Cluster au SIO

ALBERT SHIH[1]

[1]Observatoire de Paris - Meudon

21 février 2008

Type de «machines» de calcul

Mémoire partagée

- Tous les processeurs accèdent à toute la mémoire avec un même espace d'adressage.
- Deux catégories :
  - UMA = Uniform Memory Access : les machines SMP.
  - NUMA = None Uniform Memory Access : plusieurs machines SMP interconnectées.

Type de «machines» de calcul

Mémoire partagée

- Tous les processeurs accèdent à toute la mémoire avec un même espace d'adressage.
- Deux catégories :
  - UMA = Uniform Memory Access : les machines SMP.
  - NUMA = None Uniform Memory Access : plusieurs machines SMP interconnectées.

Avantages

- Espace d'adressage unique : facilité de programmation.

Type de «machines» de calcul

Mémoire partagée

- Tous les processeurs accèdent à toute la mémoire avec un même espace d'adressage.
- Deux catégories :
  - UMA = Uniform Memory Access : les machines SMP.
  - NUMA = None Uniform Memory Access : plusieurs machines SMP interconnectées.

Avantages

- Espace d'adressage unique : facilité de programmation.

Désavantages

- Manque de *scalabilité*, vitesse du bus mémoire.
- Très coûteux quand on augmente le nombre de CPU.

Type de «machines» de calcul

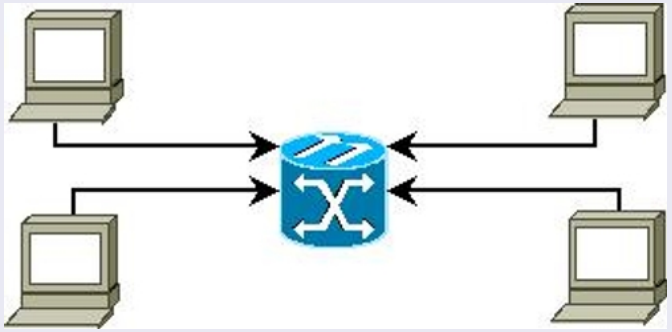
Mémoire distribuée

- Un espace mémoire différent est associé à chaque processeur.
- L'accès à la mémoire d'un autre processeur se fait via un réseau d'interconnexion.

Type de «machines» de calcul

Mémoire distribuée

- Un espace mémoire différent est associé à chaque processeur.
- L'accès à la mémoire d'un autre processeur se fait via un réseau d'interconnexion.



Type de «machines» de calcul

Mémoire distribuée

- Chaque processeur a sa propre mémoire locale, pas de notion d'adressage globale.
- Les processeurs opèrent indépendamment les uns des autres.
- Si un processeurs a besoin d'une donnée dans la mémoire d'un autre processeurs, à charge du programmeur de définir explicitement la communication par envoi/réception de message.
- Les reseaux d'interconnexion sont divers, avec des niveaux de performances très variables.

Mémoire distribuée

Avantages

- Bonne scalabilité.
- Accés rapide à la mémoire sur chaque processeur.
- Coût très inférieur.

Mémoire distribuée

Avantages

- Bonne scalabilité.
- Accés rapide à la mémoire sur chaque processeur.
- Coût très inférieur.

Désavantages

- Complexification importante de la programmation.
- Gestion et utilisation plus complexe.

Systèmes hybrides

Systèmes hybrides

- Chaque bloc de base (nœud) est une machine SMP avec 4,8 16 etc. processeurs
- Connexion des nœuds par un réseau (type Ethernet, infiniband, myrinet, etc.)
- En général désigné sous le nom : cluster.

Systèmes hybrides

Systèmes hybrides

- Chaque bloc de base (nœud) est une machine SMP avec 4,8 16 etc. processeurs
- Connexion des nœuds par un réseau (type Ethernet, infiniband, myrinet, etc.)
- En général désigné sous le nom : cluster.

Exemples

TOP 500

- BlueGene/L : 106496 coeurs de calcul (26624 nœuds) : 596 Tflops
- IDRIS (2008) : 40480 coeurs de calcul (10120 nœuds) : 139 Tflops

Impact sur utilisation

Soumission de job

- Impossible de gérer manuellement.
- Necessite un outil de soumission de job.

Impact sur utilisation

Soumission de job

- Impossible de gérer manuellement.
- Necessite un outil de soumission de job.

Partage de fichiers

- Impossible de gérer manuellement.
- Necessite un outil partage de fichiers
- Difficulté avec NFS quand le nombre de nœud augmente.

Calcul au SIO

En début 2007

- Un AlphaServer GS 1280 (Mémoire partagé type NUMA).
- 5 machines quadri-proc simple-cœur et 4 quadri-proc dual-cœurs. Système non homogène.

Calcul au SIO

En début 2007

- Un AlphaServer GS 1280 (Mémoire partagé type NUMA).
- 5 machines quadri-proc simple-cœur et 4 quadri-proc dual-cœurs. Système non homogène.

Choix

- Abandons du AlphaServer → revente de l'AlphaServer.
- Choix de la technologie X86_64 (AMD et Intel).

Calcul au SIO

En début 2007

- Un AlphaServer GS 1280 (Mémoire partagé type NUMA).
- 5 machines quadri-proc simple-cœur et 4 quadri-proc dual-cœurs. Système non homogène.

Choix

- Abandons du AlphaServer $\rightarrow$ revente de l'AlphaServer.
- Choix de la technologie X86_64 (AMD et Intel).

Motivations

- Coût (achat, maintenance et infrastructure).
- Homogénéité.

Situation du cluster actuel

Machines SIO

- Groupe de 15 serveurs.
- 5 quadri-proc mono-cœur (06/2003).
- 3 quadri-proc dual-cœurs (12/2006).
- 7 bi-proc quad-cœurs (12/2007) ← revente GS1280.
- Total : 100 cœurs de calcul.

Situation du cluster actuel

Machines SIO

- Groupe de 15 serveurs.
- 5 quadri-proc mono-cœur (06/2003).
- 3 quadri-proc dual-cœurs (12/2006).
- 7 bi-proc quad-cœurs (12/2007) ← revente GS1280.
- Total : 100 cœurs de calcul.

Second groupe

- Groupe de 11 serveurs bi-proc quad-cœurs.
- Total : 88 cœurs de calcul.

Situation du cluster actuel

Machines SIO

- Groupe de 15 serveurs.
- 5 quadri-proc mono-cœur (06/2003).
- 3 quadri-proc dual-cœurs (12/2006).
- 7 bi-proc quad-cœurs (12/2007) ← revente GS1280.
- Total : 100 cœurs de calcul.

Second groupe

- Groupe de 11 serveurs bi-proc quad-cœurs.
- Total : 88 cœurs de calcul.

Important

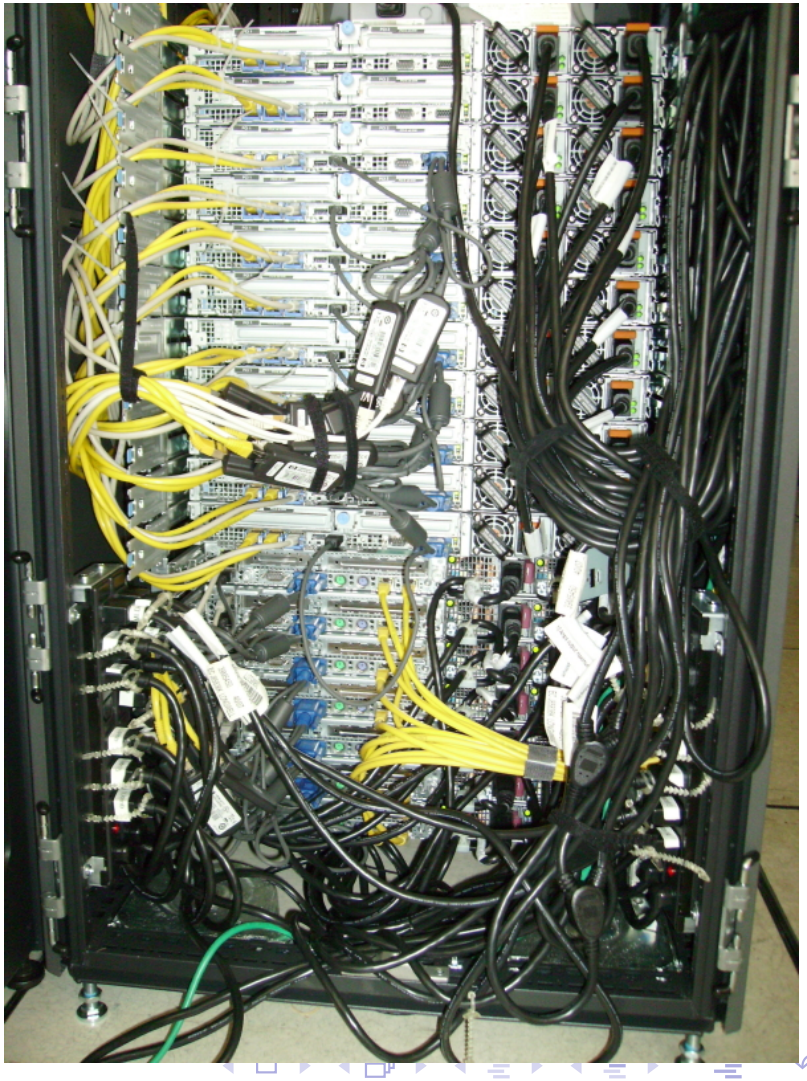
- Le cluster → à voir comme **une** machine.

Quelques chiffres

Sur l'ensemble des machines

- 3 PDU.
- 81 ports réseaux.
- $\approx$ 110 cables réseaux, $\approx$ 450m cables réseaux.
- 60 cables d'alimentation électrique.
- 416 Go de Ram.
- $\approx$ 1.5Tflops.
- $\approx$ 10Kw de consommation électrique.
- $\approx$ 600 kg de matériel.

Photos



Soumission de job

Torque/Maui

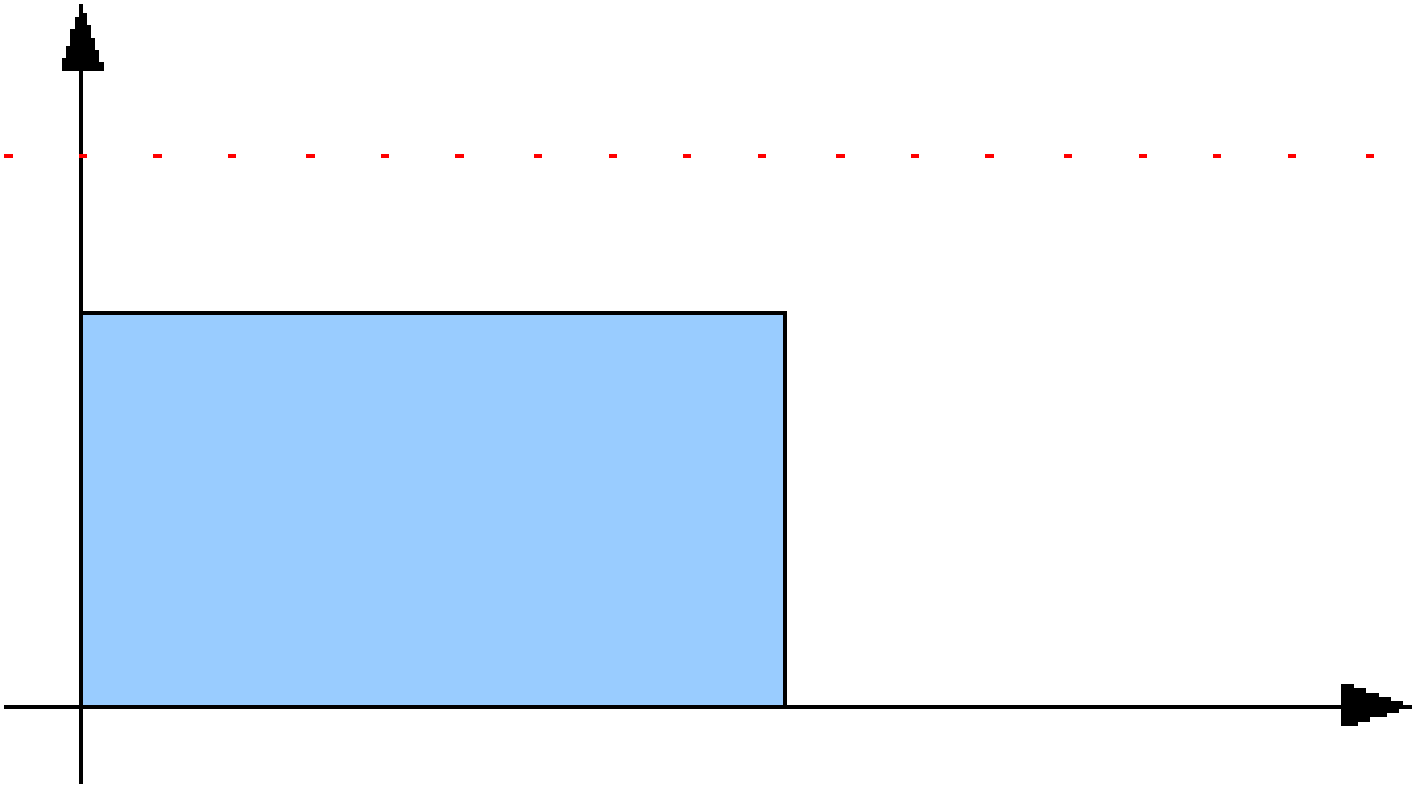
- Utilisation historique à l'Obs de PBS.
- Choix du successeur de PBS : Torque/Maui
- Système complexe, très configurable.
- Necessite que **tout** le monde passe par le système de soumission.
- Pour une utilisation optimale fournir les renseignements (temps, CPU etc. . .) le plus précis possible.

Soumission de job

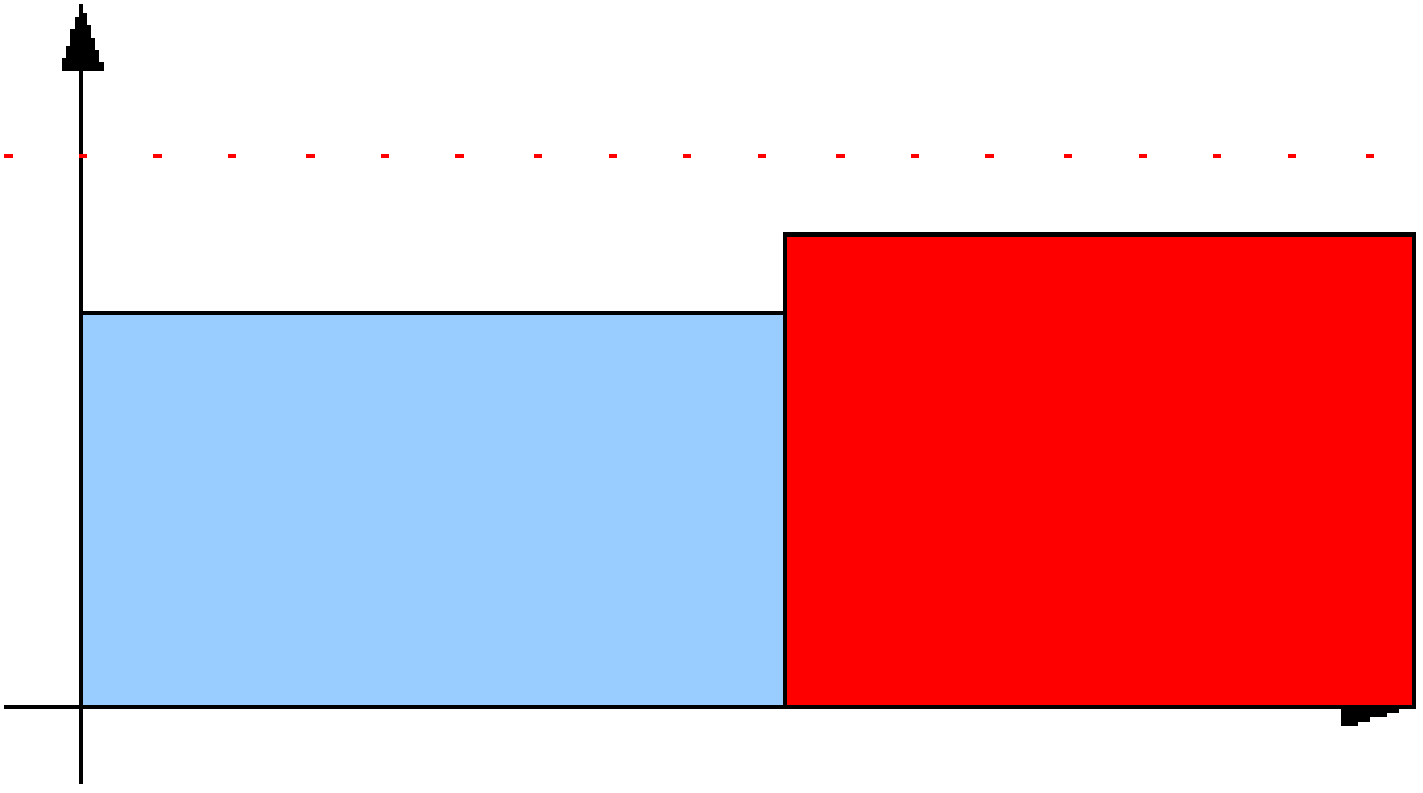
Pourquoi les informations sont importantes ?

- Gestion de processus (ordonnanceur/*scheduler*) $\longrightarrow$ rangement.
- Problème complexe.
- Impossibilité (ou mauvais *rangement*) si les informations ne sont pas correctes/absentes

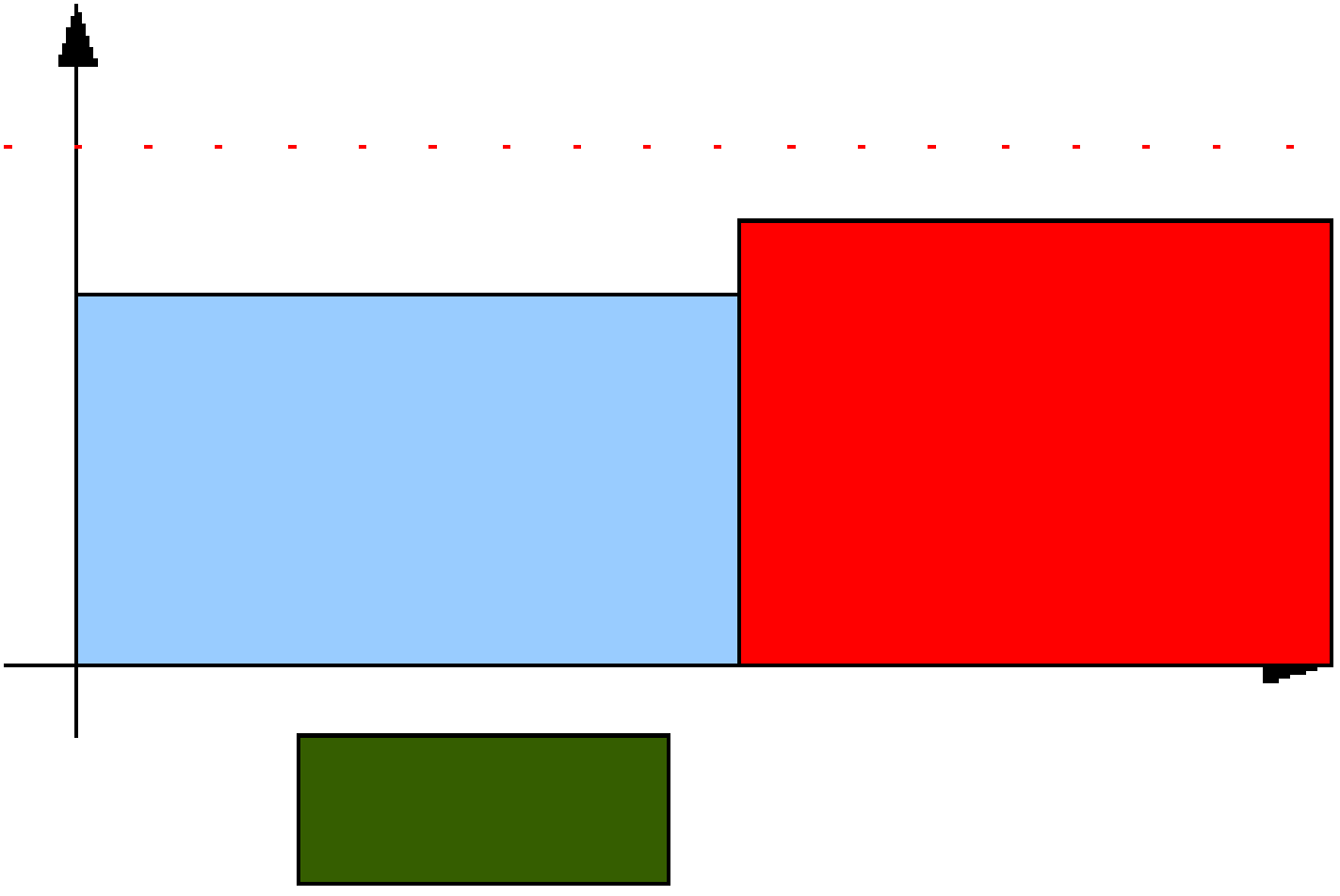
Soumission de job



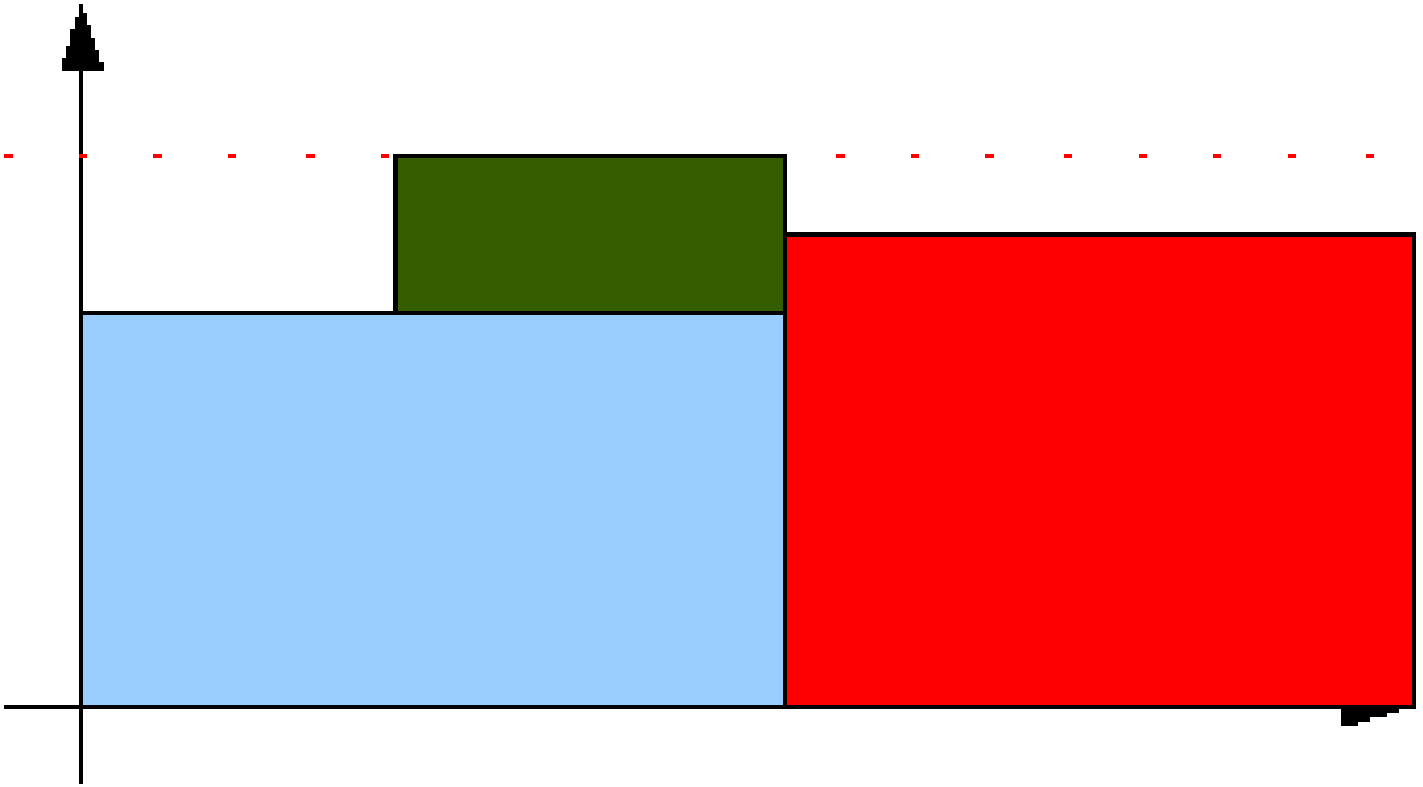
Soumission de job



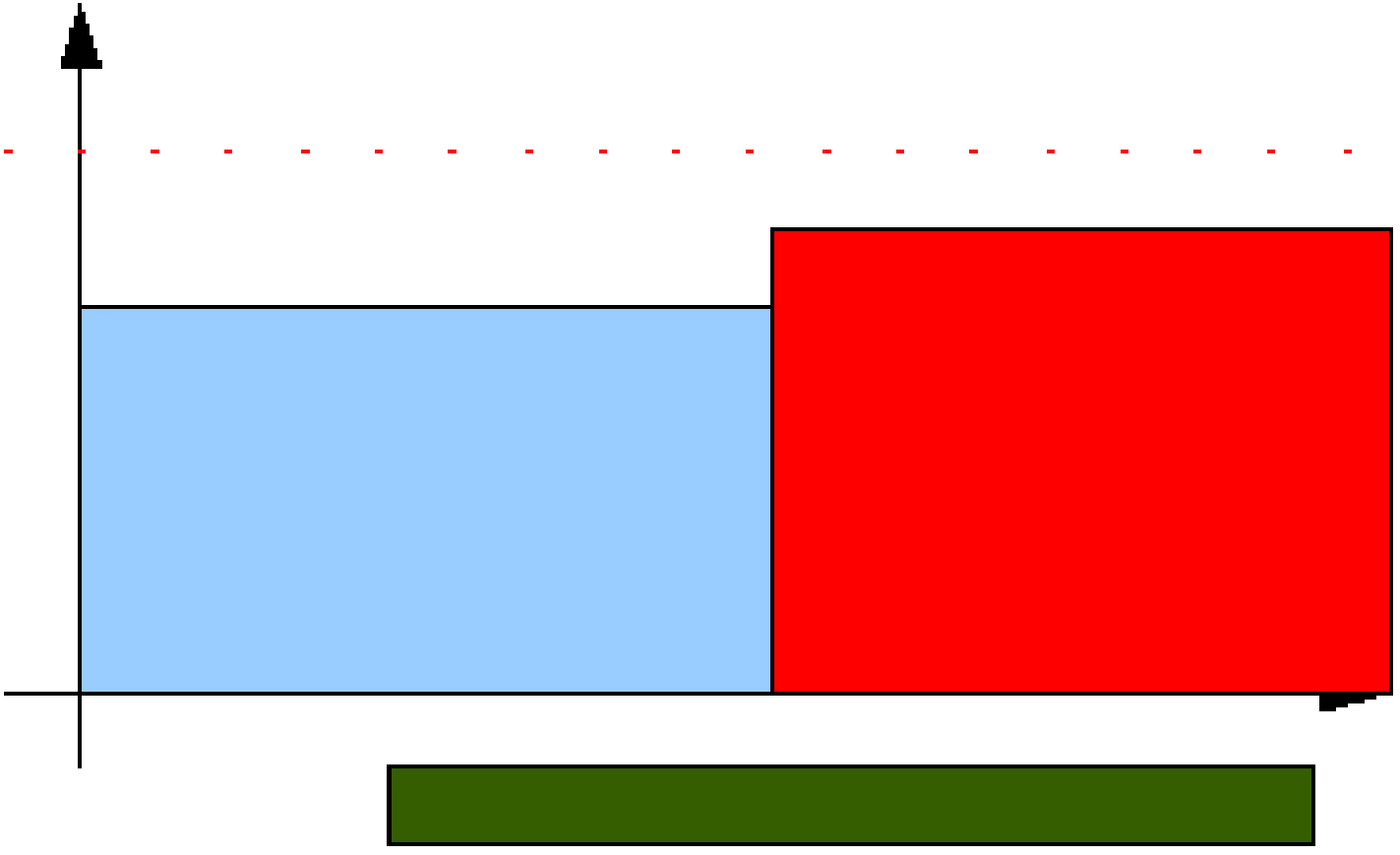
Soumission de job



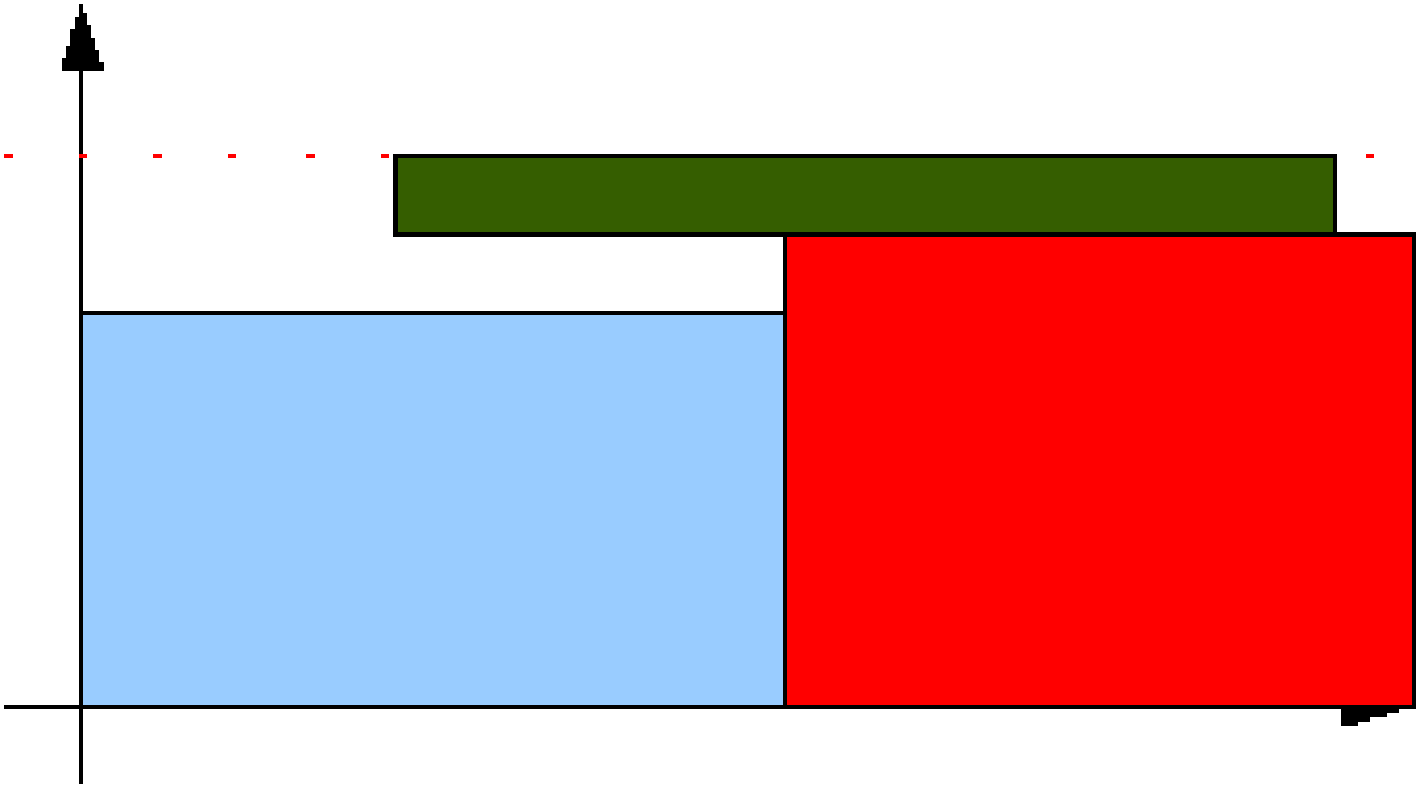
Soumission de job



Soumission de job



Soumission de job



Partages de fichiers

Serveurs NFS

- Historiquement partage depuis un serveur par *NFS* des espaces de données.
- Point de blocage quand le nombre de nœud augmente fortement.
- Sans NFS (`/travail`) complexité dans l'utilisation. Impossible de passer l'échelle.
- Necessité d'un système *scalable* sans point d'engorgement.

Partages de fichiers

glusterfs

- En test sur les nouvelles machines (quadri10-16,mld01-11)
- Agrégation des espaces de données.
- Visibilité unique pour l'utilisateur.
- Deployement sur les autres machines à court terme.

Partages de fichiers

glusterfs

- En test sur les nouvelles machines (quadri10-16,mld01-11)
- Agrégation des espaces de données.
- Visibilité unique pour l'utilisateur.
- Deployement sur les autres machines à court terme.

Exemple

-
```
<quadri10> /travail# df -h
Filesystem Size Used Avail Use% Mounted on
glusterfs 3.8T 91G 3.5T 3% /travail
<quadri10> /travail#
```

Conclusions

Situation actuelle

- Très utilisés.
- `https://sionet.obspm.fr/ganglia`
- Torque/Maui fonctionne bien.
- Pour l'instant pas de possibilité de suspension.

Conclusions

Situation actuelle

- Très utilisés.
- `https://sionet.obspm.fr/ganglia`
- Torque/Maui fonctionne bien.
- Pour l'instant pas de possibilité de suspension.

Évolutions

- Augmentation du nombre de nœud ? Problème de climatisation.
- Autre type d'ordonnanceur ?

Documents

Documents

- http://sio.obspm.fr/fichiersHTML/bench.html
- http://sio.obspm.fr/fichiersHTML/calcul.html
- http://sio.obspm.fr/fichiersHTML/PBS.html
- Liste de diffusion mpopm@sympa.obspm.fr

Documents

Documents

- http://sio.obspm.fr/fichiersHTML/bench.html
- http://sio.obspm.fr/fichiersHTML/calcul.html
- http://sio.obspm.fr/fichiersHTML/PBS.html
- Liste de diffusion mpopm@sympa.obspm.fr

Références

- http://www.resinfo.cnrs.fr/IMG/pdf/Josy_calcul_louvet.pdf
- http://www.resinfo.cnrs.fr/spip.php?article1
- http://calcul.math.cnrs.fr

Questions

?