Meso-Centre de Calcul intensif hébergé à l'Observatoire de la Côte d'Azur
http://crimson.oca.eu

Historique

- Avant 2005 :
  - Moyens de calcul mutualisés au niveau de l'OCA.
- À partir de 2005 :
  - l'OCA héberge un centre de calcul mutualisé par 7 laboratoire OCA+UNS
- À partir de 2008 :
  - l'OCA héberge la composante « calcul intensif » de la plateforme mutualisé de l'UNS

Missions

- Par rapport aux centres nationaux :
  - Accompagnement, mise au point et validation
- Par rapport aux équipes :
  - Calculs hors de portée de machines d'équipes
  - Éviter les acquisitions superflues
    - mutualisation de moyens humains
    - économie de fluides
    - optimisation de l'argent public
  - Formation

Organisation

- Projet animé par :
  - un responsable scientifique
  - un responsable technique
- Techniquement géré par une équipe du Service Informatique & Télécom de l'OCA

  → Mutualisation
- Comité de Pilotage
- Comité des utilisateurs

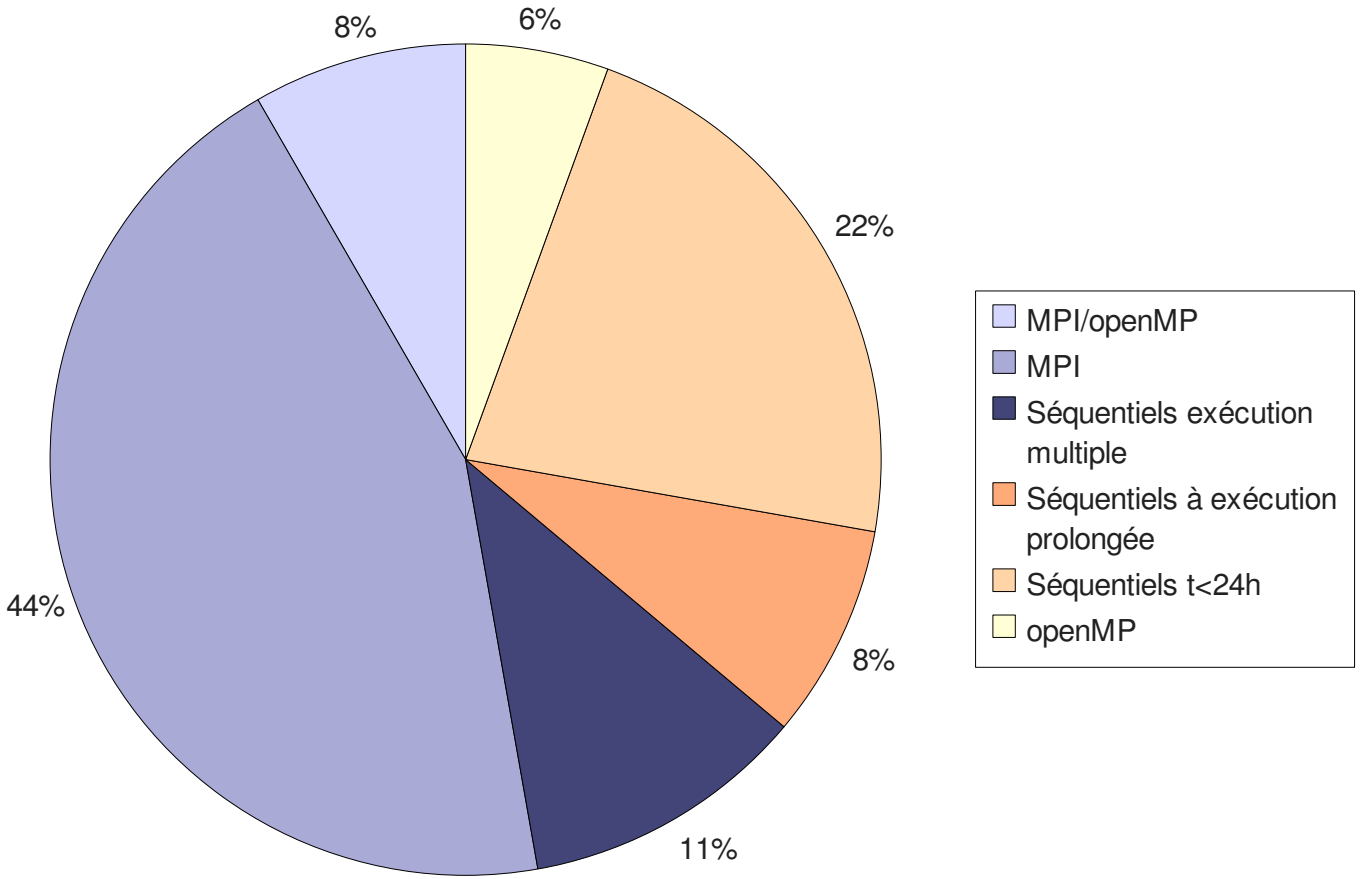
Des Comités : Pour quoi faire ?

- Tous les laboratoires partenaires ont un représentant dans chaque comité
- Comité de pilotage :
  - Valide les choix techniques et procédures
  - Participe aux ouvertures des plis lors des appels d'offres
  - Réunions annuelles
- Comité des utilisateurs
  - Arbitre l'accès aux ressources
  - Fait éventuellement remonter des problèmes

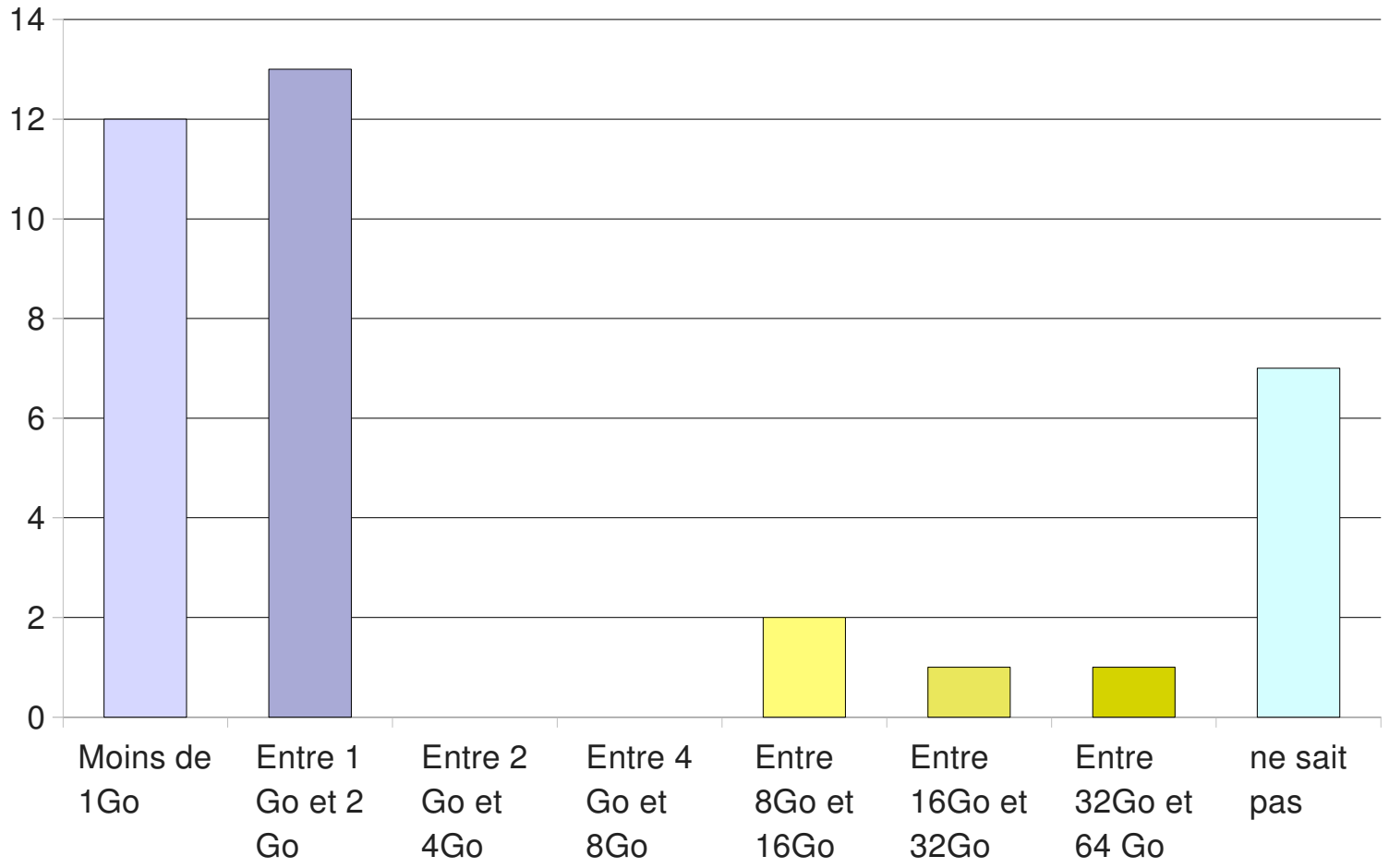
2005
Par où commencer ?

- Étude des besoins
  - Tournée des laboratoires pour trouver des utilisateurs potentiels *et* disponibles
  - Distinguer les *besoins* et les *envies*
  - Établir la répartition des différents types d'applications.
    - parallèle (MPI vs OMP) ou séquentiels (isolés ou par lots)
    - utilisation mémoire

2005
Types de codes



2005
Mémoire



2005 Benchmark

- Choix :
  - Nous nous sommes basés sur les benchmarks utilisateurs (et non sur des benchmarks standards)
  - Échantillon représentatif de codes utilisateurs validés par le comité.
    - la représentativité se mesure « au doigt mouillé »
- Les benchmarks permettent de sélectionner:
  - L'architecture (CPU, interconnect, stockage).
  - Le compilateur

2005
Paramètres fixés (1)

- Choix pour une plate-forme homogène :
  - Solution tout-terrain résultant d'un compromis prenant en compte la diversité des besoins...
  - ...plutôt qu'un ensemble de solutions spécialisées.
- **Préférence** pour des solutions open-source
- Linux pour l'OS
- Pour le reste: totale liberté donnée au fournisseur (engagement sur les résultats).

2005
Paramètres fixés (2)

- Préférence pour une plateforme pratiquement homogène:
  - Nœuds de calcul X86 (AMD selon bench...) 2G/cœur connectés en réseau de calcul haute performance (IB selon bench).
  - + 1 nœud « gonflé » 32G
  - Un « bon » compilateur (Pathscale selon bench)
  - Quid de l'espace disque ?
  - Validation sur les performances d'applications utilisateur pré-sélectionnées.

2005
Appels d'offres

- Appel d'offre en plusieurs lots:
  - Aménagement de la salle
  - Onduleur
  - Mise à niveau électrique
  - Climatisation
  - Cluster

→ Plus jamais ça....

2005
The winner is...

- Offre *constructeur* retenue:
  - 200 cœurs de calcul (AMD dual Core) a 2G/cœur 2.3Ghz
  - Interconnect IB SDR
  - Gestionnaire LSF
  - Supervision Ganglia
  - HP-MPI
  - Compilateur PathScale
  - Debugger ddt
  - Stockage ésotérique

2005
Mais...

- Installation dans les délais (1/3 paiement)
- Mise online à peu près dans les délais
- 6 mois de retard avant validation.
  - Solution disque totalement inadaptée à MPI (notamment aux I/O MPI2).
    - Validation conditionnée par les perfs
    - Il reste 2/3 du paiement

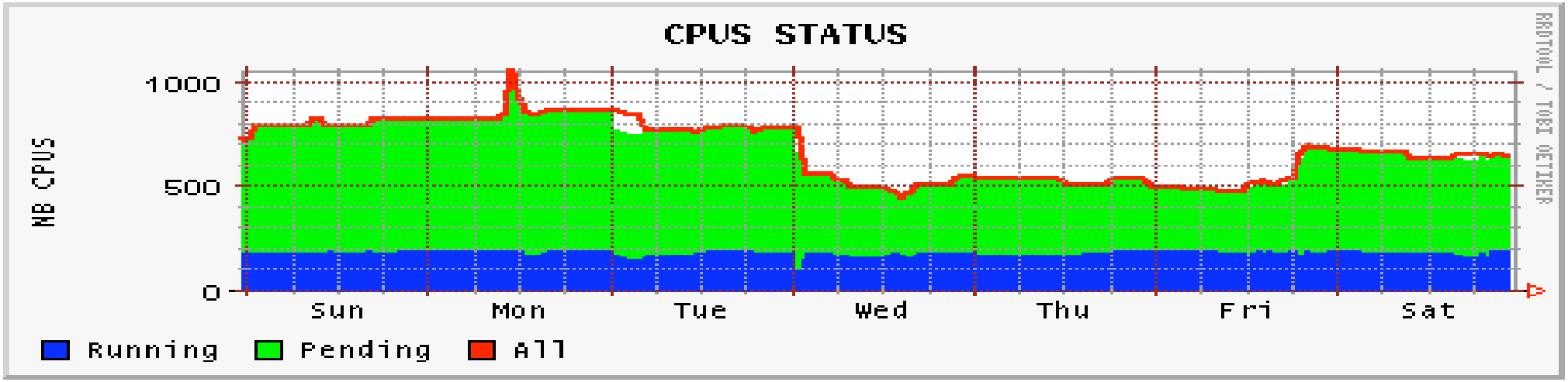
    → remplacement sans frais par une solution Lustre

2005-2008
...De plus...

- Le support, initialement excellent, migre du national à l'européen :
  - devient, en pratique, limité à l'échange standard de pièces détachées.
  - l'environnement « supporté » par le vendeur n'est pas à son catalogue.

2005-2008 Et les utilisateurs dans tout ça ?

- Globalement satisfaits :
  - Taux d'occupation autour de 90%
  - Assez bien répartis sur les laboratoires
  - Doivent parfois être poussés vers les centres nationaux



2008
Première extension

- L'étude des besoins se base maintenant sur 2 ans de surveillance (nagios, ganglia,...)
  - Besoins en mémoire revus à la baisse
  - Besoin en nombre de cœurs :
    - pour applications parallèles
    - pour travaux séquentiels lancés en masse
- Souhait de favoriser la simplicité pour l'utilisateur

  → l'environnement peut être modifié, mais doit rester homogène

2008
Appel d'offre

- Considérablement simplifié:
  - pas de mise à jour de l'environnement
  - acquisition totalement homogène
- Liberté laissée sur l'environnement logiciel (hors compilateur)
- Les candidats peuvent revoir les choix d'environnement logiciel...
- ...mais doivent prendre en charge l'existant

2008
Le choix

- Solution intégrateur
  - 752 cœurs AMD connectés en IB DDR
  - Linux « constructeur » → CentOS
  - LSF → SGE
  - System Imager → KickStart
  - Lustre → GPFS

2008
Les Soucis

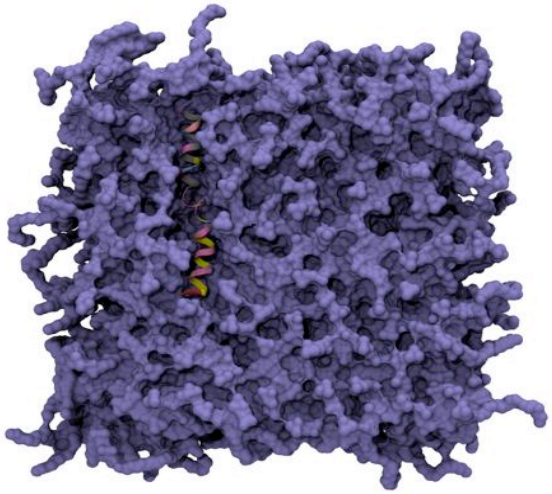
- Dual core → Quad core
  - goulot d'étranglement mémoire sur les codes spectraux
- LSF → SGE
  - prise en main
  - mise à jour des documentations
  - accompagnement des utilisateurs
  - mise en place de files prenant en compte les problème d'accès mémoire.
- Instabilité GPFS (cohabitation SDR/DDR ?)

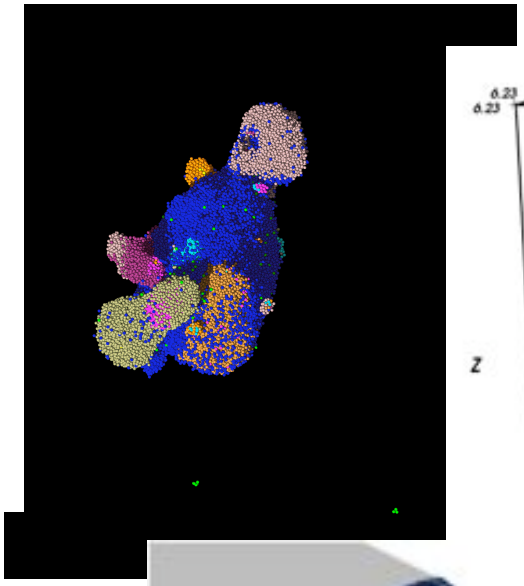
Quid de l'humain?

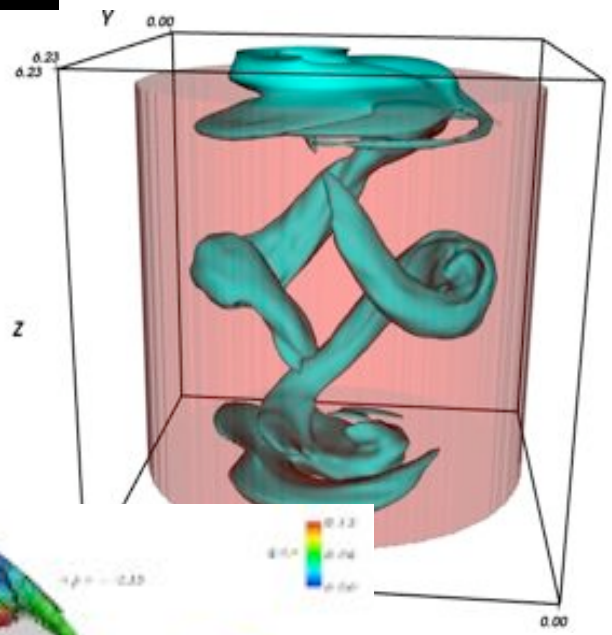
- Les utilisateurs sont répartis dans 8 laboratoires distribués géographiquement
  - certains sont à 50 km...
  - ...d'autres dans le couloir

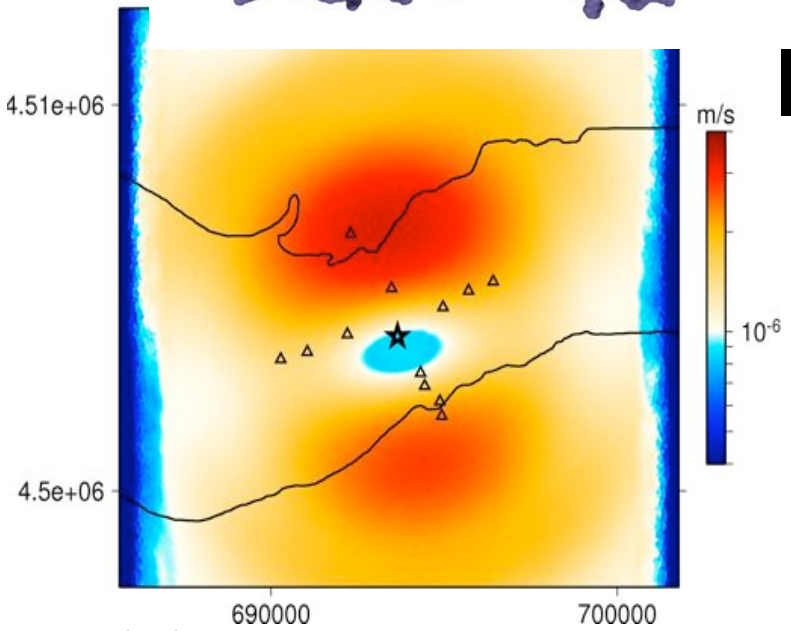
    → comment gommer la distance ?
  - Prise en charge du support via un helpdesk et déplacement *si nécessaire*
  - Problématique surtout avec les utilisateurs proches.
- Une centaine de comptes, pour une dizaine d'utilisateurs à un instant donné.

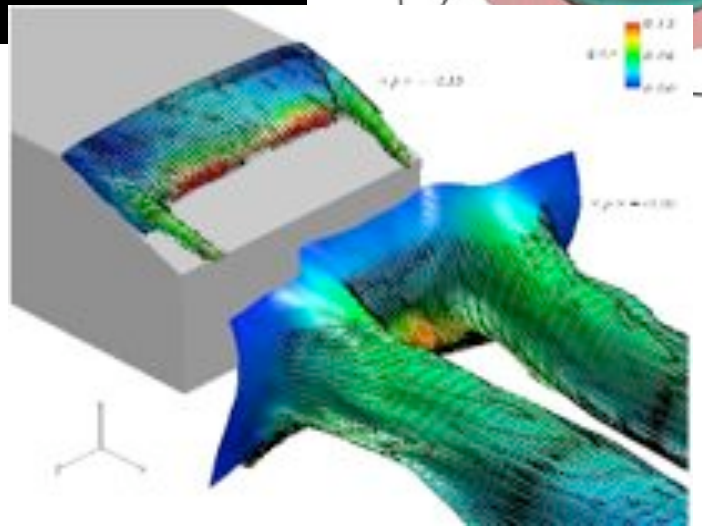
Pas de thématique spécifique











Support Développement

- 1 ingénieur vs beaucoup d'utilisateurs
  - nécessité d'un minimum de méthode de la part des utilisateurs.
    - beaucoup de problèmes sont en fait d'assez bas niveau
    - intervention sur code uniquement si le code est versionné (→ mise en place d'une « forge » dépassant finalement les cadres du mésocentre)
    - mise en place de quelques formations spécialisées (C/C++, outils de développement)

Éligibilité

- Chaque laboratoire a un représentant qui, seul, décide de quel membre de son laboratoire a accès au cluster :

  En pratique, aucun refus sur les cas raisonnables

- Chaque utilisateur est affecté à un projet...
- ...utilisé uniquement pour la « comptabilité »
- Il est possible d'ouvrir l'accès à des collaborateurs **dans le cadre de travaux communs**.

  → les publications doivent être co-signées par des membres d'un laboratoire partenaire.

Tendances récentes

- Jobs séquentiel soumis par lots :
  - un utilisateur lance plusieurs centaines d'instances d'une même application sur des jeux de données différents.
    - notamment analyse de données observationnelles
  - typiquement le genre d'utilisation :
    - trop grosse pour des machines d'équipe
    - peu de chance de passer sur des centres nationaux
  - peut servir de « bouche-trou »

Mutualisation (1)

- Permet de réaliser des économies d'échelle importantes en terme de ressources humaines.
  - de part la relative uniformité impliquée
  - de part la mise en commun de moyens humains entre l'équipe CRIMSON et le SIT.
  - s'applique
    - aux méthodes et outils (assez simple)
    - aux matériels (moins simple)

Mutualisation (2)

- Actuellement nous avons intégré les équipement suivants
  - espace disque de 50To
  - mini cluster de calcul/visualisation et post-traitement
- Compromis à trouver :
  - l'équipe tire parti d'un service commun

    → la communauté doit en tirer un avantage
  - pour le moment négocié au coup par coup

Mutualisation (3)

- Résistance de la part des équipes :
  - le coût complet de possession n'est généralement pas « compris »
  - besoin de se sentir propriétaire
  - contraintes ésotérique des financeurs
- La surveillance systématique peut parfois servir d'argumentaire: certaines machines d'équipe ne sont utilisées qu'à 10%
- Mais la situation s'arrange lentement.

Mutualisation (4)

- Notre prochaine étape sera de mettre en place des files de soumission avec préemption :
  - ainsi un utilisateur pourra partager sa machine en sachant que ses travaux seront prioritaires
  - problème : fonctionne mieux avec des codes supportant la notion de checkpoint.

Renouvellement

- Le renouvellement complet de la plateforme est actuellement en cours pour un montant de 700K
- Nouvelle problématique :
  - choix du site
  - consommation
  - environnement
  - la puissance brute n'est plus le critère principal
- Appel d'offre intégré en un seul lot