EuroHPC PL

EuroHPC PL

Narodowa Infrastruktura Superkomputerowa dla **EuroHPC**

Large-scale computing laboratory - Cyfronet

Marek Magryś

Fundusze Europejskie
Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego

Task: I.1 Supercomputing platform

Task objectives:

- The purpose of the task is to design, build and make available a supercomputing platform for scientific research on solutions to meet the current and future needs of Polish society, the scientific community and at the same time increase the competitiveness and innovation of enterprises by providing them with access to a unique computing infrastructure, enabling them to perform large-scale calculations and achieve a significant acceleration of the research and development process.
- The platform will consist of an infrastructure that will include supercomputers, data storage systems and the necessary system software.

Task: I.2.3 SCM-enabled computing platform

Task objectives:

- The task will provide access to a test platform equipped with storage class memory (SCM), which can be used in future computing systems to achieve two goals: increasing data processing speed and increasing the system's fault tolerance.

Performance measure - FLOPS

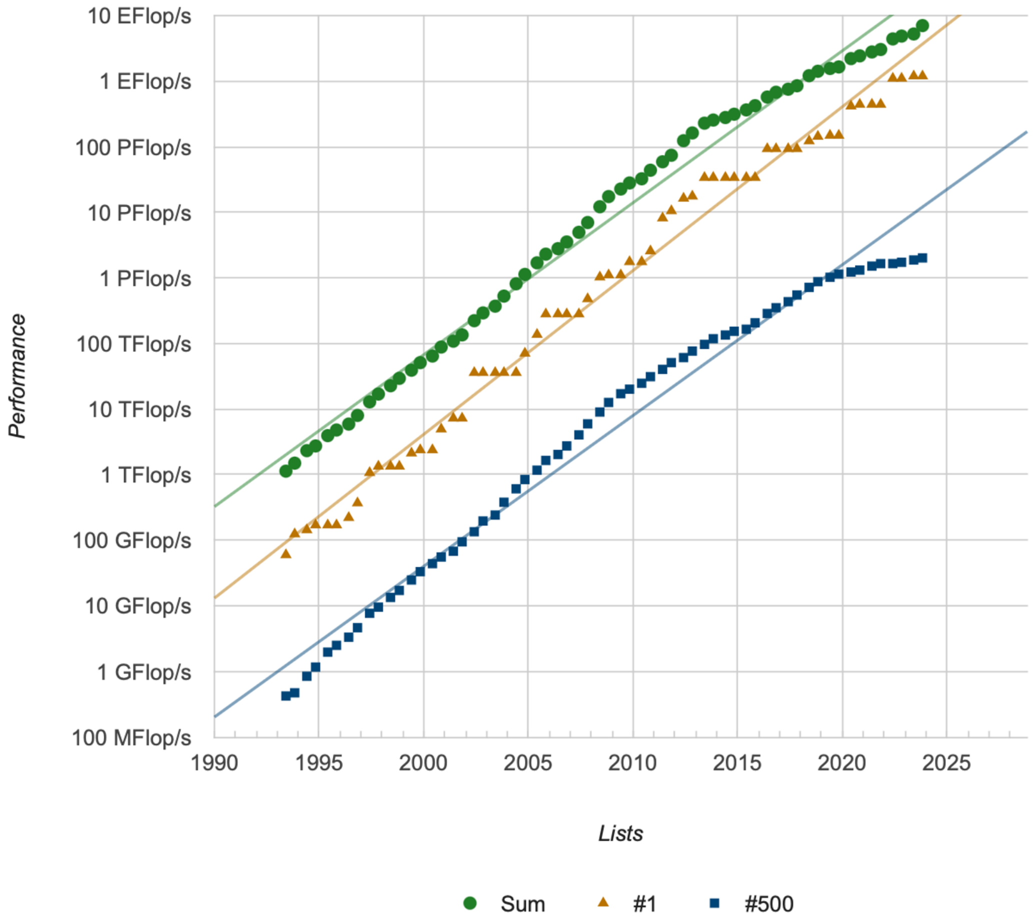
- FLOPS - **FL**oating point **O**perations **P**er **S**econd
- theoretical power - R_{peak}, real power - R_{max}
- 1 FLOPS == 1 add/multiply operation of (64 bit numbers) per second
  - k = kilo = 10^3 = 1 000
  - M = mega = 10^6 = 1 000 000
  - G = giga = 10^9 = 1 000 000 000
  - T = tera = 10^{12} = 1 000 000 000 000
  - P = peta = 10^{15} = 1 000 000 000 000 000
  - E = eksa = 10^{18} = 1 000 000 000 000 000 000
- **Playstation 5: ca. 640 GFLOPS, Apple M2: ca. 1 TFLOPS**

8 * 3 * 16 = 384 GFLOPS

cores * frequency [GHz] * IPC = R_{peak}

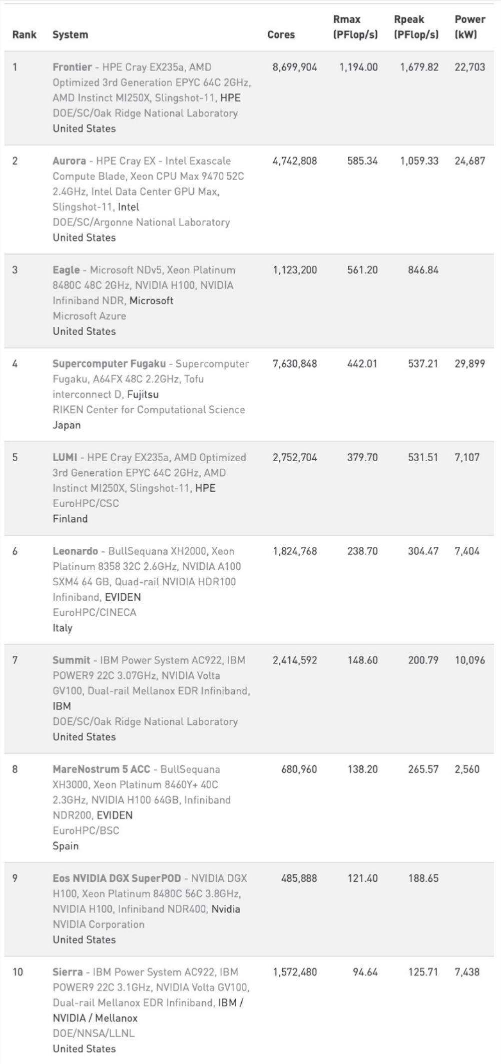
TOP500
The List.

Projected Performance Development



The fastests supercomputers in the world

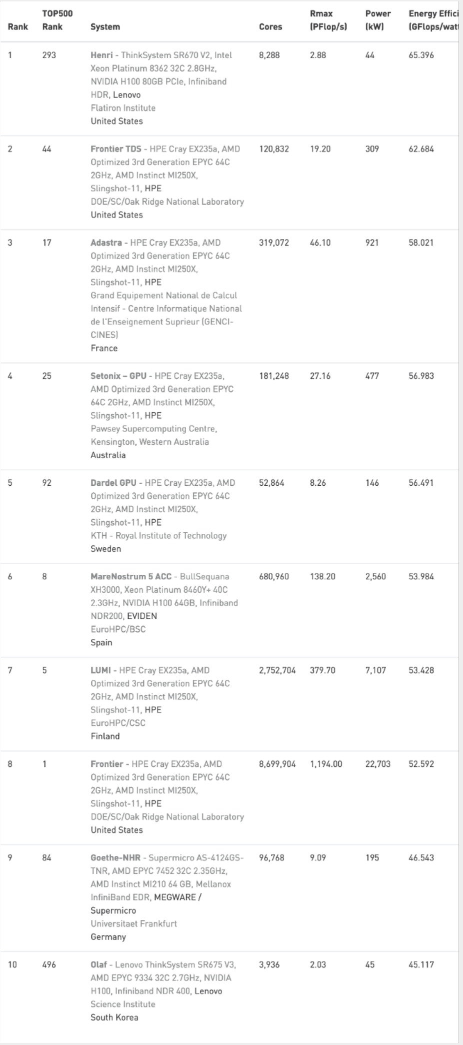
TOP 500
The List.



	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	AMD
🇺🇸	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	intel
🇺🇸	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		NVIDIA
🇯🇵	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	AMD
🇪🇺	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	NVIDIA
🇺🇸	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	NVIDIA
🇪🇺	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	NVIDIA
🇺🇸	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		NVIDIA
🇺🇸	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	NVIDIA

The GREEN 500

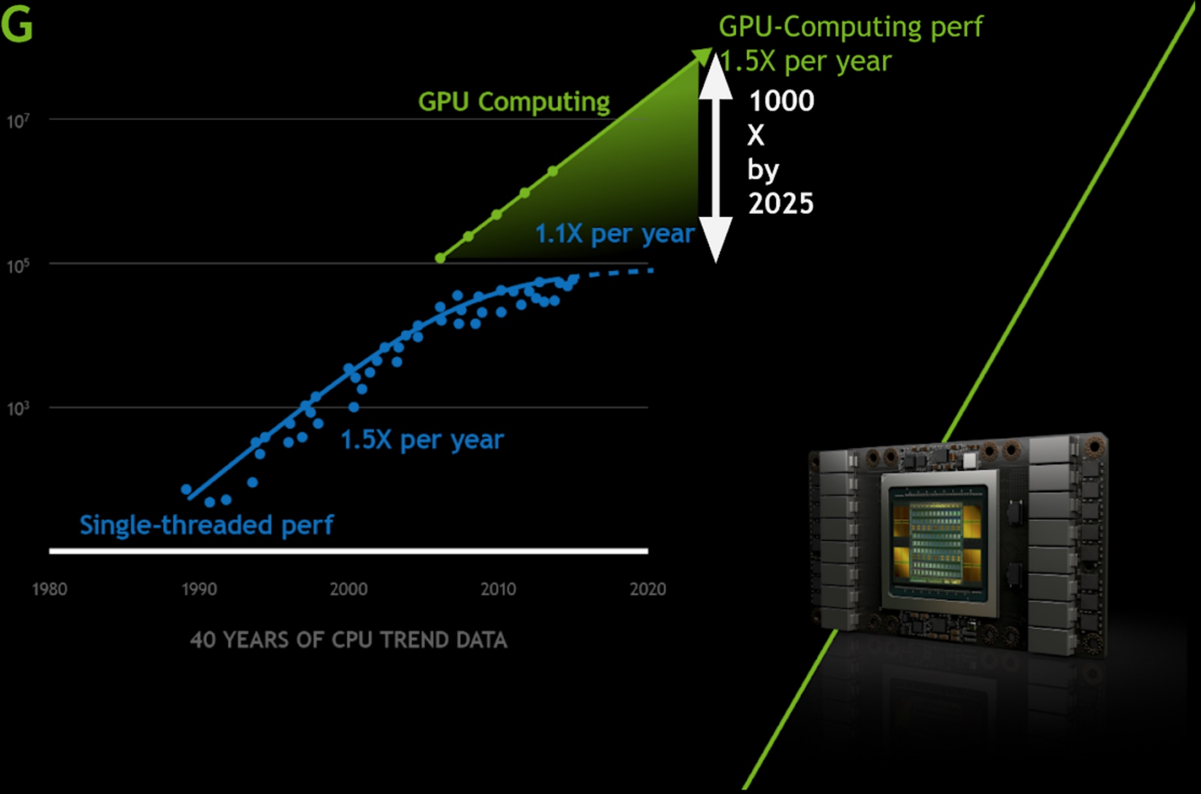
Is computing speed everything ?



Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Effici (GFlops/wat
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

THE TIME HAS COME FOR GPU COMPUTING

For 30 years, the dynamics of Moore's law held true. Microprocessor performance advanced at a rate
of 50 percent per year as more and more transistors were fit onto a single chip. But that approach is hitting the limits of semiconductor physics, and, today, CPU performance only grows by 10 percent per year. NVIDIA GPU computing has given the industry a path forward — and will provide a 1,000X speed-up by 2025.



Source: NVIDIA

Development of HPC over time (recent)

lat 50 1973/2023 CYFRONET

- 2006 - Baribal, 384 GFLOPS
  - large SMP
- 2010 - Zeus, 374 TFLOPS
  - commodity cluster with multiple partitions
- 2015 - Prometheus, 2,4 PFLOPS
  - fastest system in Poland’s history
- 2021 - Ares, 4,0 PFLOPS
  - innovative liquid cooling and heat re-use
- 2022 - Athena, 7,7 PFLOPS
  - fastest system in Poland for HPC and AI
- 2023 - Helios
  - ?

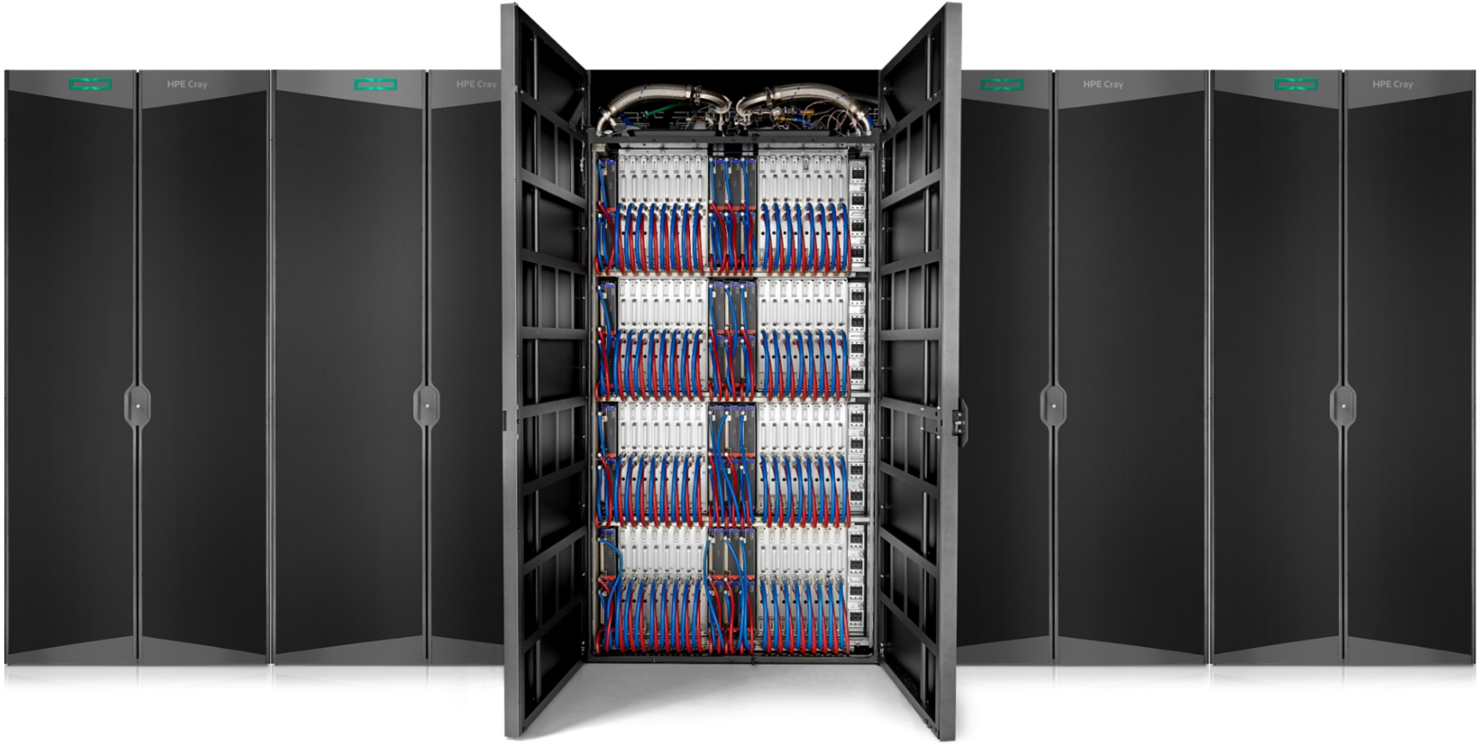
Faeton

288 TFLOPS HPC FP64

- 60 Intel servers
  - 2x Intel Xeon 8352s, 1 TB RAM, 2x 100 GbE
- 4 large memory nodes
  - 2x Intel Xeon 8352s, 1 TB RAM + 8 TB Intel Optane, 2x 100 GbE
- 12 disk nodes
  - 2x Intel Xeon 8352s, 512 GB RAM + 1 TB Intel Optane, 100 TB NVMe
  - DAOS
- OS/Middleware:
  - Rocky Linux 9, Openstack, Openshift
- Main usage:
  - interactive work (pre- and postprocessing)
  - data analytics
  - front-end for HPC/AI services

Faeton CPU

- Intel Xeon Platinium 8352s “Ice Lake”
  - 32 cores (64 threads)
  - 48 MB L3 cache
  - 2,2 GHz (3,4 GHz turbo)
  - 205 W TDP
  - Intel 10nm
  - two CPUs per node
- Memory
  - 8 DDR4 channels
  - Intel Optane support
  - SGX memory enclaves up to 512 GB
  - Faeton memory config:
    - 1 TB base configuration
    - additional 8 TB Optane (4 nodes)
    - up to 300 GB/s
- PCI-Express 4.0



Source: HPE

Announcing HPE Supercomputing Solution for Generative AI with NVIDIA Quad GH200

Hewlett Packard Enterprise

NVIDIA

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

CYFRONET

Los Alamos
NATIONAL LABORATORY

I NCSA

University of BRISTOL

Helios

35
PFLOPS
HPC FP64

1,8
EFLOPS
AI FP8

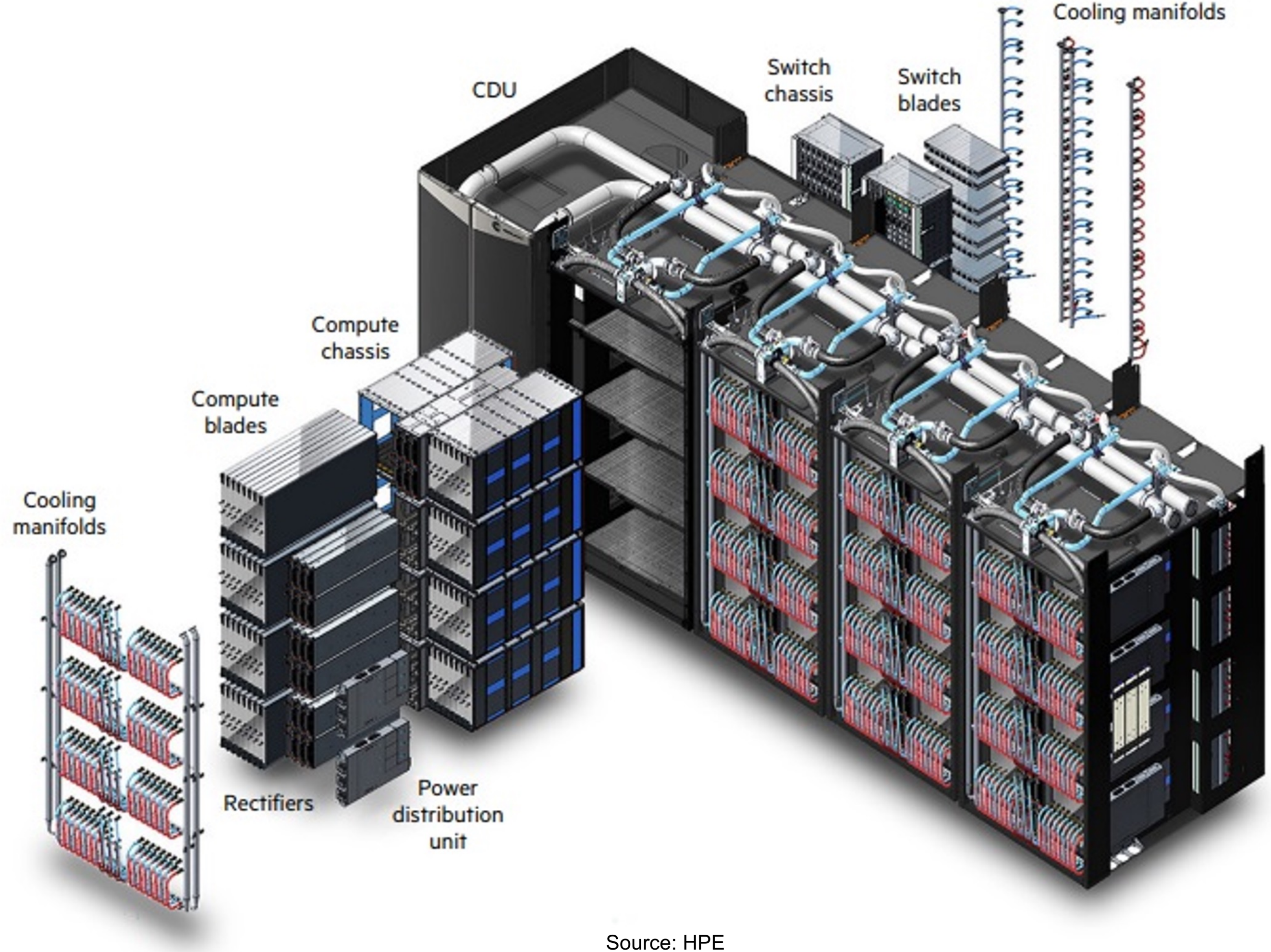
- CPU partition (75264 cores, 196 TB RAM)
  - 272 standard nodes
    - 2x AMD 9654, 384 GB DDR5, 1x Slingshot 200 Gb/s
  - 120 nodes with double memory
    - 2x AMD 9654, 768 GB DDR5, 1x Slingshot 200 Gb/s
- GPU partition (440 accelerator modules)
  - 110 nodes
    - 4x NVIDIA Grace Hopper GH200 CPU+GPU, 4x Slingshot 200 Gb/s
- INT partition (24 accelerators)
  - 6 nodes
    - 2x AMD 9654, 1536 GB DDR5, 4x NVIDIA H100 HGX 94 GB, 30 TB NVMe, 2x Slingshot 200 Gb/s

Helios

- HPE Cray EX4000
  - full direct liquid cooling
    - >99% energy to liquid
    - DLC for servers, switches, power supplies
    - up to 400 kW per rack
    - 3,5 tons per rack
  - Slingshot network
    - 200 Gb/s per port
    - low latency, RDMA, adaptive routing
    - compatible with Ethernet
  - platform used to build world's largest systems
    - Frontier, El Capitan, Aurora, LUMI, ALPS, Shaheen III, Setonix

Source: HPE

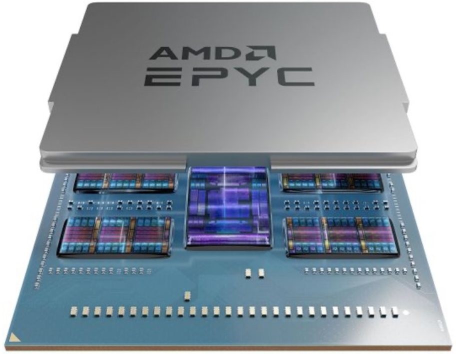
Helios



Source: HPE

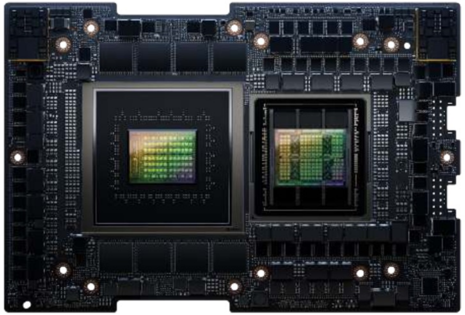
Helios CPU

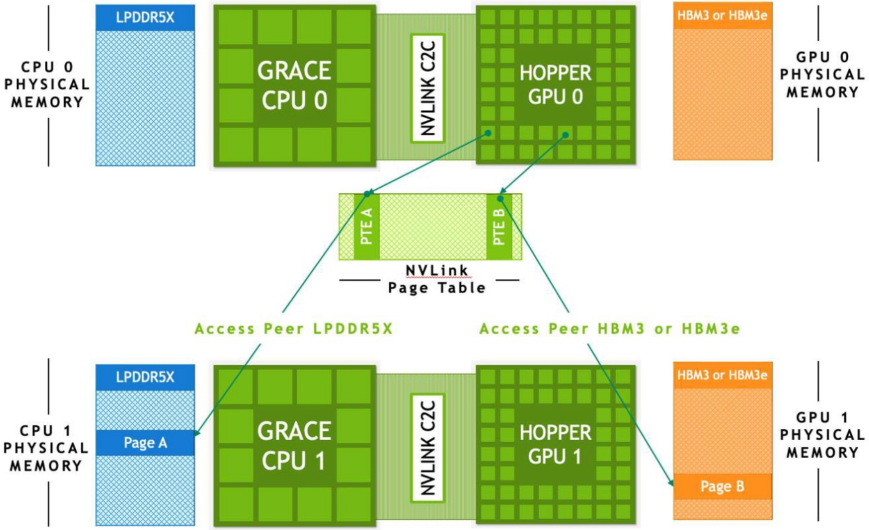
- AMD EPYC “Genoa” 9654
  - 96 cores (192 threads)
  - 384 MB L3 cache
  - 2,4 GHz (3,7 GHz turbo)
  - 360 W TDP
  - TSMC 5nm
  - dual-CPU nodes
- Memory
  - 12 DDR5 channels
  - Helios config:
    - 384 GB or 768 GB
    - do 750 GB/s
- PCI-Express 5.0



Helios GPU

- NVIDIA GH200 “Grace Hopper Superchip”
  - CPU and GPU in a single module
    - 900 GB/s CPU<->GPU
    - coherent memory
    - programmable TDP
    - 1 kW TDP
  - 72 ARM Neoverse v2 cores
    - 128 GB LPDDR5 RAM
    - >500 GB/s
  - Hopper GPU
    - 67 TFLOPS Tensor FP64
    - 3958 TFLOPS Tensor FP8
    - 96 GB HBM3
    - ~ 4 TB/s





Helios - storage

- HPE Cray Clusterstor E1000
  - two Lustre filesystems for /scratch:
    - total 1,5 PB capacity
    - 1800 GB/s reads, 800 GB/s writes
    - 1 M IOPS metadata operations
  - two Lustre filesystems for /project:
    - total 16 PB capacity
    - 185 GB/s reads, 166 GB/s writes
    - S3 gateways for easy integration with data flows
  - independent TDS (Test and Development System)
    - miniature /scratch filesystem
    - 100 TB capacity
- additional NFS servers for software and system tools

Source: HPE

Helios - performance

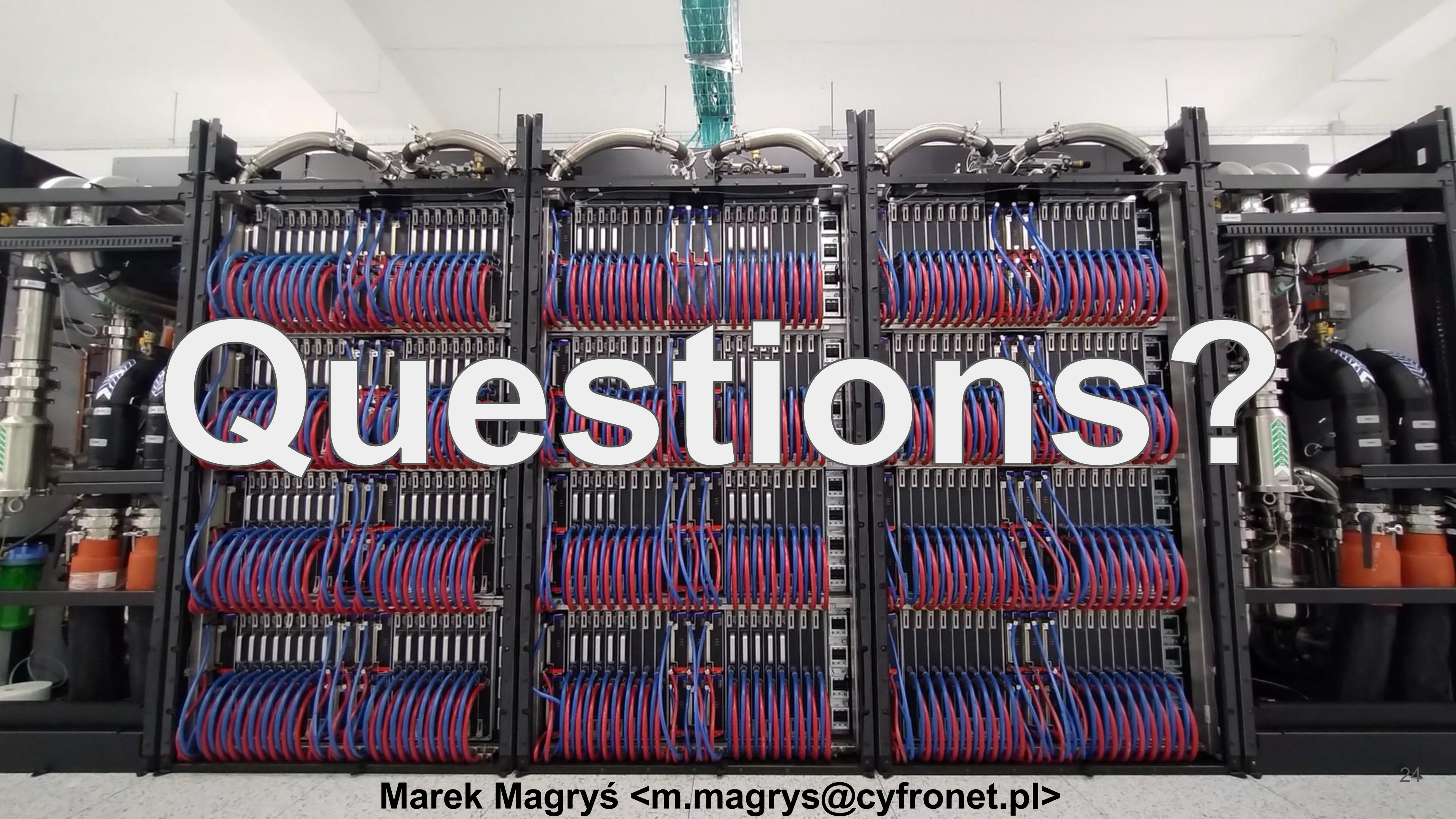
- TOP500 (Linpack benchmark)
  - CPU partition
    - installed in October 2023
    - 2,89 PFLOPS (#291 in November 2023)
    - first tests in HPE factory in Kutná Hora, CZ
    - we got a better result after submission (>3 PFLOPS)
  - GPU partition
    - delivery in December 2023
    - >20 PFLOPS
    - would be in Top40 on current list
    - we target June 2024 for debut
- Green500
  - 60-85 GLOPS/Watt looks doable but requires time

REFRION

REFRION

CYFRONET AGH

Hewlett Packard
Enterprise
CoolIT



Questions?

Marek Magryś <m.magrys@cyfronet.pl>