Réunion STIC-HEP, 25 février 2005 à Grenoble
Compte rendu par FM
Notes : FM, MJ, DL
http://lpsc.in2p3.fr/congres/LCGFrance/LCGhome.html

Présents:
Johann Collot (Dir. LPSC), Fabio Hernandez (LCG-France,CCIN2P3), Denis Linglin (Dir. CCIN2P3), Eric Fede (EGEE, CPPM), François Etienne (IN2P3 et CPPM), Edith Knoups (CPPM), Charles Loomis (EGEE, LAL), Julia Andreeva (CERN), François Lediberder (IN2P3), Brigitte Plateau (Dir. ID-IMAG), Dominique Boutigny (LAPP), Fréderic Desprez (Dir. LIP-ENS Lyon), Lionel Schwarz (Stockage CCIN2P3), Pierre Girard (CCIN2P3), Rolf Rumler (EGEE, CCIN2P3), Hélène Cordier (LCG/EGEE, CCIN2P3), Alan Su (ENS Lyon), Alain Lecluse (IBCP), Franck Cappello (INRIA-LRI), Philippe Marty (INRIA-LRI), Cyril Labbe (IMAG), Yves Denneulin (IMAG), Guillaume Huard (IMAG), Fairouz Malek (LCG-France, LPSC), Jerôme Fulachier (LPSC), Fabian Lambert (LPSC), Michel Jouvin (LAL).
Par téléphone pour la partie C: Guy Wormser (LaL), Dominique Pallin (LPCC)
Excusés: Thierry Priol (INRIA-IRISA-Rennes), Nadine Neyroud (LAPP), Sophie Nicoud (CNRS UREC),

Labos IN2P3: CCIN2P3 (Lyon) LPSC (Grenoble), CPPM (Marseille), LAPP (Annecy), LAL (Orsay), LPCC (Clermont)
Labos STIC: LRI (Orsay), IMAG (Grenoble), IBCP (Lyon)

AGENDA : http://lpsc.in2p3.fr/cdsagenda/fullAgenda.php?ida=a0548

10:00 ***Le contexte (15')*** *Brigitte Plateau, François Etienne*
10:15 ***A. Problématiques de traitement des données du LHC (IN2P3)***

Vue d'ensemble du projet LCG (20') Fabio Hernandez (CCIN2P3)
Modèle de calcul LCG (T0, T1, ...), volume des données, etc...
10:35 ARDA (30') Julia Andreeva (CERN)
Les services proposés par le middle-ware (soumission de jobs, gestion de données, replication, ...)
11:05 Les services d'EGEE (30') Eric Fede (CPPM)
Description des services de grille du projet EGEE
11:35 Infrastructure opérationnel de LCG (1h00') Pierre Girard (CCIN2P3)
Rôle des centres des ressources, core infrastructure centre, regional centres, gestion des incidents opérationnels, etc...
12:35 ***B. Middleware de Grille développé par STIC***
GRID 5000 (30') Franck Cappello
13:05 Déjeuner

14:00 Projets au STIC (1h00') Franck Cappello

15:00 Pause café

15:15 C. Discussions sur les sujets d'intérêt commun (2h00')

IN2P3+STIC

- Discussion sur les intérêts en commun
- Plan de travail de recherche interdisciplinaire
- Document de collaboration
- Demandes de postes au CNRS
- Mise en place d'un réseau de contacts
- Présentation du projet au CS? GDR ?

17:15

Contexte :

Brigitte et François introduisent la réunion en informant l'auditoire de l'historique du déclenchement de cette rencontre. La première des raisons est celles d'assurer un avenir dans la recherche en Informatique pour les chercheurs CR1 qui seront embauchés par la section 07 dans des laboratoires de l'IN2P3 dans le cadre des grilles de calcul. Pour Brigitte, La grille est un outil privilégié de recherche pour les STIC et un outil de production pour l'IN2P3. Il est sans doute possible de faire des choses ensemble. Il faudra aujourd'hui, dégager les perspectives scientifiques.Pour Denis, cette réunion est une première et elle devrait être le début d'un processus jusque là vain.

A. Problématiques de traitement des données du LHC (IN2P3)

1. F. Hernandez (CCIN2P3, LCG-France) : Vue d'ensemble du projet LCG

Grande masse de donnée.
Hiérarchie de centre de ressources par Tier (rang) :

- *Tier 0 : acquisition et filtrage des données brutes, reconstruction*
- *Tier 1 : stockage des données brutes et reconstruites (ESD), reconstruction, haute disponibilité des données, serveur de données pour les Tier 2*
- *Tier 2 : stockage de données sur disque uniquement (accessible via middleware), simulation, analyse (y compris interactive)*

Tous les centres ne servent pas toutes les expériences (CC IN2P3 = 4 expériences).
Nombre de Tier 1s prévisionnels par expérience :

- *Atlas : 10*
- *LHCB : 3*
- *CMS : 6*
- *Alice : 6*

LCG : 3 composantes principales

- *Europe : 117 sites : challenge de la gestion cohérente d'un si grand nombre de site*
- *USA : Grid3*
- *Pays nordiques : NorduGrid*

LCG France :

- *Mise en place d'un Tier 1 à Lyon*
- *Favoriser l'émergence de Tier2 dans les labos IN2P3/DAPNIA, intégration des Tier 2*
- *Budget : 16,5 M€ d'ici 2008*
- *Somme des 4 expériences : 8,7 MSI2000 (~3000 bi-pro), 5PB disque (~1500 serveurs disque), 5PB MSS (4 silos de 25000 cartouches à 200 GB, pas un gros pb), réseau 10 Gb avec le CERN, 1Gb avec les Tier2 (déjà presque opérationnel)*

F. Le Diberder : LCG est le projet majeur en informatique de l'IN2P3 qui captera l'essentiel de nos ressources dans les années à venir.

2. J. Andreeva (CERN): ARDA project

Middleware package in Europe coming from EDG : currently LCG-2 (Feb 04)

US Grids running different middleware: common Virtual Data Toolkit (VDT)

Data challenge in 04 provided valuable experience that drives development of LCG-3.

- *LCG-3 developed by EGEE in collaboration with VDT (US).*
- *Applications and users more involved in development : ARDA project*
- *Progressive move from LCG-2 to 3 : mixture of both during a certain period*

ARDA: interface to GRID enable HEP analysis applications.

- *Complex and evolving requirement : difficult to solve needs for all experiments*
- *Users communities and non European Grid providers involved*
- *Based on gLite : provide feedback to gLite developers based on experiment use cases*
- *Handle job submission : split jobs in optimal units, interface between experiments metacatalog and Grid data management system, selection of available resources, experiment specific SW deployment*
- *Job processing : monitor jobs progress, access to logs, retrieval of outputs*

Workload Management System (WMS)

- *ARDA evaluated 2 WMS : task queue (pull model) from Alien, WMS (pull and push) from EDG*
- *Critica : WMS integration with other middleware components (file catalog, package messengers…)*

- *Every experiment can have its own strategy for data discovery : flexible integration mechanism needed*
- *Job splitting/clustering needed*

Data Management:
- *High performance catalog critical both for insertion and queries*
- *2 catalog evaluated : Catalog from Alien and Fireman*
- *gLiteIO stability under high load and concurrency, graceful error recovery*
- *Space management and permissions (ACLs)*
- *Reliable file transfer*
- *Data movement/location service : need to handle collection of files, integrate with exp catalog/bookkeeping systems*

Package management:
- *Multiple approaches possible from pre-installation (static) of SW by site managers to installation on demand (with removal at end of job)*
- *On demand installation doesn't scale well with very large software distribution (Atlas = 3GB)*
- *Current gLite package manager derived from gLite*
- *Still a lot of work needed to handle all the use cases*

Complementary project started at CERN (3D) to work on distributed databases in Grid environments

3. E. Fede (CPPM), Services EGEE

Présentation basée sur l'état des lieux fait par F. Hemmer (CERN) lors de la revue LCG de mi-février.

Développement EGEE essentiellement dans JRA1. Première version de gLite prévue fin mars.

Architecture gLite orientée service
- *Service le plus léger possible*
- *Réutilisation de l'existant chaque fois que possible*
- *LCG basé essentiellement sur EDG avec des apport de VDT (VOMS Datatag et EDG) et quelques autres middleware*
- *Intégration de composants Alien (gLiteIO, package manager)*
- *Supprimer les SPOF des services centralisés*
- *Coexistence avec LCG-2 et OSG (US) critique pour EGEE*

WMS :
- *Basé sur le WMS de EDG*
- *Support de d'instance multiple du même job avec des jeux de données différents*
- *Support des modèles push et pull*

- *Futur : implémentation Web service, WMS distribué*

Job management :
- *CE*
- *Logging et bookkeeping*
- *Job provenance : trace des jobs soumis sur une longue période*
- *Grid Accounting : optimisation des régles d'utilisation*
- *VOMS*
- *Advance reservation : en cours de développement*

Data Management:
- *SE : basé sur SRM, Posix file API, transfert avec GridFTP*
- *File Replica et Catalog : résolution nom logique compatible SRM, catalogue distribué (en cours de développement)*
- *File transfer : relations transactionnelles avec les catalogues*
- *gLiteIO : basé sur AliEn*
- *Metadata catalog : service EGEE assez limité, interface avec les catalogues des expériences*

Monitoring et Information Service :
- *Fourniture d'infos fiables sur l'état de la grille*
- *Basé sur R-GMA : modèle producteur/consommateur*
- *API compatible Web service*
- *API instrospection : découverte des API de la grille*
- *Réplication du schéma (en cours)*

Service d'accès à la grille via le Web sous la forme d'un Web service

Package Manager

Sécurité :
- *Présente dans tous les services…*
- *Accès aux données : via des ACLs, notion de « rôle » intégré dans les VOMS*
- *Intégration VOMS et R-GMA en cours*

LCG ne développe pas de service : utilisation de composants disponibles « sur le marché »
- *EGEE est le principal candidat et fournisseur*

4. P. Girard (CC IN2P3), Infrastructure Opérationnelle

EGEE/LCG ROC deputy and RC coordination.

- *EGEE et LCG partagent la même infrastructure de grille et utilisent le même middleware*

- *La plupart des sites offrent leurs ressources aux 2 projets*

Les 2 projets se distinguent par leurs objectifs
- *Grille dédiée vs grille généraliste*
- *Politique d'administration : pas de notion de Tier dans EGEE*

Contribution bilatérale
- *LCG est moteur pour EGEE : doit être prêt en 2008*
- *EGEE contribue à la mise en place de l'exploitation de LCG*

Principes d'exploitation LCG/EGEE
- *2 Global Operation Centers (GOC) / 1er niveau EGEE : RAL et Taipei : gèrent l'enregistrement des « sites » (contacts, description, type de service.)*
- *2ème niveau : Core Infrastructure Center (CIC), Regional Operation Center (ROC), Ressource Center (RC). CIC et RCs ont une relation « logique ». ROC et RCs ont une relation « physique ».*
- *Site : fournit des ressources (ex : CE) et/ou des services (ex : resource broker)*
- *CIC : gestion des VOs, replica catalog, resource brokers, MIS, monitoring et accounting, site and users support, CIC « on duty » tournant chaque semaine, validation des nouvelles versions. Fonctionnent ensemble, 1 seul CIC vu de la grille.*
- *En Europe, 5 CICs : CERN, CC IN2P3, INFN, RAL, Russie (en cours)*
- *RC : fournisseurs de ressources (CE, SE), équipe de gestion du système, des réseaux, du stockage, expert du middleware grille…*
- *Communautés organisés en VO*
- *Challenge : équipe d'exploitation de la grille répartie sur 100 sites !*

CC IN2P3 : CIC + ROC + RC
- *ROC : représentant des sites d'une région géographique : assistance aux sites (y compris intégration de nouveaux sites), coordination du déploiement, intermédiaire entre CIC et RC pour le suivi des problèmes, sécurité, monitoring régional*
- *Suivi des sites de la région (fédération) France : ~10 sites (y compris CGG, CINES)*
- *Site de gestion du workflow d'exploitation : http://cic.in2p3.fr*
- *Coordination avec les autres ROCs et les CICs*

5.Discussion sur la partie A:

Collaboration STIC sur le middleware ?
Le problème, c'est que la France ne participe pas aux développements de middleware…Grâce aux futurs chercheurs, on pourra s'y lancer. Il reste et il restera des sujets. Mais il s'agit d'un milieu concurrentiel.

Dans EGEE, les VO LCG sont automatiquement supportées. Donc tout site EGEE peut tourner LCG (mais n'est pas obligé de le faire).

B. Middleware de Grille développé par STIC (F. Cappello)

Enjeu : passer des modèles mathématiques (abstraits) aux systèmes réels : besoins d'émulateurs qui puissent traiter la taille d'une grille réelle et d'applications pour la validation.

Démarrage à partir des outils de simulation existants : SimGrid, MicroGrid, Bricks…
Aucun totalement satisfaisant pour prendre en compte systèmes et réseaux
Grid5000 : tentative de construire une plate-forme expérimentale pour progresser dans la simulation et la validation. Pas une grille mais un outil d'étude des grilles et de leur middleware.
D'autres projets de nature similaire : DAS-3 (Pays Bas, 1000 CPUs sur 4 sites, middleware Globus), collaboration à l'étude.

1. GRID5000

- *Il y a une certaine homogénéité de la grille : 2/3 Intel 64 bits, mais on agrège aussi Itanium, G5, etc dans le dernier tiers.*
- *Budget : 7,6 M€ (Ministère, ACI grid et MD, INRIA, CNRS, Conseils régionaux).*
- *Un comité technique de 28 membres (F. Simon/ renater), plein de bénévoles*
- *Steering (11 p.) : Cappello, Priol, …*
- *Construire la plateforme répartie sur 8 sites (de 256 à 1000 CPUs) : un projet en soi…Interconnexion MPLS niveau 2 par RENATER.*
- *Pas d'homogénéité totale des configurations dans Grid5000 : minimum de 2/3 des ressources par site conforme au standard (Pentium Linux).*
- *Permet des expériences sur les réseaux, les relations middleware/OS, les API de programmation, l'architecture des applications et la mesure de la performance, la scalability (mise à l'échelle?) et la tolérance aux pannes.*
- *Besoin d'une plateforme hautement reconfigurable : les « utilisateurs » doivent pouvoir complètement reconfigurer les machines en fonction des besoin de leur expérience.*
- *Démarrage en Oct. 2003, première machine hiver 2004, premiers runs début février 2005 (1250 CPUs), actuellement en phase de calibration (prévu : 3 mois, expériences à blanc pour valider l'utilisation et la reproductibilité), prévision : 2500 CPUs en 2006, 5000 CPUs en 2007.*
- *Financement ministériel ACI est terminé. Contacts en cours avec les organismes pour pérenniser l'infrastructure.*
- *Sécurité : Grid5000 permet aux utilisateurs de reconfigurer complètement les systèmes (y compris le kernel) avec le risque d'introduire des trous de sécurité utilisables pour des attaques DoS ou autres ou des impacts sur la sécurité locale.*

- *Choix du confinement : accès via gateway et strong authorization, pas de possibilité de sortir. Le confinement est incontournable en l'absence de middleware.*
- *Besoin de maintenir une possibilité de connexion avec l'extérieur pour des collaborations particulières.*
- *Interconnexion des sites via des liaisons RENATER dédiées (différentes de la liaison RENATER ordinaire du site) et MPLS niveau 2.*
- *Contrôle de la plateforme: Utilisateurs doivent pouvoir déployer les piles logicielles dont ils ont besoin.*
- *Possibilité de reboot/restart à distance.*
- *Possibilité d'exécution « pas à pas » : utilisation du checkpoint/restart.*
- *Gateway traduisant des ordres remote en actions locales.*
- *Une machine par site ayant une configuration 'inattaquable' (toujours possible de revenir à un état connu) et capable de redémarrer l'ensemble de la config locale.*
- *modes d'utilisations : shared, reserved (utilisateurs peuvent remplacer l'OS sur les machines réservées), batch (usage exclusif), calendrier permettant de switcher d'un mode à l'autre les ressources.*
- *Besoin de prise en compte d'usages locaux spécifiques.*
- *Pas de quota d'utilisation par groupe de recherche actuellement : sera probablement nécessaire à l'avenir.*

® *Projet de plateforme européenne de recherche sur les grilles.*
® *Actuellement surtout du financement pour les contacts entre plateformes expérimentales (réseau d'excellence CoreGrid, animé par l'INRIA).*

2. Activités Grid dans STIC

- *Protocoles et logiciels pour réseaux très haut débit (40 Gb/s) :* ***ENS Lyon****. Concerne aussi les réseaux intra machine : myrinet, infiniband.*
- *Adaptive Online Data Compression : niveau de compression adapté en fonction de l'encombrement réseau:* ***LORIA****.*
- *Gestion de données et sécurité : transfert haute performances (parallélisation), certification de résultats (détection probabiliste,* ***IMAG****), authentification LDAP distribuée.*
- *High perf communication framework for Grid (****LABRI*** *Bordeaux,* ***IRISA*** *Rennes).*
- *Détection de fautes dans les grilles (Regal :* ***INRIA****/LIP6) : infrastructure permettant de prédire l'état d'une machine pour construire des outils de consensus permettant la réplication de services state full.*
- *Kadeploy2 : outil de reconfiguration rapide d'un grand cluster (de l'OS à l'application) (****IMAG****).*
- *V-clusters : problématique de machine virtuelle appliquée aux grilles. Donnée un environnement grille par utilisateurs. Isolation performances et sécurité (****LRI****).*
- *Single System Image OS for clusters : Kerrighed (****IRISA****).*

- *Gestionnaire de resources : OAR **(IMAG)***
- *XtremWeb : middleware for desktop grid (**LRI**).*
- *GRID-RPC / DIET (Distributed Interactive Engiennering Toolbox) (GRAAL, **INRIA**).*
- *JuxMem : Grid Data Sharing Service (Jxta + DSM) : persistence, data consistency, data localisation, scale 1K to 10K nodes (**IRISA**).*
- *MPICH-V (LRI) : multi protocols fault tolerant MPI : MPI ne fournit rien pour la tolérance au pannes (avec des 10K machines, pannes très fréquentes : qqs minutes à qq heures). Couche en dessous de MPI qui fournit le service de tolérance aux pannes avec différentes stratégies transparentes pour l'utilisateur.*
- *Parallel Corba (**INRIA/IRISA**) : CORBA performant en environnement distribué sans changement de l'ORB.*
- *ProActive : outils de programmation parallèle et distribué en JAVA (**IRISA**). Fournit un paradigme MPI en environnement Java.*
- *Ordonnancement de tâches // sur grille (**LORIA**) : déterminer meilleure allocation de ressources et meilleur placement pour un ensemble de tache. Basé sur les graphes de tâches.*
- *Algorithmique et ordonnancement pour plateformes hétérogènes et distribuées (GRAAL**, INRIA**).*

® *STIC a une connaissance et du recul sur l'ensemble de la pile logicielle.*
® *Leadership européen sur la recherche en grille.*
® *Mise en place du plus grand instrument d'étude des grilles : Grid5000.*
® *Contacts avec Japon et US.*

C. Discussions sur les sujets d'intérêt commun

- *Développement Middleware : pas en France, mais on pourrait quand même. Pourquoi pas?*

®*Guy : les partages de responsabilité se sont faits par paquets. Un optimum entre 1 et 50 paquets = un petit nombre. Dans EGEE, 3 pôles middleware : CERN, Italie, UK, dans la suite de EDG et donc, pas la France. En France, on n'avait pas une masse critique de compétences et la communauté STIC ne s'est guère manifestée, aussi ne s'est-on pas battu pour rejoindre le club middleware.*

Chaque pôle s'est spécialisé (depuis la période EDG) :
CERN = management données
Italie = RB (Resource Broker)
UK = stockage de masses et système d'information
2 autres pôles se sont créés avec EGEE :
- *sécurité : pays nordiques*
- *qualité et contrôle : France (STIC → UREC, IN2P3)*

® *DL : les problèmes d'accès et transport de données sont une piste importante.*

- *Pourquoi peu de collaborations. Jusqu'à présent ?*

®Guy : le choix initial de Globus a déplu à STIC, qui était sur des alternatives comme dans le cas d'ETOILE. En UK au contraire, au démarrage d'E-science, on a dit : on finance ceux qui travaillent autour de Globus. Et donc rien d'étonnant au résultat. En contrepartie, la communauté STIC France est très présente dans COREGRID.

- *B. Plateau - Un cadre et des sujets :*
  - identifier les verrous technologiques : gestion de données mais il y en a d'autres.
  - aspect grille de production; STIC dans grid5000. Echanges sur des projets communs.

®FLD/FE : ce serait bien d'aller au CERN pour travailler sur le middleware. Un CR1 basé au CERN. L'autre CR1 localisé peu importe où dans un labo IN2P3.
®DL : Outre le fait que parachuter quelqu'un au CERN comporte des risques, passer de la recherche à la production nécessite un partenariat et des supports locaux de développement qu'on ne trouve pas n'importe où. Quelques laboratoires de l' IN2P3 ont ces ressources qu' il faut exploiter.
®Guy : on avait parlé de thèses en co-tutelle. Un étudiant, dirigé par Cécile Germain et le LRI, va démarrer au CERN sur l'étude de la grille comme objet informatique (macroscopique et microscopique).
®FE : Bourse CIFRE à Marseille, sur les grilles, étudiant ssu de l'école ingénieur d'informatique.
®DL : Bourse CIFRE en discussion pour un projet de grille industrielle, hébergé en partie au CC.

- *Fabio : les grands domaines de collaboration que je vois:*
  - développer du MW en France ou rejoindre – temporairement - une équipe ailleurs (Cern, Italie).
  - gridifier les applications : le domaine applicatif = les couches proches des utilisateurs, non MW.
  - recherche sur l'infrastructure des sites : on y a plus d'autonomie sur les solutions. Comment gérer et distribuer les masses de données. Comment exploiter.
  - les technologies de machines virtuelles pour satisfaire 40 expériences et leurs environnements spécifiques.

®Fede : il reste de la place partout
®Rolf : dans le déploiement du MW, il y a des étapes. Les Italiens le pratiquent : il y a le core MW et les aménagements.
® Jouvin: aide à la gridification, au portage des applications. GridRPC, web services
® Linglin:
La gestion, le transport et la distribution de données sont un vaste chantier possible.
On peut regarder nos besoins et faire notre shopping list, on peut aussi faire l'inverse.
Si je reprends les sujets listés par Franck, la plupart nous intéressent :

- *protocoles et logiciels pour réseaux très haut débit Pascale Primet*
- *compression de données ? oui, si on a des goulots d'étranglement*
- *gestion de données et sécurité Denneulin*
- *détection des fautes dans les grilles Pierre Sens LIP6*
- *Kadeploy2 IMAG*

- *Machines virtuelles Cappello*
- *Kerrighed, single system Image OS for clusters IRISA*

Etc.. On peut ainsi se faire un tableau des sujets HEP-LCG-France versus sujets STIC-INRIA et voir dans quelles cases on peut mettre un nom (les futurs CR1 ou d'autres, côté STIC) et une collaboration possible.

- *Cappello*

On a une collaboration avec le LAL. Incompréhension au début. On a seulement des prototypes, on teste ensemble, puis on met en production. On ne tenait pas assez compte du côté production et le LAL ne réalisait pas que nos produits n'étaient que des prototypes.

Conclusions:

- *DL : Faire une matrice des sujets STIC/INRIA versus HEP/LCG-France*
- *Thème organisationnel ? Un GDR ? Guy pense plus à un forum CNRS qu'à un GDR, déjà un peu lourd.*
- *Guy : L'aspect évaluation est important aussi : personne n'évalue EDG ou EGEE en France. Les projets interdisciplinaires n'ont pas d'évaluation. Penser à un espace d'évaluation.*
- *FLD : on y pense*
- *Créer un site web de collaboration, de forum ? oui, il suffit que quelqu'un s'en occupe.*
- *Passage de LCG au CS de l'in2p3.*

®*Les directeurs de labo peuvent envoyer un complément avec la stratégie.*

®*Rédiger 2 pages de résumé, qui seront envoyés à la C07 et passés aux candidats CR1.*

®*Inciter une collaboration entre LCG-France et GRID5000. Ceci peut d'ores et déjà se mettre en place.*