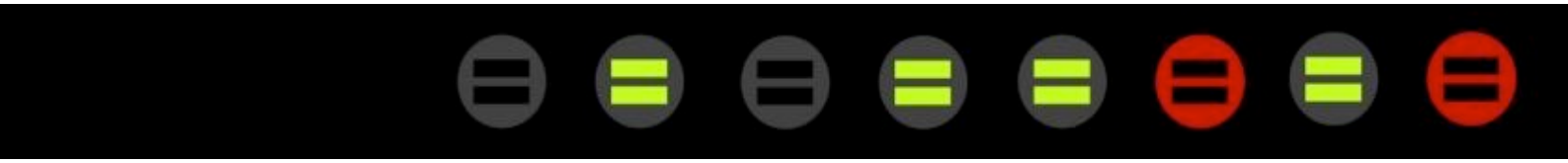
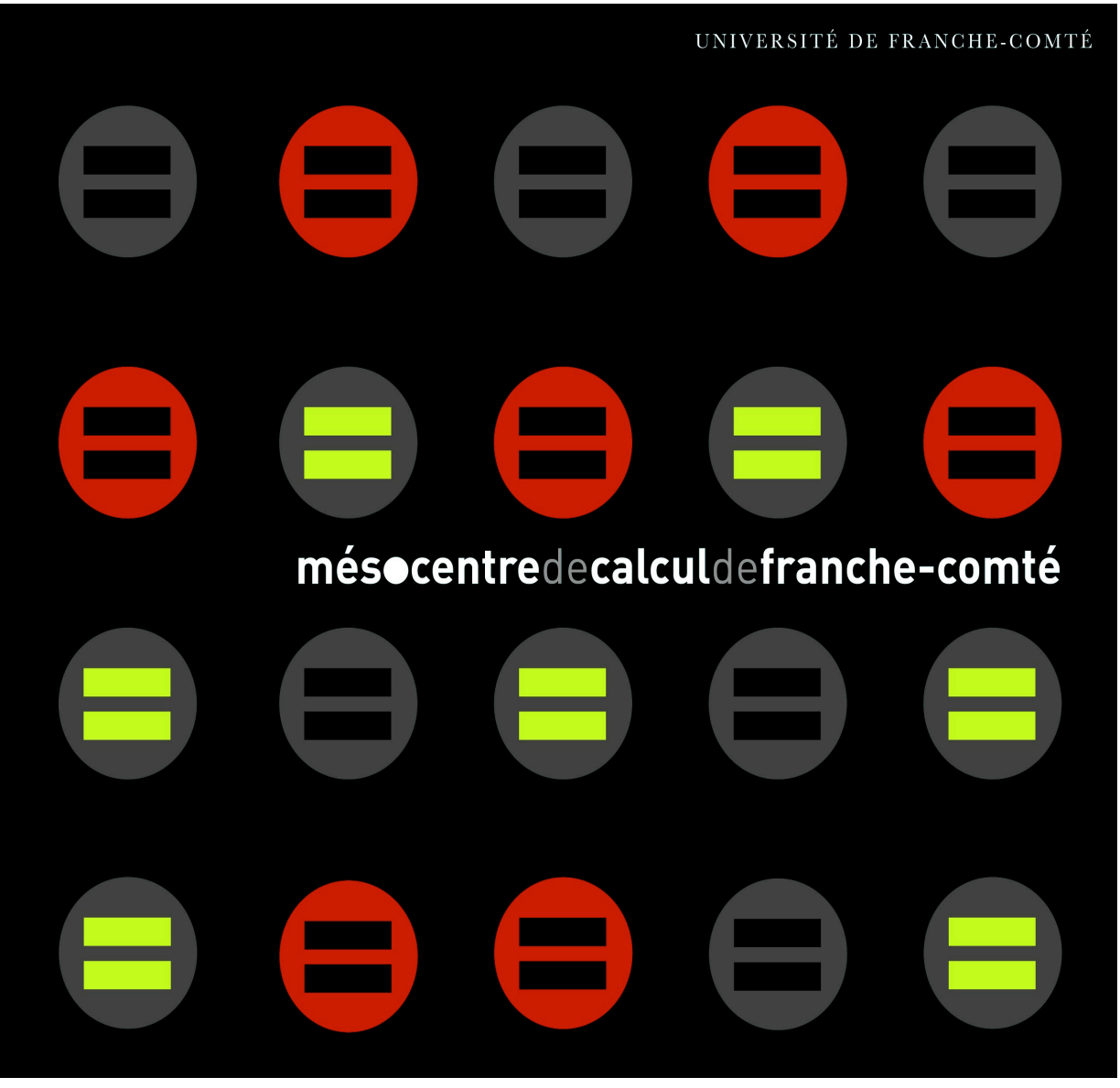
mésocentredecalculdefranche-comté





UNIVERSITÉ DE FRANCHE-COMTÉ

UFC

ensmm

utbm
université de technologie
Belfort-Montbéliard

Franche-Comté
Conseil régional

mésocentredecalculdefranche-comté

UFC
UNIVERSITÉ DE FRANCHE-COMTÉ

Expérience d'un jeune mésocentre

L. Philippe

10 juin 2010

mésocentredecalculdefranche-comté

Plan

- Situation de départ / contraintes
- Solution installée
- Utilisation actuelle
- Bilan et évolutions

Projet

- Finnancements acquis
- Pré-projet
- Pas de personnel
- Peu d'expertise
- Locaux insuffisement climatisés

UFC
UNIVERSITÉ DE FRANCHE-COMTÉ

UFC
UNIVERSITÉ DE FRANCHE-COMTÉ

utbm
université de technologie
Belfort-Montbéliard

ensmm

Franche-Comté
Conseil régional

UFC
UNIVERSITÉ DE FRANCHE-COMTÉ

Utilisateurs

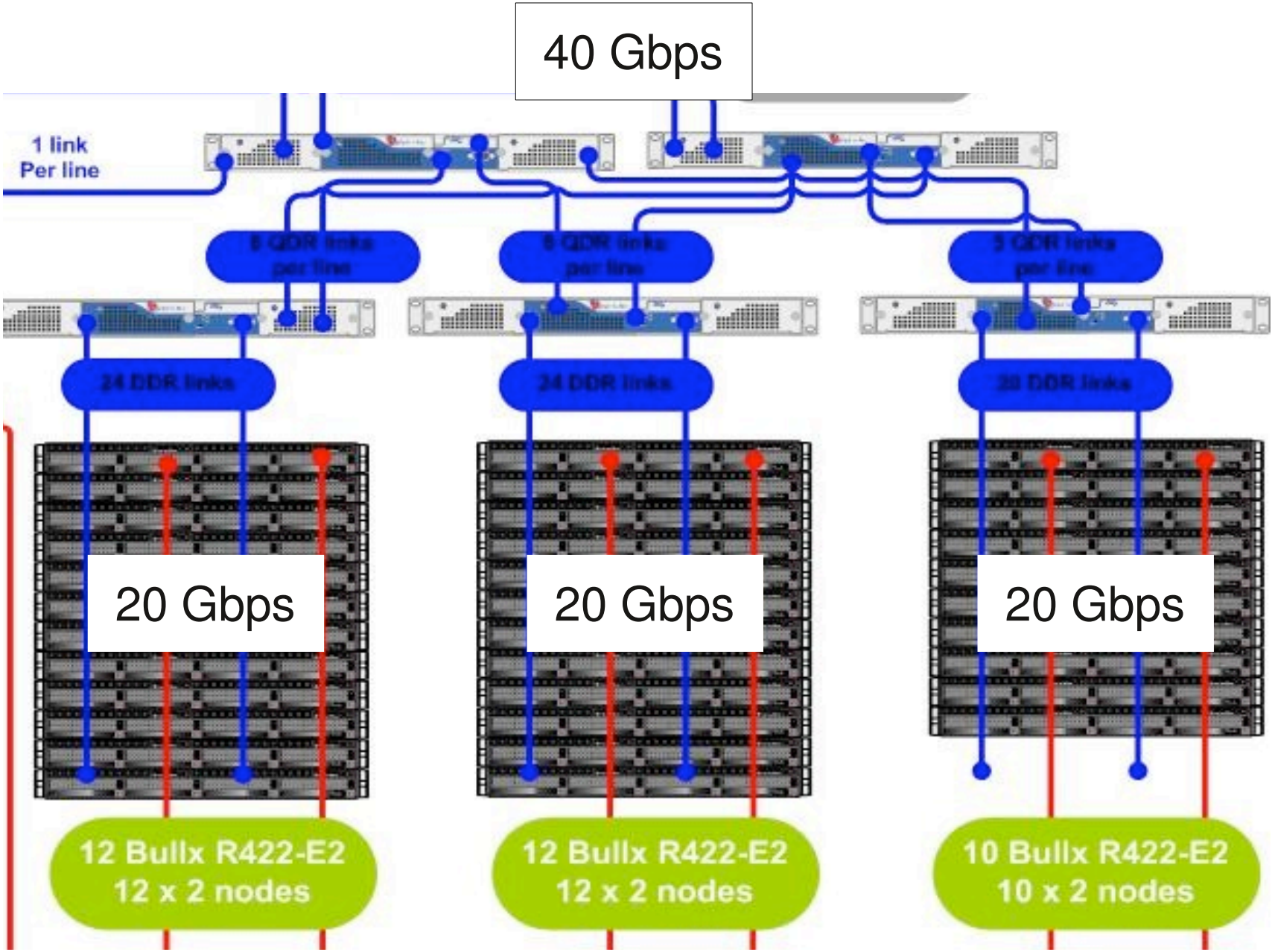
- Etablissements franc-comtois + industriels
- Quelques utilisateurs avertis :
  - Utilisation de clusters locaux
  - Centres de calcul
- Beaucoup d'utilisateurs non avertis
  - Septiques
  - Ne savent pas ce que peuvent en faire
  - Windows
- Participation à l'éllaboration du CCTP

Choix d'une solution clef en main

Cluster - Calcul

- Machine Bull
- 68 noeuds de calcul : 5,5 Tflops
  - R422-E2 : bi-proc, Xeon X5560, 2,66 Ghz
  - 544 cœurs
  - 32 avec 24 Go, 1 avec 96 Go, 35 avec 12 Go
- 1 (2) noeud GPGPU Telsa s1070 : 4 Tflops
- Réseau InfiniBand DDR / QDR :
  - Non bloquant DDR
- Non ondulés

Réseau IB


40 Gbps
1 link
Per line
24 DDR links
24 DDR links
20 Gbps
20 Gbps
20 Gbps
12 Bullx R422-E2
12 x 2 nodes
12 Bullx R422-E2
12 x 2 nodes
10 Bullx R422-E2
10 x 2 nodes

Cluster - Admin

- 5 noeuds hors files d'attentes:
  - 1 noeud de connexion
  - 2 noeuds interactifs
  - 1 noeud serveur NFS
  - 1 admin
  - Ondulés
- Baies de stockage :
  - Système de fichiers partagés NFS 20TB
  - Système de backup 10TB
- OS Bullx Cluster Suite (Base RHEL 5)
- Système de batch : SGE

UFC
UNIVERSITÉ DE FRANCHE-COMTÉ

Portes froides

- Installation
- Deux armoires climatisées
- Faux plancher
- Commandes :
  - Nagios
  - Redémarrage
- Efficace
- Evolution ?

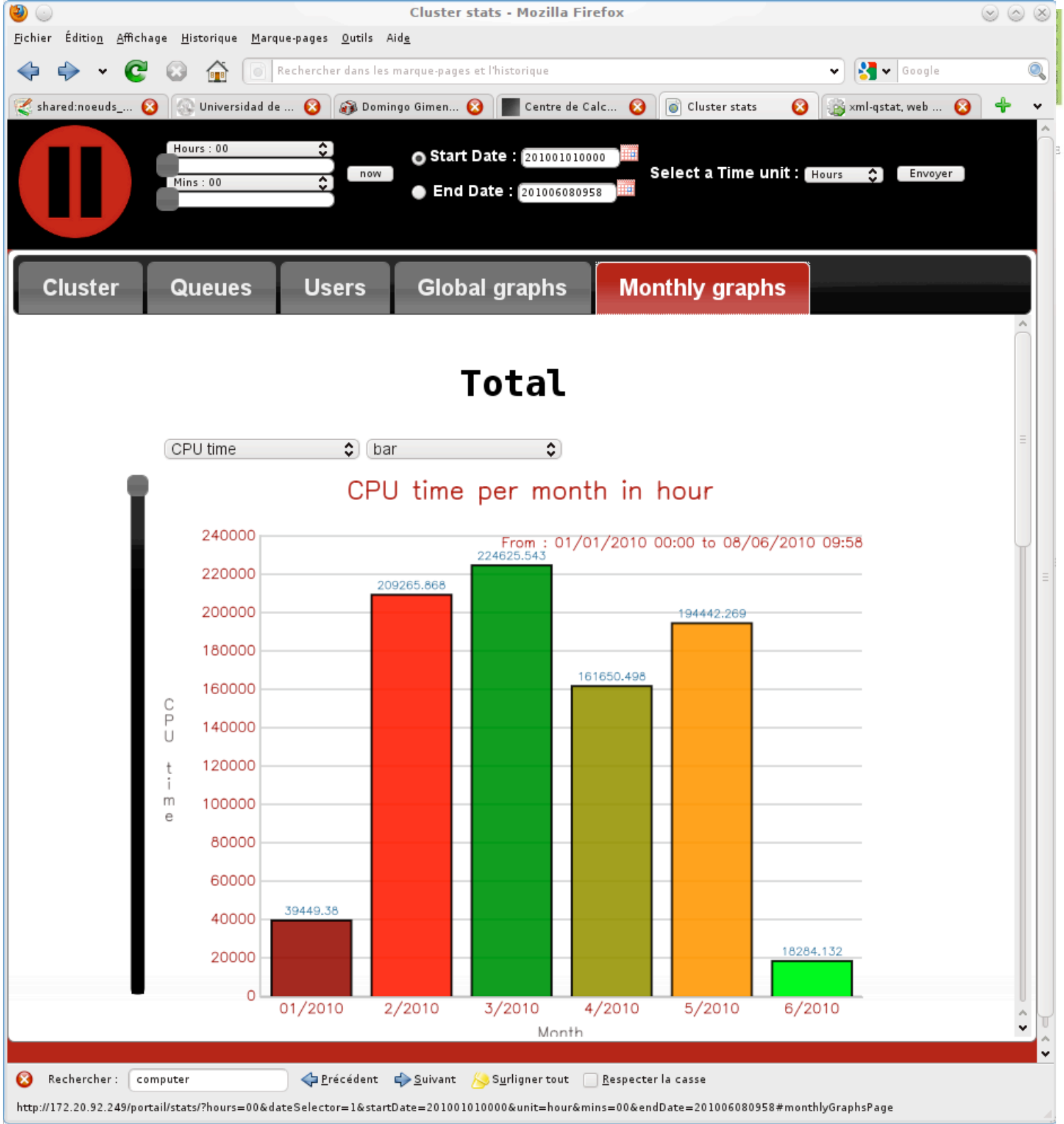
Logiciels administration

- Bull cluster suite :
  - Redhat Enterprise
  - Cluster DB : référence mais statiqu
  - BSM : Bull System Manager
    = Nagios
- Ganglia
- Xmlqstat : détails sur les jobs
- Outil maison exploite SGE

xmlqstat

mésocentredecalculdefranche-comté

FC
FRANCHE-COMTÉ

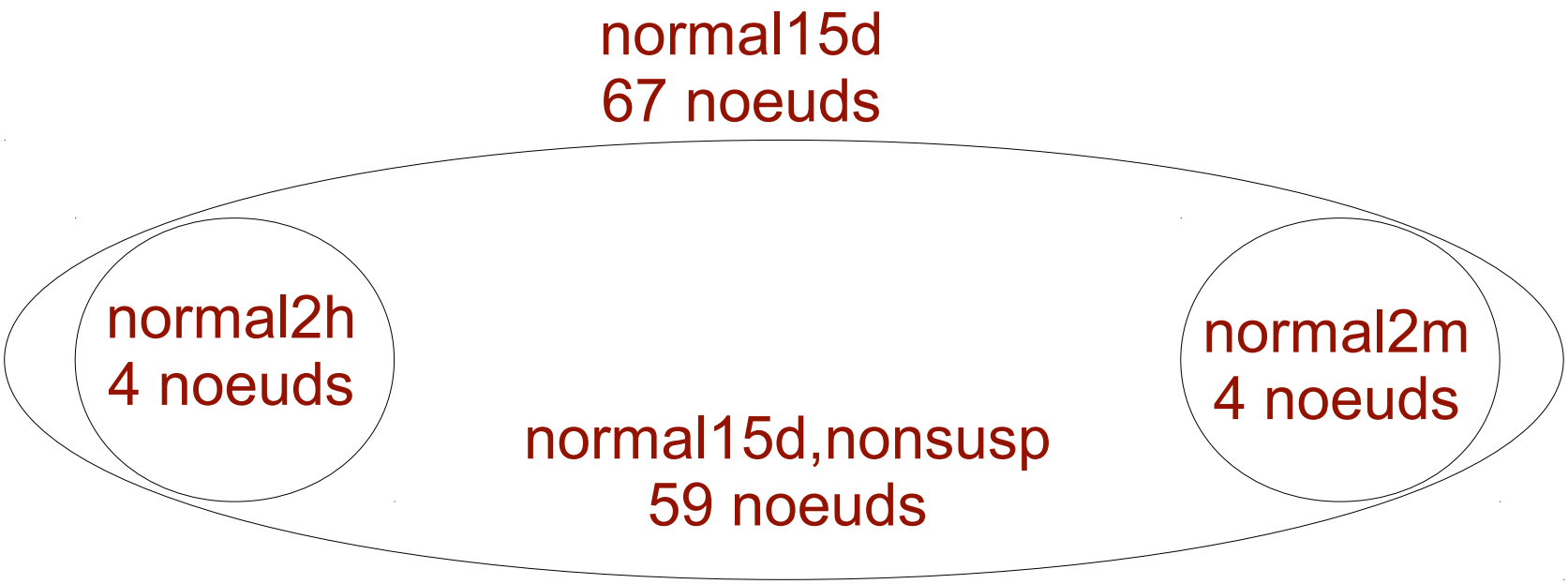


Files d'attente

- Gestionnaire SGE
- Contraintes :
  - Permettre l'exécution de jobs courts
  - Demande pour une exécution longue : 2 mois
  - Noeuds spécifiques :
    - Tesla
    - 96 Go de mémoire
  - Jobs parallèles
  - Soumissions interactives

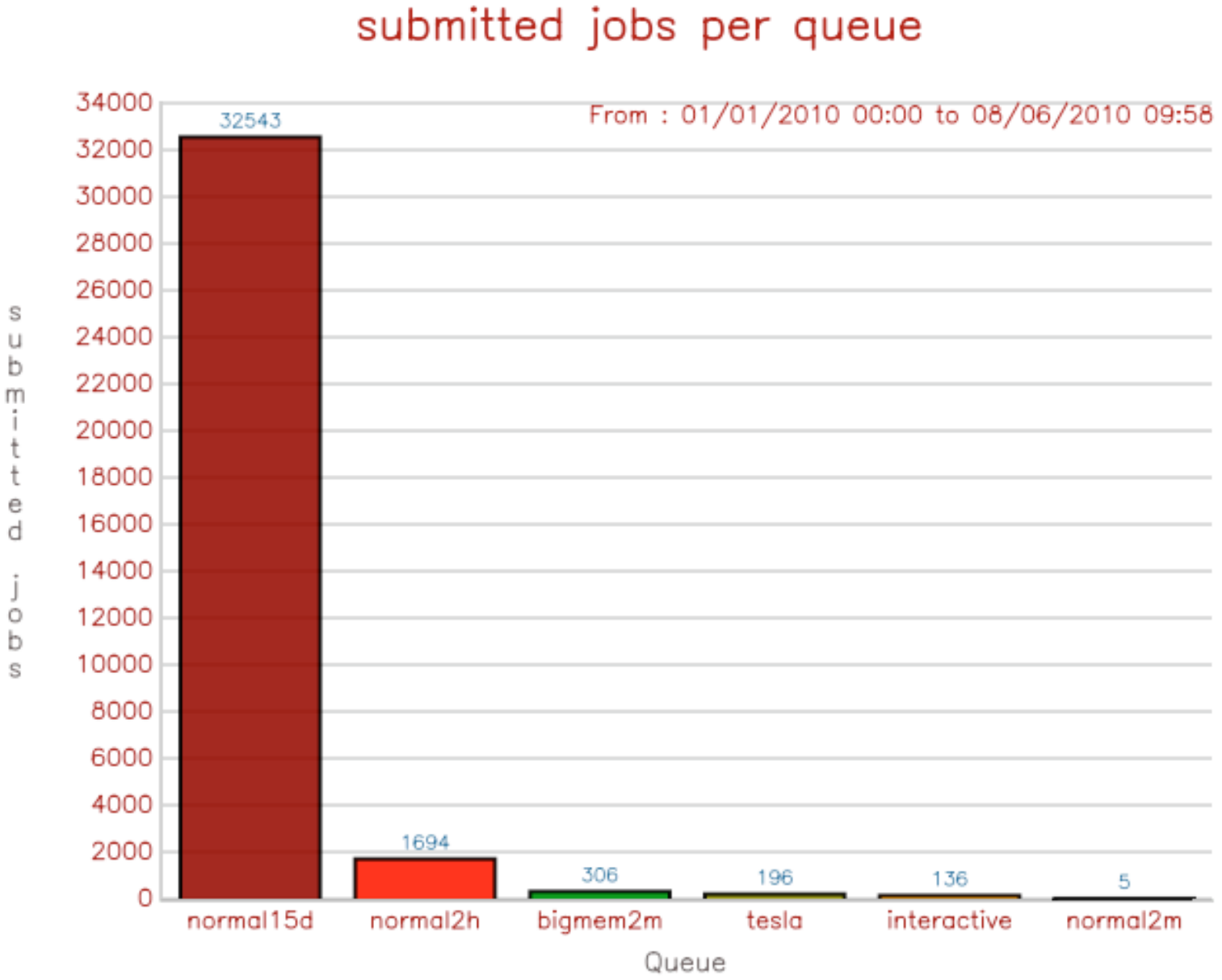
Files d'attente

- Files spécifiques : tesla, 96 Go, inter
- Files suspensives : pb jobs parallèles



- Environnements parallèles : openmp, mpi

Files d'attente



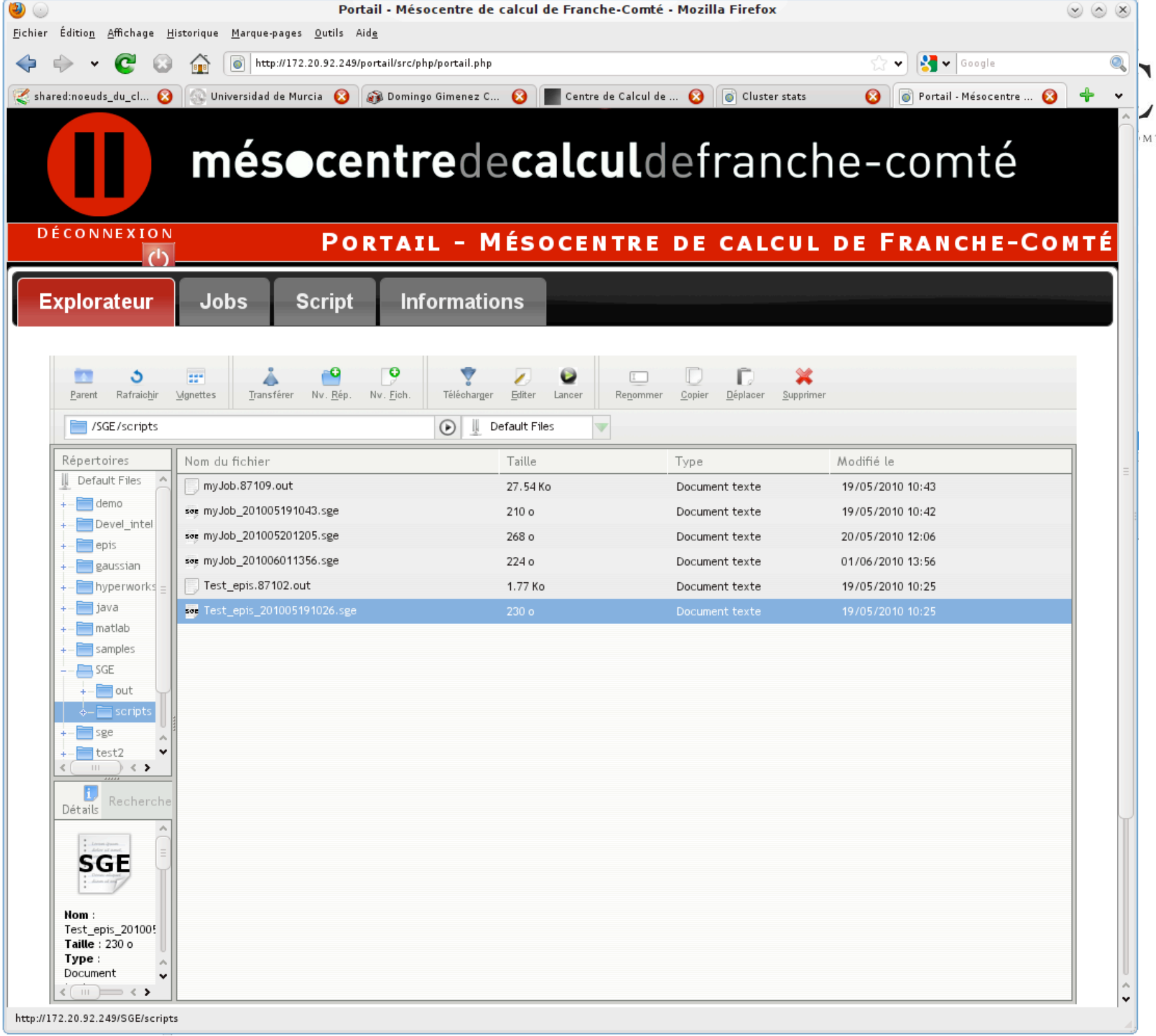
Accès

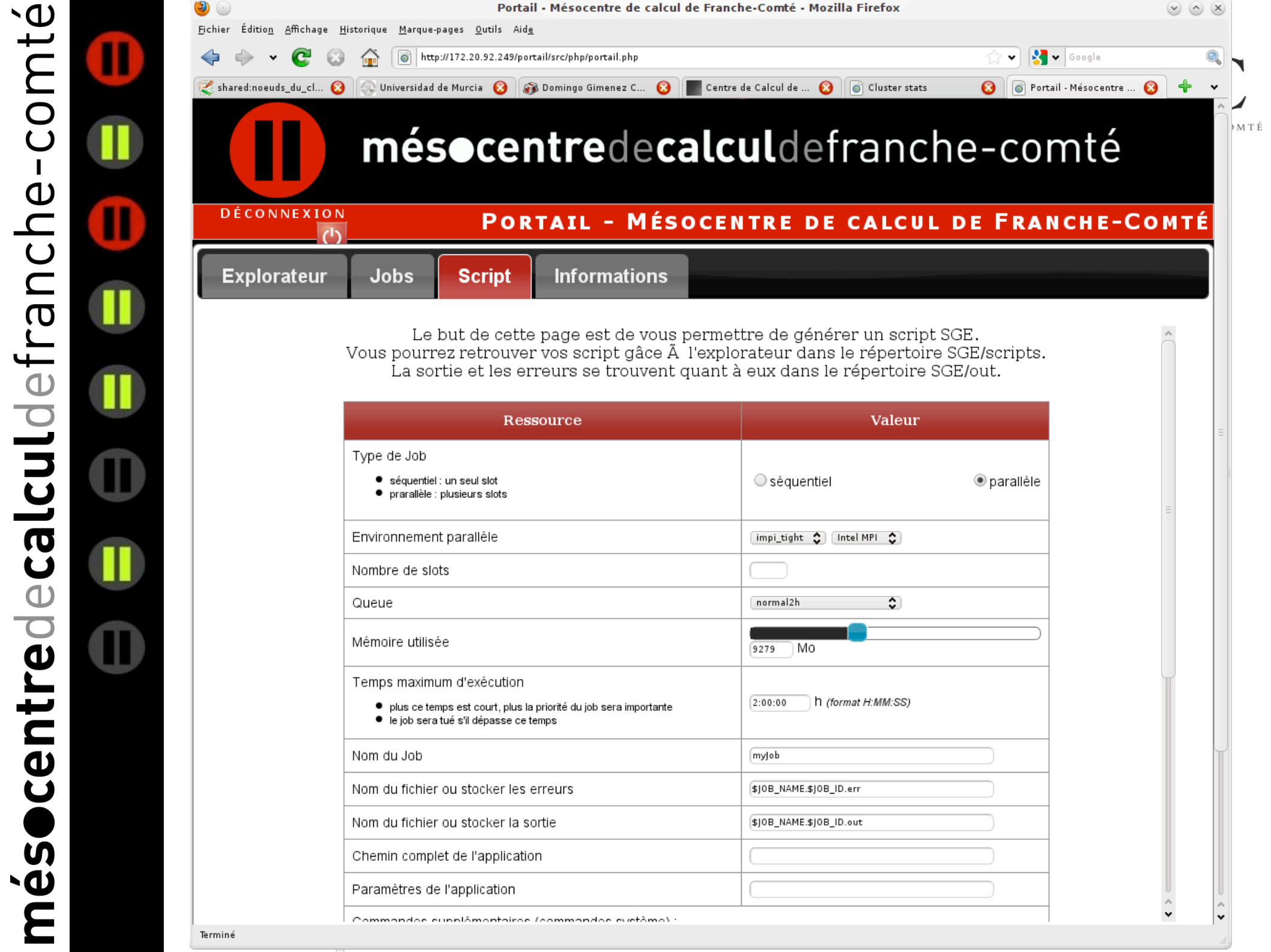
- Réseau privé université
  - SSH
  - VPN
- Annuaire LDAP
- Modules pour paramétrer environnement:
  - Logiciels
  - MPI
  - Langages

Logiciels installés

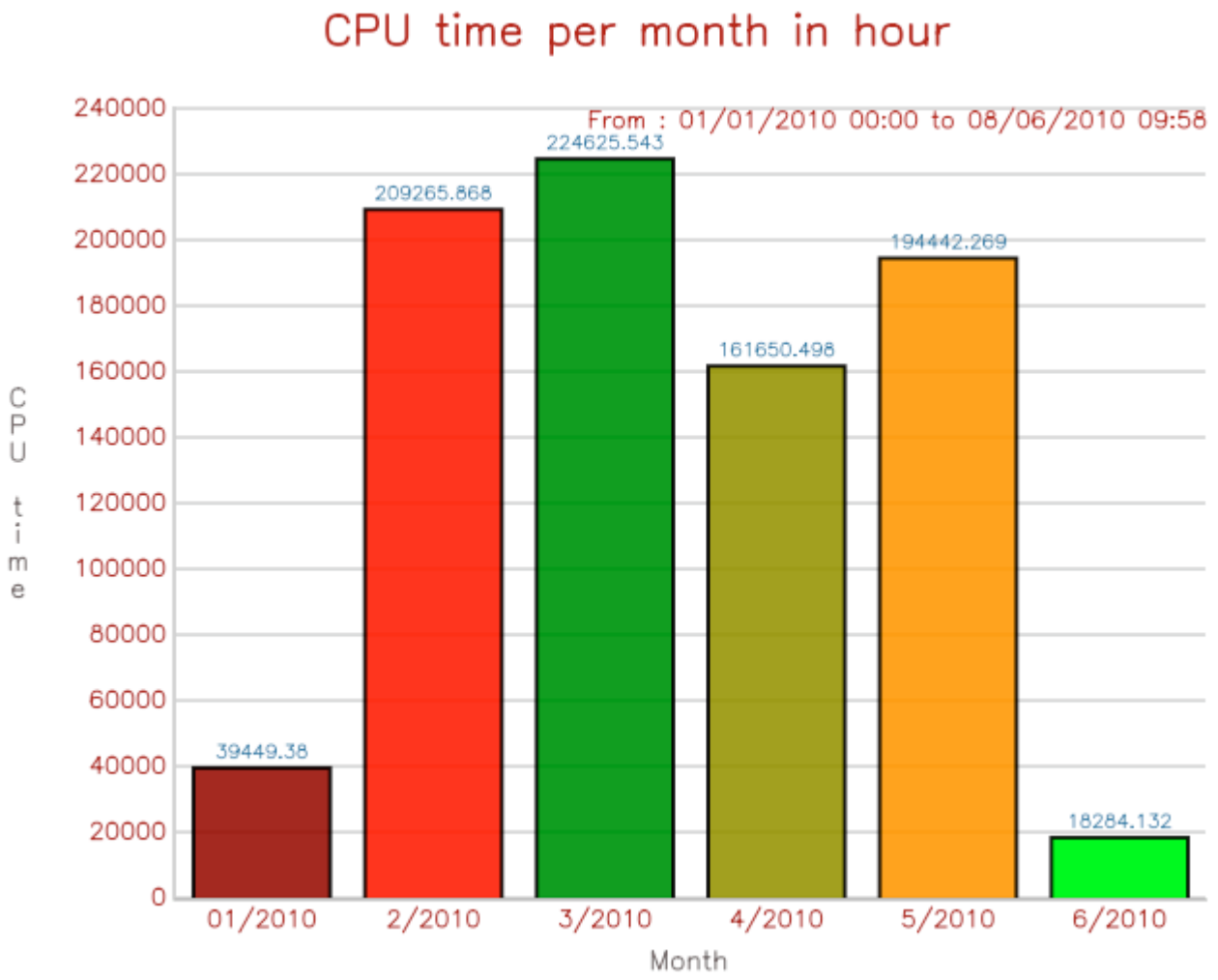
- Libres :
  - Distribution, bibliothèques (Blas, Lapack, Petsc, ...)
  - Gamess, Ab-Initio, Namd, Meep, Espresso, FFTW, R, Grace, ...
- Commerciaux :
  - Intel Cluster toolkit
  - Molpro, Gaussian, VASP, MAGMA
  - Matlab :
    - 7 toolboxes + compilateur
  - Comsol multiphysics :
    - Structural mechanics

mésocentredecalculdefranche-comté

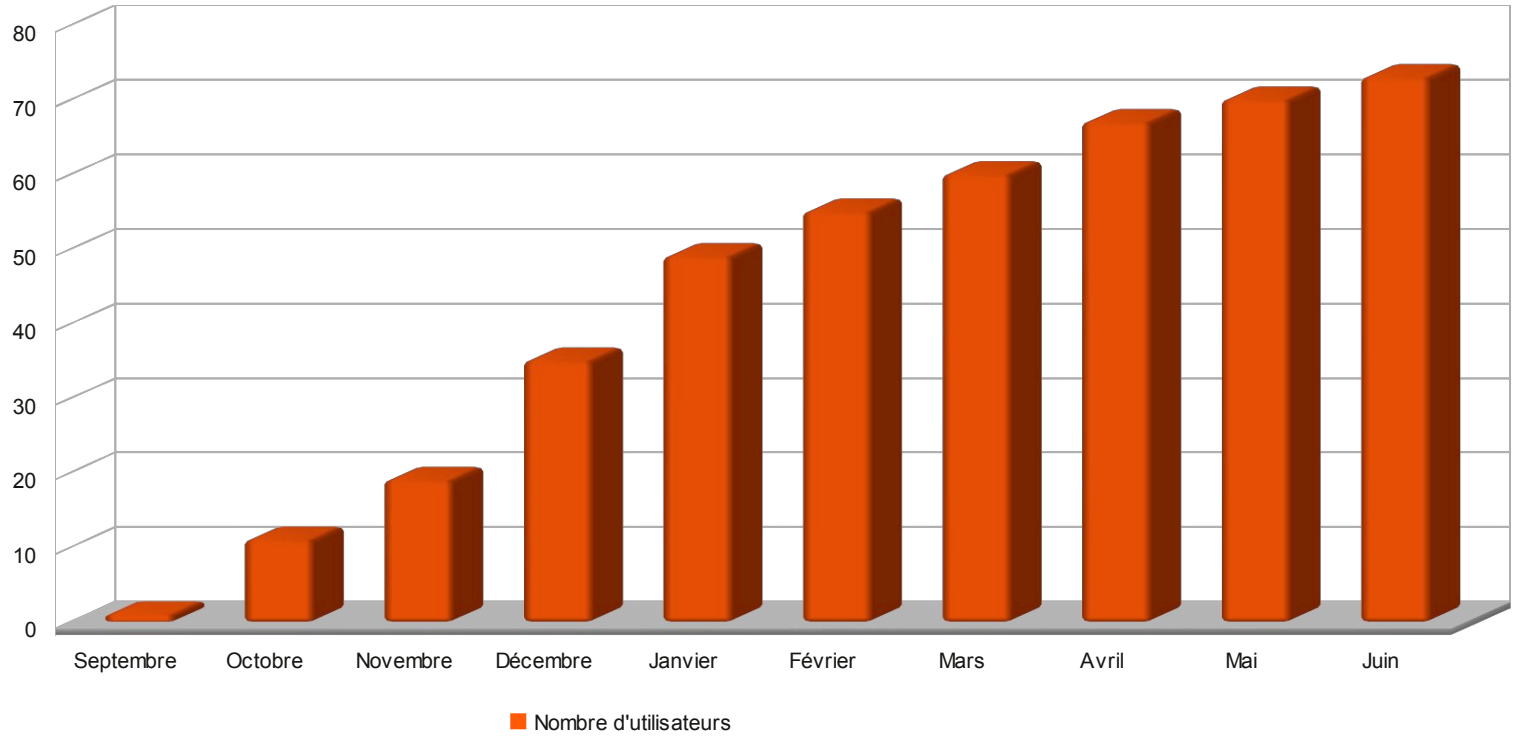



Portail - Mésocentre de calcul de Franche-Comté - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils Aide
http://172.20.92.249/portail/src/php/portail.php
mésocentredecalculdefranche-comté
DÉCONNEXION
PORTAIL - MÉSOCENTRE DE CALCUL DE FRANCHE-COMTÉ
Explorateur
Jobs
Script
Informations
Le but de cette page est de vous permettre de générer un script SGE.
Vous pourrez retrouver vos script gâce Ã l'explorateur dans le répertoire SGE/scripts.
La sortie et les erreurs se trouvent quant à eux dans le répertoire SGE/out.
Ressource
Valeur
Type de Job
séquentiel : un seul slot
prarallèle : plusieurs slots
séquentiel
parallèle
Environnement parallèle
impi_tight
Intel MPI
Nombre de slots
Queue
normal2h
Mémoire utilisée
9279
Mo
Temps maximum d'exécution
plus ce temps est court, plus la priorité du job sera importante
le job sera tué s'il dépasse ce temps
2:00:00
h (format H:MM:SS)
Nom du Job
myJob
Nom du fichier ou stocker les erreurs
$JOB_NAME.$JOB_ID.err
Nom du fichier ou stocker la sortie
$JOB_NAME.$JOB_ID.out
Chemin complet de l'application
Paramètres de l'application
Terminé

Utilisation du cluster



Evolution du nombre d'utilisateurs



- Environ 30 réels

Problèmes

- Matériels :
  - Panne courant
  - Composants: disques, mémoires, carte mère
  - Mise à jour SGE
- MPI :
  - 3 versions : Bull MPI, Intel MPI, OpenMPI
  - Problèmes :
    - Orphan daemon : fin d'exécution (Namd, VASP)
    - Recompilation des codes avec OpenMPI
- Fichiers Windows = dos2unix

Bilan

- Fonctionnel rapidement
- Nombre d'heures
- Nombre d'utilisateurs :
  - Croissance rapide
  - Utilisation des anciens clusters
  - Contacts industriels
- Formation :
  - 5 jours : Optimisation, OpenMP, MPI
  - 1 à 2 jours : sujet spécifique
- Financements ?

Evolutions

- Très vite confronté aux problèmes d'évolution
- Réseau :
  - Ajouts limités : 72 noeuds
  - Baisser la bande passante:
    - Facteur bloquage ½ = 90 noeuds
    - Sous-clusters = hétérogénéité
    - Adaptation des files ?
  - Achat nouveau switch :
    - 3 x 36 = 180 noeuds
    - 96 ports = 216 noeuds

Evolutions 2

- Matériel :
  - Noeud interactif 32 coeurs
  - Ajout de noeuds de calcul
  - Espace de stockage 40 To
- Logiciels :
  - Système : changer de distribution
  - Tests :
    - Machines virtuelles
    - Wine