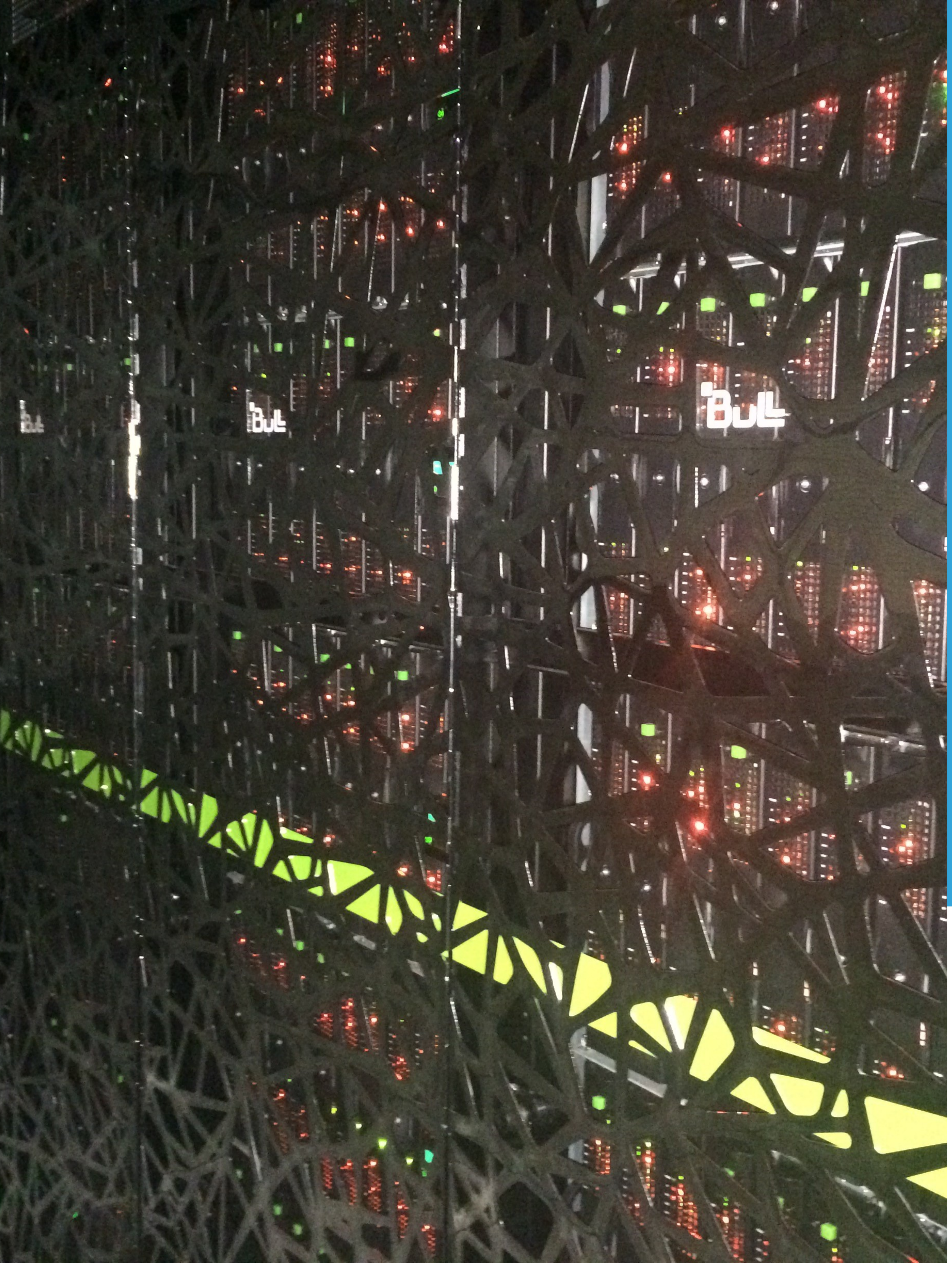
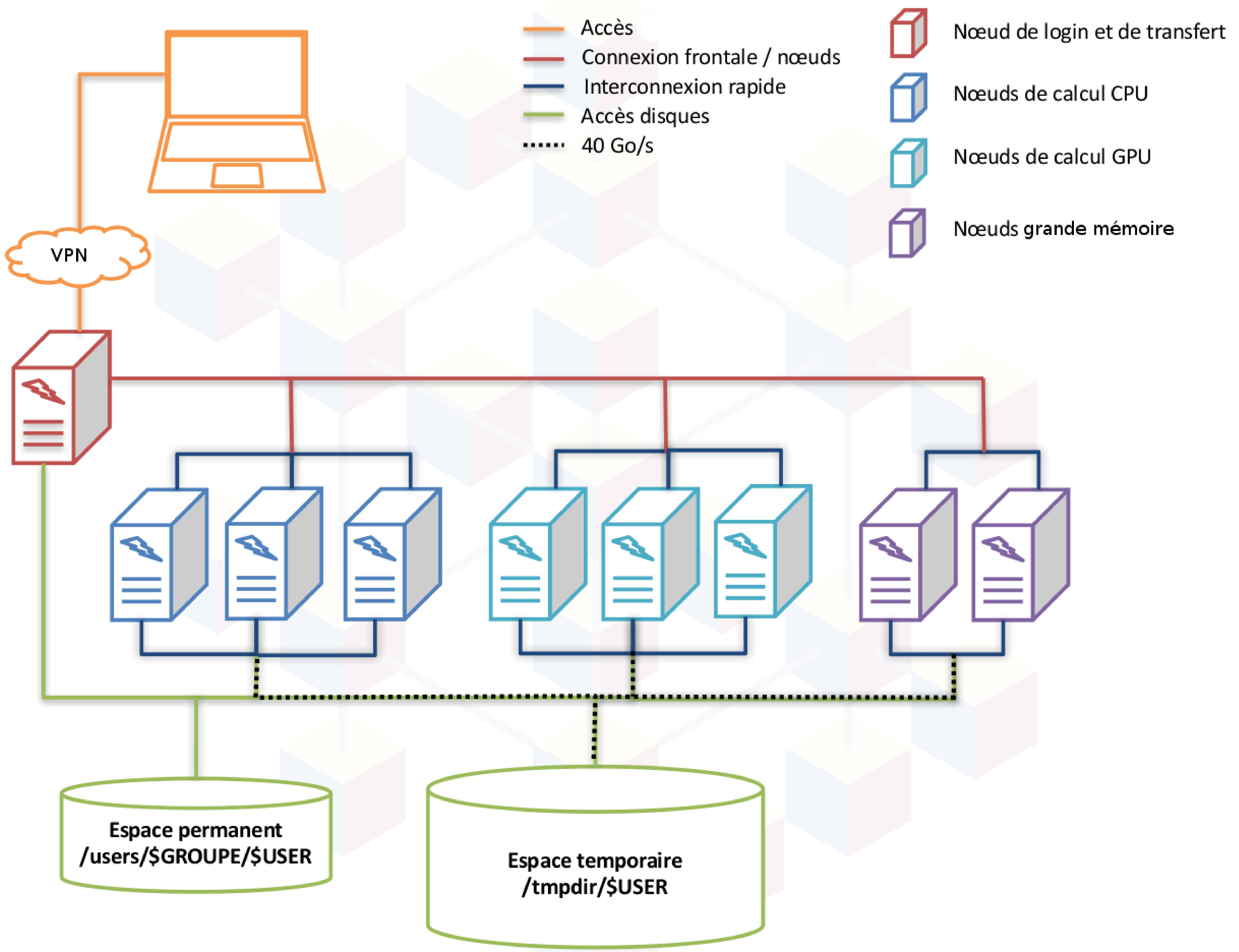


CONTRÔLER VÉRIFIER LE PLACEMENT

Emmanuel Courcelle

CALMIP (UMS 3667)
Espace Clément Ader
www.calmip.univ-toulouse.fr

UN CLUSTER DE CALCUL



Un exemple : Olympe (CALMIP)

3 frontales

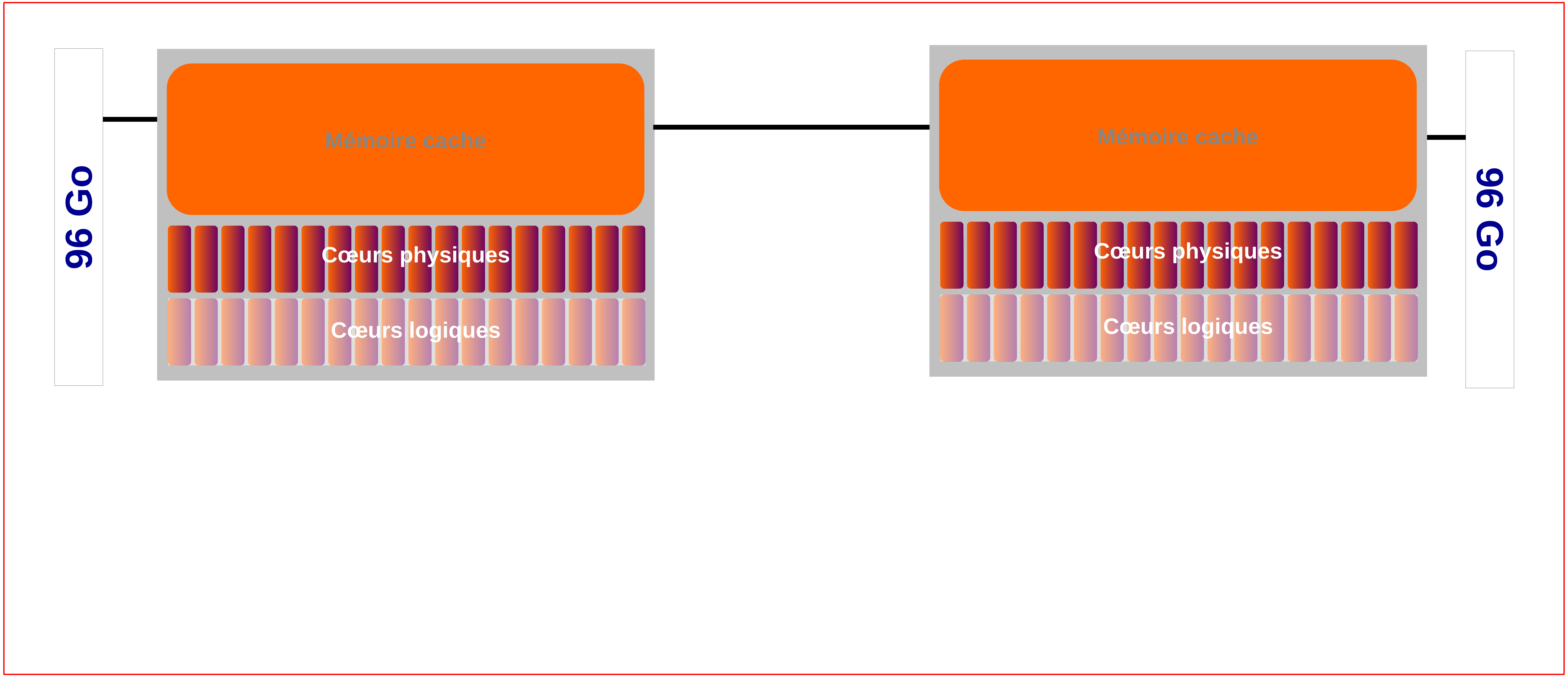
360 nœuds bi-processeurs
- 2 x 18 cœurs
- 192 Go

2 nœuds grande mémoire
- 2 x 18 cœurs
- 1500 Go

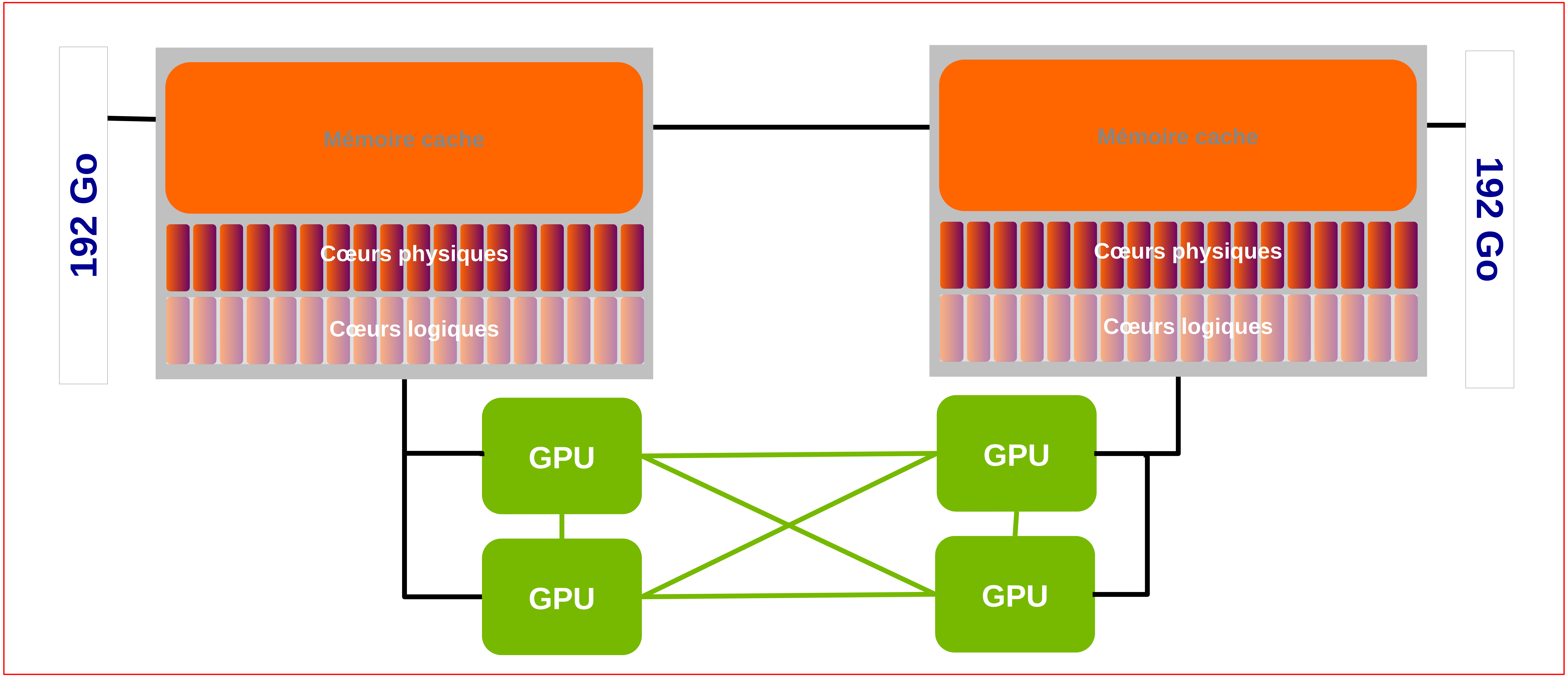
12 nœuds GPU
- 2 x 18 cœurs CPU
- 384 Go
- 4 GPUs connectés :
  - Avec les CPUS en PCI (2 x 2)
  - Entre eux en nvLink

Les nœuds sont **attribués** à un **utilisateur exclusif** durant un **temps limité**

UN NŒUD DE CALCUL

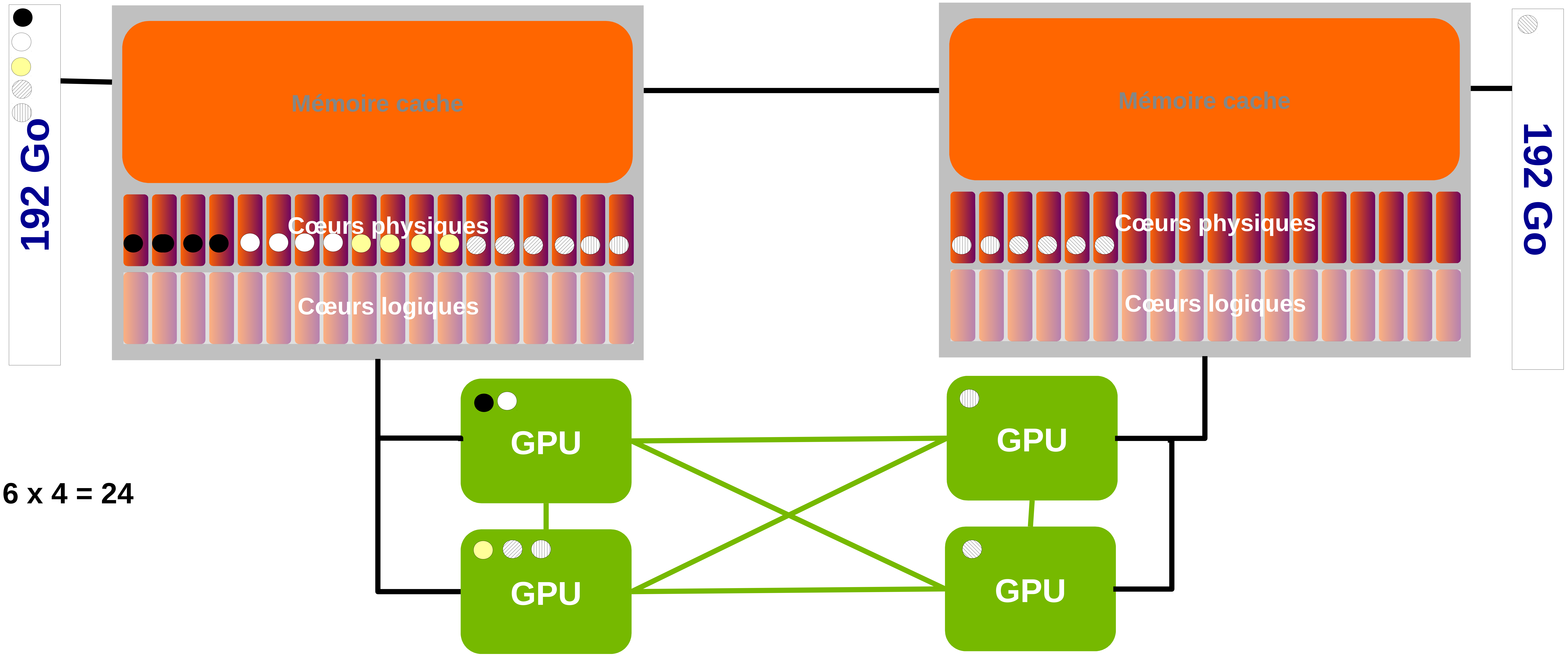


UN NŒUD DE CALCUL avec GPUs



UN PLACEMENT NON OPTIMAL

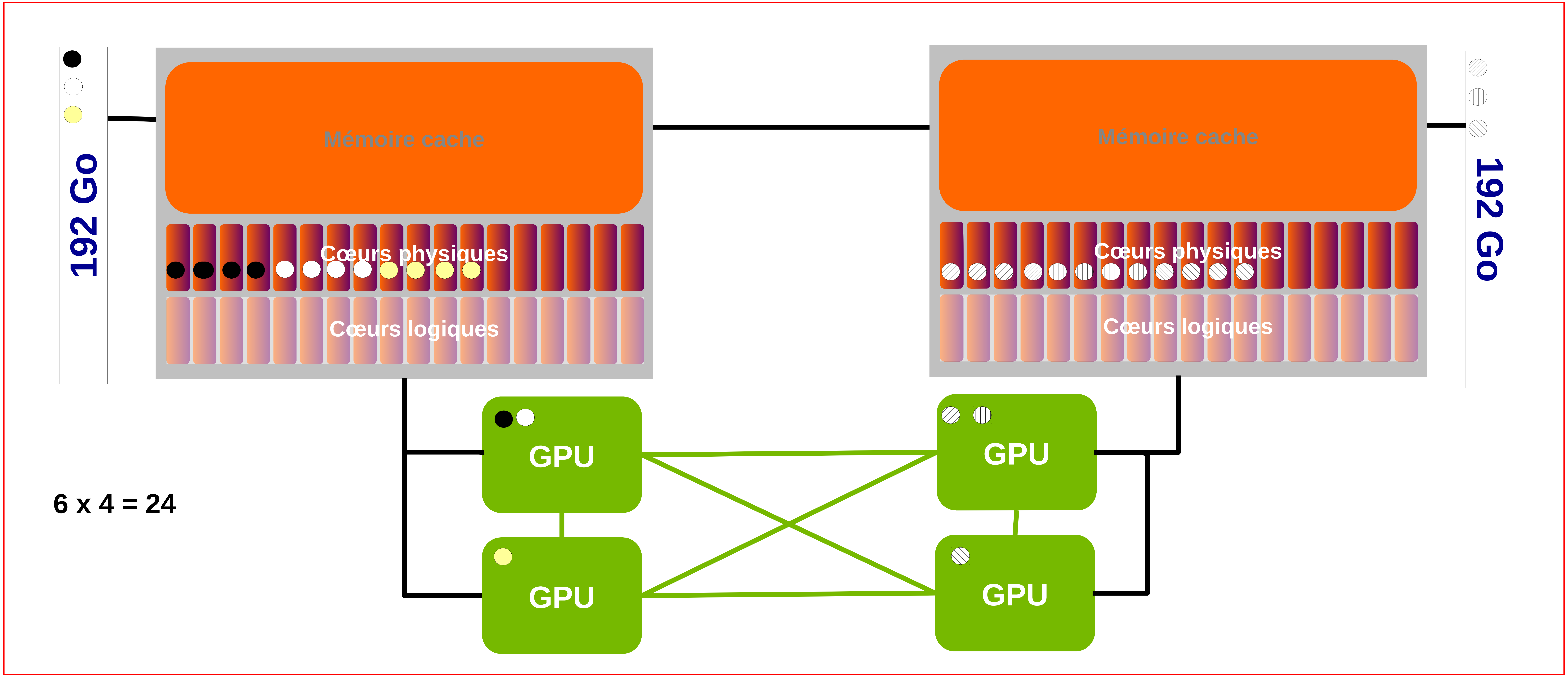
● Processus A (4 threads) ○ Processus B (4 threads) 🟡 Processus C (4 threads) ▨ Processus D (4 threads) ▥ Processus E (4 threads) ▧ Processus F (4 threads)



6 x 4 = 24

UN MEILLEUR PLACEMENT

● Processus A (4 threads) ○ Processus B (4 threads) 🟡 Processus C (4 threads) ▨ Processus D (4 threads) ▥ Processus E (4 threads) ▧ Processus F (4 threads)



6 x 4 = 24

LE PROGRAMME placement

Contrôler le placement (cœurs) lors de l’exécution d’une application

Vérifier **visuellement** que les instructions de placement seront pertinentes

Vérifier que le placement est correct lorsque l’application est **en exécution**

Plusieurs moyens de contrôle du placement

1/ **Variables d’environnement** (1 processus, multithreadé)

```
export GOMP_CPU_AFFINITY="0 1 18 19"
export KMP_AFFINITY="granularity=fine,explicit,proclist=[0,1,18,19]"
```

2/ **numactl**

Pour un code mpi/openmp , une commande différente pour chaque rang !

Rang 0 → `numactl --physcpubind=0-3 my_exe`

Rang 1 → `numactl --physcpubind=18-21 my_exe`

3/ **srun (slurm)**

Des switches de ligne de commande

```
srun --cpu_bind=mask_cpu:0xf,0x3c0000 my_exe
```

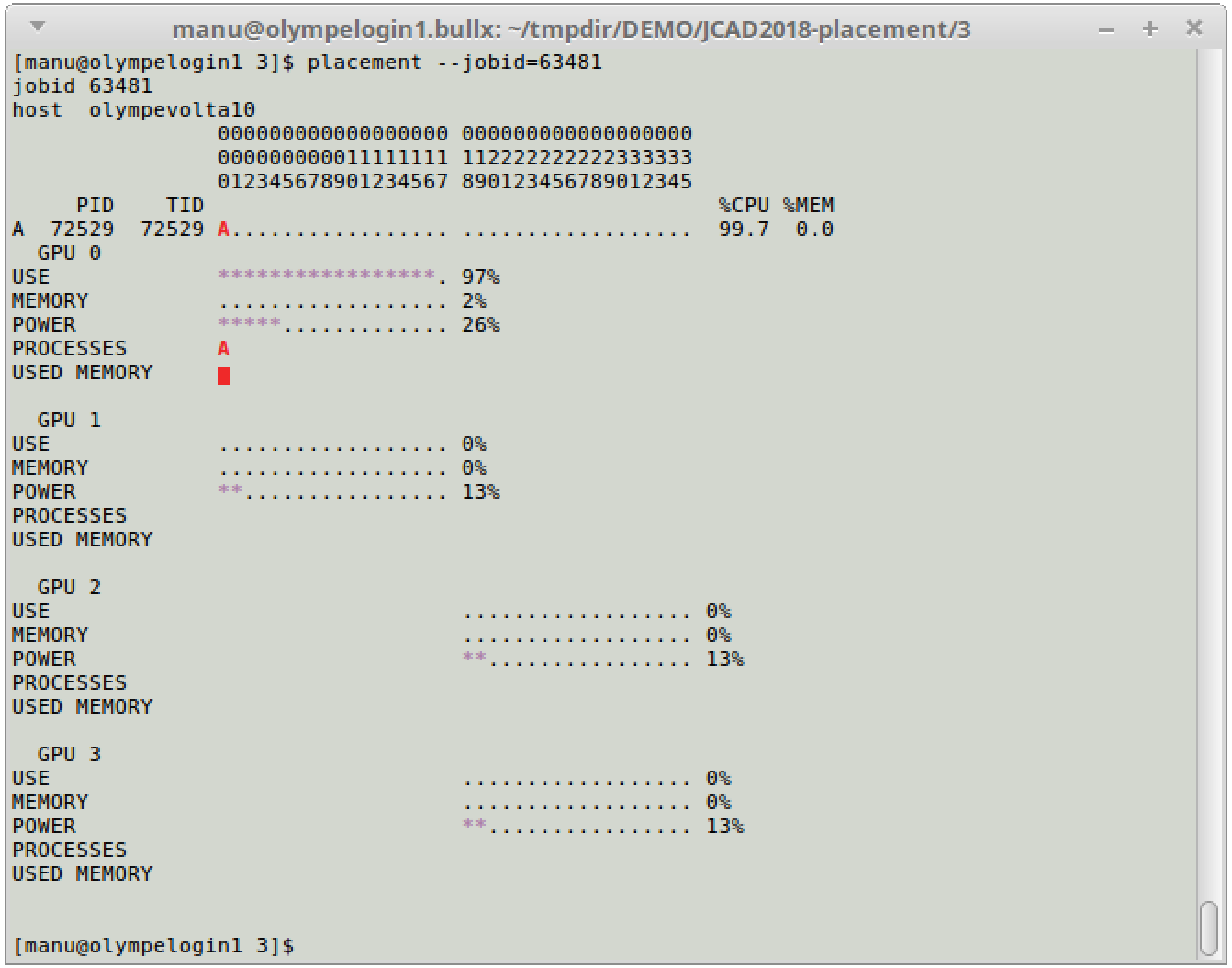
Important pour les perfs !

Exemples de la démo :

+ 8 % lorsque le placement est correct

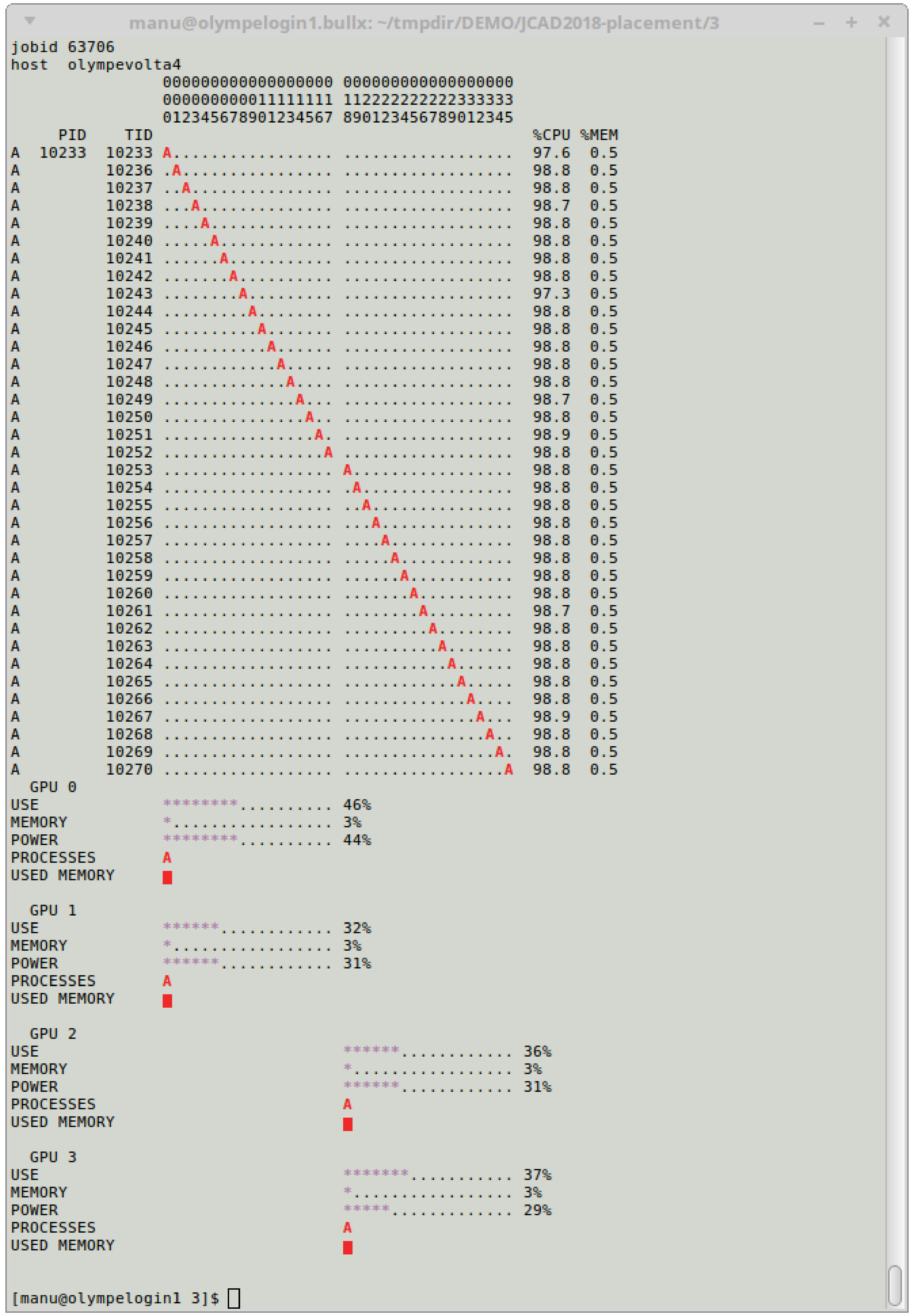
+ 15 % lorsqu’on utilise les GPU

Galerie de photos



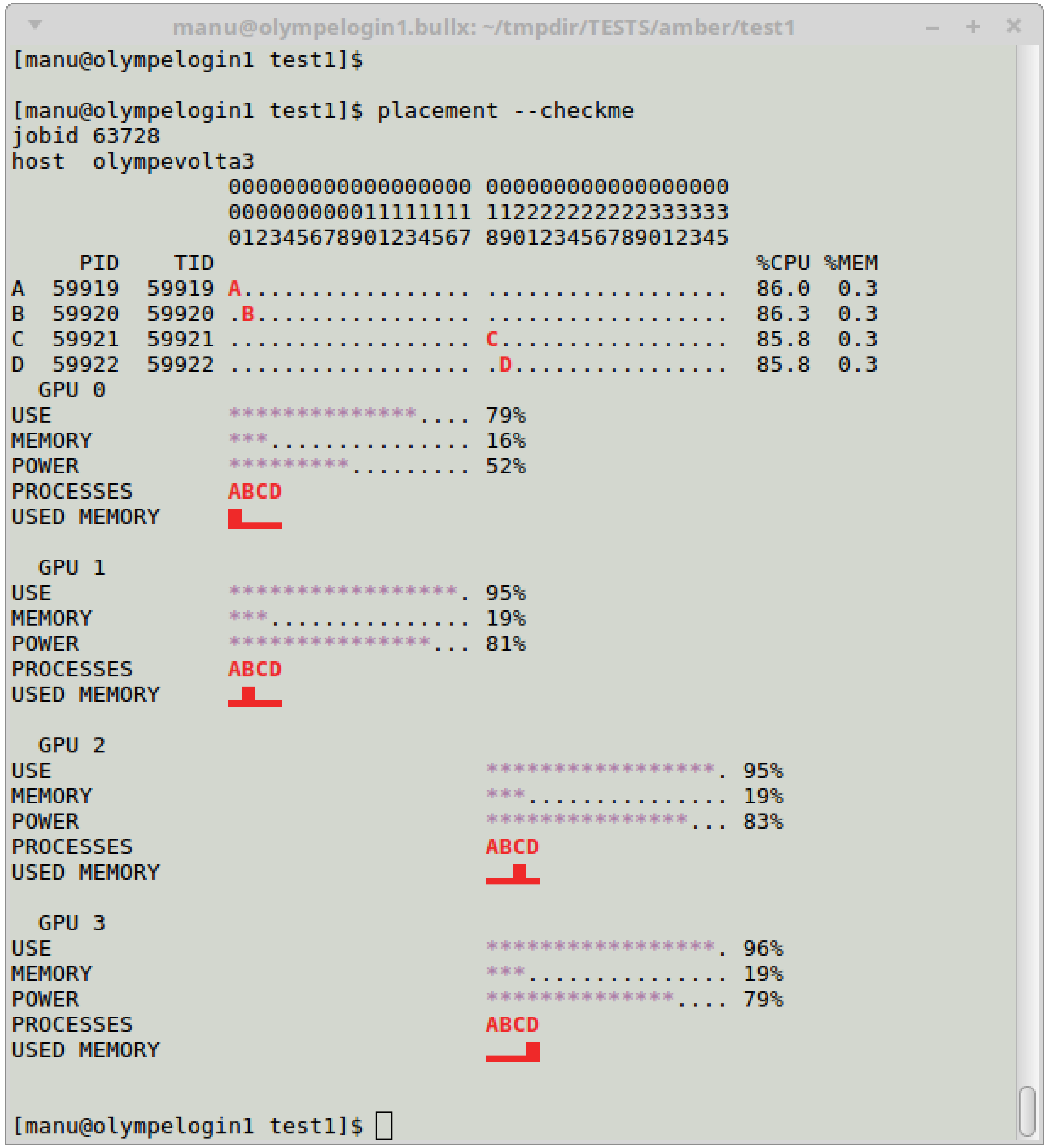
```
[manu@olympelogin1 3]$ placement --jobid=63481
jobid 63481
host  olympevolta0
                000000000000000000 000000000000000000
                000000000011111111 112222222222333333
                012345678901234567 890123456789012345
     PID    TID                                        %CPU %MEM
A  72529  72529 A................. ..................  99.7  0.0
  GPU 0
USE             *****************. 97%
MEMORY          .................. 2%
POWER           *****............. 26%
PROCESSES       A
USED MEMORY     █

  GPU 1
USE             .................. 0%
MEMORY          .................. 0%
POWER           **................ 13%
PROCESSES
USED MEMORY

  GPU 2
USE                                .................. 0%
MEMORY                             .................. 0%
POWER                              **................ 13%
PROCESSES
USED MEMORY

  GPU 3
USE                                .................. 0%
MEMORY                             .................. 0%
POWER                              **................ 13%
PROCESSES
USED MEMORY


[manu@olympelogin1 3]$
```

Galerie de photos



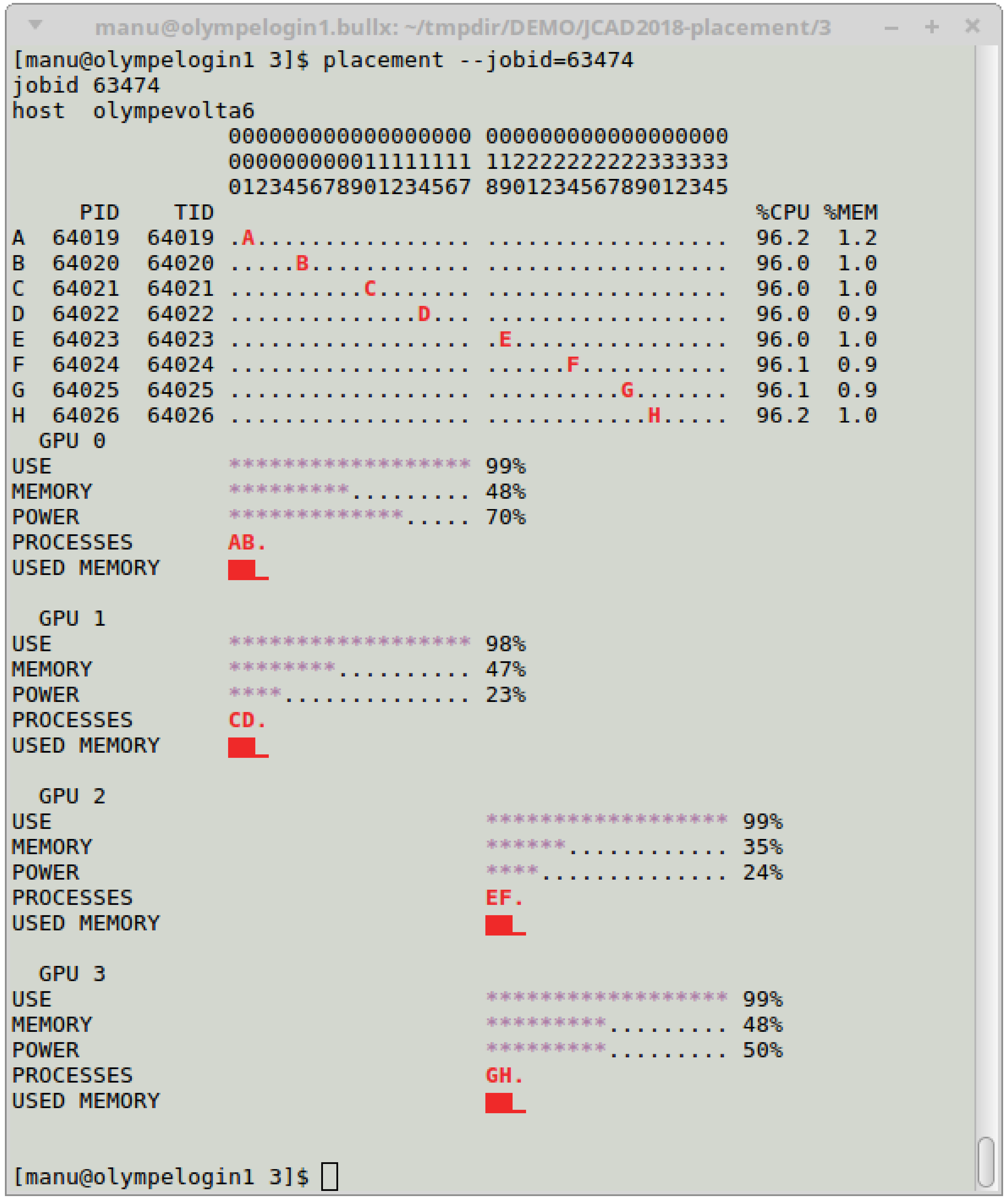
```
jobid 63706
host  olympevolta4
                000000000000000000 000000000000000000
                000000000011111111 112222222222333333
                012345678901234567 890123456789012345
     PID    TID                                        %CPU %MEM
A  10233  10233 A................. ..................  97.6  0.5
A         10236 .A................ ..................  98.8  0.5
A         10237 ..A............... ..................  98.8  0.5
A         10238 ...A.............. ..................  98.7  0.5
A         10239 ....A............. ..................  98.8  0.5
A         10240 .....A............ ..................  98.8  0.5
A         10241 ......A........... ..................  98.8  0.5
A         10242 .......A.......... ..................  98.8  0.5
A         10243 ........A......... ..................  97.3  0.5
A         10244 .........A........ ..................  98.8  0.5
A         10245 ..........A....... ..................  98.8  0.5
A         10246 ...........A...... ..................  98.8  0.5
A         10247 ............A..... ..................  98.8  0.5
A         10248 .............A.... ..................  98.8  0.5
A         10249 ..............A... ..................  98.7  0.5
A         10250 ...............A.. ..................  98.8  0.5
A         10251 ................A. ..................  98.9  0.5
A         10252 .................A ..................  98.8  0.5
A         10253 .................. A.................  98.8  0.5
A         10254 .................. .A................  98.8  0.5
A         10255 .................. ..A...............  98.8  0.5
A         10256 .................. ...A..............  98.8  0.5
A         10257 .................. ....A.............  98.8  0.5
A         10258 .................. .....A............  98.8  0.5
A         10259 .................. ......A...........  98.8  0.5
A         10260 .................. .......A..........  98.8  0.5
A         10261 .................. ........A.........  98.7  0.5
A         10262 .................. .........A........  98.8  0.5
A         10263 .................. ..........A.......  98.8  0.5
A         10264 .................. ...........A......  98.8  0.5
A         10265 .................. ............A.....  98.8  0.5
A         10266 .................. .............A....  98.8  0.5
A         10267 .................. ..............A...  98.9  0.5
A         10268 .................. ...............A..  98.8  0.5
A         10269 .................. ................A.  98.8  0.5
A         10270 .................. .................A  98.8  0.5
  GPU 0
USE             ********.......... 46%
MEMORY          *................. 3%
POWER           ********.......... 44%
PROCESSES       A
USED MEMORY     █

  GPU 1
USE             ******............ 32%
MEMORY          *................. 3%
POWER           ******............ 31%
PROCESSES       A
USED MEMORY     █

  GPU 2
USE                                ******............ 36%
MEMORY                             *................. 3%
POWER                              ******............ 31%
PROCESSES                          A
USED MEMORY                        █

  GPU 3
USE                                *******........... 37%
MEMORY                             *................. 3%
POWER                              *****............. 29%
PROCESSES                          A
USED MEMORY                        █


[manu@olympelogin1 3]$ 
```

Galerie de photos



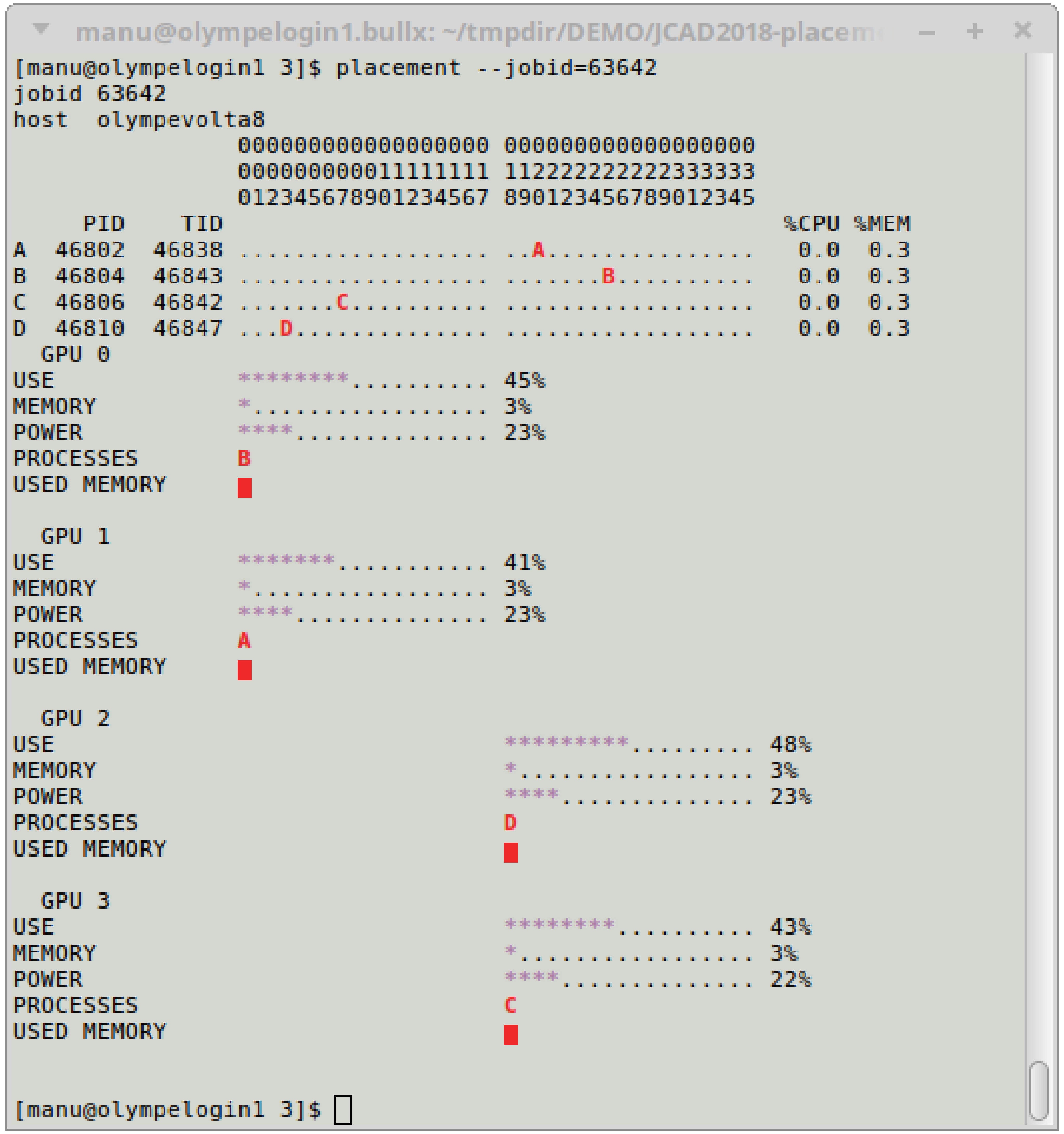
```
[manu@olympelogin1 test1]$

[manu@olympelogin1 test1]$ placement --checkme
jobid 63728
host  olympevolta3
                000000000000000000 000000000000000000
                000000000011111111 112222222222333333
                012345678901234567 890123456789012345
     PID    TID                                        %CPU %MEM
A  59919  59919 A................. ..................  86.0  0.3
B  59920  59920 .B................ ..................  86.3  0.3
C  59921  59921 .................. C.................  85.8  0.3
D  59922  59922 .................. .D................  85.8  0.3
  GPU 0
USE             **************.... 79%
MEMORY          ***............... 16%
POWER           *********......... 52%
PROCESSES       ABCD
USED MEMORY

  GPU 1
USE             *****************. 95%
MEMORY          ***............... 19%
POWER           ***************... 81%
PROCESSES       ABCD
USED MEMORY

  GPU 2
USE                                *****************. 95%
MEMORY                             ***............... 19%
POWER                              ***************... 83%
PROCESSES                          ABCD
USED MEMORY

  GPU 3
USE                                *****************. 96%
MEMORY                             ***............... 19%
POWER                              **************.... 79%
PROCESSES                          ABCD
USED MEMORY


[manu@olympelogin1 test1]$
```

Galerie de photos



```
[manu@olympelogin1 3]$ placement --jobid=63474
jobid 63474
host  olympevolta6
                000000000000000000 000000000000000000
                000000000011111111 112222222222333333
                012345678901234567 890123456789012345
     PID    TID                                        %CPU %MEM
A  64019  64019 .A................ ..................  96.2  1.2
B  64020  64020 .....B............ ..................  96.0  1.0
C  64021  64021 ..........C....... ..................  96.0  1.0
D  64022  64022 ..............D... ..................  96.0  0.9
E  64023  64023 .................. .E................  96.0  1.0
F  64024  64024 .................. ......F...........  96.1  0.9
G  64025  64025 .................. ..........G.......  96.1  0.9
H  64026  64026 .................. ............H.....  96.2  1.0
  GPU 0
USE             ****************** 99%
MEMORY          *********......... 48%
POWER           *************..... 70%
PROCESSES       AB.
USED MEMORY     ██_

  GPU 1
USE             ****************** 98%
MEMORY          ********.......... 47%
POWER           ****.............. 23%
PROCESSES       CD.
USED MEMORY     ██_

  GPU 2
USE                                ****************** 99%
MEMORY                             ******............ 35%
POWER                              ****.............. 24%
PROCESSES                          EF.
USED MEMORY                        ██_

  GPU 3
USE                                ****************** 99%
MEMORY                             *********......... 48%
POWER                              *********......... 50%
PROCESSES                          GH.
USED MEMORY                        ██_

[manu@olympelogin1 3]$ 
```

Galerie de photos



```
[manu@olympelogin1 3]$ placement --jobid=63642
jobid 63642
host  olympevolta8
                000000000000000000 000000000000000000
                000000000011111111 112222222222333333
                012345678901234567 890123456789012345
     PID    TID                                        %CPU %MEM
A  46802  46838 .................. ..A...............   0.0  0.3
B  46804  46843 .................. .......B..........   0.0  0.3
C  46806  46842 .......C.......... ..................   0.0  0.3
D  46810  46847 ...D.............. ..................   0.0  0.3
  GPU 0
USE             ********.......... 45%
MEMORY          *................. 3%
POWER           ****.............. 23%
PROCESSES       B
USED MEMORY     █

  GPU 1
USE             *******........... 41%
MEMORY          *................. 3%
POWER           ****.............. 23%
PROCESSES       A
USED MEMORY     █

  GPU 2
USE                                *********......... 48%
MEMORY                             *................. 3%
POWER                              ****.............. 23%
PROCESSES                          D
USED MEMORY                        █

  GPU 3
USE                                ********.......... 43%
MEMORY                             *................. 3%
POWER                              ****.............. 22%
PROCESSES                          C
USED MEMORY                        █

[manu@olympelogin1 3]$ 
```

Installez, Essayez, Contribuez

Placement fonctionne :

Avec slurm

Avec d’autres gestionnaires de batch ?

→ **Contribuez !** 🙂

Sans gestionnaire de batch

Avec des processeurs intel (2 threads hardware / cœur)

Avec d’autres processeurs ?

→ **Contribuez !** 🙂

https://github.com/calmip/placement