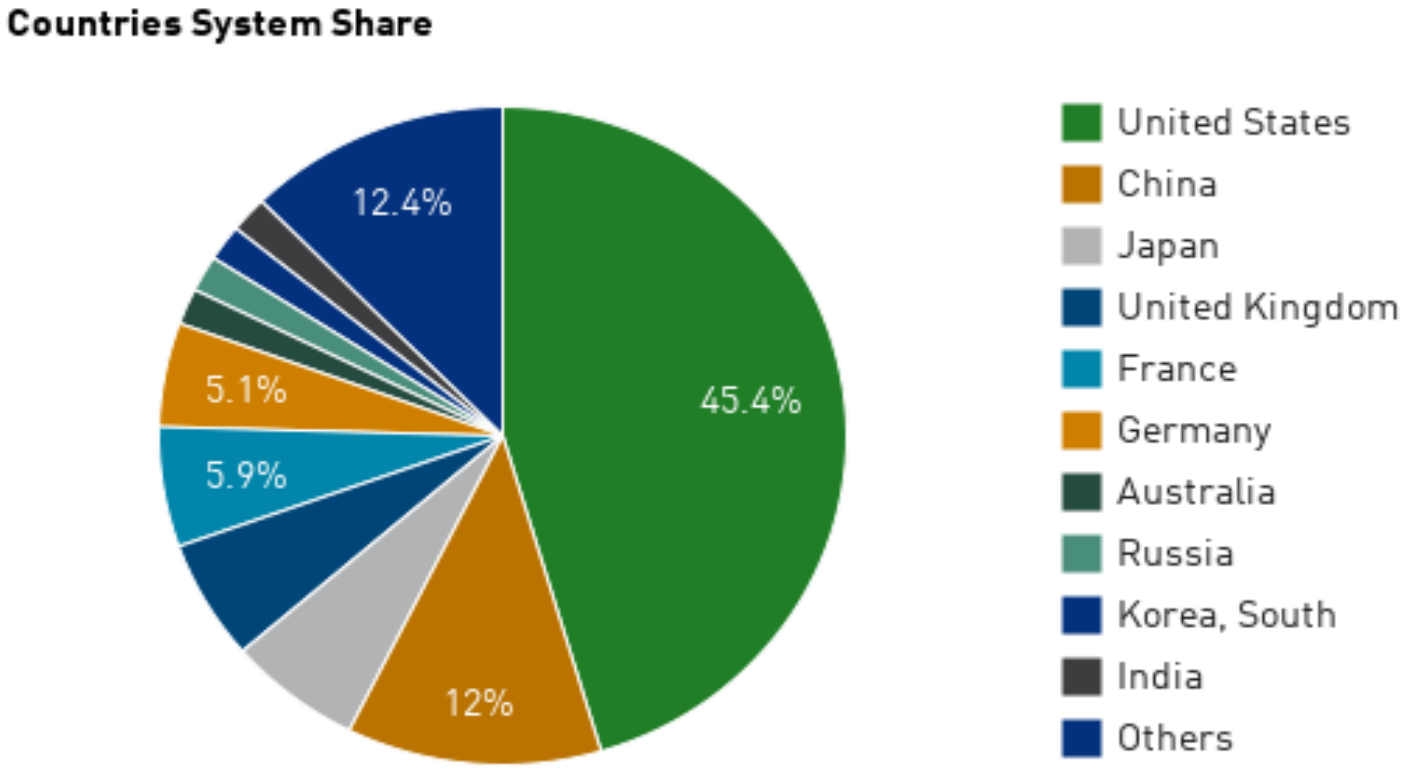
CINES

Introduction aux architectures parallèles et au supercalculateur Occigen

Emilie Boulard

Représentation de la somme de Gflops (Rpeak) par pays



Exemple de constructeurs :

HP

IBM

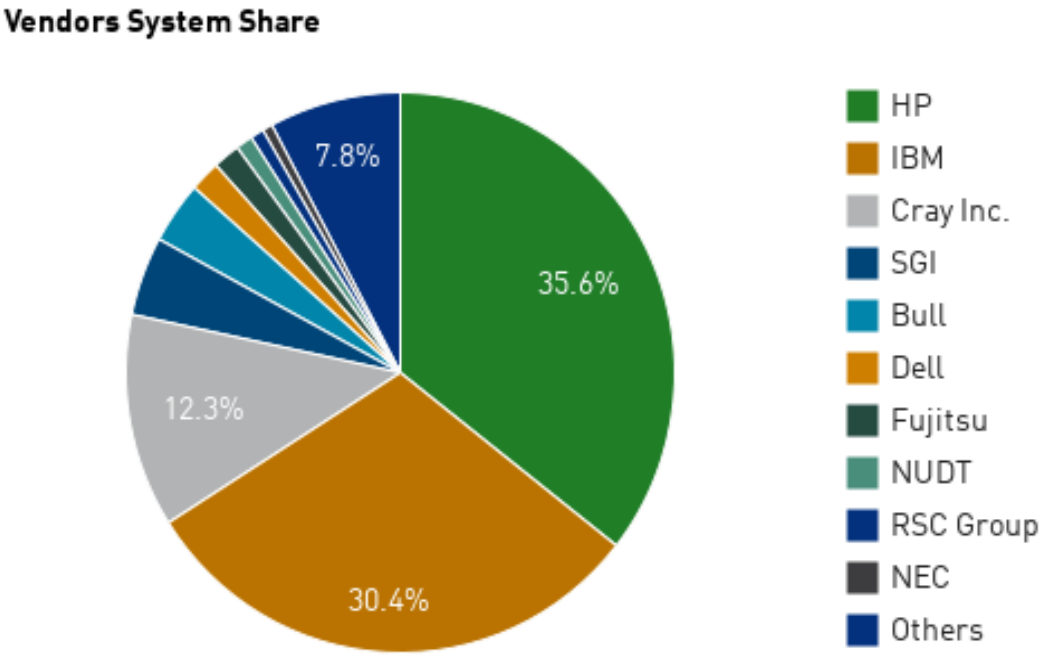
CRAY

SGI

Bull

Dell

...

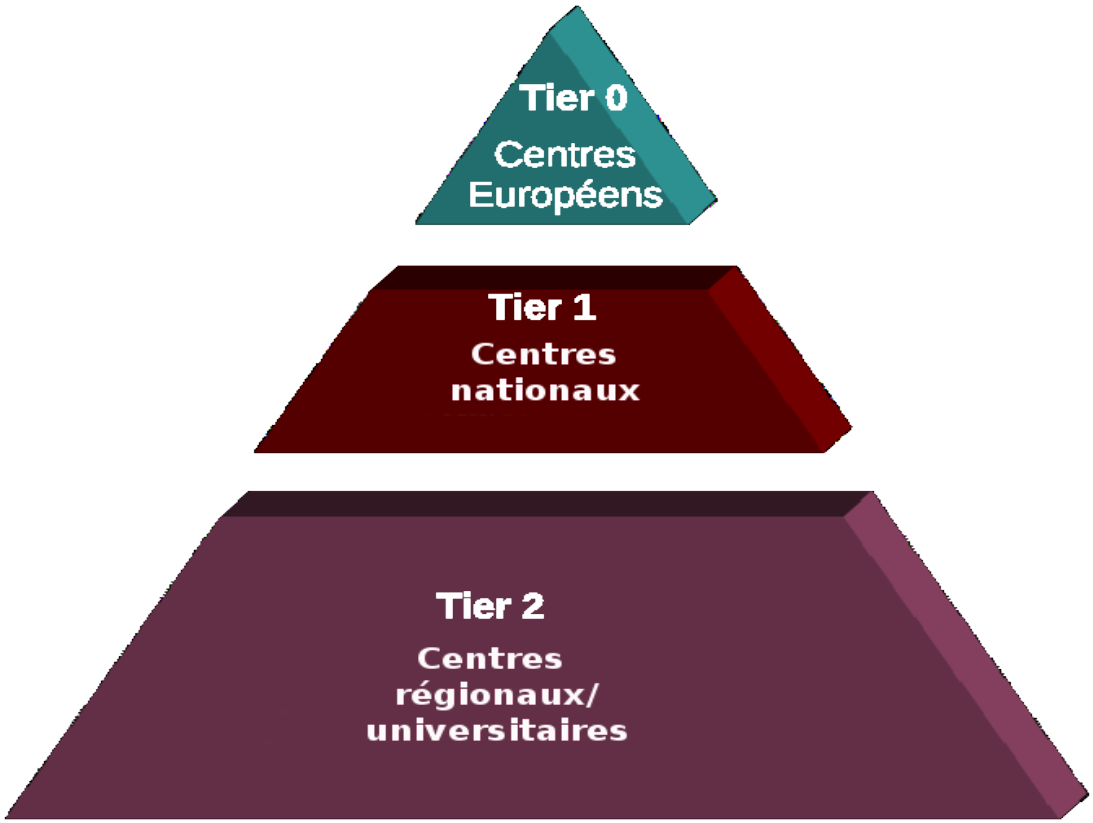


TOP 500

INES Centre Informatique National de l'Enseignement Supérieur

Organisation des centres de calcul:

- 3 types de centres : Tier2, Tier1, Tier0.
- En France gérés par GENCI.



Centres de calcul en Europe :

CINECA

BSC

Sara

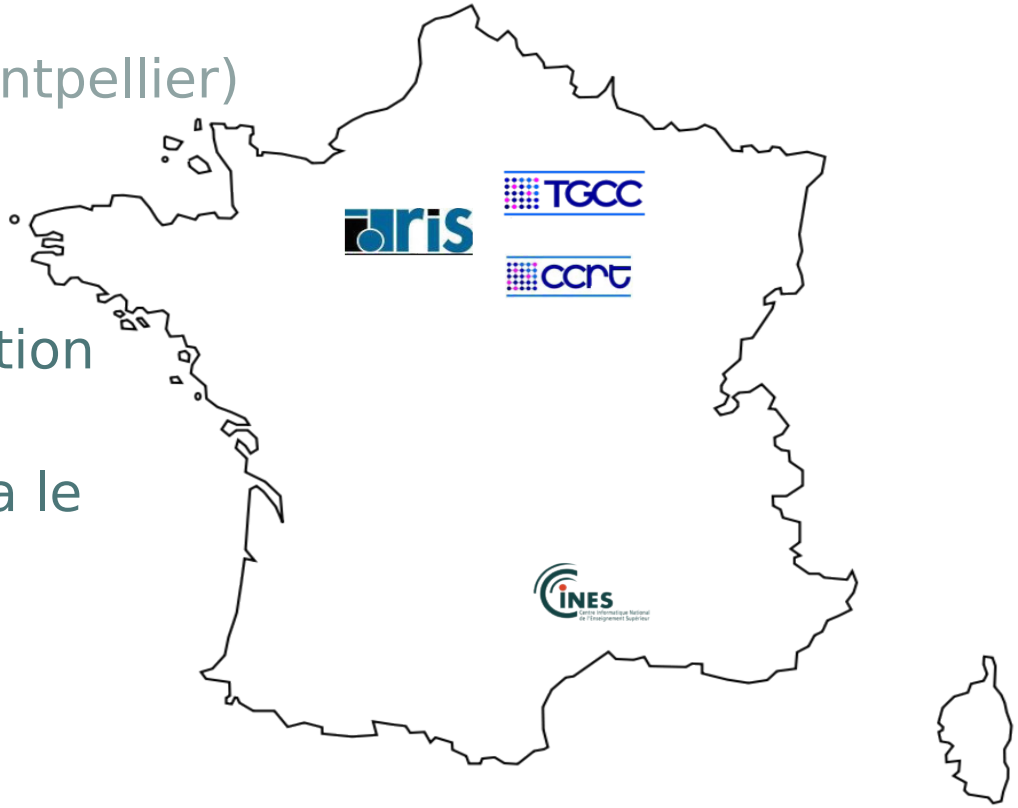
JÜLICH

CSC

...

Centres de calcul en France:

- 3 centres nationaux:
  - TGCC du CEA (Bruyère Le Chatel)
  - IDRIS du CNRS (Orsay)
  - CINES du MENESR (Montpellier)



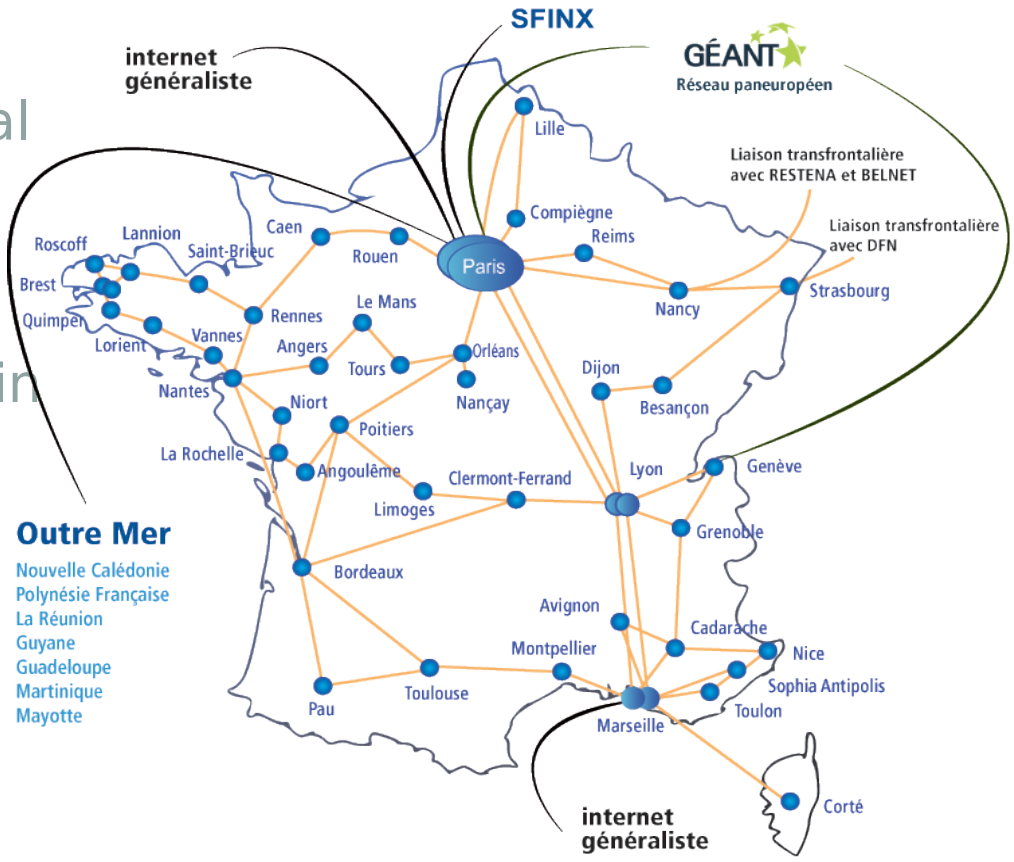
- Financement et coordination réalisées par GENCI.
- Attribution des heures via le DARI.

Centres de calcul en France:

- ➔ Mésocentres
  - Niveau régional.
  - Ensemble de ressources humaines, matérielles, logicielles pour le HPC.

Infrastructure réseau:

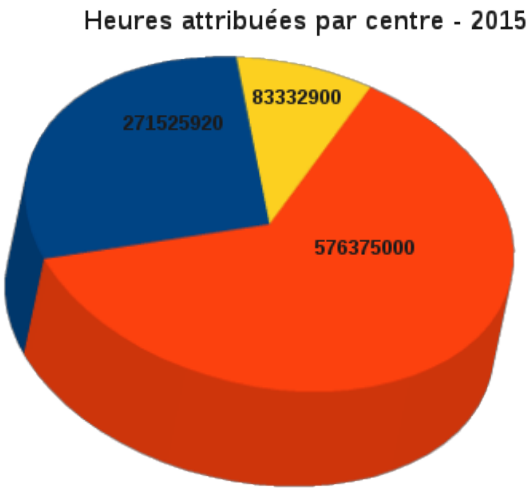
- Réseau National de télécommunications pour la Technologie l'Enseignement et la Recherche.
- Switch nationnal de RENATER hébergé au CINES.
- Différents niveaux :
  - International
  - National
  - Régional
  - Métropolitain

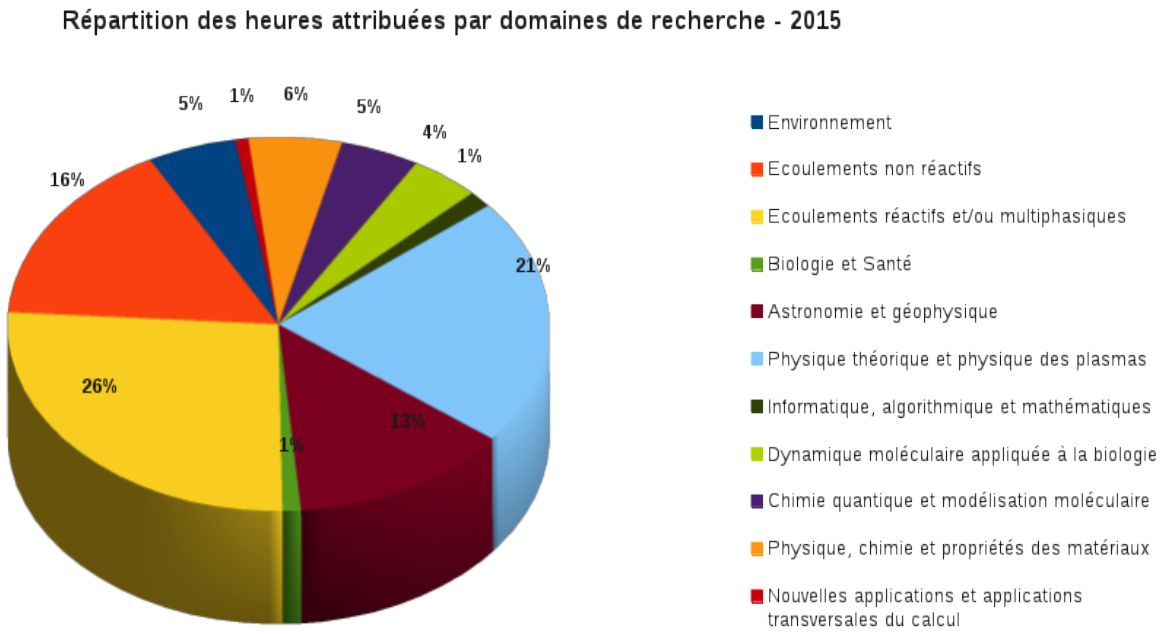


DARI:

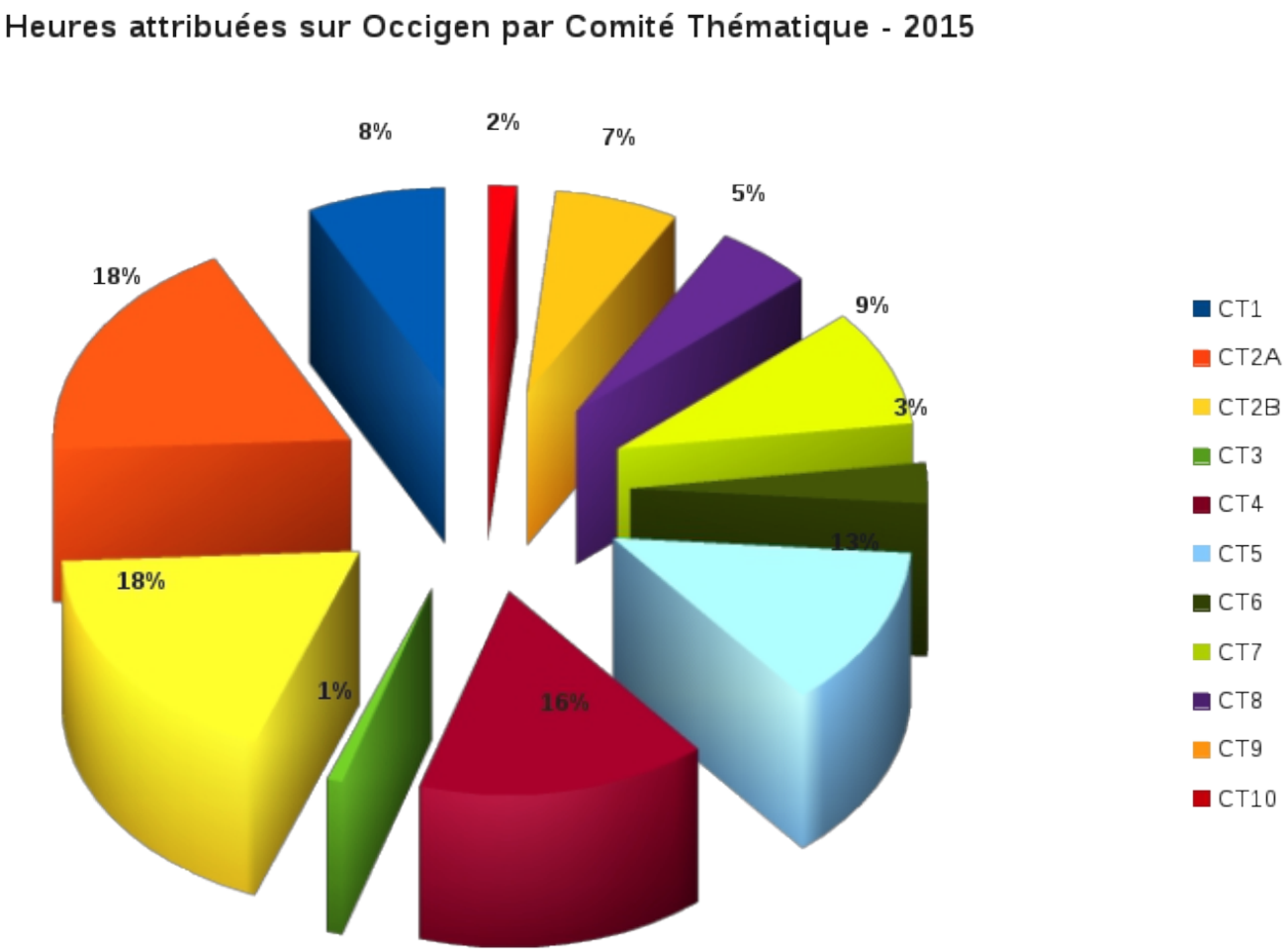
- Accès : www.edari.fr (développée par le CINES)
- 2 appels par an.
- Nombre d'heures attribuées via le DARI en 2015: 931 233 820 heures.
- Attribution basée sur une expertise scientifique et technique.
- Maintien d'un rythme de soumission régulier : au moins 30% des heures doivent être consommées avant fin mai.
    - Si en dessous des 30%: écretage de 40% de l'attribution initiale.
    - Possibilité de prévenir en amont afin d'éviter l'écretage.

Le DARI en quelques chiffres:

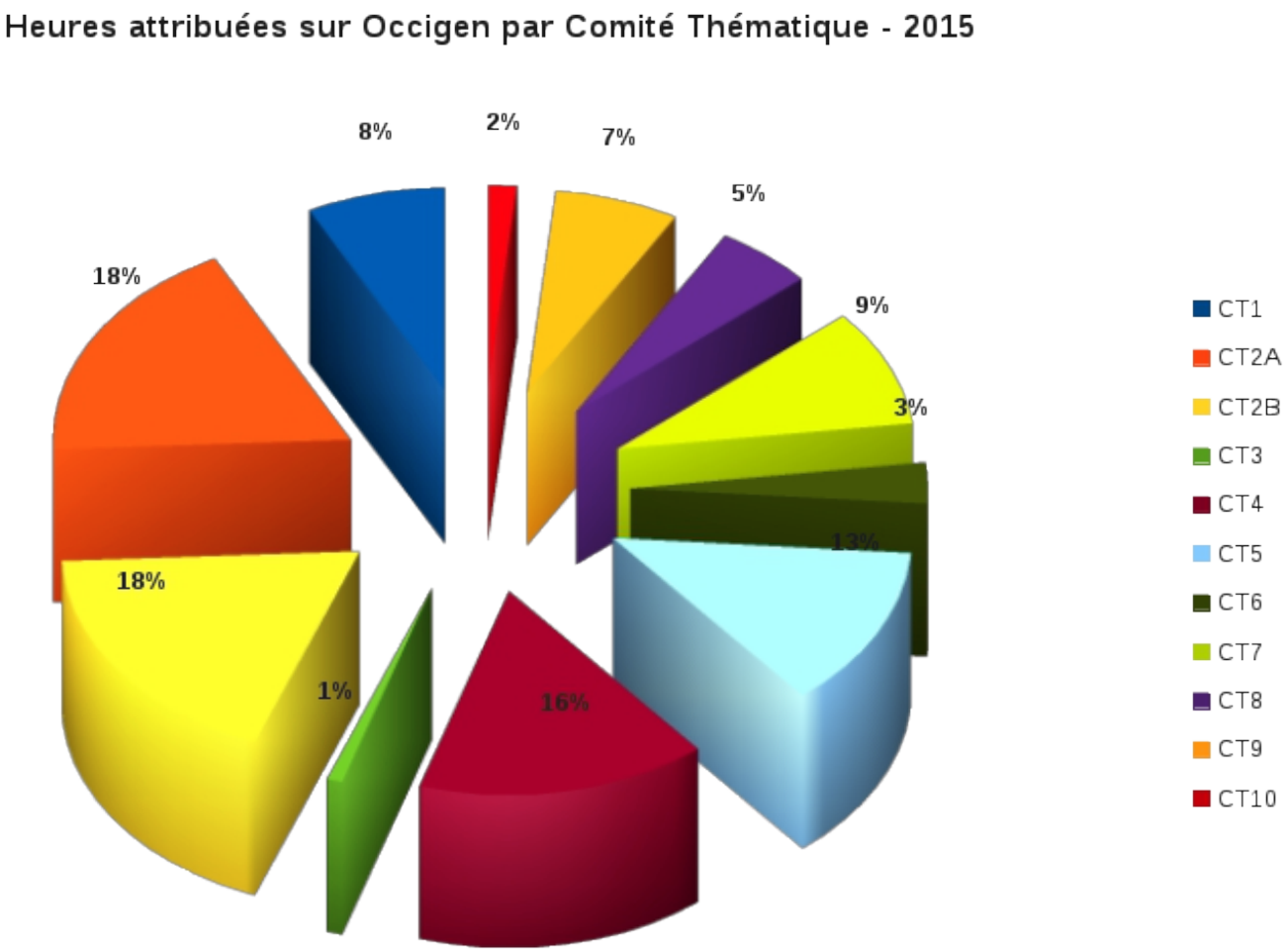




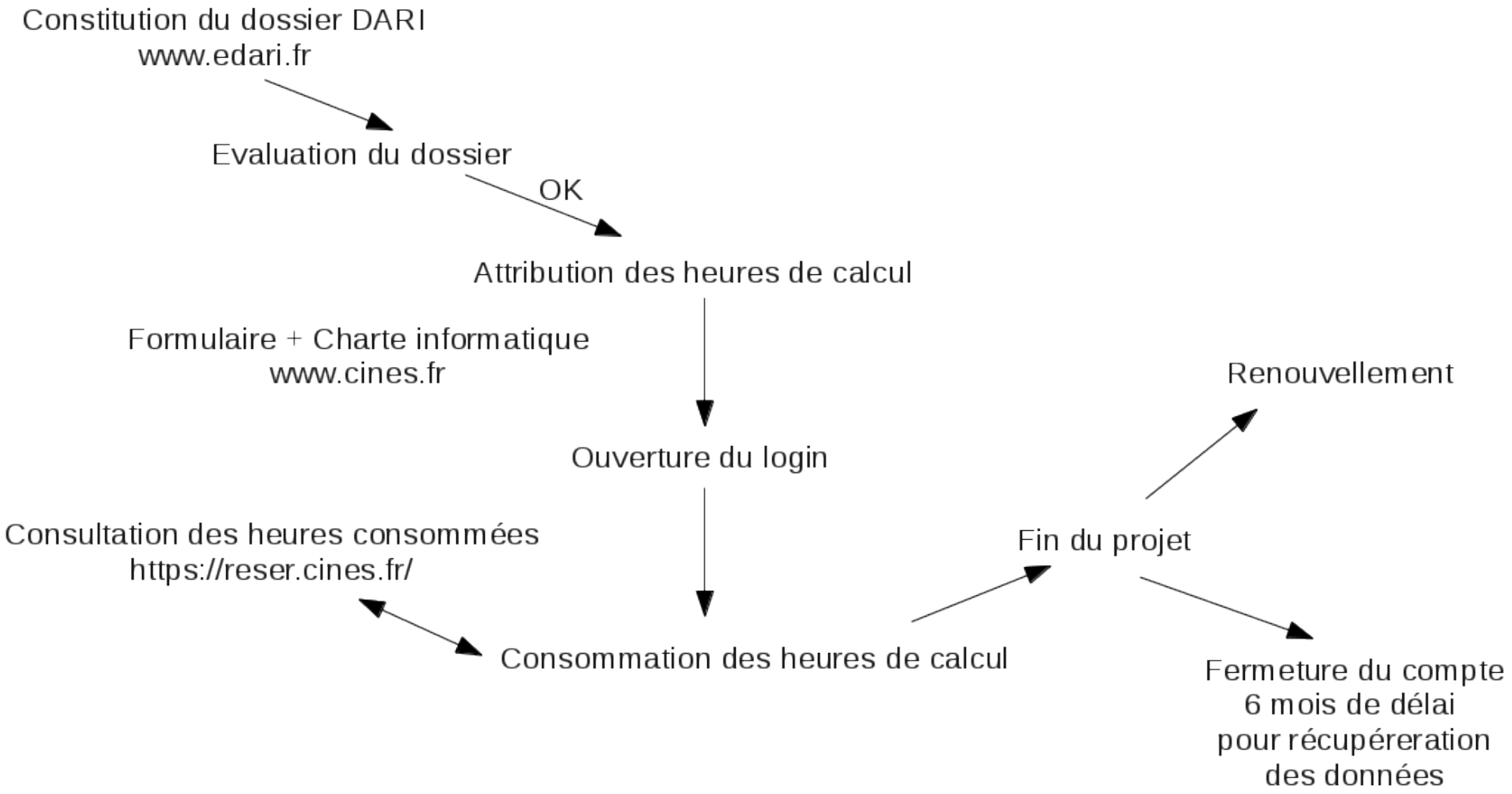
Le DARI en quelques chiffres:



Le DARI en quelques chiffres:



Déroulement d'un projet au CINES:



Les moyens associés au calcul au CINES:

- **Expertise:** 3 équipes dédiées au calcul
  - Support
  - Gestion des ressources de calcul (DARI).
  - Administration système

- **Calcul:**
  - YODA: IBM Power 7 (6.6 Tflop/s)
  - Cristal: Bull (3 Tflop/s)
  - Occigen: Bull (2.1 Pflop/s)

- **Données:** Narcisse
  - Store : système de fichier Lustre pour la conservation des données de calcul (2 Po).
  - Robots IBM avec stockage sur bande (2x 2Po).
  - Sauvegarde quotidienne.

La politique de protection des ressources de calcul:

- Procédures strictes lors de la création de comptes:
  - Nécessité d'obtenir une autorisation d'ouverture de login.
  - Charte de bon usage à signer et à respecter.
- Cloisonement:
  - Cloisonement réseau interne et externe.
  - Cloisonement des répertoires personnels.
- Protection contre l'usurpation d'identité:
  - Filtrage des accès via une IP associée à un login.
  - Interdiction du prêt de login.
  - Connexion par mot de passe uniquement.

Yoda:

- Cluster IBM P1600.
- 8 noeuds Power 7 (3.3 GHz), 32 coeurs.
- 128 Go de mémoire par noeud.
- Réseau d'interconnexion Infiniband DDR (débit 20 Gbits/s)
- Espaces disques partagés : /home et /scratch (débit ~ 3 Go/s – montage GPFS) et /store (montage NFS).

Cristal:

- Machine de pré/post traitement et de visualisation.
- 1 noeud de connexion: cristal2 (Bullx S6030).
- 1 noeud Sandy bridge: cristal3 avec 4 GPUs (Quadro 7000 – 6Go/GPU), 32 coeurs pour 1To de mémoire.
- 8 noeuds Sandy bridge: cristal4-11 avec 2 GPUs (Quadro 6000 – 6Go/GPU), 16 coeurs pour 128Go de mémoire.
- Réseau d'interconnexion Infiniband FDR (débit 56 Gbits/s).

Occigen :

- 7 noeuds de connexion.
- 2106 noeuds de calculs.

- Réseau d'interconnexion Infiniband FDR (débit 56 Gbits/s).

- Espace disque partagé:
  - /home (montage PanFS – débit 5Go/s)
  - /scratch (montage Lustre – débit aggrégé 106Go/s)
  - /store (montage Lustre – débit aggrégé 50 Go/s).

Architecture matérielle:

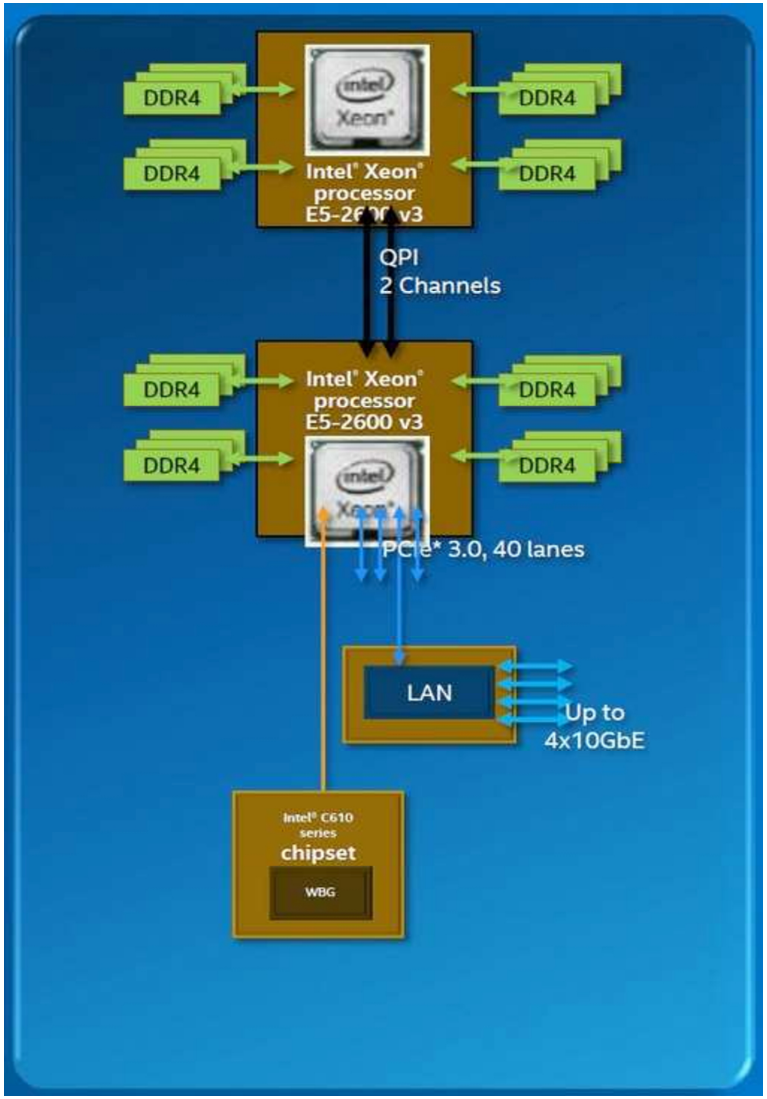
- 24 racks de calculs (2106 noeuds → 50 544 coeurs):
  - 1 rack = 5 chassis.
  - 1 chassis = 9 lames.
  - 1 lame = 2 noeuds.
  - 1 noeud = 2 processeurs.
  - 1 processeur = 12 coeurs.

Processeur Intel Haswell EP E5-2690 (comprenant 12 coeurs) | Lame B720 (comprenant 2 noeuds) | Chassis DLC (comprenant 9 lames) | Rack (comprenant 5 chassis)

- 5 racks de stockage: espace /scratch
  - Racks montés sur système de fichiers partagés Lustre
  - Accès via réseau Infiniband.
  - Capacité 5Po

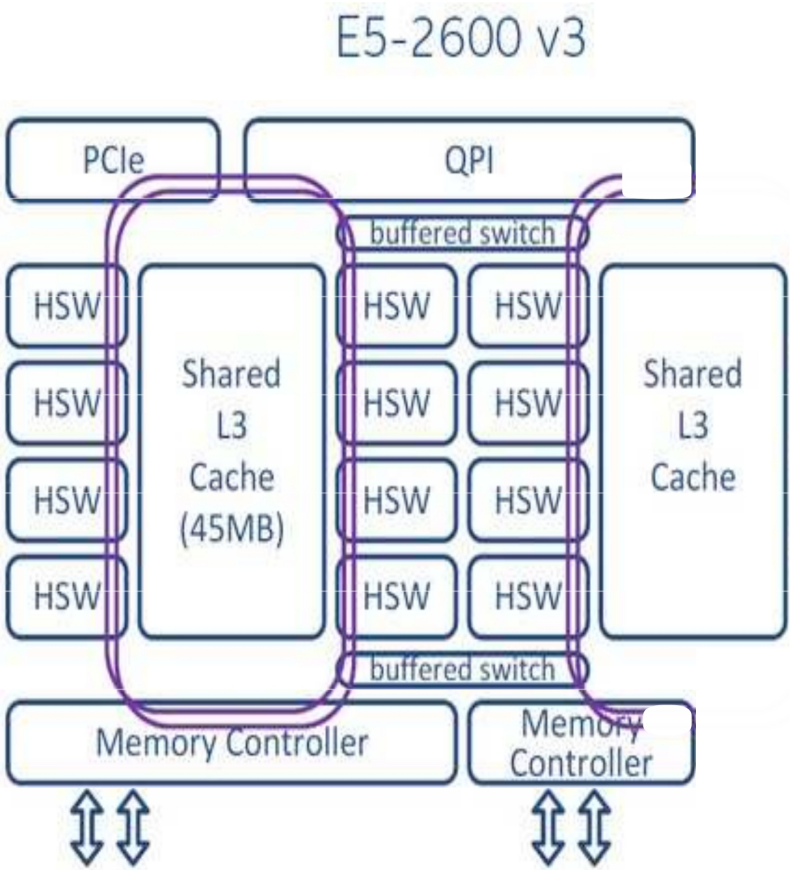
Architecture matérielle: Lame

- Lame B720:
  - 2 noeuds de calculs, composés de 2 processeurs chacun.
  - ½ noeuds disposent de 64 Go de mémoire DDR4.
  - ½ noeuds disposent de 128 Go de mémoire DDR4
- Calcul en dépeuplé possible.
- Bande passante mémoire par noeud: entre 109 et 116 Go/s.



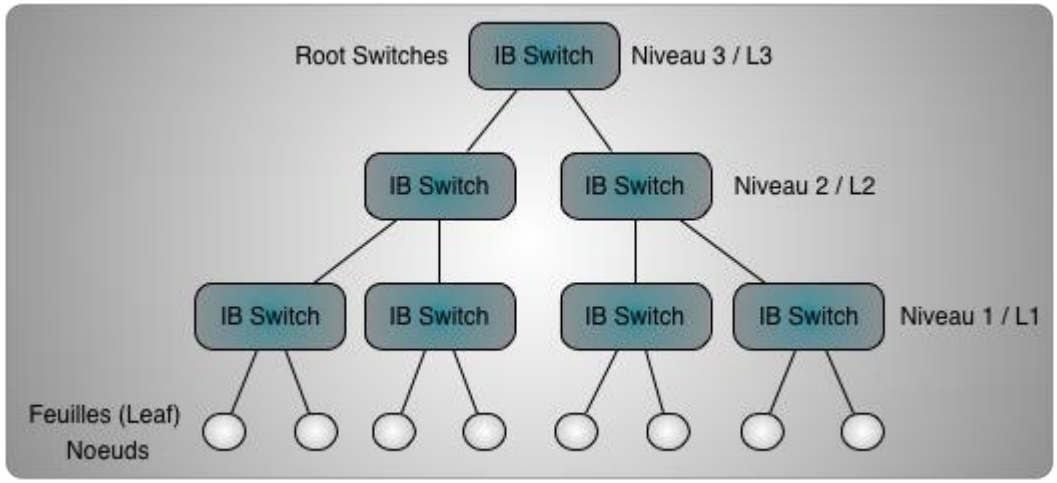
Architecture matérielle: CPU

- Intel Haswell EP v3 E5-2690:
  - 12 coeurs.
  - 2.6 GHz (500 Gflops/s)
- 3 niveaux de cache:
  - L1 : 32 Ko données
    32 Ko instructions
  - L2 : 256 Ko
  - L3 : 30 Mo partagé (2.5 Mo/coeur).
- Bande passante mémoire 68.2GB/s.



Architecture matérielle: Réseau

- Infiniband FDR
  - Débit 56 Go/s
- Topologie: Fat Tree
- Bande passante bi-directionnelle max : 4.5 Go/s.
- Latence inférieure à 3 µs



Architecture matérielle: systèmes de fichiers

- Interne:
  - /home:
    - PanFS, passerelle Ethernet 10 GB / Infiniband.
    - 240To.
    - Débit: 5.3 Go/s.
  - /scratch:
    - Lustre, Infiniband.
    - 5 Po.
    - Débit cumulé : 106 Go/s.

- Externe:
  - /store:
    - Lustre / Infiniband.
    - 2Po
    - Débit cumulé: 50 Go/s

Les espaces de données:

- 3 espaces disponibles : /home, /scratch et /store.
- Espaces personnel + un espace partagé au sein du groupe.
- Accès rapide via des variables : $SCRATCHDIR, $SHAREDHOMEDIR...

Nom	Quota	Accessibilité	Perf	Sauvegarde
/home	40 Go	Login + calcul	+	oui
/scratch	4 To	Login + calcul	+++	non
/store	100 000 files	Login + calcul	++	oui

Nom	Utilisation
/home	Trousse à outils : compilation, stockage executable, librairies nécessaires aux jobs.
/scratch	Espace de travail : lancement des jobs, écritures temporaires des fichiers de sorties des jobs.
/store	Espace archivage : stockage des résultats et données.

Accès et transfert de données:

- Accès : SSH → occigen.cines.fr
  - Windows:
    Putty, cygwin, Xming.
  - Linux:
    ssh login@occigen.cines.fr
    (option -X si besoin).

- Transferts:
  - Pour accéder à une machine externe:
    autorisation nécessaire.
  - Clients git/svn disponibles.
  - Transfert possible:
    via scp, sftp, gftp depuis machine locale.

```
Last login: Fri Dec 19 15:56:03 2014 from 195.83.179.24
********************************************************************************
*            Bienvenue au CINES  sur  OCCIGEN                                  *
*                                                                              *
* Cluster BULLx DLC : 2106 noeuds B720, 50544 coeurs,  202 To RAM              *
*                                                                              *
*  -1053 noeuds B720 : 24 coeurs Intel E5-2690 @2.6GHz,  64Go RAM              *
*  -1053 noeuds B720 : 24 coeurs Intel E5-2690 @2.6GHz, 128Go RAM              *
*  -Espace temporaire /scratch(Lustre) : 5 Po  @  ~120Go/s                     *
*                                                                              *
* Red Hat 6.4  (kernel :  2.6.32-504.1.3.el6.x86_64)                          *
*                                                                              *
********************************************************************************

*** - Votre mot de passe doit contenir 12 caracteres minimum                 ***
*** - Pour acceder a vos espaces de travail, utilisez les variables          ***
*** $HOMEDIR $SCRATCHDIR $STOREDIR $SHAREDHOMEDIR $SHAREDSCRATCHDIR          ***
***    Exemple d'utilisation : cd $SCRATCHDIR                                ***
***                                                                          ***
***  Les anciens utilisateurs de JADE disposent encore de $DATADIR           ***

12/1/2015 : Vous disposez du repertoire /store partage entre tous
            les calculateurs du CINES et aussi accessible depuis
            les noeuds de calcul.

Si vous rencontrez le moindre problème, n'hésitez pas à contacter
nos équipes support au 04 67 14 14 99 ou par mail svp@cines.fr
```

CINES Centre Informatique National de l'Enseignement Supérieur

Slurm: Fonctionnement général

- But : assurer un fort taux de remplissage de la machine et assurer un partage équitable des ressources.

- Partage équilibré des ressources via 2 mécanismes:

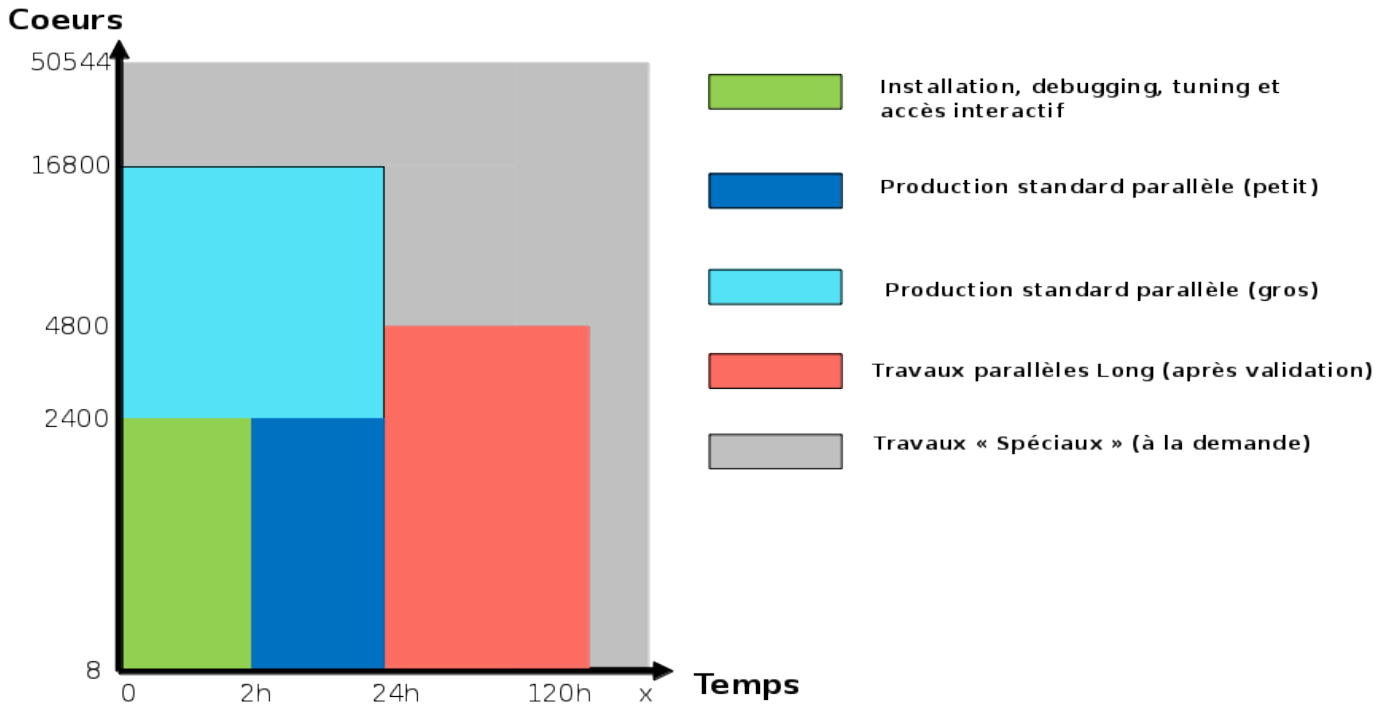
  - Fair share: participe au calcul de la priorité du job.
  - Backfill: optimise le remplissage de la machine sans impacter la priorité établie des jobs.

- Temps d'attentes impossible à prévoir.

- Durée par défaut d'un job : 24h.
  - Si checkpoint/restart/chaînage du job impossible : demande auprès de svp@cines.fr pour une durée > 24h.

Slurm: Différents types de jobs

- Améliore le partage équilibré des ressources.
- Transparent pour l'utilisateur.
- Bonne pratique: définir précisément le temps d'exécution de son job.



Slurm: différentes partitions

- Par défaut: « all » les jobs sont exécutés normalement.

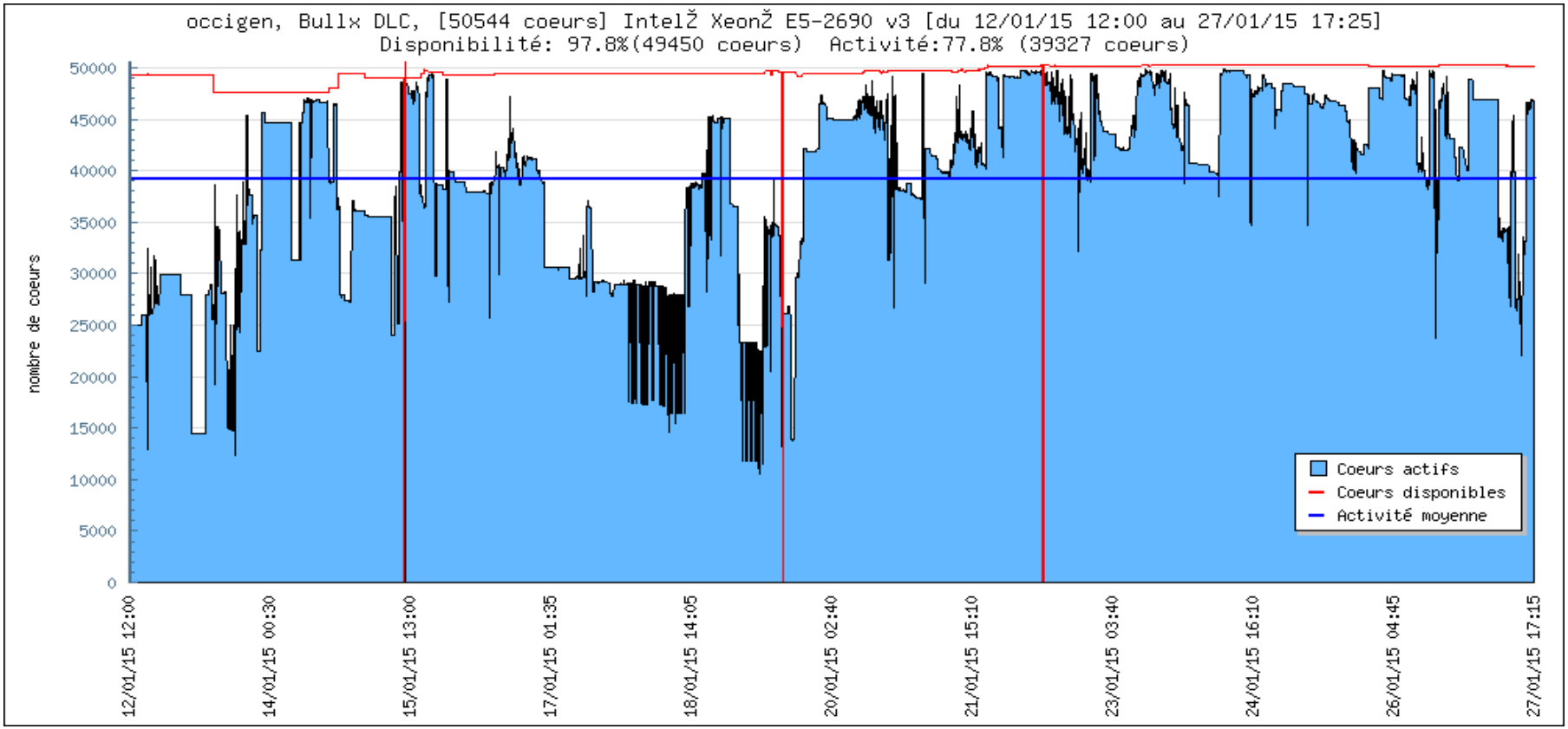
- En cas de dépassement de quota:
  - BLOCKED_SCRATCH
  - BLOCKED_WORK
  - BLOCKED_STORE

Si votre job est dans cette partition, il sera bloqué tant que vous serez en dépassement de quota sur l'un des 3 espaces disques.

- Bonne pratique: nettoyer régulièrement les fichiers inutiles, sauvegarder/rapatrier les données importantes.

Slurm: Aperçu de l'activité d'Occigen

➔ Entre le 12 Janvier 2015 et le 27 Janvier 2015:



Slurm : Exécution

➔ **Plusieurs modes de soumission:**

- Mode batch:
  sbatch script.slurm
- Mode interactif:
  salloc -N 1 -n 10 -t 10:00; srun “command”; exit.

➔ **Les commandes de suivi:**

- Suivi des jobs: squeue -u *login*
  - Etat des jobs : CA, PD, CD, R, S, F, NF
- Détail d'un job: scontrol show job *jobID*

```
[                :~$]squeue -u login
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
             59244       all  jobname    login PD       0:00     32 (Priority)
             58848       all  jobname    login PD       0:00     16 (Priority)
             58948       all  jobname    login PD       0:00     16 (Priority)
             59115       all  jobname    login PD       0:00      8 (Priority)
             59023       all  jobname    login  R    6:49:06     32 occigen[1707-1719,1738-1749,1752-1753,2044-2048]
             58797       all  jobname    login  R   21:12:13     32 occigen[1012-1013,1801-1808,1818,1823,1826,1903-1904,1909-1913,1915,2021-2025,2029,2250,2812-2815]
```

Slurm : Exécution

➔ **Les commandes d'action sur les jobs:**

- Suspendre un job de la queue: scontrol hold *jobID*
- Remettre un job dans la queue: scontrol release *jobID*
- Supprimer un job: scancel *jobID*
- Chaîner des jobs : variable
  #SBATCH - -dependency=afterok:*jobID*

➔ **Décompte des heures consommées:**

- Noeud en usage exclusif.
- Heures consommées =
  nbre heures ELAPSED x *nbre noeud* x *nbre coeurs/noeud*