Denis PUGNÈRE
Guillaume BAULIEU
Yoan GIRAUD

CNRS / IN2P3 / IPNL

IN2P3
Institut national de **physique nucléaire**
et de **physique des particules**

Impacts organisationnels et techniques d'un noeud de grille dans un laboratoire du CNRS

JRES-2009
Nantes

Avant propos

Ce n'est pas un tutoriel sur la grille

Cette présentation est un retour d'expérience :

- Voici ce qu'on a (bien / mal) fait,
- comment nous l'avons fait,
- quels sont les impacts ?

PLAN

- Introduction
- Différences entre cluster et noeud de grille
- Le projet
- Préparation
  - Installation
  - Exploitation
- Ressources humaines
- Contraintes
- Ré-aménagements
- Conclusion

Organisation du calcul académique en France

- Centres de calcul (CC-IN2P3, CEA, CINES, IDRIS)
- Mésocentres* : 31 sites
- Grille de recherche* (GRID5000) : 18 laboratoires
- Grille de production* : 15 laboratoires
- Combien de clusters locaux ?

Quelles implications de l'installation noeud de grille (de production) dans un laboratoire ?

* : liste au 01/11/2009, voir références sur l'article

Contexte

- Institut de Physique Nucléaire de Lyon
- Laboratoire mixte CNRS / Université-Lyon1
- 220 personnes
- Activités :
  - Étude des propriétés des composants subatomiques de la matière ainsi que leurs interactions,
  - Étude de la physique des particules et des astroparticules, la matière nucléaire

Origine du projet

- Renouvellement de la « fermette » de calcul
- Cluster ou noeud de grille ?
  - Plusieurs équipes impliquées dans le projet LHC
  - D'autres ont des besoins de calculs que la grille peut combler
  - La plupart des calculs exécutés sur les clusters sont transposables sur la grille
- La grille : c'est quoi ?
  - C'est beau !

Crédit : CERN Photolab (BUL-PHO-2003-010)

Origine du projet

- Renouvellement de la « fermette » de calcul
- Cluster ou noeud de grille ?
  - Plusieurs équipes impliquées dans le projet LHC
  - D'autres ont des besoins de calculs que la grille peut leur combler
  - D'autres non
- La grille : c'est quoi ?
  - C'est beau !
  - C'est compliqué ? Pas plus que le cloud-computing
  - C'est … différent d'un cluster

Cluster / noeud de grille

Cluster local	Noeud de grille
Interface de **lancement** de tâches	Interface de **lancement** de tâches
Interface de **visualisation de l'état** des tâches en cours d'exécution ou en attente,	Interface de **visualisation de l'état** des tâches en cours d'exécution ou en attente,
Interface de **récupération des résultats** issus des tâches	Interface de **récupération des résultats** issus des tâches **locales** ou sur les **autres noeuds**
	Structuration des collaborations en **Virtual Organisations** : les différents membres d'une communauté **partagent leurs ressources** (espace de stockage, *CPU*, données, logiciels, environnement de développement et d'exécution, outils...)
	Interface mutualisée d'accès à la **comptabilité** (accounting) de la consommation des ressources
	Interface mutualisée et les **services** associés permettant de ***monitorer*** l'état de l'**ensemble des noeuds** (**disponibilité** des services, **capacité** des ressources...).
	Optimisation de l'usage des ressouces

L'intergiciel : Glite

- **Les machines de service :**
  - **CE** (**Computing Element**) : Composé d'un gestionnaire de tâches et d'un ordonnanceur,
  - **BDII** (**Berkeley Database Information Index**) : permet de publier les caractéristiques et l'état du site vis à vis de la grille,
  - UI (**User Interface**) : serveurs d'accès pour lancer les calculs sur la grille,
  - **LFC** (**LCG File Catalog**) : serveurs de catalogues,
  - **SE** (**Storage Elements**) : les serveurs de stockage composés généralement de têtes de stockages (**Head Nodes**) et de serveurs d'entrée sorties (**Disk Nodes**),
  - les serveurs nécessaires au fonctionnement de certaines expériences du LHC (en particulier les **Vobox** CMS et Alice),
- **le cluster de calcul (Workers-Nodes).**

Calendrier

- Départ → réalisation : 1 an
  - Décision septembre 2006
    - 1ers bons de commandes en novembre 2006
    - Aménagements en mars 2007
  - Certifié pour la production en juin 2007
  - Premiers jobs de la VO Alice en sept 2007
  - Augmentation progressive de la capacité de calcul

Préparation de l'infrastructure

- Aménagement de la salle serveurs :
  - faux plancher (ou pas) ?
  - Détection incendie ? Coupe circuit ?
  - Extinction incendie ?
- Alimentation électrique :
  - Distribution électrique : simple ou double arrivée ?
  - Onduleur ?
  - Groupe électrogène ?
- Climatisation ?
- Réseau ?

Modèles de coûts

- Coûts d'infrastructure très variables d'un site à l'autre
- Coûts capacité de calcul et stockage assez comparables, exemple : modèle de cout LCG
  - CPU : en € / KSI2K-HEP
  - Stockage : en € / Go utile
  - Pris en compte : cartes d'alimentation, le rack, les interfaces du commutateur réseau correspondant, le serveur de disques pour le stockage, la garantie de 3 ans pour le *CPU* et de 4 ans pour le disque
- Non pris en compte : machines de service, consommation électrique, climatisation, onduleur...

Phases d'installation et d'exploitation

- **Installation**
  - **Acquisition des connaissances : (auto-)formation, veille, ateliers, réunions**
  - **Installation / déploiement**
- **Exploitation quotidienne :**
  - **administration, modifications, pannes, déboggage...**
  - **Mise a jour de l'ntergiciel : nouvelles versions, nouvelles fonctionnalités**
  - **Installations :**
    - **Nouveaux services de grille**
    - **Soft des expériences**

Spectre des connaissances

- Programmation (shell, perl, python),
- Réseau : agrégation de liens, limites des protocoles, routage,
- Sécurité : filtrage, certificats, rôles, authentification,
- Stockage de grande capacité : optimisation, limites des systèmes de fichiers,
- Calcul distribué, déploiement en nombre, monitoring, gestion de systèmes de batch,
- Software de chaque *Virtual Organisation*,
- Le mode de fonctionnement spécifique de chaque *Virtual Organisation*

Besoins en formation

- Formation spécifique des administrateurs de noeuds de grille → très utile
- Partage des connaissances entre les administrateurs des différents sites (très utile) : réunions, forums bi-annuels... → accumulation d'expérience(s)
- g**gle est (m|t)on ami -Quand la doc est à jour- : mais est-ce que je documente aussi ce que je fais pour les autres ?
- Freins à la compréhension :
  - Dissémination des services : monitoring, des projets EGEE & LCG, des expériences (Alice, CMS), des outils (tickets, système d'information)...
  - Quantité d'acronymes !

Installation d'un noeud de grille

… Il faut installer un Bi-Di-Aïe-Aïe !

Euhhhh, un quoi ?

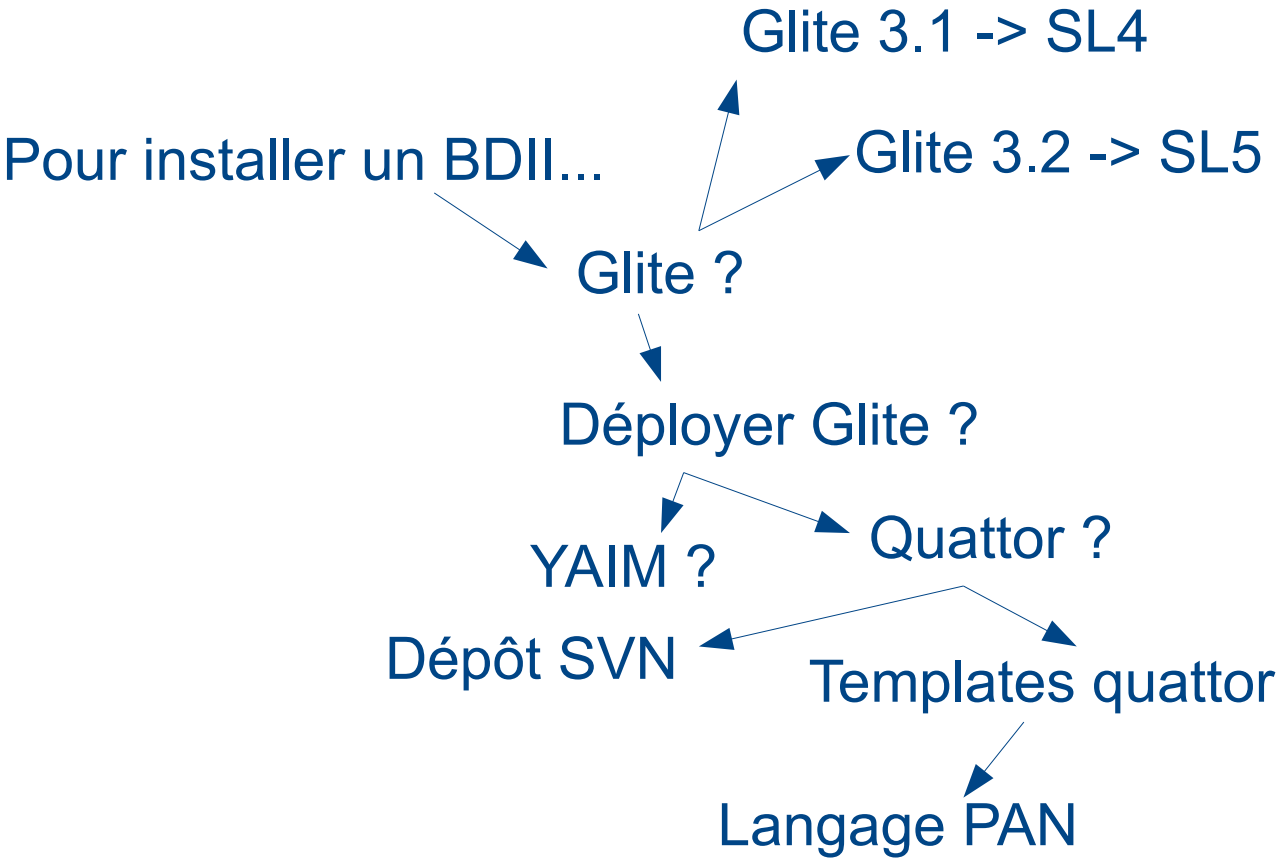
… un BD2I

Aaah, un BDII !!
(Je crois qu'il veut parler du Berkeley Datab... Info... ...machin)

Ok, installons un BDII

.
..
…
Quelques jours après

Installation d'un noeud de grille



Installation d'un noeud de grille

Contraintes du projet EGEE

- Certification des sites
- SLA (accord entre le projet et le site)
  - 70% de disponibilité / mois (> 21j par mois)
  - 75% de stabilité (stabilité = disponibilité / (disponibilité + arrêts pour maintenance imprévus)
  - réponse au tickets dans les 4h,
  - résolution des problèmes < 5j
- Disponibilité et fiabilité mensuelle de tous les sites mesurée et publiée par le projet

Contraintes du laboratoire

- **Comment assurer la surveillance du site comme le demande les projets EGEE/LCG pendant les périodes de congés ?**
- **Doit-on définir des astreinte ?**
- **1,4 ETP sur 3 à 4 personnes à temps partiel sur ce projet qui ont aussi d'autres contraintes**

Organisation

- **Wiki interne**
- **Procédures opérationnelles (arrêt / redémarrage du site...)**
- **Roulement dans la prise de congés**
- **Partage des connaissances le + possible en interne et externe**
- ***Monitorer* TOUT ce qui est possible : température, compresseurs de la climatisation, consommation électrique, réseau, charge serveurs, services réseaux, espace disque, nombre de processus...**

Petits arrangements...

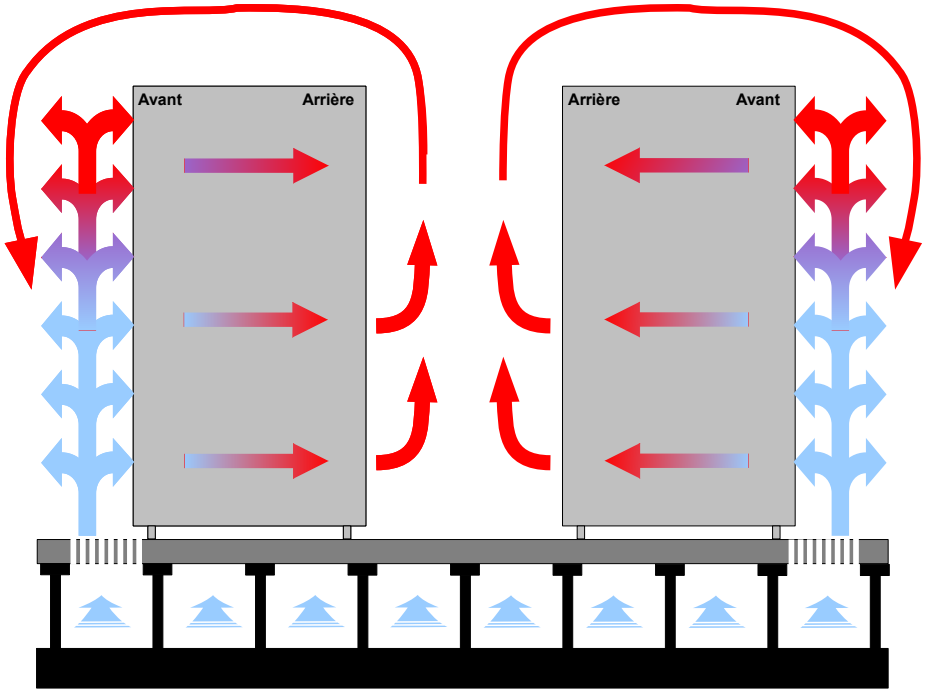
- **Les besoins/spécifications des uns et des autres sont quelques fois contradictoires :**
  - **Soft experience X sous SL4 mais pas certifié SL5 ; experience Y demande à passer en SL5**
  - **2 implémentations d'un même protocole de stockage (xrootd) : natif et plugin xrootd. Plugin déployé mais ne sera pas utilisé.**
- **Administrateur de noeud de grille = interface avec les projets EGEE, LCG, les expériences locales, les utilisateurs, le laboratoire**

Ré-aménagements

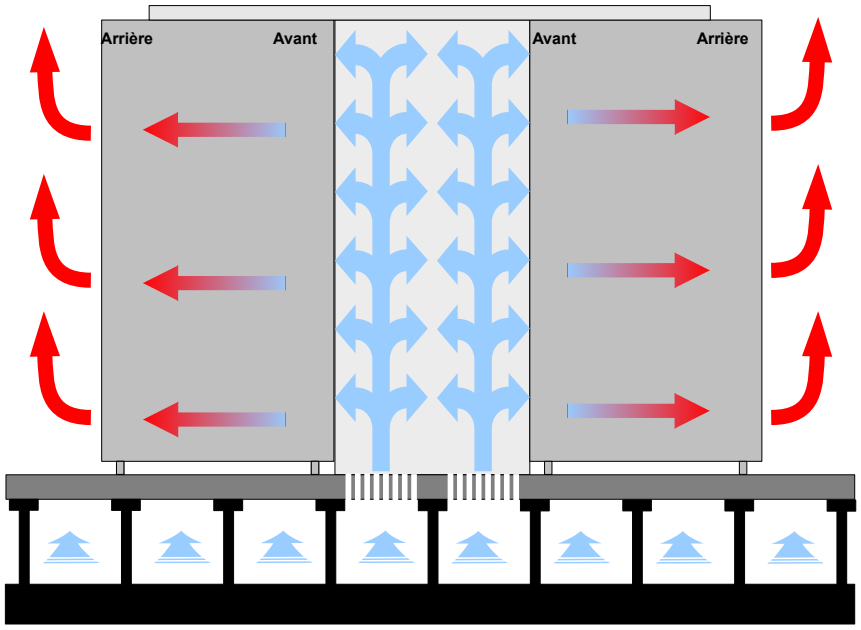
- **Aménagement initial de la salle serveurs (mars 2007)**
  - **coupure 1h vendredi + arrêt tout le samedi**
- **Retournement de l'ensemble des baies (décembre 2008)**
  - **1 jour d'arrêt total du site**

Ré-aménagements

Retournement de l'ensemble des baies : Création d'un couloir froid cloisonné



Avant



Après

Ré-aménagements

- Aménagement initial de la salle serveurs (mars 2007)
  - coupure 1h vendredi + arrêt tout le samedi
- Retournement de l'ensemble des baies (décembre 2008)
  - 1 jour d'arrêt total du site
- Changement de VLAN de tous les services grille : problème de performance / saturation de liens (septembre 2009)
  - 2 jours d'arrêt du noeud de grille
- Prochain chantier : Changer le coeur de réseau 1Gb/s → 10Gb/s : Durée ???

Conclusion

- Ne pas négliger l'infrastructure
- L'investissement financier est important et ne se limite pas à l'achat de machines
- Difficultés : quantité de matériels différents (augmente la quantité de travail de maintenance)
- Apporte des contraintes (organisation, surveillance...)
- Nécessite d'apprendre beaucoup de choses (réseau, stockage, soft des manips...) : compétences acquises très variées
  - Compétences au service du laboratoire et des expériences
  - Proximité des administrateurs et des chercheurs : efficacité
- Le laboratoire bénéficie de l'infrastructure pour les services standards
- C'est intéressant ! (quand même)

Questions