Pledge Tier1 e Tier2 & Survey Risorse WP3 DataCloud

G.Donvito – INFN-BARI

D. Cesini – INFN-CNAF

Tier1

Resources@T1 2023-2024

ALL VO No Cloud	2023	2024	Delta
Pledge CPU (HS06)	660000	792000 (plan) 703000(with OF) 844000 (w/o OF)	132000
Pledge disk (TBN)	69576	82949	13373
Pledge tape (TB)	158282	193581	35299

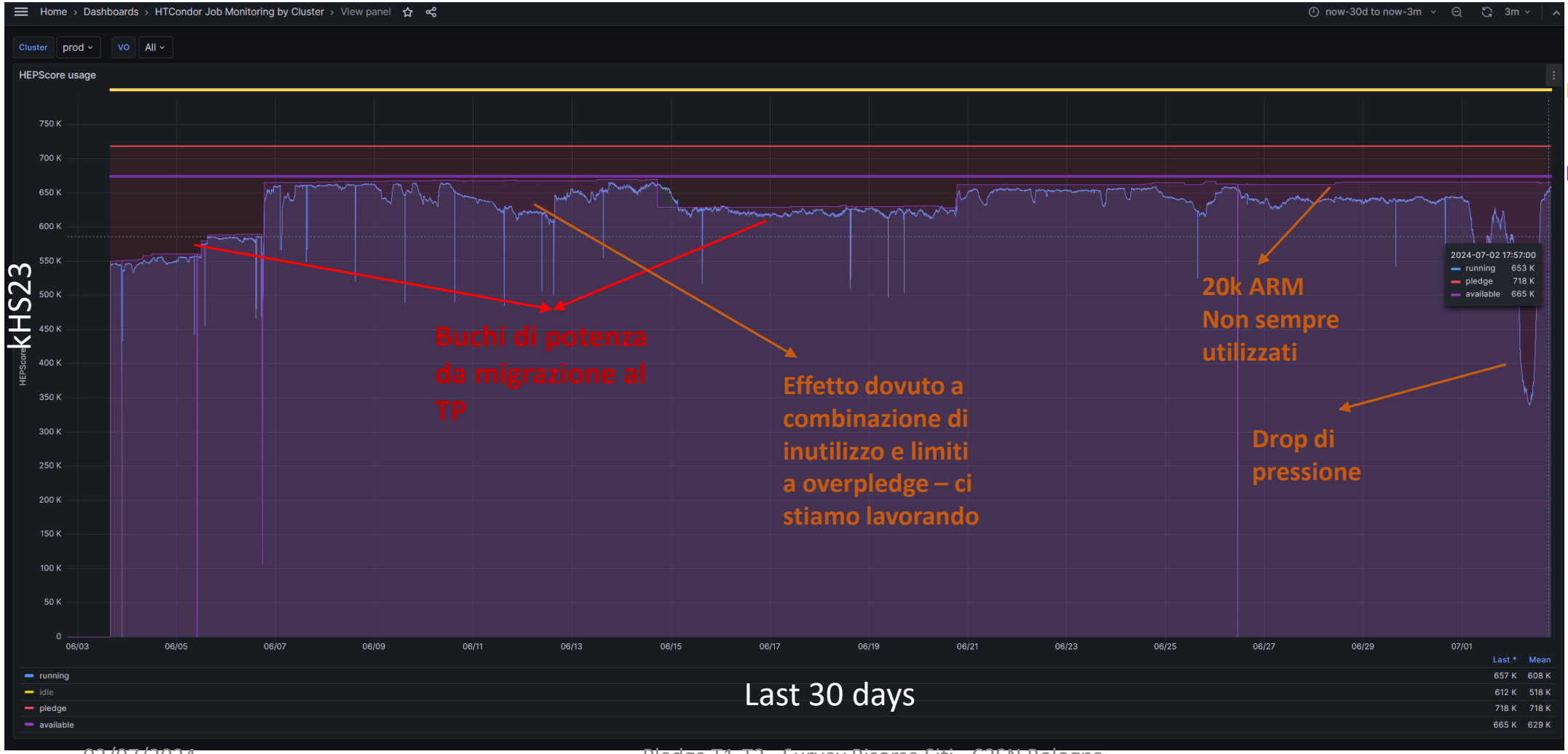
Abbiamo comunicato ad WLCG un ritardo di 30 giorni (01/05)

Abbiamo comunicato ad WLCG un ritardo di 45 giorni (15/06)

T1 CPU

CPU - Farm

- Pledge 2024: 703kHS06 (w/o OVERLAP- 843k) ➔ **CNAF TOTAL PLAN 792kHS06 - Potenza installata Totale: 665KHS06**
- Ultima Gara installata: CPU 2022 in Mar/Apr 2023 ➔ Leonardo non ancora pervenuto, in attesa della configurazione dei bridge IB/ETH da parte di CINECA
- **Tutta la Farm ad HTCondor 23**

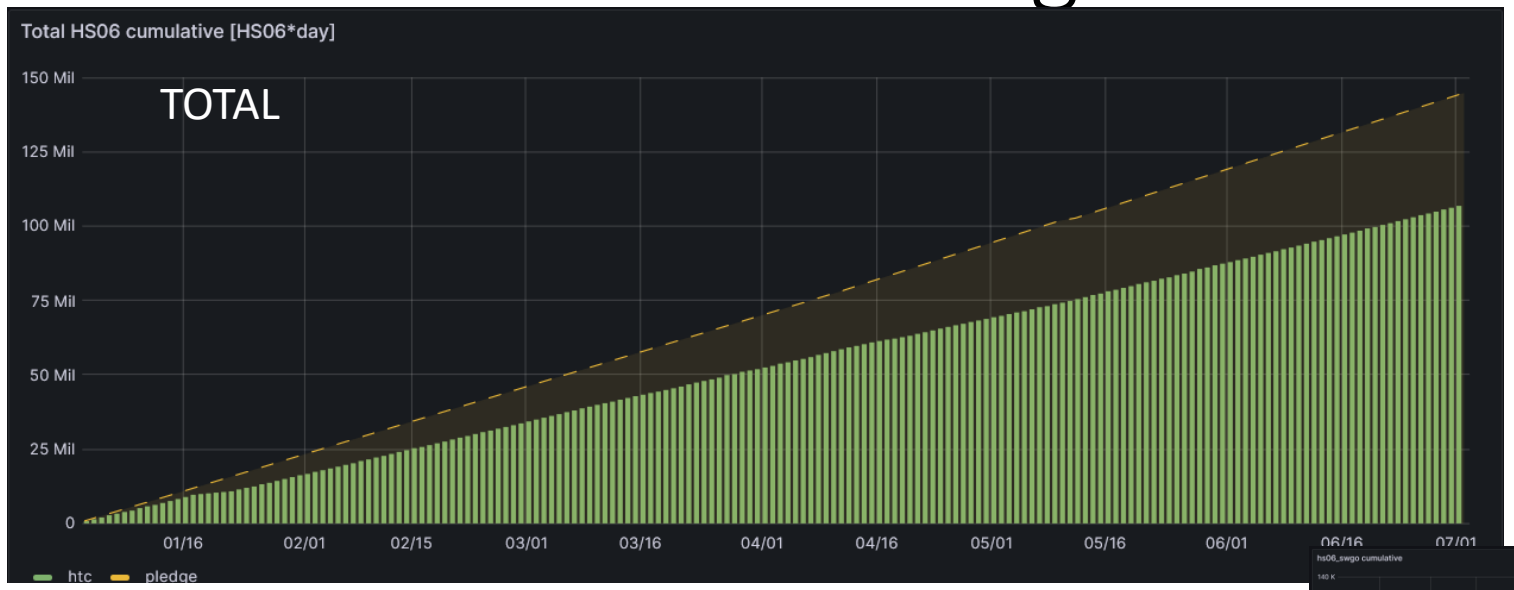


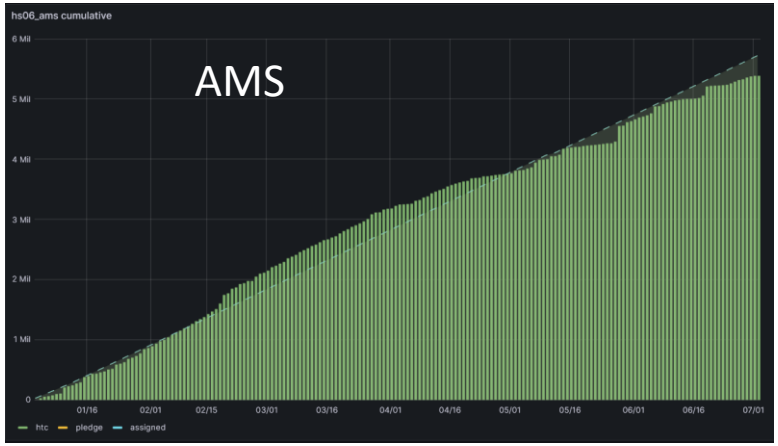
PLEDGE PLAN 792KHS

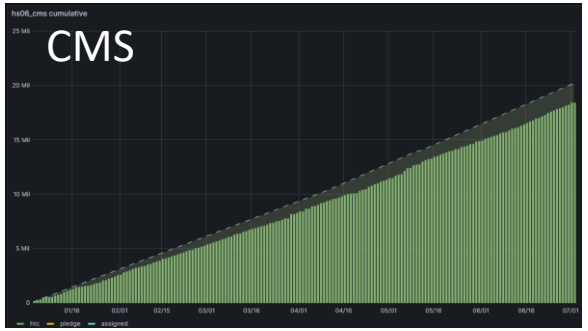
PLEDGE with OF 703KHS
Installed 665KHS

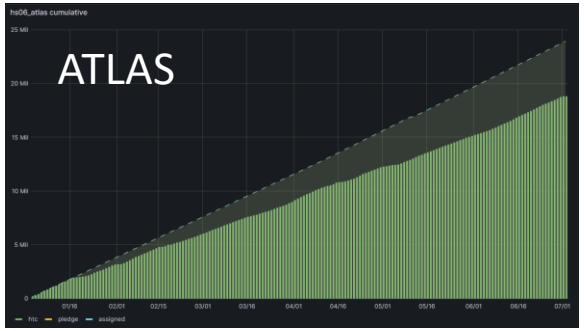
Last 30 days

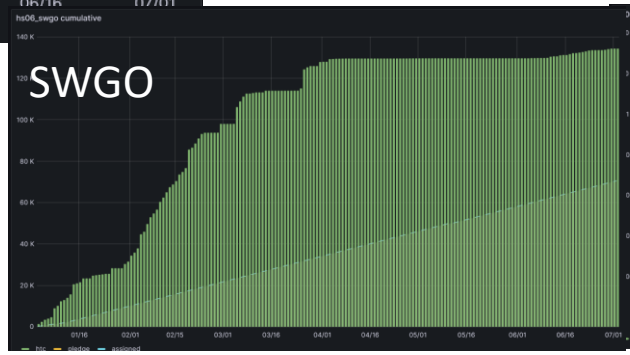
HS06 Integrati – ultimi 6 mesi

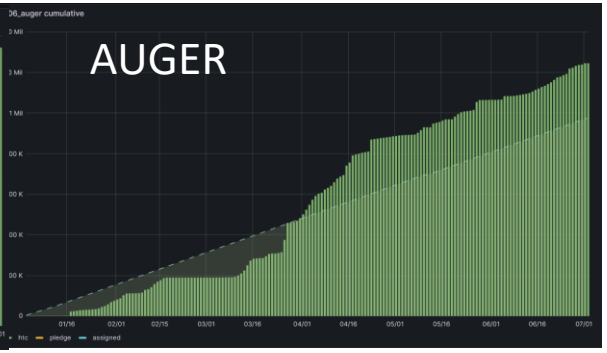


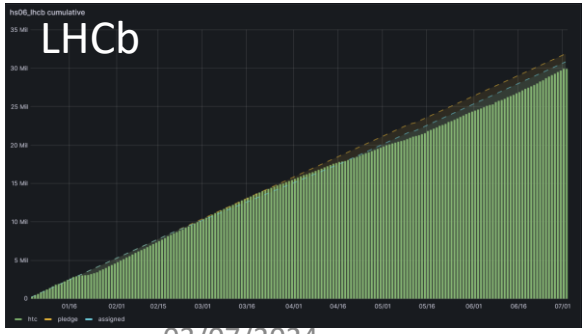


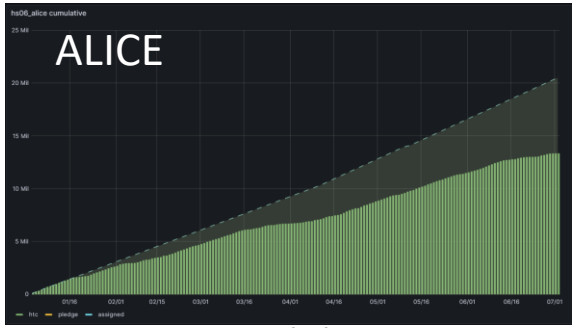


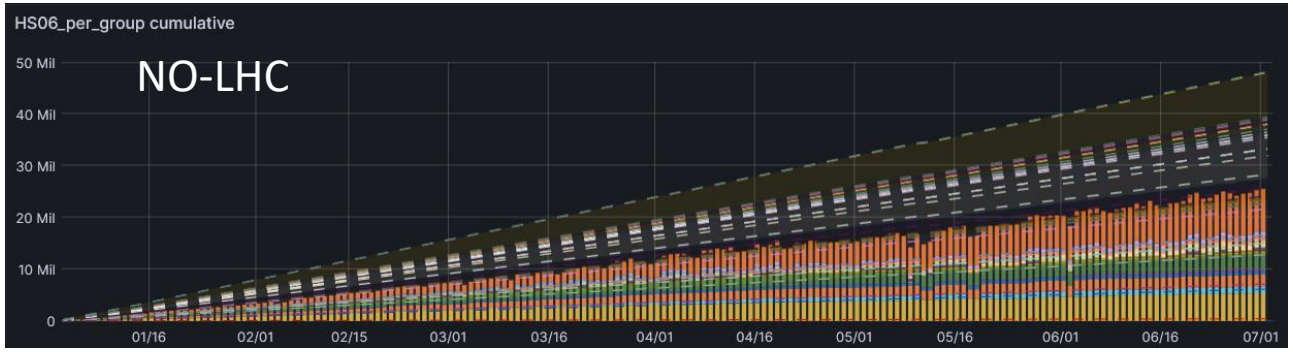




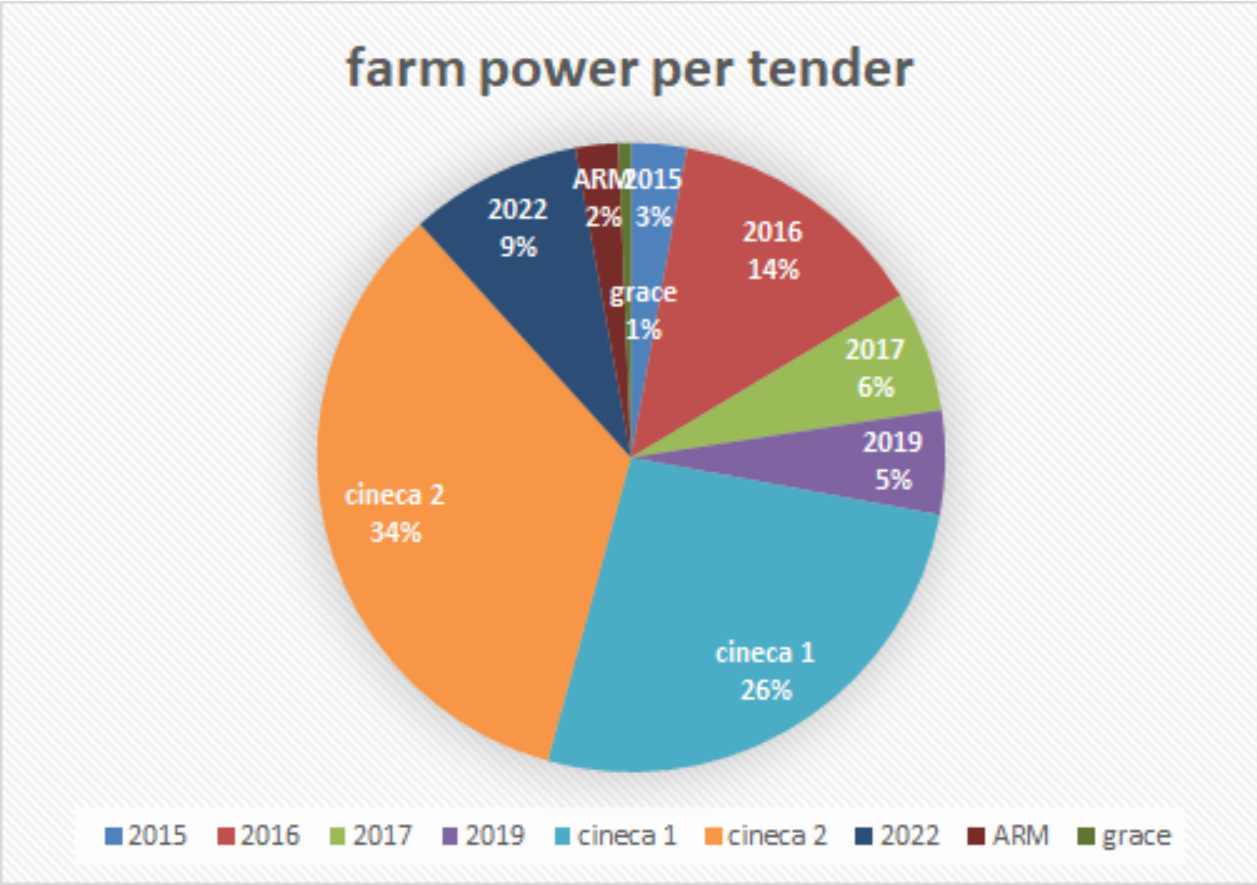








Composizione della farm

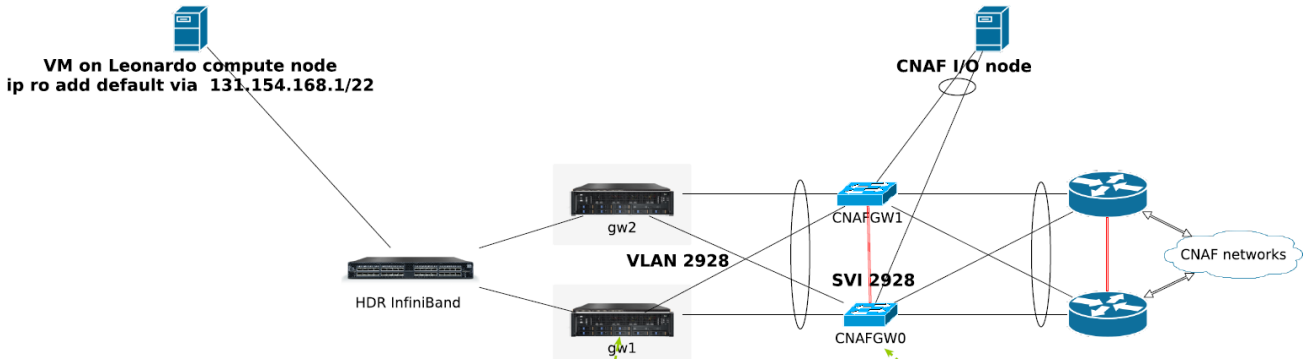


- Attualmente il 60% della potenza è installata al CINECA
  - Manca ancora Leonardo (max 300 nodi)
  - 2880 HS06/nodo
- Aggiunti i nodi ARM
  - Ampere (4 nodi)
    - 3754HS06/nodo
    - 3.74 hs06/W (vs 2.64 HS06/W gara2022)
  - Grace (1 nodo)
    - 4459 HS06/nodo
    - 4.67 HS06/W (vs 2.64 HS06/W gara2022)
  - Solo ATLAS ha dato l'ok per considerare nei pledge 2025+ fino al 30% di risorse su ARM

Gara 2022, 2019, 2017 portate al tecnopolo
2015 e 2016 speriamo di poterle dismettere

Set-up Leonardo GP

- Srategia che stiamo implementando:
  - WN creati tramite job SLURM “infiniti” che instanziano machine virtuali WholeNode gestite da noi
    - Immagini VM create da noi e disponibili via shared fs
  - PCI passthrought per scheda infiniband
  - IP pubblico su interfaccia infiniband
  - Accesso inboud/outbound via NVIDIA Skyway collegati direttamente a nostri apparati
  - Pilot con alcuni nodi Galileo@casalecchio aveva dato esito positivo (ormai mesi fa)
    - La configurazione con Leonardo è abbastanza cambiata
      - Abbiamo 2 Skyway
      - MTU 2042 invece che 4096



02/07 – Call con rete CINECA

03/07 – setup con CINECA al TP dei primi 200Gbs su 1.6Tbs

Leonardo down

- LEONARDO: slow down in the $SCRATCH area 8/03 - 20/03 --> 13 gg
- Leonardo: unexpected power issues 21/03 - 29/03 --> 9 gg
- Leonardo: issues with the $WORK and $DRES areas 03/04 - 03/05 --> 30 gg
- Leonardo partial unavailability for maintenance from April 15 to 19 15/04 - 19/04 --> 5 gg
- Leonardo: unexpected issues 7/05 8/05 2 gg
- LEONARDO scheduled maintenance 14/05 - 16/05 e 21/05 - 23/05 --> 6 gg
- Leonardo scheduler problem 3/06 --> 1 gg
- LEONARDO partial maintenance 4/06 - 5/06 --> 2 gg
- LEONARDO: $SCRATCH area issues 4/06 --> 1 gg
- Leonardo: short interruption of Slurm service tomorrow and job mail notification 12/06 --> 1 gg
- LEONARDO: partial Booster and full DCGP maintenance June 24th and 25th 24/06 - 25/06 --> 2 gg
- Leonardo: SLURM unexpected issues 21/06 --> 1 gg
- Leonardo: production down 30/06 - 1/07 --> 2 gg
- Leonardo short stop of job submission for July 3rd 3/07 --> 1 gg

- 06/03 al 03/07 ➔ 119 gg
  - 33 giorni con Leonardo DOWN
  - 43 di DOWN dei fs che non sappiamo se e quanto ci impatta direttamente

- **Rimane il fatto che Leonardo è down per quasi 1/3 del tempo, almeno da questa breve statistica**
  - **Non è detto che tutti i down elencati avrebbero creato problem ai nostril job “infiniti”**

- **Numero di volte che ci hanno contattato per concordare l'intervento: 0**

Leonardo down

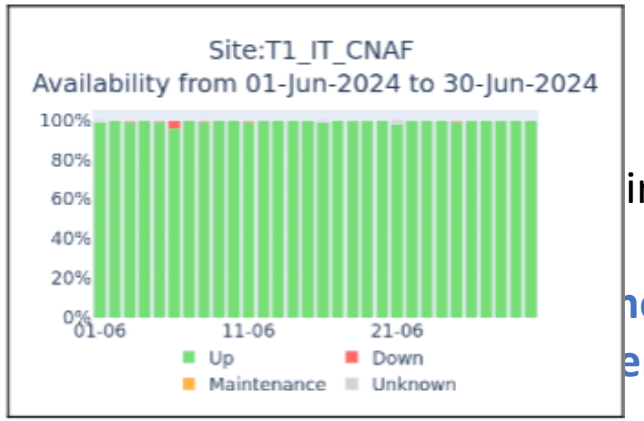
- LEONARDO: slow down in the $SCRATCH area 8/03 - 20/03 --> 13 gg
- Leonardo: unexpected power issues 21/03 - 29/03 --> 9 gg
- Leonardo: issues with the $WORK and $DRES areas 03/04 - 03/05 --> 30 gg
- Leonardo
  - 19/04 --
- Leonardo
- LEONARD
- Leonardo

- LEONARDO partial maintenance 4/06 - 5/06 --> 2 gg
- LEONARDO: $SCRATCH area issues 4/06 --> 1 gg
- Leonardo: short interruption of Slurm service tomorrow and job mail

WLCG Target Availability for each site is 97.0%. Target for 8 best sites is 98.0%

Availability Algorithm: (CREAM-CE + ARC-CE + HTCONDOR-CE + GLOBUS) * (all SRMv2 + all SRM + all GRIDFTP)

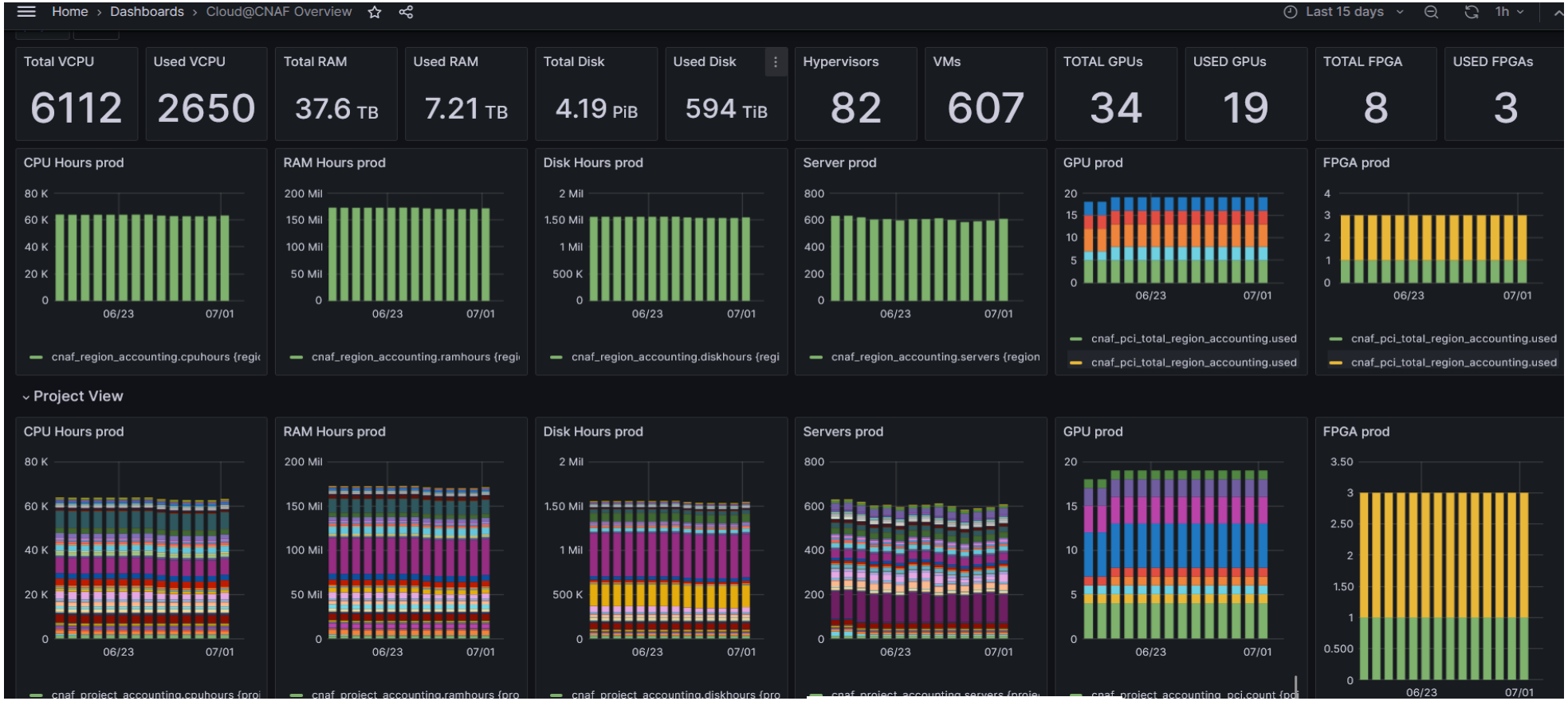
- 06/03 al 03/07 ➔ 119 gg
  - 33 giorni con Leonardo DOWI
  - 43 di DOWN dei fs che non sa irettamente



T1_IT_CNAF Avail: 100.0% Unkn: 0.0%

- **Rimane il fatto che Leonardo è do neno da questa breve statistica**
  - **Non è detto che tutti i down em ai nostril job "infiniti"**
- **Numero di volte che ci hanno contattato per concordare l'intervento: 0**

Stato Cloud@CNAF



- Parte dei pledge «HTC» assegnati su Cloud@CNAF per accesso interattivo o piccoli cluster dedicati
  - AGATA, NTOF, etc..
- VIRGO Low Latency Cluster on K8s

- Circa 100 tenant configurati
  - Cloud@CNAF
  - INFNCLOUD
- Pledge assegnato a tutti gli esperimenti dei referaggi 2022 e 2023 con label “CLOUD”

	CPU (HS06)	Disk (TB-N)
QUAX	100	130
AMS-02	200	
HERD	1.000	100
SWGO	40	
Fermi	1.100	
AUGER	80	
Cygno	160	10
Totale	**2.680**	**240**

	Crescita netta	
	CPU (HS06)	Disk (TB-N)
Cygno	2.800	125
Darkside	200	100
NEWS	50	10
QUAX	400	130
Totale	**3.450**	**365**
Totale effettivo	**2.760**	**365**

Il totale effettivo con OF per ora non implementato

T1 DISK and TAPE

Disk storage in produzione

Installed: 53.64 PB, Pledge 2024: 82.1 PB, Used: 48.8 PB

	Storage system	Model	Net capacity, TB	Experiment	End of support
	ddn-10, ddn-11	DDN SFA12k	10120	ALICE, AMS	12/2022 (+10 spare hdd)
2015	os6k8	Huawei OS6800v3	3400	GR2, Virgo	07/2024
2016	md-1,md-2,md-3,md-4	Dell MD3860f	2308	DS, Virgo, Archive	12/2024
	md-5, md-6 e md-7	Dell MD3820f	50	metadati, home, SW	11/2023 e 12/2024
2017	os18k1, os18k2	Huawei OS18000v5	7800	LHCb	7/2024
2018	os18k3, os18k5, os18k5	Huawei OS18000v5	11700	CMS	6/2024
	ddn-12, ddn-13	DDN SFA 7990	5840	GR2,GR3	2025
	ddn-14, ddn-15	DDN SFA 2000NV	24	metadati	2025
	os5k8-1,os5k8-2	Huawei OS5800v5	8999	ATLAS	2027
	Cluster CEPH	12xSupermicro SS6029	3400	ALICE, cloud, etc.	2027

Da Rimpiazzare con AQ 23-24

Da spostare al Tecnopolo

Mancano 14PB da Pledge Storage 2022 – bloccati nella relativa Gara – Collaudo non superato

AQ 2023-2024 - primo AS da 64PBN in fase di collaudo – fondi per secondo AS in arrivo a Settembre (ICSC) (anche per eventuale 6/5)

Acquisti storage recenti e futuri

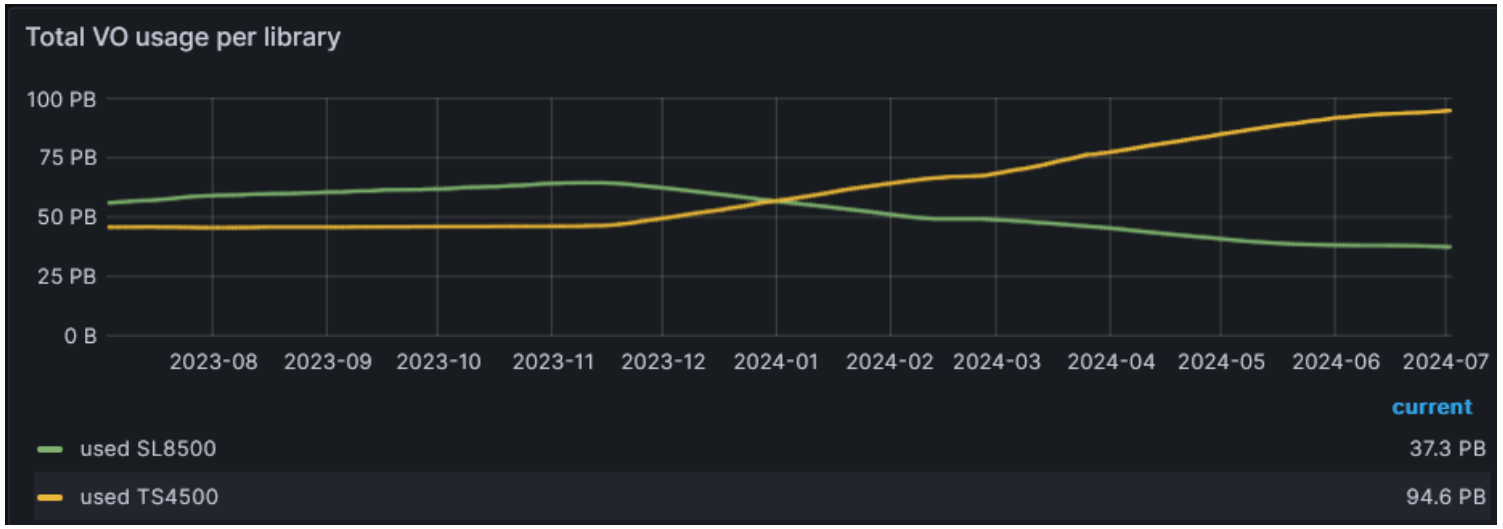
- Gara storage 2022 (14PB netti)
  - Nuova proposta con apparati DDN SFA7990X
  - In attesa per la consegna entro giugno

- AQ storage 2023-2024 (Terabit+ICSC)
  - Huawei OceanStore Micro 1500/1600
    - 8 sistemi di 10PB + 40 server
  - Installazione e collaudo in corso del primo AS 64PB
  - Secondo AS da 16 PB a Settembre
  - 6/5? occorre decidere se acquistarlo su ICSC
- Tape Library (ICSC)
  - Nuova libreria Installata, manca cablaggio FC per completare il collaudo
- Gare nastri (ICSC)
  - Acquistati 14PB (JE e JF)

  - URGENTE: Nuova gara di acquisto tape JF da 96PB
    - Spediamo in AC questa settimana
      - Anche se fondi disponibili da Settembre
    - Pledge+Overpledge+ICSC+Repack vecchia Oracle da dismettere

Stato Tape

- Pledge 2024: 190PB
- Installato: 182PB
- Usato: 132PB
- Nuova Libreria già installata direttamente al Tecnopolo, non ancora in produzione
- Circa 200 cassette libere ➔ 4PB
- 14PB da installare nelle due librerie IBM
  - Una da mettere online al TP



Tier-2

Survey Risorse
Siti Datacloud

Survey Risorse Siti Datacloud

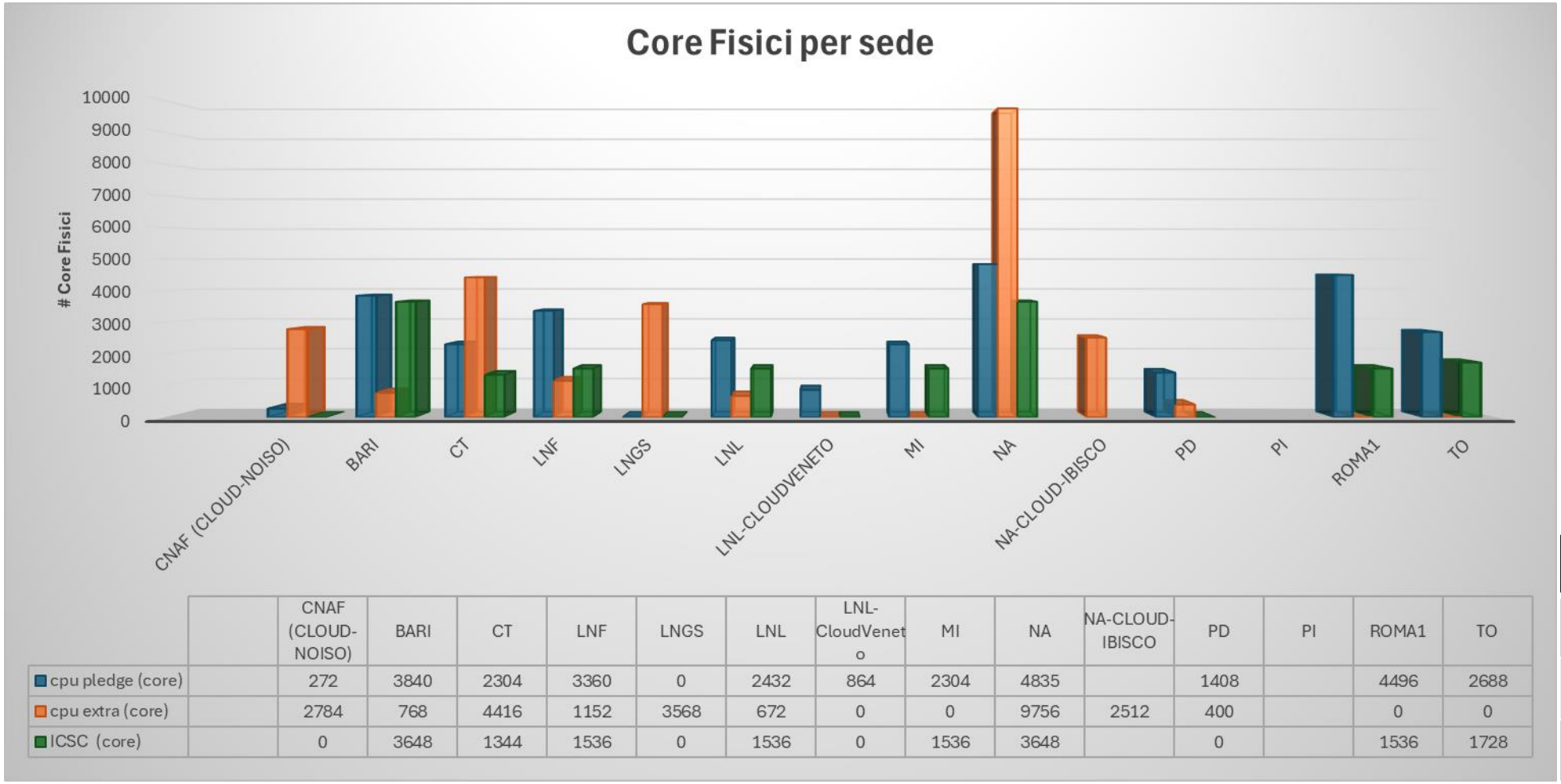
	CPU Pledge	Note (a chi sono pledged)	CPU Extrapledge TOT	CPU Extrapledge progetto 1	CPU Extrapledge progetto n	GPU Extrapledge
RISORSE CPU/GPU Tipo 1						
HS06						
HEPScore						
Num Socket						
Num Core						
CPU Model						
Num Nodes						
Num Rack						
Batch system						
Cloud						
Data Acquis.						
Maint Expiry date						
Interconnect Type						
RAM/core						
eur/core						
GPU Type						
GPU number						
eur/GPU						

	Disk Pledge	Note	Disk Extrapledge TOT	Disk Extrapledge progetto 1	Disk Extrapledge progetto n
RISORSE DISK Tipo 1					
TB_N					
Num Server					
RU					
Num controller					
FS Type					
Cloud					
Data Acquisizione					
Maint Expiry date					
Interconnect Type					
eur/TB_N					

- **Tier1 (Cloud)**
- **9 Tier2 (8 risposte)**
  - **Cloud-Veneto**
  - **NA-CLOUD-IBISCO**
- **LNGS**
- **PD**

HS06	
HEPScore	
Num Core	
CPU Model	
Num Nodes	
Num Rack	
Batch system	
Cloud	
Data Acquis.	
Maint Expiry date	

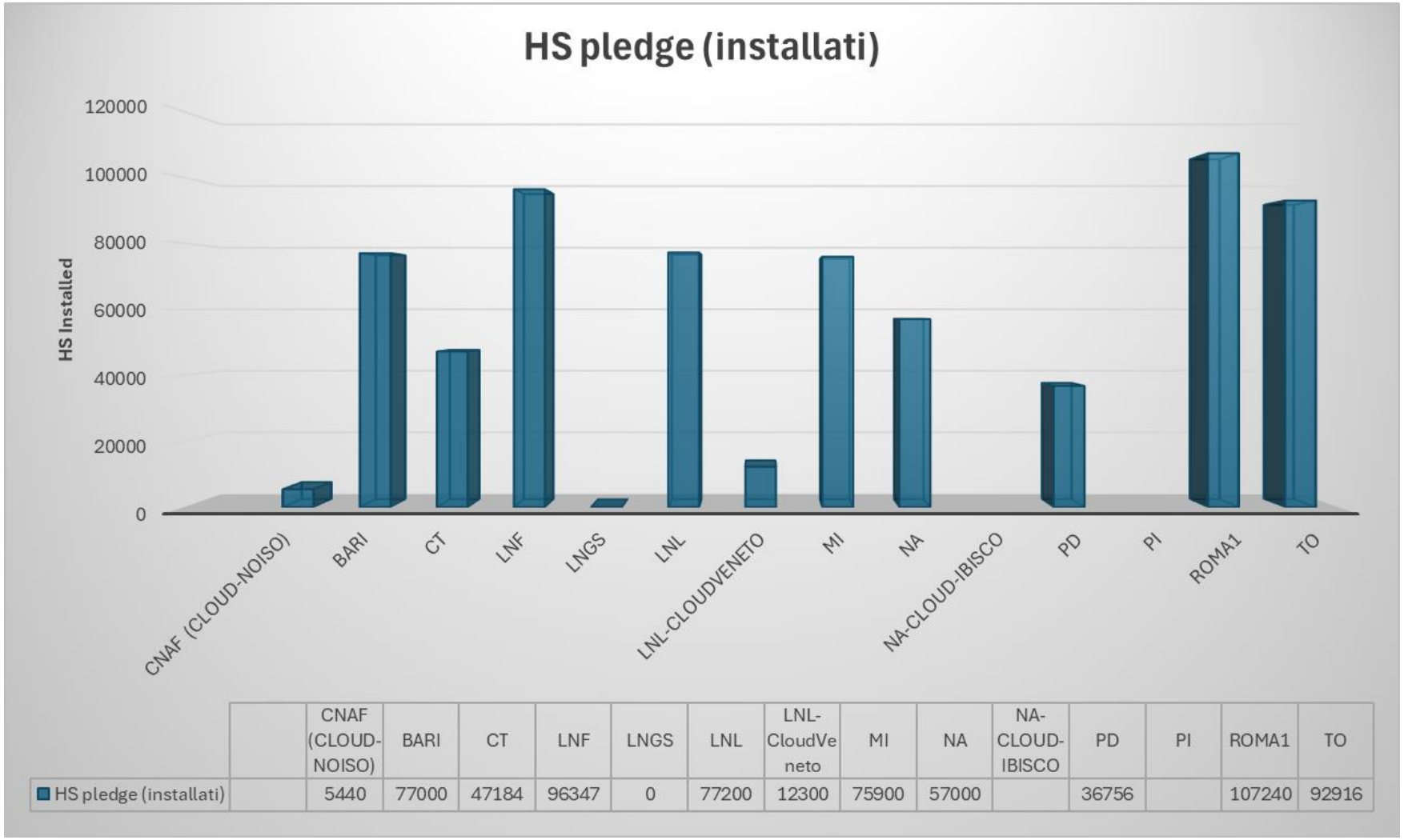
RISORSE DISK Tipo N			
TB_N			
Num Server			
RU			
Num Controller			
FS Type		Total	Used
Cloud	KW cooling		
Data Acquis.	KW IT		
Maint Expiry date	RU		
Interconnect Type			
eur/TB_N			



		CNAF (CLOUD-NOISO)	BARI	CT	LNF	LNGS	LNL	LNL-CloudVeneto	MI	NA	NA-CLOUD-IBISCO	PD	PI	ROMA1	TO
cpu pledge (core)		272	3840	2304	3360	0	2432	864	2304	4835		1408		4496	2688
cpu extra (core)		2784	768	4416	1152	3568	672	0	0	9756	2512	400		0	0
ICSC (core)		0	3648	1344	1536	0	1536	0	1536	3648		0		1536	1728

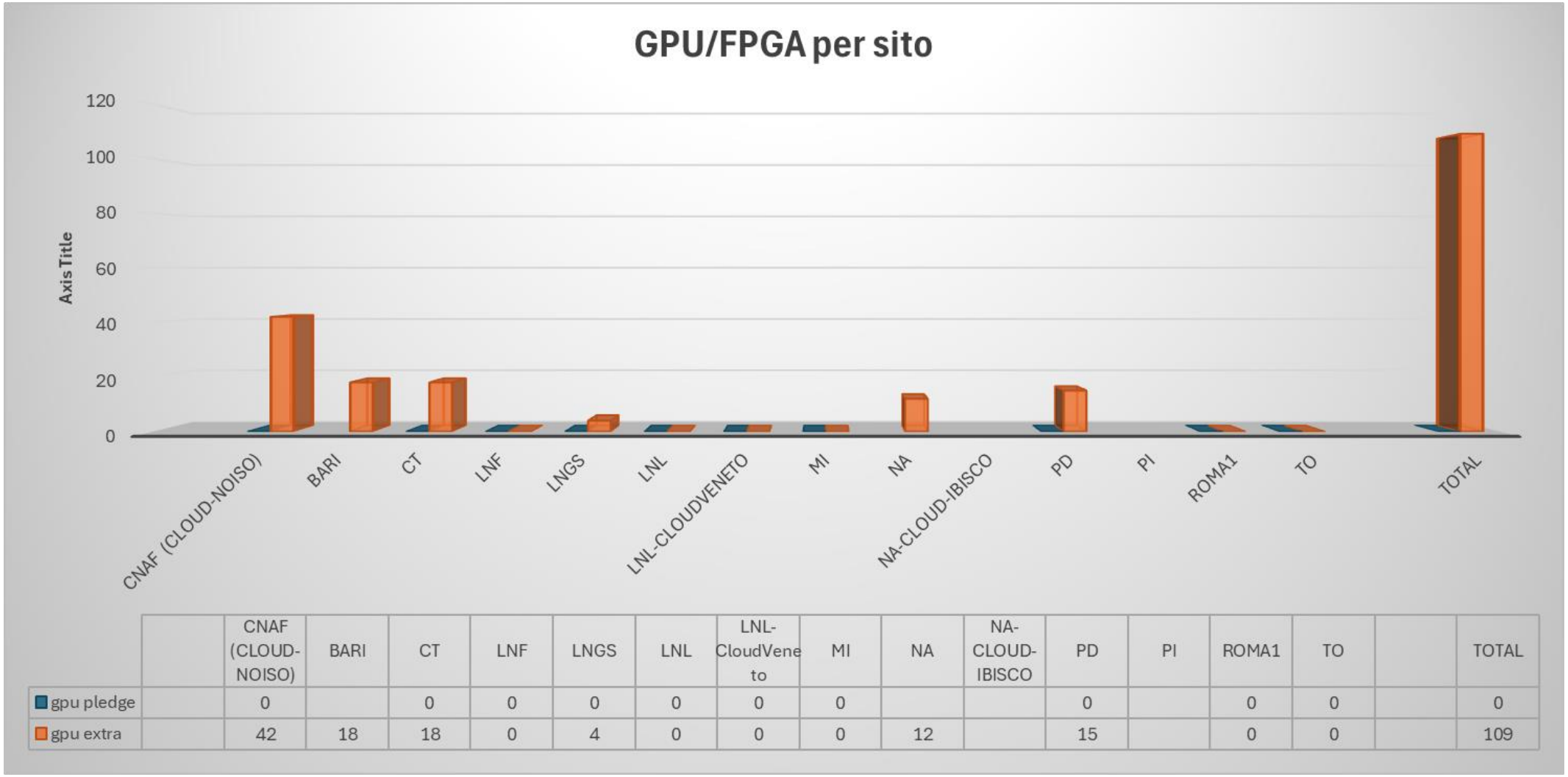
	TOTAL
Pledged	28803
Extra	26028
ICSC	16512

ICSC expected = 18048
1728 da pisa

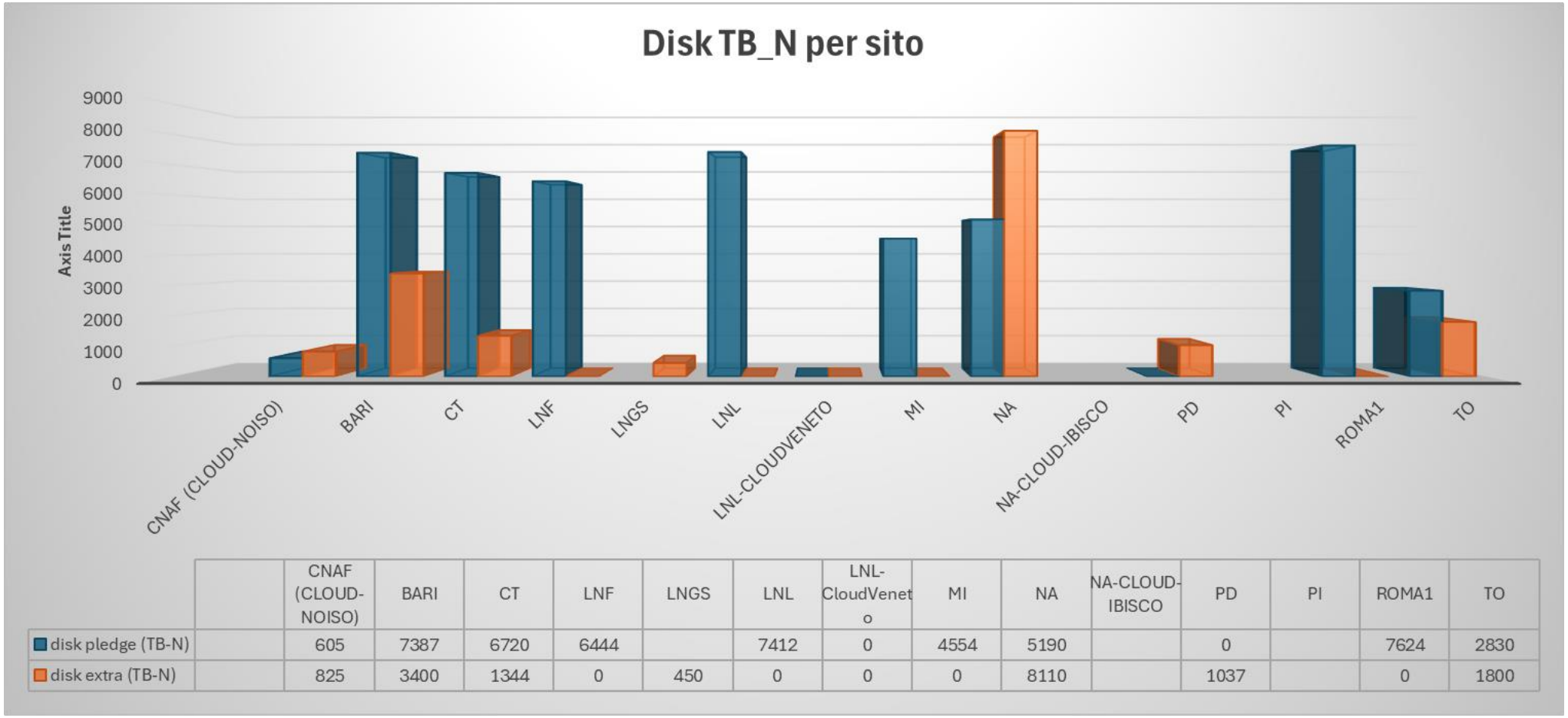


	CNAF (CLOUD-NOISO)	BARI	CT	LNF	LNGS	LNL	LNL-CloudVeneto	MI	NA	NA-CLOUD-IBISCO	PD	PI	ROMA1	TO
HS pledge (installati)	5440	77000	47184	96347	0	77200	12300	75900	57000		36756		107240	92916

TOTAL=685283
CRIC(WLCG T2)= 567275



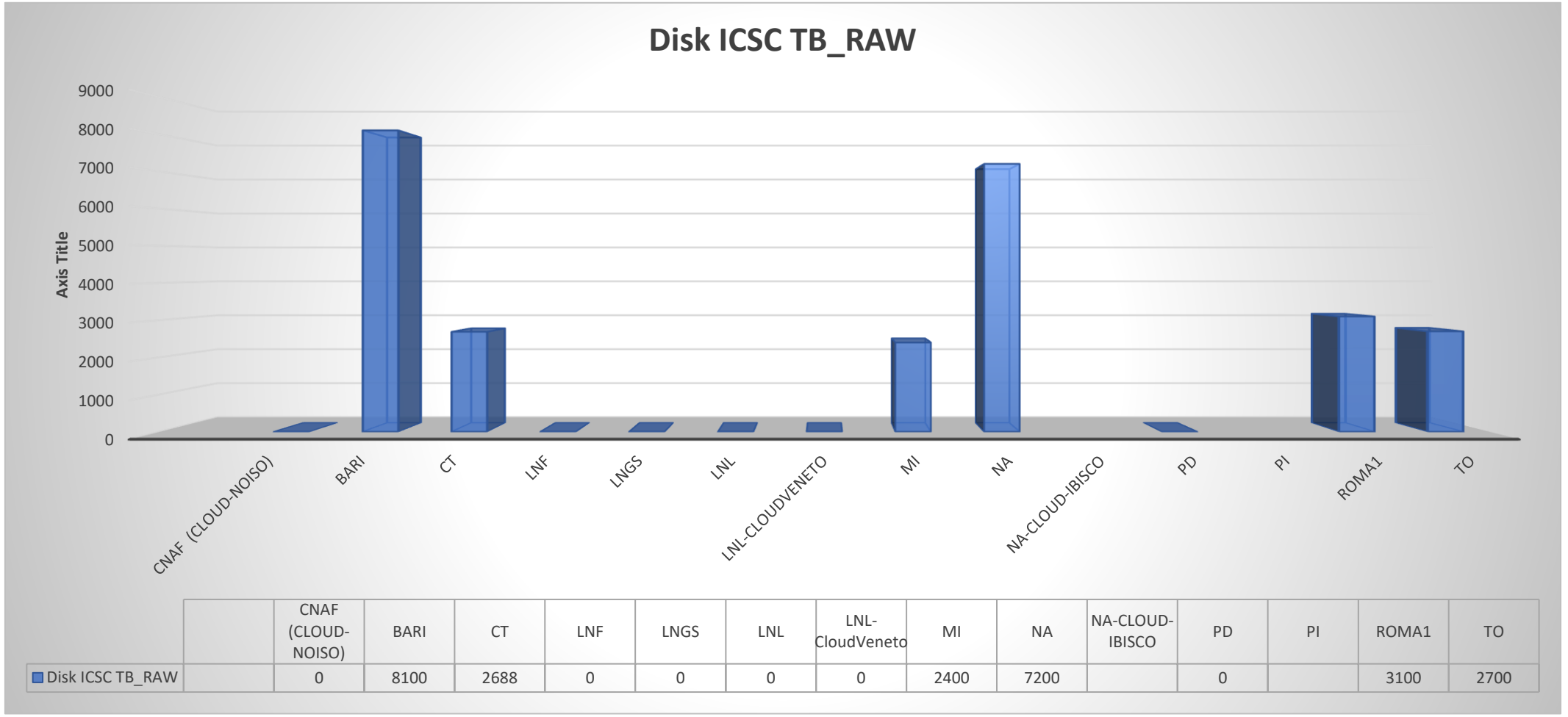
		CNAF (CLOUD-NOISO)	BARI	CT	LNF	LNGS	LNL	LNL-CloudVeneto	MI	NA	NA-CLOUD-IBISCO	PD	PI	ROMA1	TO		TOTAL
gpu pledge		0		0	0	0	0	0	0			0		0	0		0
gpu extra		42	18	18	0	4	0	0	0	12		15		0	0		109



	CNAF (CLOUD-NOISO)	BARI	CT	LNF	LNGS	LNL	LNL-CloudVeneto	MI	NA	NA-CLOUD-IBISCO	PD	PI	ROMA1	TO
disk pledge (TB-N)	605	7387	6720	6444		7412	0	4554	5190		0		7624	2830
disk extra (TB-N)	825	3400	1344	0	450	0	0	0	8110		1037		0	1800

	TOTAL TB-N
Pledged	48766
Extra	16966

CRIC Total Pledge T2 (WLCG): 4660TB-N



		CNAF (CLOUD-NOISO)	BARI	CT	LNF	LNGS	LNL	LNL-CloudVeneto	MI	NA	NA-CLOUD-IBISCO	PD	PI	ROMA1	TO
Disk ICSC TB_RAW		0	8100	2688	0	0	0	0	2400	7200		0		3100	2700

Non da tutti incluso nelle risposte non essendo ancora in prod
Probabilmente richiesta da chiarire

Considerazioni sul survey Datacloud WP3

- Dobbiamo chiarire meglio le richieste
  - i.e. core fisici vs threads, pledge vs non-pledge, TB-N vs TB-raw
  - Limiti temporali, cosa includere e cosa no
- Gestire via spreadsheet è molto laborioso
  - Difformità risposte, aggregazione manuale
  - Inserimento nuovi dati e aggiornamento
- Occorre un tool dedicato, magari semplice e «brutto» ma che faccia il lavoro che serve
  - Eventualmente da integrare con tool accounting WP1
- Effettuata analisi solo dei dati sulla quantità di risorse, manca quella sul resto dei dati
  - RAM, Maintenance expiry, costi, infrastruttura, protocolli di accesso
- Non abbiamo incluso nelle richieste:
  - dati sulle risorse di rete
  - le risorse certificate ISO
  - Le risorse da «HPC bubbles»

Gara "HPC Bubbles"

- Accordo Quadro Nazionale
  - Listino prezzi per nodi + accessori
  - 2 anni di validità
  - Lotto1
    - CPU, GPU, FPGA
  - Lotto2
    - Storage
  - Sedi Coinvolte: CNAF, BARI, MI-BI, PI, TO, LNGS, NA, RM1, PD/LNL
- **Stato gara**
  - **Ordini inviati (a parte 6/5)**

Quantità nodi con fondi Terabit-ICSC-DARE

	Nodo CPU	Nodo GPU	Nodo FPGA Xilinx	Nodo FPGA Terasic	Nodo storage
BA	24	6	0	0	32
CNAF	26	30	2	2	52
MIB	0	0	2	2	0
NA	18	1	2	0	8
PD	6	6	0	0	0
PI	20	0	0	0	0
RM1	12	0	0	0	0
TO	14	6	0	0	0
LNGS	0	6	0	0	12
CT	12	0	0	0	8
LNF	12	0	0	0	0
LNFESA	8	6	0	0	6
LNL	4	0	0	0	0
MI	4	0	0	0	0
TOTALE	160	61	6	4	118

Core: 30 kcore fisici
Circa 34 HS/core

GPU: 244 NVIDIA H100
40 FPGA
InfiniBAnd 400Gbs

45 PB RAW

Claudio Grandi

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