



JUPITER: GPUS FOR EXASCALE

2024-11-05 | ANDREAS HERTEN | JÜLICH SUPERCOMPUTING CENTRE



Member of the Helmholtz Association



EuroHPC
Joint Undertaking



Bundesministerium
für Bildung
und Forschung

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen



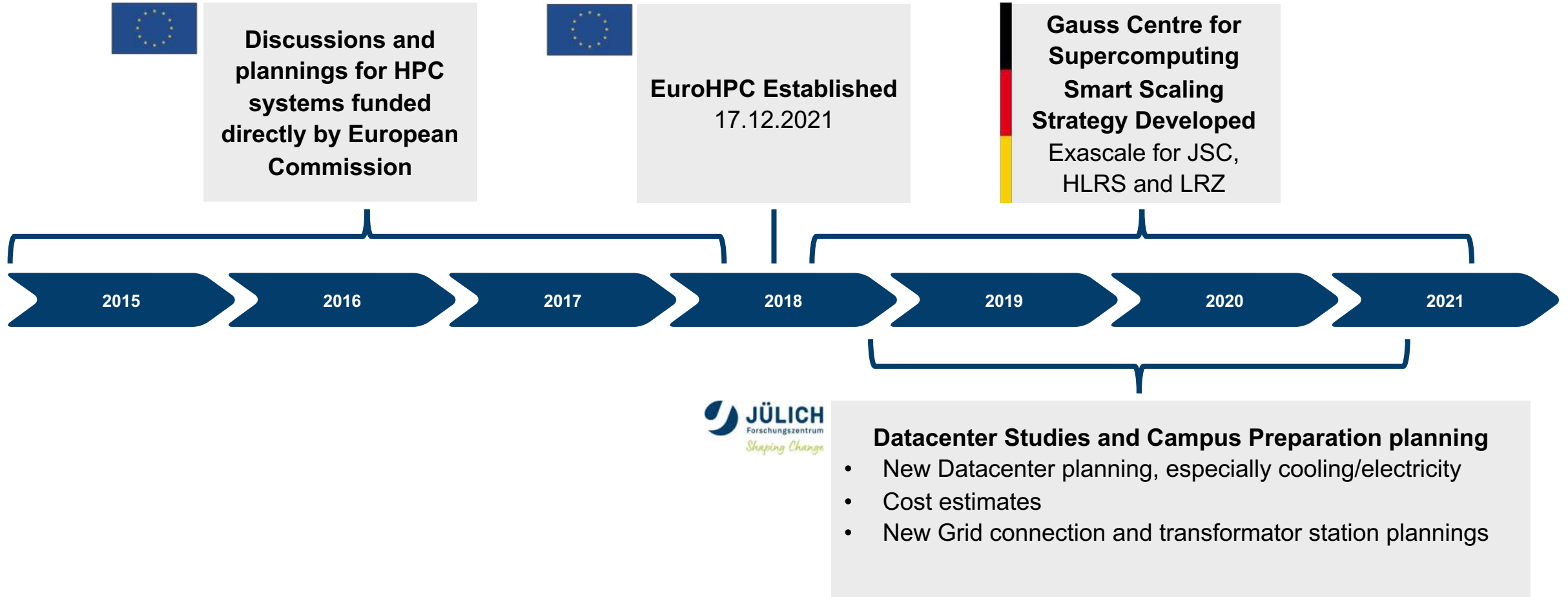
GCS
Gauss Centre for Supercomputing

JÜLICH
Forschungszentrum
Shaping Change

OUTLINE

- JUPITER
 - Components
 - MDC
 - JEDI
 - Procurement
 - GH200
 - Results
- GPU Programming
 - CPU vs GPU
 - GPU Core Features
 - CUDA

A LONG TIME AGO ...



JUPITER

- ParTec/Eviden Supercomputer Consortium
- Implementing Modular Supercomputing Architecture
- JUPITER **Booster**: High scalability; 1 EFLOP/s HPL, >70 EFLOP/s FP8
- JUPITER **Cluster**: High versatility; 0.5 B/FLOP balance
- Network: 200/400 Gigabit NVIDIA Mellanox InfiniBand NDR
- Storage: 29 PB Flash + 310 PB Spinning + Tape
- 17 MW Linpack Power Consumption
- Direct Liquid Cooled to enable heat-reuse

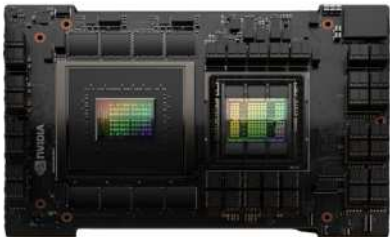


SYSTEM DETAILS

JUPITER MODULES

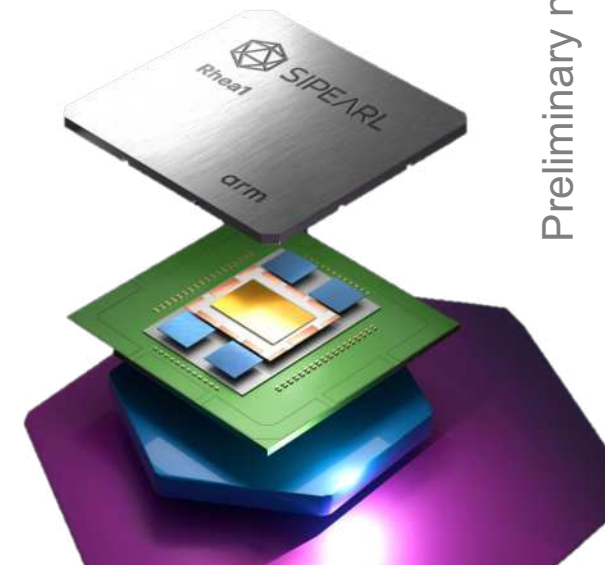
JUPITER Booster

- ~125 Racks BullSequana XH3000
- Node design
 - ~6000 nodes
 - 4× NVIDIA CG1 per node
- CG1: NVIDIA Grace-Hopper
 - 72 Arm Neoverse V2 cores (4×128b SVE2); 120 GB LPDDR5
 - H100 (132 SMs); 96 GB HBM3
 - NVLink C2C (900 GB/s)



JUPITER Cluster

- ~14 Racks BullSequana XH3000
- Node design
 - ~1300 nodes
 - 2× SiPearl Rhea1 per node
- Rhea1
 - 80 Arm Neoverse V1 cores (2×256b SVE)
 - 256 GB DDR5, 64 GB HBM2e



JUPITER – BOOSTER COMPUTE NODE ARCHITECTURE

- 4× NVIDIA Grace-Hopper in SXM5 Board (4× 680W)
- 4× NVIDIA InfiniBand NDR200
- 480 GB LPDDR5X / 360 GB HBM3 (usable)
- NVLink 4
 - GPU-GPU 150 GB/s per dir, CPU-GPU 450 GB/s per dir, CPU-CPU 100 GB/s per dir
- CG4 Motherboard (4× CG1 GH module + 4× CX7 HCA assembly)
 - All NVIDIA, except the BMC

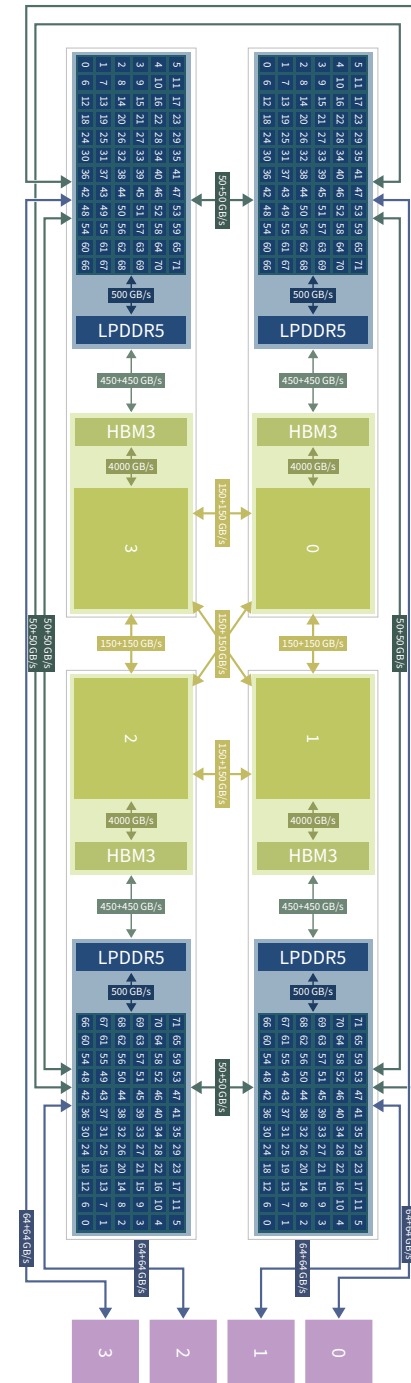
Node Specs

- **ARM Neoverse V2** CPU Specs
 - SVE2/NEON (4x 128 bit vector op)
 - 72 cores @ ~2.4GHz (~3.2 GHz turbo)
 - 120 GB LPDDR5X (8 channels)
 - ≥450 GB/s
 - ~150 ns latency

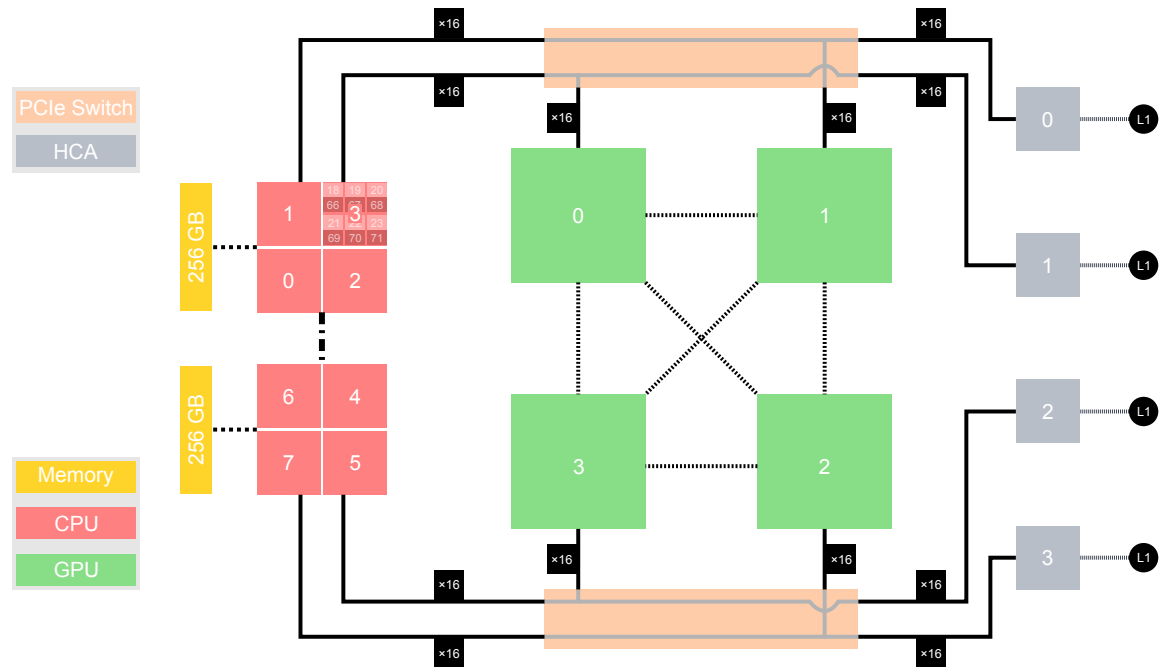
CPU Specs

- **H100** GPU Specs
 - ~50 TFLOP/s (HPL single GPU)
 - 96 GB HBM3
 - 4000 GB/s
 - ~450 ns latency

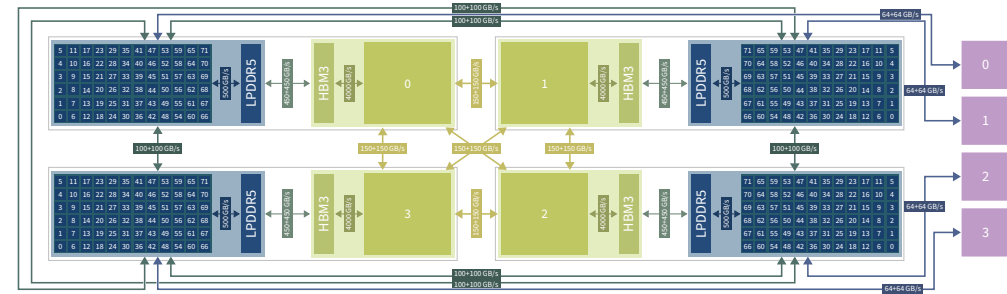
GPU Specs



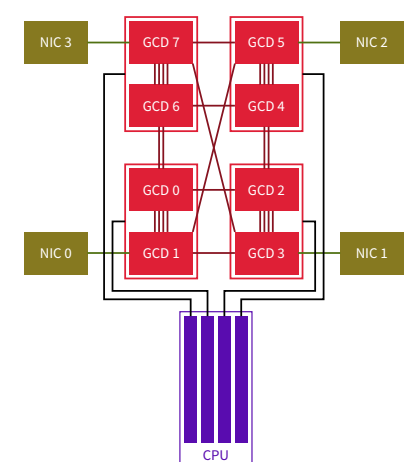
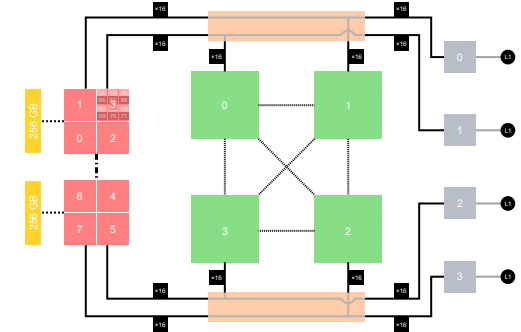
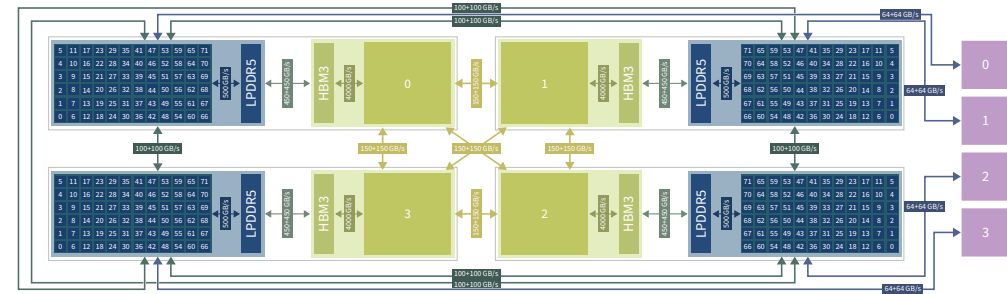
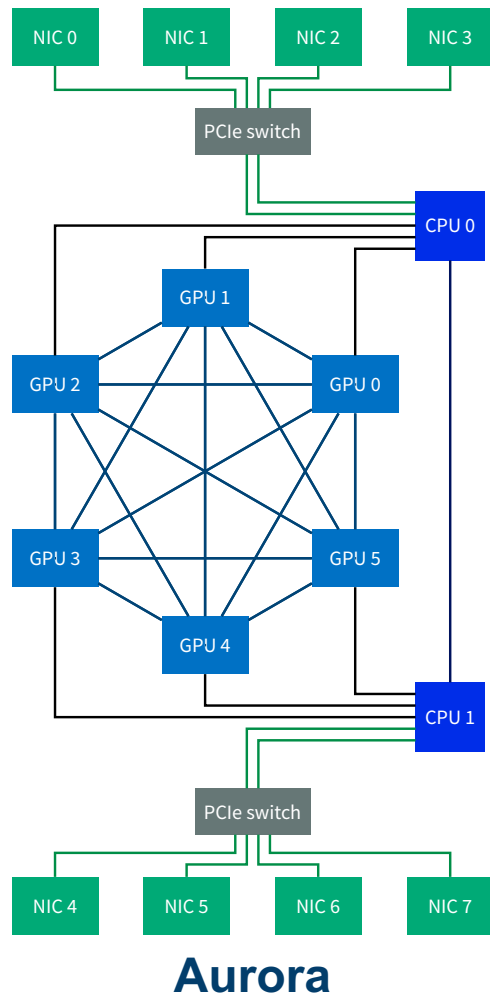
NODE COMPARISON



JUWELS Booster

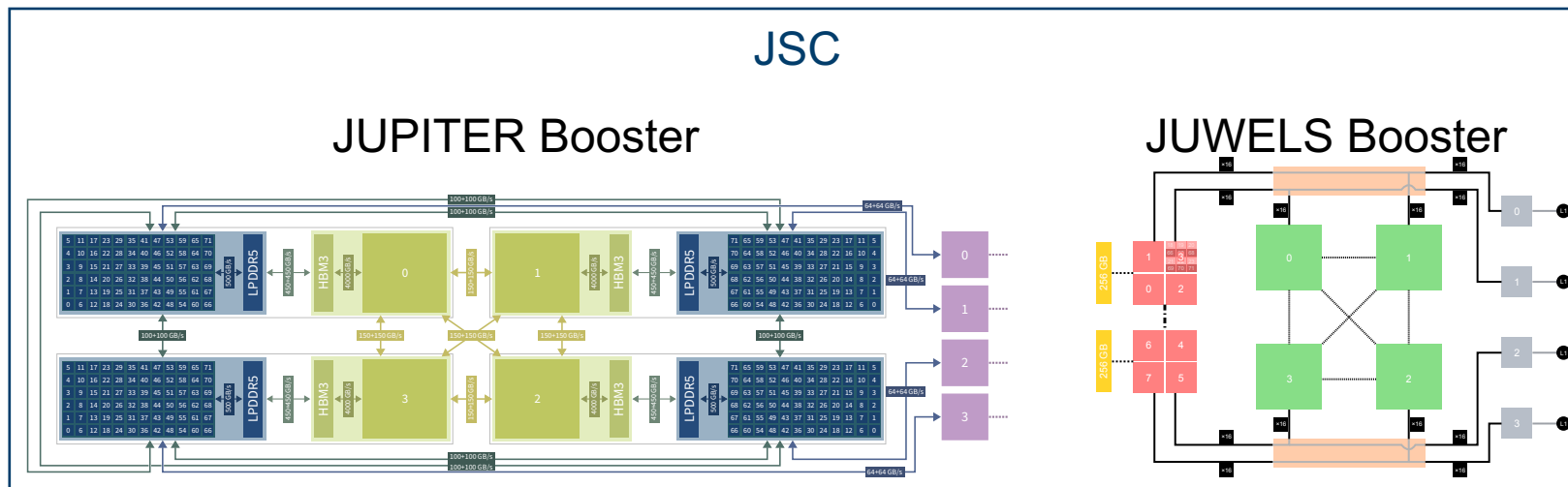
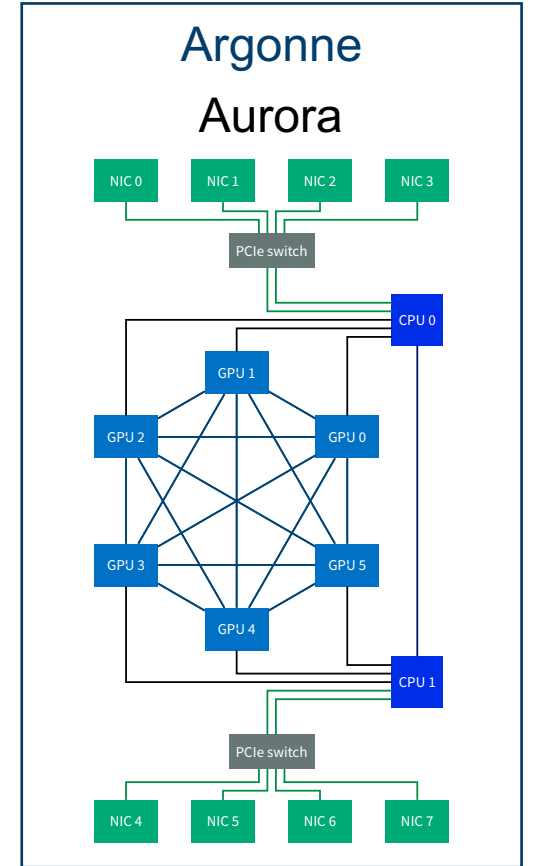
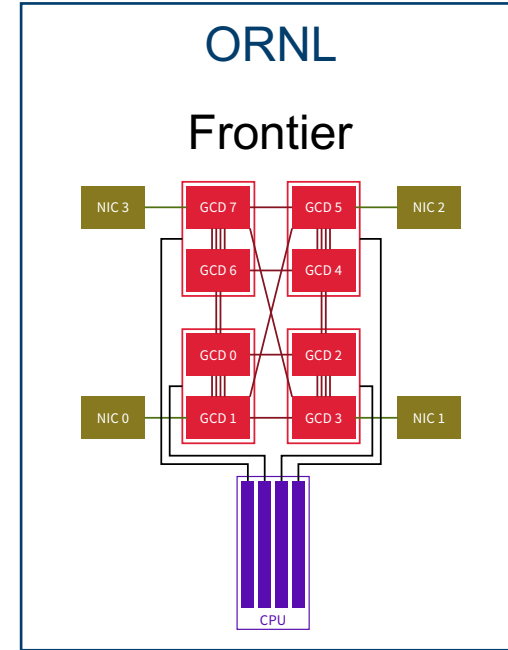


NODE COMPARISON



NODE COMPARISON

- JUWELS Booster: 2× CPU, 4× GPU, 4× IB
- JUPITER Booster: 4× CPU+GPU, 4× IB
- Frontier: 1× CPU, 4×(2× GPU), 4× Slingshot
- Aurora: 2× CPU, 6× GPU, 8× Slingshot
- El Capitan: 4× APU



JUPITER – CLUSTER COMPUTE NODE ARCHITECTURE

- 2× SiPearl Rhea1
- 1× NVIDIA InfiniBand NDR200
- 512 GB DDR5 (36 nodes with 1024 GB)
- CCIX

Node Specs

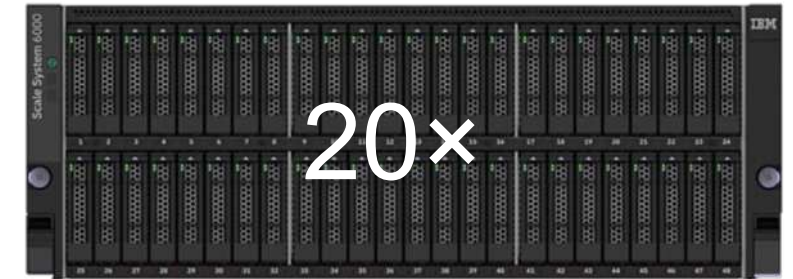
- ARM Neoverse V1 Zeus
 - 2 x 256 SVE per core
- 2.5 GHz (~3.0 GHz turbo)
- 64 GB HBM2e per Socket
 - 1.64 TB/s
- 256 GB DDR5
- PCIe Gen5

CPU Specs

JUPITER – STORAGE (SCRATCH)



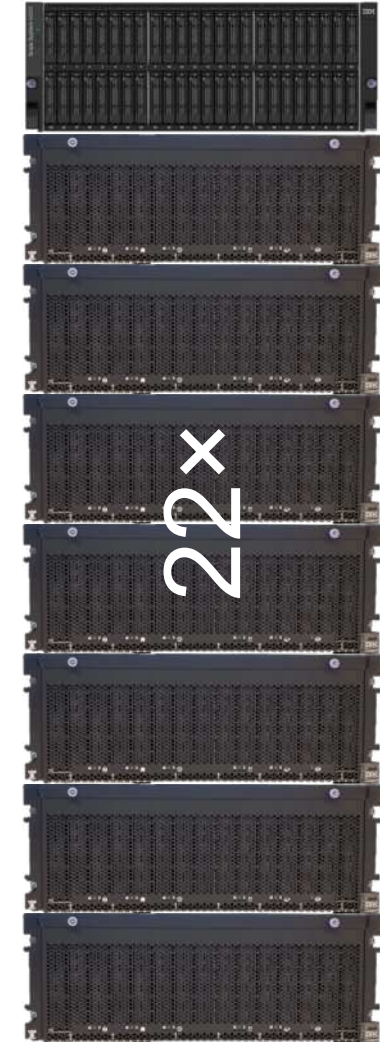
- Gross Capacity: 29 PB; Net Capacity: 21 PB
- Bandwidth: 2.1 TB/s Write, 3.1 TB/s Read
- 20× IBM SSS6000 Building Blocks (40 servers)
 - 2× NDR400 per server
 - 48× 30 TB NVMe drives per block
 - IBM Storage Scale (aka Spectrum Scale/GPFS)
- Manager and Datamover Nodes
- Exclusive for JUPITER
 - Integrated into InfiniBand fabric



JUPITER – STORAGE (EXASTORE)

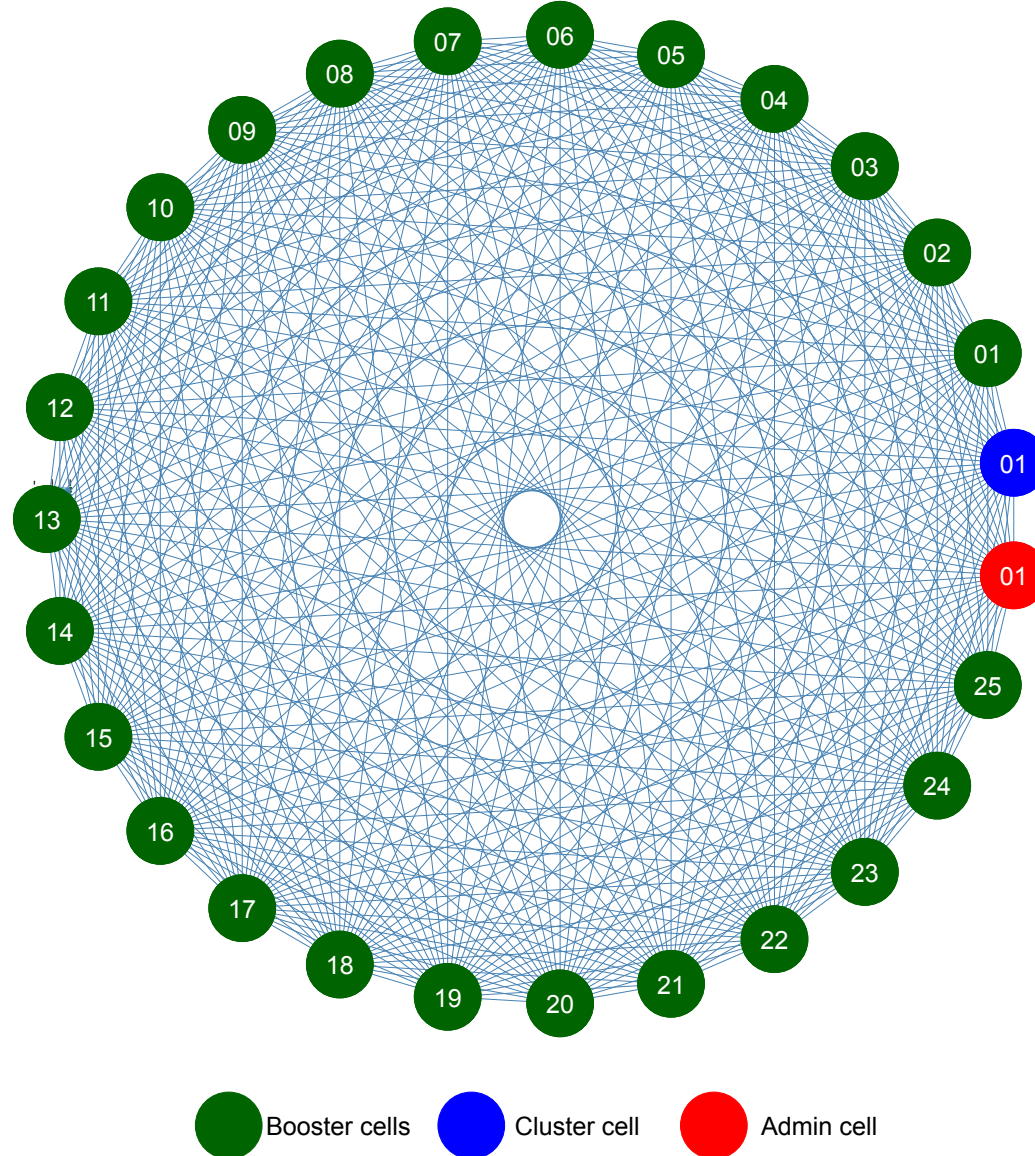
In kind contribution from JSC, not part of the JUPITER procurement

- Gross Capacity: 308 PB; Net Capacity: 210 PB
- Bandwidth: 1.1 TB/s Write, 1.4 TB/s Read
- 22× IBM SSS6000 Building Blocks (44 servers)
 - 2× NDR200 per server
 - 7× JBOD enclosures, each with 91x 22 TB Spinning Disks per block
 - IBM Storage Scale (aka Spectrum Scale/GPFS)
- Manager and Datamover Nodes
- Exclusive for JUPITER
 - Integrated into InfiniBand fabric



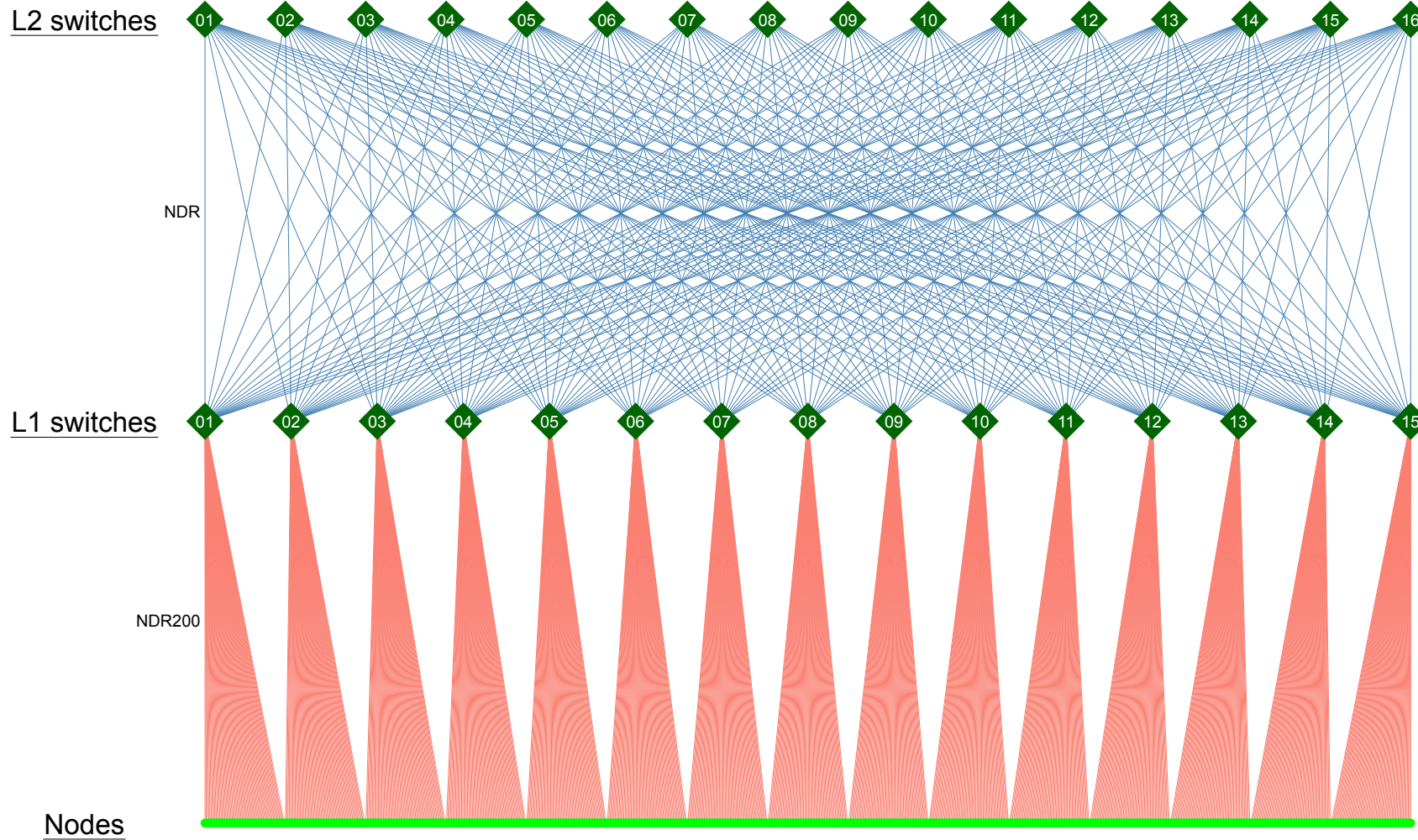
JUPITER – INTERCONNECT

One Network to Rule Them All



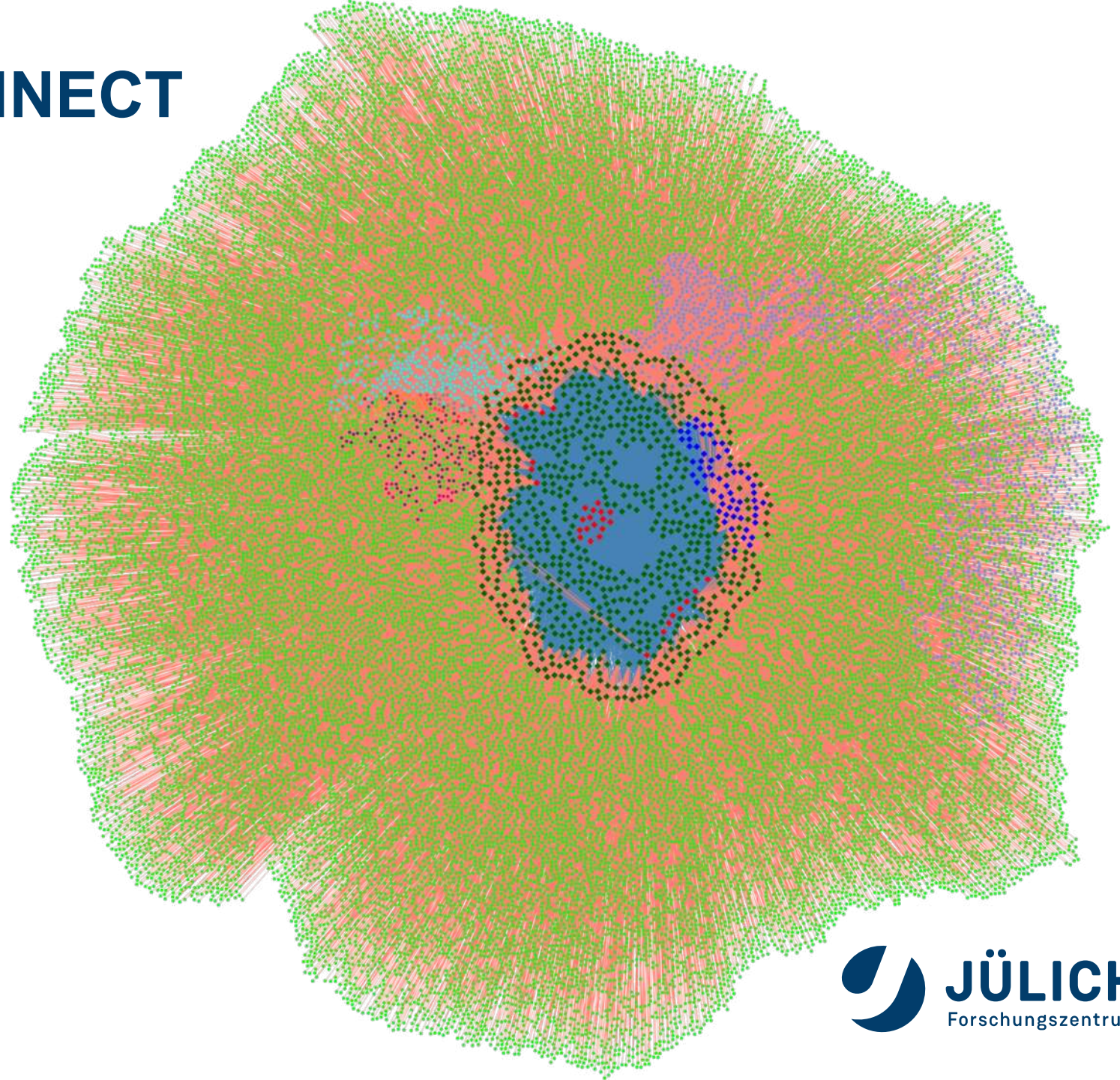
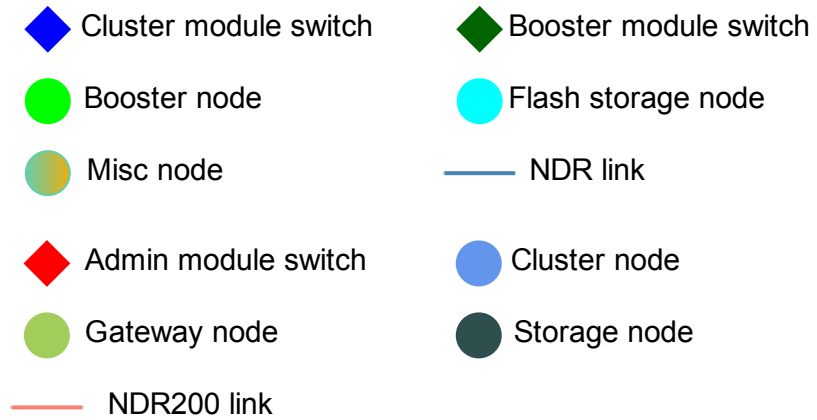
JUPITER – INTERCONNECT

One Network to Rule Them All



JUPITER – INTERCONNECT

One Network to Rule Them All



STATUS

POWER TRANSFORMER SUBSTATION AND LINES

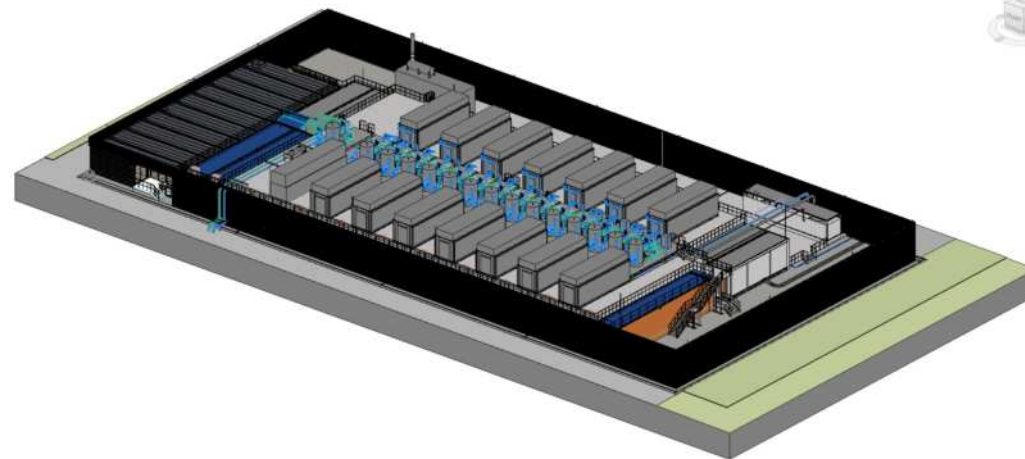
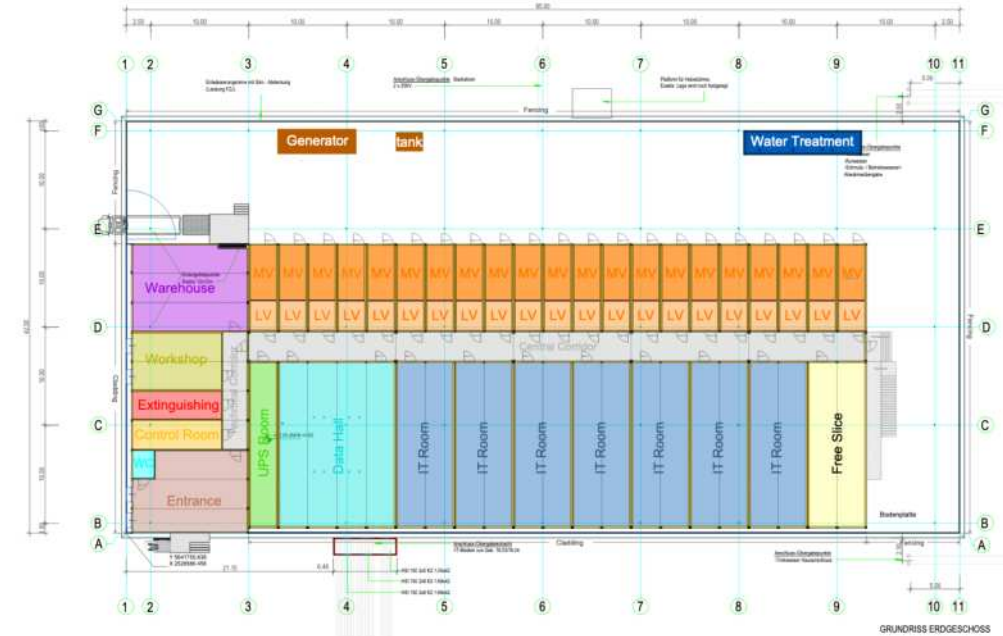
Upgrade of transformers 110 kV / 35 kV from 2 x 40 MVA to 2 x 60-80 MVA and upgrade 110kV power line



MODULAR DATA CENTER FOR JUPITER

EVIDEN

- Vendor: **Eviden**
- Area: ~2300 m²
- 1× Datahall (storage, management)
- 7× IT modules (20 racks per module)
- UPS, generator
- Entrance area
- Workshop, warehouse
- 15 × 2.5 MW power stations

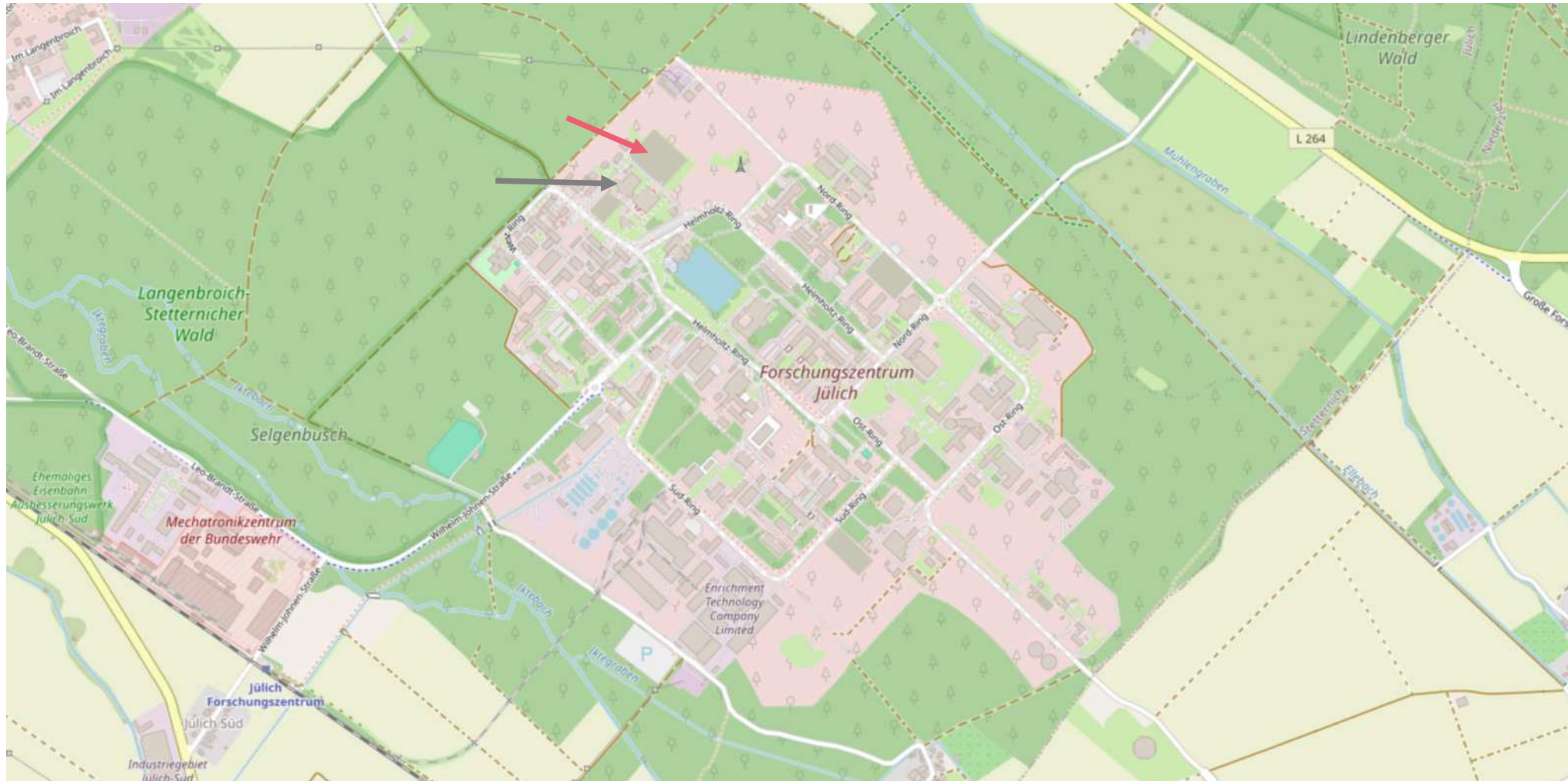


MODULAR DATA CENTER FOR JUPITER

EVIDEN



CONCRETE FOUNDATION



CONCRETE FOUNDATION

Construction of concrete slab 85 m x 42 m x 0.5 m





MDC SHIPMENT START

10./11.9.2024



DATAHALL ARRIVAL

28.10.2024



JUPITER INSTALLATION IN ANGERS (EVIDEN FACTORY)

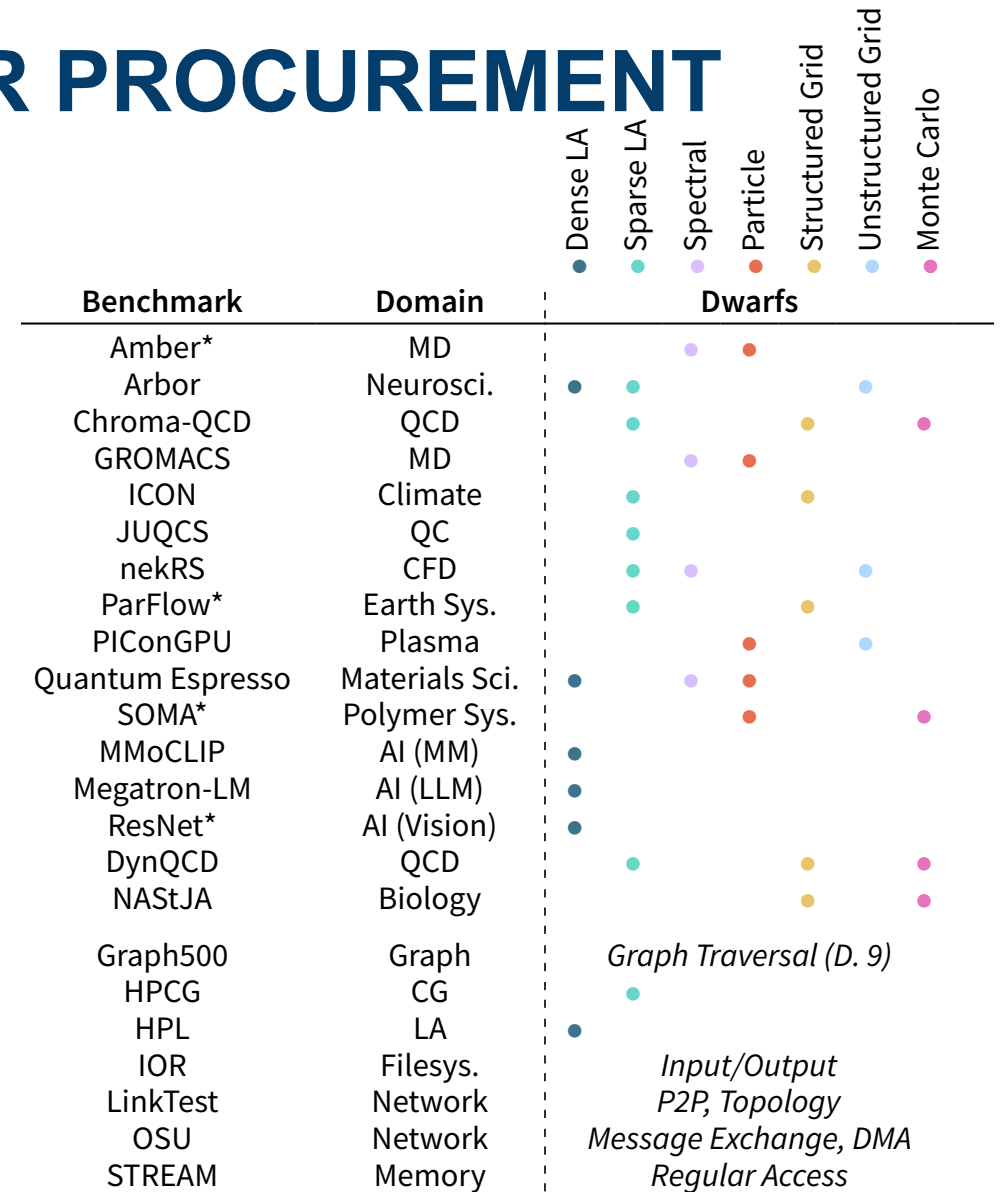
- 10 XH3000 racks, 480 nodes
- Hardware tests and benchmark preparation
- JUPITER Management Server preparation
- **Afterwards**
 - Integration into containers
 - Shipment to Jülich
 - “Plug in and run”



APPLICATIONS

APPLICATIONS FOR THE JUPITER PROCUREMENT

- Selection criteria
 - Current workload
 - Future workload
 - Relevance
 - Balance with other applications
 - Domains
 - Programming models
 - Programming languages
 - Profile
- High Scalability up to Exascale



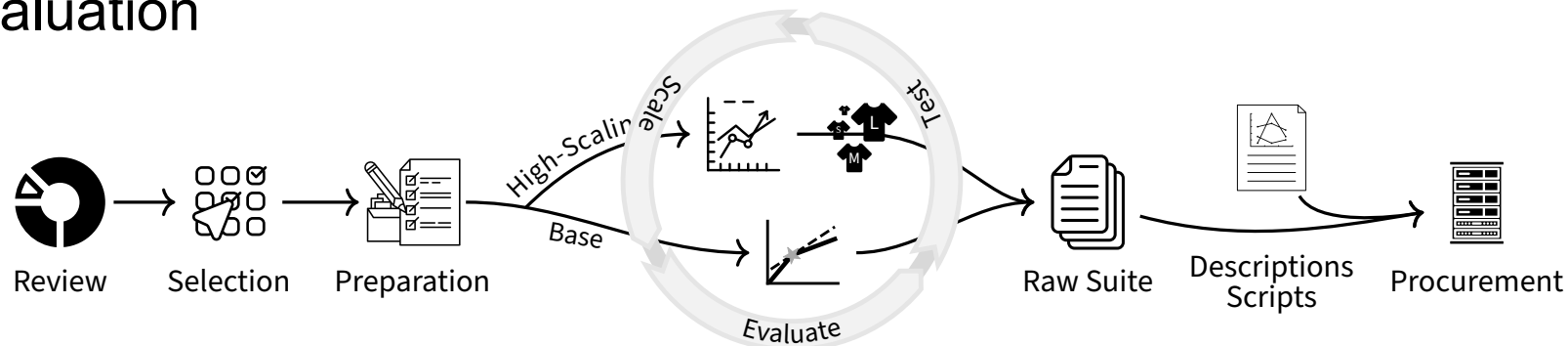
APPLICATIONS FOR THE JUPITER PROCUREMENT

- Selection criteria
 - Current workload
 - Future workload
 - Relevance
 - Balance with other applications
 - Domains
 - Programming models
 - Programming languages
 - Profile
- High Scalability up to Exascale

Benchmark	Booster			Cluster	MSA
	GPU	GPU High-Scale	CPU	CPU	
Arbor	✓	✓			
Chroma	✓	✓			
Gromacs	✓				
ICON	✓				
JUQCS	✓	✓			✓
nekRS	✓	✓			
ParFlow	✓				
PICongPU	✓	✓			
Quantum ESPRESSO	✓				
AI-MMoCLIP	✓				
AI-NLP	✓				
dynQCD				✓	
NAStJA				✓	
Graph500			✓		
HPCG	✓			✓	
HPL	✓			✓	
IOR			✓	✓	
LinkTest			✓	✓	✓
OSU	✓		✓	✓	
STREAM	✓			✓	

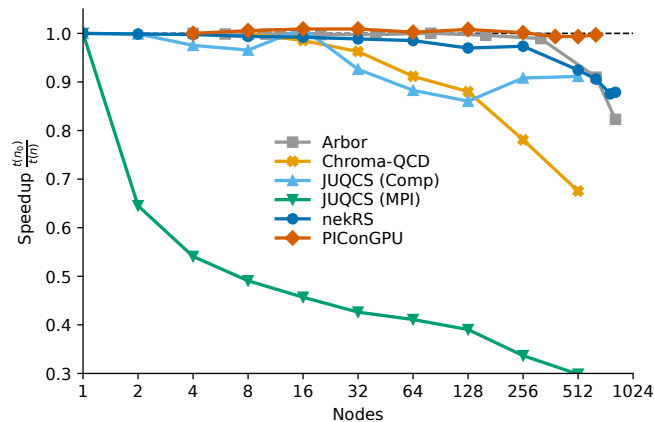
EVALUATION

- Criteria
 - Requirements to project planning, etc.
 - Technical requirements to overarching design and details
 - Performance of applications, benchmarks
 - Total cost of ownership (TCO): How much science for money
 - Further categories (Synthetic Benchmarks, High-Scaling Applications)
- Quantified evaluation



PAPER AT SC24

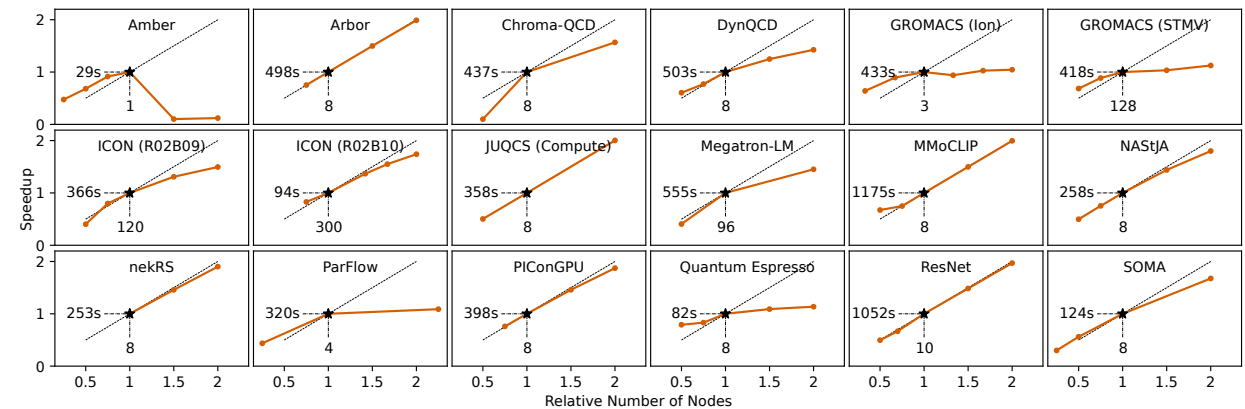
- Paper for Technical Program at SC24
- All benchmarks released as open source
→ <https://github.com/FZJ-JSC/jubench>
(meta-repo)
- Results, discussions, experiences, ...



Application-Driven Exascale: The JUPITER Benchmark Suite

Andreas Herten , Sebastian Achilles , Damian Alvarez , Jayesh Badwaik , Eric Behle , Mathis Bode ,
Thomas Breuer , Daniel Caviedes-Voullième , Mehdi Cherti , Adel Dabah , Salem El Sayed ,
Wolfgang Frings , Ana Gonzalez-Nicolas , Eric B. Gregory , Kaveh Haghighi Mood , Thorsten Hater ,
Jenia Jitsev , Chelsea Maria John , Jan H. Meinke , Catrin I. Meyer , Pavel Mezentsev , Jan-Oliver Mirus ,
Stepan Nassyr , Carolin Penke , Manoel Römmel , Ujjwal Sinha , Benedikt von St. Vieth , Olaf Stein ,
Estela Suarez , Dennis Willsch , Ilya Zhukov

Jülich Supercomputing Centre
Forschungszentrum Jülich
Jülich, Germany





GH200 TEST NODES

- GH200 Prototype
- 2× Grace-Hopper superchips
 - 1 Grace CPU (72 cores), 480 GB LPDDR5X RAM
 - 1 H100 GPU
 - TDP 700-1000 W
- Slightly different variant compared to JUPITER node design



ENABLEMENT: JEDI, JUREAP

-  JEDI: JUPITER test system
 - 48 nodes; JUPITER design
 -  **Top 1 Green500!**
- Usage
 - System management preparations
 - Application porting
 - JUREAP; Research and Early Access Program



JUREAP

Seeding Exascale in Europe!



jureap@fz-juelich.de • <https://www.fz-juelich.de/en/ias/jsc/jupiter/jureap>

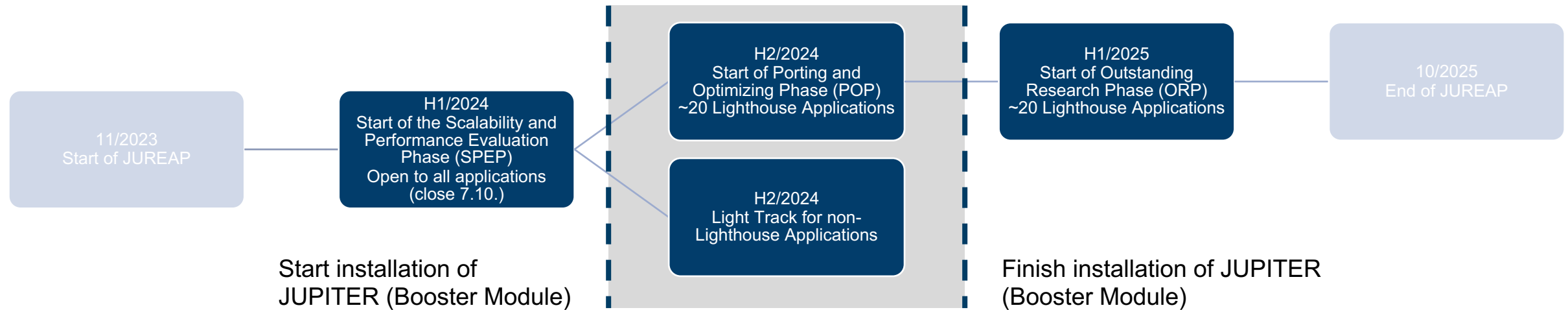
JUPITER Research and Early Access Program

OVERVIEW

Timeline

Current state:

- GCS Exascale Pioneer Call just closed
- Evaluations ongoing
- Light Track in parallel



Phase 1: Scalability and Performance Evaluation Phase (SPEP)

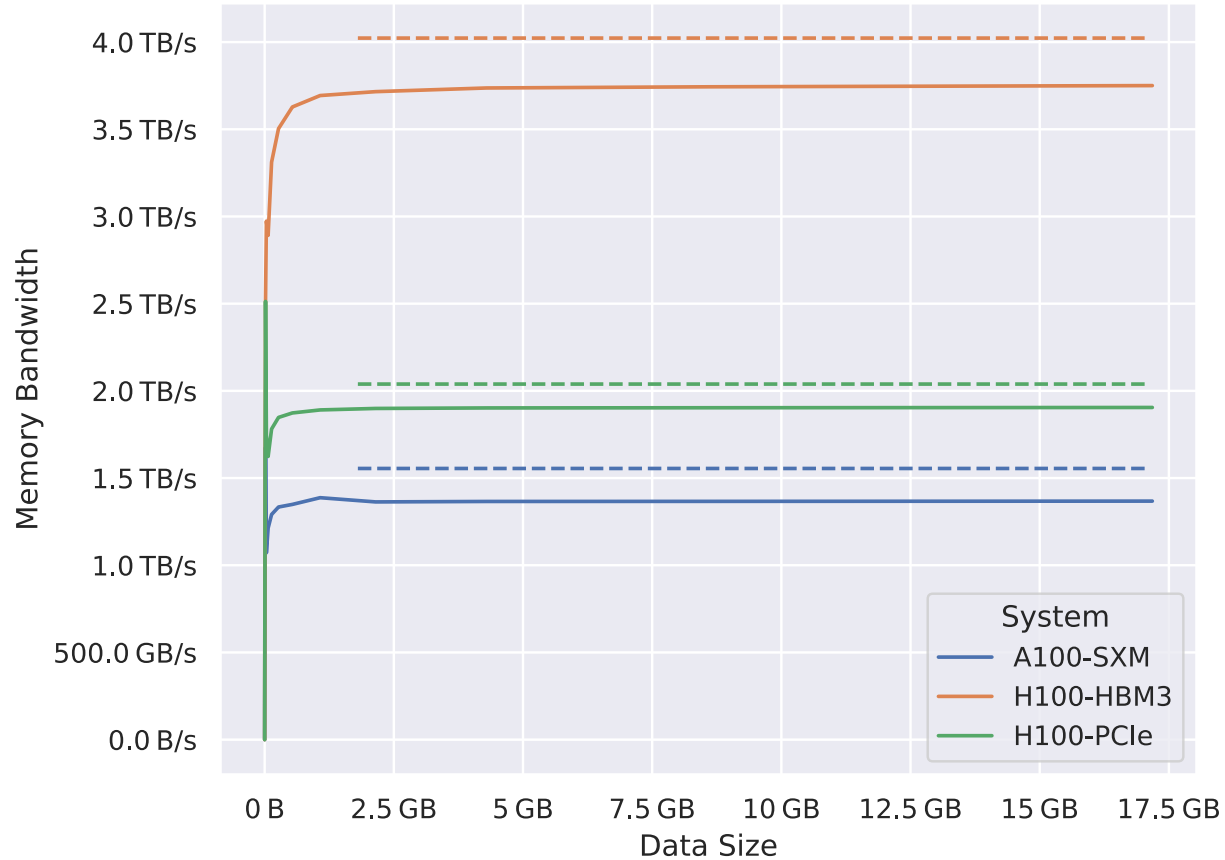
Phase 2: Porting and Optimizing Phase (POP)

Phase 3: Outstanding Research Phase (ORP)

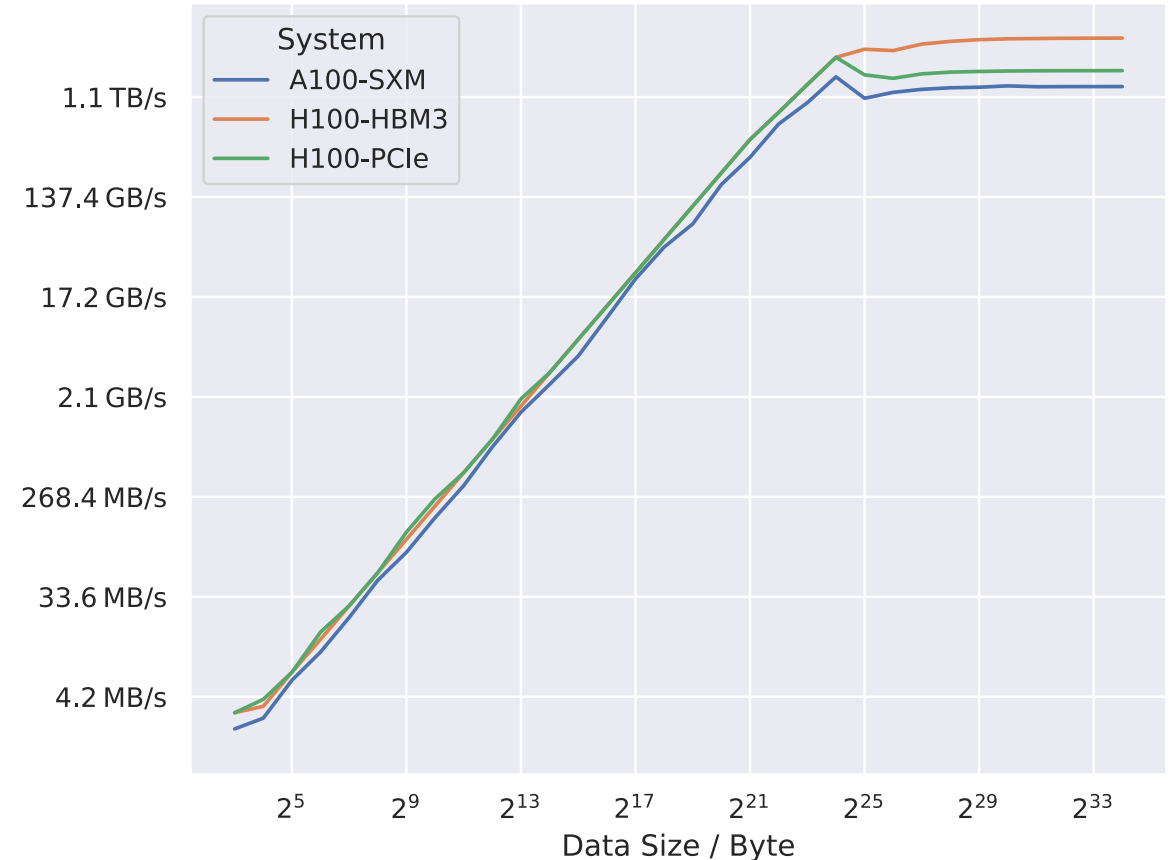
GPU STREAM

GPU STREAM Variant Scan for GPU Generations/Flavors

Linear View



Double-Logarithmic View

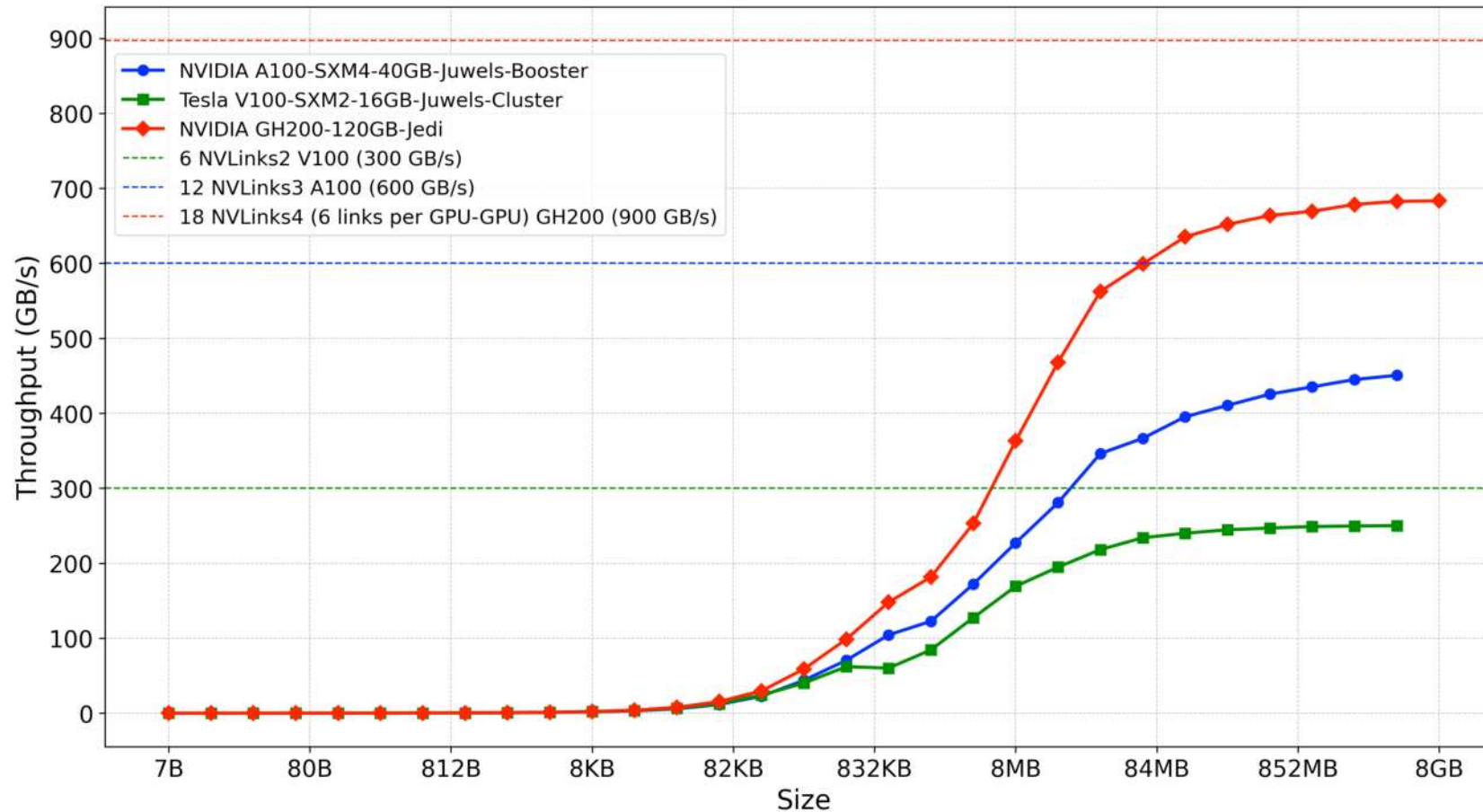


<https://github.com/AndiH/CUDA-Cpp-STREAM>

NCCL TESTS (GPU-GPU)

By Javad Kasravi / JSC

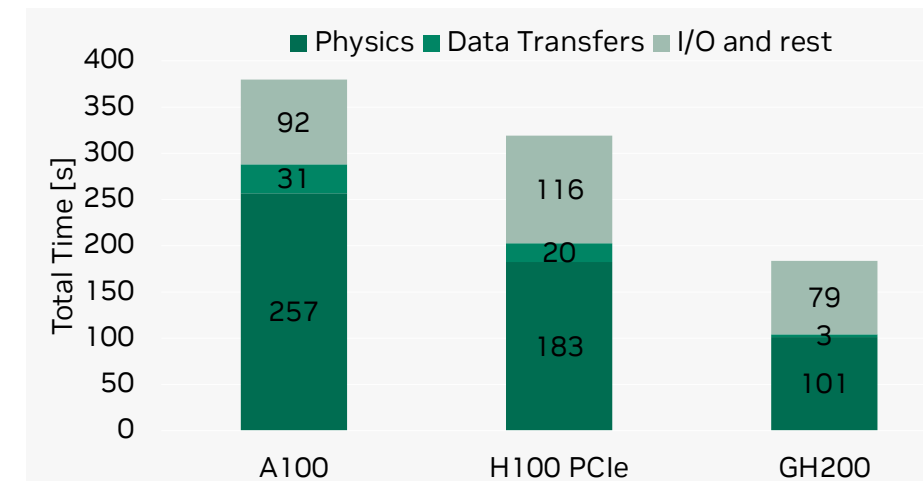
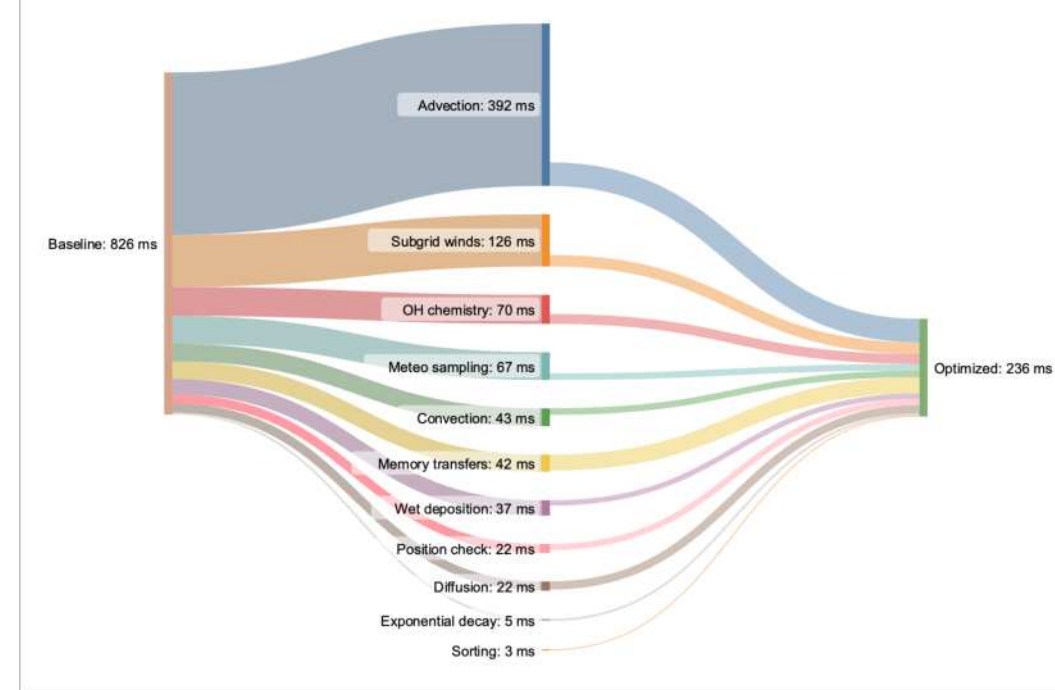
Intra-node connection speed (single node)



MPTRAC

By Lars Hoffmann / JSC

- Lagrangian particle dispersion model:
atmospheric transport processes
(troposphere/stratosphere)
→ volcanic emissions
- Continuously optimized for GPUs
Recently: Significant speedup on A100
- First test on GH200



See also [GTC talk by Mathias Wagner](#)

FIRST GPU EXPERIENCES (H100)



LQCD benchmark: Great mem utilization

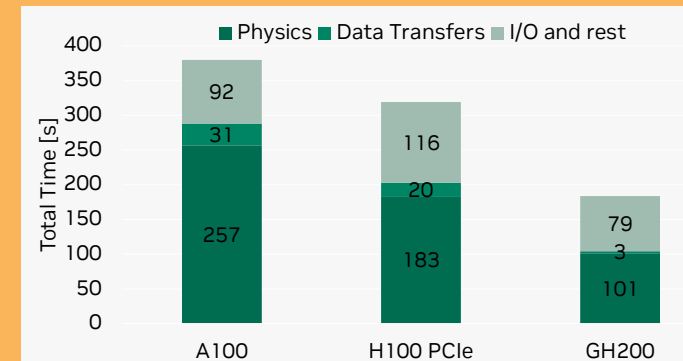
ChASE: >2× vs. A100
across all solvers

ICON: 1.6× vs. A100
in first benchmark (R2B4)

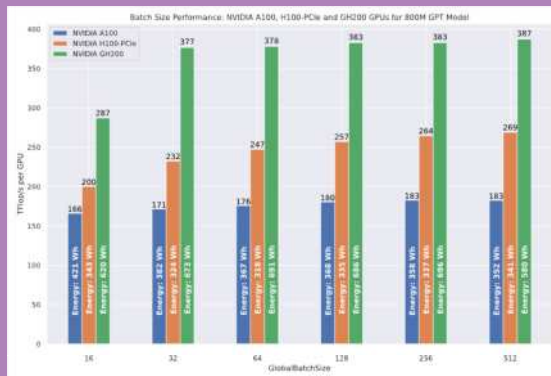
nekRS: 2.1× vs. A100
for RBC benchmark

Arbor: 1.97× vs. A100
for Busyring benchmark

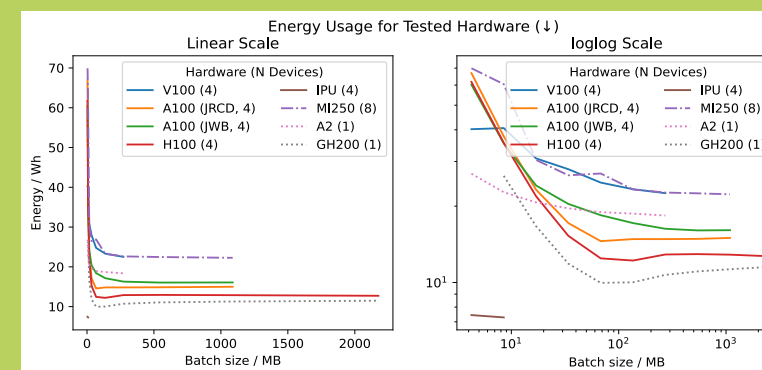
JUQCS: 2.6× vs. A100
for 31 Qubits



MPTRAC: >2× vs A100



LLM benchmark: >2× vs A100



MAELSTROM AP1: >6× energy efficiency vs. A100

FIRST CPU INVESTIGATIONS (*GRACE*)

- Focus mostly on GPU currently
 - Some first results on Grace hardware
- Very competitive performance, especially wrt TDP (but still early)

DynQCD: 1.5× vs. EPYC Rome 7742 (2×64 cores)

- Best: Grace-Clang, ACfL
- Slightly worse: GCC
- Investigating FMLA instructions
- (*Auto-Vectorization works well!*)

NAStJA:

- 2.3× vs. EPYC Rome 7402 (2×24 cores)
- 5.6× vs Intel Skylake 8168 (2×24 cores)

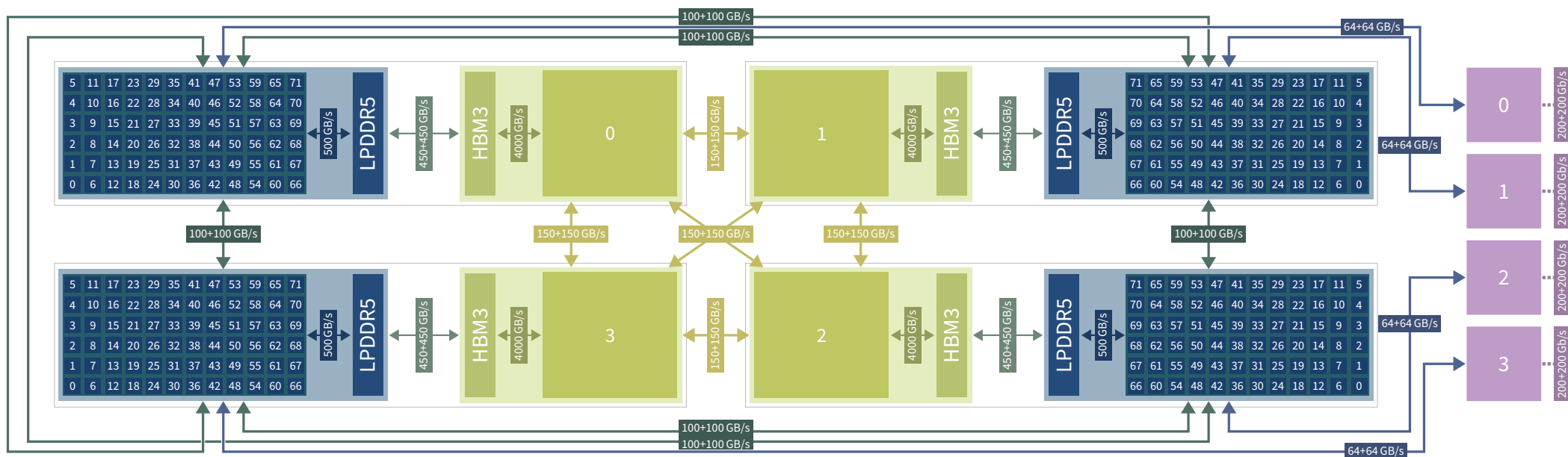
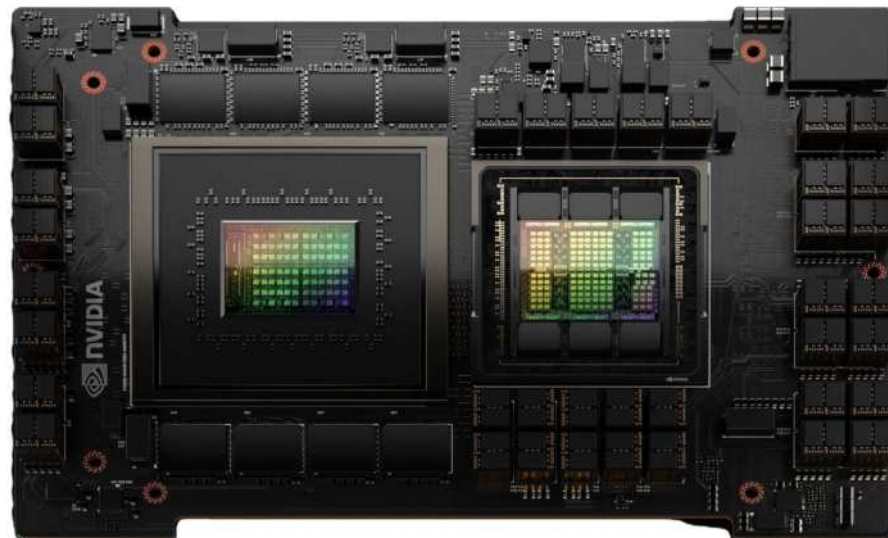
JUQCS: 1.35× vs. EPYC Rome 7402 for 31 Qubits (2×24 cores)

FLEUR:

- 1.2× vs. Intel Skylake 8168 (2×24 cores, 400 W TDP)
- 0.8× vs. EPYC Rome 7742 (2×64 cores, 450 W TDP)
- 1.5× vs. Intel SPR-HBM (2×32 cores, 700 W TDP)

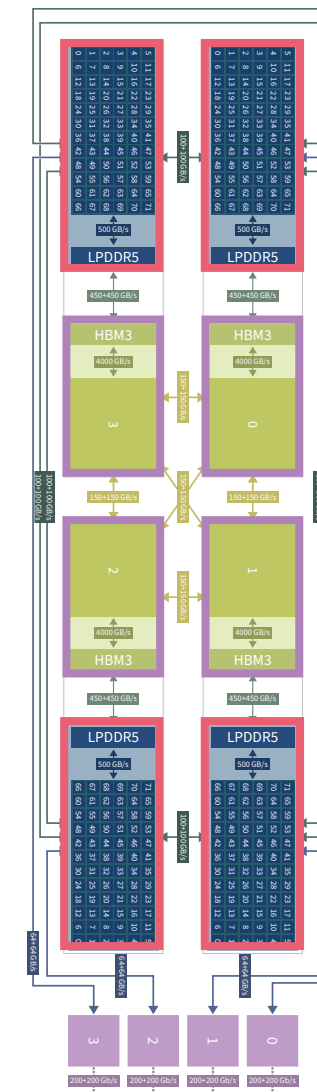
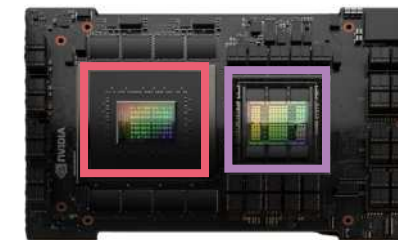
MAX DRIVING THE EXASCALE TRANSITION

NVIDIA GH200

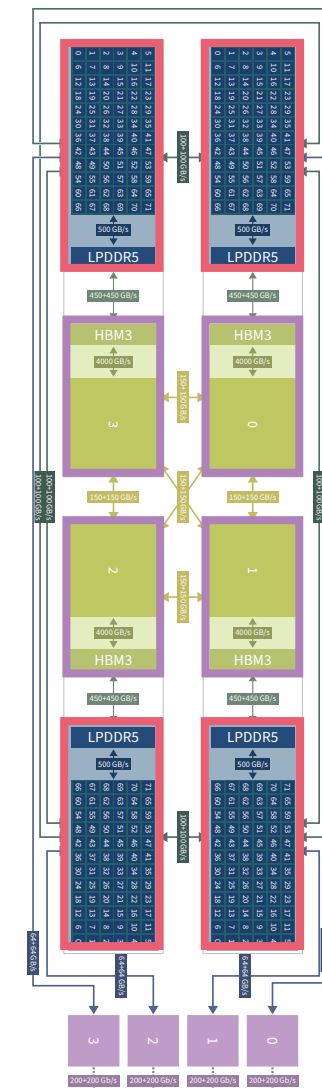
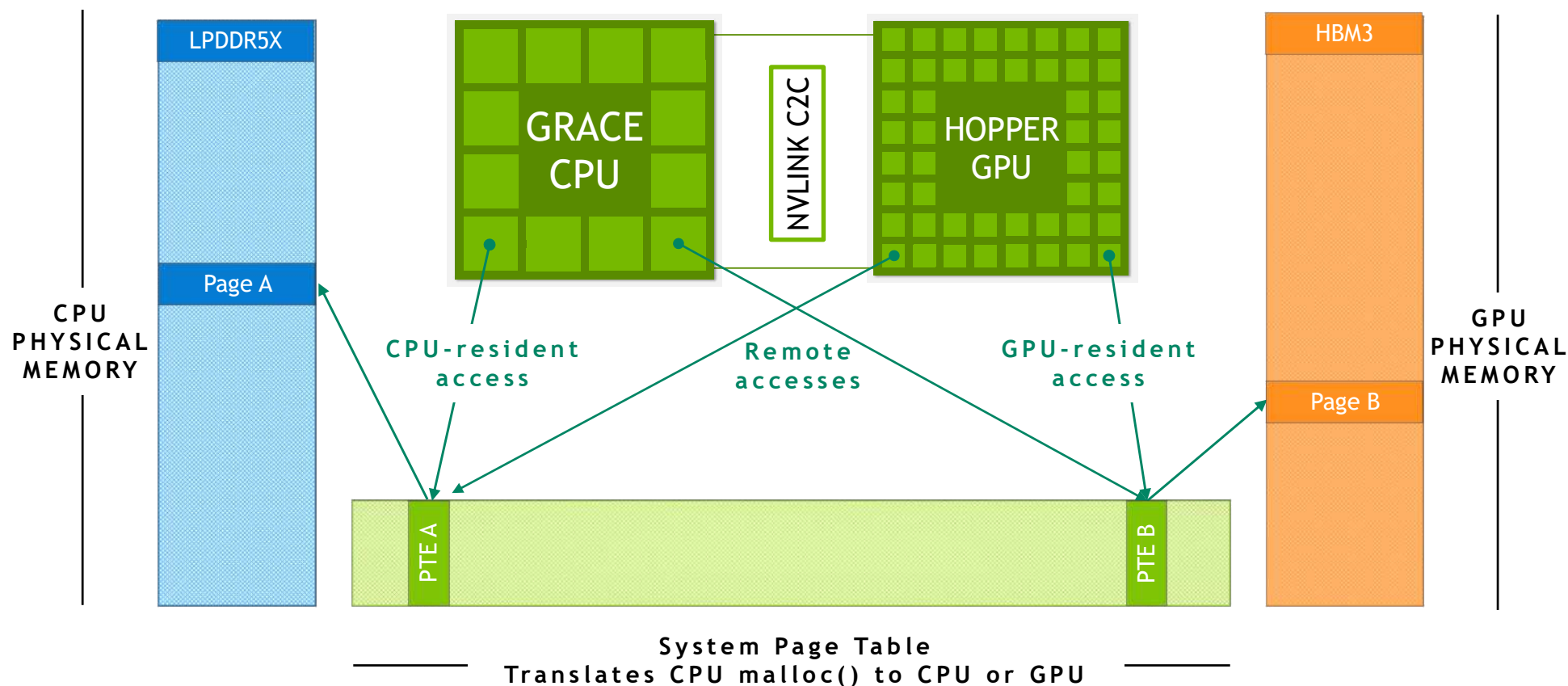
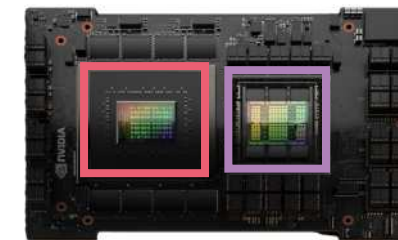


NVIDIA GH200

- “Superchip”: CPU and GPU in a package
- 4 × **CPU**: NVIDIA Grace, 72 cores (4×128b SVE2): ~ 3.6 TFLOP/s FP64; 120 GB 500 GB/s (L1: 64 kB+64 kB; L2: 1 MB; L3 (shared) 117 MB)
- 4 × **GPU**: NVIDIA Hopper, 132 multiprocessors (128 cores): ~60 TFLOP/s FP64; 96 GB 4000 GB/s (L1: 256 kB; L2 (shared) 60 MB)
- Memory-consistent connections: **CPU-GPU** (900 GB/s), **GPU-GPU** (300 GB/s), **CPU-CPU** (200 GB/s); NUMA domains accessible
- Package: 680 W shared for CPU+GPU; currently, CPU-focused (max 300 W)

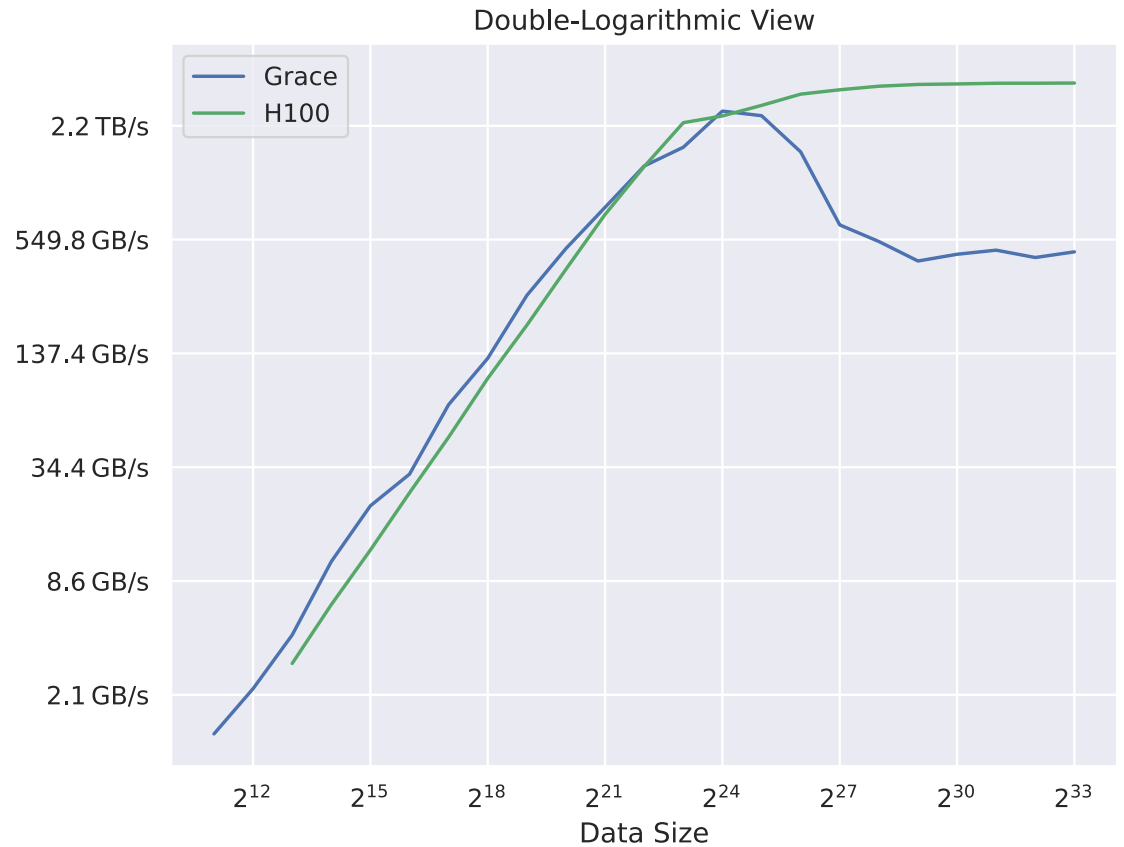
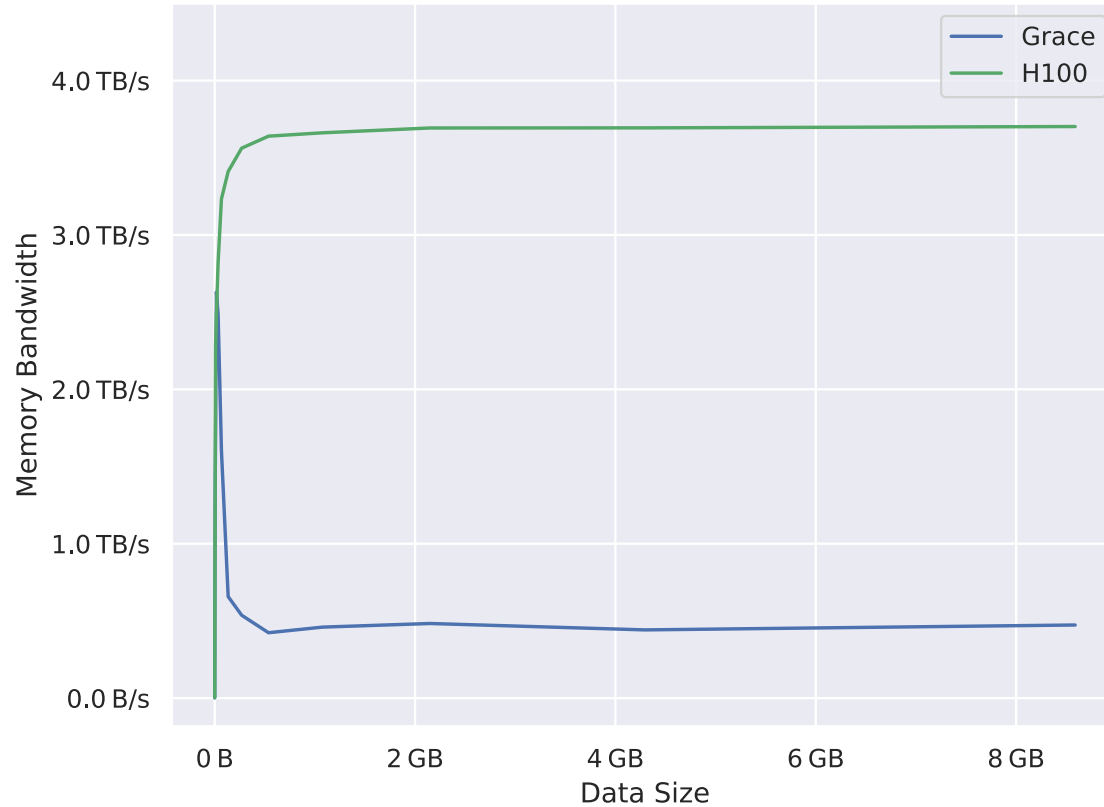


NVIDIA GH200



MEMORY PERFORMANCE

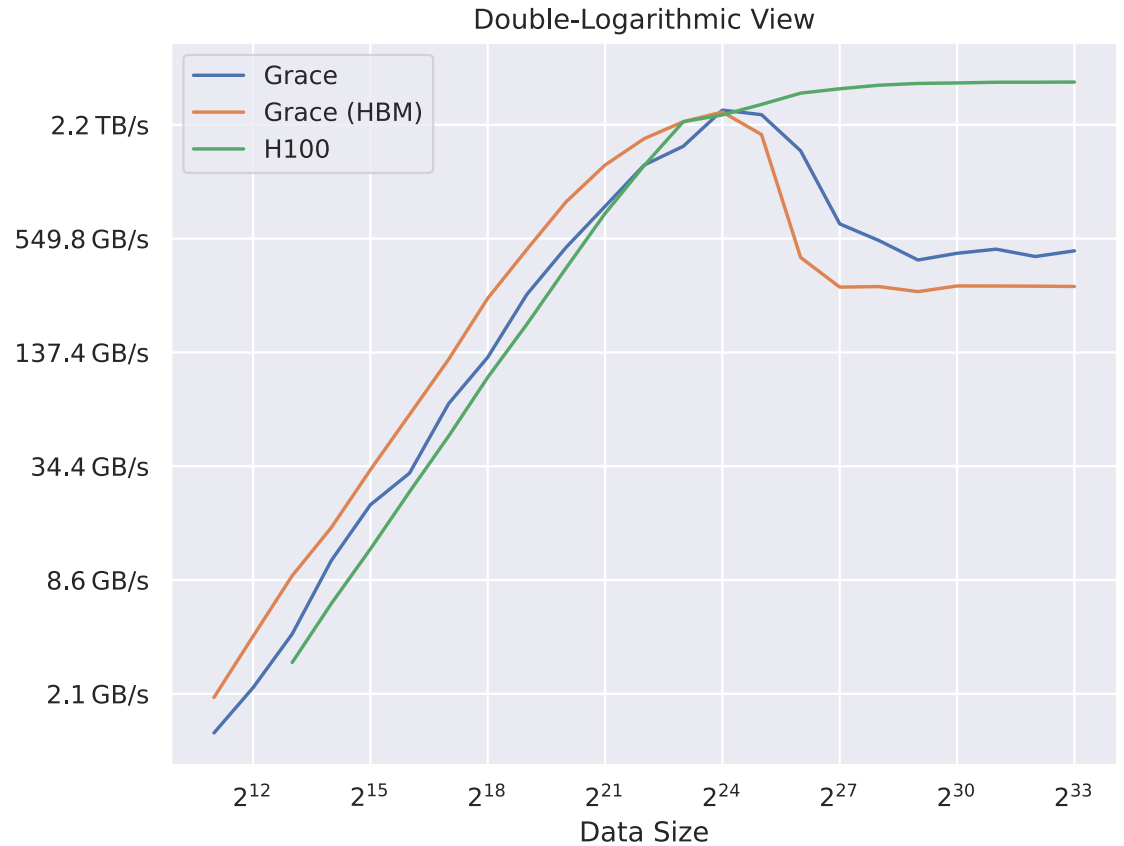
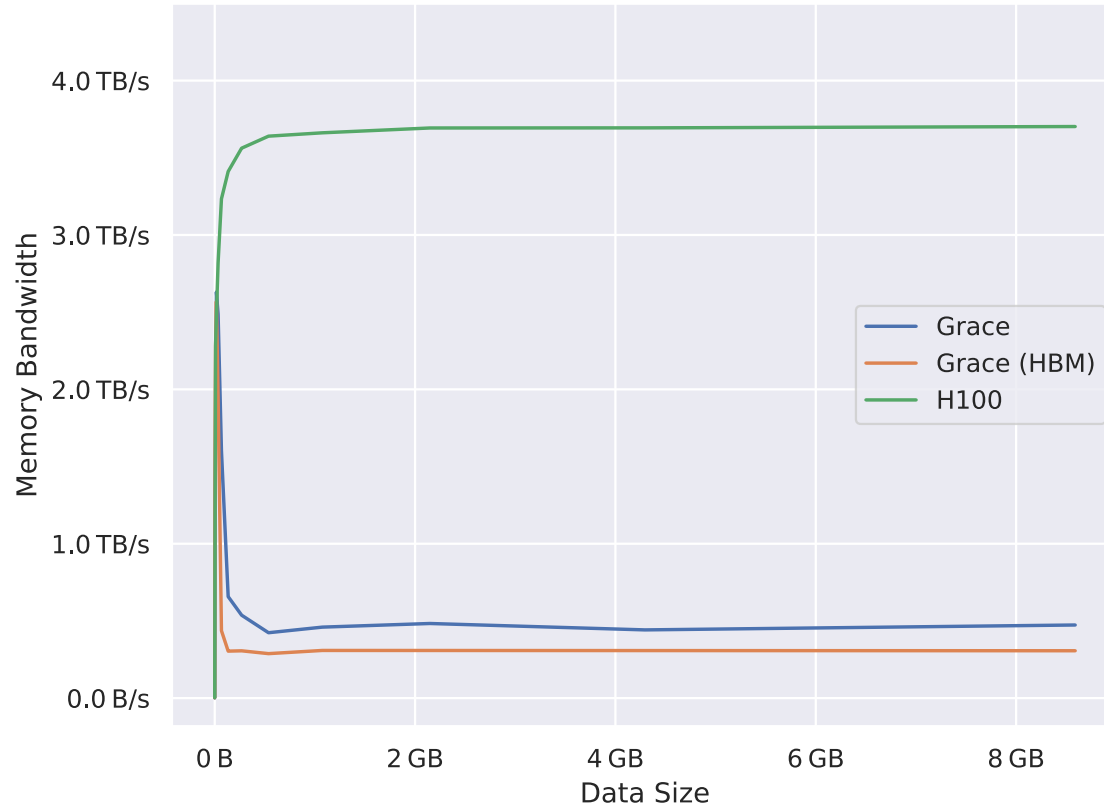
GPU STREAM Variant Scan for GH200 Superchip



Plots of variant of STREAM memory benchmark (BabelStream) using one GH200 Superchip.
Memory size (x axis) increasing in powers of two, from 2^{13} to 2^{33} .
Values in Byte/s (1 kB = 1000 B). Software versions: CUDA 12.2.0, driver 560.35.03.

MEMORY PERFORMANCE

GPU STREAM Variant Scan for GH200 Superchip



Plots of variant of STREAM memory benchmark (BabelStream) using one GH200 Superchip.
Memory size (x axis) increasing in powers of two, from 2^{13} to 2^{33} .
Values in Byte/s (1 kB = 1000 B). Software versions: CUDA 12.2.0, driver 560.35.03.

GPUS

CPU VS. GPU



CPU



GPU

CPU VS. GPU



CPU

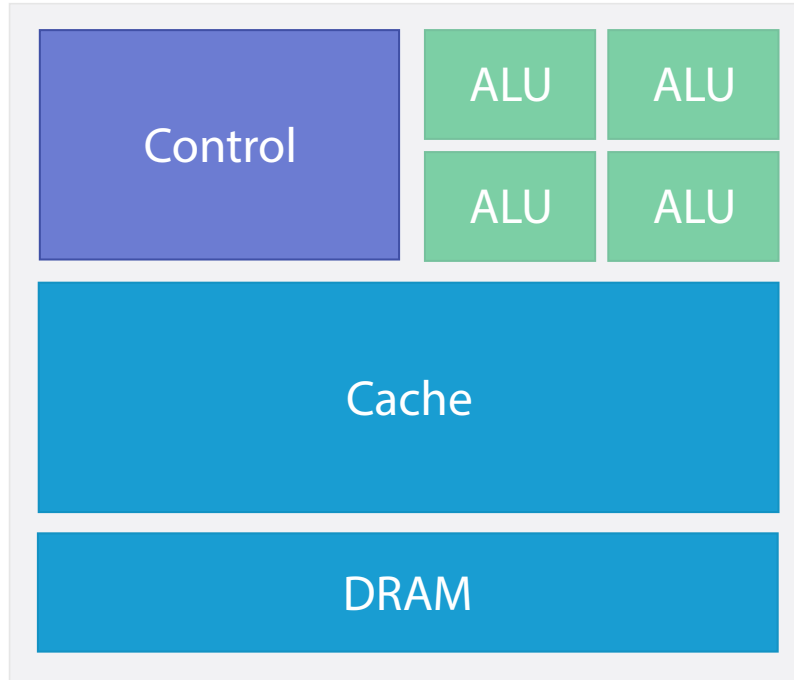


GPU

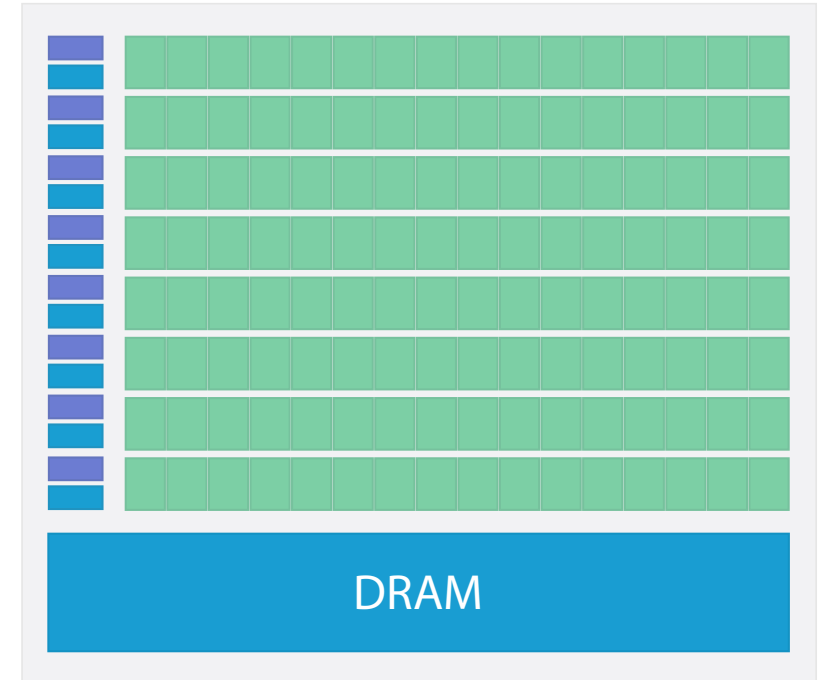


H100

CPU VS. GPU

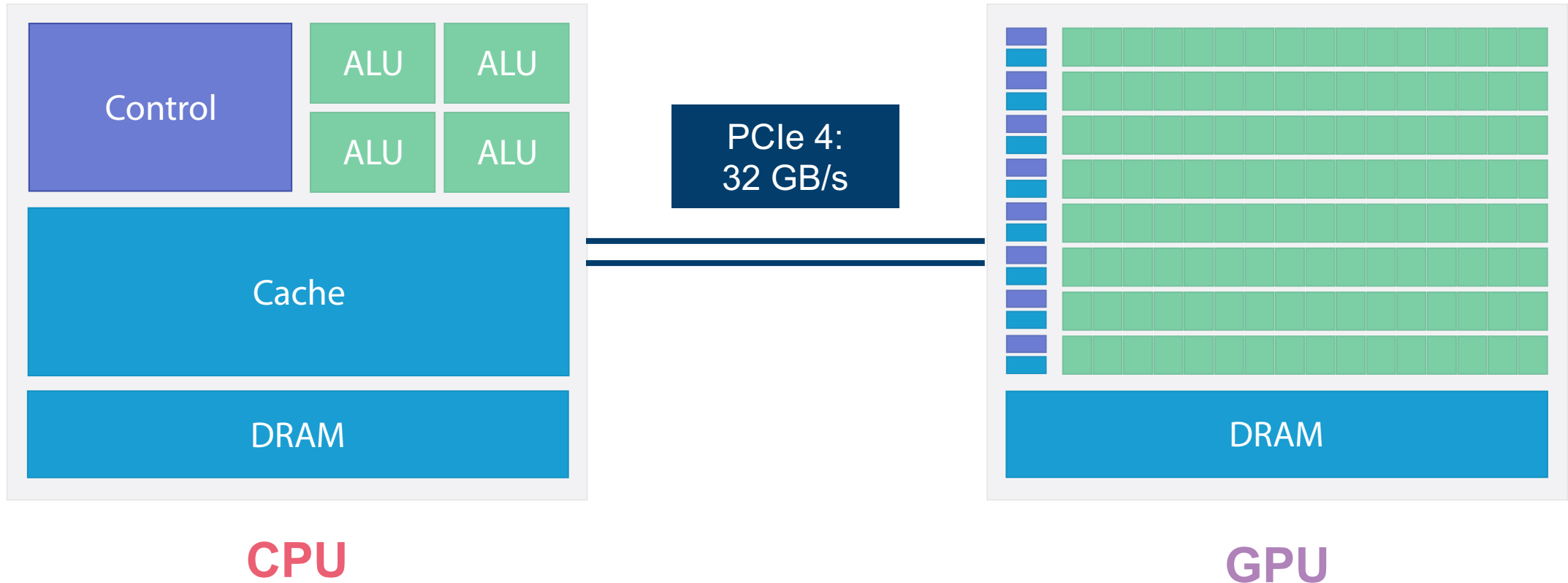


CPU

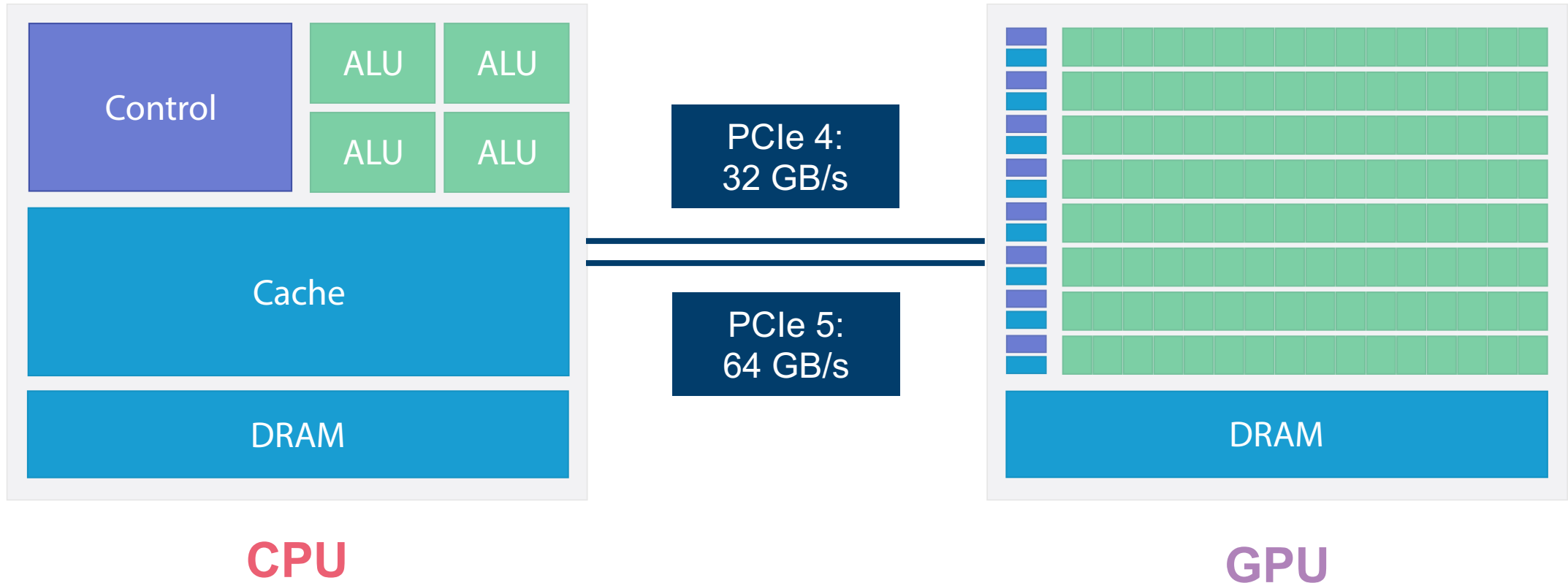


GPU

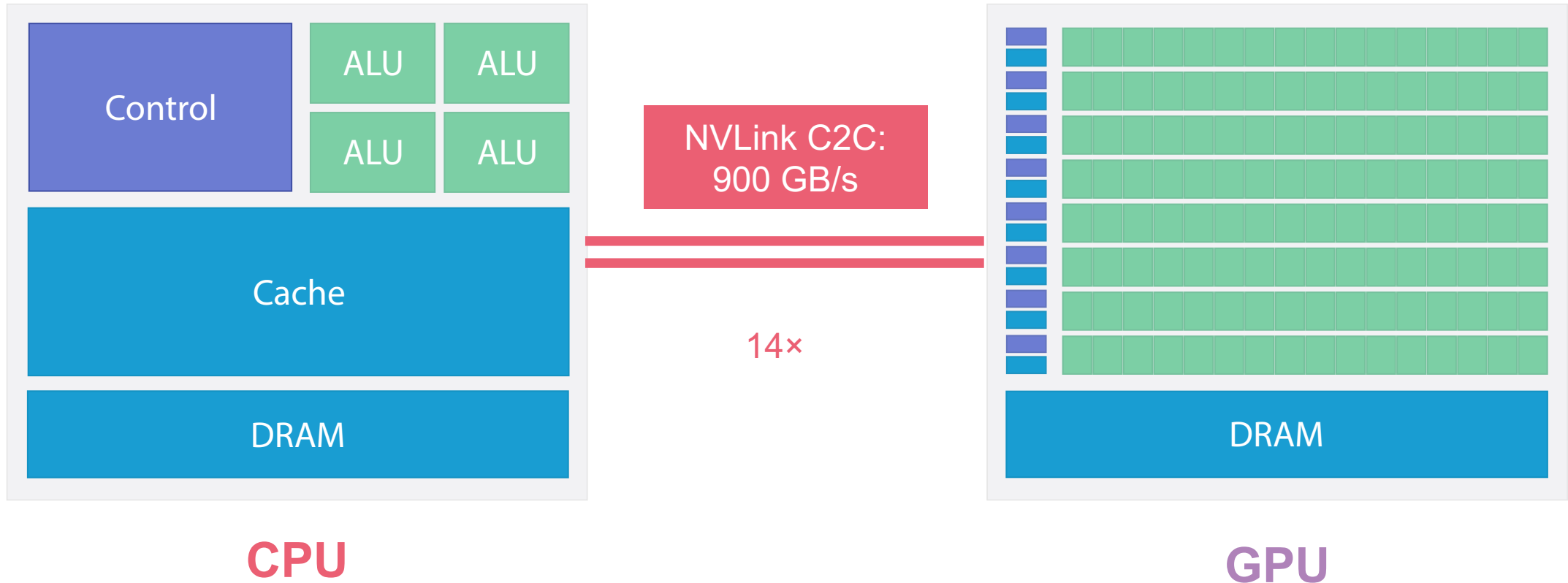
CPU VS. GPU



CPU VS. GPU

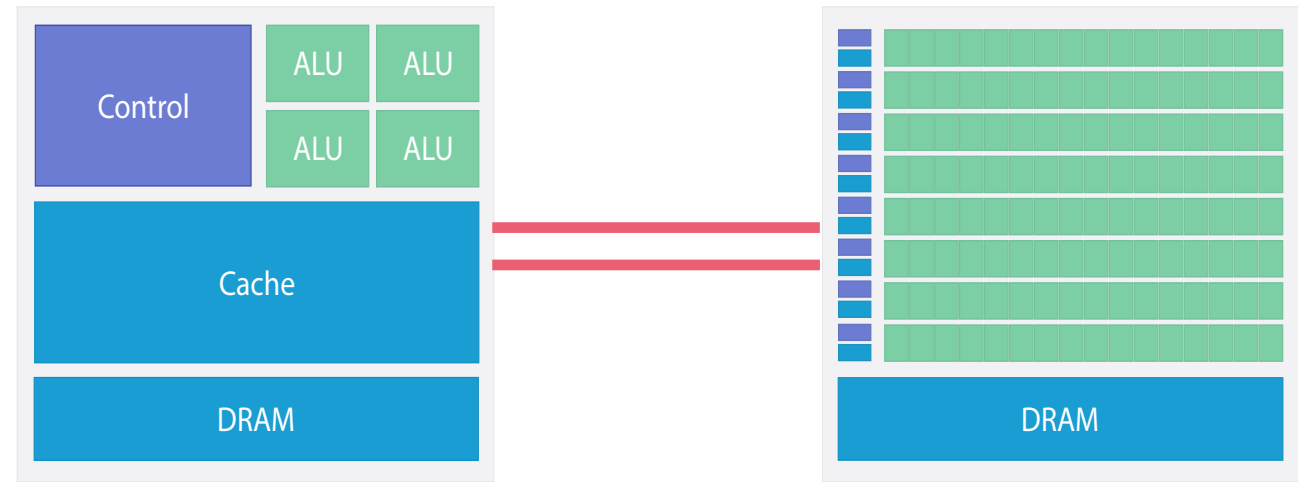


CPU VS. GPU



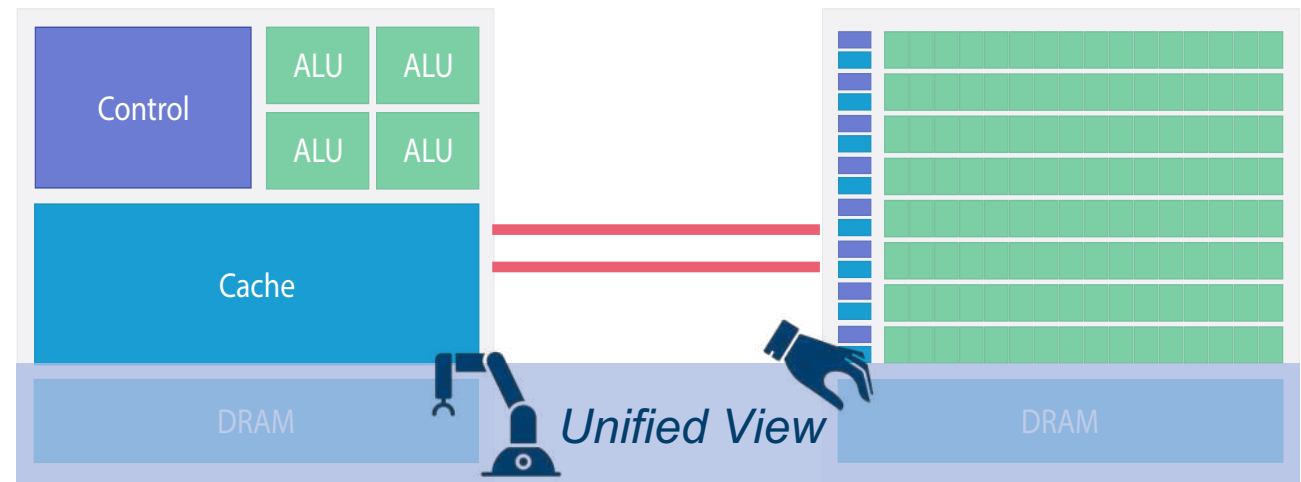
MEMORY

- Physically-different memory spaces
- Transfer memory via CPU-GPU bus
→ **Bottleneck**



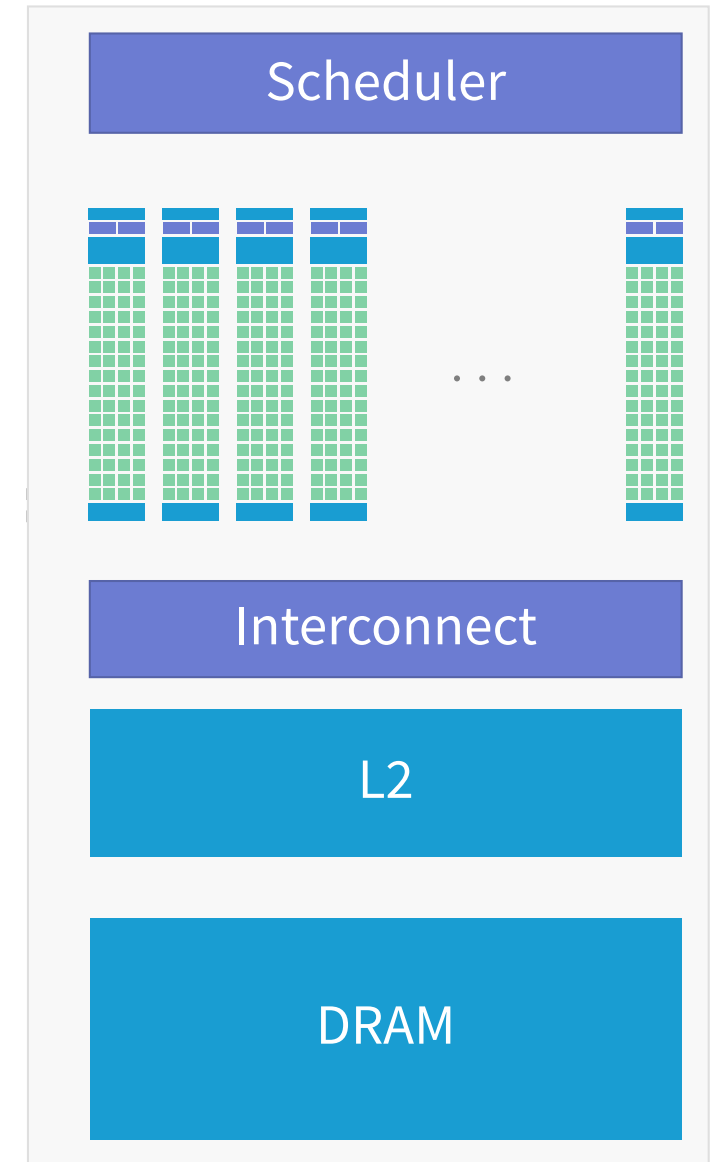
MEMORY

- Physically-different memory spaces
- Transfer memory via CPU-GPU bus
→ **Bottleneck**
- Transfer: **Manual** or **automatic**
 - **Manual**: Explicit API methods to move data (in bulk) at well-defined program locations
 - **Automatic**: Allocate memory with capable APIs → transfer on demand
 - Different levels of automatic-ness
 - Different overheads
 - **GH200**: Most converged *Unified Memory* implementation (hardware, software)



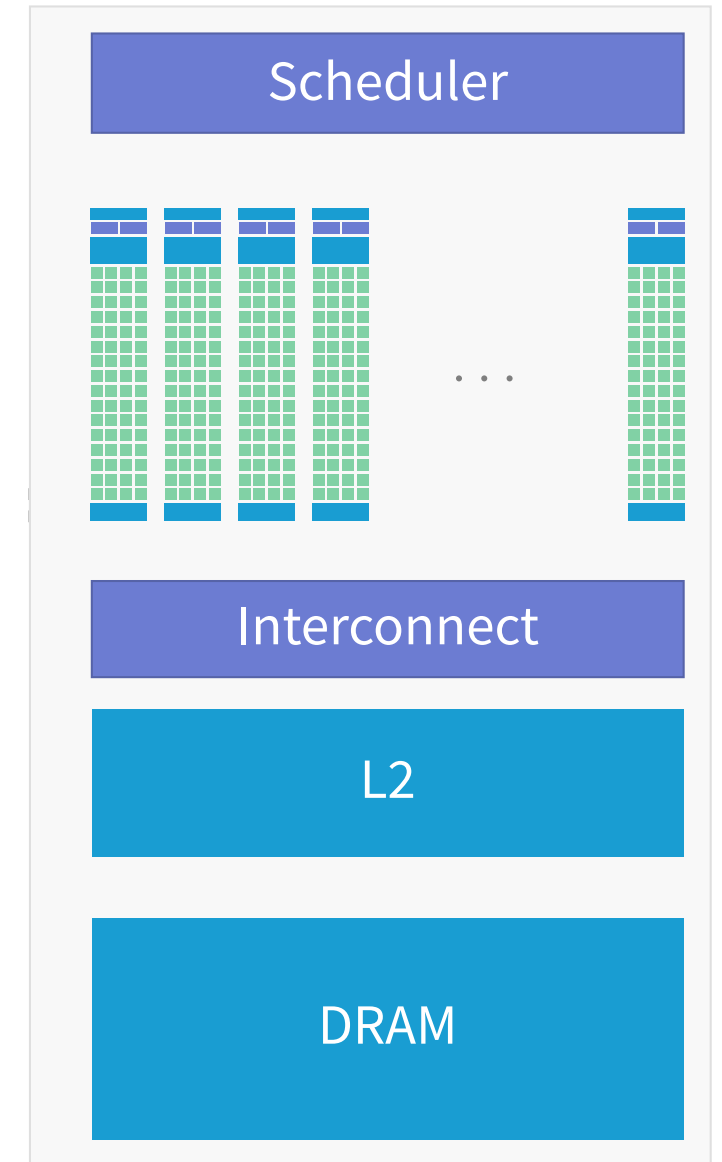
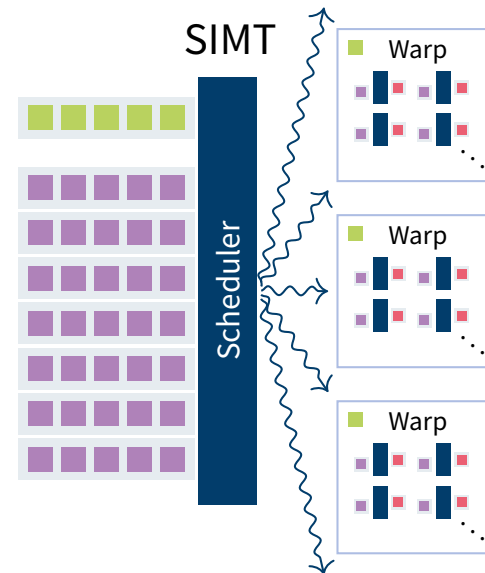
GPU OPERATION MODE

- Load data to GPU memory
- Load instructions to scheduler
- Execute on multiprocessor
- Retrieve data from GPU memory



GPU OPERATION MODE

- Load data to GPU memory
- Load instructions to scheduler
- Execute on multiprocessor
- Retrieve data from GPU memory
- Operation method:
Single Instruction, Multiple Threads
 - Mental model: operate with *threads* on individual data elements
 - Parallel function: `kernel<<<, >>>`
 - Kernel executed on multiprocessor



THREAD EXECUTION

- Explicit for loop → implicit threads
- CPU Core \cong GPU Multiprocessor
- 32 threads execute in lock-step (AMD: 64)
- Overlap compute, transfer
- → Expose parallelism in code

```
void scale(float scale, float * in, float * out, int N) {  
    for (int i = 0; i < N; i++)  
        out[i] = scale * in[i];  
}
```



```
__global__ void scale(float scale, float * in, float * out, int N) {  
    int i = threadIdx.x + blockIdx.x * blockDim.x;  
    if (i < N)  
        out[i] = scale * in[i];  
}
```

GPU BLOCK DIAGRAM



- Shared L2 Cache
- Building blocks: multiprocessors (80)

GPU BLOCK DIAGRAM



- Shared L2 Cache
- Building blocks: multiprocessors (80)

Compute elements for FP64, FP32, Int, Matrix

GPU BLOCK DIAGRAM



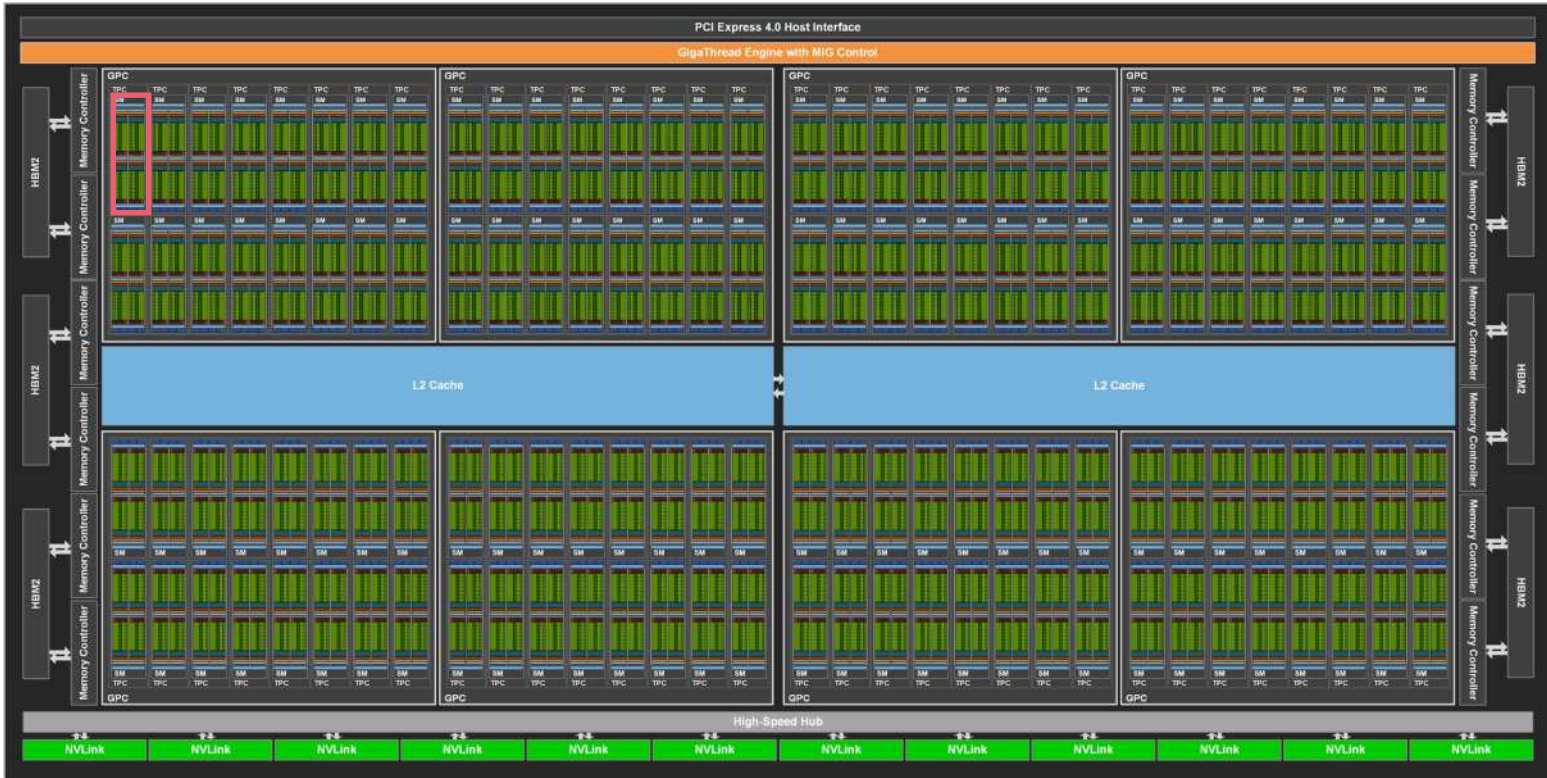
- Shared L2 Cache
- Building blocks: multiprocessors (80)

V100



Compute elements for FP64, FP32, Int, Matrix

GPU BLOCK DIAGRAM



- 108 multiprocessors
- 1.48 GHz (before: 1.53 GHz)

A100

TC (FP64): 64 FMAs / cyc

GPU BLOCK DIAGRAM

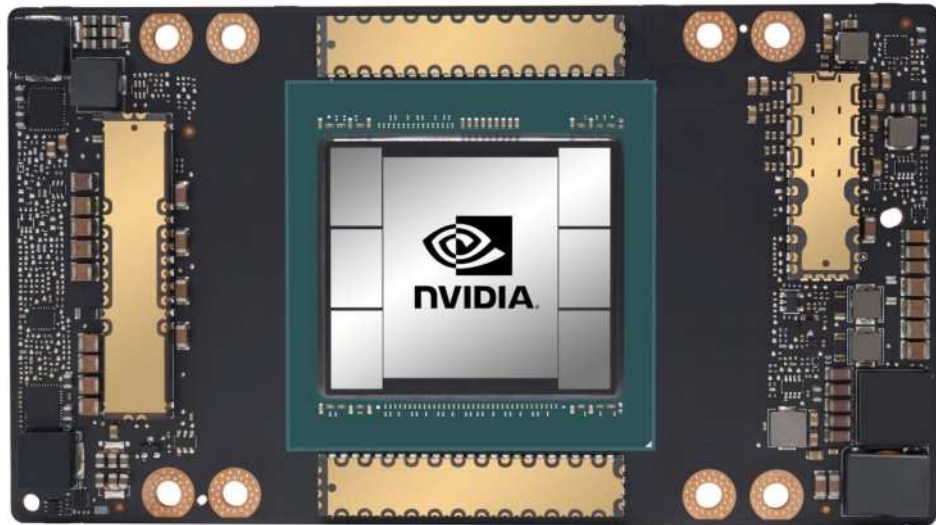


- 132 multiprocessors (PCIe: 114)
- 1.83 GHz

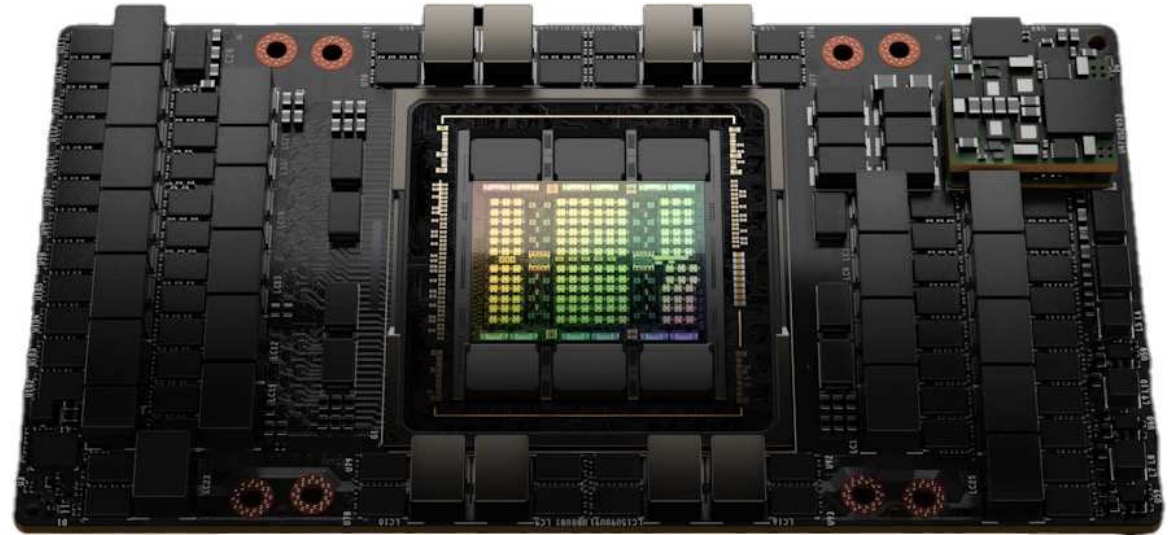
H100

TC (FP64): 2x64 FMAs / cyc

PICTURES

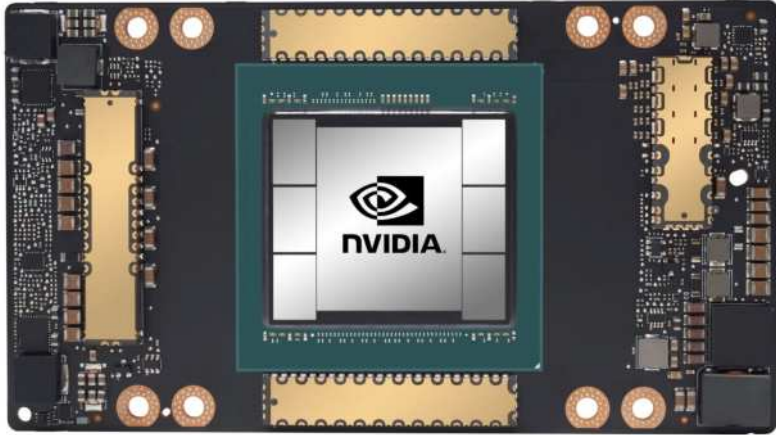


A100

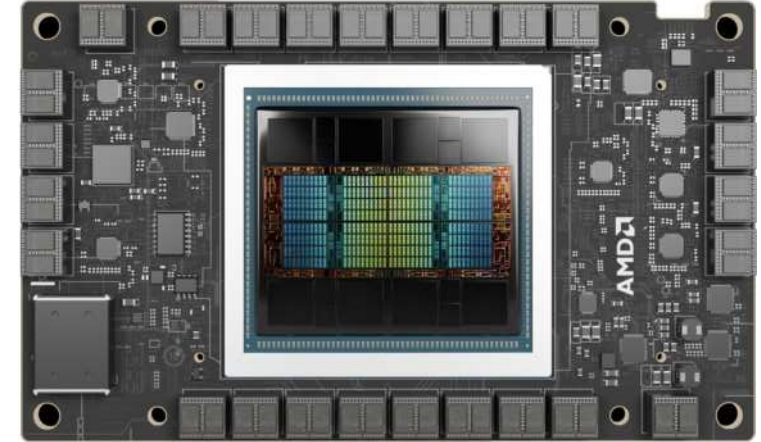


H100

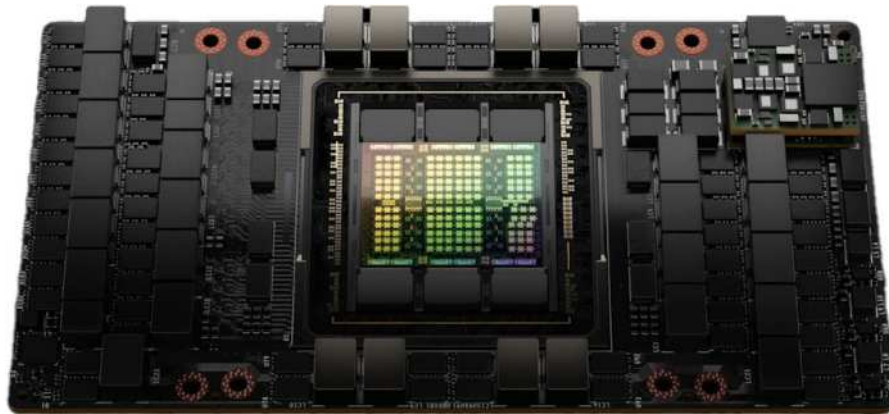
PICTURES



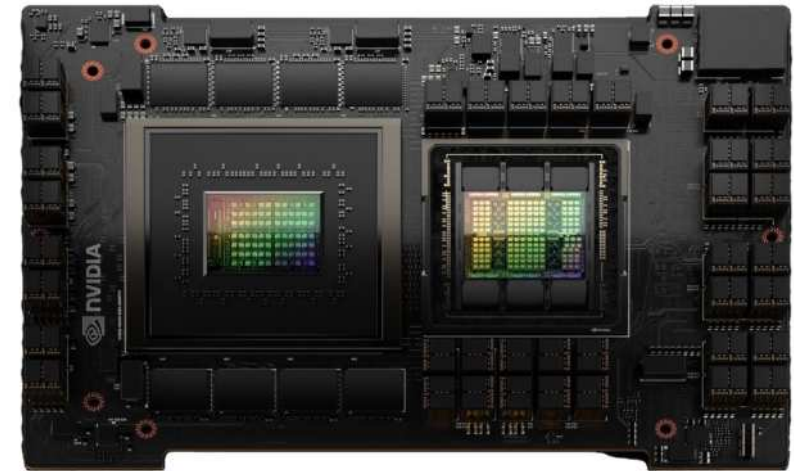
A100



MI300X



H100



GH200

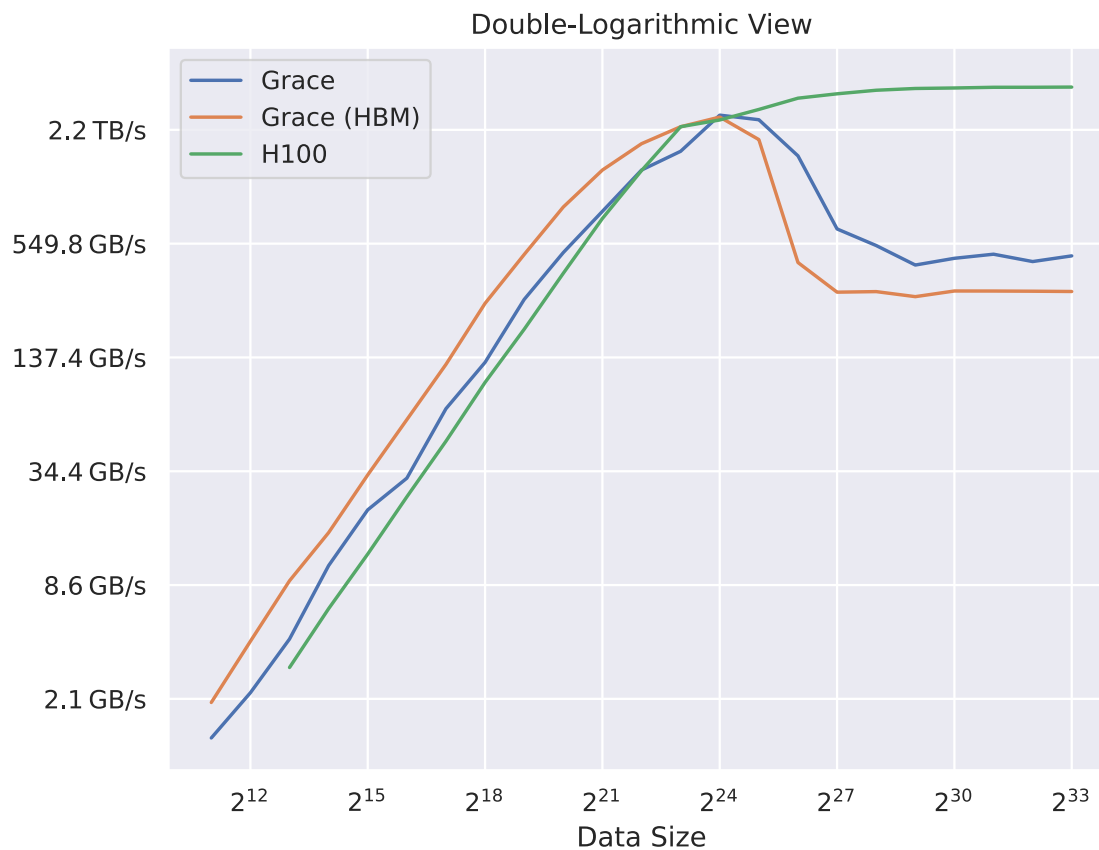
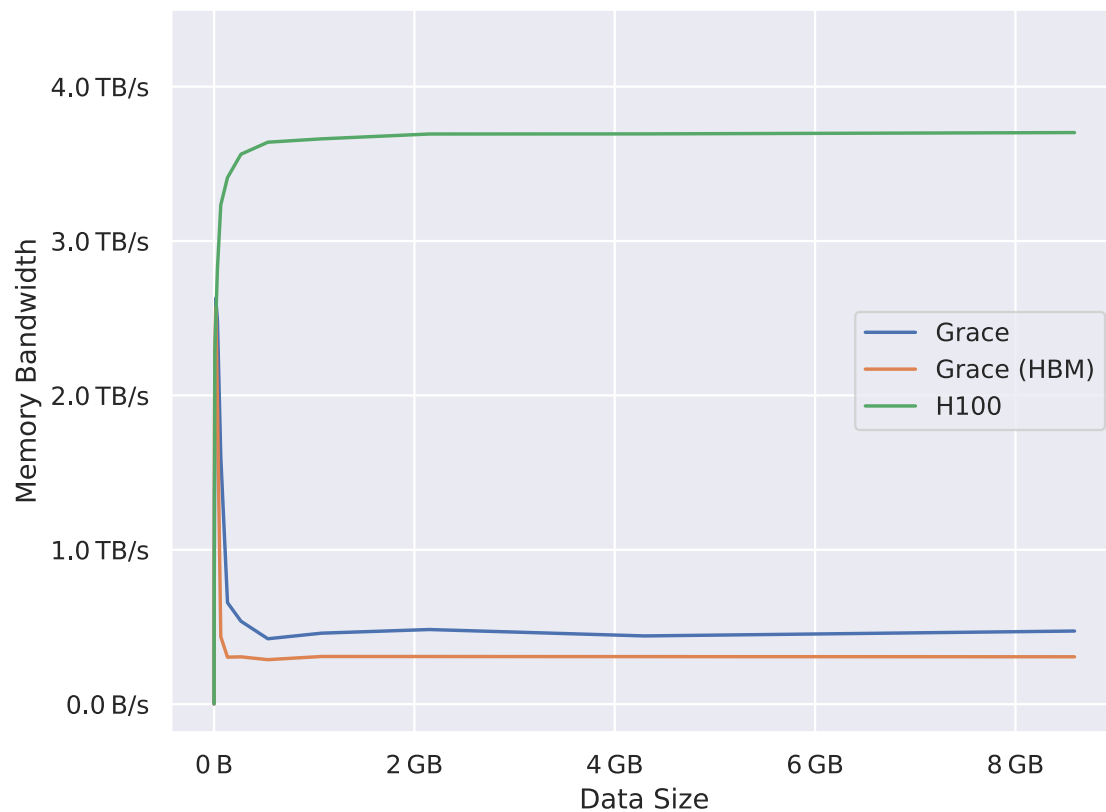
PERFORMANCE

- Performance is a matter of precision, type of compute
- Sparsity: 2x

	FP64 (Vec)	FP64 (Matrix)	FP32* (Matrix)	FP16 (Matrix)	Memory
	<i>TFLOP/s</i>				<i>TB/s</i>
A100	9.7	19.5	156	312	1.6
H100	33.5	67	495	989	3.3
GH200					4
MI300X	82	163	654	1307	5.3

MEMORY PERFORMANCE

GPU STREAM Variant Scan for GH200 Superchip



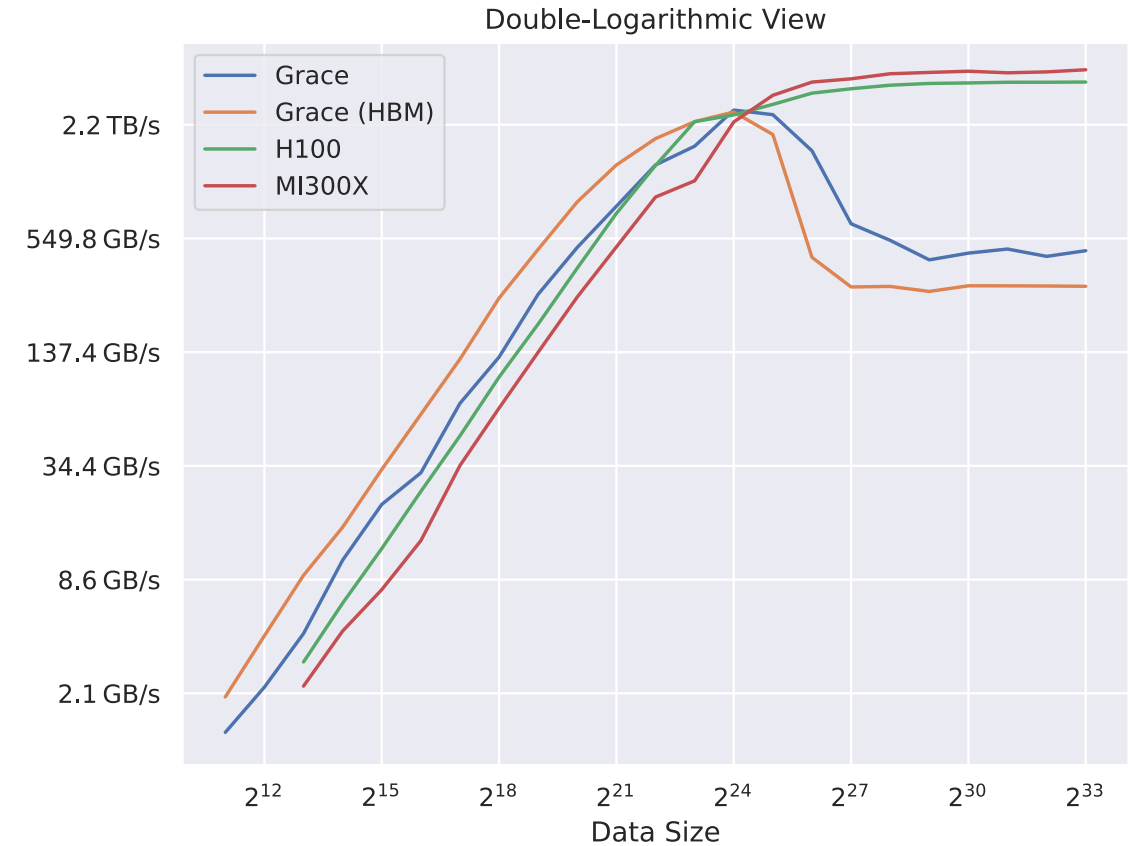
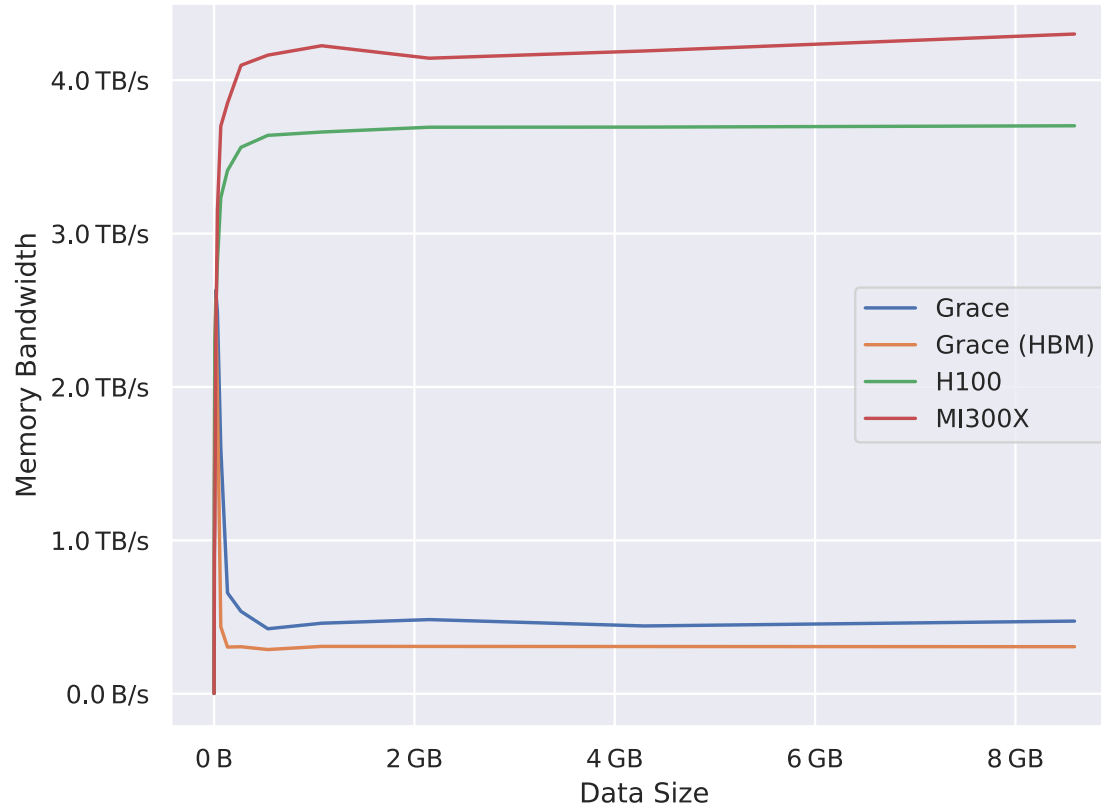
Plots of variant of STREAM memory benchmark (BabelStream) using one GH200 Superchip and one MI300X GPU.

Memory size (x axis) increasing in powers of two, from 2^{13} to 2^{33} .

Values in Byte/s (1 kB = 1000 B). Software versions: CUDA 12.2.0, driver 560.35.03; ROCm 6.8.5.

MEMORY PERFORMANCE

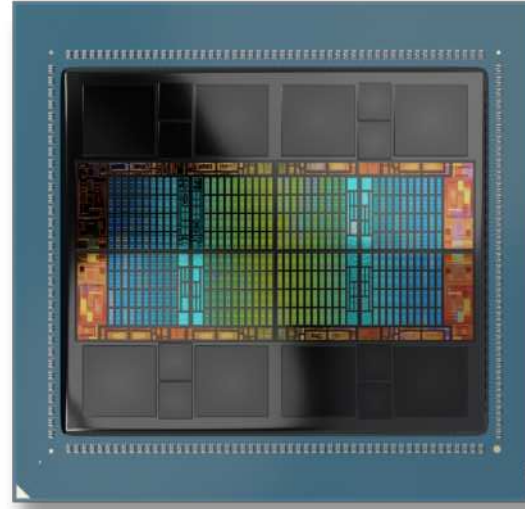
GPU STREAM Variant Scan for GH200/MI300X



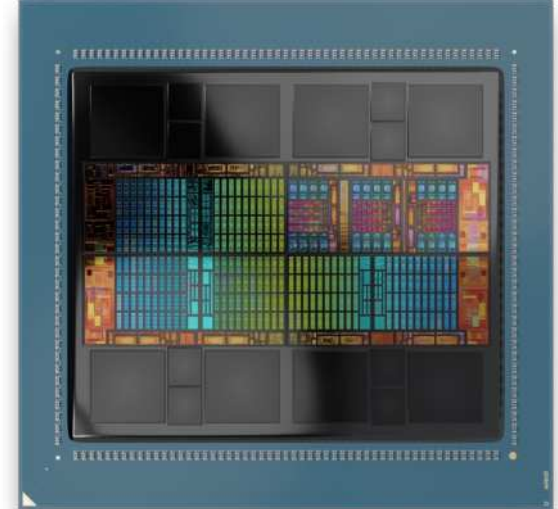
Plots of variant of STREAM memory benchmark (BabelStream) using one GH200 Superchip and one MI300X GPU.
Memory size (x axis) increasing in powers of two, from 2^{13} to 2^{33} .
Values in Byte/s (1 kB = 1000 B). Software versions: CUDA 12.2.0, driver 560.35.03; ROCm 6.8.5.

AMD MI300A, MI300X

- AMD's current flagship GPU
- Two variants
 - MI300X: Classical GPU; 128 GB HBM3
 - MI300A: APU with integrated CPU chiplet (Zen4, 24 cores); 192 GB HBM3



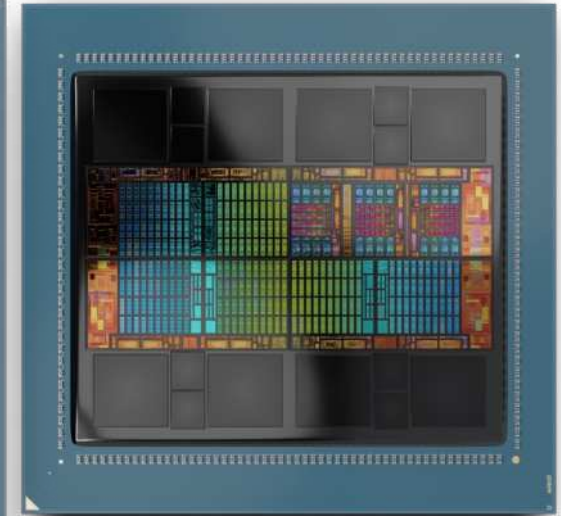
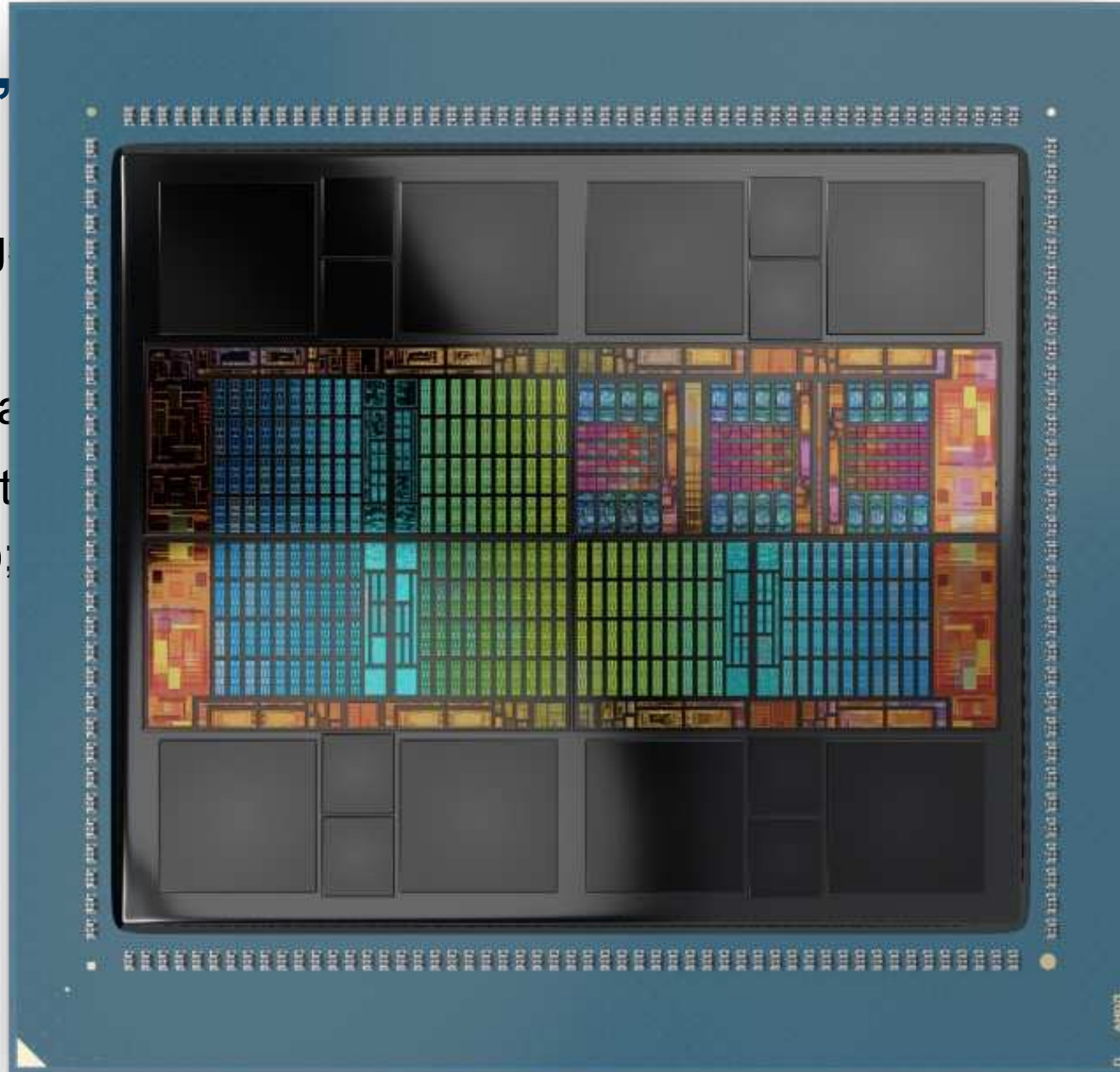
MI300X



MI300A

AMD MI300A,

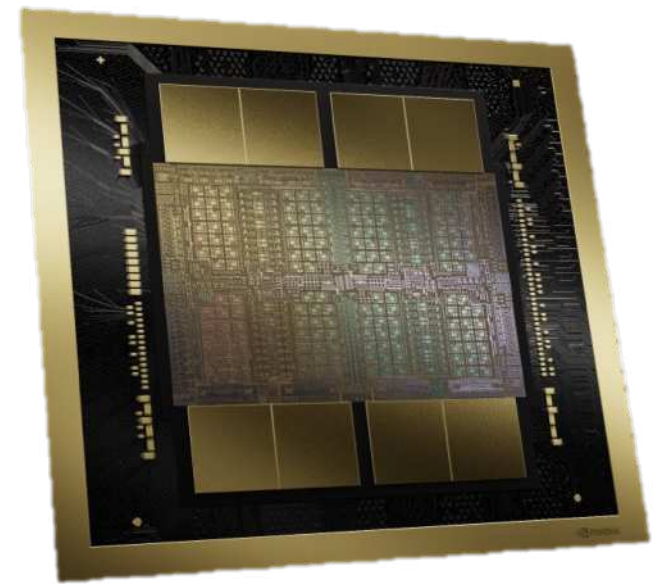
- AMD's current flagship
- Two variants
 - MI300X: Classical GPU
 - MI300A: APU with Zen4 (24 cores)



MI300A

NVIDIA BLACKWELL GPU

- Latest NVIDIA GPU, shipping soon™
- Grace-Blackwell GB200: 1 Grace, 2 Blackwell
- Blackwell: Fused GPU pair



	FP64 (Vec)	FP64 (Matrix)	FP32* (Matrix)	FP16 (Matrix)	Memory
	<i>TFLOP/s</i>				<i>TB/s</i>
A100	9.7	19.5	156	312	1.6
H100	33.5	67	495	989	3.3
GH200					4
B100	45	45	1250	2500	8?
MI300X	82	163	654	1307	5.3

PROGRAMMING GPUS

- Many programming models for GPUs, CPUs
- Different levels of abstraction, portability, performance-attainability, open-ness

Many Cores, Many Models: GPU Programming Model vs. Vendor Compatibility Overview

	CUDA		HIP		SYCL		OpenACC		OpenMP		Standard		Kokkos		ALPAKA		Python
	C++	Fortran	C++	Fortran	C++	Fortran	C++	Fortran	C++	Fortran	C++	Fortran	C++	Fortran	C++	Fortran	
NVIDIA	●1	●2	◐3	★4	▲5	↗6	●7	●8	◐▲9	◐●10	●11	●12	▲13	★14	▲15	↗16	●▲17
AMD	◐18	★19	●20	★4	▲21	↗6	▲22	▲★23	●▲24	●25	▲◐★26	↗27	▲28	★14	▲29	↗16	★30
Intel	◐▲31	↗32	▲33	↗34	●35	↗6	★36	★37	●38	●39	●◐40	●41	▲42	★14	▲43	↗16	◐44

Paper / HTML version at <https://go.fzj.de/gpumodels>



SUMMARY

- JUPITER: First European Exascale system; EuroHPC JU, BMBF, MKW-NRW; at JSC
- Booster: 24 000 Grace-Hopper CPU/GPU superchips
- Cluster: SiPearl Rhea1 CPU
- Applications, usability core to the design; large benchmarking suite, JEDI, JUREAP
- GPU: Massive parallel performance, throughput
- Programming: *tommorrow*

Thank you for your attention!
a.herten@fz-juelich.de

Talk features self-created imagery, as well as imagery from colleagues, and from Eviden, NVIDIA, SiPearl, AMD, IBM, OSM; plus individually marked imagery.

JUPITER

The Arrival of
Exascale in Europe

fz-juelich.de/jupiter | [#exa_jupiter](https://twitter.com/#!/exa_jupiter)



Ministry of Culture and Science
of the State of
North Rhine-Westphalia



GCS
Gauss Centre for Supercomputing

JOINING FORCES



Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen



EVIDEN



fz-juelich.de/jupiter