Infrastructures Grille

C. SAGUEZ
Ecole Centrale de Paris
Grande Voie des Vignes
92295 Châtenay-Malabry Cédex
christian.saguez@ecp.fr

Dans cet article, nous présentons quelques grands enjeux associés au développement de la technologie GRID. Après avoir exposé les besoins auxquels doivent répondre les nouvelles architectures informatiques en terme de capacité de stockage, de traitement des informations et de puissance de calcul, nous décrivons les principes de base du concept de GRID. Dans une dernière partie, nous illustrons ceux-ci par la description de trois grands projets actuellement en cours : les projets européens DATAGRID et EUROGRID et la plate forme RNTL e-Toile.

Introduction

Un des éléments importants dans l'évolution des technologies de l'information, lié à la convergence des télécommunications, de l'informatique et de l'audiovisuel, est que l'ensemble de l'information dont nous disposons ou que nous générons (données, textes, images, vidéo…) est maintenant numérique. Ainsi nous nous trouvons face à une explosion de la quantité d'informations directement accessible via les réseaux de communication. La maîtrise et le traitement de ces informations est devenu un enjeu stratégique de tout premier ordre pour toutes les organisations tant industrielles que publiques ou associatives.
Face à une concurrence de plus en plus grande, les entreprises doivent faire preuve du maximum de réactivité dans la conception de nouveaux produits ou services, l'adaptation de leurs outils de production ou la mise en œuvre de solutions spécifiques à la demande. Dans ce contexte, les outils informatiques de modélisation et de simulation sont à la base de leur compétitivité.
Ces deux remarques démontrent l'importance de disposer, dans les meilleures conditions, de moyens adaptés pour le stockage, la transmission et le traitement de l'information et pour la mise en œuvre de la puissance de calcul nécessaire.
Or, aujourd'hui, on constate que, suivant la loi de Moore, la puissance de traitement des processeurs croit très rapidement et est facilement accessible au travers de PC ou de stations de travail. Ces puissances de traitement sont globalement extrêmement sous-utilisées, souvent à moins de 15%, et ainsi les entreprises ou organisations disposent de réserve importante de puissance qu'elles souhaitent utiliser au moins en rationalisant leurs investissements informatiques.
Par ailleurs, des réseaux de communication à haut débit ont été installés ou sont en cours de déploiement, à l'exemple de l'initiative GEANT permettant de disposer, au niveau des organismes de recherche, d'un réseau à 10 Gigabits.
Cette organisation des moyens disponibles et des besoins en terme de puissance de traitement et de capacité de stockage a été à l'origine du concept de grille.

Dans cet article, après avoir, dans une première partie, présenté les besoins et enjeux au travers des grilles d'information, des grilles de calcul et de l'ingénierie concourante, on décrit le concept de grille de calcul et les défis technologiques associés. La mise en œuvre de cette technologie est illustrée par trois exemples de projets en cours de réalisation : les projets européens DATAGRID et EUROGRID et la plate forme RNTL e-Toile.

I – Besoins et enjeux

I-1 Grille d'information

La maîtrise et la disponibilité des informations par l'accès à de très grandes bases d'informations réparties géographiquement et à des capacités de traitement adaptées (data-mining par exemple) est un enjeu essentiel pour toutes les organisations. Ces informations peuvent être d'origine et de type très variés. A titre d'exemple, on peut indiquer les bases de données expérimentales soit issues de grandes expériences scientifiques (physique des particules, biologie, observation spatiale,…) soit résultant d'actions spécifiques telles que les données commerciales ou économiques, les bases de données textuelles provenant de publications ou de centre de documentation, maintenant largement interconnectés, les bases de données techniques,…
Les quantités de données stockées ou générées peuvent atteindre maintenant des grandeurs de l'ordre du petabyte ou plus, nécessitant des capacités de stockage

inaccessibles pour un organisme seul. Par ailleurs, il apparaît souvent inutile de devoir rapatrier en un même lieu l'ensemble des données alors que seule une faible partie est nécessaire pour le traitement envisagé.
Pour tous ces points les infrastructures Grille apportent une solution efficace. Par exemple l'initiative SETI, pour la recherche de signaux extra-terrestres, a réussi, en récupérant sur des milliers de PC les cycles processeurs inutilisés, à générer une puissance de 33,79 Téraflops. Le projet Decrypthon a permis de même l'établissement de la carte de 500 000 protéines du vivant.. Le projet DATAGRID présenté par la suite au paragraphe III illustre également ce point.

I-2 Grille de calcul

Les techniques de simulation et les outils de conception de nouveaux produits ou systèmes apportent une différenciation concurrentielle forte en permettant réduction des coûts et du temps de mise au point. Avec les outils de PLM (Product Lifecycle Management) on s'oriente vers le concept d'entreprise numérique. La complexité des systèmes considérés, la nécessité de coupler plusieurs phénomènes (thermique, structure, fluides…) ou de prendre en compte des géométries 3D complexes induisent des besoins de puissance de calcul pouvant atteindre plusieurs téraflops. L'objectif est de résoudre soit des systèmes couplés d'équations aux dérivées partielles par diverses méthodes numériques (différences finies, volumes finis, éléments finis, Monte-Carlo, méthodes particulaires…) soit de grands systèmes algébro-différentiels et de disposer de capacités de pré et post traitement adaptées intégrant des techniques de réalité virtuelle.
Parmi les grands défis nécessitant de telles puissances de calcul, on peut citer les modèles météo et les études sur le changement climatique global, les simulations ab-initio de matériaux dans les domaines scientifiques, les simulations et outils de conception en aéronautique, automobile, chimie ou nucléaire pour ne citer que quelques exemples industriels, sans oublier les domaines de la finance, notamment avec les calculs de risque et de la biologie.

Cette problématique sera illustrée au chapitre III par le projet européen EUROGRID.

I-3 Ingénierie concourante

De plus en plus souvent, dans le cadre de grands projets, les entreprises doivent collaborer entre elles à partir de sites géographiquement répartis, mettre en commun des bases de données ou de connaissances, des descriptions 3D d'objets complexes, des outils de modélisation ou de simulation… Ces problématiques se retrouvent en particulier dans des secteurs tels que l'aéronautique, l'espace ou l'automobile entre maître d'œuvre, équipementiers et fournisseurs. L'objectif est, par une mise en commun en permanence de moyens, de données et d'outils dans le cadre d'un système informatique virtuel (une grille), de réduire considérablement les temps et coûts de conception. Ainsi le GRID apparaît comme une architecture des systèmes d'information particulièrement bien adaptée aux nouvelles organisations du travail et des entreprises.

II – Le concept de grille

Le concept de grille a été mis au point pour répondre à ces différents besoins en optimisant au maximum l'utilisation des moyens de traitement et de stockage disponibles.
Il a pour objet de fournir, de manière transparente et sure, à des communautés d'intérêt (organisation virtuelle), l'accès à des moyens de traitement et de stockage hétérogènes distribués géographiquement, permettant de disposer de capacités difficilement accessibles , individuellement ou incompatibles avec les propres moyens financiers d'une structure telles que des puissances de calcul de plusieurs téraflops ou des capacités de stockage de l'ordre du petabyte.

Un système de grille repose sur les éléments suivants :

- des moyens matériels, systèmes de traitement et de calcul (PC, stations de travail, clusters,…) et systèmes de stockage
- des mécanismes de communication par des réseaux haut-débit reliant les différents centres
- des services GRID réunis au sein d'un Middleware
- des boites à outils génériques (outils de visualisation, bibliothèques de données …)
- des logiciels d'applications spécifiques adaptés à l'architecture grille.

Le Middleware est la brique de base regroupant l'ensemble des éléments logiciels pour la mise en œuvre d'une grille. Il comprend notamment les fonctions suivantes :

- le partage et l'allocation des différentes ressources de la grille suivant des critères techniques de performance, mais également des critères économiques et d'éventuelles contraintes utilisateurs
- l'exécution, l'ordonnancement et l'administration de la grille, intégrant toutes les fonctions de monitoring et de gestion en termes de facturation notamment
- l'ensemble des procédures de sécurisation de la grille, notamment les outils d'authentification des utilisateurs, la gestion des restrictions d'accès, la confidentialité des données et des résultats,…
- les outils collaboratifs permettant aux divers acteurs de travailler ensemble et d'échanger documents, données, logiciels, résultats…, en garantissant la cohérence de ceux-ci au cours de l'ensemble des manipulations
- les outils d'évaluation des performances et de mesure de la qualité de service
- les outils de développement et les interfaces utilisateurs pour le déploiement des différentes applications.

Ces Middlewares s'appuient sur des protocoles standards de l'Internet tels que FTP (File Transfer protocol), LDAP (Ligth Directory Acccess Protocol), HTTP (Hypertext Transfert Protocol). Parmi les Middlewares les plus utilisés actuellement il faut citer les outils GLOBUS, LEGION et UNICORE.
Le développement de ces Middlewares fait l'objet d'une très grande activité tant au niveau recherche qu'industriel notamment dans le cadre du GGF (Global GRID Forum). Une initiative très importante OGSA (Open Grid Services Architecture) a été récemment lancée pour assurer la convergence entre les technologies GRID et les technologies Web services, notamment WSDL (Web Services Description Language) et ainsi définir des standards pour des systèmes de services distribués pour des services sophistiqués correspondant aux nouveaux modes d'organisation des entreprises et des organisations.

III – Quelques exemples

Aujourd'hui de très nombreux projets de grille sont en cours de développement dans le monde. Il s'agit d'un axe majeur des actions proposées dans le cadre du 6ème Programme Cadre de Recherche et de Développement de l'Union Européenne et de nombreuses expérimentations industrielles sont en cours.

Dans ce paragraphe nous présentons trois initiatives importantes de grille actuellement en Europe.

III-1 Le projet européen DATAGRID

Le projet DATAGRID est un projet européen (5ème PCRD) pour la mise en place d'une grille de calcul et de traitement de données pour l'analyse de données issues de grandes expériences scientifiques.
Le projet, issu d'une initiative du CERN, réunit 6 partenaires principaux (CERN, CNRS, ESRIN, INFN, NIKHEF, PPARC) et 15 partenaires associés dont trois industriels (CS Communication et Systèmes, DATAMAT et IBM-UK) ;
Le projet a pour principaux objectifs :

- de développer un middleware open-source fondé sur l'outil GLOBUS ,
- de déployer des testbeds à grande échelle,
- de valider le concept de grille sur différents démonstrateurs.

Trois grandes applications ont été retenues pour valider le projet :

i) Physique des particules

L'objectif est de traiter les données qui seront fournies par les expériences avec le nouveau collisionneur LHC en cours de montage au CERN.
Le LHC est un accélérateur permettant la ,collision entre protons et ions à des énergies jamais atteintes aujourd'hui, ceci devant recréer les conditions initiales de l'univers après le « Big Bang ». Les différents détecteurs du LHC fournissent une énorme quantité de données (environ 3,5 Petabytes par an) données devant être stockées, accessibles et traitées par la communauté mondiale des physiciens des particules regroupant en Europe plus de 250 instituts et dans le reste du monde plus de 200 instituts.
Cette problématique requiert des besoins de stockage et de traitement inaccessibles à une seule organisation et seule une architecture de type grille peut y répondre.

ii) Bio-informatique

Le domaine de la bio informatique est caractérisé par son interdisciplinarité (biologie moléculaire, calcul scientifique, technologies de l'information, gestion de données…)
Il manipule de très grandes bases de données réparties à travers le monde – bases de données connaissant une croissance exponentielle depuis plusieurs années.
Les architectures GRID doivent permettre de répondre aux problèmes d'organisation des données d'accès à celles-ci (en prenant compte les aspects de distribution et de replicas) et de traitement de ces données (data-mining). Un premier prototype regroupant sur ce concept huit sites européens est actuellement opérationnel.

iii) Observation de la terre

L'objectif de cette application est le stockage, la distribution et le traitement des données fournies par les satellites d'observation ERS 1/2 et ENVISAT. A titre d'exemple le satellite d'observation de la terre ENVISAT fournira un volume de 500 Gbytes de données par jour. Le concept GRID doit permettre d'accroître la disponibilité de ces données, de fournir un moyen de retraiter les archives de données et d'autoriser l'utilisation de traitements complexes de fusion de données , d'analyse et de modélisation de celles-ci. Un premier système fonctionne sur plus de 10 sites en Europe regroupant plus de 4000 chercheurs.

III-2 Le projet européen EUROGRID

Le projet EUROGRID est un projet européen (5ème PCRD) pour le développement des technologies grille autour du calcul haute performance. Le projet dont le coordinateur est la société allemande PALLAS, regroupe 6 centres de calcul haute performance dont le centre IDRIS du CNRS et deux grands utilisateurs (GIE, EADS, CCR et Deutsche Wetterdienst)
Le projet s'appuie sur le Middleware UNICORE de la société PALLAS. Les principales applications mises en œuvre sont :

- la simulation en recherche biomoléculaire,
- les modèles de prédiction météorologique, notamment les modèles atmosphériques régionaux,
- le couplage de codes d'IAO, notamment pour des applications dans le domaine de l'aéronautique,
- la simulation multiphysique.

Au niveau européen il convient également de signaler l'initiative GRIDSTART dont l'objectif est de consolider les avancées européennes dans les technologies GRID et de stimuler le développement des grilles dans tous les domaines scientifiques, industriels et grand public. Ils associent la plupart des projets européens (AVO, CROSSGRID, DAMIEN, DATAGRID, DATATAG, EGSO, EUROGRID, GRIA, GRIDLAB, GRIP).

III-3 La plate-forme RNTL e-Toile

Le projet e-Toile est une plate-forme financée par le RNTL. Il réunit de grands acteurs français de la recherche et de l'industrie dans le domaine des grilles (CNRS, INRIA, ENS Lyon, CEA, EDF, France Télécom, SUN France, CS-Communication et Systèmes). Les objectifs du projet peuvent se résumer comme suit :

- mettre à la disposition de la communauté scientifique française une plate-forme d'expérimentation d'une grille de calcul s'appuyant sur RENATER et VTHD. Cette plate-forme doit permettre de valider les travaux de recherche sur le middleware et de tester des applications dans les domaines du calcul intensif et du traitement de grandes quantités de données ;
- développer un middleware prototype intégrant les travaux les plus récents en réseaux actifs, DSM (Distributed Shared Memory), allocation de ressources et sécurité. Ces travaux se font en étroite complémentarité avec l'ACI GRID du ministère de la recherche

- favoriser la valorisation de cette technologie dans les grands domaines scientifiques et industriels en préparant une démarche de type GRID Service Provider.

Trois applications nécessitant l'accès à de grandes puissances de calcul seront testées :

- Des problèmes d'optimisation combinatoire à partir de la bibliothèque Bob développée par le laboratoire PRISM (université de Versailles Saint Quentin). Il s'agit de traiter des arbres de quelques milliards de sommets.
- Un logiciel de dynamique moléculaire avec potentiels empiriques lissés par des calculs ab initio et un atelier de neutronique composé d'un modèle de données métier et d'une solution générique d'enchaînement de codes de calcul développé par EDF.
- Deux applications dans les domaines de la physique nucléaire pour la simulation de données de l'expérience ALICE et des sciences du vivant pour des problèmes de dynamique moléculaire.

Conclusion

Les grilles constituent un enjeu fondamental dans l'évolution des architectures des systèmes d'information. En permettant une utilisation rationnelle des moyens informatiques et en fournissant des capacités de stockage et de puissances de calcul inaccessibles à la plupart des acteurs, ces technologies ouvrent de nouvelles voies importantes au développement des technologies de l'information et de la communication.
Elles sont très étroitement associées à l'évolution de l'organisation des entreprises en favorisant les travaux coopératifs entre équipes et sociétés géographiquement réparties et en étant à la base des futures organisations virtuelles.
Des modèles économiques adaptés sont déjà proposés par des sociétés spécialisées nouvelles, notamment dans le cadre des métiers de GRID Service Provider, intermédiaires entre demandeurs de capacités et propriétaires des moyens informatiques.
L'implication des grands acteurs du secteur des TIC et le niveau élevé des investissements engagés démontrent là aussi les espoirs importants mis dans ces technologies.
Il est essentiel que les différents acteurs français, avec une impulsion forte des pouvoirs publics, saisissent au mieux cette opportunité dans un secteur où la France dispose d'une compétence scientifique et technique très importante.

Bibliographie

- I. Foster, C Kesselman, S Tuecke « The Anatomy of the GRID » (Int. J. Supercomputer Application, 2001)
- I. Foster, C Kesselman, J.M. Mick, S. Tuecke « The physiology of the GRID »
- I. Foster, C Kesselman « The GRID : Blue print for a New Computing Infrastructure » (Morgan Kaufmann, 1999)