Elettra
Sincrotrone
Trieste

Vi presentiamo Kalculus
Elettra 2.0 HPC Hybrid (CPU/CPU) Cluster

Elettra Sincrotrone Trieste

Acquisizione

- ✓ Il nuovo cluster HPC GPU/CPU, battezzato "Kalculus", è un acquisto effettuato su NADI, budget Elettra 2.0 nell'ambito del WP2 del Progetto Diesel 2.0.
- ✓ Il suo costo è stato di 104.248 Euro IVA esclusa (127.182,56 IVA inclusa), divisi in 7.900 Euro per lo chassis, 8.589 Euro l'una per le 8 lame CPU e 12.370 Euro l'una per le 2 lame GPU. Il rimanente è stato speso per i cavi di collegamento alla rete.
- ✓ L'R.d.A. è stata siglata dal RUP il 08/10/2020 e la procedura di gara (R.d.O. su MEPA) si è conclusa con l'ordine emesso il 12/01/2020, la consegna è stata effettuata il 08/01/2021 e l'installazione a rack l'11/03/2021.

Elettra Sincrotrone Trieste

Le Lame

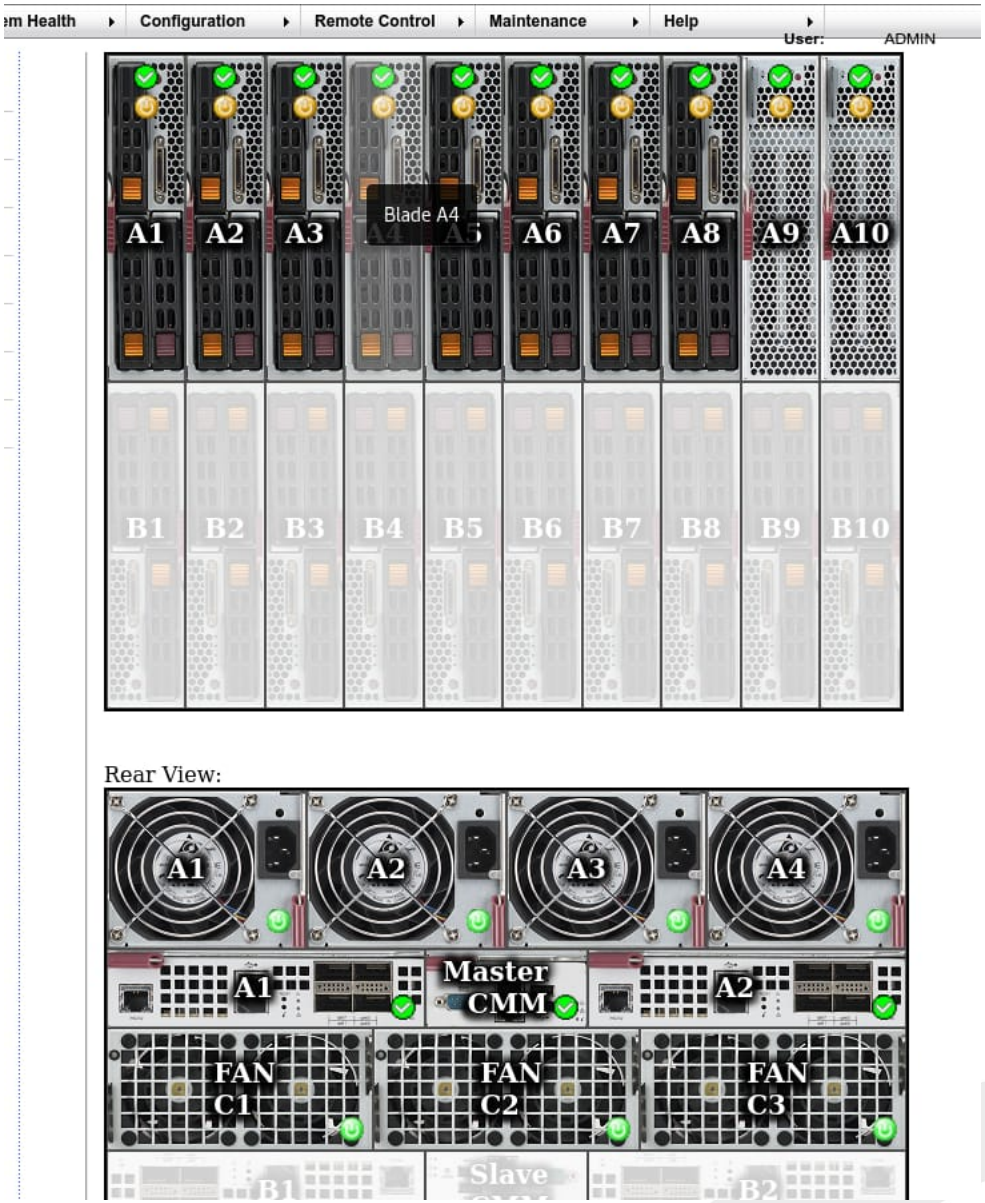
- ✓ Ognuna delle 8 lame CPU è dotata di:
  - 2 processori Intel Xeon Gold 6248R con 24 core (48 thread) a 3.0 GHz (Turbo a 4.0 GHz) e 35.75 MB di cache;
  - 12 moduli di RAM DDR4-2933 da 64 GB l'uno;
  - comunica attraverso 2 porte di rete da 25 Gb/s;
  - consuma al massimo circa 500 W.
- ✓ Ognuna delle 2 lame GPU è dotata di:
  - 1 processore AMD EPYC 7F32 con 8 core (16 thread) a 3.7 GHz (Turbo a 3.9 GHz) e 128 MB cache;
  - 1 processore NVIDIA Tesla A100 con 6912 CUDA core, 432 Tensor core e 40 GB di RAM video "HBM2" ad una velocità di 1.5 TB/s. Si tratta della GPU state-of-the-art al momento dell'acquisto per AI e HPC, con processo produttivo a 7 nm e 54 miliardi di transistor (la generazione precedente ne aveva solo 21 miliardi);
  - 8 moduli di RAM DDR4-3200 da 32 GB l'uno;
  - comunica attraverso 2 porte di rete da 25 Gb/s.
  - consuma al massimo circa 550 W.

Elettra Sincrotrone Trieste

Dati e Management Console Picture

SYSTEM SPECIFICATIONS

	NVIDIA A100 for NVLink		NVIDIA A100 for PCIe
Peak FP64	9.7 TF		9.7 TF
Peak FP64 Tensor Core	19.5 TF		19.5 TF
Peak FP32	19.5 TF		19.5 TF
Tensor Float 32 (TF32)	156 TF | 312 TF*		156 TF | 312 TF*
Peak BFLOAT16 Tensor Core	312 TF | 624 TF*		312 TF | 624 TF*
Peak FP16 Tensor Core	312 TF | 624 TF*		312 TF | 624 TF*
Peak INT8 Tensor Core	624 TOPS | 1,248 TOPS*		624 TOPS | 1,248 TOPS*
Peak INT4 Tensor Core	1,248 TOPS | 2,496 TOPS*		1,248 TOPS | 2,496 TOPS*
GPU Memory	40GB	80GB	40GB
GPU Memory Bandwidth	1,555 GB/s	2,039 GB/s	1,555 GB/s
Interconnect	NVIDIA NVLink 600 GB/s** PCIe Gen4 64 GB/s		NVIDIA NVLink 600 GB/s** PCIe Gen4 64 GB/s
Multi-Instance GPU	Various instance sizes with up to 7 MIGs @ 10 GB		Various instance sizes with up to 7 MIGs @ 5 GB
Form Factor	4/8 SXM on NVIDIA HGX™ A100		PCIe
Max TDP Power	400 W	400 W	250 W



Elettra Sincrotrone Trieste

Caratteristiche generali

- ✓ Nel complesso il sistema ci fornirà quindi:
  - 400 core (800 thread) CPU e 6.5 TB di RAM su 8 lame;
  - 13824 CUDA core, 864 Tensor core e 80GB di memoria GPU HBM2 su 2 lame;
  - interconnessione fra le lame a 25 Gb/s;
  - connessione verso la LAN e gli altri cluster di virtualizzazione e storage a 100 Gb/s.
- ✓ La potenza complessiva assorbita dovrebbe aggirarsi attorno a 6 kW, da misurare una volta in produzione.
- ✓ Il cluster HPC attualmente in produzione è composto dal doppio delle lame e fornisce complessivamente circa la metà delle risorse del cluster nuovo: 252 core (504 thread) per il calcolo CPU e 2 TB di RAM. Non fornisce risorse di calcolo basate su GPU.

Elettra
Sincrotrone
Trieste

Sviluppi

- ✓ Lo chassis può ospitare ancora 10 lame per il calcolo, indifferentemente se CPU o GPU.
- ✓ Il progetto è quello di espanderlo mantenendo la proporzione fra CPU e GPU (8/2), andando ad installare l'hardware più potente reso disponibile dall'evoluzione tecnologica. Ad esempio la GPU con 80 GB di RAM.
- ✓ Il sistema operativo che vi verrà installato permetterà un uso flessibile del cluster HPC, consentendo sia un accesso alle risorse con un sistema "a code" per il calcolo parallelo distribuito, sia il suo utilizzo in modalità "tradizionale" sottomettendo dei processi di calcolo su un singolo nodo (lama).

Elettra
Sincrotrone
Trieste

Alcune foto

Elettra
Sincrotrone
Trieste

Altre foto