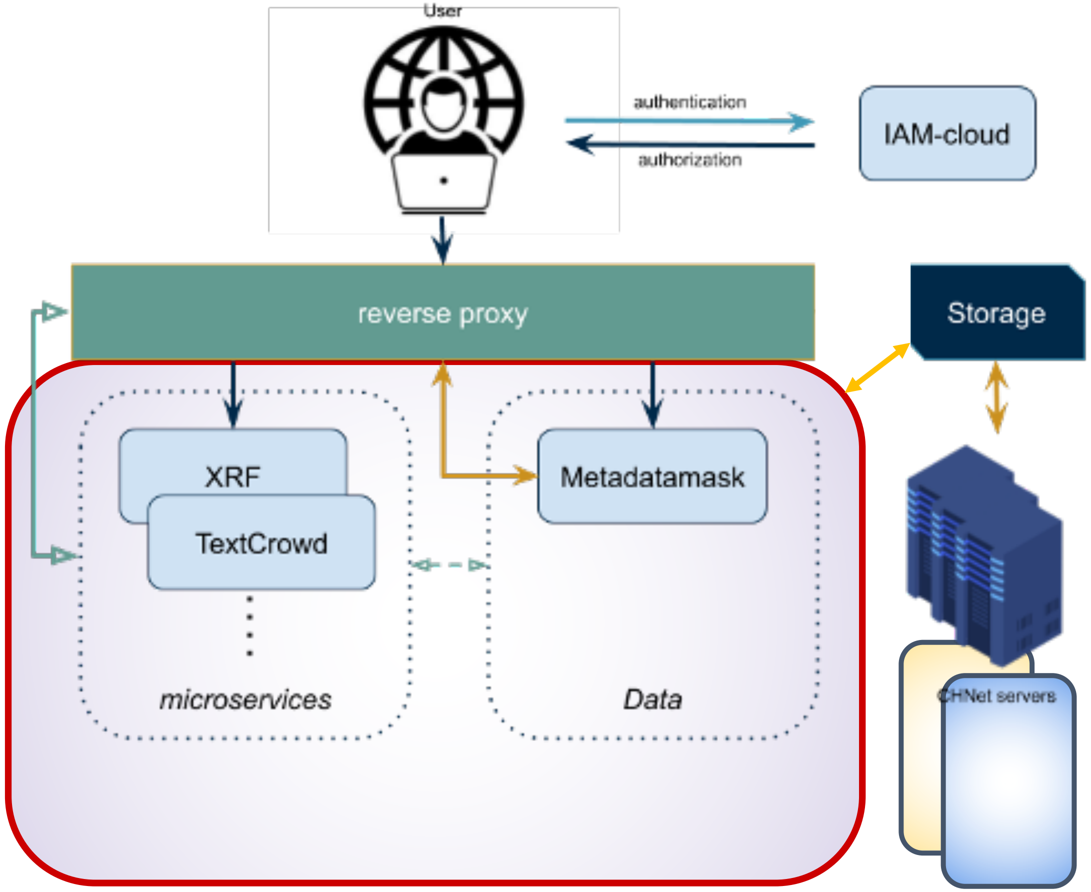
Bonus track: CHNeT and 4CH project

- CHNet **(Cultural Heritage Network)** is the network of the Italian **National Institute for Nuclear Physics (INFN)** devoted to Cultural Heritage
- Leveraging on skills and techniques developed for Nuclear Physics.
  - giving indications to restorers about the correct procedures to be applied for restoration/conservation;
  - characterising materials and manufacturing techniques;
  - studying the provenance of the raw materials (in order to retrace ancient trade routes or to use original materials during restorations);
  - giving indications about material authenticity.
- EU **4CH** project started Jan 2021
  - INFN lead partner
  - Prototyping a European Competence Centre for the conservation of Cultural Heritage
  - In the medium term the data and the services will be hosted at Tecnopolo



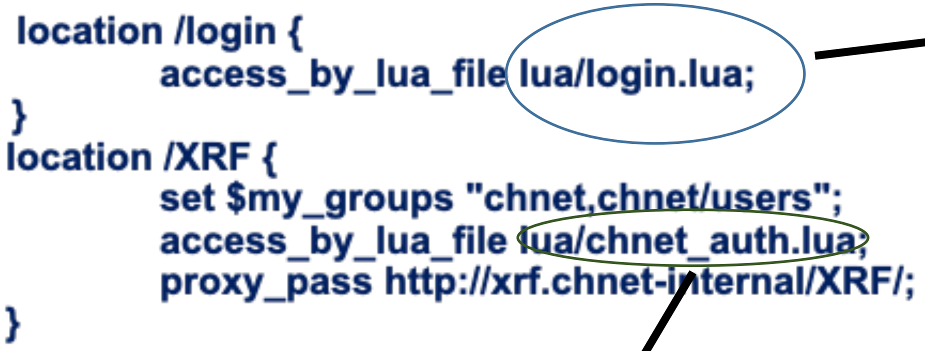
The portal

- The goal is to create a digital infrastructure for the services and all the data acquired by INFN-CHNet laboratories, allowing to access them and, through online tools, reuse them for further analysis
- Modular architecture based on containers
  - Each application structured on frontend and backend containers
  - Reverse Proxy container in front of all services as unique entry point



The reverse proxy

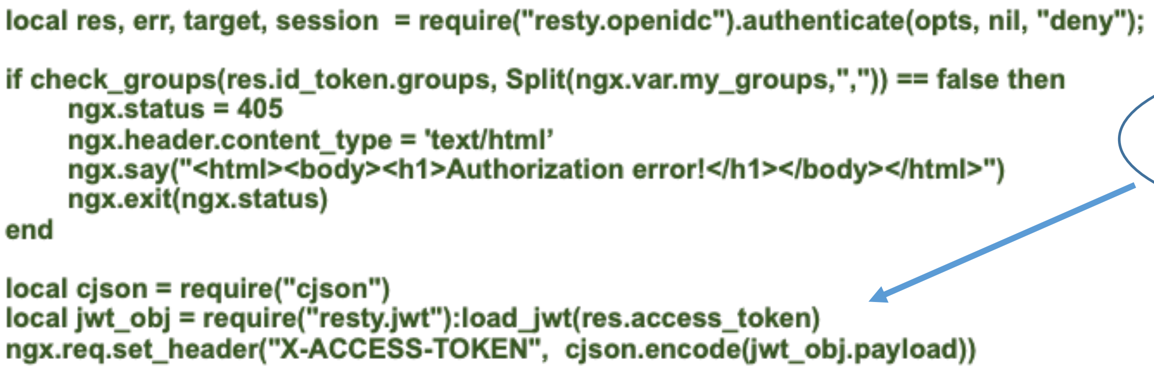
- Implementation is based on Nginx (openresty) + lua
- lua-resty-openidc module to support OpenID Connect and Oauth 2.0



```
location /login {
        access_by_lua_file lua/login.lua;
}
location /XRF {
        set $my_groups "chnet,chnet/users";
        access_by_lua_file lua/chnet_auth.lua;
        proxy_pass http://xrf.chnet-internal/XRF/;
}
```

```
local opts = {
  redirect_uri = ngx.var.redirect_uri,
  discovery = ngx.var.discovery,
  client_id = ngx
  client_secret =
  scope = "openi
  lua_ssl_verify_depth = 2",
  lua_ssl_trusted_certificate =….,
  ssl_verify = "no",
  renew_access_token_on_expiry = true,
  revoke_tokens_on_logout = true
}
local res, err = require("resty.openidc").authenticate(opts)
```

Id and access tokens saved in session

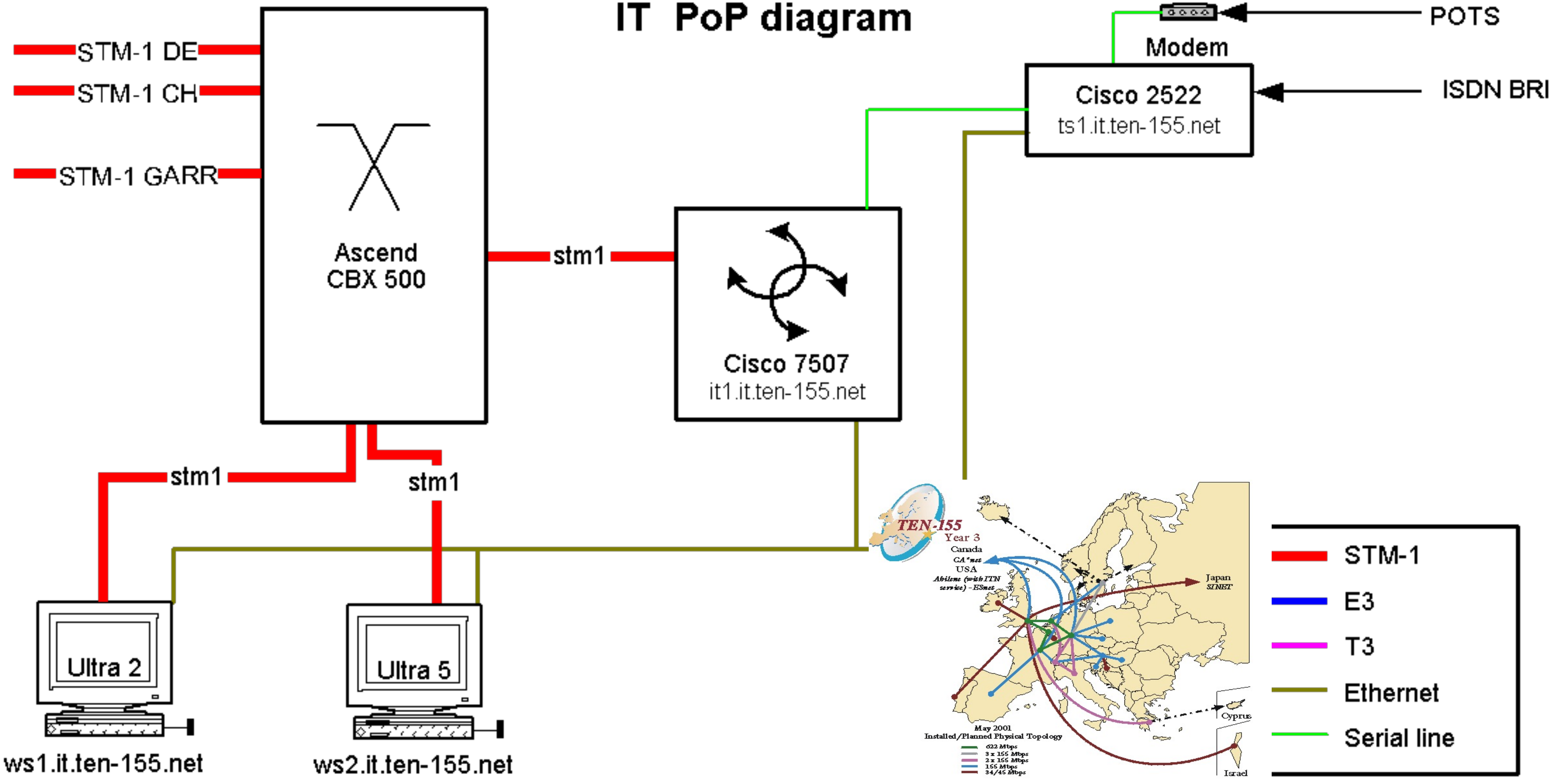


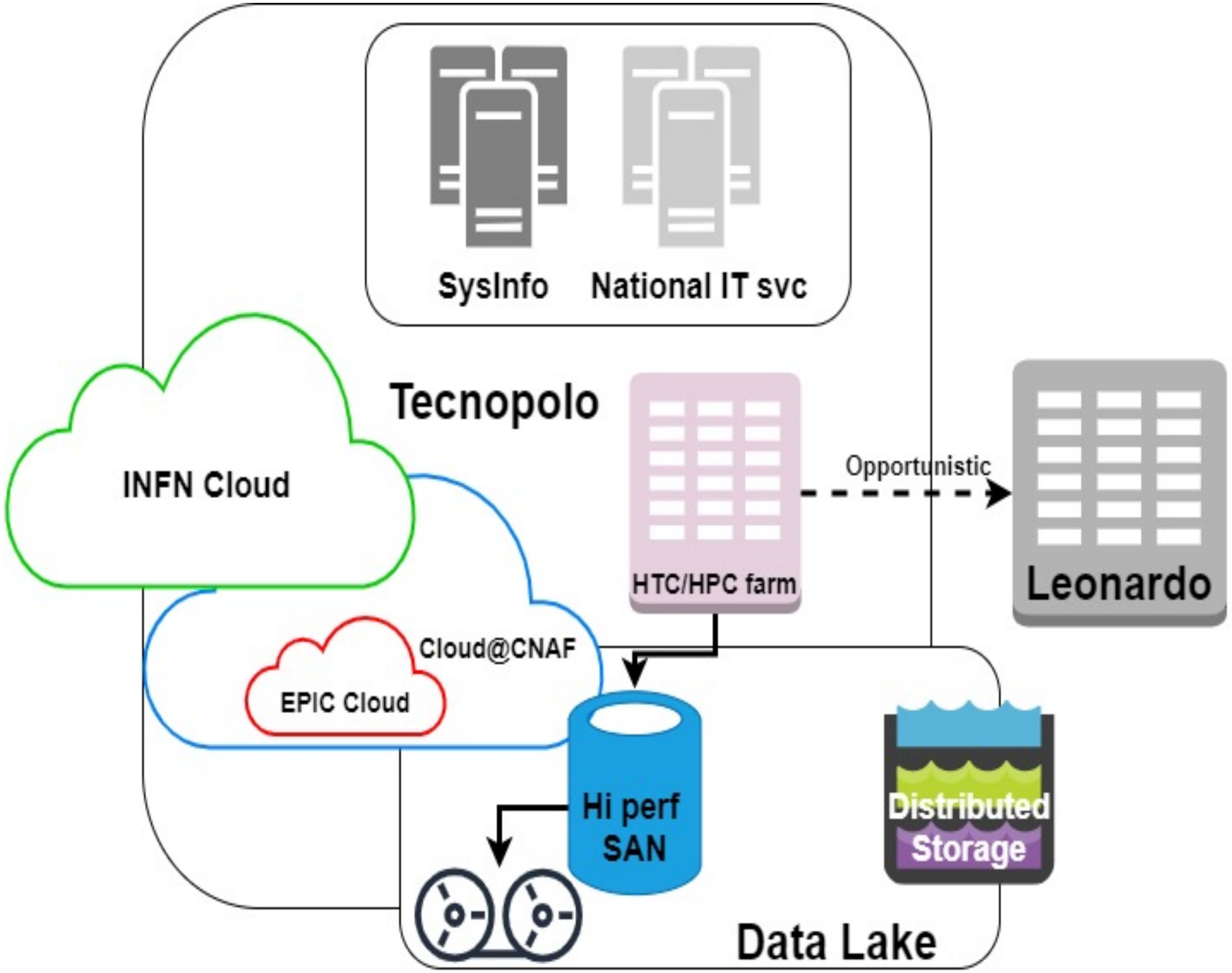
```
local res, err, target, session  = require("resty.openidc").authenticate(opts, nil, "deny");

if check_groups(res.id_token.groups, Split(ngx.var.my_groups,",")) == false then
     ngx.status = 405
     ngx.header.content_type = 'text/html'
     ngx.say("<html><body><h1>Authorization error!</h1></body></html>")
     ngx.exit(ngx.status)
end

local cjson = require("cjson")
local jwt_obj = require("resty.jwt"):load_jwt(res.access_token)
ngx.req.set_header("X-ACCESS-TOKEN",  cjson.encode(jwt_obj.payload))
```

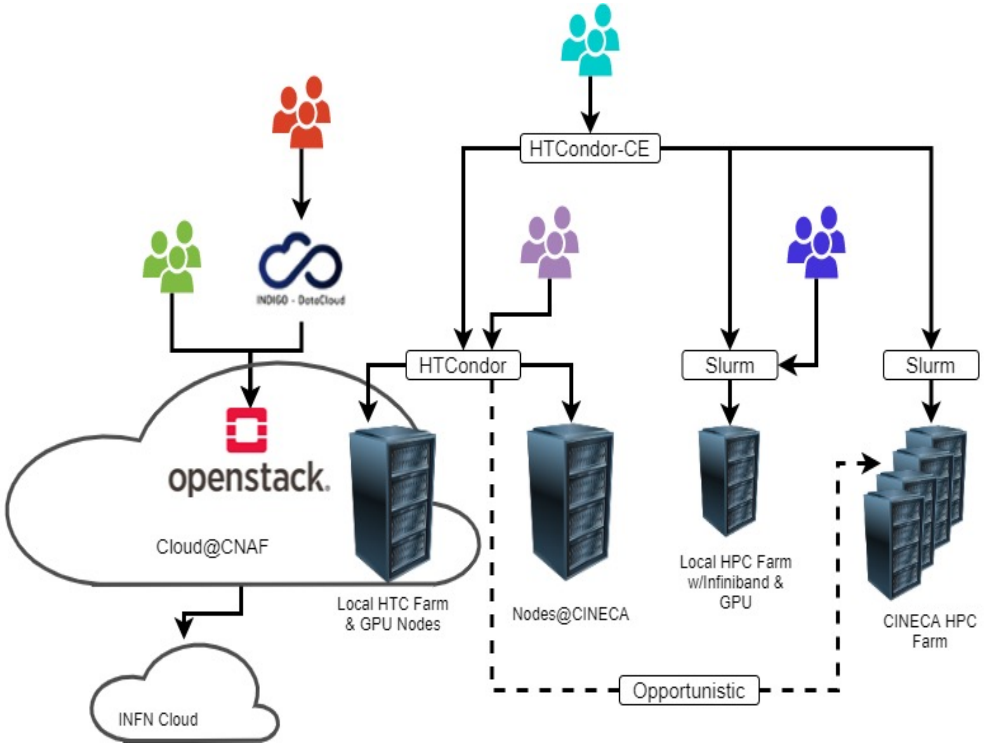
Access token passed to upstream as header

Backup slides

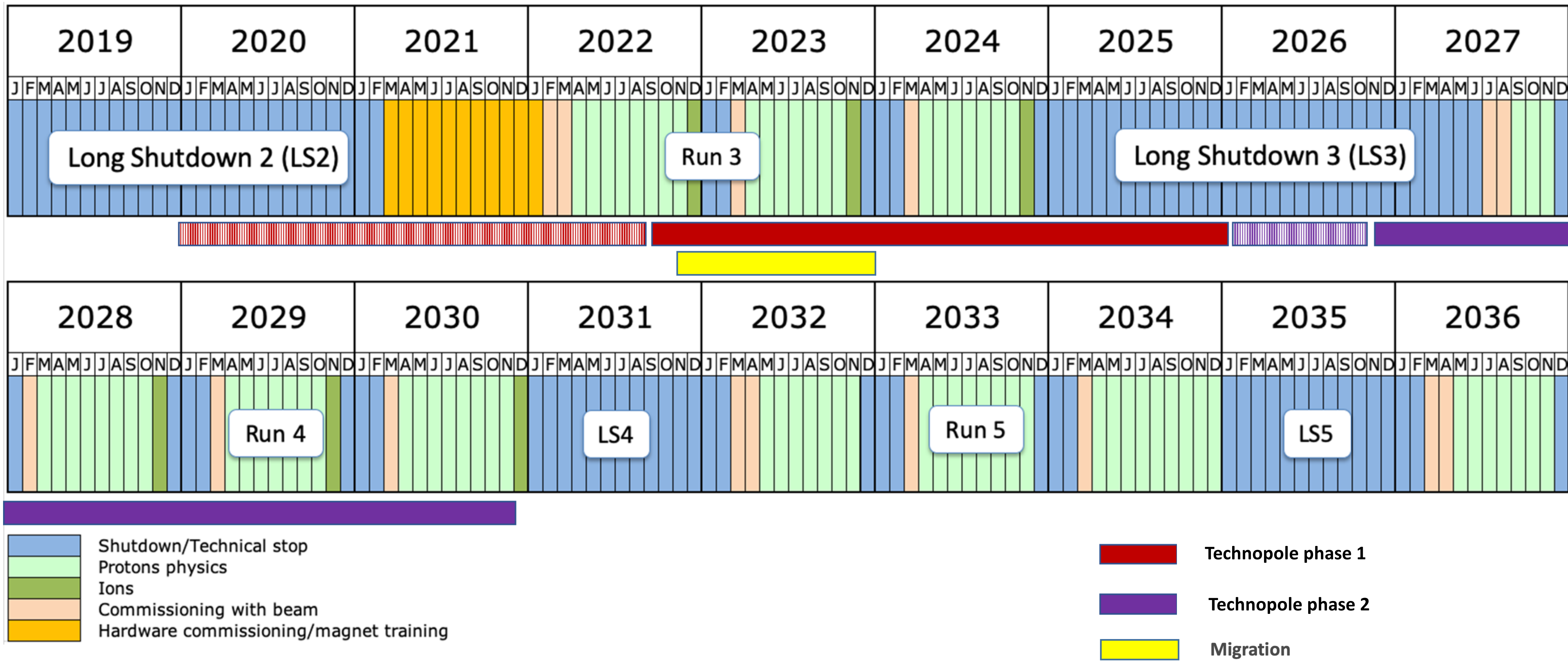
IT PoP diagram




SysInfo
National IT svc
Tecnopolo
INFN Cloud
HTC/HPC farm
Opportunistic
Leonardo
Cloud@CNAF
EPIC Cloud
Hi perf SAN
Distributed Storage
Data Lake


HTCondor-CE
INDIGO - DataCloud
HTCondor
Slurm
Slurm
openstack.
Cloud@CNAF
Local HTC Farm & GPU Nodes
Nodes@CINECA
Local HPC Farm w/Infiniband & GPU
CINECA HPC Farm
Opportunistic
INFN Cloud

Tecnopolo vs. LHC schedule



Technological plants

Central infrastructures (sized for both phase 1 and phase 2)

  - Primary electrical and hydraulic distribution in data halls B5 and C2
  - Located in halls G1, G3 and on the second floor of C2
  - 9460 m^2 in the halls (G1,G3,C2) + 3525 m^2 of tunnels to data halls

- Secondary infrastructures
  - Power distribution, hydraulic distribution for temperate and cold water, the ventilation and conditioning system, security systems and access control
  - Located within data center (or near it)
- Power distribution: 4 main branches (3+1 redundancy) consisting of 4x3200 A busways
  - Phase 1: up to a maximum of 3 MW under UPS with (3+1) redundancy (4x1 MW UPS)
- Hydraulic distribution (3+1 redundancy)
  - Temperate water (40-50 ^{0}C) for racks with DLC (40-80 kW/rack)
  - Chilled water (19-26 ^{0}C) for racks with rear door and/or plenum cooling (up to 30 kW/rack)

The main characteristics of the two phases

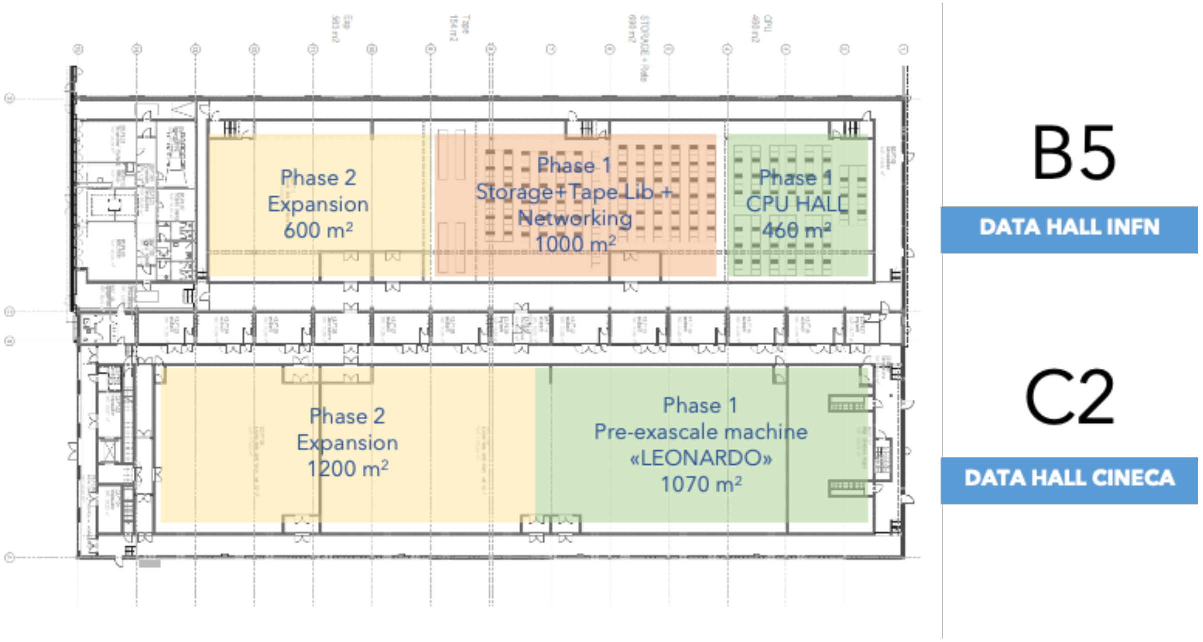
IT power [MW]	Phase 1	Phase 2
Years	2021-2026	2027-2030
INFN	2	3-10
CINECA	8	10
Total	10	13-20

Tier1@CNAF:

- *800 m^2 for IT*
- *Power:1.4 MW*

PUE@Tecnopolo: 1.1

PUE@CNAF: 1.6



IT surface [m^2]	Phase 1	Phase 2	Future Expansion (requested)
Years	2021-2026	2027-2030	Beyond 2030
INFN	1560	2060 (+500)	3060 (+1000)
CINECA	1000	2200 (+1200)	3200 (+1000)
Total	2560	4260 (+1700)	6260 (+2000)

Surface for Technical Infrastructures

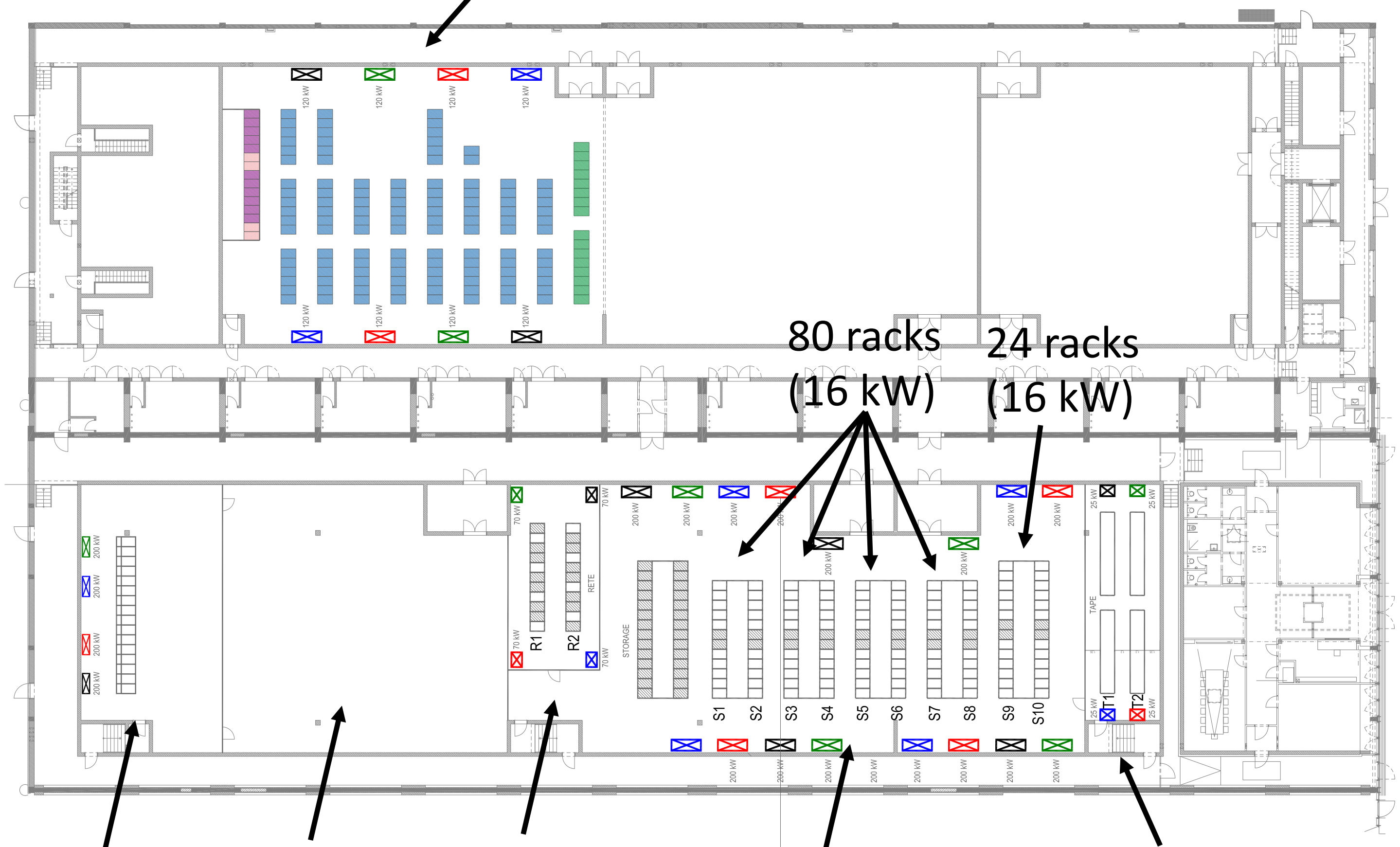
Plants [m^2]	Tunnels [m^2]
9460	3525

The layout of the data halls

Leonardo

CINECA data hall

INFN data hall



80 racks (16 kW)

24 racks (16 kW)

CPU area

Expansion area

Network core

Storage/services area

Tape libraries

Requirements

2025:

- CPU (~1 MHS06) – part of pledges from Leonardo + ~3-8 racks
- Disk (~100-120 net PB) – from 25 to 60 racks depending on the storage model
- Tape – space to install up to 4 libraries
- Services – 24 racks

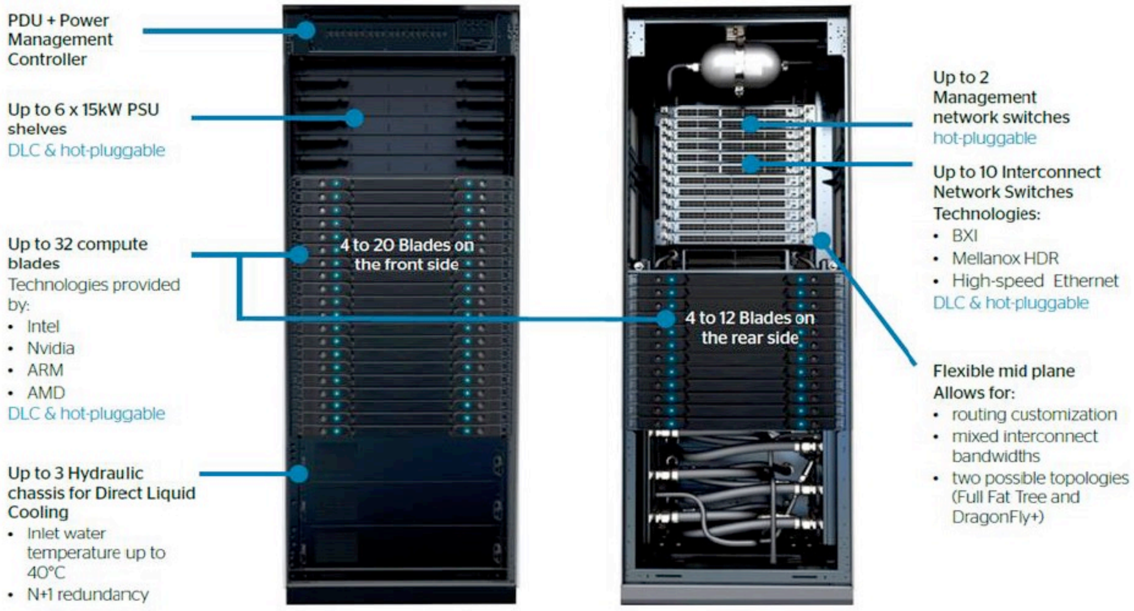
Storage area with 16 kW racks

2 options for racks in CPU area:

- 40 kW racks (with cooling plenum)
- Racks with DLC (80-90 kW)

Leonardo: the pre-exascale machine

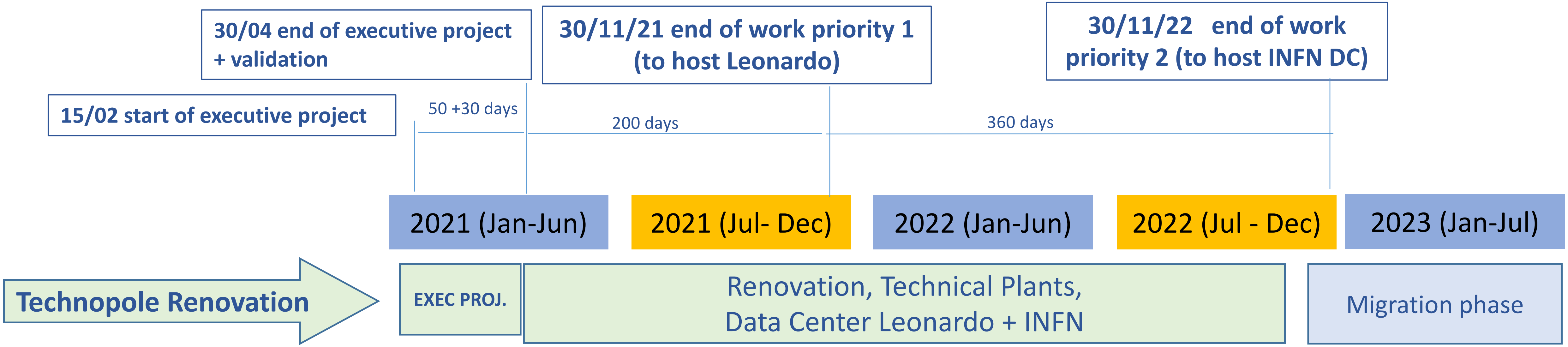
- The tender was assigned to ATOS-BULL (August 2020)
  - Delivery scheduled for the end of 2021
- 2 partitions
  - General Purpose (~3 MHS06)
    - 2 **Sapphire Rapids** CPUs/node
      - No Ethernet: card
      - Interconnection to Tier1 via IB
      - Skyway IB-Ethernet (2x1.6 Tbits)
  - HPC (“boost”)
    - 1 CPU (Ice Lake) + 4 Nvidia GPUs/node
- A fraction of the GP partition will be used for our experiments
  - Opportunistic use also possible


PDU + Power Management Controller
Up to 6 x 15kW PSU shelves
DLC & hot-pluggable
Up to 32 compute blades
Technologies provided by:
• Intel
• Nvidia
• ARM
• AMD
DLC & hot-pluggable
Up to 3 Hydraulic chassis for Direct Liquid Cooling
• Inlet water temperature up to 40°C
• N+1 redundancy
4 to 20 Blades on the front side
4 to 12 Blades on the rear side
Up to 2 Management network switches
hot-pluggable
Up to 10 Interconnect Network Switches
Technologies:
• BXI
• Mellanox HDR
• High-speed Ethernet
DLC & hot-pluggable
Flexible mid plane
Allows for:
• routing customization
• mixed interconnect bandwidths
• two possible topologies (Full Fat Tree and DragonFly+)

1 rack = 96 WNs ~200 kHS06
80 kW (DLC)

- Possibility to buy additional 3-4 racks for the Tier1 is under evaluation

The road to Tecnopolo



30/04 end of executive project + validation

30/11/21 end of work priority 1 (to host Leonardo)

30/11/22 end of work priority 2 (to host INFN DC)

15/02 start of executive project

50 +30 days

200 days

360 days

2021 (Jan-Jun)

2021 (Jul- Dec)

2022 (Jan-Jun)

2022 (Jul - Dec)

2023 (Jan-Jul)

Technopole Renovation

EXEC PROJ.

Renovation, Technical Plants,
Data Center Leonardo + INFN

Migration phase

- Use of the buildings granted by RER to INFN&CINECA (35 years)
- Preliminary renovation works started
- Executive project to be finalized end of Q1 2021
- Agreement in discussion with TERNA (ENEL in the 1st phase) for the provisioning of electric power
- Framework agreement is being finalized between INFN and CINECA
  - Management of technological plants
  - Use of Leonardo computing time

Summary

- 2021-2022: preparation of the data halls and technological plants
  - 3 MW of power (phase 1, 2022-2026)
  - Up to 10 MW of power (phase 2, 2027-)
- End of 2022: start of migration
  - Exploit Leonardo for part of CPU pledges
- Migration planned to be done with no downs (or almost)
  - Storage systems moved one by one with previous data copy (no down)
  - Down of one library at a time (O(10 days)) will not prevent data being transferred from CERN
  - Down scheduled for network routing change (O(1 h))
- Current data center will remain in production until the end of 2023