



AGENDA

1. Aspetti generali e architetture dei Data Center;
2. Panoramica delle Infrastrutture tecnologiche;
3. Caratteristiche delle sale server e posizionamento degli apparati;
4. Infrastrutture di rete LAN e SAN;
5. Concetti base delle GPU dal punto di vista funzionale ed energetico;
6. Cambiamenti delle infrastrutture tecnologiche e delle sale server, per ospitare i sistemi per l'IA;
7. L'architettura della rete LAN in presenza di sistemi per l'IA;



■ 1,5

65

Valore data center
654 milioni di euro

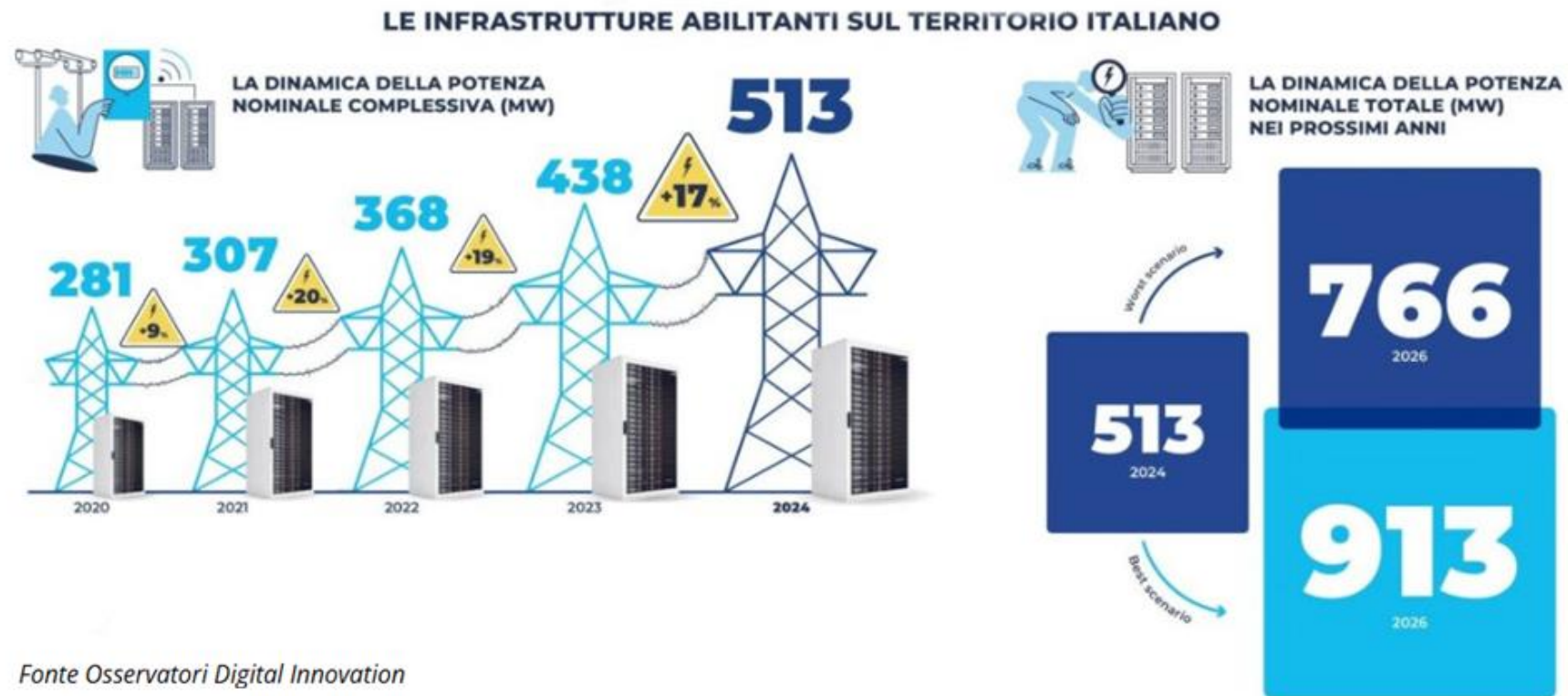
184 MW
Milano
che è la prima
città italiana

**Francoforte,
prima in Europa**

Withub

Efficienza energetica nei Data Center

1. Aspetti generali e architetturali dei Data Center



LA CRESCENTE RICHIESTA DI SERVIZI DIGITALI - EMAIL, SOCIAL MEDIA, E-COMMERCE, STREAMING, VIDEOCONFERENZE, 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, CHE DEVE ESSERE COSTANTE NEL TEMPO E DI QUALITÀ.

1. Aspetti generali e architetturali dei Data Center

BUILDING:

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

POWER CONDITIONING E DISTRIBUTION:

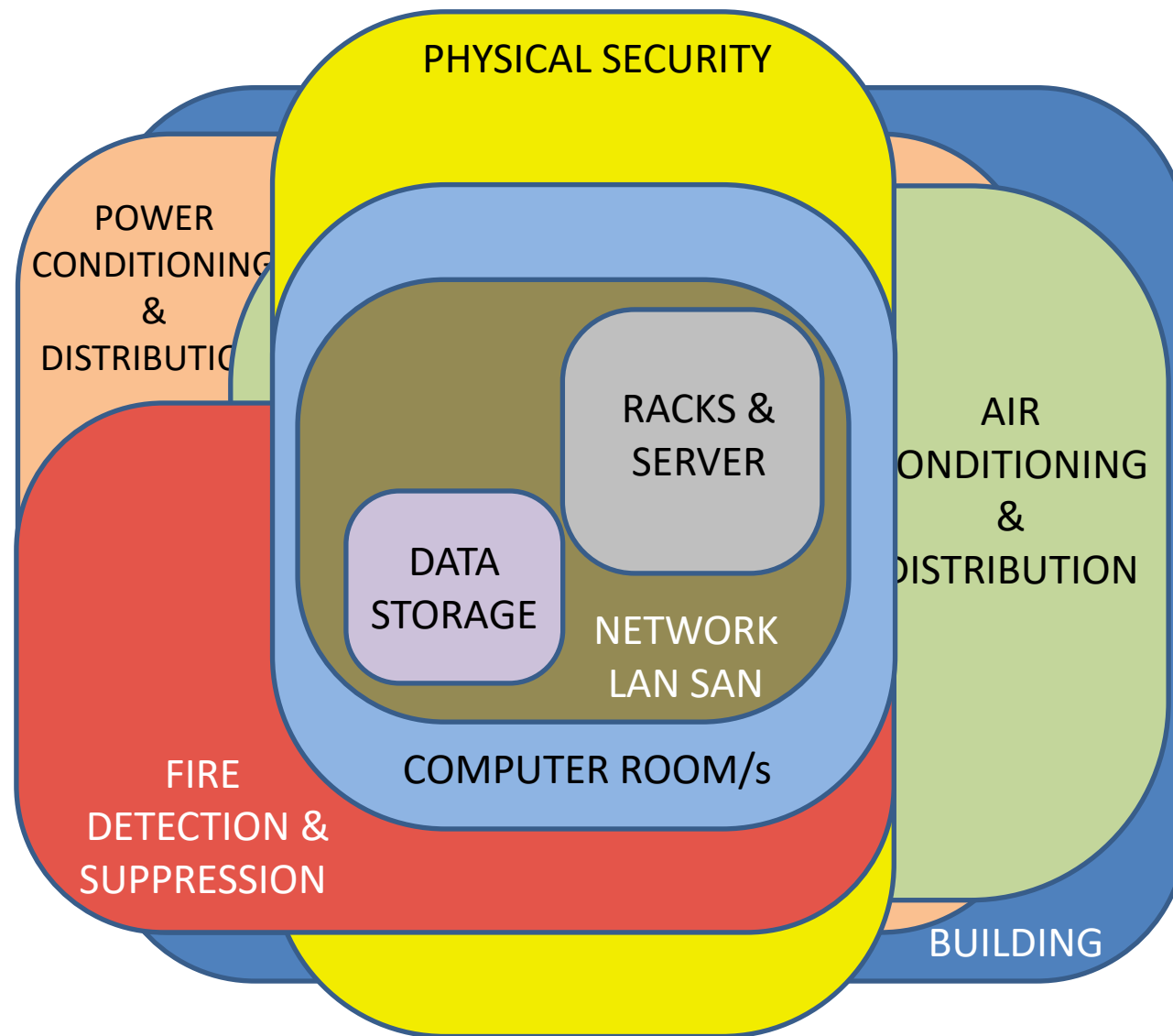
- Forniture in Media Tensione
- UPS
- Gruppi Elettrogeni
- Cabine di trasformazione
- Distribuzioni alle utenze elettriche in bassa tensione

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 architetturali dei Data Center

Ridondanza $2N : 2(N + 1)$ – Fault Tolerant

Il Data Center ha **impianti interamente duplicati, ognuno di capacità sufficiente (anche con componenti ridondati), che sono contemporaneamente attivi**, quindi in caso di guasto o manutenzione programmata **l'impianto alternativo consente il funzionamento del Data Center senza interruzioni di servizio** (potranno accadere interruzioni di servizio giusto per allarmi antincendio per esempio).

Tier 4 richiede per esempio apparati ICT con doppia alimentazione.

Con questo modello si prevedono disservizi dell'ordine di minuti per anno

Ridondanza $2N$ – Contemporaneamente manutenibile

Il Data Center ha **impianti duplicati ognuno di capacità sufficiente, ma un solo impianto attivo alla volta**, quindi in caso di guasto o manutenzione programmata **può essere attivato l'impianto alternativo**, riducendo ulteriormente il fuori servizio del Data Center.

Con questo modello si prevedono disservizi inferiori poche ore per anno

Ridondanza $N + 1$ – Componenti ridondati

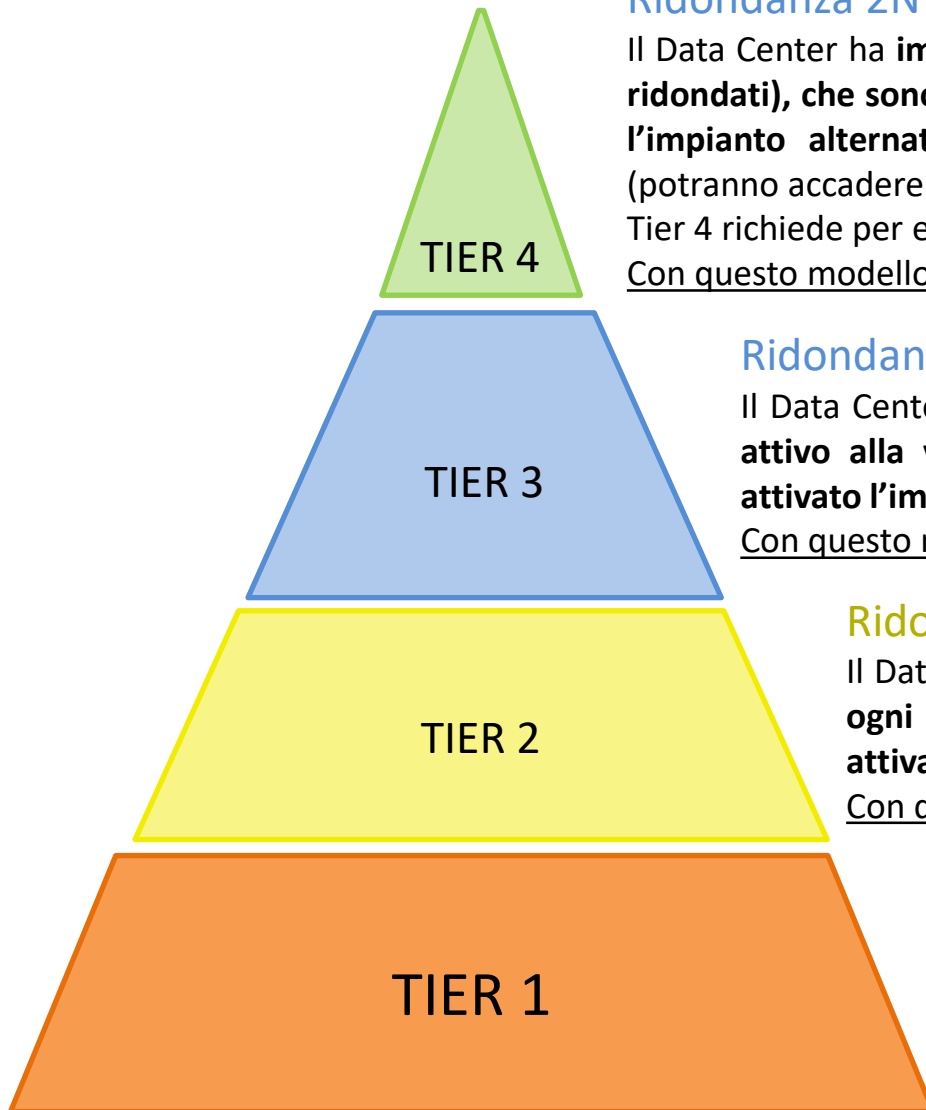
Il Data Center ha gli impianti con componenti ridondati (**almeno uno in più del necessario per ogni componente**), quindi in caso di guasto o manutenzione programmata **possono essere attivati i componenti ridondati riducendo il fuori servizio del Data Center**.

Con questo modello si prevedono disservizi inferiori ad un giorno per anno

Ridondanza N – Nessun componente ridondato

Ogni componente o impianto del Data Center è singolo, quindi **ogni guasto crea un fuori servizio del Data Center** più o meno grave, così come le attività di manutenzione programmata, per il tempo necessario all'intervento.

Con questo modello si prevedono fuori servizio dell'ordine dei giorni per anno



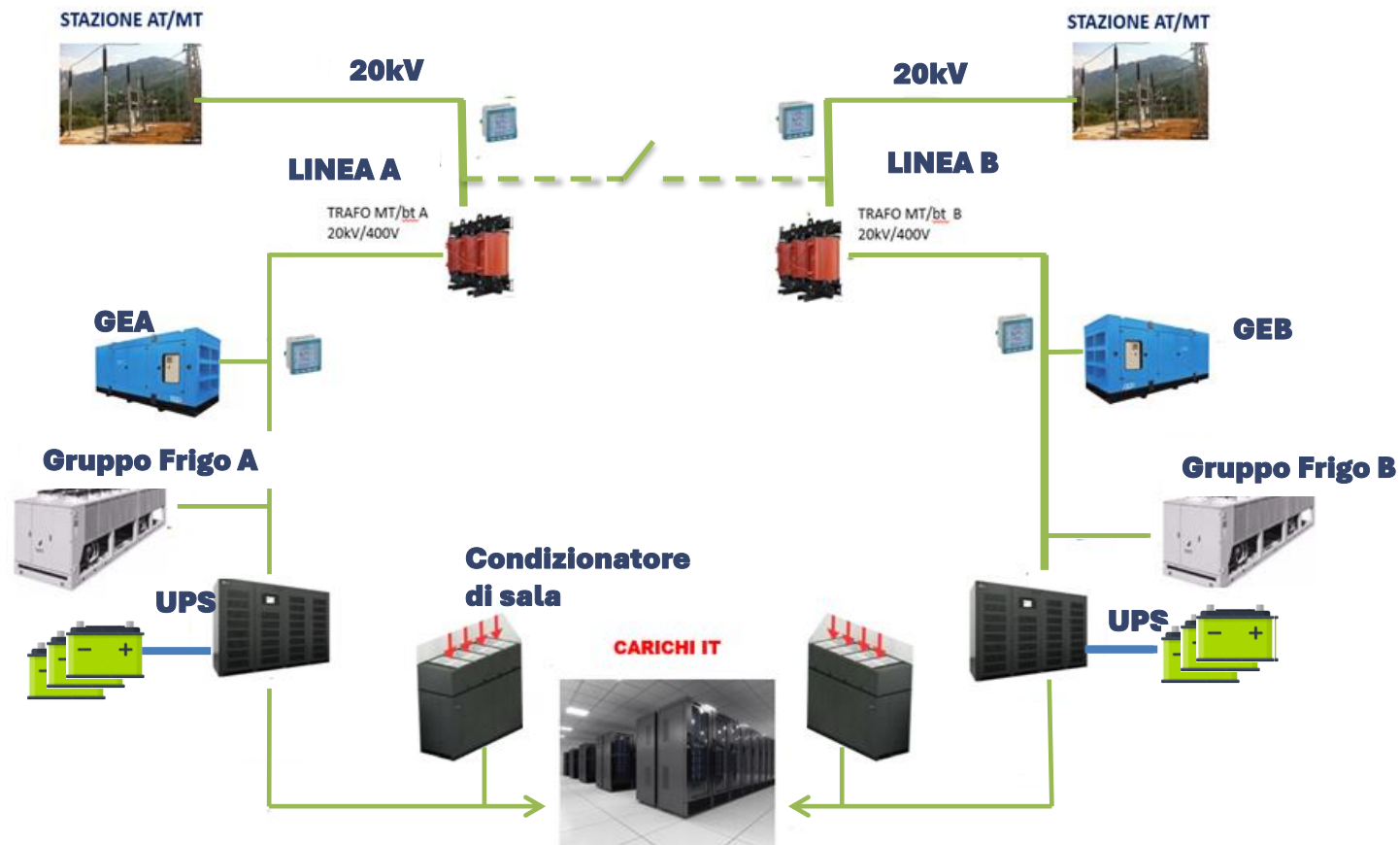
Ing. Mario D'Ettorre



AGENDA

1. Aspetti generali e architetture dei Data Center;
- 2. Panoramica delle Infrastrutture tecnologiche;**
3. Caratteristiche delle sale server e posizionamento degli apparati;
4. Infrastrutture di rete LAN e SAN;
5. Concetti base delle GPU dal punto di vista funzionale ed energetico;
6. Cambiamenti delle infrastrutture tecnologiche e delle sale server, per ospitare i sistemi per l'IA;
7. L'architettura della rete LAN in presenza di sistemi per l'IA;
8. Conclusioni

2. Panoramica Impianto elettrico



I GRANDI DATA CENTER SONO TIPICAMENTE DI CLASSE TIER III / IV QUINDI CON IMPIANTI DUPLICATI OGNUNO DI CAPACITA' SUFFICIENTE PER IL FUNZIONAMENTO DEL 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 centro di calcolo 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 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

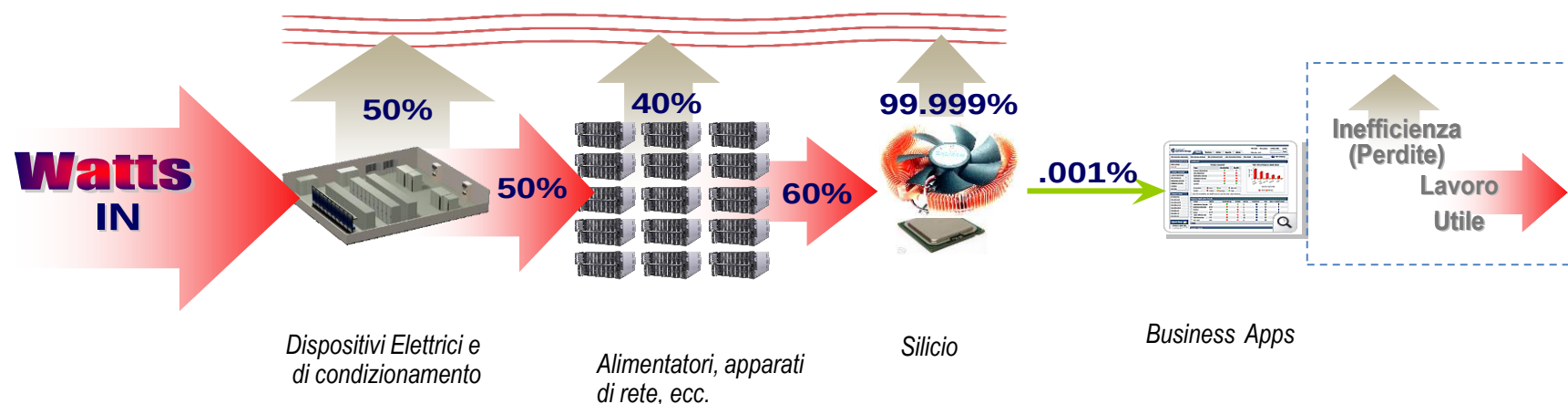
Ad oggi sono state introdotte altre metriche:

WUE: Water Usage Effectiveness

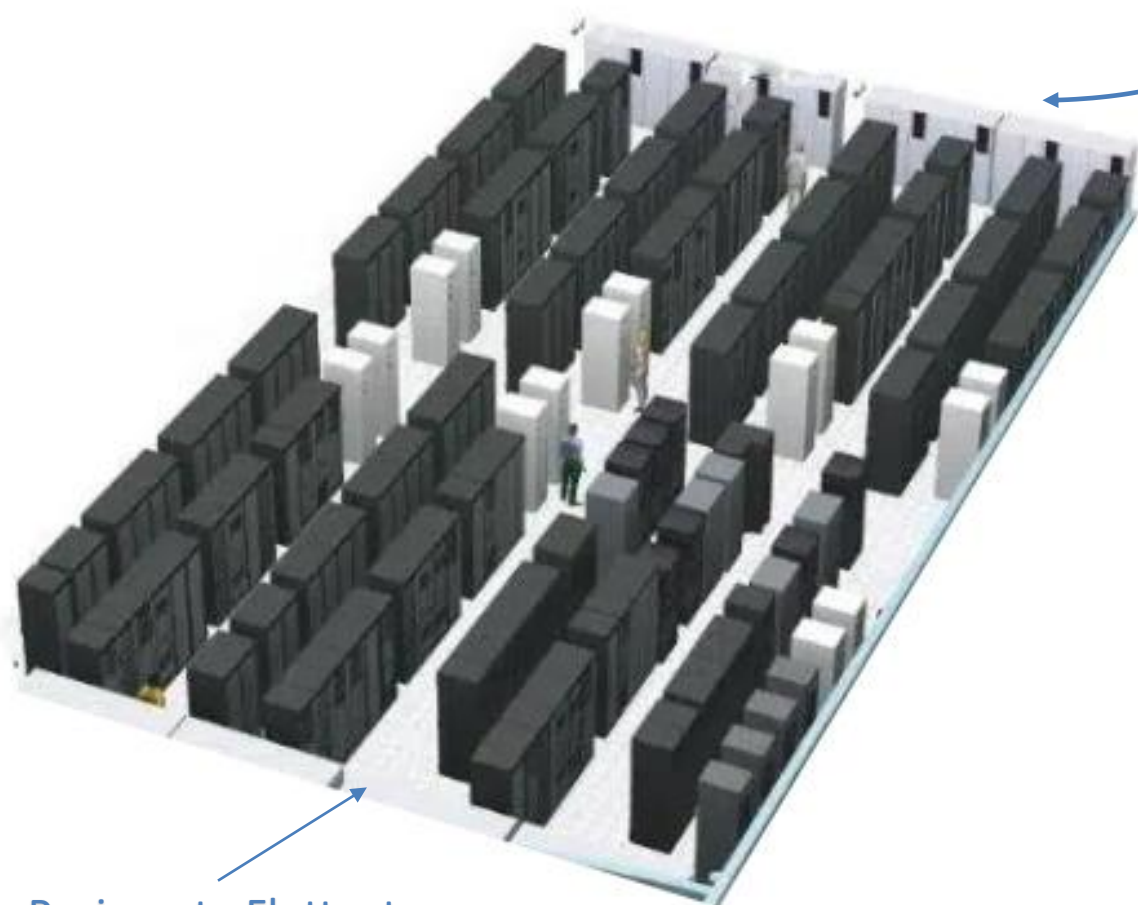
DCRE: Data Center Reusable Effectiveness che comprende tutte le risorse consumate dai Data Center

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.

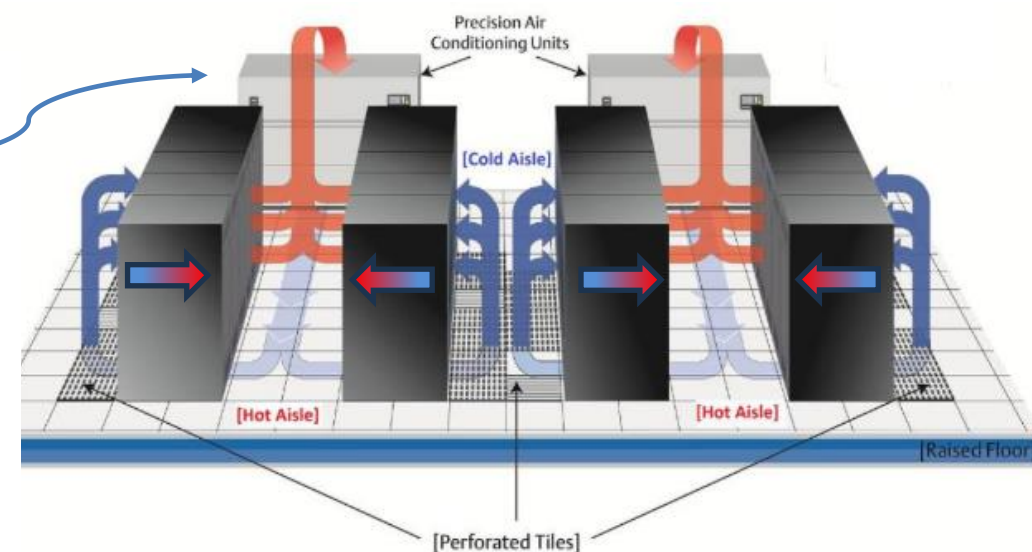
Della parte di energia che arriva ai sistemi ICT il 99,999% è disperso in calore e solo lo 0,001% è utilizzato per la elaborazione delle informazioni.



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

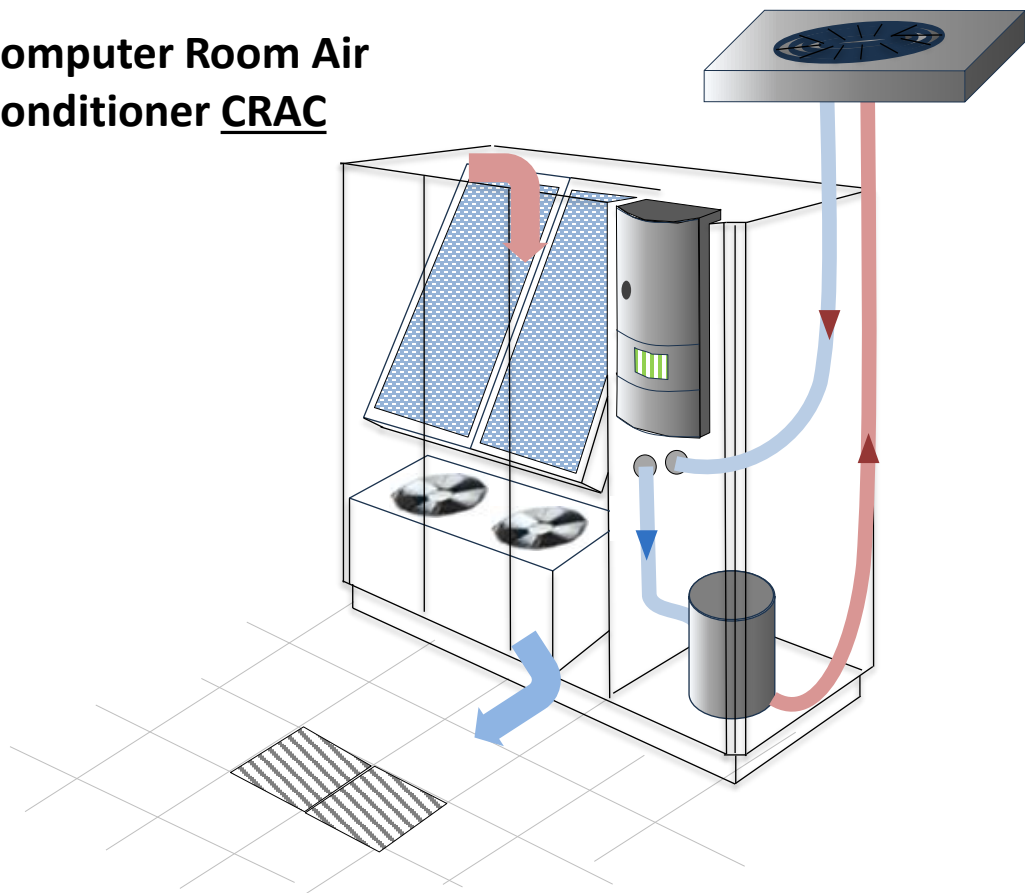


Pavimento Flottante



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.

Computer Room Air Conditioner CRAC



Condizionamento di precisione a Gas Refrigerante (ad espansione diretta DX) costituito da tante unità indipendenti, ognuna col proprio compressore ed unità esterna

Le unità sono disposte nel perimetro delle sale

L'aria calda viene aspirata dall'alto attraversa la griglia fredda e viene soffiata al di sotto del pavimento flottante (plenum)

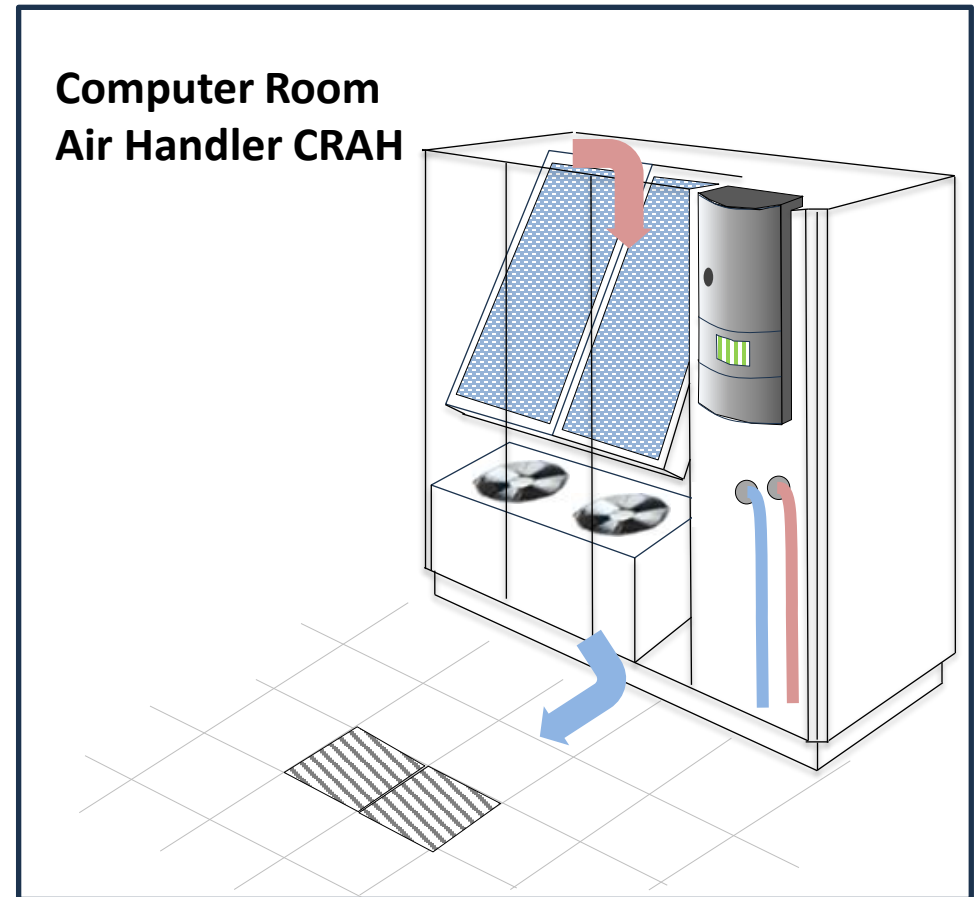
Più adatto per data center medio piccoli.

Impianto di condizionamento ad acqua refrigerata.

Si preferisce per grandi Data Center:

- sono più efficienti su larga scala
- L'acqua può essere prelevata da fonti di acqua fredda (acqua di falda)
- Minore impatto ambientale per la riduzione di gas refrigeranti
- Free Cooling che con un ulteriore impianto consente di raffreddare l'acqua se la temperatura esterna è bassa (es. inverno)

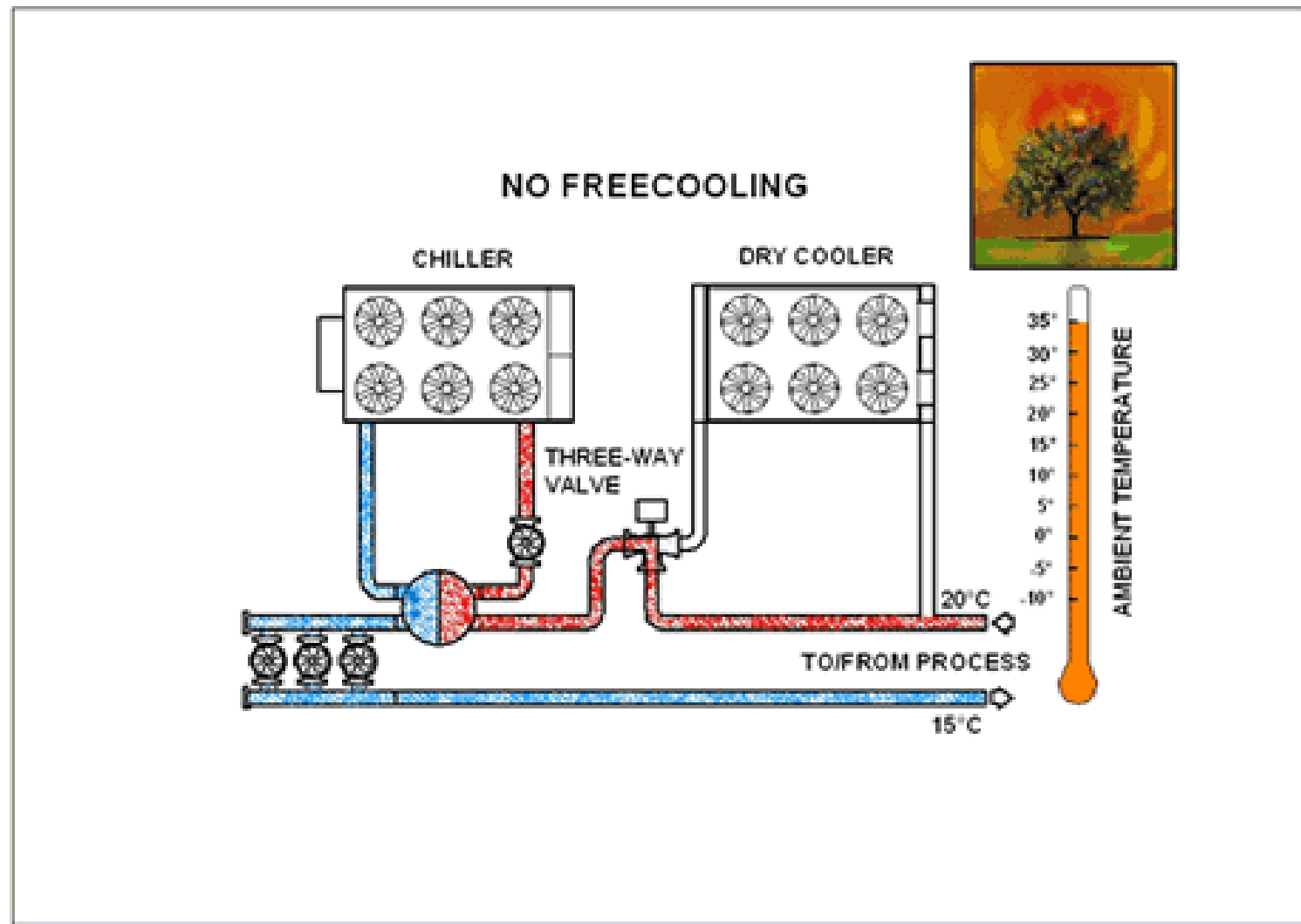
Chiaramente è richiesto un impianto complesso, con tubazioni di mandata e ritorno per i CRAH, pompe per la circolazione del liquido, un grande refrigeratore (chiller) e unità di evaporazione o torri evaporative all'esterno.



FUNZIONAMENTO DEL FREE COOLING



DRY COOLER



3. Caratteristiche delle sale server

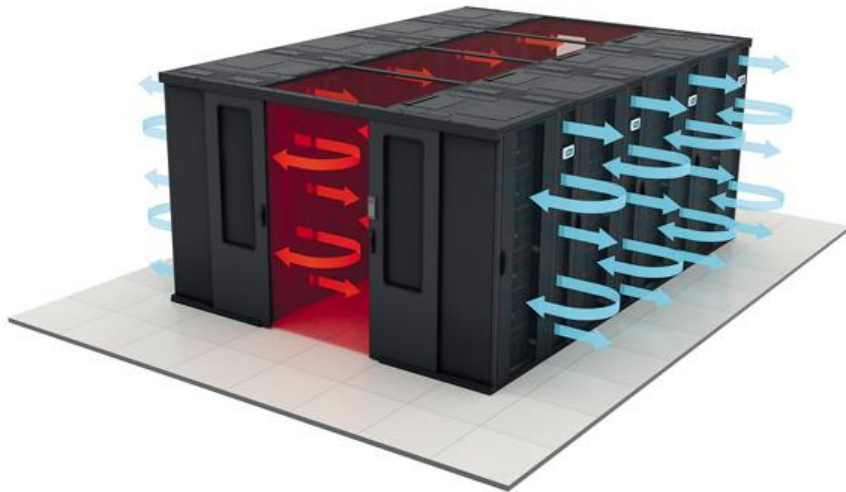
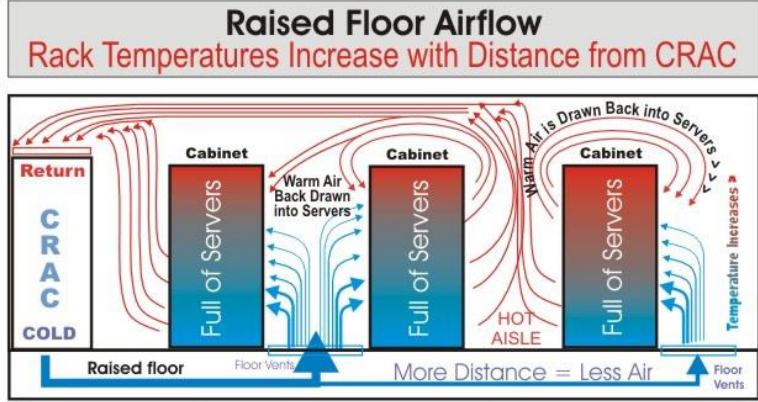
Il condizionamento perimetrale è poco efficiente:

- 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 CRAH sottili e alti quanto i rack e sono inseriti nella fila di rack a tutta altezza

Disegno condizionamento «in row»



3. Caratteristiche delle sale server



Ing. Mario D'Ettorre

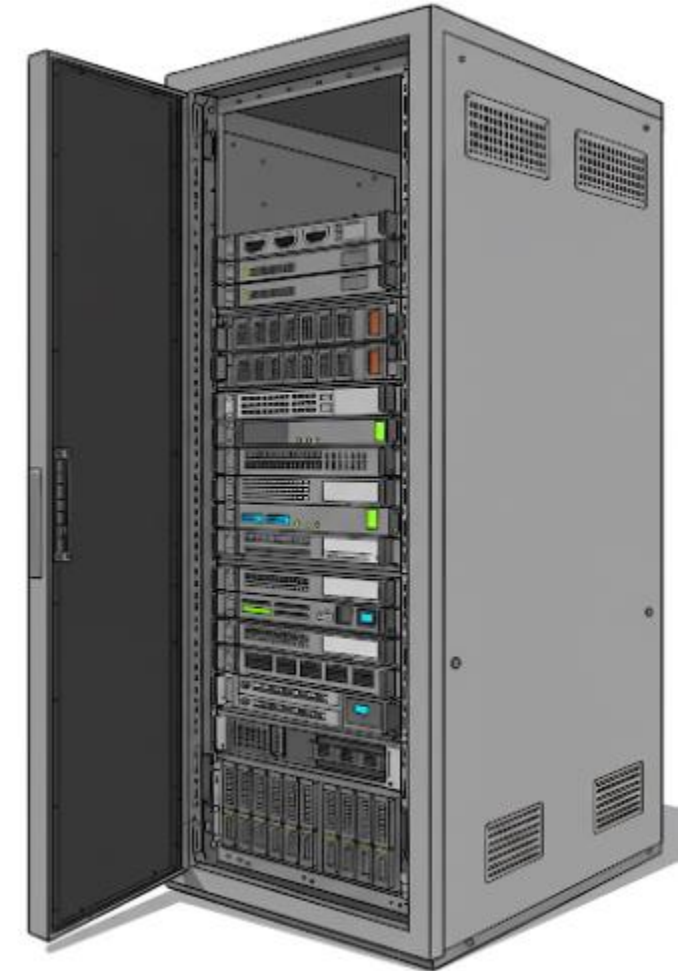
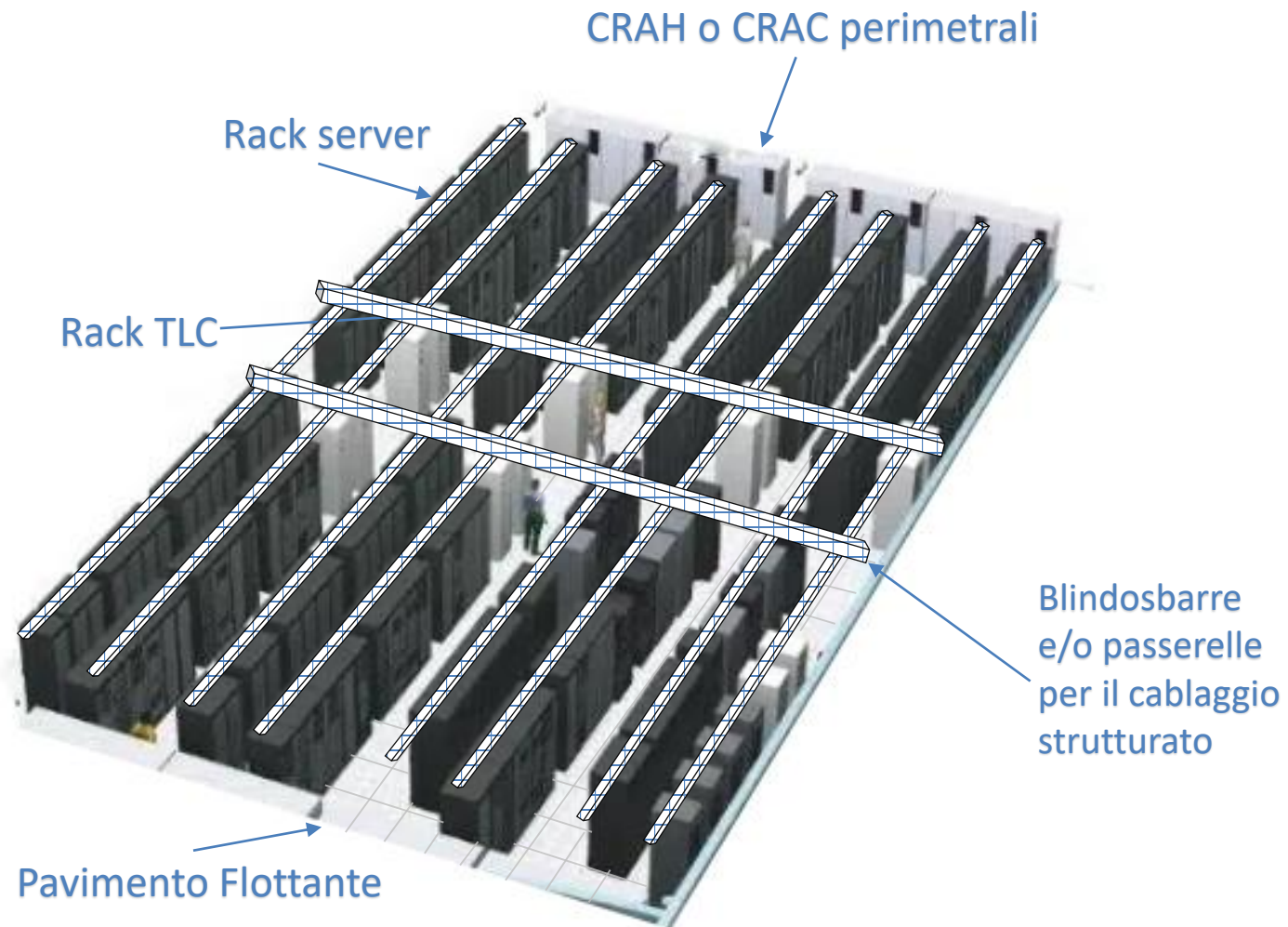
Efficienza energetica nei Data Center



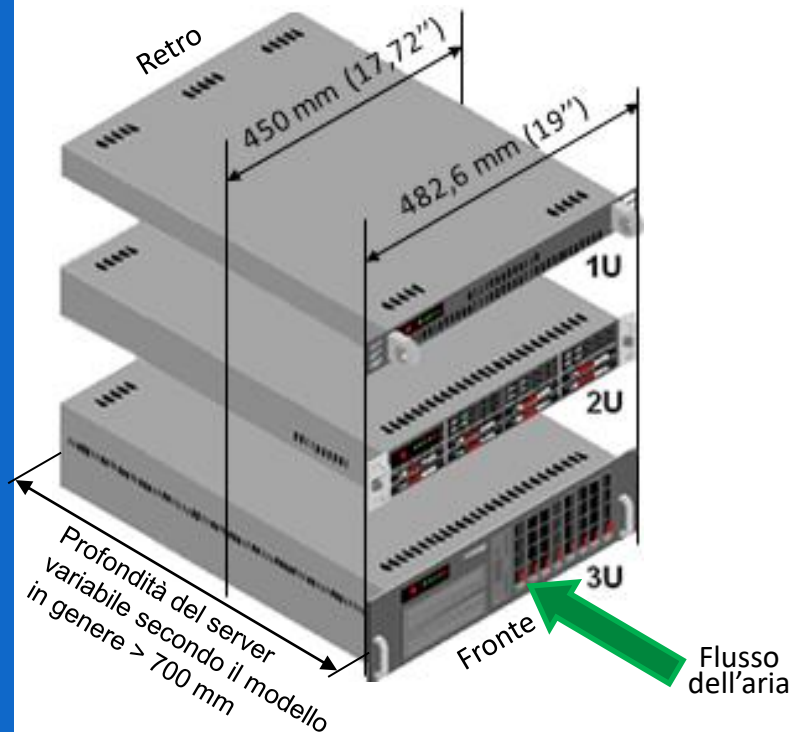
AGENDA

1. Aspetti generali e architetture dei Data Center;
2. Panoramica delle Infrastrutture tecnologiche;
- 3. Caratteristiche delle sale server e posizionamento degli apparati;**
4. Infrastrutture di rete LAN e SAN;
5. Concetti base delle GPU dal punto di vista funzionale ed energetico;
6. Cambiamenti delle infrastrutture tecnologiche e delle sale server, per ospitare i sistemi per l'IA;
7. L'architettura della rete LAN in presenza di sistemi per l'IA;
8. Conclusioni

3. Caratteristiche delle sale server



3. Caratteristiche delle sale server

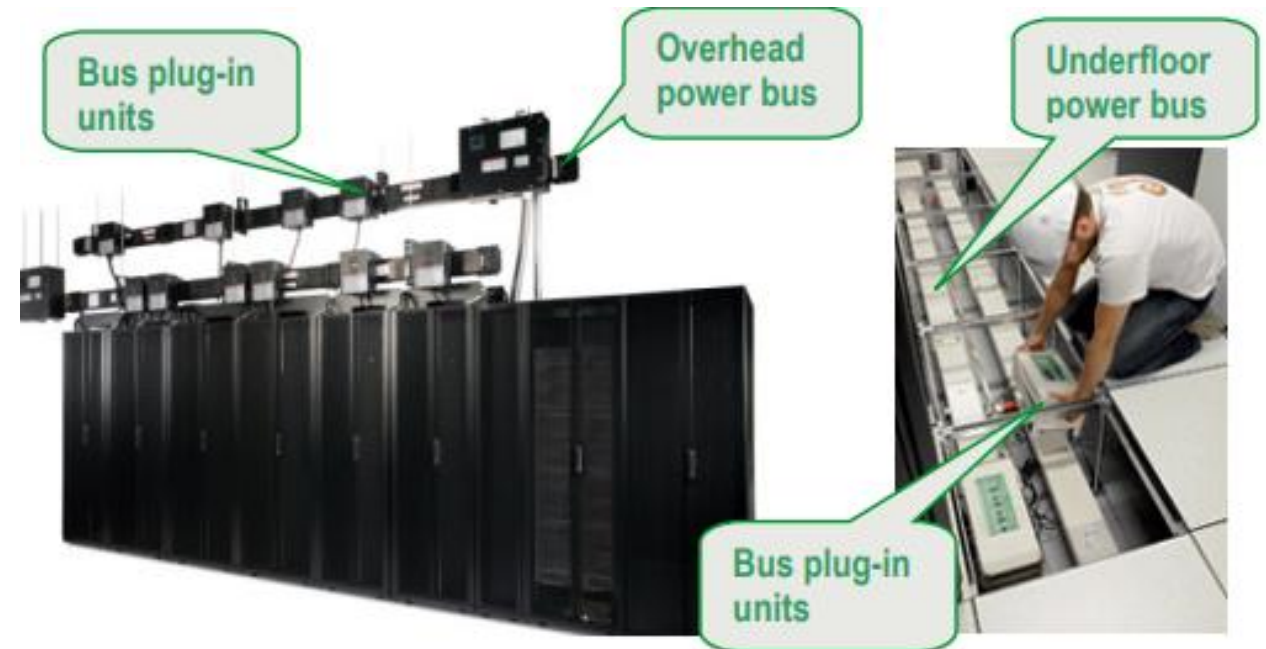
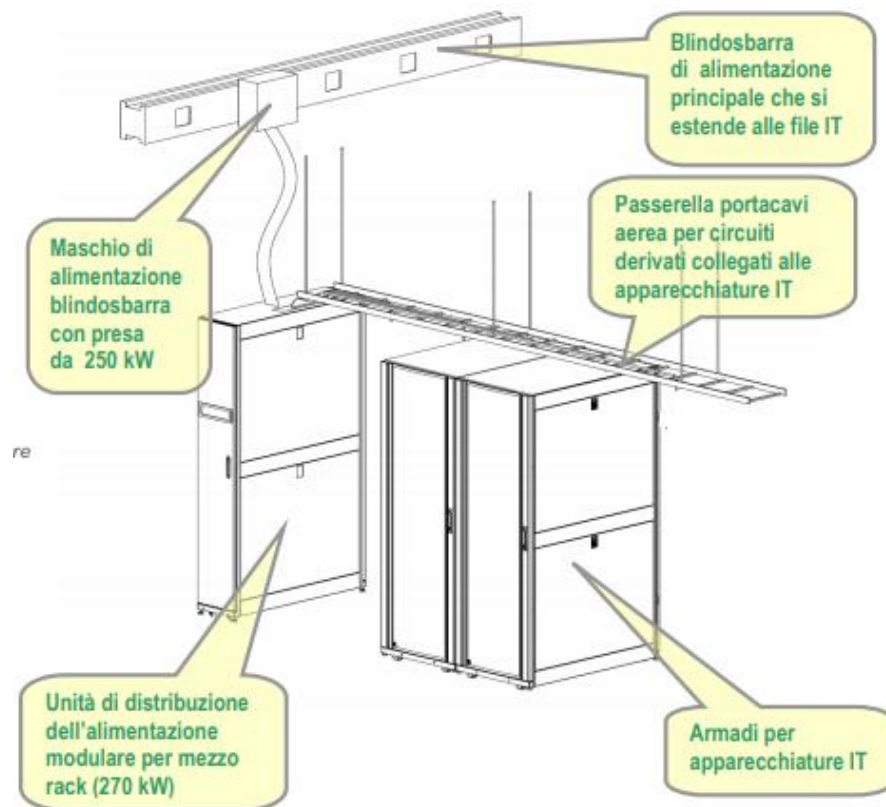


- La potenza usata con le densità attuali rack è 7 - 10 kW per rack (@ 32 A monofase – 16 A trifase)
- Il flusso dell'aria di raffreddamento è sempre dal fronte al retro
- I cavi di interconnessione sono **sempre** sul retro del rack
- I server sono **montati su slitte** che consentono l'estrazione per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

Ing. Mario D'Ettorre

Efficienza energetica nei Data Center

3. Caratteristiche delle sale server



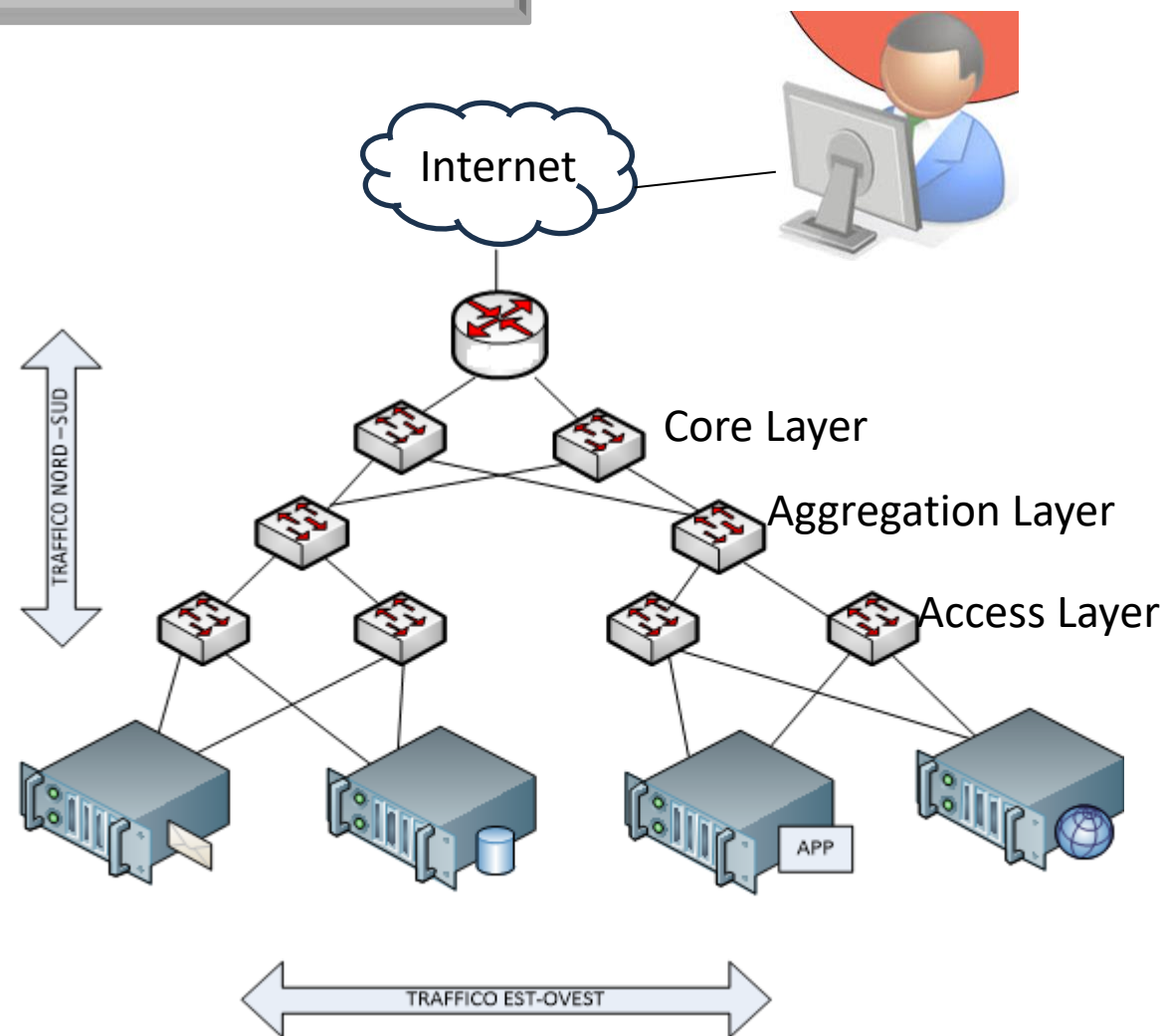


AGENDA

1. Aspetti generali e architetture dei Data Center;
2. Panoramica delle Infrastrutture tecnologiche;
3. Caratteristiche delle sale server e posizionamento degli apparati;
4. Infrastrutture di rete LAN e SAN;
5. Concetti base delle GPU dal punto di vista funzionale ed energetico;
6. Cambiamenti delle infrastrutture tecnologiche e delle sale server, per ospitare i sistemi per l'IA;
7. L'architettura della rete LAN in presenza di sistemi per l'IA;
8. Conclusioni

Nel Data Center i server sono collegati mediante una infrastruttura di rete che consente lo scambio dei dati tra loro e verso l'esterno.

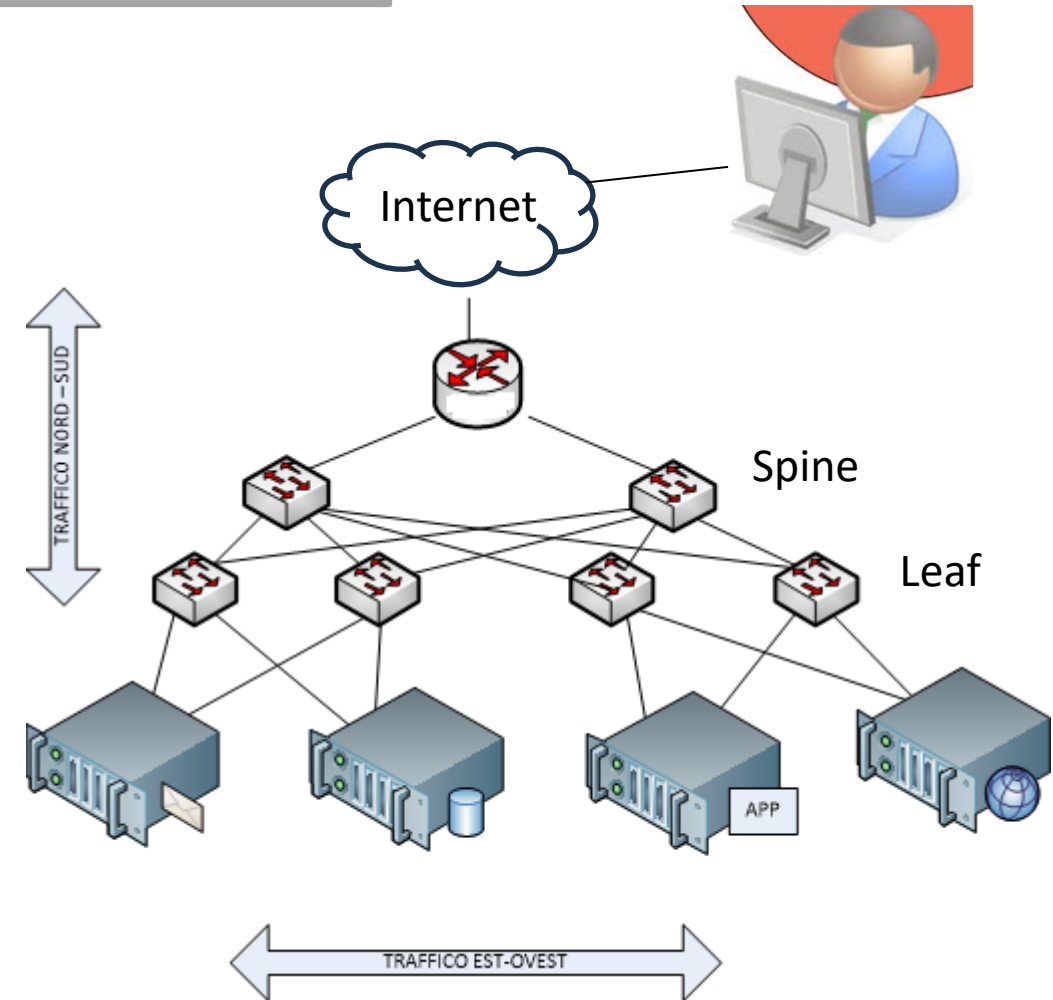
Architettura classica a tre livelli che facilita il traffico esterno – interno detto traffico Nord – Sud rispetto al traffico tra server detto traffico Est – Ovest



Nota: lo schema è un modello semplificato

Architettura Spine-Leaf a tre stadi:

- migliora drasticamente il flusso del traffico Est-Ovest,
- garantisce un numero massimo prevedibile di hop (tipicamente due) tra qualsiasi coppia di server



Nota: lo schema è un modello semplificato

4. Infrastrutture di rete LAN – Protocolli di rete

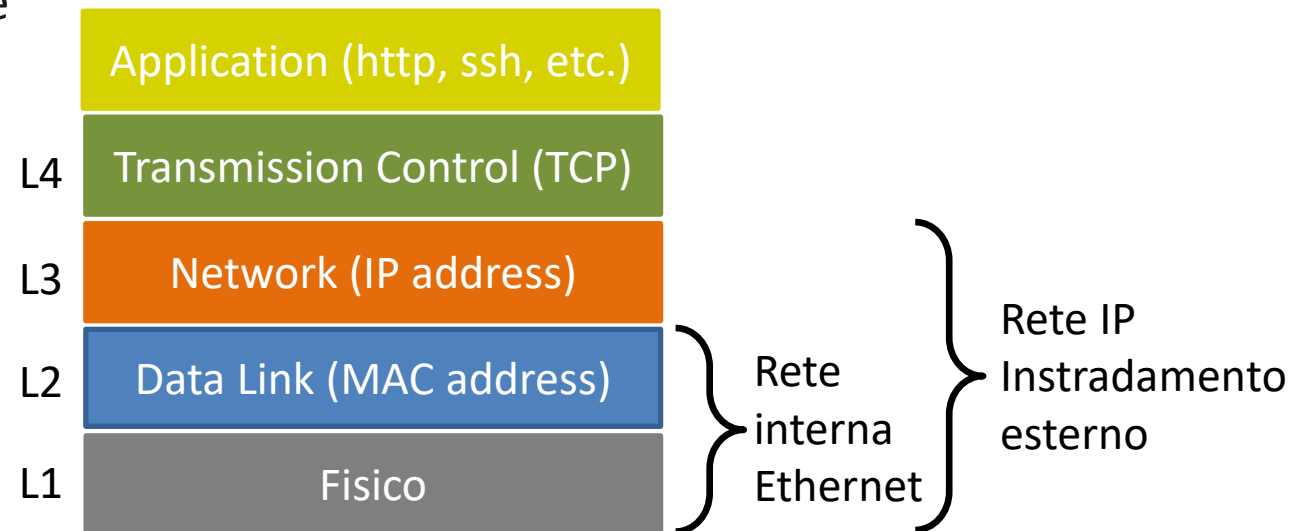
L'architettura logica della rete nei data center convenzionali è basata sul modello a strati in cui ogni strato svolge una funzione specifica di comunicazione:

L4 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

L3 incapsula i dati e gli assegna un indirizzo univoco (IP address) valido a livello di rete e si preoccupa dell'instradamento dei pacchetti tra sottoreti diverse

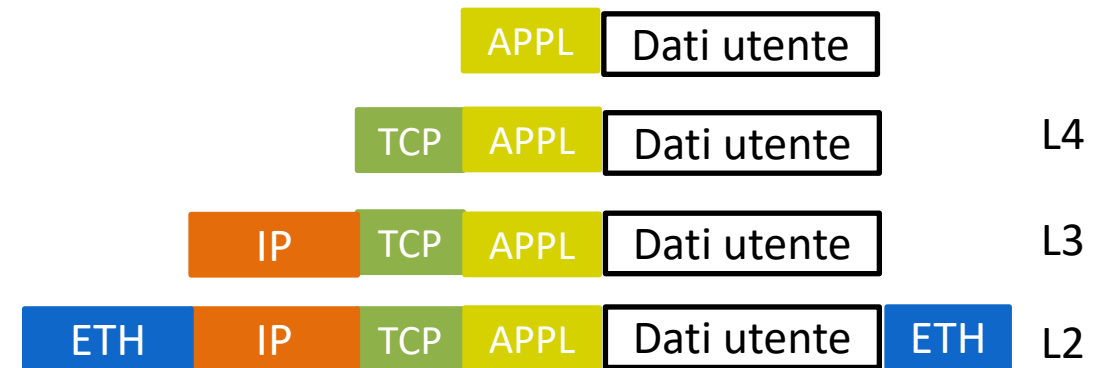
L2 incapsula i dati e gli fornisce un indirizzo di destinazione univoco (MAC address) che vale solo sul segmento di rete interna di destinazione finale

L1 si occupa della trasmissione/ricezione vera e propria traducendo i dati che arrivano da L2 in segnali elettrici e stabilisce un collegamento diretto (via cavo o via radio) tra i sistemi



L'incapsulamento consiste nel fatto che ogni livello aggiunge suoi dati caratteristici in trasmissione che vengono poi tolti in ricezione

L'architettura a livelli aggiunge una grande flessibilità ma a scapito delle prestazioni



Questo protocollo si adatta bene ai flussi di dati tipici che presentano dimensioni e durate diverse, da piccole query a grandi trasferimenti batch e sono in gran parte indipendenti l'uno dall'altro.

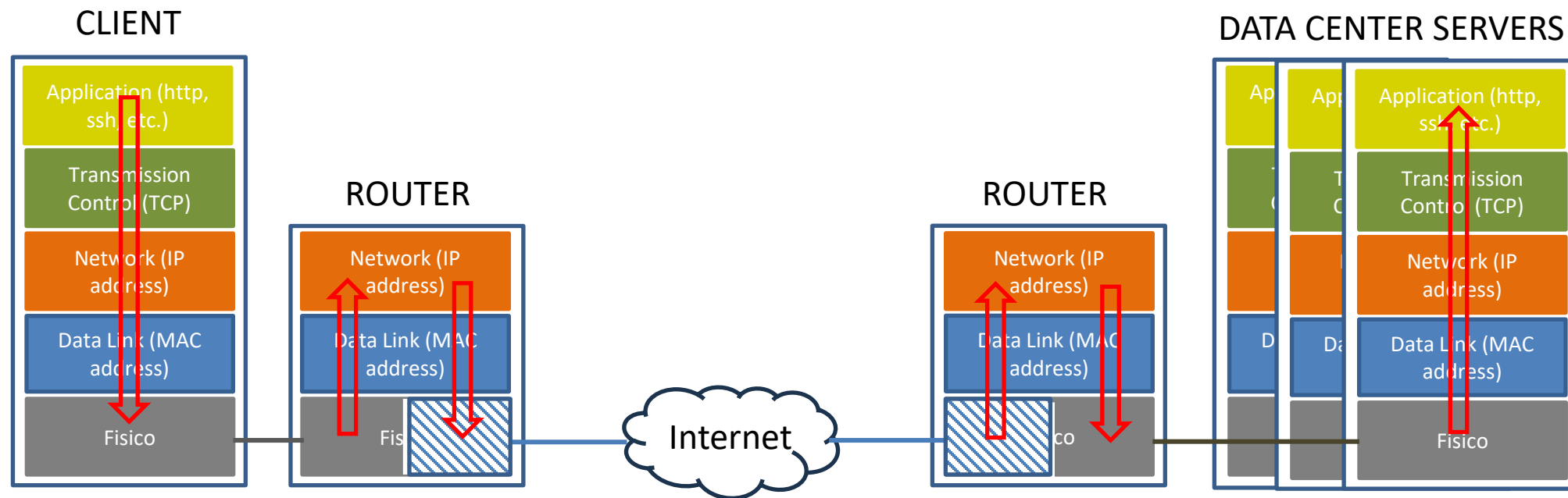
I pacchetti di dati possono arrivare con latenze diverse dell'ordine dei microsecondi (μs) per effetto dei meccanismi di ritrasmissione del TCP in casi di perdite o errori nei pacchetti di dati.

Il protocollo TCP/IP sfrutta il kernel del sistema operativo aggiungendo **latenza di elaborazione** ad ogni transazione

Per i Cluster AI le esigenze di rete determinano un cambio radicale !!

Il protocollo IP (Internet Protocol) fornisce l'indirizzamento logico (IP address) e l'instradamento dei pacchetti (routing) tra reti diverse. Per questo nei router troviamo implementati solo i livelli L1, L2, L3

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



Storage e Storage Area Network (SAN)

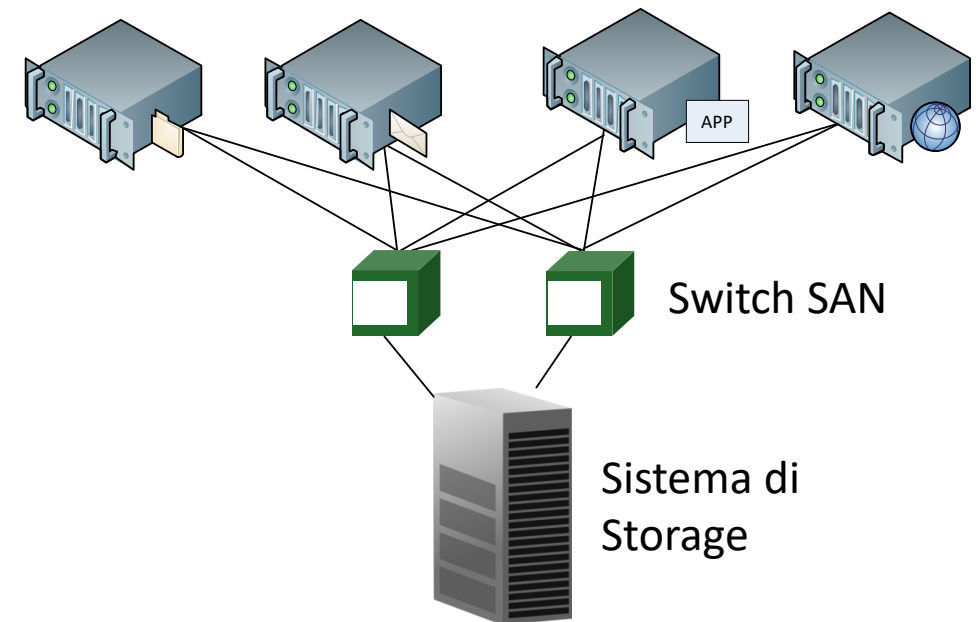
Nei Data Center si rende necessaria la presenza di una infrastruttura centralizzata ad alte prestazioni e alta disponibilità, progettata per fornire ai server capacità di memorizzazione condivisa, scalabile e resiliente

Questa infrastruttura è costituita da sistemi di storage e da una rete di trasmissione dati specializzata.

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 per ambienti mission-critical.

La Storage Area Network (SAN) è una rete ad alta velocità e bassa latenza distinta dalla tradizionale Local Area Network (LAN) e specializzata per interconnettere i server e i sistemi di storage.

L'architettura della rete SAN prevede il protocollo specializzato Fiber Channel (FC) e degli switch dedicati



Il protocollo Fiber Channel (FC) a differenza del TPC/IP consente **la consegna di blocchi di dati in ordine e senza perdite, minimizzando la latenza del protocollo.**

Si può utilizzare anche la rete LAN (in questo caso si parla di Network Attached Storage – NAS) **ma non si ottengono le stesse prestazioni della rete SAN** in quanto il NAS di fatto è un server che mette a disposizione i suoi file e quindi troviamo sia la latenza del protocollo TCP/IP che la latenza del file system

Contrariamente a quanto avviene nel caso del NAS, i server accedono allo storage sulla SAN 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.

I tempi di accesso ai dati da parte del server oltre che dal protocollo sono determinati dalle caratteristiche dello storage che una volta erano Hard Disk rotativi (HDD) mentre adesso si usano memorie allo stato solido Solid State Device (SSD) che hanno tempi di accesso di gran lunga inferiori

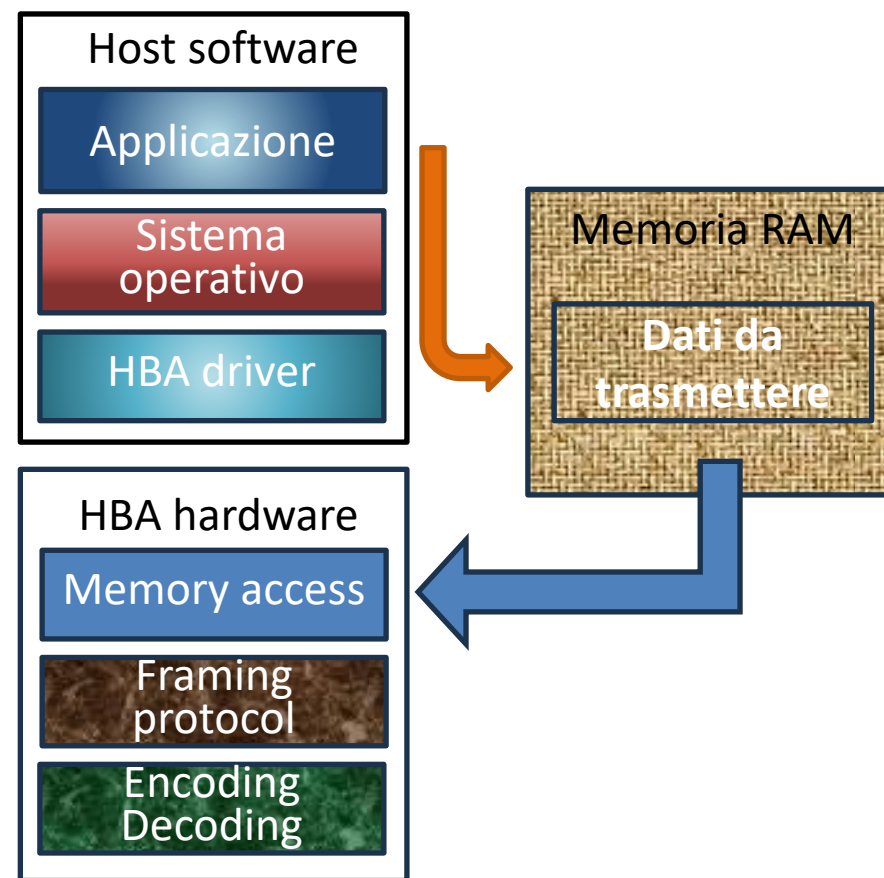
Per ridurre i tempi di accesso allo storage le interfacce fisiche sono state progettate per la prima volta con la tecnica dell'accesso diretto alla memoria RAM del server

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

Il Sistema Operativo passa al Driver HBA il puntatore ai dati da trasmettere. L'hardware dell'HBA preleva i dati dalla memoria del server li formatta secondo il protocollo FC e li invia sulla rete

La HBA si distingue quindi dalla normale Network Interface Card (NIC) della rete LAN del TCP/IP su Ethernet, che non possiede queste specializzazioni hardware e software

L'accesso diretto alla memoria e l'elaborazione del protocollo senza occupare la CPU viene indicato come RDMA ed è un elemento fondamentale per ridurre la latenza, che ritroveremo in ambito AI





AGENDA

1. Aspetti generali e architetture dei Data Center;
2. Panoramica delle Infrastrutture tecnologiche;
3. Caratteristiche delle sale server e posizionamento degli apparati;
4. Infrastrutture di rete LAN e SAN;
5. Concetti base delle GPU dal punto di vista funzionale ed energetico;
6. Cambiamenti delle infrastrutture tecnologiche e delle sale server, per ospitare i sistemi per l'IA;
7. L'architettura delle reti LAN/SAN in presenza di sistemi per l'IA;
8. Conclusioni

Graphics Processing Units (GPU) sono strutturate per eseguire calcoli simili su grandi quantità di dati, come sono per esempio i valori dei pixel di una immagine (luminanza e colore RGB)

Le GPU contengono da centinaia a migliaia di core di elaborazione che lavorano in parallelo, ottimizzati principalmente per eseguire operazioni aritmetiche ad alta velocità

Perché sono richieste le GPU ?

I modelli linguistici di grandi dimensioni (LLM) per l'AI generativa richiedono operazioni non lineari, moltiplicazioni tra matrici di grandi dimensioni ed altre funzioni che se svolte in maniera sequenziale, come sarebbe fatto nelle CPU richiederebbero ore e ore di elaborazione

La presenza delle GPU richiede un cambiamento notevole nell'architettura dei Data Center:

1. Tutti questi core che lavorano in parallelo richiedono molta potenza elettrica per poter funzionare e di questa potenza elettrica il 95% e 99% di essa viene trasformata in calore
2. Le GPU lavorano ad alta velocità e in modo altamente sincronizzato e quindi le reti LAN e SAN devono essere ad elevata larghezza di banda e latenza minima

5. Concetti base delle GPU

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

Componenti Chiave dei Nodi di Calcolo (Server GPU)

1. **GPU (Graphics Processing Units):** Rappresentano gli acceleratori primari per i calcoli AI. Modelli di punta includono le GPU NVIDIA H100, A100, L4 e le AMD Instinct
2. **CPU (Central Processing Units):** Gestisce il sistema operativo e l'orchestrazione dei task computazionali
3. **Memoria di Sistema (RAM):** È utilizzata dalla CPU per il sistema operativo, le applicazioni e come area di buffering per il trasferimento dei dati da e verso le GPU.
4. **Memoria GPU (HBM - High Bandwidth Memory):** Si tratta di una memoria ad altissima velocità **integrata direttamente nelle GPU**. La sua elevata larghezza di banda è cruciale per le prestazioni, in quanto ospita i dati e i parametri del modello AI durante le fasi di calcolo intensivo.
5. **Storage locale NVMe SSD (Non-Volatile Memory Express Solid State Drives):** Unità a stato solido collegate sul bus di sistema che sostituiscono gli hard disk tradizionali o i moderni SSD con interfaccia SATA. Sono essenziali per non rallentare le operazioni
6. **Interfacce di rete** ad alte prestazioni tipo RDMA (Infiniband o nuove tecnologie tipo RDMA over Converged Ethernet che sfruttano il noto protocollo)

Addestramento AI

E' la fase più impegnativa poiché l'addestramento viene fatto in parallelo sulle GPU in funzione dalla grandezza del modello:

Parallelismo dei Dati (Data Parallelism): Modello replicato su ogni GPU, dataset partizionato (mini-batch), elaborazione dei dati in parallelo sul modello, aggregazione finale dei dati.

Parallelismo del Modello (Model Parallelism): Modello partizionato su più GPU (per modelli molto grandi).

Criticità: Efficienza della comunicazione inter-GPU e accesso rapido ai dati per evitare "GPU starvation" (GPU inattive in attesa di dati) – Presenza di flussi «elefante» di dati e traffico many-to-many tra i nodi

Inferenza AI

Richiede minori prestazioni, il traffico inter-GPU è sostenuto solo durante la generazione della risposta, le prestazioni della rete rimangono importanti per ottimizzare la rapidità della risposta



AGENDA

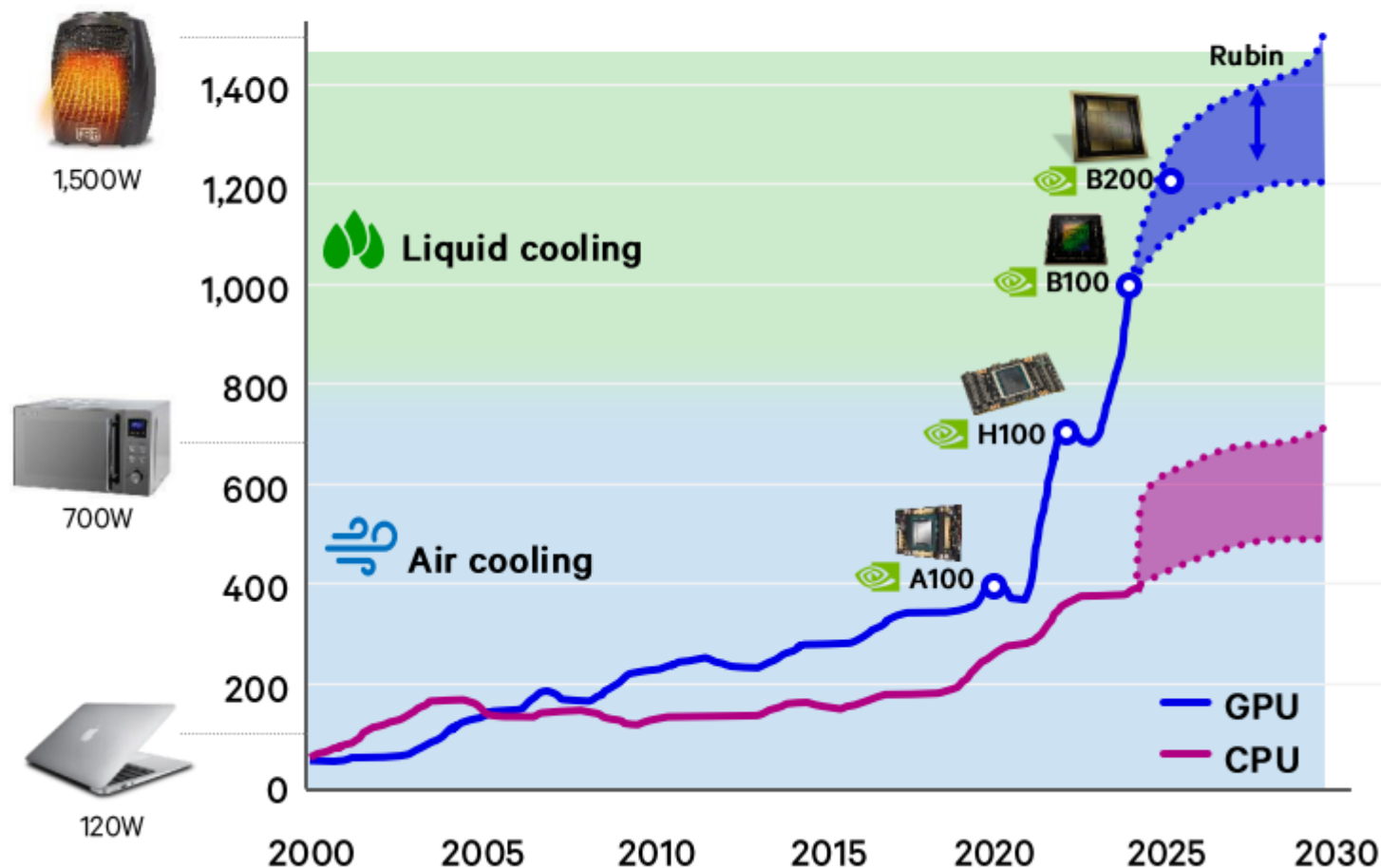
1. Aspetti generali e architetture dei Data Center;
2. Panoramica delle Infrastrutture tecnologiche;
3. Caratteristiche delle sale server e posizionamento degli apparati;
4. Infrastrutture di rete LAN e SAN;
5. Concetti base delle GPU dal punto di vista funzionale ed energetico;
6. Cambiamenti delle infrastrutture tecnologiche e delle sale server, per ospitare i sistemi per l'IA;
7. L'architettura delle reti LAN/SAN in presenza di sistemi per l'IA;
8. Conclusioni

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

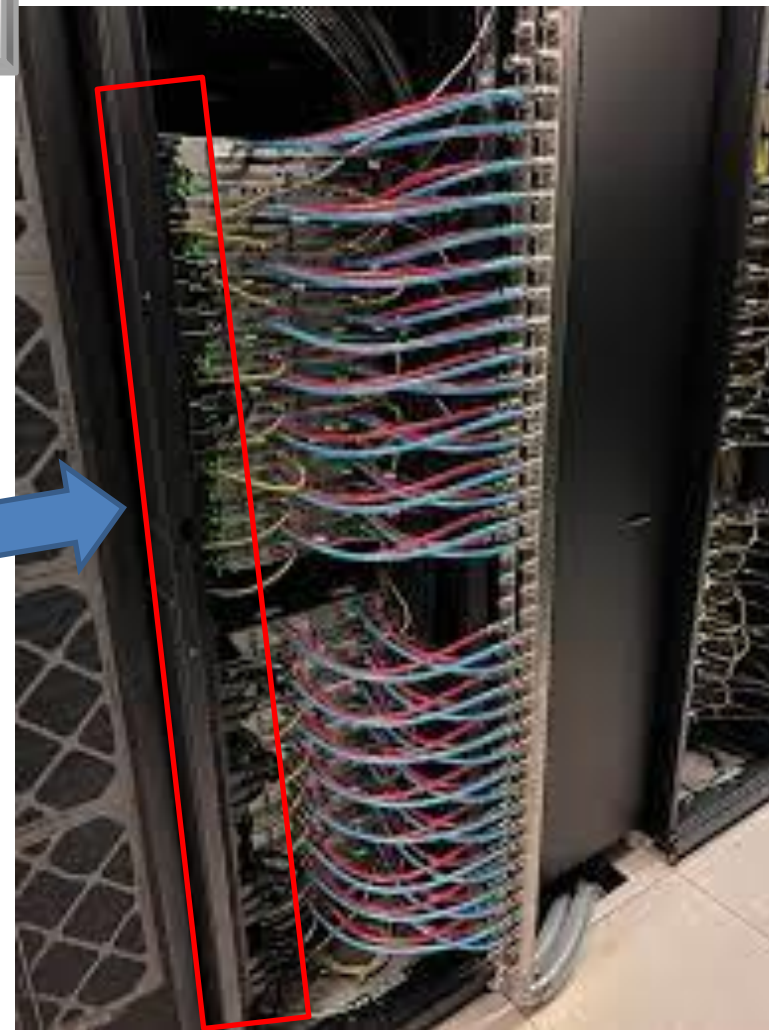


I rack per AI dedicati a grandi modelli LLM richiedono una potenza 40 – 80 kW destinata ad aumentare.

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

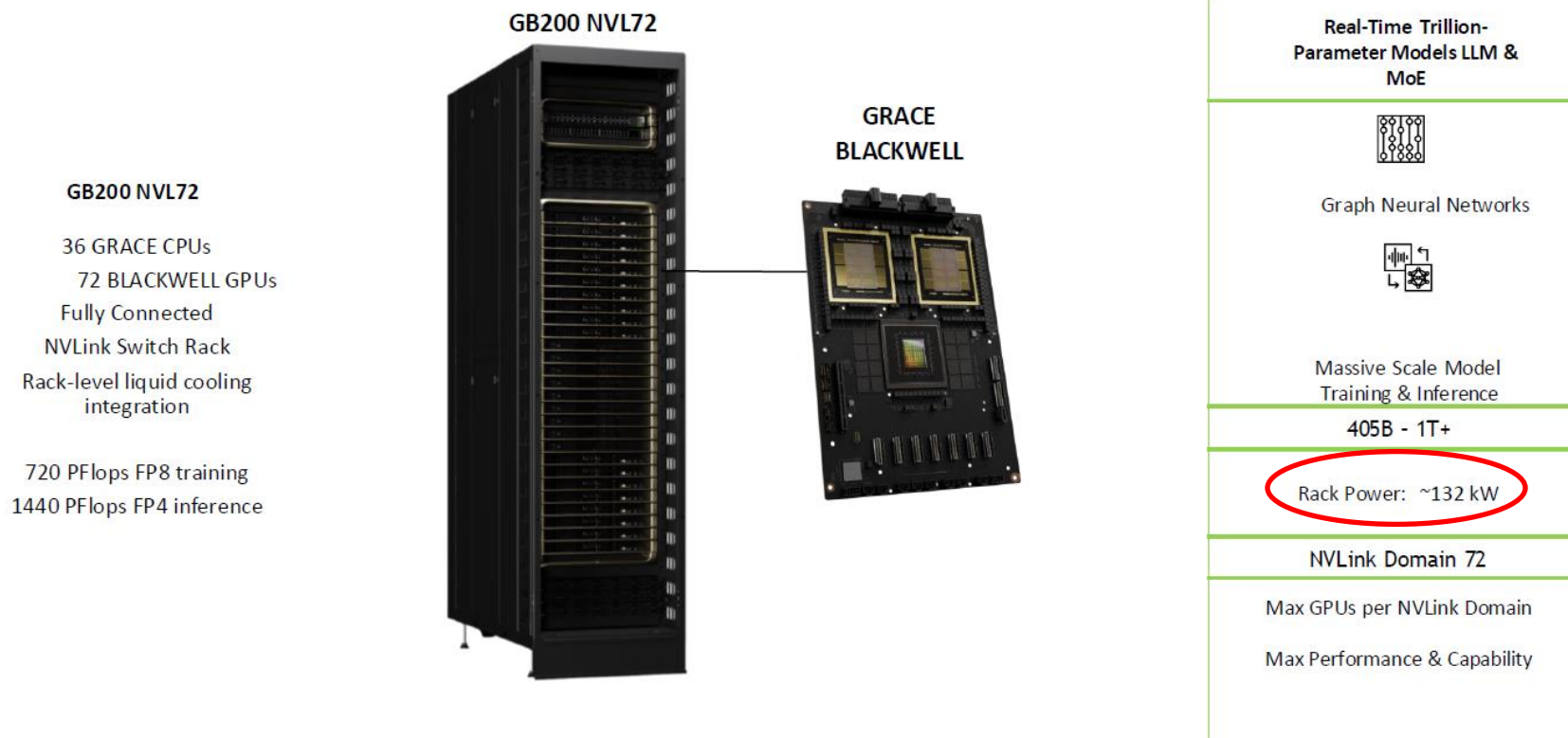


Manifold per la distribuzione del liquido di raffreddamento ai server

6. Infrastrutture per AI

The Future of NVIDIA AI Infrastructure

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





AGENDA

1. Aspetti generali e architetture dei Data Center;
2. Panoramica delle Infrastrutture tecnologiche;
3. Caratteristiche delle sale server e posizionamento degli apparati;
4. Infrastrutture di rete LAN e SAN;
5. Concetti base delle GPU dal punto di vista funzionale ed energetico;
6. Cambiamenti delle infrastrutture tecnologiche e delle sale server, per ospitare i sistemi per l'IA;
7. L'architettura delle reti LAN/SAN in presenza di sistemi per l'IA;
8. Conclusioni

I carichi di lavoro basati sulla AI impongono requisiti radicalmente diversi per la rete locale (LAN) dei data center rispetto alle applicazioni convenzionali.

Il modello di traffico attuale:

Vasta gamma di applicazioni con flussi indipendenti di traffico prevalentemente di tipo Nord-Sud (traffico server – utenti), latenze dell'ordine dei uS e tolleranza alla perdita di pacchetti gestite dal protocollo TCP/IP sono accettabili e in genere non creano impatti all'utente.

Modello di traffico dell'AI:

Operazioni di comunicazione collettiva costante e sincronizzata con varianti di multicast o broadcast prevalentemente traffico Est – Ovest (server to server). Le operazioni utilizzate nell'addestramento All-Reduce (sincronizzazione dei gradienti) e All-Gather (raccolta dati) richiedono che ogni GPU del cluster comunichi potenzialmente con tutte le altre.

I dataset in gioco sono molto grandi soprattutto con l'uso di immagini o video, quindi la latenza e la perdita di pacchetti sono fattori che impattano gravemente sull'efficienza del processo di addestramento

Le caratteristiche principali richieste dalla rete LAN per AI:

- Enorme Richiesta di Banda (100 Gbps comuni, 400Gbps/800Gbps e oltre a tendere)
- Latenza Ultra bassa
- Tolleranza Zero alla perdita di pacchetti

rendono necessario l'uso di architetture tipo Leaf-Spine e il passaggio a protocolli di rete con accesso diretto alla memoria e meccanismi di consegna di dati in ordine e senza perdite, indicati genericamente come Remote Direct Memory Access (RDMA).

Gli standard più usati sono Infiniband di tipo proprietario e RoCEv2 (RDMA over Converged Ethernet) basato sul protocollo Ethernet che consente la coesistenza del traffico RDMA con il traffico Ethernet standard.

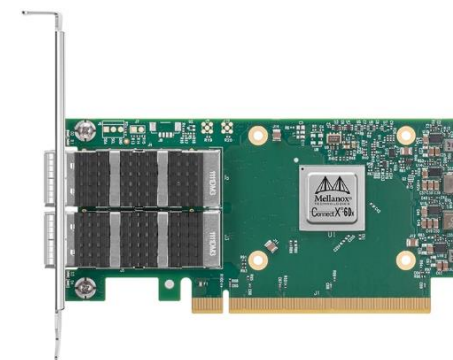
L'utilizzazione di questi standard richiede ovviamente la presenza di apparati di rete dedicati come schede di rete con supporto RDMA, switch di rete adeguati e specifiche configurazioni.

Da notare che il consumo elettrico aumenta con la velocità dell'apparato e protocollo usato

7. Infrastrutture di rete LAN e SAN impatti della AI

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

Esigenze di Storage dei Carichi di Lavoro AI

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

Addestramento fase più critica per lo storage caratterizzata da letture sequenziali massive di grandi dataset per alimentare gli algoritmi di apprendimento. Necessario un throughput elevato per evitare la sotto-utilizzazione delle GPU. Il training comporta 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 Comporta letture casuali di porzioni di dati più piccole, i parametri del modello da utilizzare e i dati in input per la previsione.

Per soddisfare le richieste di I/O della AI sono necessarie le unità di storage a stato solido (SSD) tipo All-Flash-Array (AFA) oltre a una infrastruttura SAN particolarmente performante

7. Infrastrutture di rete LAN e SAN impatti della AI

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.

Per mantenere le elevate prestazioni dello storage SSD e del protocollo NVMe, come livello di trasporto conviene usare oltre ai FC protocolli come RoCEv2 o Infiniband. Per esempio può essere usato anche il TCP/IP ma con prestazioni inferiori.





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)	○ Basso (~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







<https://www.youtube.com/watch?v=Jf8EPSBZU7Y&t=746s>

Ing. Mario D'Ettore

Efficienza energetica nei Data Center

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
- Gruppi elettrogeni e sistemi di continuità assoluta (UPS) di capacità adeguata
- Impianto di Raffreddamento più complesso, a liquido Direct to CHIP e raffreddamento classico ad aria per il resto dei componenti dei server
- Infrastruttura di rete adeguata a livello di cabling (esclusivamente in fibra ottica) e apparati di rete e con architetture progettate per offrire **altissima larghezza di banda, latenza ultra-bassa** e funzionamento **senza perdite su larga scala** (RDMA - Remote Direct Memory Access o Infiniband allo stato attuale)
- Infrastruttura 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