Consortium des Équipements de Calcul Intensif

CÉCI News by the Sysadmins

David Colignon, **Ariel Lozano** [ULiège]
Juan Cabrera, Frédéric Wautelet [UNamur]
Sébastien Skozlowskij [UMons]
Raphaël Leplae, Michaël Waumans [ULB]
Damien François, Bernard Van Renterghem, Olivier Mattelaer, Thomas Keutgen [UCL]

25th April 2019 - CÉCI Scientific Day @ ULB

Outline

- Cluster Upgrades
- CÉCI Slurm Federation
- Common software modules in the federated clusters

Clusters upgrade: Dragon2 @ UMons

- Hard:
  - **17 nodes** with 2 x 16 core Intel Xeon Gold 6142@2.6 Ghz
    - **15 nodes** with **192GB RAM** (6GB/core)
    - **2 nodes** with **384GB RAM** (12GB/core)
  - **2 nodes** with 2 x 12 core Intel Xeon Gold 6126@2.6 Ghz 192GB RAM and **2 GPU Nvidia Tesla V100** 16GB VRAM
  - Total of **592 cpu cores** and **4 Nvidia Tesla V100**
  - Skylake Xeons with AVX, AVX2, AVX-512 extensions
  - Local scracth of 3.3 TB
  - 10 Gb/s ethernet between nodes
- Soft:
  - **Second** cluster on the CÉCI **Slurm Federation**
  - Intel Parallel Studio XE 2018 and 2019
  - Common modules setup

Already in production!

Clusters upgrade: Hercules2 @ UNamur

- Hard:
  - **30 nodes** with 32 core AMD Epyc@2.0 GHz (Naples family)
    - **24 *skinny* nodes** single-socket **256GB RAM** (8GB/core)
    - **4 *medium* nodes** single-socket **512GB RAM** (16GB/core)
    - **2 *fat* nodes** dual-socket **2TB RAM** (32GB/core)
    - Local scratch from 5TB
  - **32 nodes** with 16 core Intel Sandy Bridge@2.0 GHz from Hercules will be added
  - Total of **1536 cpu cores** (1024 “AMD Epyc” + 512 “Intel Sandy Bridge”)
  - 10 Gb/s ethernet between nodes
- Soft:
  - **Third** cluster on the CÉCI **Slurm Federation**
  - Intel Parallel Studio XE 2018 and 2019
  - Common modules setup

The new High Memory CÉCI cluster

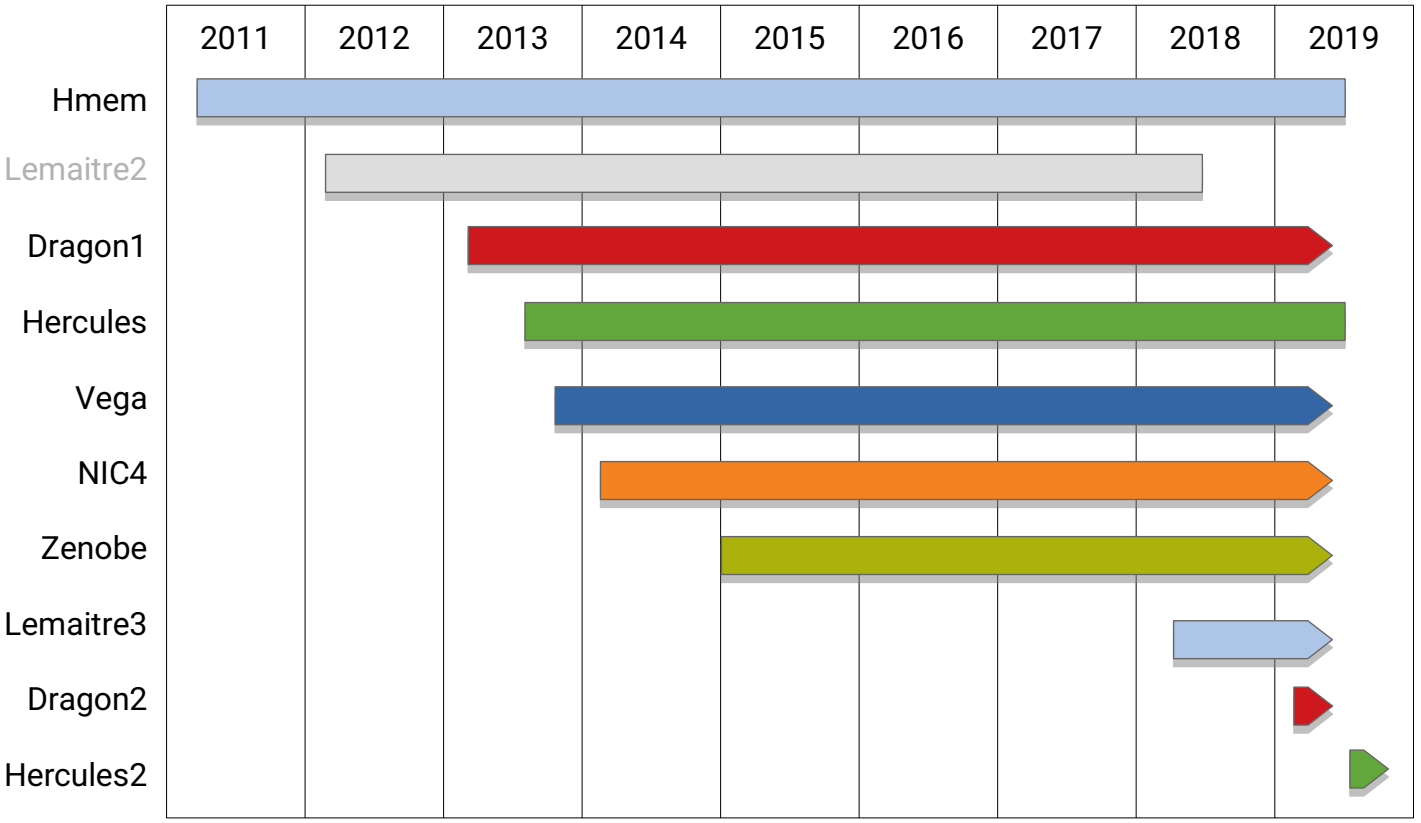
Clusters upgrade: old clusters

- **Dragon1** will stay on production as long as it keeps being used
  - an eventual OS and setup upgrade might be planned

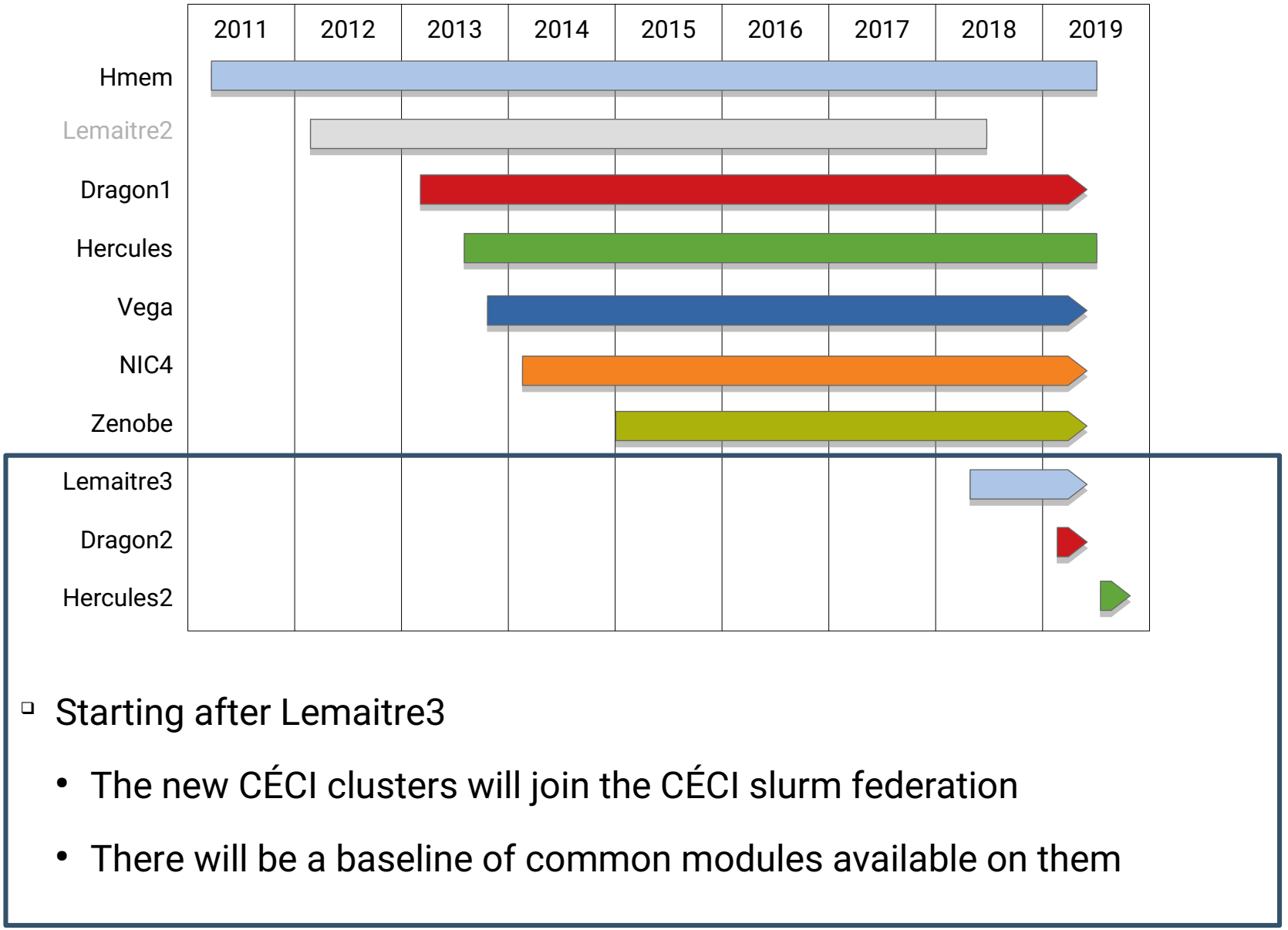
- **Hercules** to **Hercules 2** transition
  - In the following weeks all the **new AMD Epyc** nodes will be added to Hercules
  - The installation for **Hercules 2** is planned for **July 2019**
  - **Hercules 2** will replace Hercules

- **Hmem** will in principle follow Dragon1 path
  - but you should consider moving your high mem workflows to **Hercules 2** when available

Clusters upgrade: timeline



Clusters upgrade: timeline



- Starting after Lemaitre3
  - The new CÉCI clusters will join the CÉCI slurm federation
  - There will be a baseline of common modules available on them

CÉCI Slurm Federation

- Lemaitre3 and Dragon2 in the Federation at the moment

```
[alozano@lemaitre3 ~]$ sacctmgr show federation
Federation    Cluster ID             Features     FedState
---------- ---------- -- -------------------- ------------
      ceci    dragon2  2                           ACTIVE
      ceci  lemaitre3  1    globalscratch,opa      ACTIVE
```

- The main feature is the ability to submit a job to all the federated clusters at once

CÉCI Slurm Federation

- Using `sbatch` as usual the job will be submitted **only to the current cluster** (as it always worked)

```
[alozano@lemaitre3 myjob]$ sbatch myjob.sh
Submitted batch job 67814013 on cluster lemaitre3
[alozano@lemaitre3 myjob]$ squeue -u alozano
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          67814013     batch    myjob  alozano  R       0:04      1 lm3-w080
```

CÉCI Slurm Federation

- Using `sbatch` as usual the job will be submitted **only to the current cluster** (as it always worked)

```
[alozano@lemaitre3 myjob]$ sbatch myjob.sh
Submitted batch job 67814013 on cluster lemaitre3
[alozano@lemaitre3 myjob]$ squeue -u alozano
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          67814013     batch    myjob  alozano  R       0:04      1 lm3-w080
```

- But now to monitor them `squeue` will always show **all my jobs** running in the federated clusters from any of them

```
[alozano@dragon2 ~]$ squeue -u alozano
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          67814013     batch    myjob  alozano  R       0:52      1 lm3-w080
```

CÉCI Slurm Federation

- To submit a federated job the ideal is to work from the common storage as this area is already mounted on all clusters and nodes

```
[alozano@dragon2 ~]$ cd $CECIHOME/myjob
[alozano@dragon2 myjob]$ ls
myinput1.dat  myinput2.dat  myjob.sh
```

- Then we can safely copy data from/to `$SLURM_SUBMIT_DIR` to the temporary work directories on the clusters

```
[alozano@dragon2 myjob]$ cat myjob.sh
#!/bin/bash
#SBATCH --job-name=myjob
#SBATCH --time=00:10:00
#SBATCH --ntasks=1
#SBATCH --mem-per-cpu=100

mkdir -p "$LOCALSCRATCH/$SLURM_JOB_ID"
cp -r "$SLURM_SUBMIT_DIR/{myinput1.dat,myinput2.dat}" "$LOCALSCRATCH/$SLURM_JOB_ID"

...
```

CÉCI Slurm Federation

- If we want the job submitted to any cluster in the federation we **must specify that** to `sbatch` with the `--cluster=all` argument

```
[alozano@dragon2 myjob]$ sbatch --clusters=all myjob.sh
Submitted batch job 134236552 on cluster dragon2
[alozano@dragon2 myjob]$ sbatch --clusters=all myjob.sh
Submitted batch job 134236554 on cluster dragon2
[alozano@dragon2 myjob]$ squeue -u alozano
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
         134236552     batch    myjob  alozano  R       0:04      1 lm3-w080
         134236554     batch    myjob  alozano  R       0:04      1 drg2-w002
```

- The same as for the local jobs, all the federated jobs will be visible from the other clusters

```
[alozano@lemaitre3 myjob]$ squeue -u alozano
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
         134236552     batch    myjob  alozano  R       0:12      1 lm3-w080
         134236554     batch    myjob  alozano  R       0:12      1 drg2-w002
```

CÉCI Slurm Federation

- If we want the job submitted to any cluster in the federation we **must specify that** to `sbatch` with the `--cluster=all` argument

```
[alozano@dragon2 myjob]$ sbatch --clusters=all myjob.sh
Submitted batch job 134236552 on cluster dragon2
[alozano@dragon2 myjob]$ sbatch --clusters=all myjob.sh
Submitted batch job 134236554 on cluster dragon2
[alozano@dragon2 myjob]$ squeue -u alozano
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
         134236552     batch    myjob  alozano  R       0:04      1 lm3-w080
         134236554     batch    myjob  alozano  R       0:04      1 drg2-w002
```

Tip: if I want to see the jobs running or submitted only to a specific cluster you can add an argument as e.g.

`squeue --cluster=lemaitre3 -u mylogin`

```
[alozano@lemaitre3 myjob]$ squeue -u alozano
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
         134236552     batch    myjob  alozano  R       0:12      1 lm3-w080
         134236554     batch    myjob  alozano  R       0:12      1 drg2-w002
```

CÉCI Slurm Federation

- How it works...

1. When a federated job is submitted, copies of it are submitted to each eligible cluster
2. Each cluster will then independently attempt to schedule the job
3. When a cluster determines the job can be allocated it communicates with the origin cluster to verify that no other cluster is attempting to allocate resources at the same time
4. If that is the case the job is scheduled and the origin cluster is notified the job has started

```
[alozano@dragon2 myjob]$ sbatch --clusters=all myjob.sh
Submitted batch job 134236574 on cluster dragon2
[alozano@dragon2 myjob]$ squeue -u alozano
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
         134236574     batch    myjob  alozano PD       0:00      1 (None)
```

CÉCI Slurm Federation

- The setup is in progress (pending to sanitize env vars, assure common modules on both clusters)

- In the following weeks it will be finished and detailed documentation will be made available

Common software modules

- Setup of the modules on the new clusters

```
[alozano@lemaitre3 ~]$ module av
-------------------------------------------- Meta Modules --------------------------------------------
   dot     releases/elic-2017b        releases/2016b (S)        releases/2018a (S)    tis/2018.01 (S,L)
   null    releases/2016a      (S)    releases/2017b (S,L,D)    releases/2018b (S)    use.own

---------------------------- TIS: Toolchain Independent Software (2018.01) ----------------------------
   EasyBuild/3.5.1    MCR/R2014a    MCR/R2017a                          freesurfer/6.0.0
   Java/1.8.0_31      MCR/R2014b    MCR/R2017b                          gurobi/gurobi800
   Java/1.8.0_92      MCR/R2015a    MCR/R2018a                   (D)    julia/0.6.3
   Java/1.8.0_121     MCR/R2015b    NCBI-BLAST-database/20170306        julia/1.0.0      (D)
   MCR/R2013a         MCR/R2016a    crystal/17-v1.0.1                   xpress/xp850
   MCR/R2013b         MCR/R2016b    crystal/17-v1.0.2            (D)

------------------------------------------ Releases (2017b)  ------------------------------------------
   ABINIT/8.4.4-intel-2017b                      Python/2.7.14-foss-2017b
   ABINIT/8.10.2-intel-2017b              (D)    Python/2.7.14-GCCcore-6.4.0-bare
   ANTLR/2.7.7-intel-2017b                       Python/3.6.3-foss-2017b
   Boost/1.65.1-foss-2017b                       Python/3.6.3-intel-2017b                      (D)
   Boost/1.66.0-intel-2017b               (D)    Qhull/2015.2-foss-2017b
   CDO/1.9.2-intel-2017b                         Qt/4.8.7-foss-2017b
   CGAL/4.11-foss-2017b-Python-2.7.14            R/3.4.3-foss-2017b-X11-20171023
   CP2K/5.1-intel-2017b                          Ruby/2.5.0-intel-2017b
   Doxygen/1.8.13-GCCcore-6.4.0                  SCOTCH/6.0.4-foss-2017b
   Eigen/3.3.4                                   SCOTCH/6.0.4-intel-2017b                      (D)
   FFTW/3.3.6-gompi-2017b                        SQLite/3.20.1-GCCcore-6.4.0
   FFTW/3.3.6-intel-2017b                 (D)    SWIG/3.0.12-foss-2017b-Python-2.7.14
   FLUENT/14.0                                   SWIG/3.0.12-foss-2017b-Python-3.6.3
   FLUENT/18.2                            (D)    SWIG/3.0.12-intel-2017b-Python-3.6.3          (D)
   GCC/6.4.0-2.28                                ScaLAPACK/2.0.2-gompi-2017b-OpenBLAS-0.2.20
   ...
```

Common software modules: releases

- Modules are organized around the concept of **toolchain** and **releases** coming from the EasyBuild framework, the two main toolchains provided are `foss` (free open source software) and `intel`
- A **toolchain** is a collection of **compiler and numerical libraries** often used together and **known to interoperate** smoothly
- The `foss` toolchain comprises a **bundle** of
  - GNU Compiler Collection (gcc, g++, gfortran)
  - OpenMPI library
  - OpenBLAS + ScaLAPACK
  - FFTW library
- Two **releases** per year are provided for each **toolchain bundle** which corresponds to a set of specific versions of its components, e.g. `foss/2017b` comprises
  - GCC/6.4.0
  - OpenMPI/2.1.1
  - OpenBLAS/0.2.20, ScaLAPACK/2.0.2
  - FFTW/3.3.6

Common software modules

- Setup of the modules on the new clusters

```
[alozano@lemaitre3 ~]$ module av
------------------------------------------ Meta Modules ------------------------------------------
   dot     releases/elic-2017b        releases/2016b (S)        releases/2018a (S)    tis/2018.01 (S,L)
   null    releases/2016a      (S)    releases/2017b (S,L,D)    releases/2018b (S)    use.own

--------------------------- TIS: Toolchain Independent Software (2018.01) ---------------------------
   EasyBuild/3.5.
   Java/1.8.0_31
   Java/1.8.0_92
   Java/1.8.0_121                                                                              )
   MCR/R2013a
   MCR/R2013b

------------------------------------------ Releases (2017b) ------------------------------------------
   ABINIT/8.4.4-intel-2017b                      Python/2.7.14-foss-2017b
   ABINIT/8.10.2-intel-2017b             (D)     Python/2.7.14-GCCcore-6.4.0-bare
   ANTLR/2.7.7-intel-2017b                       Python/3.6.3-foss-2017b
   Boost/1.65.1-foss-2017b                       Python/3.6.3-intel-2017b                          (D)
   Boost/1.66.0-intel-2017b              (D)     Qhull/2015.2-foss-2017b
   CDO/1.9.2-intel-2017b                         Qt/4.8.7-foss-2017b
   CGAL/4.11-foss-2017b-Python-2.7.14            R/3.4.3-foss-2017b-X11-20171023
   CP2K/5.1-intel-2017b                          Ruby/2.5.0-intel-2017b
   Doxygen/1.8.13-GCCcore-6.4.0                  SCOTCH/6.0.4-foss-2017b
   Eigen/3.3.4                                   SCOTCH/6.0.4-intel-2017b                          (D)
   FFTW/3.3.6-gompi-2017b                        SQLite/3.20.1-GCCcore-6.4.0
   FFTW/3.3.6-intel-2017b                (D)     SWIG/3.0.12-foss-2017b-Python-2.7.14
   FLUENT/14.0                                   SWIG/3.0.12-foss-2017b-Python-3.6.3
   FLUENT/18.2                           (D)     SWIG/3.0.12-intel-2017b-Python-3.6.3              (D)
   GCC/6.4.0-2.28                                ScaLAPACK/2.0.2-gompi-2017b-OpenBLAS-0.2.20
   ...
```

There is a *Meta Modules* section with modules to handle the different releases

One of them will be always preloaded as default **2017b** in this case

Common software modules

- Setup of the modules on the new clusters

```
[alozano@lemaitre3 ~]$ module av
---------------------------------------------- Meta Modules ----------------------------------------------
   dot     releases/elic-2017b        releases/2016b (S)        releases/2018a (S)    tis/2018.01 (S,L)
   null    releases/2016a      (S)    releases/2017b (S,L,D)    releases/2018b (S)    use.own

----------------------------- TIS: Toolchain Independent Software (2018.01) -----------------------------
   EasyBuild/3.5.1    MCR/R2014a    MCR/R2017a                          freesurfer/6.0.0
   Java/1.8.0_31      MCR/R2014b    MCR/R2017b                          gurobi/gurobi800
   Java/1.8.0_92      MCR/R2015a    MCR/R2018a                   (D)    julia/0.6.3
   Java/1.8
   MCR/R201
   MCR/R201

------------------------------------------- Releases (2017b)  -------------------------------------------
   ABINIT/8.4.4-intel-2017b                      Python/2.7.14-foss-2017b
   ABINIT/8.10.2-intel-2017b              (D)    Python/2.7.14-GCCcore-6.4.0-bare
   ANTLR/2.7.7-intel-2017b                       Python/3.6.3-foss-2017b
   Boost/1.65.1-foss-2017b                       Python/3.6.3-intel-2017b                      (D)
   Boost/1.66.0-intel-2017b               (D)    Qhull/2015.2-foss-2017b
   CDO/1.9.2-intel-2017b                         Qt/4.8.7-foss-2017b
   CGAL/4.11-foss-2017b-Python-2.7.14            R/3.4.3-foss-2017b-X11-20171023
   CP2K/5.1-intel-2017b                          Ruby/2.5.0-intel-2017b
   Doxygen/1.8.13-GCCcore-6.4.0                  SCOTCH/6.0.4-foss-2017b
   Eigen/3.3.4                                   SCOTCH/6.0.4-intel-2017b                      (D)
   FFTW/3.3.6-gompi-2017b                        SQLite/3.20.1-GCCcore-6.4.0
   FFTW/3.3.6-intel-2017b                 (D)    SWIG/3.0.12-foss-2017b-Python-2.7.14
   FLUENT/14.0                                   SWIG/3.0.12-foss-2017b-Python-3.6.3
   FLUENT/18.2                            (D)    SWIG/3.0.12-intel-2017b-Python-3.6.3          (D)
   GCC/6.4.0-2.28                                ScaLAPACK/2.0.2-gompi-2017b-OpenBLAS-0.2.20
   ...
```

When a given release is loaded the *Releases* section shows the software available with its corresponding toolchains

Common software modules

- Setup of the modules on the new clusters

```
[alozano@lemaitre3 ~]$ module av
------------------------------------------ Meta Modules -------------------------------------------
   dot     releases/elic-2017b        releases/2016b (S)        releases/2018a (S)    tis/2018.01 (S,L)
   null    releases/2016a      (S)    releases/2017b (S,L,D)    releases/2018b (S)    use.own

---------------------------- TIS: Toolchain Independent Software (2018.01) ----------------------------
   EasyBuild/3.5.1    MCR/R2014a    MCR/R2017a                          freesurfer/6.0.0
   Java/1.8.0_31      MCR/R2014b    MCR/R2017b                          gurobi/gurobi800
   Java/1.8.0_92      MCR/R2015a    MCR/R2018a                   (D)    julia/0.6.3
   Java/1.8.0_121     MCR/R2015b    NCBI-BLAST-database/20170306        julia/1.0.0      (D)
   MCR/R2013a         MCR/R2016a    crystal/17-v1.0.1                   xpress/xp850
   MCR/R2013b         MCR/R2016b    crystal/17-v1.0.2            (D)

-----
   AB
   AB
   AN
   Bo                                                                                        (D)
   Boost/1.66.0-intel-2017b                (D)    Qhull/2015.2-foss-2017b
   CDO/1.9.2-intel-2017b                          Qt/4.8.7-foss-2017b
   CGAL/4.11-foss-2017b-Python-2.7.14             R/3.4.3-foss-2017b-X11-20171023
   CP2K/5.1-intel-2017b                           Ruby/2.5.0-intel-2017b
   Doxygen/1.8.13-GCCcore-6.4.0                   SCOTCH/6.0.4-foss-2017b
   Eigen/3.3.4                                    SCOTCH/6.0.4-intel-2017b                  (D)
   FFTW/3.3.6-gompi-2017b                         SQLite/3.20.1-GCCcore-6.4.0
   FFTW/3.3.6-intel-2017b                  (D)    SWIG/3.0.12-foss-2017b-Python-2.7.14
   FLUENT/14.0                                    SWIG/3.0.12-foss-2017b-Python-3.6.3
   FLUENT/18.2                             (D)    SWIG/3.0.12-intel-2017b-Python-3.6.3      (D)
   GCC/6.4.0-2.28                                 ScaLAPACK/2.0.2-gompi-2017b-OpenBLAS-0.2.20
   ...
```

There is a *TIS* section for software **non** depending on a toolchain or release (mainly prebuilt software distributed as binaries, or built with the system libraries and compilers)

Common software modules

- We can switch the default release by loading another one

```
[alozano@lemaitre3 ~]$ module load releases/2018b

The following have been reloaded with a version change:
  1) releases/2017b => releases/2018b
```

Common software modules

- We can switch the default release by loading another one

```
[alozano@lemaitre3 ~]$ module load releases/2018b

The following have been reloaded with a version change:
  1) releases/2017b => releases/2018b

[alozano@lemaitre3 ~]$ module av
------------------------------------------- Meta Modules -------------------------------------------
   dot     releases/elic-2017b        releases/2016b (S)      releases/2018a (S)      tis/2018.01 (S,L)
   null    releases/2016a      (S)    releases/2017b (S,D)    releases/2018b (S,L)    use.own

---------------------------- TIS: Toolchain Independent Software (2018.01) ----------------------------
   EasyBuild/3.5.1        MCR/R2014a    MCR/R2017a                          freesurfer/6.0.0
   Java/1.8.0_31          MCR/R2014b    MCR/R2017b                          gurobi/gurobi800
   Java/1.8.0_92          MCR/R2015a    MCR/R2018a                    (D)   julia/0.6.3
   Java/1.8.0_121  (D)    MCR/R2015b    NCBI-BLAST-database/20170306        julia/1.0.0      (D)
   MCR/R2013a             MCR/R2016a    crystal/17-v1.0.1                   xpress/xp850
   MCR/R2013b             MCR/R2016b    crystal/17-v1.0.2             (D)

------------------------------------------ Releases (2018b) ------------------------------------------
   Boost/1.67.0-foss-2018b                  Python/2.7.15-GCCcore-7.3.0-bare                    (D)
   CGAL/4.11.1-foss-2018b-Python-2.7.15     Qt5/5.10.1-foss-2018b
   CMake/3.9.6                              SAMtools/1.9-foss-2018b
   Eigen/3.3.4                              SCOTCH/6.0.5-foss-2018b
   FFTW/3.3.8-gompi-2018b                   SQLite/3.24.0-GCCcore-7.3.0
   GCC/7.3.0-2.30                           ScaLAPACK/2.0.2-gompi-2018b-OpenBLAS-0.3.1
   GLib/2.54.3-GCCcore-7.3.0                X11/20180604-GCCcore-7.3.0
   GMP/6.1.2-GCCcore-7.3.0                  foss/2018b
   GROMACS/2018.3-foss-2018b                gompi/2018b
   HDF5/1.10.2-foss-2018b                   icc/2018.3.222-GCC-7.3.0-2.30
   ...
```

Common software modules

- You can **search** for a specific software to see **all available versions** with the `module spider software` command

```
[alozano@lemaitre3 myjob]$ module spider boost

----------------------------------------------------------------------------------------------------
  Boost:
----------------------------------------------------------------------------------------------------
    Description:
      Boost provides free peer-reviewed portable C++ source libraries.

     Versions:
        Boost/1.60.0-foss-2018a
        Boost/1.61.0-foss-2016b
        Boost/1.61.0-intel-2016a-Python-2.7.11
        Boost/1.63.0-foss-2016b-Python-3.5.2
        Boost/1.65.1-foss-2017b
        Boost/1.65.1-foss-2018a
        Boost/1.66.0-foss-2018a
        Boost/1.66.0-intel-2017b
        Boost/1.66.0-intel-2018a-Python-3.6.4
        Boost/1.67.0-foss-2018b
```

Common software modules

- The new CÉCI clusters will share this setup for the installed modules
- For the default release it will be assured to have the same set of software installed among all the federated clusters
- This is a necessary condition, as every cluster must be able to handle the same `module load` section in your slurm script

Thanks for listening!