



Ordine degli Ingegneri della Provincia di Roma

Commissione Data Center & Cloud

Corso Base sull'Intelligenza Artificiale

Lezione 8

“Architetture HW e sostenibilità energetica per la IA”

Ing. Mario D’Ettorre

Sabato 31 Gennaio 2026





AGENDA

1. Aspetti Generali e architetture dei Data Center;
2. CPU e GPU differenze ed interazioni per i compiti AI;
3. Digressione elettronica;
4. Infrastrutture di rete LAN e SAN;
5. L'architettura della rete LAN in presenza di sistemi per l'IA;
6. Conclusioni

1. Aspetti generali e architetture dei Data Center

I data center in Italia

La mappa attuale



Aree dove costruire nuovi impianti

In milioni di metri quadri

Bitonto e Giovinazzo (Bari)

8

Porto Torres (Sassari)

2,7

Cagliari, Capoterra, Assemini

1,5

Lecce

1,5

Verona

1,5

Aree con data center che si possono ampliare

In migliaia di metri quadri

Tortolì (Nuoro)

345

Acerra (Napoli)

250

Bari

250

Porto Torres (Sassari)

67

Piegara (Perugia)

65

Nel 2023

Valore data center
654 milioni di euro

Potenza

184 MW
Milano
che è la prima città italiana

430 MW

791 MW

Francoforte,
prima in Europa

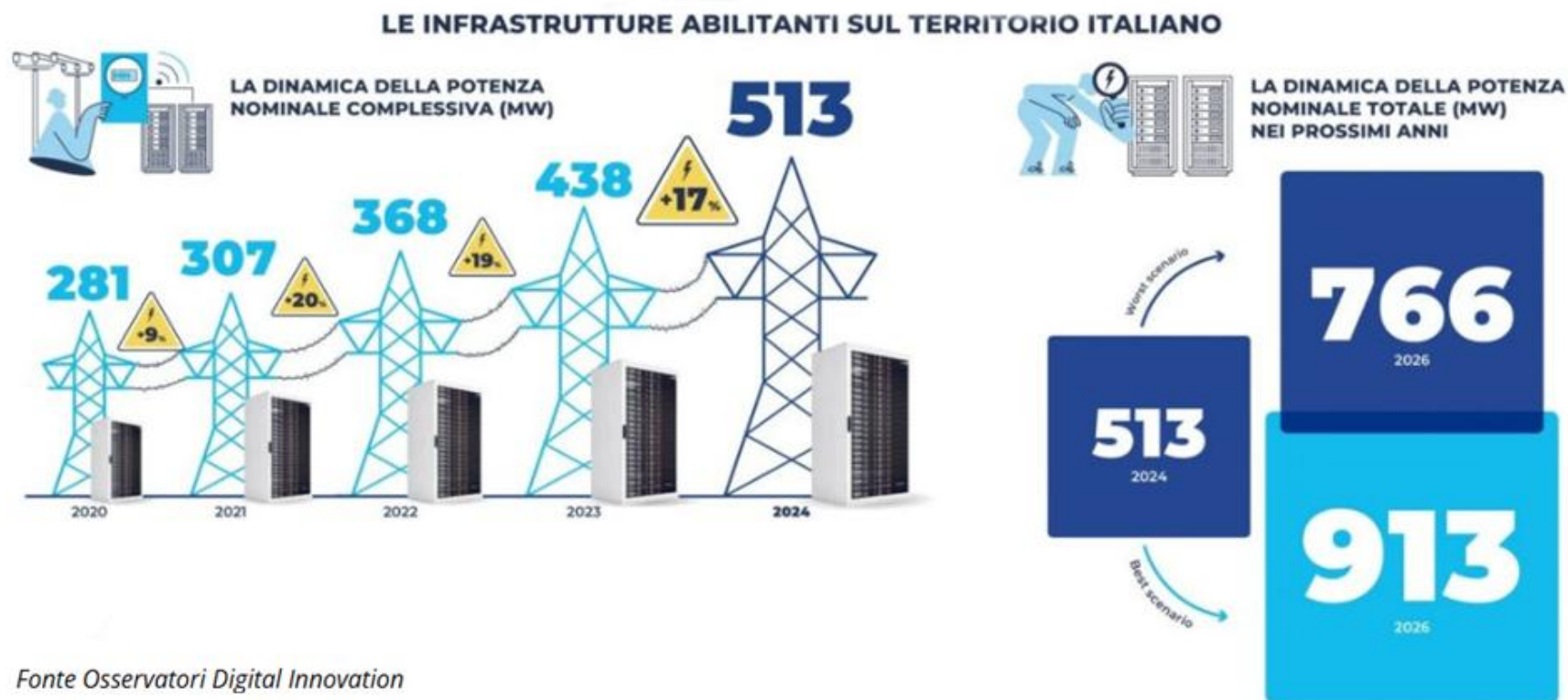
Withub

Fonte: Il Messaggero 2 Gennaio 2025 - Giacomo Andreoli

Ing. Mario D'Ettore

Corso IA, HW e sostenibilità energetica

1. Aspetti generali e architetturali dei Data Center



LA CRESCENTE RICHIESTA DI SERVIZI DIGITALI - SOCIAL MEDIA, E-COMMERCE, STREAMING, CLOUD COMPUTING, HOME BANKING, E-LEARNING E DA ULTIMO L'INTELLIGENZA ARTIFICIALE - DETERMINA LA RICHIESTA DI NUOVI DATA CENTER E DI CONSEGUENZA LA CRESCITA DELL'ENERGIA RICHIESTA PER ALIMENTARLI.

L'ENERGIA DEVE ESSERE PRESENTE SENZA INTERRUZIONI

Ing. Mario D'Ettorre

Corso IA, HW e sostenibilità energetica

1. Aspetti generali e architetture dei Data Center

BUILDING:

- Collocazione del sito
- Portata dei solai
- Altezza dei piani

POWER CONDITIONING E DISTRIBUTION:

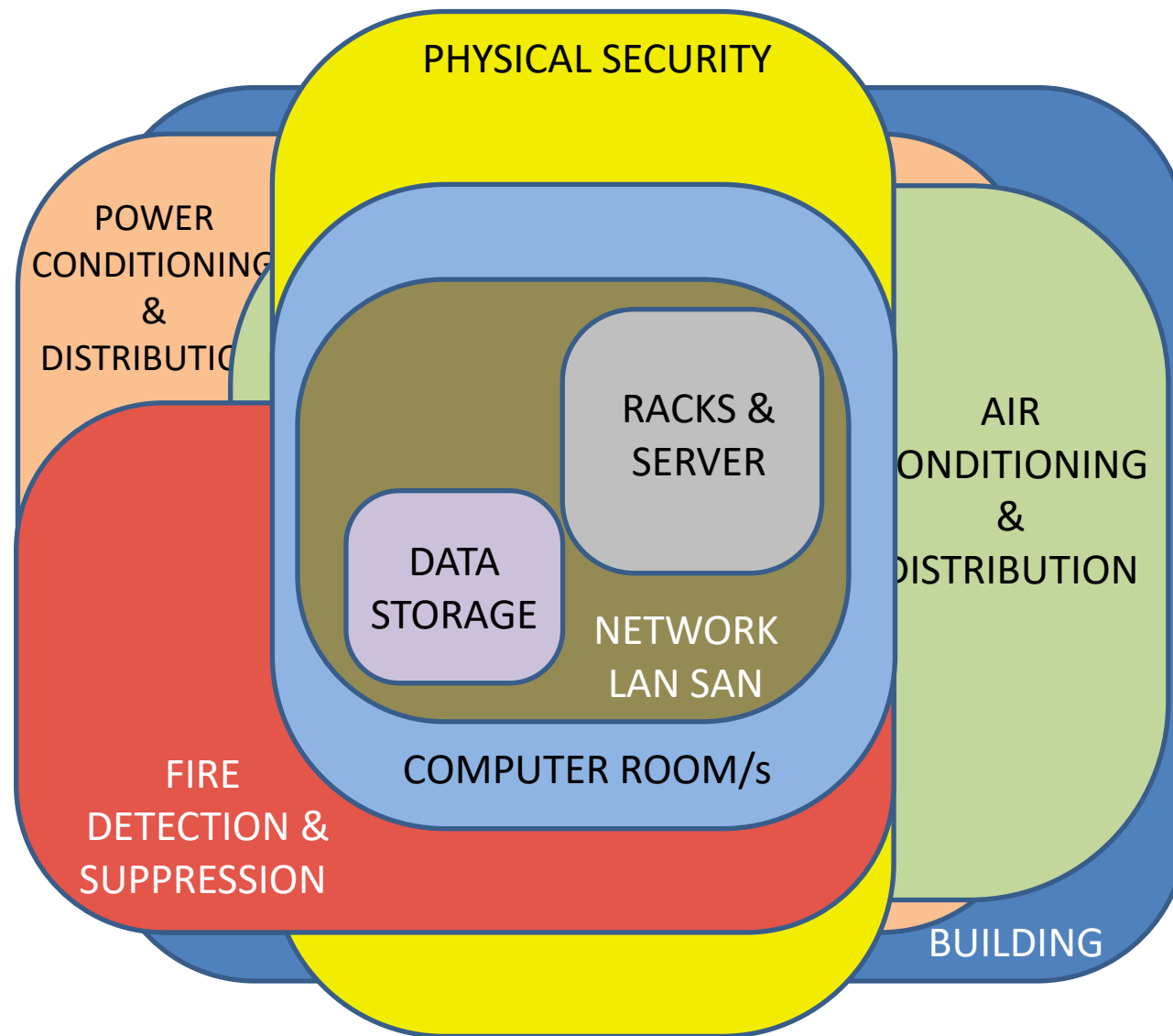
- Forniture in Media Tensione
- Cabine di trasformazione
- UPS
- Gruppi Elettrogeni
- Distribuzioni alle utenze elettriche in bassa tensione

FIRE DETECTION & SUPPRESSION:

- Impianto rilevazione incendi
- Impianto estinzione incendi
- Vie di fuga

AIR CONDITIONING & DISTRIBUTION

- Gruppi frigoriferi
- Piping
- Unità di condizionamento nelle sale server



PHYSICAL SECURITY:

- Controllo accessi al sito
- Controllo accessi alle sale server ed ai singoli rack
- Controllo Perimetrale

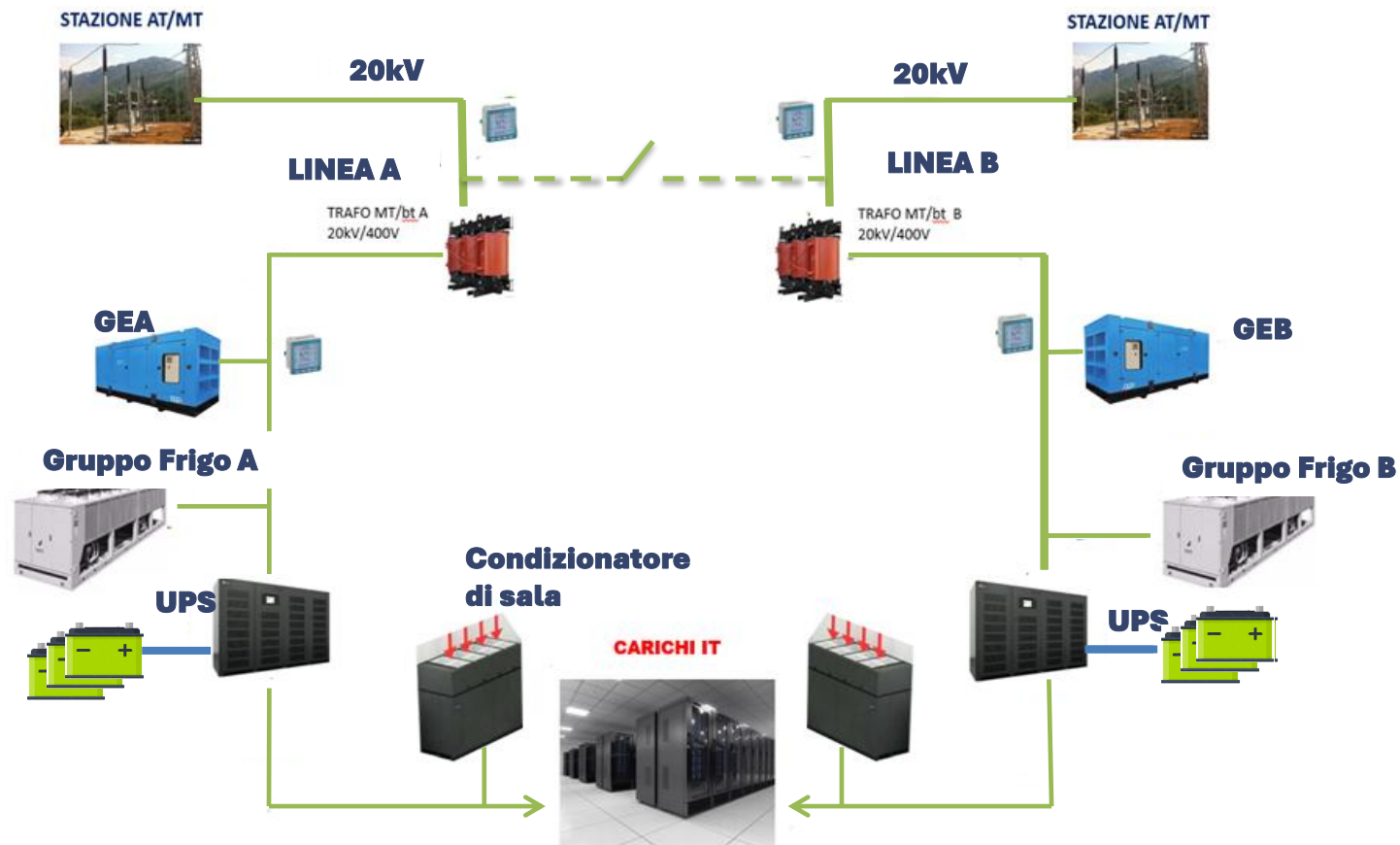
COMPUTER ROOM/s

- Allestimento dei rack: corridoi caldi e freddi
- Condizionamento perimetrale o ad Isole
- Pavimento flottante

NETWORK LAN SAN

- Cablaggio strutturato in rame fino a 10 Gbps fibra ottica per velocità superiori
- Disposizione degli switch

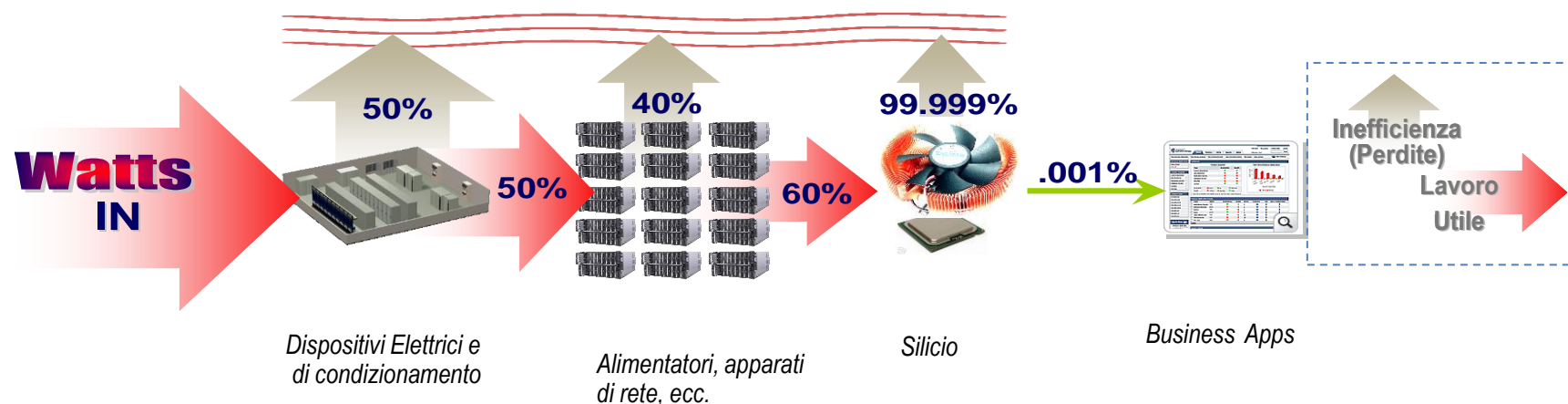
1. Aspetti generali e architetture dei Data Center



I GRANDI DATA CENTER ENTERPRISE HANNO IMPIANTI COMPLETAMENTE DUPLICATI OGNUNO DI CAPACITA' SUFFICIENTE PER IL FUNZIONAMENTO DEL DATA CENTER (DENOMINATI DI CLASSE TIER III / IV)

I Data Center sono dei divoratori di energia, che per la maggior parte viene utilizzata dagli impianti di supporto o dispersa sotto forma di calore.

Della parte di energia che arriva ai sistemi ICT il 99,999% è disperso in calore e solo lo 0,001% è utilizzato per la elaborazione delle informazioni.



Il calore generato deve essere estratto dai sistemi e dissipato, altrimenti si provocano malfunzionamenti e nei casi peggiori anche la distruzione dei circuiti.

1. Aspetti generali e architetturali dei Data Center

Nel 2007 nasce il consorzio Green Grid che definisce il **PUE** (Power Usage Effectiveness) attualmente armonizzato nella norma ISO/IEC 30134-2.

Tale parametro consente di **misurare l'efficienza energetica di un data center** ed è data dal rapporto fra tutta la potenza assorbita dal Data Center rispetto a quella assorbita dai soli apparati IT:

$$\text{PUE} = \frac{\text{Potenza totale}}{\text{Potenza IT}}$$

Minore è il valore del PUE (Più vicino ad 1) e maggiore è l'efficienza del Data Center

Il PUE si migliora ottimizzando l'efficienza di tutti gli impianti di servizio del Data Center, ma principalmente degli UPS e del Condizionamento che sono impianti energivori e sempre attivi.

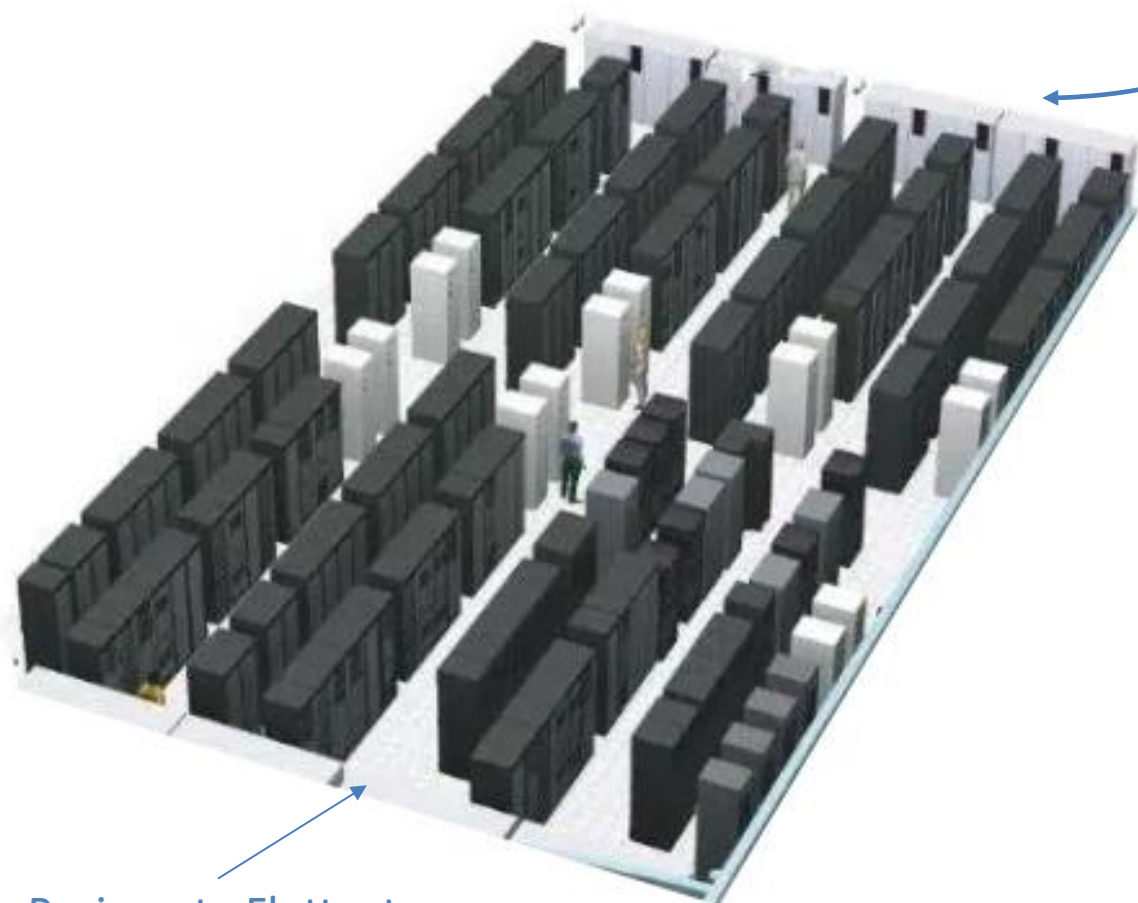
Un approccio oggi ampiamente utilizzato è quello di aumentare la temperatura di esercizio dell'impianto di condizionamento dai 18 °C - 20 °C del passato a temperature più alte, per esempio 26 °C, che sono accettabili dagli attuali sistemi IT

Ad oggi sono state introdotte altre metriche:

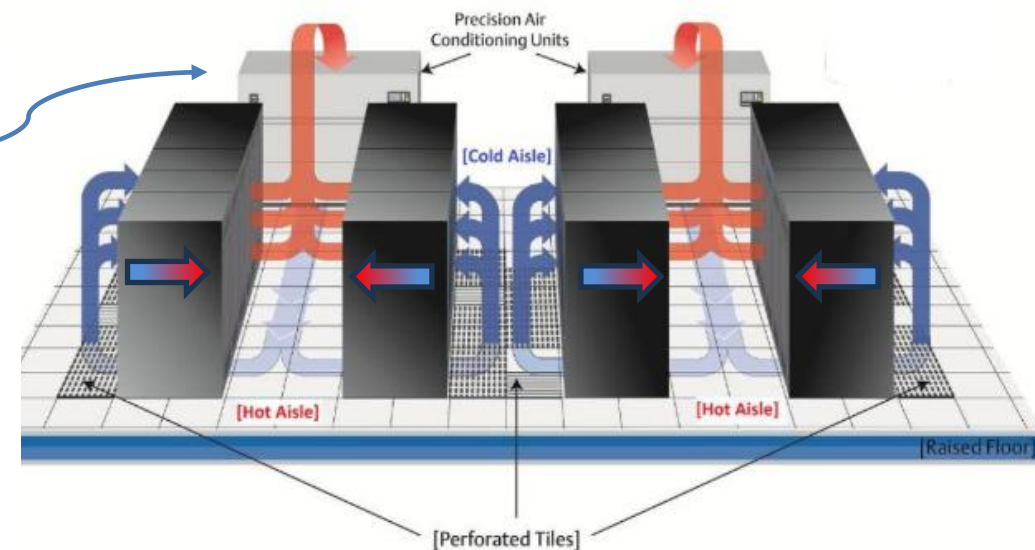
WUE: Water Usage Effectiveness

DCRE: Data Center Reusable Effectiveness che comprende tutte le risorse consumate dai Data Center

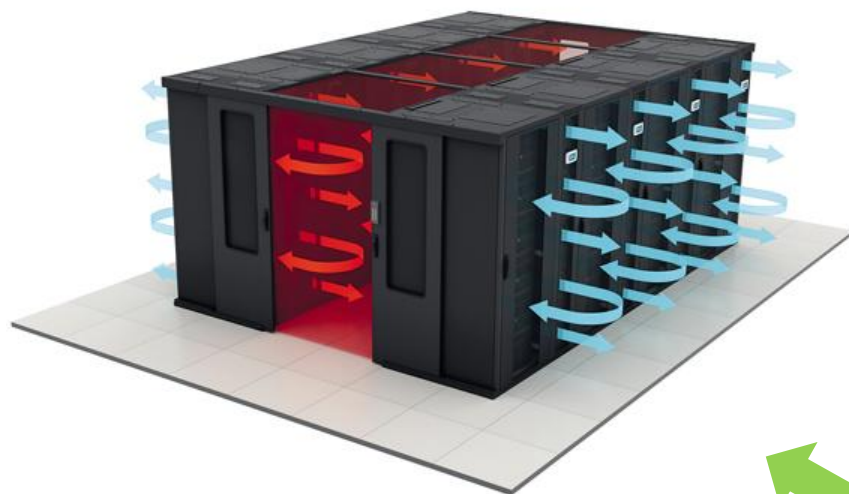
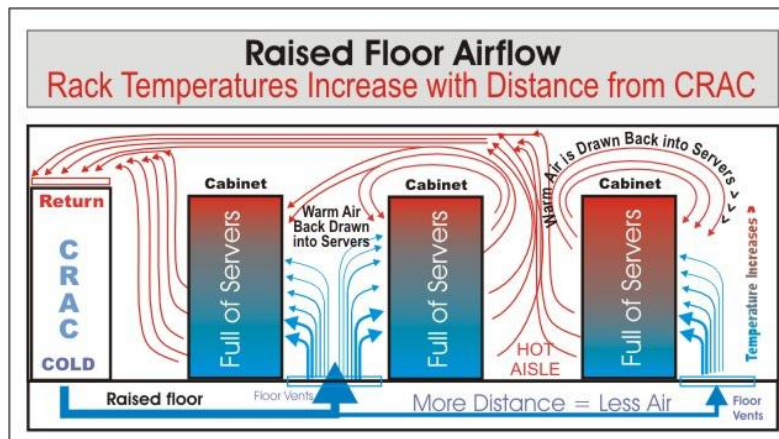
1. Aspetti generali e architettonici dei Data Center



Pavimento Flottante



Schema di raffreddamento classico con **condizionamento perimetrale** e aria fredda convogliata al di sotto del pavimento flottante «Plenum» e rack disposti back-to-back in modo da creare corridoi caldi e freddi.



Il condizionamento perimetrale è poco efficiente:

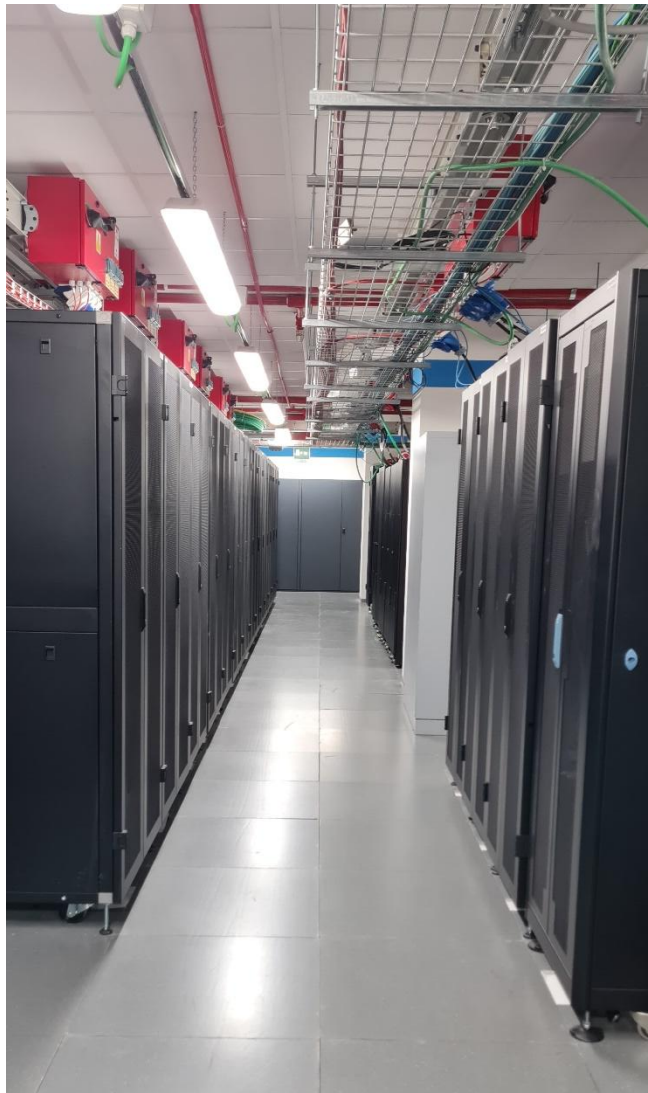
- i flussi di aria fredda non sono uniformi
- notevoli dispersioni di aria in tutto l'ambiente della sala server
- il plenum di aria fredda viene disturbato da elementi presenti al di sotto del pavimento flottante (cavi, tubi, ecc)

Nel tempo sono state realizzate nuove soluzioni di condizionamento più efficienti di cui le ultime generazioni sono denominate ad «**isole chiuse**» nelle quali il **corridoio caldo** o il **corridoio freddo** vengono chiusi e separati fisicamente dal resto della sala.

Un esempio è condizionamento «in row» nel quale i condizionatori (CRAH) sono sottili e alti quanto i rack e sono inseriti nella fila di rack a tutta altezza

Disegno condizionamento «in row»

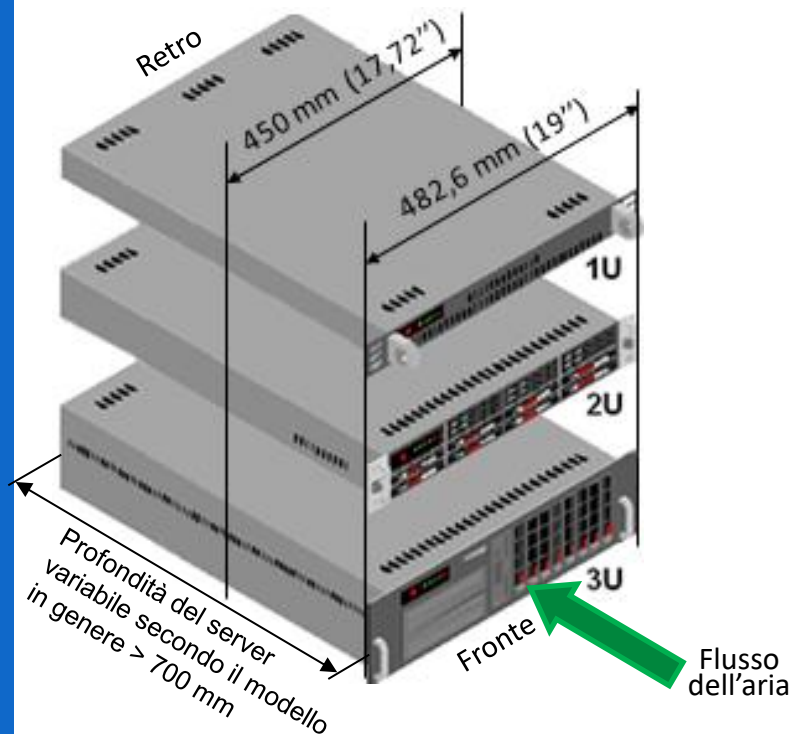
1. Aspetti generali e architetturali dei Data Center



Ing. Mario D'Ettorre

Corso IA, HW e sostenibilità energetica

1. Aspetti generali e architetturali dei Data Center



RETRO DI
UN RACK



- La potenza usata con le densità attuali rack è 7 - 10 kW per rack (@ 32 A monofase – 16 A trifase)
- Il flusso dell'aria di raffreddamento è sempre dal fronte al retro
- I cavi di interconnessione sono sempre sul retro del rack
- I server sono montati su slitte che consentono l'estrazione per le manutenzioni

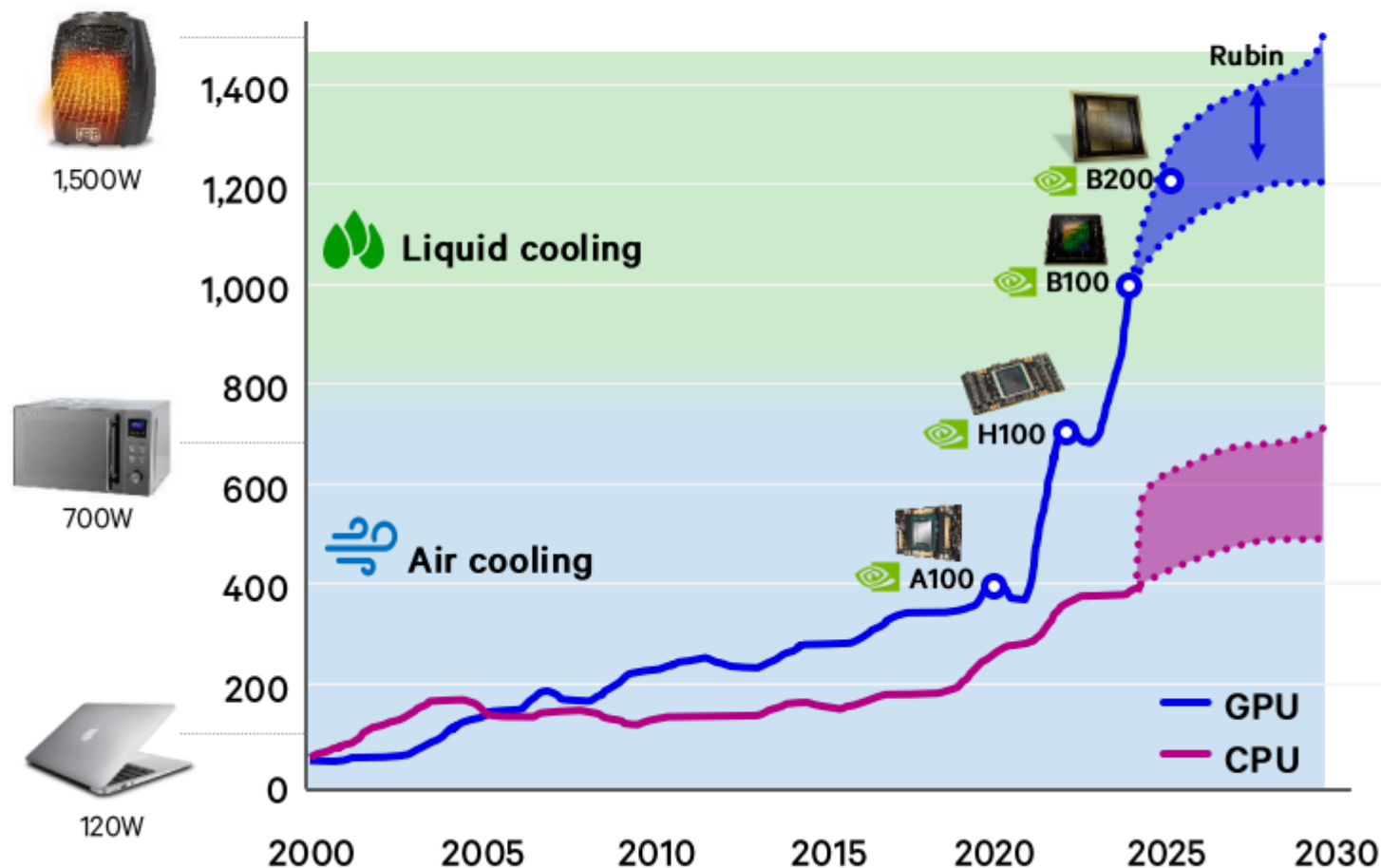
Ing. Mario D'Ettore

Corso IA, HW e sostenibilità energetica

LA DENSITÀ DI
POTENZA DELLE GPU
GIÀ OGGI RICHIEDE IL
RAFFREDDAMENTO A
LIQUIDO «ON CHIP» E
LA RICHIESTA SARÀ
SEMPRE PIÙ SPINTA
CON PROBABILE
RICORSO ALLA
IMMERSIONE TOTALE
NEL LIQUIDO

CPU and GPU power consumption forecast Thermal Density Power - TDP (watts)

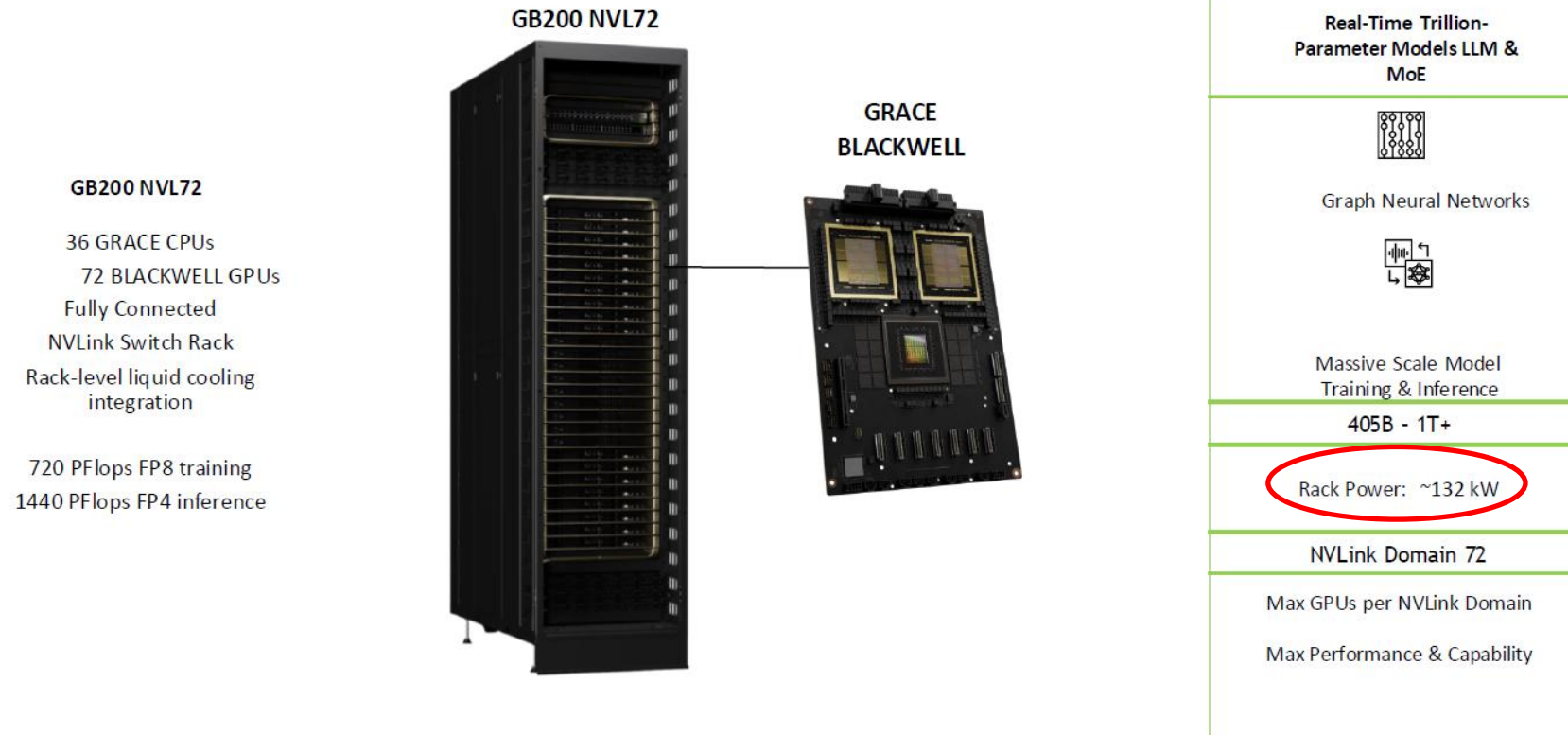
OMDIA



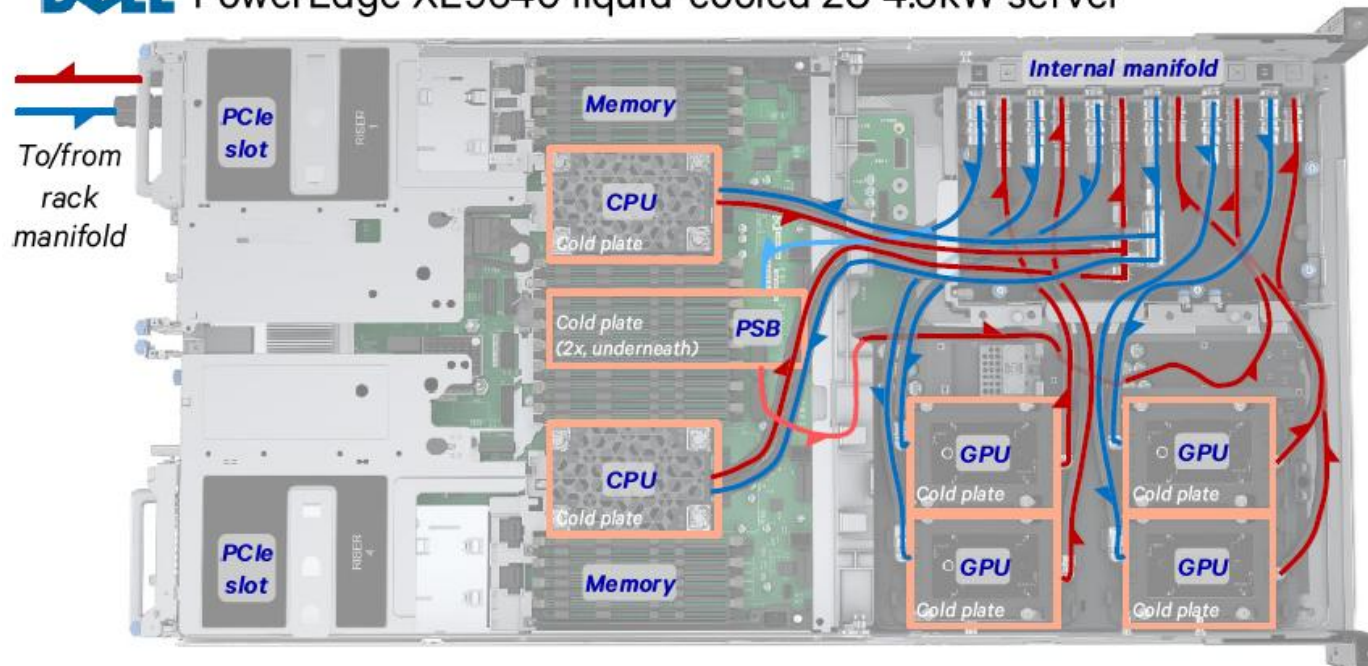
1. Aspetti generali e architetturali dei Data Center

The Future of NVIDIA AI Infrastructure

Accelerated computing innovations enabling efficient deployment of trillion-parameter-scale AI



DELL PowerEdge XE9640 liquid-cooled 2U 4.6kW server



Liquid-cooled components

- ✓ GPUs
- ✓ CPUs
- ✓ PSB



Air-cooled components

- ✓ PSUs
- ✓ Memory
- ✓ Storage
- ✓ PCIe cards



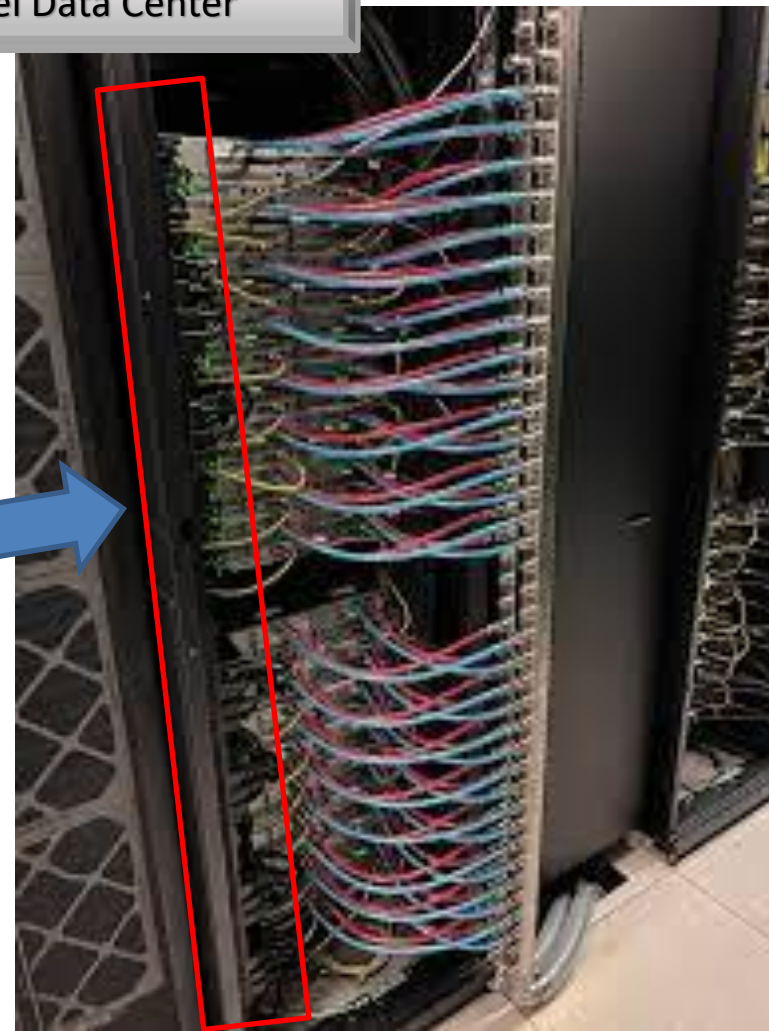
In general, cold plates capture 70-90% of server heat depending on server and cold plate design.

I rack per AI dedicati a grandi modelli LLM richiedono una potenza 40 – 80 kW destinata ad aumentare.

Le alte potenze in gioco richiedono profonde modifiche alle infrastrutture tecnologiche:

- Impianto di raffreddamento:
 - disponibilità di raffreddamento a liquido «direct to chip» per le GPU e CPU o a tendere per immersione
 - incremento del condizionamento convenzionale per i flussi di aria fredda che devono raffreddare gli altri componenti dei server (memoria, chip interni di supporto, alimentatori, ecc.)
- Impianto elettrico
 - Cabine di trasformazione, UPS, G.E.
 - Quadri elettrici
 - Cablaggi adeguati alle nuove correnti
 - PDU nei rack adeguate alle nuove richieste di prese elettriche

PDU



Manifold per la distribuzione del liquido di raffreddamento ai server

Data Center per implementazione massiva della IA

- Necessità di potenza elettrica adeguata – allo stato attuale anche 80 kW per rack, a tendere anche 130 -150 kW per rack
- Impianto elettrico adeguato per veicolare ai rack server le potenze richieste (blindobarre di grande potenza)
- Gruppi elettrogeni e sistemi di continuità assoluta (UPS) di capacità adeguata
- Impianto di Raffreddamento ad acqua refrigerata più complesso, a liquido Direct to CHIP e raffreddamento classico ad aria per il resto dei componenti dei server.
- Allestimento delle sale server ad Isole chiuse per una maggiore efficienza energetica
- Infrastruttura di rete adeguata a livello del cablaggio di rete, eliminando il rame e utilizzando esclusivamente fibra ottica magari di tipo monomodale che non ha limitazioni di lunghezza



AGENDA

1. Aspetti Generali e architetture dei Data Center;
2. CPU e GPU differenze ed interazioni per i compiti AI;
3. Digressione elettronica;
4. Infrastrutture di rete LAN e SAN;
5. L'architettura della rete LAN in presenza di sistemi per l'IA;
6. Conclusioni



Perché sono richieste le GPU ?

L'AI generativa richiede calcoli simili su grandi quantità di dati che la GPU consente di eseguire in parallelo ottimizzando i tempi di risposta.

Infatti le stesse operazioni se fossero eseguite sulle normali CPU in maniera sequenziale richiederebbero decine se non centinaia di ore.

Le fasi dell'AI sono Addestramento (Training) e Inferenza (Inference).

Impatto delle fasi sulle GPU

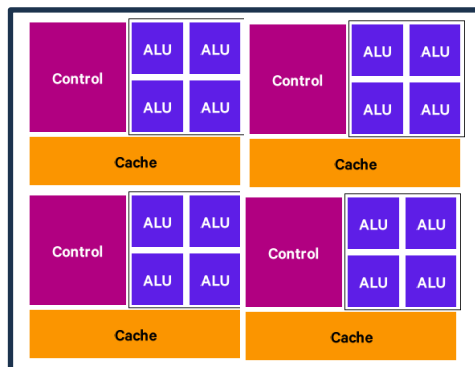
Aspetto	Training	Inferenza
Complessità	Alta (backprop, ottimizzazione)	Media (solo forward pass)
Precisione	FP32/FP16/TF32	FP16/INT8
Carico computazionale	Massivo, multi-GPU	Leggero se batch piccolo, intenso se su larga scala
Operazioni chiave	Fwd + Bwd pass, ottimizzatori	Solo Fwd pass



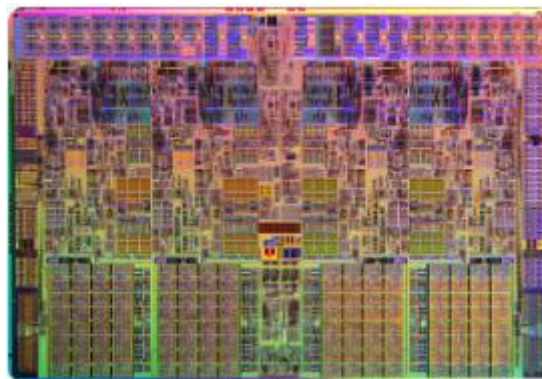
	CPU	GPU
Scenario	Executing a single GPT-3 inference operation taking 350 TFLOP.	
Capability	<i>A processor core can typically execute 1-2 instructions per cycle. Processor clock rates have been stable around 3 GHz.</i>	<i>NVIDIA A100 GPU has 512 tensor cores that can perform a 4x4 matrix multiplication in a single cycle, with a nominal capacity of 312TFLOPs.</i>
Total processing time	$\frac{350 \text{ TFLOP}}{3 \text{ GHz} * 1 \text{ FLOP}} = \sim 32 \text{ hours}$	$\frac{350 \text{ TFLOP}}{312 \text{ TFLOPs}} = \sim 1 \text{ second}$

Source: Andressen Horowitz

CPU



Chip fisico



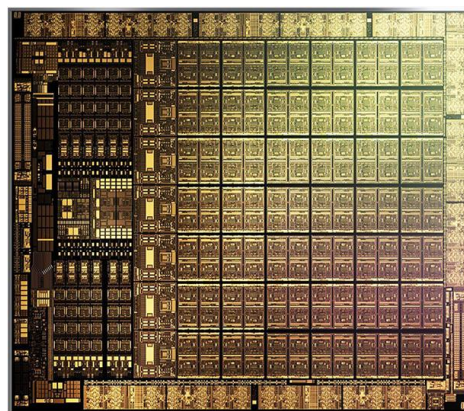
- **Componente per usi generali progettato per esecuzione di compiti sequenziali**
- Costituito di pochi Core molto evoluti e in grado di eseguire istruzioni complesse
- Controlla il flusso del programma, gestendo la GPU in modo efficiente
- Finalizza e restituisce i risultati agli utenti

Se la CPU non riesce a tenere il passo, le GPU rimangono inattive, sprecando energia e cicli di elaborazione.

GPU



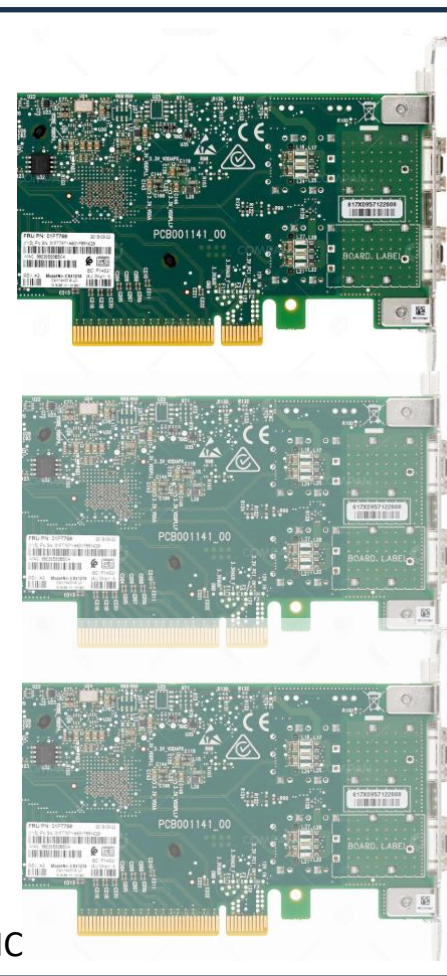
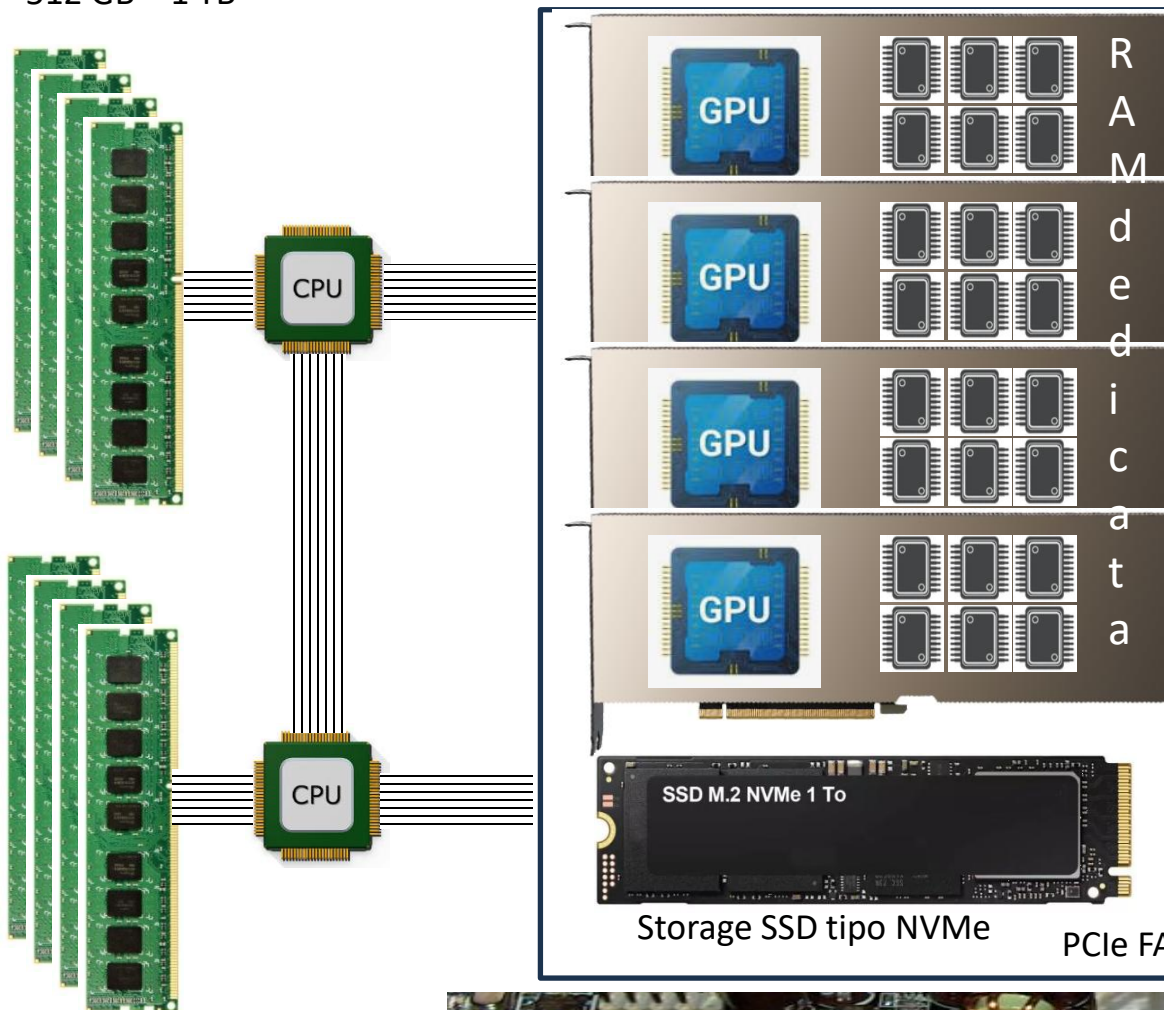
Chip fisico



- **Componente costituito da centinaia di core semplici che lavorano in parallelo**
- I core sono semplici, ottimizzati per eseguire operazioni aritmetiche ad alta velocità
- **Dato che tutti i core sono contemporaneamente attivi il componente richiede molta potenza elettrica**

Server per AI

RAM ECC
512 GB – 1 TB



Scheda LAN
di accesso
10/25 GBE

Scheda per link
GPU <-> GPU
per cluster GPU
200/400 GBE
(eventuale)

Scheda
collegamento
alla SAN
(eventuale)



PSU 2 ~ 3 kW



Ing. Ma

Corso IA, HW e sostenibilità energetica

Il Duetto CPU-GPU nell'IA – Una Prospettiva di Rete Interna

L'efficacia dei sistemi di IA dipende in modo critico dalla fluidità ed efficienza con cui CPU e GPU interagiscono.

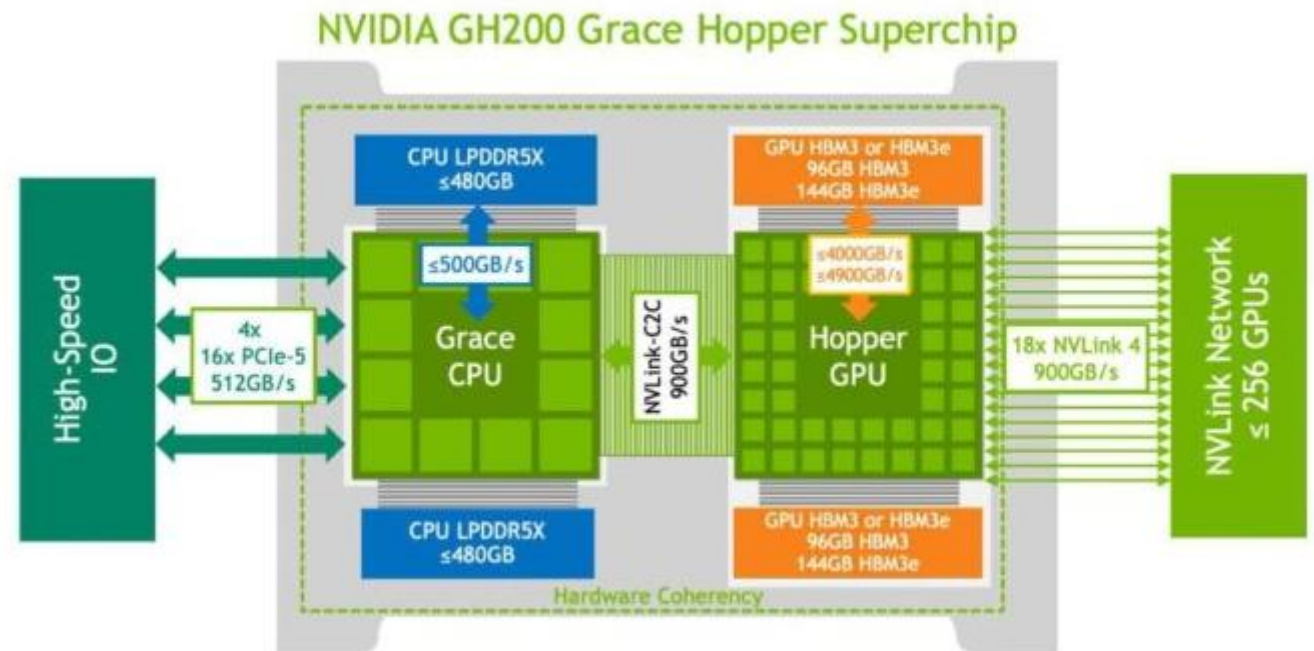
Nei server standard l'interazione avviene attraverso il bus PCIe:

- Il PCIe è costituito di 16 canali (lane) seriali full duplex ad alta velocità.
- La versione PCIe 5.0 prevede 16 lane per ~4GB/s per un totale di ~64GB/s se si sfruttano tutte le 16 lane

Per superare i limiti di larghezza di banda di PCIe, NVIDIA ha sviluppato NVLink, un'interconnessione ad alta velocità per la comunicazione tra GPU e, in architetture specifiche come Grace Hopper, per la comunicazione CPU-GPU.²⁴ NVLink offre una connessione punto-punto a latenza ultra-bassa.³⁵

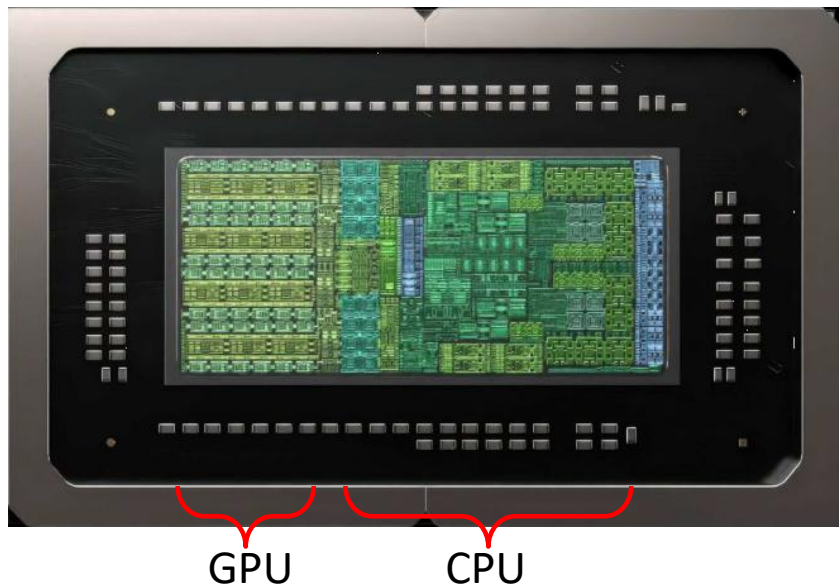


NVIDIA Grace Hopper At QCT Computex 2023 2



NVIDIA GH200 Diagram September 2024

Superchip CPU-GPU nell'IA



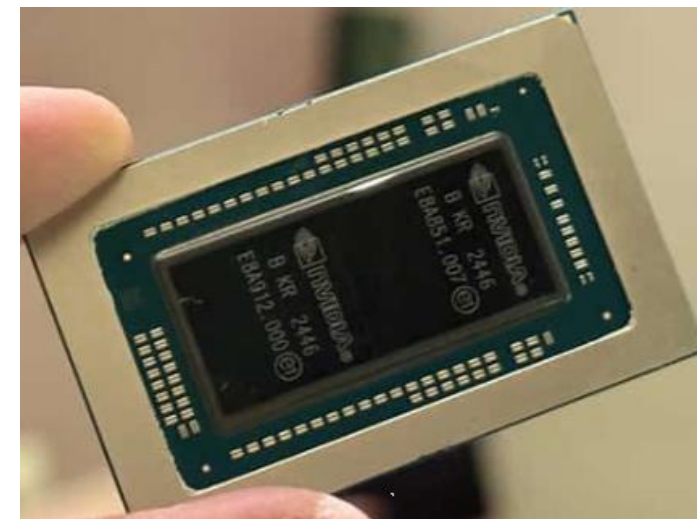
NVIDIA ha creato il superchip GB10 Grace Blackwell (CPU – GPU) dove la CPU e la GPU sono integrati insieme e interconnessi tramite l'interfaccia chip-to-chip NVLink-C2C.

Il Superchip presenta i core CUDA di ultima generazione di Nvidia e i core tensori di quinta generazione, offrendo una potenza fino a 31 TFLOPs FP32 e 1.000 TOPS FP4

La CPU Nvidia Grace è ad alta efficienza e include 20 core ad alte prestazioni basati sull'architettura ARM.

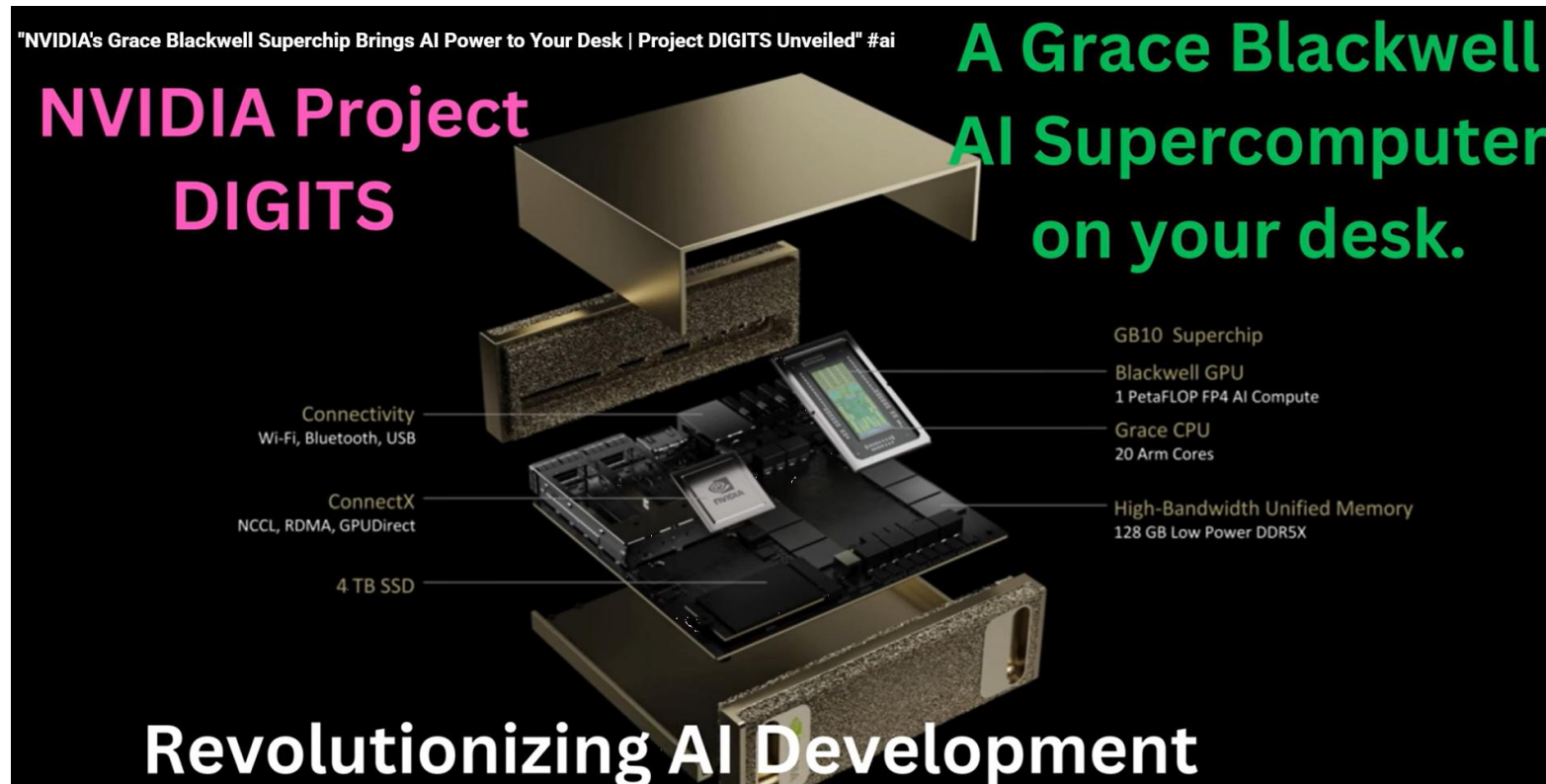
Il GB10 Grace Blackwell Superchip è realizzato con una tecnologia a 3 nm adottando criteri e tecnologie per ridurre i consumi in modo da poter essere utilizzato in sistemi desktop-like

Il Thermal Design Power THD dichiarato è di soli 140 W.



Il sistema NVIDIA DGX SPARK che contiene il Superchip GB10

- La **DGX Spark** è pensata per lo sviluppo locale di modelli di intelligenza artificiale.
- **Usa lo stesso stack software DGX** (CUDA, cuDNN, NCCL, framework AI, container NVIDIA) delle piattaforme più grandi
- Le applicazioni sviluppate possono essere testate e portate direttamente senza modifiche sostanziali.



Oltre a NVIDIA il sistema è prodotto da Acer, Asus, Dell, Gigabyte, HP, Lenovo, MSI.

Ing. Mario D'Ettore

Corso IA, HW e sostenibilità energetica

Confronto con le GPU Blackwell di fascia alta

Technical Specifications ¹		
	GB200 NVL72	HGX B200
Blackwell GPUs Grace CPUs	72 36	8 0
CPU Cores	2,592 Arm Neoverse V2 Cores	-
Total FP4 Tensor Core	1,440 petaFLOPS	144 petaFLOPS
Total FP8/FP6 Tensor Core	720 petaFLOPS/petaOPS	72 petaFLOPS/petaOPS
Total Fast Memory	Up to 30TB	Up to 1.4TB
Total Memory Bandwidth	Up to 576TB/s	Up to 62TB/s
Total NVLink Bandwidth	130TB/s	14.4TB/s
Individual Blackwell GPU Specifications		
FP4 Tensor Core	20 petaFLOPS	18 petaFLOPS
FP8/FP6 Tensor Core	10 petaFLOPS	9 petaFLOPS
INT8 Tensor Core	10 petaOPS	9 petaOPS
FP16/BF16 Tensor Core	5 petaFLOPS	4.5 petaFLOPS
TF32 Tensor Core	2.5 petaFLOPS	2.2 petaFLOPS
FP32	80 teraFLOPS	75 teraFLOPS
FP64/FP64 Tensor Core	40 teraFLOPS	37 teraFLOPS
GPU Memory Bandwidth	186GB HBM3e 8 TB/s	180GB HBM3e 7.7 TB/s
Multi-Instance GPU (MIG)	7	
Decompression Engine	Yes	
Decoders	7 NVDEC ² 7 NVJPG	
Max Thermal Design Power (TDP)	Configurable up to 1,200W	Configurable up to 1,000W
Interconnect	5th Generation NVLink: 1.8TB/s PCIe Gen5: 128GB/s	



Confronto GPU per AI: Potenza assorbita e prestazioni

GPU	Architettura	TDP (Watt)	Precisione AI TOPS/FLOPS	Core (CUDA/SP/AI)	Contesto d'uso
NVIDIA H100 SXM	Hopper	700 W	4,000+ TOPS INT8 ~60 TFLOPS FP64	16896	Training + Inference LLM
NVIDIA A100 80GB	Ampere	400 W	~312 TFLOPS FP16 624 INT8 TOPS	6912	Training LLM, HPC
NVIDIA L40S	Ada Lovelace	350 W	~1,466 INT8 TOPS	18176	Inference + grafica AI
NVIDIA RTX 4090	Ada Lovelace	450 W	~1,321 INT8 TOPS ~82 TFLOPS FP16	16384	AI prosumer / local inferenza
AMD MI300X	CDNA 3	750 W	~1,000 TFLOPS FP16 1,600 INT8 TOPS	19456	AI training su larga scala
Intel Gaudi 2	Custom Habana	600 W	~1,000 INT8 TOPS	32	AI inferenza e training
NVIDIA Jetson Orin AGX	Ampere (embedded)	15–60 W	~275 INT8 TOPS	2048	Edge AI

Ing. Mario D'Ettore

Corso IA, HW e sostenibilità energetica



AGENDA

1. Aspetti Generali e architetture dei Data Center;
2. CPU e GPU differenze ed interazioni per i compiti AI;
3. Digressione elettronica;
4. Infrastrutture di rete LAN e SAN;
5. L'architettura della rete LAN in presenza di sistemi per l'IA;
6. Conclusioni

Come sono realizzati i moderni chip digitali ?

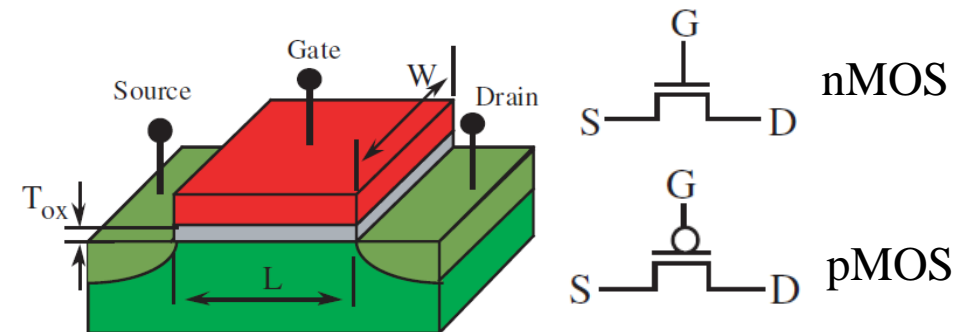
L'evoluzione tecnologica dei circuiti digitali ha una lunga storia che inizia dalle valvole termoioniche.

- Il transistor ha consentito di iniziare la miniaturizzazione e la realizzazione dei circuiti integrati su singolo chip di silicio. Questi circuiti sono stati usati estensivamente per tutti gli anni 60 e 70;
- Negli anni successivi si è avuta la svolta con l'avvento del **transistor Metallo-Ossido-Semiconduttore (MOS)**, che ha dato luogo alla creazione di circuiti integrati a Larga e Larghissima scala (LSI e VLSI)*

I transistor MOS sono tutt'ora la tecnologia dominante e **non ci sono tecnologie alternative pronte per la produzione di massa.**

I transistor MOS sono costituiti da un canale e da un Gate di controllo isolato fisicamente dal canale. Applicando un potenziale elettrico al Gate, il canale si chiude conducendo corrente elettrica.

I transistor MOS sono di due tipologie, **canale N (nMOS)** che conduce con potenziale positivo superiore ad una certa soglia e **canale P (pMOS)** che conduce con potenziale positivo inferiore ad una certa soglia.



* Un processore Intel i5 2nd Gen contiene 1,16 mld di transistor

La porta CMOS la base dei circuiti logici attuali

Unendo un pMOS con un nMOS si ottiene il circuito o «porta» CMOS perchè costituita da due transistor complementari.

La porta CMOS si comporta come una funzione logica NOT assegnando al livello 0 un valore di tensione nullo e a livello 1 il valore di tensione VDD.

Consumo di corrente e potenza nei circuiti CMOS

1. Consumo in condizioni statiche

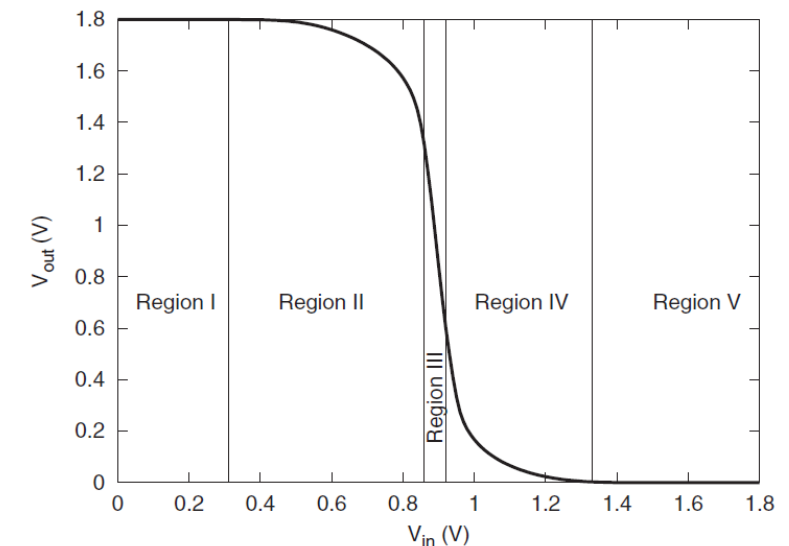
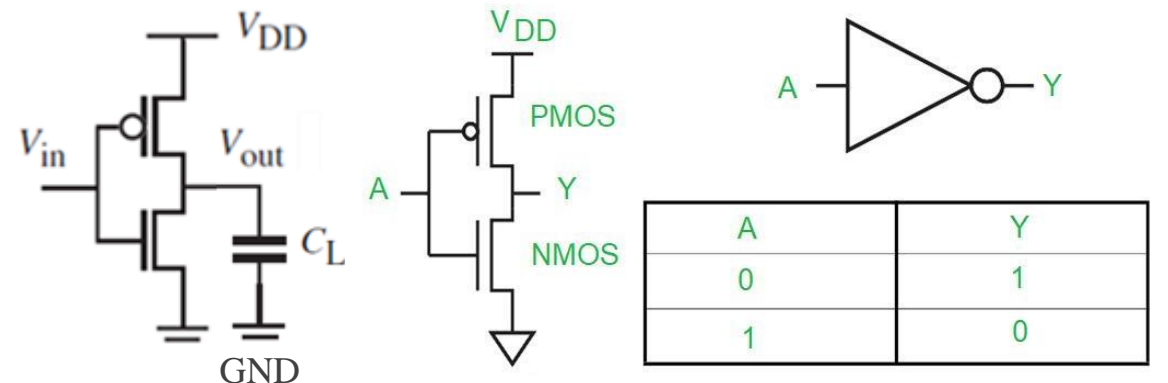
- In una porta CMOS solo uno tra nMOS e pMOS è attivo
- In condizioni statiche non esiste un percorso diretto VDD -> GND
- In condizioni statiche la corrente assorbita è quasi nulla e quindi lo stesso vale per la dissipazione di potenza

2. Consumo in condizioni dinamiche

- Durante le commutazioni logiche c'è una dissipazione di potenza dei transistor determinata dalla capacità di uscita C_L, dalla tensione di alimentazione e dalla frequenza di lavoro:

$$P_d = \frac{C_L V_{DD}^2}{T} = C_L V_{DD}^2 f$$

T è il periodo nel quale la tensione in uscita cambia da 0 a VDD e viceversa e $f = 1/T$ la frequenza





Conclusioni

- Nessuna tecnologia “rivoluzionaria” a breve termine
 - Oggi non esistono **tecnologie alternative mature** in grado di ridurre drasticamente il consumo rispetto al **CMOS**. Esistono evoluzioni del CMOS (es. GAAFET) che hanno efficienza migliore del 15~20% ma hanno una maggiore complessità realizzativa
- Cosa si può ancora ottimizzare
 - Riduzione della **tensione di alimentazione** (V_{DD}) interna dei chip
 - Materiali e geometria dei transistor
 - Tecniche architetturali
 - *Dynamic Voltage and Frequency scaling*
 - *Spegnimento selettivo dei blocchi*
 - *Architetture logiche più efficienti (i.e. TPU)*
- La potenza da dissipare continua ad aumentare
 - Aumento delle funzionalità on-chip
 - Aumento dei transistor e della densità di integrazione (attualmente 2~3 nm)
 - Frequenze operative sempre più elevate

“Oggi non esistono tecnologie in grado di cambiare radicalmente il quadro dei consumi: il CMOS può solo essere ottimizzato. L’aumento del numero di transistor e delle frequenze rende inevitabile la crescita della potenza totale dei chip.”

«La sfida diventa disporre della potenza necessaria nel tempo»



AGENDA

1. Aspetti Generali e architetture dei Data Center;
2. CPU e GPU differenze ed interazioni per i compiti AI;
3. Digressione elettronica;
4. Infrastrutture di rete LAN standard e con Cluster GPU
5. La rete SAN e lo Storage;
6. Conclusioni

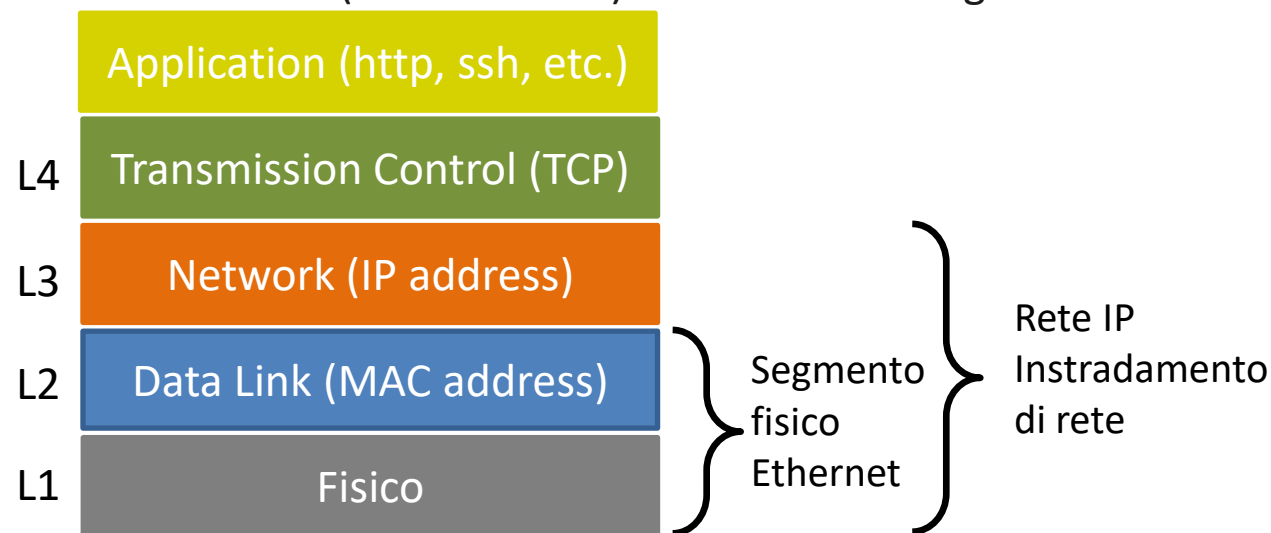
Come funziona la rete LAN

L'architettura logica della rete LAN è basata sul modello a livelli ISO/OSI in cui ogni livello svolge una funzione specifica di comunicazione. Livelli adiacenti scambiano dati tramite le loro interfacce.

La trasmissione dei dati avviene quindi in una serie di passaggi di dati da livelli superiori a livelli inferiori in trasmissione e da livelli inferiori a livelli superiori in ricezione.

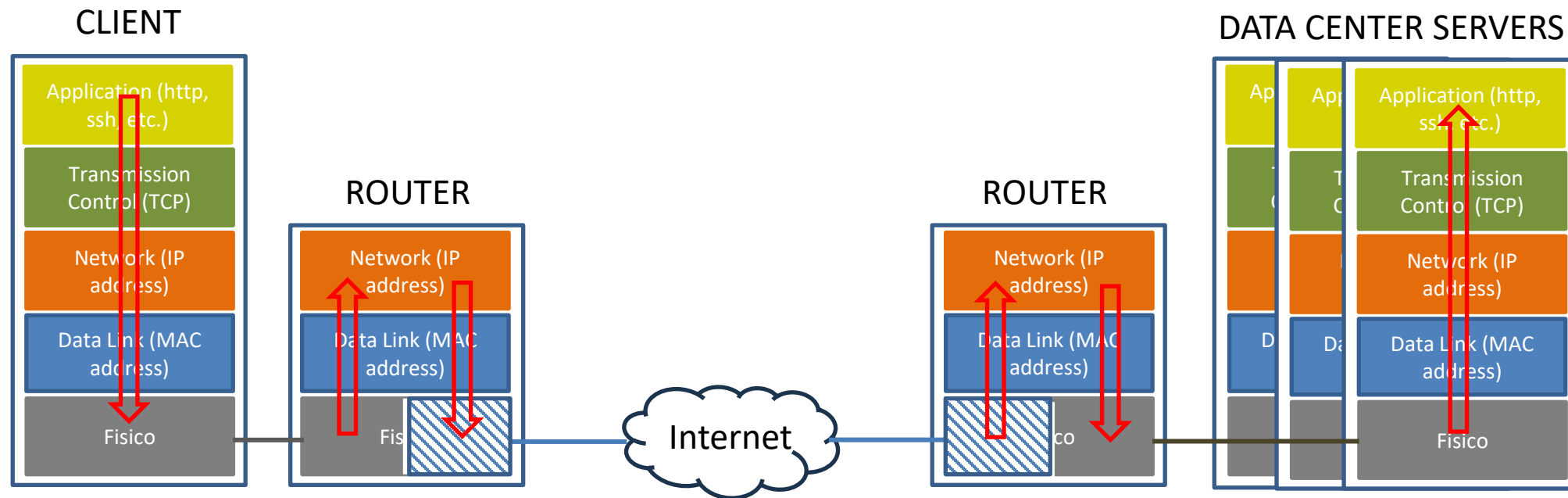
Dato un pacchetto di dati da trasmettere ogni livello N in trasmissione aggiunge e in ricezione toglie dei propri dati di controllo al pacchetto dei dati.

- L1 si occupa della trasmissione/ricezione fisica traducendo i dati da e verso L2 in segnali elettrici e stabilisce un collegamento diretto (via cavo o via radio) tra i sistemi;
- L2 incapsula i dati e gli fornisce un indirizzo di destinazione univoco (MAC address) valido solo sul segmento fisico che collega i sistemi;
- L3 – IP incapsula i dati e gli assegna un indirizzo univoco (IP address) valido a livello di rete globale e si preoccupa dell'instradamento dei pacchetti tra sottoreti diverse
- L4 - TCP in ricezione controlla i pacchetti ricevuti, riordina la sequenza se i pacchetti non sono in ordine e se ci sono errori ne richiede la ritrasmissione



Per effetto di questi meccanismi i pacchetti di dati possono arrivare a livello Applicativo con latenze diverse dell'ordine dei microsecondi (μs) e questo non è accettabile per il trasferimento dei dati tra GPU quindi occorrono protocolli diversi più efficienti

Esempio di trasmissione che dimostra il lavoro generato dalla interazione tra i vari livelli che crea il ritardo di elaborazione



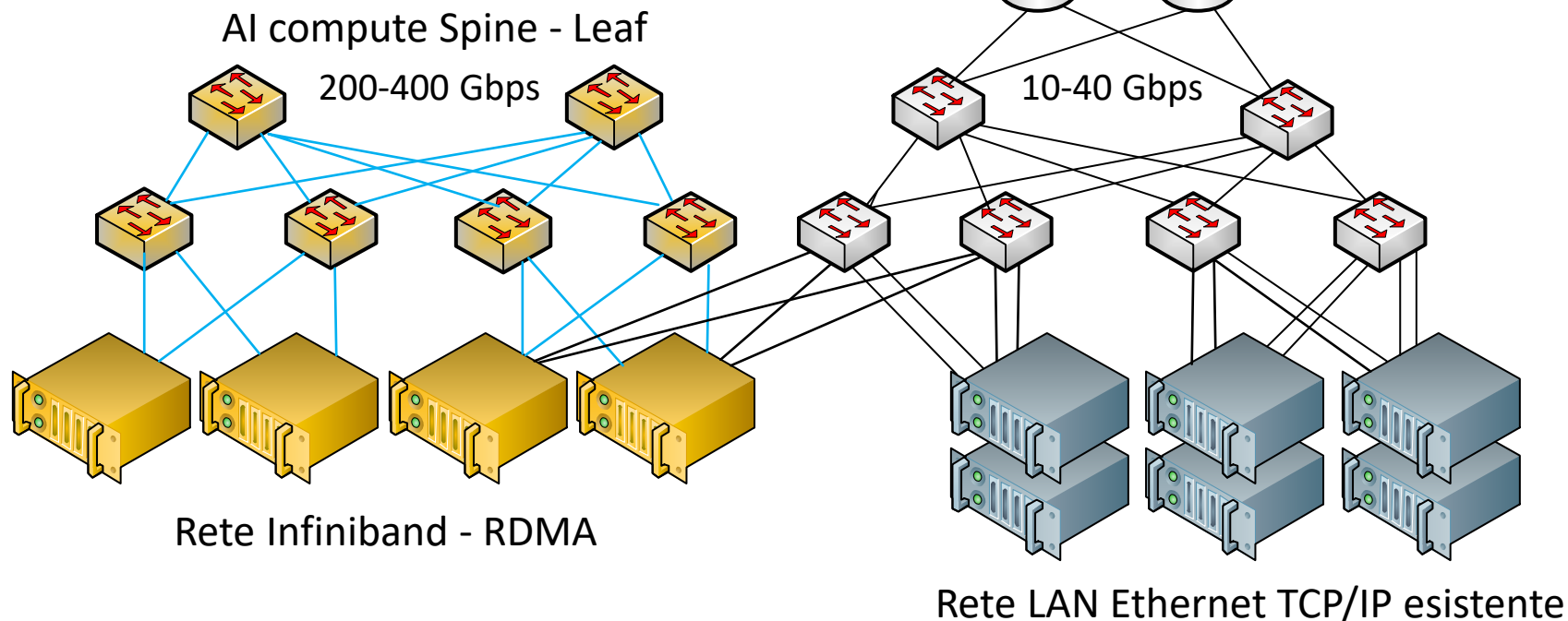
Ing. Mario D'Ettorre

Corso IA, HW e sostenibilità energetica

Per i grandi sistemi dove i modelli LLM sono grandissimi con 100+ miliardi di parametri si ricorre a Cluster di GPU costituiti di N nodi ognuno dei quali è un server per AI

Un GPU cluster **richiede la presenza di una infrastruttura di rete per lo scambio dei dati tra i nodi in modo sincronizzato e con latenza ultrabassa**, come avviene tra le GPU all'interno dei nodi. Questa rete è diversa dal TCP/IP e utilizza hardware specifico; quindi il Data Center avrà due tipologie di rete, con apparati di switching e schede di rete diverse

Vengono utilizzate tecnologie di accesso diretto alla memoria indicate Remote Direct Memory Access **RDMA** e attualmente si chiamano **Infiniband** e **RoCE** la differenza è che Infiniband è un protocollo proprietario mentre RoCE (RDMA over Converged Ethernet) è un protocollo basato su Ethernet.



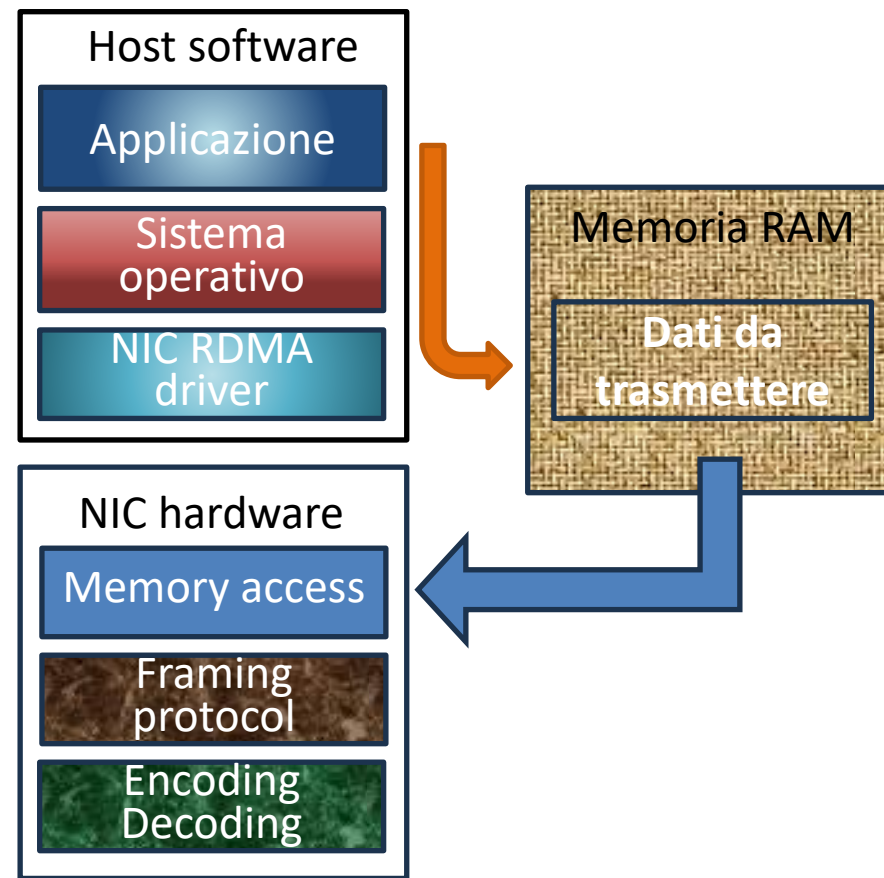
L'accesso diretto alla memoria del protocollo RDMA è molto più veloce perché sfrutta hardware dedicato, liberando la CPU di questo compito e quindi riducendo drasticamente i tempi di elaborazione del protocollo .

Questa tecnologia consente di leggere/scrivere dati direttamente nella RAM di un'altra macchina

Il Sistema Operativo passa al Driver della scheda di rete (Network Interface Controller - NIC) il puntatore ai dati da trasmettere. L'hardware della NIC preleva i dati dalla memoria del server li formatta secondo il protocollo richiesto e li invia sulla rete

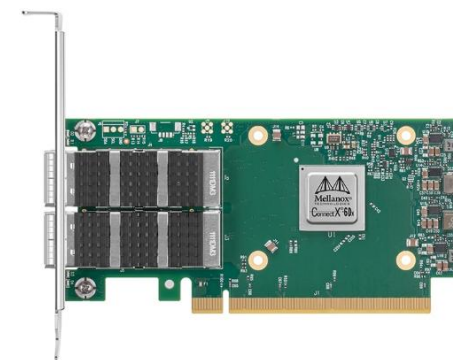
La normale NIC della rete LAN del protocollo TCP/IP su Ethernet è più semplice, non possiede queste specializzazioni hardware e richiede l'interazione con la CPU

Sui server troveremo quindi schede di rete convenzionali e schede di rete compatibili con il nuovo protocollo



NIC – Consumo elettrico medio Schede di rete

Tipo di NIC	Velocità	Consumo tipico (Watt)
Ethernet standard	10 Gb	4-8 W
Ethernet standard	25 Gb	8-12 W
RoCEv2 (Mellanox CX5)	25Gb	14-17 W
RoCEv2 (Mellanox CX6)	100 Gb	18-25 W
InfiniBand (HDR)	100 Gb	25-35 W



Consumo switch InfiniBand per velocità

Velocità	Generazione	Modello Indicativo	Porte	Consumo tipico
100 Gb	HDR	Mellanox SB7800	36	~136 W
400 Gb	NDR	NVIDIA QM9790	32	~640 W



Da considerare ovviamente moltiplicato per il numero di schede di rete e di switch presenti



AGENDA

1. Aspetti Generali e architetture dei Data Center;
2. CPU e GPU differenze ed interazioni per i compiti AI;
3. Digressione elettronica;
4. Infrastrutture di rete LAN standard e con Cluster GPU
5. La rete SAN e lo Storage;
6. Conclusioni

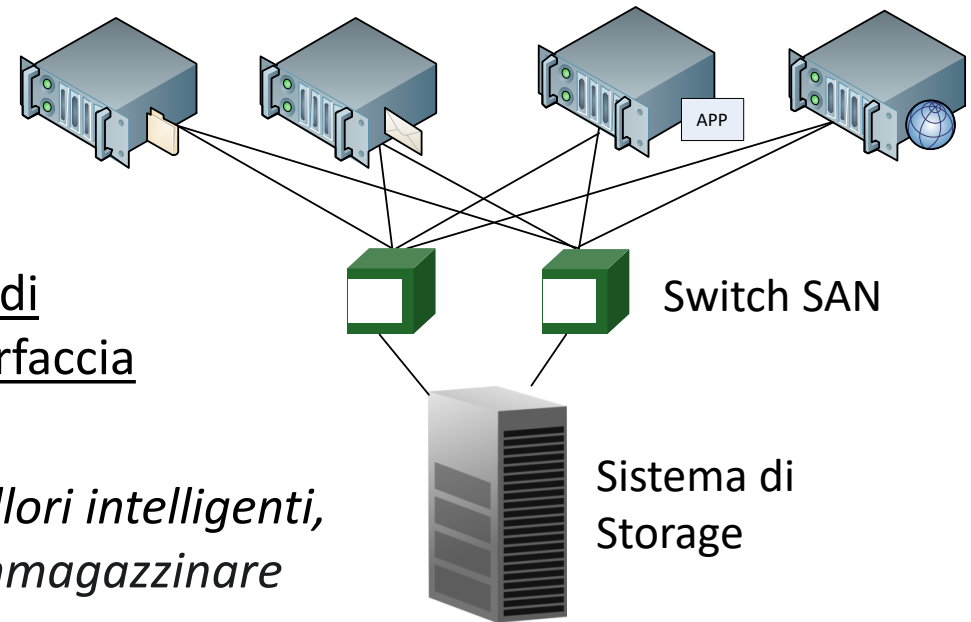
Storage e Storage Area Network (SAN)

Nei Data Center tradizionali si rende necessaria la presenza di una infrastruttura di Storage ad alte prestazioni e alta disponibilità, progettata per fornire ai server grandi capacità di memorizzazione condivisa, scalabile e resiliente.

L'infrastruttura è composta da sistemi di storage e da una rete di trasmissione dati dedicata con propri apparati e schede di interfaccia denominate Host Bus Adapter (**HBA**)

Un sistema di storage è un insieme di dischi (HDD/SSD), controllori intelligenti, software di gestione e funzionalità avanzate, progettato per immagazzinare PetaByte di dati in alta disponibilità e affidabilità.

Ogni server accede allo storage della SAN condiviso come se fosse un disco fisico collegato localmente, di fatto i sistemi di storage vengono visti dal server come dischi locali (es. C:, D: di Windows) mentre invece sono resi disponibili tramite la SAN.





Storage e SAN per i Carichi di Lavoro AI

I workload della AI differiscono significativamente dalle applicazioni tradizionali nelle varie fasi:

Preparazione dei Dati/ETL (Extract, Transform, Load)

- Attività intensive sia di lettura che di scrittura, spesso su dataset molto vasti per l'estrazione dei dati grezzi, trasformazione in formati coerenti e puliti e caricamento nello storage.

Addestramento

- **Lectture sequenziali massive di grandi dataset** (petabyte di immagini, video, testo, dati tabellari) per alimentare gli algoritmi di apprendimento. **Necessario un throughput elevato per evitare che le GPU rimangano idle in attesa dati.**
- **Scritture periodiche significative** per il checkpointing dei modelli, ovvero il salvataggio dello stato del modello a intervalli regolari per la resilienza e la ripresa. Questi checkpoint possono essere di grandi dimensioni e richiedono prestazioni di scrittura elevate per minimizzare l'impatto sulla durata complessiva del training.

Inferenza

- Letture casuali di porzioni di dati più piccole, i parametri del modello da utilizzare e i dati in input per la previsione. Richiede bassa latenza (uS), per consentire decisioni o risposte in tempo reale

Per soddisfare queste richieste anche la SAN deve essere ad alte prestazioni e bassa latenza e le unità di storage devono essere a stato solido (SSD) tipo All-Flash-Array (AFA)

Ing. Mario D'Ettorre

Corso IA, HW e sostenibilità energetica



La SAN dei data center convenzionali è costruita per garantire **la consegna di blocchi di dati in ordine e senza perdite, minimizzando la latenza del protocollo.**

Si utilizza un protocollo diverso dal TCP/IP **denominato Fiber Channel (FC) perché il cablaggio viene realizzato esclusivamente in fibra ottica** per avere una trasmissione immune ai disturbi che possono provocare errori nei dati.

Il protocollo FC già è stato pensato fin dall'inizio a bassa latenza con le schede HBA che effettuano l'accesso diretto alla memoria

Il protocollo FC adesso è in concorrenza con il moderno RoCE, però può vantare il fatto della base di conoscenze ed installazioni ormai consolidata.

Si può prevedere una convergenza fisica delle reti del cluster GPU e della SAN utilizzando il protocollo RoCE e configurando una separazione logica delle due reti; in questo modo non è più necessario usare apparati di switch e interfacce di rete distinte.



Per le unità SSD è stato creato il protocollo Non Volatile Memory express – NVMe - per collegare lo storage flash direttamente al bus PCIe interno al sistema, come disco locale e sostituisce i vecchi protocolli SATA o SAS progettati per i dischi magnetici.

Unità SSD con protocollo NVMe sono infatti ormai diffusi all'interno dei PC e quindi a maggior ragione all'interno dei nodi del Cluster AI per lo storage locale.

Il protocollo NVMe è un protocollo di interfaccia verso lo storage SSD, non è un protocollo di rete, però i suoi comandi e dati possono essere incapsulati in un protocollo di rete di trasporto per realizzare la SAN.



I comandi NVMe possono essere trasportati su protocollo FC noto come NVMe-oFC, si possono usare anche protocolli come RoCE o Infiniband. Per esempio può essere usato anche il TCP/IP ma con prestazioni inferiori però può funzionare con le NIC standard.

N.B.: Il protocollo NVMe-oFC può sfruttare le infrastrutture FC già esistenti nei Data Center che spesso hanno già la compatibilità con il trasporto di comandi NVMe.

7. Infrastrutture di rete LAN e SAN impatti della AI

Tabella comparativa: NVMe/TCP vs NVMe/RoCEv2 vs NVMe/InfiniBand

Caratteristica	NVMe/TCP	NVMe/RoCEv2	NVMe/InfiniBand
Tipo di trasporto	TCP/IP su Ethernet (L3)	RDMA over Converged Ethernet (L2/L3)	RDMA over InfiniBand (L2 proprietario)
Latenza	● Media-bassa (~100–200 µs)	● Bassa (~10–20 µs)	● Molto bassa (~1–2 µs)
Throughput	● Buono (dipende dalla CPU)	● Ottimo	● Eccellente (line-rate nativo)
CPU usage	● Alto (stack TCP/IP)	● Basso (kernel bypass)	● Minimo (RDMA puro, kernel bypass totale)
Configurazione	● Semplice	● Complessa (rete lossless, QoS, PFC)	● Complessa (ambiente dedicato InfiniBand)
Compatibilità HW	● NIC standard	● NIC RDMA-capable (es. Mellanox)	● Richiede HCA InfiniBand
Interoperabilità	● Massima (TCP/IP standard)	● Solo tra nodi RDMA su Ethernet	● Solo InfiniBand
Costo infrastruttura	● Basso	● Medio-alto	● Alto
Scalabilità	● Elevata	● Elevata (con RDMA)	● Elevata, ma richiede fabric dedicato
Utilizzo tipico	Cloud, storage general-purpose	AI/ML, database, low-latency workloads	HPC, supercomputing, AI exascale

Nota: RoCEv2 ha anche il livello IP (L3) per consentire il routing dei pacchetti

Ing. Mario D'Ettorre

Corso IA, HW e sostenibilità energetica



AGENDA

1. CPU e GPU differenze ed interazioni per i compiti AI;
2. Digressione elettronica;
3. Principi di networking;
4. Infrastrutture di rete LAN standard e con Cluster GPU;
5. La rete SAN e lo Storage;
6. Conclusioni.

Data Center per implementazione massiva della IA

- Server con molteplici GPU che necessitano di potenza elettrica adeguata – allo stato attuale anche 80 kW per rack, a tendere anche 130 -150 kW per rack
- Impianto elettrico adeguato per veicolare ai rack server le potenze richieste
- Probabile impianto di Raffreddamento più complesso, a liquido Direct to CHIP e raffreddamento classico ad aria per il resto dei componenti dei server
- Infrastrutture di rete standard TCP/IP per il traffico client server e ad alte prestazioni di tipo RDMA per il cluster di GPU
- Infrastruttura di Storage adeguata a livello di prestazioni utilizzando apparati e interfacce compatibili con uno dei protocolli NVMe-oFC, NVMe-RoCEv2 o NVMe-Infiniband scegliendo quello più adeguato alle proprie esigenze
- Apparati di storage ad alte prestazioni con storage a stato solido per supportare le richieste dei Cluster di GPU



Grazie

Ing. Mario D'Ettorre

dettorremario@gmail.com