Future Computing Architecture

3rd and 4th Lessons

Marco Briscolini, PhD

marco.briscolini@gmail.com

Cell: 3357693820

03/19/2025

Piano del Corso – 16 ore in 8 moduli

Descrizione generale delle architetture HPC e AI e loro componenti di base
- Le previsioni di mercato AI&HPC nel mondo
- Componenti principali: parte computazionale, rete di interconnessione, sottosistema storage
- Concetti di metrica delle varie componenti (misurazione della capacita' computazionale, trasmissione dati, lettura/scrittura dati)
- Metriche riconosciute a livello mondiale (Top500, Green500, IO500)
- Concetti introduttivi sull'analisi della complessita' computazionale di un ambito applicativo

Architetture di calcolo e loro evoluzione
- Architetture omogenee e accelerate
- Concetti generali sui microprocessori (CPU)
- Concetti generali sugli acceleratori (Graphical Processor Unit)
- Integrazione CPU-GPU e trasmissione dati

Reti a alte prestazioni per architetture HPC e AI e loro evoluzione
- Reti con protocollo Infiniband e alcune topologie correlate
- Reti di tipo Ethernet a alte prestazioni
- Protocolli RDMA e RoCE

Sottosistemi storage a alte prestazioni e loro evoluzione
- Concetti generali sulla gerarchia dei sottosistemi storage
- Sistemi a disco magnetico e a stato solido
- Connessione di sistemi storage su SAN, Infiniband, Ethernet, nVME over Fabric, e altro

Architetture storage a alte prestazioni
- Architetture di sottositemi storage
- Filesystem paralleli per lettura/scrittura a alte prestazioni

06

Problematiche di efficientamento energetico per sistemi HPC a grande scala (architetture pre e exascale)
- Il concetto di PUE e di efficienza energetica a parita' di potenza computazionale
- Come le varie architetture si caratterizzano in termini di "Potenza di Calcolo"/Watt
- Utilizzo di tecniche di gestione del carico di lavoro per ottimizzare l'efficienza energetica
- Soluzioni di raffreddamento a aria, a acqua diretta e immersivo
- Concetti generali sul disegno e la realizzazione di Data Center efficienti

07

Accenni sulle architetture innovative in ambito AI&HPC
- Architetture AI scalabili
- Interconnessione tra sistemi AI
- AI/HPC/Q-C architettura integrata per carichi computazionali complessi

08

Accenni al disegno e alla progettazione di un'architettura HPC
- Definizione di specifiche di progetto
- Valutazione preliminare dell'architettura ottimale
- Disegno di massima dell'architettura

- Concetto di rispondenza e verifica alle specifiche di progetto

Piano del Corso – Lesson 3

Reti a alte prestazioni per architetture HPC e AI e loro evoluzione

- Reti con protocollo Infiniband e alcune topologie correlate
- Reti di tipo Ethernet a alte prestazioni
- Protocolli RDMA e RoCE

Cosa compone generalmente un'architettura AI&HPC di ultima generazione

Elaborazione:

un numero di nodi di calcolo (≥16) di varia tipologia
- Nodi single o dual CPUs
- Nodi oltre dual CPUs
- Nodi con acceleratori GPU

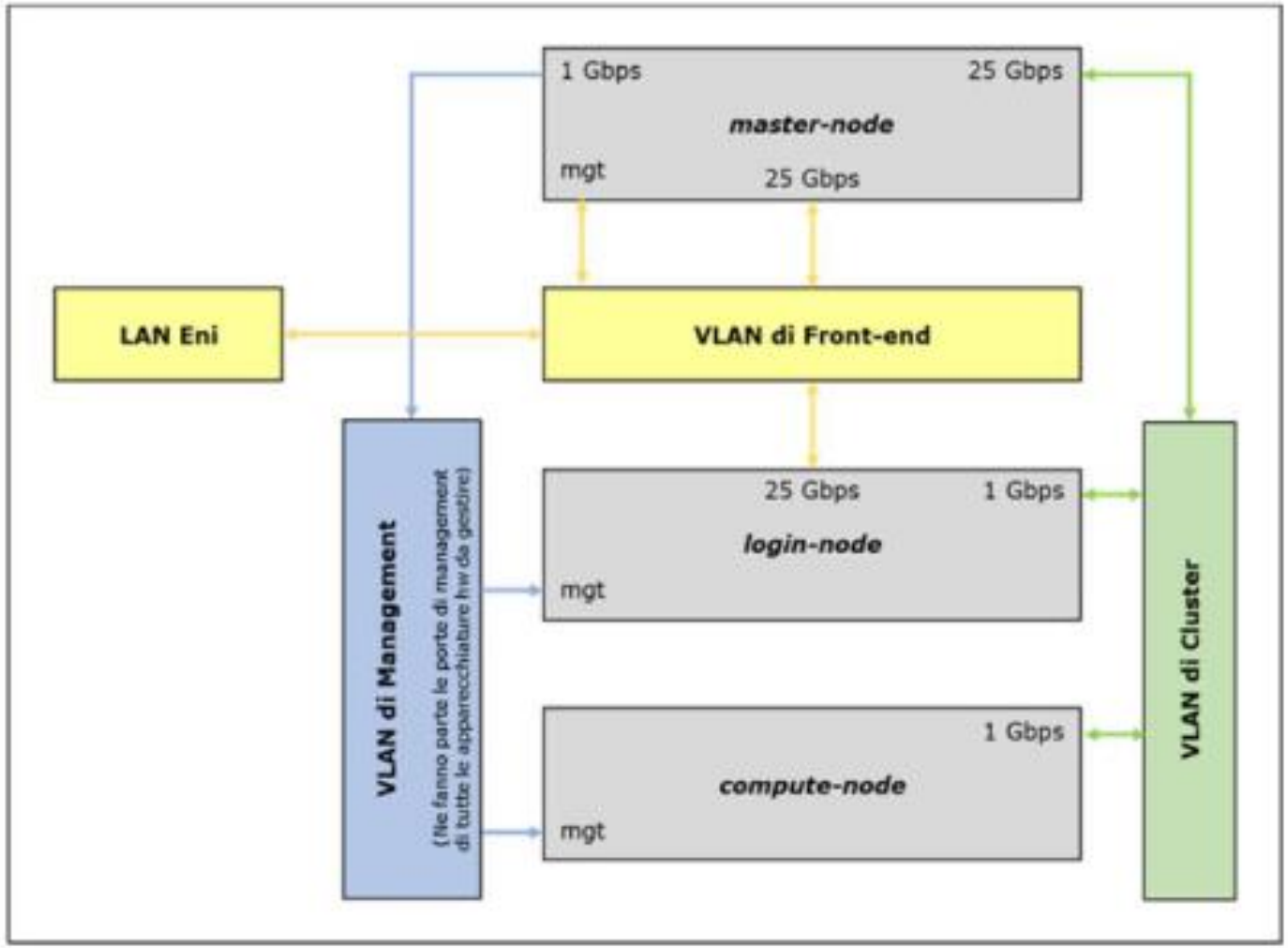
Comunicazione:
Rete a media-elevate BW (≳10Gbs) e media-bassa latenza (≲100μsec)
- Varie topologie per supportare un'elevata scalabilita' orizzontale
- Rete di gestione generalmente su protocollo Ethernet per limitato scambio dati

Sottosistema storage:
- Varie tipologie di sottosistemi storage (flash, hdd, tape)
- Multi tier storage per gestione dati efficiente

Sistemi di servizio:
- Nodi di gestione dell'architettura
- Nodi per la gestione del carico di lavoro

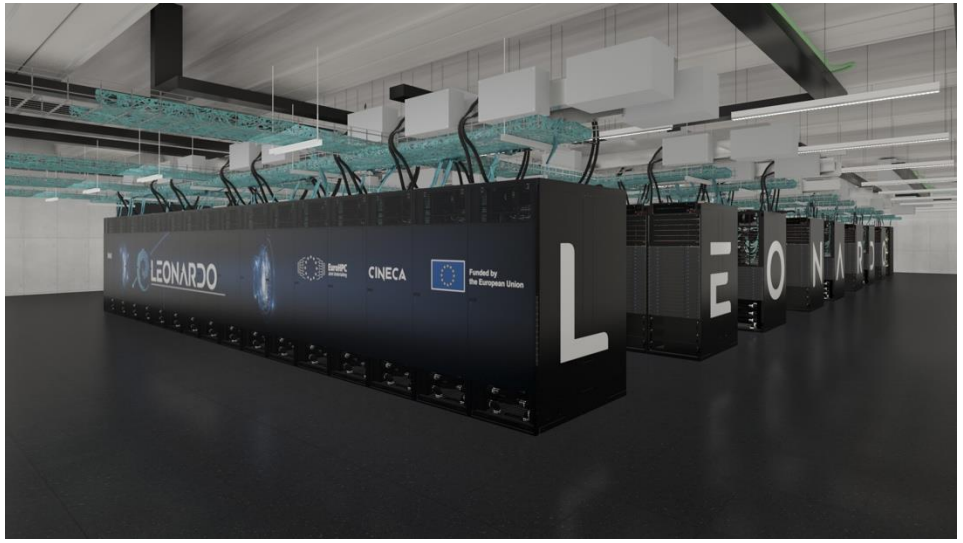
https://io500.org/

Esempio di schema a blocchi di un'architettura AI&HPC



https://io500.org/

CINECA - Leonardo, the World's Largest AI hybrid Supercomputer **Italy**

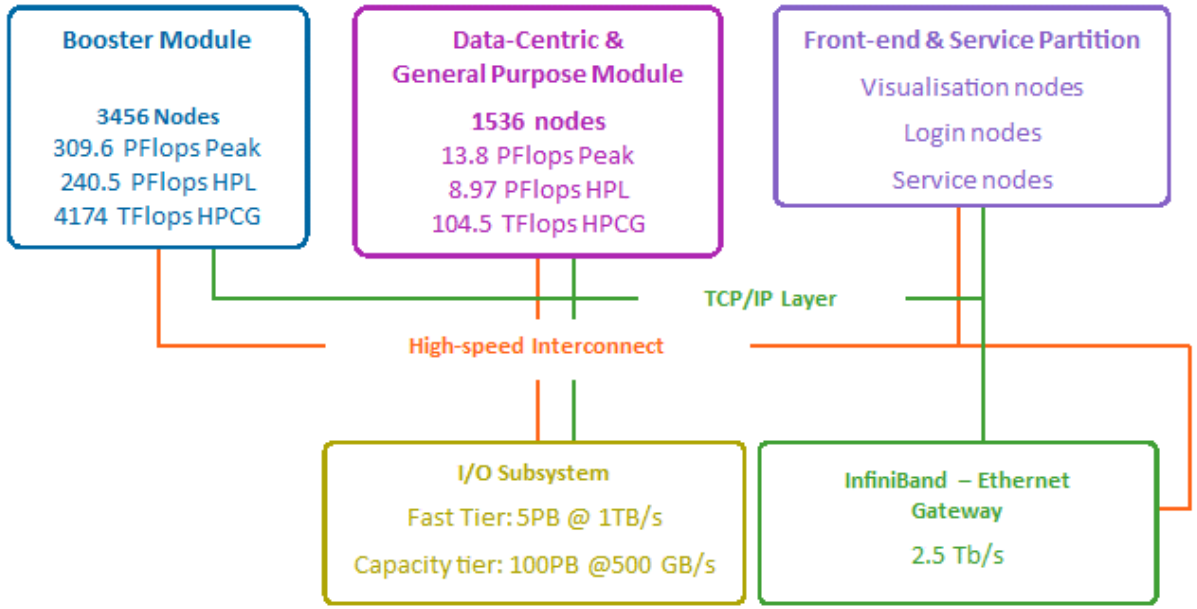


10 ExaFlops FP16 AI workload

320 PFlops (Rpeak)

250 PFlops (Linpack)

>4 PFlops (HPCG)



"CINECA plays a critical part in evolving both the research and industrial community in accelerated HPC application development. The Leonardo supercomputer is the result of our long-term commitment to pushing the boundaries of what a modern exascale supercomputer can be."
Sanzio Bassini, Director of the HPC department at CINECA

https://it.wikipedia.org/wiki/Leonardo_%28supercomputer%29

Tipologia di rete in un'architettura AI&HPC

Rete di gestione a banda stretta e latenza elevata: BW ~ 1Gbs, λ ~100msecs
- Rete Ethernet a 1Gbs con una o due connessioni fisiche/logiche per nodo
- Gestione on-off nodi, propagazione firmware, analisi funzionamento nodi
- Propagazione sistema operativo, ambienti di programmazione, altri SW
- Gestione e verifica funzionamento altre componenti varie (switch, rack, PDUs, etc)

Rete a banda intermedia: BW ~ 10Gbs, λ ~100msecs (opzionale)
- Rete Ethernet a ~10Gbs con una o due connessioni fisiche/logiche per nodo
- Sincronizzazione parallel filesystem
- Gestione workload scheduler
- Trasmissione protocolli di virtualizzazione, ambienti Cloud, propagazione SW, etc
- Limitata capacita' per comunicazioni parallel processing (MPI)

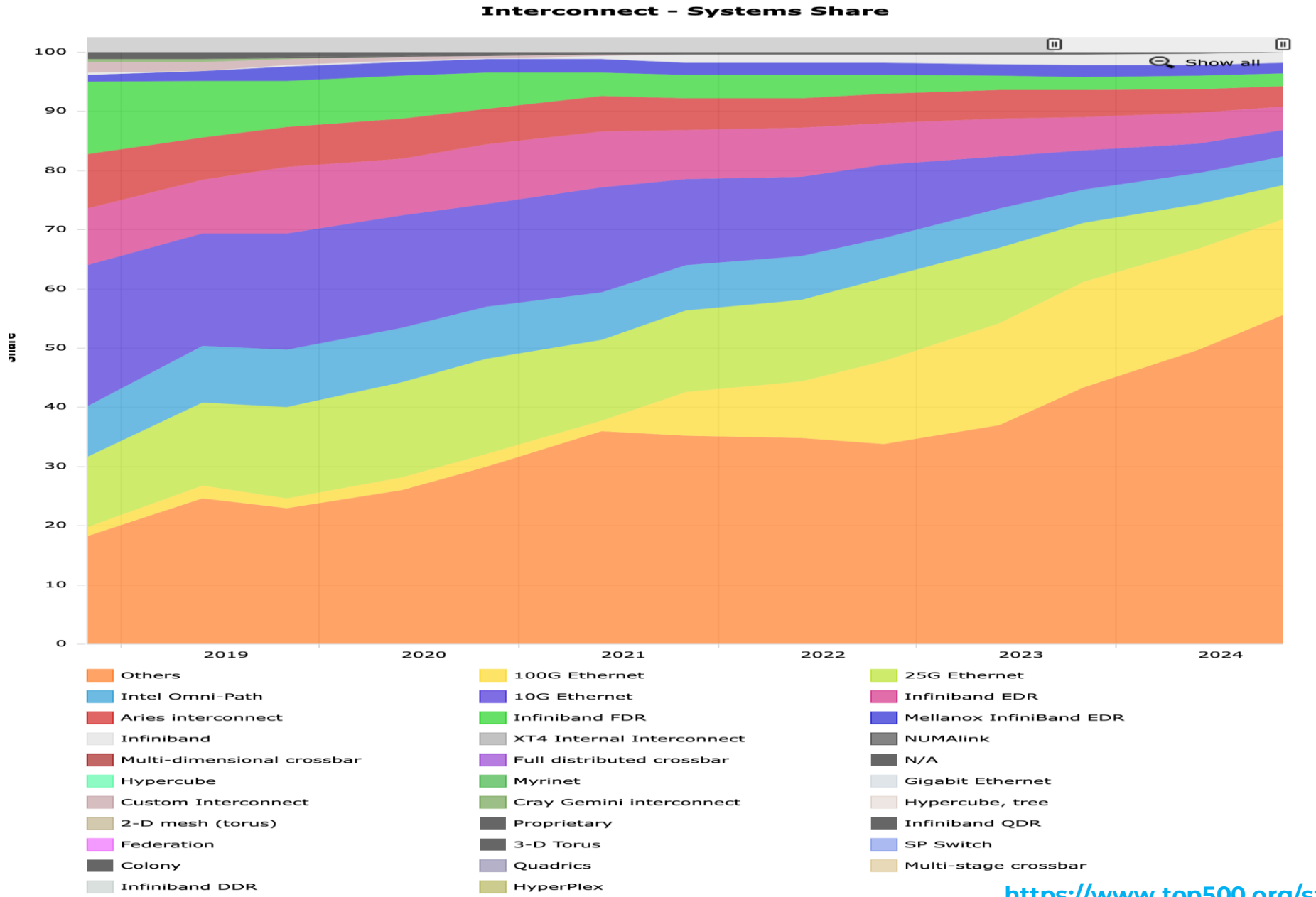
Rete a banda larga e bassa latenza: BW ≥ 100Gbs, λ ≤ 5µsecs
- Varie topologie per supportare un'elevata scalabilita' orizzontale
- Rete Ethernet e protocolli TCPIP e RoCE
- Rete Infiniband e protocollo RDMA
- Trasferimento dati per IO a late prestazioni
- Trasferimento dati per programmazione parallela (MPI)

https://community.fs.com/it/article/roce-vs-infiniband-vs-tcp-ip.html

https://access.redhat.com/documentation/it-it/red_hat_enterprise_linux/8/html/configuring_infiniband_and_rdma_networks/understanding-infiniband-and-rdma_configuring-infiniband-and-rdma-networks

https://io500.org/

Differenti reti a alte prestazioni in un'architettura AI&HPC – 2012-25



https://www.top500.org/statistics/overtime/

Tecnologia delle reti generalmente utilizzate in un'architettura AI&HPC

Desc/Tipo	Ethernet	Infiniband	SAN	Note
Bandwidth Gbs	≥1	≥100	≥8	Valori di riferimento
Latenza Secs	~100msec	~µsec	~100msec	Dati msec MPI µsec
Protocollo	TCP-IP, RoCE	RDMA, IPoIB	SAN	Vari utilizzi
Utilizzo	Gestione, Dati, MPI, Cloud	Dati e MPI	Dati	Connessione sistemi storage

https://community.fs.com/it/article/roce-vs-infiniband-vs-tcp-ip.html

https://access.redhat.com/documentation/it-it/red_hat_enterprise_linux/8/html/configuring_infiniband_and_rdma_networks/understanding-infiniband-and-rdma_configuring-infiniband-and-rdma-networks

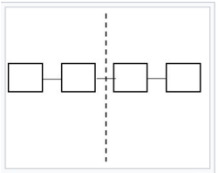
https://io500.org/

La banda passante di una rete e alcune misurazioni – bisection bandwidth

Bisection bandwidth

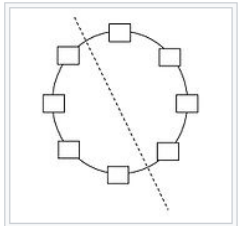
In computer networking, if the network is bisected into two equal-sized partitions,
the **bisection bandwidth** of a network topology is the bandwidth available between the two partitions.
Bisection should be done in such a way that the bandwidth between two partitions is minimum.
Bisection bandwidth gives the true bandwidth available in the entire system.
Bisection bandwidth accounts for the bottleneck bandwidth of the entire network.
Therefore bisection bandwidth represents bandwidth characteristics of the network better than any other metric.

For a linear array with n nodes bisection bandwidth is one link bandwidth. For linear array only one link needs to be broken to bisect the network into two partitions.



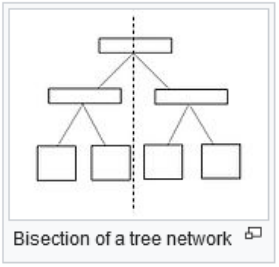
Bisection of linear array network

For ring topology with n nodes two links should be broken to bisect the network, so bisection bandwidth becomes bandwidth of two links.



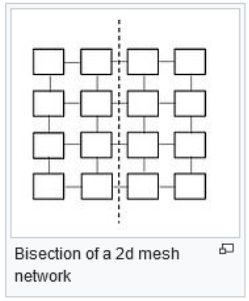
Bisection of a ring network

For tree topology with n nodes can be bisected at the root by breaking one link, so bisection bandwidth is one link bandwidth.



Bisection of a tree network

For Mesh topology with n nodes, $\sqrt{n}$ links should be broken to bisect the network, so bisection bandwidth is bandwidth of $\sqrt{n}$ links.



Bisection of a 2d mesh network

https://en.wikipedia.org/wiki/Bisection_bandwidth

La banda passante di una rete e alcune misurazioni – fattore di oversubscription

Oversubscription rate

E’ un indicatore generalmente usato per tenere in conto il decadimento di banda passante tra i vari livell che compongono una rete.

1:1 oversubscription indica che non c’e’ decadimento della banda passante tra un livello e quello superior

2:1 oversubscription indica che la banda si dimezza passando dal livello inferiore al superiore

Per architetture di elevate dimensioni spesso si definiscono delle zone con 1:1 oversubscription ma che sono connesse a altre con livelli 2:1 o 3:1 o superiori di oversubscription

https://en.wikipedia.org/wiki/Bisection_bandwidth

Componenti di una rete Infiniband

L'architettura InfiniBand definisce quattro classi di nodi che compongono la rete di comunicazione:

- Host Channel Adapter (HCA): risiede sui server o comunque su piattaforme dotate di Cpu e gestisce la richieste di trasmissione.
- Target Channel Adapter (TCA): dualmente all'HCA, risiede tipicamente sui sistemi di immagazzinamento dati con lo scopo di ricevere le richieste dei server.
- Switch: si occupa della gestione del traffico all'interno di una medesima sottorete.
- Router: instrada il traffico tra le diverse sottoreti.

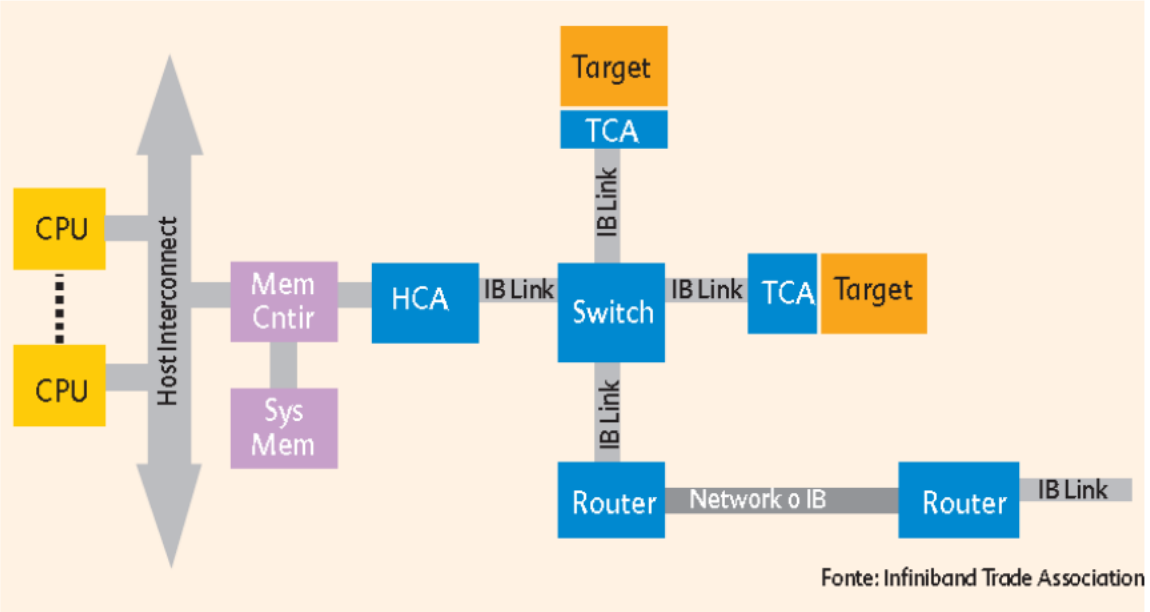


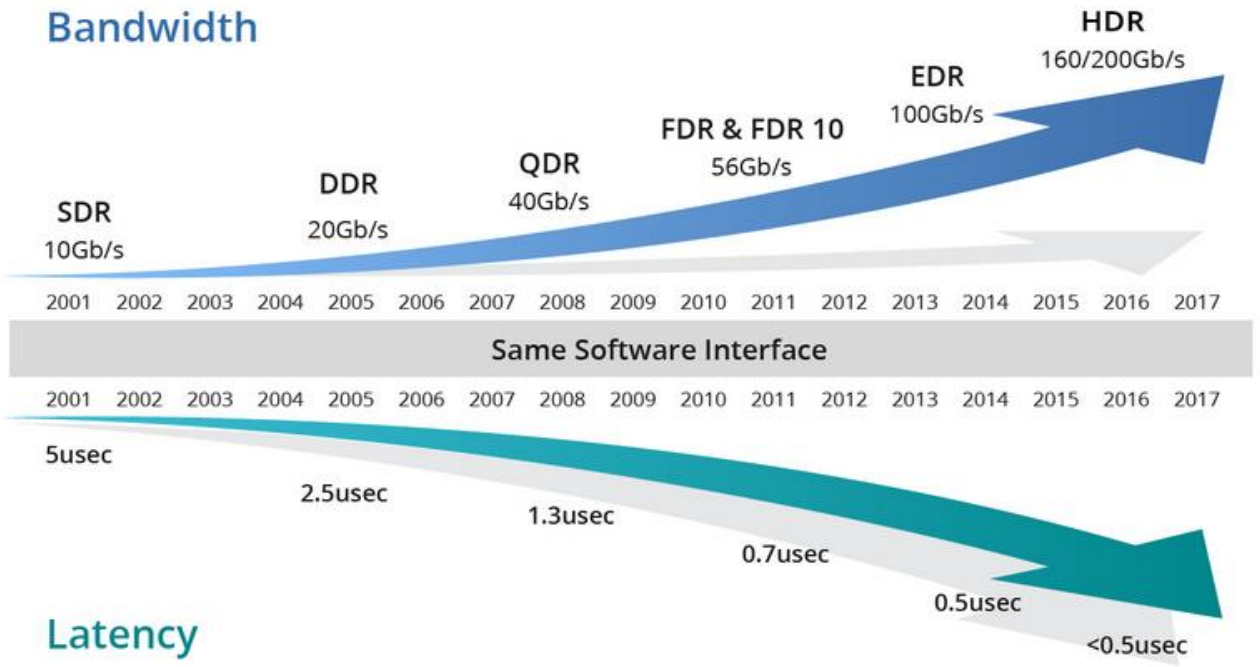
Figura 10 - Connessioni e attori InfiniBand

Tesi di laurea: "Analisi e ottimizzazione della rete InfiniBand del sistema HPC CRESCO di ENEA", Andrea Petricca, UniRoma1 a.a. 2008-09

Sviluppo rete Infiniband

Velocità massime teoriche in differenti configurazioni

	Singolo (SDR)	Doppio (DDR)	Quadruplo (QDR)
1X	2.5 Gbit/s	5 Gbit/s	10 Gbit/s
4X	10 Gbit/s	20 Gbit/s	40 Gbit/s
12X	30 Gbit/s	60 Gbit/s	120 Gbit/s



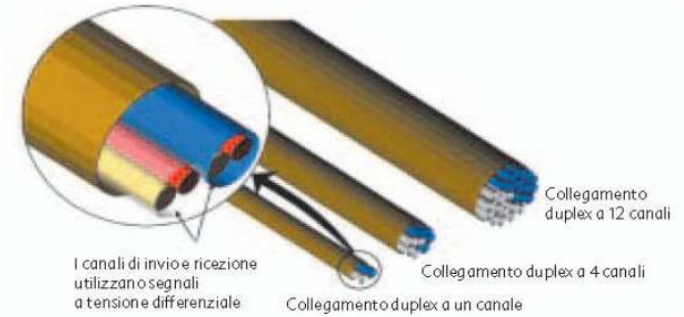


Figura 8 - Coppie 1X, 4X e 12X di un cavo InfiniBand

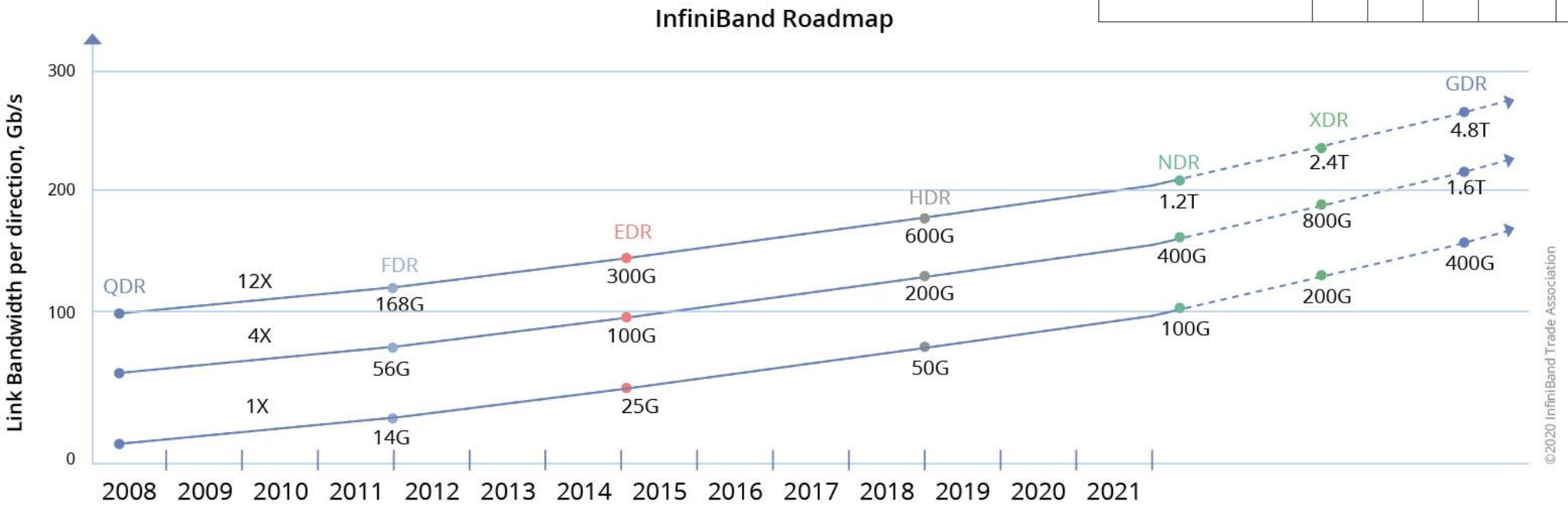
https://www.eneagrid.enea.it/papers_presentations/papers/TesiAPetriccaInfiniband.pdf

https://community.fs.com/it/article/why-hpc-data-centers-need-infiniband-interconnection.html

https://it.wikipedia.org/wiki/InfiniBand

Sviluppo rete Infiniband

	SDR	DDR	QDR	FDR10	FDR	EDR	HDR	NDR	XDR
Adapter latency (microseconds)	5	2.5	1.3	0.7	0.7	0.5			
Signaling rate (Gbit/s)	2.5	5	10	10.3125	14.0625	25.78125	50	100	250
Speeds for 4x links (Gbit/s)	8	16	32	40	54.54	100	200		
Speeds for 8x links (Gbit/s)	16	32	64	80	109.08	200	400		
Speeds for 12x links (Gbit/s)	24	48	96	120	163.64	300	600		
Year	2001	2005	2007	2011	2011	2014	2017	after 2020	future



https://community.fs.com/it/article/infiniband-insights-powering-highperformance-computing-in-the-digital-age.html

Sviluppo tecnologico della rete Infiniband

Performance [edit]

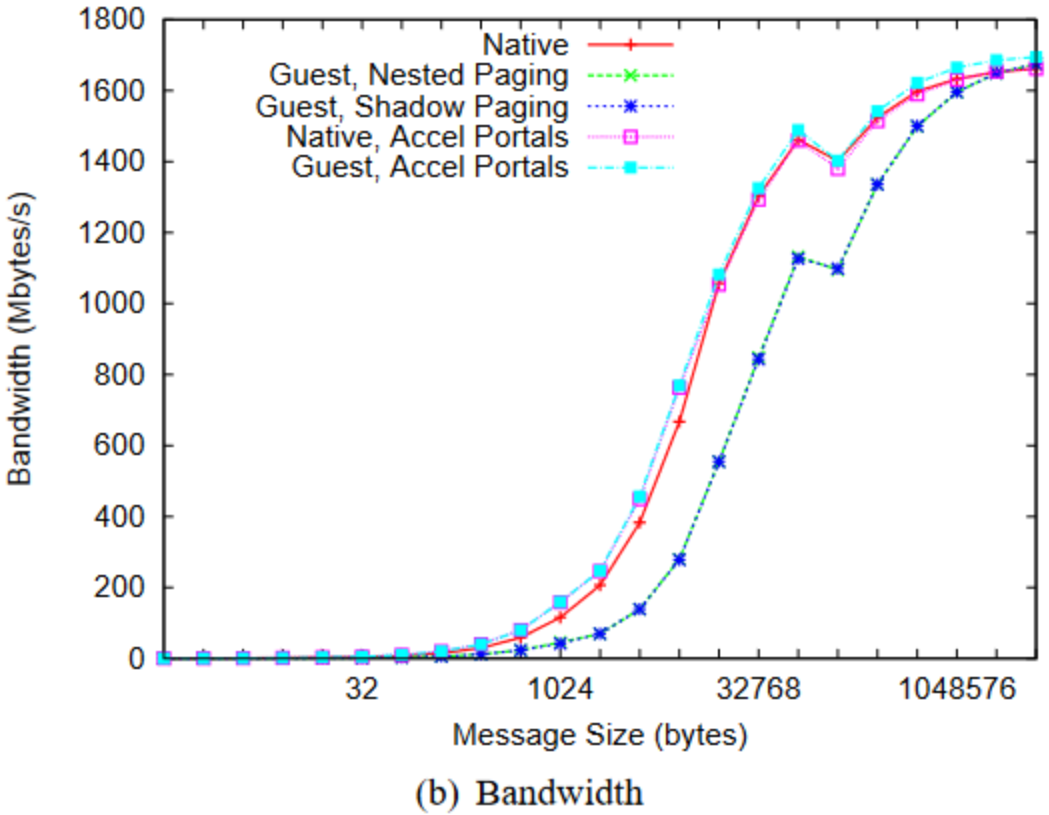
Original names for speeds were single-data rate (SDR), double-data rate (DDR) and quad-data rate (QDR) as given below.[12] Subsequently, other three-letter acronyms were added for even higher data rates.[19]

InfiniBand unidirectional data rates

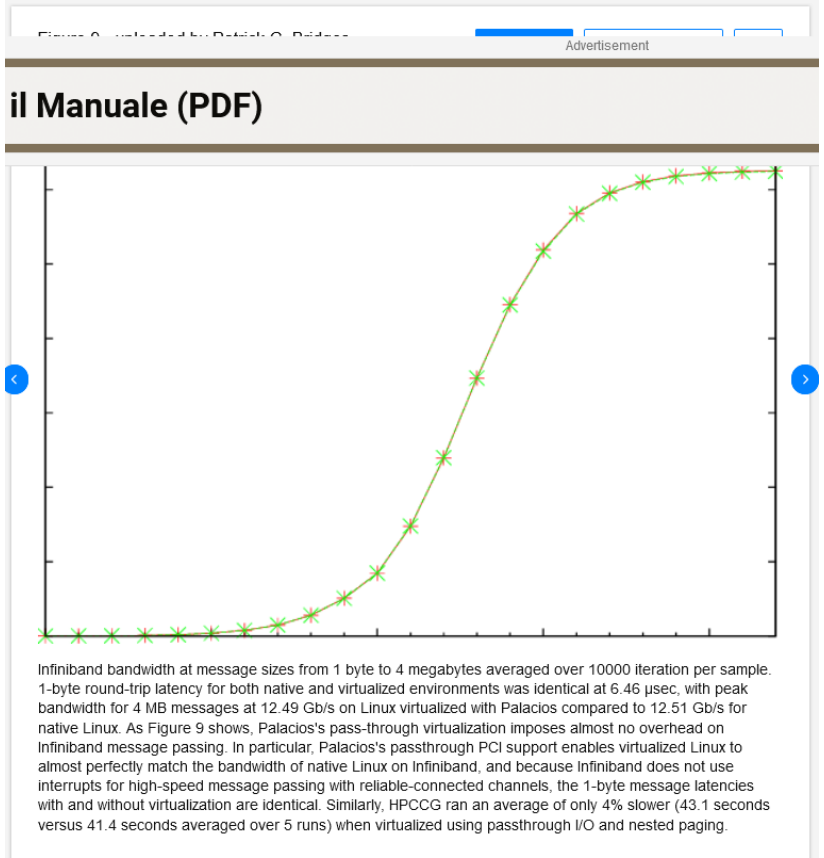
	Year[20]	Line code		Signaling rate (Gbit/s)	Throughput (Gbit/s)[21]				Adapter latency (µs)[22]
					1x	4x	8x	12x	
SDR	2001, 2003	NRZ	8b/10b[23]	2.5	2	**8**	16	24	5
DDR	2005			5	4	**16**	32	48	2.5
QDR	2007			10	8	**32**	64	96	1.3
FDR10	2011		64b/66b	10.3125[24]	10	**40**	80	120	0.7
FDR	2011			14.0625[25][19]	13.64	**54.54**	109.08	163.64	0.7
EDR	2014[26]			25.78125	25	**100**	200	300	0.5
HDR	2018[26]	PAM4		53.125[27]	50	**200**	400	600	<0.6[28]
NDR	2022[26]		256b/257b[i]	106.25[29]	100	**400**	800	1200	?
XDR	2024[30]		[*to be determined*]	200	200	**800**	1600	2400	[*to be determined*]
GDR	TBA	[*to be determined*]		400	400	**1600**	3200	4800	

https://en.wikipedia.org/wiki/InfiniBand

Bandwidtv vs MB in Infiniband: a classical behaviour



(b) Bandwidth

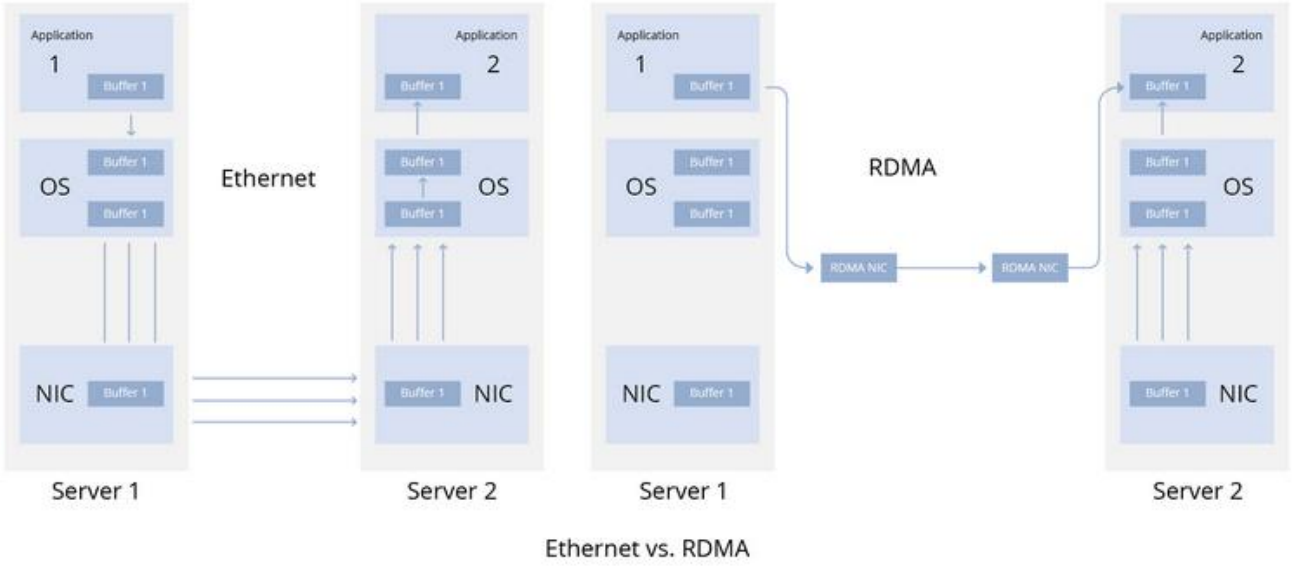


file:///C:/Users/a911569/Downloads/Minimal-overhead_virtualization_of_a_large_scale_s.pdf

https://www.researchgate.net/figure/Infiniband-bandwidth-at-message-sizes-from-1-byte-to-4-megabytes-averaged-over-10000_fig3_221137841

Comunicazione con protocollo Remote Direct Memory Access - RDMA

Nella struttura TCP/IP convenzionale, i dati viaggiano dalla scheda di rete alla memoria principale e poi subiscono un ulteriore trasferimento allo spazio di memorizzazione dell'applicazione. Al contrario, i dati dello spazio applicativo seguono un percorso simile: passano dallo spazio applicativo alla memoria principale prima di essere trasmessi a Internet attraverso la scheda di rete. Questa complessa operazione di I/O richiede una copia intermedia nella memoria principale, allungando il percorso di trasferimento dei dati, imponendo un carico alla CPU e introducendo una latenza di trasmissione.



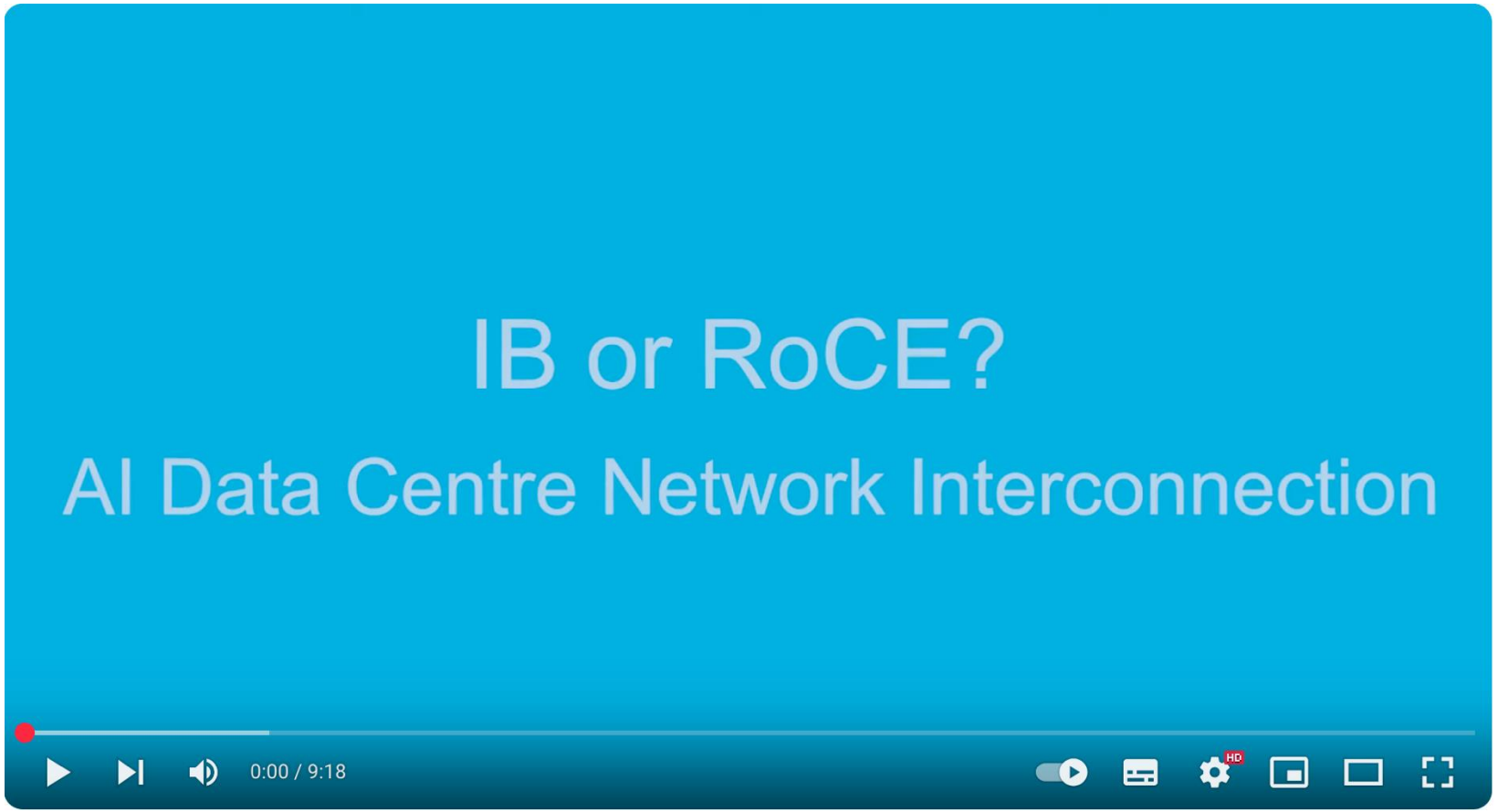
Ethernet vs. RDMA

RoCE – Remote over Converged Ethernet
Abilita RDMA su protocollo Ethernet

https://community.fs.com/it/article/infiniband-insights-powering-highperformance-computing-in-the-digital-age.html

RDMA è una tecnologia che "elimina gli intermediari". Funzionando con un meccanismo di bypass del kernel, RDMA facilita la lettura e la scrittura diretta dei dati tra le applicazioni e la scheda di rete, riducendo la latenza di trasmissione dei dati all'interno dei server a quasi 1 microsecondo.
Inoltre, il meccanismo di zero-copy di RDMA consente all'estremità ricevente di accedere direttamente ai dati dalla memoria del mittente, evitando di coinvolgere la memoria principale. Ciò si traduce in una sostanziale riduzione del carico della CPU, migliorando in modo significativo l'efficienza complessiva della CPU.
Come sottolineato in precedenza, l'adozione diffusa di InfiniBand può essere in gran parte attribuita all'impatto trasformativo di RDMA sull'efficienza del trasferimento dei dati.

Comunicazione con protocollo Remote Direct Memory Access – RDMA vs RoCE



https://www.youtube.com/watch?v=eGoP2wPoaEM

https://youtu.be/eGoP2wPoaEM?si=hkJVIJLF8GDCpRY7

Esempi piu’ comuni di tipologia di rete scalabili

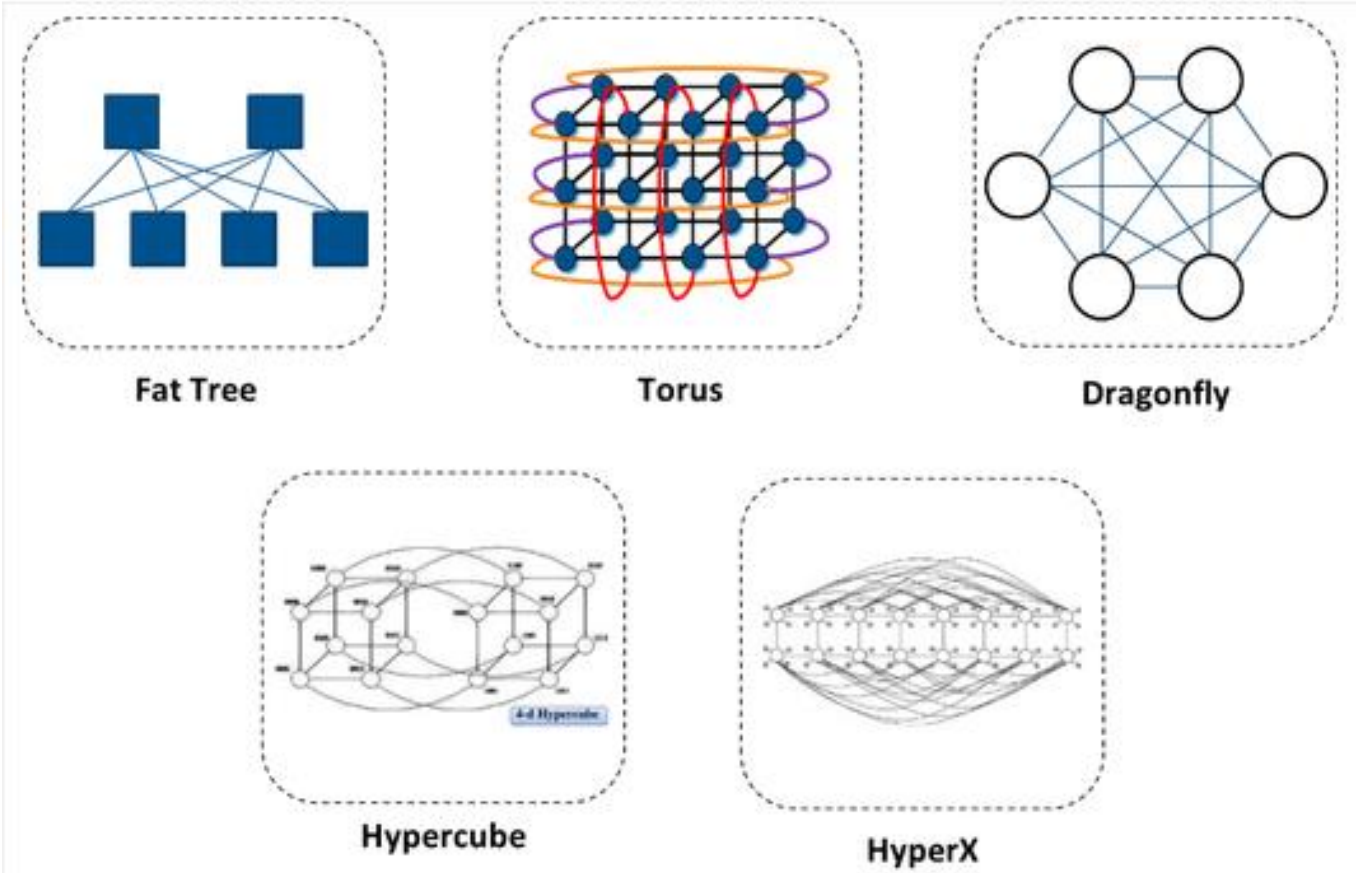


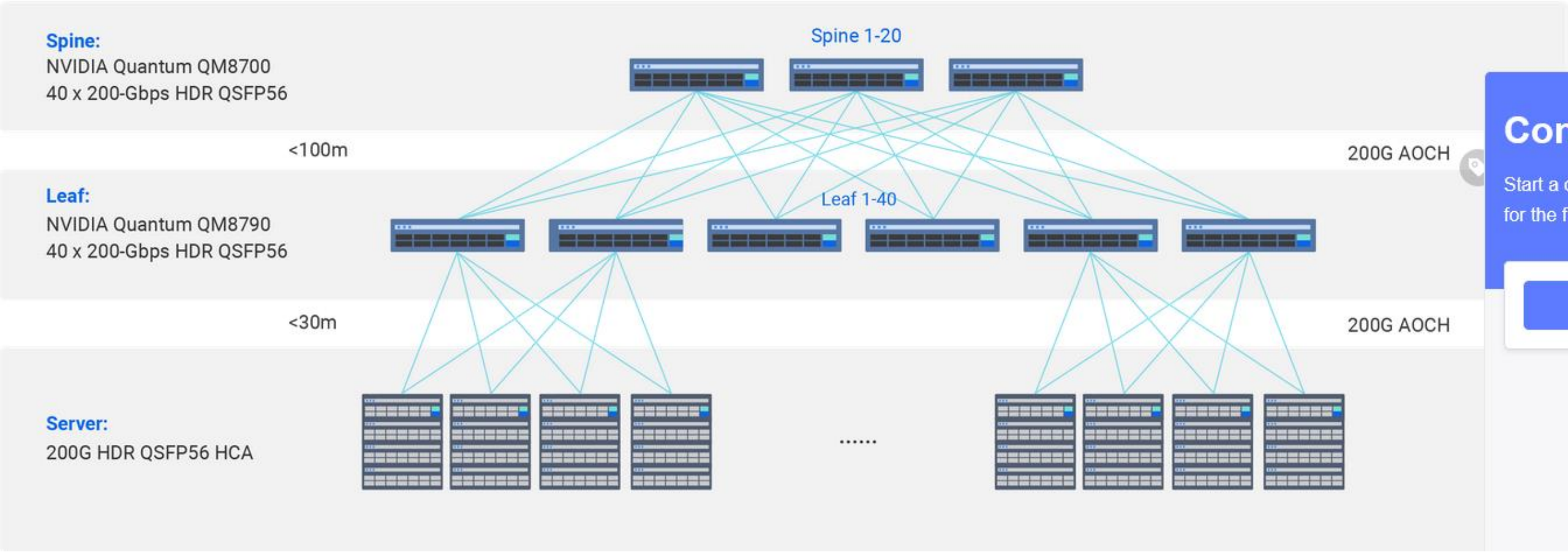
Figure 1 – Network Topologies

https://www.hpcwire.com/2019/07/15/super-connecting-the-supercomputers-innovations-through-network-topologies/

Esempi piu’ comuni di tipologia di rete scalabili

Unmatched Performance with NVIDIA Quantum HDR Switch and InfiniBand

Based on the NVIDIA Quantum HDR switch, ConnectX-6/7 InfiniBand smart network cards, and a flexible 200Gb/s InfiniBand end-to-end solution, this setup offers high throughput bandwidth and ultra-low network transmission latency. It delivers exceptional performance in High-Performance Computing (HPC), AI, and massive-scale cloud infrastructure, while reducing costs and complexity.

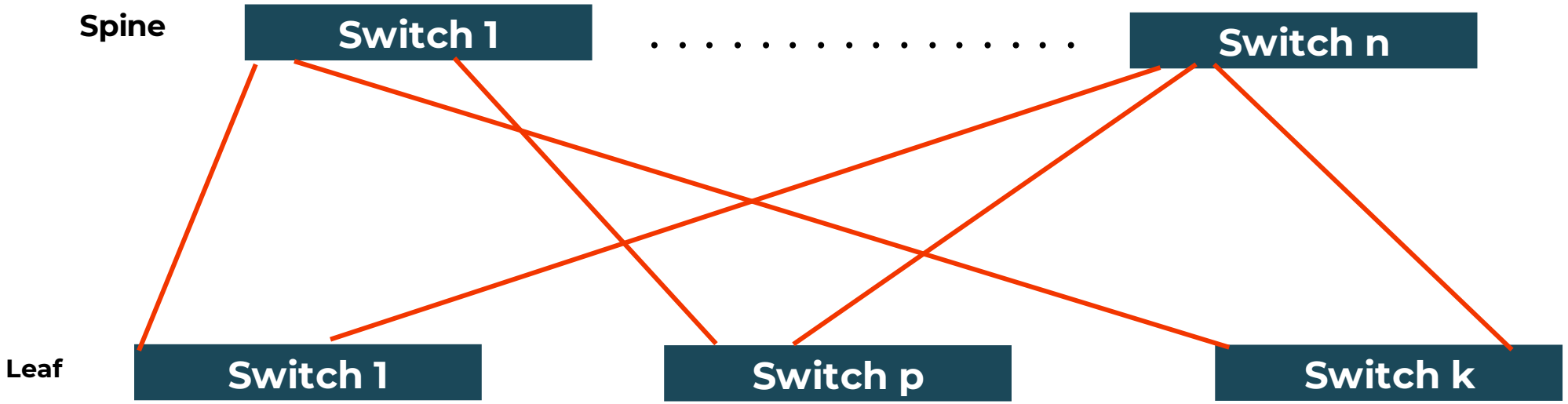


https://www.naddod.com/products/32638.html?gad_source=1&gclid=CjwKCAiArfauBhApEiwAeoB7qIcW8S18rJd0oFyxsH-ADyqY6RJS74p4ZO4b3SpX0drzT3bh3w-dQBoC084QAvD_BwE

Disegno schematico di una rete a vari livelli

Topologia fat-tree a due o piu' livelli

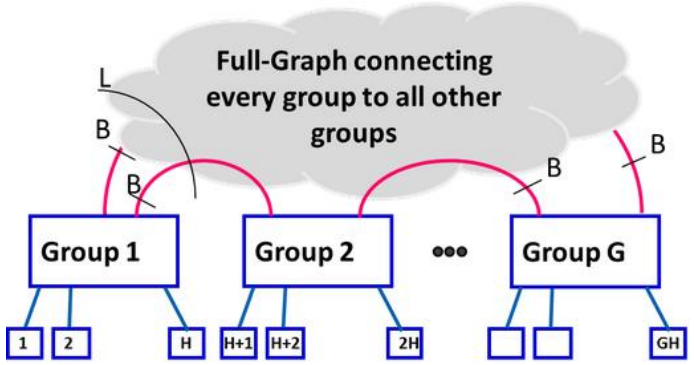
Molto comune in cluster di piccole-medie dimensioni (decine di nodi)
Disegno lineare e facilita' di realizzazione
Alcuni potenziali colli di bottiglia nello scambio di dati per effetti di data contention sugli switch spine
Costi eccessivi per cluster di grandi dimensioni (alcune centinaia di nodi)

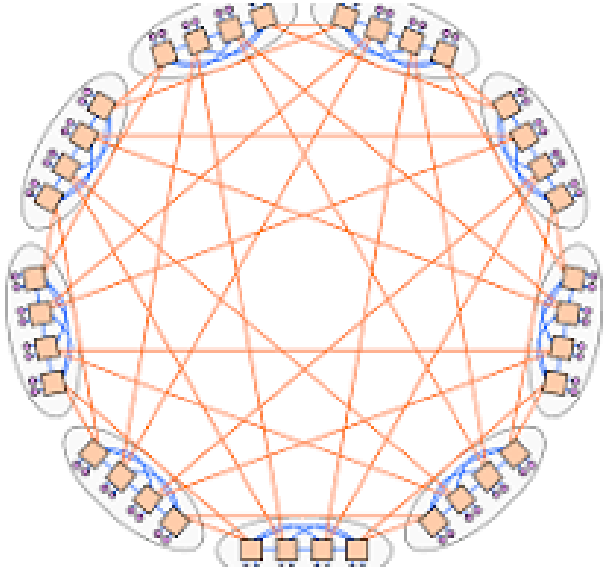


Disegno schematico di una rete a vari livelli

Topologia Dragonfly

Molto comune in cluster di grandi dimensioni (centinaia di nodi)
Disegno meno intuitivo e necessita' di una realizzazione piu' attenta
Elevata scalabilita' e elevate simmetria nel disegno e nel raggruppamento
Ottimizza i costi di connessione nei cluster di grandi dimensioni





https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35155.pdf

https://www.osti.gov/servlets/purl/1510703

Piano del Corso – Lesson 4

Sottosistemi storage a alte prestazioni e loro evoluzione

Concetti generali sulla gerarchia dei sottosistemi storage

Sistemi a disco magnetico e a stato solido

Connessione di sistemi storage su SAN, Infiniband, Ethernet, nVME over Fabric, e altro

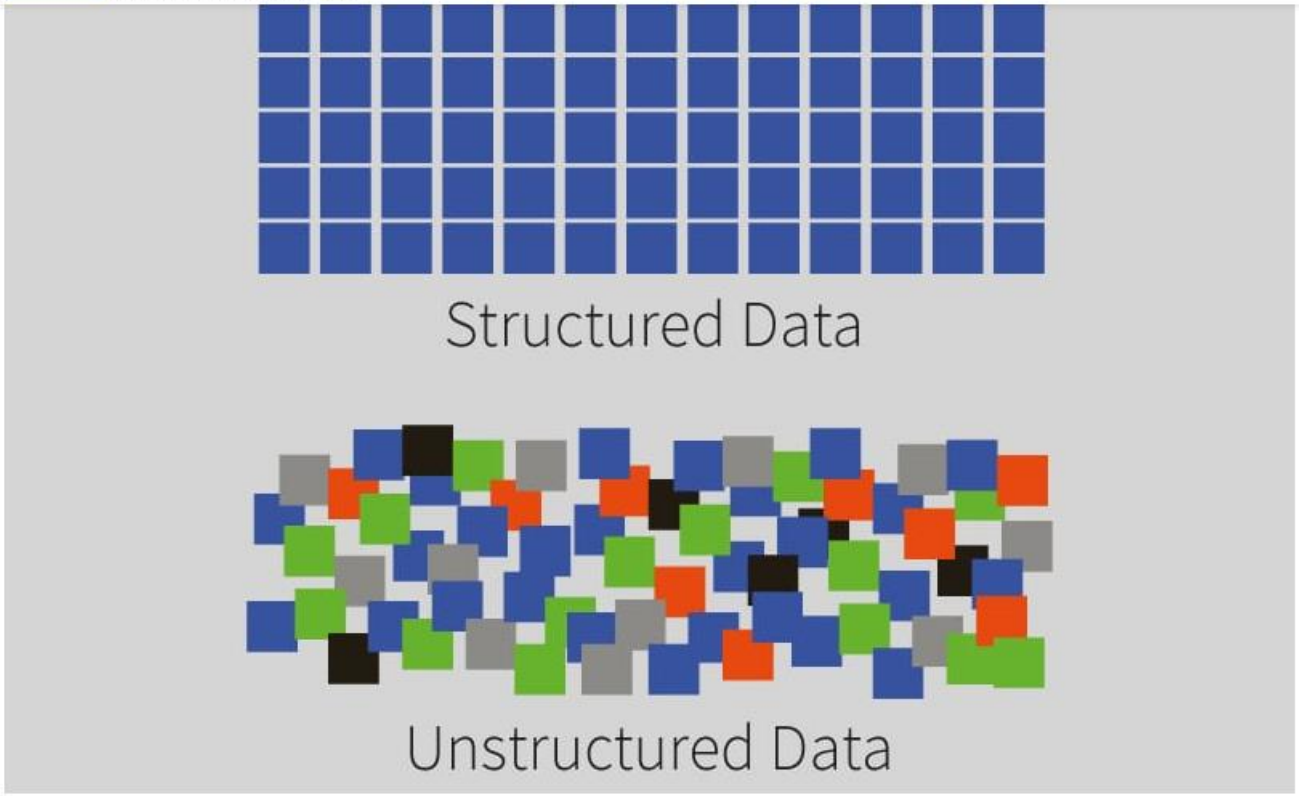
La cescita esponenziale dei dati nel mondo

Data is Growing:

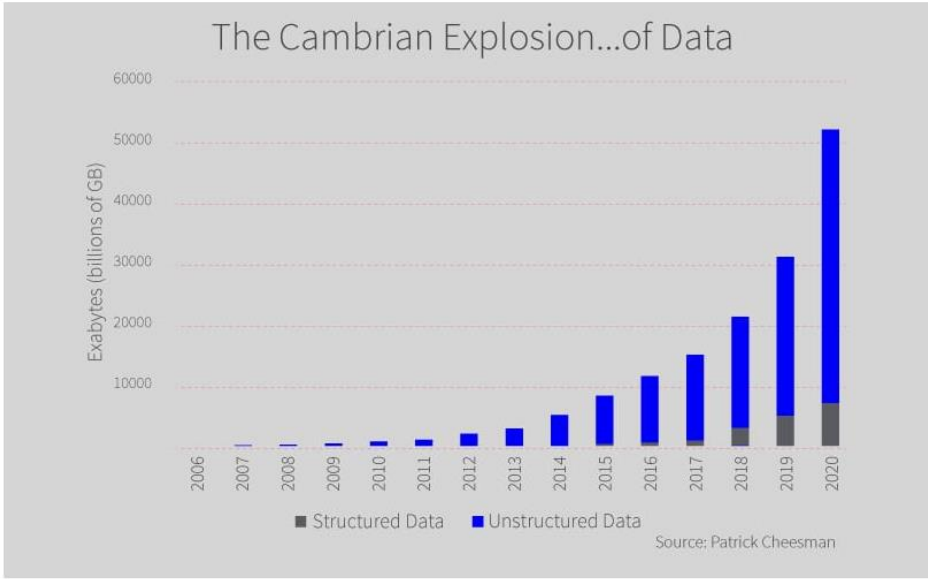
The amount of digital data in the universe is growing at an exponential rate, doubling every two years, and changing how we live in the world. If we look at data we can divide them in: **Structured Data** (highly organized

edge OPTICAL SOLUTIONS

WELCOME PRODUCTS SUPPORT BLOG PAR

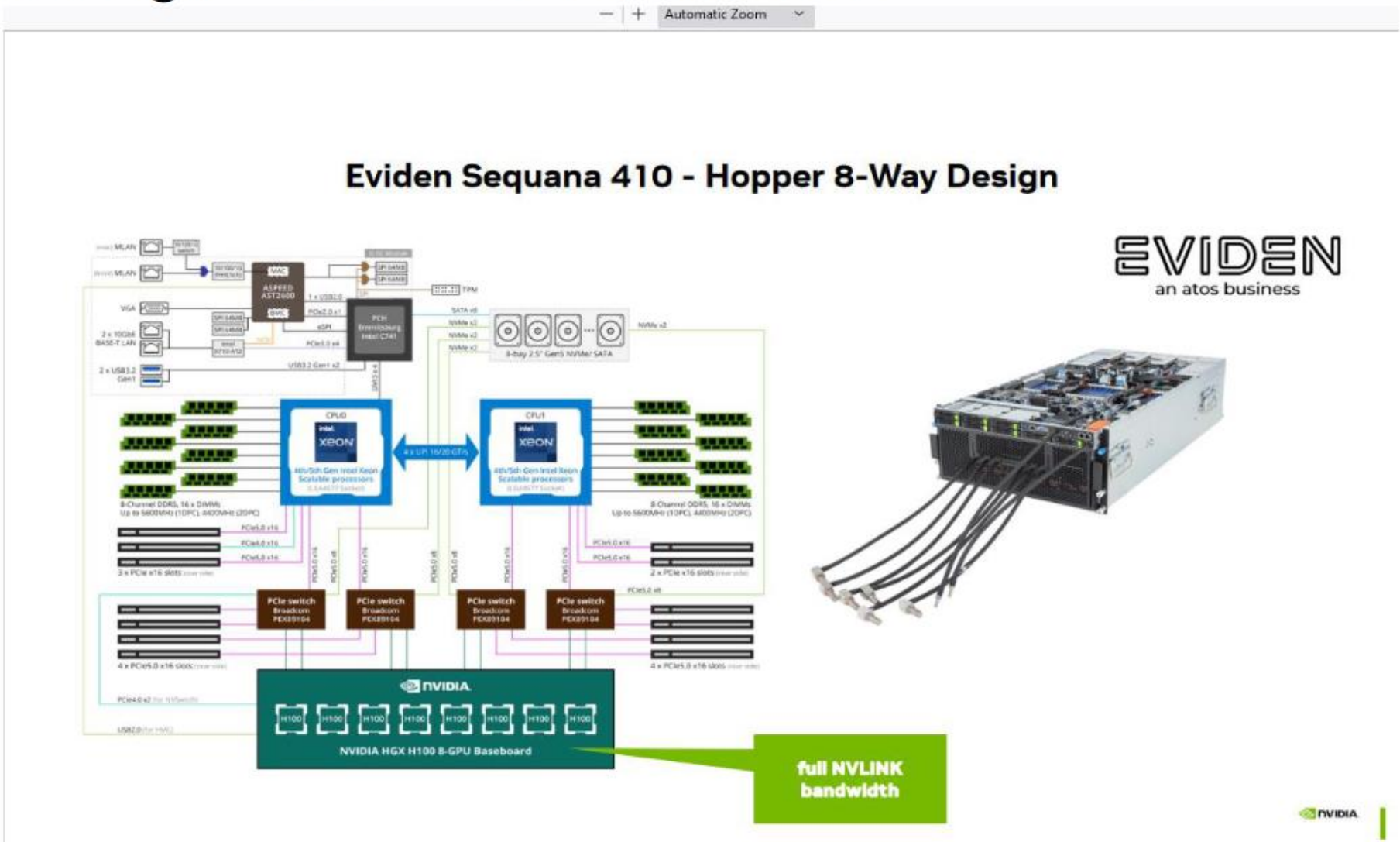


The volume of unstructured data has exploded in the past decade:



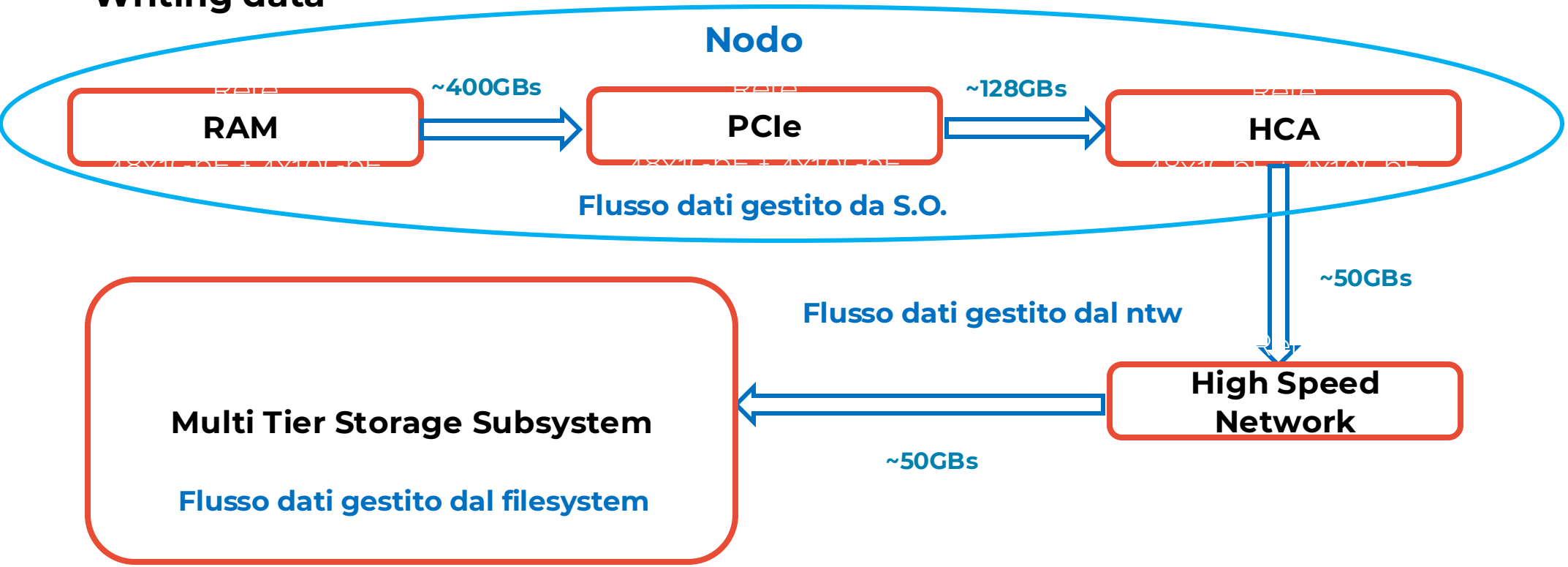
https://edgeoptic.com/storage-protocols-comparison-fibre-channel-fcoe-infiniband-iscsi/

System design – 2xCPUs + 8xGPUS in a node



Il flusso dati: dal nodo al sottosistema storage e viceversa

Writing data



I dati di transfer rate sono teorici, nella realta' i valori misurati sono circa 80% dei valori teorici o anche inferiori

Reading data: flusso inverso

Il flusso dati: banda pasante di trasferimento (transfer rate)

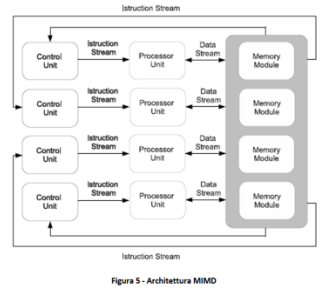


Figura 5 - Architettura MIMD

Stream benchmark – memory BW in a node - CPU to RAM data transfer rate

The performance unit for each operation is measured in MB/s, indicating the amount of data moved between the CPU and per second during that specific operation. The following figures show the system aggregate memory bandwidth for various DIMM sizes, performance per watt, and performance per dollar for STREAM–Triad across PowerEdge R7615 and R7625 servers with 4th Gen AMD EPYC 9654 and 9654P processors:

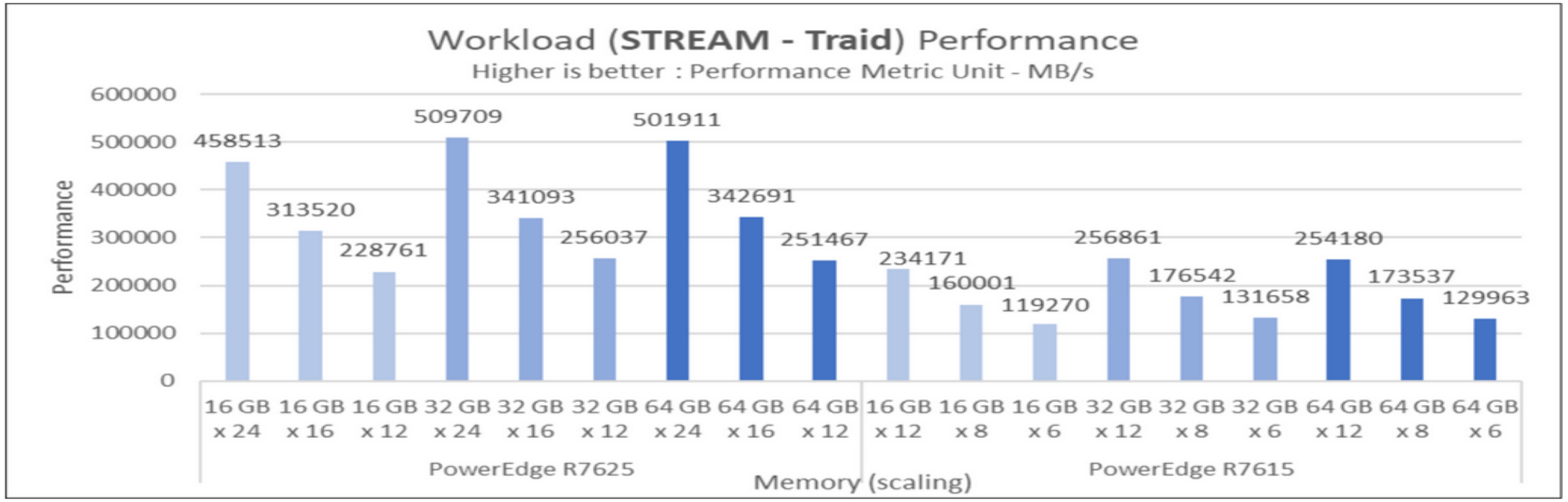


Figure 6. System aggregate memory bandwidth trends with different DIMM configurations and memory capacity for 4th Gen AMD EPYC processor-based PowerEdge R7615 and R7625 servers with default BIOS settings

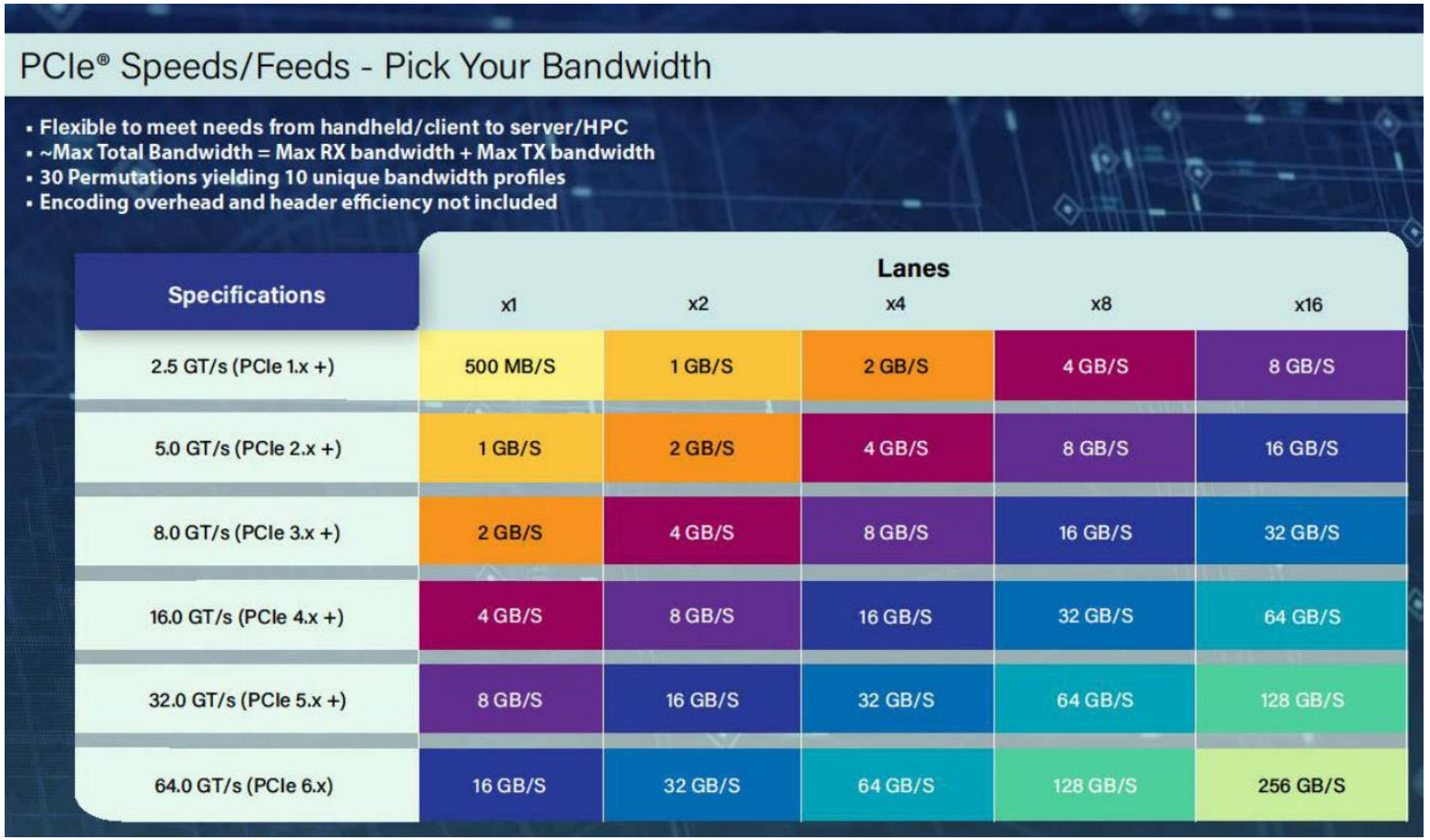
https://infohub.delltechnologies.com/en-US/l/workload-based-ddr5-memory-guidance-for-next-generation-poweredge-servers/memory-bandwidth-stream/

Il flusso dati: banda pasante di trasferimento (transfer rate)

PCIe – data transfer rate

Peripheral Component Interconnect Express

Il PCI-SIG ha completato la messa a punto della specifica PCI Express 6.0, raddoppiando nuovamente la bandwidth rispetto alla versione precedente. Il traguardo è stato raggiunto passando alla tecnologia di signaling PAM4 e adottando Forward Error Correction (FEC) per garantire l'integrità del segnale.



PCIe® Speeds/Feeds - Pick Your Bandwidth

- Flexible to meet needs from handheld/client to server/HPC
- ~Max Total Bandwidth = Max RX bandwidth + Max TX bandwidth
- 30 Permutations yielding 10 unique bandwidth profiles
- Encoding overhead and header efficiency not included

Specifications	Lanes x1	x2	x4	x8	x16
2.5 GT/s (PCIe 1.x +)	500 MB/S	1 GB/S	2 GB/S	4 GB/S	8 GB/S
5.0 GT/s (PCIe 2.x +)	1 GB/S	2 GB/S	4 GB/S	8 GB/S	16 GB/S
8.0 GT/s (PCIe 3.x +)	2 GB/S	4 GB/S	8 GB/S	16 GB/S	32 GB/S
16.0 GT/s (PCIe 4.x +)	4 GB/S	8 GB/S	16 GB/S	32 GB/S	64 GB/S
32.0 GT/s (PCIe 5.x +)	8 GB/S	16 GB/S	32 GB/S	64 GB/S	128 GB/S
64.0 GT/s (PCIe 6.x)	16 GB/S	32 GB/S	64 GB/S	128 GB/S	256 GB/S

https://www.hwupgrade.it/news/skvideo/pci-express-60-ultimato-bandwidth-raddoppiata-rispetto-al-pci-express-50_103873.html

Il flusso dati: banda pasante di trasferimento (transfer rate)

Attribute	Infiniband	Fibre Channel	FCoE	iSCSI
Bandwidth (Gbps)	2.5/5/10/14/25/50	8/16/32/128	10/25/40/100	10/25/40/100
Adapter Latency*	25us	50us	200us	Wide range
Switch latency per	100-200 ns	700 ns	200 ns	200 ns
	channel adapter	bus adapter	converged network adapter	network interface card
Switch Brands	Mellanox, Intel	Cisco, Brocade	Cisco, Brocade	HPE, Cisco, Brocade
Interface Card Brands	Mellanox, Intel	Qlogic, Emulex	Qlogic, Emulex	Intel, Qlogic
Deployment amount	X	XX	X	XXX
Reliability	XX	XXX	XX	XX
Ease of management	X	XX	XX	XXX
Future upgrade path	XX	XXX	XX	XXX

edge OPTICAL SOLUTIONS

WELCOME PRODUCTS SUPPORT BLOG PARTNE

*Please note as we didn't had a chance to get above adapter latency figures and switch latency per hop figures from neutral source they are a bit relative, but can give approximate idea about each protocol performance in this area.

https://edgeoptic.com/storage-protocols-comparison-fibre-channel-fcoe-infiniband-iscsi/

Architetture sottosistemi storage e loro caratteristiche principali

Dischi a stato solido SSD e nVME

- **NVMe concepts**
  NVMe (non-volatile memory express) is a host controller interface and storage protocol created to accelerate the transfer of data between enterprise and client systems and solid-state drives (SSD) over a computer's high-speed Peripheral Component Interconnect Express (PCIe) bus

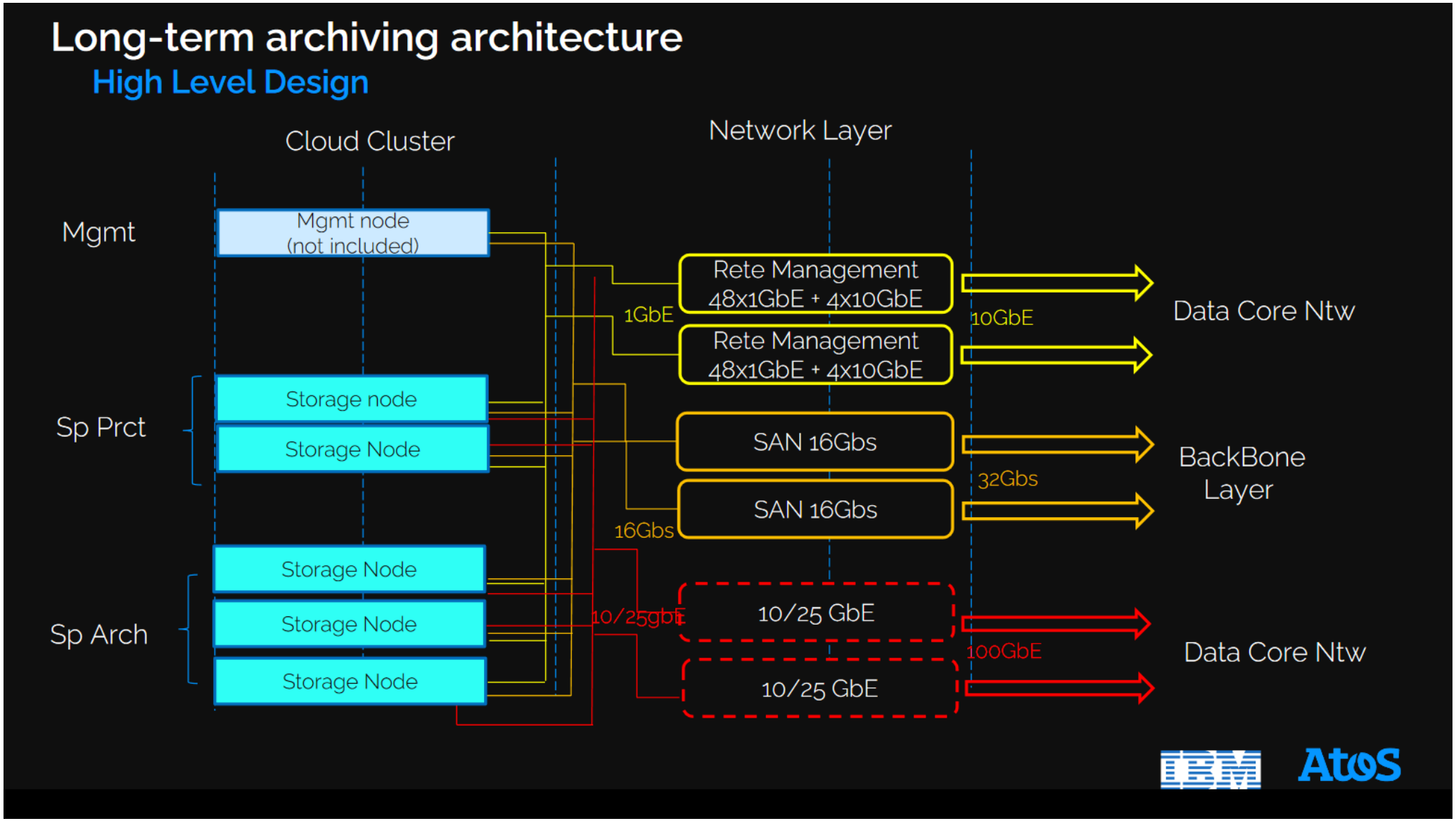
Dischi magnetici: Hard Disk Drive (HDD)

Nastri magnetici: Tape

Architetture sottosistemi storage e loro caratteristiche principali

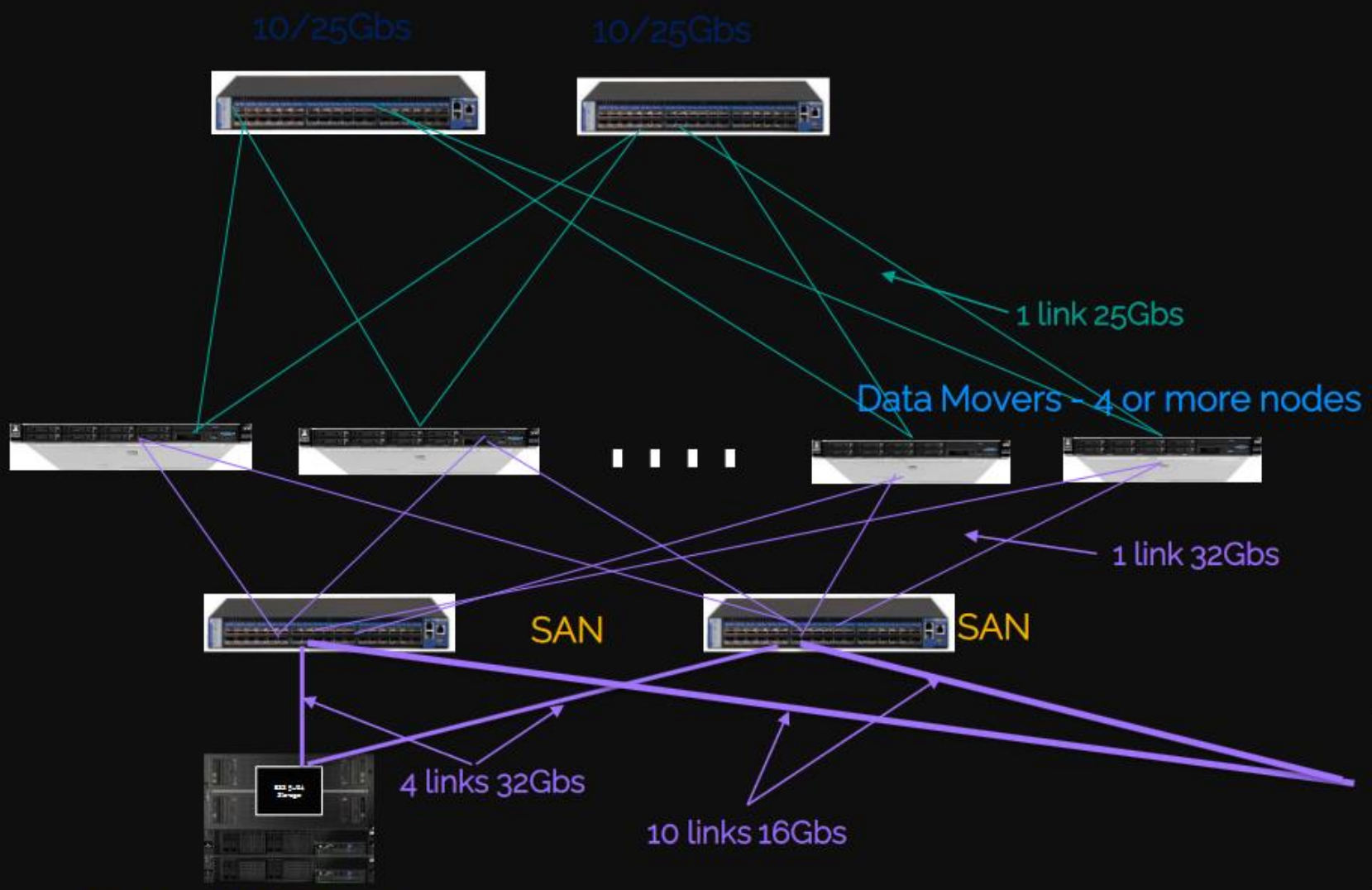
Descr/Tipo	Flash	HDD	Tape	Nota
Capacita' dati per Unita	~10TB	~20TB	~50TB	Le capacita' variano anno/anno
Tipologia di accesso al dato	random	sequenziale	sequenziale	Maggior utilizzo
IOPS e BW	~100000 ~10GBs	~100 ~1GBs	n.a. ~1GBs	Valori molto differenti a seconda del media
Watt per unita'	~10W	~10W	~1W	Valore medio considerando anche lo stato idle
Costo/GB	~0,1USD	~0,03USD	~0,003USD	Costo/GB si riduce anno/anno
Unita' per sottosistema	~decine	~centinaia	~migliaia	Architetture capacitive
Capacita' tipo per sottosistema	~PB	~10PB	~50PB	Valori indicativi

Cosa compone generalmente un'architettura AI&HPC di ultima generazione



https://io500.org/

Long term archiving based on IBM Spectrum Archive



Flash ~150TB net

6xTS1160
4xTS1150

Tape Drive History and Roadmap

LTO Generations	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO10	LTO11	LTO12
New Format Capacity (Native)	1.5 TB (L5)	2.5 TB (L6)	6 TB (L7)	12.0 TB	Up to 24 TB	Up to 48 TB	Up to 96 TB	Up to 192 TB
Other Format Capacities (Native)	800 GB (L4) (400 GB L3 R/O)	1.5 TB (L5) (800 GB L4 R/O)	2.5 TB (L6) (1.5 TB L5 R/O)	9 TB (M8) 6 TB (L7)	Up to 12 TB (L8) (6 TB L7 R/O)	Up to 24 TB (L9) (12 TB L8 R/O)	Up to 48 TB (L10) (24 TB L8 R/O)	Up to 96 TB (L11) (48 TB L10 R/O)
Native Data Rate	140 MB/s	160 MB/s	300 MB/s	Up to 360 MB/s	Up to 708 MB/s	Up to 1100 MB/s		

2008 2010 2011 2013 2014 2015 2017 2017 2018

TS1100 Generations	TS1130	TS1140	TS1150	TS1155	TS1160	TS1170
New Format Capacity (Native)	1 TB (JB) 640 GB (JA)	4 TB (JC) 1.6 TB (JB)	10 TB (JD) 7 TB (JC)	15 TB (JD)	20TB (JE) 15 TB (JD) 7 TB (JC)	Up to 50 TB (JF) Up to 30 TB (JE) 15 TB (JD)
Other Format Capacities (Native)	700 GB (JB) 500 GB (JA) 300 GB (JA)	1 TB (JB) 700 GB (JB) (All JA R/O)	4 TB (JC)	7 TB (JC) 4 TB read only (JC)	10 TB (JD) 7 TB (JC) 4 TB (JC)	10 TB (JD)
Native Data Rate	160 MB/s	250 MB/s	360 MB/s	360 MB/s	400 MB/s	Up to 1000 MB/s

Summary and comments – 3rd & 4th Lessons

- Network in a AI&HPC architecture
- Ethernet and Infiniband
- Network topologies
- Evaluation criteria: BW and latency
- Infiniband road-map
- TCP-IP, RDMA and RoCE
- Introduction of storage architectures
- An example on how to design a small cluster

Thank You

Marco Briscolini, PhD

marco.briscolini@gmail.com
Cell: 3357693820

03/19/2025