Il (centro di) calcolo dell'INFN

Luca dell’Agnello – INFN-CNAF

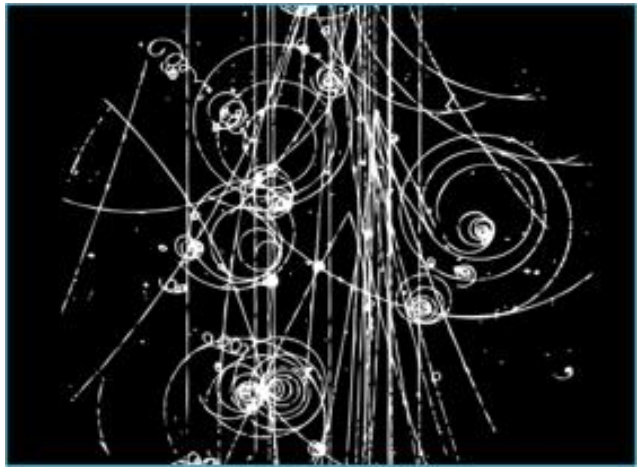
15 luglio 2025

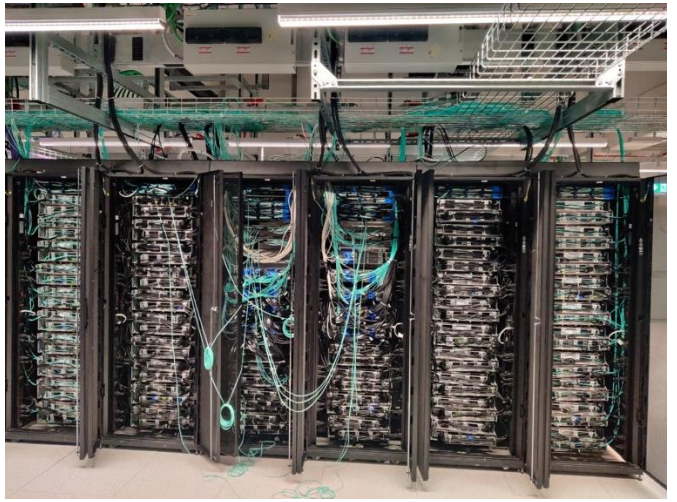
INFN

- L’INFN, l’Istituto Nazionale di Fisica Nucleare, è uno dei principali enti di ricerca italiani
  - La missione dell’INFN è la ricerca e lo studio dei componenti fondamentali della materia e delle leggi fisiche dell’universo
    - Esperimenti agli acceleratori (es. a LHC al CERN), raggi cosmici, astroparticelle, Onde gravitazionali
    - Fisica Teorica
  - Finanziato del MUR (Ministero dell'Università e della Ricerca)
- E’ distribuito su tutto il territorio nazionale con varie strutture
  - 20 sezioni ospitate nei Dipartimenti di Fisica delle Università
  - 4 laboratori (LNF, LNGS, LNL, LNS)
  - 3 Centri Nazionali dedicati a compiti specifici

Il CNAF (1/2)

- Il CNAF è il Centro Nazionale che ha come missione lo studio e lo sviluppo delle applicazioni informatiche e telematiche
- Fondato nel 1962 come centro per l'analisi dei fotogrammi prodotti nelle camera a bolle
  - L'acronimo originale era «Centro Nazionale Analisi Fotogrammi»
- La missione del CNAF è cambiata alcune volte ma sempre legata alla gestione dei dati per gli esperimenti dell'INFN
- Protagonista della nascita ed evoluzione della rete GARR per circa un decennio (1992-2001)
- Abbiamo iniziato ad occuparci di calcolo scientifico nel 2000
  - Realizzazione del Data Center Tier1 per gli esperimenti a LHC
  - Realizzazione, in collaborazione con CERN ed altre organizzazioni delle prime infrastrutture per il calcolo distribuito
    - Grid, cloud





Il CNAF (2/2)

- La sede principale è presso il Dipartimento di Fisica dell'Università di Bologna
  - Al Tecnopolo è ospitato il Data Center Tier1 dell'INFN
- Il Tier1 è il principale Data Center dell'INFN
  - Ospita circa metà della potenza di calcolo e della capacità di storage dell'INFN
  - Offre un servizio H24 per il calcolo degli esperimenti gestendo i servizi e supportando i ricercatori nell'utilizzo delle risorse di calcolo
- Al CNAF vengono studiate e sviluppate soluzioni IT innovative per migliorare l'usabilità e l'efficienza del centro e l'utilizzo di sistemi distribuiti su scala geografica
- Presso altre 10 sedi INFN vi sono altrettanti Tier2, Data Center più piccoli e con requisiti meno stringenti

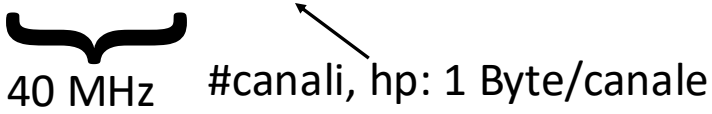
Una breve digressione

Perché serve il calcolo

- Gli esperimenti di fisica, in particolar modo quelli agli acceleratori come LHC, producono una grande quantità di eventi «fotografati» dai rivelatori
- Questi eventi vengono memorizzati in formato digitale e quindi rielaborati (anche più volte) ed analizzati (anche usando tecniche di ML)

Quanti dati vengono prodotti a LHC? (1/2)

- I dati prodotti da LHC dipendono dalla frequenza delle collisioni e dalla tipologia del rivelatore
- 4 esperimenti principali (Alice, Atlas, CMS, LHCb)
  - Collisioni p-p a 40 MHz ($40*10^6$ collisioni/s)
  - Ogni esperimento ha ~10^8 canali di lettura
- Dati prodotti da LHC:
  - #exp * frequenza collisioni * #canali * MB/canale
  - 4 exp* 40 *10^6 *10^8 = 4* $4*10^{15}$ = 4*4 PB/s (non sostenibile!)



- Con vari sistemi di riduzione (trigger), vengono selezionati solo gli **eventi** interessanti. Ad es. per ATLAS/CMS il flusso di dati scende a ~10 GB/s

Quanti dati vengono prodotti a LHC? (2/2)

- Quantità di dati da memorizzare in un anno dopo la riduzione:
  - #eventi/sec * tempo di presa dati * dimensione di un evento «raw»
    - Es.(CMS): 8 KHz * 6 * 10^6 s * 1 MB/evento = ~5 * 10^{10} MB = 50 PB

Tempo di presa dati (i secondi in un anno sono ~$\pi * 10^7$)

  - I dati «raw» vengono tipicamente salvati su nastro in 2 copie in 2 posti «distanti»
- Oltre ai dati «raw», vi sono poi le simulazioni MC, i dati ricostruiti e le analisi per un totale di ~100 PB/anno per esperimento
- *Nel Run-4 o HL-LHC (2030-2033) prevista produzione 300 PB/anno per esperimento*

Parentesi: ma cosa è un PB?

- 1 bit – può assumere valore 0 o 1
- 1 B = stringa di 8 bit: può codificare 2^8 = 256 caratteri
- 1 MB = 10^6 B
  - Dimensione di un evento CMS: 1 MB
- 1 GB = 10^9 B
  - Dimensione di un film 4k (1 h): 7-16 GB
  - Memoria di un telefonino: 100-500 GB
- 1 TB = 1000 GB = 10^{12} B
  - Disco rigido di un laptop: 500 GB - 1 TB
- 1 PB = 1000 TB = 10^{15} B
  - Dati «raw» prodotti in un anno da CMS/ATLAS: 50 PB
- 1 EB = 1000 PB = 10^{18} B
  - Dati prodotti da LHC nel 2023: 1 EB
- 1 ZB = 1,000,000 PB = 10^{21} B
  - Dati prodotti nel mondo nel 2021

Margaret Hamilton con le schede perforate con il software di controllo dell'Apollo 11 (64 KB)

Un po’ di notazioni

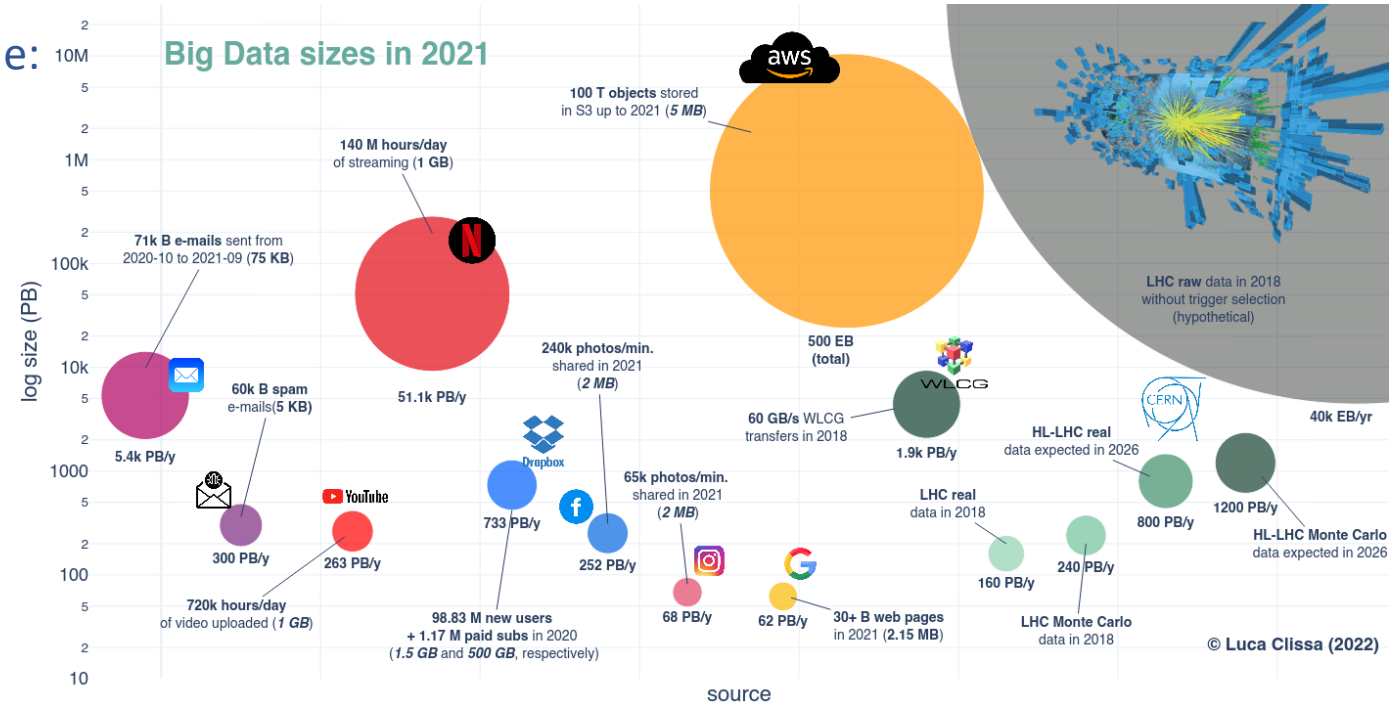
Prefissi del Sistema internazionale [modifica | modifica wikitesto]

Prefissi del Sistema internazionale di unità di misura

Prefisso	Simbolo	Notazione scientifica	Numero decimale	Scala lunga [note 1]	Scala corta [note 2]	Adozione [note 3]
quetta	Q	10^{30}	1 000 000 000 000 000 000 000 000 000 000	Quintilione	*Nonillion*	2022[1]
ronna	R	10^{27}	1 000 000 000 000 000 000 000 000 000	Quadriliardo	*Octillion*	2022[1]
yotta	Y	10^{24}	1 000 000 000 000 000 000 000 000	Quadrilione	*Septillion*	1991[2]
zetta	Z	10^{21}	1 000 000 000 000 000 000 000	Triliardo	*Sextillion*	1991[2]
exa	E	10^{18}	1 000 000 000 000 000 000	Trilione	*Quintillion*	1975[3]
peta	P	10^{15}	1 000 000 000 000 000	Biliardo	*Quadrillion*	1975[3]
tera	T	10^{12}	1 000 000 000 000	Bilione	*Trillion*	1960[4]
giga	G	10^{9}	1 000 000 000	Miliardo	*Billion*	1960[4]
mega	M	10^{6}	1 000 000	Milione	*Million*	1960[4]
chilo	k	10^{3}	1 000	Mille	*Thousand*	1795
etto	h	10^{2}	100	Cento	*Hundred*	1795
deca	da	10^{1}	10	Dieci	*Ten*	1795

Big Data: non solo LHC

- Altri esperimenti di fisica ed astrofisica di nuova generazione pongono sfide interessanti in termini di quantità di dati
  - CTA: 12 PB/anno compressi (prodotti all'origine: fino a 1 EB/anno)
  - DUNE (USA): 30 PB/anno
  - SKA: fino a 2 PB/giorno!!!
- Ma anche in altri campi scientifici…
  - 1 singolo genoma: ~100 GB
  - Survey su 1M di persone: 100 PB
- Ed i servizi web (social, Google, Netflix)?
  - Netflix: 140 Milioni di ore/giorno (file da 1 GB)
- E-mail: 300 PB/anno



Come vengono conservati i dati (1/2)

- Per la conservazione dei dati a lungo termine (dati non riproducibili come i «raw» o che richiedono per la generazione molte ore di CPU) si usano librerie a nastri magnetici
  - Robot con capacità di centinaia di PB
- Pro: affidabilità, costo «contenuto»
- Contro: bassa velocità di accesso ai dati
  - In streaming: 400 MB/s.
  - Spesso necessario saltare da un punto all'altro di un nastro o da un nastro all'altro
- I dati vengono prima copiati sequenzialmente su disco e poi acceduti dai ricercatori

Una tape di nuova generazione contiere 50 TB

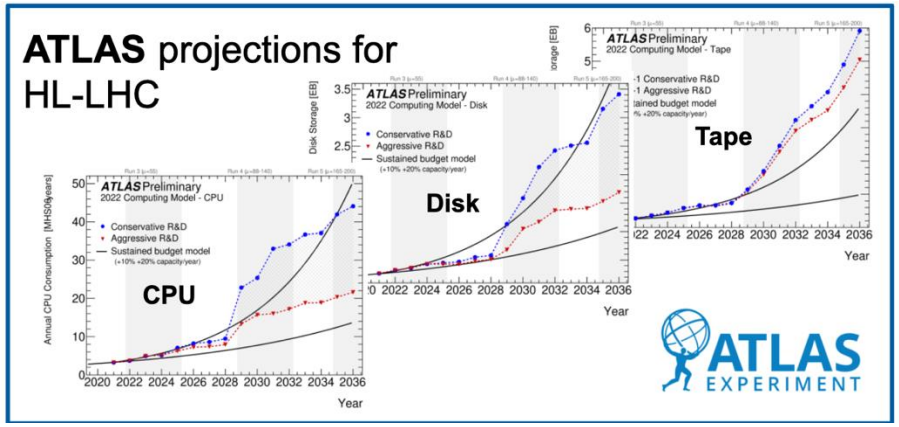
Come vengono conservati i dati (2/2)

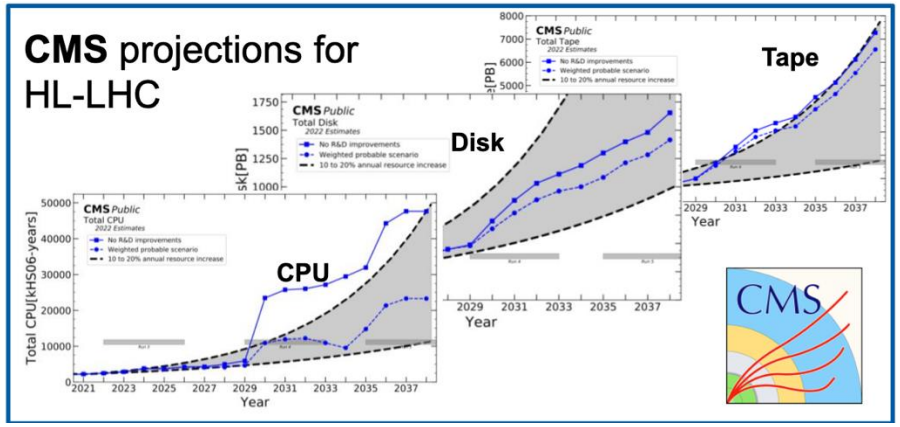
- Per la loro elaborazione i dati sono tenuti (o copiati) su disco
- Pro: accesso rapido
  - Sistemi storage con «controller» ad alte prestazioni possono accedere a più dischi in parallelo (es. sistema da 8 PB netti con banda passante di 30 GB/s)
- Contro: costi più elevati (x9), minore affidabilità a lungo termine

Un disco di tipo "enterprise" contiene 20 TB
Solitamente parte della capacita' viene usata per ridondanza (RAID)

Quanti computer servono per tutti i dati di LHC?

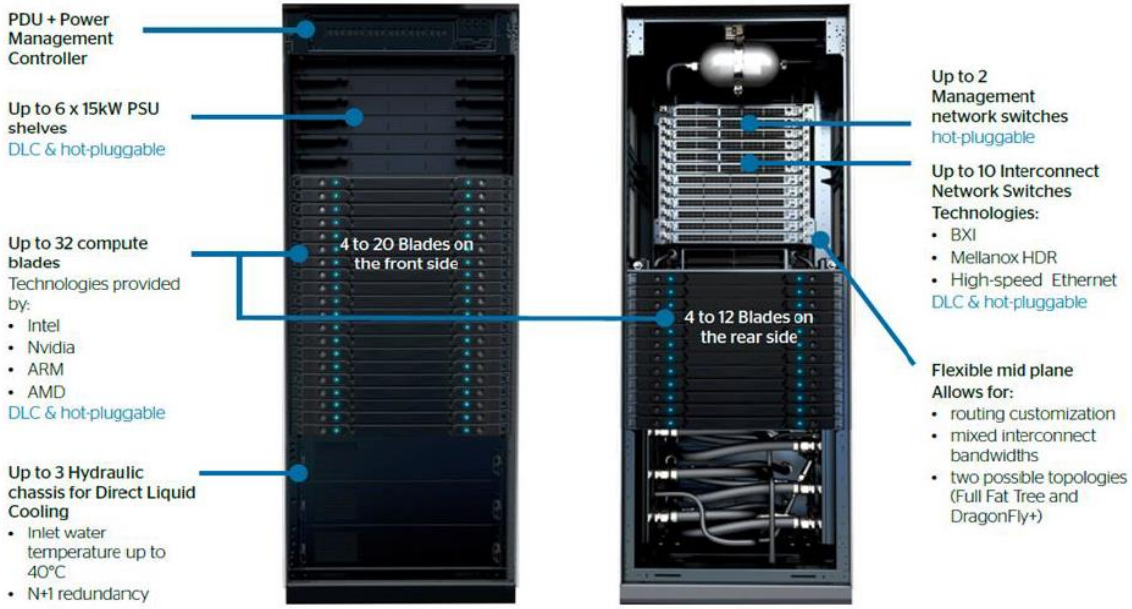
- Per produrre le simulazioni, ricostruire i dati raw, effettuare le analisi servono veramente «tante» ore di calcolo
  - Ad es., sempre per gli esperimenti a LHC, sono necessari ~200,000-300,000 CPUcore/anno
- E il peggio deve ancora arrivare! ☺
  - Previsto un aumento «robusto» (x6-7) con il Run4 (dal 2030)





Parentesi: cosa è un CPUcore

- Un processore multicore è caratterizzato dall'essere costituito da una moltitudine di unità di elaborazione indipendenti («core»), integrate sullo stesso chip.
  - Ad es. i computer della GP di Leonardo hanno ciascuno 2 CPU con 112 core (quindi ogni macchina può eseguire ~220 processi contemporaneamente)
- Con l'espressione «300,000 CPUcore/anno» si intende l'uso esclusivo per un anno di 300,000 core che equivalgono a ~1500 computer



1 rack = 96 computer, ~21500 core, 80 kW (DLC)

Schema di un rack di Leonardo (GP)

Fine della digressione

Come si gestice tutto questo?

- Un grande data center?
  - Un edificio con ~1,000,000 computing core e 200,000 HDD
  - Avrebbe potuto funzionare (Google)
  - Pro:
    - (probabile) economia di scala
  - Contro:
    - «single point of failure»
    - Non «appealing» per le nazioni partecipanti finanziare il tutto senza ritorno economico per il proprio territorio
    - Probabilmente difficile trovare personale sufficiente nella stessa zona
- Infrastruttura distribuita
  - Pro:
    - Costi e crescita di know-how distribuiti
    - Realizzazione ridondanza dei servizi più facile
    - Ritorno locale dell'investimento
  - Contro:
    - Necessaria un'infrastruttura di calcolo distribuito
- Nasce la Grid

Parentesi: cosa è la Grid?

- Goal: fare in modo che i ricercatori/le ricercatrici vedano tutti i siti come un unico insieme di risorse
- Premessa implicita: tutto ciò, per funzionare, necessita di una rete di interconnessione veloce ed affidabile
- Schematicamente si tratta di fattorizzare il problema:
  - Livello fisico - le «macchine» sono distribuite in molti (>100) data center
    - Problema: ogni data center ha i propri strumenti per il controllo degli accessi, la modalità di accesso alle risorse di calcolo ed ai dati
  - Livello logico – vengono concordati protocolli ed interfacce comuni
    - Un sistema di »trust» per permettere la mutua Autenticazione e Autorizzazione
    - Un insieme di protocolli per l'accesso ai dati, la loro movimentazione fra siti, il supporto, l'accounting
- Per costruire la Grid è stato necessario definire queste interfacce e sviluppare i tool («middleware»)
  - Tutto sviluppato come open-source (in Linux)

La collaborazione WLCG

WLCG - Worldwide LHC Computing Grid

- Nasce nel 2005 come infrastruttura globale di calcolo distribuito per gestire i dati prodotti da LJC
- Comprende 170 centri di calcolo in oltre 40 nazioni
  - A seconda della loro dimensione e delle funzionalità offerte, i centri si distinguono in Tier-0 (CERN), 14 Tier-1 e 155 Tier-2
  - In Italia, oltre al CNAF (Tier-1) ci sono 10 Tier-2
- La collaborazione ora include anche esperimenti non LHC o non HEP (es. astroparticelle)
- Sono state sviluppate, nell'ambito di WLCG, soluzioni di avanguardia per il calcolo distribuito
  - Servizi per l'autorizzazione all'accesso alle risorse (VOMS, IAM)
  - Servizi per la distribuzione di grandi moli di dati fra i vari data center (FTS, Rucio)
  - Sistemi per l'accesso ai sistemi di calcolo e storage (CE, StoRM, dCache, EOS)
  - Sistemi di monitoraggio ed accounting
  - …….

Il Tier-1 (1/5)

- Dal 2003, il CNAF ospita il Tier-1 italiano, fornendo risorse, supporto e servizi necessari alle attività di storage, distribuzione, processamento e analisi dei dati
- Attualmente, oltre 70 comunità scientifiche utilizzano il Centro
  - Non solo LHC e non solo nel campo della Fisica
- La nuova sede (Tecnopolo) è stata inaugurata a maggio 2024
- L'infrastruttura (spazio, potenza elettrica e condizionamento) è in grado di soddisfare richieste per il prossimo decennio
  - Tecnologia moderna che permette di raggiungere un PUE (~1.2-1.3) basso indispensabile per abbattere i consumi.
  - Al momento disponibili fino a 3 MW di potenza
  - Prevista l'espansione a 10 MW per il Run4

Cosa è il PUE? Indica il rapporto tra l'energia elettrica totale usata e quella per la sola parte IT

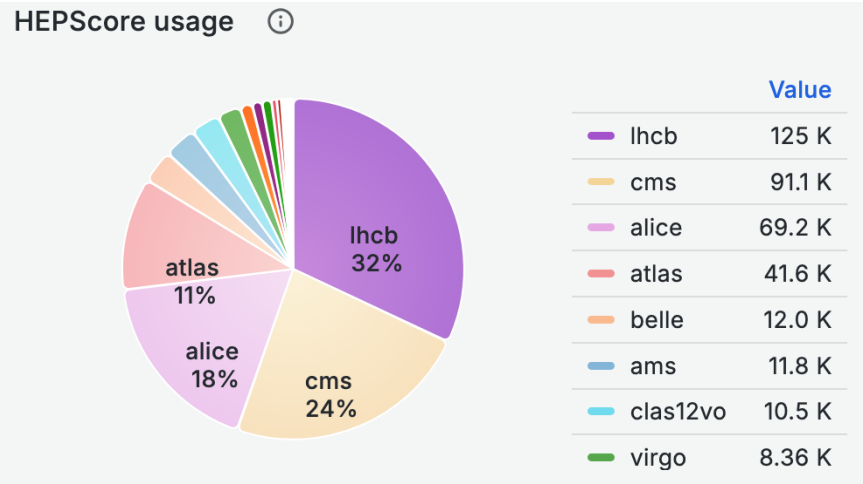
Il Tier-1 (2/5)

- Potenza di calcolo erogata con un ~2.000 nodi di calcolo (macchine sia fisiche che virtuali) organizzati in un sistema detto «farm»
  - I ~60.000 core della farm sono gestiti da un batch system (HTCondor)
- Parte della potenza di calcolo è fornita dal CINECA tramite Leonardo
  - I core forniti sono integrati in maniera trasparente nella nostra farm
  - Collegamento ad alta velocità fra il nostro data center ed il CINECA

Un batch system gestisce la farm prendendo in carico le richieste di esecuzione dei programmi (job) degli utenti e smistandoli sui vari computer tenendo conto di priorità, quote, necessità dei singoli job (es. #core, quantità di memoria)

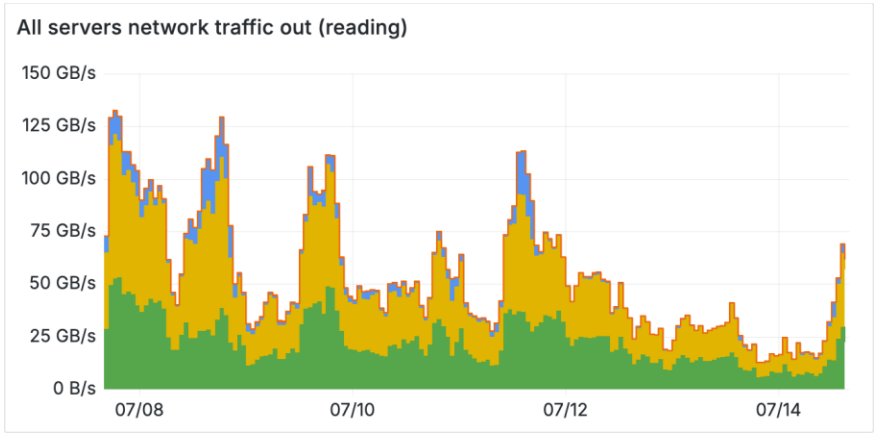
Il Tier-1 (3/5)

- L’interfacciamento alla farm avviene sia direttamente (i.e. collegandosi su un server centrale al CNAF) che «via Grid» (il ricercatore/ricercatrice sottomette la richiesta ad un server centrale che smista il job nel sito con i dati necessari o più scarico).
- E’ disponibile anche una cloud: in prospettiva ci sarà una convergenza con l’inclusione della farm nella cloud INFN
  - Disponibili anche server con GPU integrati nella cloud



Il Tier-1 (4/5)

- ~100 PB di disco organizzato in file-system paralleli accessibili direttamente dalla farm (protocollo Posix)
- ~200 PB di nastri distribuiti su 3 librerie integrate con i file-system
- L'accesso ai dati avviene dalla farm o dall'esterno usando i sistemi di trasferimento dati disponibili sulla Grid
  - Job su farm remote possono accedere in lettura ai dati presenti al CNAF

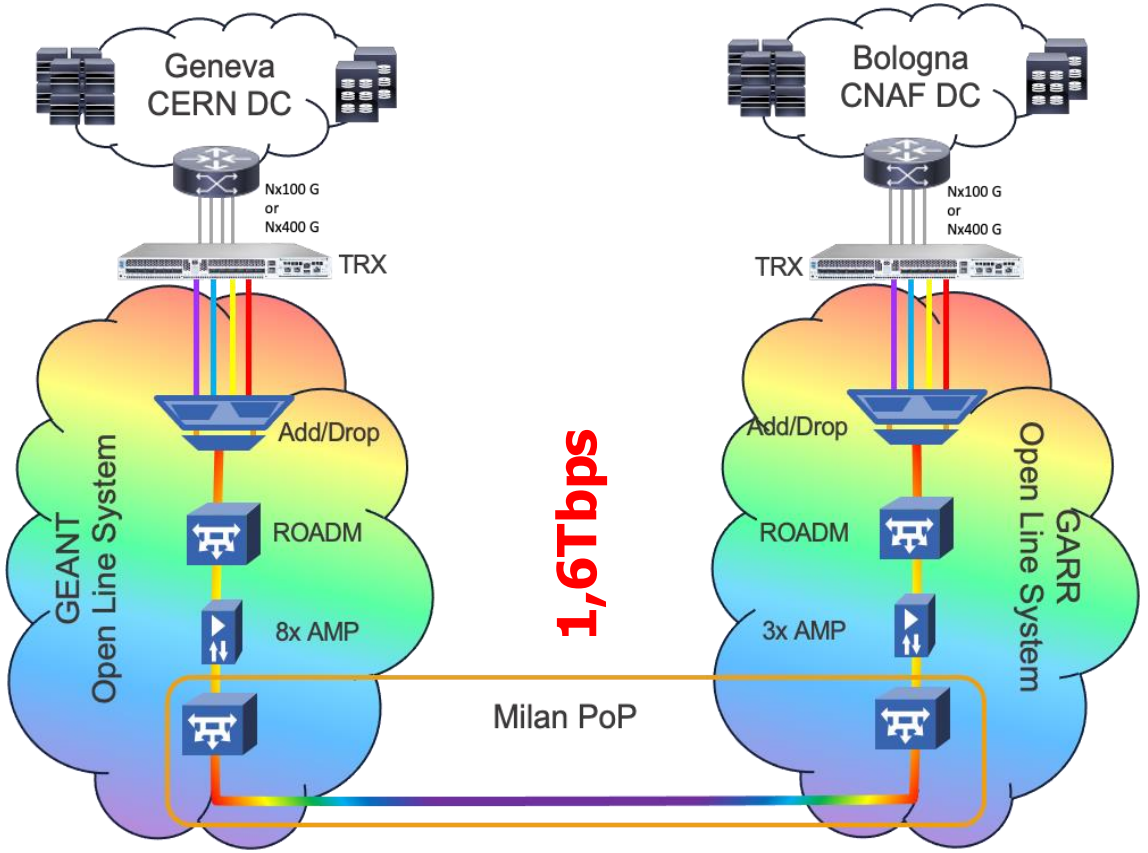


IBM Spectrum Scale

ceph

Il Tier-1 (5/5)

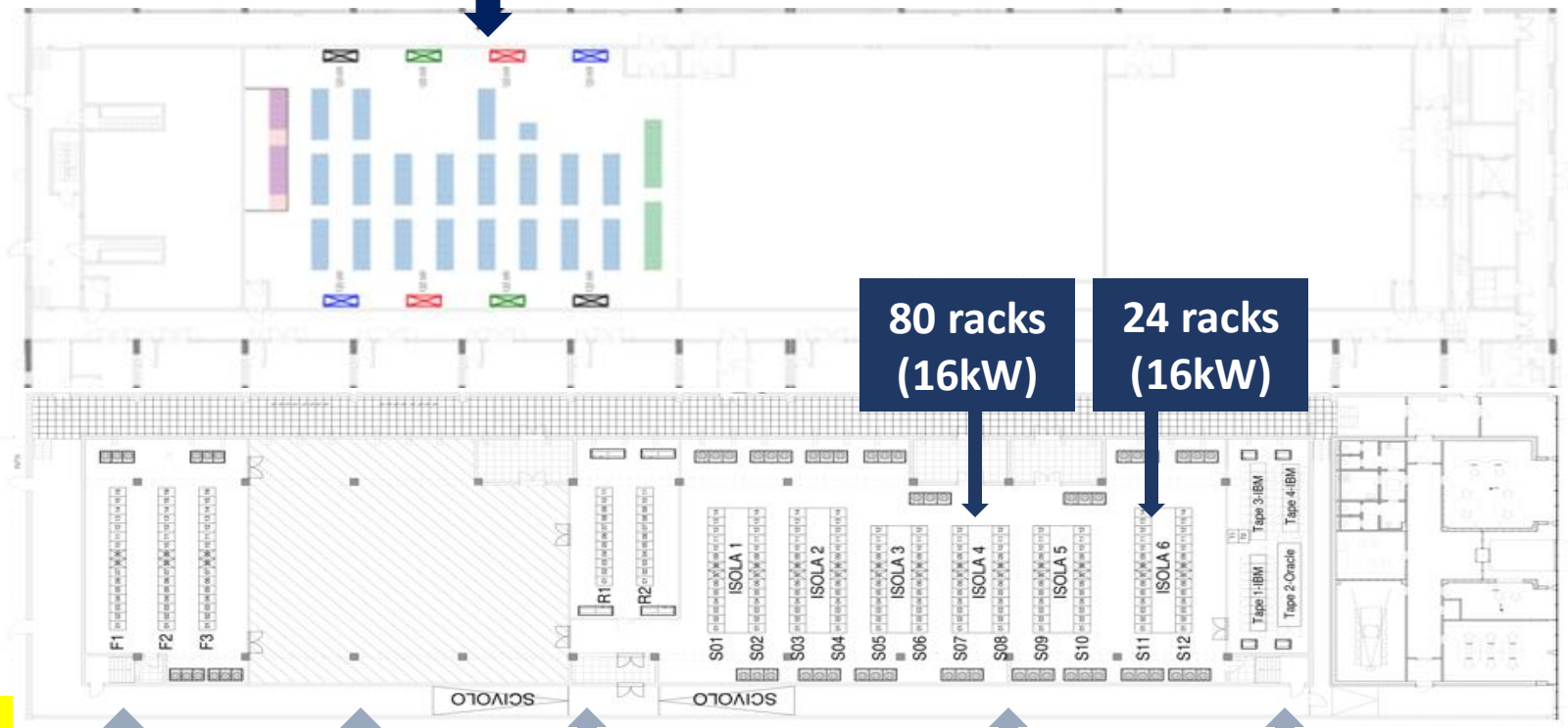
- Il CNAF è parte di LHCOPN/ONE, una rete ad alte prestazioni (almeno 2x100 Gbps) che interconnette tutti i Tier
  - Indispensabile per la copia dei dati dal CERN ai Tier-1 e fra i Tier-2 ed i Tier-1
- Nel 2023 è stato realizzato il primo light path ottico diretto in Europa su una distanza di 1000 km (di fibra ottica) fra CNAF e CERN
  - Può scalare fino a 1.6 Tbps



Le prossime sfide

- Gli esperimenti di prossima generazione (già con HL-LHC) richiederanno enormi risorse di calcolo e storage (e rete) per la crescente complessità degli eventi e la quantità di dati raccolti (quasi un ordine di grandezza maggiore a LHC)
- Risorse di calcolo necessarie x20 rispetto a oggi ma il budget a disposizione resterà sostanzialmente costante
- Nuovi esperimenti richiederanno risorse paragonabili a HL-LHC
- È necessario apportare innovazioni in vari settori
  - Tecnologia
    - Parallelizzazione (HP06/$$), eterogeneità (HP06/Watt)
  - Infrastruttura
    - Data Lake: consolidamento dello storage in pochi centri, CPU da centri HPC o cloud commerciali, riduzione costi operations (QoS)
  - Computing e Analysis Model degli esperimenti

Il Data Center



CINECA Data Hall

INFN Data Hall

DLC 80kW

C2

B5

Risorse Attuali
80.000 CPU core
100PB Disco
200PB TAPE

Trend di crescita ~ 20% annuo

CPU area

EXPANSION area

NETWORK core

STORAGE/SERVICE area

Tapes libraries