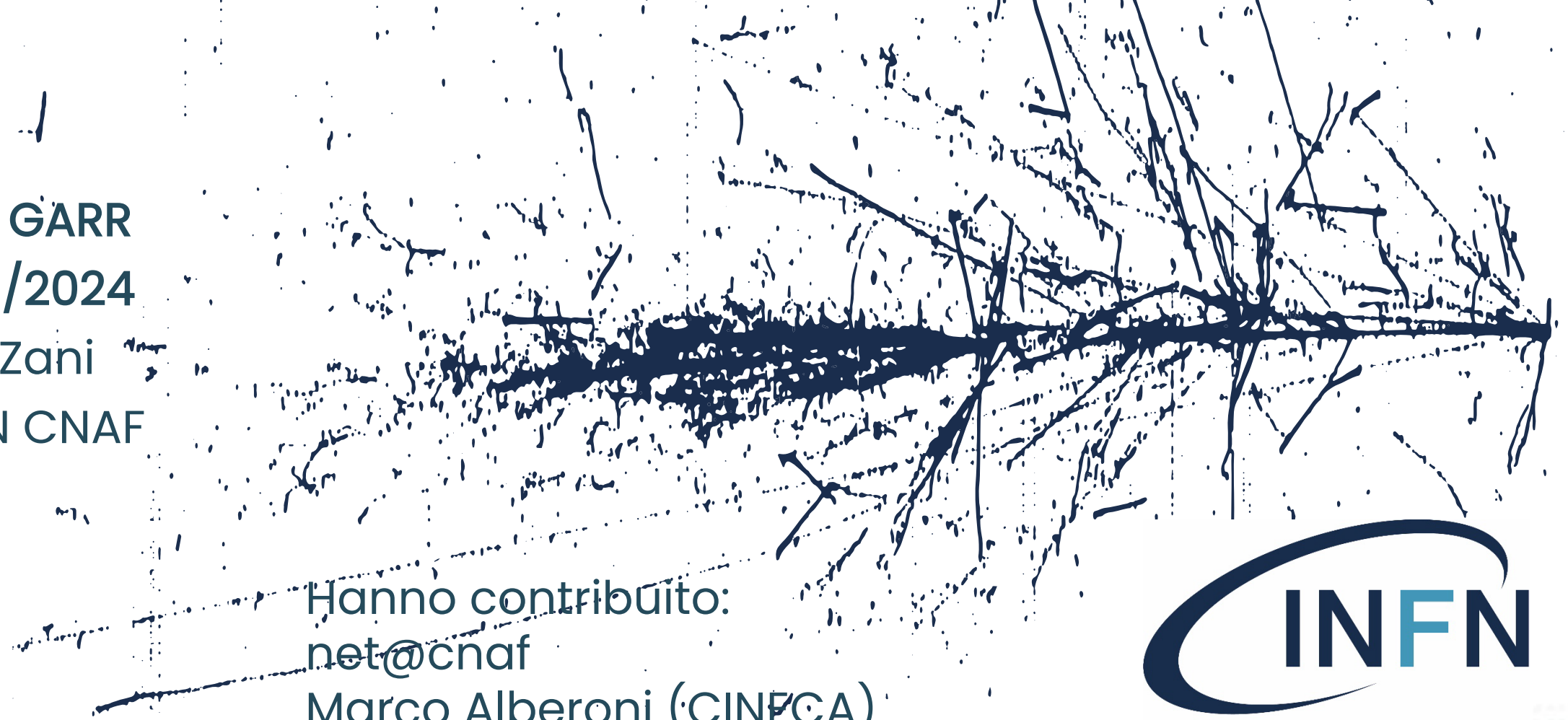


WS GARR
5/11/2024

S.Zani
INFN CNAF

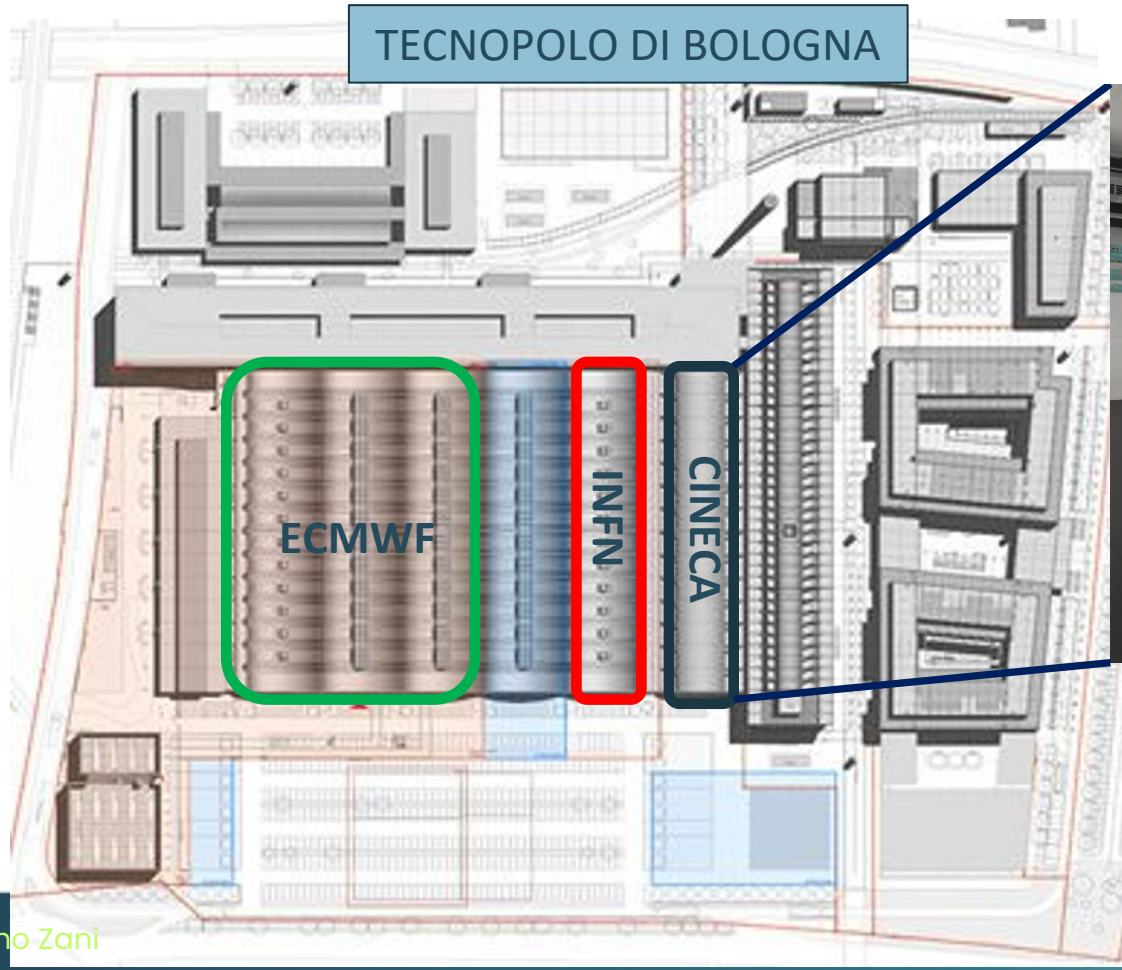
Hanno contribuito:
net@cnaf
Marco Alberoni (CINECA)
Marco Sbrighi (CINECA)

Istituto Nazionale di Fisica Nucleare

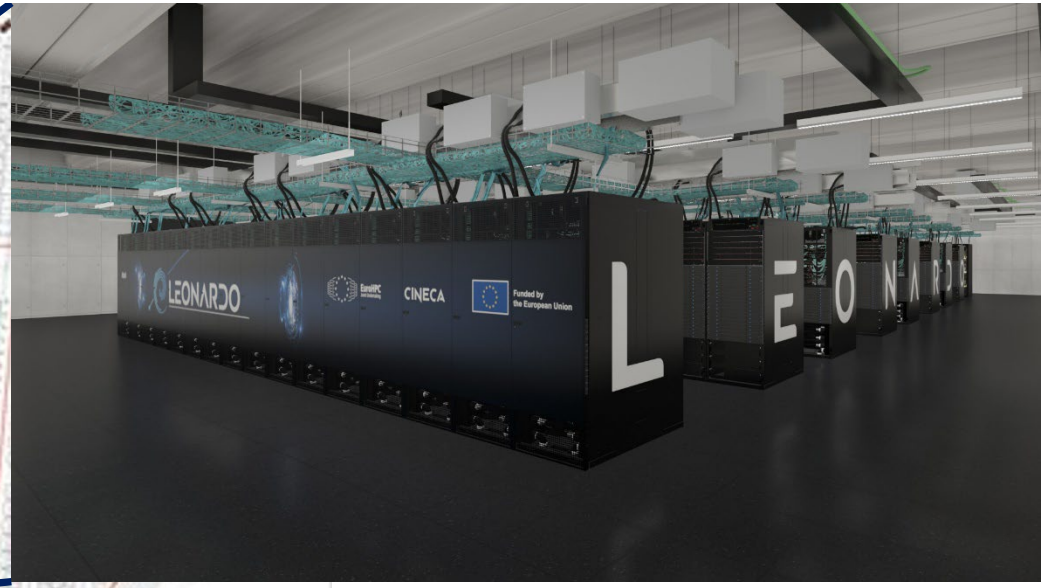
HTC e HPC interconnessione di reti diverse

Uso “opportunistico” di un supercalcolatore (aspetti di rete)

Utilizzo dei nodi del super calcolatore Leonardo *(HPC)* come «Worker Node» standard del TIER1 dell’INFN *(HTC).*



LEONARDO
Pre Exascale Supercomputer di CINECA, INFN, SISSA



- Il Datacenter del CNAF e quello di CINECA sono adiacenti (80 metri di fibra).
- Leonardo è basato su **processori x86**

Il modello di calcolo High Throughput Computing dell'INFN

Istituto Nazionale di Fisica Nucleare

Fatta eccezione per i Fisici Teorici dell'INFN che fanno uso di applicazioni di tipo HPC (Inter process communication, bassissima latenza fra i processori) **la maggior parte dei workflow dell'INFN prevede elaborazioni di un elevato numero di processi indipendenti che accedono a grandi quantità di dati a rate elevati (HTC).**

*"The mission of the WLCG project is to provide **global computing resources** to store, distribute and analyse the **~200 Petabytes of data expected every year** of operations from the Large Hadron Collider (LHC) at CERN"*

Il modello di calcolo distribuito degli esperimenti e la necessità di **spostare grandi quantità di dati fra i diversi data center** hanno un fortissimo impatto sulla evoluzione delle reti anche a livello globale.

Le “Bocche di fuoco”
I datacenter HTC (High Throughput Computing) dell’INFN

Istituto Nazionale di Fisica Nucleare

Ogni Job che “Gira” su ogni core deve avere accesso ad alta velocità a **grandi quantità** di dati locali o custoditi in altri centri WLCG. La rete locale di un centro di calcolo per LHC deve garantire un throughput aggregato tra nodi di calcolo e storage dell’ordine di decine di **Tbps** (Terabit per secondo)

I **clustering file system**, delle **LAN** e l’efficienza dei **data mover** sono le «Bocche di fuoco» dei datacenter dell’INFN e la rete GARR (a cui sono connesse) ha un ruolo fondamentale per garantire il throughput a livello di dorsale nazionale e verso le reti della ricerca mondiali.

MIGLIAIA DI SERVER (Commodity)
Decine di migliaia di CPU CORE

A differenza dei centri HPC:
Latenza dell’ordine dei ms contro 100+ ns di un DC HPC (No inter process communication)

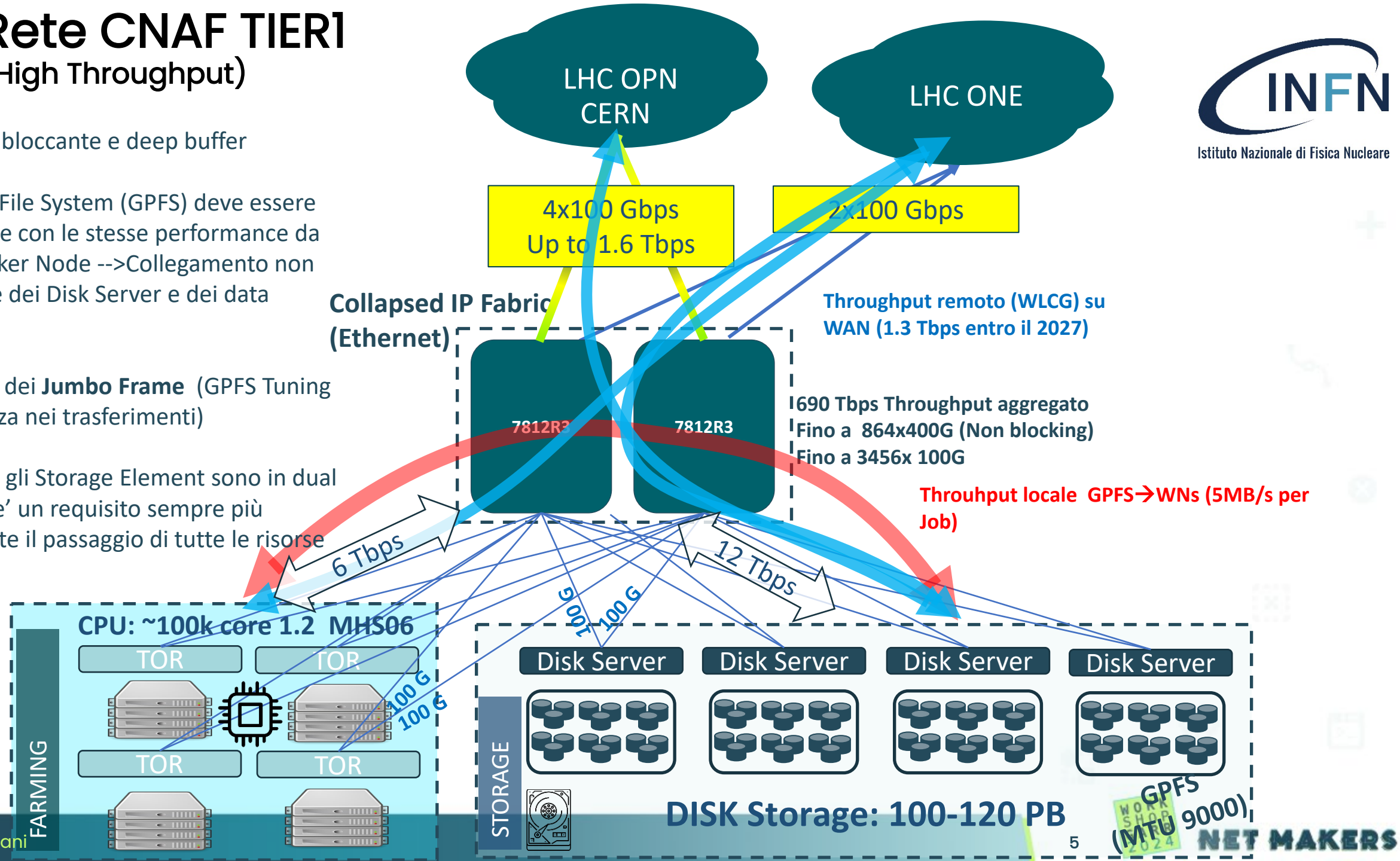
Rete CNAF TIER1
(High Throughput)

Rete non bloccante e deep buffer

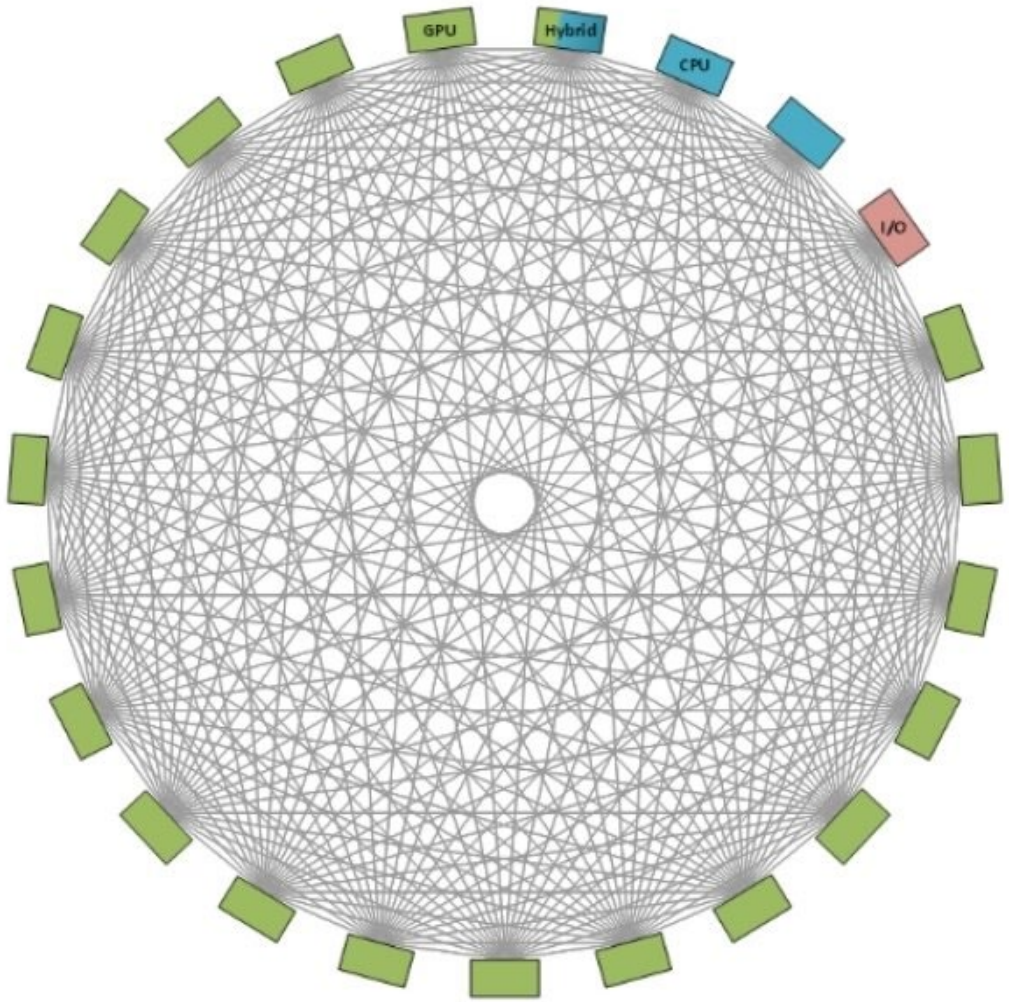
Il Cluster File System (GPFS) deve essere accessibile con le stesse performance da ogni Worker Node -->Collegamento non bloccante dei Disk Server e dei data mover

Supporto dei **Jumbo Frame** (GPFS Tuning e efficienza nei trasferimenti)

IPv6:Tutti gli Storage Element sono in dual stack ed e' un requisito sempre più Importante il passaggio di tutte le risorse su IPv6



Rete di Leonardo (Dragonfly+)



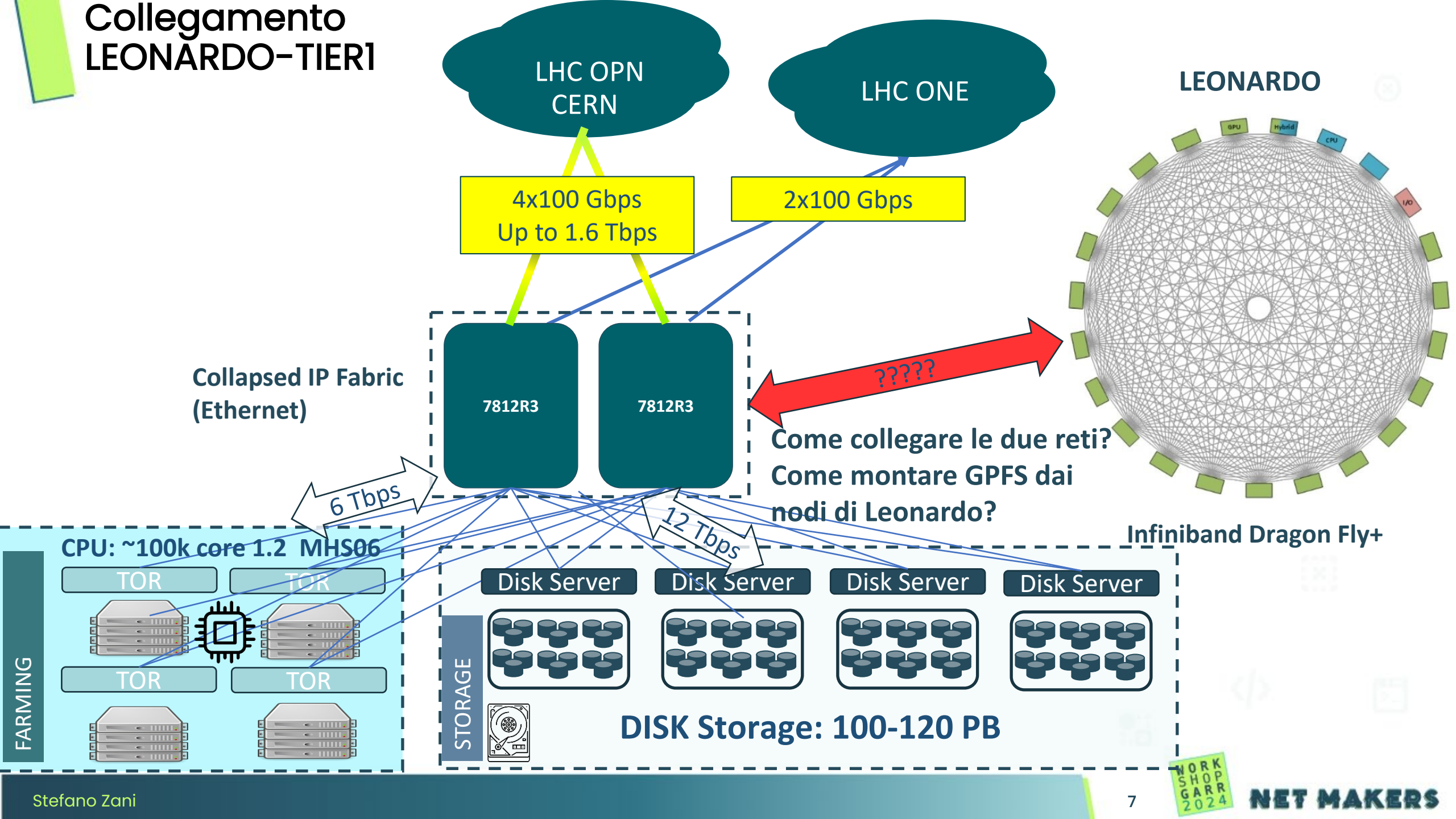
LEONARDO

Atos BullSequana XH2000
3456 nodes with 4x GPU A100-64
1536 nodes with 2x CPU General Purpose
Partition **Sapphire Rapids (56 core)**

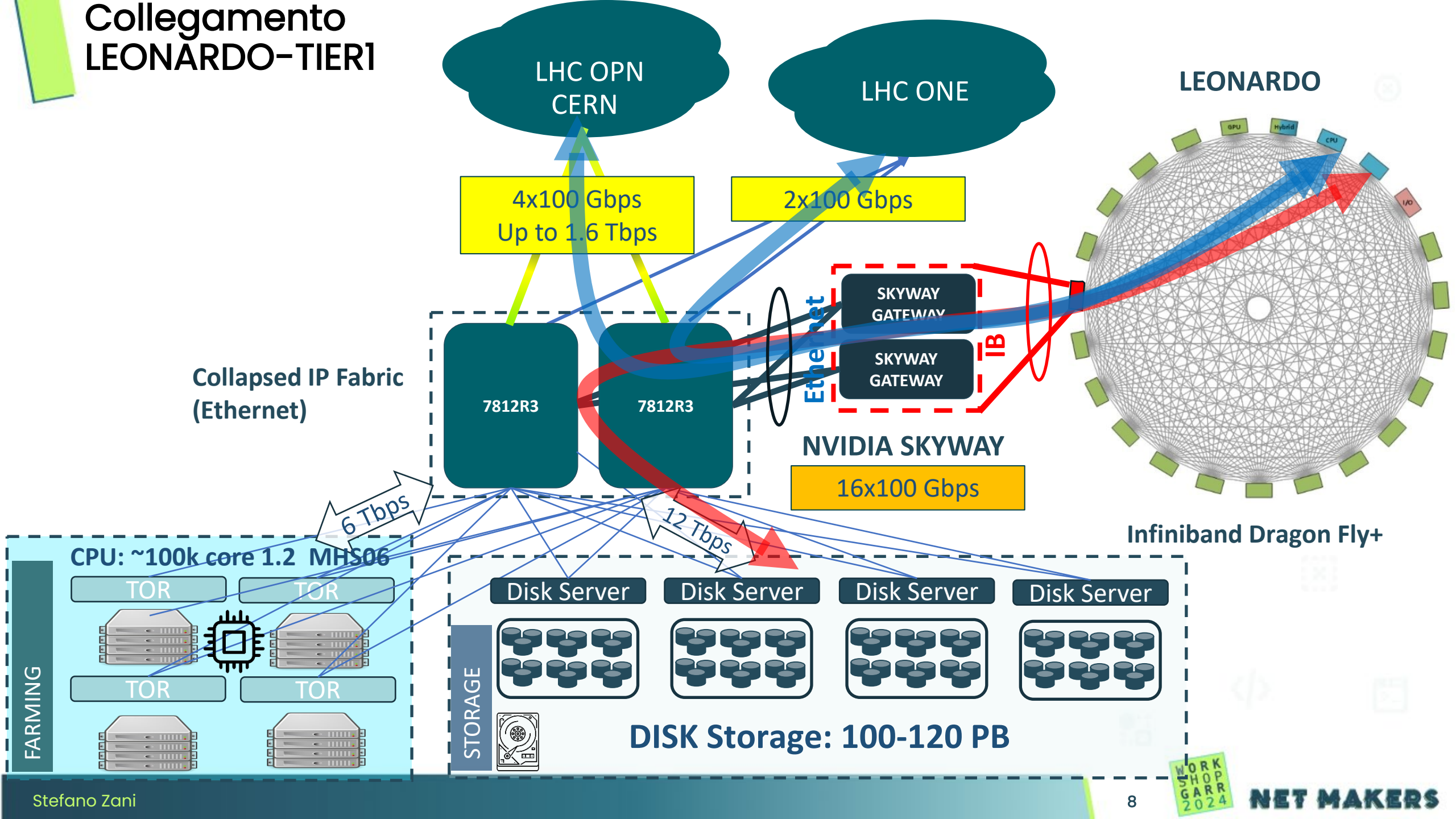
Rete basata su NVIDIA **Infiniband (HDR 200Gbps)**
Topologia: Dragonfly + ottimizzata per ridurre la latenza fra i nodi al minimo

NO Ethernet sui Nodi di calcolo

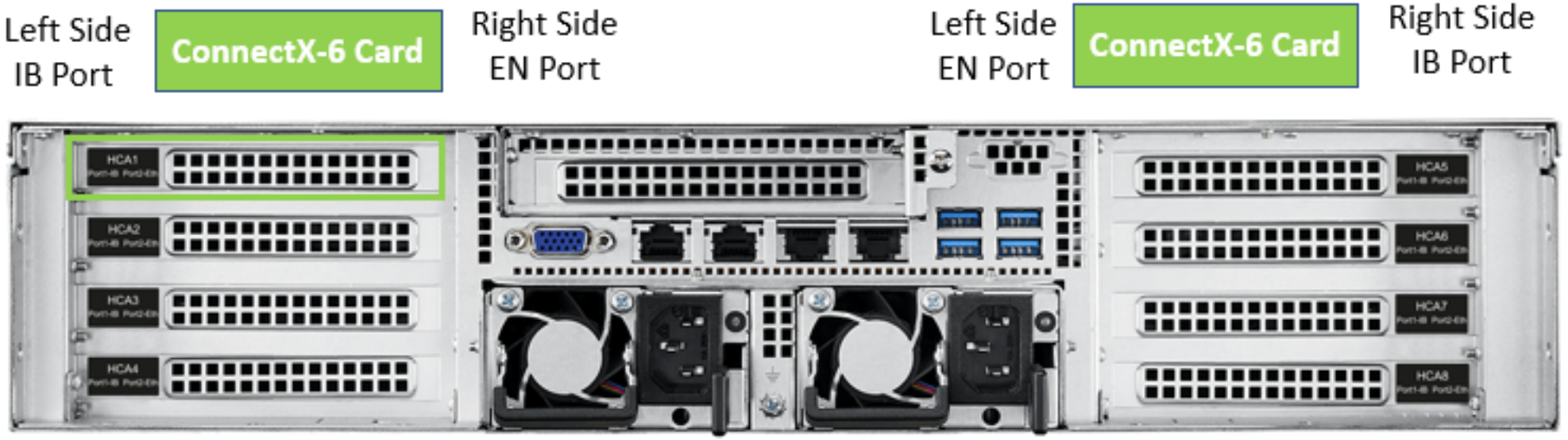
Collegamento LEONARDO-TIER1



Collegamento LEONARDO-TIER1

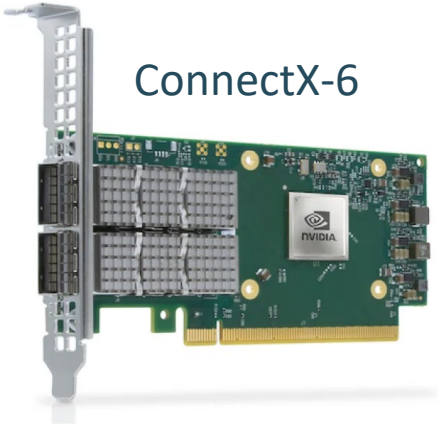


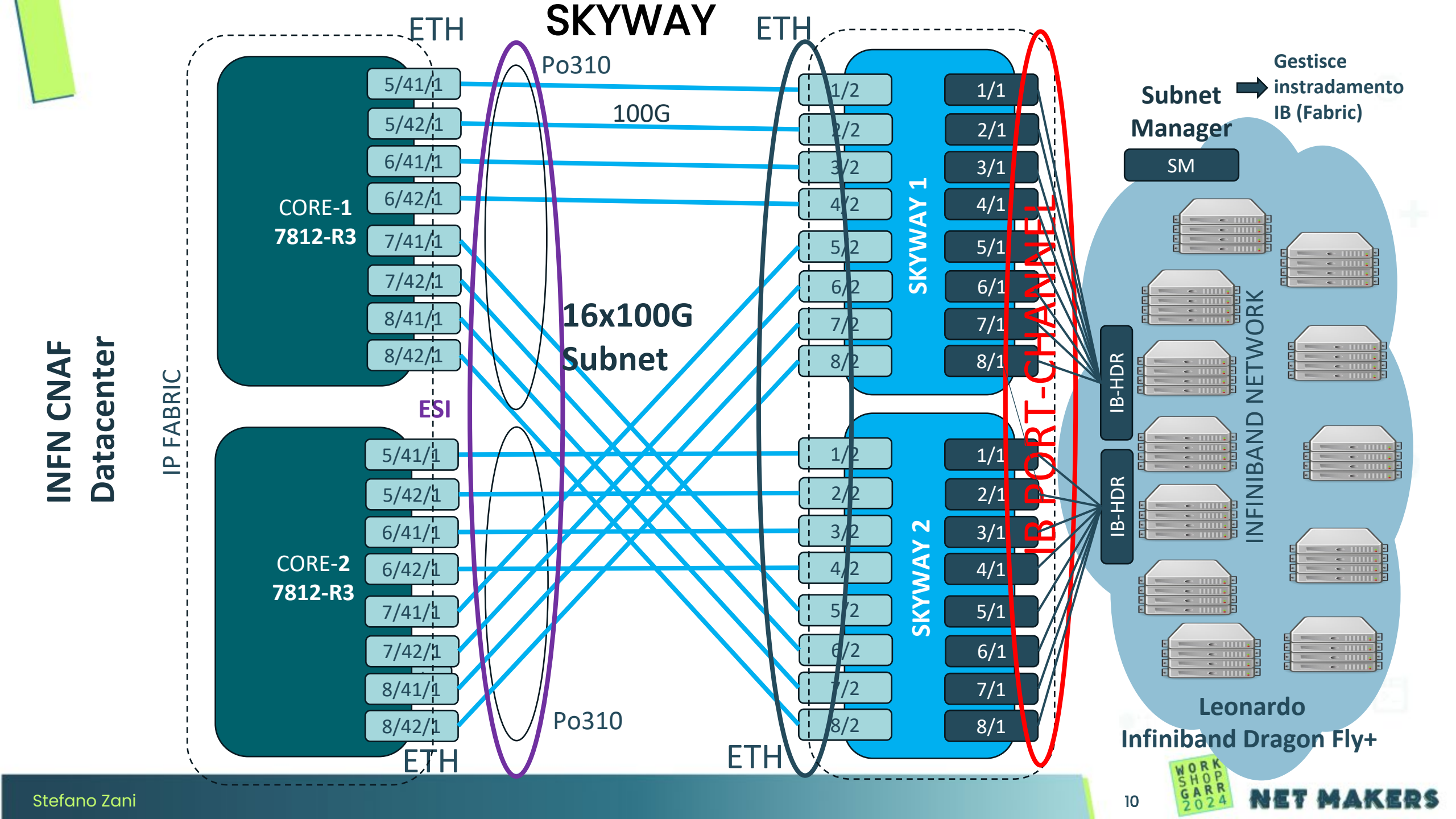
Skyway

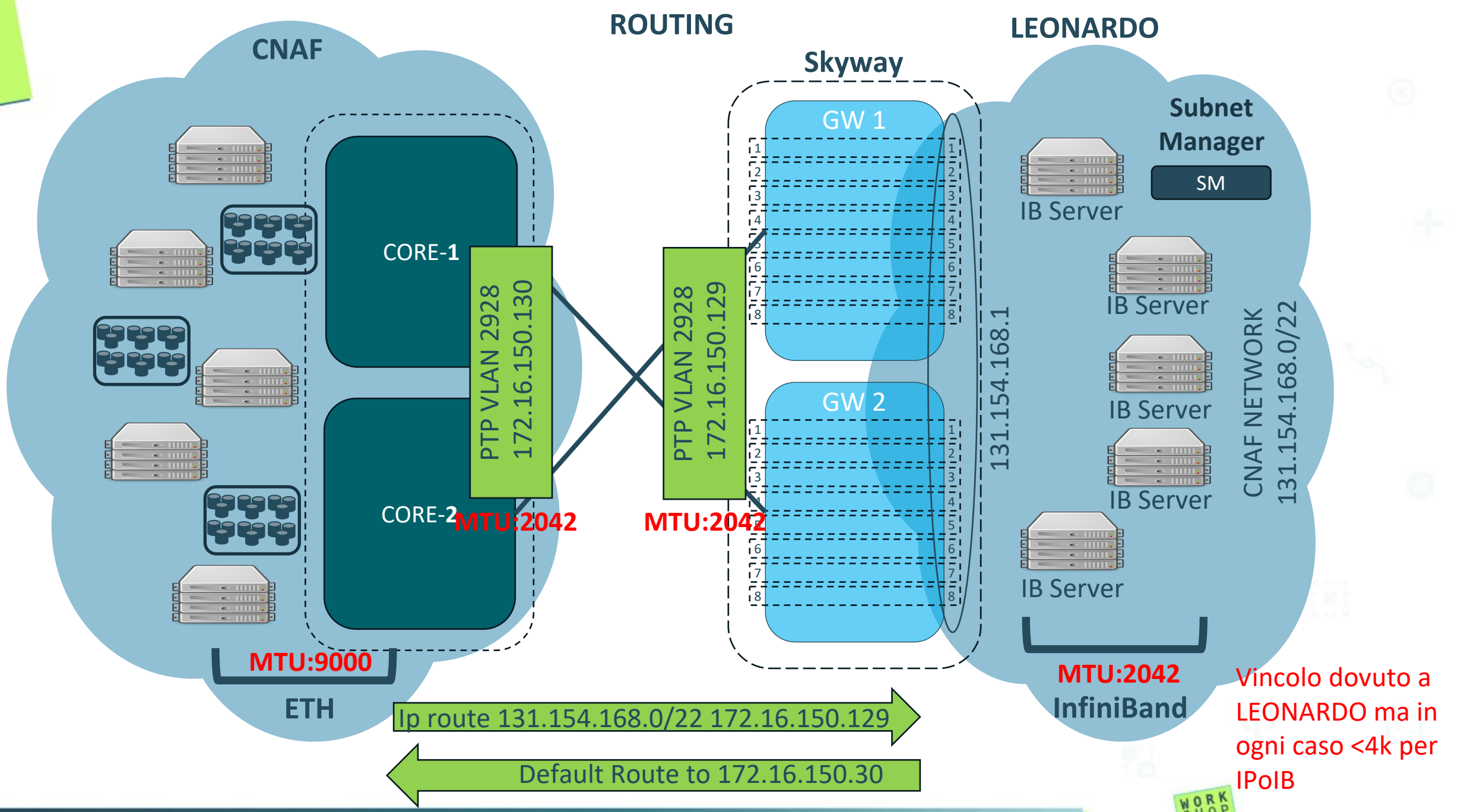


Gateway Infiniband/Ethernet

8 schede ConnectX6 HDR dual head (8x200/100Gb Ethernet, 8x200G Infiniband)
I flussi che entrano da una interfaccia Infiniband escono dalla interfaccia Ethernet della stessa scheda.
Il forwarding da Infiniband ad Ethernet e vice versa viene eseguito in HW dalla scheda stessa.




INFN CNAF
Datacenter
IP FABRIC
ETH
SKYWAY
ETH
Po310
100G
16x100G
Subnet
ESI
CORE-1
7812-R3
CORE-2
7812-R3
5/41/1
5/42/1
6/41/1
6/42/1
7/41/1
7/42/1
8/41/1
8/42/1
1/2
2/2
3/2
4/2
5/2
6/2
7/2
8/2
SKYWAY 1
SKYWAY 2
1/1
2/1
3/1
4/1
5/1
6/1
7/1
8/1
IB PORT-CHANNEL
Po310
ETH
ETH
Subnet
Manager
Gestisce
instradamento
IB (Fabric)
SM
IB-HDR
IB-HDR
INFINIBAND NETWORK
Leonardo
Infiniband Dragon Fly+



Vincolo dovuto a LEONARDO ma in ogni caso <4k per IPoIB

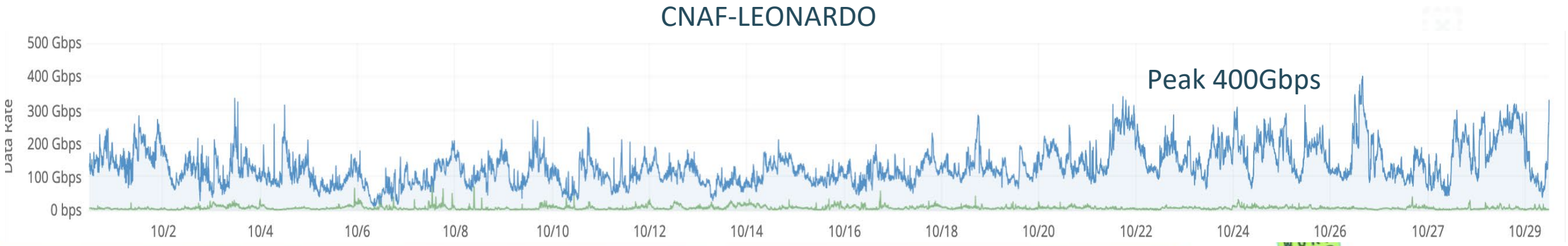
Attuale utilizzo dei nodi di Leonardo

200 Nodi della GP@Leonardo

- Dual 56 cores sockets Intel Sapphire Rapids
- 112 cores/nodo →TOTALE CORE: 22400
- (16 x 32) GB DDR5 4800 MHz
- 1680 Gflops/nodo (peak)
- 2880 HS06/node
- TOTALE HS06: 576 kHS06

L'effcienza dei Job in esecuzione su Leonardo è la stessa di quelli che girano in locale quindi per i nostri workload la soluzione funziona .

1.6 Tbps di interconnessione via Skyway A fronte di un throughput a regime stimato dell'ordine di 900 Gbps (considerando 5MBps per Job single core)



Principali Criticità di questo collegamento fra IB e Ethrenet

Essendo necessario attraversare questi apparati per raggiungere il mondo IP, rappresentano un point of failure ed un potenziale collo di bottiglia.

Non è possibile usare Jumbo frame con MTU 9000 (Limitata a 2k per ottimizzazione IB)

Per ora non è supportato IPv6 (In roadmap) ed è sempre più urgente: I siti WLCG devono passare totalmente ad IPv6 rapidamente viste anche le pressioni degli USA

> “This IPv6 policy must require that, no later than Fiscal Year 2023, all new networked Federal information systems must be IPv6-enabled at the time they are deployed. **Plus, the policy will state the agency's strategic intent to phase out the use of IPv4 for all systems**”.

- https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-07.pdf
- https://www.energy.gov/cio/department-energy-internet-protocol-version-6-ipv6-policy
- https://www.directives.doe.gov/directives-documents/200-series/0200-1-border-a-chg2-ltdchg

Quali tecnologie per HPC e AI datacenter ?

Infiniband è ancora leader nell'interconnessione di super calcolatori dedicati ad applicazioni che necessitano di **Ultra Low Lantency**

PRO

- 120ns (attraversamento switch port-port)
- 600ns End to End (ConnectX6)

CONTRO

Costoso

Vendor Lock-in

Quali tecnologie per HPC e AI datacenter ?

Ethernet però sta evolvendo velocememente ed oggi ha un vantaggio in termini di banda (800 Gb on market e 1600 Gb in arrivo)

Sulla latenza e la gestione della congestione ci sta lavotando

- **RDMA over converged Ethernet (RoCE v2)**
  - **PFC**:Priority-based Flow Control
  - **ECN**: Explicit Congestion Notification
  - **DCQCN:** (Data Center Quantized Congestion Notification)
- **Ultra Ethernet Consortium (UEC) https://ultraethernet.org/**
  - **Multi-path packet spraying** per migliorare in termini di latenza e Jitter della rete in caso di congestione.
  - **Flexible ordering**
  - **Congestione handling**

Infiniband vs RoCE v2

Se ci fosse una convergenza su Ethernet l'integrazione con gli HTC datacenter potrebbe essere semplificata

	InfiniBand	RoCE v2
End-to-End Delay	2us	5us
Flow Control Mechanism	Credit-Based Flow Control Mechanism	PFC/ECN, DCQCN
Forwarding Mode	Forwarding based on Local ID	IP-based Forwarding
Load Balancing Mode	Packet-by-Packet Adaptive Routing	ECMP Routing
Recovery	Self-Healing Interconnect Enhancement for Intelligent Data Centers	Route Convergence
Network Configuration	Zero Configuration through UFM	Manual Configuration

Considerazioni generali

G7 Conference on Large Research Infrastructures

27-30 Ottobre 2024

In tutti i paesi del G7 si sta **investendo** in "Grandi infrastrutture di ricerca" e fra le altre ci sono "Grandi Infrastrutture di Calcolo".

Il focus per quanto riguarda le grandi infrastrutture di calcolo è su **AI e Quantum**

L'Europa sta finanziando le "**AI Factory**" che saranno centri HPC dedicati alle applicazioni di Intelligenza Artificiale.

https://digital-strategy.ec.europa.eu/en/news/eu-boosts-european-ai-developers-ai-factories-call-proposals

https://eurohpc-ju.europa.eu/index_en

Conclusioni

I centri di calcolo HPC per AI, saranno enormi cluster di GPU interconnesse con reti a bassa latenza e soprattutto a riddottissimo Jitter ottimizzati per il trainig di reti neurali.

I nostri modelli di calcolo che utilizzeranno questo tipo di applicazioni, potranno beneficiare direttamente a pieno delle performance di questi Datacenter.

Per tutti i workload differenti, dovremo accedere in modo efficiente alle risorse che verranno messe a disposizione interfaccandosi ad alta velocità e nel modo migliore possibile.

Titoli di coda...

Questo mezzo ha 600 Cavalli

Anche questo mezzo ha 600 Cavalli

Titoli di coda...

Avere a disposizione strumenti molto potenti è esaltante

Non tutti i mezzi “Potenti” possono essere usati efficacemente per tutti gli impieghi.

WORKSHOP GARR 2024

NET MAKERS

Grazie per l'attenzione

per le domande: https://wooclap.com/e codice WSGARR24