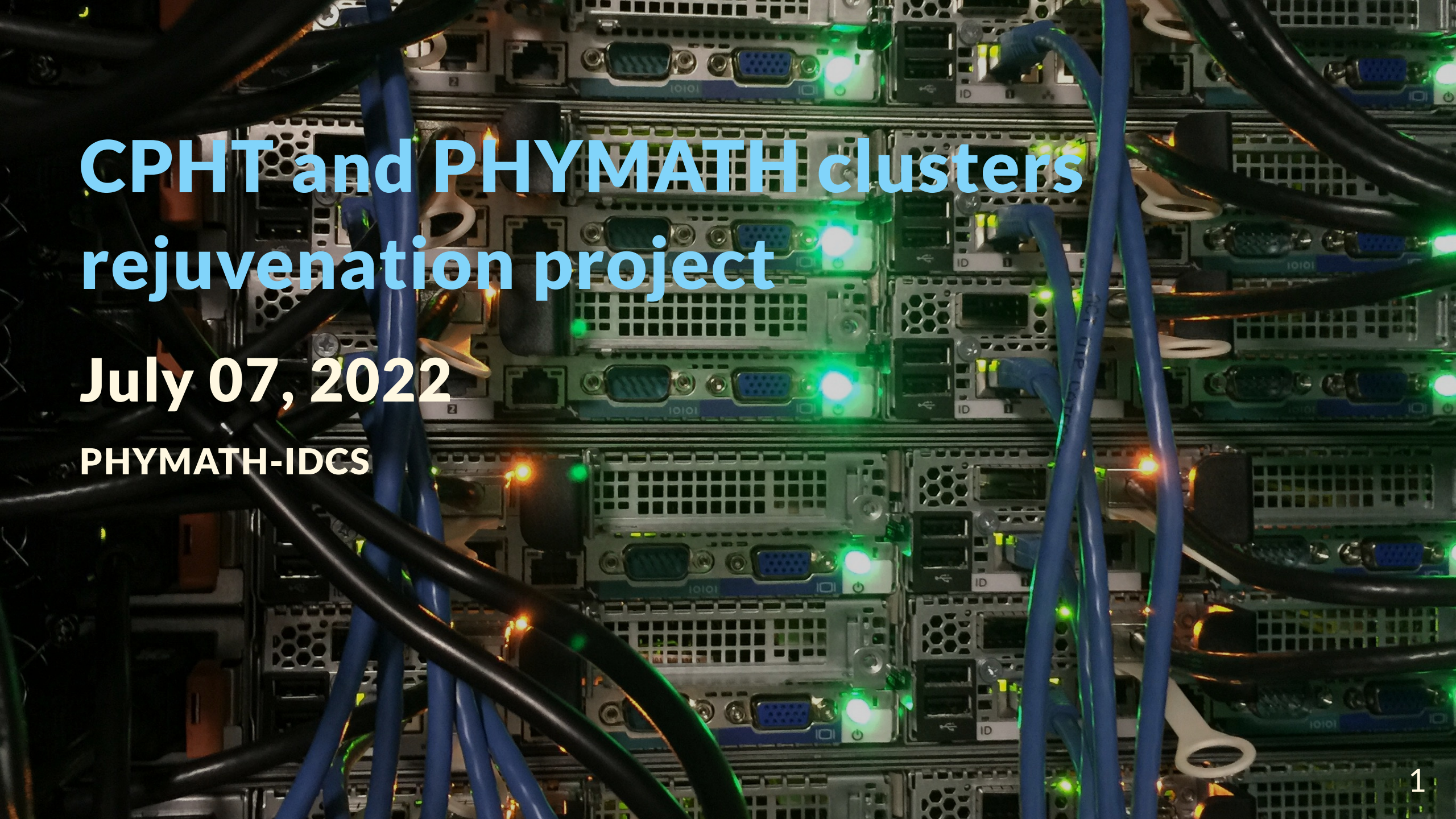


CPHT and PHYMATH clusters rejuvenation project

July 07, 2022

PHYMATH-IDCS

- Operations started with *NeoTeckno* on mid-March
- The production deadline has passed (end of May)
- Some examples of difficulty we encountered :
  - creating an image compatible with all nodes ;
  - interconnection of cluster (BMC/IDRAC) administration and provisioning ethernet networks to those of Cholesky
  - BMC/IDRAC network configuration of nodes (manual operation for some nodes)
  - Identification of problems encountered during the deployment of the node image (eg. failed hard disk)
- 4-day service consumed by *NeoTeckno*

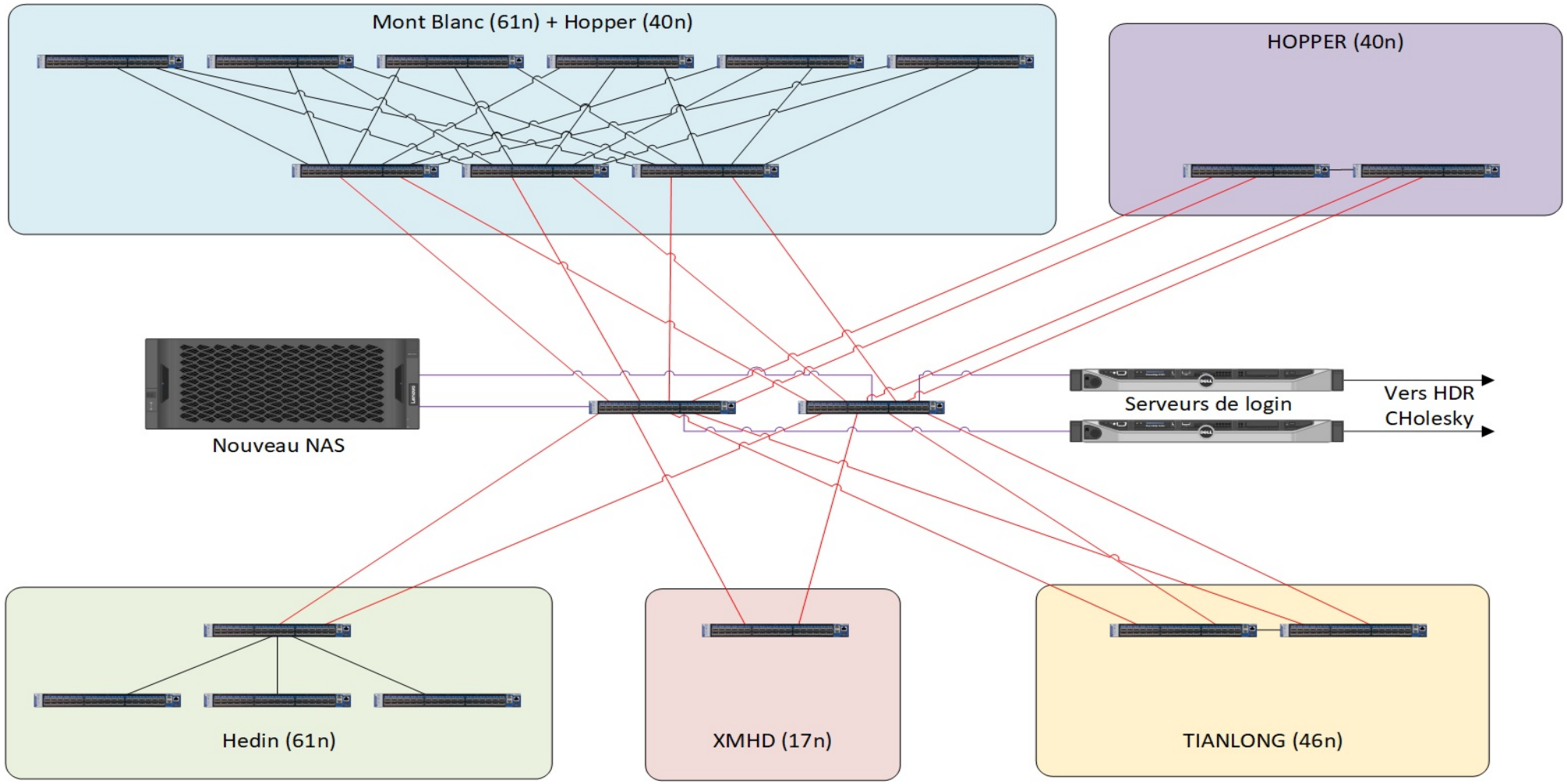
2. Operations performed for login nodes

- Installation and configuration :
  - 2 new login/master bare metal nodes :
    - `cholesky-login01.mesocentre.idcs.polytechnique.fr`
    - `cholesky-login02.mesocentre.idcs.polytechnique.fr`
    - available soon (maybe on July) with the name `cholesky-login` - DNS load-balancing between this 2 login nodes - replacing the current login frontend (VM)

2. Operations performed for file storage

- Installation and configuration
  - 1 NAS file server (100 TB raw, 88 TB usable for DATA)
    - access from **new login** and compute nodes **only on IB40 network (Hedin, Montblanc, Tianlong, Hopper nodes)**
    - export NFS data directory
    - NFS mount access from nodes : `/mnt/nas/workdir/$USER`
    - user quota : **200 GB** (maybe increased on requested)

Operations performed for IB40 network



Operation performed for nodes

- OS image creation (nearly identical as Cholesky nodes)
- image deployed actually on **hedin**, **tianlong** and **hopper** nodes (montblanc in progress)
- some nodes have hardware problems (hard drive, IB40 interface) : needs to be replaced
- access to HOME (and WORKDIR) user directories on BeeGFS cholesky file storage **only with ethernet network (not IB)**
- use WORKDIR on NAS (IB40) : `/mnt/nas/workdir/$USER`

Today

- **hedin**, **tianlong**, **hopper** nodes** are ready for calculation :
  - SLURM job submission from new login nodes ;
  - SLURM partitions are the same before migration
    - accessible only from identified SLURM accounts

```
hedin             up 1-00:00:00     10  down* hedin[008,017,019,033,035-037,042-043,048]
hedin             up 1-00:00:00     25   idle hedin[001-007,009,011,013,015,021,023,025,027,029,031,039,041,046,050,052,054,056,058]
hedin_memlarge    up 1-00:00:00      6  down* hedin[012,014,032,038,040,049]
hedin_memlarge    up 1-00:00:00     19   idle hedin[010,016,018,020,022,024,026,028,030,034,044-045,047,051,053,055,057,059-060]
tianlong          up 1-00:00:00      4  down* tianlong[009,017-019]
tianlong          up 1-00:00:00     28   idle tianlong[001-008,010-016,020-032]
tianlong_cpu24    up 1-00:00:00      1  down* tianlong034
tianlong_cpu24    up 1-00:00:00     11   idle tianlong[033,035-044]
```

** Hardware problems on nodes

Process to calculate

- Procedure is the same as submetting on Cholesky
- 1 GITLAB project (gitlab.idcs.polytechnique.fr) = 1 SLURM account
  - GITLAB hedin project = hedin SLURM account
  - GITLAB tianlong project = tianlong SLURM account
- Only project coordinator(s) can add or remove users :
  - M. Ferrero for tianlong account
  - S. Biermann for hedin account

Software Modules

- **Keeping an unique software module tree** for Cholesky nodes and CPHT and Hopper nodes
- **Intel MPI** (with **libfabric 1.15**) and **OpenMPI 4.x** are also compatible with Intel Q-Logic QDR40 transport layer : **needs to add flags to mpirun binary**
- Use our **mpilaunch** wrapper instead of mpirun that choose dynamically the good transport layer (HDR IB Mellanox for Cholesky nodes or Intel Q-Logic IB for CPHT and Hopper nodes)
- Modules depending on OpenMPI : **needs to be re-compiled (in progress)**

Software Modules (conclusion)

- For OpenMPI use `openmpi/4.1.4` :

```
$ module load gcc/10.2.0
$ module load openmpi/4.1.4
$ mpilaunch
```

- For Intel MPI use `intel_mpi/2019.9` :

```
$ module load intel_compiler/19.1.3.304
$ module load intel_mpi/2019.9
$ mpilaunch
```

TODO - IN PROGRESS

- Some *beta-testers* need to test **Hedin**, **Tianlong** and **Hopper** partitions to validate before going into production :
  - somes IB bandwith tests with OpenMPI have be done and the results seem consistent
- Integration in progress of **Montblanc** nodes
- User documentation : https://meso-ipp.gitlab.labos.polytechnique.fr/user_doc/