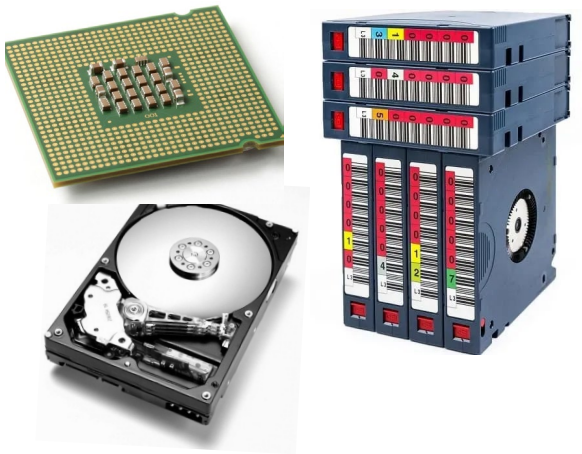
T1 Highlights

Referaggi C3SN 02-03 Settembre 2026

Pledge Risorse T1 2026/2027
HTC,HPC,Cloud - NO EPIC

TOTAL	Pledge HTC 2026	Pledge Cloud 2026	richieste 2027 HTC (delta)	Richieste Cloud/Bubbles 2027 (delta)	Rimpiazzi
CPU Pledge (kHS23)	1108387 TOT 886710 w OF	32000 TOT 25500 w OF	CSN1: 18.8 CSN2: 50.71 + 3 GPU CSN3:10+12.9 TOT: 82	CSN2: 3.410core + 1 GPU (include virgo?) CSN3: 2UI dedicate LEA: 1GPU dedicata CSN4: 8.45Mch su bubble 150kch su bubble Inf 65kGPUh su bubble Inf + CSN5	110
Disk Pledge (TBN)	113.400 TOT + 10.000 ICSC	1050	CSN1: 5700 CSN2: 3158 CSN3: 410 + 2500 TOT: 11.8	CSN2: - 136 LEA: 1 veloce + 10 CSN4: 80TB su bubble	9000
Tape Pledge (TB)	287.059 TOT + 36.030 ExtraP	0	CSN1: 20000 CSN2: 6490 CSN3: 350 + 5400 TOT: 32240	0	0

Storage veloce – da capire costo in caso di richieste non triviali



CMS e ATLAS: no CPU aggiuntive

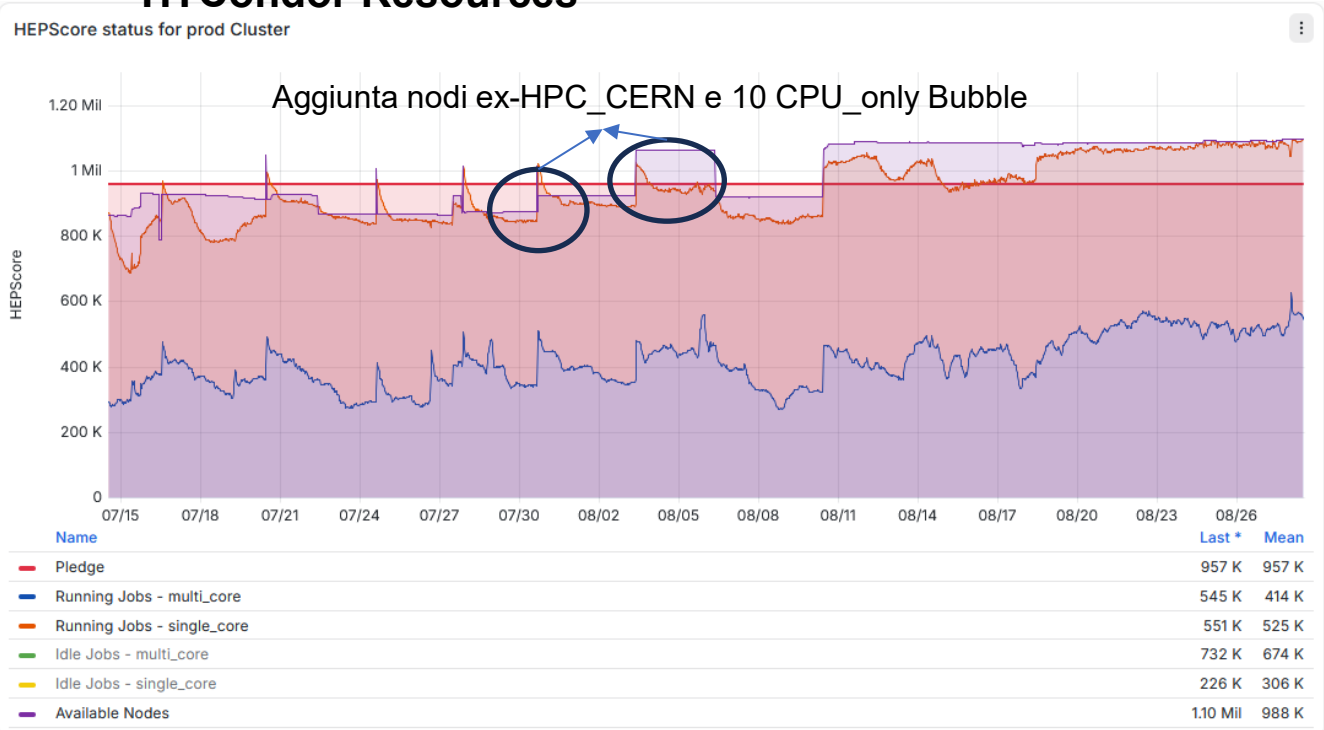
CSN5 Preview (EPIC+noEPIC):

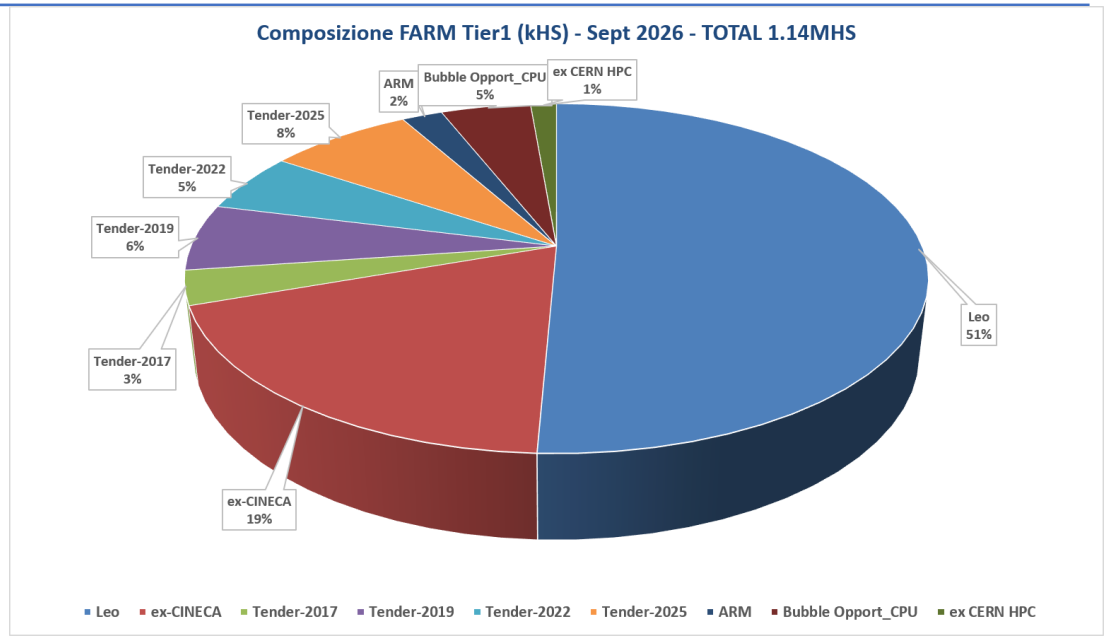
- 2244 core annuali
- 26653 kcore-hour ➔ 3kcore annuali
- **32 + 4(epic) GPU annuali**
- **100000 GPU-hr ➔ 12 GPU annuali**
- 350 + 200(epic) TB DISK HDD
- **102 + 1(epic) TB DISK NVMe**
- NO TAPE

CPU – Farm, HPC, Cloud

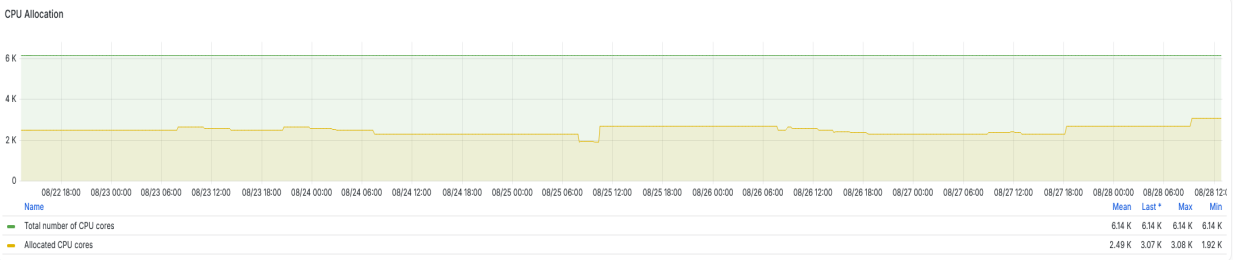
CPU: FARM & CLOUD & HPC

HTCondor Resources

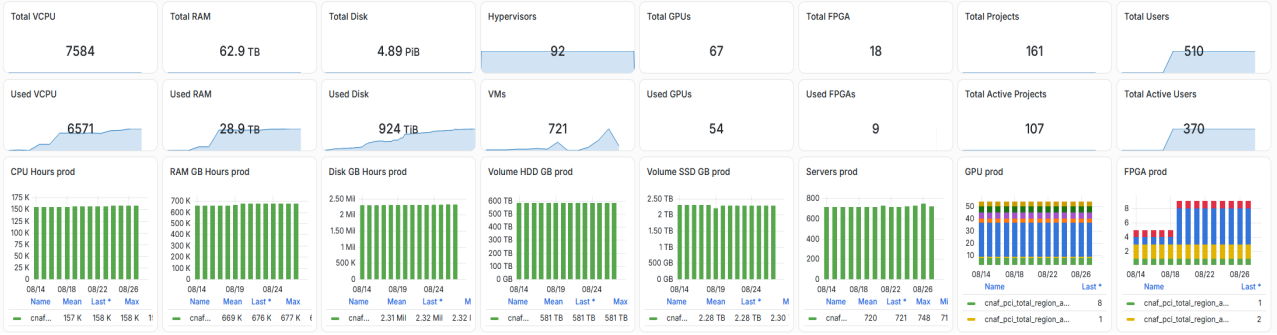




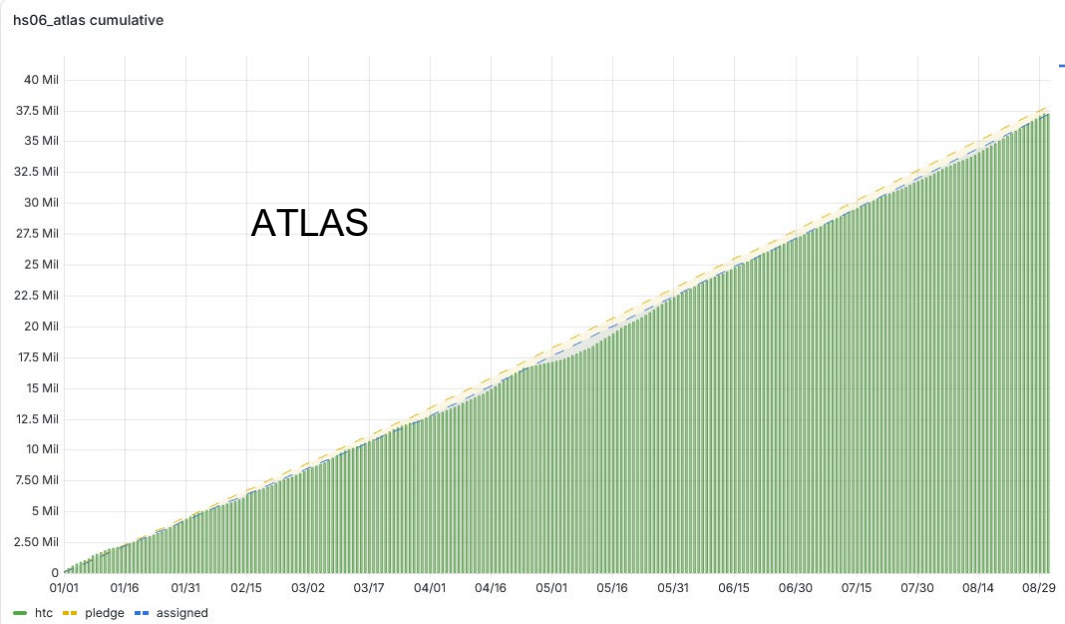
SLURM CPU Allocation @ HPC Bubbles

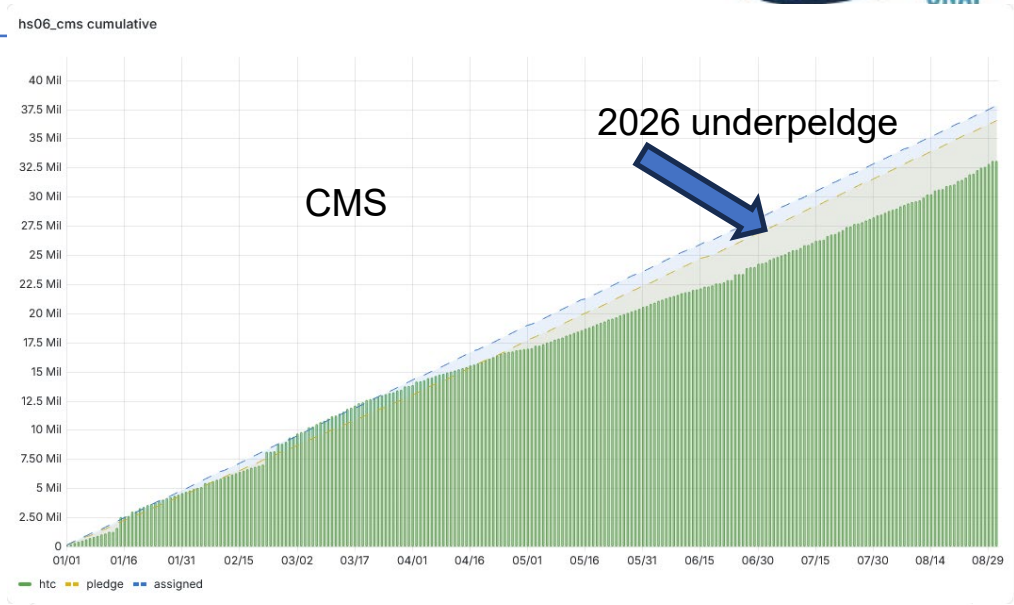


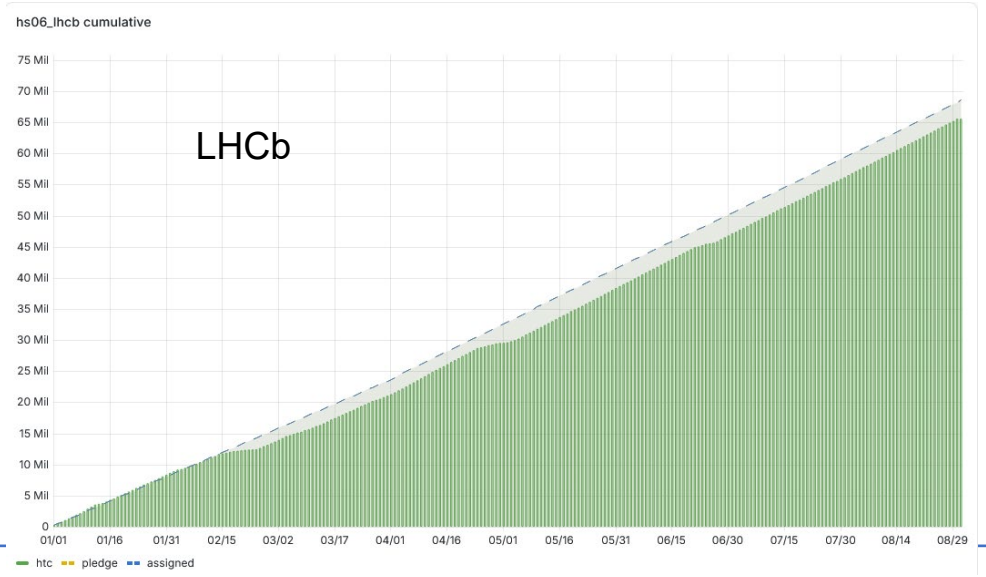
CLOUD RESOURCES (NO EPIC)

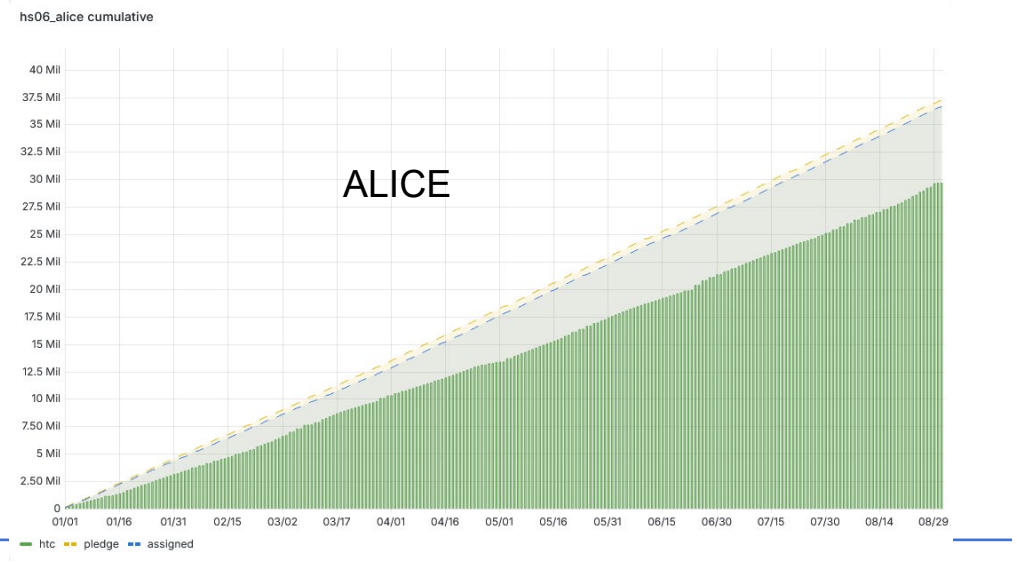


LHC Cumulative CPU Accounting

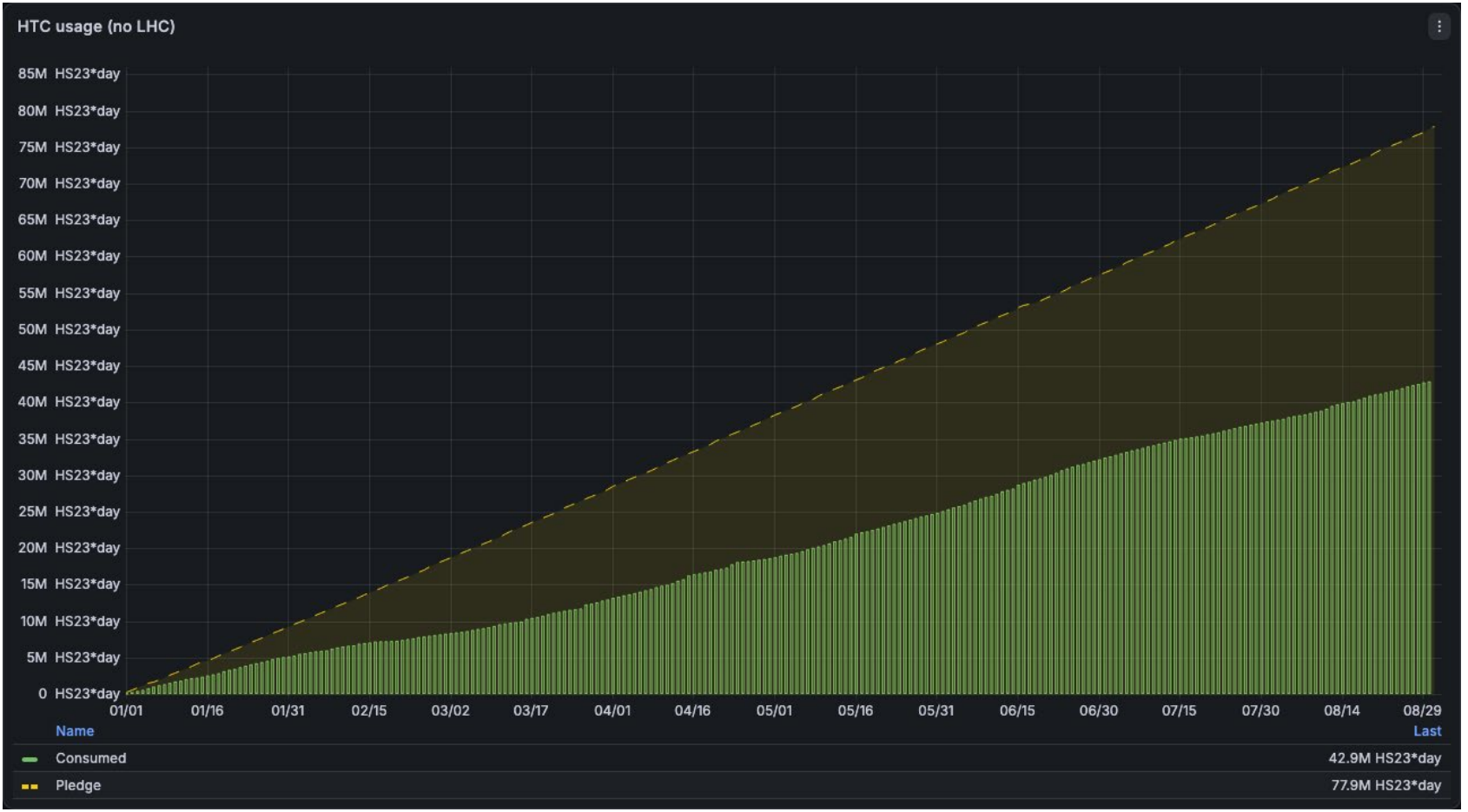








NO-LHC Cumulative CPU Accounting

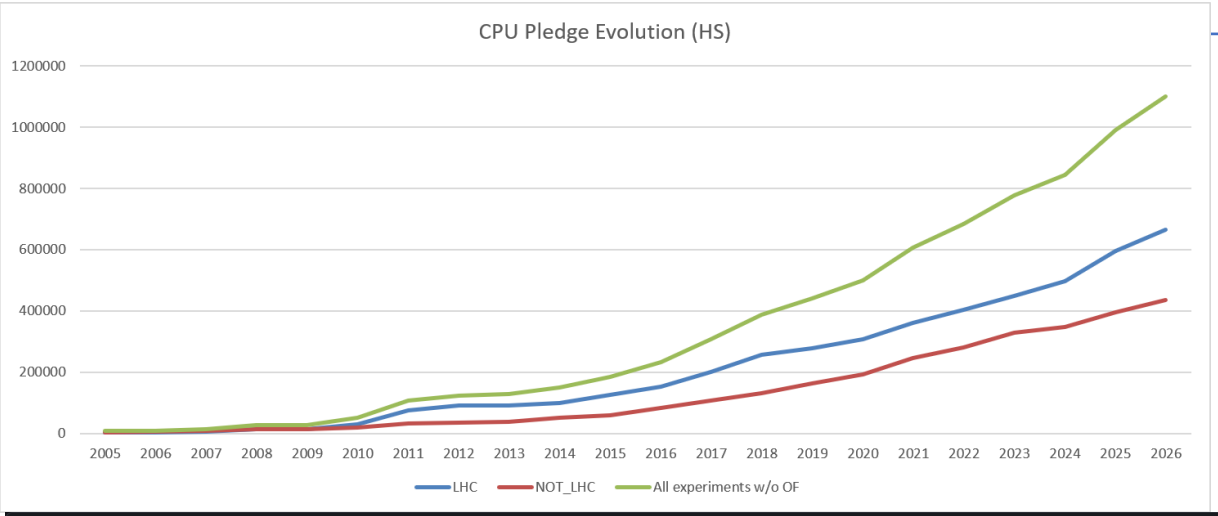


Pledge NO-LHC usato al 55% in media

Rende disponibile:
(98+196+46) * 0.45 = 153kHS

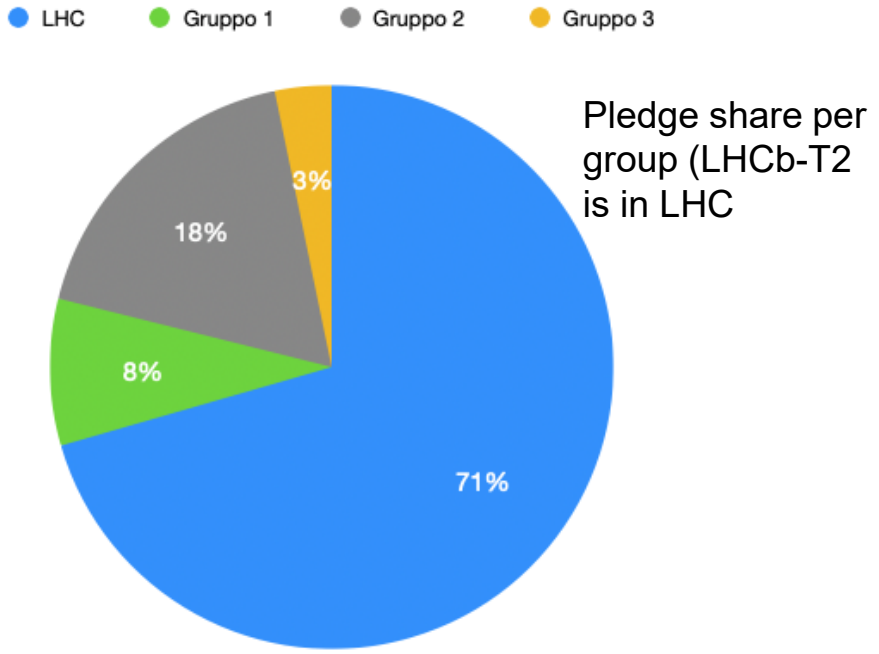
Ridotto di OF ➔ 122kHS

Overlap Factor



	LHC	Gruppo 1	Gruppo 2	Gruppo 3	
Pledge totale	773K HS23	92K HS23	196K HS23	36K HS23	1.096K HS23
OF part	155K HS23	18K HS23	39K HS23	7K HS23	219K HS23

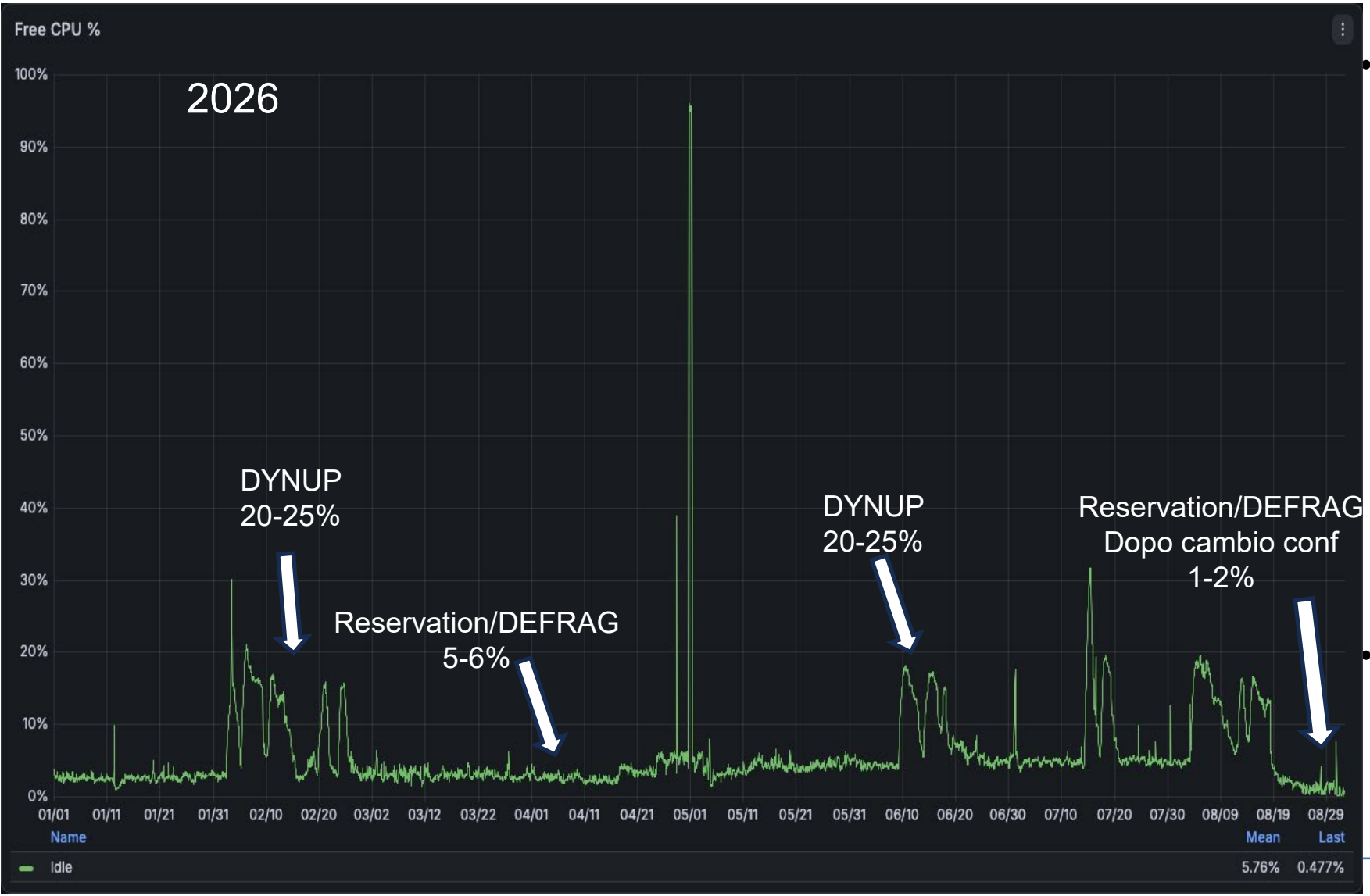
Overlap Factor
20%



Pledge share per group (LHCb-T2 is in LHC

- Proposte
  - Ridurre l'OF?
  - Escludere LHC dal calcolo dell'OF?
  - Saltare un anno di OF?

CPU % IDLE due to Reservation/DEFRAG and DYNUP



- Percentuali di CPU che perdiamo per:
  - DYNUP – aggiornamenti rolling della farm
    - i.e. security patches
  - RESERVATION per multicore job
    - Migliorato dopo interazione con CMS
      - Thanks D.Spiga!!
    - ridotto la quantità di nodi da drainare in simultaneal
    - Ridotto la quantità di core da liberare prima che il nodo accetti nuovi job
- Non contiene CPU perse per
  - Leonardo DOWN
  - WN Maintenance
    - i.e. vecchi nodi rotti, etc

Risorse in EPIC@CNAF

N.B. - Disco in Replica3: spazio Total e Used da dividere per 3

In arrivo (ultima gara HBD):

- 2 switch HUAWEI CE8855-32CQ4BQ

32 x 100Gb/s porte
4 x 200Gb/s porte

- **3 server Supermicro**

1x AMD EPYC 9645 **96-core/192T**
768GB (24 x 32GB) di RAM
2 x 480GB SSD per SO
32 x 15,3 TB NVME E3.S
4 x 100Gb/s

CPU Highlights

- Potenza installata: 1136.3kHS
  - Inclusi 53kHS opportunistici da HPC Bubbles
    - Nodi CPU Only
  - Non è previsto acquisto di CPU nel 2026
- CPU cloud quasi fully booked
  - Circa 1000 coreHT disponibili
  - Possiamo aggiungere nodi della HPC Bubble
- Richiesto per il 2027 il rimpiazzo di 106kHS:
  - 40 nodi gara 2017 per 36kHS
  - 55 nodi gara 2019 per 70kHS
- Nuove CPU: 4 €/HS23 è ormai un miraggio ➔ **15 €/HS23 per preventivi**
  - Sottomessa proposta su bandi SINERGY con ICSC
    - Chiesto circa 1 MEuro per CPU
    - Saranno use case medici quindi per dati sensibili
      - **Necessaria dinamicità per usare risorse non-EPIC in workflow EPIC e viceversa**
    - Precedente proposta (circa 10Meu) su bandi IPCEI non andata a buon fine

- **Uso «opportunistico» dell'HPC Bubble del CNAF in HTCondor per gestire gli over-pledge CPU**
  - **Circa il 60% dei nodi CPU-only**
- Per HPC bubble ➔ no allocazione fissa ma tetto budgettario
  - Significa togliere i limiti di allocazione sul numero dei nodi
    - Account slurm disponibile
- Possibile ridurre/eliminare OF?
  - almeno per gli LHC
- Possibile uso di risorse EPIC per use case NON EPIC?

Stato Risorse Tier1 – CPU (no EPIC)

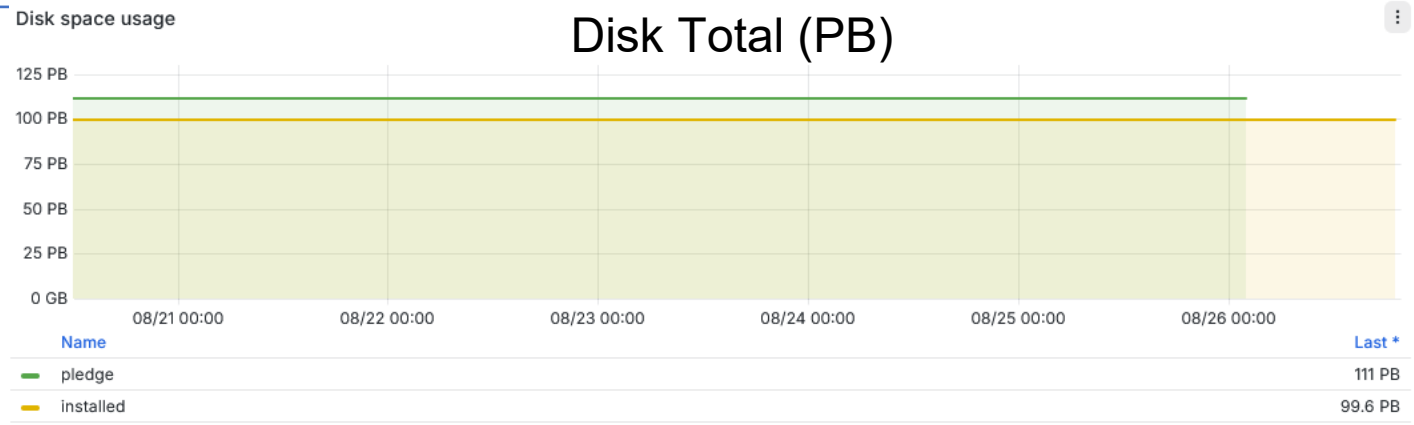
	Pledge26 (kHS)	Extra Pledge 2026 (kHS)	Installato 2026 (kHS)	Finanziato 2026 (kHS)	Acquistato 2026 (kHS)	Disponibile 2026 (kHS)	Previsione delta 2027 (kHS)	Dismissioni (kHS)	Da acquisire (kHS)
CPU Grid	1108.4 (no OF) 886.7(w OF)	15.8	1136.3 Di cui 53kHS oppurtunistici	0	0	1136.3-1108.4-15.8 = 12.1(no OF) 1136.3-886.7-15.8 = 233.8 (w OF)	82 (no OF) 233.6 (w OF)	110	82+110-12.1 = 98.6 (no OF) + 53(opportunistic) ➔ **151.6** 65.6+110-233.8 = **-58.2** (w OF)
CPU Cloud									

DISK

Stato Risorse T1 – DISCO GPFS

Disco (Pledge2026=113.0PBN+10PBN ICSC, In prod=99.6PBN)

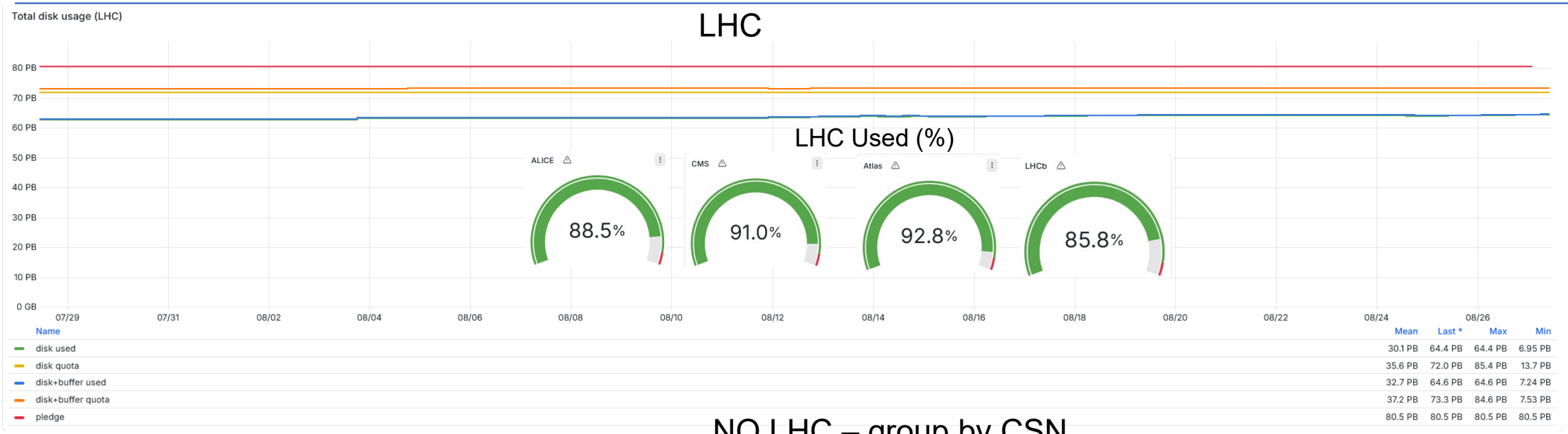
- **AQ 2023-2024 – 80PB**
  - Huawei OD-1600 - 73PB effettivi
    - In produzione
    - manutenzione fino 2031
- **Migrati i sistemi con manutenzione attiva**
  - Lenovo-DDN gara 2022 – 14PB
    - In produzione
    - manutenzione fino 2031
  - Huawei os5k8 – 9PB
    - In produzione
    - Manutenzione non rinnovabile, chiesto il rimpiazzo per il 2027
  - DDN gara 2019
    - 3 su 6PB in produzione
    - Manutenzione scaduta
    - Rinnovo troppo costoso
    - Dimezzato numero sistemi in prod per avere parti spare
    - Finanziato rimpiazzo nel 2026



- **Gara disco 2025**
  - **AQ biennale per risorse 2025-2026-2027**
  - **AQ con capacità 50PB**
  - **GE14555 indizione il 27.02.2026**
  - **GE14723 aggiudicazione il 24.07.2026**
    - **Non c'è stand-still**
    - **Contratto AQ in fase di finalizzazione**
  - **Un solo partecipante ITALWARE – ribasso dello 0.01% rispetto ai 120€ della base di gara**
    - **Sistemi HUAWEI**
  - **Q4 2026 speriamo di avere il primo Ordine in produzione**
    - **7 sistemi per 25PB**
- **Disco veloce per aumentare capacità buffer**
  - **800TB Netti su due sistemi Huawei acquistati a fine 2025**
  - **Installato**

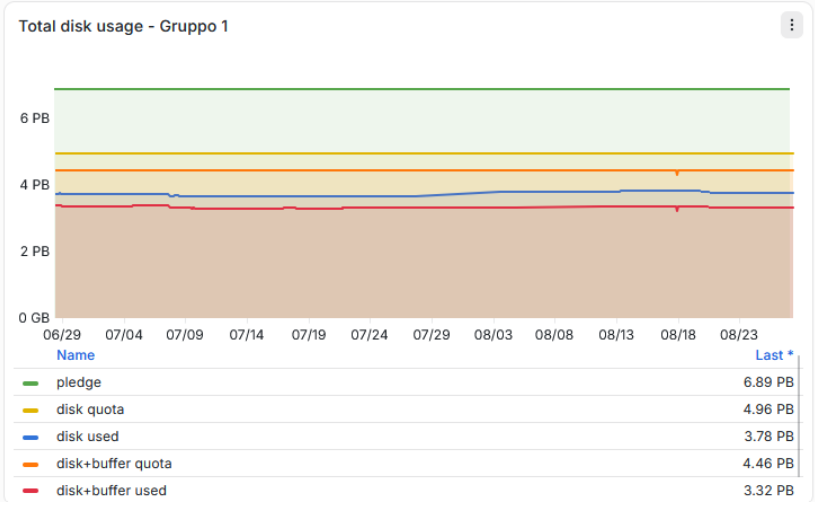
Disk assigned vs usage vs pledge

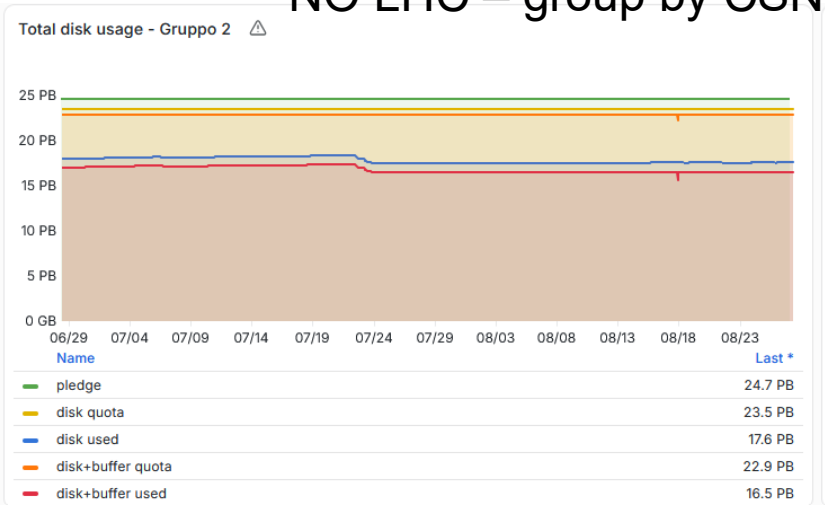
LHC

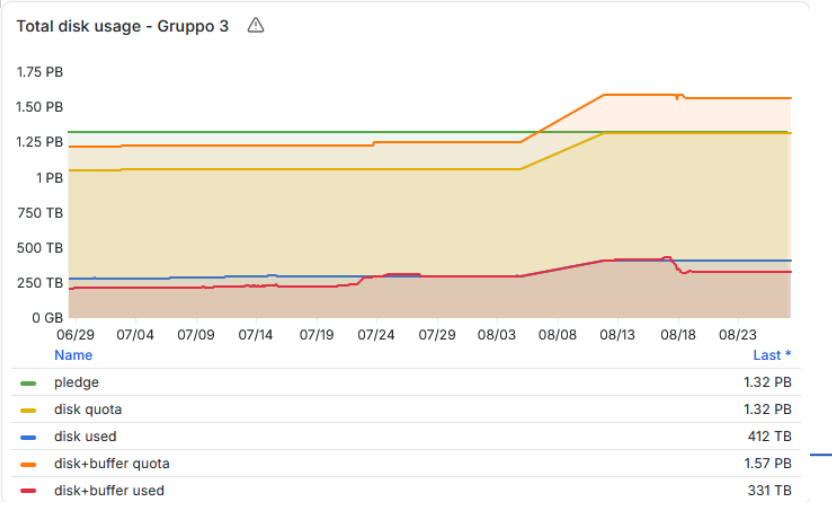


LHC Used (%)

NO LHC – group by CSN







Cluster CEPH@CNAF

Cluster	Nodi	Mon	MDS/RGW	HDD					SSD					
				Disk	Tot (PiB)	Free (PiB)	Used (PiB)	Used %	Disk	Tot (PiB)	Free (PiB)	Used (PiB)	Used (%)	Maint. Exp.
Terabit; SERVIZI:CephFS per i cluster HPC, Tier3, collog, influx, os-server e Kubernetes (Replica 3); - RBD per Proxmox storage (Replica 3)	24	3	4	576	8.2	6.7	1.5	18.3	48	0.487	0.49	0.02	0.04	1/12/2030
Cluster Cloud@CNAF: - SERVIZI: RBD per Openstack	14	3		336	4.8	3.9	0.913	19.0	28	0.1	0.96	0.004	4.4	9 nodi 01/10/27 - 5 nodi 27/03/2029 (Mon su VM)
Cluster S3. SERVIZI: S3 CNAF. (Replica 3) - Swift per Openstack. (Replica 3) - S3 GPFS AFM. (EC 4+2)	12	3	10	290	4.6	3.8	0.799	17.4	22	0.023	0.02	0.002	8.7	sept. 2026
EPIC - SERVIZI: RBD per Openstack	29	3		540	8.4	7.1	1.2	14.29	214	1.5	1.4	0.067	4.5	8 nodi 10/2028 - 16 nodi 01/2030 - 5 nodi 05/2030

- Cluster CEPH di Cloud@CNAF
  - performance in calo per dischi a fine carriera
  - (non) Richiesto rimpiazzo per 9 server (500k€)
  - Richiesto rinnovo manutenzione per tutti i nodi (21k€) + 20k€ per sostituzione dischi rotti

Stato Risorse Tecnopolo – DISK

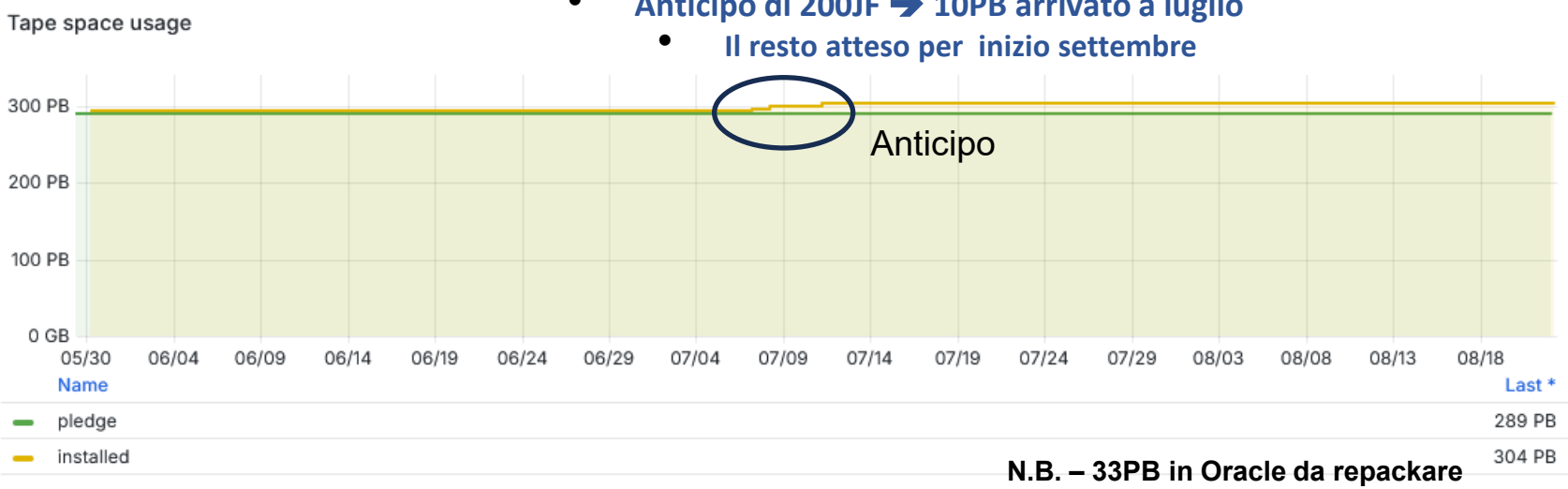
	Pledge 2026 (PB)	Extra Pledge 2026 (PB)	Installato2026 (PB)	Finanziato 2025 (PB)	Finanziato 2026 (PB)	Acquistato 2026 (PB)	Disponibile 2026 (PB)	Previsione delta 2027 (PB)	Rimpiazzi (PB)	Da acquisire 2027 (PB)	NOTE
DISK	113.4	10 (2024 ICSC) 1.79 (LHCb)	73+14+9 +3=99	20	13.07	25	99 -113.4 - 10 + 25 = 0.6	32	3(2025) 9(2026)	113.4 – 99 – 25 + 8.07 + 32 + 9 = **38.47**	2025 acquistato 0 Capacità residua AQ ➔ 25PB

TAPE

Stato Risorse Tecnopolo - TAPE

- **TAPE(Pledge2026=287PB +10PB ICSC)**
  - **Vecchia libreria IBM TS4500 migrata al tecnopolo**
    - In produzione – 110PB
  - **Nuova libreria IBM TS4500 in produzione al tecnopolo**
    - In produzione – 104PB
  - **Vecchia libreria Oracle SL8500 portata al tecnopolo**
    - 80PB
    - In dismissione ma online
    - Repack bloccato (mancano 33PB)
    - 47PB liberi
      - Ma meglio non usarli

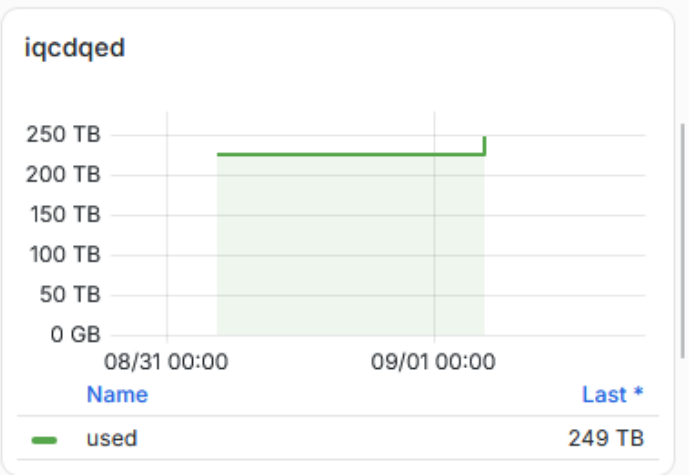
- **Gara 2025**
  - **AQ biennale per risorse 2025-2026-2027**
    - **AQ con capacità 160PB**
      - **1000 cassette JE (12.35 €/TB)**
      - **2800 cassette JF (6.81 €/TB)**
      - **100 Cleaning tape**
  - **GE14448 indizione il 14.11.2025**
  - **DP14669 per aggiudicazione effettuata il 01.06.2026**
    - **Ratifica GE14670 del 11.06.2026**
    - **7.55 €/TB i.i.**
  - **Contratto sottoscritto il 22.07-2026**
  - **Ordine di 80PB inviato il 29.07.2026**
  - **Anticipo di 200JF ➔ 10PB arrivato a luglio**
    - **Il resto atteso per inizio settembre**

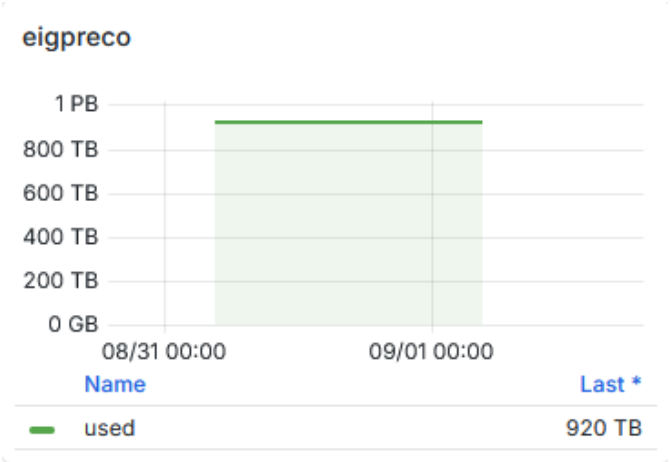


N.B. – 33PB in Oracle da repackare

Stato Risorse Tier1 – TAPE-2

	Pledge 2026 (PB)	Extra Pledge 2026 (PB)	Installato 2026 (PB)	Finanziato 2025 (PB)	Finanziato 2026 (PB)	Acquistato 2026 (PB)	Disponibile 2026	Previsione delta 2027	Repack	Da acquisire 2027	NOTE
Tape	287.06	36.03 10 (ICSC) 21.03 (LHCb) 5 (CMS)	214 110+104 +80(da dismettere)	61.25	20.82 (non considerava 20PB di overpledge)	80	-29.09 (110+104+80 – 287.06 – 36.03 = -29.09)	32.24	33 (incluso in pledge)	61.33 (32.24 + 29.09)	- 2025 acquistato 0 - Capacità residua AQ ➔ 80PB





Gruppo IV via ICSC – da far rientrare nelle pledge standard?

Verso HI-LUMI

Rate per HI-LUMI

- DC27 e tape per Run4 a DOMA
  - Slide DOMA CMS (Katy)
  - DC27 sarà all'inizio del 2027
  - per il 50% dei rate attesi per Run4
    - DC24 era per il 25%
      - il 25% è stato toccato poco e male
  - Inoltre, coinvolgerà tape
- I Tier1 devono già adesso cominciare ad attrezzarsi, non solo con la capacità storage
  - "potentially difficult to achieve depending on how sites decide to provision in preparation for Run-4, still 3 years in the future" (slide 21)
- CMS sta facendo l'esercizio di simulare le necessità di tape in vari scenari di Run4 (fattore 2 rispetto a DC27).

Tape Staging/Archiving Overview

Run-4								
	T0		Re-Reco		Re-Mini		MC (incl. PU)	
Data Flow	archiving	staging	archiving	staging	archiving	staging	archiving	staging
CERN	70.6	--	8.3	20.7	0.0	0.0	0.6	--
KIT	4.0	--	0.8	2.1	0.7	4.4	1.3	--
PIC	2.0	--	0.4	1.0	0.3	2.3	0.8	--
IN2P3	4.0	--	0.8	2.1	0.7	4.4	1.3	--
CNAF	4.0	--	0.8	2.1	0.7	4.4	1.3	--
RAL	2.8	--	0.6	1.4	0.5	3.1	1.3	--
Serbia	2.0	--	0.4	1.0	0.3	2.3	0.8	--
NCBJ	2.0	--	0.4	1.0	0.3	2.3	0.8	--
JINR	3.2	--	0.7	1.7	0.6	3.7	1.3	--
FNAL	16.0	--	3.3	8.2	2.7	17.9	1.4	--
Summed T1s	39.9	--	8.24	20.6	6.7	44.8	10.3	--

Throughput in units of GB/s

Certainly some flows are not included, e.g. staging for user analysis jobs

CMS si aspetta di scrivere dati dal Tier0 su tape a 4 GB/s, e NEL MENTRE (non necessariamente tutti gli scenari sovrapposti, ma alcuni anche sì) di migrare e richiamare ai rate che vedete in tabella, incluso lo scenario terrificante Re-Mini in cui richiamano a 4.4 GB/s 23 PB in 2 mesi.

Questi numeri riguardano solo CMS!!!

1 tape drive ➔ 400MB/s (picco)

Necessità di buffer NVMe adeguati

Necessità di server di data management ulteriori

Rate per HI-LUMI

- DC27 e tape per Run4 a DOMA
  - Slide DOMA CMS
  - DC27 sarà all'iniz
  - per il 50% dei rat
    - DC24 era per i
      - il 25% è st
  - Inoltre, coinvolge
- I Tier1 devono già attrezzarsi, non sc storage
  - "potentially difficu sites decide to pro still 3 years in the f
- CMS sta facendo l necessità di tape in vari scenari di Run4 (fattore 2 rispetto a DC27).

Da ultimo C3SN – il DC27 non dovrà impattare economicamente sui siti, si cercherà di fare il possibile con quello che c'è – in attesa di avere numeri definitivi circa i rate di HI-LUMI

Tape Staging/Archiving Overview

Run-4	T0		Re-Reco		Re-Mini		MC (incl. PU)	
Data Flow	archiving	staging	archiving	staging	archiving	staging	archiving	staging
				20.7	0.0	0.0	0.6	--
				2.1	0.7	4.4	1.3	--
				1.0	0.3	2.3	0.8	--
				2.1	0.7	4.4	1.3	--
				2.1	0.7	4.4	1.3	--
				1.4	0.5	3.1	1.3	--
				1.0	0.3	2.3	0.8	--
				1.0	0.3	2.3	0.8	--
				1.7	0.6	3.7	1.3	--
				8.2	2.7	17.9	1.4	--
				20.6	6.7	44.8	10.3	--

user analysis jobs

dal Tier0 su tape a 4 GB/s, e NEL
e tutti gli scenari sovrapposti, ma
hiamare ai rate che vedete in
ficante Re-Mini in cui richiamano a

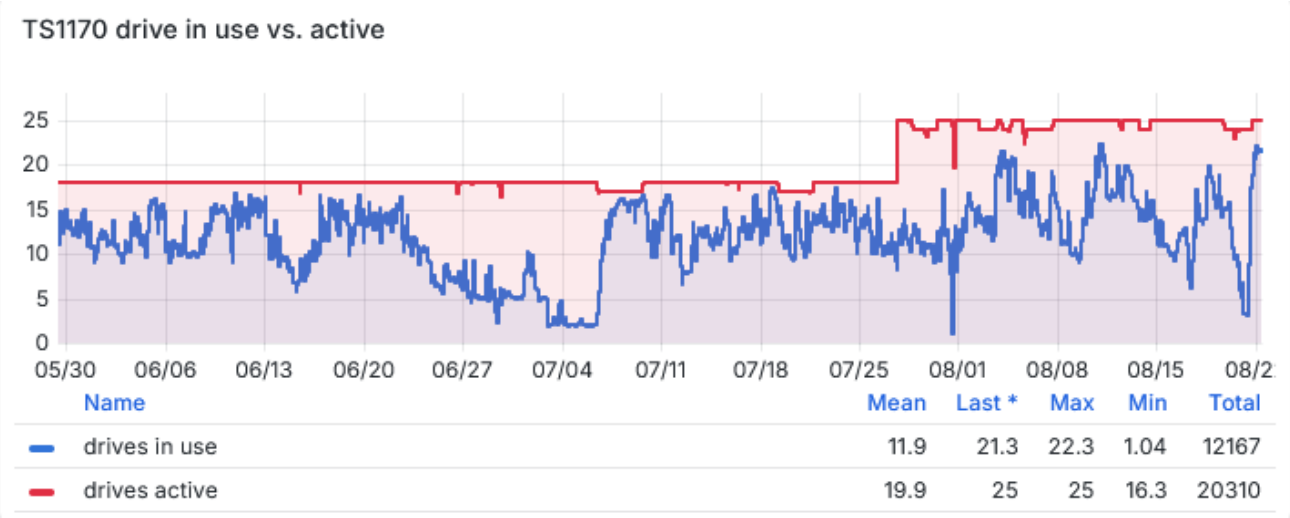
o CMS!!!

1 tape drive ➔ 400MB/s (picco)

Necessità di buffer NVMe adeguati

Necessità di server di data management ulteriori

Necessità di tape drive e cablaggio



- 7 drive JF acquistati
  - 15k/drive i.i.
  - Prezzo scadeva il 30/06
    - Aumenti del 40%
  - Finanziati da datacloud/GE nel 2026
- Altri 4 drive finanziati da PON CREST
  - TD e/o gara da fare
- Se confermati i rate della slide precedente siamo ancora lontani
- NON C'E' solo LHC
  - Es. Belle, JUNO potrebbero richiedere più performance e sistemi dedicati

Necessità risorse IT-non-pledge

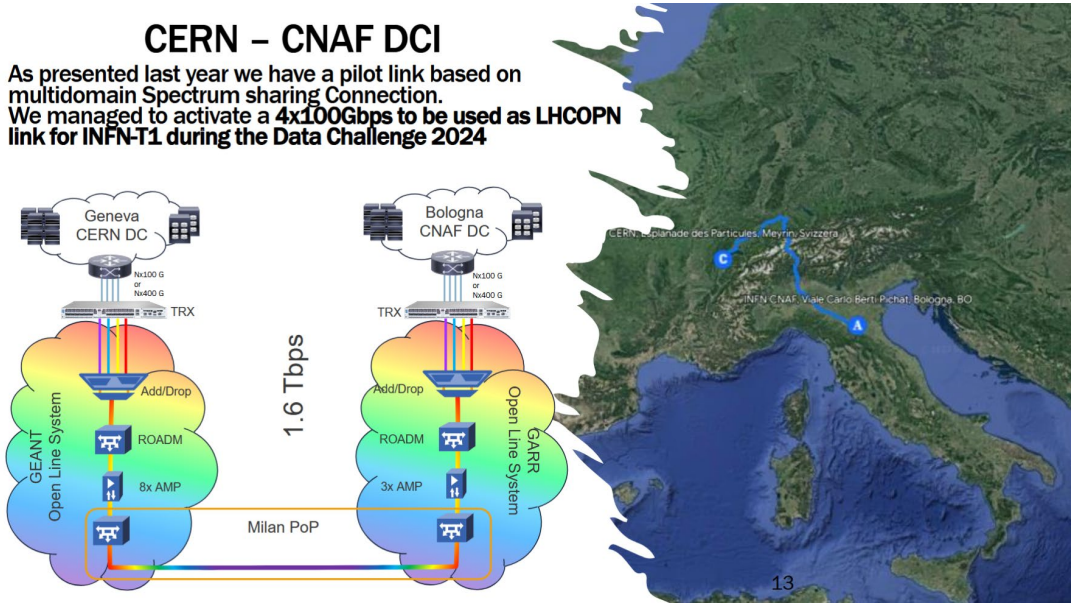
Reparto	Inventariabile (k€)	Consumo (k€)	Manutenzioni (k€)	Licenze (k€)
Rete	80	45	53	5
Farming	80	0	31	0
Storage	865 ➔ 365	30	161	280
Infra	Gestite via TICC			
TOTALE	**525**	**75**	**245**	**285**
	1130			

- Richiesta a Datacloud
  - Manutenzioni e SW sono di fatto incomprimibili e in parte non note
    - Nuova negoziazione per gpfs e tsm con ibm in corso

BACKUP

Network scale UP per HL-LHC

- Connettività Geografica
  - Link ottico CNAF-CERN siamo pronti a supportare fino ad 1.6 Tbps su OPN con tecnologia attuale
  - Nel corso del 2026 Passeremo a 800Gbps
    - **Data intervento 15/09**



- Connettività Locale (I/O interno)
  - Gli attuali CORE permettono una espansione molto importante
  - abbiamo 6 slot utilizzabili per espansione per ogni CORE per un potenziale di 216 porte a 400 Gbps (o 864 porte a 100Gb) per apparato: circa 86,4 Tbps per apparato
  - può crescere orizzontalmente aggiungendo uno spine level auspicabilmente con tecnologia 1,6T acquistando nuovi concentratori
  - Esauriti gli switch di management
    - Richiesti a datacloud
- Nel 2025 JUPLUS ha finanziato l'upgrade del cablaggio strutturale al TP
  - Dovrebbe permetterci di gestire RUN4

HPC Bubbles

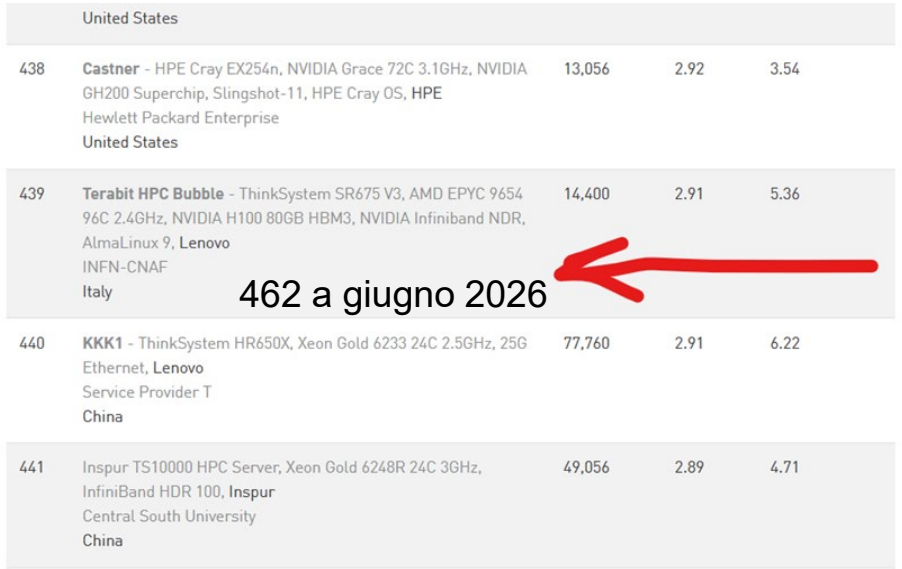
CPU Node Lenovo Lenovo ThinkSystem SR665 V3	192 physical cores - Dual AMD AMD EPYC 9654 96C 360W 2.4GHz 1.5TB RAM DDR5 IB NDR 400G - NVIDIA ConnectX-7 NDR OSFP400 1-Port PCIe Gen5 x16 InfiniBand Adapter 20TBL (SSD) + OS disks
GPU node Lenovo ThinkSystem SR675 V3	Like CPU nodes + 4x NVIDIA H100 SXM5 con 80GB HBM3
FPGA Node Lenovo ThinkSystem SR675 V3	32 physical cores - AMD EPYC 9124 16C 200W 3.0GHz Processor RAM 768GB DDR5 IB NDR 440G 4 x XILINX U55C o 4 x TerasicP0701
Storage Nodes (CEPH Bricks) DELL PowerEdge R760xd2	64 physical core – Dual Intel Xeon Gold (Sapphire Rapids), model 6428N, 1.8GHz, 32Core/64Thread, 60M Cache, DDR5-4800, 185W TDP 1TB RAM DDR5 IB Mellanox 400G 384 TBL HDD + 25.6 TBL NVMe
Accessories	Switch IB NVIDIA/Mellanox NDR 400Gbit/s Switch ETH – NVIDIA Modello SN3420 - 12x QSFP28 100GbE + 48x SFP28 25GbE IB and ETH Cables Mainteinance 3+2 years

Acquired resources on three different NRRP projects

	CPU Nodes	GPU Nodes	Physical CPU cores	H100 GPU Number	FPGA Xilinx Nodes	Nodo FPGA Terasic Nodes	Storage Nodes/PB _Raw
BA	24	6	5760	24	0	0	32/12.3
CNAF	26	30	10752	120	2	2	52/19.9
MIB	0	0	0	0	2	2	0
NA	18	1	3648	4	2	0	8/3.1
PD	6	6	2304	24	0	0	0
PI	20	0	3840	0	0	0	0
RM1	12	0	2304	0	0	0	0
TO	14	6	2688	24	0	0	0
LNGS	0	6	1152	24	0	0	12/4.6
CT	12	0	2304	0	0	0	8/3.1
LNF	12	0	2304	0	0	0	0
LNFESA	8	6	2688	24	0	0	6/2.3
LNL	4	0	768	0	0	0	0
MI	4	0	768	0	0	0	0
TOTAL	**160**	**61**	**41341**	**244**	**6**	**4**	**118/45.3**

Top500.org November 2025

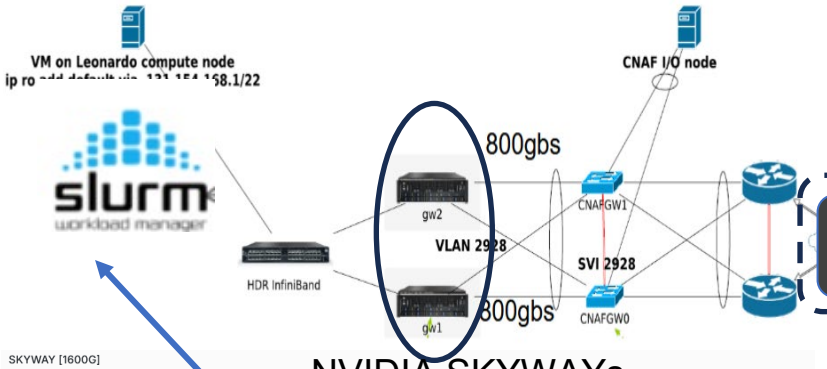
- Terabit Bubble@CNAF ==> 437
  - 16 CPU Nodes + 21 GPU nodes
    - 7104 CPU cores + 84 NVIDIA H100 GPU

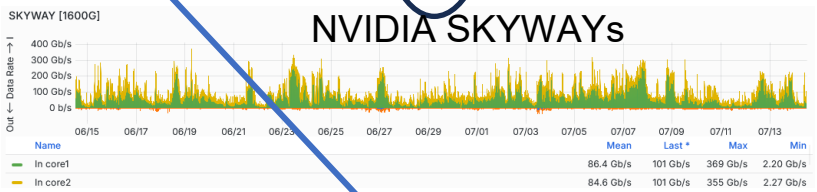


One bubble@CNAF has been installed into the ISO-certified (27001) zone

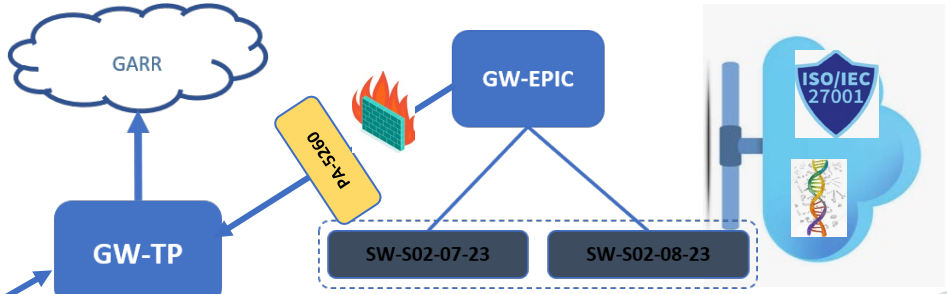
- 3648 core; 36 NVIDIA H100 GPUs

CPU Resources



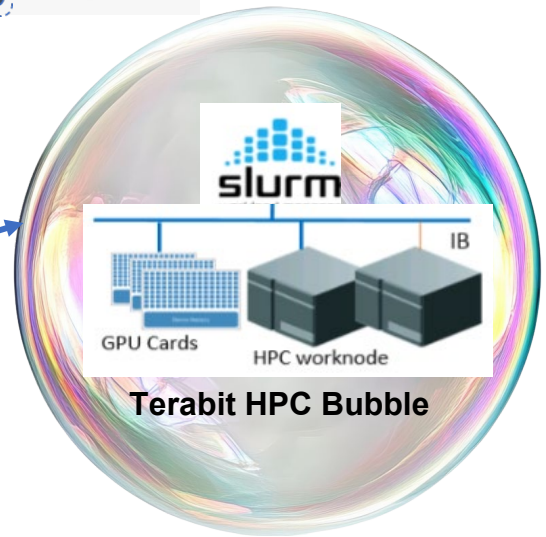


NVIDIA SKYWAYs



CORE 1

CORE 2

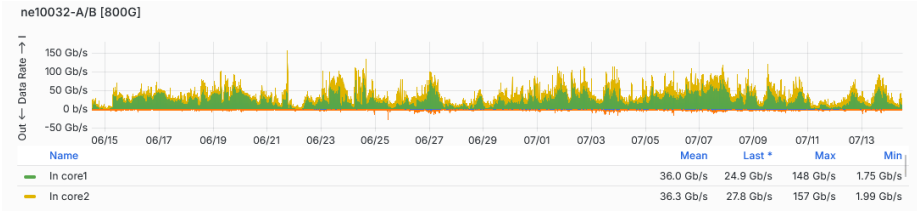


Terabit HPC Bubble

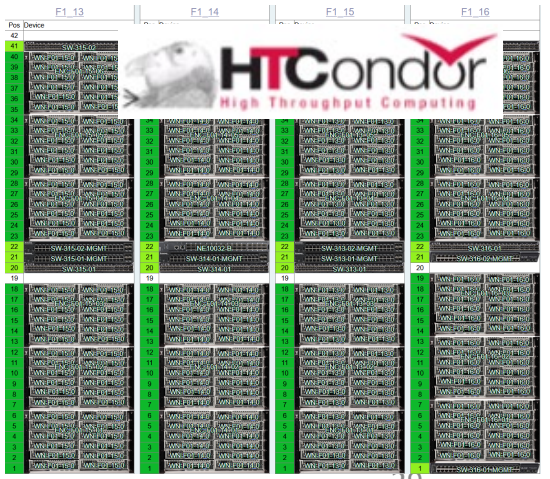
SLURM jobs instantiate whole-node VMs, added as HTCondor WNs

2xAggregatori di Marconi, fuori manutenzione da N anni
Richiesto rimpiazzo in preventivi (per la sola parte HPC Bubble)

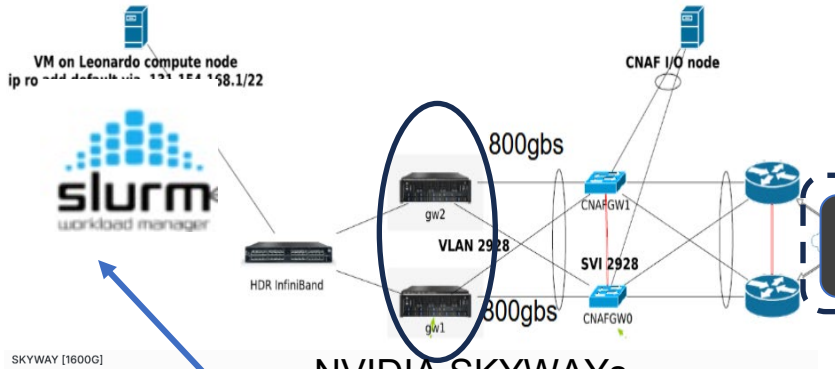
HTC Cluster



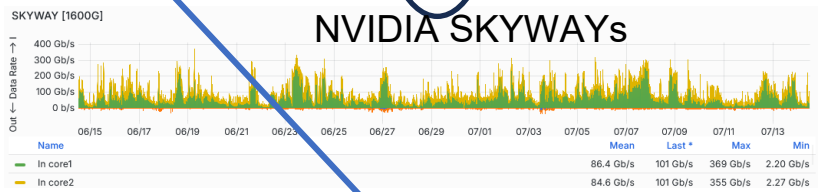
Ex-CINECA Worker Nodes

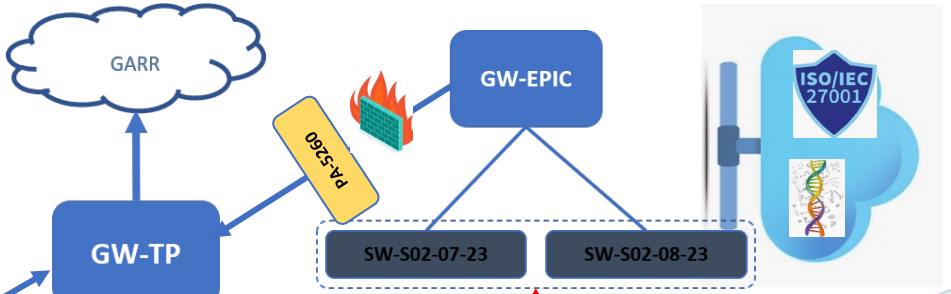


CPU Resources



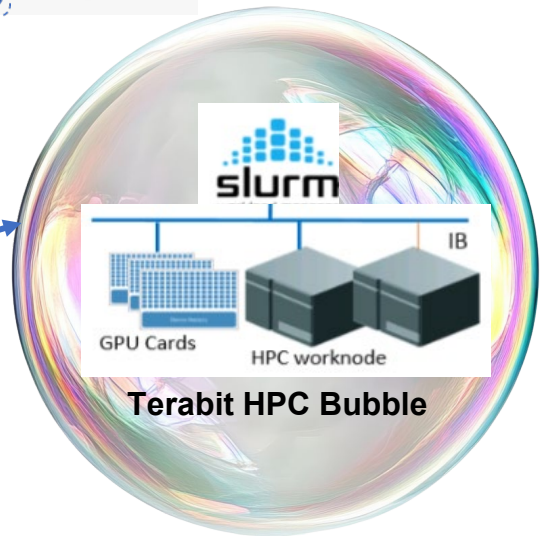
NVIDIA SKYWAYs





CORE 1 CORE 2

In fase di studio Nx100G
Con propagazione VLAN T1
Per permettere l'uso dinamico di risorse dentro EPIC da parte di applicazioni non EPIC

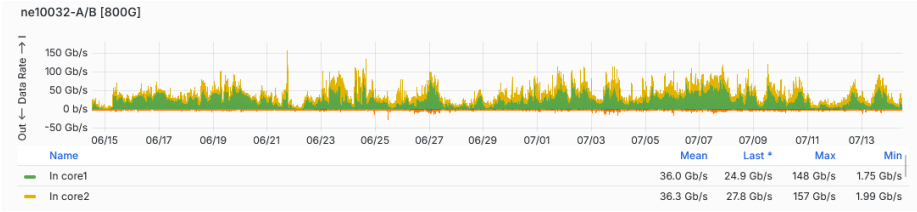


Terabit HPC Bubble

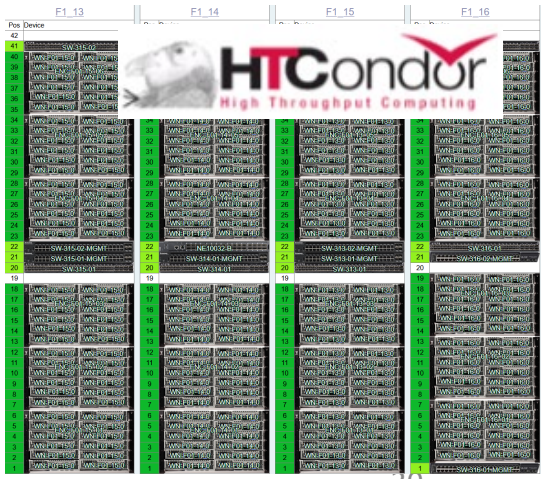
SLURM jobs instantiate whole-node VMs, added as HTCondor WNs

2xAggregatori di Marconi, fuori manutenzione da N anni
Richiesto rimpiazzo in preventivi (per la sola parte HPC Bubble)

HTC Cluster



Ex-CINECA Worker Nodes



To get some space outside the Tier1.....

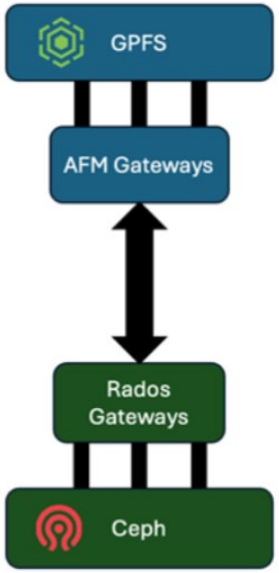
Storage loans at Naples

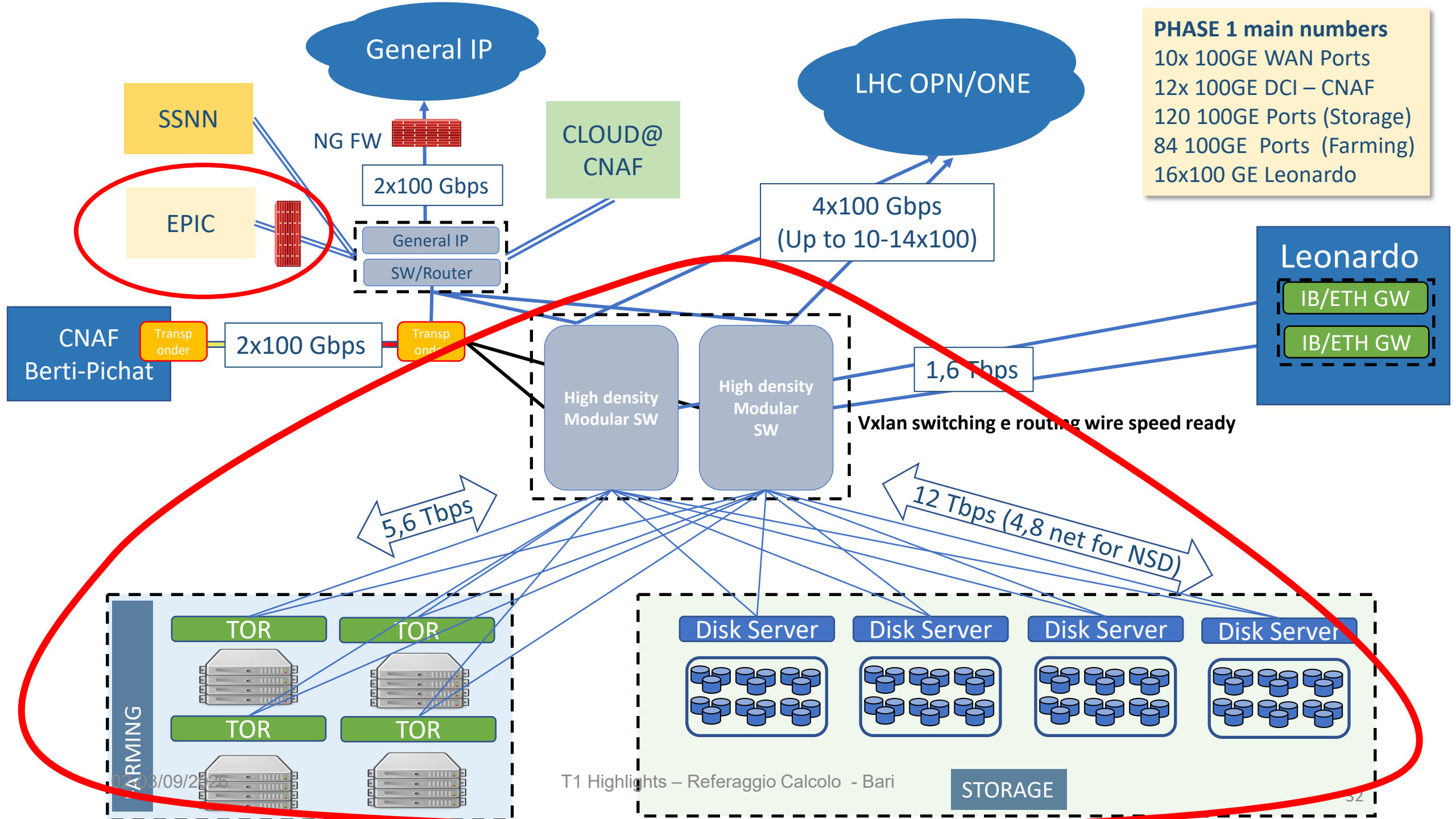
- 1 OceanDisk 1500 and 4 servers (from INFN-NA Tier2 site)
  - 1.7 PB usable space
  - 20 Gbps connectivity
  - Used as backup (1.4 PB) for
    - Borexino 305.6 TB
    - Cosmo_wnext 484.5 TB
    - Cuore 587.5 TB
    - Nucleus 4.2 TB
  - RSYNC completed (can we remove those 4 from gpfs_data?)
- "T2" storage area for JUNO
  - Accessible via WebDAV as SA and POSIX via AFM (/storage/gpfs_afm/juno/)
  - Local cache 42 TB (NVMe)
  - 300 MB used so far…

Grazie ai colleghi di INFN-NA per la disponibilità ed il supporto!

Use CNAF CEPH S3 as backend for GPFS

- Local "cache" file system (gpfs_s3, 21 TB, NVMe)
  - Borexino
  - FLASH
- Linked to the local S3 (Ceph, 1.5PB?) cluster
  - 2 GW nodes on GPFS side (2x100Gbps)
  - 4 GW nodes on S3 side (2x25Gbps)




General IP
LHC OPN/ONE
SSNN
NG FW
2x100 Gbps
CLOUD@ CNAF
EPIC
General IP
SW/Router
4x100 Gbps
(Up to 10-14x100)
PHASE 1 main numbers
10x 100GE WAN Ports
12x 100GE DCI – CNAF
120 100GE Ports (Storage)
84 100GE Ports (Farming)
16x100 GE Leonardo
Leonardo
IB/ETH GW
IB/ETH GW
CNAF
Berti-Pichat
Transp onder
2x100 Gbps
Transp onder
High density Modular SW
High density Modular SW
1,6 Tbps
Vxlan switching e routing wire speed ready
5,6 Tbps
12 Tbps (4,8 net for NSD)
FARMING
TOR
TOR
TOR
TOR
Disk Server
Disk Server
Disk Server
Disk Server
STORAGE