Future Computing Architecture

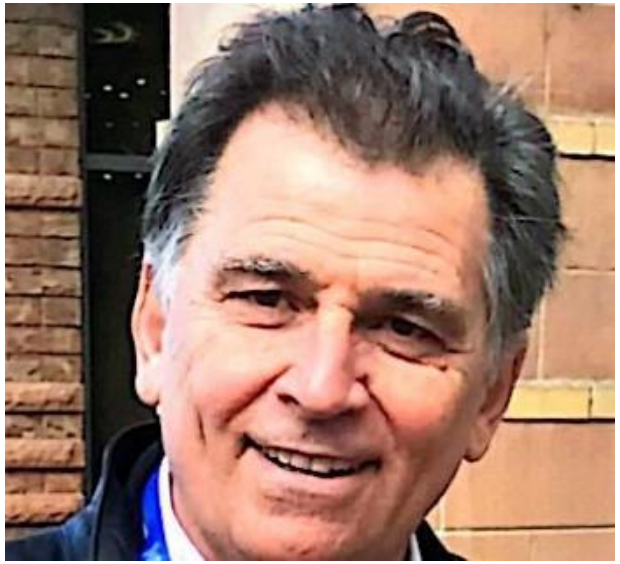
1st & 2nd lessons

Marco Briscolini, PhD

e-mail: marco.briscolini@gmail.com
mob: 3357693820

03/12/2025

Marco Briscolini, PhD



- Contact information:

M+ 39 335 7693820
marco.briscolini@gmail.com

Previous experience

E4 COMPUTER ENGINEERING — **2024 - AI&HPC Bus Dev Manager**

EVIDEN — **2022-24 - AI&HPC Italy Sales Leader**

IBM — **2019-21 – AI&HPC Italy Sales Leader**

Lenovo — **2015-18 - HPC Architect and Sales**

IBM — **1987-2014**

Architect and Sales
Research engineer in comp. sci. and HPC
Software dev for visualization
HPC and simulation software dev
Tutoring & teaching activities

Education

Aerospace Engineering Doctorate, La Sapienza

“Laurea in Fisica”, La Sapienza

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 1

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

Siti di interesse per temi legati a AI&HPC&QC

http://192.168.1.1/html/connectionless.html

https://www.hpcwire.com/

https://www.scientific-computing.com/

https://top500.org/

https://research.ibm.com/

https://www.lenovo.com/it/it/servers-storage/solutions/hpc/

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

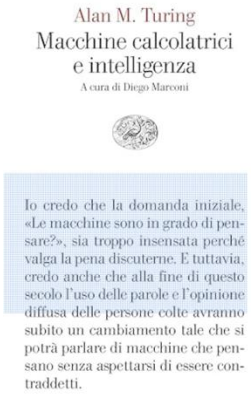
https://eviden.com/solutions/advanced-computing/

https://www.hpe.com/us/en/compute/hpc/supercomputing/cray-exascale-supercomputer.html

https://www.dell.com/it-it/dt/apex/compute-hci/high-performance-computing.htm#

Alan Turing and John von Neumann – I padri delle moderne architetture di elaborazione

Alan Turing – 1912-1954



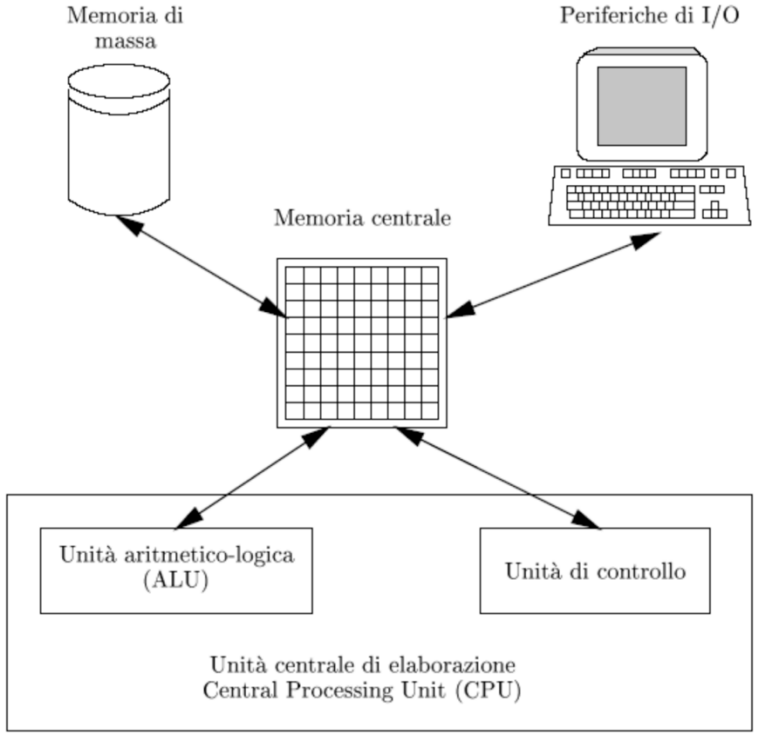
John von Neumann – 1903-1957

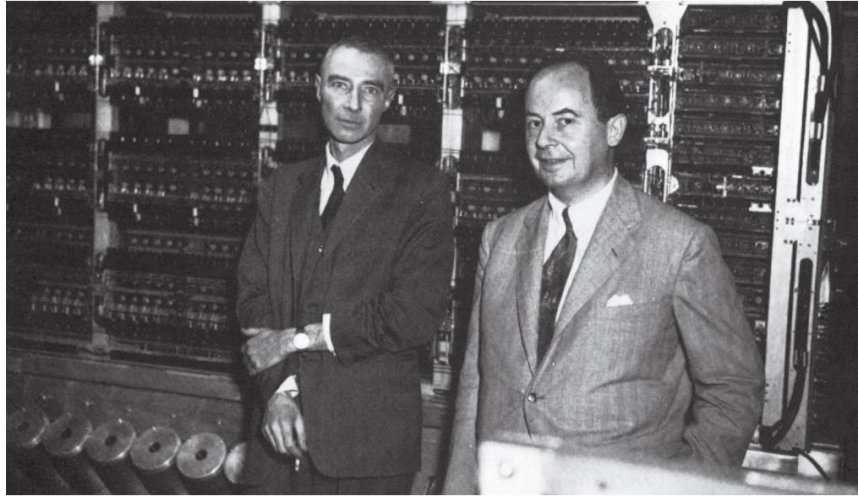
Alan Turing – il modello astratto di un calcolatore moderno

In informatica, una **macchina di Turing** (o più brevemente **MdT**, in inglese **TM**, da *Turing machine*) è un modello matematico computazionale che descrive una macchina astratta[1][2] che manipola (legge e scrive) i dati contenuti su un nastro di lunghezza potenzialmente infinita, secondo un insieme prefissato di regole ben definite. A dispetto della sua apparente semplicità, questo modello è in grado di simulare la logica di qualunque algoritmo eseguibile su un computer reale[3].

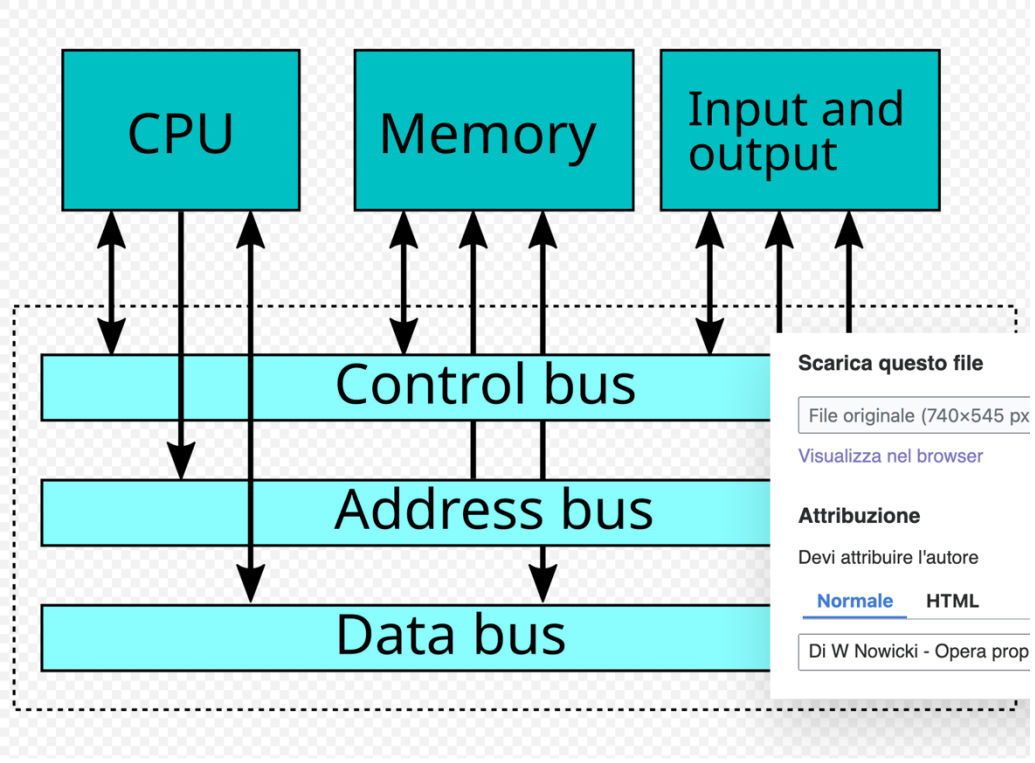
Il teorema di Turing **asserisce l'esistenza di problemi non decidibili, per i quali cioè non esiste alcun algoritmo in grado di dare una risposta in tempo finito su tutte le istanze del problema**. La dimostrazione di questo risultato si deve ad Alan Turing, che lo provò in un articolo del 1937.

Architettura di von Neumann e sua evoluzione



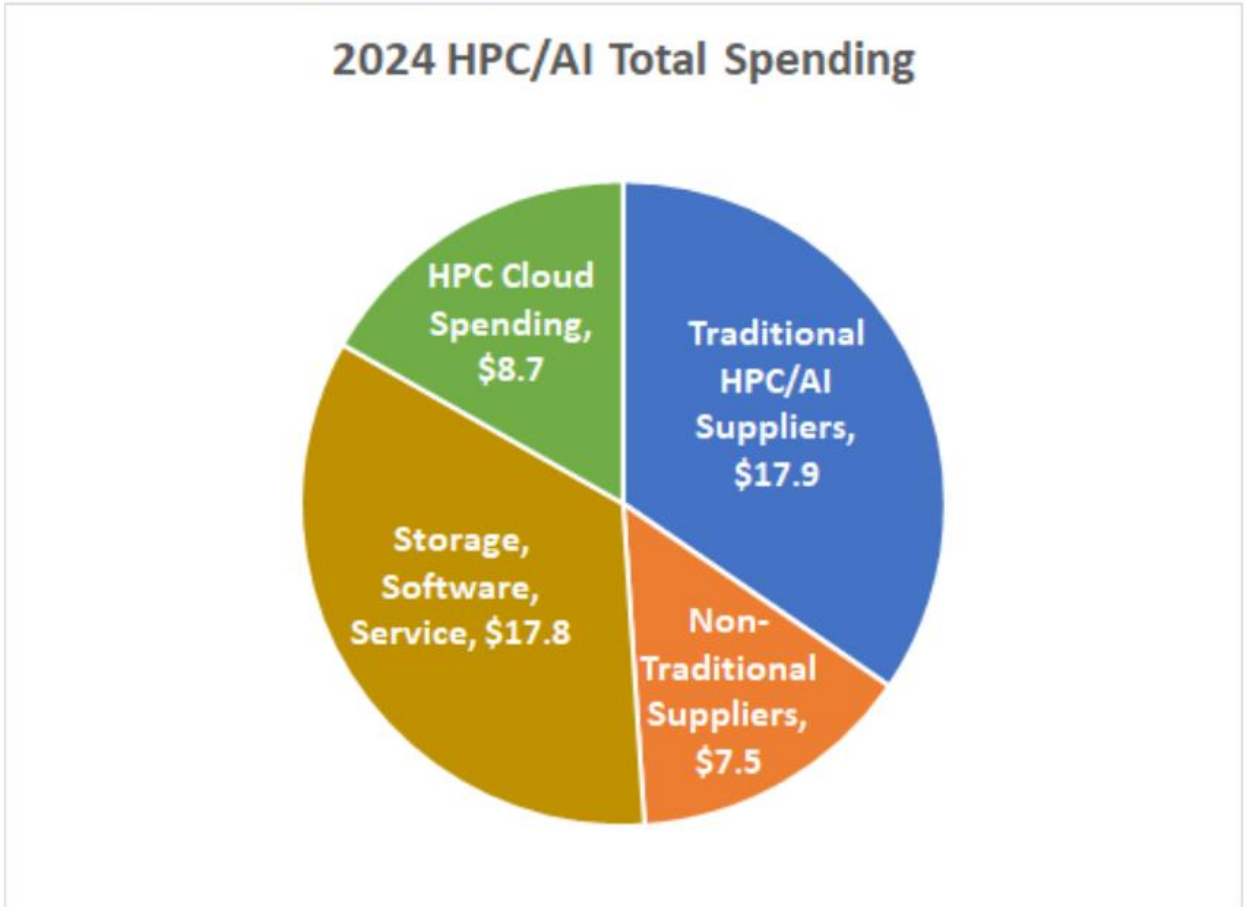


Il calcolatore EDVAC – 1949
Electronic Discrete Variable Automate Computer



The Overall HPC/AI Market in 2024

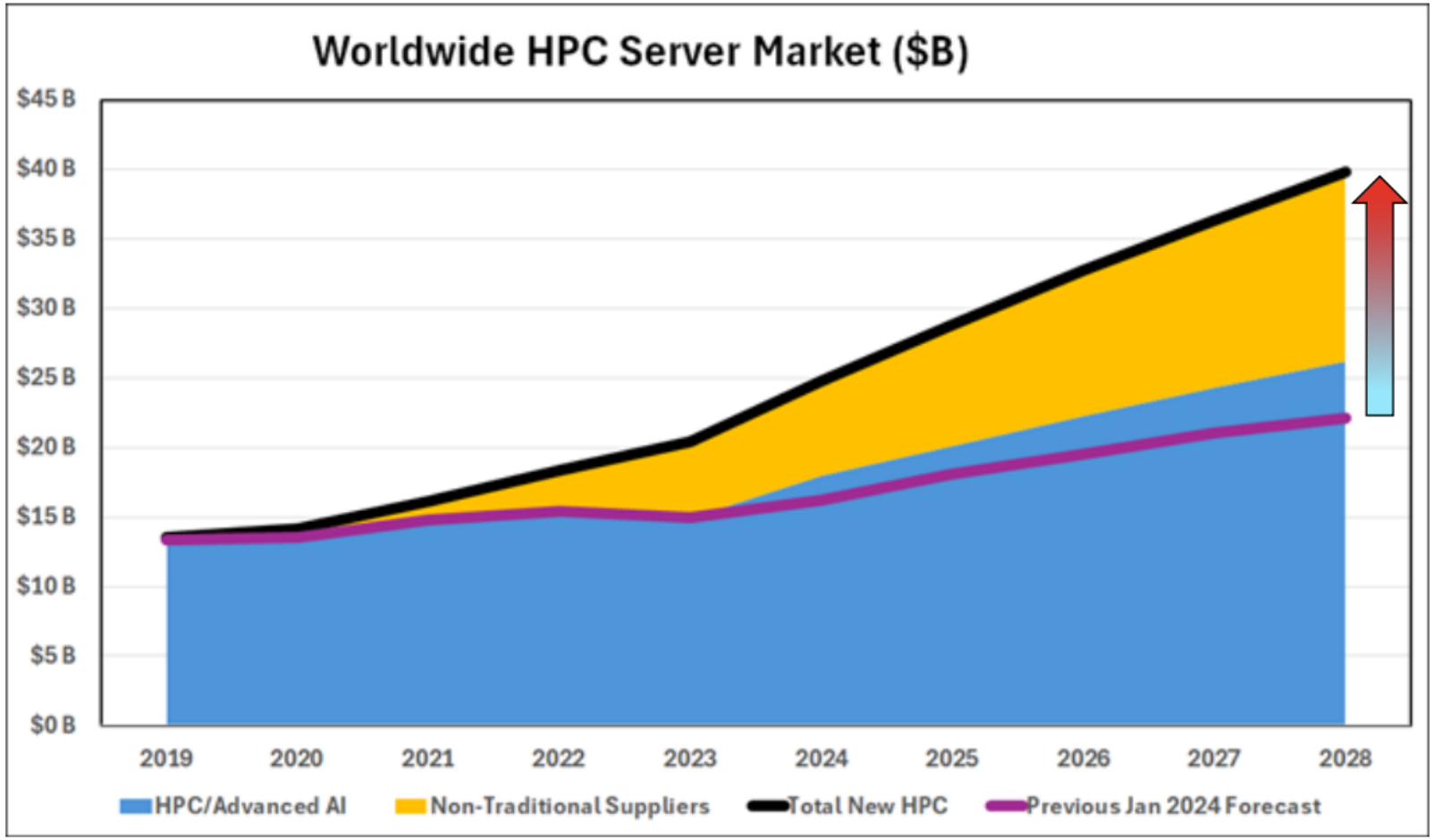
2024 HPC/AI Spending is projected to reach $51.9 billion ($US)



- **$25.4 billion in on-premises servers**
- **$8.7 billion in spending to run HPC/AI workloads in the cloud**

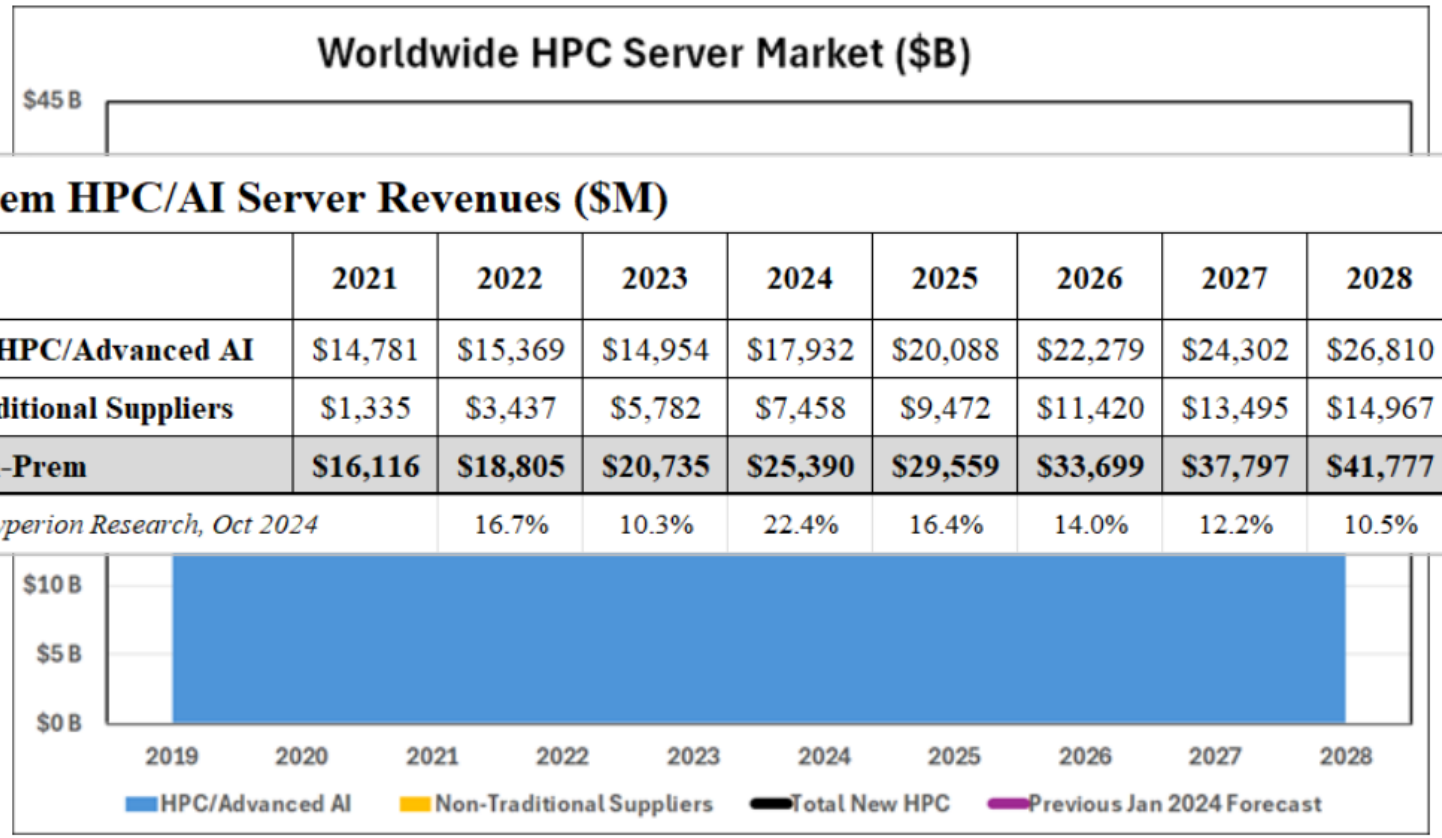
Updated View of the On-Prem Sever Market

- *Hyperion Research just announced a 36.7% increase in the HPC/AI server market size (growing at 15% CAGR)*
- *Now tracking non-traditional AI/HPC suppliers*



Updated View of the On-Prem Sever Market

- *Hyperion Research just announced a 36.7% increase in the HPC/AI server market size (growing at 15% CAGR)*
- *Now tracking non-traditional AI/HPC suppliers*



On-Prem HPC/AI Server Revenues ($M)

	2021	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
Historic HPC/Advanced AI	$14,781	$15,369	$14,954	$17,932	$20,088	$22,279	$24,302	$26,810	12.4%
Non-Traditional Suppliers	$1,335	$3,437	$5,782	$7,458	$9,472	$11,420	$13,495	$14,967	21.0%
Total On-Prem	**$16,116**	**$18,805**	**$20,735**	**$25,390**	**$29,559**	**$33,699**	**$37,797**	**$41,777**	**15.0%**
Source: Hyperion Research, Oct 2024		16.7%	10.3%	22.4%	16.4%	14.0%	12.2%	10.5%	

HPC Spending forecast

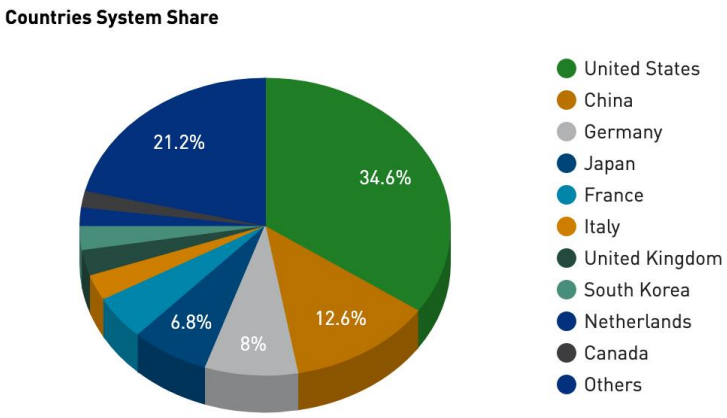
The Broader Market ($ Millions)

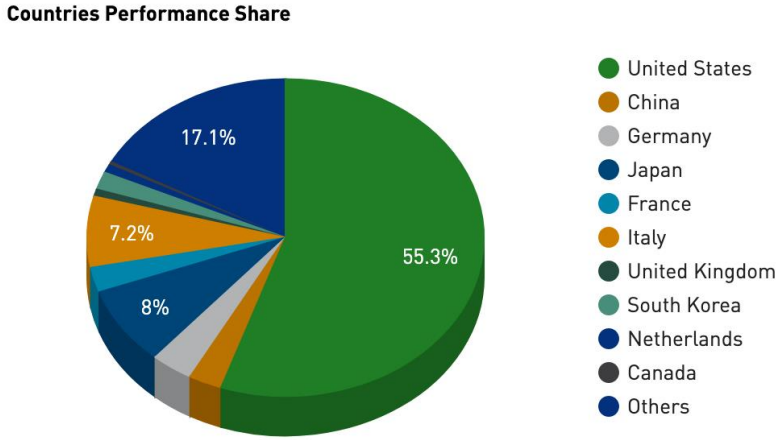
2022 total HPC spending reached $US 37 Billion
2026 is projected to exceed $US52 Billion

The Broader HPC Market	2022
On-premises servers	$15,441
Storage	$6,408
Middleware	$1,790
Applications	$5,092
Service	$2,224
Total On-premises	$30,956
Cloud Spending	$6,304
Total	**$37,260**
Source: Hyperion Research, 2023	

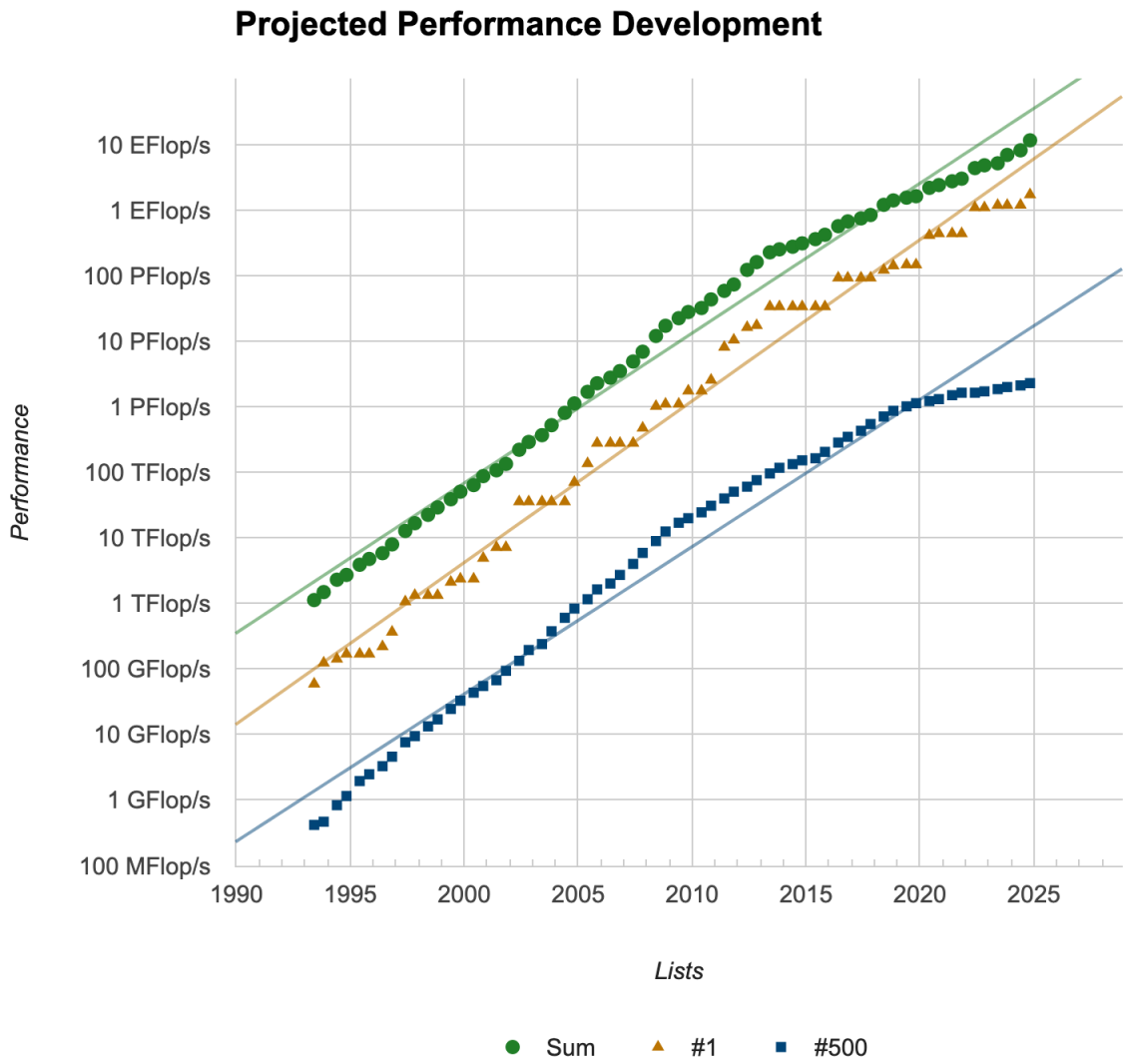
The Broader HPC Market	2026
On-premises servers	$20,576
Storage	$9,068
Middleware	$2,281
Applications	$6,349
Service	$2,308
Total On-premises	$40,582
Cloud Spending	$11,613
Total	**$52,195**
Source: Hyperion Research, 2023	

HPC Systems per Countries





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



https://top500.org/statistics/perfdevel/

HPL 500

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, **HPE** DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, **Intel** DOE/SC/Argonne National Laboratory United States	4,742,808	585.34	1,059.33	24,687
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, **Microsoft** Microsoft Azure United States	1,123,200	561.20	846.84	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, **Fujitsu** RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, **HPE** EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107
6	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, **EVIDEN** EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404
7	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, **IBM** DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 40C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR200, **EVIDEN** EuroHPC/BSC Spain	680,960	138.20	265.57	2,560
9	**Eos NVIDIA DGX SuperPOD** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400, **Nvidia** NVIDIA Corporation United States	485,888	121.40	188.65	
10	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, **IBM / NVIDIA / Mellanox** DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438

https://top500.org/lists/top500/2023/11/

HPCG 500

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	HPCG (TFlop/s)
1	4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, **Fujitsu** RIKEN Center for Computational Science Japan	7,630,848	442.01	16004.50
2	1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, **HPE** DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	14054.00
3	5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, **HPE** EuroHPC/CSC Finland	2,752,704	379.70	4586.95
4	6	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, **EVIDEN** EuroHPC/CINECA Italy	1,824,768	238.70	3113.94
5	7	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, **IBM** DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	2925.75
6	12	**Perlmutter** - HPE Cray EX 235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-11, **HPE** DOE/SC/LBNL/NERSC United States	888,832	79.23	1905.00
7	10	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, **IBM / NVIDIA / Mellanox** DOE/NNSA/LLNL United States	1,572,480	94.64	1795.67
8	13	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, **Nvidia** NVIDIA Corporation United States	555,520	63.46	1622.51
9	18	**JUWELS Booster Module** - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, **EVIDEN** Forschungszentrum Juelich (FZJ) Germany	449,280	44.12	1275.36
10	50	**AOBA-S** - SX-Aurora TSUBASA B401-8, Vector Engine Type 30A 16C 1.6GHz, Infiniband NDR 200, **NEC** Cyberscience Center, Tohoku University Japan	64,512	17.22	1089.00

https://top500.org/lists/hpcg/2023/11/

Green500

Green500 Data

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/ watts)
1	293	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
2	44	**Frontier TDS** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	120,832	19.20	309	62.684
3	17	**Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	319,072	46.10	921	58.021
4	25	**Setonix – GPU** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Pawsey Supercomputing Centre, Kensington, Western Australia Australia	181,248	27.16	477	56.983
5	92	**Dardel GPU** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE KTH - Royal Institute of Technology Sweden	52,864	8.26	146	56.491
6	8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 40C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR200, EVIDEN EuroHPC/BSC Spain	680,960	138.20	2,560	53.984
7	5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	7,107	53.428
8	1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	22,703	52.592
9	84	**Goethe-NHR** - Supermicro AS-4124GS-TNR, AMD EPYC 7452 32C 2.35GHz, AMD Instinct MI210 64 GB, Mellanox InfiniBand EDR, MEGWARE / Supermicro Universitaet Frankfurt Germany	96,768	9.09	195	46.543
10	496	**Olaf** - Lenovo ThinkSystem SR675 V3, AMD EPYC 9334 32C 2.7GHz, NVIDIA H100, Infiniband NDR 400, Lenovo Science Institute South Korea	3,936	2.03	45	45.117

https://top500.org/lists/green500/2023/11/

IO500

#	INFORMATION							IO500		
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW (GIB/S)	MD (KIOP/S)	REPRO.
1	ISC23	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory and Tsinghua University	SuperFS	300	36,000	210,254.98	4,847.48	9,119,612.35	
2	ISC23	JNIST and HUST PDSL	Cheeloo-1 with OceanStor Pacific	Huawei	OceanFS2	10	9,600	137,100.02	2,439.37	7,705,448.04	
3	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	31,200	43,218.79	11,362.27	164,391.73	
4	SC22	Sugon Cloud Storage Laboratory	ParaStor	Sugon	ParaStor	10	2,560	8,726.42	718.11	106,042.93	-
5	SC22	SuPro Storteck	StarStor	SuPro Storteck	StarStor	10	2,560	6,751.75	515.15	88,491.65	-
6	SC22	Tsinghua Storage Research Group	SuperStore	Tsinghua Storage Research Group	SuperFS	10	1,200	5,517.73	179.60	169,515.95	-
7	SC23	LRZ	SuperMUC-NG-Phase2	Lenovo	DAOS	90	6,480	4,585.68	1,054.72	19,937.45	
8	ISC22	National Supercomputing Center in Jinan	Shanhe	PDSL	flashfs	10	2,560	3,534.42	207.79	60,119.50	-
9	SC22	Cloudam HPC on OCI	HPC-OCI	Cloudam	BurstFS	64	1,920	3,033.03	278.48	33,033.54	-
10	SC21	Huawei HPDA Lab	Athena	Huawei	OceanFS	10	1,720	2,395.03	314.56	18,235.71	-
11	SC21	Olympus Lab	OceanStor Pacific	Huawei	OceanFS	10	1,720	2,298.69	317.07	16,664.88	-
12	SC21	Huawei Cloud		PDSL	Flashfs	15	1,560	2,016.70	109.82	37,034.00	-
13	ISC21	Intel	Endeavour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65	-
14	ISC20	Intel	Wolf	Intel	DAOS	52	1,664	1,792.98	371.67	8,649.57	-
15	ISC22	University of Cambridge	Cumulus	Dell/Intel	DAOS	200	2,000	1,107.17	283.19	4,328.68	-
16	SC22	Meadowgate Technologies	Meadowgate	INTEL HPE	DAOS	10	1,280	1,014.24	213.15	4,826.12	-
17	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456	988.99	176.37	5,545.61	-
18	SC21	BPFS Lab	Kongming		BPFS	10	800	972.60	96.26	9,827.09	-
19	SC19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625	938.95	174.74	5,045.33	-
20	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	
21	ISC20	TACC	Frontera	Intel	DAOS	60	1,440	763.80	78.31	7,449.56	-
22	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	
23	ISC22	Oracle Cloud Infrastructure	Oracle Cloud with WEKA on RDMA	WekaIO	WEKA	373	7,460	625.95	233.17	1,680.38	-

https://io500.org/

Metriche di riferimento – capacita', elaborazione, IO, Reti

Unita'/Tipologia	Capacita'	Prest Comp	Prest IO
Byte	1	OneFLOPs	OneBYTE/s
KB	X 1000	KiloFLOPs	KiloBYTE/s
MB	X 1000	MegaFLOPs	MegaBYTE/s
GB	X 1000	GigaFLOPs	GigaBYTE/s
TB	X 1000	TeraFLOPS	TeraBYTE/s
PB	X 1000	PetaFLOPs	PetaBYTE/s
EB	X 1000	ExaFLOPs	ExaBYTE/s

Unita'/Tip	Transfer Rate
Bit	Onebit
Kb	Kilobits
Mb	Megabits
Gb	Gigabits
Tb	Terabits
Pb	Petabits
Eb	Exabits

PB e' la scala caratteristica degli attuali sistemi di maggiori dimensioni (primi 20 al mondo)

EX e' la scala del primo e secondo sistema al mondo

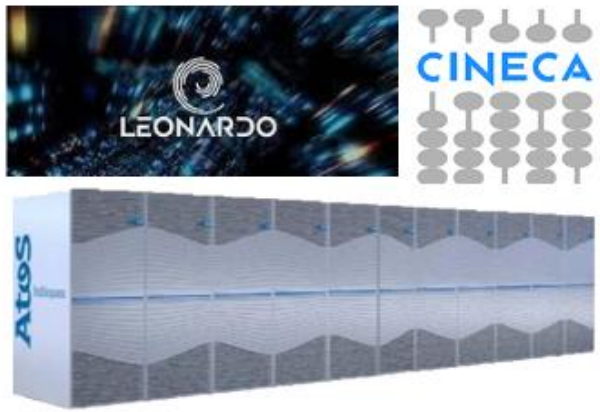
Tbs e' la scala di riferimento delle reti nei maggiori sistemi al mondo

https://io500.org/

European leadership – Promoting European Technology

6 EuroHPC systems out of 8 awarded to Atos (2/3 pre-exa, 4/5 peta)

EuroHPC Joint Undertaking



CINECA – Leonardo (Italy) 320 PFlops

BSC – MareNostrum5 (Spain) 300 PFlops

LuxProvide – Meluxina (Lux) 18.7 PFlops

IZUM – Vega (Slovenia) 10.11 PFlops

NCSA – Discoverer (Bulgaria) 6 PFlops

MACC – Deucalion (Portugal) 5 PFlops

EVIDEN

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

Elaborazione:
un numero elevato (≥10) di nodi di calcolo di varia tipologia
- Nodi single o dual CPUs
- Nodi oltre dual CPUs
- Nodi con acceleratori GPU

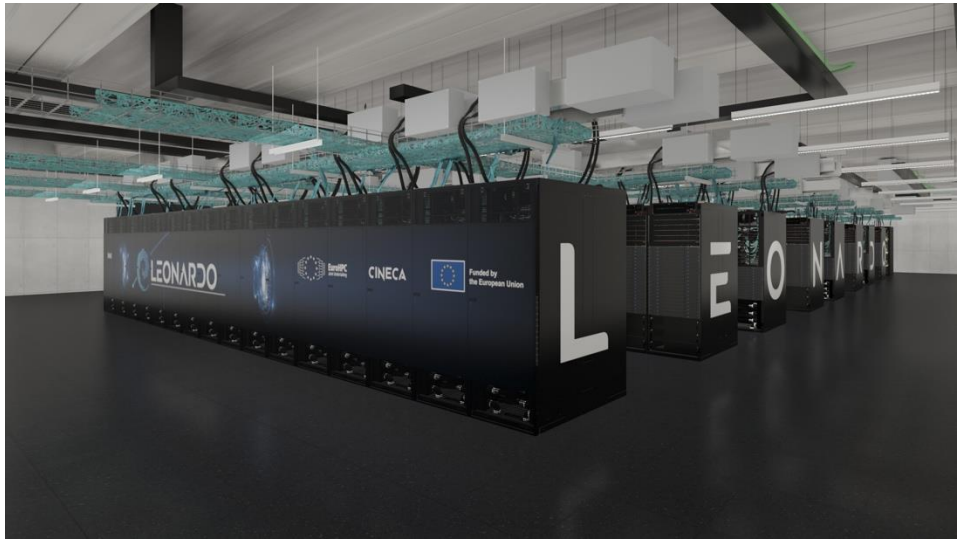
Comunicazione:
Rete a alta efficienza e bassa latenza
- Varie topologie per supportare un’elevata scalabilita’ orizzontale
- Rete di gestione generalmente su protocollo Ethernet per limitato scambio dati

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/

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

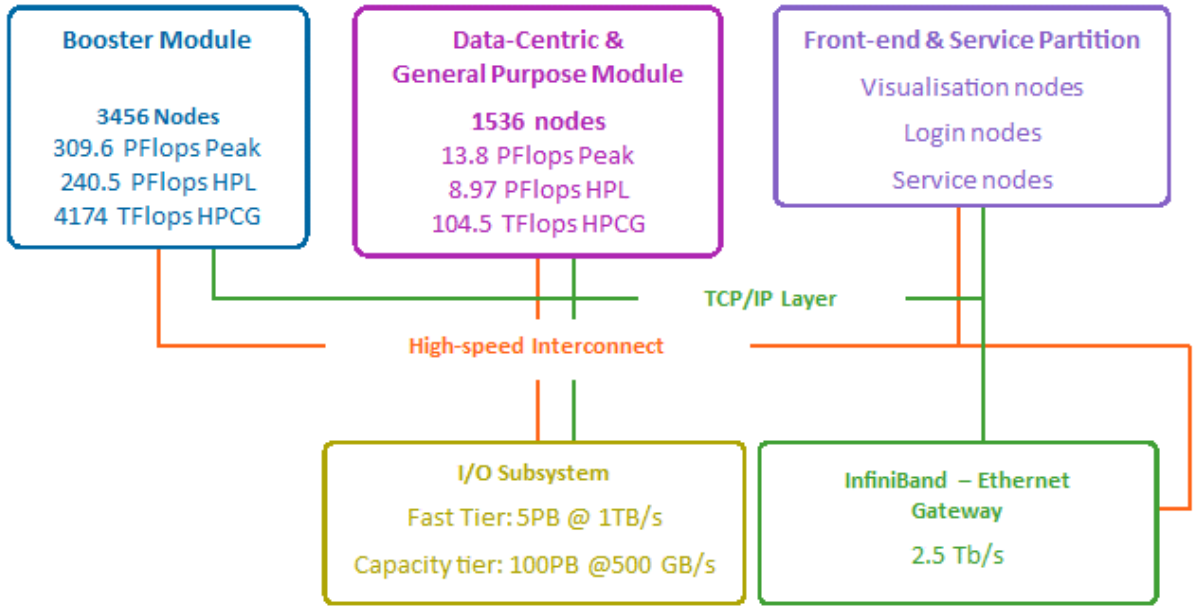


10 ExaFlops FP16 AI workload

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

22

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

Misure prestazionali di riferimento per un sistema HPC: HPL, HPCG, IO

HPL: rappresenta la risoluzione di una matrice piena

Elevato rapporto tra il carico computazionale per punto di risoluzione rispetto alla comunicazione

Enfatizza le capacita' di elaborazione del processore rispetto alle capacita' di comunicazione dati tra nodi concorrent

Generalmente circa 70% di efficienza rispetto a theor peak perf -→ RMAX/RPEAK ~ 0,7

Rappresenta ragionevolmente bene applicazioni con ampio uso di FFT, alcune simulazioni CFD, applicazioni EDA per campi magnetici, simulazioni geofische, simulazioni in material science

HPCG: rappresenta la risoluzione di una matrice sparsa

Basso rapporto tra il carico computazionale per punto di risoluzione rispetto alla comunicazione

Enfatizza le capacita' di comunicazione dati tra nodi concorrenti rispetto alle capacita' di elaborazione del nodo

Generalmente circa 5% di efficienza rispetto a theor peak perf -→ RMAX/RPEAK ~ 0,05

Rappresenta ragionevolmente bene applicazioni in ambito meteo/clima, alcune simulazioni CFD in geometrie complesse, simulazioni a elementi finiti in meccanica strutturale e in sistemi Multiphysics

IO500: rappresenta l'accesso a dati non strutturati e di diversa tipologia

Accesso random a dati di piccolo e medie dimensioni

Enfatizza le capacita' di comunicazione dati tra nodi concorrenti e le caratteristiche del sottosistema storage connesso

Rappresenta ragionevolmente bene il traffic I/O in ambito bioinformatico, analisi dati non strutturati per knowledge graph, simulazioni in AI and ML

https://top500.org/

Tipologia di architettura

SIMD: single instruction multiple data

MIMD: multiple instruction multiple data

Essential CPU Components List and Their Functions

There are several key CPU components that work together to ensure proper functioning of the processor. These components play distinctive roles in processing instructions, controlling data flow, and conducting other tasks within the computer system:

Find Study Materials Create Study Materials

communication between other CPU components and peripherals.

- **Arithmetic Logic Unit (ALU):** The ALU is responsible for performing mathematical and logical operations, such as addition, subtraction, multiplication, and comparison of numeric values.
- **Registers:** Registers are small, fast storage areas within the CPU that temporarily store data or instructions being used. They include the program counter (PC), instruction register (IR), and various general-purpose registers.
- **Cache Memory:** Cache is a small, high-speed memory area within the CPU that stores frequently used data and instructions, reducing the time taken to fetch them from main memory and improving processing speed.
- **System Clock:** The system clock generates a continuous series of electrical pulses that control the pace at which instructions are executed. Faster clock speeds result in faster processing.
- **Bus:** The bus is a set of wires that facilitate the transfer of data and instructions between different components within the CPU as well as other devices in the computer system.

https://www.studysmarter.co.uk/explanations/computer-science/computer-organisation-and-architecture/cpu-components/

La rivoluzione del 1975 : CRAY-1 e l'introduzione sul mercato delle architetture vettoriali

CRAY-1: 8 registri vettoriali a 64bits che supportava 8x DAXPY per ciclo

Da CRAY-1 il settore HPC ha avuto una svolta fondamentale

IBM a fine anni '80 introdusse la vector facility (vector array processor applicato ai processor scalari)

Digital Equipment tento' negli anni '90 una soluzione analoga ma con poco successo di mercato

Intel introdusse nel 2008 per l'architettura x86 l'unita' vettoriale AVX (Advance Vector Extension) che attualmente supporta 512bit FP e 32 registri vettoriali

Gli acceleratori grafici (GPU) nella sostanza sono dei sistemi vettoriali

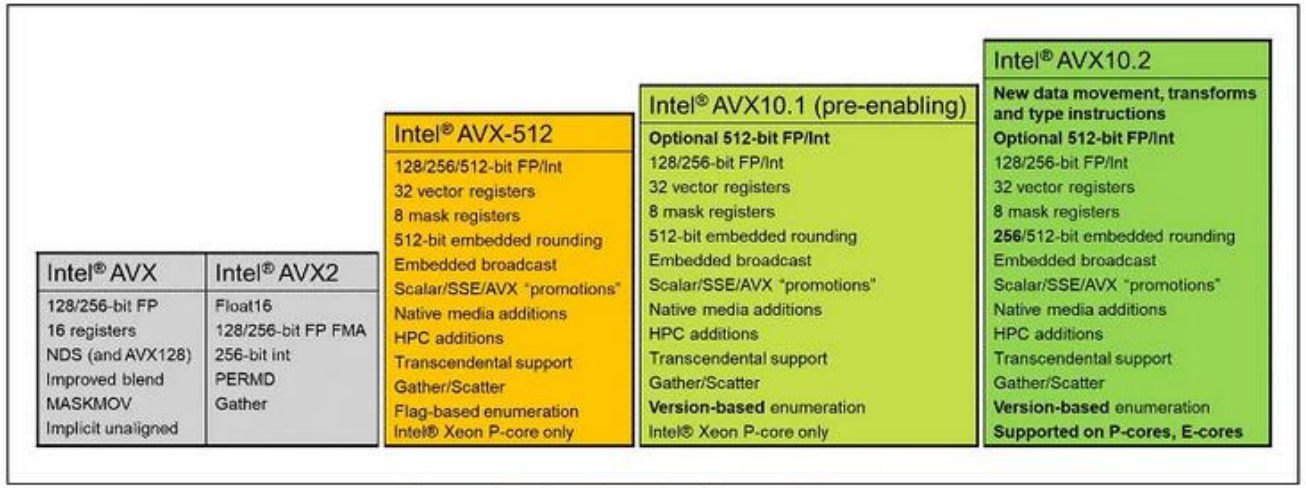


Figure 1-2. Intel® ISA Families and Features

https://comeaprire.com/definizioni/una-panoramica-completa-dei-processori-vettoriali/

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

Operazioni elementari per la misura delle prestazioni: SAXPY, DAXPY, CAXPY and ZASPY

Necessita' di definire un metodo di misura delle prestazioni che fosse rappresentativo di algoritmi reali e adatto a differenti architetture

La somma di due vettori e il prodotto con uno scalare comporta:

Load da RAM ai registri del processore
Multiply and add sulla pipeline
Store del risultato nel vettore y

I processor consentono alcuni overlap a ogni ciclo:
Load sui registri
Mult+Add+Store su pipeline vettoriale

CPU theorethical peak performance:

Clock(GHz) *fops *cores → x GFLOPs

Es: 3*32*32 = 3TFs

Function

The computation is expressed as foll

$$\begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \leftarrow \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} + \alpha \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

Purpose

These subprograms perform the following computation, using the scalar α and vectors ***x*** and ***y***:

$y \leftarrow y+\alpha x$

Table 1. Data Types

Data Types	
alpha, ***x***, ***y***	**Subprogram**
Short-precision real	SAXPY
Long-precision real	DAXPY
Short-precision complex	CAXPY
Long-precision complex	ZAXPY

https://www.ibm.com/docs/en/essl/6.2?topic=vss-saxpy-daxpy-caxpy-zaxpy-multiply-vector-by-scalar-add-vector-store-in-vector

Metriche tipiche di un Sistema HPC a alte prestazioni di generazione attuale

Elaborazione:
PETAFLOPS (2008 https://it.wikipedia.org/wiki/IBM_Roadrunner) to Exaflops (2023 https://www.hdblog.it/hardware/articoli/n517740/el-capitan-supercomputer-2-exaflops-amd-epyc/)

Comunicazione:
GigaBits (Gbs) to Tbs

Storage:
Prestazioni: GigaByte/s (GBs) to TBs
Capacita': PB to TB to Exabyte

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

https://io500.org/

Calcolo sequenziale e parallelo: i limiti della scalabilita' parallela

Efficienza parallela = (T(1)/p) /T(p) → tempo teorico / tempo reale

Elapsed time (p) = T*s + T*(p)

T*s = tempo di orologio per la sola parte sequenziale non parallelizzabile
T*(p) = tempo di orologio per la parte parallelizzabile

T(1) = T*s + T*(1)

T(p) = T*s + α(p)x T*(1)/p -→ T(p) ~ T*s when p >> 1

Eff Par = (T*s + α(p)x T*(1)/p)/(px T(p)) = T*s/(px T(p)) * (1 + α(p)x T*(1)/(T*s x p))

Complessita’ computazionale

3D Turbulent Flow → gradi di liberta’ ~ Re**(9/4) Re ≥ 3000

https://pubs.aip.org/aip/pof/article-abstract/21/12/125103/256210/The-number-of-degrees-of-freedom-of-three?redirectedFrom=fulltext

Image Processing, Seismic Processing, etc -→ Fast Fourier Transform -→ NlogN

https://math.stackexchange.com/questions/1704788/complexity-of-fft-algorithms-cooley-tukey-bluestein-prime-factor

N to N direct interactions -→ N**2

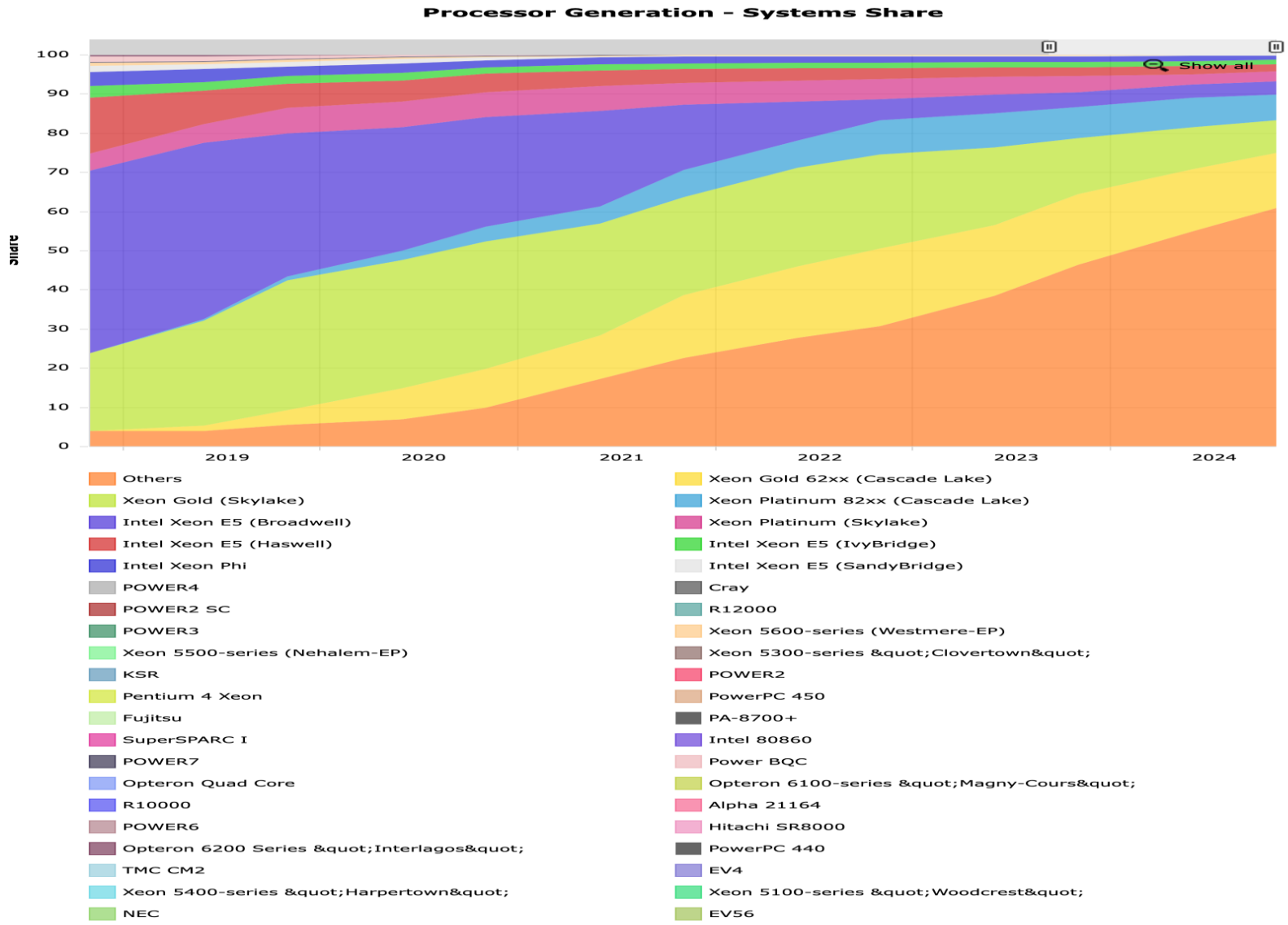
Material science, bioinformatic, n-body problems, etc

Piano del Corso – Lesson 2

Architetture di calcolo e loro evoluzione

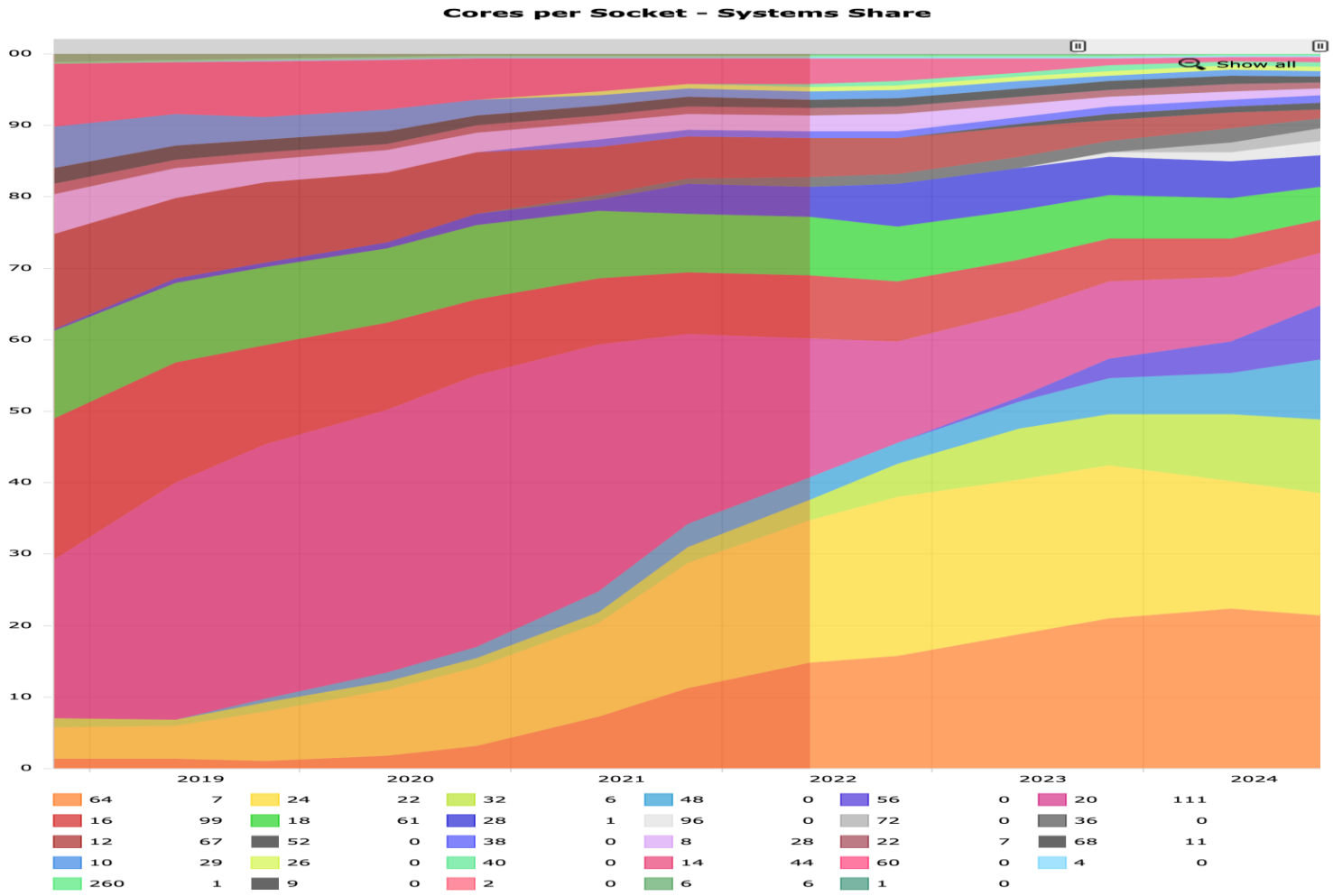
- Architetture omogenee e accelerate
- Concetti generali sui microprocessori (CPU)
- Concetti generali sugli acceleratori (Graphical Processor Unit)
- Integrazione CPU-GPU e trasmissione dati

Processor generation – 2018-25



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

Core per socket 2019-24



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

Processori e loro limiti alla scalabilita’ prestazionale

Theorethical Peak Perf = clock * fops * cores

Generalmente fops = 16 ÷ 32 64bits (Intel e AMD (recenti annunci))
iops = 32 ÷ 64 32bits

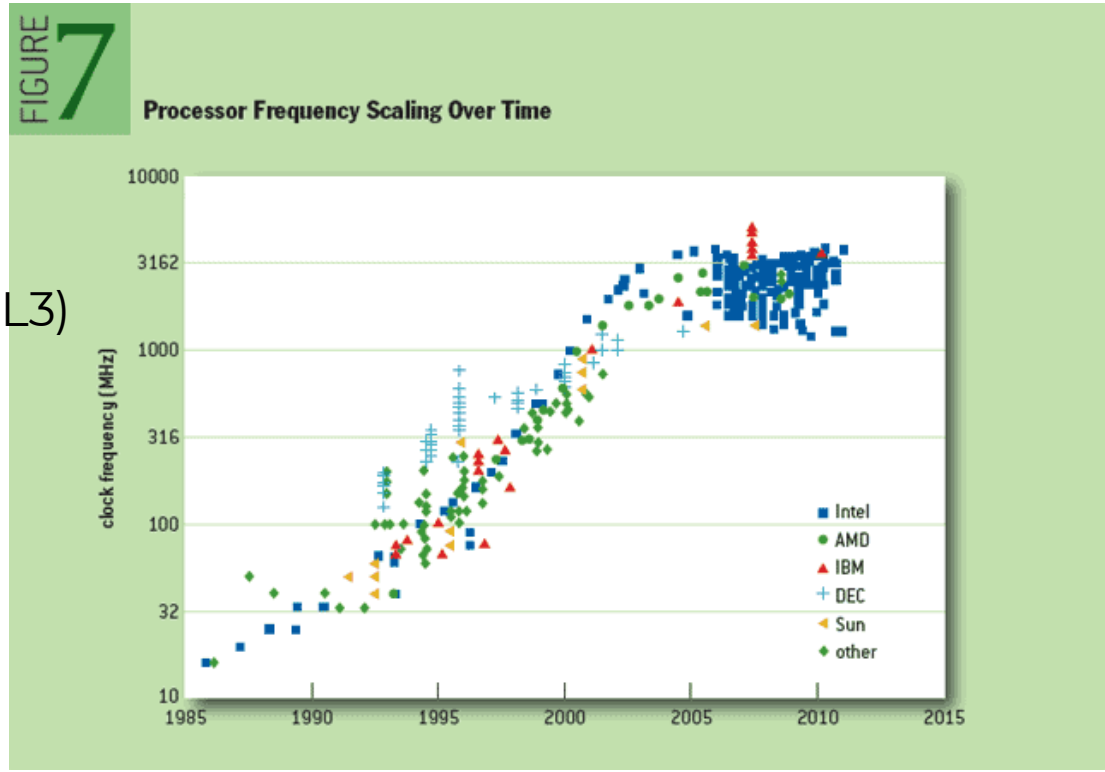
La frequenza del processore non puo’ crescere oltre un limite per i consumi e la dissipazione di calore

Attuali clock ≤ 4GHz → TDP ≤ 450°C

Aumento del numero di cores per CPU per aumentare le prestazioni

Limite al numero di cores dovuti alla memory BW (L2 e L3)

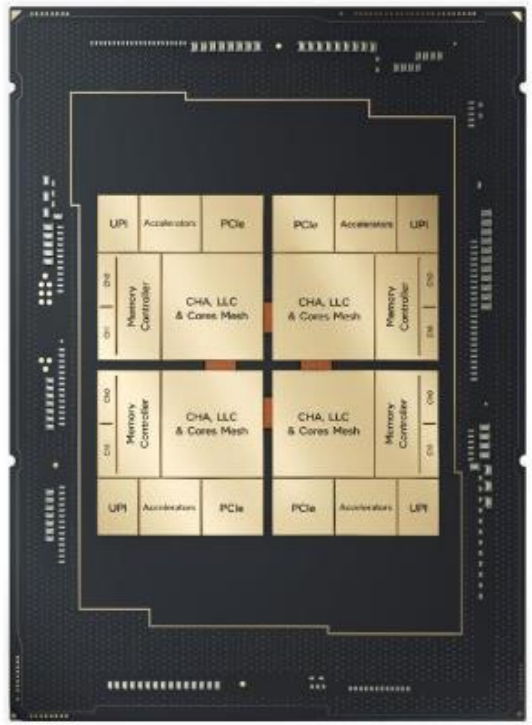
Il clock non e’ univoco: AVX per alcune architetture potrebbe avere un clock circa 70% del valore nominale

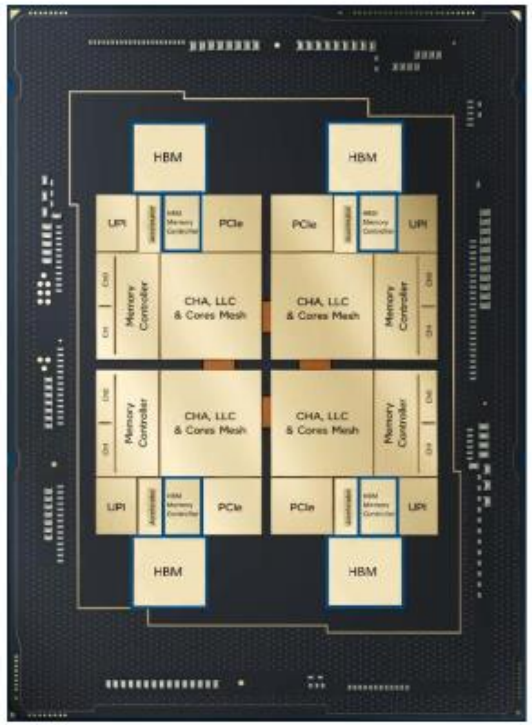


Intel Sapphire Rapids

4th Gen Intel® Xeon® Scalable Processors

- Up to 56x Cores
- 64GB HBM2e (MAX Series)
- 8x Channels of DDR5 4800MT/s
- Up to 4x UPI link at 16GT/s
- Increased I/O Bandwidth with PCIe Gen 5.0
- Intel® Speed select technology
- Intel® Advance Matrix Extension (AMX) and AVX-512
- Foundational support for CXL 1.1
- TDP up to 350W

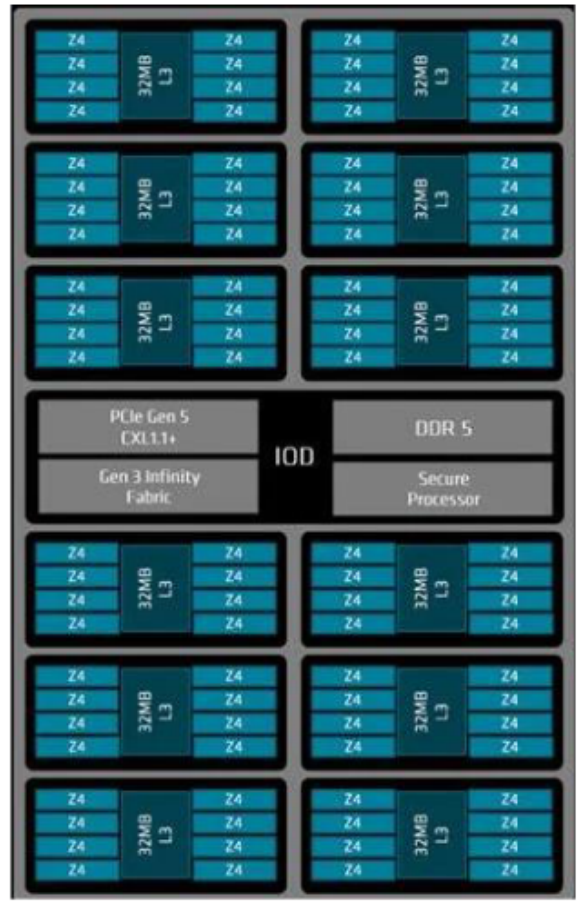




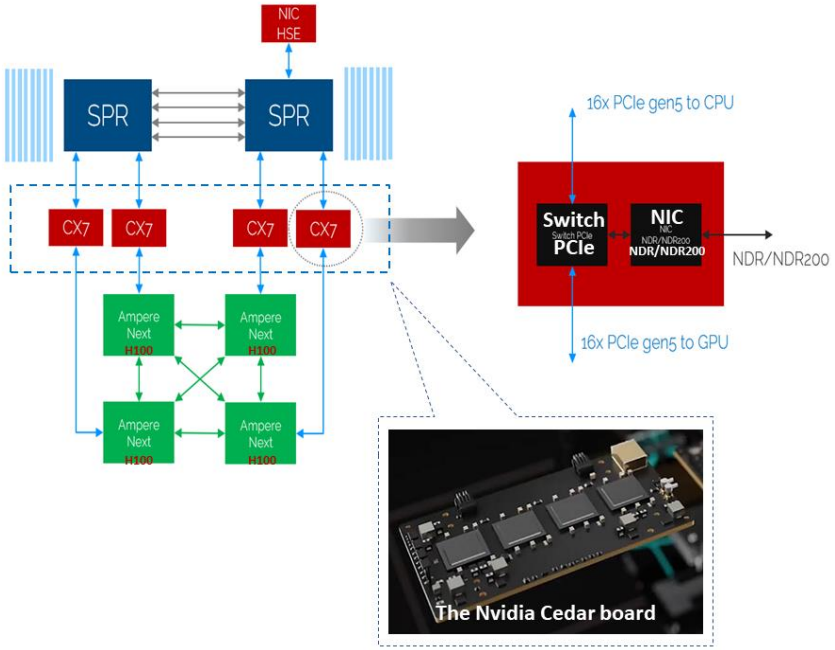
AMD EPYC Genoa & Bergamo

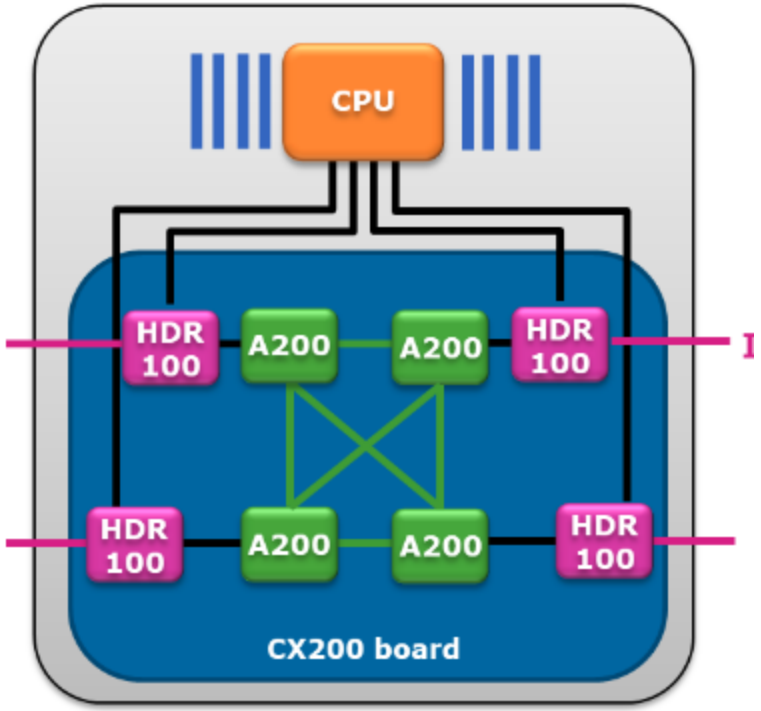
4th Gen AMD EPYC™ Processors

- Up to 96x – 128x Zen4 cores
- Maximum boost clock up to 4.4GHz
- 5nm technology
- Up to 12TB of memory per socket
- 12x memory channels of DDR5 4800MT/s
- Up to 160x PCIe 5.0 lanes
- Foundational support for CXL 1.1+
- AVX-512 support
- TDP up to 400W



System design – 1x CPU + 4x GPUs

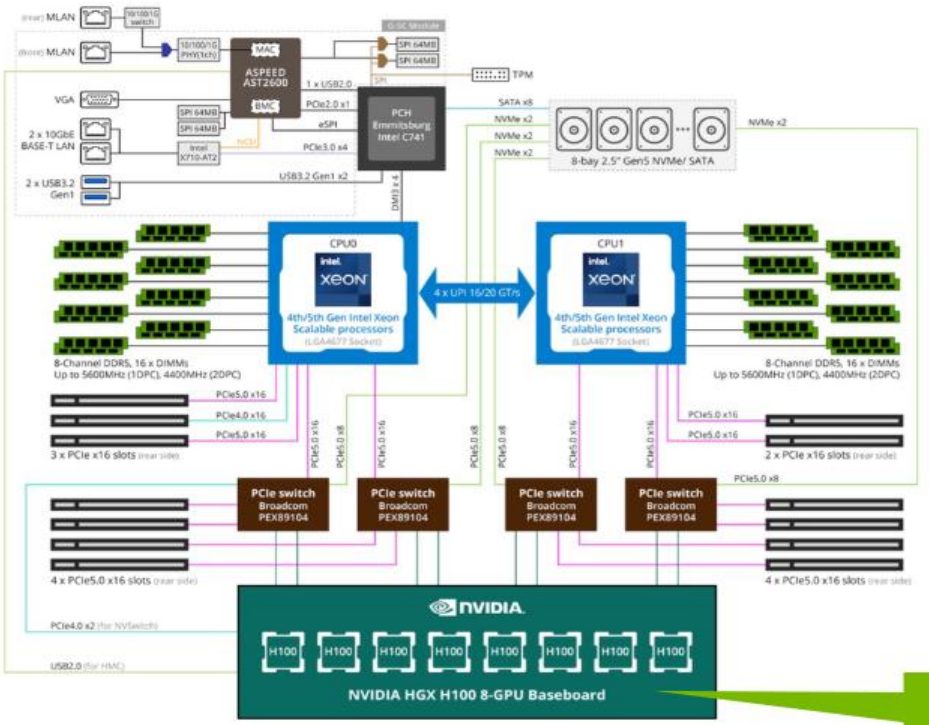




System design – 2xCPUs + 8xGPUS in a node

Automatic Zoom

Eviden Sequana 410 - Hopper 8-Way Design



EVIDEN
an atos business

full NVLINK bandwidth

NVIDIA

Le GPU consentono un aumento sostanziale delle prestazioni in un nodo con consumi compatibili

H100, GH200, H200: 34TFs FP64 – 67TFs FP32

https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

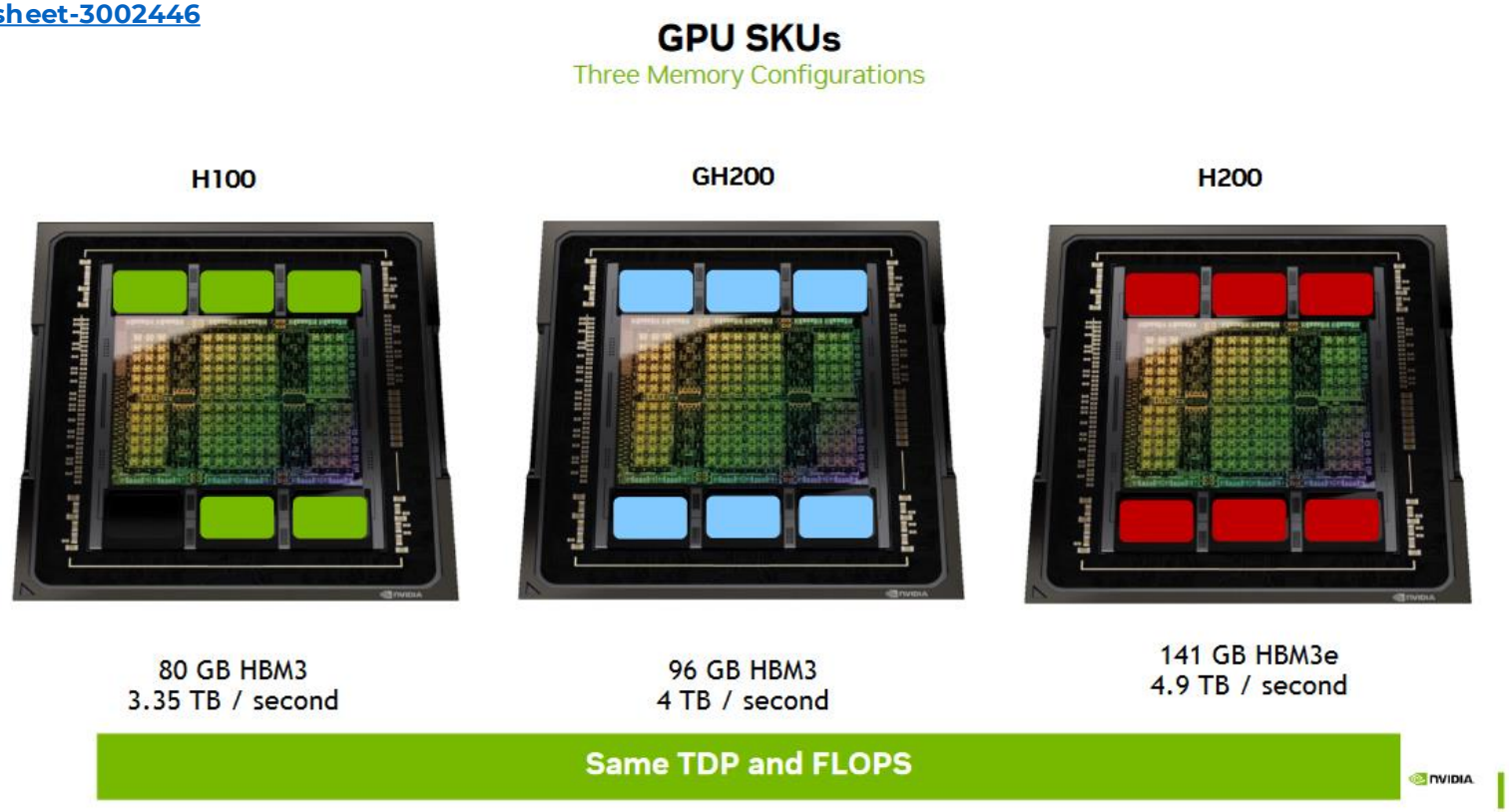
https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23-h200-datasheet-3002446

TDP ~ 700W

~48GFs/W theoretical on a GPU

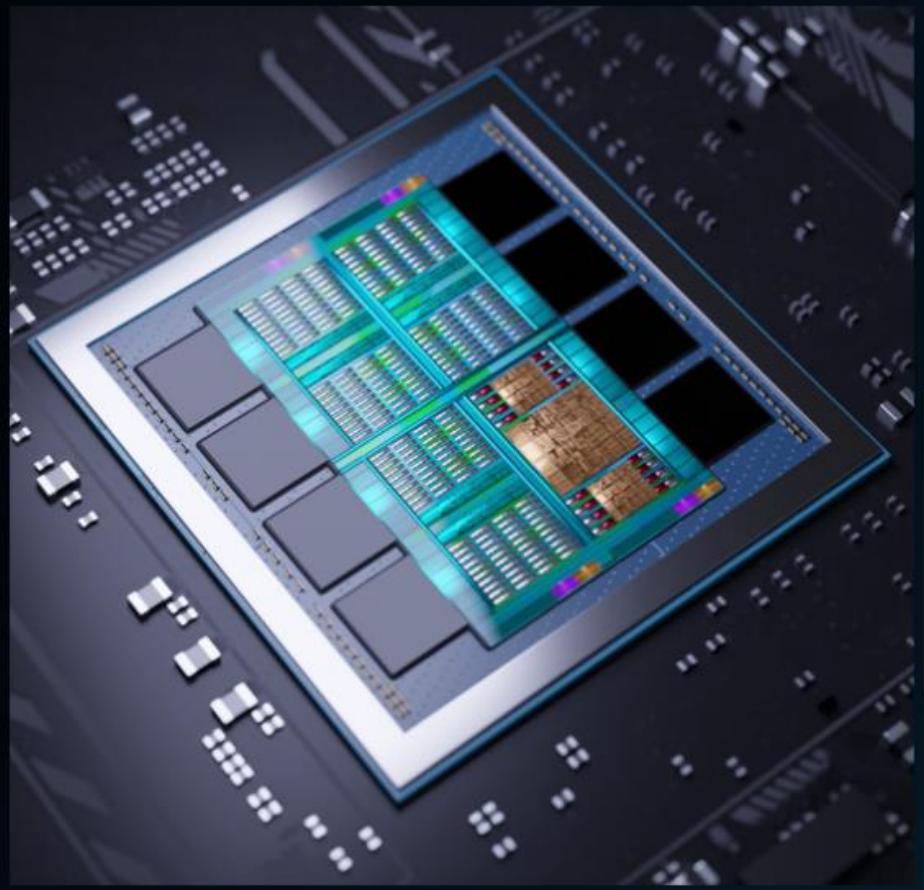
vs

~12GFs/W theoretical on a CPU



MI300A APU - Delivered Performance (TF) (Projected)

Up to 850W TDP



	TDP (W)	550W	700W	850W
	GPU (CDNA 3) Cores	228 CU		
	CPU (Genoa) Cores	24 Cores		
	Memory Size	128GB HBM3		
	STREAM (TB/s)	4.0	4.0	4.5
	Infinity Fabric BW – Full Mesh (TB/s)	896 GB/s		
HPC	FP64 HPL (TFLOPS)	55	66	70
HPC	FP64 Matrix / DGEMM (TFLOPS)	72	85	93
HPC	FP32 Matrix / SGEMM (TFLOPS)	99	111	119
HPC	FP64 Vector / FMA64 (TFLOPS)	45	51	54
ML	FP32 Vector / FMA32 (TFLOPS)	72	85	92
ML	BFLOAT16 (TFLOPS)	636	723	771
ML	FP16 (TFLOPS)	598	696	759
ML	FP8 (TFLOPS)	1,085	1,296	1,414
ML	INT8 (TOPS)	987	1,356	1,456

MI300X/MI300A products and technology are subject to US export licensing requirements outlined by ECCN codes 4A003.c and 4E001.a of the US Export Administration Regulations (EAR)

Esempio di un’architetture di calcolo di classe Exascale con GPUs

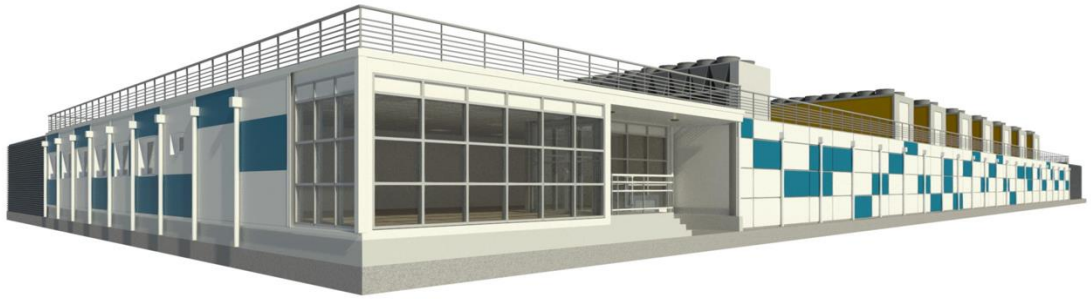
4x H100 per nodo: 136TFs FP64 theoretical peak perf (up-to-day NVIDIA GPU)

HPL efficiency ~70%

~10.000 nodi ~ 1,4 ExaFs RPEAK ~ 1ExaFs RMAX (up-to-day NVIDIA GPU)

~ 3500 nodi ~ 1,4 ExaFs RPEAK ~ 1ExaFs RMAX (new generation NVIDIA GPU)

~14 ÷ 20MW power consumption at HPL (a rough estimate)



JUPITER@Julich in a modular solution

https://www.fz-juelich.de/en/news/archive/press-release/2024/next-milestone-for-jupiter-2013-high-tech-base-for-the-european-exascale-supercomputer

https://www.nextplatform.com/2023/10/05/details-emerge-on-europes-first-exascale-supercomputer/

Summary and comments – I&II Lessons

- HPC architecture evolution: scalar → vector → cluster of nodes

- SIMD vs MIMD microprocessor architecture

- peak vs sustained performances

- CPU vs GPU: performance and power consumption

- key building blocks for an AI&HPC architecture

- Vertical vs horizontal scalable architecture

Thank You