FROM A SINGLE CPU CORE TO THE MOON 🚀

CÉCI User Meeting – Louvain-la-Neuve - April 18th, 2024

👩‍💻 Orian Louant - 🏫 University of Liège - ✉️ orian.louant@uliege.be

Hello, I’m Orian

- *Logisticien de Recherche* CÉCI and member of the LUMI User Support Team
- Sometimes, I do some system administration but is far from main activity
- My work is mainly focused on application support
- My interests are in parallel programming in general. I’m more interested in the software side of HPC than the hardware side

What this talk is about

Advertize to you all the amazing hardware you can have access to as a CÉCI user and as a Belgian researcher

Non-parallel and shared-memory parallel workloads

🪲 Hercules2 - 🐉 Dragon1/2

Clusters optimized for jobs with limited parallelism or no parallelism at all

- Focus on single-core performance
- Long maximum walltime time

Distributed-memory massively parallel workloads

🌌 Lemaitre3/4 - 🏠 NIC5

Clusters optimized for jobs able use leverage a large number of CPU cores

- Fast interconnect between the nodes
- Short maximum walltime time

A few years ago, PRACE introduced the European model for HPC where systems are divided in three levels:

Tier-2	the computing capacity that is available at research institutions
Tier-1	the computing capacity that exceeds the capacity of an institution in terms of needs and costs and which is provided at the level of a region or country
Tier-0	the very large-scale computing infrastructures

As a CÉCI user, you have access to every supercomputer tier levels:

Tier-2	CÉCI core business
Tier-1	Lucia, operated by Cenaero, accessible with your CÉCI account
Tier-0	LUMI, operated by CSC (Finland, Belgium, is a member of the consortium) as well as other EuroHPC pre-exascale supercomputers

LUCIA

THE WALLOON TIER-1 SUPERCOMPUTER

ABOUT LUCIA

Lucia is the Walloon Tier-1 HPC cluster and features two main compute partitions:

- A CPU partition with AMD EPYC Milan CPUs
- A GPU partition with NVIDIA A100 GPUs

Lucia delivers 2.72 PFLOPS of sustained HPL-linpack performance (#322 in Nov. 23 Top500)

- Exploited by Cenaero, a private non-profit research center
- Universities and accredited research centers have an 85% share of the compute time
- Access is granted to CÉCI users via projects

    - **Access:** https://www.ceci-hpc.be/projetstier1.html
    - **Documentation:** https://doc.lucia.cenaero.be/
    - **Support (if you have an account):** https://support.lucia.cenaero.be/

LUCIA CPU NODES

300 nodes (38400 cores) with

- 2x AMD EPYC 7763 64-core CPUs
- 270 “standard” nodes with 256 GB of RAM
- 30 “medium” nodes with 512 GB of RAM
- 1x Infiniband HDR-100 interconnect

7 large memory nodes with

- 2x AMD EPYC 7513 32-core CPUs
- 2 TB of RAM
- 1x Infiniband HDR-100 interconnect

1 extra-large memory node with

- Same as the large memory nodes but with 4 TB of RAM

LUCIA DE BROUCKÈRE
(1904-1982)

Chemist. The first woman to receive an academic appointment at a Belgian Faculty of Science in 1937 (Université Libre de Bruxelles)

LUCIA GPU NODES

50 nodes with

- 1x AMD EPYC 7513 32-core CPU
- 4x NVIDIA A100 40GB GPUs
- 256 GB of RAM
- 2x Infiniband HDR-200 interconnect

2 nodes with

- 2x AMD EPYC 7513 32-core CPUs
- 8x NVIDIA A100 80GB GPUs
- 2 TB of RAM
- 2x Infiniband HDR-200 interconnect

Lucia data center is located at A6k-E6k, next to the central train station of Charleroi

ROLE OF LUCIA IN THE WALLOON HPC ECOSYSTEM

Lucia is optimized for larger jobs than NIC5 or Lemaitre3/4:

- On the tier-2 clusters you might need to wait a long time to get a entire compute node allocated to you
- On Lucia is expected your job will use a (multiple) compute node(s)

The level of support you can expect is not the same as for the other CÉCI clusters:

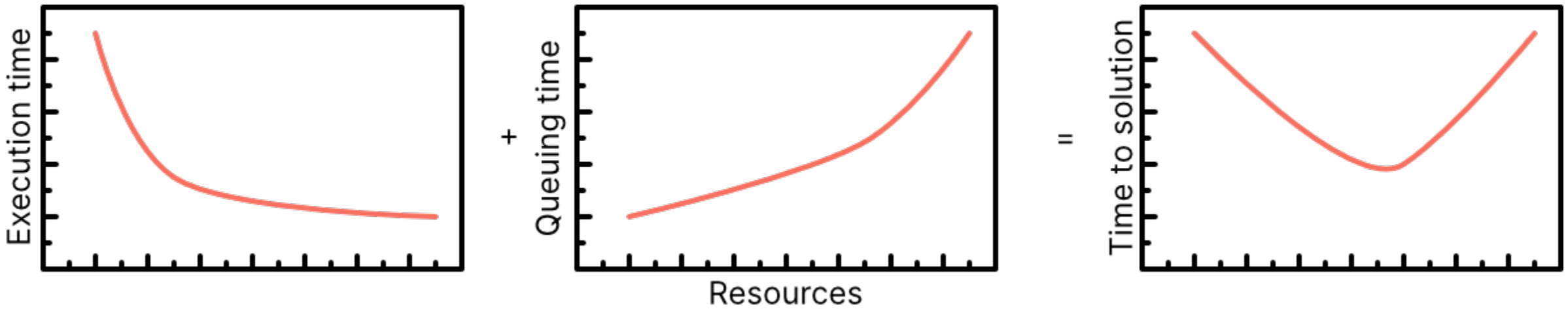
- It's expected that Lucia users have previous experience with HPC system:
  - You know the basics of the Linux command line and Slurm
  - You or someone in your group can install the software you need

- The Cenaero HPC team is very small. The CÉCI team try to help as much as possible in the support of our users but what we have limited access (we have no root access)

BIGGER IS NOT ALWAYS BETTER

Don’t submit big jobs for the sake of it:

- When running large jobs, make sure that you code can run efficiently at a large scale
- Ideally, you should always do some basic scalability study
- Two jobs using 8 nodes at 75% parallel efficiency are better than one job using 16 nodes at 45% efficiency



OPTIMIZE THE RESOURCES USAGE

Most of the nodes in the `batch` and `medium` partitions are in user exclusive mode: when a job start on a node, only a job of the same user is allowed to start on that node

- Ideal job for Lucia should use a multiple of the number of cores on the compute nodes (128 cores)
- If you do capacity computing with job arrays, the size of the array should be able to fill a compute node
- If your job (or job array) is small, use the **`shared`** partition

The GPU nodes can be shared by multiple users. It's recommended to

- Request slice of resources corresponding to what is available per GPU: 8 cores and 60 GB of RAM
- If you need more memory per GPU, use the **`ia`** partition (up to 250GB of RAM per GPU)

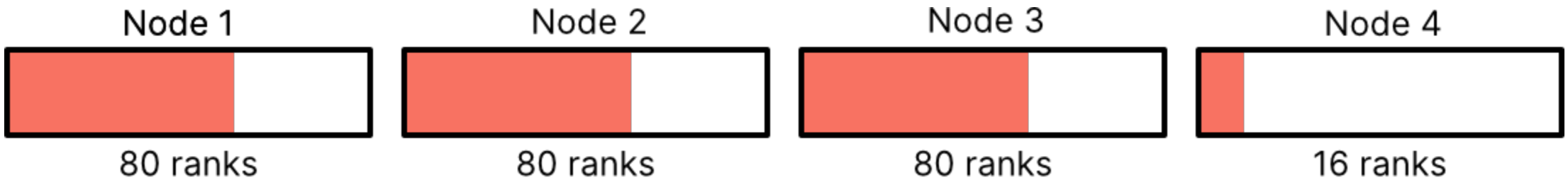
CPU NODES ARE IN USER EXCLUSIVE MODE

- In the `batch` partition, the sweet spot is to request a memory allocation <= 1.8 GB per core
- In the `medium` partition, the sweet spot is between 1.8 and 3.8 GB per core

Real life example extracted from the queue this week-end:

- Job running in the `batch` partition, requesting 256 tasks and 3000MB of memory per task
- 4 nodes (512 cores) allocated to the jobs with 256 cores really used
- Imbalanced allocation (last node almost empty)

Turns out after inspection that the job require 1.7 GB of memory per core: could have run with half the resources!



Recently, Cenaero informed us of their plan to switch to job exclusive mode for the `batch` and `medium` partitions

If you are currently using Lucia for capacity computing (large job array with jobs using an handful of cores), you probably need to change your workflow to adapt to this change.

The CÉCI already have some materials available regarding the use of workflow management tools:

- CÉCI documentation about workflow management
- PRACE workshop on HPC Workflow
- Recordings of the PRACE workshop on HPC Workflow (YouTube)

You should also consider the use of tools like HyperQueue, a scheduler for sub-node tasks

LUMI

THE SUPERCOMPUTER OF THE NORTH

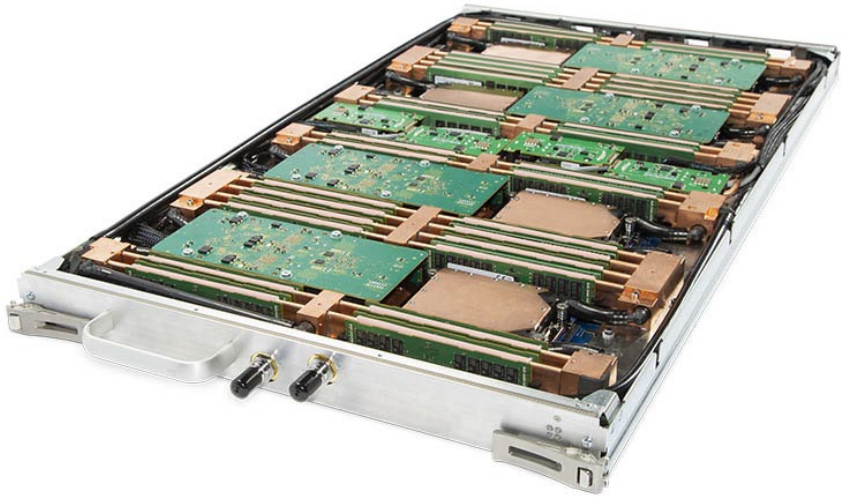
LUMI

- LUMI is currently the fastest supercomputer in Europe (5th in the world in the Nov. 23 TOP500 ranking) with a sustained HPL-linpack performance of 379.70 PFLOPS
- LUMI was acquired by the EuroHPC Joint Undertaking (50%) and a consortium of 11 countries (50%)
- ~200M€ total budget with compute time allocated in proportion of the investment of each stakeholder
- Installed in Kajaani, Finland
- **Belgium** is the second-largest contributor in the consortium with a **7.651% share**

LUMI CPU nodes

2048 nodes (262 144 cores) with

- 2x AMD EPYC 7763 64-core CPUs
- 188 nodes with 256 GB of RAM
- 128 nodes with 512 GB of RAM
- 32 nodes with 1 TB of RAM
- 1× 200Gbps Slingshot 11 interconnect



Slurm partition	Max job size	Max jobs	Max walltime	Job exclusive
small	4 nodes (512 cores)	220 (200 running)	3 days	No
standard	256 nodes (32 768 cores)	60 (50 running)	2 days	Yes

LUMI GPU nodes

2978 nodes with

- 1x AMD EPYC 7A53 64-core CPUs
- 4x AMD Instinct MI250X 128GB GPUs
- 512 GB of RAM
- 4x 200Gbps Slingshot 11 interconnect

Slurm partition	Max job size	Max jobs	Max walltime	Job exclusive
small-g	4 nodes (32 GCDs)	210 (200 running)	3 days	No
Standard-g	512 nodes (4 096 GCDs)	105 (100 running)	2 days	Yes

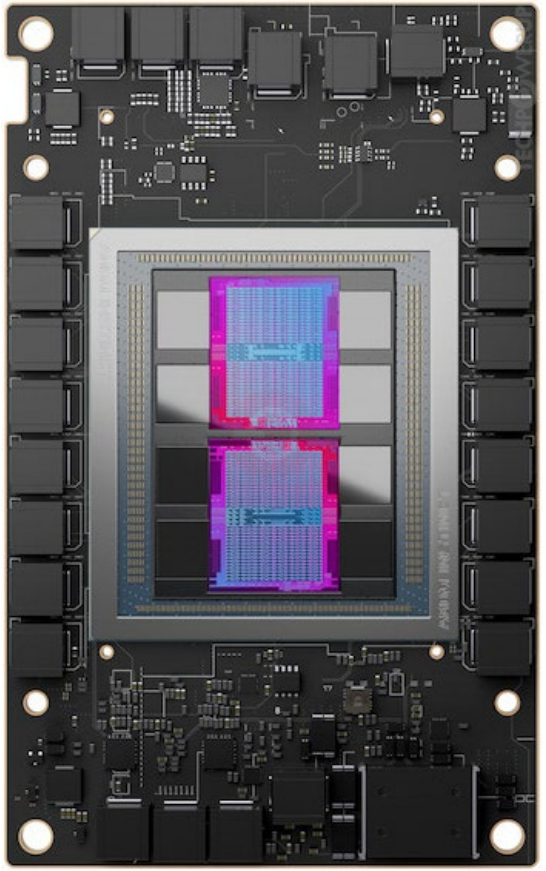
THE AMD MI250X ACCELERATOR

The MI250X accelerator embrace a chiplet design. One board features two Graphics Compute Dies (GCDs) with:

- 110 compute units
- 64 GB of HBM2e memory
- 1.6 TB/s of memory bandwidth

From Slurm point of view, one MI250X board appears as two GPUs

Type of operation	Peak perfomance (TFLOPS)
Single precision Matrix (FP32)	95.7
Double precision Matrix (FP64)	95.7
Single precision (FP32)	47.9
Double precision (FP64)	47.9
Half precision (FP16 and bfloat16)	383



https://docs.lumi-supercomputer.eu/hardware/lumig/

PROGRAMING MODELS

CUDA being a proprietary technology, it's not available on LUMI:

- AMD provides a similar stack which includes HIP, the Heterogeneous-computing Interface for Portability:
  - the kernel language is identical to the CUDA one
  - the runtime functions are different but porting is often limited to a search and replace
  - AMD warps size is 64 threads while NVIDIA is 32
- Most of NVIDIA libraries have an AMD equivalent
- AMD ROCm compiler also supports OpenMP target offloading

C++ portability layers like Kokkos, RAJA and Alpaka have HIP backends

SyCL with AdaptiveCpp or Intel DPC++ compiler (with Codeplay HIP plugin)

SOME ADDITIONAL THINGS TO KNOW

With a larger scale comes a bigger potential for failure:

- More components means more potential for failure
- With larger jobs, more components are involved and the mean time between failure can be lower than what you are used to

About MPI:

- Because LUMI uses an “exotic” interconnect the only really supported MPI implementation is the one provided by the vendor: Cray MPICH
- While all MPI implementations implement the same standard some behaviours may change

The LUMI filesystem performance is very bad if you are using a lot of small files:

- Metadata operation quickly becomes a bottleneck
- Dataset with millions of small files should be converted to a more HPC friendly format (HDF5, SquashFS, FFCV, ...)

LUMI IN THE BELGIAN HPC LANDSCAPE

The CPU partition of LUMI (LUMI-C) is rather small:

- If you consider the Belgian share, it's about half the size of Lucia
- It's even smaller if you consider the Walloon share of LUMI
- Should not be considered as a second Tier-1 machine

However, LUMI-C offer opportunities Lucia cannot provide: with a large number of nodes, you can routinely run jobs requiring tenths of nodes (up to 256 nodes)

The GPU partition (LUMI-G) is where the majority of the compute power of LUMI is:

- The Belgian share represents ~5x the number of GPUs offered on Lucia (10x GCDs)
- It's the largest part of the investment. If your code can use GPU, you should focus on targeting LUMI-G
- Like LUMI-C, you can run larger jobs (up to 512 nodes) that are impossible to run Lucia

GETTING ACCESS

Preparatory (4 month) or development access (1 year):

- 25k GPU-hours
- 500k CPU-core-hours
- Or a proportion of both

Regular access (1 year) with a maximum of

- 500k GPU-hours
- 10M CPU-core-hours
- Or a proportion of both

Application forms and cut-off dates:

- https://www.enccb.be/GettingAccess

Preparatory/development projects

Can be submitted continuously and are reviewed once a month

Next cut-off dates for regular projects

- 3rd of June
- 7th of October

🇪🇺 THE EUROHPC-JU SUPERCOMPUTERS

THE EUROHPC JOINT-UNDERTAKING

The EuroHPC Joint-Undertaking is an organization of the European Union and of the EuroHPC JU participating countries that aims to coordinate the efforts and pool supercomputing resources. It is jointly funded by its members with a budget of around 7 billion euros for the period 2021-2027.

As European researchers, you have access, for free, to EuroHPC supercomputers:

- 5 mid-range: 4-15 PFLOPS
- 3 pre-exascale: 100-400 PFLOPS
- 2 Exascale and 5 mid-range supercomputers planned in the near-future
- 6 Quantum computers planned

Access is granted by submitting a project (**including access to LUMI**):

- https://eurohpc-ju.europa.eu/access-our-supercomputers/eurohpc-access-calls_en

EuroHPC
Joint Undertaking

PRE-EXASCALE SUPERCOMPUTERS

LEONARDO
Italy

246.54 PFLOPS
(#6)

Intel Ice-Lake and Sapphire Rapids
NVIDIA Ampere A100

MARENOSTRUM 5
Spain

178.30 PFLOPS
(#8)

Intel Emerald and Sapphire Rapids
NVIDIA Hopper H100

EuroHPC
Joint Undertaking

MID-RANGE SUPERCOMPUTERS

MELUXINA
Luxembourg

12.89 PFLOPS
(#71)

AMD EPYC and FPGA
NVIDIA Ampere A100

KAROLINA
Czech Republic

9.59 PFLOPS
(#113)

AMD EPYC
NVIDIA Ampere A100

DEUCALION
Portugal

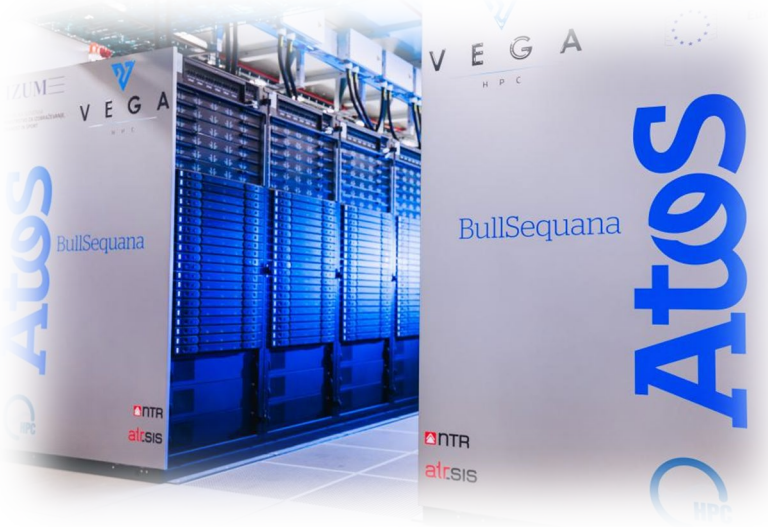
7.22 PFLOPS
(?)

A64FX (ARM) and AMD EPYC
NVIDIA Ampere A100

VEGA
Slovenia

6.92 PFLOPS
(#198)

AMD EPYC
NVIDIA Ampere A100



DISCOVERER
Bulgaria

4.52 PFLOPS
(#166)

AMD EPYC

JUPITER – THE FIRST EUROPEAN EXASCALE SUPERCOMPUTER

JUPITER (Joint Undertaking Pioneer for Innovative and Transformative Exascale Research) will be the first European supercomputer with an HPL performance of 1 ExaFLOPS:

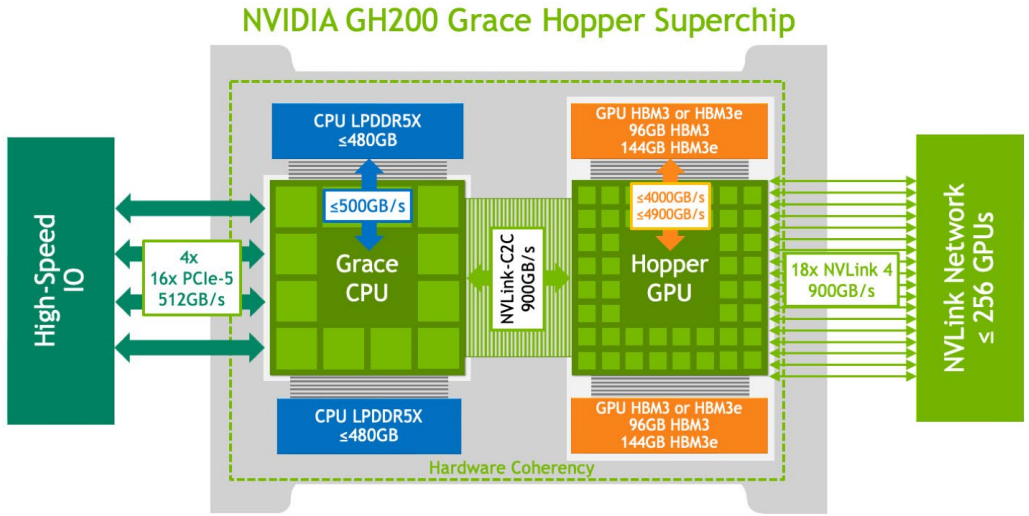
- Installation to start this year at the Julich supercomputing center (Germany)
- ~6000 GPU computes with NVIDIA GH200 superchips
- The CPU partition will feature the first European HPC Processor developed in the framework of the European Processor Initiative (EPI)

A second exascale system is planned (Jule Verne consortium – France and The Netherland)

JUPITER GPU NODES

The JUPITER GPU nodes will feature 4 NVIDIA Grace-Hopper Superchips:

- Hopper GPU with 96 GB of HBM3 memory
- Grace CPU has 72 ARM Neoverse V2 cores and 120 GB of the LPDDR5X memory
- The CPU and GPU are connected via a chip-to-chip NVLink interconnect



The CPUs in a node are interconnected via a dedicated CPU NVLink (cNVLink, 100 GB/s) and the GPUs via NVLink4 (150 GB/s)

Each node has 4x InfiniBand NDR 200 Gbps interconnects

https://www.fz-juelich.de/en/ias/jsc/jupiter/tech

JUPITER CPU NODES

JUPITER will feature a CPU partition with a CPU developed in the framework of the European Processor Initiative (EPI)

- The EPI general goal is to reach European independence in High Performance Computing Processor Technologies and build an Exascale machine based on European processors

JUPITER will have ~1300 CPU nodes with

- 2x SiPearl Rhea-1 CPUs
  - ARM Neoverse V1 cores
  - On package HBM2e memory (like the Intel Sapphire Rapids)
- 512 GB of DDR5 memory

- The next iteration (Rhea-2) will be used for the second European exascale machine

https://www.european-processor-initiative.eu/

TAKE HOME MESSAGES

As a CÉCI user you have access to a large and varied offering of computing resources and have access to an entire ecosystem of European HPC cluster:

- from a few hundred CPU cores to several thousand
- from a few dozen GPUs to thousands of GPUs

The future might be more heterogeneous in terms of hardware that it used to be

- x86 might not be the only CPU architecture with ARM becoming to be more relevant in HPC
- Accelerator hardware will also be more diverse with AMD (and Intel) offering hardware competing with NVIDIA for some use cases