Computing on Heterogenous Systems with GPU-accelerators

NRIS

Norwegian research infrastructure services

Hicham Agueny
Scientific Computing Group
IT-department, UiB/NRIS

Outline

I. **Hardware topology [CPU vs GPU]** [**GPU**: Graphics Processing Unit]

II. **Overview of GPU-programming models**

III. **Benchmark**

IV. **Overview of NRIS services**

V. **Supercomputer LUMI**

Short introduction

IT-Group, UiB

Scientific Computing Group

Diverse backgrounds

Services

- Monitoring local & *NRIS HPC systems
- Cloud services *NREC, *NIRD
- Training courses in HPC
- Partners in projects
- Projects e.g. building a scientific software stack for HPC
- Consulting
- Help desk

About me
GPU-services

GPU-Graphics Processing Units

Team leader of GPU-team,NRIS

- **Porting codes to GPU**
- **Data analysis**
- **GPU-training courses**
- **GPU-based tutorials**
- **Consulting**

Services are built based on users needs

*NRIS - Norwegian Research Infrastructure Services
*NREC - Norwegian Research and Education Cloud
*NIRD – Norwegian Infrastructure for Research Data

Heterogenous systems

Heterogenous system is a system (HPC system) composed of:

- Different type of hardware and software that are connected.
- Ex. A system with different types of processors .e.g. **CPUs** and **GPUs**.

Advantage:

- Increasing performance
- Scalability

Supercomputer

Cluster

Computer (or node/server)

Performance of a computer

- **The performance of a processor is measured by the quantity:**

FLOPS (**F**loiting-**P**oint of **O**pertaions **P**er **S**econd).

- It is a measure of the speed of a computer to perform arithmetic operations.

For a single processor:

FLOPS = (Clock speed)x(cores)x(FLOPs/cycle)

=Peak performance

FLOP is a way of encoding real numbers (i.e. FP64 or FP32, FP16…)

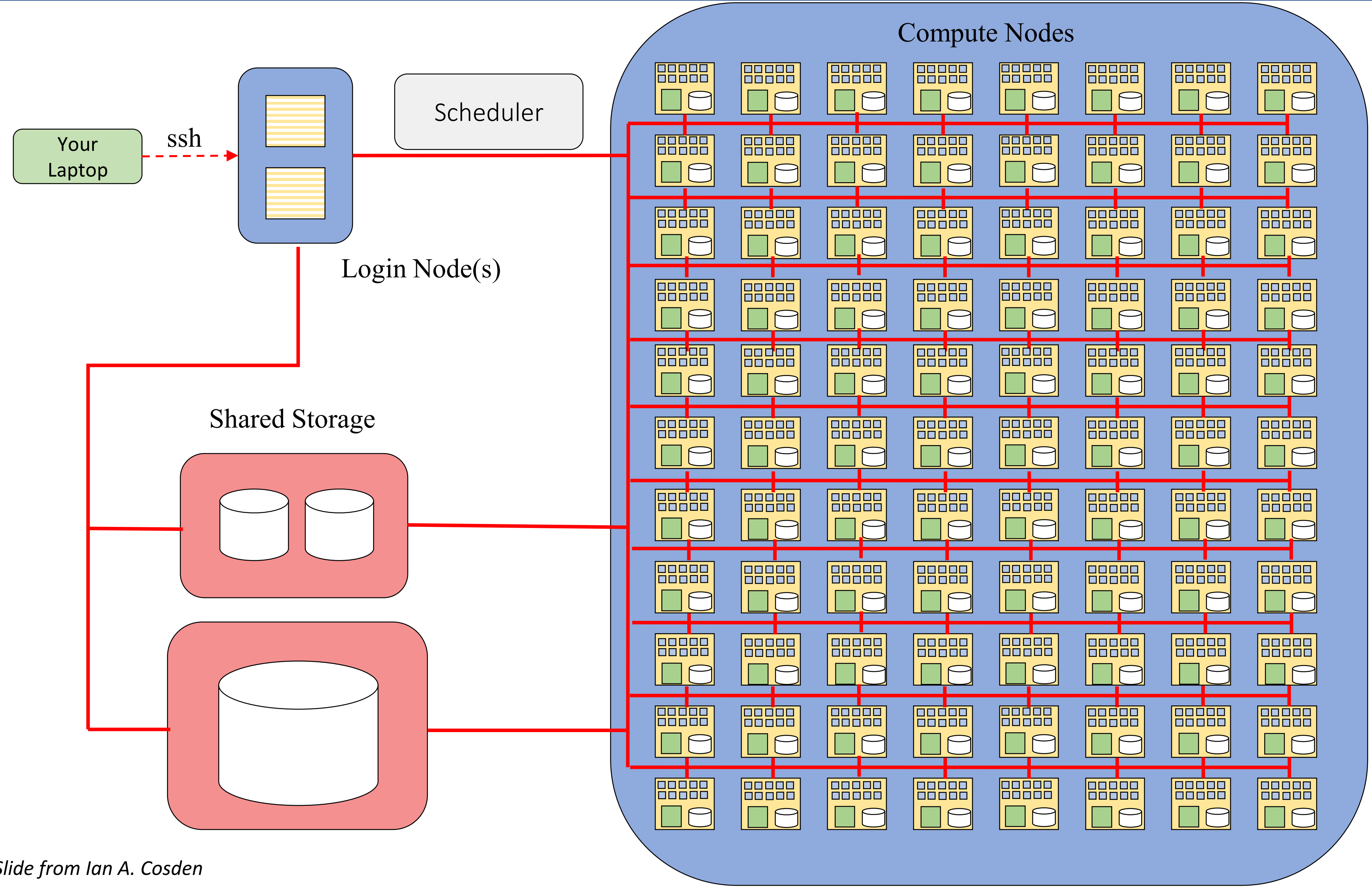
Exp**. 1 PetaFLOPS** = 10^15 calculations per second.

Supercomputer

Cluster

Computer (or node/server)

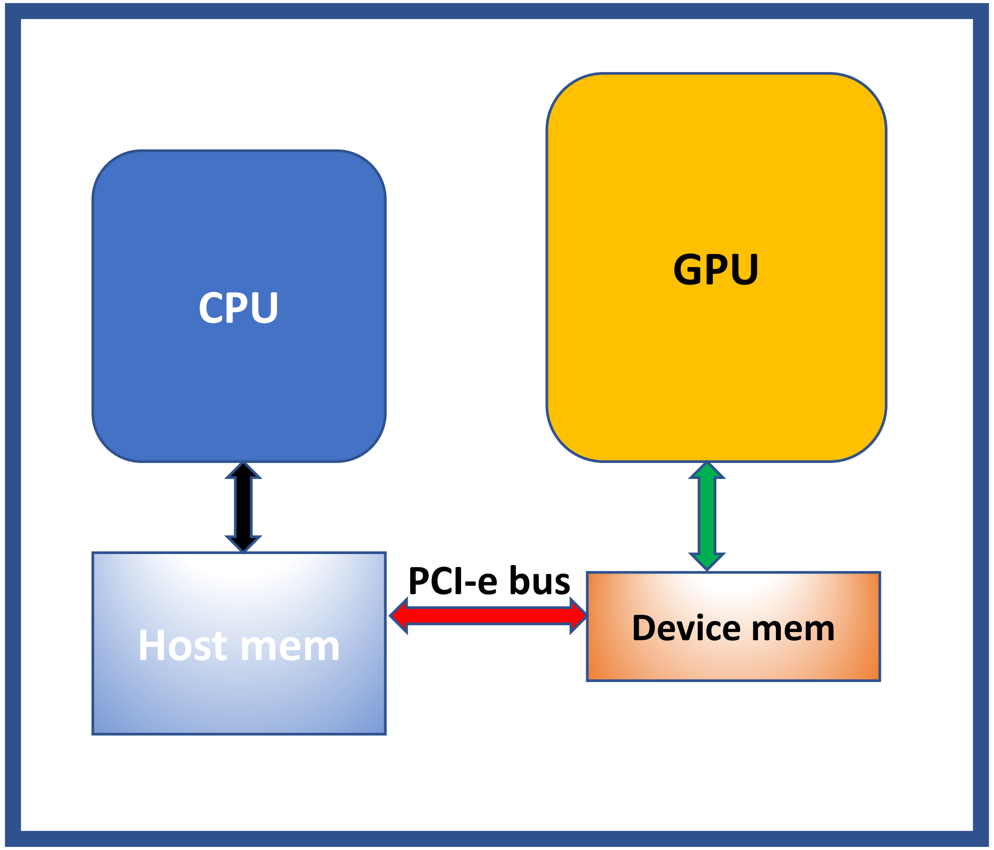
Architecture of Cluster



Slide from Ian A. Cosden

https://indico.cern.ch/event/814979/contributions/3401193/attachments/1831477/3105158/comp_arch_codas_2019.pdf

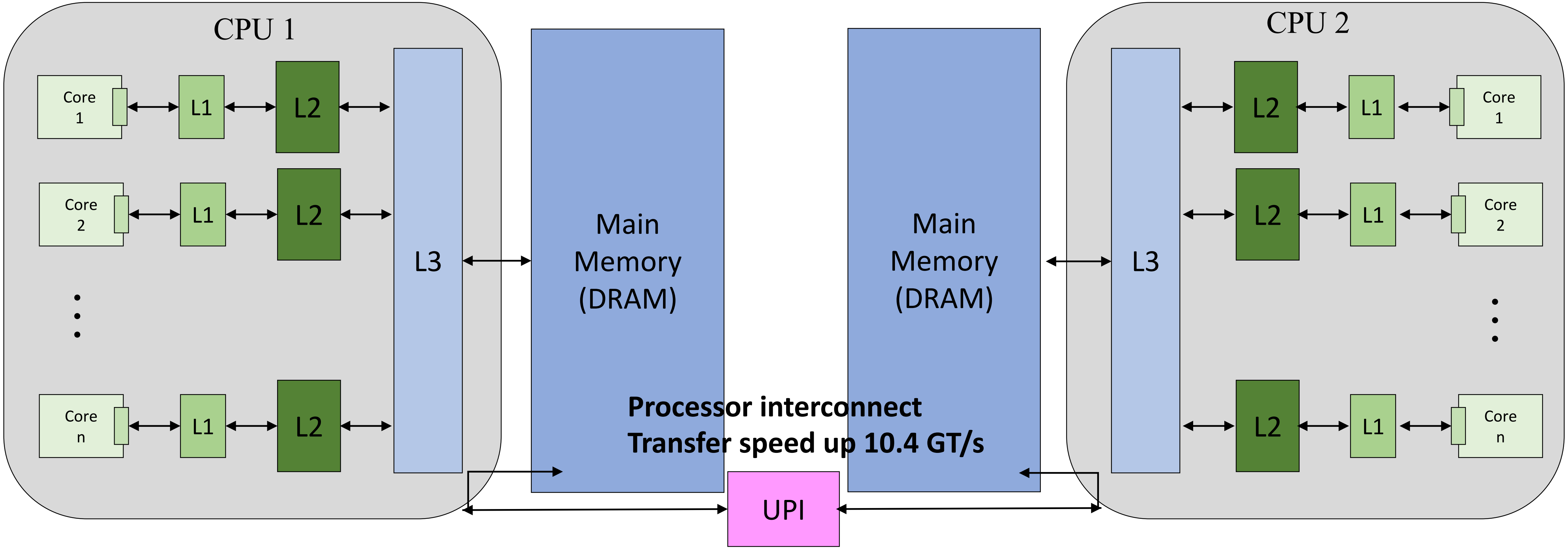
Single node



CPU-Architecture

CPU-Architecture

Dual Socket Intel Xeon CPU



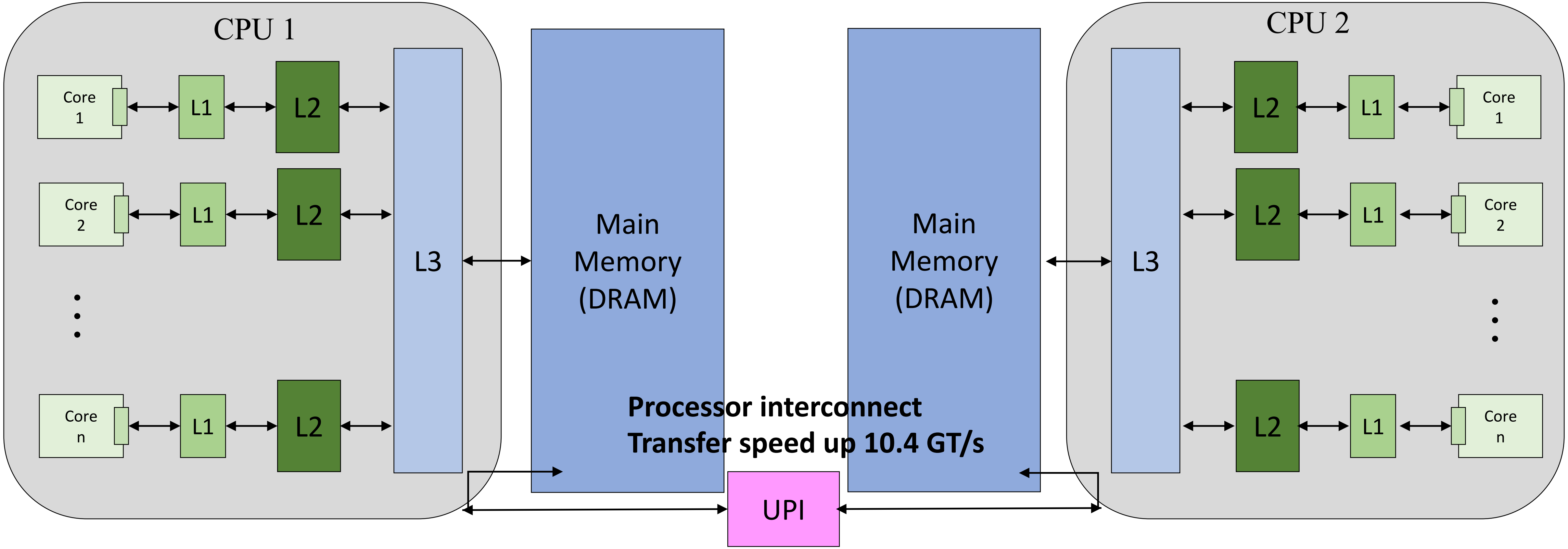
Access L1 Memory is Faster than L2

	Registers	L1 Cache	L2 Cache	L3 Cache	DRAM	Disk
Speed	1 cycle	~4 cycles	~10 cycles	~30 cycles	~200 cycles	10ms
Size	< KB per core	~32 KB per core	~256 KB per core	~35 MB per socket	~100 GB per socket	TB

https://indico.cern.ch/event/814979/contributions/3401193/attachments/1831477/3105158/comp_arch_codas_2019.pdf

CPU-Architecture

Dual Socket Intel Xeon CPU

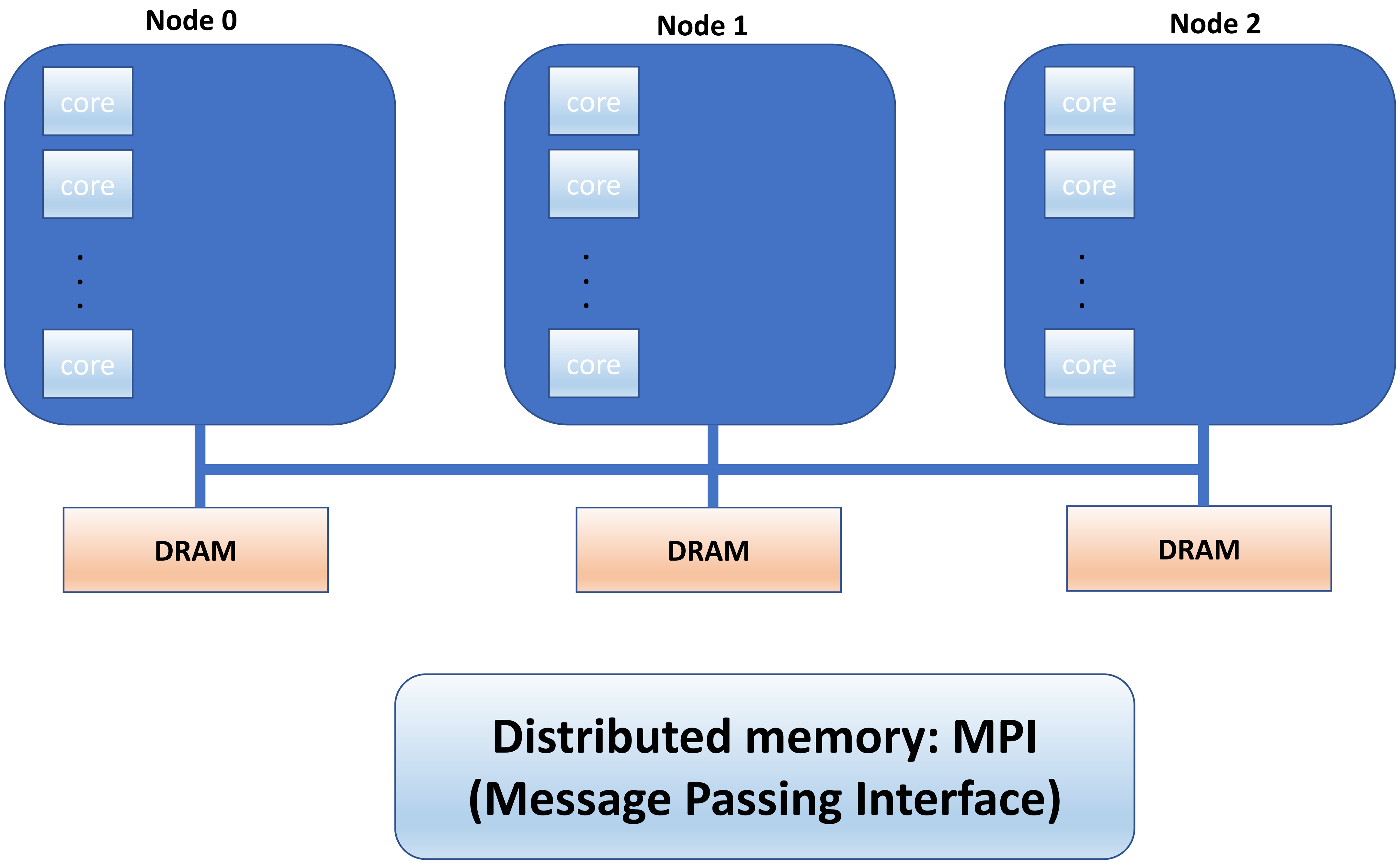


Access L1 Memory is Faster than L2

	Registe	RAM	Disk
Speed	1 cycle	0 cycles	10ms
Size	< KB per core	0 GB per ocket	TB

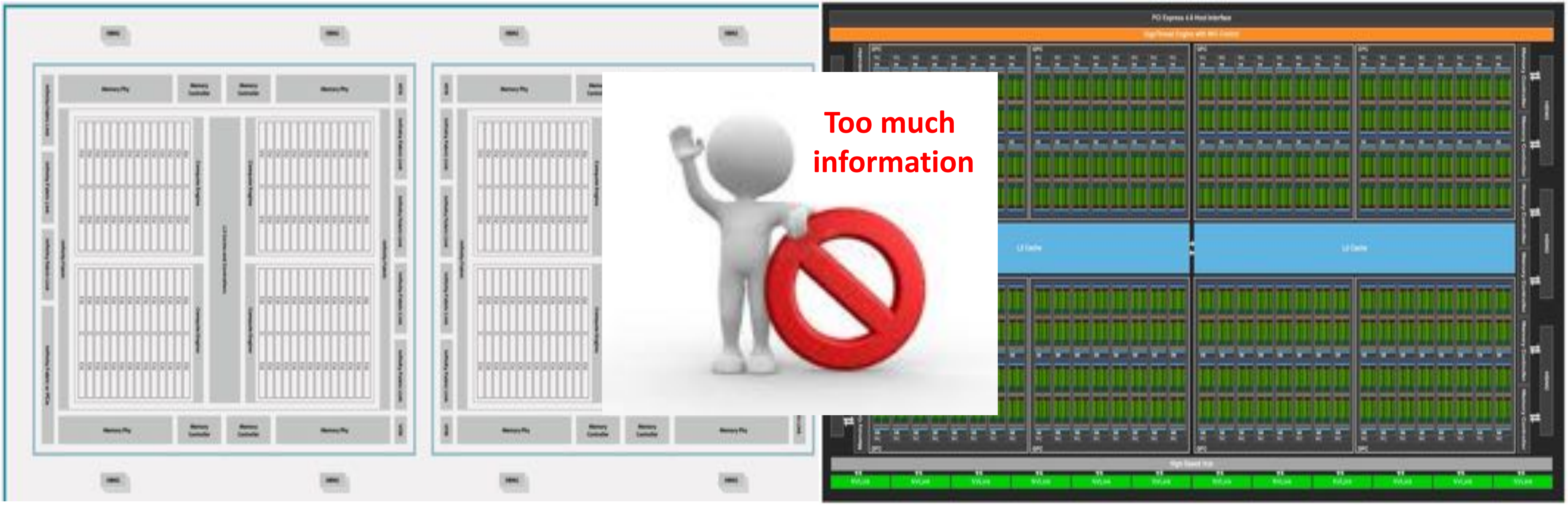
Shared memory: OpenMP

https://indico.cern.ch/event/814979/contributions/3401193/attachments/1831477/3105158/comp_arch_codas_2019.pdf


Node 0
Node 1
Node 2
core
core
core
core
core
core
core
core
core
DRAM
DRAM
DRAM
Distributed memory: MPI
(Message Passing Interface)

GPU-Architecture

GPU-Architecture



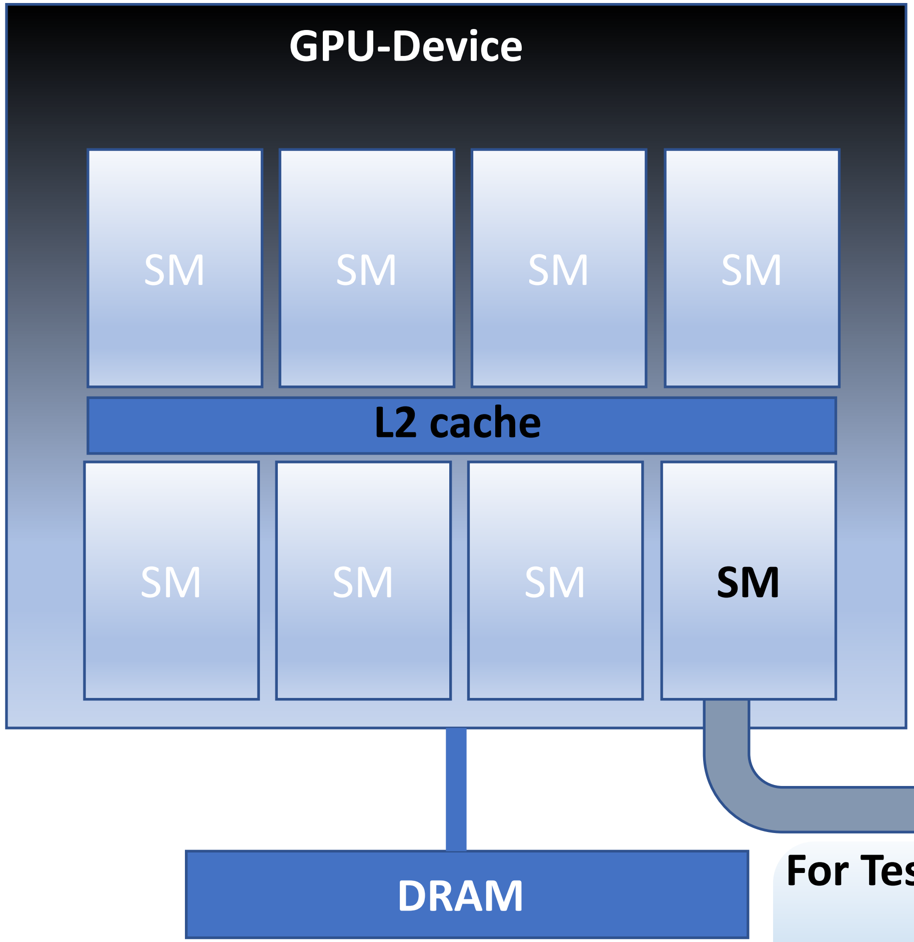
AMD GPU MI250X

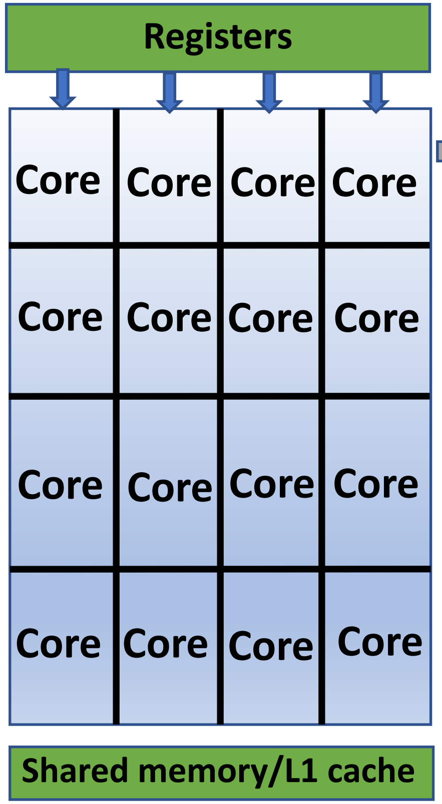
NVIDIA GPU GA100

https://www.amd.com/system/files/documents/amd-cdna2-white-paper.pdf

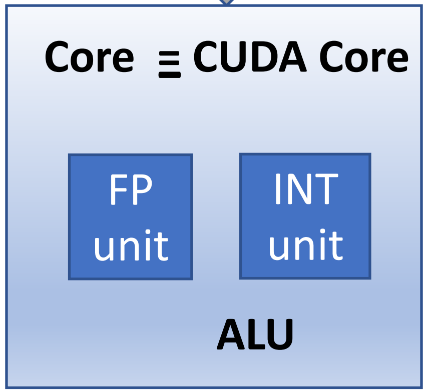
https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

Architecture of NVIDIA-GPU devices (simplified version)





Data are stored temporarily. Registers are private to each thread.



All threads share L1 cache.

Streaming Multiprocessor (SM)

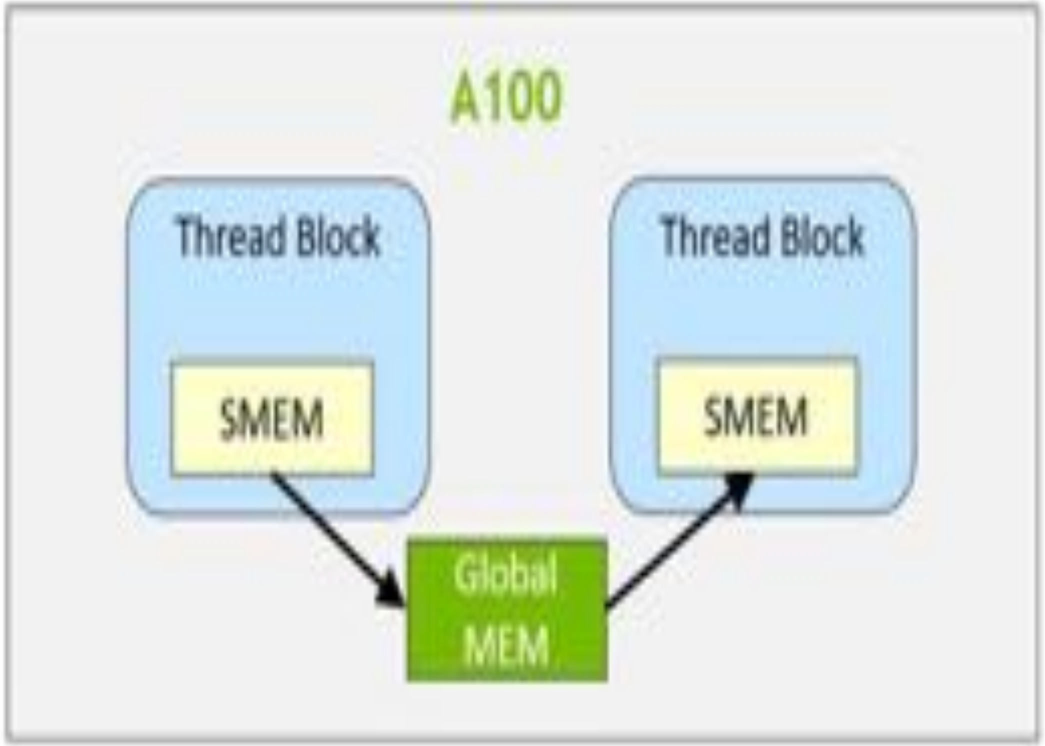
- Core includes ALU. Various logical operations for computations (FP32, FP64) e.g. addition, multiplication and …
- ALU executes SIMD instructions. (processing multiple data with the same operation concurently (in parallel))

For Tesla A100 GPU: 108 SMs, each SM has:

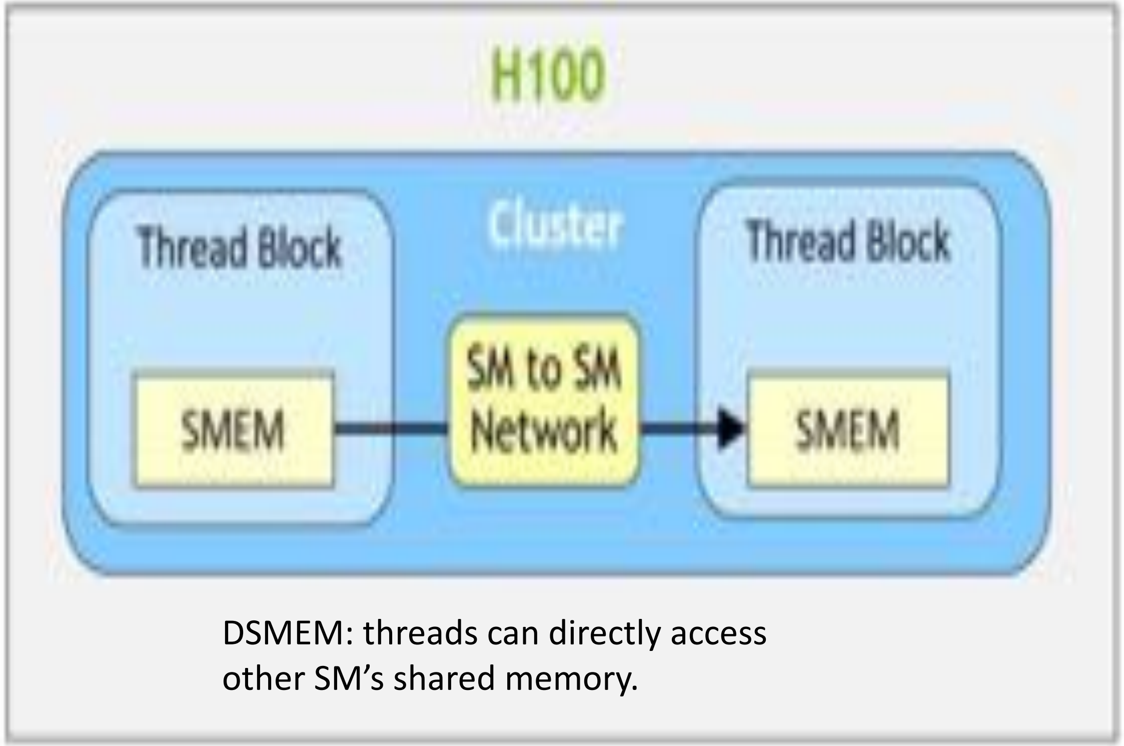
64 FP32 cores ===== A total of 6912 FP32 CUDA cores
64 INT32 cores ===== A total of 6912 INT32 CUDA cores.

SM-to-SM Network

There is a need to access glob mem to pass data.



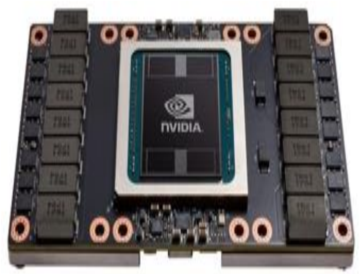
No need to access glob mem to pass data.

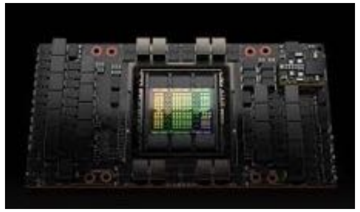


https://resources.nvidia.com/en-us-tensor-core

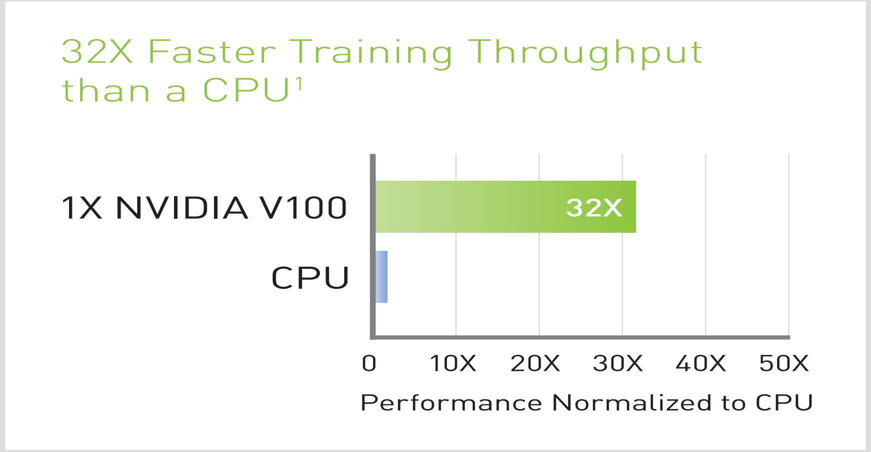
A100 vs H100

NVIDIA GPU charcteristics

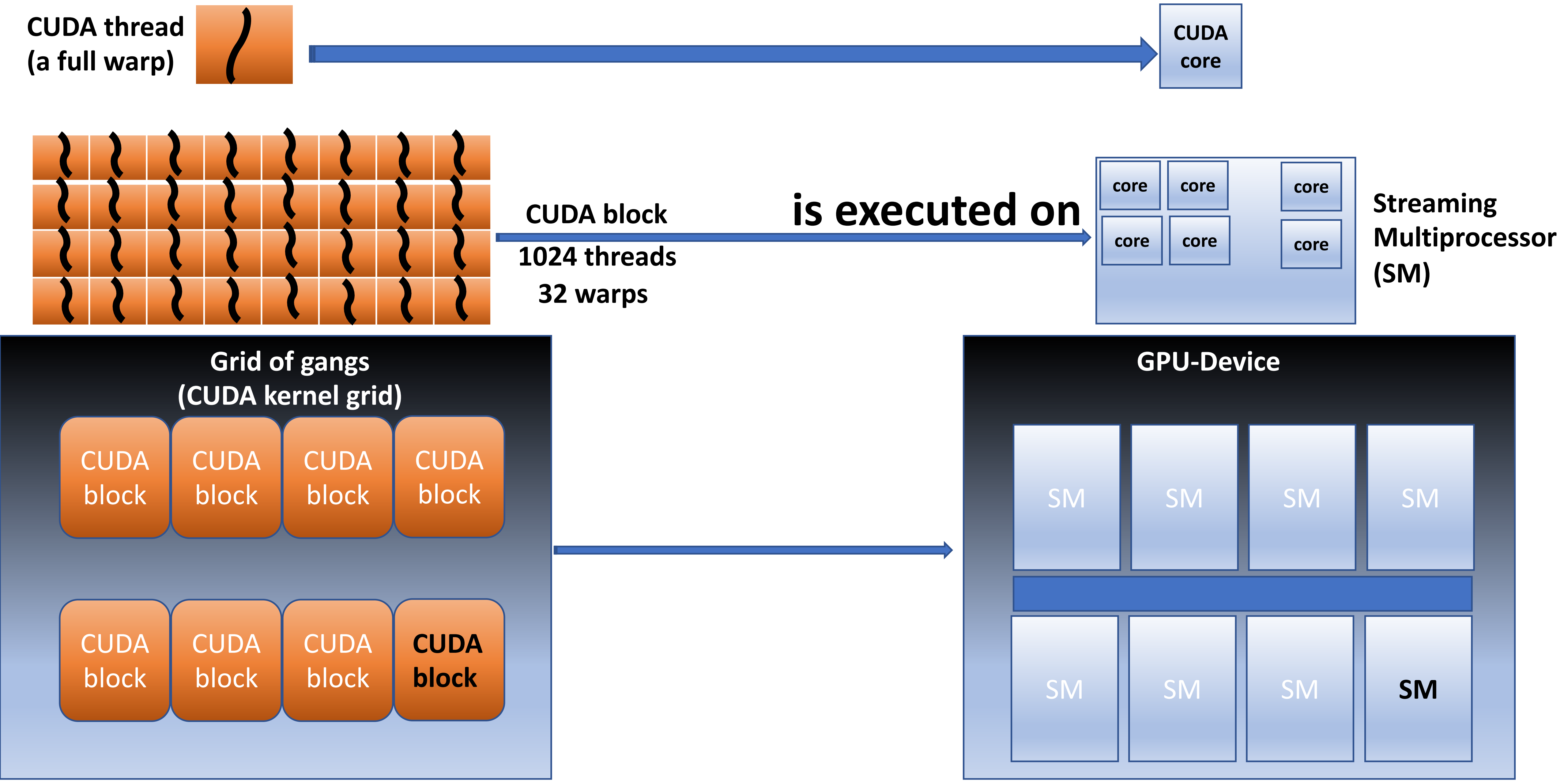




Architecture	NVIDIA P100 (Pascal)	NVIDIA GV100 (Volta)	NVIDIA GA100	NVIDIA GH100
SMs	56	84	128	144
FP32 CUDA cores per SM	64	64	64	128
NVIDIA CUDA cores	3584	5376	8192	18432
Tensor cores/GPU	NA	672	512	576
Peak performance	9.3 TFLOPS	15.7 TFLOPS	39 TFLOPS	66.9 TFLOPS
Transistors	15.3 billion	21.1 billion	54.2 billion	80 billion



Execution on GPU



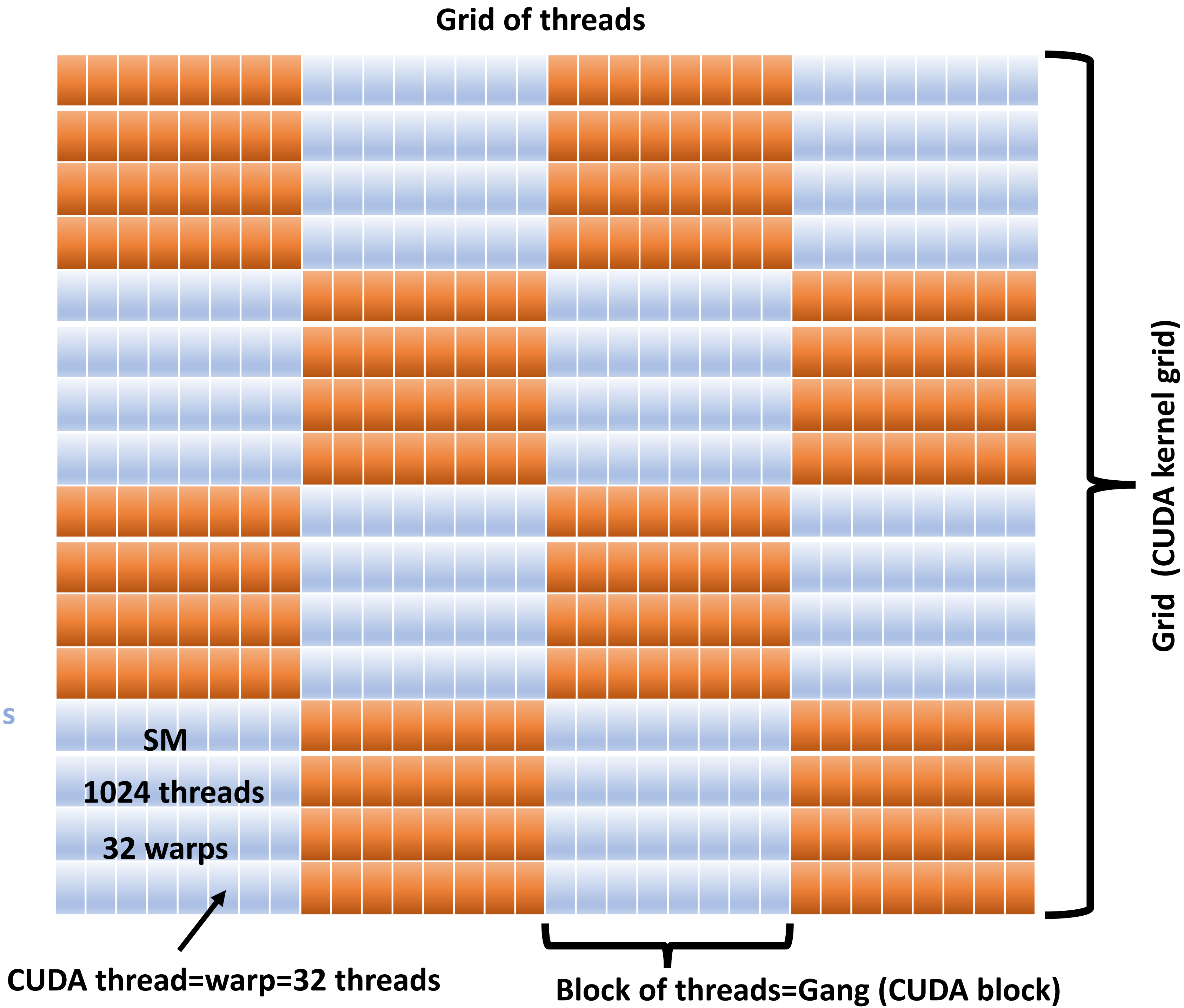
Mapping a matrix into a GPU-device

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

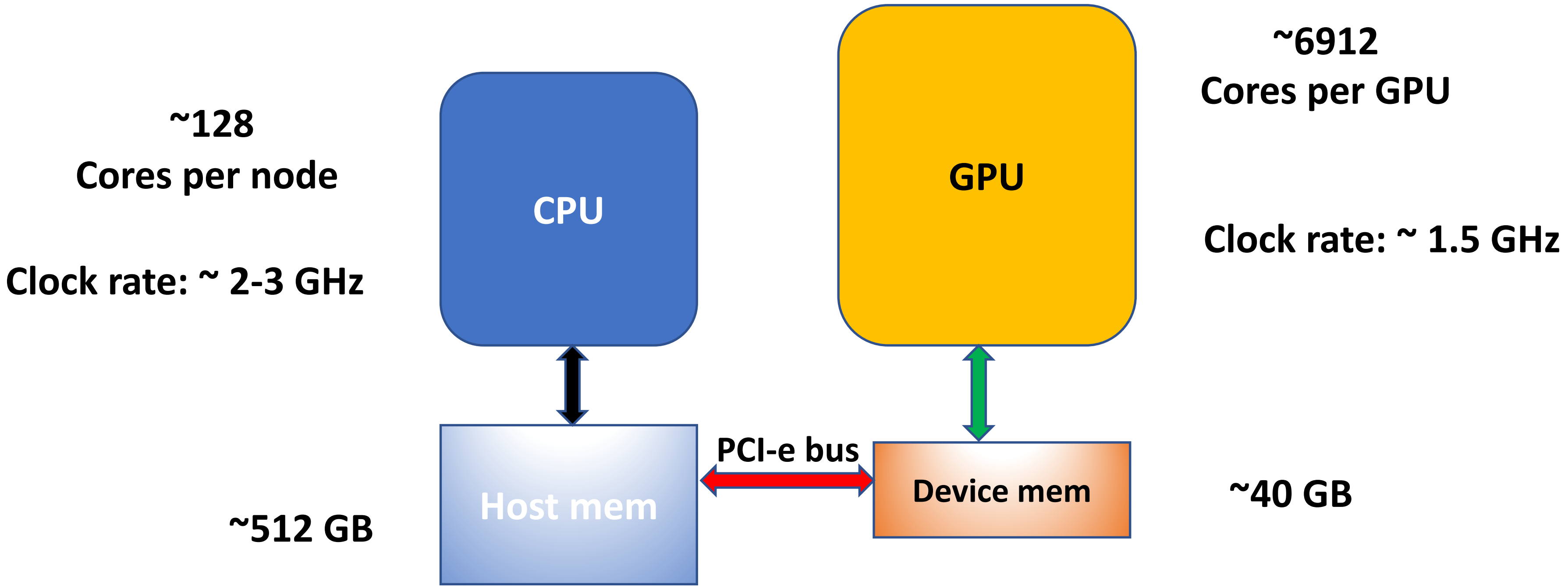
Each thread executes a different piece
Of data item (data parallelism).

Each CUDA block has 1024 threads
(32 warps. A warp=collection of 32 threads
are executed simultaneously by a SM)

Execution on hardware
CUDA thread
CUDA block
CUDA grid



CPU vs GPU



GPUs are designed for:
High troughput
Low latency

Bottelneck:
Data transfer between CPU and GPU
Profiling

Outline

I. Hardware topology [CPU vs GPU] [**GPU**: Graphics Processing Unit]

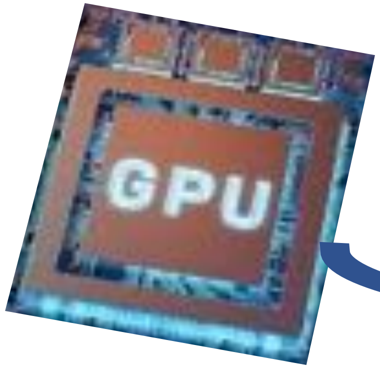
II. Overview of GPU-programming models

III. Benchmark

IV. Overview of NRIS services

V. Supercomputer LUMI

Heterogenous programming models



GPU-Device

CPU-Host

Thousands of cores

GPU: Graphics Processing Unit
CPU: Central Processing Unit

Multiple cores

Programming model

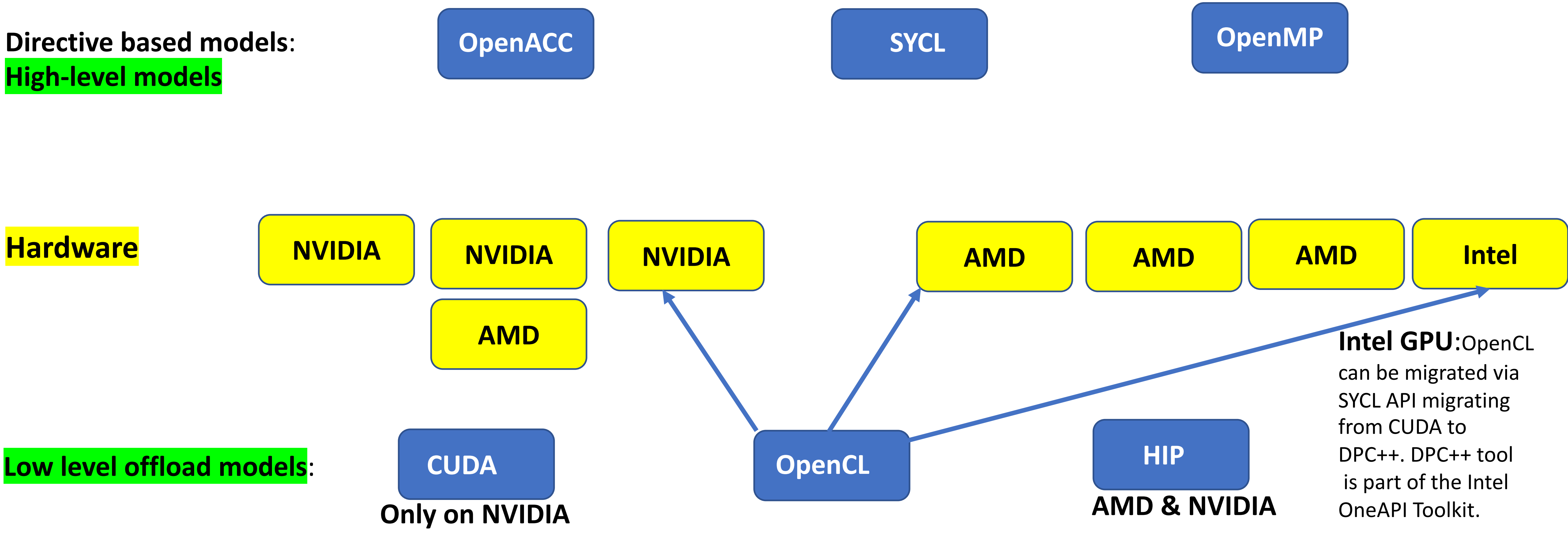
that has GPU-support

To offload a code region to a GPU-device

```
do i=1,n
   do j=1,m
      A(i,j) = B(i,j) + C(i,j)
   enddo
enddo
```

Heterogenous programming models

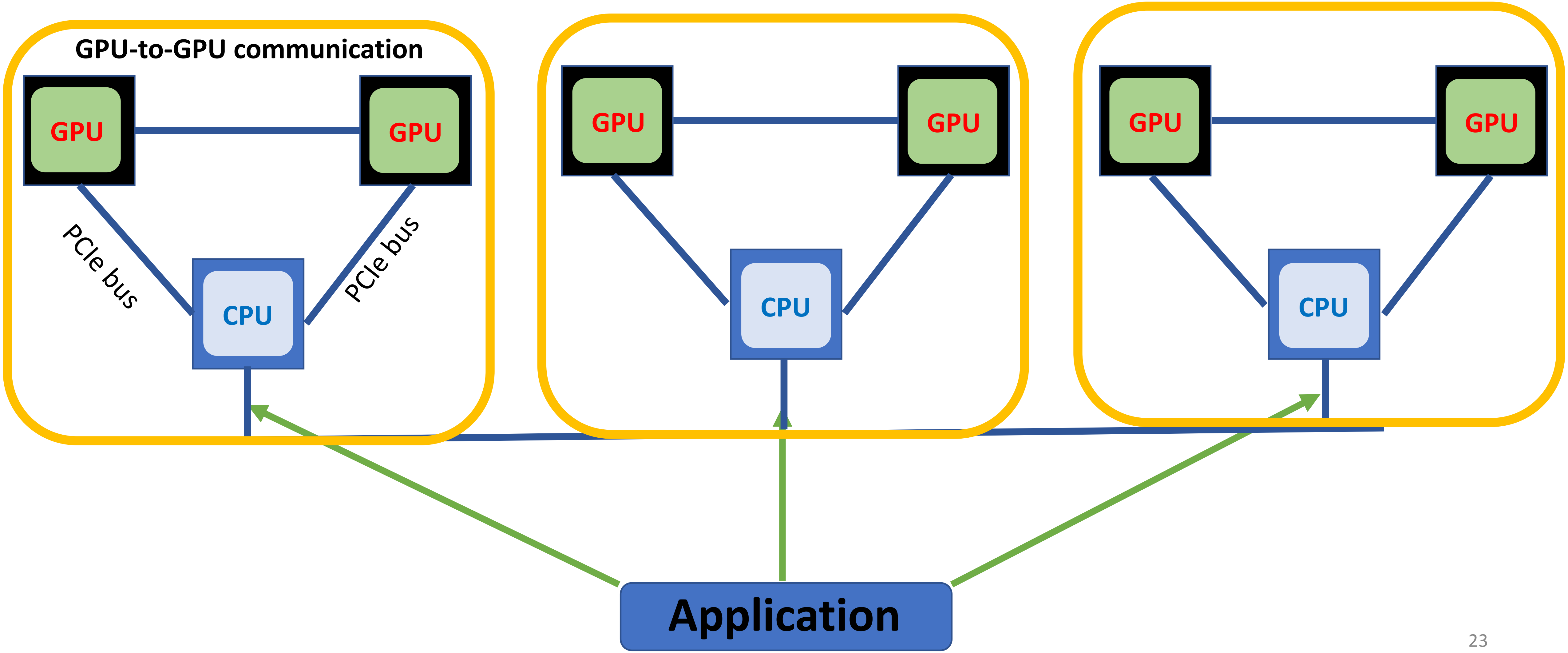
Programming models require the GPU-support



Hybrid multi-GPU programming: Combining MPI/OpenMP threading & OpenACC/OpenMP offloading

Multi-GPU programming

MPI+GPU programming model



Outline

I. Hardware topology [CPU vs GPU] [**GPU**: Graphics Processing Unit]

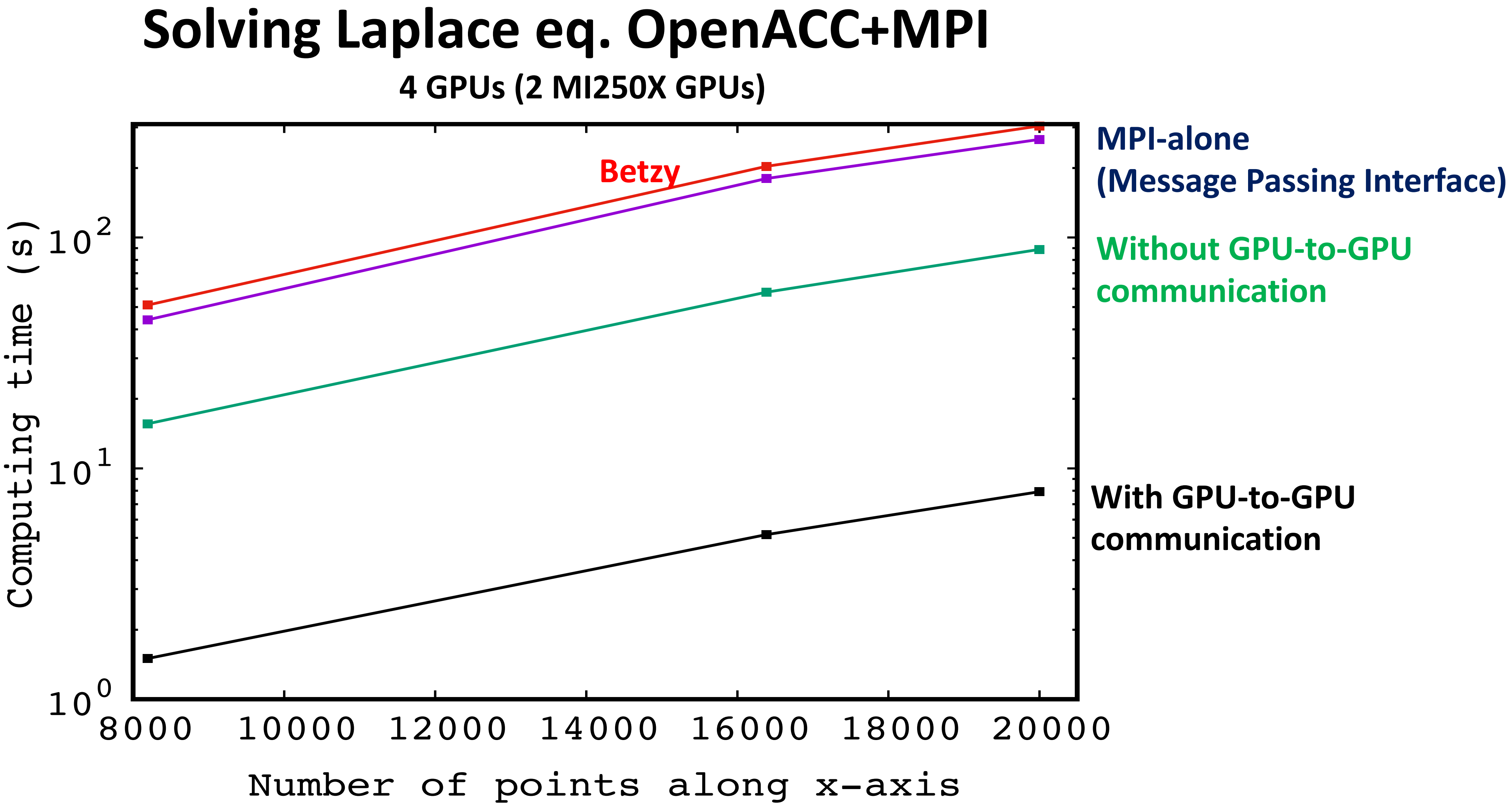
II. Overview of GPU-programming models

III. **Benchmark**

IV. **Overview of NRIS services**

V. **Supercomputer LUMI**

Benchmark

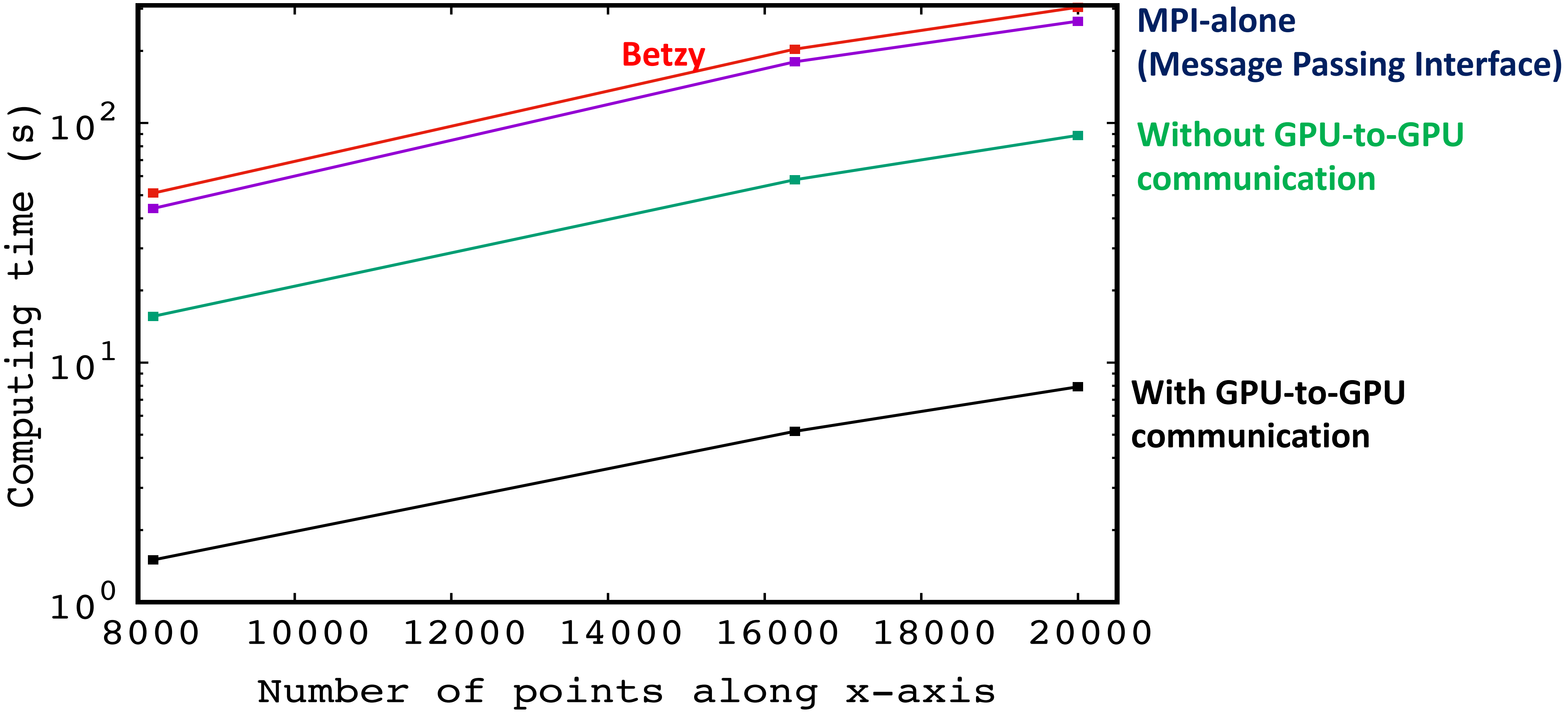


Solving Laplace eq. OpenACC+MPI
4 GPUs (2 MI250X GPUs)
Betzy
Computing time (s)
10^2
10^1
10^0
8000
10000
12000
14000
16000
18000
20000
Number of points along x-axis
MPI-alone
(Message Passing Interface)
Without GPU-to-GPU
communication
With GPU-to-GPU
communication

Benchmark

Solving Laplace eq. OpenACC+MPI

4 GPUs (2 MI250X GPUs)

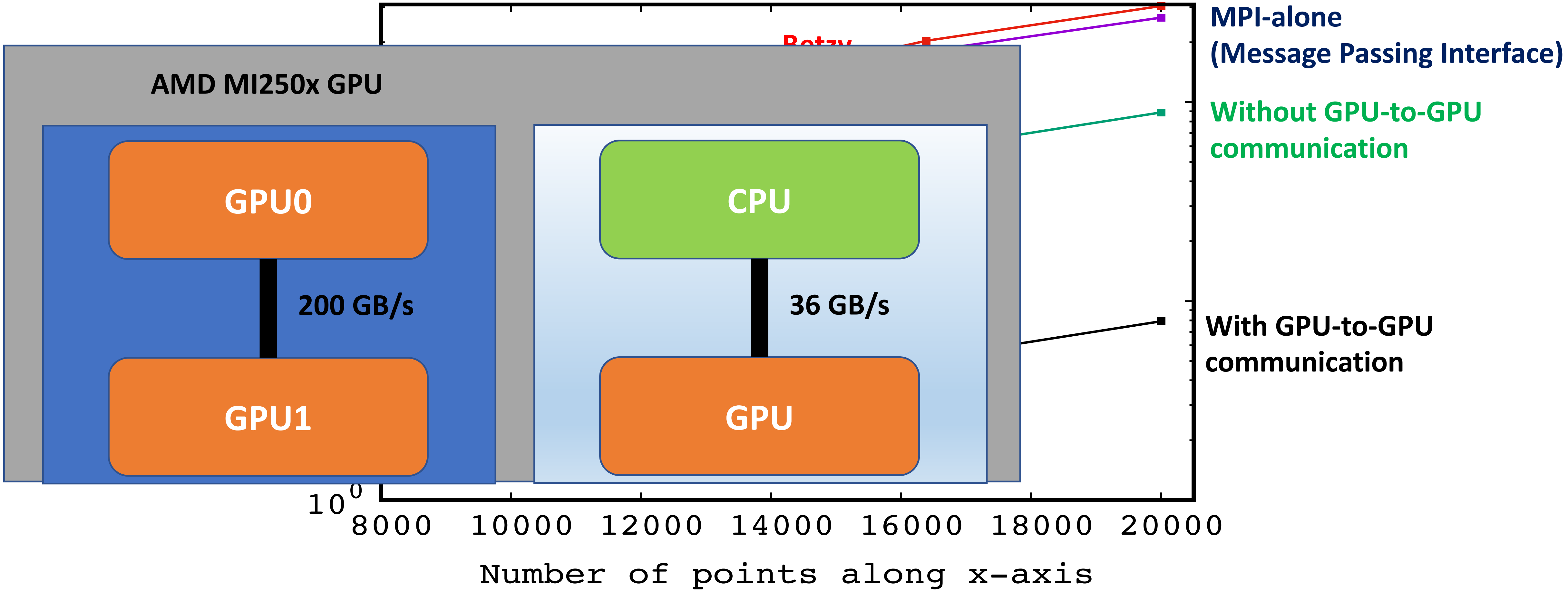


Benefit of GPUs: performance increases by a factor of 30

Benchmark

Solving Laplace eq. OpenACC+MPI

4 GPUs (2 MI250X GPUs)



Benefit of GPUs: performance increases by a factor of 30

Outline

I. Hardware topology [CPU vs GPU] [GPU: Graphics Processing Unit]

II. Overview of GPU-programming models

III. Benchmark

IV. Overview of NRIS services

V. Supercomputer LUMI

Overview of services -NRIS (Norwegian Research Infrastructure Services)

Organizations

NRIS

Collaboration

- UiB
- UiO
- NTNU
- UiT
- Sigma2

Services

National HPC and storage facilities

- High-performance computing (HPC)
- Data Storage
- Cloud services

Advanced user support

Long-term support, e.g:

- Optimisation
- GPU-acceleration
- Establish portals

Other services (Sigma2)

- Data management plan
- Research data archive
- Course resources
- Sensitive data

Focus on research rather than IT-tasks: Excellent research

Further details can be found here https://www.sigma2.no/services-overview

National HPC and storage -NRIS

Services

High-performance computing (HPC)

***NIRD Data Storage**

NIRD Cloud services (Kubernetes)

User benefits

--Access to national supercomputers (4 HPC systems) **Betzy** (172032 cores), **Fram** (32256 cores), **saga** (16064 cores) and **LUMI.**

Different teams

- Infra team
- Software team
- GPU team
- LUMI team
- Support service team

***NIRD** – National Infrastructure for Research Data

National HPC and storage -NRIS

Services

High-performance computing (HPC)

***NIRD Data Storage**

NIRD Cloud services (Kubernetes)

User benefits

--Access to national supercomputers (4 HPC systems) **Betzy** (172032 cores), **Fram** (32256 cores), **saga** (16064 cores) and **LUMI.**

--New ***NIRD** platform (IBM) installed at **Lefdal Mine Datacenter**, Nordfjordeid (**available on Dec. 2022**).

Lefdal Mine Datacenter

***NIRD** – National Infrastructure for Research Data https://www.sigma2.no/procurement-project-nird2020

https://www.sigma2.no/data-centre-facility

National HPC and storage -NRIS

Services

High-performance computing (HPC)

***NIRD Data Storage**

NIRD Cloud services (Kubernetes)

User benefits

--Access to national supercomputers (4 HPC systems) **Betzy** (172032 cores), **Fram** (32256 cores), **saga** (16064 cores) and **LUMI.**

--New *__NIRD__ platform (IBM) installed at **Lefdal Mine Datacenter**, Nordfjordeid (**available on Dec. 2022**).

--Capacity of **32 TB (initially)** and up to **70 TB (future)**

--Store, process & share large datasets (work on the same shared area).

--With support of AI/ML & data analysis.

--The facility is integrated with HPC systems (large datasets computation).

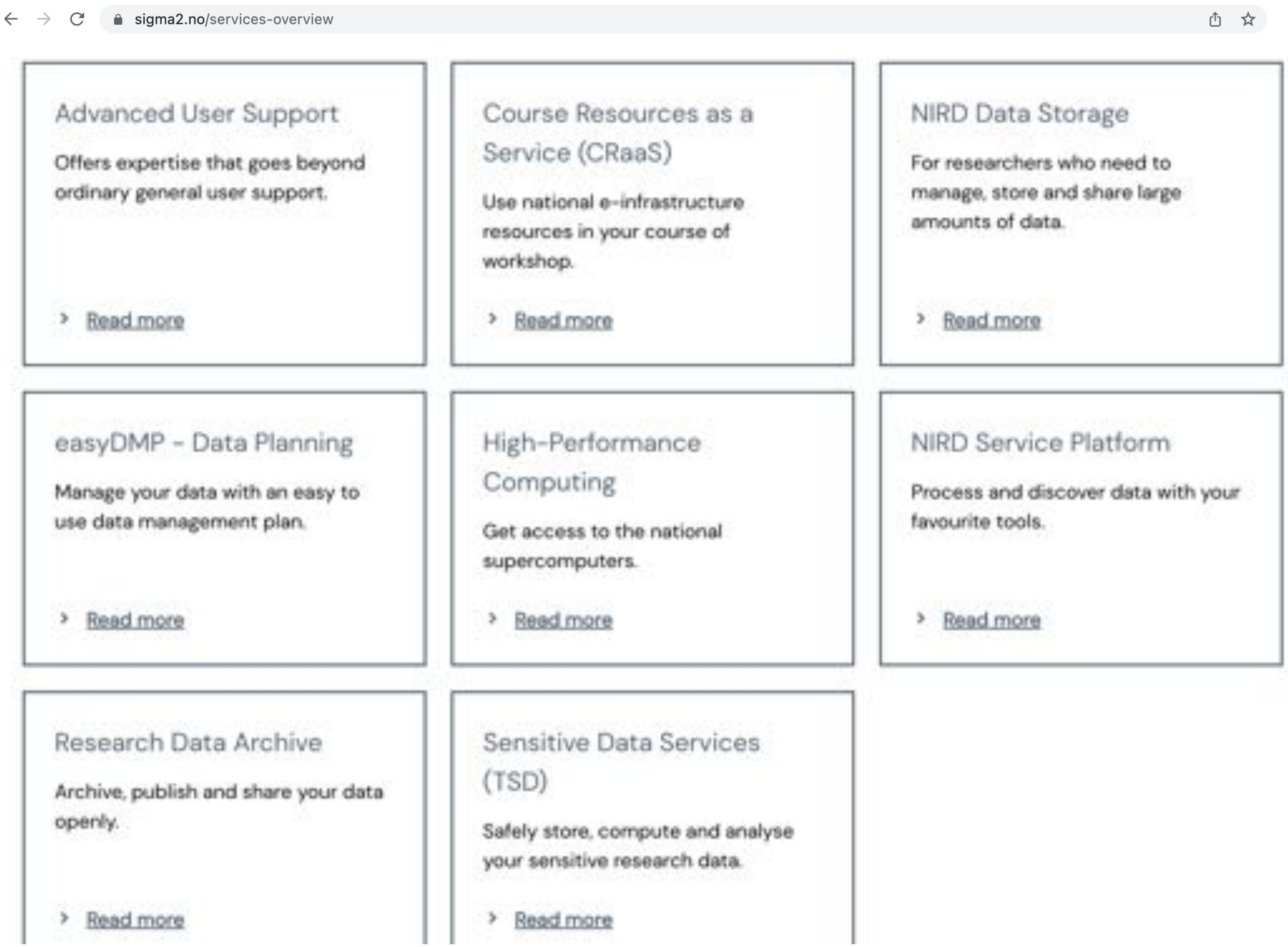
--Web services, Data visualization, pre/post-processing, data sharing,...

Lefdal Mine Datacenter

***NIRD** – National Infrastructure for Research Data

Screenshot of NRIS services

https://www.sigma2.no/services-overview



Outline

I. Hardware topology [CPU vs GPU] [**GPU**: Graphics Processing Unit]

II. Overview of GPU-programming models

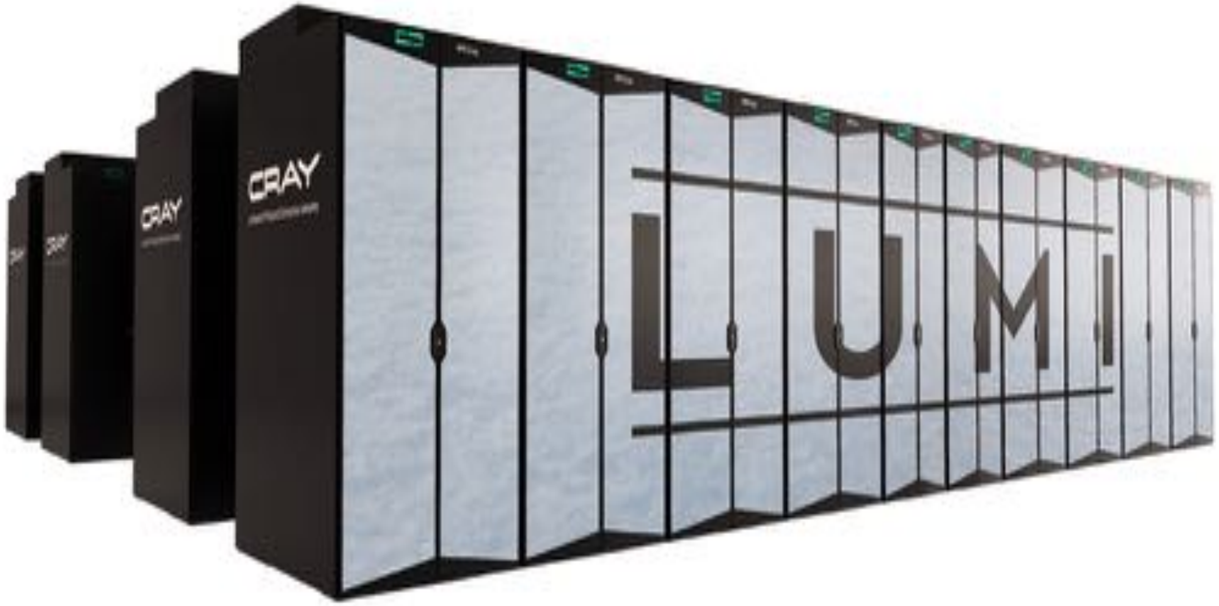
III. Benchmark

IV. Overview of NRIS services

V. Supercomputer LUMI

About the supercomputer LUMI

What is LUMI ? It is the 3rd fastest supercomputers in the world

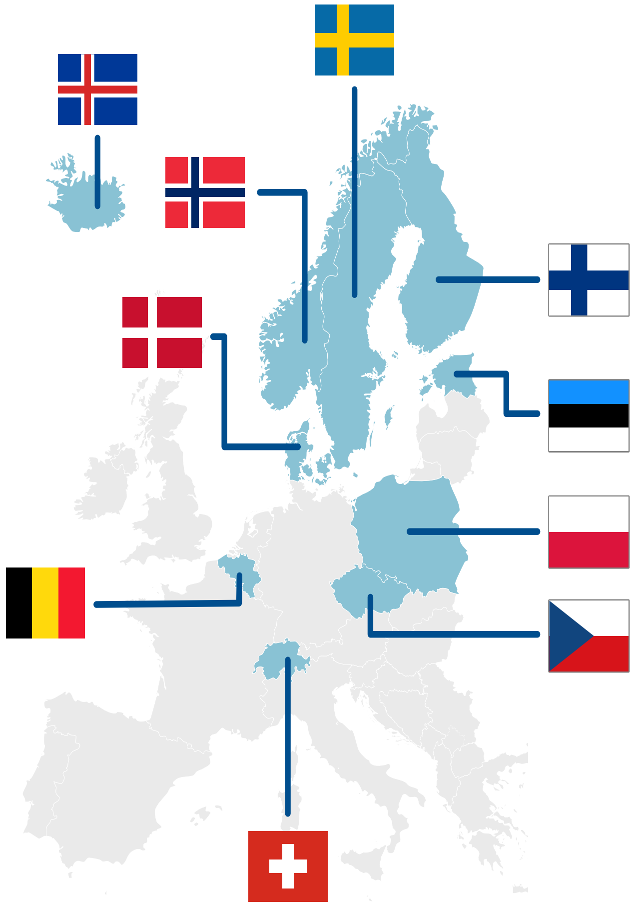


Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, **HPE** DOE/SC/Oak Ridge National Laboratory **United States**	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, **Fujitsu** RIKEN Center for Computational Science **Japan**	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, **HPE** EuroHPC/CSC **Finland**	1,110,144	151.90	214.35	2,942
4	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR	2,414,592	148.60	200.79	10,096

https://top500.org/lists/top500/list/2022/06/

About the supercomputer LUMI

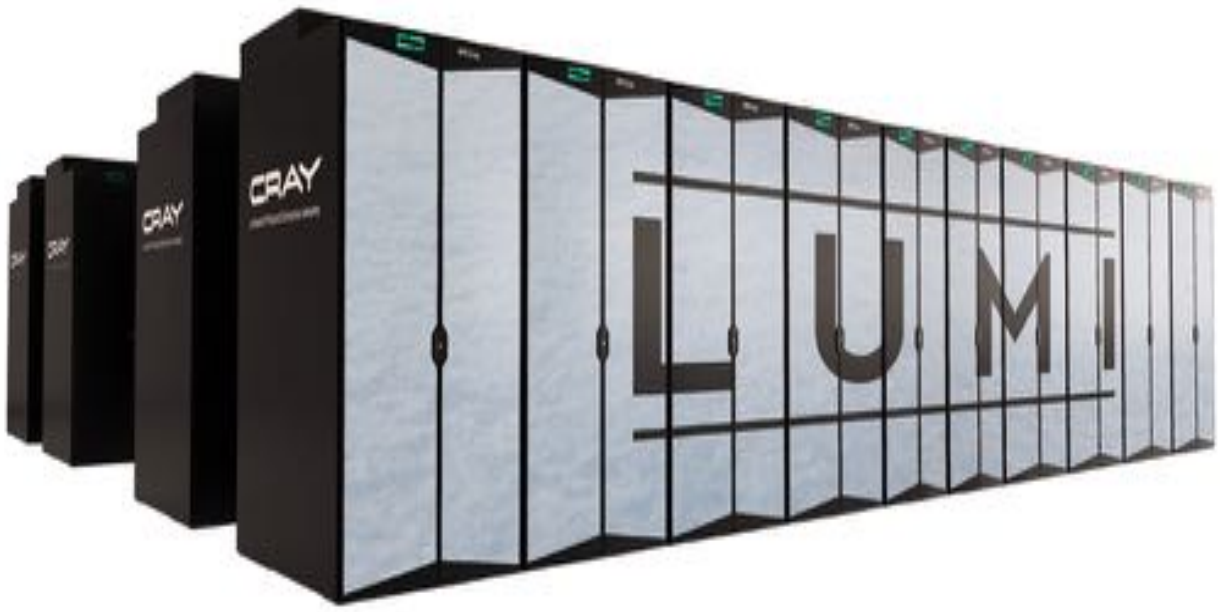
What is LUMI ?



- **LUMI** (Large Unified Modern Infrastructure): **SNOW**
- **LUMI** is located in a data center in Kajaanni, **Finland**.
- Funded by the **EuroHPC JU** (50%) and a **consortium of 10 countries.**
- **LUMI consortium**: **Finland**, Belgium, The Czech republic, Denmark, Estonia, **Norway**, Poland, Sweden, Switzerland and Iceland.

About the supercomputer LUMI

What is LUMI ?



- **LUMI** (Large Unified Modern Infrastructure): **SNOW**
- **LUMI** is located in a data center in Kajaanni, **Finland**.
- Funded by the **EuroHPC JU** (50%) and a **consortium of 10 countries.**
- **LUMI consortium**: **Finland**, Belgium, The Czech republic, Denmark, Estonia, **Norway**, Poland, Sweden, Switzerland and Iceland.

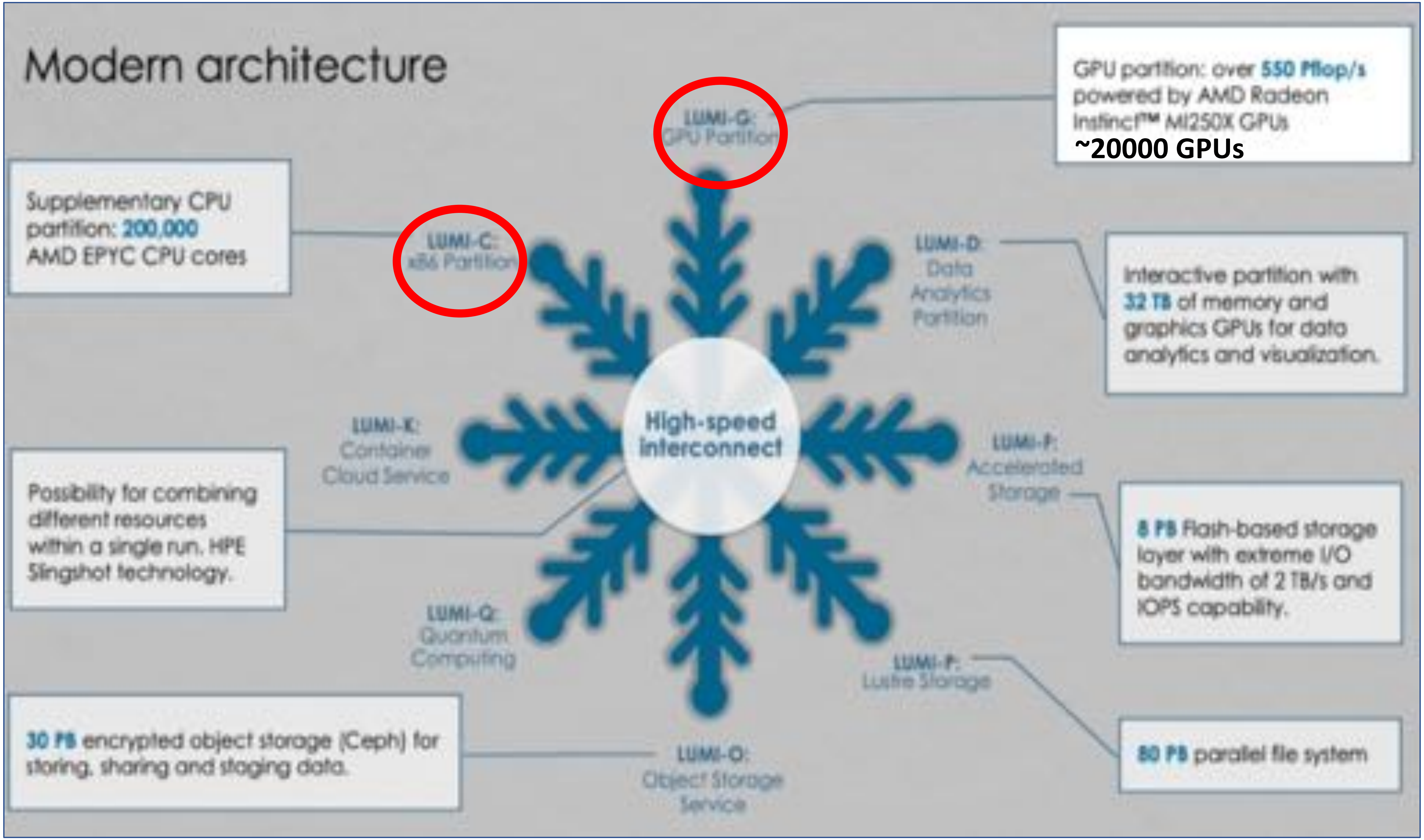
Computing power equivalent to

1 500 000

Modern laptop computers

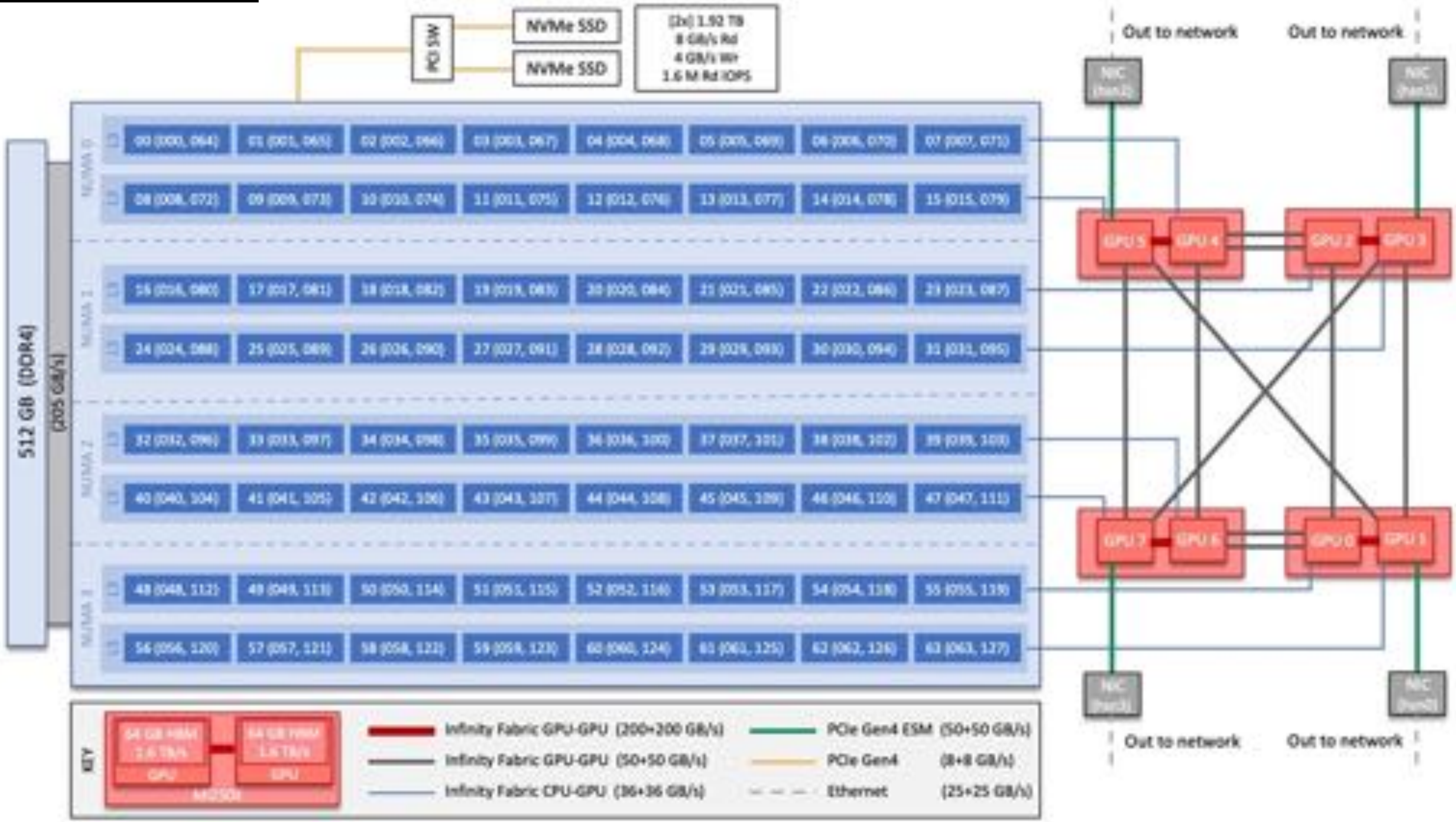
About the size of a tenis court

Weight around 150 000 Kg



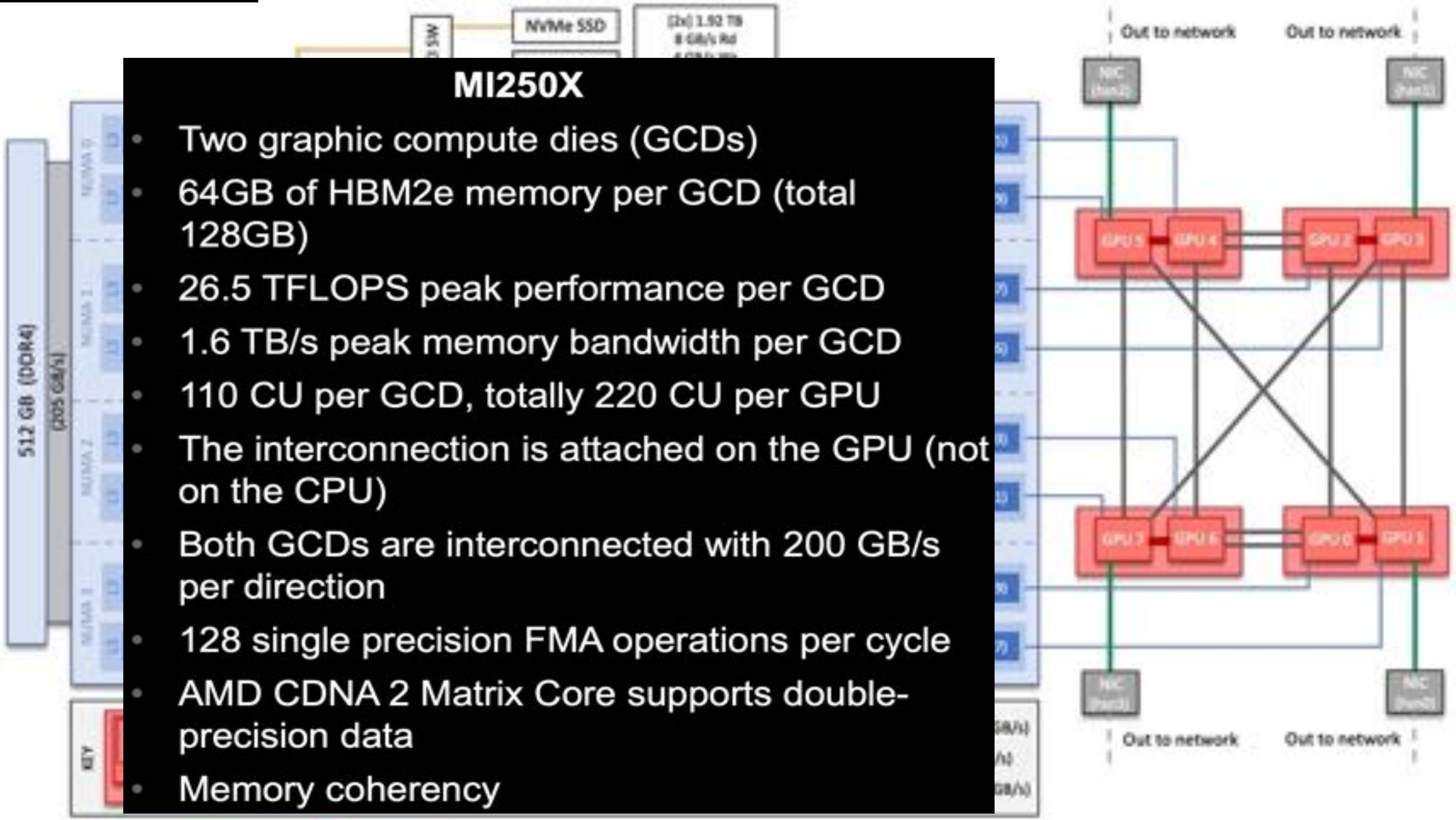
Taken from Tomasz Malkiewicz, CSC, presentation

LUMI-G partition: Architecture of AMD MI250x GPU



https://docs.olcf.ornl.gov/systems/crusher_quick_start_guide.html

LUMI-G partition: Architecture of AMD MI250x GPU



MI250X

- Two graphic compute dies (GCDs)
- 64GB of HBM2e memory per GCD (total 128GB)
- 26.5 TFLOPS peak performance per GCD
- 1.6 TB/s peak memory bandwidth per GCD
- 110 CU per GCD, totally 220 CU per GPU
- The interconnection is attached on the GPU (not on the CPU)
- Both GCDs are interconnected with 200 GB/s per direction
- 128 single precision FMA operations per cycle
- AMD CDNA 2 Matrix Core supports double-precision data
- Memory coherency

https://docs.olcf.ornl.gov/systems/crusher_quick_start_guide.html

LUMI-G for AI/ML applications

- **AMD ROCm** instead of **CUDA**.
- ROCm is an **open software** platform for HPC & GPU-computing.
- **ROCm is comprised** of open technologies:
  - ML Frameworks (PyTorch/TensorFlow/Jax)
  - ***Libraries** (MIOpen / Blas / RCCL), programming model (HIP)
  - *Tools, guidance and insights are **shared freely** across the ROCm **GitHub community** and forums.
- GPU-accelerated applications with AMD ROCm

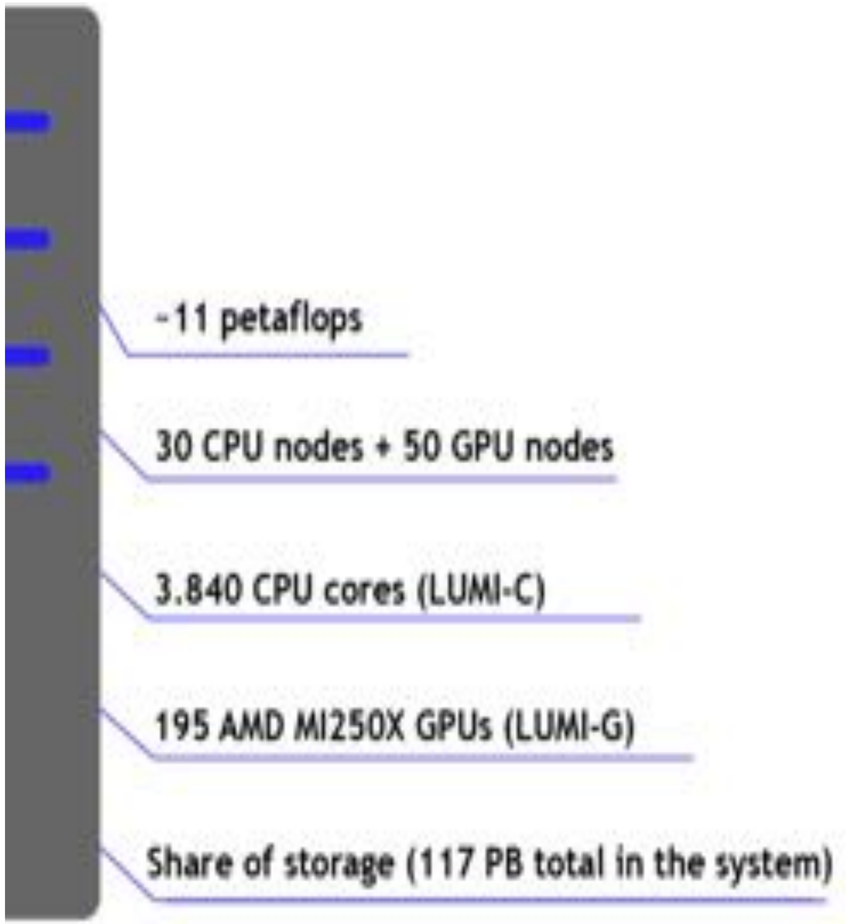
  https://www.amd.com/system/files/documents/gpu-accelerated-applications-catalog.pdf

https://rocmdocs.amd.com/en/latest/

Norway’s share of LUMI is 2%

Key figures – Sigma2 share

CPU-core-hours	34 003 333
GPU-hours	1 771 000
TB-hours	16 862 500
Central disk	A share of the total 117 PB
Theoretical Performance (Rpeak)	~11 PFLOPS



What Supercomputers can be used for ?

To solve majore challenges in the world.

Supercomputer platfrom

HPC

AI/ML

Data analytics

To solve complicated problems in physical sciences, engineering, business & humanities, such that:

- Exploring the boundaries of quantum chemistry (first LUMI project).
- Predicting the structure of proteins using data-driven methodologies (ML, DL).
- Understanding the functionality **of COVID-19 virus** & potential cures.
- Desiging new molecues with unique functionality for modern technology.
- Delivering reliable weather and climate predictions.
- New data-driven business.
- Nature language processing (e.g. smart speaker).

Best Practice Guide LUMI

Jussi Heikonen, CSC, Finland
Georgios Markomanolis, CSC, Finland
Cristian-Vasile Achim, CSC, Finland
Ezhilmathi Krishnasamy, University of Luxembourg, Luxembourg
Abdulrahman Azab, University of Oslo, Norway
Pedro Ojeda-May, HPC2N, Umeå University, Sweden
Michele Martone, LRZ, Germany
Marcin Krotkiewski, University of Oslo, Norway
Hicham Agueny, University of Bergen, Norway
Maria Guadalupe Barrios Sazo, University of Oslo, Norway
Ole Widar Saastad (Editor), University of Oslo, Norway

12 January 2023

PRACE

https://prace-ri.eu/training-support/best-practice-guides/best-practice-guide-lumi/

Conclusion

Conclusion

- **Lifetime** of **hardware** is less than **five years.**
- **Software can be used for decades**.
- **Software investments** provide more **flexibility**….

Bottlenecks in GPUs:

Scientific breakthroughs

- **Data transfer between CPU and GPU.**
- **Access to the global memory to pass data.**
- **Portability.**

I stop HERE

☺