GAIA

Grille : Applications, Instrumentation, Analyse

La grille : un moyen et un objet d'étude pour les sciences expérimentales

June 13, 2005

1 Introduction

Les technologies de grilles de calcul et de stockage ont aujourd'hui atteint un niveau de maturité suffisant pour permettre l'établissement d'infrastructures de production comme l'a démontré le projet Européen EGEE (http://www.eu-egee.org/) en déployant avec succès une grille d'une dimension inégalée (12000 processeurs et 5 PétaOctets de stockage répartis dans 130 noeuds de ressources informatiques sur 3 continents reliés entre eux par des réseaux à très haut débit) utilisée quotidiennement par différentes communautés d'utilisateurs.

Cette infrastructure européenne constitue un véritable centre de calcul virtuel, distribué géographiquement et ouvert à tous types d'applications scientifiques. Ce centre permet de fédérer les communautés scientifiques en répondant à leurs besoins grandissants en terme de stockage, de traitement et d'analyse de leur données.

Cependant, et contrairement à la plupart de ses voisins Européens, la France ne s'est pas encore dotée d'une grille nationale utilisable en production par ses scientifiques.

Le projet **GAIA** (*Grille : Applications, Instrumentation, Analyse*) vise à installer en France une infrastructure de grille de calcul, la remettant au niveau de ses voisins européens et offrant des fonctionnalités novatrices.

Concrètement, le quadruple objectif de *GAIA* est :

- la mise en place opérationnelle d'un centre de calcul virtuel (infrastructure matérielle et logicielle) à partir des ressources établies dans les centres de calcul participants ;
- la mise en service de ce centre pour une centaine de chercheurs, représentant une dizaine de domaines applicatifs (tels que ceux s'appuyant actuellement sur EGEE) pour la production de résultats scientifiques inaccessibles jusqu'à ce jour ;
- l'instrumentation de la grille, offrant à la communauté informatique un observatoire unique d'un système distribué à très grande échelle, et en condition de production ;

- l'exploitation des masses de données recueillies par l'observatoire de la grille, à fins de modélisation et de contrôle de ce système complexe.

GAIA cherche ainsi à établir une coopération étroite et gagnant-gagnant entre les communautés des sciences expérimentales, utilisatrices de la grille, et la communauté de recherche en informatique, intéressée à la production de grilles avancées.

Cette coopération sera mise en place selon trois axes principaux :

Construire Renforcer l'infrastructure matérielle de la partie française de la grille EGEE, et créer les conditions de déploiement à grande échelle de cette infrastructure. Cet axe, visant à rendre l'infrastructure accessible à des communautés scientifiques diversifiées et permettant son usage intensif, est essentiel pour la pluri-disciplinarité du projet ;

Observer Instrumenter finement l'infrastructure, notamment les aspects réseaux. Cet axe vise à fournir à la communauté informatique les données nécessaires pour modéliser, comprendre et analyser *in vivo* une grille de production.

Optimiser Exploiter les données recueillies par instrumentation, pour identifier les limitations opérationnelles de l'infrastructure (notamment relatives à la gestion des volumes d'information mis en jeu par les sciences expérimentales). Cet axe, intéressé à la fouille des données de la grille, repose sur l'apprentissage statistique. Son but est de mettre la communauté de recherche en informatique en mesure de concevoir la prochaine génération des grilles.

La demande financière est de 2 MEuros, dont 50% serviront à recruter des chercheurs en informatique et 50% à financer le matériel nécessaire.

Domaine d'étude en informatique

2 Un observatoire de grille

Une grille de calcul et de stockage constitue un système complexe, combinant un ensemble de composants, réseaux, processeurs et accès aux données.

La très grande taille et le caractère distribué d'un tel système complexe limite une approche analytique du fonctionnement de la grille : son état à un instant donné n'est connu que de facon approchée ; son utilisation en production conduit à l'apparition de conditions de charge non contrôlées et parfois imprévues.

La grille est ainsi vue comme un objet d'étude : il s'agit d'un phénomène, certes artificiel, mais dont les lois de comportement ne sont pas connues. L'étude de ce phénomène vise deux buts, respectivement à court et moyen terme : l'aide à la gestion et l'aide à la conception.
En gestion de production tout d'abord, l'étude cherchera à identifier et à récupérer les jobs en échec, ainsi que la dynamique de l'usage des ressources.
En conception, l'étude fournira des modèles réalistes pour le comportement d'un système distribué à très grande échelle ; de tels modèles guideront la conception de politiques d'allocation optimale des ressources, ainsi que l'algorithmique applicative.

Cette activité a une complémentarité évidente avec la collaboration GRID'5000, et vise à interagir fortement avec elle.

2.1 Monitoring et recueil de traces

Objet : Constitution d'un observatoire de la grille.

[...]

Cette activité recherchera une collaboration forte avec le NoE CoreGrid, en particulier son Institute on Grid Information and Monitoring Services.

2.2 Fouille des données de la grille

Objet : Des données d'observation au(x) modèle(s) de la grille.

La complexité des composants individuels de la grille, de leurs interactions, et la structure fondamentalement décentralisée de l'utilisation de la grille, rendent réaliste une approche de type *boite noire* pour la modélisation de ce système complexe.

Une telle modélisation requiert un dialogue pluri-disciplinaire approfondi, intégrant :

- l'expertise de la communauté STIC/systèmes parallèles et distribués, relative à l'évaluation de tels systèmes. Le point clé concerne l'identification des descripteurs et paramètres d'ordre du système, commandant son passage à l'échelle.

- l'expertise de la communauté IN2P3, relative à l'acquisition et la gestion de grandes masses de données (particulièrement spatio-temporelles), ainsi que les paramètres d'ordre décrivant les applications traitées et leurs spécificités opérationnelles.

- l'expertise de la communauté STIC/apprentissage statistique et fouille de données, permettant l'exploitation des données recueillies à fins de modélisation robuste et flexible, ainsi que l'exploitation d'un tel modèle à fins d'identification des conditions critiques d'utilisation du système.

3 Accès et traitement des données

[...]

4 Fouille de données scientifiques

Il importe de noter que le système complexe constitué par la grille, d'une part, et d'autre part les systèmes complexes étudiés en physique (e.g. l'expérience Auger), en biologie (e.g. étude du comportement dynamique des macro-molécules biologiques) ou en médecine (e.g. analyse d'images médicales pour le dépistage ou le suivi d'évolution de tumeurs), présentent des caractéristiques similaires et des structures comparables.

En particulier, ces problématiques ne disposent pas d'une *vérité terrain*, interprétation exacte unique qui puisse servir de référence pour calibrer les algorithmes d'analyse, de fouille et d'interprétation des données. Le problème devient en conséquence de déterminer l'espace de recherche (classes d'interprétations - problème de sélection de modèles), les critères d'intérêt (qu'est-ce qu'une interprétation intéressante), et un mode d'interaction pertinent (trouver un compromis entre le temps nécessaire pour proposer une solution, et la qualité de cette solution, qui permette à l'expert d'entretenir un dialogue satisfaisant avec le système - problématique des algorithmes any-time).

Un aspect fortement innovant du projet *GAIA* concerne ainsi la réalisation de fonctionnalités de modélisation, également pertinentes pour les utilisateurs *et pour les administrateurs* de la grille.

Nous situerons brièvement les dimensions stratégiques de la recherche proposée (applicative, fondamentale et prospective), avant de décrire de manière détaillée les fonctionnalités qui seront étudiées et réalisées dans le projet.

- Dimension applicative.
  Au delà des algorithmes répondant aux problématiques déjà identifiées par les partenaires, qui seront détaillées ci-dessous, le projet *GAIA* cherchera à mettre à la disposition de la communauté des utilisateurs un *Centre d'Expertise*, en s'inspirant d'initiatives comparables au niveau européen.

Ce Centre d'Expertise aura pour fonction de diriger efficacement les utilisateurs vers les experts et les approches les plus appropriées du domaine de l'apprentissage statistique et de la fouille de données.

Il favorisera également la veille et la réactivité scientifique, en facilitant la détection des grandes tendances et les évolutions des applications considérées.

- Dimension fondamentale.
  Un axe de recherche fondamentale, en coopération avec le réseau d'excellence PASCAL (http://www.pascal-network.org), s'intéresse aux impacts du modèle de calcul et de stockage proposé par les grilles sur l'apprentissage statistique et la fouille de données, du point de vue d'une part du passage à l'échelle des algorithmes, d'autre part des opportunités nouvelles apportées par une puissance de calcul virtuellement illimitée.
  - estimation : sélection de modèles, caractérisation des distributions des données, étude de convergence**?**;
  - algorithmes any-time : intervalles de confiance, convergence en horizon fini
  - apprentissage incrémental (on-line) : rafraîchissement des résultats en fonction de l'évolution des données
  - détection d'anomalies/ruptures.

  Les avancées fondamentales à rechercher concernent**?**: - La confrontation des bornes statistiques existantes au niveau non-asymptotique (bornes généralement estimées trop conservatives) et asymptotiques (estimées trop optimistes), avec la convergence empirique observée sur des données de très grande taille.
  - La qualification des algorithmes existants en fonction des paramètres d'ordre des problèmes (volume, distribution des données, critères de spécificité ou de sensibilité cherchés), en vue de déterminer les algorithmes les plus performants dans une région des paramètres d'ordre donnée.
  - L'étude théorique de l'apport des portages de méthodes sur des architectures très distribuées, prenant éventuellement en compte le type de distribution des données pour proposer des bornes affinées.

- Dimension prospective.

  Les données générées et traitées dans le cadre des sciences physiques se fondent majoritairement sur des structures spatio-temporelles. Cependant, l'exemple de la bio-informatique et de la médecine montent la nécessité de faire face à des données structurées complexes. Il nous paraît absolument nécessaire d'anticiper la montée, en complexité et non pas seulement en volume, des données pertinentes dans la masse d'expériences que produira ou traitera une grille. Les annotations des données (calibration, modes de production, environnement logiciel?) sont effectuées le plus naturellement sous forme textuelle, quel que soit leur mode de conservation ultérieur. L'utilisation de telles annotations contrôle largement la

possibilité de partager les données pour certaines applications (imagerie médicale), et leur réutilisabilité à long terme dans tous les cas. Un axe de recherche prospective concernera donc la jonction avec les standards de fait des données structurée (XML).