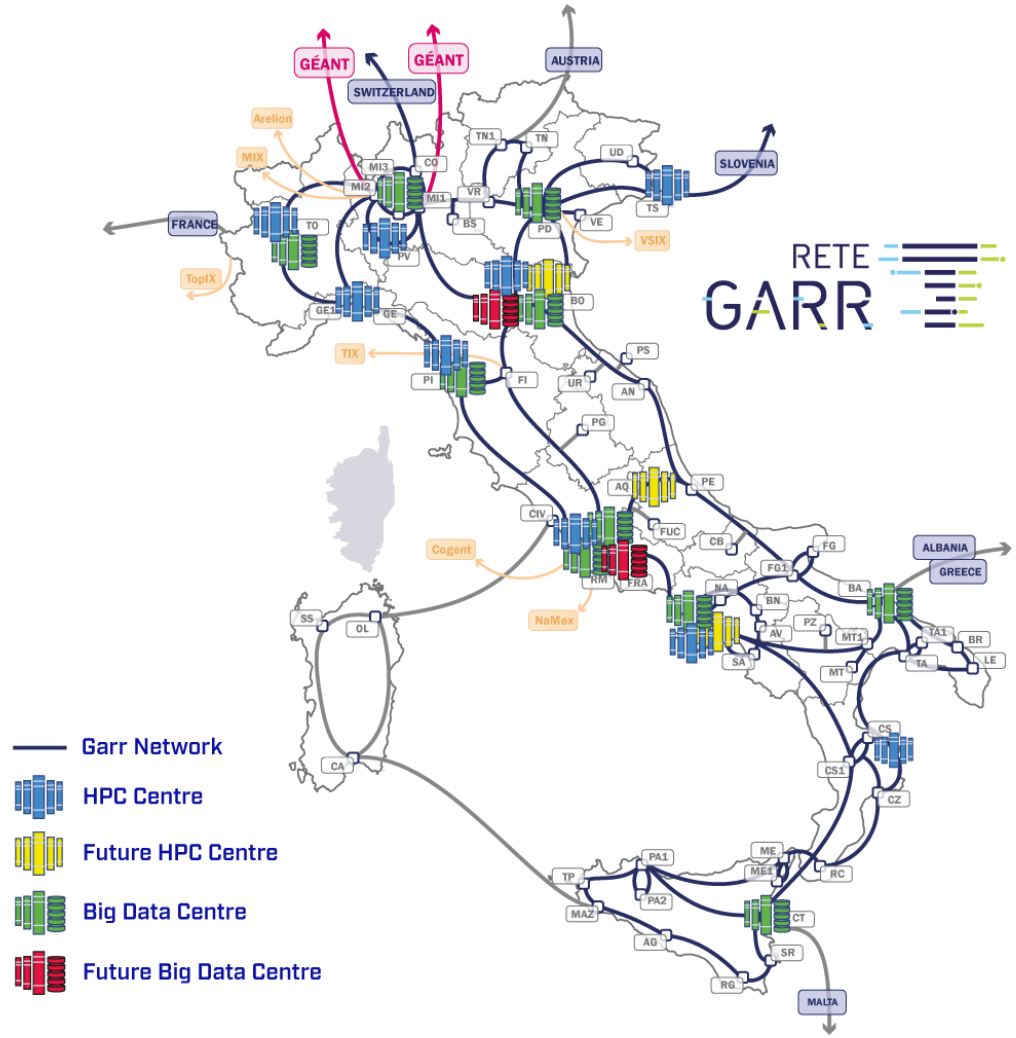
DataCloud – WP3
Risorse HTC e HPC

D.Cesini – INFN-CNAF

Outline

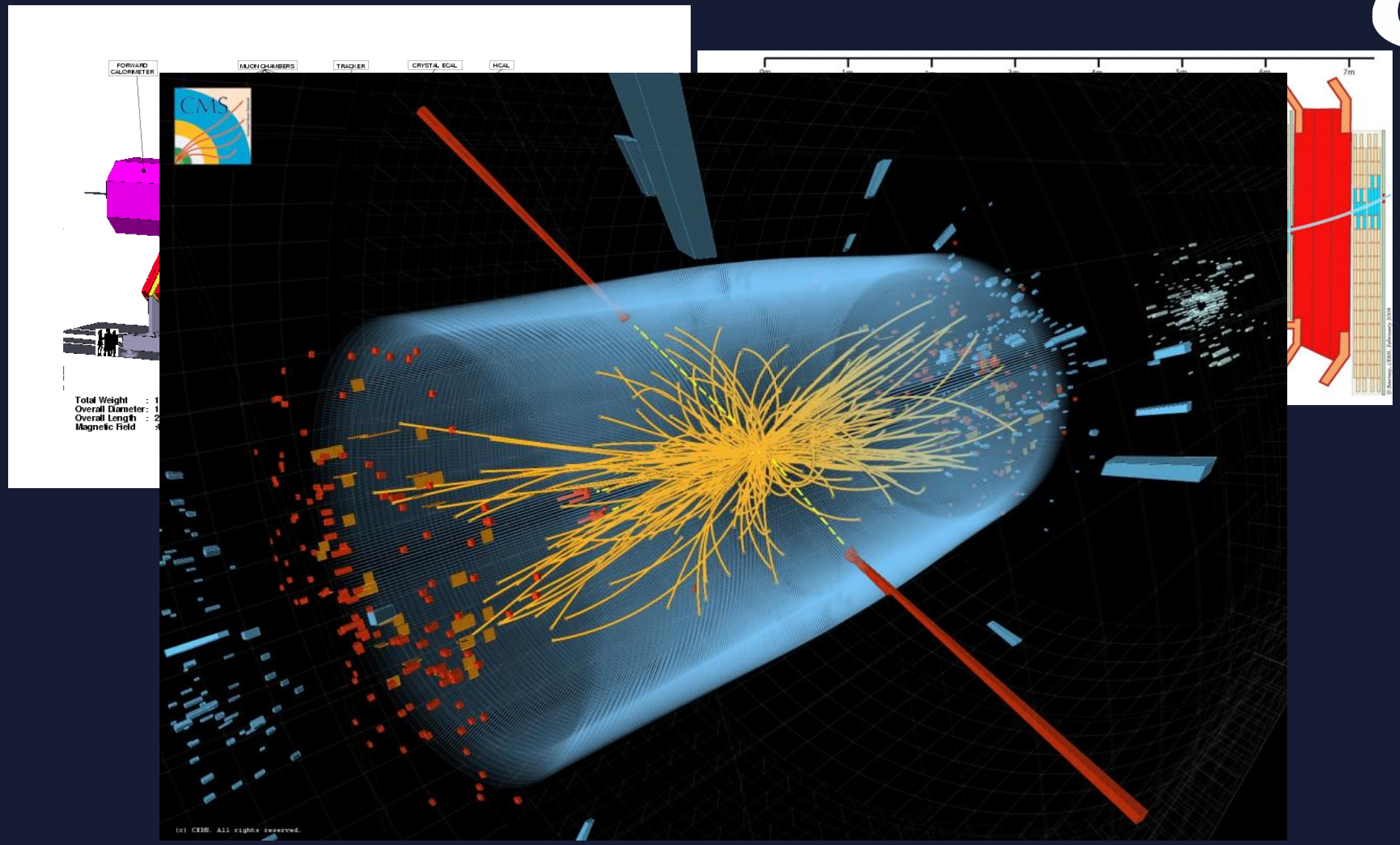
- HTC vs HPC
- La Grid
  - Perché federare le risorse
  - Evoluzione verso il Datalake
- Risorse pledged Al Tier-1 e ai Tier-2
  - CPU, DISK, TAPE
- HPC
- Le HPC Bubble

INFN Datacloud distributed resources



HTC and HPC - definition

- High Throughput Computing (HTC)
  - The focus is on the execution of many copies of the *same program* at the *same time*
    - not in the speedup of individual jobs
  - Many copies of the same program run *in parallel* or *concurrently*
  - Maximize the **throughput**
- High Performance Computing (HPC)
  - speed up the individual job as much possible so that results are achieved more quickly
- HTC infrastructures tend to deliver large amounts of computational power over a long period of time.
  - In contrast, High Performance Computing (HPC) environments deliver a tremendous amount of compute power over a short period of time.
- The interest in HTC is in how many jobs complete over a long period of time instead of how fast an individual job can complete.


FORWARD CALORIMETER
MUON CHAMBERS
TRACKER
CRYSTAL ECAL
HCAL
Total Weight
Overall Diameter
Overall Length
Magnetic Field
CMS
(c) CERN. All rights reserved.

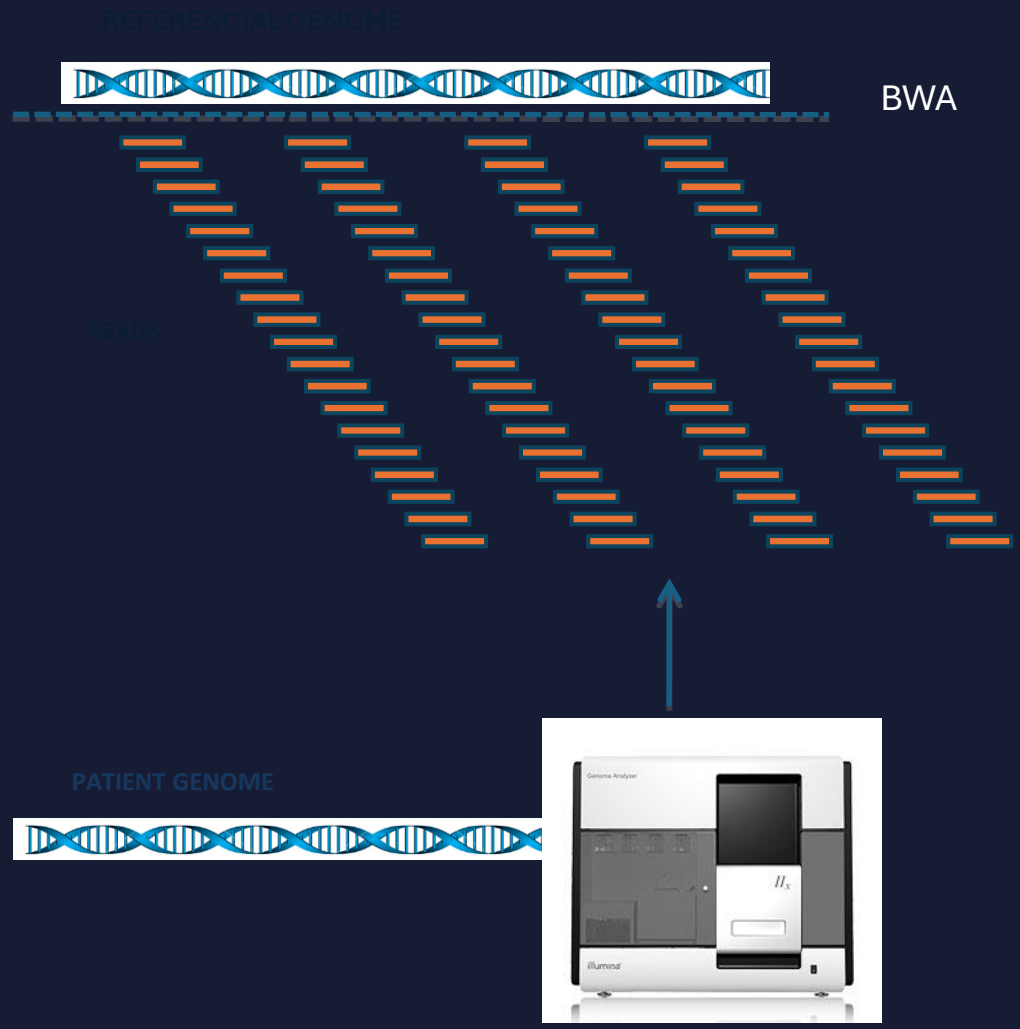
Massively Parallel Genome Sequencing

Used in the study of cancer Diseases

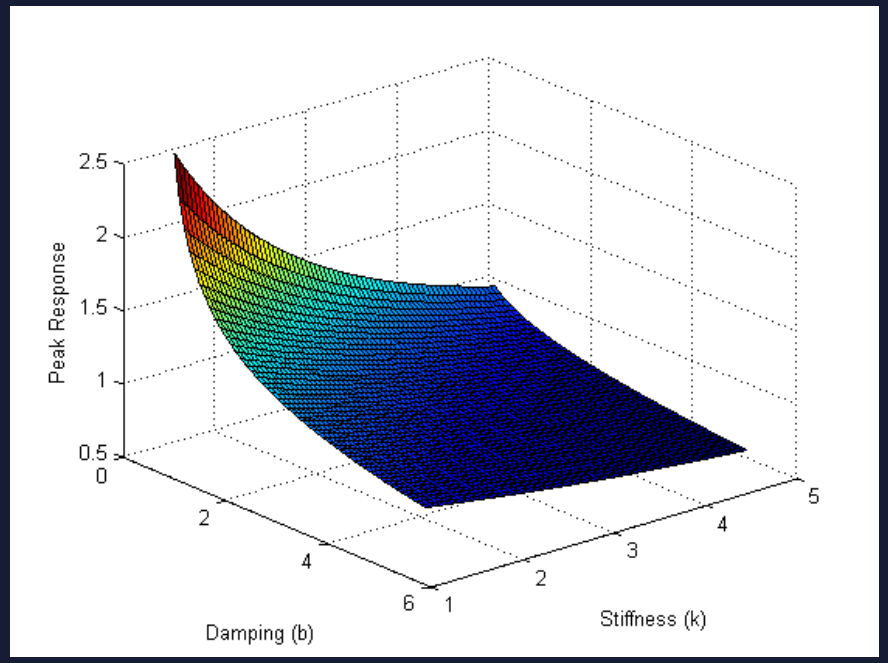
Allows massive amount of DNA or RNA fragments to be sequenced in a single experiment.

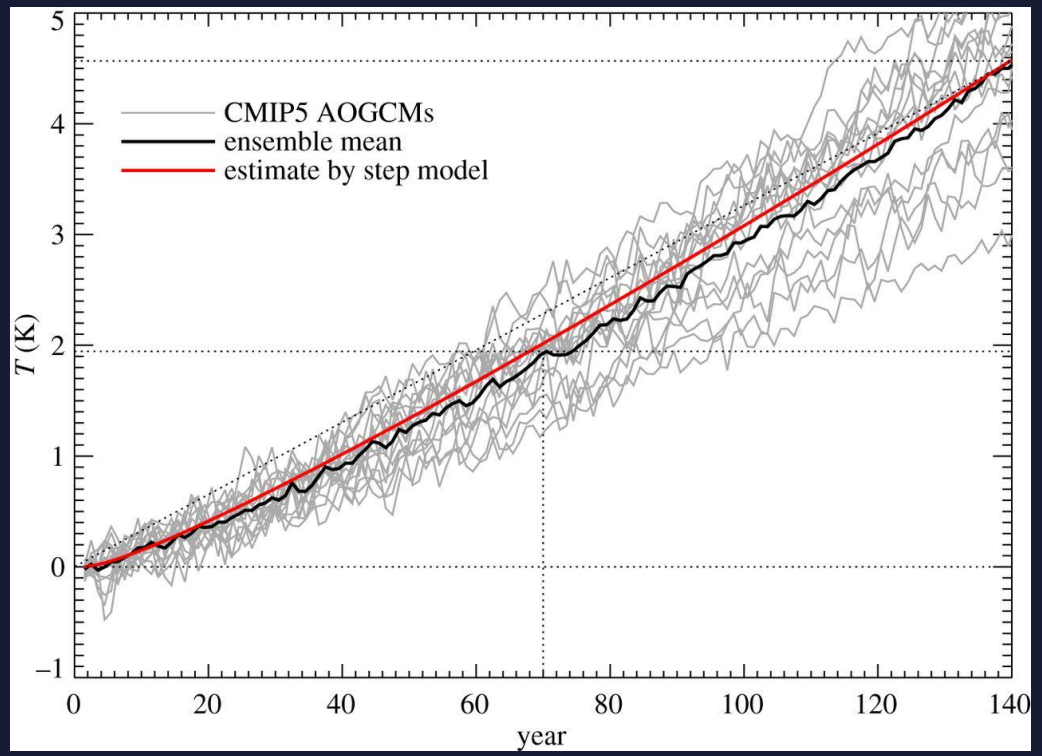
For the *massively parallel sequencing it is used* **BWA** tool (Burrows-Wheeler Aligner) for indexing and alignment

- Memory request~ 3,5 GB
- Total time ~ 50h using the group local resources



Parameter sweep and ensemble simulations





Grids and distributed systems

- What is a Grid?
- Grid types
- Anatomy of a Grid
- Accessing a Grid

© Grant Faint

What is a Grid? - Early definition

Ian Foster

I.Foster, C.Kesselman: The Grid: Blueprint for a New Computing Infrastructure", 1998

Carl Kesselman

"A computational Grid is a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities"

What is a computational Grid? the 3 points checklist

A Grid is a system that.....

1) **Coordinates resources that are not subject to centralized control**
2) **Uses standard, open, general-purpose protocols and interfaces**
3) **Delivers nontrivial qualities of service** (Ian Foster, 2002)

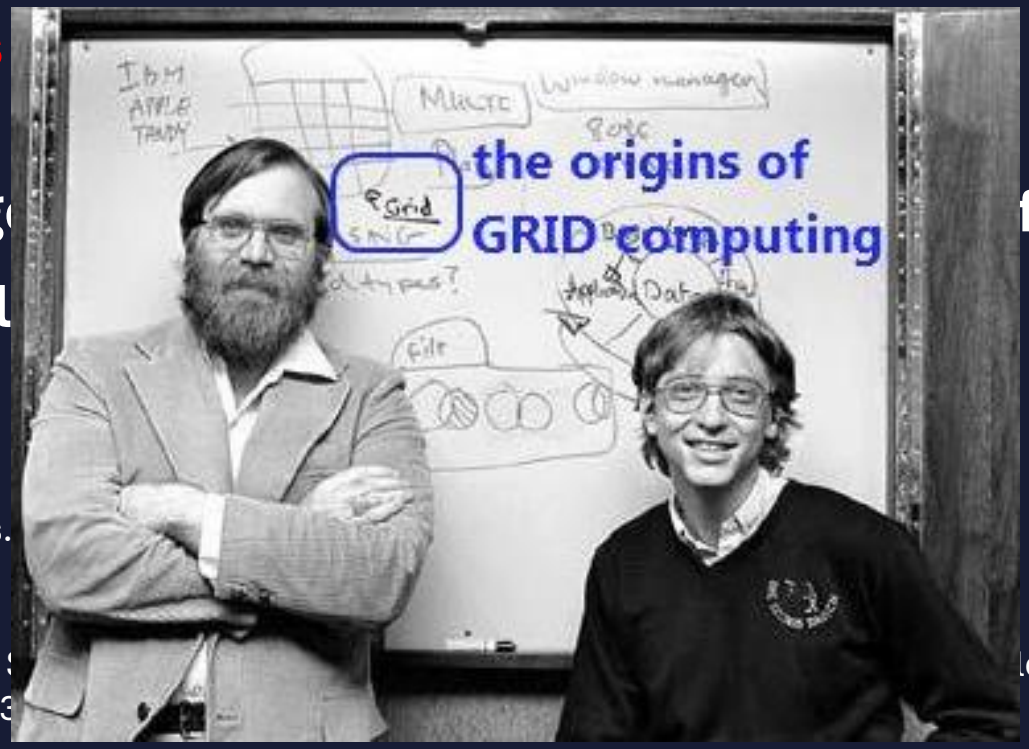
[1] Foster, I. and Kesselman, C. eds. The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, 1999, 259-278
[2] Ian Foster, Carl Kesselman, and Steven Tuecke. 2001. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. Int. J. High Perform. Comput. Appl. 15, 3 (August 2001), 200-222. DOI=10.1177/109434200101500302
[3] What is the Grid? A Three Point Checklist. I. Foster, GRIDToday, July 20, 2002.

What is a computational Grid? the 3 point checklist

A Grid is a system that.....

1) **Coordinates resources control**
2) **Uses standard, open, g faces**
3) **Delivers nontrivial qual**

(Ian Foster, 2002)



[1] Foster, I. and Kesselman, C. eds.
Morgan Kaufmann, 1999, 259-278

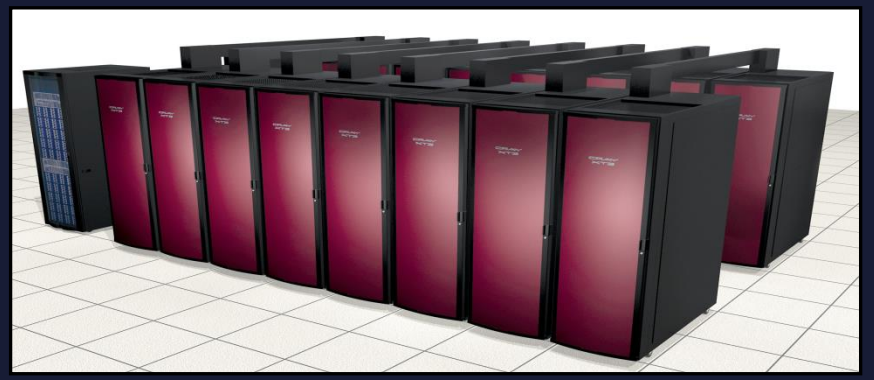
[2] Ian Foster, Carl Kesselman, and
e Virtual Organizations. Int.
J. High Perform. Comput. Appl. 15, 3

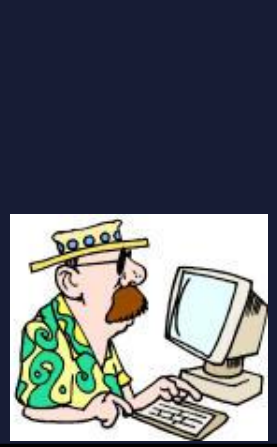
[3] What is the Grid? A Three Point Checklist. I. Foster, GRIDToday, July 20, 2002.

Grid: No centralized control

The user in general has full ownership of a desktop workstation.

A Cluster is a shared resource – Only the administrator has full control of the system
The physical layer is still well defined.





I submit my jobs to “the GRID” and they get processed: somehow, somewhere, after some time.

There is no GRID owner!

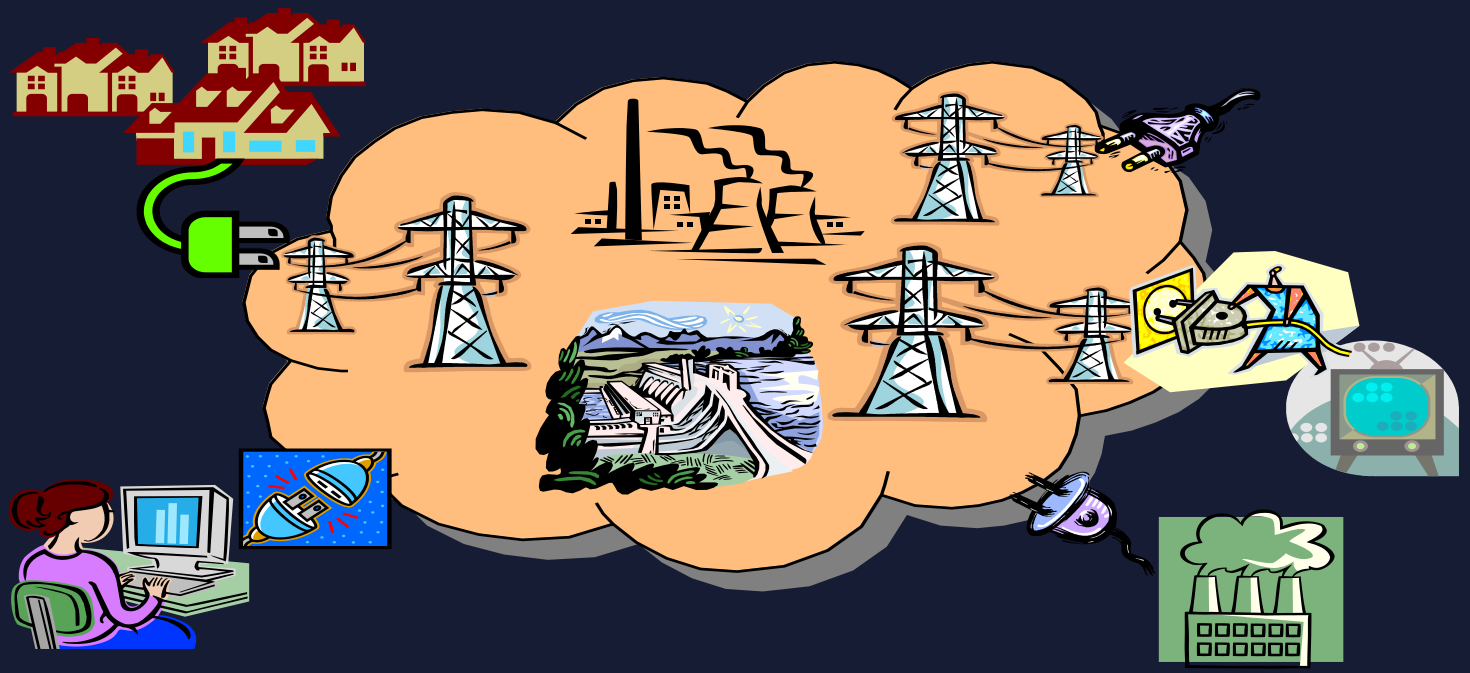
1st Law of the Grid

- **95% of the Grid is… agreement**

- Key terms
  - Coordination
  - No centralized control
  - Standards
  - Protocols
  - Interfaces

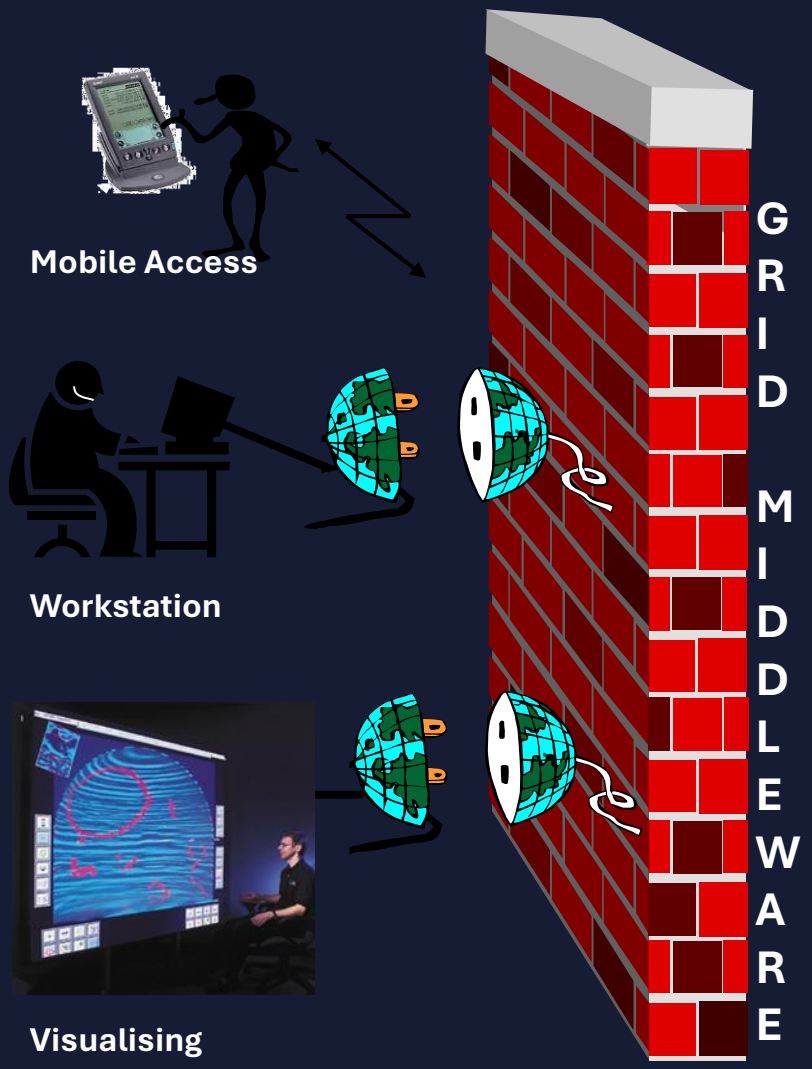
- Standards, protocols, interfaces,… aim at providing common abstractions of different implementations of similar services

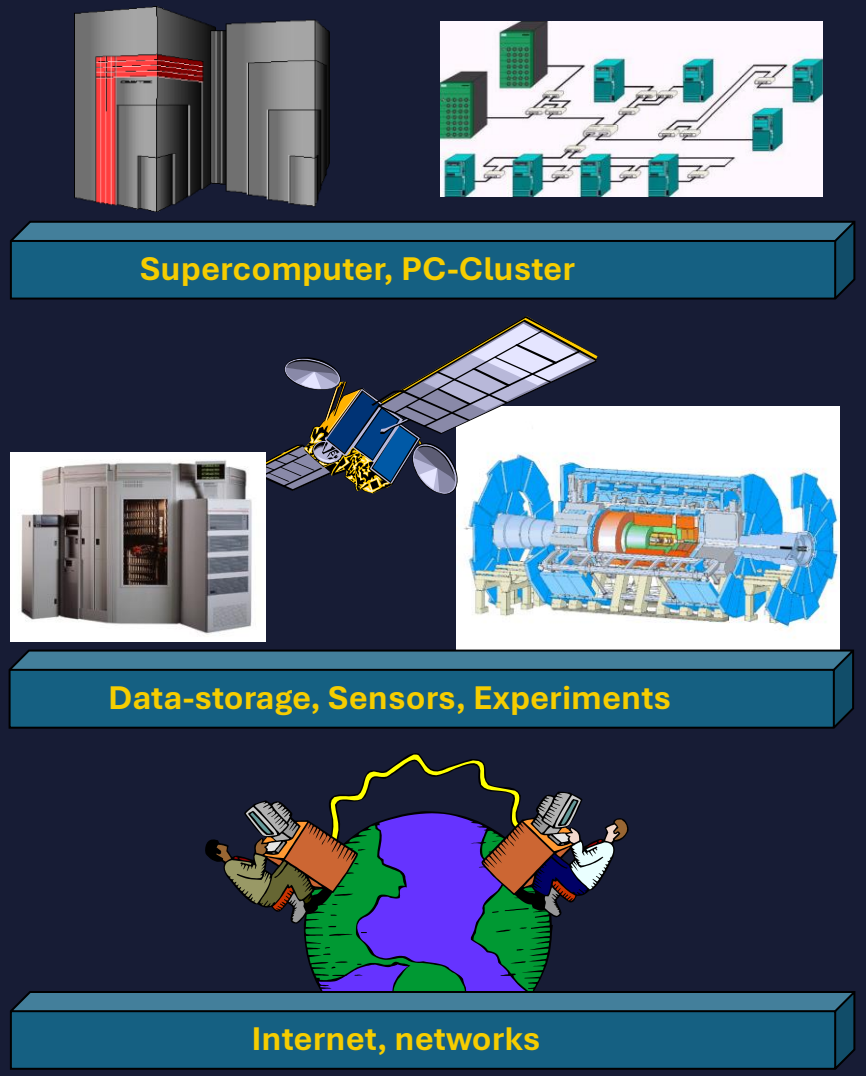
Power Grid Similarity



“We will probably see the spread of computer utilities, which, like present electric and telephone utilities, will service individual homes and offices across the country”
(Len Kleinrock, 1969)

The Grid Paradigm





Grid Types

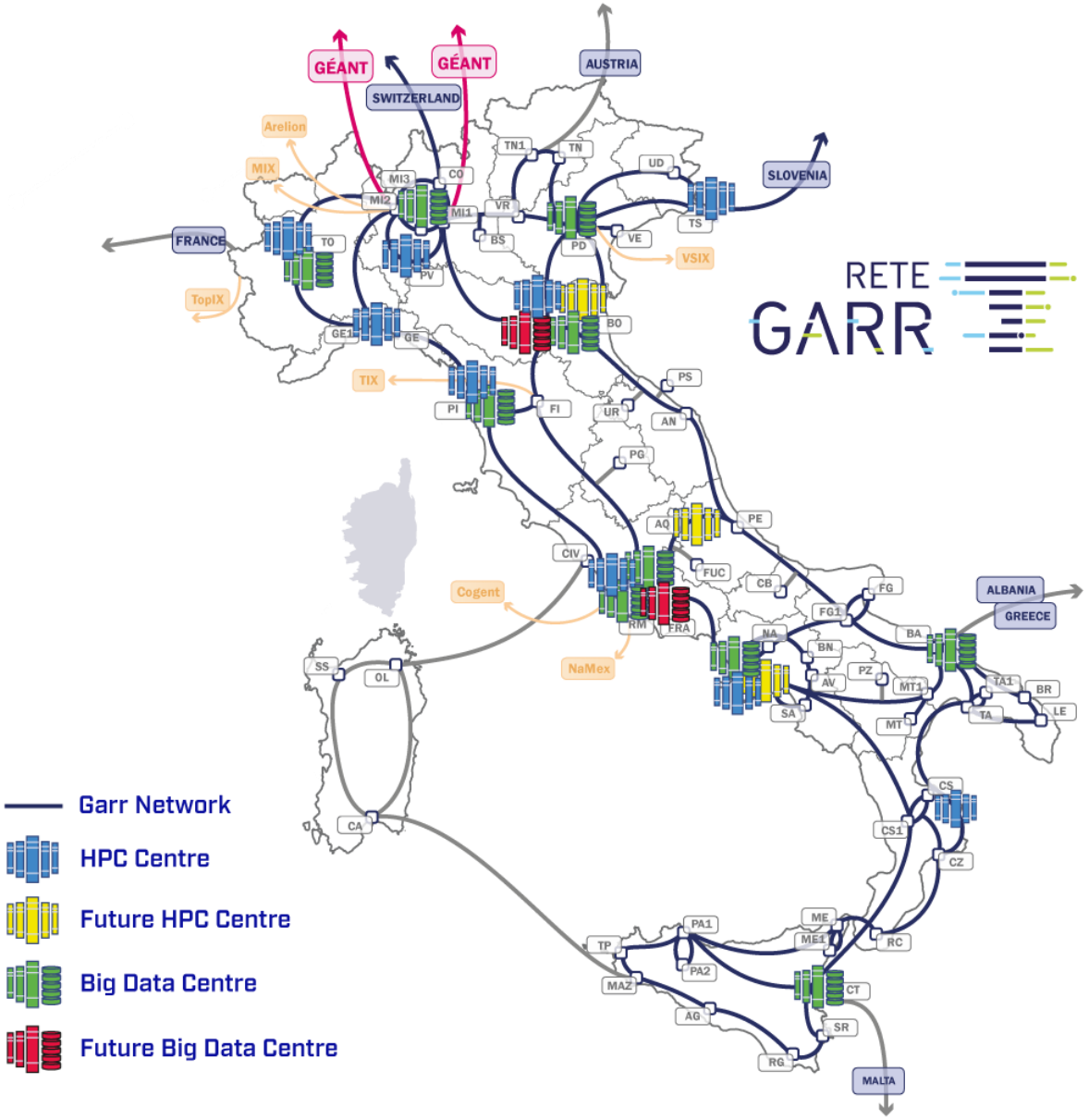
Supercomputer Based Service Grid

Cluster Based Service Grid

Volunteer Desktop Grid

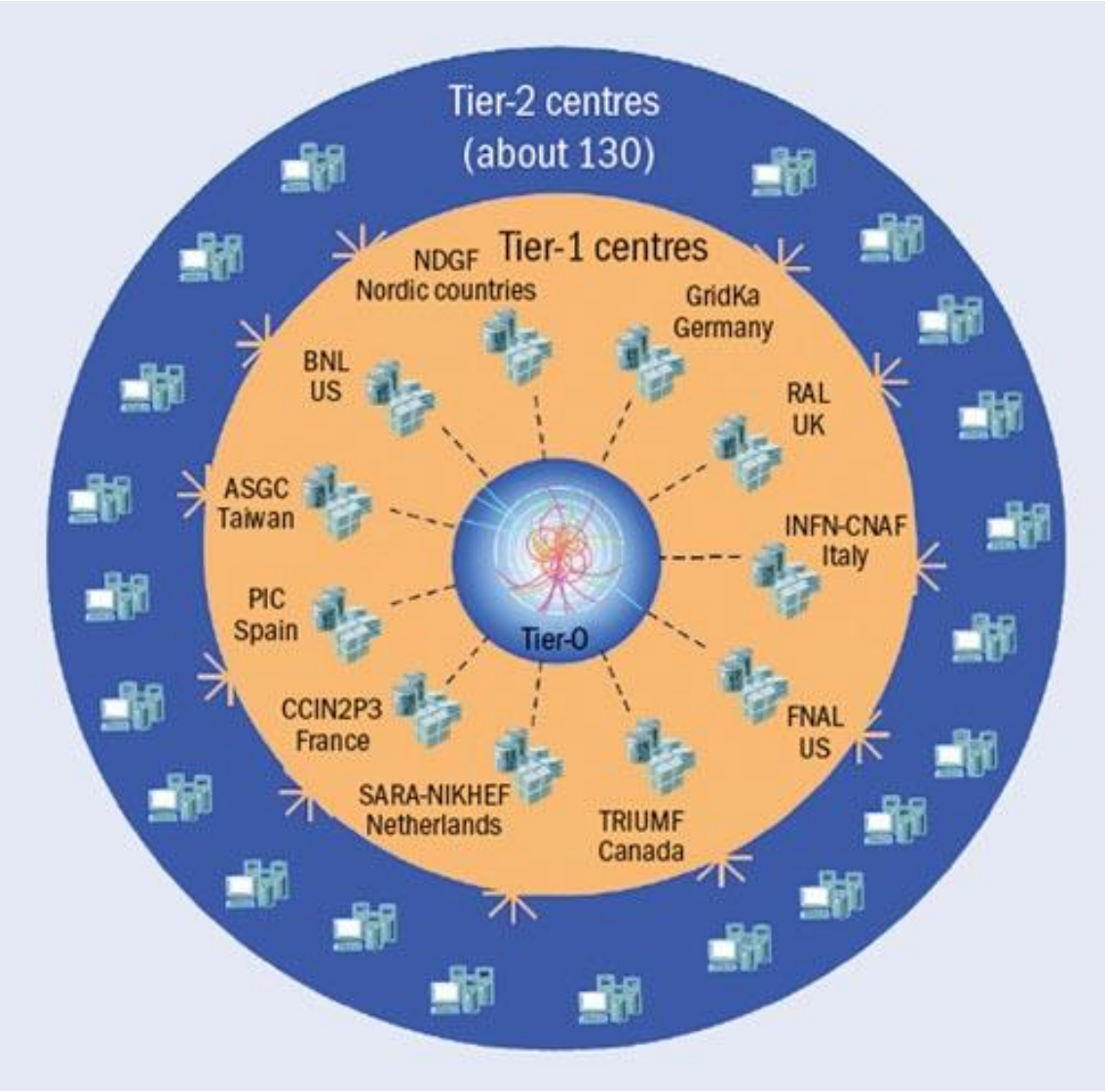
ICSC SPOKE 0
Infrastruttura
Cloud di supercalcolo

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

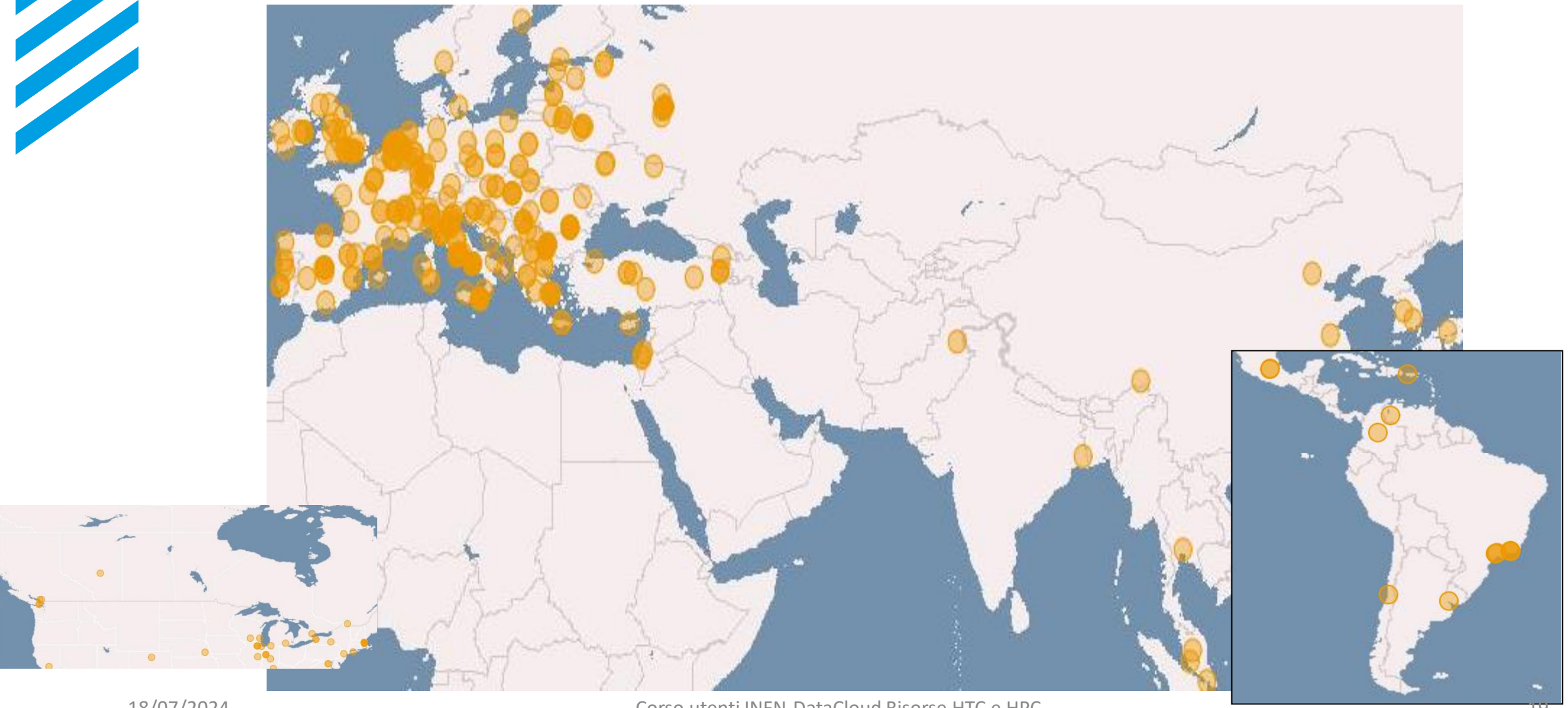


WLCG

- **Worldwide LHC computing Grid**
- Service GRID for the LHC high energy physics experiments
- Tiered structure
- Part of the European grid Infrastructure (EGI)
  - O(1M) logical CPUs
  - O(1) EB disk
  - O(1) EB tape



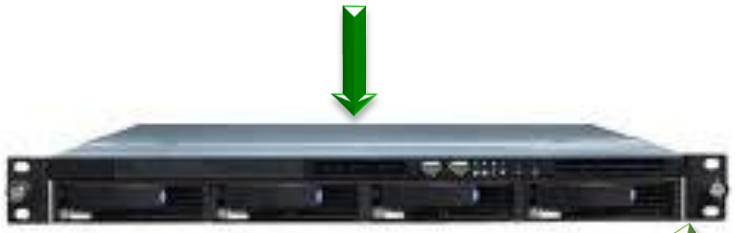
The European Grid Infrastructure



The European Grid Infrastructure

A typical Grid site

Batch System (PBS,LSF…)

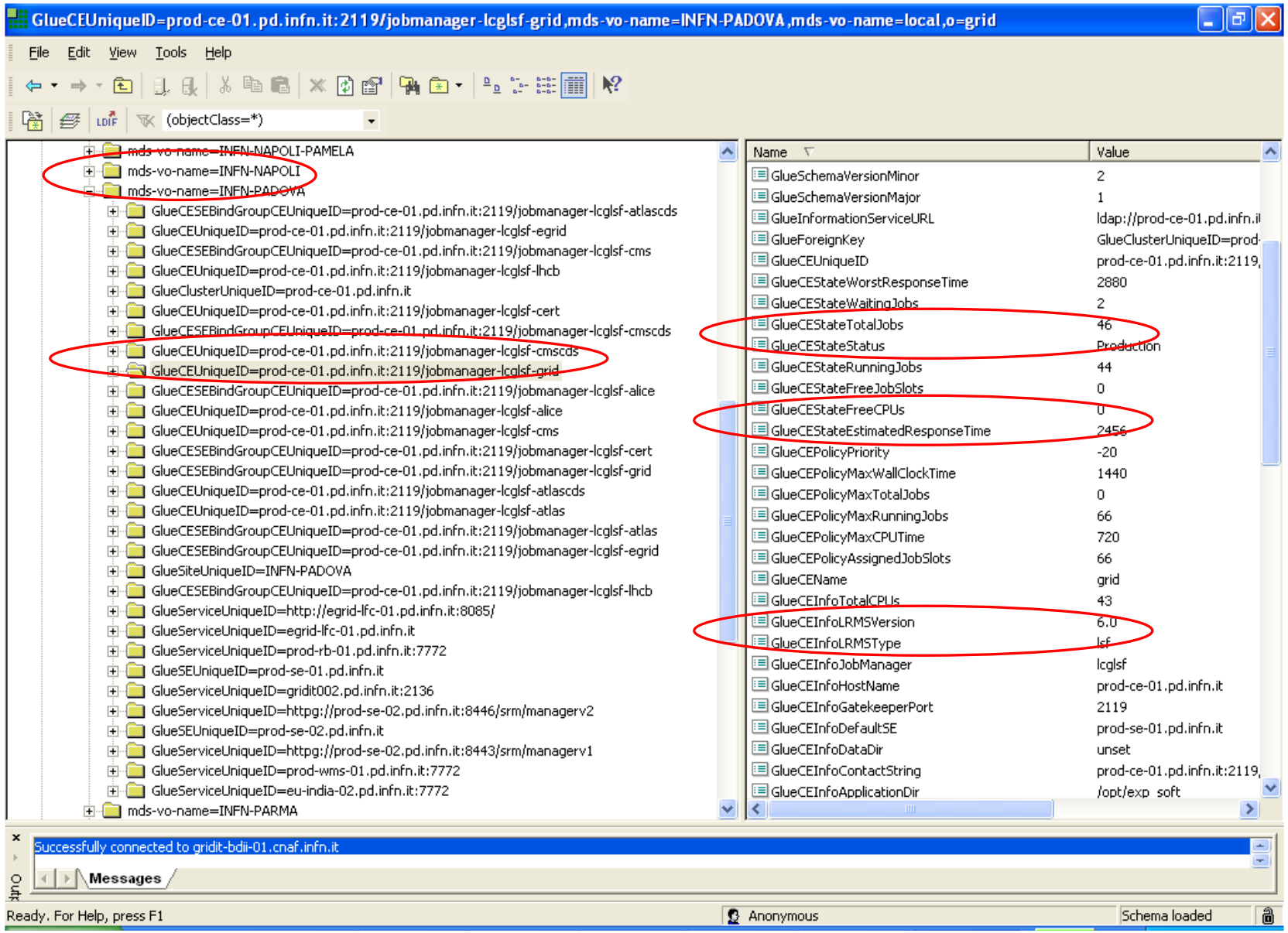


Computing Element (CE)

Storage Element (SE)

INFORMATION SYSTEM (es. BDII)

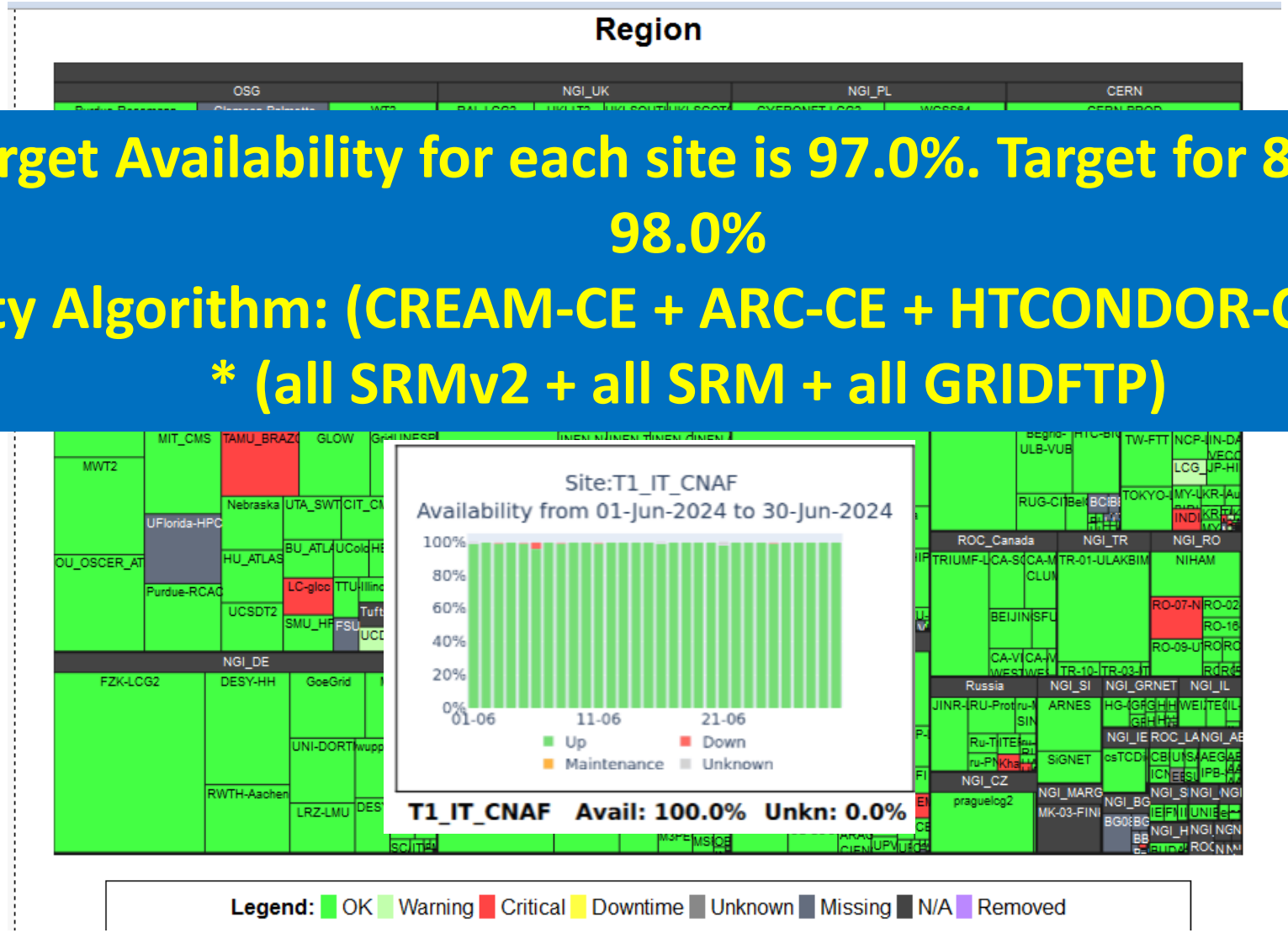
Standard Interface: The Information System



Operations Tools: Monitoring

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)



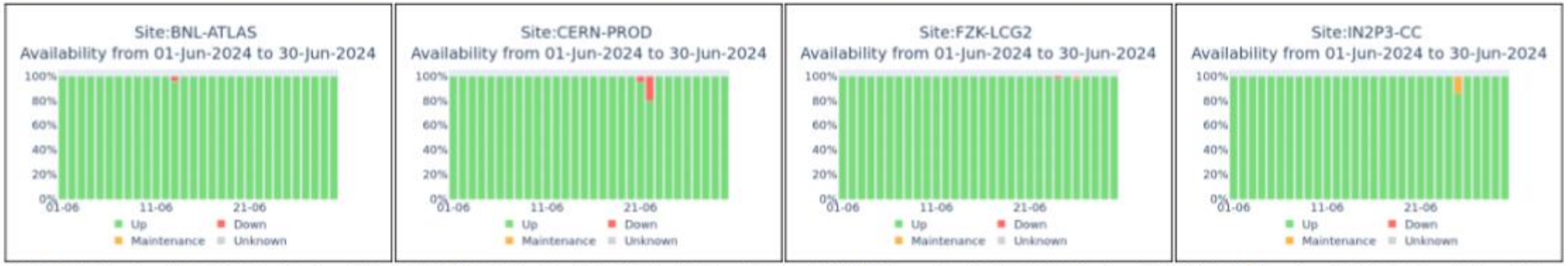
C

Availability of WLCG Tier-0 + Tier-1 Sites

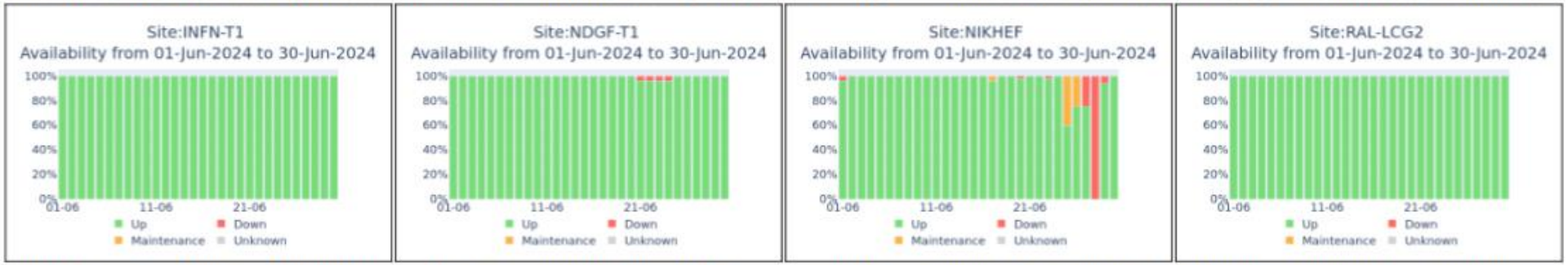
ATLAS **June 2024**

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)



BNL-ATLAS **Avail: 100.0%** **Unkn: 0%** | **CERN-PROD** **Avail: 99.0%** **Unkn: 0%** | **FZK-LCG2** **Avail: 100.0%** **Unkn: 0%** | **IN2P3-CC** **Avail: 100.0%** **Unkn: 0%**



INFN-T1 **Avail: 100.0%** **Unkn: 0.0%** | **NDGF-T1** **Avail: 99.0%** **Unkn: 0%** | **NIKHEF** **Avail: 93.0%** **Unkn: 0%** | **RAL-LCG2** **Avail: 100.0%** **Unkn: 0%**

RRC-KI-T1 **Avail: 100.0%** **Unkn: 0%** | **SARA-MATRIX** **Avail: 95.0%** **Unkn: 0.0%** | **TRIUMF-LCG2** **Avail: 100.0%** **Unkn: 0%** | **pic** **Avail: 100.0%** **Unkn: 0.0%**

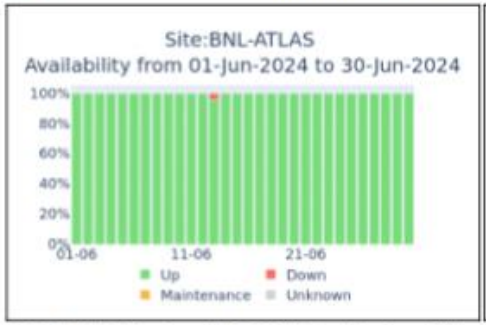
C

Availability of WLCG Tier-0 + Tier-1 Sites

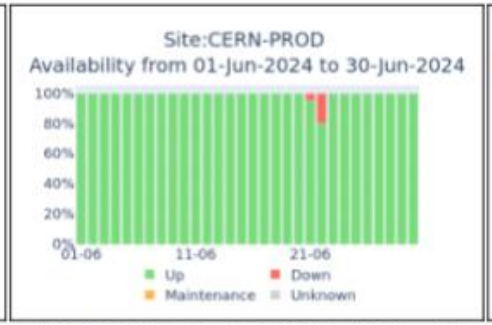
ATLAS **June 2024**

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)



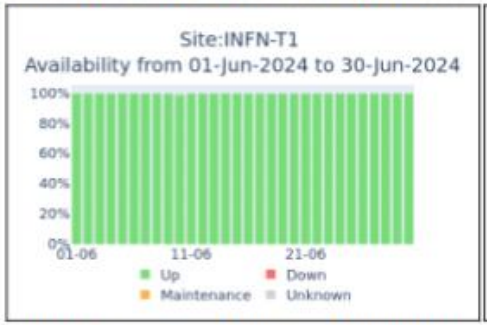
BNL-ATLAS Avail: 100.0% Unkn: 0%



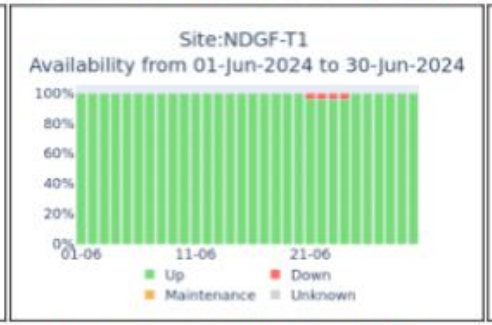
CERN-PROD Avail: 99.0% Unkn: 0%

FZK-LCG2 Avail: 100.0% Unkn: 0%

IN2P3-CC Avail: 100.0% Unkn: 0%



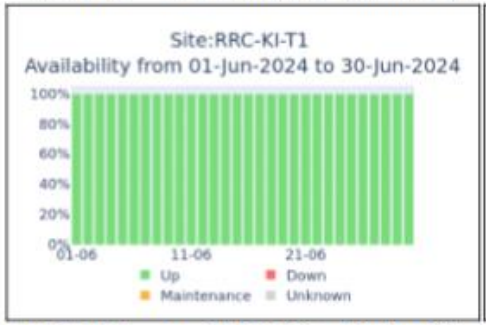
INFN-T1 Avail: 100.0% Unkn: 0.0%



NDGF-T1 Avail: 99.0% Unkn: 0%

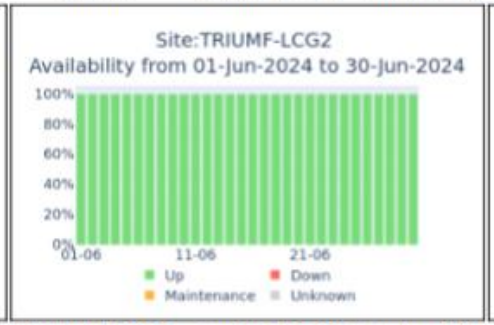
NIKHEF Avail: 93.0% Unkn: 0%

RAL-LCG2 Avail: 100.0% Unkn: 0%

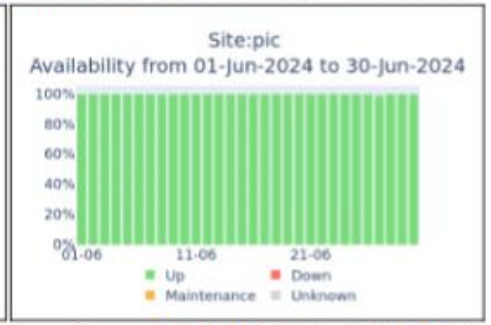


RRC-KI-T1 Avail: 100.0% Unkn: 0%

SARA-MATRIX Avail: 95.0% Unkn: 0.0%



TRIUMF-LCG2 Avail: 100.0% Unkn: 0%



pic Avail: 100.0% Unkn: 0.0%

Operations Tools: Monitoring

Tier-2 Availability and Reliability Report

ATLAS **June 2024**

Federation Summary - Sorted by Availability

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

Color coding: N/A <30% <60% <90% >=90%

Federation	Availability	Reliability
PL-POLISH-WLCG	100%	100%
IL-HEPTier-2	100%	100%
US-NET2	100%	100%
CN-IHEP	100%	100%
HK-ATLAS-T2	100%	100%
PT-LIP-LCG-Tier2	100%	100%
CA-EAST-T2	100%	100%
FR-IN2P3-CPPM	100%	100%
FR-IN2P3-LAPP	99%	99%
JP-Tokyo-ATLAS-T2	99%	100%
CH-ATLAS	99%	99%
SI-SiGNET	99%	99%
US-MWT2	99%	99%
US-AGLT2	99%	99%
FR-IN2P3-LPC	98%	98%
RU-RDIG	98%	98%
FR-GRIF	97%	97%
IT-INFN-T2	97%	97%

Federation	Availability	Reliability
UK-London-Tier2	97%	97%
ES-ATLAS-T2	97%	98%
T2-LATINAMERICA	96%	96%
DE-FREIBURGWUPPERTAL	95%	100%
CA-WEST-T2	94%	98%
DE-DESY-ATLAS-T2	94%	98%
RO-LCG	91%	91%
UK-SouthGrid	89%	96%
DE-DESY-GOE-ATLAS-T2	88%	91%
US-SWT2	88%	88%
SK-Tier2-Federation	86%	86%
DE-MCAT	85%	85%
CZ-Prague-T2	81%	96%
UK-NorthGrid	75%	75%
TR-Tier2-federation	70%	73%
UK-ScotGrid	22%	28%
CH-CHIPP-CSCS	1%	3%
SE-SNIC-T2	0%	0%

Perché infrastrutture federate

- Non è una questione di ottenere una infrastruttura più grande da una serie di centri di calcolo
- Valore aggiunto nella necessità di avvicinare le comunità
- Supporto a comunità internazionali e distribuite per natura
  - Creazione di Virtual Organization
- Failover e disaster recovery

Generated with AI · May 2024

2nd Law of the Grid

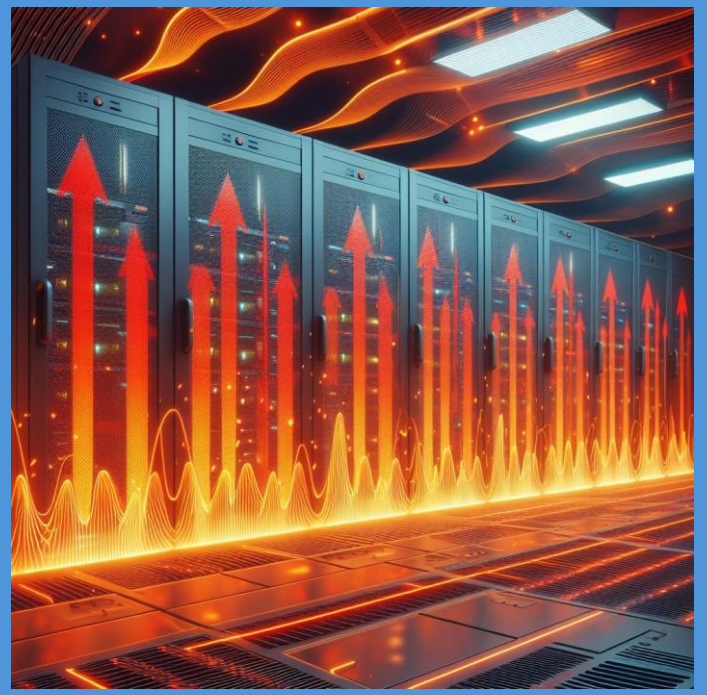
Anything that can go wrong, will

- The Grid is a very complex environment, errors will happen
- Some errors are preventable, some are manageable by the infrastructure, some can only be managed by the user

“A distributed system is one in which the failure of a computer you didn‘t even know existed can render your own computer unusable.”
Leslie Lamport

Expect the unexpected

> *When the Air-conditioning / power fails (again & again & again);*



Generated with AI · May 2024

©Jamie Shiers 2008 J. Phys.: Conf. Ser. 119 052030 – Lessons Learnt from WLCG Service Deployment

Expect the unexpected

> *When a service engineer puts a Coke into a machine to 'warm it up';*

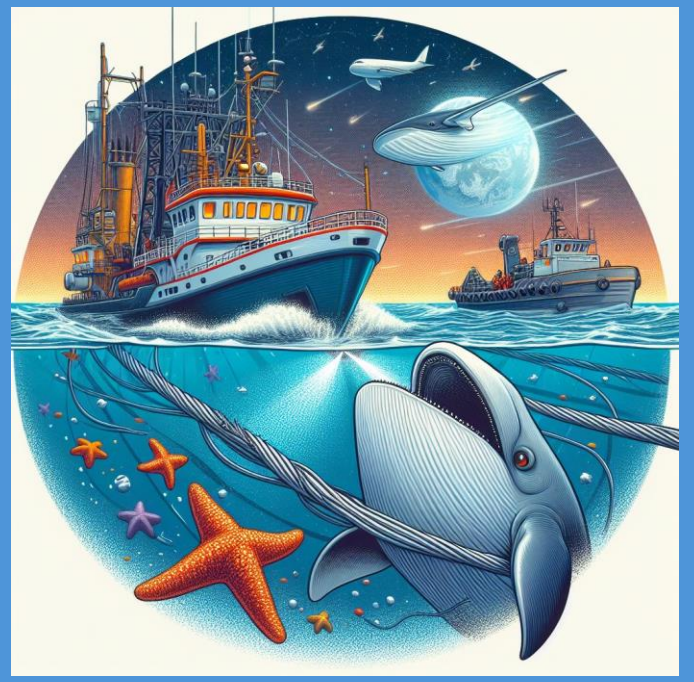
©Jamie Shiers 2008 J. Phys.: Conf. Ser. 119 052030 – Lessons Learnt from WLCG Service Deployment

Generated with AI · May 2024

Expect the unexpected

When a fishing trawler cuts a trans-Atlantic network cable;

©Jamie Shiers 2008 J. Phys.: Conf. Ser. 119 052030 – Lessons Learnt from WLCG Service Deployment



Generated with AI · May 2024

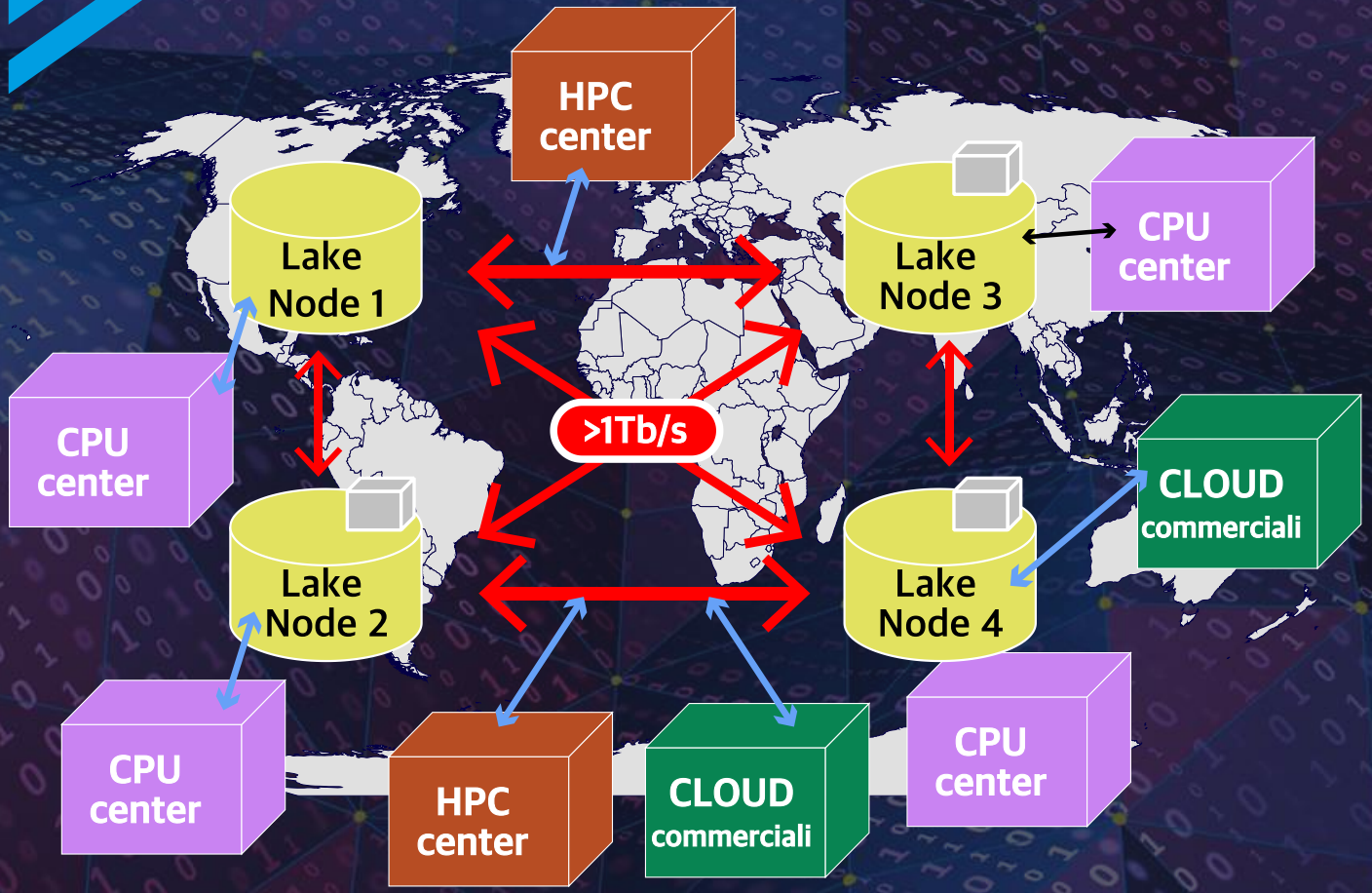
Expect the unexpected

> *When a Tsunami does the equivalent in Asia Pacific;*

Generated with AI · May 2024

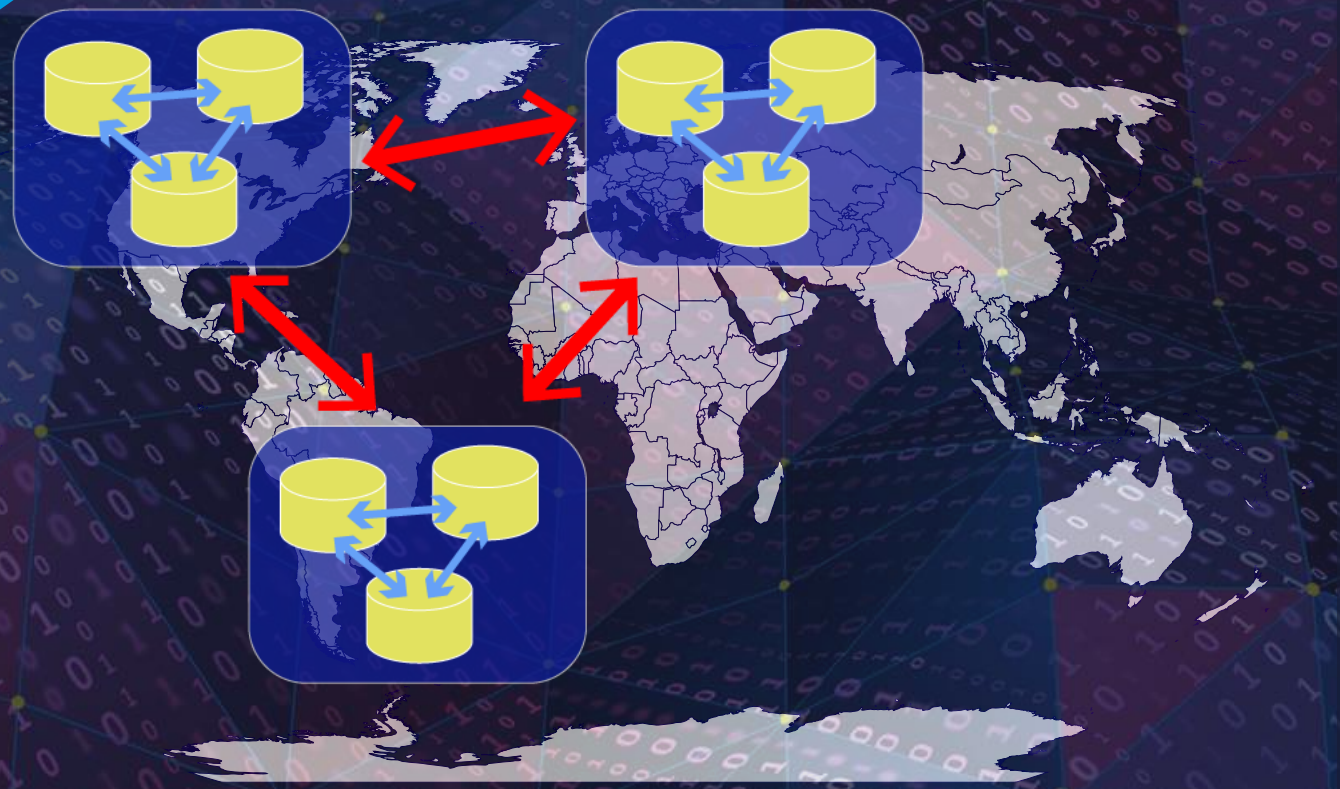
©Jamie Shiers 2008 J. Phys.: Conf. Ser. 119 052030 – Lessons Learnt from WLCG Service Deployment

Il **datalake** scientifico e la sua evoluzione



- Struttura logicamente singola
  - Interconnessa tramite reti ad alta banda multi-Tbps “inter-lake”
- Disaccoppiamento completo degli aspetti di storage e calcolo
- Elevata disponibilità e sicurezza del dato in centri dedicati
  - Minimizzazione del numero di copie
- CPU/GPU: utilizzate dovunque si trovino agganciandole al «Lake»
- Il datalake sfruttabile da altre comunità nazionali ed internazionali

Il **datalake** scientifico e la sua evoluzione



- Struttura logicamente singola
  - Interconnessa tramite reti ad alta banda “inter-lake”
- Disaccoppiamento completo degli aspetti di storage e calcolo
- Elevata disponibilità e sicurezza del dato in centri dedicati
  - Minimizzazione delle copie
- CPU/GPU: utilizzate dovunque si trovino agganciandole al «Lake»
- Il datalake sfruttabile da altre comunità nazionali ed internazionali

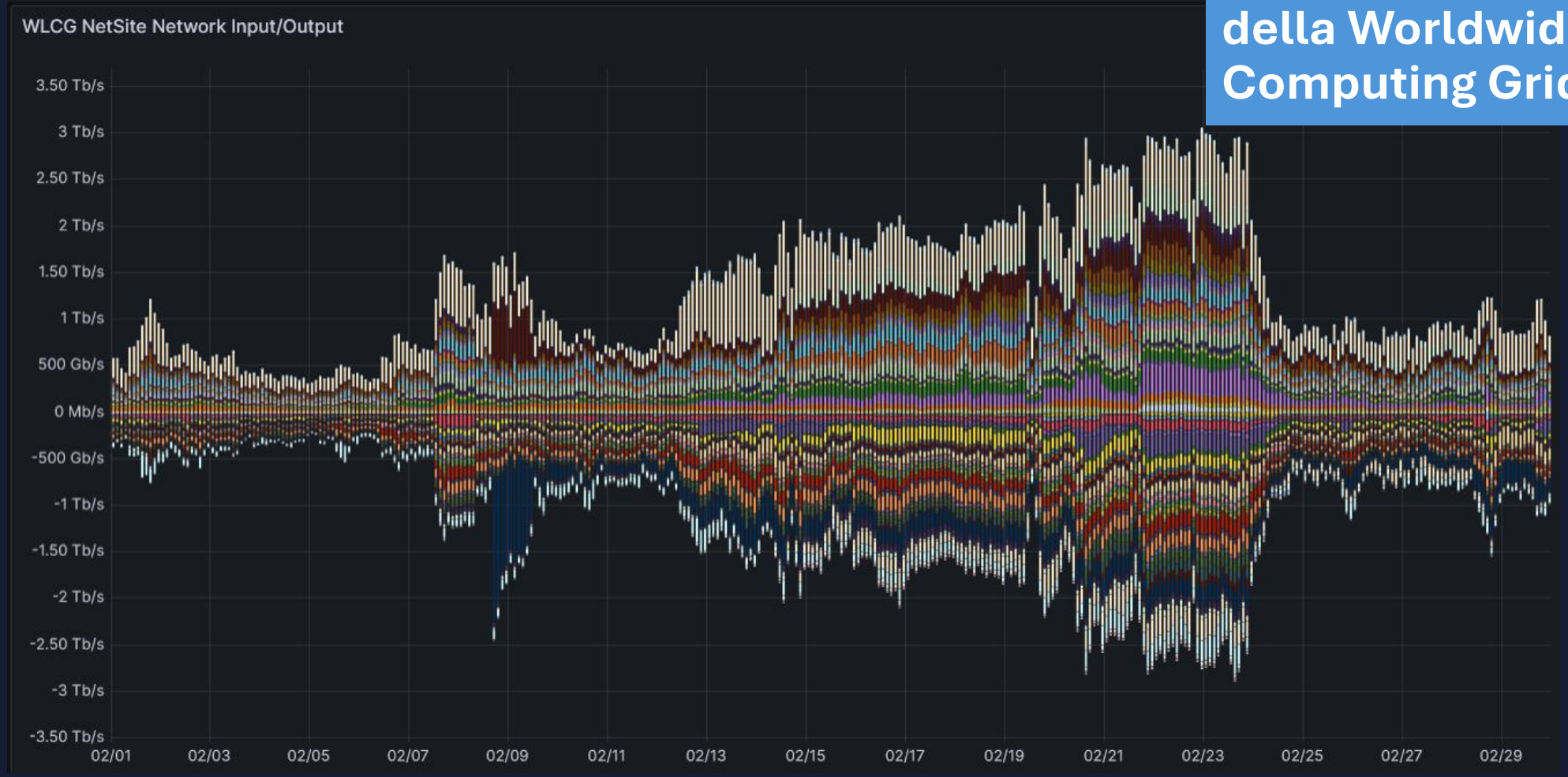
Generated with AI · May 2024

Come dovremmo fare il **Datalake <u>secondo</u> l'Intelligenza Artificiale**

Draw ***«a worldwide datalake with high speed network connections»***

Lo sviluppo dei sistemi di gestione dei dati in un mondo data-intensive

La strada verso la realizzazione del Data Lake
Il Data Challenge 2024 della Worldwide LHC Computing Grid



La gestione dei dati in un mondo sempre più **data-intensive**

Data challenge al CNAF effettuato attraverso il nuovo link ottico diretto con il CERN

- Estensione diretta delle rispettive reti su spettro condiviso multi-dominio
- Alta banda: attualmente 400 Gbps, fino a 1.6 Tbps; Latenza: 9.5s

289 Gbps

Resources@Tier1

Le risorse PLEDGED

PLEDGE @T1 2023-2024

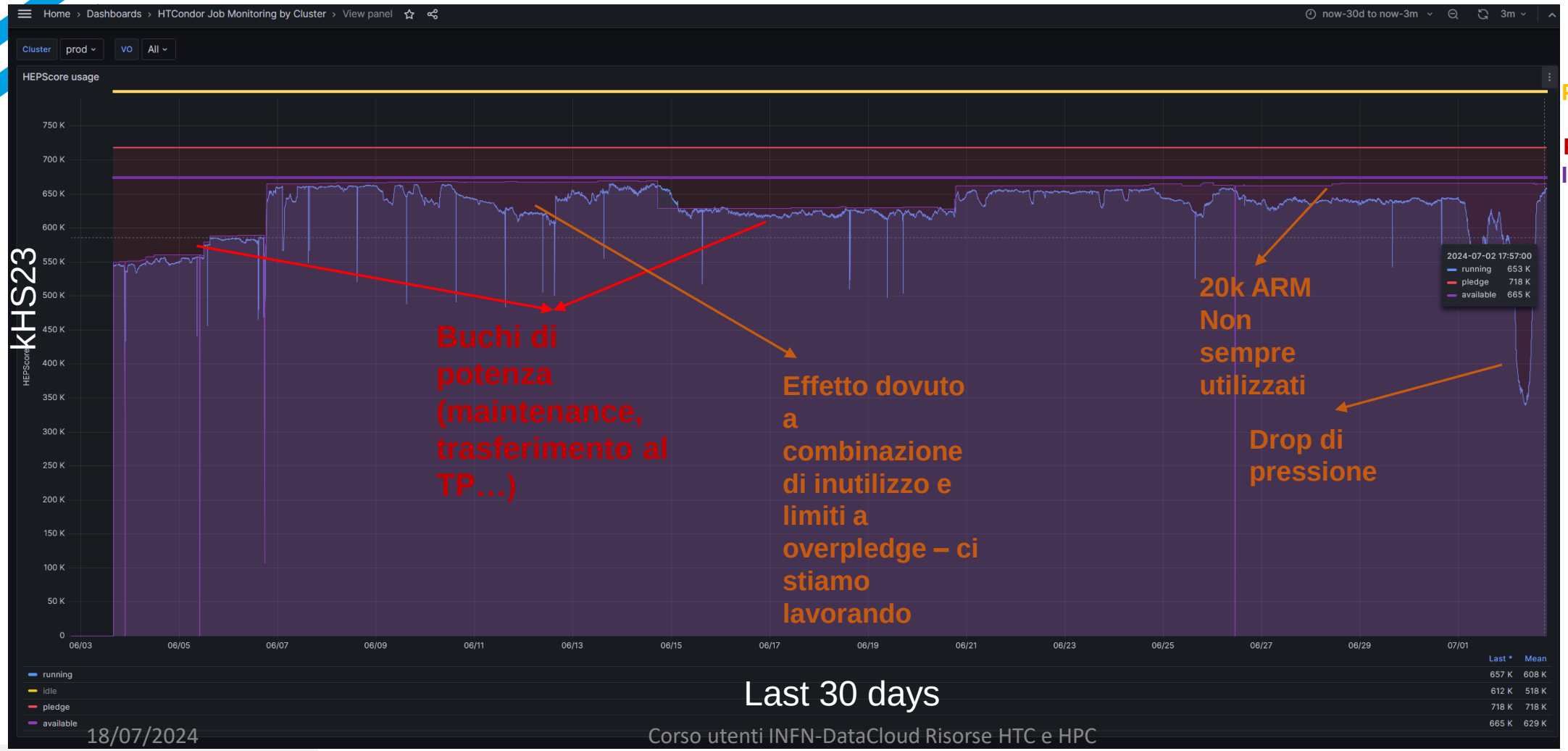
	1/4/2023			1/1/2024		
Experiment	**CPU**	**DISK**	**TAPE**	**CPU**	**DISK**	**TAPE**
ALICE	102960	11430	24600	102960	11430	24600
ATLAS	128700	12240	31770	128700	12240	31770
CMS	104000	12740	41080	104000	12740	41080
LHCB	113430	11561	25261	113430	11561	25261
LHC TIER1	**449090**	**47971**	**122711**	**449090**	**47971**	**122711**
Belle2	27000	1020	650	31000	1320	650
CDF	0	0	4000	0	0	4000
DUNE_CSN1	0	0	0	5000	1100	510
FCC	1000	100	0	1000	200	0
GMINUS	0	0	0	640	160	1200
ICAR-US	0	0	0	4000	1600	4000
KLOE	0	33	3075	0	33	3075
LHCB TIER2	62600	0	0	62600	0	0
LHCf	12000	120	0	12000	170	0
MuonCollider	3000	150	150	3000	150	150
MUONE	0	0	0	1000	100	650
NA62	3300	275	3300	3300	275	3300
PADME	4417	100	1780	4417	100	1780
Gruppo 1	**113317**	**1798**	**12955**	**127957**	**5208**	**19315**
AMS2	27833	2800	1550	31833	2800	1650
AUGER	4633	1000	300	5430	1100	300
Borexino	2069	359	80	500	359	80
BULLKID	0	0	0	125	10	0
CTA	5296	1865	700	5296	2000	2700
CUORE	3942	750	0	3000	850	100
CUPID	905	25	10	905	25	10
CYGNO	417	40	50	417	40	200
DAMPE	27306	800	200	27306	850	10
DARKSIDE	4917	2340	1920	6000	3150	1920
DUNE	3500	510	510	0	0	0
ENUBET	500	10	5	250	10	5
ET	100	50	0	500	55	0
EUCLID	0	1450	1000	0	1450	1000
FERMI-GLAST	833	15	40	350	15	20
GAPS	1433	80	0	1863	80	0
Gerda	40	70	100	40	50	50
Herd	7111	300	0	8111	450	0
hyperk	10838	152	605	10838	152	605
Icarus	4000	1300	3000	0	0	0
JUNO	19183	2100	1000	19983	3000	1000
KM3	3000	450	250	7500	450	250
LIMADOU	2147	180	10	2300	180	12
LITEBIRD	0	55	205	0	0	0
LSPE	0	21	0	0	0	0
MAGIC	0	0	0	0	0	30
NEWS	284	300	150	284	300	150
NUCLEUS	1000	170	83	500	170	83
PAMELA	747	110	150	747	110	150
QUAX	100	130	120	0	0	120
QUBIC	300	25	25	0	0	0
SPB2_MiniEUSO	200	10	0	200	10	0
SWGO	210	150	0	400	300	0
Tristan	1200	40	0	2000	40	0
Virgo	50000	856	4168	50000	900	5158
Xenon100	1250	500	3500	1250	600	3500
Gruppo 2	**185294**	**19013**	**19731**	**187928**	**19506**	**19103**

Resources PLEDGE @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

CPU Farm at T1

Pledge 2024: 703kHS06 (w/o OVERLAP- 843k) ➔ **CNAF TOTAL PLAN 792kHS06 - Potenza installata Totale: 665KHS06**
Ultima Gara installata: CPU 2022 in Mar/Apr 2023 ➔ Nodi Leonardo non ancora disponibili
Tutta la Farm ad HTCondor 23

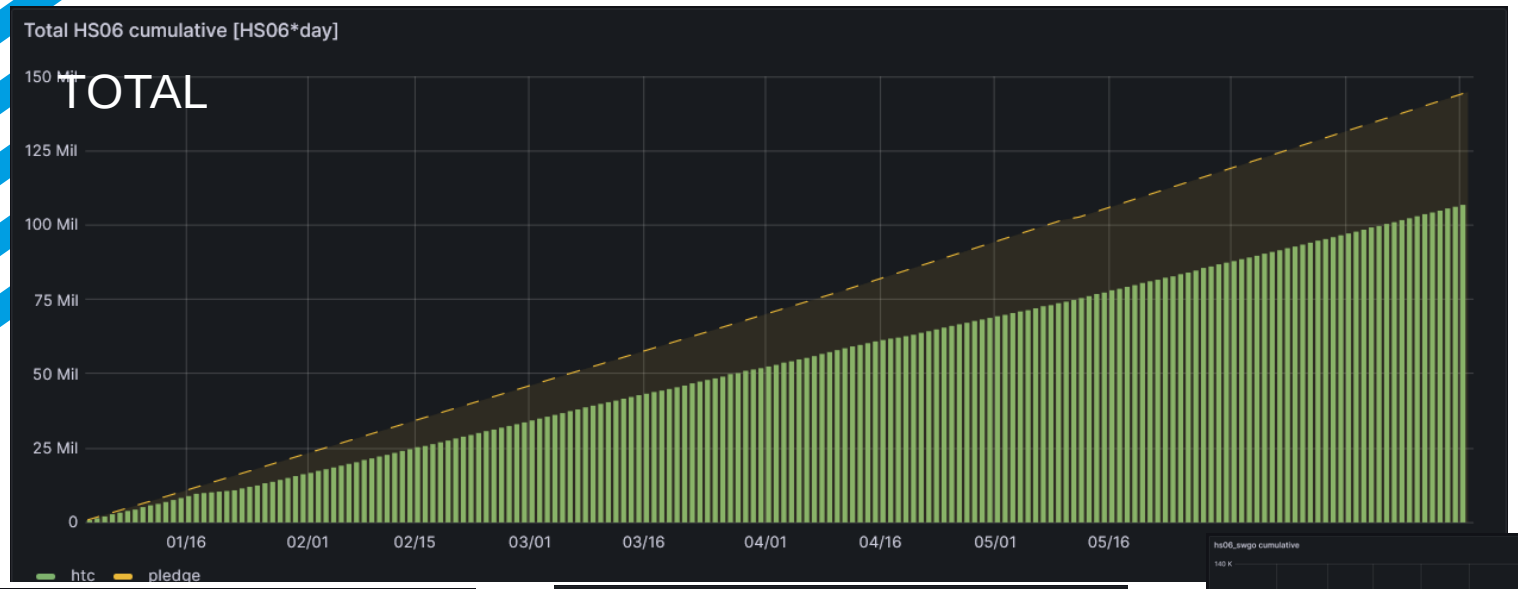


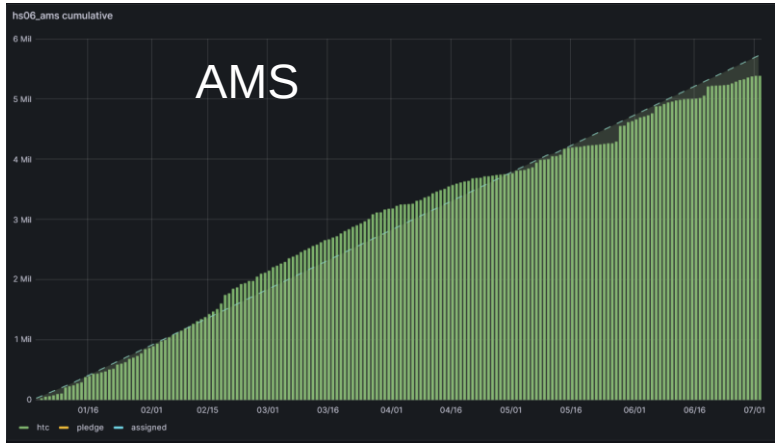
PLEDGE PLAN 792KH

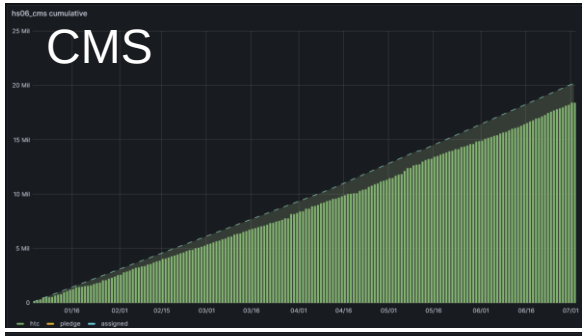
PLEDGE with OF 703I
Installed 665KHS

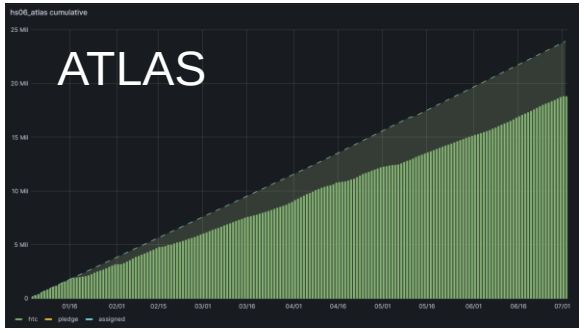
Last 30 days

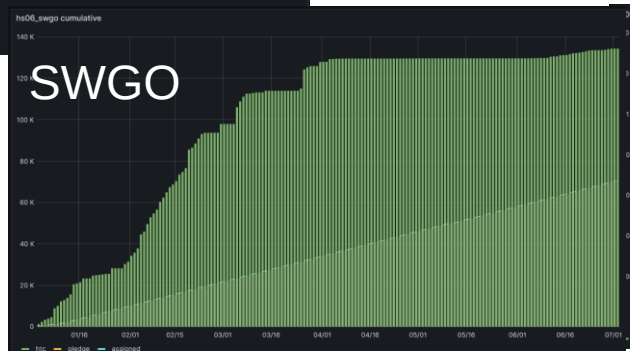
HS06 integrati

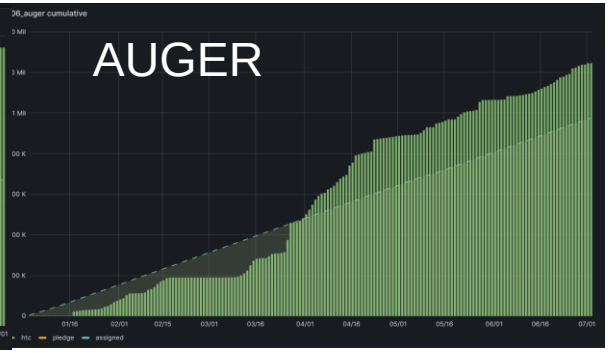


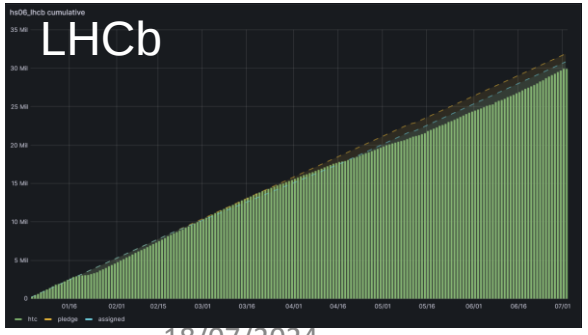


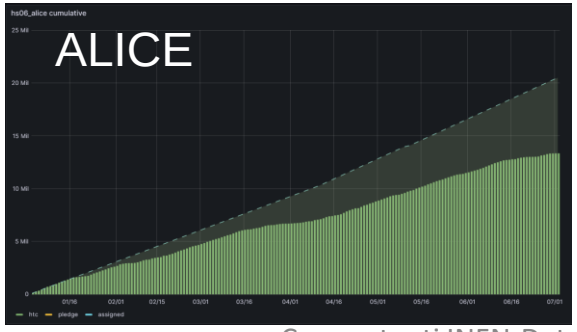


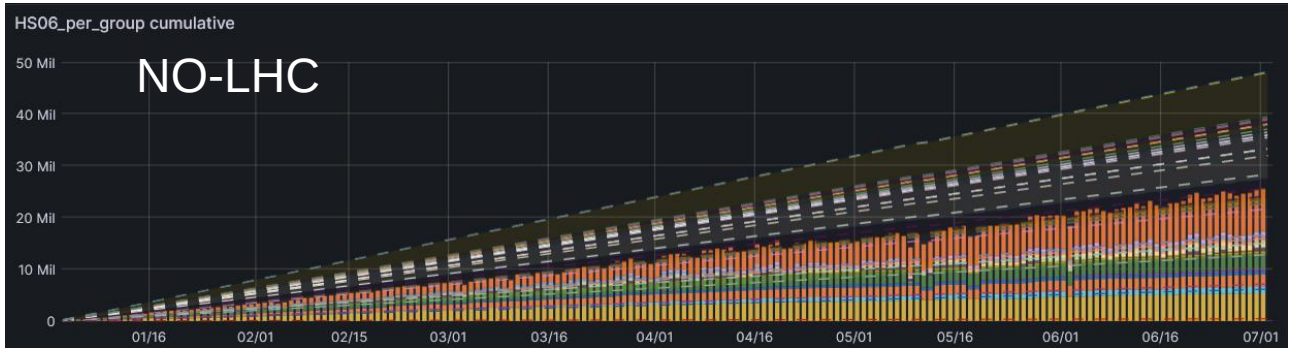




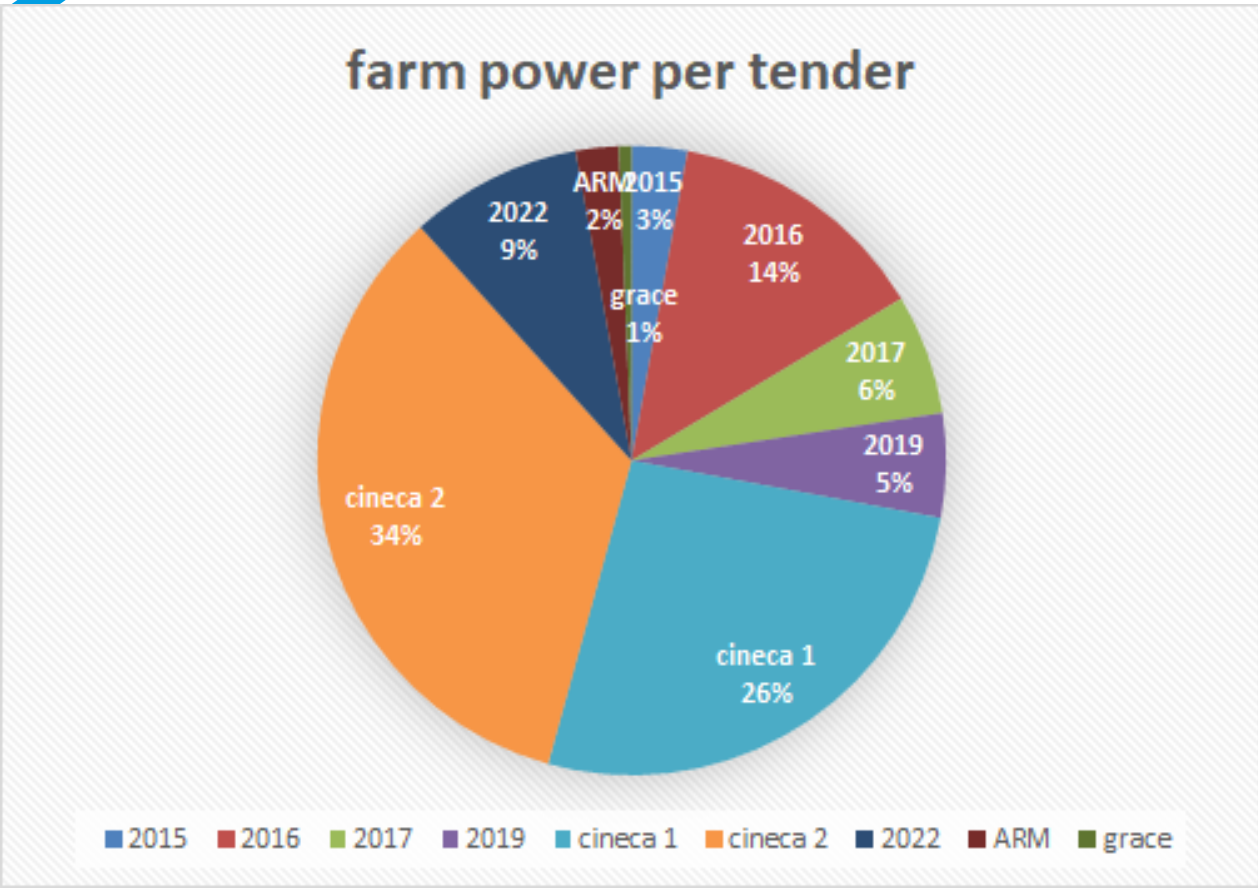








Composizione della farm al T1

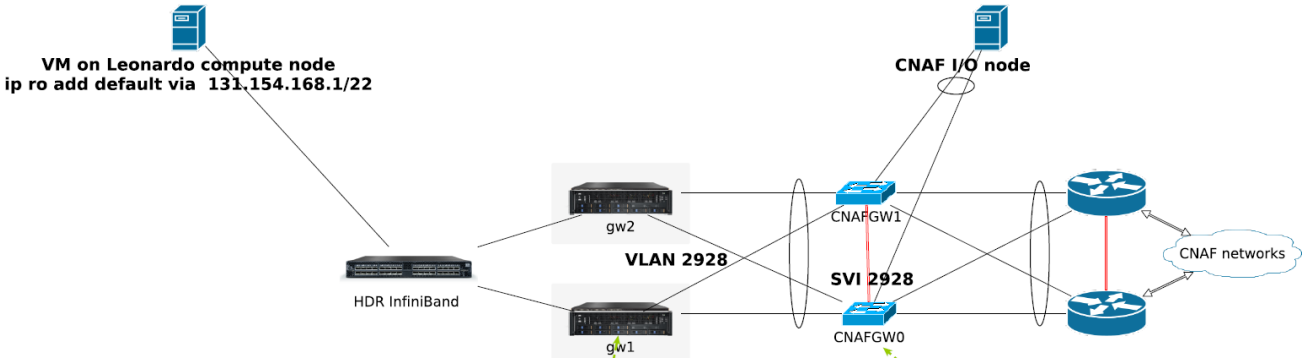


- Manca ancora Leonardo (max 200 nodi)
  - 2880 HS06/nodo
- Aggiunti i nodi ARM
  - Ampere (4 nodi)
    - 3754HS06/nodo
    - 3.74 hs06/W (vs 2.64 HS06/W gara2022)
  - Grace (1 nodo)
    - 4459 HS06/nodo
    - 4.67 HS06/W (vs 2.64 HS06/W gara2022)
  - Solo ATLAS ha dato l'ok per considerare nei pledge 2025+ fino al 30% di risorse su ARM

CPU in 2024 – Leonardo integration

- No direct CPU acquisition in 2024
- We will use up to 200 Leonardo-GP@CINECA nodes
  - Dual 56 cores sockets Intel Sapphire Rapids
  - **2800 HS06/nodo**

- 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

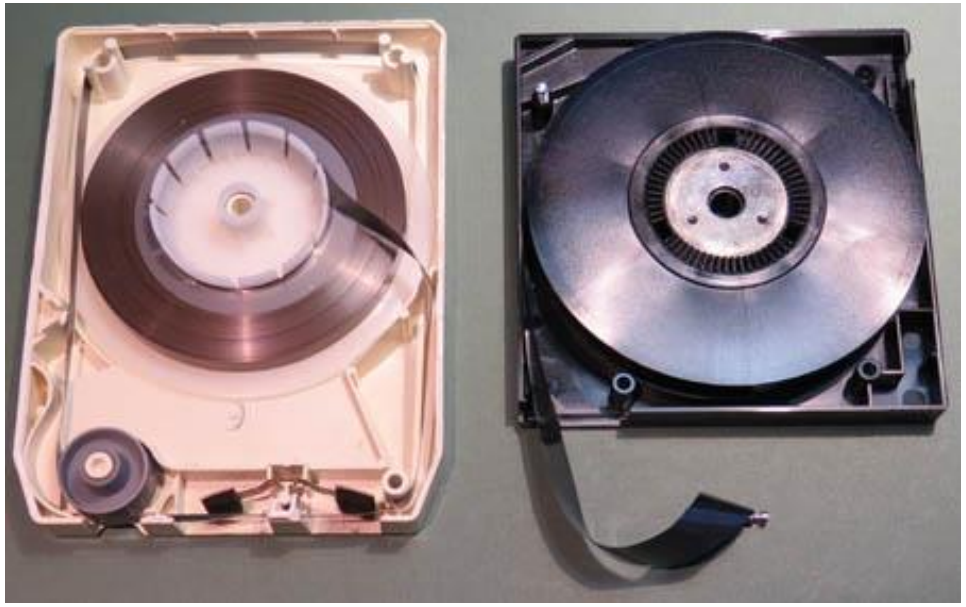
DISK and TAPE al T1

DISK USAGE @T1

TAPE@T1


Total usage - all (LHC + no-LHC)
200 PB
150 PB
100 PB
50 PB
0 GB
06/23
07/01
07/08
07/16
current
used
133.16 PB
installed
189.85 PB

TAPE devices



Back to the 80s

Commodore’s datassette: a 90-minutes tape (45 minutes on each side) will hold on the order of 150 kilobytes on each side if no compression or fast loader is used.

Back to the 80s

Zak McKracken and the Alien Mindbenders
Released: 1988 (36 years ago)
Publisher: Lucasfilm GamesInfo / Logos
Coder: David Fox
Matthew Kane
Graphics: Gary Winnick
Martin Cameron
Musician: Matthew Kane
Sound FX: Chris GriggInfo
Matthew Kane
Box Art: Steve Purcell

Bubble Bobble

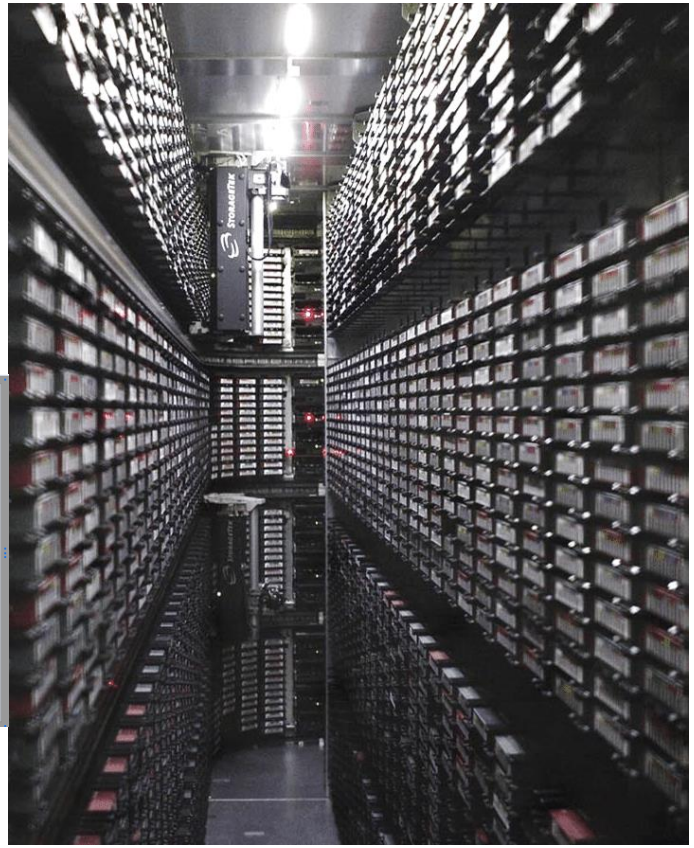
Published: 1987, Firebird
Category: Platformer - Single Screen
Players: 1 or 2, Simultaneous

Turrican II: The Final Fight
Published: 1991, Rainbow Arts
Category: Shoot'em Up - Platformer
Players: 1 Only

Tape area network

- Is the part of the SAN dedicated to the interconnection among servers, libraries and tape drives
- Tape drives can be installed in a central array and attached to the SAN, making them accessible to every server on the network

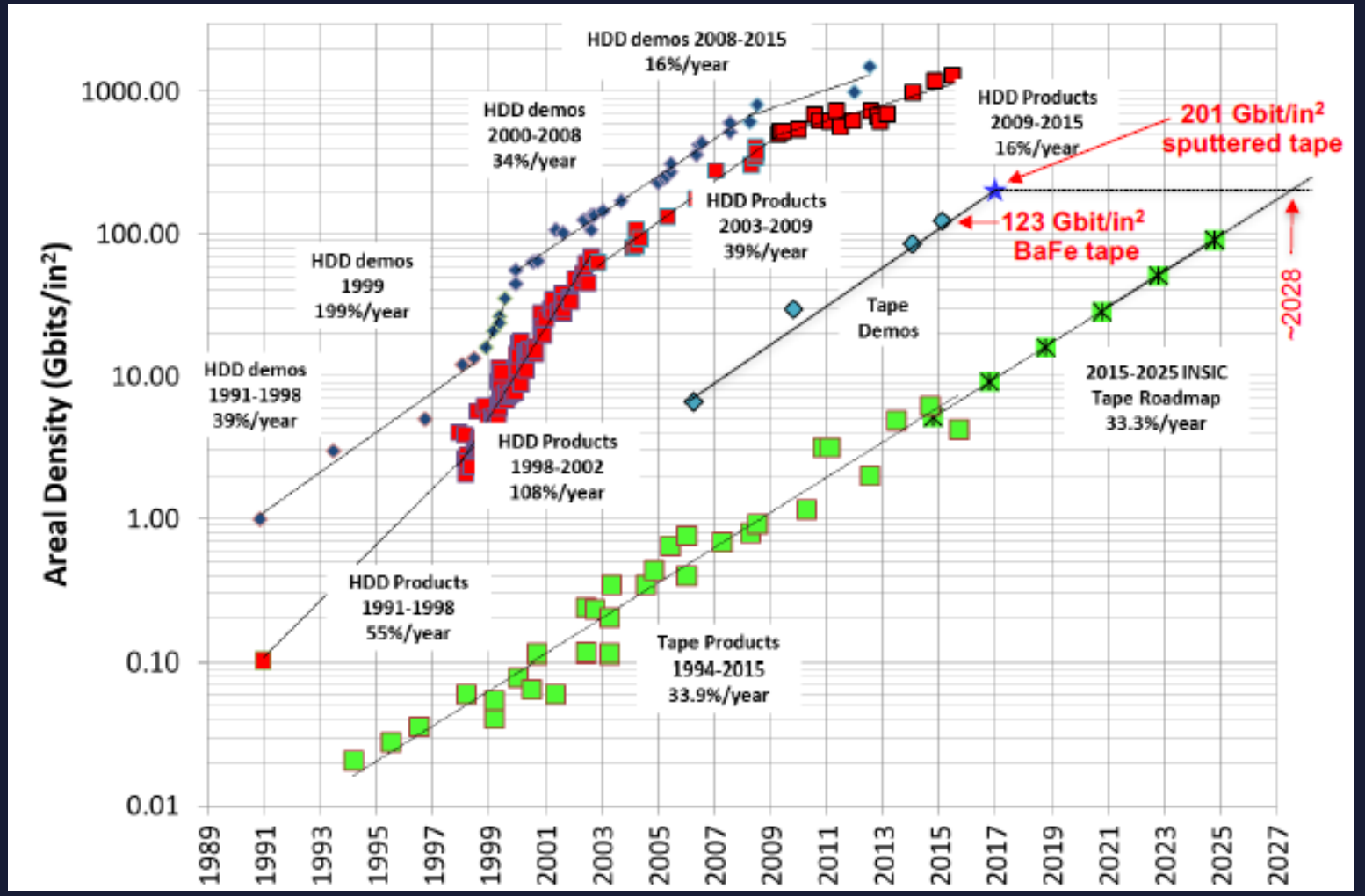
Tape Libraries at CNAF



Library	Tape drives	Max data rate/drive, MB/s	Max slots	Max tape capacity, TB	Installed cartridges	Used space, PB	Free space, PB
SL8500 (Oracle)	16*T10KD	250	10000	8.4	~10000	**36**	-
TS4500 (IBM)	19*TS1160	400	6198	20	5104	**89**	**4.6**

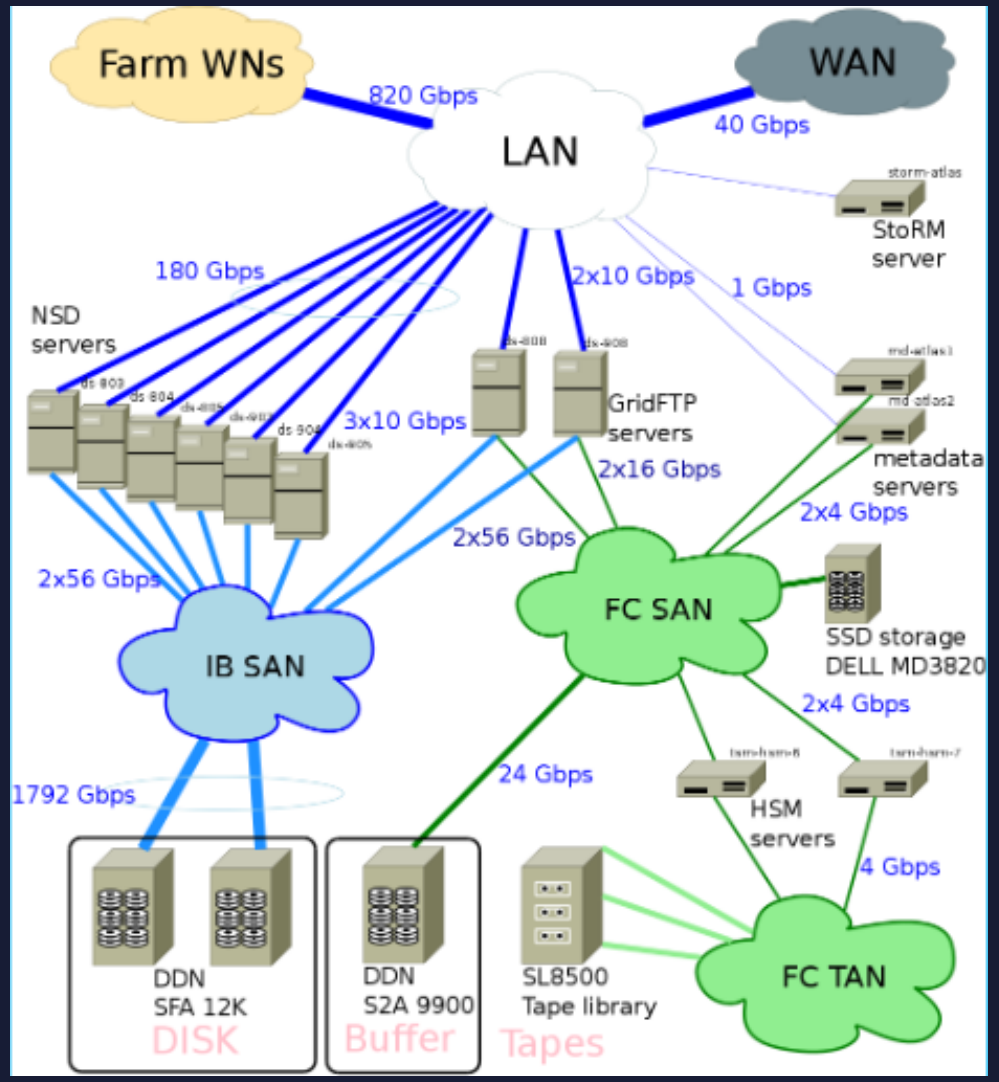
Installata al Tecnopolo una nuova TS4500 (IBM) con solo drive JF per nastri da 50TB/cartridge

Areal density scaling



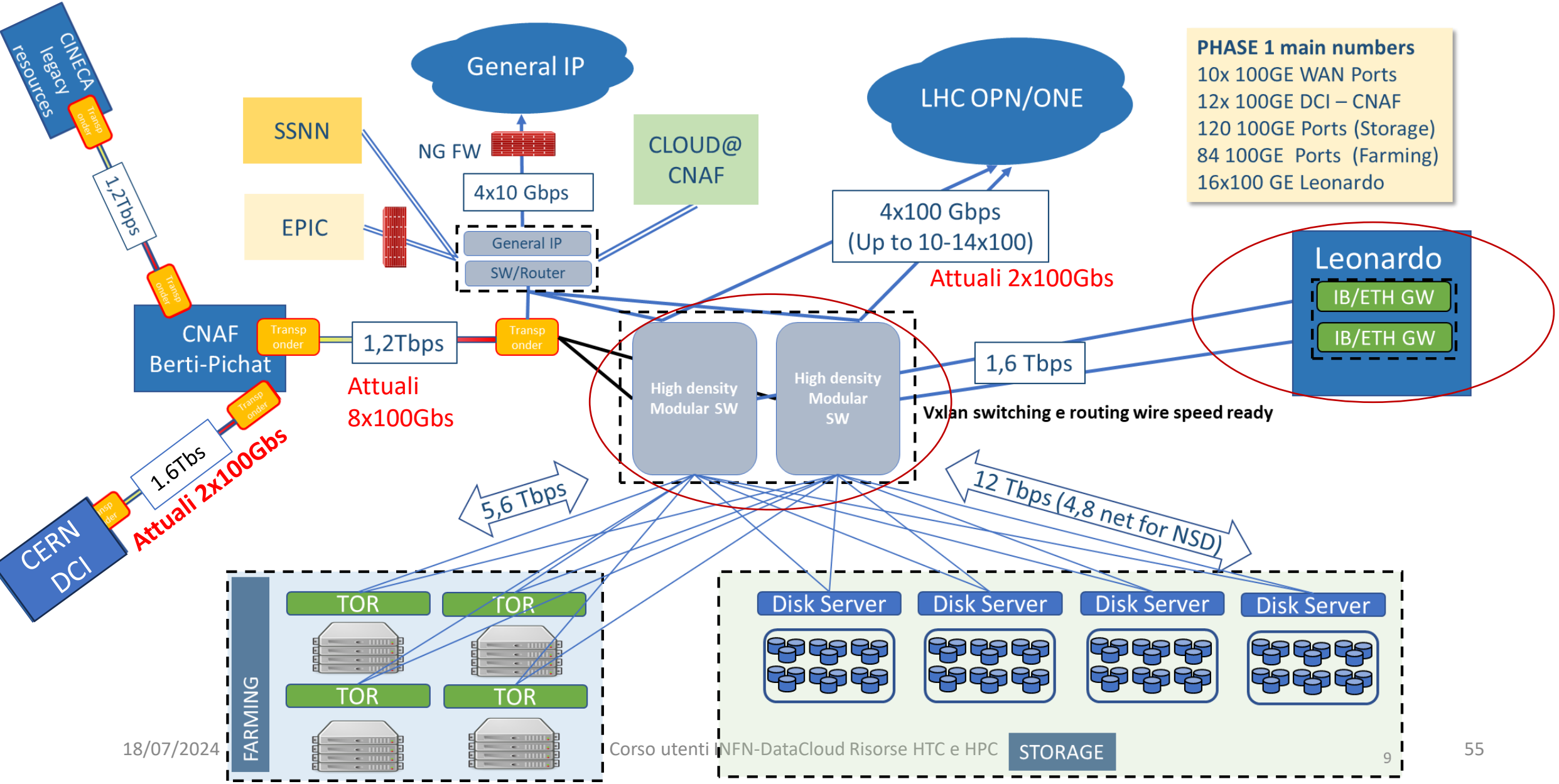
- 2015: IBM-FujiFilm demonstration of 123 Gb/in2 on BaFe tape
- 2017: IBM-Sony demonstration of 201 Gb/in2 on Sputtered Tape

A Storage Area Network at CNAF

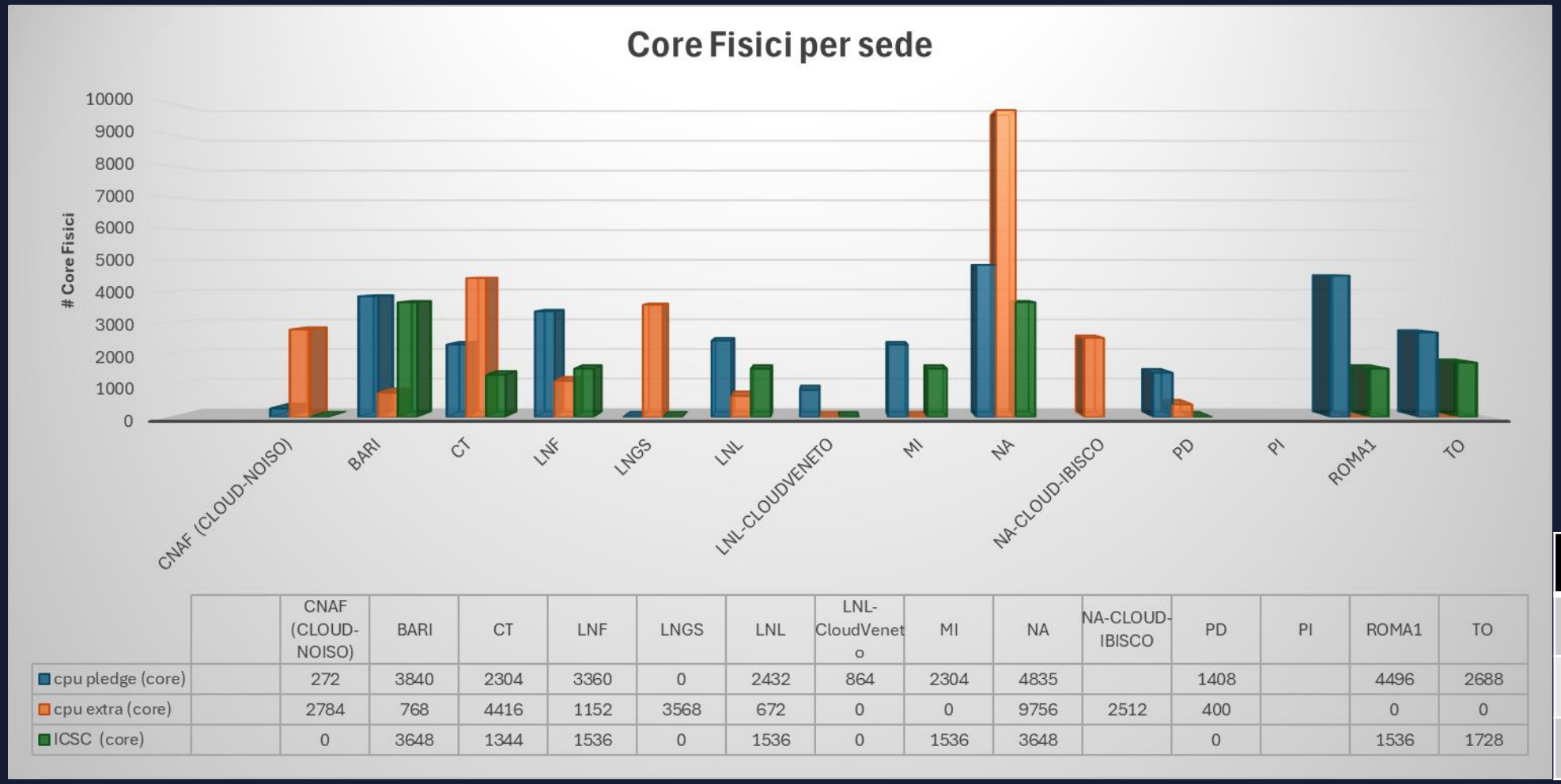


- 7 PB disk space
- 6 NSD servers (3x10 Gbps)
- 2 metadata servers (1Gbps)
- 2 GridFTP (XrootD) (2x10 Gbps)
- 2 HSM servers
- Metadata on SSD (mirrored)
- VM as Storm server
- Throughput required (5 MB/s/TB) = 21.5 GB/s
- Throughput available (6 NSD x 30 Gbps) = 22.5 GB/s

Networking Infrastructure at CNAF



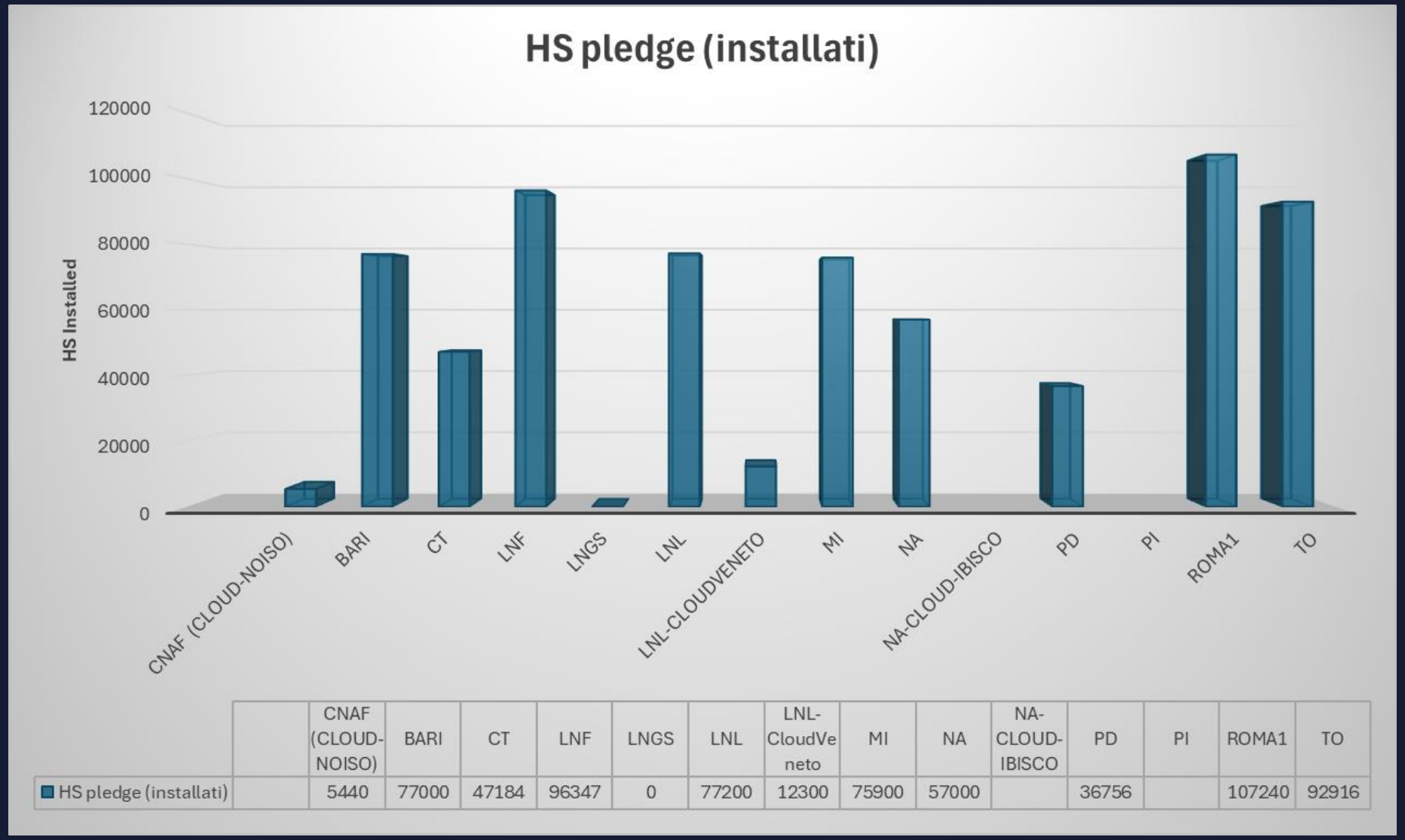
Risorse ai Tier-2



	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

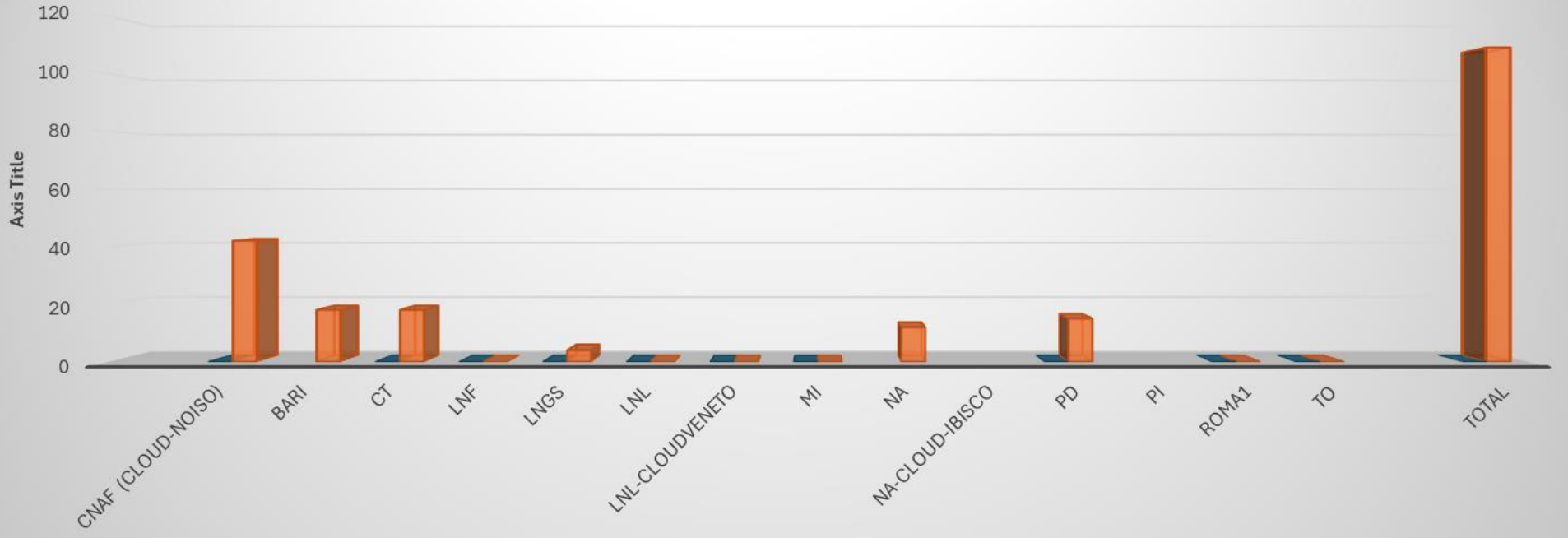
	TOTAL
Pledged	28803
Extra	26028
ICSC	16512

ICSC expected = 18048
1728 da PISA

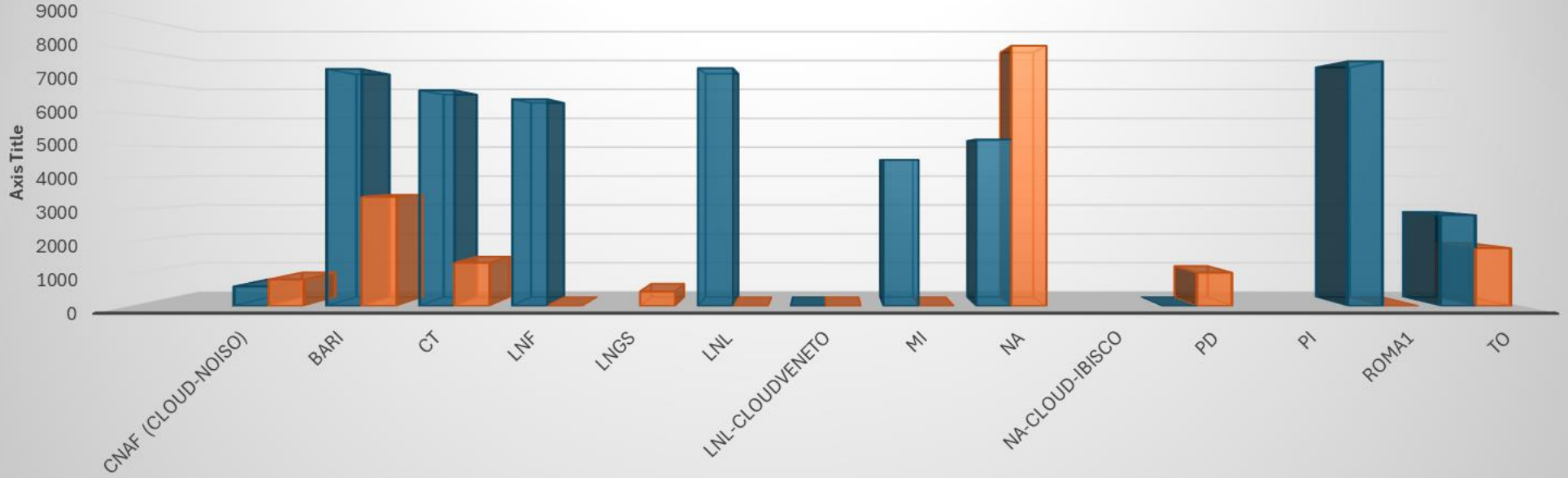


		CNAF (CLOUD-NOISO)	BARI	CT	LNF	LNGS	LNL	LNL-CloudVeneto	MI	NA	NA-CLOUD-IBISCO	PD	PI	ROMA1	TO
HS pledge (installati)		5440	77000	47184	96347	0	77200	12300	75900	57000		36756		107240	92916

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

Risorse della seconda tornata di acquisti (2024)

Tier-2, esclusi sistemi HPC

(i.e. solo quota ICSC)

Potenza CPU:

~17 HS06/coreHT

→ ~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

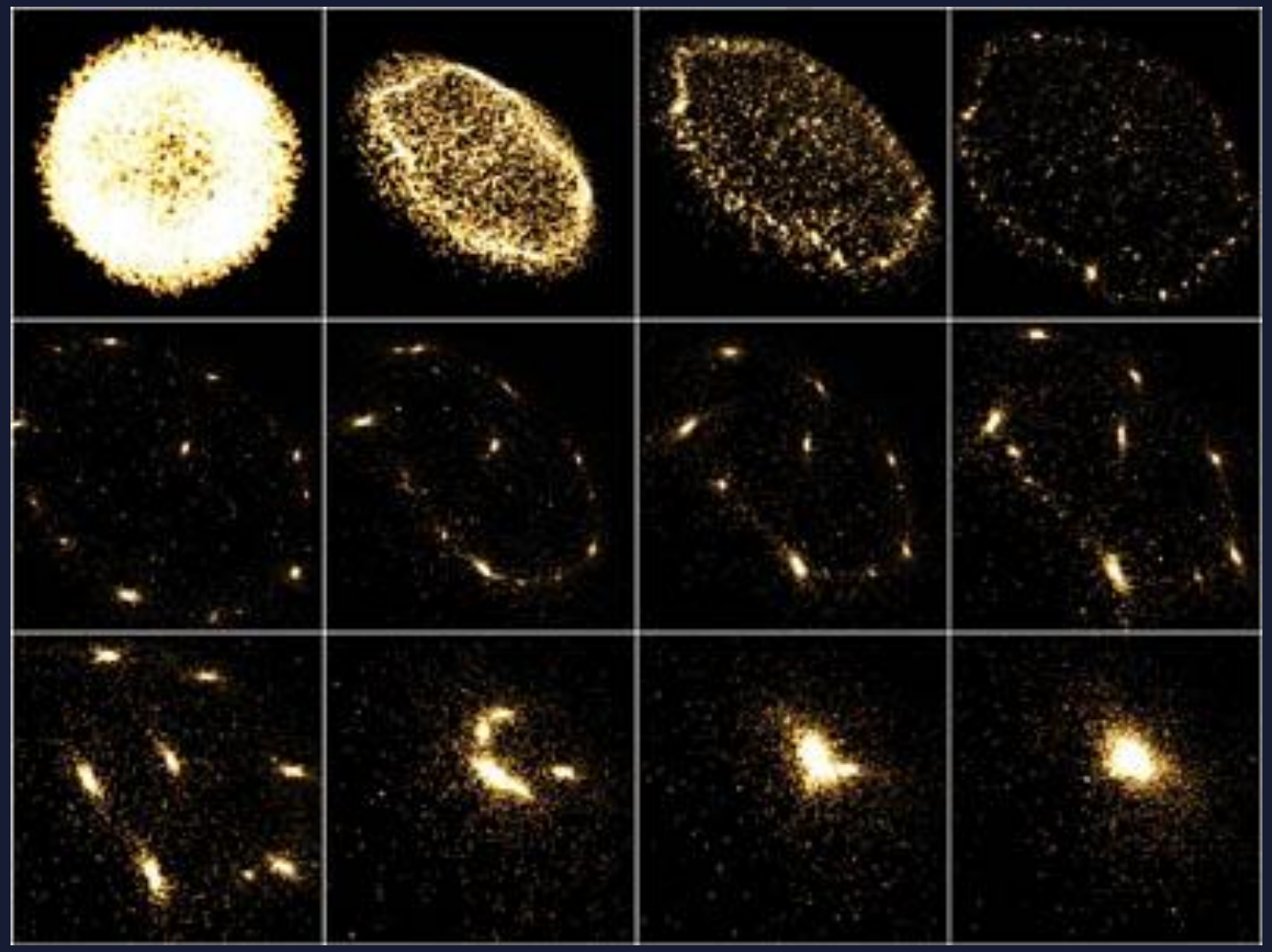
Sito	CPU (Core fisici)	Storage (PBL)
BA	2304	12.2
CT	2304	16.7
LNF	2304	2.5
LNFESA	1536	2.3
LNGS	-	4.6
LNL	784	5.8
PD	-	
MI	784	3.9
NA	2304	12.2
RM1	784	4.5
PI	2304	3.2
TO	1536	4.5
CNAF	-	-
TOT	16896	72.4

High Performance Computing
HPC

HTC and HPC - definition

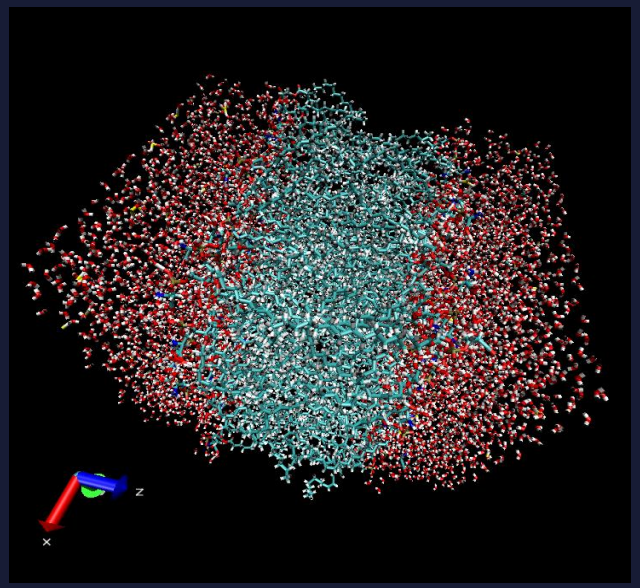
- High Throughput Computing (HTC)
  - The focus is on the execution of many copies of the *same program* at the *same time*
    - not in the speedup of individual jobs
  - Many copies of the same program run *in parallel* or *concurrently*
  - Maximize the **throughput**
- High Performance Computing (HPC)
  - speed up the individual job as much possible so that results are achieved more quickly
- HTC infrastructures tend to deliver large amounts of computational power over a long period of time.
  - In contrast, High Performance Computing (HPC) environments deliver a tremendous amount of compute power over a short period of time.
- The interest in HTC is in how many jobs complete over a long period of time instead of how fast an individual job can complete.

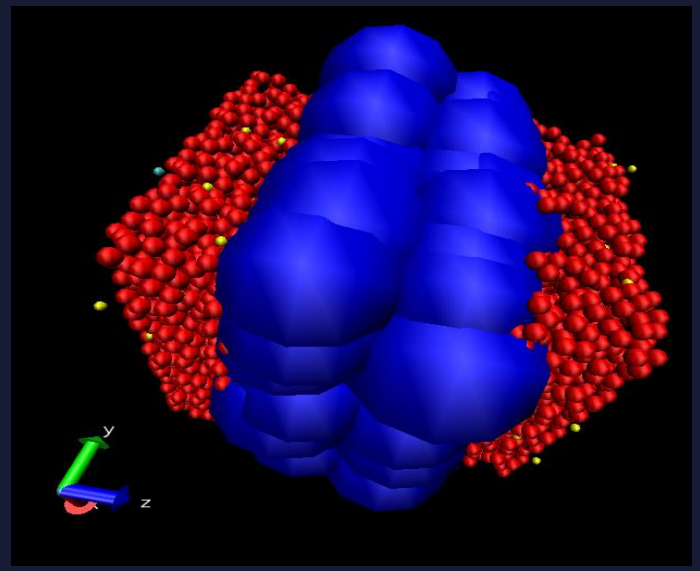
HPC Applications



HPC - Applications

- Molecular Dynamics

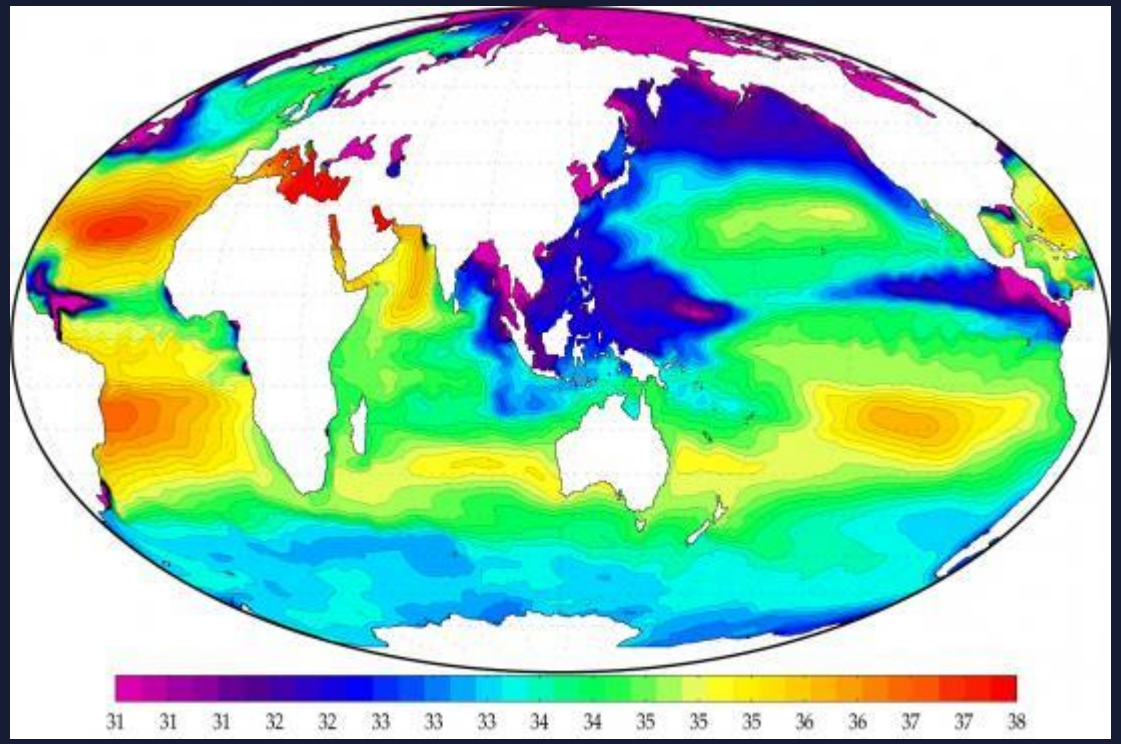


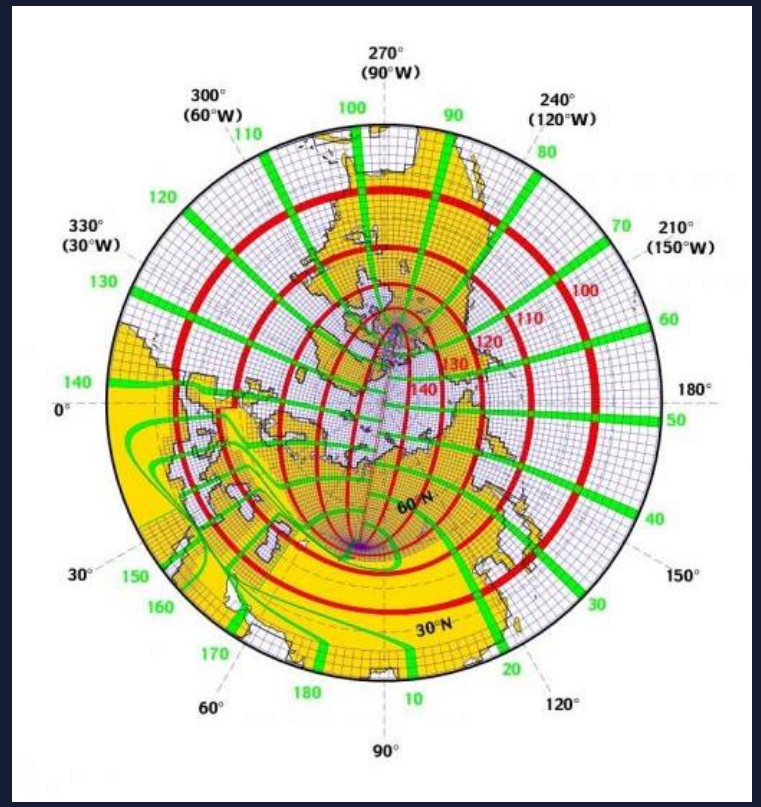


NAMD, Quantum Espresso, Gromacs, Gaussian, etc..

HPC - Applications

- Earth simulation

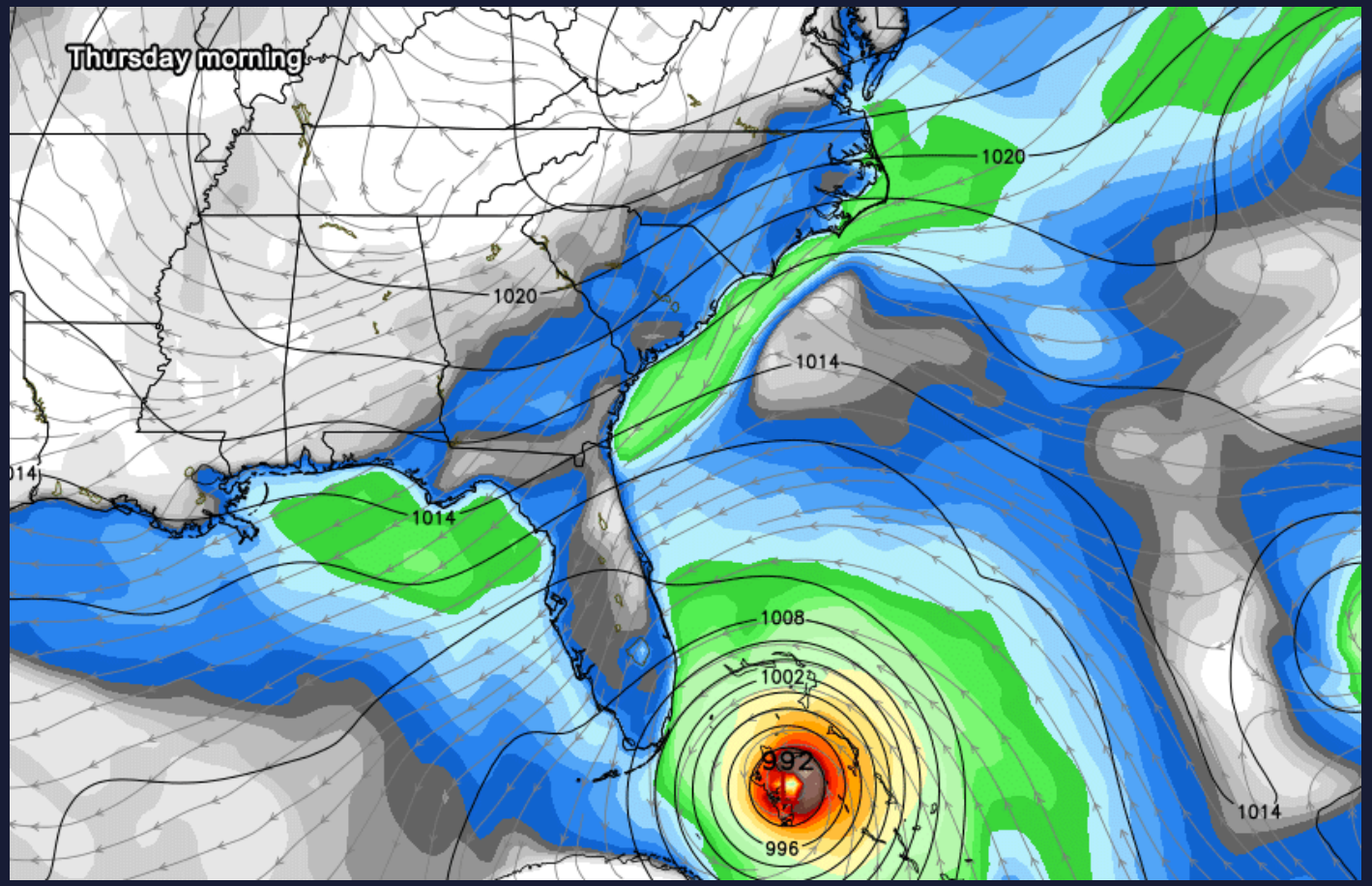




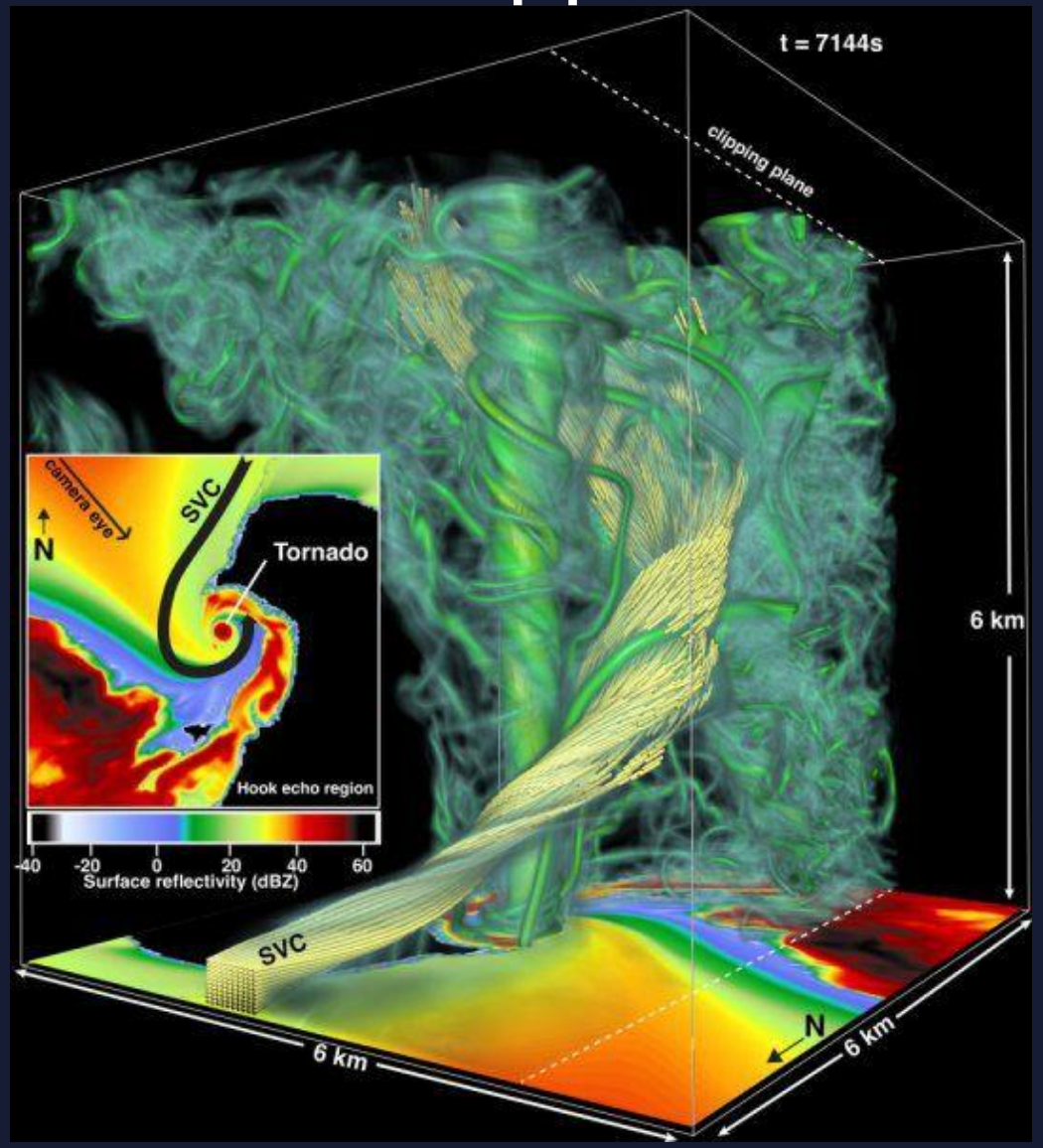
WRF, MM5, GLOBO, NEMO, etc..

HPC- Applications

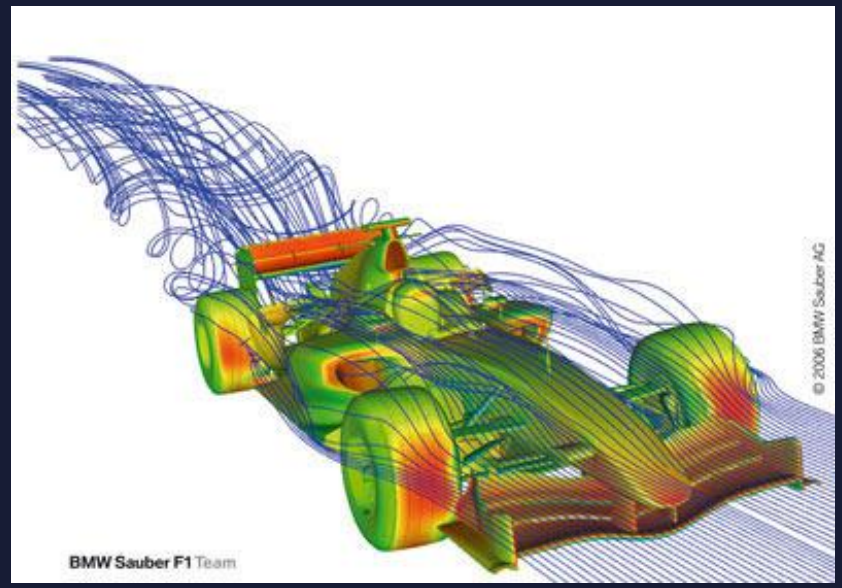
- Weather Simulation



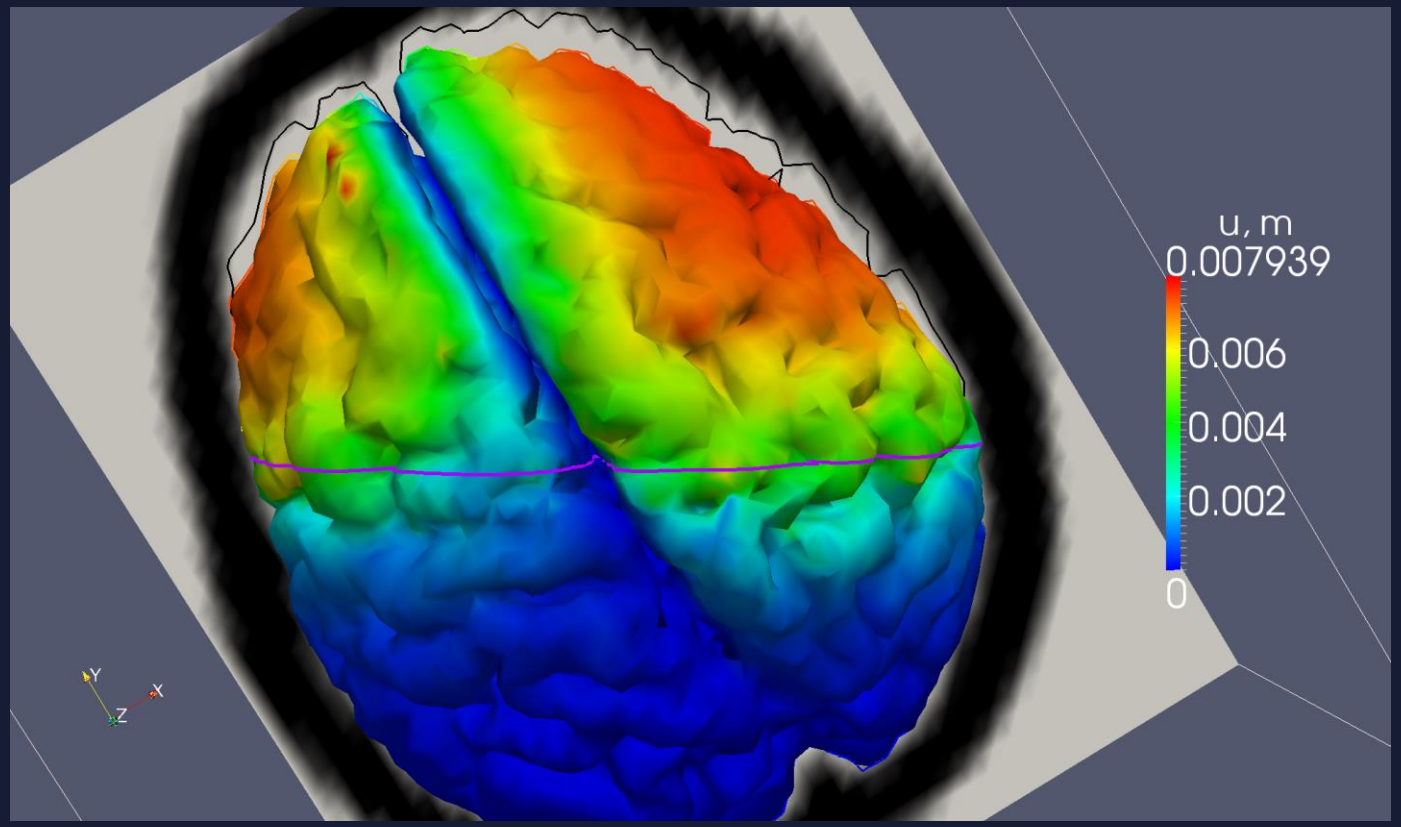
HPC - Applications



- Fluid Dynamics



HPC - Applications



- Brain Simulation

Once upon a time....

The vector machines

- Serial number 001 Cray-1™
  - Los Alamos National Laboratory in 1976
  - $8.8 million
  - 80 MFLOPS scalar, 160/250 MFLOPS vector
  - 1 Mword (64 bit) main memory
  - 8 vector registers
    - 64 elements 64bit each
  - Freon refrigerated
  - 5.5 tons including the Freon refrigeration
  - 115 kW of power
    - 330 kW with refrigeration

(*) Source: Wikipedia

Serial number 003 was installed at the National Center for Atmospheric Research (NCAR) in 1977 and decommissioned in 1989

Frontier

- Frontier, or OLCF-5, is the world's first and fastest exascale supercomputer, hosted at the Oak Ridge Leadership Computing Facility (OLCF) in Tennessee, United States
- It is based on the Cray EX and is the successor to Summit (OLCF-4).
- As of March 2023, Frontier is the world's fastest supercomputer.
- Frontier uses 9,472 AMD Epyc 7713 "Trento" 64 core 2 GHz CPUs (606,208 cores) and 37,888 Instinct MI250X GPUs (8,335,360 cores). They can perform double-precision operations at the same speed as single precision

Active	Deployment: Sep. 2021 Completion: May 2022
Operators	Oak Ridge National Laboratory and U.S. Department of Energy
Location	Oak Ridge Leadership Computing Facility
Power	22.7 MW[1]
Operating system	HPE Cray OS
Space	680 m^2 (7,300 sq ft)
Speed	1.206 exaFLOPS (Rmax) / 1.71481 exaFLOPS (Rpeak)[1]
Cost	US$600 million (estimated cost)
Purpose	Scientific research and development
Website	www.olcf.ornl.gov/frontier/

© Wikipedia

LEONARDO@CINECA

- **Petascale supercomputer located at the CINECA datacenter in Bologna, Italy.**
- **Atos BullSequana XH2000 computer**
  - **14,000 Nvidia Ampere GPUs**
  - **200 Gbit/s Nvidia Mellanox HDR InfiniBand connectivity.**
- **250 petaflops**
  - **top five in TOP500**
  - **second in Europe**

Leonardo

Active	November 24, 2022
Sponsors	European High-Performance Computing Joint Undertaking
Operators	CINECA
Location	Bologna, Italy
Architecture	13,824 Nvidia Ampere GPU cores
Power	6 MW
Space	900+ m^2
Memory	2.8 petabytes
Storage	110 petabytes
Speed	250 petaFLOPS (peak)
Cost	€240 million
Website	Leonardo Pre-exascale Supercomputer

© Wikipedia

LEONARDO@CINECA - 2

- **Booster Module**
  - The 3,456 individual nodes which make up the "booster module" are custom BullSequana X2135 "Da Vinci" blade servers, each composed of:
    - 1x Intel Xeon 8358 CPU, with 32 cores running at 2.6 GH
    - 512 GB RAM DDR4 3200 MHz
    - 4x NVidia custom Ampere GPU, 64 GB HBM2
    - 2x NVidia HDR InfiniBand network adapters, each with two 100 Gbit/s ports
    - Each node is expected to deliver 89.4 TFLOPs peak.
- **Data Centric Module**
  - The "data centric module" consists of 1536 nodes, each comprising a BullSequana X2610 compute blade with:
    - 2x Intel Sapphire Rapids CPUs, with 56 cores
    - 512 GB RAM DDR5 4800 MHz
    - 1x NVidia HDR InfiniBand network adapter, with one 100 Gbit/s port
    - 8 TB NVM storage

Clusters

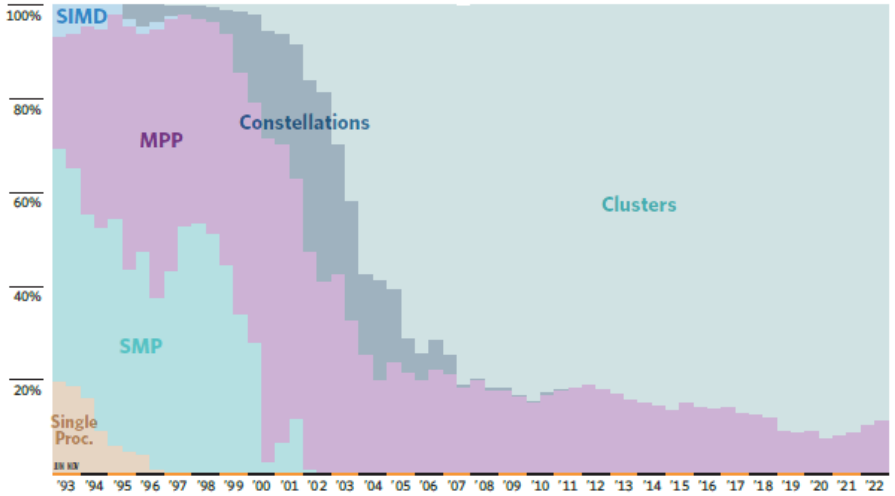
[a cluster is a] parallel computer system comprising an integrated collection of independent nodes, each of which is a system in its own right, capable of independent operation and derived from products developed and marketed for other stand-alone purposes

Dongarra et al. : "High-performance computing: clusters, constellations, MPPs, and future directions", Computing in Science & Engineering (Volume:7 , Issue: 2)

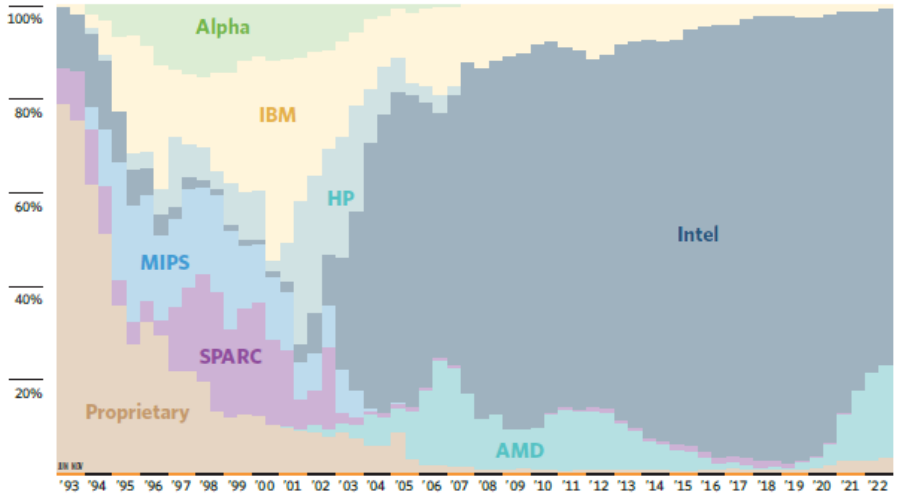
(*) Picture from: http://en.wikipedia.org/wiki/Computer_cluster

Top500.org – stats

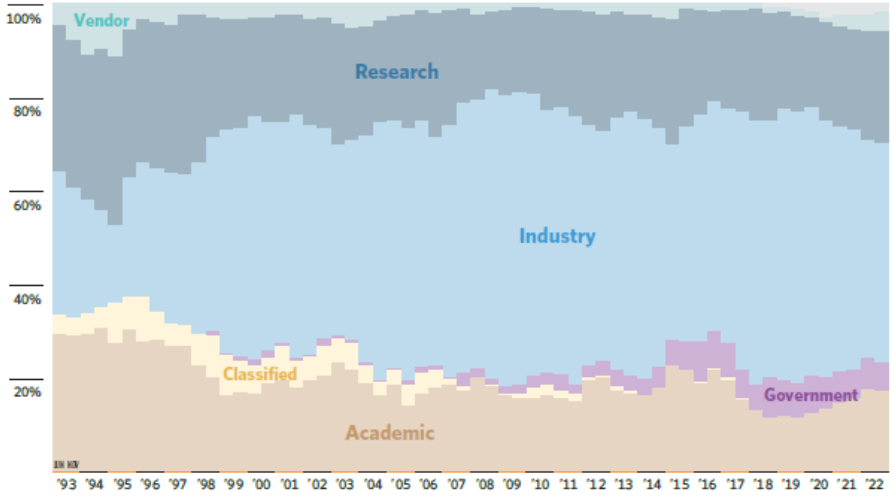
ARCHITECTURES



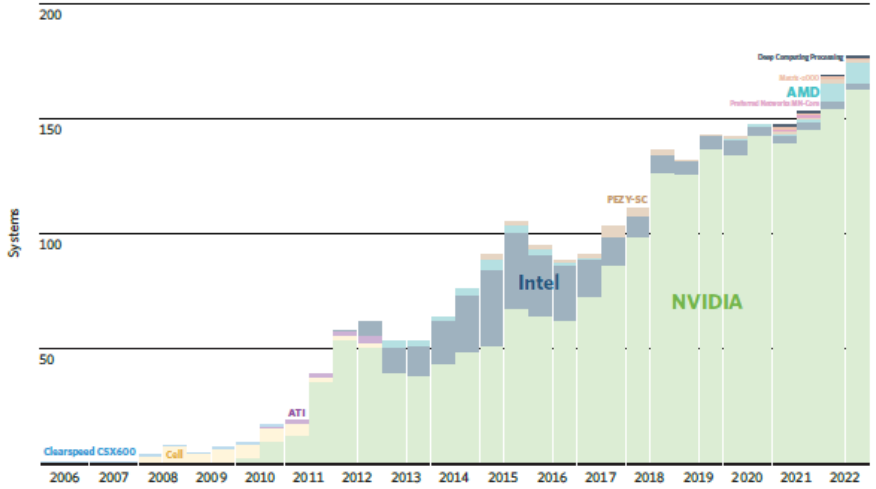
CHIP TECHNOLOGY



INSTALLATION TYPE



ACCELERATORS/CO-PROCESSORS



HPC made easy: **HPC Bubbles**

terabit

- Cluster HPC istanziati su Cloud tramite interfacce user-friendly
- Acquisizione delle bubbles di datacloud tramite progetti PNRR

Generated with AI · May 2024

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

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