CBPsmn
Centre Blaise Pascal de Simulation et de Modélisation Numérique

Josy 2023 / ZFS

Cas d'usage : Centre de calculs, bigdata

Loïs Taulelle - 2023-11-29

http://www.ens-lyon.fr/PSMN/

Pôle Scientifique de Modélisation Numérique

Usine de production du
Centre Blaise Pascal Science et Modélisation Numérique

- 4 clusters, 8 partitions
- Soumission de jobs en best effort (Slurm)
- Séquentiel, parallèle, tests
- De 1 à 768 cores (7216 max), 3 à 16GiB/core
- Walltime moyen : 8 jours (max 30)
- Interconnexion rapide (IB 56 et 100Gb/s)
- ~600+ nœuds de calcul (~30 000 cores)
- nœuds dédiés (fatnode, visu) et services
- ~750 utilisateurs pour ~41 laboratoires

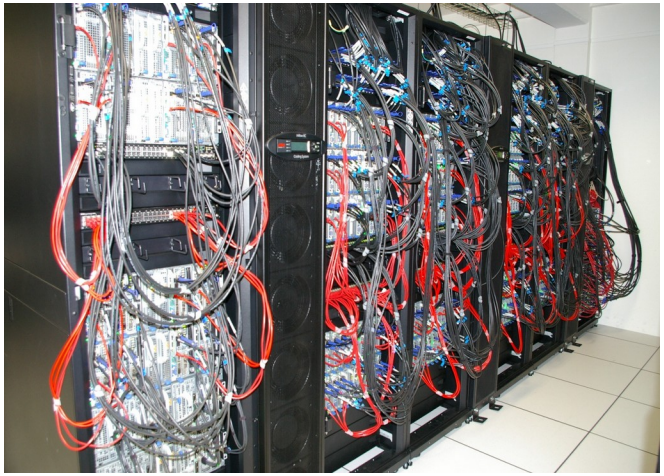
La collaboration CBP-PSMN permet l'exploration technologique en avance de phase, et offre des choix techniques habituellement inaccessible à des centres de notre taille (intra-établissement).

Une équipe réduite

H. Gilquin (*IR, CNRS/UMPA, 80%*) [1993-2023]
G. Lasseur (*IR, CNRS/UMPA, 50%*) [1993-2015]

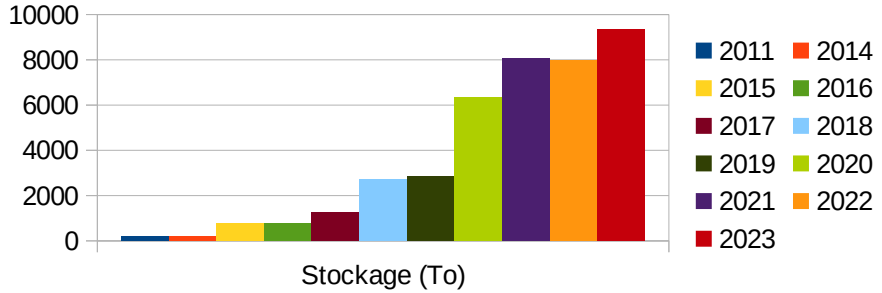
L. Taulelle (*AI, ENSL, 100 %*) [2011-...]
C. Calugaru (*IR, ENSL, 50%*) [2010-...]
M. Calvas (*AI, ENSL, 50 %*) [2017-...]

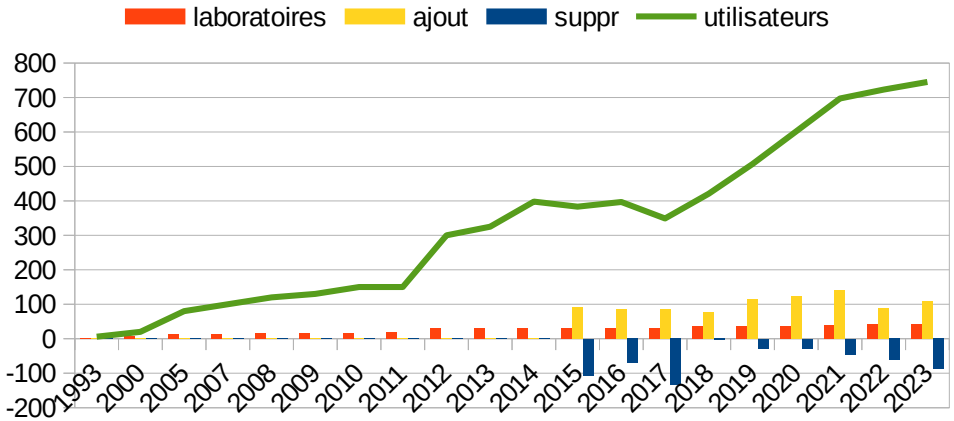
Collaboration régulière (permanente) :
E. Quemener (*IR, ENSL*) [2010-...]

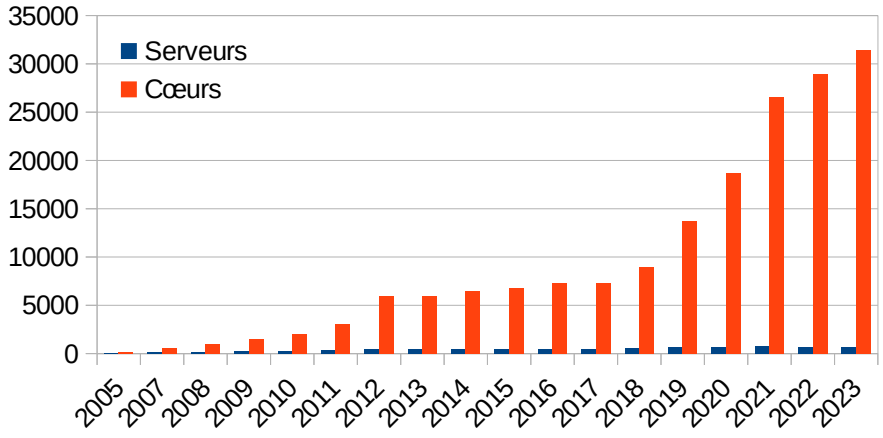


Missions & Évolutions

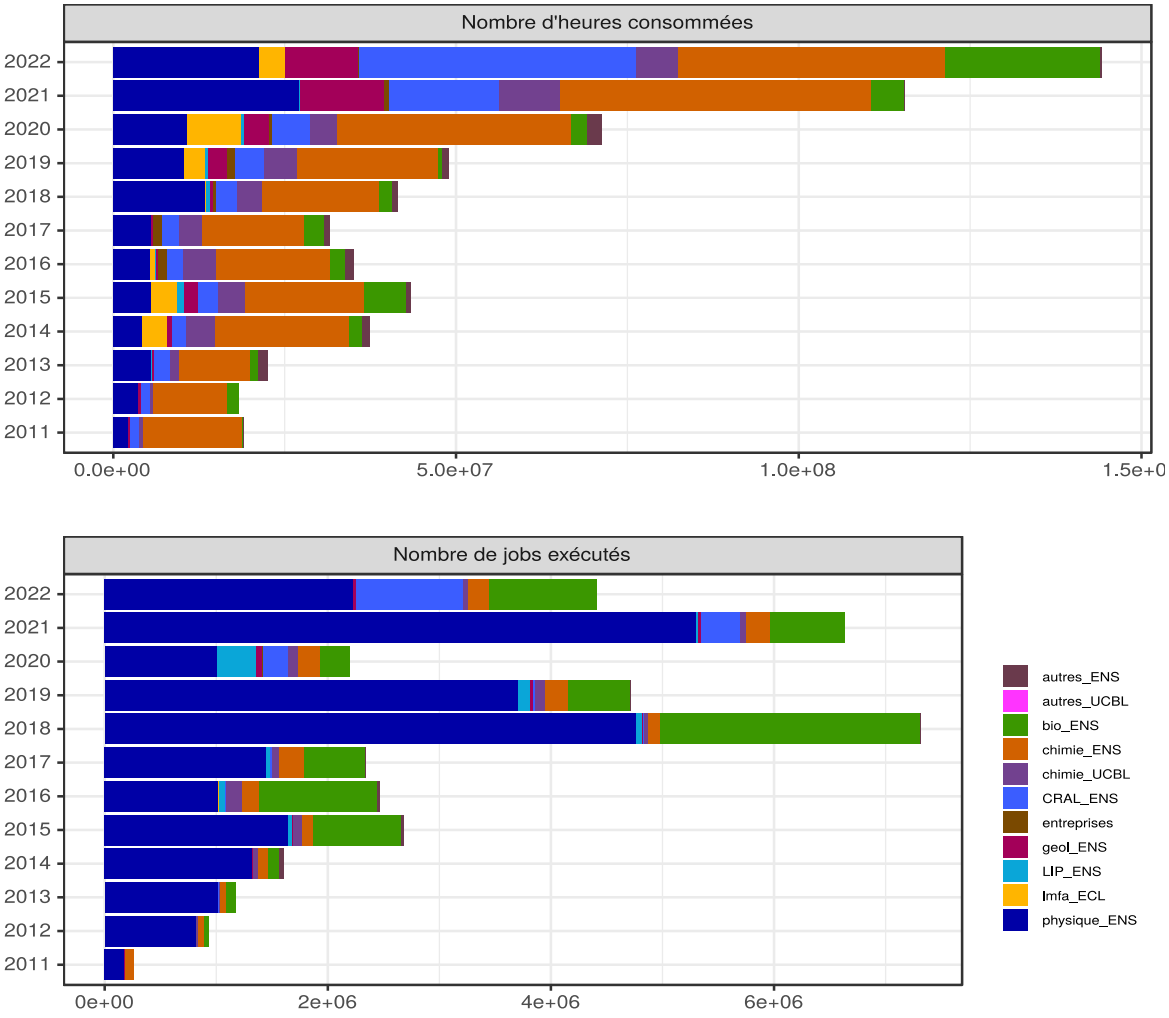
- **Mutualiser les moyens de calcul à l'ENS de Lyon**
- Faire la passerelle entre le laboratoire et le grand centre (Tier1, Tier0)
- Compléter l'offre des grands centres



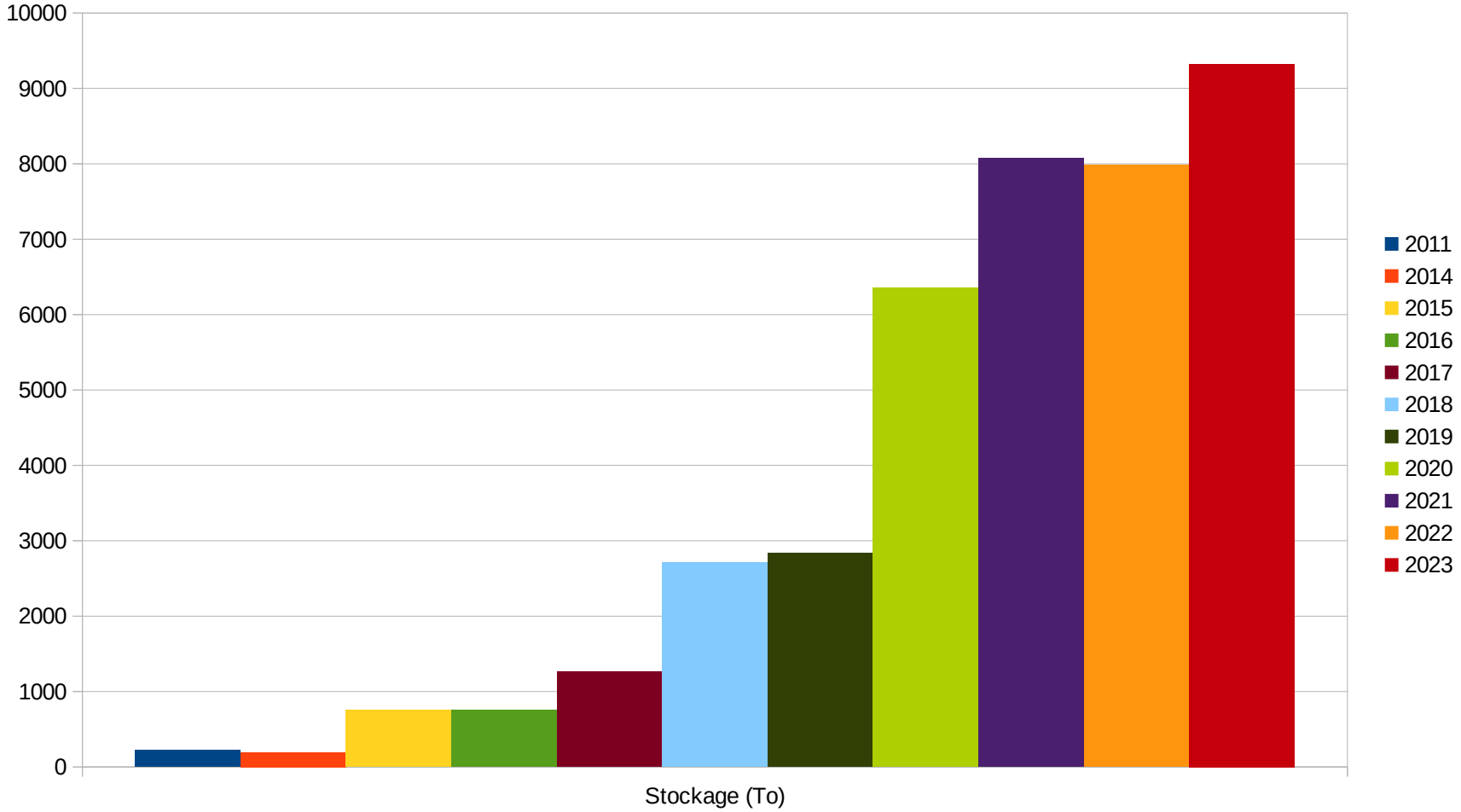




De la Science : Plein !!



D’où, des données, plein ! → Data! Big!



Donc, ZFS...

- Utilisé au PSMN depuis que ça existe (sur Sun X4500, Slowlaris, puis Debian)
- Du Dell R510 au Dell R740xd... (de 12 à 24 disques)
- Du SAS, surtout ! (respectueux des normes)
  - MD1200/1400 (12 disques)
  - LSI MD3060e (Wembley, 60 disques)
  - Seagate Exos X 5U84 (ME484, 84 disques)
- Des frontales correctement dimensionnées (ratio core/ram)...
  - **vieille** règle : 1GHz + 1GiB pour 1TiB... (*avant 2010*)
  - Ne **JAMAIS** négliger la RAM
  - Ne pas hésiter sur les cores (plus, mais lents = OK)

CBPsmn
Centre Blaise Pascal de Simulation et de Modélisation Numérique

Pourquoi ?

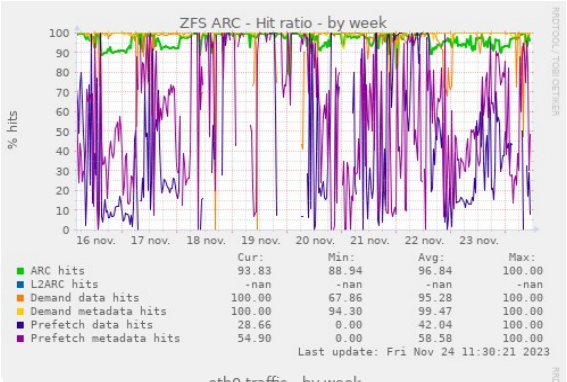
- **Prime Directive** : supporter le mode « Panic! »
- Donc, rester simple : peu de protocoles, peu de briques et des trucs éprouvés
  - SAS, ZFS et NFSv3
- Des volumétries très variables
  - Une baie = un seul pool, des volumes
- Des volumétries très changeantes
  - Les quota sont modifiables à chaud
- Des I/O importantes, mais pas toujours
  - Du CPU, de la RAM, du multi-attachement
  - Séparer les disciplines, étaler les disciplines

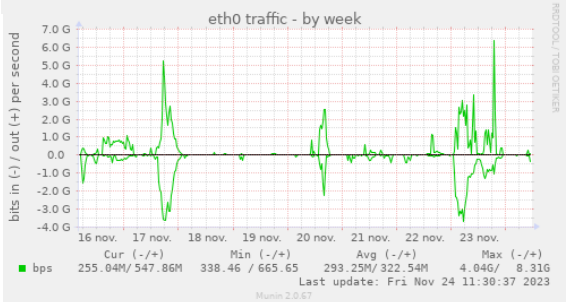
Exemples (1/1)

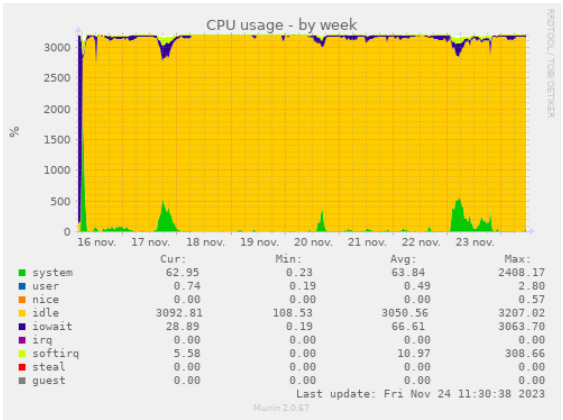
- Sun X4150 ‘heager’ → NAS (depuis 2010)
  - 2x Xeon E5440 @ 2.83GHz (8c), 24GiB RAM, 2x 300GiB OS
  - 6x 1TiB data (4,7TiB utile), 1x 1Gib/s
  - de Solaris (ZFSv4) à Debian 12 (OpenZFS 2.1)
  - ssh, nfs et rsync vers ~10 clients
- Dell R740xd ‘data11’ → NAS + DAS (2020)
  - 2x Xeon 4214R @ 2.40GHz (48c HT), 192GiB RAM, 2x 240GiB OS
  - 3x 12x 16TiB (Front + 2x MD1400, 520TiB utile), 1x 10Gib/s
  - double-attachement SAS, multipath-tools
  - Données CRAL, **2** users

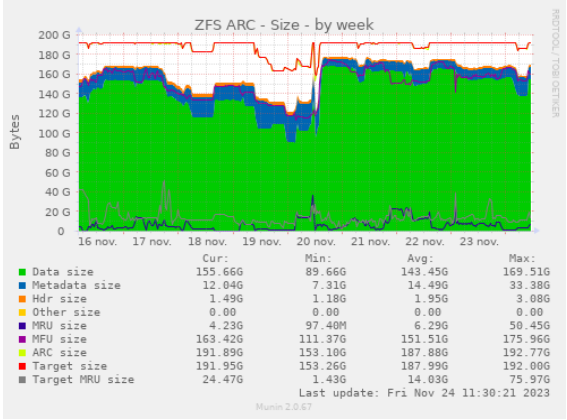
Exemples (1/2)

- Dell R730xd ‘data7’ → NAS + DAS (depuis ~2018)
  - 2x E5-2640 @ 2.60GHz (32c HT), 256GiB RAM, 2x 300GiB OS
  - 24x 1TiB data (~20TiB utile), 1x 10Gib/s
  - 1x MD3060e (60x 4TiB, multipath, ~200TiB utile)
  - 1x ME484 (84x 16TiB, multipath, ~800TiB utile)
  - homes et data des Bio (~240 users), *commence à faiblir...*









CBPsmn
Centre Blaise Pascal de Simulation et de Modélisation Numérique

Des questions ?