CÉCI Technical Talks

The CÉCI sysadmins

13th CÉCI Users Meeting

C.E.C.I

ULB | UCLouvain Université | UMONS | Université de Namur | Liège université

Vega
2112 cores

256 GB RAM

70 TiB storage

Decommissionned

Lemaitre 3
1984 cores

95 GB RAM

144 TB storage

Q2 2018

Dragon 2
592 cores

384 GB RAM

4 GPU

100 TB storage

Q1 2019

Hercules 2
1536 cores

2 TB RAM

16 GPU

60 TB storage

Q3 2019

NIC5
4672 cores

1 TB RAM

250 TB storage

Q4 2020

8656 cores total – 20 GPU

LIÈGE université | UCLouvain Université | ULB | UMONS | UNIVERSITÉ DE NAMUR

NIC5
4672 cores

1 TB RAM

250 TB storage

Q4 2020

Lemaitre 4
5120 cores

766 GB RAM

144 TB storage

Q1 2024

Lyra
1472 cores

384 GB RAM

46 GPU

3 PB storage

Q2 2024

Dragon 3
~1200 cores

512 GB RAM

120 TB storage

Q1 2025

Hercules 3
~2300 cores

3 TB RAM

16 GPU

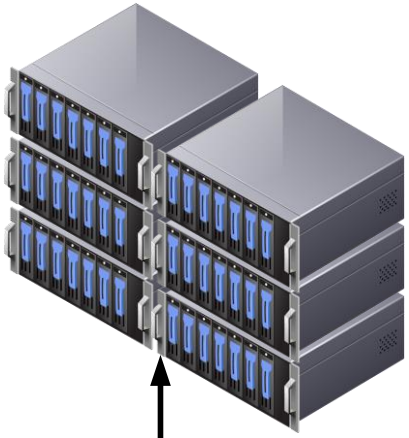
300 TB storage

Q1 2025

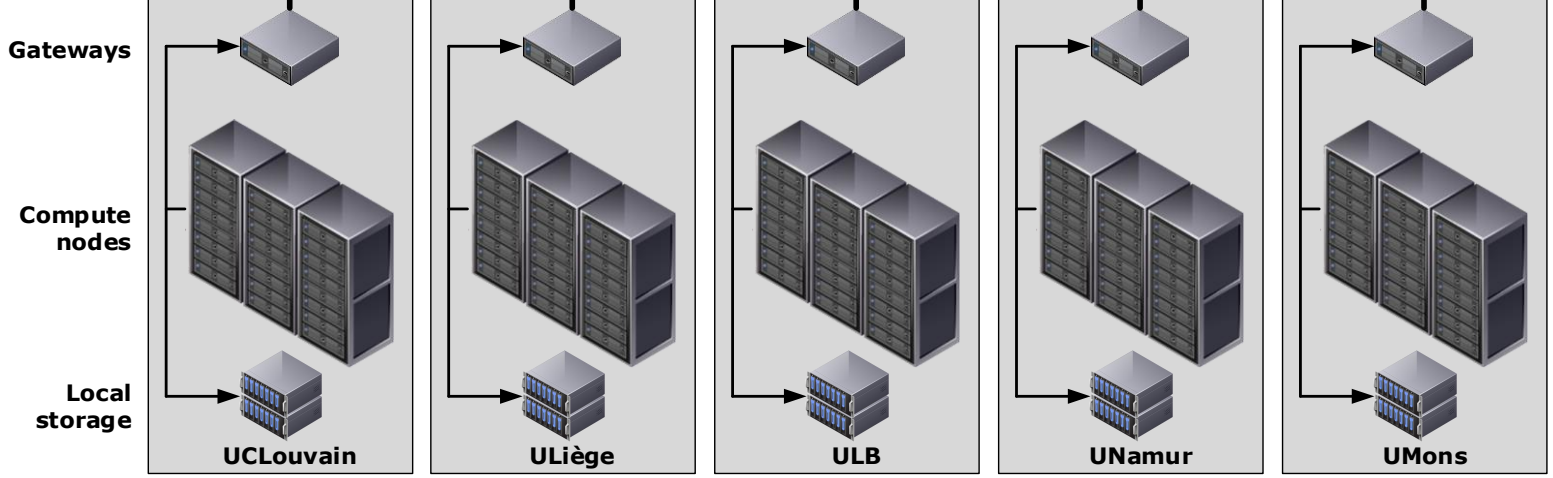
14,000+ cores total – 70+ GPU

CÉCI Common Storage

- Distributed storage solution
- Visible from all the frontends and compute nodes of all CÉCI clusters
- 230 TB net



IBM Spectrum Storage



New CÉCI common storage

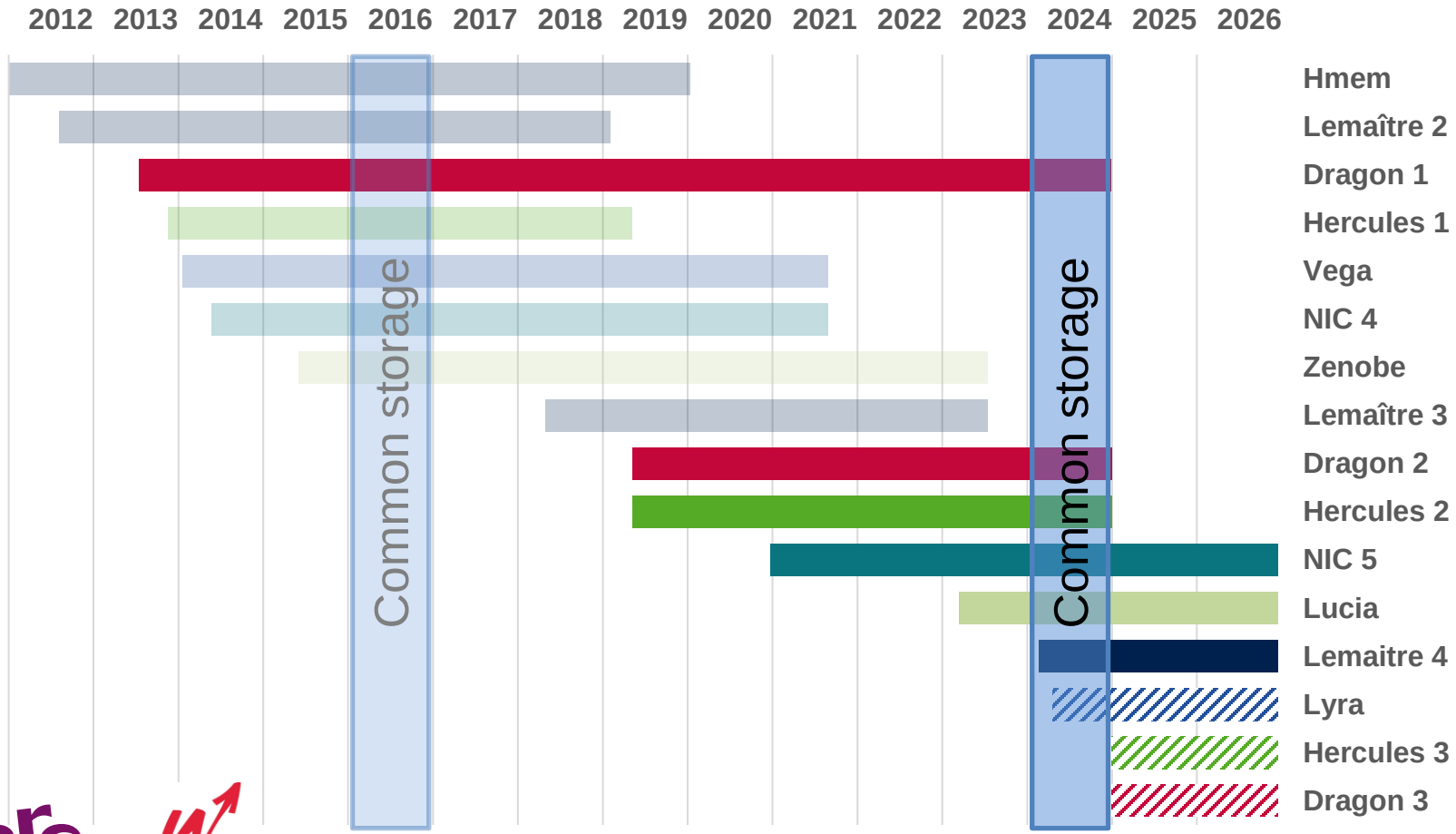
Current

- Installed in 2016
-  IBM Spectrum Scale
- 230 TB net storage on spinning rust
- Installed at UCLouvain and ULiège

New

- Expected Q3 2024
-  Red Hat ceph
- **1 PB** net storage on NVMe SSD
- Fully distributed storage

CÉCI 2024 roadmap



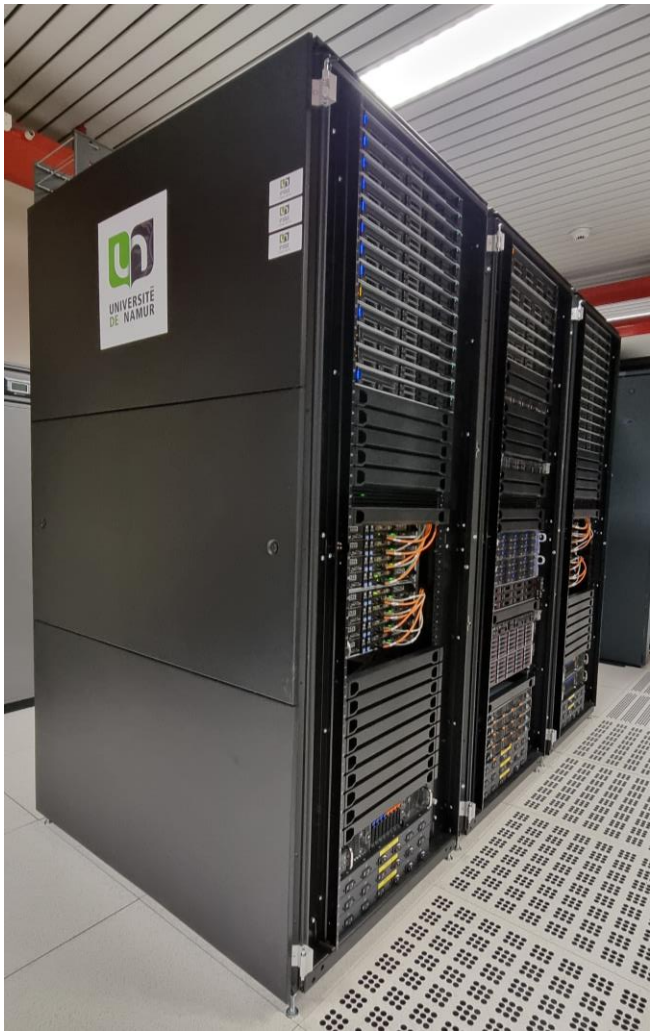
Hercules 3

The CÉCI sysadmins

CÉCI day 2024

Hercules 2

- 1536 cores
- up to 2 TiB memory
- 16 GPU
- 60 TB storage
- 10 GbE network

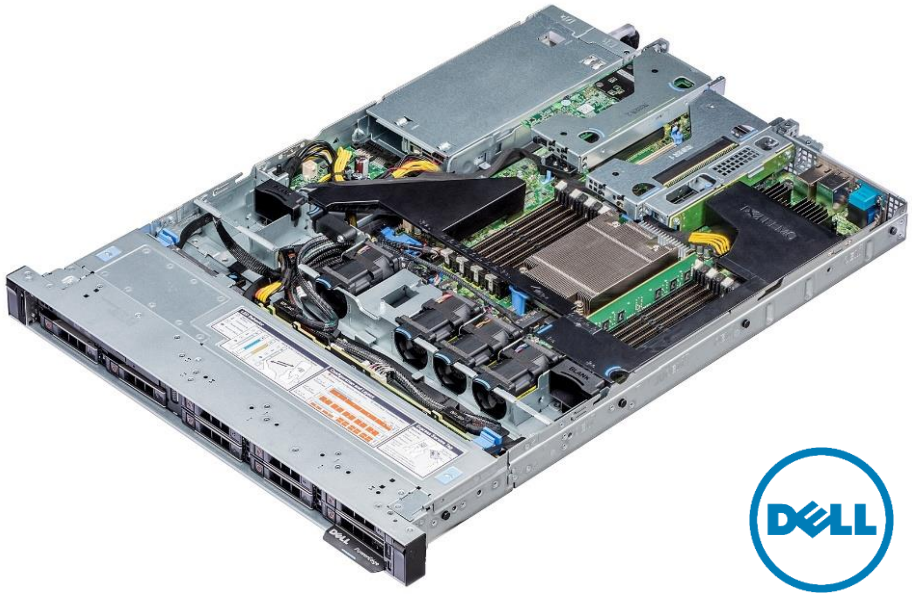


Hercules – CPU nodes

Hercules 2

#	Processor	Memory	Storage
30	2x 8-core Intel “Sandybridge”	64 – 128 GB	1 TB HDD
24	1x 32-core AMD EPYC Gen1 “Naples”	256 GB	5 TB HDD
4	1x 32-core AMD EPYC Gen1 “Naples”	512 GB	5 TB HDD
2	2x 32-core AMD EPYC Gen1 “Naples”	2 TB	8 TB HDD

In production since 2012



Hercules – CPU nodes

Hercules 2

#	Processor	Memory	Storage
30	2x 8-core Intel “Sandybridge”	64 – 128 GB	1 TB HDD
26	1x 32-core AMD EPYC Gen1 “Naples”	256 GB	5 TB HDD
4	1x 32-core AMD EPYC Gen1 “Naples”	512 GB	5 TB HDD
2	2x 32-core AMD EPYC Gen1 “Naples”	2 TB	8 TB HDD

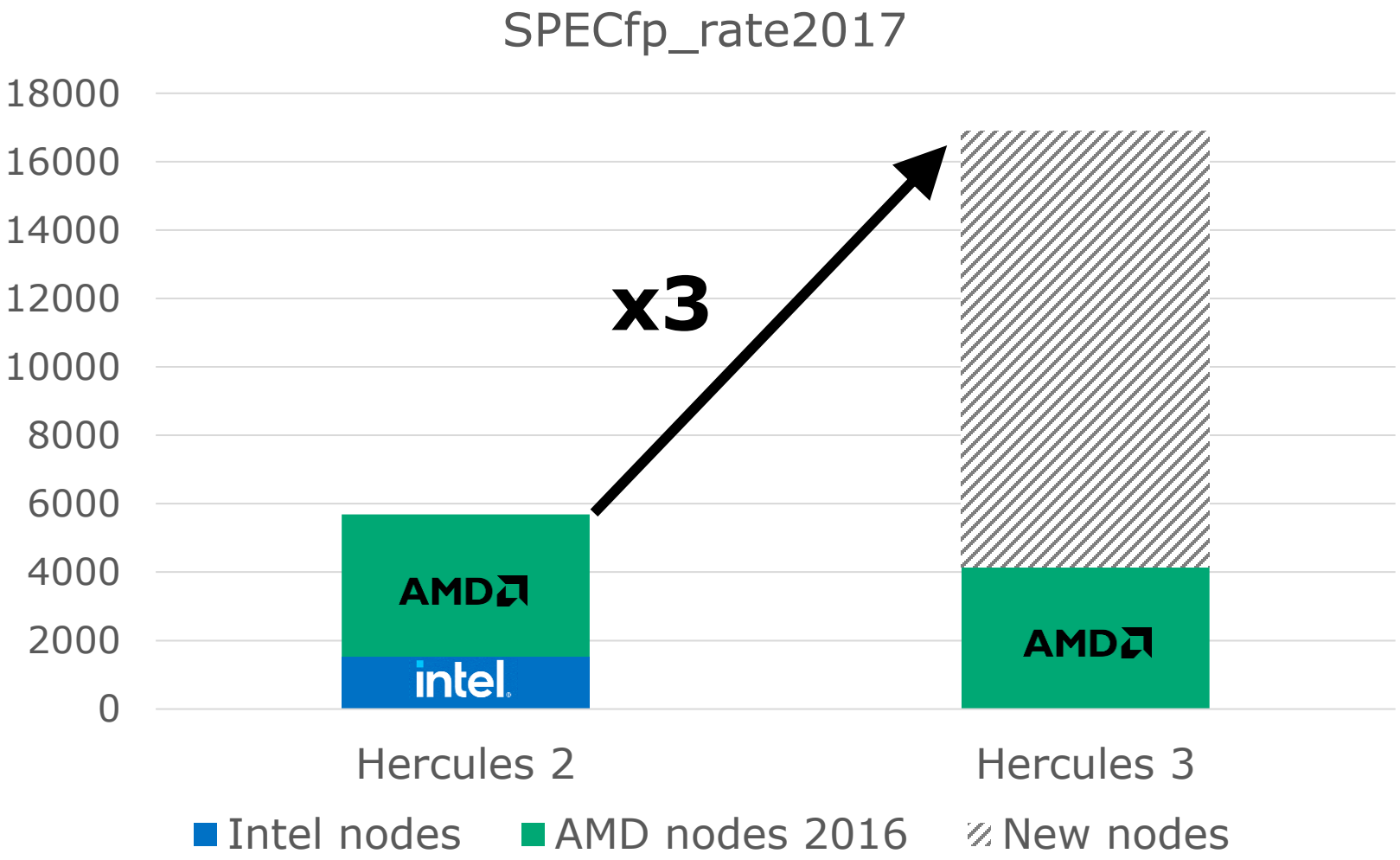
32 nodes from Hercules 2

Hercules 3

#	Processor	Memory	Storage
26	1x 32-core AMD EPYC Gen1 “Naples”	256 GB	5 TB HDD
4	1x 32-core AMD EPYC Gen1 “Naples”	512 GB	5 TB HDD
2	2x 32-core AMD EPYC Gen1 “Naples”	2 TB	8 TB HDD
18	**1x 48-core AMD EPYC Gen4 “Genoa”**	**384 GB**	**6 TB NVMe**
4	**1x 48-core AMD EPYC Gen4 “Genoa”**	**768 GB**	**6 TB NVMe**
1	**2x 48-core AMD EPYC Gen4 “Genoa”**	**3 TB**	**12 TB NVMe**

23 new nodes

Total CPU performance



Hercules – GPU nodes

Hercules 2

#	Processor	Memory	Storage	#GPU	VRAM
1	1x 32-core AMD EPYC “Rome”	256 GB	7 TB HDD	4	11 – 24 GB
1	1x 32-core AMD EPYC “Rome”	256 GB	7 TB NVMe	4	48 GB
1	2x 32-core AMD EPYC “Milan”	512 GB	200 GB SSD	8	48 GB

Hercules – GPU nodes

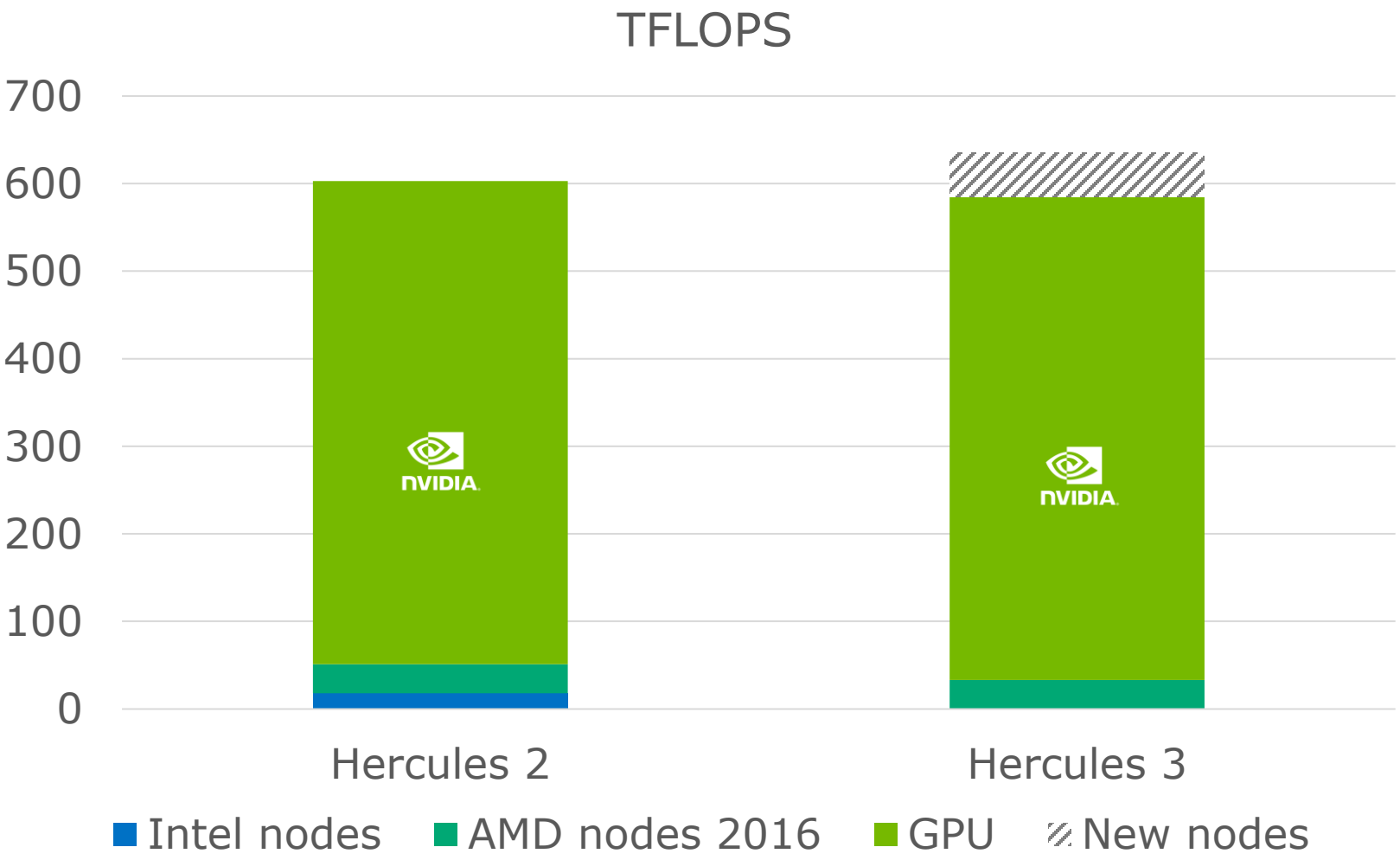
Hercules 2

#	Processor	Memory	Storage	#GPU	VRAM
1	1x 32-core AMD EPYC “Rome”	256 GB	7 TB HDD	4	11 – 24 GB
1	1x 32-core AMD EPYC “Rome”	256 GB	7 TB NVMe	4	48 GB
1	2x 32-core AMD EPYC “Milan”	512 GB	200 GB SSD	8	48 GB

Hercules 3

#	Processor	Memory	Storage	#GPU	VRAM
1	1x 32-core AMD EPYC “Rome”	256 GB	**14 TB HDD**	4	11 – 24 GB
1	1x 32-core AMD EPYC “Rome”	256 GB	**12 TB NVMe**	4	48 GB
1	2x 32-core AMD EPYC “Milan”	512 GB	**16 TB NVMe**	8	48 GB

CPU+GPU performance



Homedir storage

Hercules 2

HPE ProLiant XL420 Gen10

40 TB HDD
+3 TB SSD cache

Home directories

Quota: 1 TB/user

Hercules 3

ceph

5-nodes Ceph cluster

256 TB HDD
+64 TB NVMe cache

Home directories
Virtual Machines

Quota: **4 TB/user**

Globalscratch storage

Hercules 2

HPE StorageWorks D2700

20 TB HDD

Global scratch

Quota: 4 TB/user

Hercules 3

HPE ProLiant XL420 Gen10

60 TB HDD
+6 TB SSD cache

Global scratch
Databases

Quota: **8 TB/user**

From Hercules 2 to Hercules 3

Hercules 2	Hercules 3
1536 cores	**2368** cores
up to 2 TB memory	up to **3 TB** memory
16 GPU	16 GPU
60 TB storage	**300 TB** storage
10 GbE network	25 GbE Network

Expected in production Q1 2025

Dragon 3

UMONS

The CÉCI sysadmins

13th CÉCI Users Meeting

Dragon - Compute nodes

CPU nodes

#	Processor	Memory	Storage
15	2x 16-core Intel Xeon Gold 6142 “Skylake”	192 GB	3 TB HDD
2	2x 16-core Intel Xeon Gold 6142 “Skylake”	384 GB	3 TB HDD

GPU nodes

#	Processor	Memory	Storage	#GPU	VRAM
2	2x 12-core Intel Xeon Gold 6126 “Skylake”	192 GB	3 TB	2	16 GB

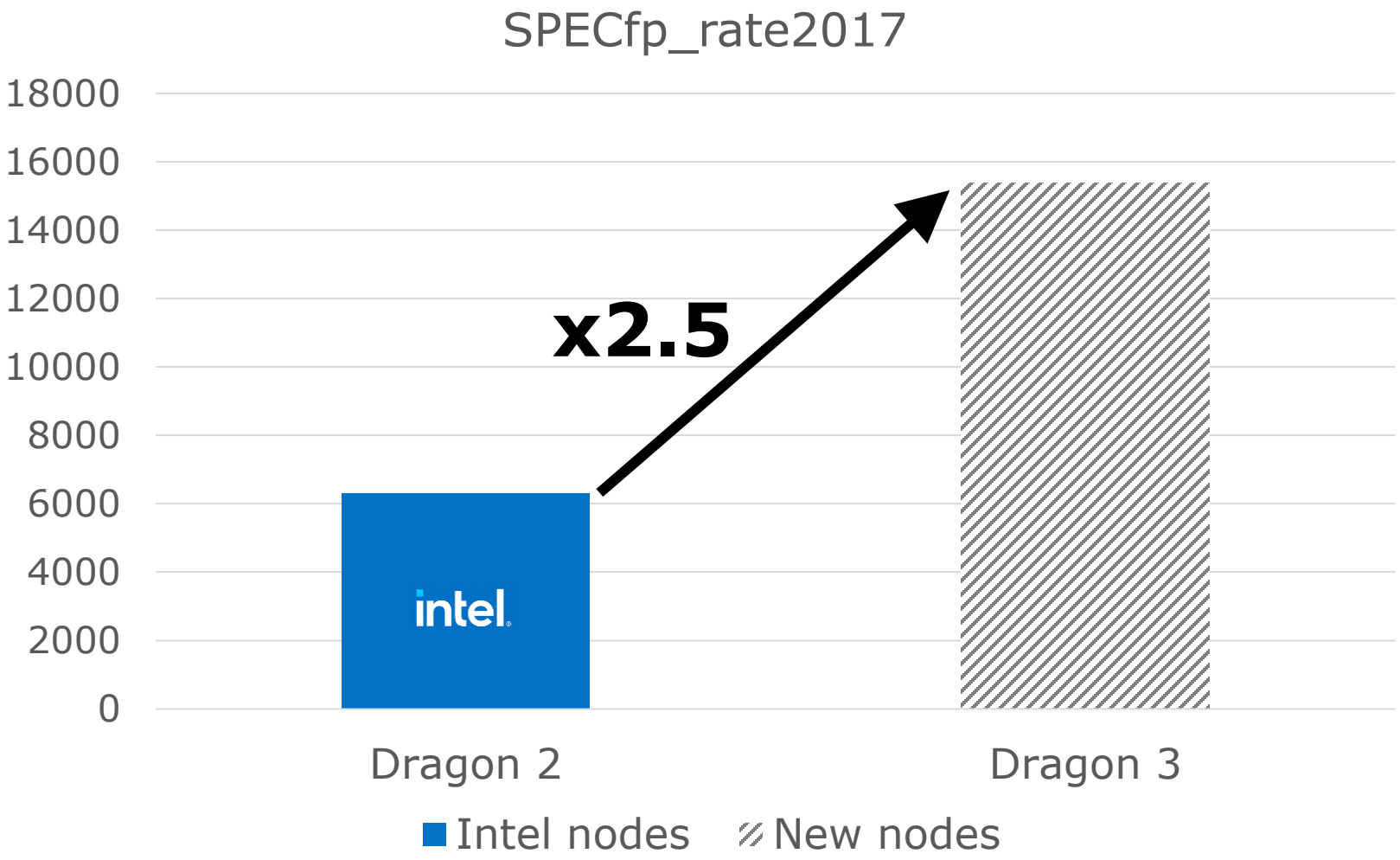
Dragon 2: 592 cores – 4 GPU

CPU nodes

#	Processor	Memory	Storage
20	2x 32-core Intel Xeon Gold 6548N “Emerald Rapids”	512 GB	6 TB NVMe

Dragon 3: 1280 cores – no GPU

Total CPU performance



From Dragon 2 to Dragon 3

Dragon 2	Dragon 3
592 cores	**1280** cores
up to 384 GB memory	up to **512 GB** memory
4 GPU	no GPU
40 TB storage	**120** TB storage
10 GbE Network	25 GbE Network

Expected in production Q1 2025

Lemaître 4

The CÉCI sysadmins

13th CÉCI Users Meeting

Georges Lemaître

Father of the Big Bang Theory...

Burroughs E101, May 1959, Digital Laboratory of UCL (credit archives.uclouvain.be)

Previous Lemaître Clusters

2005-2012 **Lemaitre 1** 256 Core, 1.3 TFlops, 120W/c

128 dual AMD Opteron 252/2.6Ghz 4GB RAM
10Gbps Infiniband
6TB /home, 1.2TB /scratch

Previous Lemaître Clusters

2005-2012 Lemaitre 1 256 Core, 1.3 TFlops, 120W/c
2012-2018 **Lemaitre 2** 1334 Core, 13.6 TFlops, 22W/c

112 dual Intel 6 core
Xeon/2.53Ghz 48GB RAM
40Gbps QDR Infiniband
40TB /home, 120TB /scratch

Previous Lemaître Clusters

2005-2012 Lemaitre 1 256 Core, 1.3 TFlops,120W/c
2012-2018 Lemaitre 2 1334 Core, 13.6 TFlops, 22W/c
2018-2024 **Lemaitre 3** 2032 Core, 75.4 Tflops, 13W/c

80 Intel
2x12 core Skylake Gold 2.3Ghz
96GB RAM
56Gbps Omni-path
35TB /home, 583TB /scratch

Current Lemaître Cluster

2005-2012	Lemaitre 1	256 Core, 1.3 TFlops, 120W/c
2012-2018	Lemaitre 2	1334 Core, 13.6 TFlops, 22W/c
2018-2024	Lemaitre 3	2032 Core, 75.4 TFlops, 13W/c
2024-	**Lemaitre 4**	5120 Core, 303 Tflops, 8W/c

40 Dual 64 Core AMD Epyc Genoa 3.7GHz, 768GB RAM

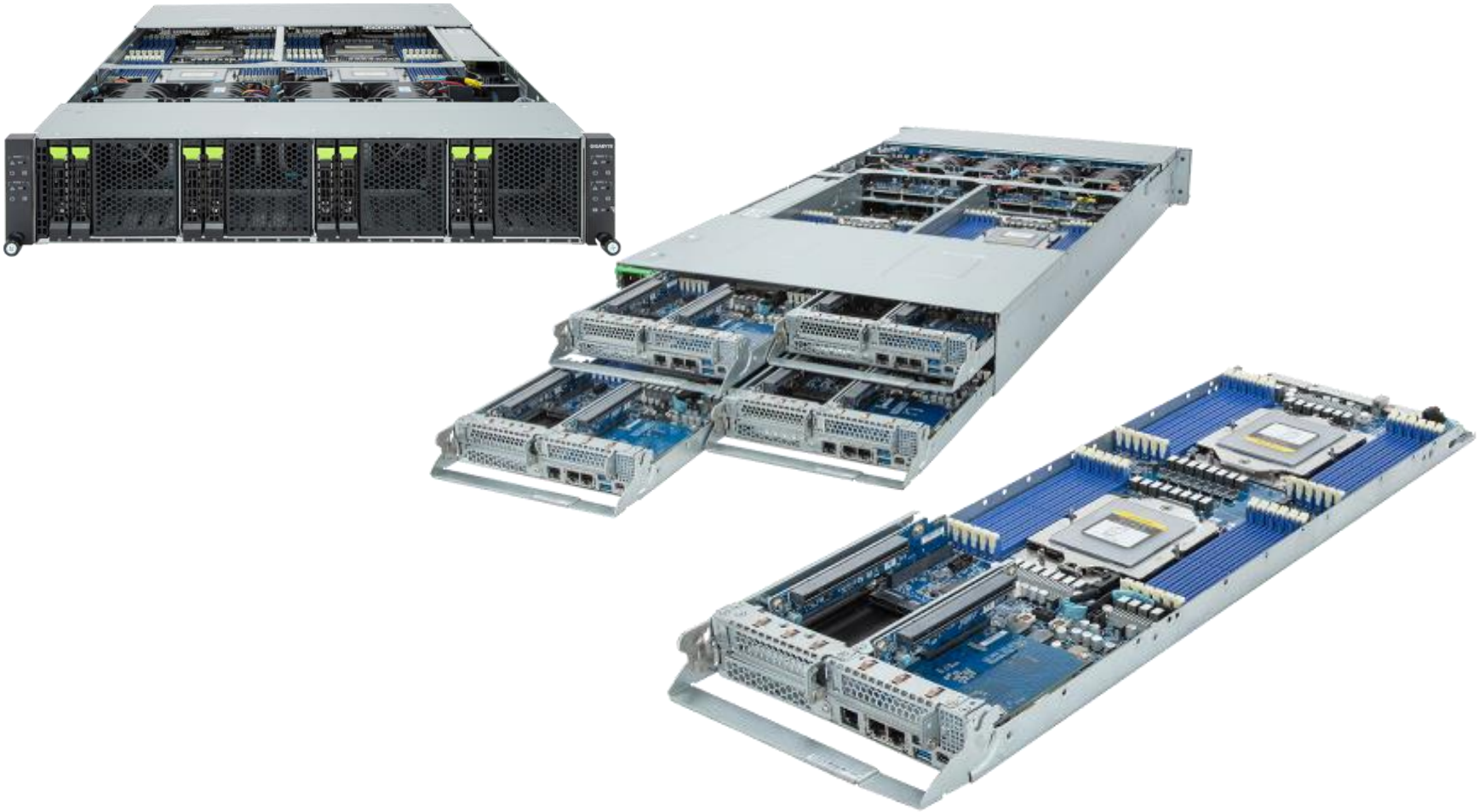
100 Gbps Infiniband HDR
40TB /home, 318TB /scratch

Inside Lemaitre 4

The front-end, the working nodes, the file servers...

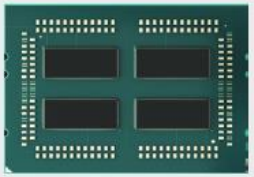
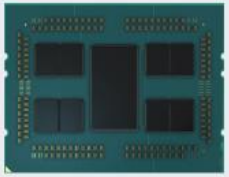
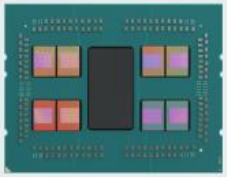
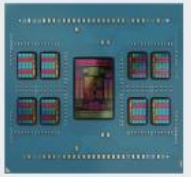
Inside Lemaitre 4

The working nodes



Inside Lemaitre 4

The AMD EPYC 9004 GENOA

	AMD EPYC 7001 'NAPLES'	AMD EPYC 7002 'ROME'	AMD EPYC 7003 'MILAN'	AMD EPYC 9004, 8004 'GENOA', 'SIENA'
				
Core Architecture	'Zen'	'Zen 2'	'Zen 3'	'Zen 4' and 'Zen 4c'
Cores	8 to 32	8 to 64	8 to 64	8 to 128
IPC Improvement Over Prior Generation	N/A	~24%[ROM-236]	~19% [MLN-003]	~14%[EPYC-038]
Max L3 Cache	Up to 64 MB	Up to 256 MB	Up to 256 MB	Up to 384 MB (EPYC 9004) Up to 128 MB (EPYC 8004)
Max L3 Cache with 3D V-Cache™ technology			768 MB	Up to 1152 MB
PCIe® Lanes	Up to 128 Gen 3	Up to 128 Gen 3	Up to 128 Gen 4	Up to 128 Gen 5 8 bonus lanes Gen 3
CPU Process Technology	14nm	7nm	7nm	5nm
I/O Die Process Technology	N/A	14nm	14nm	6nm
Power (Configurable TDP [cTDP])	120-200W	120-280W	155-280W	70-400W
Max Memory Capacity	2 TB DDR3-2400/2666	4 TB DDR4-3200	4 TB DDR4-3200	6 TB DDR5-4800

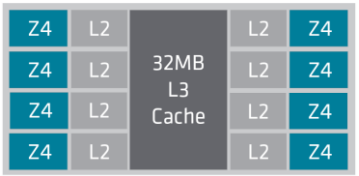
Inside Lemaitre 4

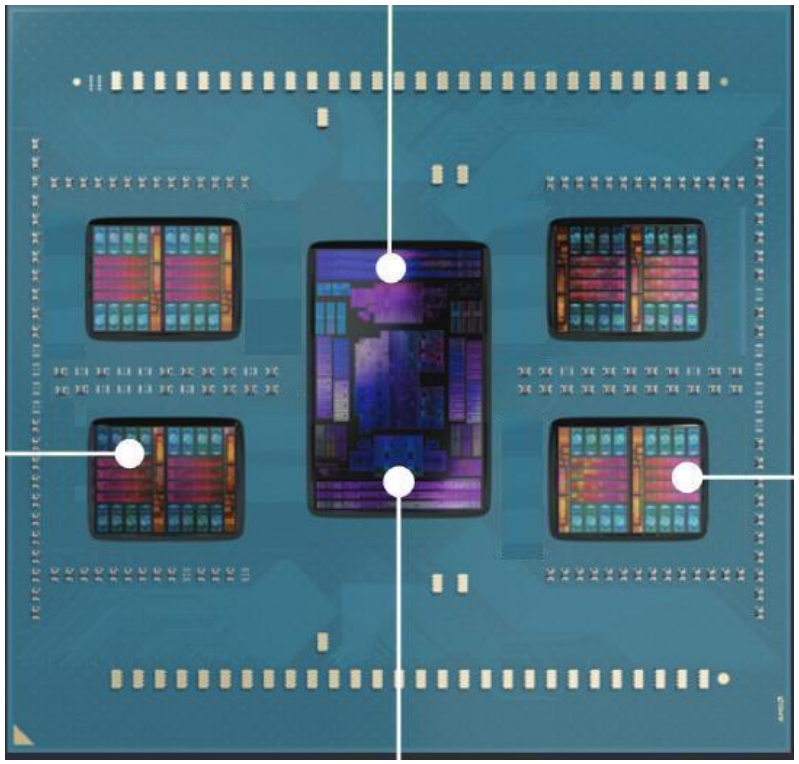
The AMD EPYC 9534 GENOA

8 zen4 cores
with 1MB L2 cache

=

a Core Complex CCX,
with 32MB L3 cache





Two CCX

=

A CPU Die CCD

4 CCD + 1 I/O Die

=

EPYC 9534

I/O die : 12 Mem DDR5, PCIe, Infinity Fabric 36Gb/s, Sata Controllers,...

Inside Lemaitre 4

The AMD EPYC 9534 GENOA

- 12 Channel DDR5 Mem (2x 12x 32GB = 768 GB per node)
- AVX512 support, on a 256-bit data path

  (Since 2017 on Intel Skylake SP)
- +123% SPECrate 2017_fp_base on previous AMD Zen3 Milan
- 2.45GHz normal frequency, 3.7GHz in Turbo mode
- NUMA architecture !

Lemaitre 4 I/O considerations

Cluster	FS	IOzone MB/s	Mdtest
Lemaitre3	/home	811	28377
Lemaitre4	/home	1718 (x2)	123659 (x4)
Lemaitre3	/scratch	3466	16937
Lemaitre4	/scratch	2224 (-64%)	158745 (x9)

- **Iozone** (write thoughput, 256KB block on one node)
- **Mdtest** IOR (iops, write/s of 100 process on 10 nodes)
- /home Lemaitre3 : 23 SAS 2TB in ZFS Raid6
- /home Lemaitre4 : 9 SATA 7.68TB **SSD** in RAID6
- /scratch Lemaitre3 Beegfs with 4 file servers
- /scratch Lemaitre4 Beegfs with 1 file server

Lemaitre 3 decommissioning

PRELEMINARY PROPOSAL, feedback welcome by email...

- **May 1st** : no new user account
- some working nodes are removed

- **July 1st** : no more running jobs
- cleaning of the /scratch
- /home in readonly

- **September 1st** : stop with a R/O /home copy on lemaitre4

Lemaitre 4 Welcome!

```
Welcome to

                        /  _ _  _ '_/_ _ (_/
                       (__(-//)(// // (-  /


              Massively parallel CISM-CECI cluster

  40 nodes: 2 x AMD EPYC 9534 64-Core processors (128 HT) 768GB RAM
            non-blocking 100Gbps Infiniband network
                  318T BeeGFS global scratch

           contact, support: egs-cism@listes.uclouvain.be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
4422/10240 CPUs available (load 56%) - 77 jobs running, 0 pending.

* Job info for user bvr: 0 job running, 0 pending.
* Diskquotas for user bvr
Filesystem       used      limit      files      limit
$HOME              53K       100G         10
$GLOBALSCRATCH   0.0kB  unlimited          0  unlimited
$CECIHOME      365.4MB    100.0GB        690     100000
$CECITRSF        0.0kB      1.0TB          1  unlimited
* Account expiration: 2033-05-12

   Don't know where to start?
            --> http://www.ceci-hpc.be/install_software.html
            --> http://www.ceci-hpc.be/slurm_tutorial.html
[bvr@lm4-f001 ~]$
```

Vega

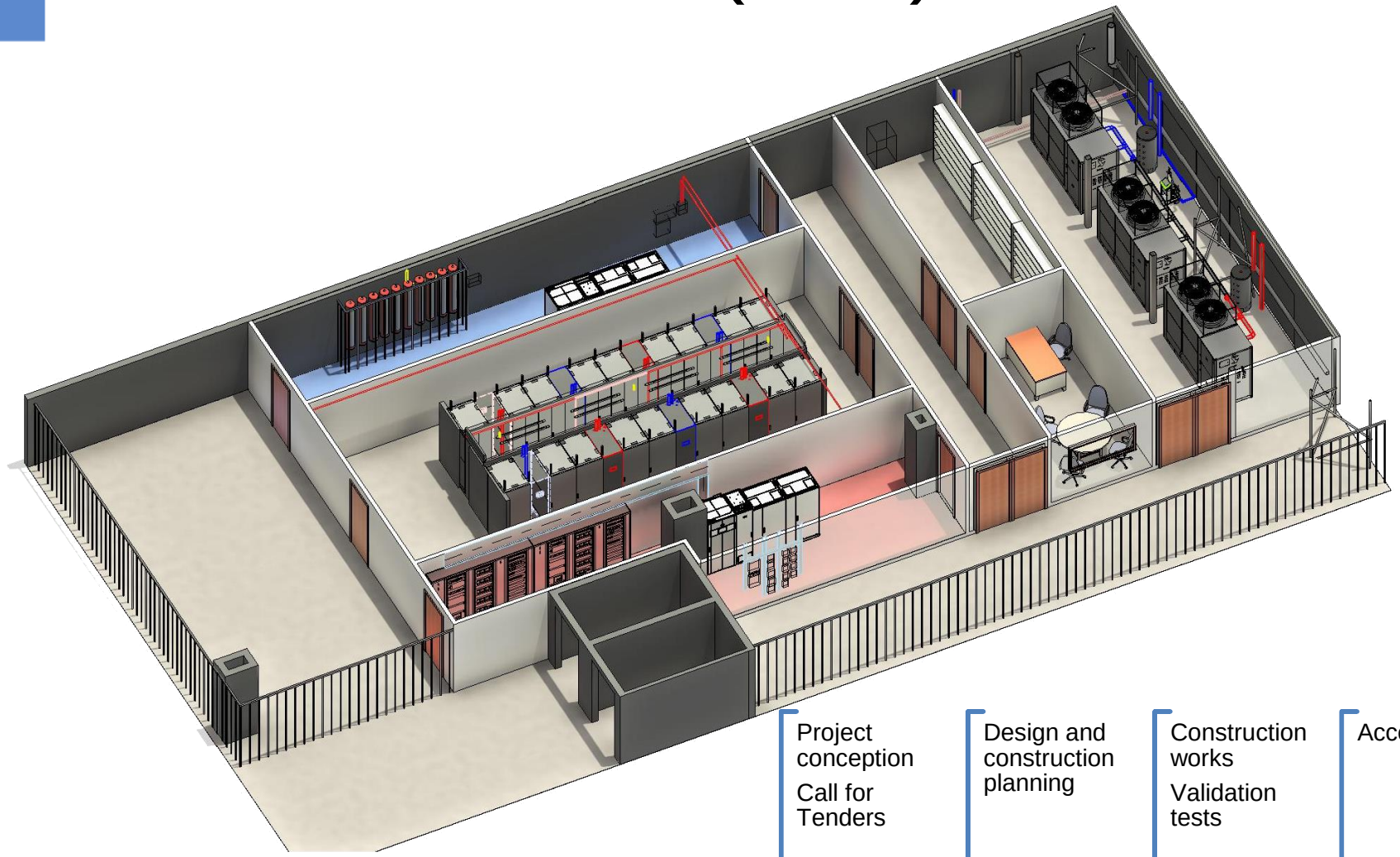
The CÉCI sysadmins

13th CÉCI Users Meeting

ULB – New CECI Cluster

- Vega Cluster decommissioned in Oct 2020
- ULB Solbosch DC not appropriate to host new cluster
- 2020 Q3 the ULB approved a long-awaited project for building a new DC

ULB – New DC (A6K)



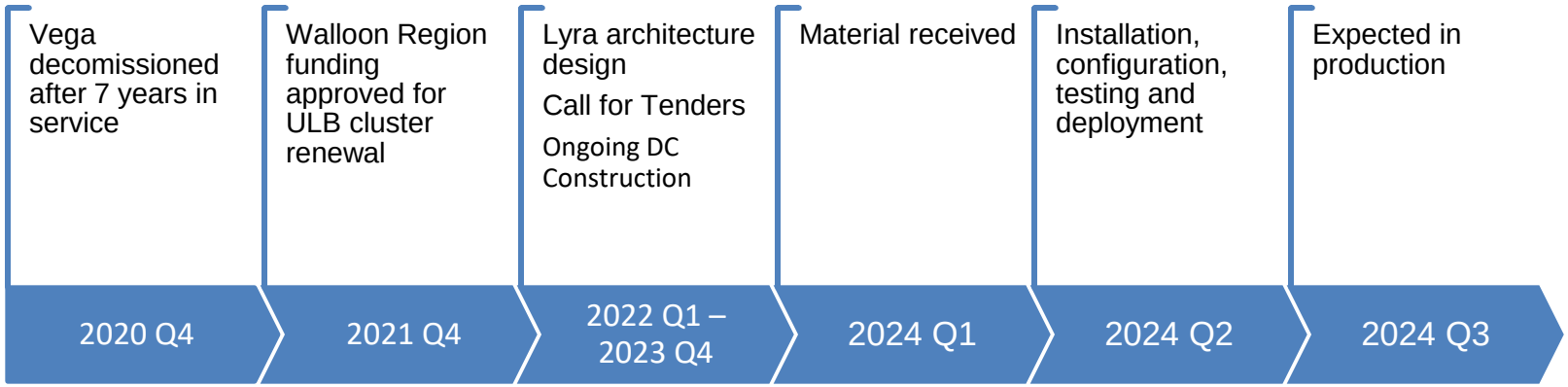
Project conception Call for Tenders	Design and construction planning	Construction works Validation tests	Acceptance
2021 Q1-Q4	2022 Q1 – Q2	2022 Q3 – 2024 Q1	2024 Q2

ULB – New DC (A6K)

- 18 racks 48U
- 250 kW of IT equipment
- Redundant power supply
- 2 power line sources with UPS
- Free cooling ~9 months/year
- PoP Belnet with redundant links
- Power Usage Effectiveness PUE ~= 1.2

ULB – New CECI cluster (Lyra)

- Walloon Region funding approved 2021 Q4
  - CECI cluster oriented for ML, AI, Big Data
  - All compute nodes will have GPUs



ULB – New CECI cluster (Lyra)

- 23x Servers:
  - 2x 32 core Epyc 9354 3.25GHz
  - 384GB RAM
  - 2x Nvidia RTX 6000 ADA
  - 6x 22TB HDD
  - 4x 2TB NVMe
  - 25GbE
- Totals
  - Cores: 1472
  - RAM: 8832 GB
  - Raw HDD: 3036 TB
  - Raw flash: 176 TB
  - GPUs: 46

ULB – New CECI cluster (Lyra)

Planned Hyperconverged architecture

- Everything will be virtualized
- 3x servers:
  - Run all management services and login nodes
  - Login nodes will have GPUs !
  - Compute nodes for debug partition
  - CEPH cluster for Home and Software (~100TB net)
- 20x servers:
  - 40x VMs as compute nodes
    - 28 Cores
    - 160GB RAM
    - 1x RTX A6000 ADA GPU
    - 2TB LOCALSCRATCH
  - Total dedicated to compute: 1120 cores / 6400 GB RAM
  - CEPH cluster for GLOBALSCRATCH (~880 TB net)

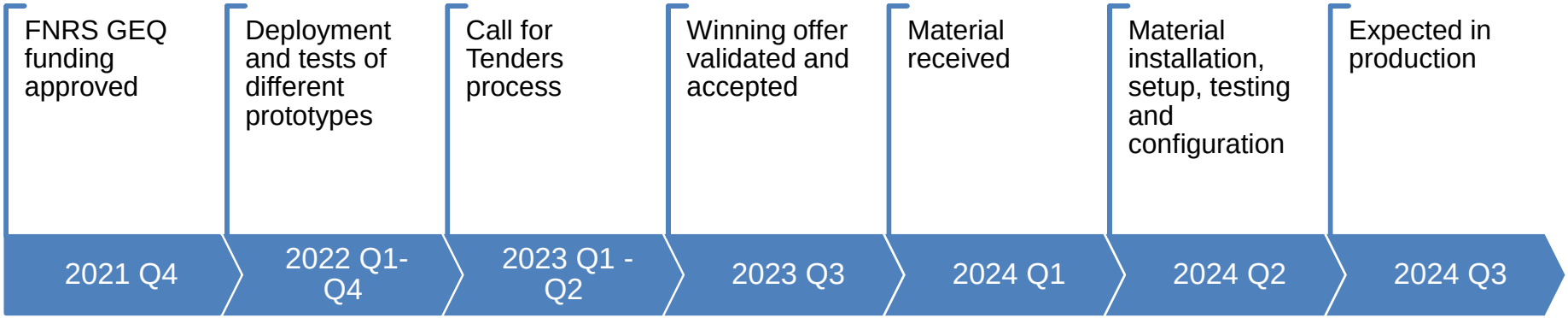
New CÉCI Common Storage

The CÉCI sysadmins

13th CÉCI Users Meeting

New CECI Common Storage

- Current common storage on production since 2017 Q2
- CECI submission to the 2021 FNRS GEQ call:
  - Storage solution, distributed and accessible by all CECI clusters
  - 1PB net minimum

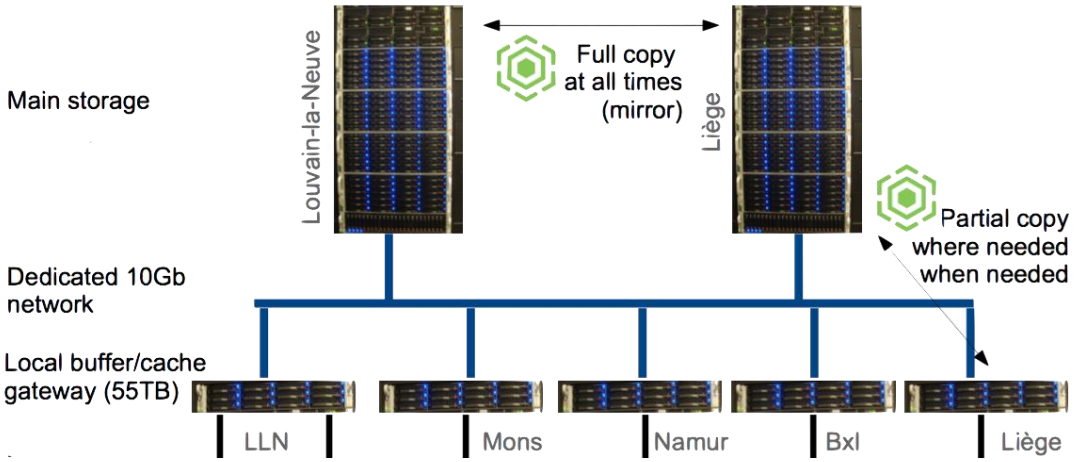


New CECI Common Storage

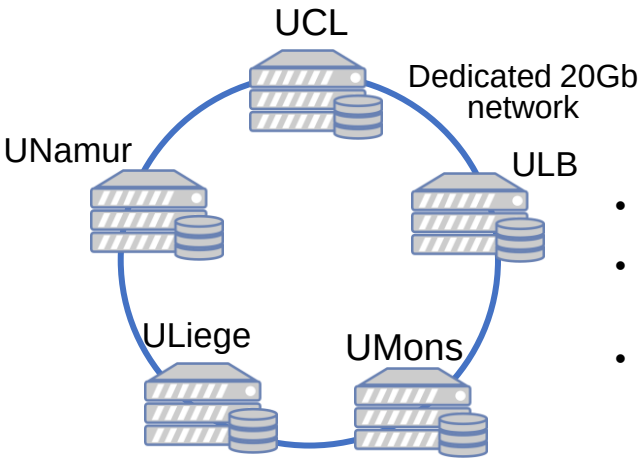
- Per site 4x storage servers:
  - 2x 16 core Epyc 9121 3GHz
  - 192GB RAM
  - 24x 7.68TB NVMe
  - 10GbE
- 20 Gbps Belnet multisite link
- Distributed CEPH cluster

New CECI Common Storage

Current GPFS storage



New CEPH storage



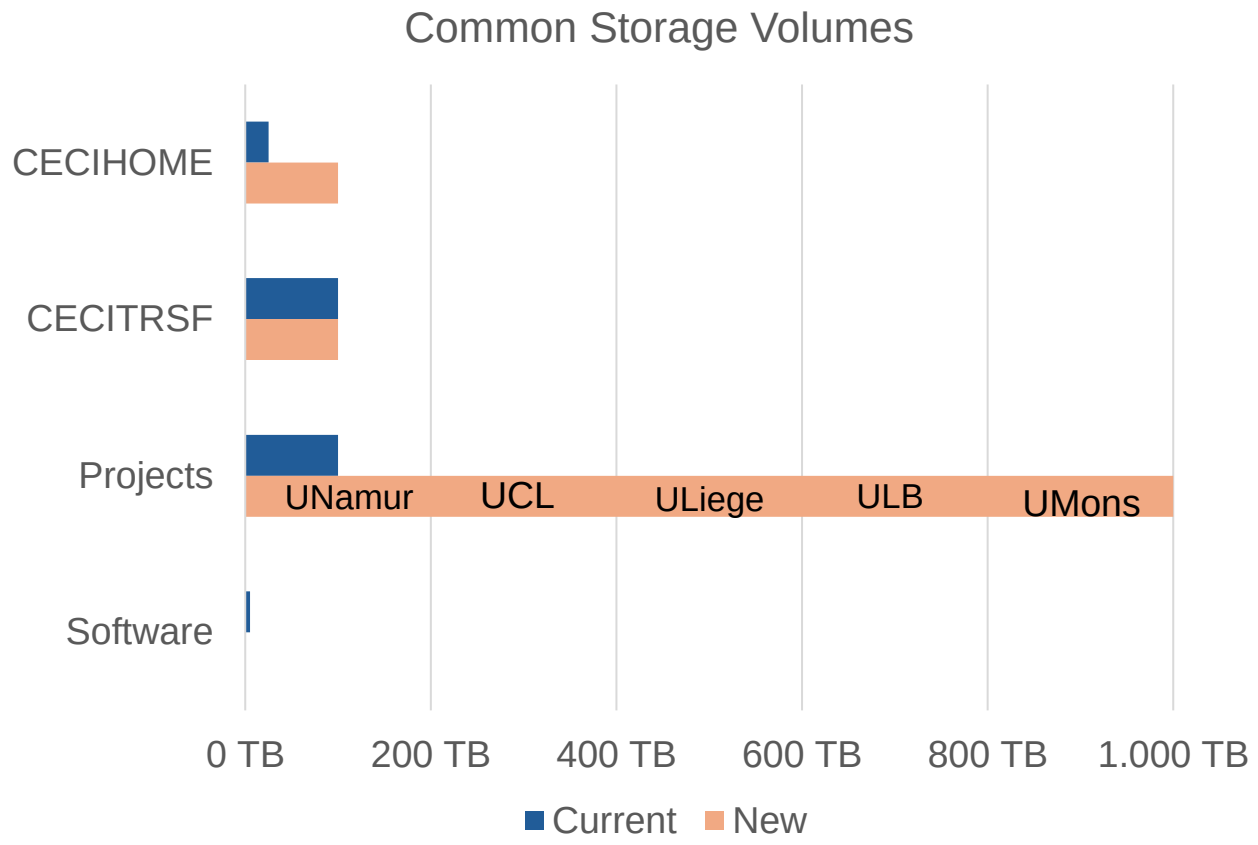
CEPH cluster

- Each site has identical equipment
- Global cephfs filesystem with 3x replication
- With this setup **~1200 TB** net

New CECI Common Storage

Planned volumes distribution

- Current storage: 230TB
- New storage: 1200TB

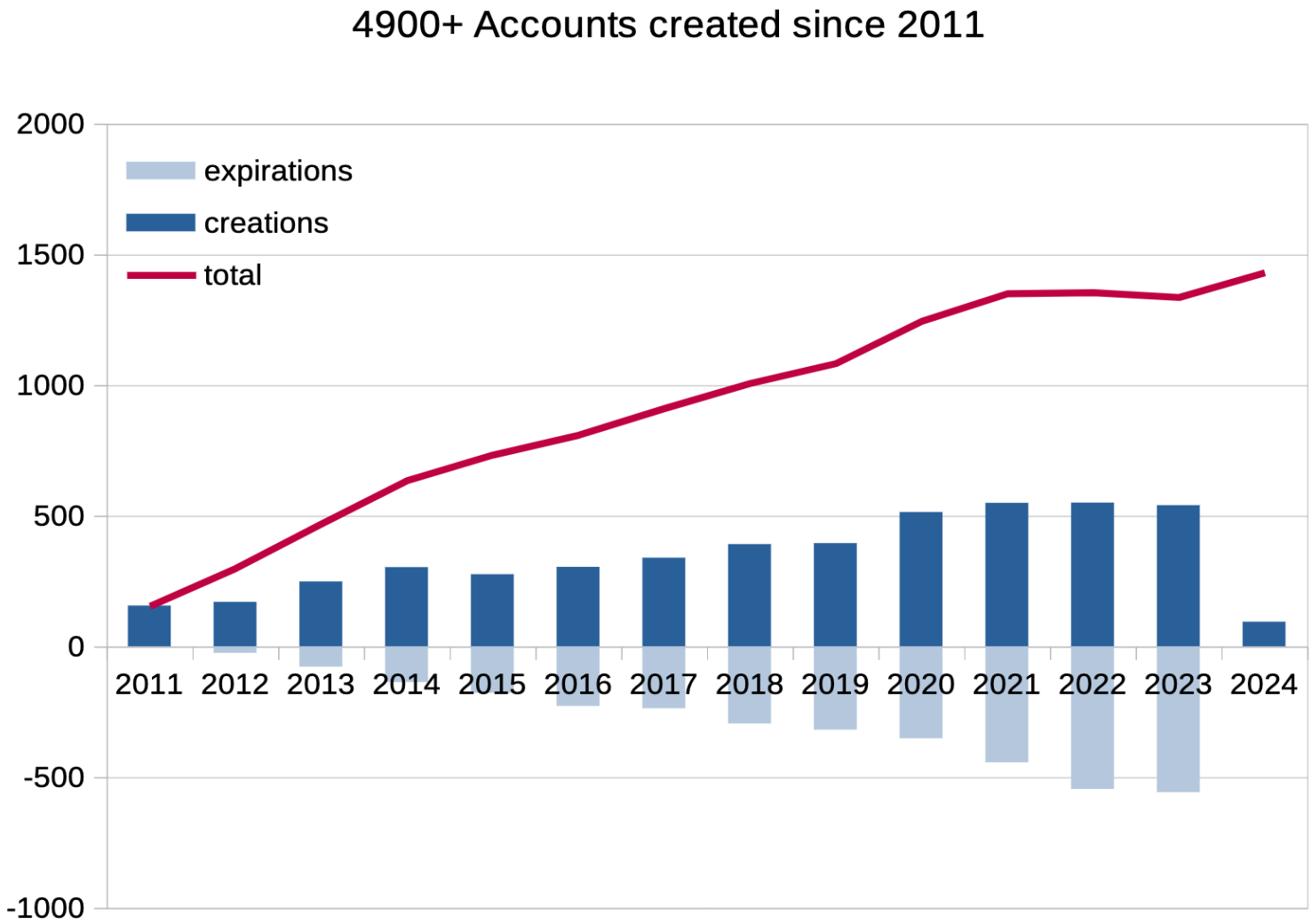


Help us help you!

The CÉCI sysadmins

13th CÉCI Users Meeting

Help "you": the CÉCI users



Help "us": the CÉCI sysadmins 💻

UCLouvain: **CISM** (4.4)
ULiège: **nicadm** (1)
UNamur: **PTCI** (1.5)
UMons: **CMN** (1)
ULB: **HPC team** (2)
CÉCI: Logisticien (1.5)

~11.5

Help "us": the CÉCI sysadmins 💻

https://www.ceci-hpc.be

Help "us": the CÉCI sysadmins 💻

UCLouvain: **CISM** (4.4)
ULiège: **nicadm** (1)
UNamur: **PTCI** (1.5)
UMons: **CMN** (1)
ULB: **HPC team** (2)
CÉCI: Logisticien (1.5)

~11.5

Our goals

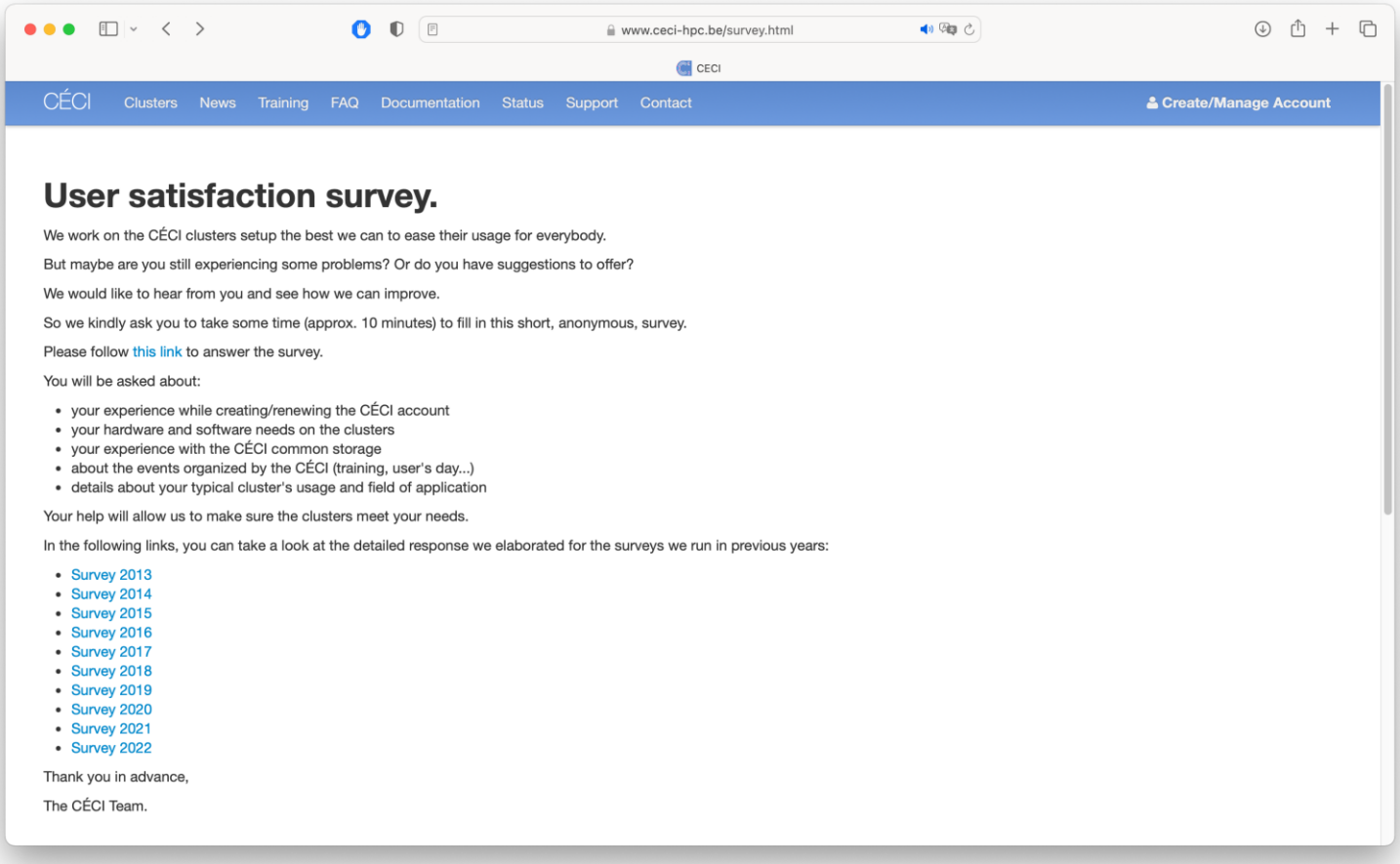
- provide as much uptime as possible
- accommodate a wide spectrum of jobs
- maximize resource utilization
- minimize turnaround for your jobs

Our messages today

- we can only act on your problems if you report them
- how you request help really impacts the possible time to response
- there are certain tasks you can do by yourself
- wasting resources has an impact on the planet and on other's research
- certain workflows can lead to problems on HPC clusters

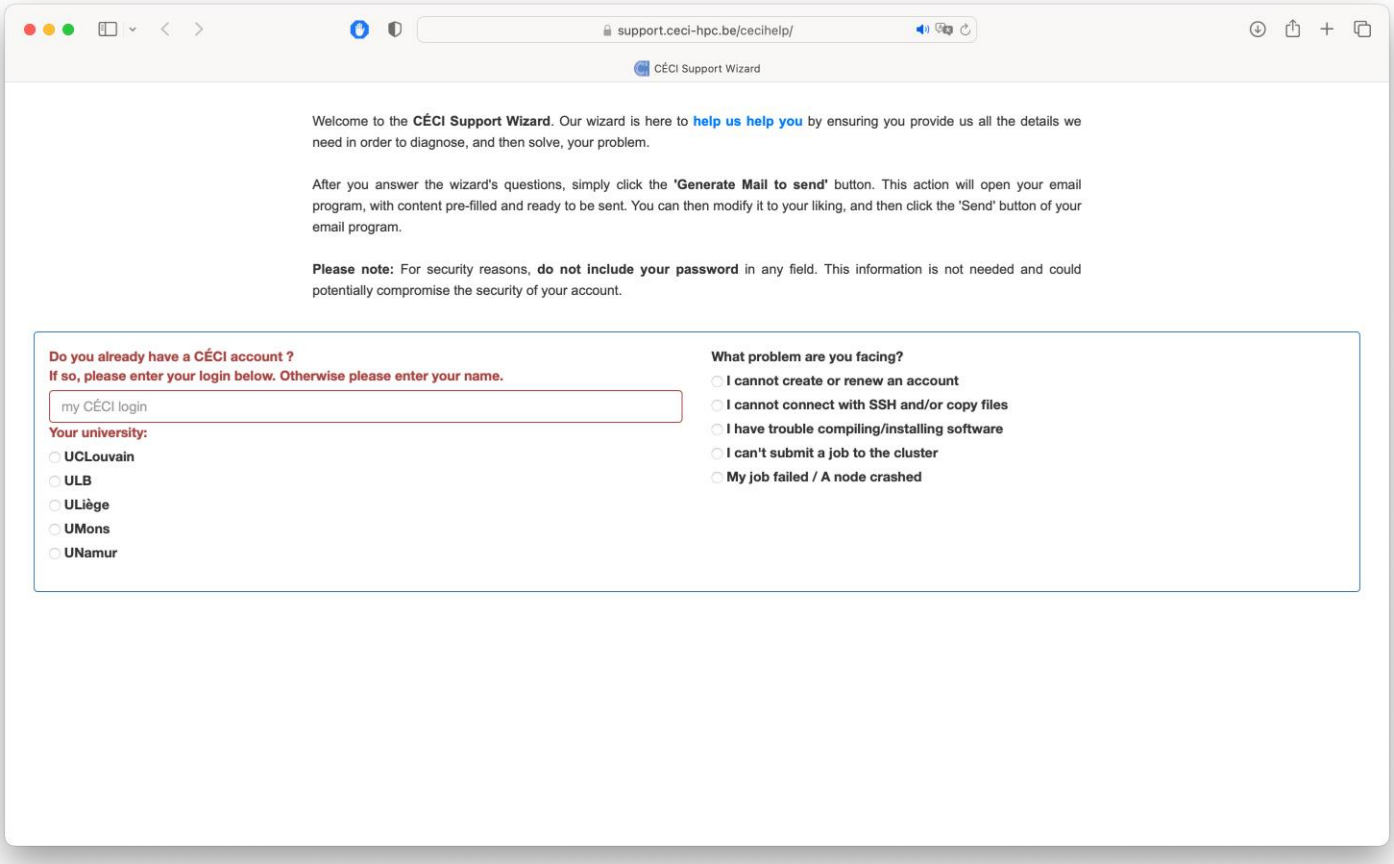
We can only act on your problems if you report them

- Via the annual satisfaction survey https://www.ceci-hpc.be/survey.html



We can only act on your problems if you report them

- Via the Support Wizard https://support.ceci-hpc.be/cecihelp/

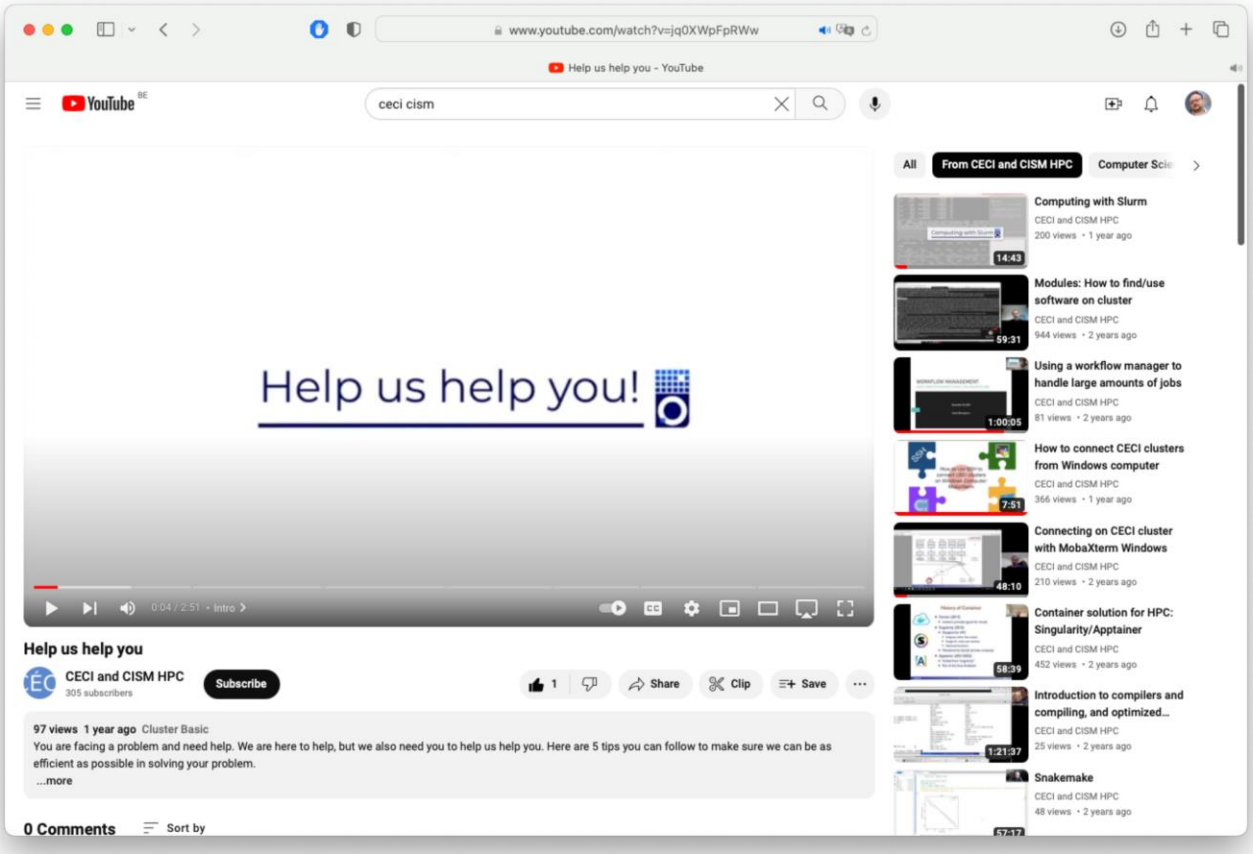


How you request help impacts the possible time to response

1. Start with **Due Diligence**
   - Read the error message, try to find words you understand
   - Check http://www.ceci-hpc.be/status.html
2. Use the **Correct Channel**: (support wizard unless stated otherwise)
3. State the **General Goal**; explain your intentions
4. Provide **Important Information**
   - W*ho*: what is your login?
   - *What*: what job ID, what file, what modules?
   - *When*: on which date, at what time?
   - *Where*: which cluster, which directory?
   - *Why*: what is the problem?
5. Give the **Exact Error** message you are facing

How you request help impacts the possible time to response

More info: http://www.ceci-hpc.be/helpushelpyou.html



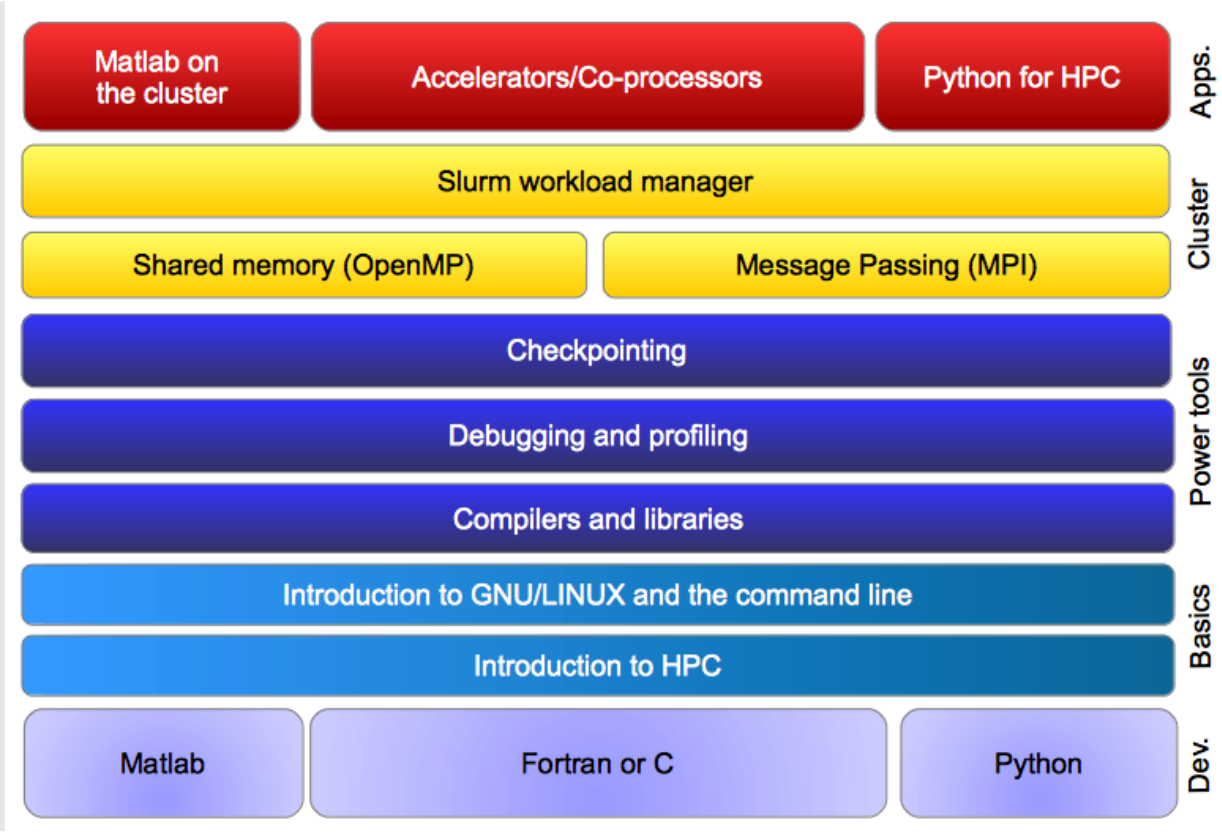
There are certain tasks you can do by yourself

- installing software (modules included)
- changing permissions back after an error
- sharing files among users
- changing group ownership (not user ownership)
- joining a tier-1 project
- learning Linux
- ...

Lemaitre3: module load gameshell

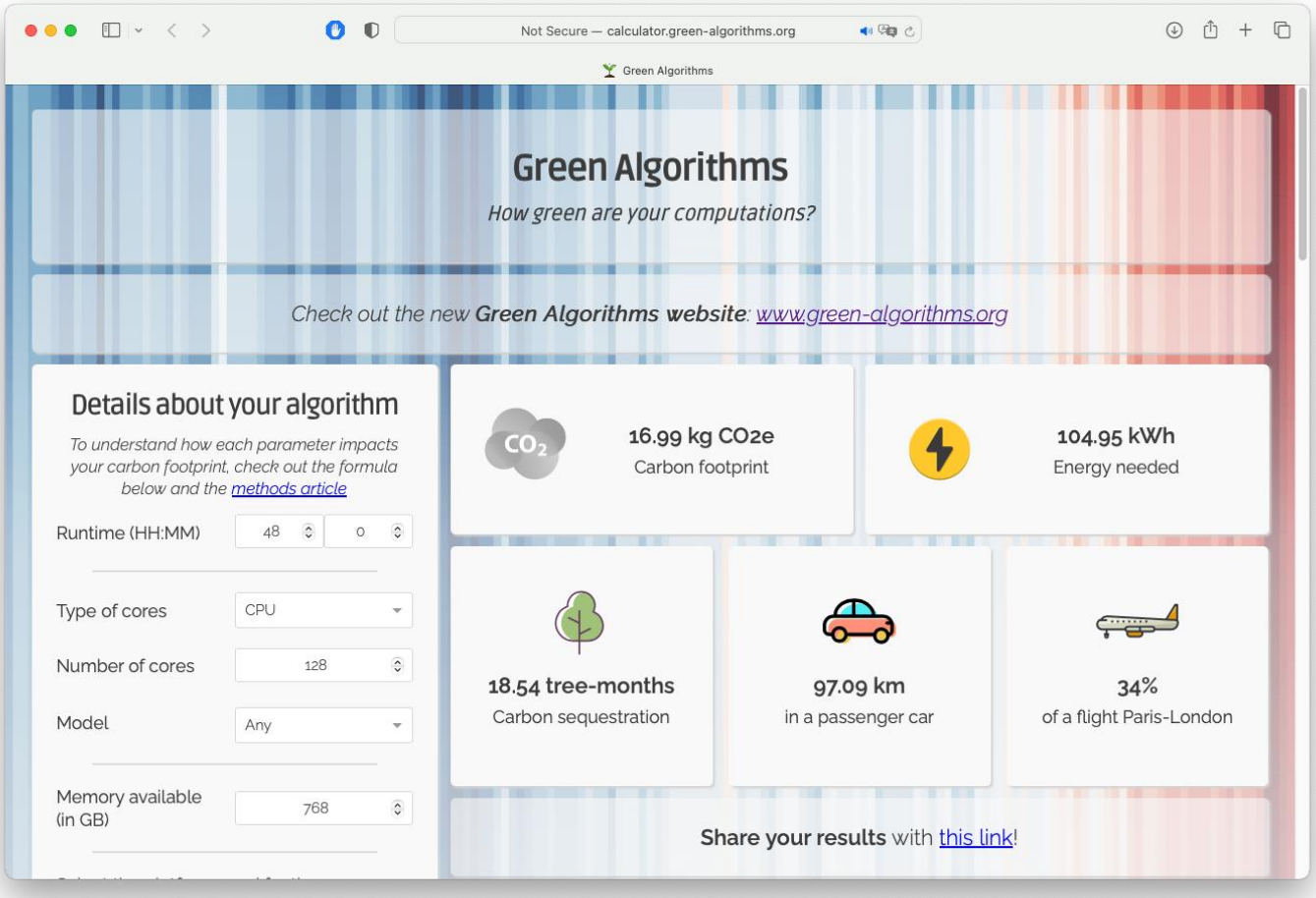
There are certain tasks you can do by yourself

https://www.ceci-hpc.be/training.html | https://www.cism.ucl.ac.be/videos



Wasting resources has an impact on the planet

More info: http://calculator.green-algorithms.org



Wasting resources has an impact on the planet

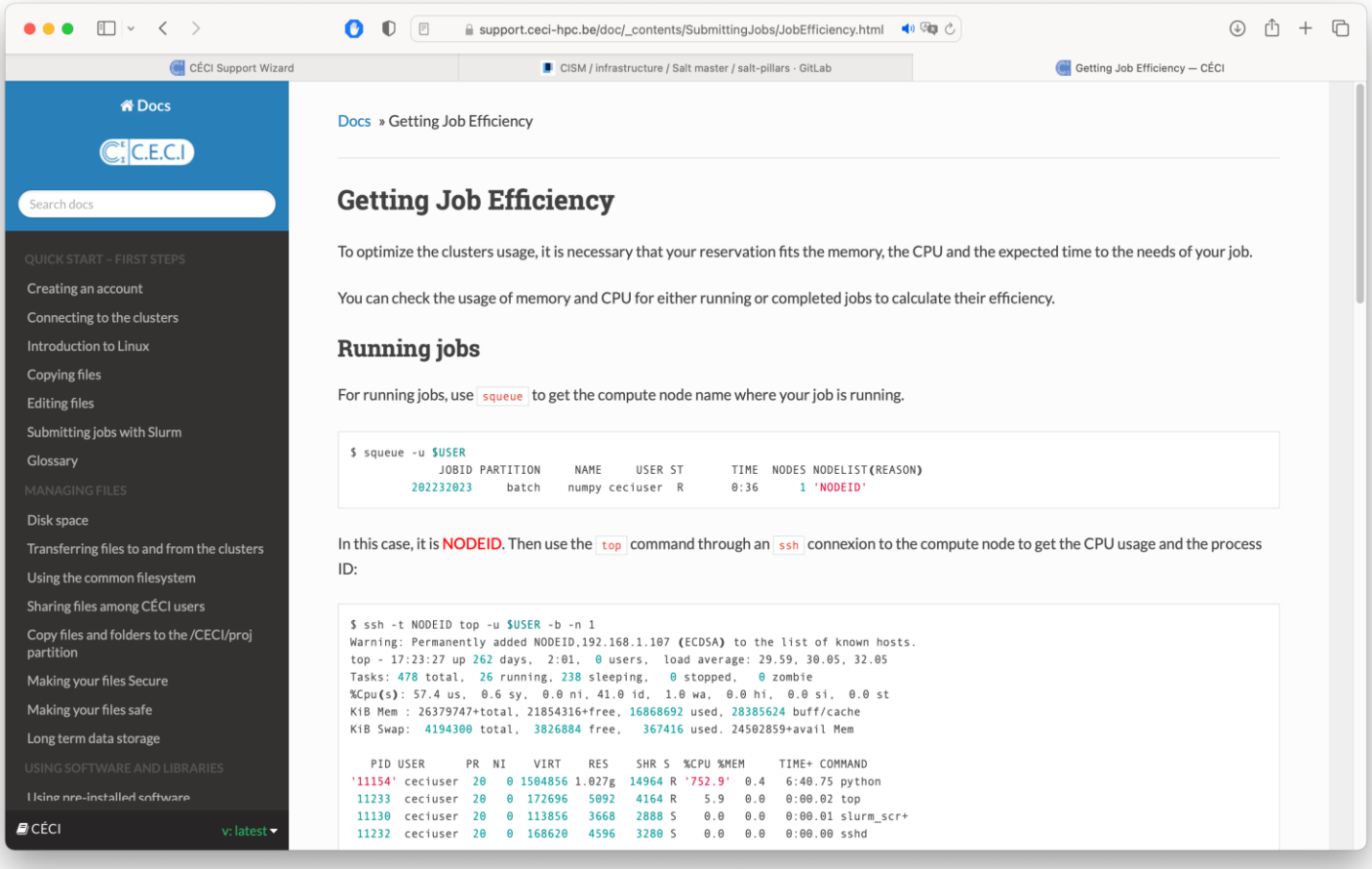
- jobs whose output is discarded because of misconfiguration
- jobs whose output is unsaved due to file manipulation error
- jobs whose results are lost because of hardware failure
- jobs not using the resources requested because of misconfiguration
- interactive jobs sitting idle
- jobs under-using resources due to bad scaling

are negatively impacting the climate and your colleagues' research

$$P(\text{idle}) \approx \tfrac{1}{2}\, P(\text{running})$$

Wasting resources has an impact on the planet

So be considerate of the resources you ask for and use them.
https://www.ceci-hpc.be/job_efficiency



Certain workflows create problems on HPC clusters

- Running anything CPU-intensive on the head node
- Issuing too many requests to the scheduler
- Not testing first on a small scale
- Excessive and/or “bad” I/O on a parallel/network filesystem
- Storing a large number of small files
- Not double checking the email options

are example of user behavior that impacts other users' experience

So please