Supercomputing 2025
St. Louis

Costantini, Longo, Scarponi, Zani, Chierici

SC25

- The International Conference for High Performance Computing, Networking, Storage, and Analysis

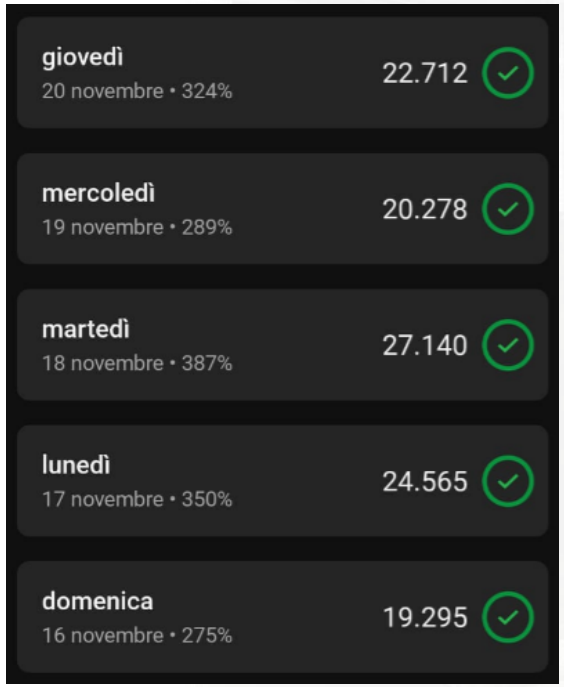
Top500

- Durante SC a novembre (e ISC a giugno) viene aggiornata la lista top500, contenete i 500 calcolatori più potenti al mondo.
- In questa lista siamo presenti anche noi, grazie al cluster terabit HPC.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
435	**Jean** - Liqid Cluster, Intel Xeon Platinum 9242 48C 2.3GHz, Infiniband HDR, Liqid Army Research Laboratory DoD Supercomputing Resource Center (ARL DSRC) United States	55,296	2.92	4.07	
436	**Castner** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE Hewlett Packard Enterprise United States	13,056	2.92	3.54	
437	**Terabit HPC Bubble** - ThinkSystem SR675 V3, AMD EPYC 9654 96C 2.4GHz, NVIDIA H100 80GB HBM3, NVIDIA Infiniband NDR, AlmaLinux 9, Lenovo INFN-CNAF Italy	14,400	2.91	5.36	
438	**KKK1** - ThinkSystem HR650X, Xeon Gold 6233 24C 2.5GHz, 25G Ethernet, Lenovo Service Provider T China	77,760	2.91	6.22	

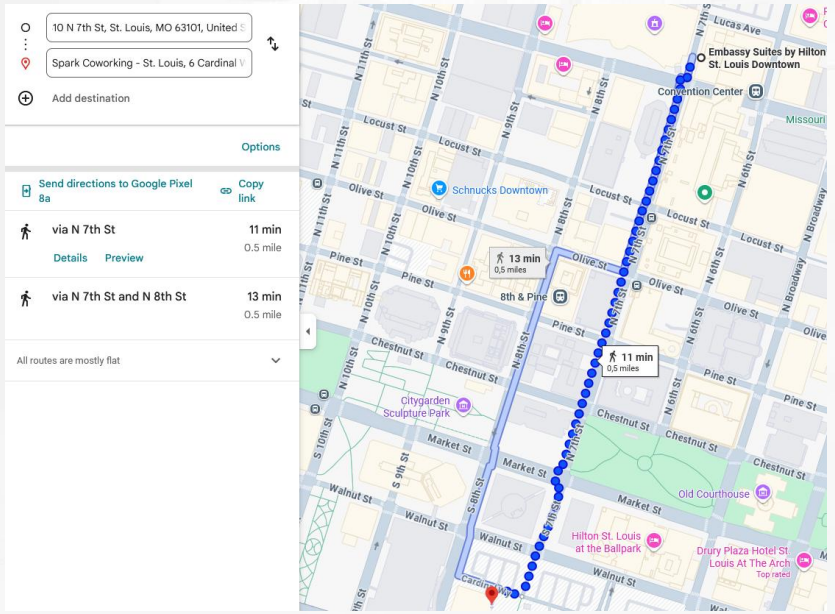
La nostra “vacanza” in USA

Passi giornalieri di Andrea



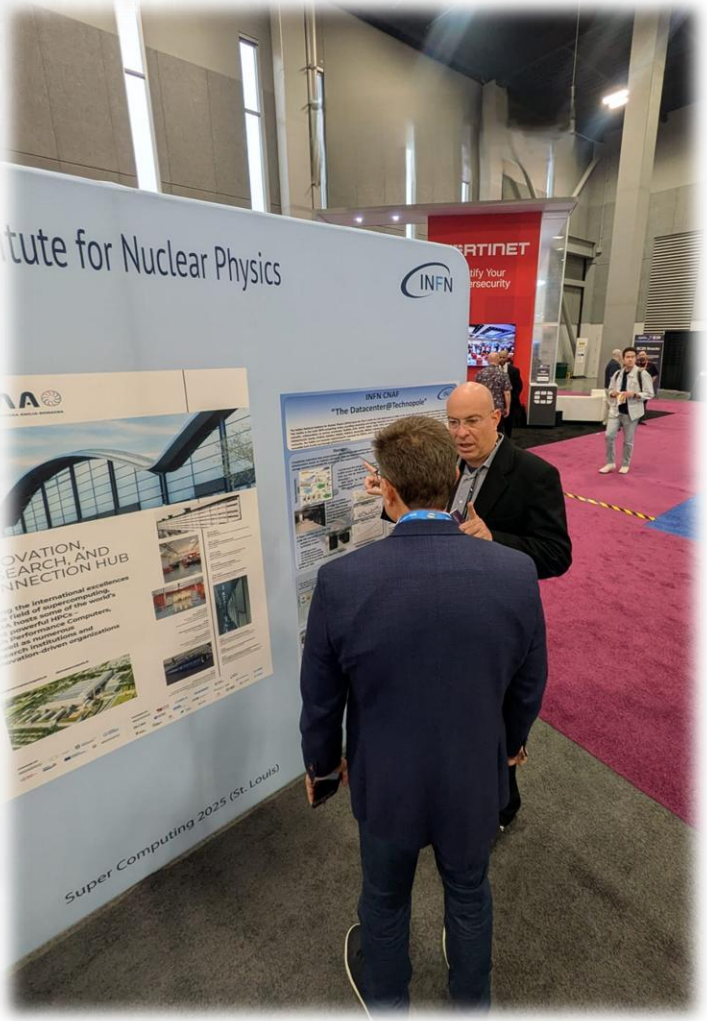
Qui non ho fatto la mia “corsetta al parco”

- Incontri quotidiani con vendor e produttori
  - Spesso in alberghi attorno al convention center e lontani tra loro



- Exibit floor immenso, tantissime cose da vedere

Il Booth INFN



SC25 in un'immagine

Tanti dubbi per il futuro…

In breve

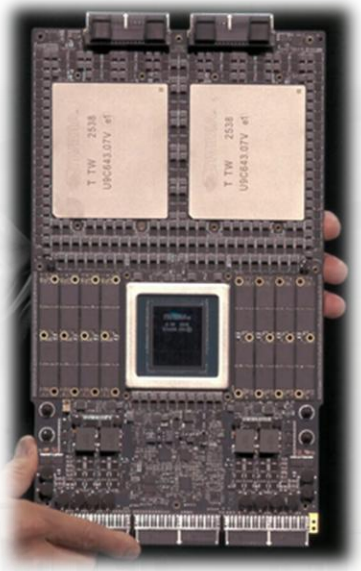
- «Eventi privati sotto NDA, no foto, slide in arrivo» VS «Exhibit floor»
- Proposte «full-Nvidia» onnipresenti – Jensen Huang lunedì presente a firmare autografi
- I numeri di certe installazioni sono imbarazzanti
  - Argonne schiererà 100k GPU nel nuovo data center (NVL72)
- Molte soluzioni su rack OCP
- Il DLC è arrivato anche sugli switch
- Gli aumenti di prezzo sono generalizzati: preoccupa molto quello delle memorie
- CNAF 2030 – Che rilevanza potremo avere in futuro e come deve evolvere il nostro data center per sostenere le prossime richieste?
  - CPU core-count sempre più alto, ha senso per noi?
  - Vendor dichiarano che abbandoneranno presto rack da 19”
  - Rack più alti di 42 unità

Eventi NDA e sessioni private

- Intel NDA
- AMD NDA
- Nvidia NDA
- Lenovo NDA e whisper room
- Dell NDA
- Riallacciati contatti con HPE
- ARISTA
- NOKIA

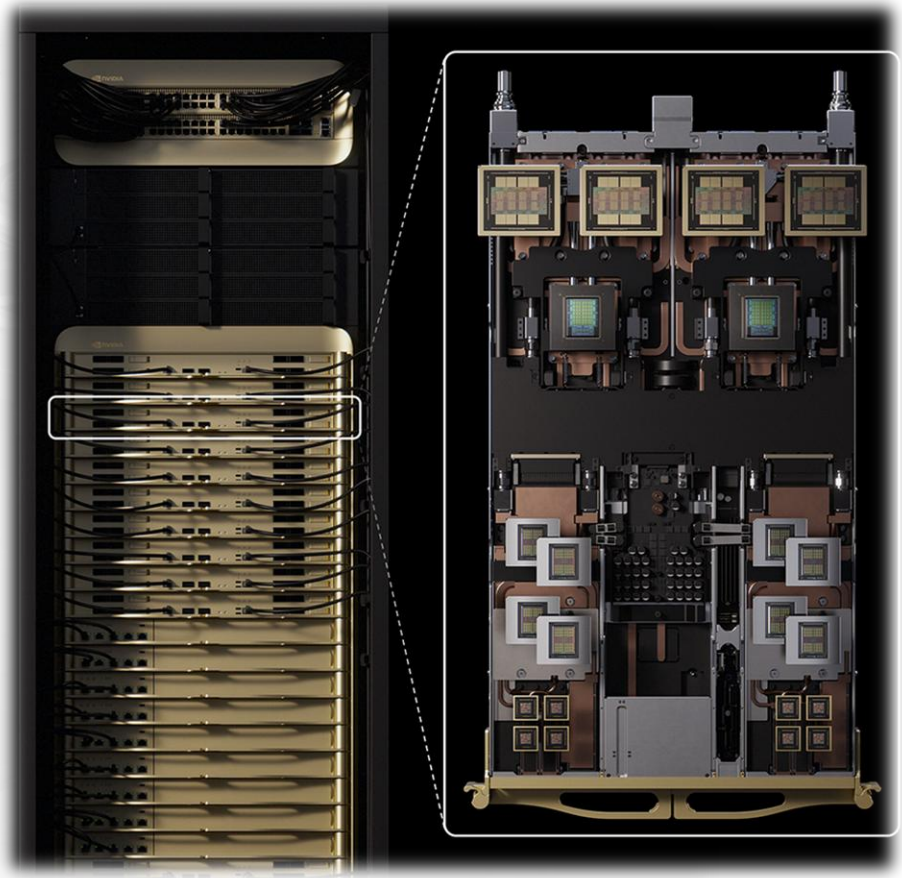
Nvidia - CPU

- Dopo Grace ci sarà Vera (Rubin)
- E' il primo design CPU tutto interno
  - Dichiarano raddoppio di prestazioni rispetto a grace
  - 88 core, 350W TDP
  - Creato per essere accoppiato alle GPU che seguiranno a Blackwell (Vera), non per competere con Intel/AMD
  - Memory controller fuori dal die principale
  - Supporto sia per LPDDR5 (preferito) che a standard DDR5
    - Si può intervenire sulla RAM con un cacciavite
  - Produzione di massa 2q2026



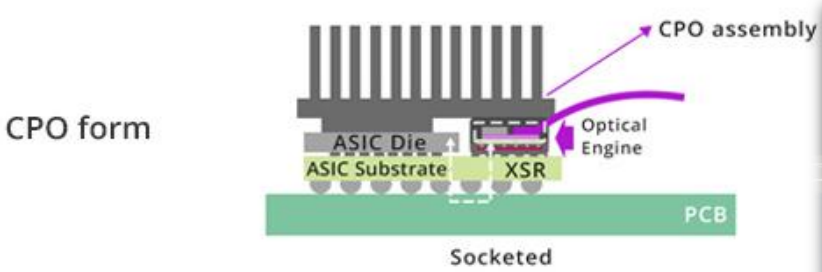
Nvidia - GPU

- Blackwell ultra (B300) target inferenza e per questo ha abbandonato alcune fp*.
  - Blackwell tradizionale (B200) resta, essendo comunque una soluzione più bilanciata.
- Rubin sarà la generazione successiva
  - «più di tutto», con HBM4, molto costosa...
- Soluzioni PCIx: RTX
  - non hanno FP64, ma mantengono FP32
  - GDDR per contenere i costi
  - Esistono due modelli, la 6000 che già conosciamo e la 4500, per lower power, infatti chiede solo 165W.
  - E' possibile usare MIG.
- Non implementano più in hardware i tensor core, ma il resto c'è ancora. Prestazioni FP64 ancora significative.



Nvidia - Network

- Switch basato su silicon photonics: quantum-x (CPO C
  Packaged Optics)
  - Niente più ottiche, usa direttamente MPO illuminati internamente da laser.
  - 144 porte 800gbps infiniband, consumo a 4kw
  - Tramite topologia fat-tree a due livelli, è possibile collegare fino a 10.000 GPU.



AMD - CPU

- Continua evoluzione dei prodotti che già conosciamo
- Prodotto attuale «Turin», poi sarà la volta di «Venice»
  - Architettura zen5 da TSMC in 3 o 4 nm, supporto ad AVX512
  - Dichiarato aumento 17% IPC
  - EPYC 9575F prodotto più interessante per noi, 64 core 400W
- Venice avrà due piattaforme distinte, SP7 e SP8
  - SP7: orientato ad HPC, 16 o 12 canali di memoria e TDP fino a 600W
  - SP8: orientato a «enterprise», 8 canali di memoria e TDP fino a 400W
- In futuro, CPU a 64 core "high-frequency" si attesteranno sui 500W

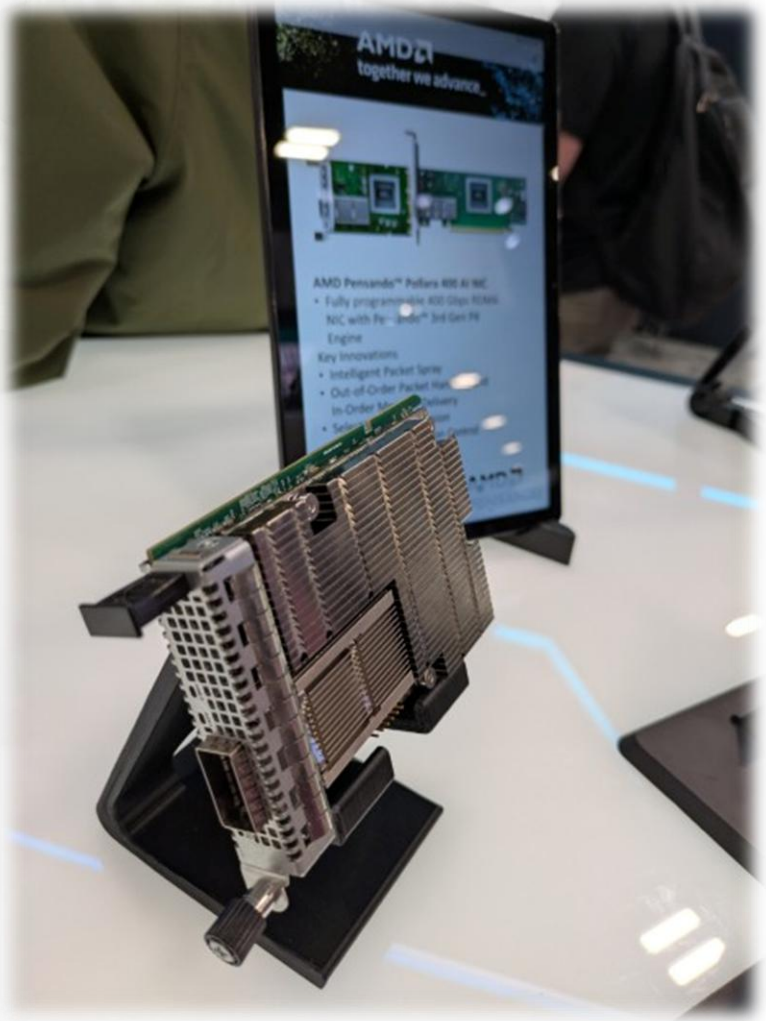
AMD - GPU

- Con «El Capitan» mantengono leadership e grande cassa di risonanza
- MI350P a 600W, 144GB HBM3e
- MI410P, prevista Q4-2026, 600w+ 180GB hbm4, solo ad acqua
- Non prevedono aria per le GPU in futuro
- Per i nostri casi d'uso GPU PCIe

- Spergiurano che con il loro software stack aggiornato, basta poco per convertire codice CUDA.
- HIP è il loro «equivalente» di CUDA, open source e accessibile a tutti
  - HIPIFY converte codice CUDA «al volo»

AMD - Network

- Prodotti ultra-ethernet
  - Acquisizione di “Pensando network”
  - Non una vera alternativa a Infiniband
  - Producono solo schede, che dichiarano essere compatibili con tutti gli switch di produttori terzi (ad esempio quelli di Broadcom)
- Prossima scheda sarà “Vulcano” con prestazioni fino a 800Gbit (claim: 800gbit AI NIC)



Intel

- CPU: nuovo approccio basato su architetture eterogenee (AMD non lo fa)
  - E-core si contrappongono ai classici P-core
  - Xeon SP fino a 86 P-Core
  - Xeon AP fino a 128 P-Core - Diamond Rapids nel 2027 fino a 192 P-core
  - Si va verso un'architettura con una sola CPU con lo stesso core-count di due
  - Opzione per MRDIMM: incremento di prestazioni ma è richiesto un esborso superiore del 20%
- GPU: «Crescent island» esce a fine 2026, come scheda PCIe a 350W
  - Seguirà nel 2028 «other island», con un TDP da 600W
  - 60GB RAM, ed è espandibile fino a 480
  - Con FP8 dichiarano di andare meglio della RTX6000, costando meno. Prestazioni vicine anche a Rubin 40 PCIe
- In tardo 2028 dovrebbe uscire «Jaguar shores», sicuramente a liquido, forse ci sarà un'alternativa anche ad aria
- Software stack che cerca di avere punti in comune con CUDA, sperano di semplificare la transizione.

Lenovo

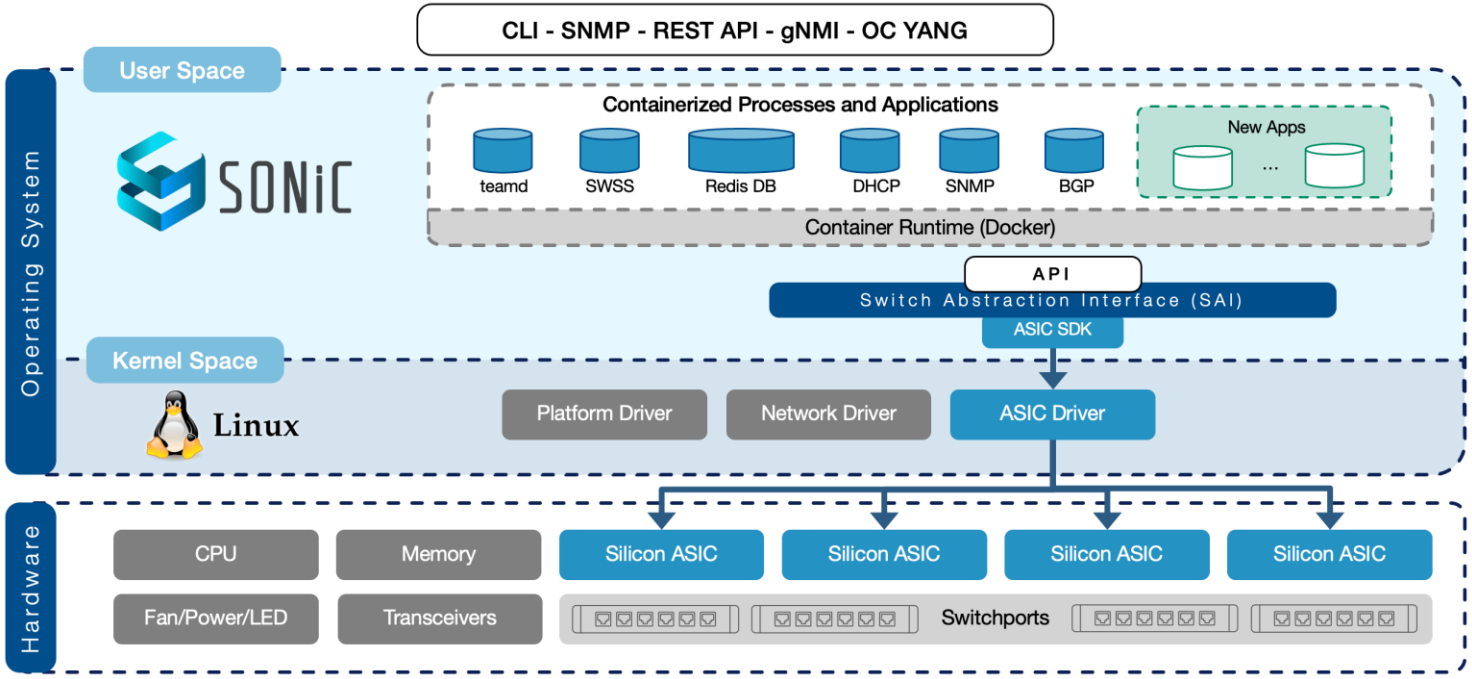
- Sistema «Neptune» evolve per accomodare tutte le novità del mercato.
  - Chassis 13U – 8 lame (Ogni lama 2 nodi)
    - Le nuove versioni con AMD Venice occupano 2 slot
    - Può essere suddivisa in 4 macchine, ciascuna con 1 socket
    - Fino a 3 chassis in un singolo rack (attenzione a peso e consumi)
- Su sistemi high-end avranno solo CPU AMD
- Per networking sponsorizzano molto OPA, consegne celeri
  - Attualmente a 400Gbps, prossimo anno raggiungeranno 800Gbps
- Nessun piano per CPU ARM diverse da Nvidia

Dell - Server

- Per i futuri acquisti (da indagare con AC e chi finanzia): selezionare due o tre vendor e allestire nella zona alta densità alcuni rack con i CDU proprietari.
  - Ogni gara può essere affidata ad uno tra questi popolando di server la zona predisposta
- Sui server DLC resta comunque 1/2% per ventole. A tendere si va a 0%
- L'attuale generazione di server con GPU SXM sarà l'ultima disponibile ad aria
  - Prevedono che entro il 2030 si arriverà a rack da 2MW
- 17^ generazione
  - 1U - 2 socket - 1 nodo – limitate SKU - rumoroso per via della dimensione delle ventole.
  - 2U – 2 socket – 1 nodo - sosterranno il massimo del TDP disponibile
- 18^ generazione da metà 2027
  - 1U solo Intel - 1 socket oppure 2 socket con processori Intel SP, meno performante
  - 2U - 2 socket, Amd disponibili h1-2027, Intel più tardi

Dell - Network

- Propongono SONIC, un sistema operativo di rete basato su debian.
- **Uno dei principali contributori allo sviluppo è Broadcom**
  - Permette di astrarre il livello hardware
  - Scritto per i così detti «white box» ora gira anche su diversi vendor
  - Disponibile sia gratuitamente che a pagamento con maggiori funzioni (che DELL Supporta)
- Sono a catalogo prodotti raffreddati a liquido, compatibili ovviamente con le CDU che propongono per i loro serve
  - Le versioni ad aria sono, come nel caso dei server, più ingombranti



HPE – soluzioni per Data Center

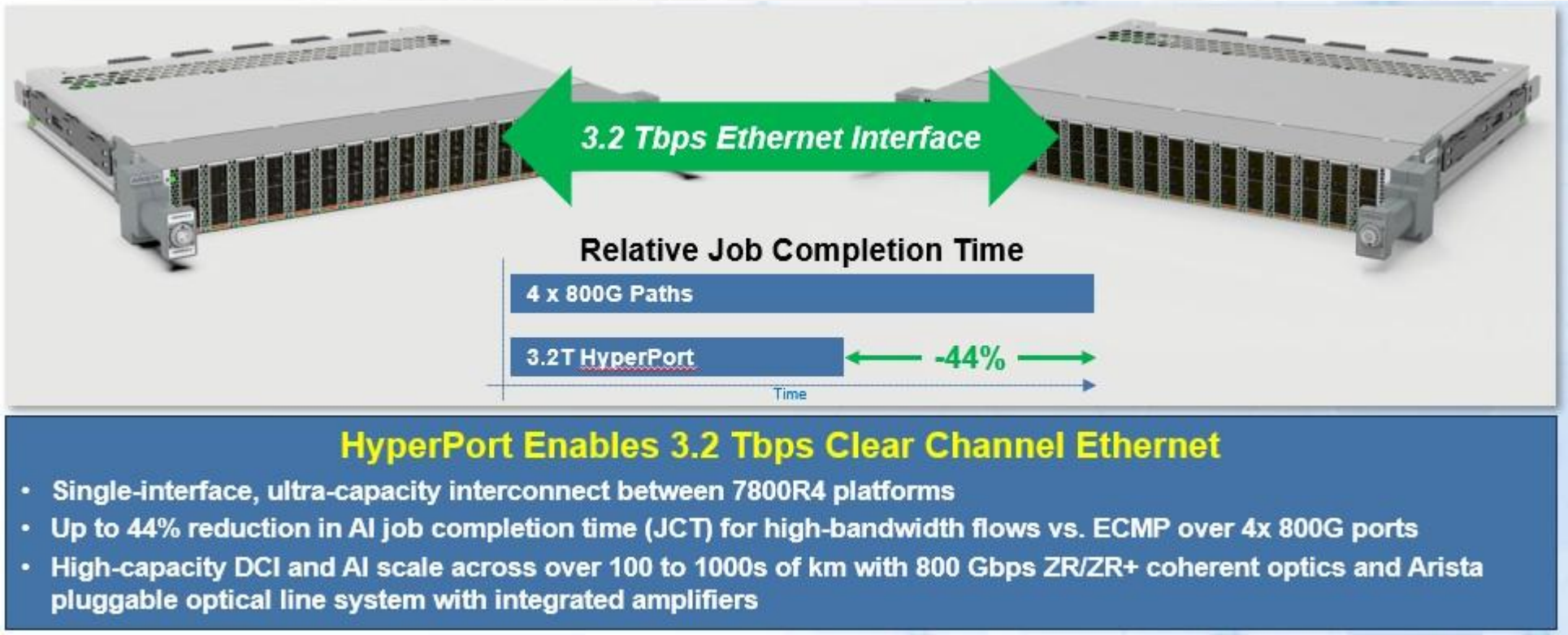
- Paolo Masera, vecchia conoscenza dai tempi di PBS
- Abbiamo discusso brevemente di futuro del Data Center
  - HPE ha grande esperienza grazie alle commesse di grosse soluzioni
- Prima della gara, necessario fare una sessione con ingegneri specializzati e fare studio di fattibilità
- Si va verso la consegna di rack pre-assemblati in sede (anche Dell ne parla)
  - Conseguenti complicazioni dal punto di vista logistico per trasporto e messa in posa
- Ci faranno avere una guida ufficiale HPE su pre-requisiti indispensabili per l'infrastruttura del Data Center

ARISTA – Evoluzione architettura

- Incontro con Darrin Thomson
- Famiglia R4 (800Gbit Ethernet) in vendita
  - schede 36x800G disponibili per le soluzioni modulari come i nostri (fino a 16 slot) (Non retro compatibile ma integrabile in mesh mista, volendo in futuro Trade In per il passaggio a nuova tecnologia)
  - Macchine 2U da 32x800G Acquistabili oggi
  - Macchine 2U da 64x800G a breve
- Q3 2026 saranno disponibili Switch da 64x1,6T (Seguendo la rapida evoluzione dei chipset Broadcom)
  - Le ottiche inseguono più lentamente. All'inizio probabilmente saranno solo breakout
  - 1600G-2XDR4 --> SMF Dual MPO-12 200G PAM4(2Km) --> 2 x 800G-XDR4
  - 1600G-2VSR4--> **MMF** Dual MPO-12 200G PAM4 (**50m**) -->2x 800G-VSR4
  - 1600G-2FR4--> SMF Dual LC 200G PAM4 (2 Km) -->2x 800G-FR4
  - 1600G-2LR4--> SMF Dual LC 200G PAM4 (10Km)-->2x 800G-LR4

ARISTA – Evoluzione architettura

- Hyper port da 3.2Tbps per l'interconnessione fra switch più efficiente di ECMP: 44% JCT **(Job Completion Time)** in quanto l'insieme delle porte viene visto come una unica porta logica.



NOKIA

- Datacenter switch basati su Chipset Broadcom (Ma ha anche linee di apparati basati su Chipset proprietari)
  - Macchine 64x400G acquistabili oggi -->64x800G Q1-Q2 2026
- SR-LINUX (Open Extensible NOS basato su kernel Linux)
- EDA (Event Driven Automation) Sistema di gestione (Possibile POC?)
  - Creazione di configurazioni di una intera fabric dalla GUI
    - Possibile creare fabric multivendor (...veramente curioso di vederlo funzionare)
  - Possibile interrogare la piattaforma usando prompt (come per una piattaforma LLM)
  - Possibile creare un digital twin della rete su cui testare le modifiche prima di applicarle
- DCI e Optical Line System
  - NOKIA ha acquistato Infinera (Transponder che ci collegano direttamente con il CERN)

NETWORKING in generale

- Accelerazione sulla evoluzione dei chipset essenzialmente per l'interconnessione ad alta velocità di GPU
- Broadcom (Integratori: Celestica, HP-Juniper, Nokia, ARISTA, …)
  - Tomahawk 5 (51.6Tbps per CHIP) 64x800Gbps
  - Tomahawk 6 (**102.4 Tbps per CHIP**) 64x1.6Tbps Q3 2026 ?
- Comparsa di soluzioni basate su **Ultra Ethernet (250 ns latency)**
- Cisco Silicon ONE Chipset
  - 64x800Gbps -->Disponibile
  - 64x1.6Tbps in 2027
  - Soluzioni raffreddate a liquido in arrivo

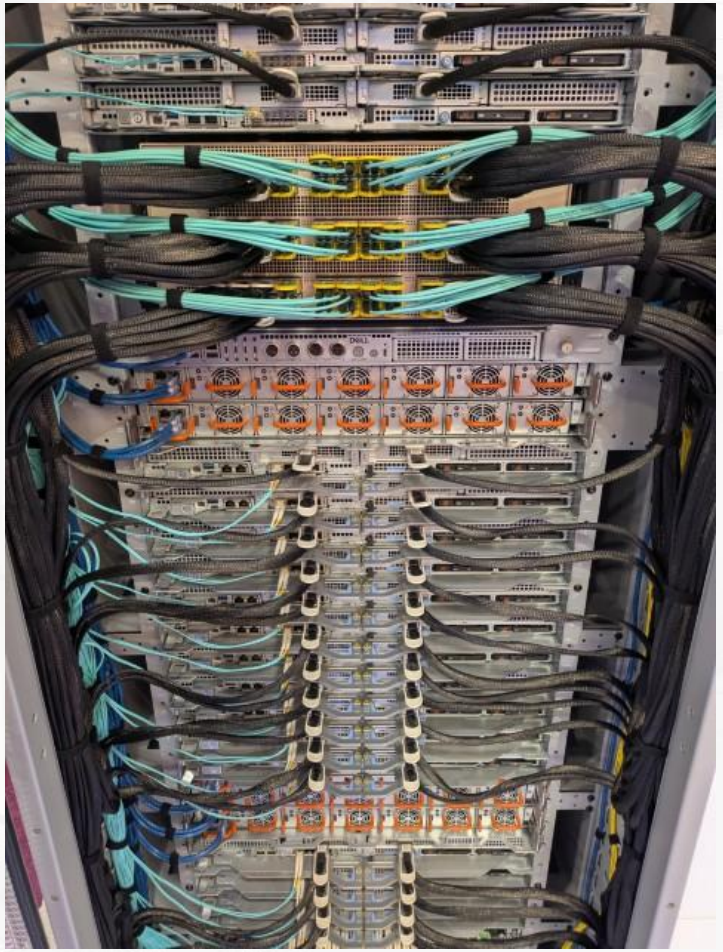
Exhibit floor

- Visitatori registrati: 16.493 (lo scorso anno erano 17.959)
- Vendor registrati: 560 (lo scorso anno erano 500)
  - Grandi assenti: NASA e NIST a causa dello shutdown
  - Tornati tutti i produttori di apparati di rete
  - Omnipath Resuscitato da Cornelis
- Preponderanza di soluzioni a liquido presentate nei vari booth
  - Molti produttori hardware vedono come unica soluzione possibile, per quanto riguarda il raffreddamento a liquido, considerare come unita' minima da fornire, il rack
- Una startup propone di condividere e allocare dinamicamente la RAM via infiniband
  - Rack di ram disponibili in base al bisogno
- RISC-V – non pervenuto

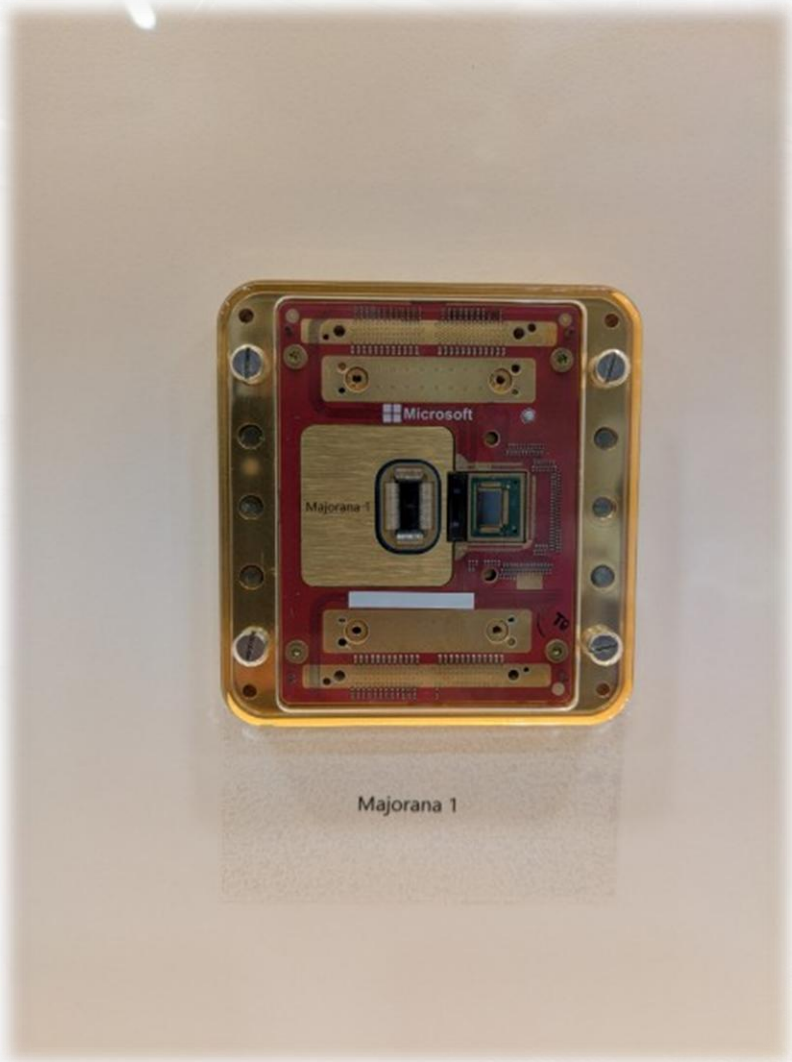

NVIDIA


NVIDIA

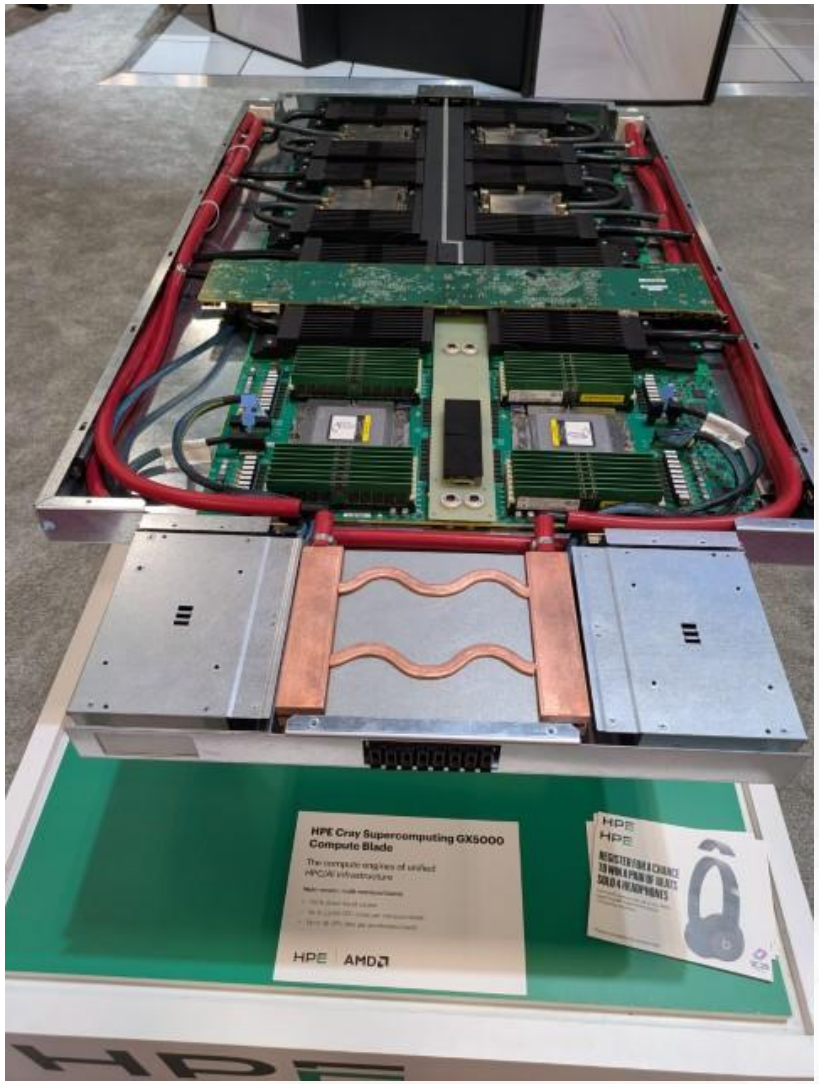

AI Factory
Compute to Cure

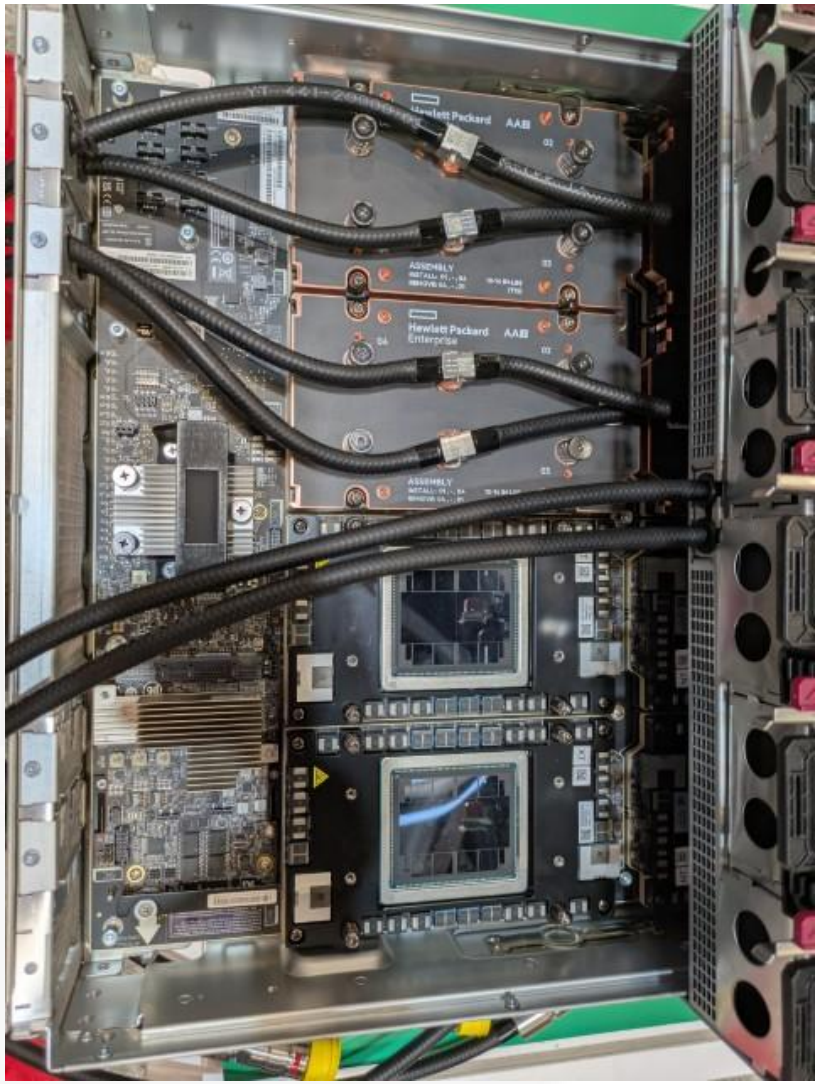



CoolIT
systems


Microsoft
Majorana 1
Majorana 1

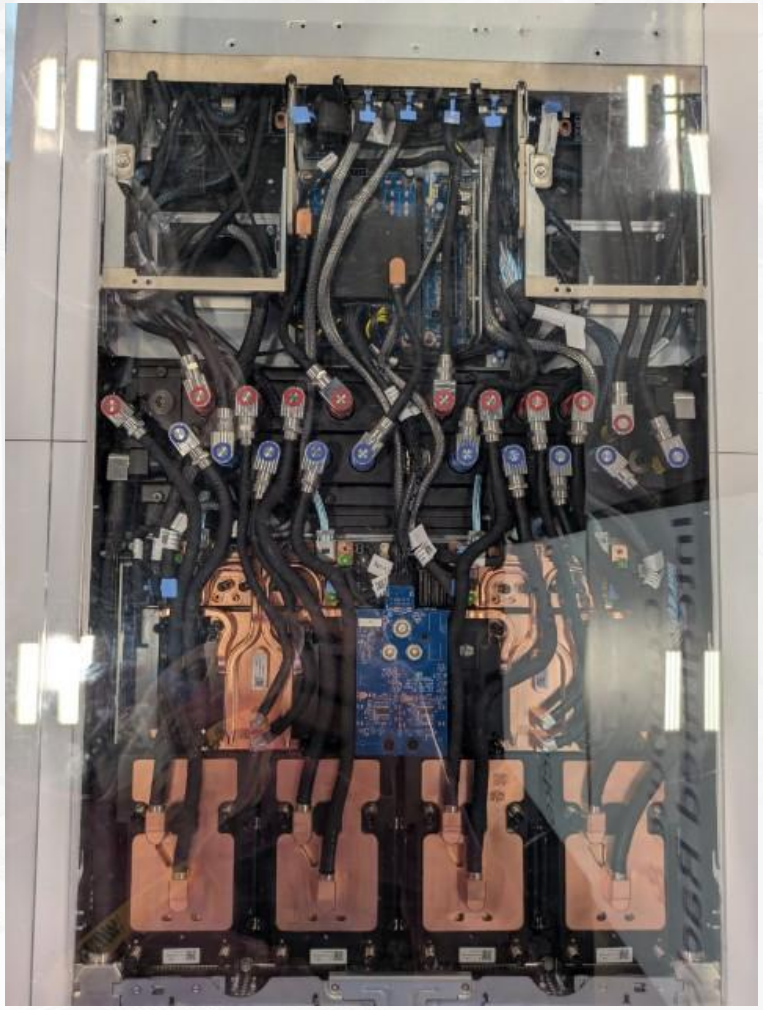

msi
CD270-S4051-X2


HPE Cray Supercomputing GX5000
Compute Blade
HPE
AMD


Hewlett Packard
AAB
Hewlett Packard Enterprise
AAB


HPE Cray


DELL PowerEdge XE9715

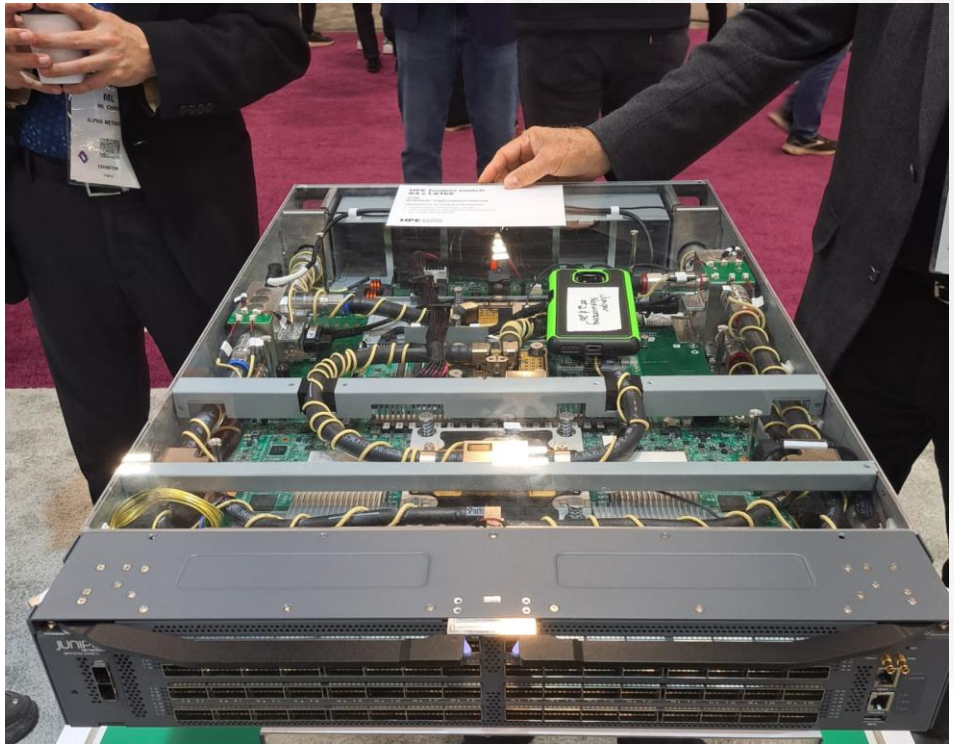



OCP
Project Deschutes
Open CDU
DELTA | Google
Specifications:
DELTA | Google

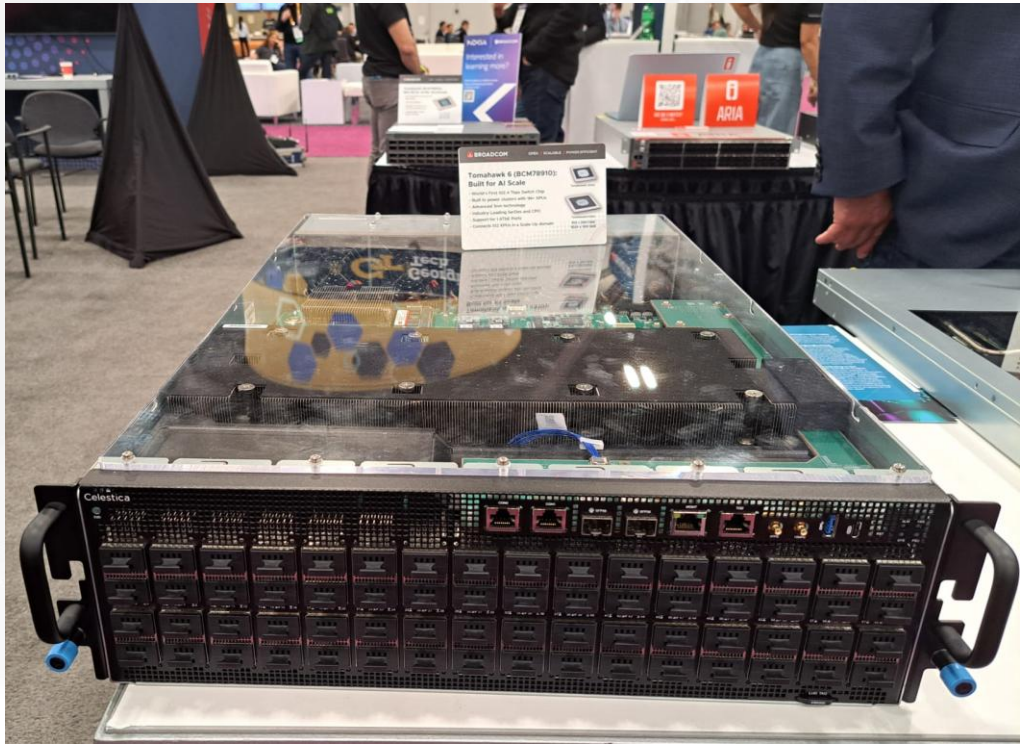

Direct Liquid Cooling
Superior Heat Dissipation
Improved Reliability & Component Lifespan
Greater Energy Efficiency
Greater Performance

Protoripi di switch con porte 1.6Tbps

Juniper 64x1.6Tb Raffreddato a liquido (Prototipo)



3U Celestica 64x1.6Tb OSFP224 Raffreddato ad aria

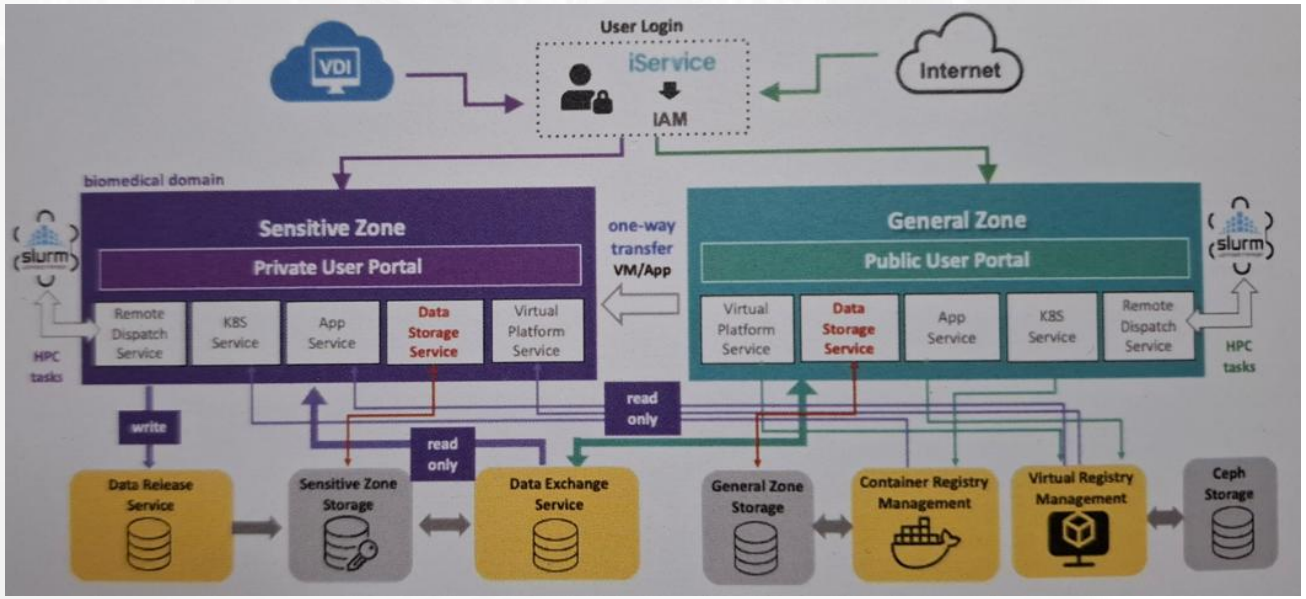


Miscellanea

- SW
  - SLURM
  - Rilascio di Slurm-bridge - https://github.com/SlinkyProject/slurm-bridge
- HW
  - NEXTSILICON
    - MAVERICK-2
      - ASIC, natively supports popular programming languages and frameworks such as C/C++, FORTRAN, OpenMP, and Kokkos, **with upcoming integrations planned for CUDA**, HIP/ROCm, and leading AI frameworks. Many applications can run on Maverick-2 without modification, simplifying porting, eliminating the need for a proprietary software
      - **Maverick-2 sets a new standard for efficient computing with more than a 4x performance-per-watt advantage over traditional GPUs and more than 20x that of high-end CPUs**
    - **RISC-V**
      - Stanno cercando di entrare in EPI, che però non li vuole (hanno BULL)

Miscellanea

- Trusted Cloud-platform
  - Iniziano a vedersi i primi contributi che trattano dati medici, e le relative piattaforme su cui agiscono
  - CHICAGO University
    - ISO27001
    - Piattaforma: web, tipo dashboard datacloud, dove prima selezioni i dati e poi quello che ci vuoi fare
  - National Center for HPC (NCHC, Taiwan)
    - ISO 27001, 27017, 27018, **27701**
    - **Risorse condivise, solo lo storage è separato**

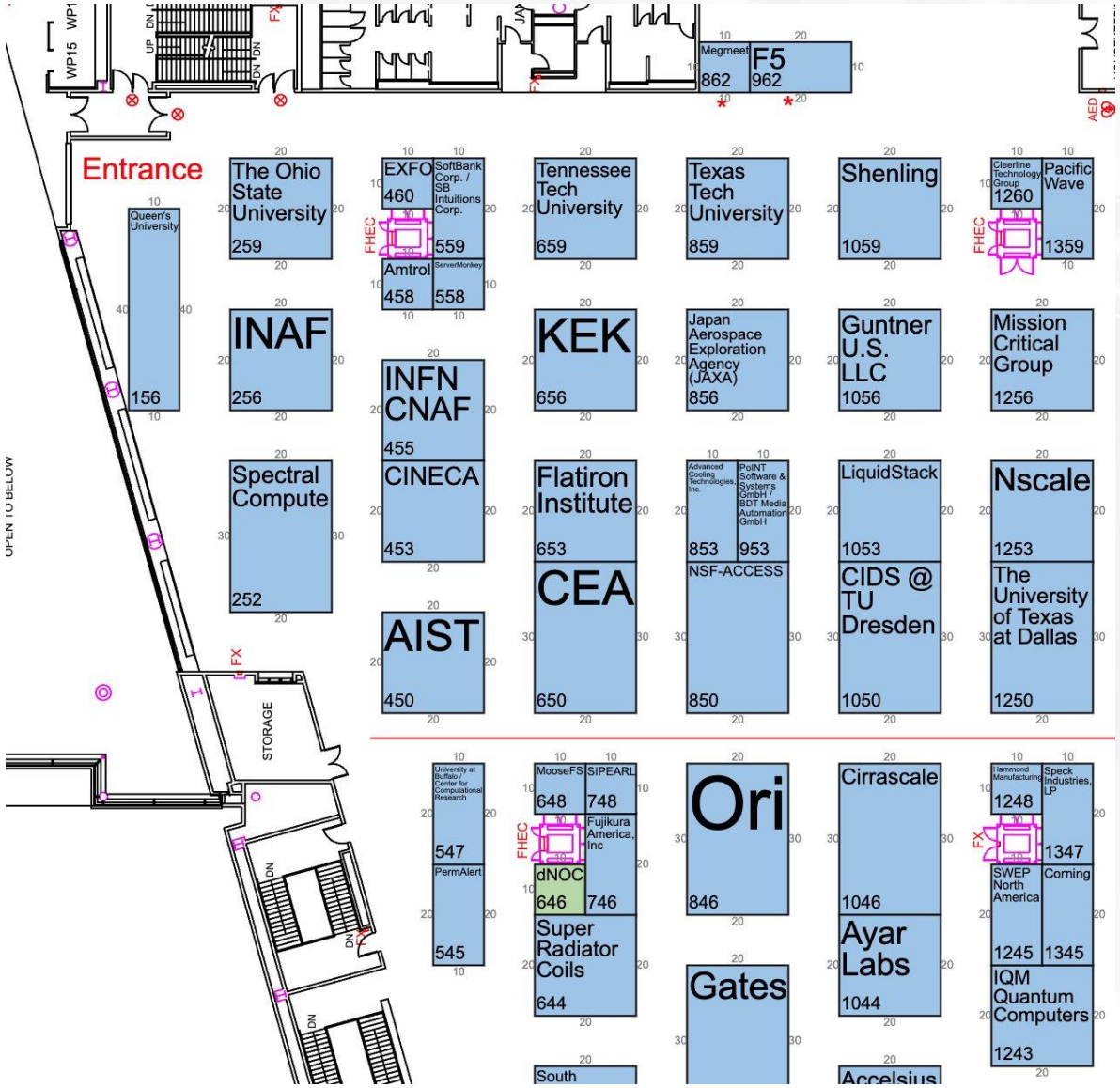


Miscellanea

- Collaborazione con INAF
  - Attività di disseminazione con interventi di 15-20 minuti
  - INFN ha partecipato il seguente contributo
    - **INFN Data processing services for high energy physics and beyond**
      - Attività core
      - Risultati PNRR e consolidamento infrastruttura
      - Progetti

SC26 - Chicago

Il Booth INFN ad SC26



16-19 NOV 2026

SC26

Chicago, IL | hpc unites.