THE NEW CÉCI COMMON STORAGE

THE CÉCI COMMON STORAGE

The goal of the CÉCI Common Storage is to provide a filesystem accessible from all login and compute nodes of all CÉCI clusters

The CÉCI philosophy is to provide complementary HPC clusters:

- With a single account, users can access the cluster that best fits their needs
- Users are encouraged to leverage multiple clusters when different parts of their workflow are better suited to the distinct characteristics of each cluster

The consequences are

- Users might want to have the same configuration/tools on all cluster
- There is a need for an easy way to transfer/share files on all clusters

THE OLD CÉCI COMMON STORAGE

- The current CÉCI shared storage is based on IBM GPFS and consists of two 450 TB main storage systems hosted in Liège and Louvain-la-Neuve.
- Those two storage systems are synchronously mirrored: any file written to one of them is automatically written to the other one.
- Five smaller storage systems serve as buffers/caches (move data where needed when needed).
- Those caches are located on each site and are tightly connected to the cluster compute nodes.
- Everything is interconnected through a dedicated 10 Gbps network.

REPLACEMENT OF THE CÉCI COMMON STORAGE

Current storage is beginning to show signs of old age:

- Close to 10 years old (2016)
- Disk failures are more frequent: we are cannibalizing unused hardware to keep it alive
- RAID controller battery failed recently which can cause a problem in case of power loss
- We can no longer guarantee the integrity of the data stored in the old common storage. If you have data you still need stored in the **trsf** partition **migrate your data asap!**

The replacement started in 2021 by answering a FNRS “Grands Équipements” call

- 500 k€ granted to the project
- Reviewers were unhappy that half of the budget was dedicated to the GPFS license: choice of an alternative solution based on CEPH (free and open source)
- Hardware acquisition started in 2023, installation and tests in 2024-2025

THE NEW COMMON STORAGE

Hardware-wise, this new storage is composed of:

4x storage nodes in each datacenter (CÉCI universities)

- 2x AMD EPYC 9124 16-core CPU
- 12x 16 GB of DDR5-4800 memory (192 GB)
- 24x 7 TiB NVMe SSDs (168 TiB)

Raw storage capacity: 3.3 PiB
Net storage capacity: 1.1 PiB

- Dedicated 10 Gbps CÉCI “ring” network (to be updated to 2x 10 Gbps).
- Uses **Ceph**, a free and open-source software-defined storage platform.
- Data is replicated (**replica 3**) to provide high availability and resilience.

THE CÉCI COMMON STORAGE

The common storage is split into two distinct spaces:

`/CECI/home` designed to work the same way as the local home filesystems you have access to on the clusters to store your own software, configuration files, and small (input) data files or (output) result files.

- Every user has a “CÉCI Home”, accessible via the environment variable `$CECIHOME`
- 100 GB and 100.000 files

`/CECI/proj` dedicated to projects, i.e., groups of researchers working towards a common goal and sharing files.

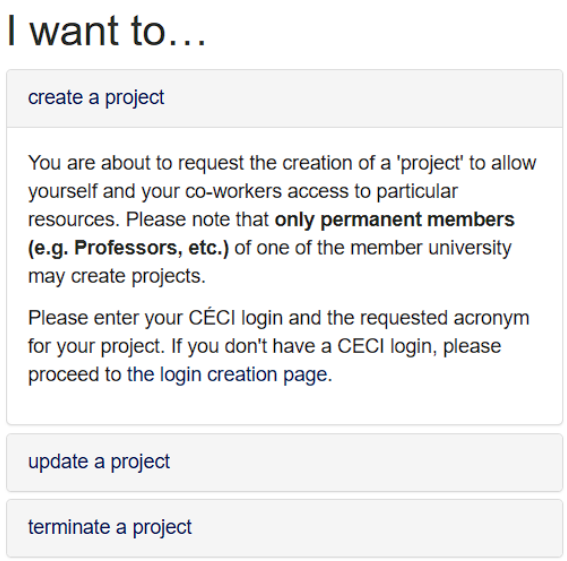
- Common Storage projects are created upon request of a **Principal Investigator**
- The process is similar to one used to request Tier-1 projects

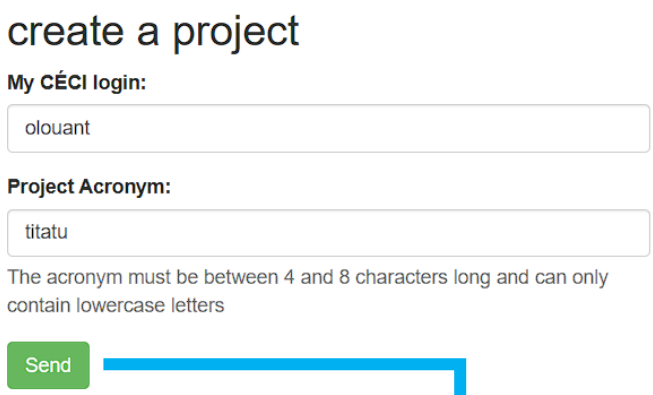
The new storage is not asynchronous like the old one: files and their content are immediately visible on all clusters

HOW TO GET A PROJECT

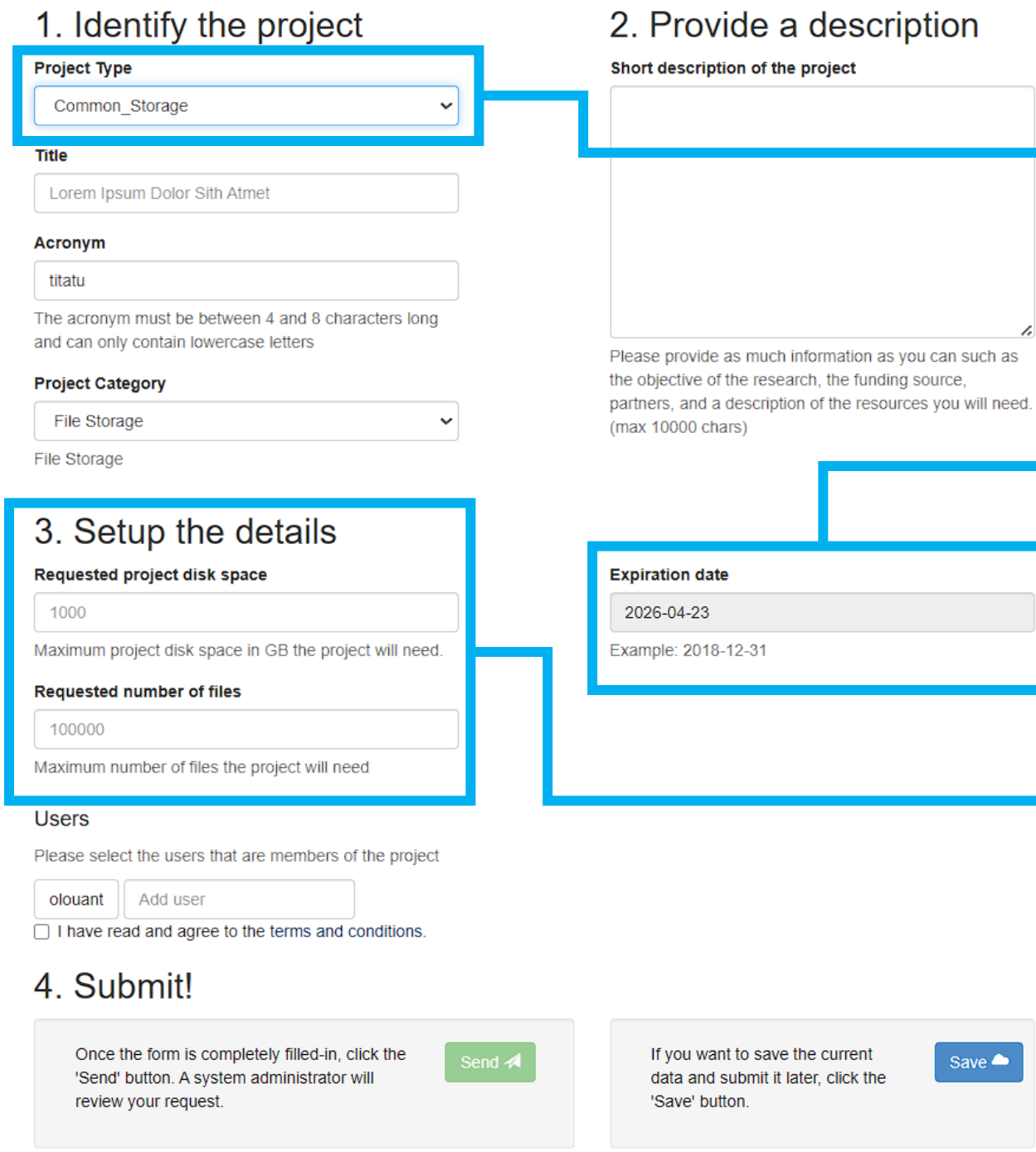
Common Storage projects are created upon request of by **Principal Investigator** holding a **permanent position** within one of the five member universities:

https://login.ceci-hpc.be/init-project/





You will receive an email with a link to create your project



Select **Common_Storage**, otherwise you will create a Tier-1 project!

- Cannot be modified
- A project lifetime is one year
- A project need to be renewed yearly

- Select the amount of storage space (in GB) and number of files.

- The number of files should be between 100.000 and 200.000 files/TB. If you asks for too many files, your request might be rejected

THE CÉCI COMMON STORAGE IS NOT

- **A scratch filesystem nor a local home**
  - Due to the distributed nature of the Ceph Common Storage the performance will be lower than the performance of the clusters `$LOCALSCRATCH` and `$GLOBALSCRATCH` filesystems
  - Avoid writing/reading a lot of small files as you will not get the same IOPS performance as the local cluster `$HOME`.
- **A long-term archival solution**
  - Even if the data is replicated which protect against disk failure, like every CÉCI cluster filesystems there is **no backup**
  - Only data that are still in use should be stored on the CÉCI infrastructure. Long-term archival should be done in your university mass storage

USAGE OF THE CÉCI COMMON STORAGE

- **Use large files or, if not possible, use archives**

  The local file systems will always be faster than the Common Storage: consider staging your data in and out in your jobs:

  - **Stage in:** at the beginning of your job decompress an archive from the Common Storage to the local filesystem
  - **Stage out:** at the end of your job compress the results and store the archive in the Common Storage
  - **Example with a 6.3 GB archive of 50.000 files:**
    - Decompression from Common Storage to NIC5 `$LOCALSCRATCH`: 47s vs 14m26s if copy without archiving
    - Compression from NIC5 `$LOCALSCRATCH` to the Common Storage: 31s vs. 1m37s if copy without archiving

- **In some cases, you might need to request metadata server pinning**

  By default, metadata is pinned to the metadata server of your home institution. If you regularly work on a cluster that is not hosted by your home institution, you can request that we pin a particular directory to the closest metadata server

- **If used properly, the Common Storage is the fastest way to transfer data between clusters**

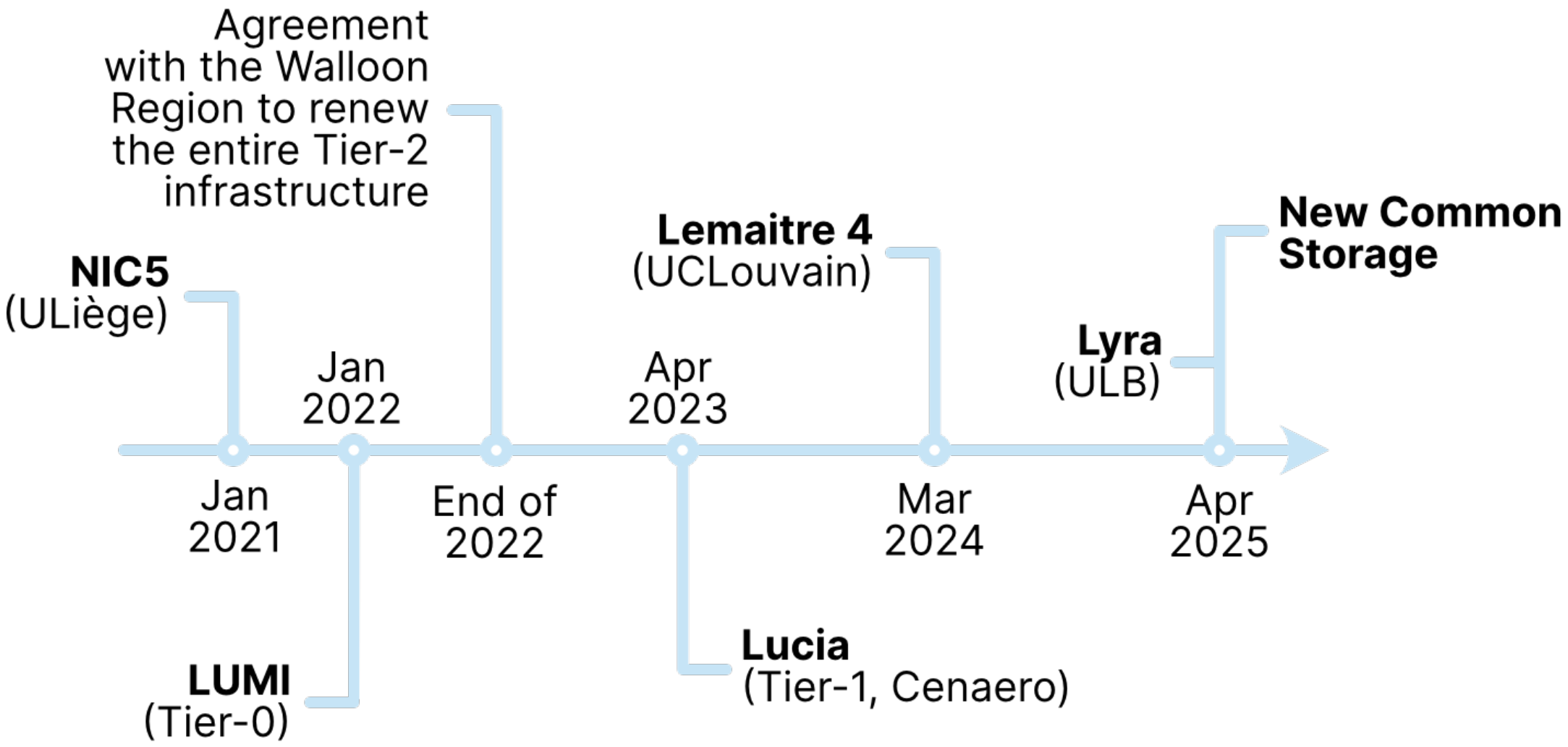
NEXT STEPS

- **The backup of your old `$CECIHOME` (`$CECIHOMEOLD`) will be deleted**
- **Decommission of the old Common Storage (June 2025)**
  - No backup of the old storage when decommissioned
  - Copy the data you need from `/CECIOLD/trfs` as soon as possible!
- **Increase available bandwidth of the Common Storage to 2x 10 Gbps**
  - Once the old storage is retired, we can reuse the 10 Gbps line for the new one
  - Should improve performance
- **The Common Storage will be available on Lucia too!**
  - Thanks to the fact that the new ULB datacenter is a few meters away from Lucia, the new Common Storage can be made available
  - No exact date yet

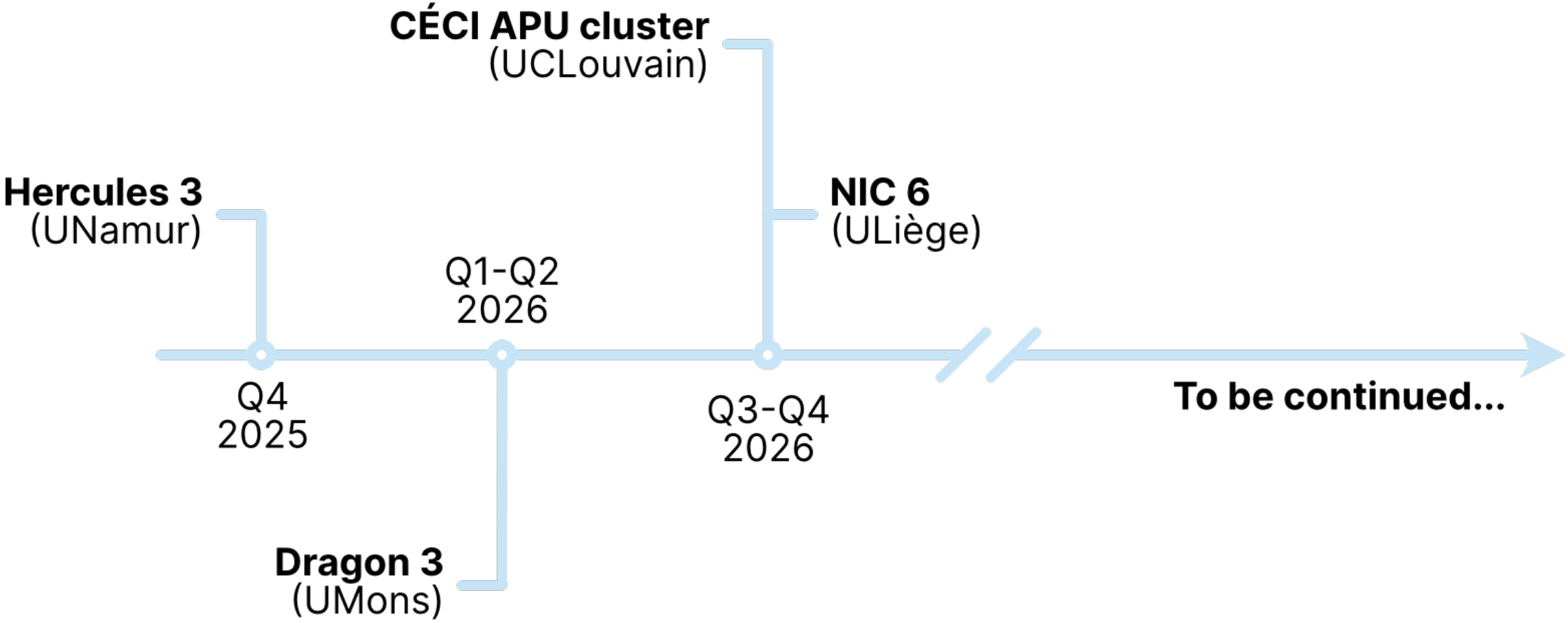
THE CÉCI INFRASTRUCTURE

THE PAST, THE PRESENT AND THE FUTURE

THE LAST 5 YEARS



THE NEXT 2 YEARS



THE CÉCI APU CLUSTER: MOTIVATIONS

State of the CÉCI GPU compute:

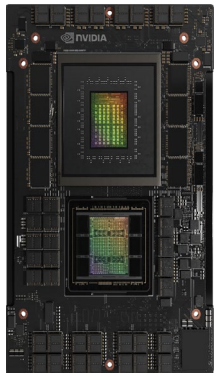
- We already have Lyra for FP32 GPU workload.
- We have FP64 GPUs on Lucia but it's not managed by the CÉCI.
- The budget for Dragon 3 doesn't allow us to replace the GPU nodes of Dragon 2: two nodes with 2x NVIDIA H100 GPUs ~ 180 k€, 1/2 the budget!

Instead of going for the classic CPU + Discrete GPU compute nodes, go for a fusion of the two:

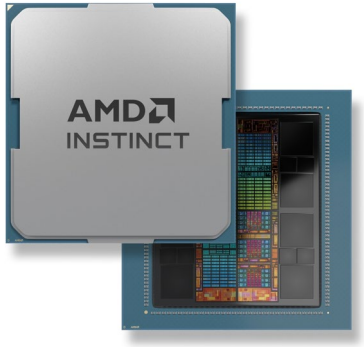
Accelerated Processing Units (APUs)

CHOOSE YOUR COLOR

TEAM GREEN
NVIDIA GRACE-HOPPER

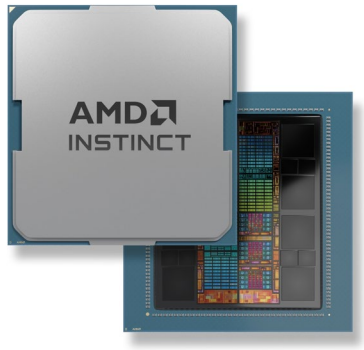


TEAM RED
AMD MI300A



SOFTWARE STACK	Very mature	Quicky improving
CPU ARCHITECTURE	ARM Neoverse V2	AMD x86 Zen4
FP64 PERFORMANCE	34 TFLOPS FP64	61 TFLOPS FP64
MEMORY	96 GB (4.0 TB/s)	128 GB (5.3 TB/s)
PRICE	High	Medium
POWER	500-600 W?	550 W (air cooled)
REALLY AN APU?	Not really	Yes

PLANNED SPECIFICATIONS



Based on a preliminary offers, the APU cluster would be:

4 to 6 APU compute nodes (16 to 24 APUs) with:

- 4x AMD MI300A 128 GB APUs
- 2x 200 Gbps Infiniband interconnect
- Large `$LOCALSCRATCH` (>30 TB)

To maximize the number of APUs, there is a possibility these APU nodes will be part of Lemaitre 4