



# The new CÉCI common storage

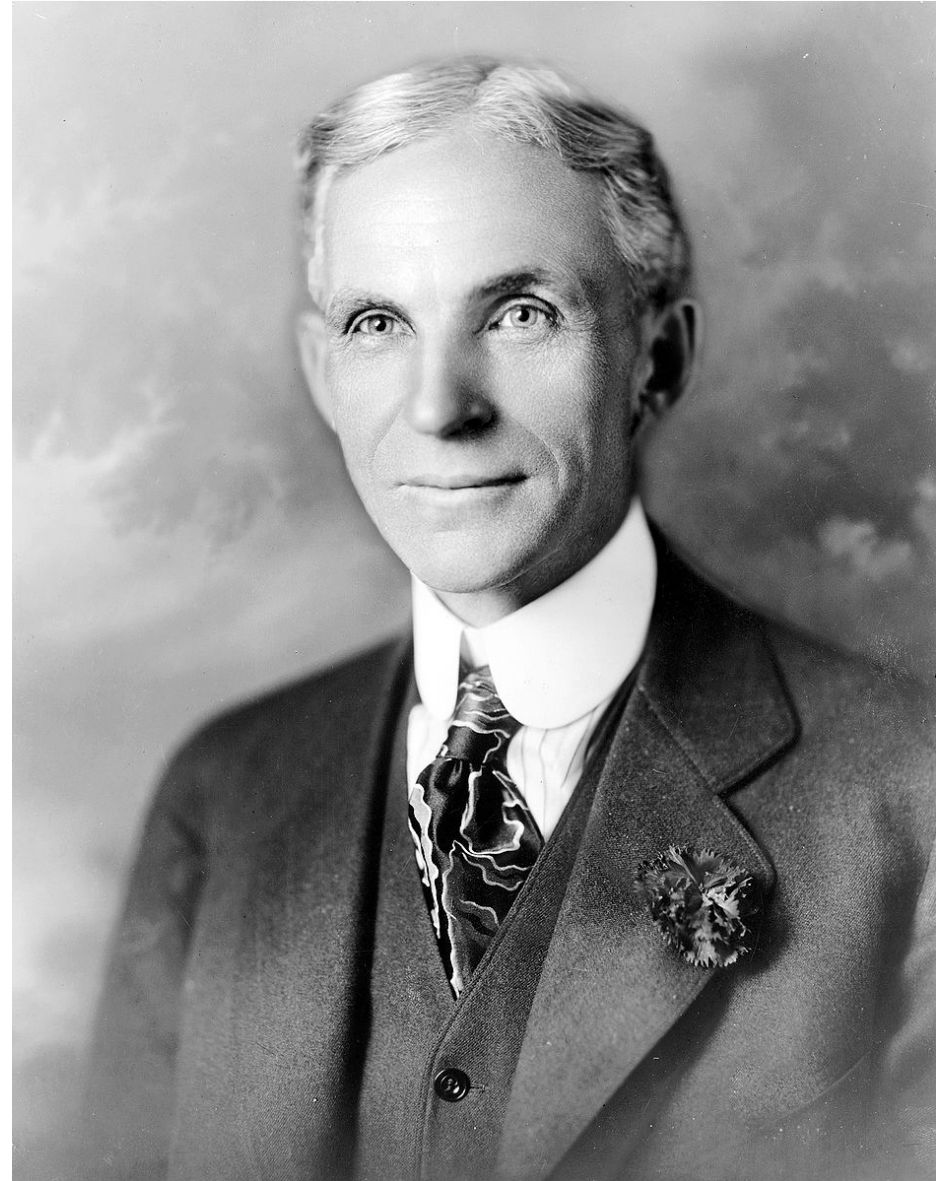
brought to you by Thomas Keutgen, David Colignon, Juan Cabrera, Frédéric Wautelet, Raphael Leplae, Bernard Van Renterghem, Sébastien Skozlowskij and Damien François

2017 CECI scientific day - Louvain-la-Neuve  
[damien.francois@uclouvain.be](mailto:damien.francois@uclouvain.be)  
[www.uclouvain.be/cism](http://www.uclouvain.be/cism)

# Henry Ford

allegedly once said:

**“If I had asked people  
what they wanted,  
they would have said  
faster horses.”**



# The CÉCI survey:

by the CÉCI users

## CÉCI cluster satisfaction survey

This survey aims at identifying the main obstacles the CÉCI users encounter when working with the clusters, so those can be alleviated.



Dear holder of a CÉCI account,

we are setting up the clusters (Hmem, Lemaitre2, Dragon1, Hercules, Vega and NIC4) the best we can to ease their usage for everybody. Maybe you are still experiencing problems ? Or you have suggestions to offer ? We would like to hear from you and see how we can help.

So please take some time to fill in this **short, anonymous**, survey.

You will be asked about:

- your experience while creating a CÉCI account,
- your hardware and software needs on the clusters,
- the details of your typical job, and
- your affiliation and field of application.

At the end, you will have the opportunity to request to be contacted for further help and to leave a suggestion or comment. Your help will allow us to make sure the clusters meet the needs of the users.

Note that this survey will **not refer** to the Tier1 cluster (Zenobe). Another survey, dedicated to Zenobe, will be setup in the near future.

Thank you in advance,

The CÉCI team

[www.cec-hpc.be](http://www.cec-hpc.be)

Next ▶

Exit and clear survey

# The CÉCI survey:

by the CÉCI users

“ more CPUs !

“ more CPUs !

“ *more CPUs*

“ more CPUs !!

“ *more CPUs*

“ more CPUs !

“ MORE CPUs !

“ more CPUs

“ more CPUs

“ ( \* - ת .

“ more CPUs !

“ more CPUs

“ more CPUs





# The CÉCI survey:

by the CÉCI users

“ Serait-il possible de mettre les 3 clusters (hmem, lemaitre2 et dragon1) **en nfs**? ”

“ Une chose utile serait d'avoir **du nfs** entre les différentes machines

“ having **shared space** for groups (like on Hydra)

“ Probably a dream but shared directories between clusters would be very interesting.

“ need of having a **common disk space** for my group/institute

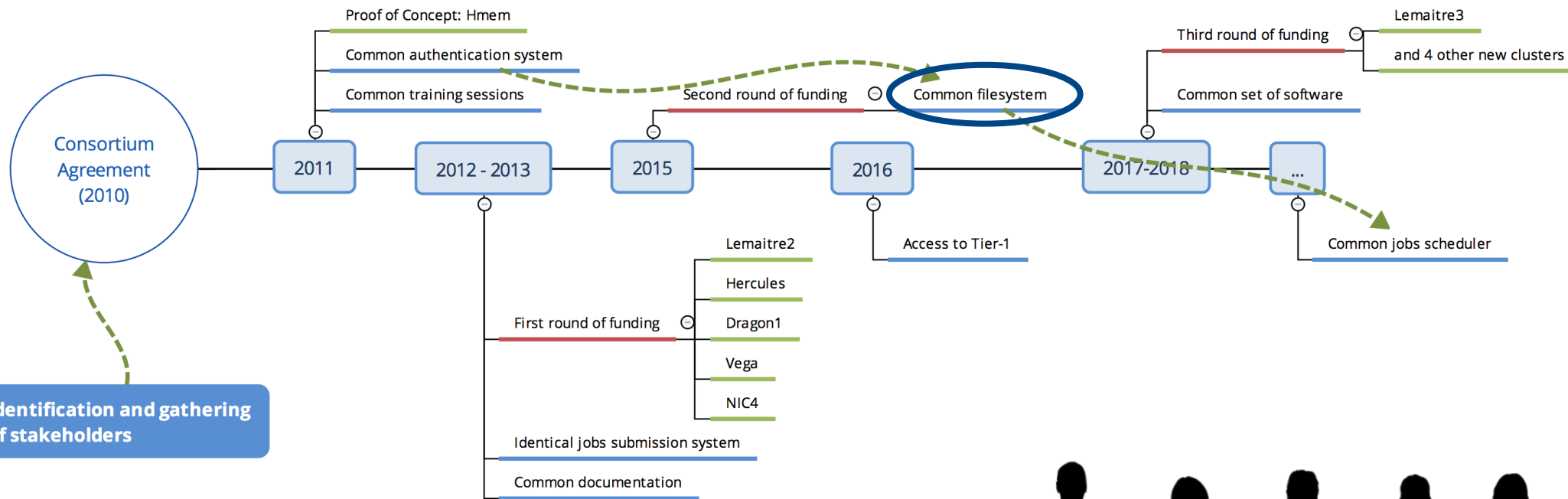
“ I would like this “**Common HOME folder**” feature for all the different clusters

“ Suggestions d'améliorations:[...] **Homogénéiser les modules** sur tous les clusters



# The CÉCI roadmap:

by the CÉCI Bureau



# What users want:

## Explicitly

single file namespace visible from all compute nodes

## Implicitly

no degradation in service level:

- availability
- bandwidth
- latency

- writing of the submission file for 500k€
- 311 pages
- dozens of contributors

- about 15 vendors consulted;
- 4 software solutions considered;
- 5 solution designs evaluated;
- 2 network solutions tested;

- description of solution
- design of benchmarks

- 4 responses
- 1 clear winner

Dec. '13 Decision of the Bureau

March '14 Submission to FNRS

July '14 Grant from FNRS (300k€)  
+ 200k€ from the five universities

June '15 Solution is elaborated

Nov. '15 Technical part of RFP written

March '16 RFP published

May '16 Offers analyzed ; market attributed

Oct. '16 Hardware installed

Jan. '17 Software (pre-)configured

Feb. '17 Sysadmin trained

March '17 Configuration finalized



# The new CÉCI common storage Features



6 clusters  
5 sites

Vega



Dragon1



NIC4



Lemaitre2



Hmem



Hercules

new



10Gbps, dedicated,  
optical network direct  
links between the sites

# 1. New network

Vega



Dragon1



NIC4



Lemaitre2



Hmem



Hercules





New hardware

## 2. New hardware



Two large storage systems:

- Liège
- Louvain-la-Neuve (DCIII)



One smaller storage system on each site



New hardware

## 2. New hardware



Two large storage systems:

- Liège
- **Louvain-la-Neuve (DCIII)**



One smaller storage system on each site

## Louvain-la-Neuve (DCIII)

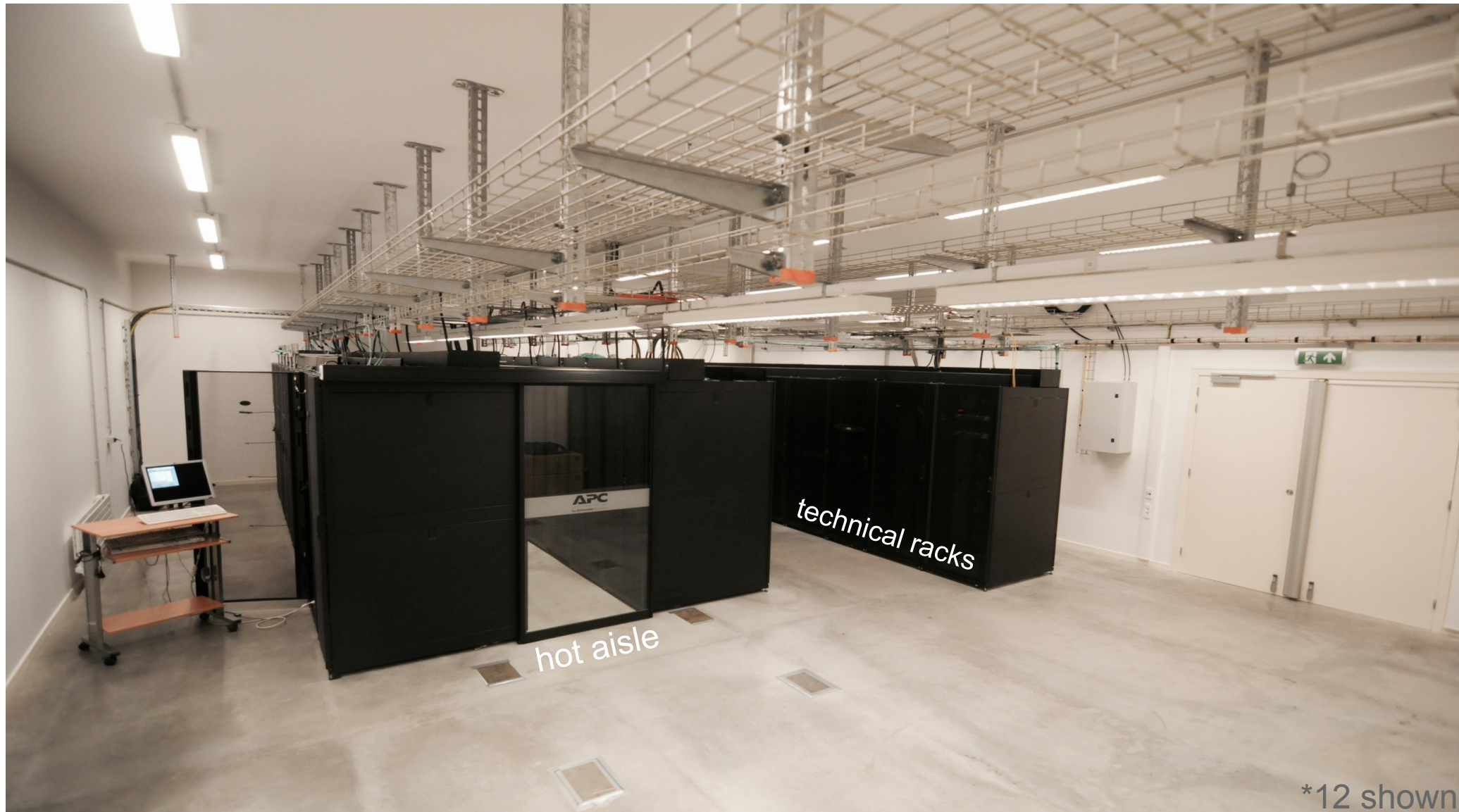
Brand new data center 200 kW cooling capacity (extension to 400kW next year)





## Louvain-la-Neuve (DCIII)

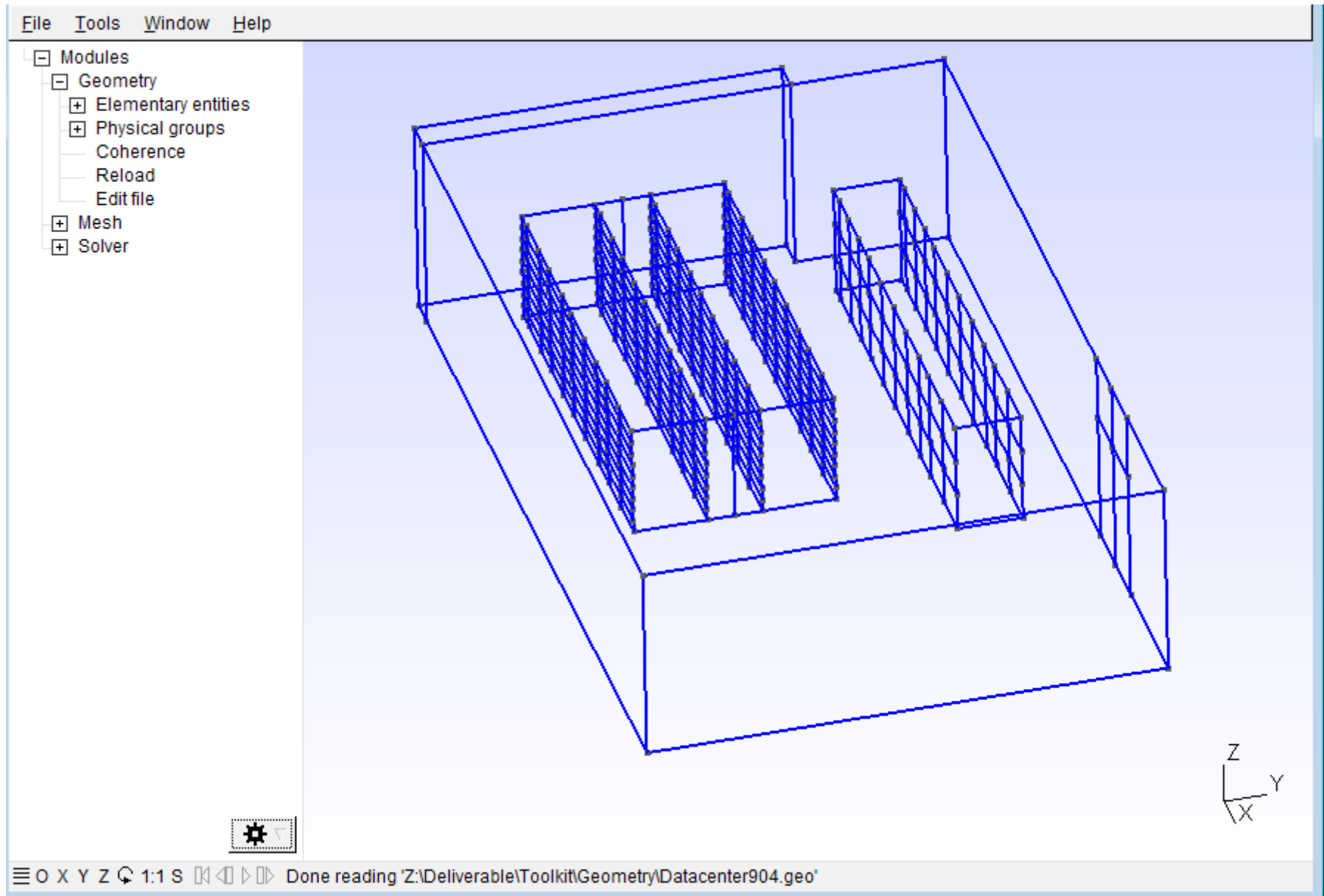
20 racks\* confined hot aisle  
UPS 3 to 4 hours autonomy now (designed for 10 minutes at 400kW)



\*12 shown

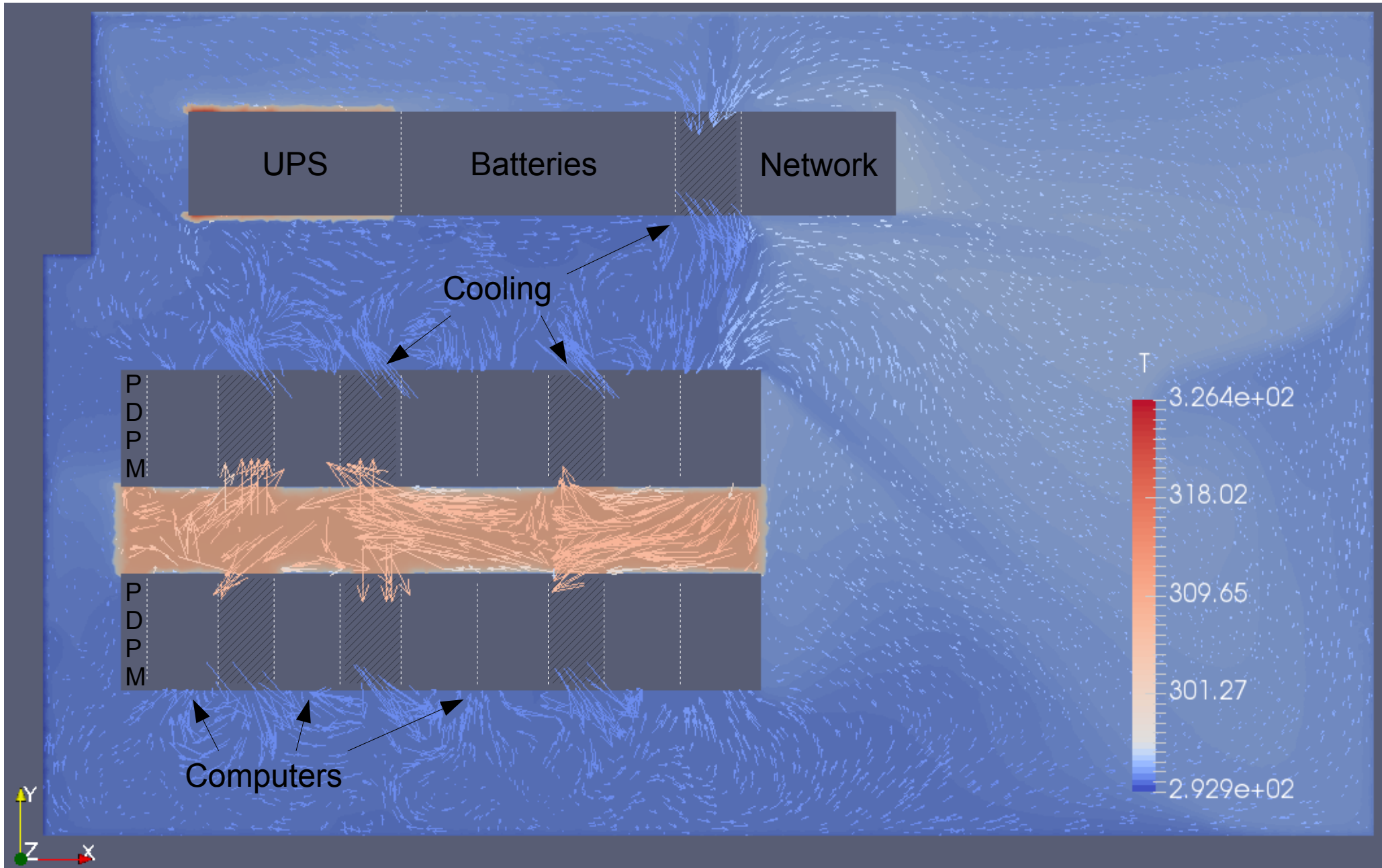
# Louvain-la-Neuve (DCIII)

Master Thesis by Vincent Flon  
(Supervisors Y. Bartosiewicz and P. Chatelain)



# Louvain-la-Neuve (DCIII)

Master Thesis by Vincent Flon  
(Supervisors Y. Bartosiewicz and P. Chatelain)





## Louvain-la-Neuve (DCIII)

10 water-based heat exchangers\*



\*3 visible



## Louvain-la-Neuve (DCIII)

Pumps and buffer tanks in the basement





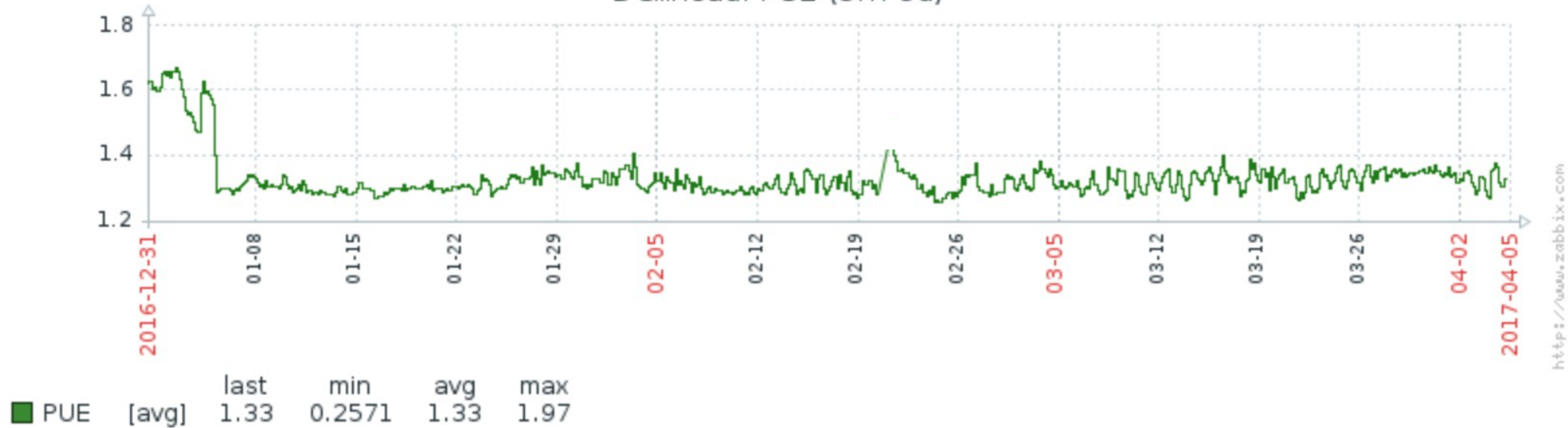
## Louvain-la-Neuve (DCIII)

One air cooler and two chillers (one more next year) on the roof

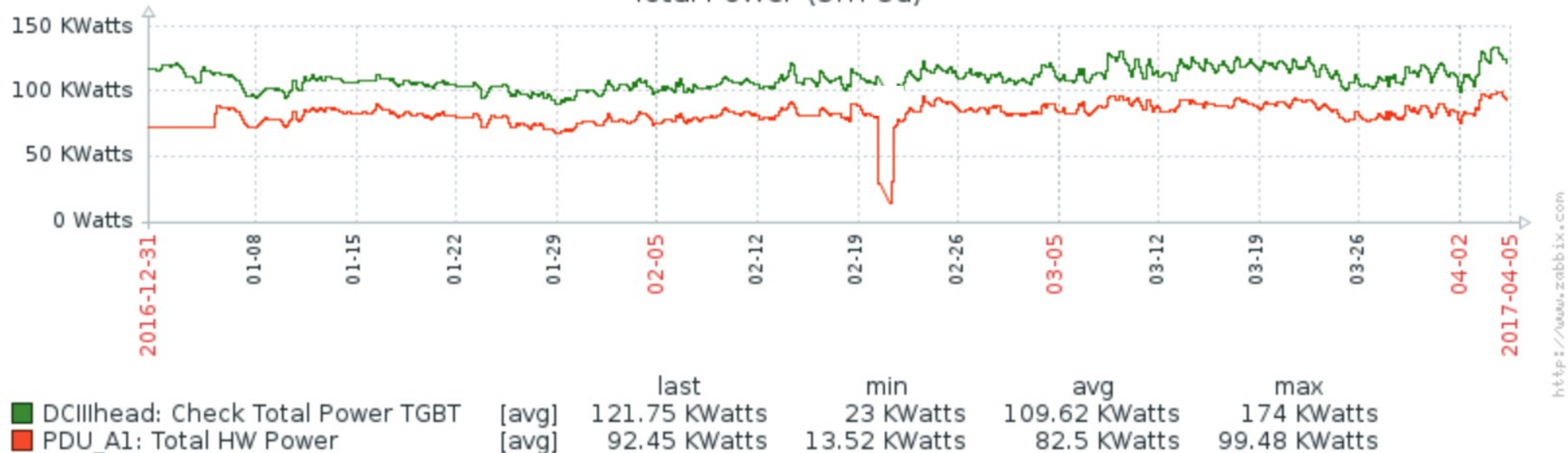


# Louvain-la-Neuve (DCIII)

DCIIIhead: PUE (3m 5d)



Total Power (3m 5d)





# Louvain-la-Neuve (DCIII)

rtbf

Info Sport Culture Audio TV Radio

Recherche

INFO

À la une Fil Info Belgique Régions Europe Monde Économie Société Médias

Direct Commission Publinfin: jeudi 06 avril

Régions Bruxelles Brabant Wallon Hainaut Liège Namur Luxembourg Flandre

Louvain-la-Neuve: l'UCL privée d'électricité durant une petite heure

SUIVEZ-NOUS

l'avenir.net

ACTU RÉGIONS SPORT BUZZ LIFE CULTURE OPINIONS PROXIMAG

EXL BRAB. WALLON NAMUR LIÈGE CHARLEROI BASSE-SAMBRE SAMBRE-MEUSE VERVIERS WALL PICARDE MONS&CENT

ACTU SPORTS

PLUS D'INFO L'ACTUALITÉ DE LOUVAIN-LA-NEUVE

ARTICLES LIÉS

IMPRIMER

À LA UNE

LES + LUS

MAIL

6 shares

RTT INFO

ACTU SPORT PEOPLE & BUZZ VOUS VIDÉOS

EN CE MOMENT Commission Publinfin Présidentielle française

Panne de courant sur le campus de l'UCL à Louvain-la-Neuve: les cours sont annulés

David Fourmanois, publié le 20 février 2017 à 15h55, mis à jour à 19h05

Sudinfo.be

Brabant wallon

Coupure de courant à Louvain-la-Neuve: problème repéré, retour progressif à la normale

Rédaction en ligne | Publié le Lundi 20 Février 2017 à 17h49

Une coupure de courant s'est produite ce lundi sur le coup de 15h10 à Louvain-la-Neuve. Pratiquement l'ensemble du site néolouvaniste est touché et les cours ont d

LE SOIR

22° min 10° -0.55% BEL 20 06/04

Actu Sports Culture Économie Débats Blogs

Actu Régions Brabant wallon

Genvat: «Pouvoir offrir un wi-fi de qualité devient un must dans les hôtels» Plus de 40.000 flashes sur nos routes en

Recommander Partager 4 Tweeter G+ 0 in Share 3

Louvain-la-Neuve: le courant rétabli sur le campus de l'UCL

Rédaction en ligne Mis en ligne lundi 20 février 2017, 20h07

Soul un bâtiment est encore concerné par la panne

Journal Alertez-nous Je me connecte Radio Newsletter

Menu DH.be

ABONNEZ-VOUS

EN CE MOMENT Moments forts du débat télévisé | Bientôt de l'électricité sans fil ? | Quelle météo cette semaine ?

L'électricité est de retour à Louvain-la-Neuve

RÉDACTION EN LIGNE Publié le lundi 20 février 2017 à 15h21 - Mis à jour le lundi 20 février 2017 à 18h40

L'alimentation électrique est rétablie sur le campus universitaire de Louvain-la-Neuve

BELGA Publié le lundi 20 février 2017 à 15h21 - Mis à jour le lundi 20 février 2017 à 18h39



## Louvain-la-Neuve (DCIII)

Full capacity tests made with heating devices





New hardware

## 2. New hardware



Two large storage systems:

- Liège
- **Louvain-la-Neuve (DCIII)**  
**Visits organized today!**



One smaller storage system on each site

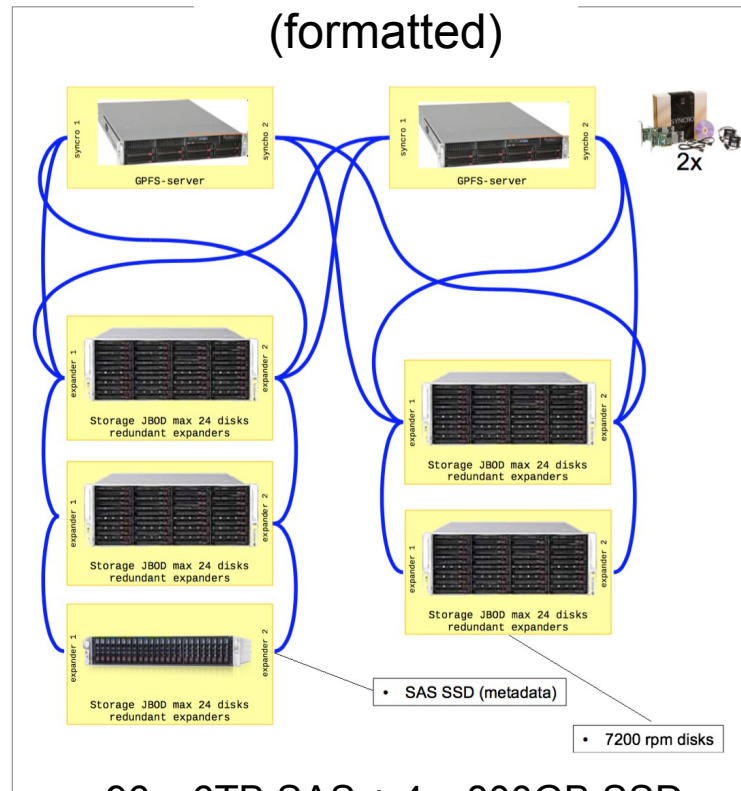




New hardware

2x

450TB netto  
(formatted)



96 x 6TB SAS + 4 x 800GB SSD  
8 x (10+2) RAID6

Each system is robust to:

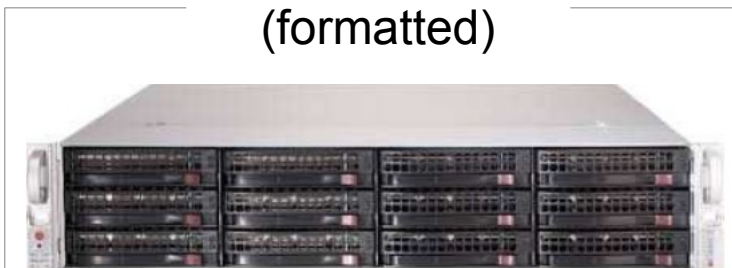
- loss of one server
- loss of one pathway
- loss of two disks

Solution is robust to:

- loss of an entire system
- loss of connectivity

5x

55TB netto  
(formatted)



12 x 6TB SAS + 2 x 480GB SSD  
1 x (10+2) RAID6

Each system is robust to:

- loss of two disks

Solution is robust\* to:

- loss of entire disk array
- loss of network access



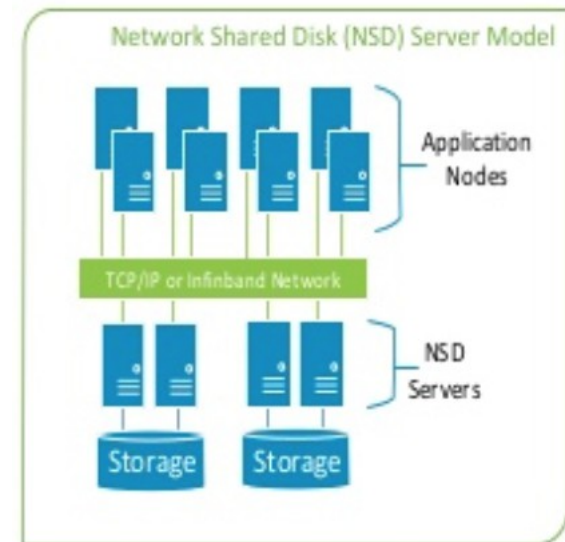


New software

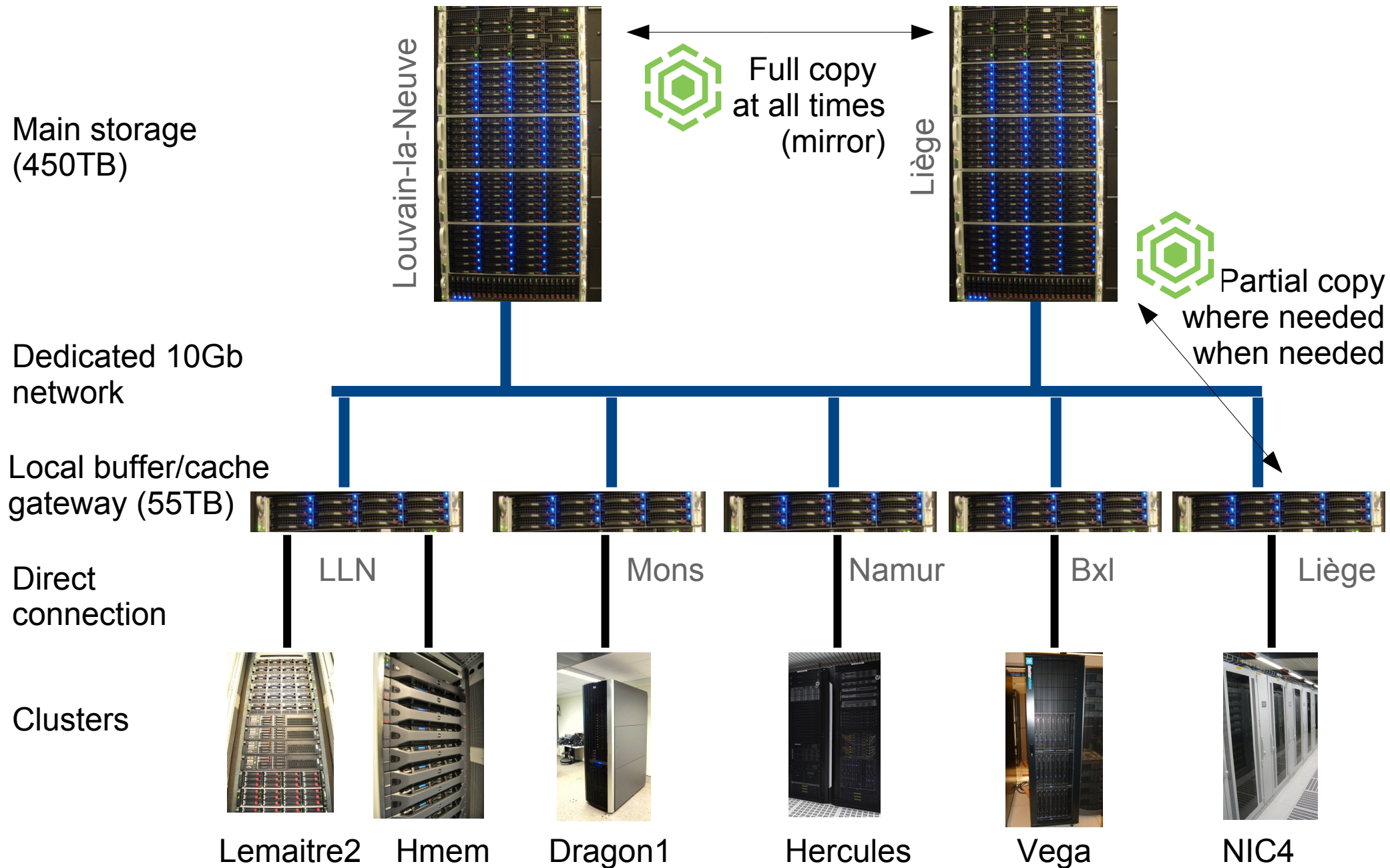
### 3. New software



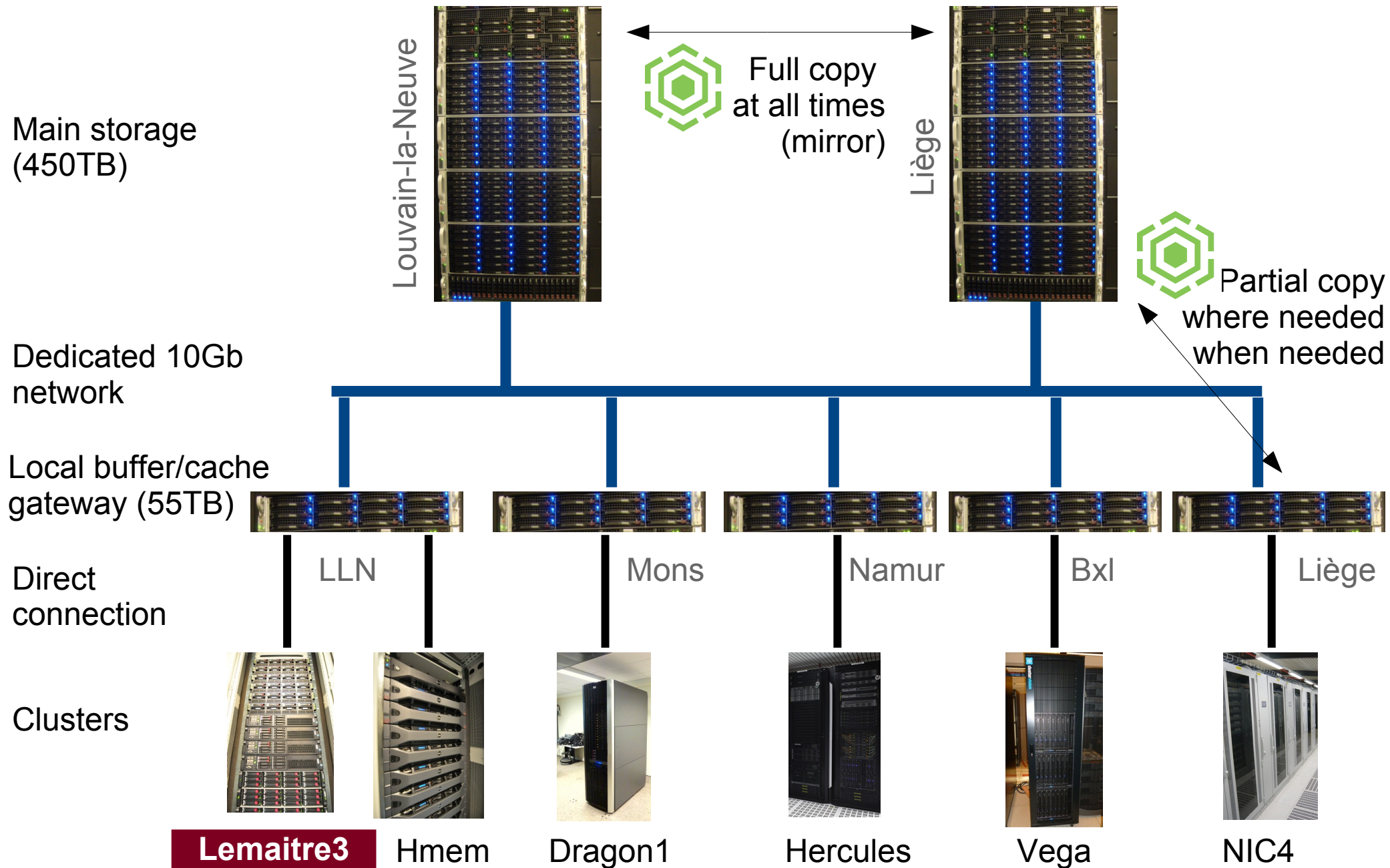
IBM  
**Spectrum**  
Scale



# Setup overview



# Setup overview



hopefully

- Next-gen processors



or



- Low-latency interconnect



or



or equivalent

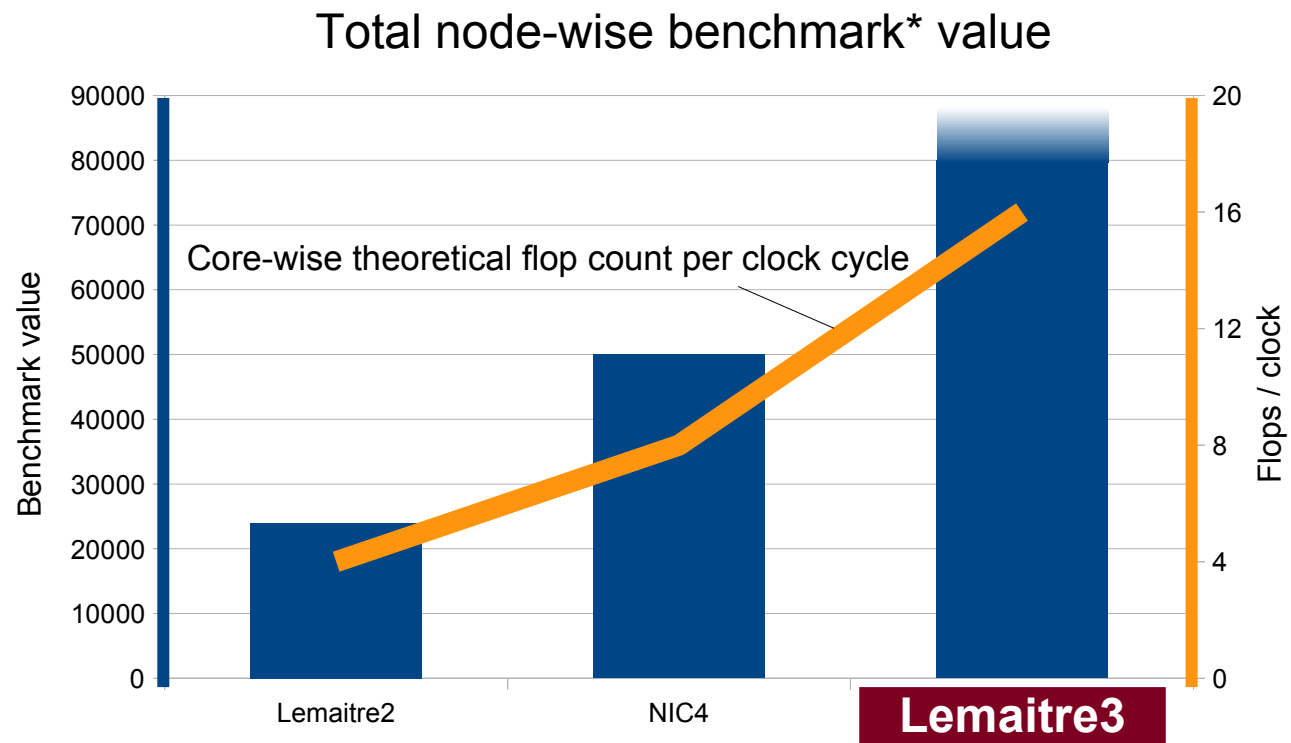
- Parallel filesystem



or



- Slurm configuration close to that of NIC4



\*Spec FP rate base 2006 per nodes times number of nodes

# Lemaitre3

YOU  
ARE  
HERE

January						
Su	Mo	Tu	We	Th	Fr	Sa
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

February						
Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28				

March						
Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

April						
Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

May						
Su	Mo	Tu	We	Th	Fr	Sa
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

June						
Su	Mo	Tu	We	Th	Fr	Sa
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

July						
Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

August						
Su	Mo	Tu	We	Th	Fr	Sa
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

September						
Su	Mo	Tu	We	Th	Fr	Sa
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

October						
Su	Mo	Tu	We	Th	Fr	Sa
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

November						
Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

December						
Su	Mo	Tu	We	Th	Fr	Sa
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

RFP  
Publication

Offer  
analysis

Winner  
notification

Procurement

Installation

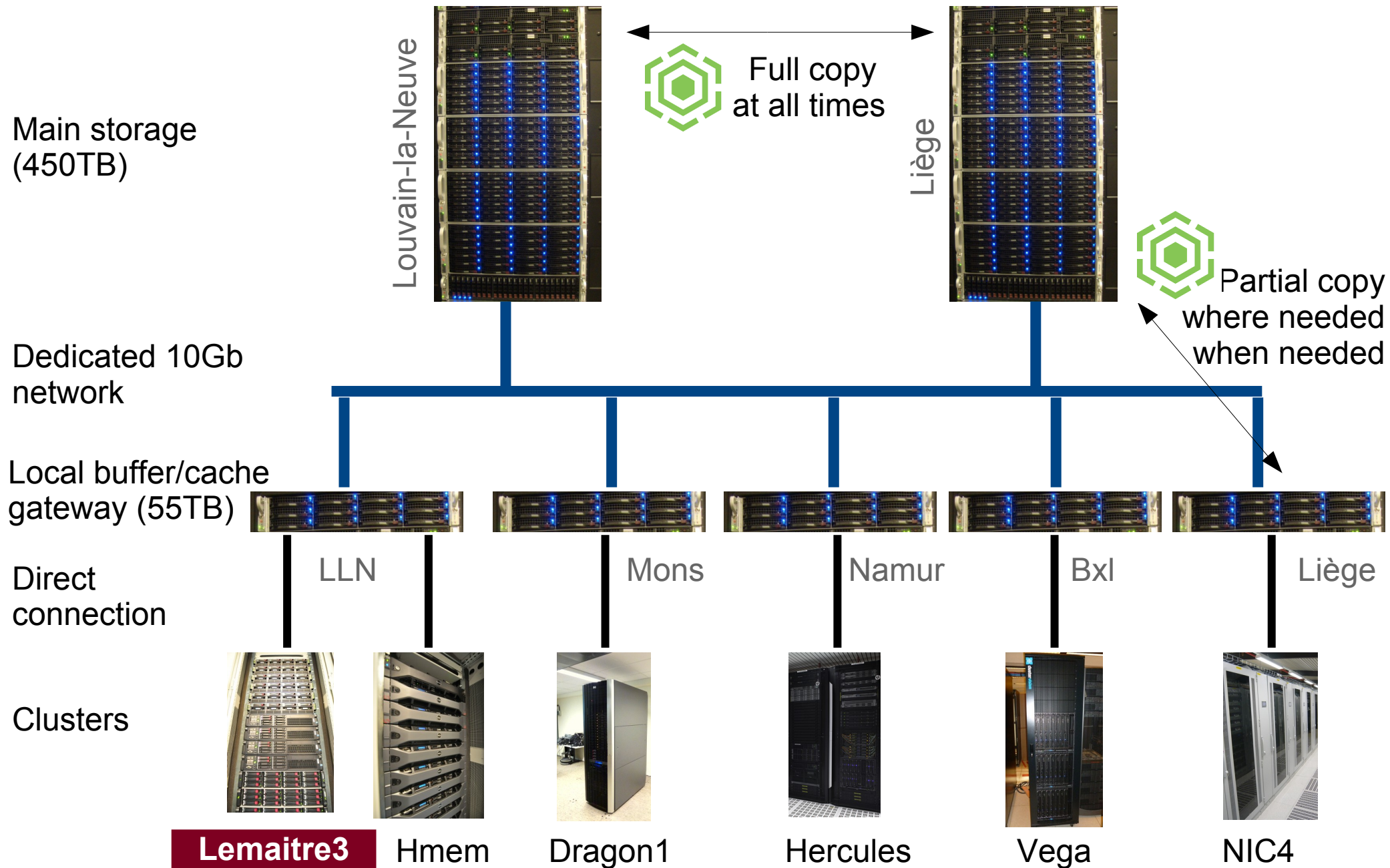
Lemaitre3

and beyond, funding permitting...

- 2017
  - Lemaitre 3: Large parallel jobs (MPI)
- 2018
  - Hercules “2”: Large-memory (TBs RAM)
  - Dragon “2”: Accelerators (GPUs, Phi's)
  - Vega “2”: HTC and Big-Data
- 2019
  - NIC “5”: Large parallel jobs (MPI)



# Setup overview





The new CÉCI  
common storage  
Short-term benefits

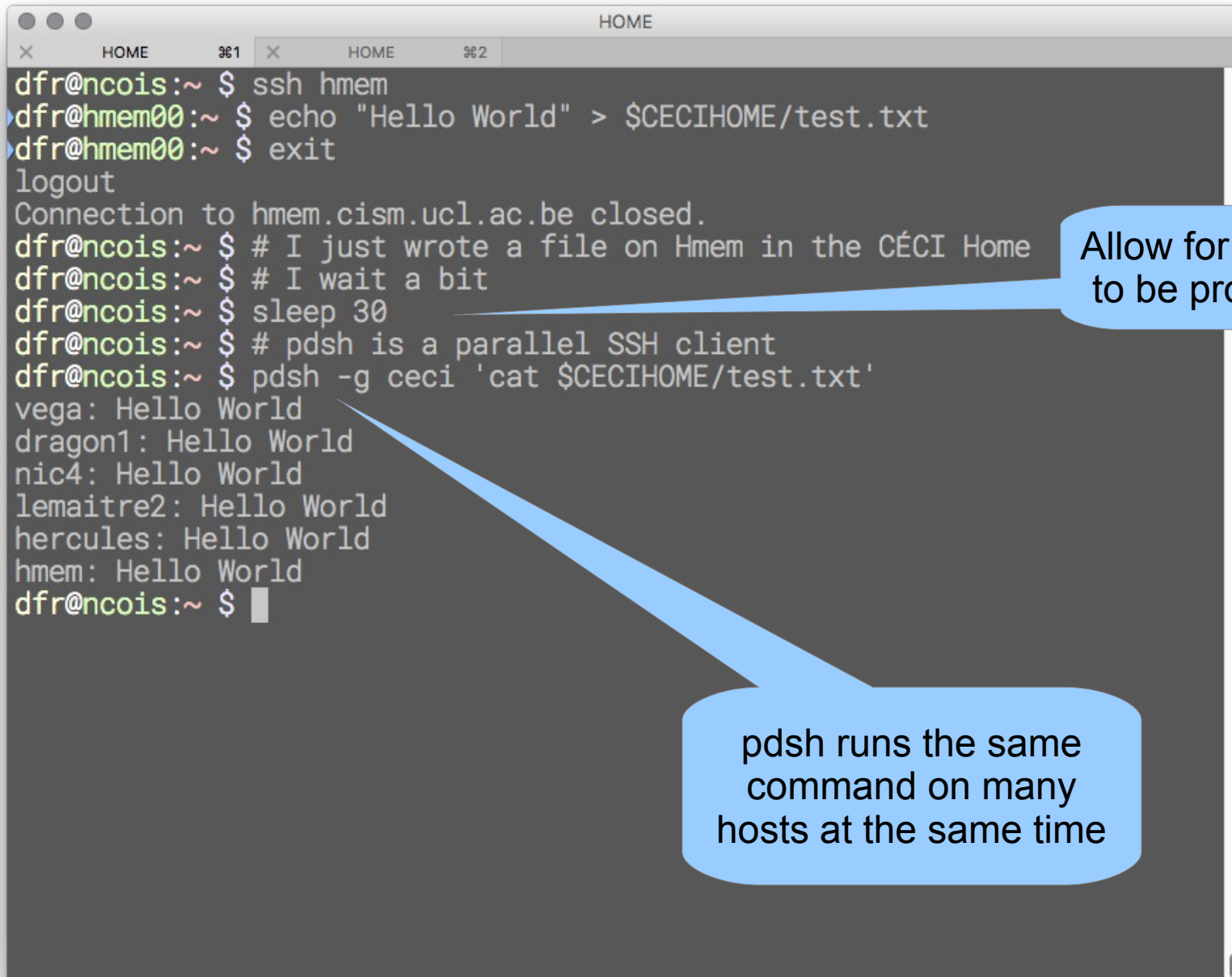
# Four spaces

- /CECI/home
  - Quota 100GB/User
  - Daily snapshots
- /CECI/proj
  - Upon request
  - Quota and duration based on request
- /CECI/trsf
  - Quota per user 100GB soft 10TB hard
  - Automatic purge of files older than 6 months
- /CECI/soft
  - Common software + modules

# Four spaces

- /CECI/home
  - Quota 100GB/User
  - Daily snapshots
- /CECI/proj
  - Upon request
  - Quota and duration based on request
- /CECI/trsf
  - Quota per user 100GB soft 10TB hard
  - Automatic purge of files older than 6 months
- /CECI/soft
  - Common software + modules

# Files written to \$CECIHOME are visible on all clusters



```
dfr@ncois:~ $ ssh hmem
dfr@hmem00:~ $ echo "Hello World" > $CECIHOME/test.txt
dfr@hmem00:~ $ exit
logout
Connection to hmem.cism.ucl.ac.be closed.
dfr@ncois:~ $ # I just wrote a file on Hmem in the CÉCI Home
dfr@ncois:~ $ # I wait a bit
dfr@ncois:~ $ sleep 30
dfr@ncois:~ $ # pdsh is a parallel SSH client
dfr@ncois:~ $ pdsh -g ceci 'cat $CECIHOME/test.txt'
vega: Hello World
dragon1: Hello World
nic4: Hello World
lemaitre2: Hello World
hercules: Hello World
hmem: Hello World
dfr@ncois:~ $
```

Allow for changes  
to be propagated

pdsh runs the same  
command on many  
hosts at the same time



# Files written to \$CECIHOME are visible on all clusters

```
HOME
x HOME %1 x HOME %2
Connection to hmem.cism.ucl.ac.be closed.
dfr@ncois:~$ # I just wrote a file on Hmem in the CÉCI Home
dfr@ncois:~$ # I wait a bit
dfr@ncois:~$ sleep 30
dfr@ncois:~$ # pdsh is a parallel SSH client
dfr@ncois:~$ pdsh -g ceci 'cat $CECIHOME/test.txt'
vega: Hello World
dragon1: Hello World
nic4: Hello World
lemaitre2: Hello World
hercules: Hello World
hmem: Hello World
dfr@ncois:~$ ssh hmem 'echo "Goodbye!" >> $CECIHOME/test.txt'
dfr@ncois:~$ sleep 30
dfr@ncois:~$ pdsh -g ceci 'cat $CECIHOME/test.txt'
vega: Hello World
vega: Goodbye!
dragon1: Hello World
dragon1: Goodbye!
nic4: Hello World
nic4: Goodbye!
hmem: Hello World
hmem: Goodbye!
lemaitre2: Hello World
lemaitre2: Goodbye!
hercules: Hello World
hercules: Goodbye!
dfr@ncois:~$
```

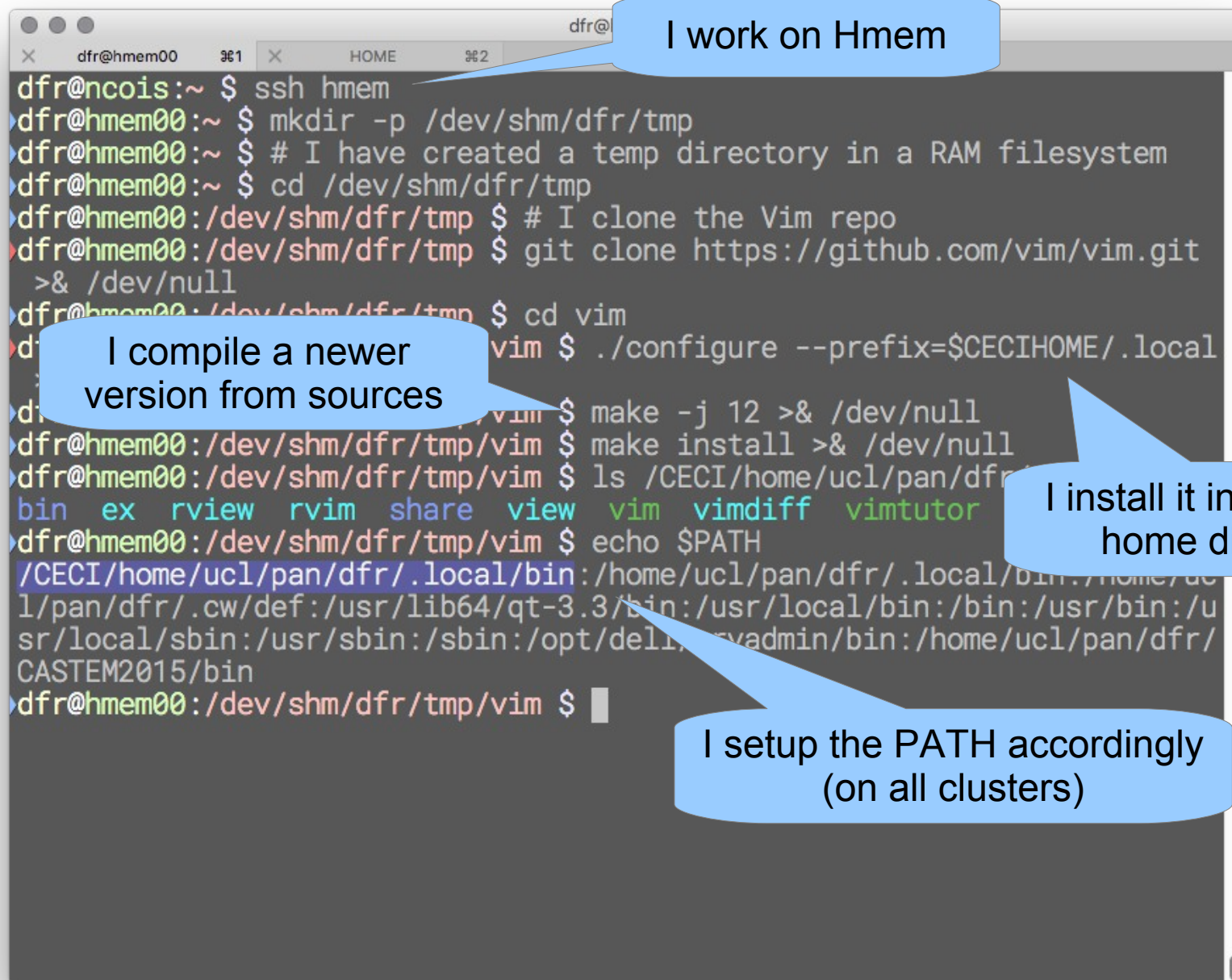
Also visible on all  
compute nodes of  
all clusters !

# All clusters have a different Vim...





# Example: install your own version of Vim on every cluster



```
dfr@ncois:~ $ ssh hmem
dfr@hmem00:~ $ mkdir -p /dev/shm/dfr/tmp
dfr@hmem00:~ $ # I have created a temp directory in a RAM filesystem
dfr@hmem00:~ $ cd /dev/shm/dfr/tmp
dfr@hmem00:/dev/shm/dfr/tmp $ # I clone the Vim repo
dfr@hmem00:/dev/shm/dfr/tmp $ git clone https://github.com/vim/vim.git
>& /dev/null
dfr@hmem00:/dev/shm/dfr/tmp $ cd vim
dfr@hmem00:/dev/shm/dfr/tmp/vim $ ./configure --prefix=$CECIHOME/.local
dfr@hmem00:/dev/shm/dfr/tmp/vim $ make -j 12 >& /dev/null
dfr@hmem00:/dev/shm/dfr/tmp/vim $ make install >& /dev/null
dfr@hmem00:/dev/shm/dfr/tmp/vim $ ls /CECI/home/ucl/pan/dfr/
bin  ex  rview  rvim  share  view  vim  vimdiff  vimtutor
dfr@hmem00:/dev/shm/dfr/tmp/vim $ echo $PATH
/CECI/home/ucl/pan/dfr/.local/bin:/home/ucl/pan/dfr/.local/bin:/home/ucl/pan/dfr/.cw/def:/usr/lib64/qt-3.3/bin:/usr/local/bin:/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/sbin:/opt/dell/svadmin/bin:/home/ucl/pan/dfr/CASTEM2015/bin
dfr@hmem00:/dev/shm/dfr/tmp/vim $
```

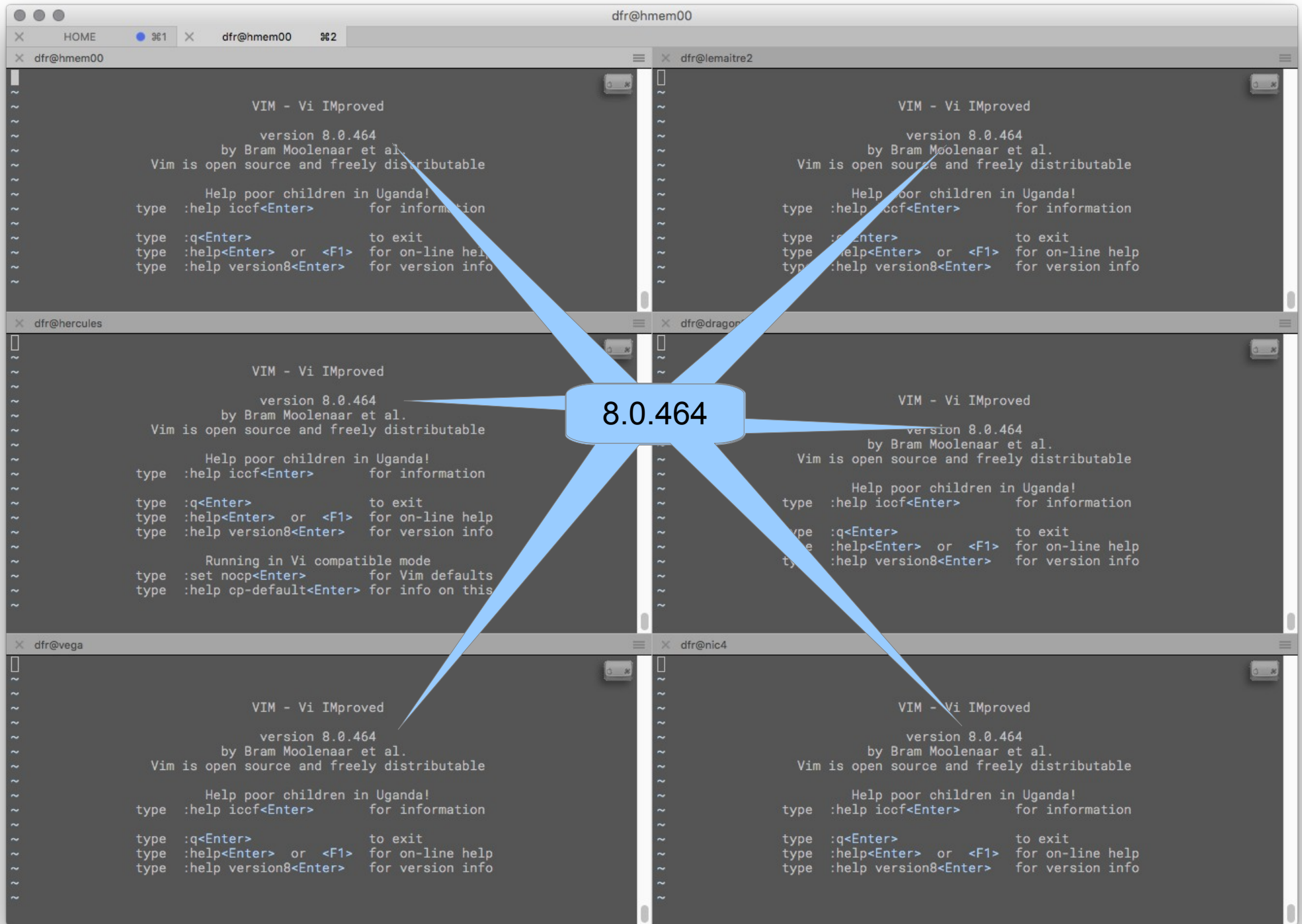
I work on Hmem

I compile a newer version from sources

I install it in my CECI home directory

I setup the PATH accordingly (on all clusters)

# Now all clusters have the same Vim...



# A catch though...

By default, compilers will tune the binary for the CPU of the machine they run on.

		run on					
		Hmem	Lemaitre2	Dragon1	Hercules	Vega	NIC4
compile on	Hmem	ok	sub-opt	sub-opt	sub-opt	sub-opt	sub-opt
	Lemaitre2	crash	ok	sub-opt	sub-opt	sub-opt	sub-opt
	Dragon1	crash	crash	ok	ok	sub-opt	ok
	Hercules	crash	crash	ok	ok	sub-opt	ok
	Vega	crash	crash	crash	crash	ok	crash
	NIC4	crash	crash	ok	ok	sub-opt	ok

# Mitigation:

## **GCC:**

- march=core2 to build binaries running everywhere
- mtune=[westmere|sandybridge|bdver1]  
to tune for the cluster you use the most

## **Intel** (multiple code paths):

- xSSE2 to build binaries running everywhere
- axSSE4.2,AVX,CORE-AVX2  
to add additional binary objects optimized for each cluster

# Mitigation:

## GCC: “Function multi-versioning”

Toggle line numbers

```
1 __attribute__ ((target ("default")))
2 int foo ()
3 {
4     // The default version of foo.
5     return 0;
6 }
7
8 __attribute__ ((target ("sse4.2")))
9 int foo ()
10 {
11     // foo version for SSE4.2
12     return 1;
13 }
14
15 __attribute__ ((target ("arch=atom")))
16 int foo ()
17 {
18     // foo version for the Intel ATOM processor
19     return 2;
20 }
21
22 __attribute__ ((target ("arch=amdfam10")))
23 int foo ()
24 {
25     // foo version for the AMD Family 0x10 processors.
26     return 3;
27 }
28 int main ()
29 {
30     int (*p)() = &foo;
31     assert ((*p) () == foo ());
32     return 0;
33 }
```

In the above example, 4 versions of function foo are created. The first version of foo with the target attribute "default" is the default version. This version gets executed when no other target specific version qualifies for execution on a particular platform. A new version of foo is created by using the same function signature but with a different target string. Function foo is called or a pointer to it is taken just like a regular function. With the new support, GCC takes care of doing the dispatching to call the right version at runtime.

# Mitigation:

## Intel compiler: “Manual processor dispatch”

### Example

```
#include <stdio.h>
// need to create specific function versions for the following processors:
__declspec(cpu_dispatch(core_2nd_gen_avx, core_i7_sse4_2, core_2_duo_ssse3, generic ))
void dispatch_func() {};      // stub that will call the appropriate specific function
version

__declspec(cpu_specific(core_2nd_gen_avx))
void dispatch_func() {
    printf("\nCode for 2nd generation Intel Core processors with support for AVX goes
here\n");
}

__declspec(cpu_specific(core_i7_sse4_2))
void dispatch_func() {
    printf("\nCode for Intel Core processors with support for SSE4.2 goes here\n");
}

__declspec(cpu_specific(core_2_duo_ssse3))
void dispatch_func() {
    printf("\nCode for Intel Core 2 Duo processors with support for SSSE3 goes here\n");
}

__declspec(cpu_specific(generic))
void dispatch_func() {
    printf("\nCode for non-Intel processors and generic Intel processors goes here\n");
}

int main() {
    dispatch_func();
    printf("Return from dispatch_func\n");
    return 0;
}
```



# Four spaces

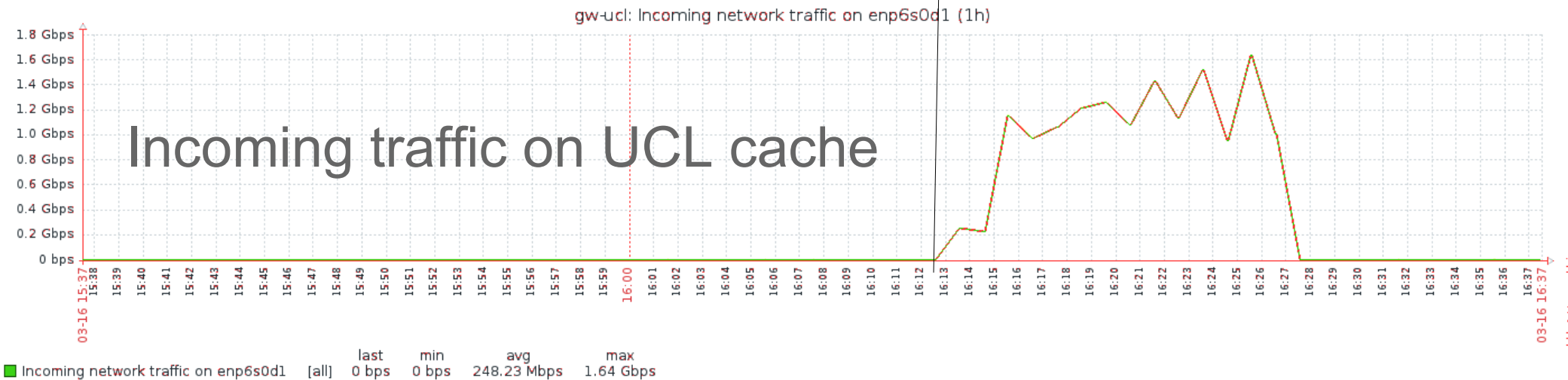
- /CECI/home
  - Quota 100GB/User
  - Daily snapshots
- /CECI/proj
  - Upon request
  - Quota and duration based on request
- /CECI/trsf
  - Quota per user 100GB soft 10TB hard
  - Automatic purge of files older than 6 months
- /CECI/soft
  - Common software + modules



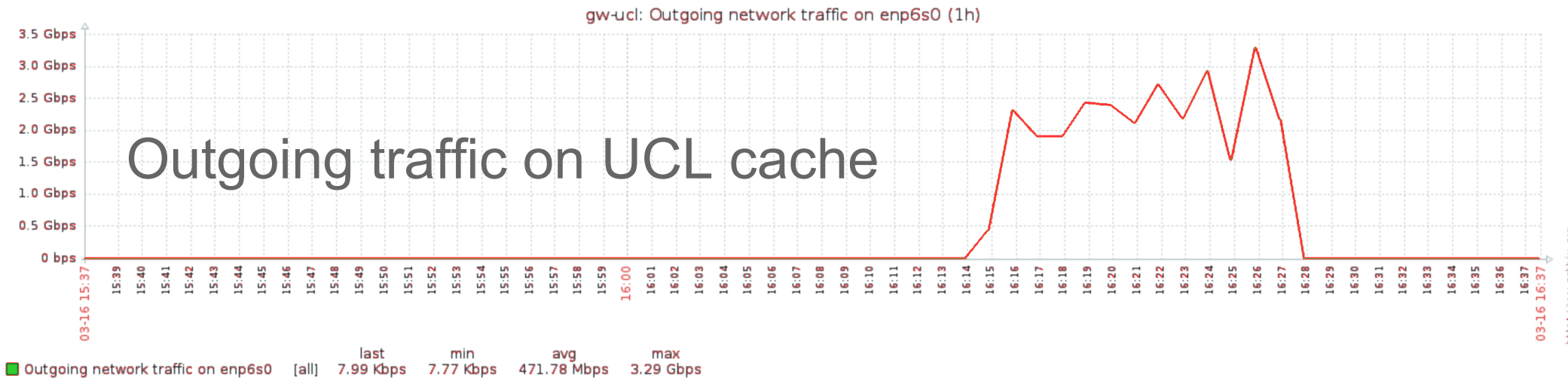
# Copy of 100GB file from the scratch of Lemaitre2 to scratch of NIC4

16:12:31 command issued

## Incoming traffic on UCL cache



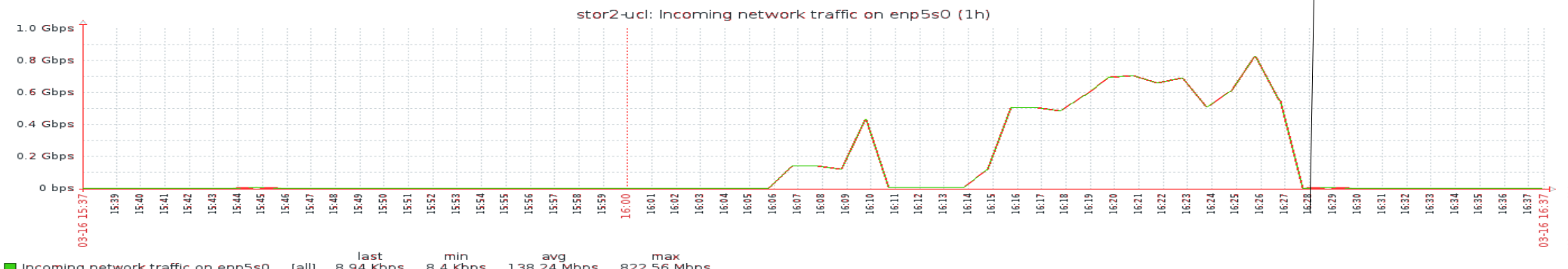
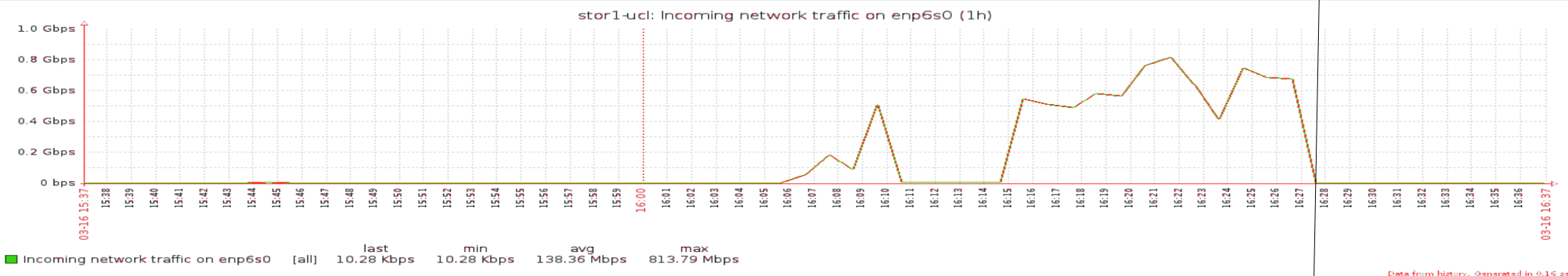
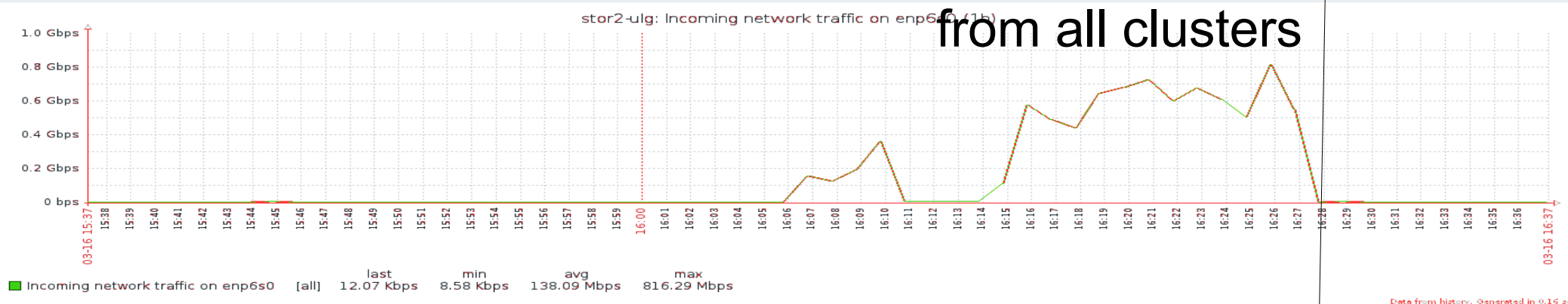
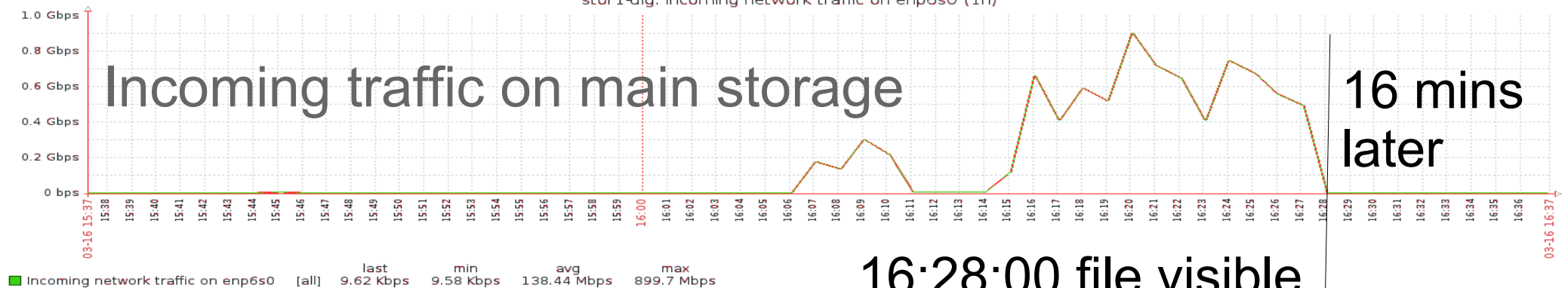
## Outgoing traffic on UCL cache



# Incoming traffic on main storage

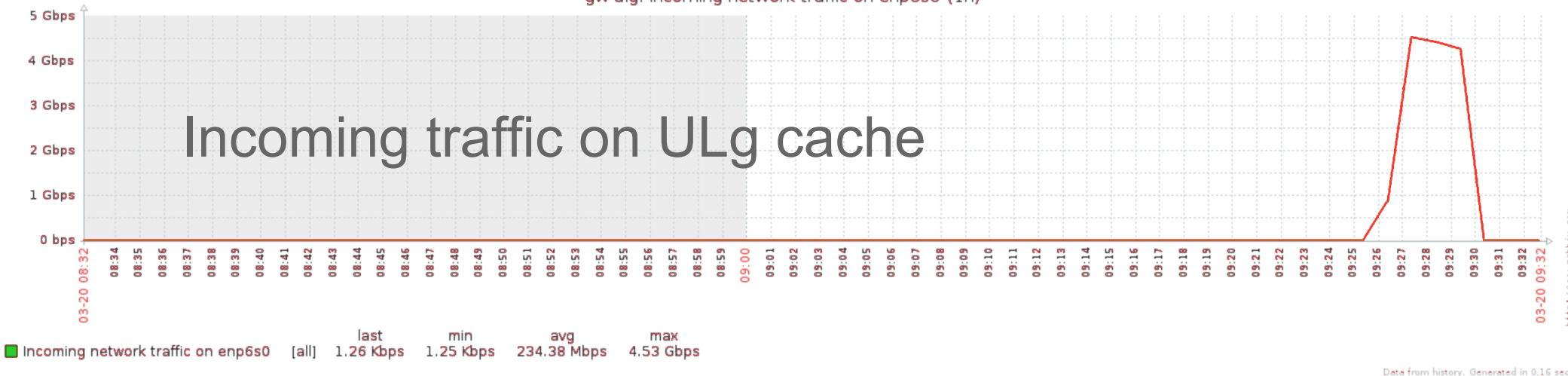
16 mins later

16:28:00 file visible from all clusters



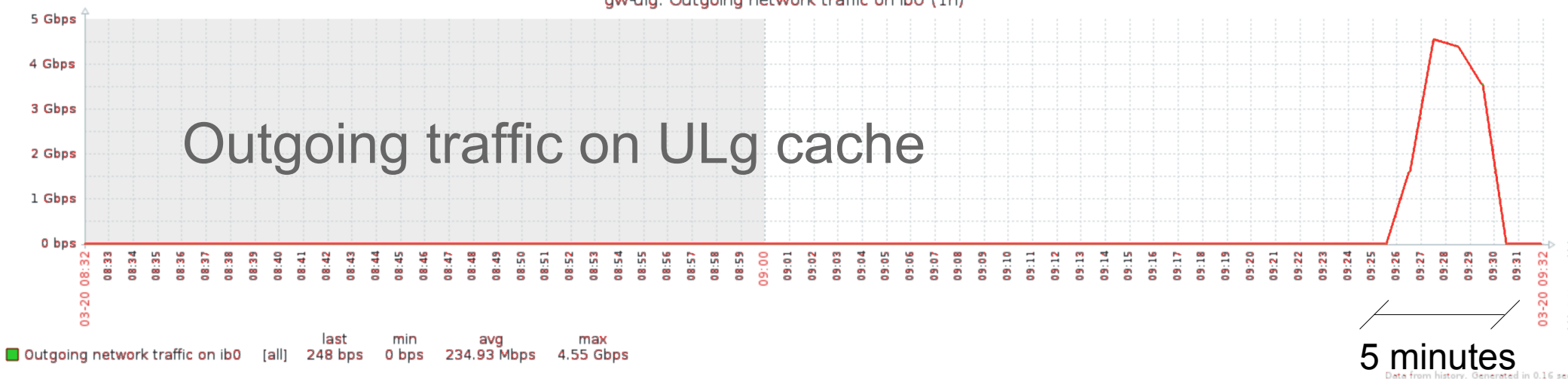
gw-ulg: Incoming network traffic on enp6s0 (1h)

## Incoming traffic on ULg cache



gw-ulg: Outgoing network traffic on ib0 (1h)

## Outgoing traffic on ULg cache



Total 21 minutes  
vs.

```
dfr@nic4 361
dfr@nic4:~$ scp dfr@lemaitre2:$GLOBALSCRATCH/bigfile.dat $
GLOBALSCRATCH
bigfile.dat 0% 79MB 6.3MB/s 4:25:25 ETA
```

UNamur (Université de Namur) su

Using the common filesystem — CÉCI

Docs

C.E.C.I

Search docs

QUICK START – FIRST STEPS

Creating an account

Connecting to the clusters

Copying files

Editing files

Slurm Quick Start Tutorial

ACCESSING THE CLUSTERS

Troubleshooting and frequent mistakes

MANAGING FILES

Transferring files to and from the clusters

Using the common filesystem

Home

Trfs

Proj

Soft

CÉCI v: latest

Docs » Using the common filesystem

# Using the common filesystem

All CÉCI clusters are connected to a central storage system that is visible to all compute nodes of all clusters. This system runs on a fast, dedicated, network. It will become the home of the users in the near future, but in this first phase, it is set up as an additional home besides the default, cluster-specific, home.

This storage system is installed at two CÉCI locations and data are replicated synchronously on both locations to ensure data safety and a certain level of high availability. Moreover, on each site, a local cache is setup to mask the latencies of the network and make sure the user experience is as smooth as possible. Those caches are replicated asynchronously with the central storage, meaning that files that are written there will appear after some delay on the other clusters. It also means that if you modify the same file from two different clusters, the result is undefined.

Warning

Do not write to the same file from two different clusters at the same time. This would corrupt the file.

The storage is split into four distinct directories:

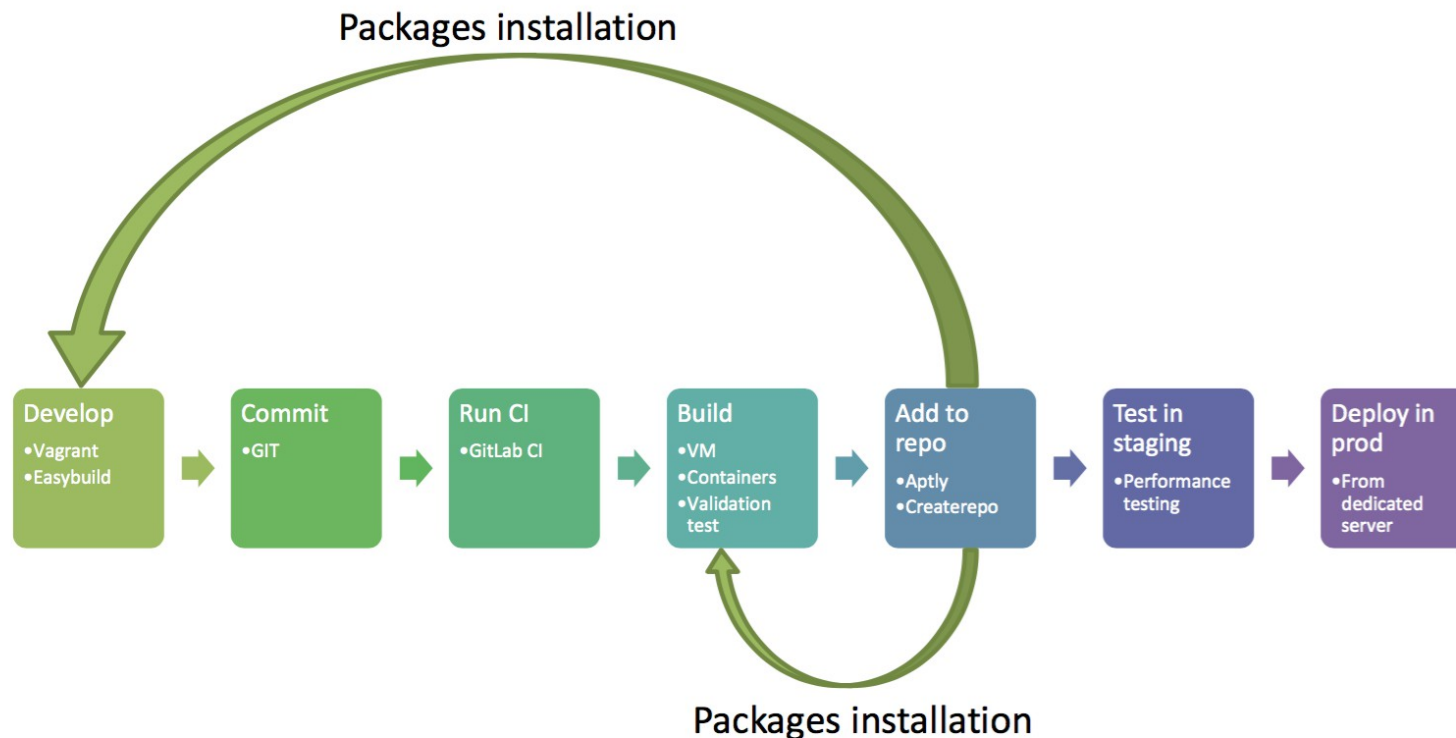




The new CÉCI  
common storage  
Long-term benefits

# Common set of software/modules

## Workflow



# Slurm Federation

## New Capabilities

- **Job Migration**
  - Pending jobs automatically migrated to less busy clusters
- **Fault Tolerance**
  - Participating clusters will take over work of a failed cluster
- **Cross-cluster Job Dependencies**
- **Unified Views**
- **Easy Administration**
  - Add/remove clusters to/from the federation with simple configuration change, no extra information required in database



The new CÉCI  
common storage

Wrap-up & future plans



# Four spaces

- /CECI/home
  - Quota 100GB/User
  - Daily snapshots
- /CECI/proj
  - Upon request
  - Quota and duration based on request
- /CECI/trsf
  - Quota per user 100GB soft 10TB hard
  - Automatic purge of files older than 6 months
- /CECI/soft
  - Common software + modules

# What's next...

- Fine-tune configuration

Your feedback: your local CÉCI system administrator

- Setup procedure to request group space
- Build common software installation

All further information through  
the CÉCI users mailing list



**Consortium des Équipements de Calcul Intensif**  
Funded by F.R.S.-FNRS

[www.cec-hpc.be](http://www.cec-hpc.be)



Try the new CÉCI  
common storage!