Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
INFN

Stato ed evoluzione dei datacenter in Datacloud

G. Donvito, C. Grandi, D. Cesini

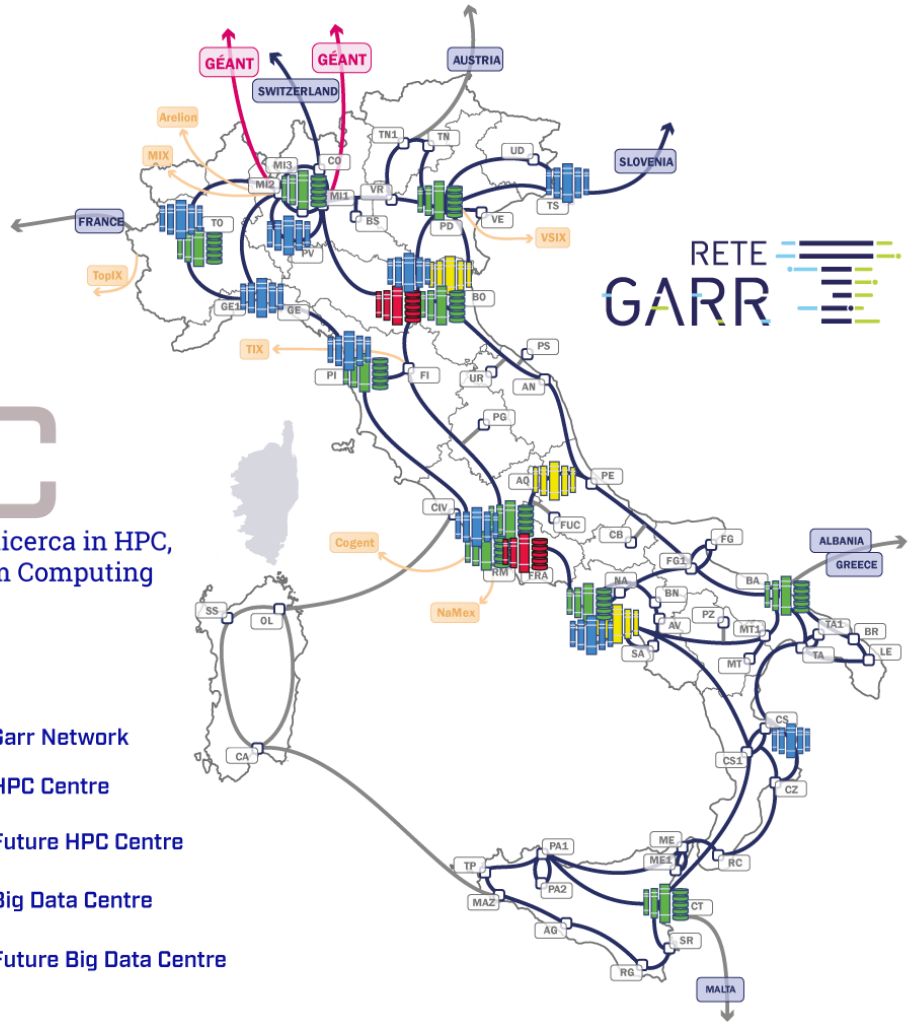
Claudio Grandi

Infrastruttura Datacloud

- 1 Tier1 (CNAF)
- 9 Tier2
- LNGS
- LNF-ESA
- MIB

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

WLCG
Worldwide LHC Computing Grid



I progetti approvati in ambito calcolo

Centri Nazionali

- **ICSC (**HPC, Big Data e Quantum computing): 320 M€, 56.5 M€ INFN, iniziato 1/9/2022, 36 mesi

Infrastrutture di Ricerca

- **Terabit**: 41 M€, 13.6 M€ INFN (oltre a 17.7 M€ per conto di GARR), iniziato 1/1/2023, 30 mesi
- **Itineris**: 155.2 M€, 5M€, iniziato 1/11/2022, 30 mesi – 400k hw

Piano complementare al PNRR MUR-salute

- **DARE**: 124 M€, 3.5 M€ INFN, iniziato 1/12/2022, 48 mesi – circa 2Meuro hw

Ecosistemi dell'Innovazione

- **Ecosister** (Emilia Romagna): 110 M€, 480 k€ INFN, iniziato 1/10/2022, 36 mesi
- **THE** (Toscana): 110 M€ , 540 k€ INFN, iniziato 1/11/2022, 36 mesi

Partenariati Estesi

- **FAIR** (Intelligenza artificiale): 114.5 M€, 1.6 M€ INFN, iniziato 1/1/2023, 36 mesi
- **NQSTI** (Tecnologie quantistiche): 117 M€ , 6.4 M€ INFN, iniziato 1/12/2022, 36 mesi

Tutti i progetti PNRR in https://home.infn.it/it/188-pnrr

Potenziamento dell'infrastruttura di calcolo

Investimenti per il rinnovo delle infrastrutture dei Tier-2 sul budget di ICSC

- Circa 17 M€ in ICSC, inclusi i nuovi centri a LNGS (HPC4DR) e LNF (Space Economy)

Investimenti per il rinnovo della rete dei centri

- Al momento previsti circa 2 M€ in ICSC

Investimenti per il potenziamento delle risorse di calcolo

- Circa 19 M€ in ICSC per hardware di tipo tradizionale HTC (alcuni sistemi HPC per LNGS, LNF)
- Circa 10 M€ in TeRABIT, dedicati prevalentemente alla implementazione delle *HPC bubbles*
- Circa 0.4 M€ in Itineris, di tipologia simile a TeRABIT
- Circa 1.9 M€ in DARE

Ci sono ritardi, aggravati dall'entrata in vigore del nuovo Codice degli Appalti l'1/7/2023…

…ma ne stiamo uscendo

Stato acquisizioni IT

Un centinaio di procedure in totale. Per il momento

Server di calcolo in convenzione CONSIP → installati

Concentratori di rete → in attesa contratto

Storage tradizionale → in fase di installazione tramite avvio anticipato, contratti in via di finalizzazione

Bubble TeRABIT (incluso storage CEPH) → Contratti stipulati, Appalti Specifici in via di esecuzione;

- Risorse ICSC a LNGS e LNF e DARE → In AS di AQ TeRABIT
- Seconda gara server calcolo ICSC → In AS di AQ TeRABIT

Server per sistemi di virtualizzazione → storage in seconda gara nazionale

Router → RDO in corso

Seconda gara storage (tradizionale e CEPH) → documentazione di gara da finalizzare

Risorse della prima tornata di acquisti

Tier-2, Tier-1 - esclusi sistemi HPC

Potenza CPU:

~17 HS06/core

→ ~600 kHS06

Storage disco:

TBN = ~0.73*TBL

→ ~30 PBN

→ Consegnati in fase di installazione

Sito	CPU (Core fisici)	Storage (PBL)
BA	7296	8.1
CT	2688	2.7
LNF	3072	4.8
LNFESA	-	
LNGS	-	
LNL	3072	4.7
PD	-	
MI	3072	2.4
NA	7296	7.2
RM1	3072	4.1
PI	3456	3.2
TO	3072	2.7
CNAF	0 (uso di Leonardo)	80 (PBN)
TOT	36096	40 + 80PBN

HPC Bubbles

 Nodo CPU	192 core fisici 1.5TB RAM DDR5 IB NDR 400G 20TBL (SSD) + dischi di sistema
 Nodo GPU	Come CPU + 4x NVIDIA H100 SXM5 con minimo 80GB e memoria HBM2e
 Nodo FPGA	32core RAM 768GB DDR5 IB NDR 440G 4 x XILINX U55C o 4 x TerasicP0701
 Nodo Storage (CEPH Bricks)	64 core fisici 1TB RAM DDR5 384 TBL HDD + 25.6 TBL NVMe
 Accessori	Switch IB, Switch ETH Cavi IB, Cavi ETH Transceiver vari Assistenza 3+2

Risorse HPC bubbles

Accordo Quadro Nazionale

- Listino prezzi per nodi + accessori
- 2 anni di validità
- Lotto1
  - CPU, GPU, FPGA
- Lotto2
  - Storage (1 nodo 380 TB raw; ≥ 250 TBN)
- Contratti stipulati per entrambi i lotti

Visto il buon risultato si userà il 6/5 e si compreranno qui parte delle risorse previste per la seconda tornata ICSC (incluse nella tabella)

Sito	Nodi CPU	Nodi GPU	Nodi FPGA	Nodi Storage
CNAF	26	30	4	52
BA	24	6	0	32
MI-BI	0	0	4	0
PI	8	0	0	0
TO	6	6	0	0
LNGS	0	6	0	12
NA	18	1	2	8
RM1	12	0	0	0
PD/LNL	10	6	0	0
LNF	20	6	0	6
CT	12	0	0	8
MI	4	0	0	0
TOTALE	**160**	**61**	**10**	**118**

Risorse della seconda tornata di acquisti

Tier-2, esclusi sistemi HPC

(i.e. solo quota ICSC)

Potenza CPU:

Assumendo ~17 HS06/core

→ ~287 kHS06

Storage disco (gara dedicata da effettuare):

Tradizionale: TBN = ~0.73*TBL

CEPH: TBN = ~0,67*TBL

→ ~50 PBN

16PBL da gara HPC bubble

Incluso potenziamento INFNCLOUD backbone a CNAF e BARI

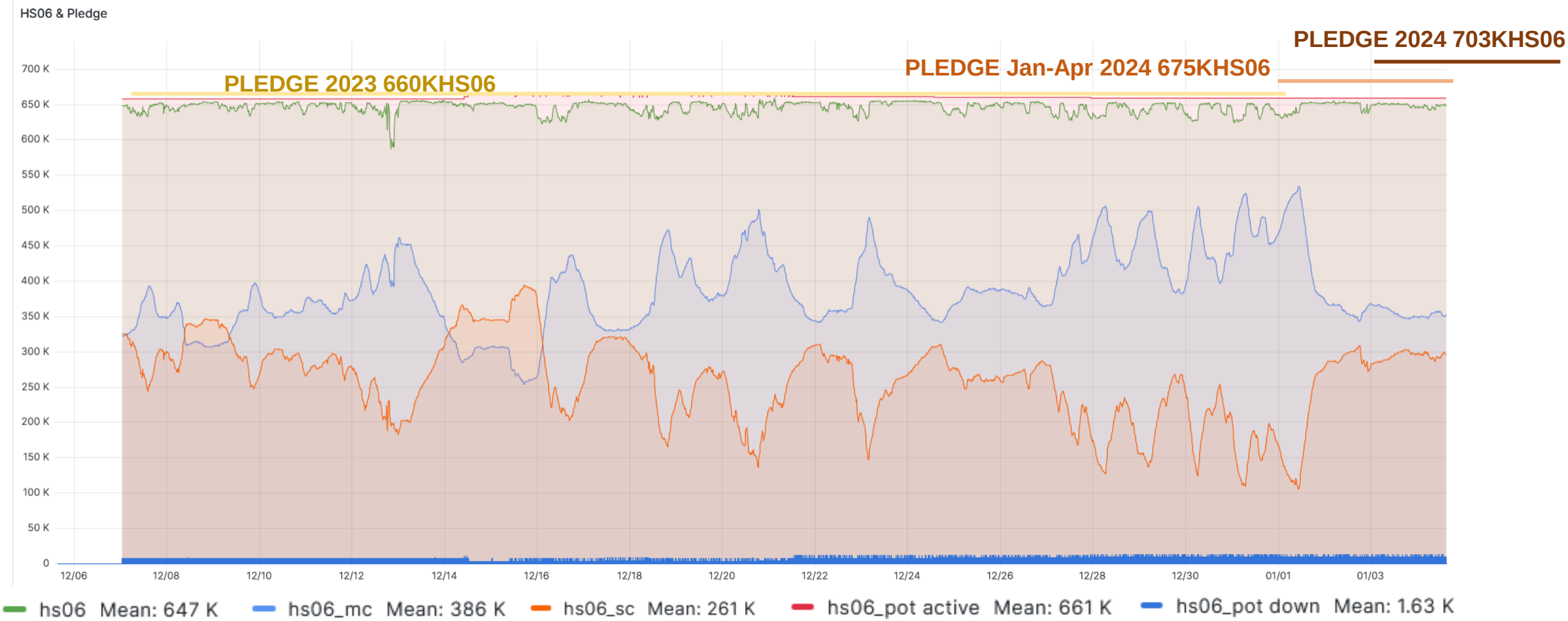
Sito	CPU (Core fisici)	Storage (PBL)
BA	2304	12.2
CT	2304	16.7
LNF	2304	2.5
LNFESA	1536	2.3
LNGS	-	4.6
LNL	784	5.8
PD	-	
MI	784	3.9
NA	2304	12.2
RM1	784	4.5
PI	2304	3.2
TO	1536	4.5
CNAF	-	-
TOT	16896	72.4

Aggiornamento risorse al Tier1

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
INFN

Inaugurazione nuovo datacenter
10 Maggio 2024

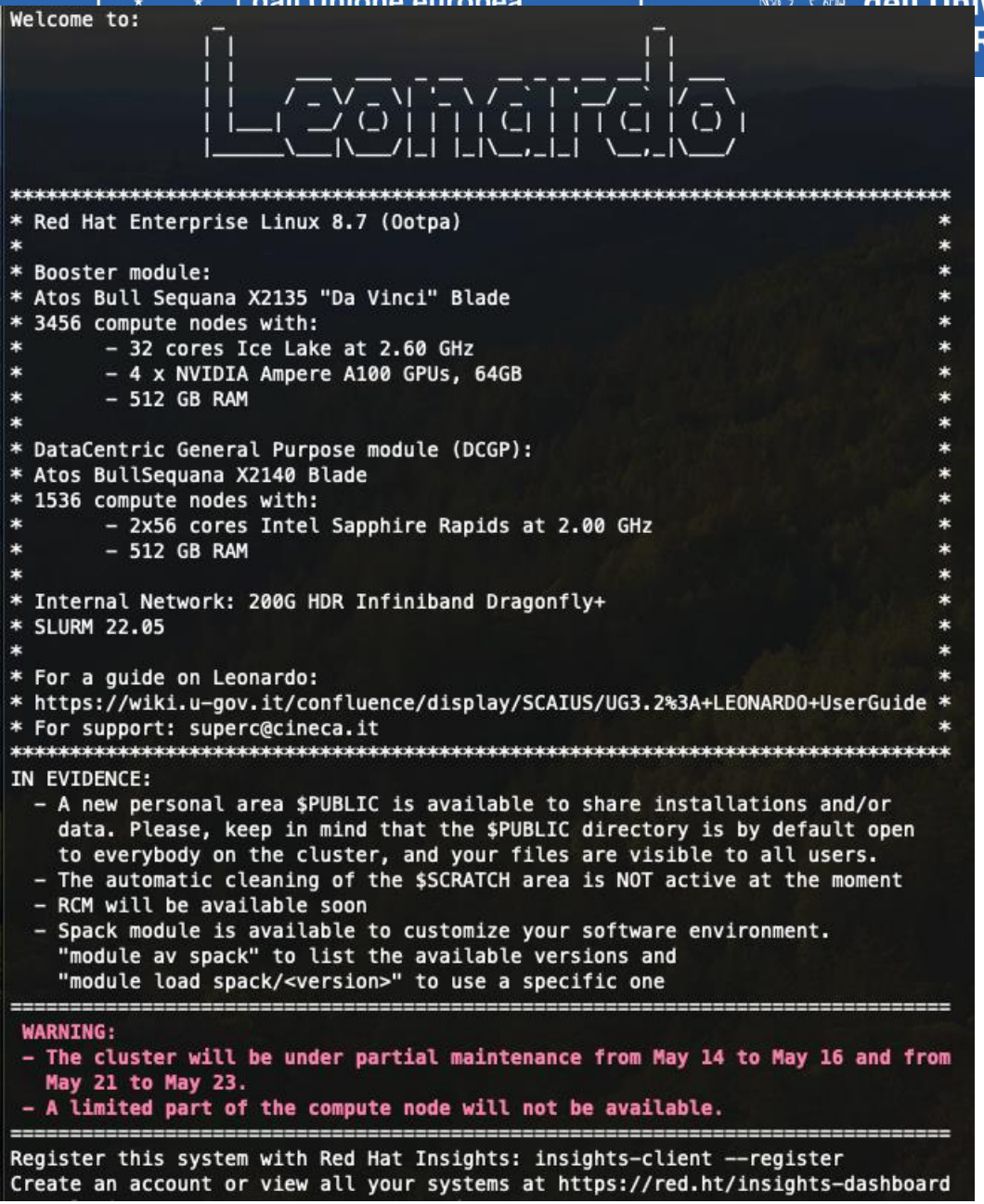
CPU Usage @T1 ALL VOs – no cloud



CPU in 2024 – Leonardo integration

- No direct CPU acquisition in 2024
- We will use up to 300 Leonardo-GP@CINECA nodes
  - Dual 56 cores sockets Intel Sapphire Rapids
  - 112 cores/node
  - (16 x 32) GB DDR5 4800 MHz
  - 1680 Gflops/node (peak)
  - HS06
    - 2800 HS06/node VM-WholeNode
  - NVIDIA Mellanox HDR DragonFly++ 200Gb/s
  - No Ethernet
- Integration Plan
  - “inifnite” SLURM jobs to launch VM containing “our” Condor WN
- PCI pass-throught to see the IB cards on Leonardo
- Mellanox Skyway IB-ETH bridges to reach our LAN
  - 16 x 100Gbs

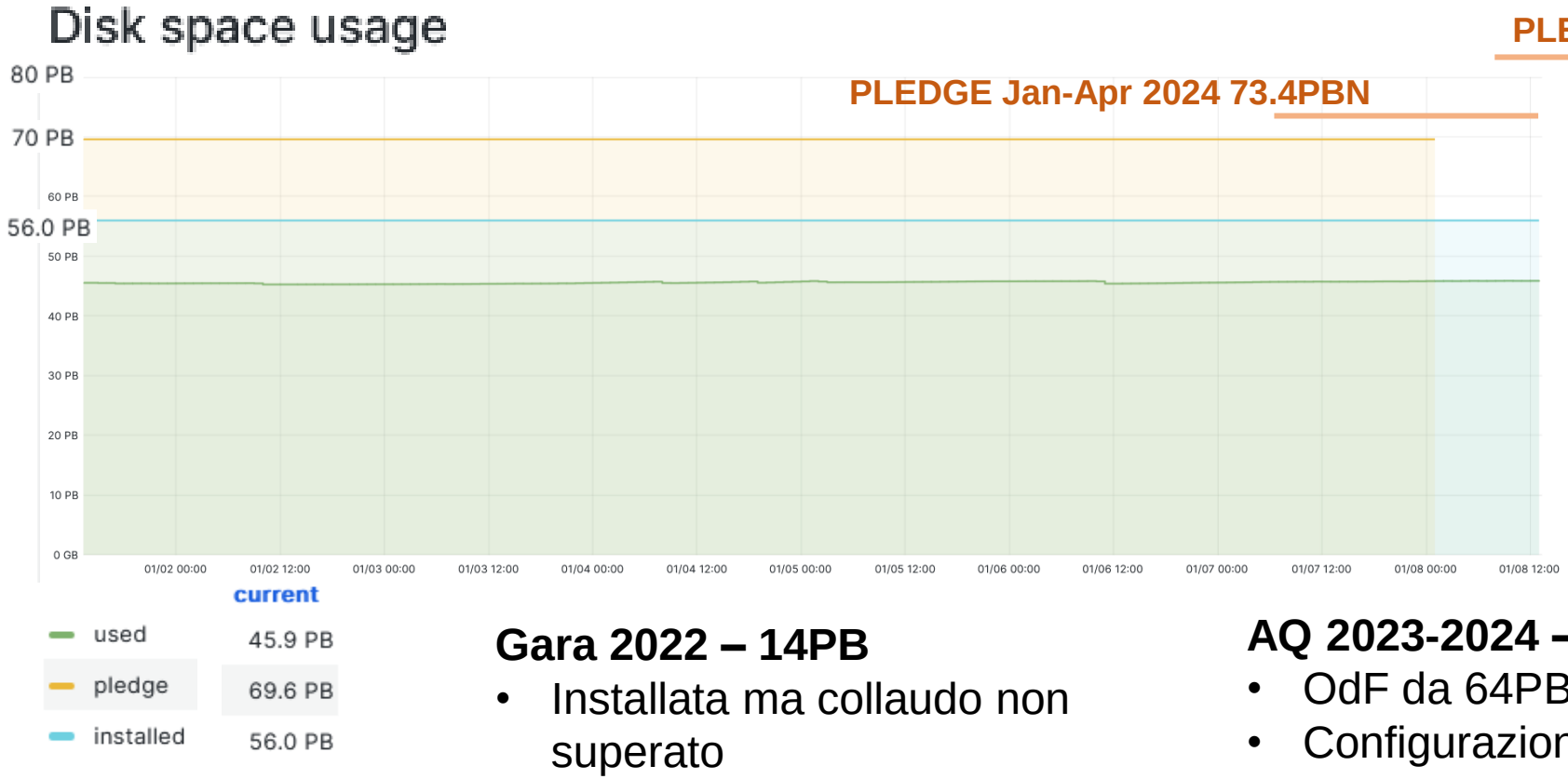
- Standard 2U appliance
- 1.6Tb/s solution
- 8-port HDR/HDR100/EDR InfiniBand
- 8-port 200/100Gb/s Ethernet




```
[a07cna00@login05 ~]$ sbatch -J cn-leo-001 CNAF-VM-new/cnaf.job
Submitted batch job 4553022
[a07cna00@login05 ~]$ squeue --me -O jobid,name,state,timeused,nodelist,reason
JOBID               NAME                STATE               TIME                NODELIST
4553022             cn-leo-001          PENDING             0:00
4519027             cn-leo-003          RUNNING             2-21:13:12          lrdn4379
[a07cna00@login05 ~]$
```

```
[a07cna00@login05 ~]$ squeue --me -O jobid,name,state,timeused,nodelist,reason
JOBID               NAME                STATE               TIME                NODELIST
4519027             cn-leo-003          RUNNING             2-21:15:16          lrdn4379
4553022             cn-leo-001          RUNNING             2:01                lrdn4122
```

Disk Usage @T1 ALL VOs – no cloud



DISK
- Circa 20PBN underpledge
- Mancano risorse delle gare 2022 e 2023
- Estesi contratti di manutenzione per sistemi più vecchi non dismessi (35PB)

Gara 2022 – 14PB
- Installata ma collaudo non superato
- Accordo per un upgrade su sistemi

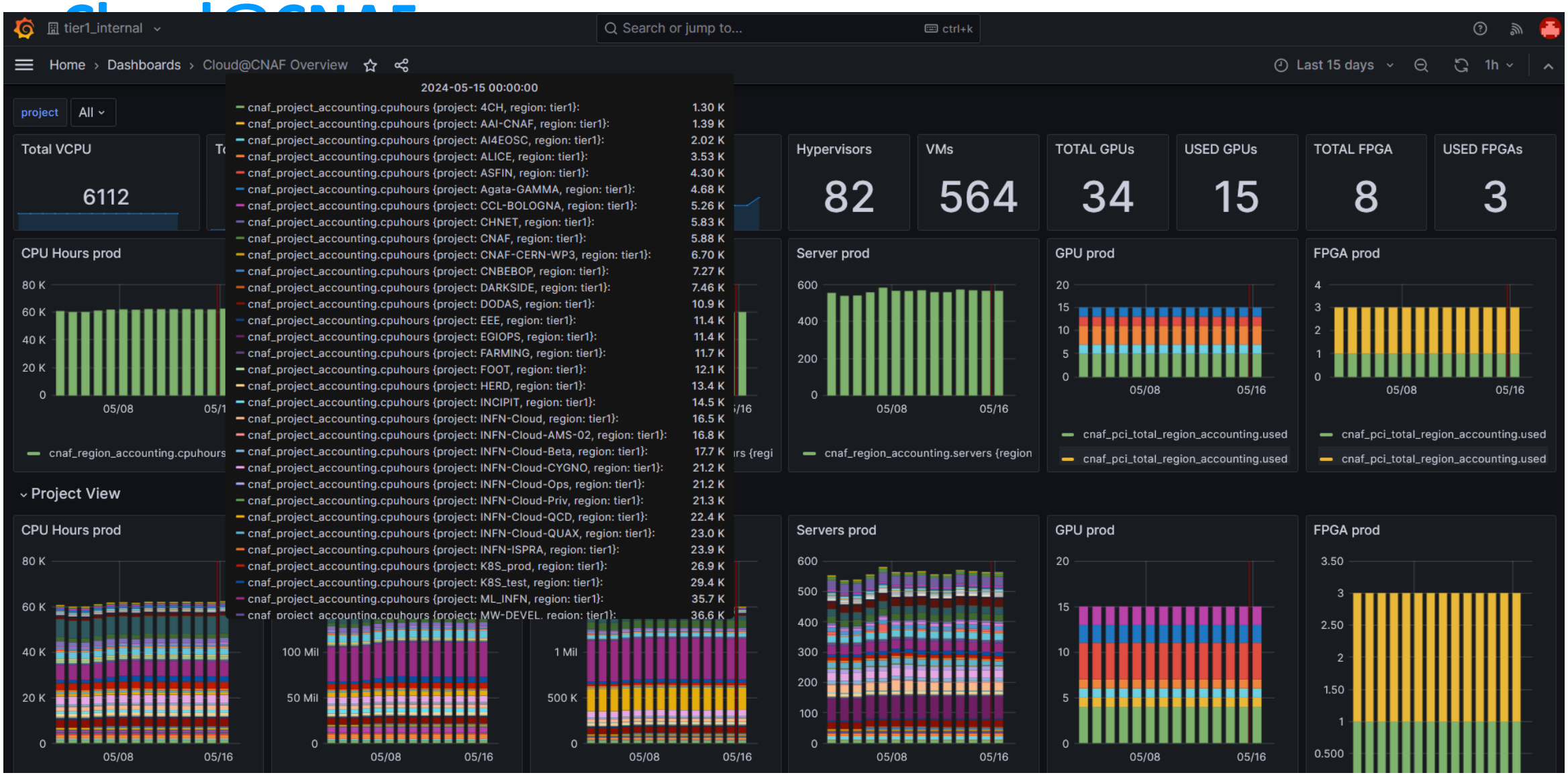
AQ 2023-2024 – (64+16 PBN)
- OdF da 64PBN 2023 installato
- Configurazione e collaudo da ultimare

Cloud@CNAF

GPU+FPGA

TIER1_PCI[V100]=8
TIER1_PCI[T4]=8
TIER1_PCI[A30]=1
TIER1_PCI[A100]=9
TIER1_PCI[RTX5000VGA]=5
TIER1_PCI[U50M]=7 fpga
TIER1_PCI[U250M]=1 fpga
TIER1_PCI[MI200]=3 AMD

90 Tenant
Di cui 11 per INFNCLOUD

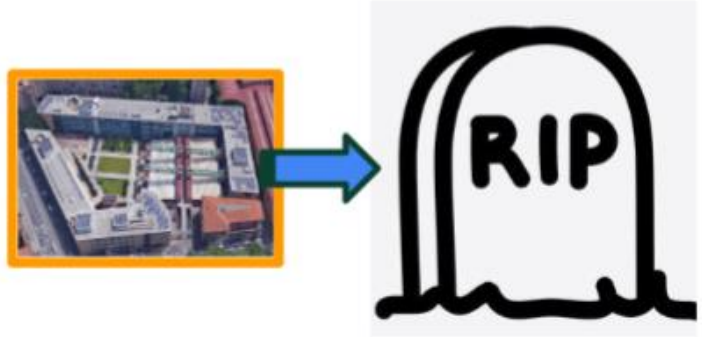


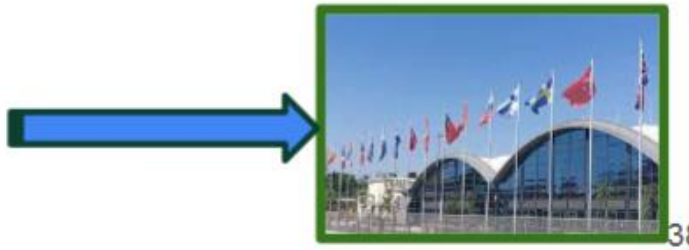
EPIC Cloud@CNAF (ISO27001)

CNAF Tape Libraries and Drives

- **1 x Oracle SL8500**
  - **1 tape library with 16 tape drives T10000D** (8.5TB/cartridge)
  - 80PB installed, 64PB USED
  - Repack on the other libraries needed
    - After completion of repack this library will be dismissed
- **2 x IBM TS4500**
  - **1 tape library with 19 tape drives TS1160** (20TB/cartridge)
    - 102 PB Installed, 50PB USED
    - cannot be further extended due to physical constraints in the current room
    - This library will be moved to the new data center
  - **1 tape library with 18 tape drives TS1170** (50TB/cartridge)

    Installation to be completed this week

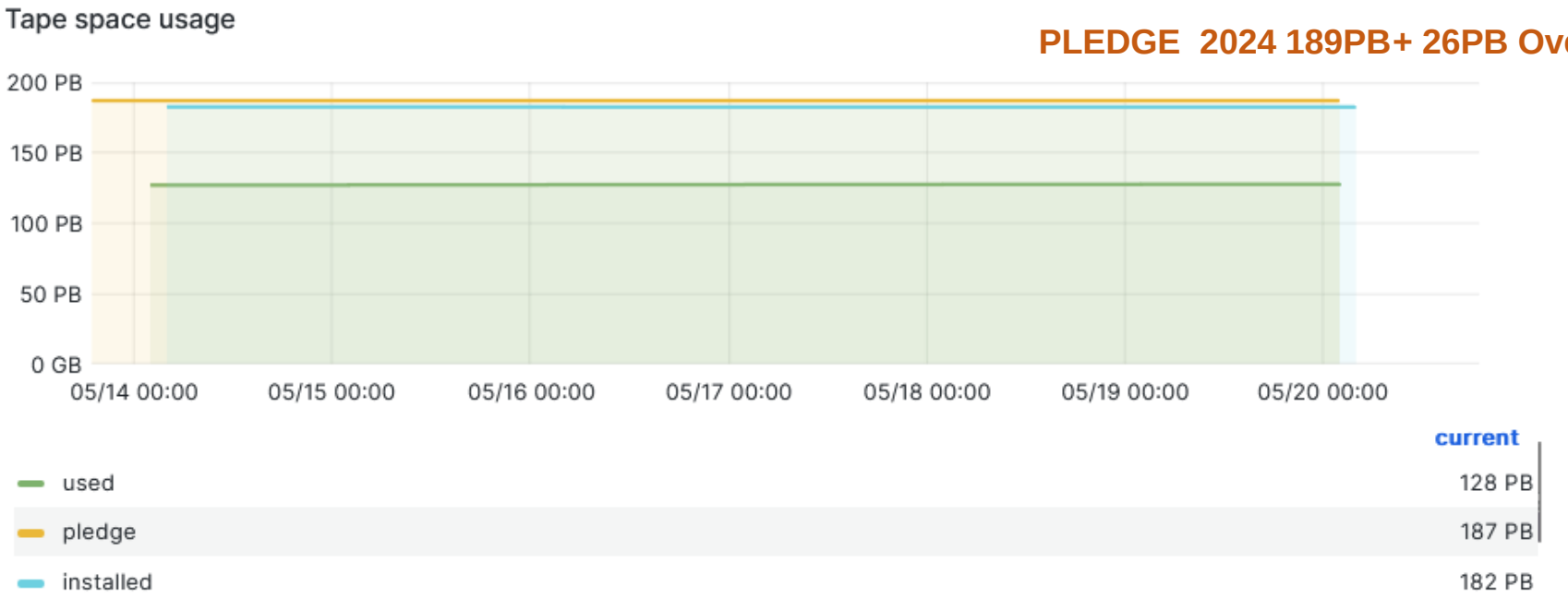




38

Tape Usage @T1 ALL VOs

Tape space usage



PLEDGE 2024 189PB+ 26PB Overpledge

- TAPE
  - RDO da 15PB scade il 28/05
    - 50% in nuovo tecnologia
  - In preparazione AQ da 200PB per 2 anni

Altro sui Tier-2
il survey di WP3 di Datacloud

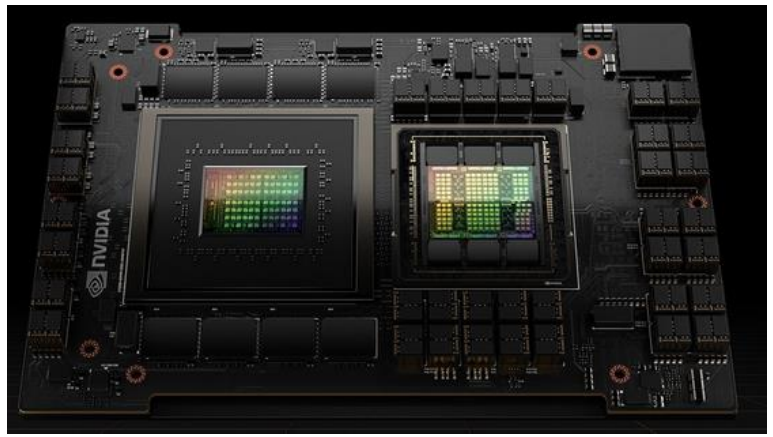
WP3 Survey

- A cosa serve:
  - Survey di tutte le risorse, anche extrapledge accumulate negli anni da progetti vari e ancora operative.
  - Correlazione dati da accounting come richiesto da C3SN
  - Dettagli tecnici per cpu, disco, tape, rete, infrastruttura
  - Dettagli su tipo di finanziamento, costo, allocazione
- Stato:
  - Prima iterazione effettuata
  - Raccolte quasi tutte le risposte
  - In fase di analisi
  - Raccolto anche feedback su come porre le domande del survey

Risorse "eterogenee"

- 4 ARM nodes in production al CNAF
  - 256 cores
  - 1 TB RAM
  - 2x4TB disk
  - **CMS, ATLAS, ALICE**
- 3 AMD GPU al CNAF
- **7 FPGA XILINX U50**
- **1 FPGA XILINX U250**
- + GPU and FPGA acquired with HPC Bubbles

- NVIDIA Grace Hopper Superchip @CNAF e BARI



Key Features

- 72-core NVIDIA Grace CPU
- NVIDIA H100 Tensor Core GPU
- Up to 480GB of LPDDR5X memory with error-correction code (ECC)
- Supports 96GB of HBM3 or 144GB of HBM3e
- Up to 624GB of fast-access memory
- NVLink-C2C: 900GB/s of coherent memory

- NVIDIA Grace Superchip @CNAF e BARI

Key Features

- Up to 144 high-performance Arm Neoverse V2 Cores with 4x128b SVE2
- High-performance NVIDIA Scalable Coherency Fabric with 3.2 terabytes per second (TB/s) bisection bandwidth
- Up to 960 gigabytes (GB) of LPDDR5X memory with error-correction code (ECC) with up to 1TB/s of memory bandwidth
- 900GB/s NVLink-C2C