

Future Computing Architecture

Lessons 7th & 8th

Marco Briscolini, PhD

marco.briscolini@gmail.com

Cell: 3357693820

04/16/2025

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 capacità computazionale, trasmissione dati, lettura/scrittura dati)
Metriche riconosciute a livello mondiale (Top500, Green500, IO500)
Concetti introduttivi sull'analisi della complessità 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 sottosistemi 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 parità 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 8th

Accenni al disegno e alla progettazione di un'architettura HPC

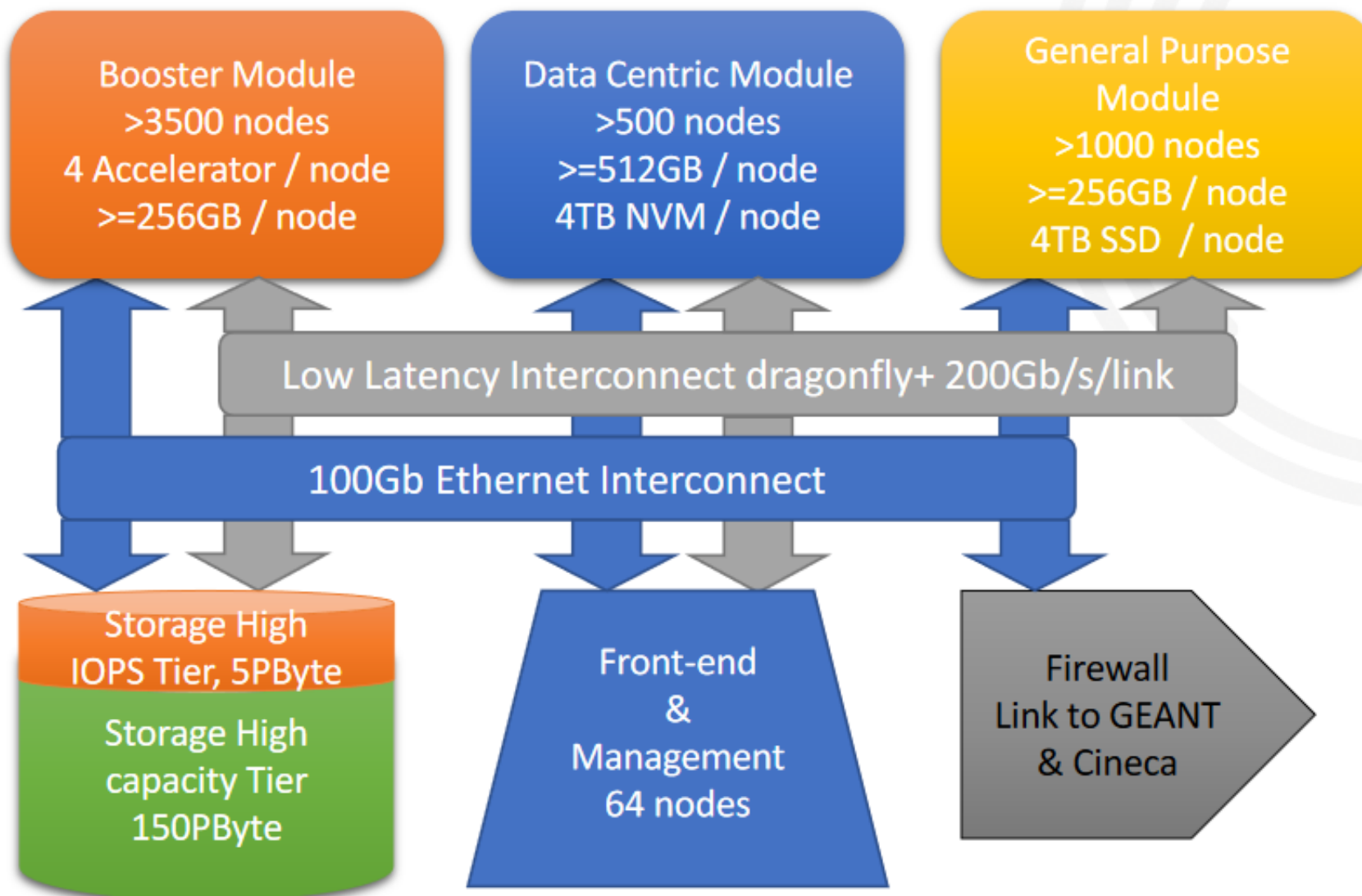
Definizione di specifiche di progetto

Valutazione preliminare dell'architettura ottimale

Disegno di massima dell'architettura

Concetto di rispondenza e verifica alle specifiche di progetto

Leonardo High Level description



Specifiche di progetto e requisiti base

Progetto basato su criteri prestazionali

Prestazione aggregata delle partizioni di calcolo: CPU (General Partition) e GPU (Booster partition)

Prestazione del sottosistema storage

Prestazione della rete di interconnessione

Criteri di efficienza energetica: TCO (Total Cost of Ownership) optimization

General Partition – specifiche minimali

Oggetto	Dettagli	Specifica unitaria	Aggregato	Spec Aggr	Nota
Nodo CPU	#cores >=36	CPU peak perf >=2TFs	2xCPU	HPL >= 2TFs	
	MEM		Max BW	>=128GB	Distribuzione DIMM bilanciata
	High Perf Ntw		Almeno un link	>=200Gbs	Si possono aggregare piu' links
	Mgmt Ntw		Un link	>=10Gbs	Gestione del Sistema
	Internal HDD	SSD o NVMe	Uno o piu' HDDs	>=1TB	RAID come opzione
FormFactor	Alta Densita'		1/2 wide		4N2U preferito

Booster Partition – specifiche minimali

Oggetto	Dettagli	Specifica unitaria	Aggregato	Spec Aggr	Nota
Nodo GPU		CPU peak perf >=2TFs	2xCPU		
CPU	MEM		Max BW	>=128GB	Distribuzione DIMM bilanciata
GPUs			>=4GPUs	Interconnesse	
	High Perf Ntw		Un link/GPU	>=400Gbs	Si possono aggregare piu' links
	Mgmt Ntw		Un link	>=10Gbs	Gestione del Sistema
	Internal HDD	SSD o NVMe	Uno o piu' HDDs	>=1TB	RAID come opzione
FormFactor	Alta Densita'		1U		1N1U

Storage – specifiche minimali

Oggetto	Dettagli	Specifica unitaria	Aggregato	Perf Aggr	Nota
Block Storage 1 st tier	NVMe	1PB net	>=5PB net	>=100GBs	Scale-out
Block Storage 2nd tier	HDD	5PB	>=50PB net	>=50GBs	Scale-out
Block Storage 3rd tier	Tape	20PB	>=20PB	10Gbs	Tape library
High Perf Ntw		>=100Gbs	>=4xlinks	>=400Gbs	Per sottosistema (eccetto Tape)
Mgmt Ntw		10GBs		10Gbs	Per sottosistema

Network – specifiche minimali

Oggetto	Dettagli	Specifica	Aggregato	Spec Aggr	Nota
High Perf Ntw	Infiniband	$\geq 200\text{Gbs}$ per nodo	$\leq 2:1$ oversub	Dipende dal numero di links	Fat-tree topology
Mgmt Ntw	Ethernet	10Gbs per link	$\leq 4:1$ oversub	Dipende dal numero di links	Flat
Fat-tree Infiniband	2:1 oversub	1 spine - 2 leaf a albero Splitter cable (800Gbs \rightarrow 4x200Gbs)	$96 \times 200\text{Gbs} / 2 = 9,6\text{Tbs}$		2:1 oversub e' ragionevole data la dimensione del Sistema. Numero di connessioni 200Gbs complessive: $12 \times 4 \times 2 = 96$ links verso nodi e storage $2 \times 10 = 20$ links da leaf a spine

Disegno e consumi – specifiche minimali

Oggetto	Dettagli	Specifica rack	Aggregato	Spec Aggr	Nota
Consumi IT		$\leq 70\text{kW IT}$	$\leq 2\text{mW IT}$		
Raffreddamento (1)	AIR	100%			
Raffreddamento (2)	DLC + AIR	Air $\leq 30\%$			
Raffreddamento (3)	DLC full	Air $\leq 5\%$			
PUE		1,4 Air 1,1DLC			Ipotesi di massima
Costo energia	0,3€/kWh				
TCO (1)					
TCO (2)					
TCO (3)					

Componenti aggiuntive

Specifiche di Progetto non sempre legate esclusivamente alle prestazioni

Nodi di front-end per attività interattive (job scheduling, compilazione, visualizzazione)

Nodi di mgmt riservati al superuser per la gestione del Sistema
(spegnimento/accensione, diagnosi errori HW e SW, verifica consumi,
manutenzione prewdittiva e programata, etc.)

Rete di front-end per la connessione alla LAN locale

Ambienti SW per la gestione e l'accesso alle risorse

General Partition - Risultato ottimale utilizzando specifiche di minimo

Oggetto	Dettagli	Single node peak perf	HPL single nodo	Spec Aggr	Nota
General Partition CPU only	2xCPU 36c@2GHz per CPU	Peak perf 2cpu*32c*32fp*2GHz ~ 4TFs	~80% peak perf ~3TFs with 16xDIMMs (full mem bw) ~60% peak perf ~2TFs with 8xDIMMs (half mem bw)	One rack 72 nodes ~200TFs or ~140TFs supposing linear scalability	Considerando che le DIM hanno una dimensione di 32GB min, l'ipotesi di popolare solo 4DIMMs per CPU e' accettabile
Booster Partition CPUs+GPUs	2xCPU + 4xGPUs (H100)	Dato non rappresentativo per le GPUs	~4x33TFs ~120TFs	One rack 16 nodes 2PFs supposing linear scalability	Nodi con GPUs hanno una prestazione HPL significativamente superiore
HPF Ntw Inifiniband NDR 800Gbs	1xlink 200Gbs per nodo CPU	4xlinks 200Gbs per nodo GPUs			Banda aggregate fat-tree 1:1 oversub (72+16*4)*200 = 14Tbs ~7Tbs in 2:1 oversub
Mgmt Ntw Ethernet 10Gbs	1xlink 10Gbs	1xlink 10Gbs			4:1 oversub ~220Gbs


Booster Partition - Risultato ottimale utilizzando specifiche di minimo

Oggetto	Dettagli	Single block peak perf	#HDDs per block	Spec Aggr	Nota
Storage HDD	block storage con filesystem parallel	50GBs per block storage	1xHDD SAS ~ 300MBs >= 200HDDs for max BW 200x 20TB/HDD ~4PB grezzi ~3PB netti (protezioni RAID determinano circa 20% di spazio non utilizzabile)	2xblocks → ~100GBs ~8PB grezzi ~6PB netti	In genere il filesystem determina un carico che reduce le prestazioni di circa 20%, quindi sono necessary piu' dischi del valore nominale per raggiungere la prestazione
Storage Flash	block storage con filesystem parallel architettura scale-out granulare	50GBs per block	1xNVMe ~ 3GBs 24xNVMe ~> 50GBs 24x10TB/NVMe ~240TB grezzi ~200TB netti	10xblocks → ~5TBs ~2,4PB grezzi ~2PB netti	Il sottosistema flash ha prestazioni molto maggiori ma una minore capacita'. Si usa per forinre spazio volatile (scratch) a alte prestazioni a costi accettabili

Storage Partition - Risultato ottimale utilizzando specifiche di minimo

Oggetto	Dettagli	Single block peak perf	#HDDs per block	Spec Aggr	Nota
Storage HDD	block storage con filesystem parallel	50GBs per block storage	1xHDD SAS ~ 300MBs >= 200HDDs for max BW 200x 20TB/HDD ~4PB grezzi ~3PB netti (protezioni RAID determinano circa 20% di spazio non utilizzabile)	2xblocks → ~100GBs ~8PB grezzi ~6PB netti	In genere il filesystem determina un carico che reduce le prestazioni di circa 20%, quindi sono necessary piu' dischi del valore nominale per raggiungere la prestazione
Storage Flash	block storage con filesystem parallel architettura scale-out granulare	50GBs per block	1xNVMe ~ 3GBs 24xNVMe ~> 50GBs 24x10TB/NVMe ~240TB grezzi ~200TB netti	10xblocks → ~5TBs ~2,4PB grezzi ~2PB netti	Il sottosistema flash ha prestazioni molto maggiori ma una minore capacita'. Si usa per forinre spazio volatile (scratch) a alte prestazioni a costi accettabili

Diseano e consumi – specifiche minimali

Oggetto	Dettagli	Specifica nodo	Consumo rack	Nota
Consumi IT – CPU	CPU	~ 500W IT	72*500W = 36kW IT	Nei limiti del raffreddamento a aria
Consumi IT - GPU	GPU	~3kW IT	16x3kW = 48kW IT	Nei limiti del raffreddamento aria-acqua
Consumi complessivi – (1)	AIR	PUE >= 1,4	Consumo ~ (36+48)*1,4 ~120kW	Ipotesi solo aria
Consumi complessivi – (2)	AIR 30%+DLC 70%	PUE Air >= 1,4 PUE DLC <= 1,1	84*(0,3*1,4+0,7*1,1) ~ 1,2*84 ~ 100kW	Ipotesi aria acqua
Consumi complessivi (3)	DLC	PUE DLC <= 1,1	84*1,1 ~ 92kW	Ipotesi solo acqua
Costo energia	0,3€/kWh	kWh = 1h*kW		Media nazionale
TCO (1) Total Cost of Ownership		120*1h = 120kWh	120*24*365*0,3 ~ 320K€/anno	OPEX (Operating Expenses) di riferimento
TCO (2)		100*1h = 100kWh	100*24*365*0,3 ~ 260K€/anno	Risparmio ~60K€/anno 
TCO (3)		92*1h = 92kWh	92*24*365*0,3 ~ 240K€/anno	Risparmio ~80K€/anno

Nella gran parte dei casi una soluzione mista aria-acqua e' quella che da il maggior beneficio economico tenendo in conto sia il risparmio energetico che il maggior investimento infrastrutturale rispetto al raffreddamento solo aria. Quest'ultimo ormai semre piu' spesso non e' piu' utilizzabile dati i consumi

Conclusioni

Specifiche di Progetto non sempre legate esclusivamente alle prestazioni

Criterio guida consiste nel determinare la soluzione ottimale date le specifiche di Progetto

- costo/prest per le componenti HW e SW

- scelta anche con soluzioni di prossima generazione se nei tempi di realizzazione

- flessibilita' e affidabilita' dell'architettura

- Efficienza energetica

- Spazi fisici necessary

- Tempi di realizzo

Proporre soluzioni innovative anche se non esplicitamente richieste se possibili in termini di costo e tempi di realizzazione

Summary and comments – Lessons 7th and 8th

- Technology trends**
- Scalable AI architecture**
- AI-QC-HPC as an integrated architecture for complex simulations**

- Example of list of specifications for a AI&HPC procurement**

EVIDEN

Thank You

