Faciliter l'utilisation des grilles de calcul dans le domaine biomédical: Le projet KnowARC

Henning Müller[1], Adrien Depeursinge[1], Antoine Geissbuhler[1]

[1]*Service d'Informatique Médicale (SIM)*
Université et Hôpitaux Universitaires de Genève, Suisse

Abstract

Objectives: *The availability of most medical patient data in digital form creates the possibility to use these data in various research projects to improve future care (including textual and visual information analysis and retrieval). Many medical institutions do not possess an adapted computing infrastructure or a budget for such an infrastructure. Still, many institutions have a large number of desktop PCs that could serve for biomedical research during the time they are little used without the need for expensive investments. The KnowARC project aims at building a middle ware for such as grid network.*

Methods: *This article reviews requirements for computing grids in hospital environments, particularly in the University Hospitals of Geneva, and then presents the solutions that the KnowARC project of the European Union plans to undertake to solve the current problems. Methods currently employed in common grid middleware distributions are also reviewed and compared with the goals of the KnowARC project.*

Results: *The computing infrastructure at the University Hospitals of Geneva is described as well as the domains with a necessity for computing and storage capacities. A list of requirements for a grid middleware to employ in such an environment is developed. Finally, the proposed solutions and ideas of the KnowARC project are described in detail to present the project to a larger community.*

Conclusions: *Grids networks are expected to become an important factor to supply the computational needs that several domains in biomedical research require. A continuous process will be necessary to feed in the requirements of the biomedical domain to developers of grid middleware, so the outcome is along the line of the needs.*

Keywords

Biomedical research, grid networks, supercomputing, distributed storage, distributed computing

1 Introduction

Le concept des *computer grids* ou grilles de calcul et de stockage a commencé dans les années 80 avec l'idée de partager des ressources de calcul et de stockage dans le monde entier [1]. Essentiellement dans le domaine de la physique de haute énergie, les nécessités d'infrastructures de stockage et de calcul très fortement distribué ont été définies selon les besoins des accélérateurs de particules comme par exemple le LHC (Large Hydron

Collider) au CERN[1] (Centre européen de la recherche nucléaire) qui produisent d'énormes quantités de données analysées postérieurement dans plusieurs centres.

Plus récemment l'idée du «computing on demand» ou calcul sur demande a motivé l'industrie à commencer des activités dans les domaines du grid. L'idée de créer une structure comparable au World Wide Web (WWW), non pas pour des données mais pour la force de calcul et l'espace de stockage a créé une forte fantaisie sur le potentiel de telles techniques. Malheureusement les grilles de calculs ne sont pas rapidement devenues des techniques applicables et appliquées partout. En effet, la complexité technique a souvent été un frein car l'adaptation complexe des applications à une telle infrastructure et surtout la maintenance onéreuse des infrastructures des grilles ont souvent bloqué les projets avant leur démarrage. De plus, les besoins dans les divers domaines sont très différents. La physique des particules requiert souvent des calculs de très longue durée avec des données très volumineuses ainsi que des clusters importants d'ordinateurs dédiés à ces tâches. Dans le domaine biomédical, beaucoup de tâches sont encore envisageables sur des petits serveurs ou même des ordinateurs de bureau et le but est plutôt de créer des applications interactives avec des interfaces facilement utilisables pour des données de tailles souvent moyennes. L'inexistence de clusters dans ce domaine engendre aussi que les applications sont pensées en termes d'infrastructures accessibles au lieu de ce qui aurait été possible avec des infrastructures plus importantes. De plus, pour la plupart des institutions la recherche en informatique est de moindre importance et l'accent est mis sur le fonctionnement de la structure clinique.

Depuis quelques années, des propositions pour l'utilisation des grid dans la santé sont apparues, notamment par l'association healthgrid[2] et son cercle de conférences. Plusieures applications existent pour le traitement d'images et la génomique. Pour les hôpitaux, ceci représente une technique potentielle importante pour le future [5].

2 Etat de l'art

2.1 Les grilles de calcul

Les idées de partage de ressource de calcul et de stockage distribué dans le monde entier a commencé dans les années 80 avec par exemple le projet Condor [1]. La plupart des projets ont commencé dans la physique des particules où de grands volumes de données ont toujours été produits. Des exemples de tels projets sont Unicore [2] et Globus [3]. Globus, en étant open source est devenu, avec sa multitude de fonctions et blocs, une base pour plusieurs autres projets. L'Union Européenne a aussi rapidement compris l'importance des telles infrastructures et a financés plusieurs projets comme le projet EDG (European Data Grid), localisé en grand partie au CERN et pour les besoins du CERN. Ces projets ont développé des infrastructures complexes pour le gestion d'un grand nombre d'ordinateurs distribués dans des clusters dans différents pays. Le projet qui succède à EDG[3] est maintenant EGEE[4] (Enabling Grids for E-science in Europe) qui a démarré en 2004.

La communauté NorduGrid[5] a commencé en 2000 à évaluer les besoins de la physique locale pour un middleware qui devrait connecter les ressources accessibles dans la région. Une évaluation des systèmes existants ainsi qu'une analyse des besoins ont été effectuées.

[1] http://www.cern.ch/
[2] http://www.healthgrid.org/
[3] http://www.edg.org/
[4] http://www.eu-egee.org/
[5] http://www.nordugrid.org/

Basée sur ces deux critères, une partie des systèmes existants a été reprise et de nouveaux composants ont été développés de façon pragmatique pour ce qui manquait. Le résultat est l'ARC (Advanced Resource Connector) middleware, qui est utilisé courramment depuis 2002 dans une soixantaine de sites dans le monde entier [5].

Dans le domaine de la santé l'association healthGrid a organisé des conférences sur les grilles dans la santé depuis 2003 afin de favoriser la progression du domaine. Au niveau suisse, le projet SwissBioGrid[6] a motivé des instituts à appliquer les technologies grid aussi dans la perspective de profiter des ressources mises à disposition dans le CSCS[7] (Swiss National Super Computing Center) qui a aussi choisi d'utiliser ARC comme middleware.

2.2 Les infrastructures pour le calcul dans les hôpitaux et universités

À l'opposé des infrastructures dans la physique des particules, les domaines biomédicaux ont rarement de grandes ressources de calcul directement à leur disposition pour la recherche. Des petits serveurs sont disponibles pour des projets mais rarement une infrastructure pour le partage entre entités est prévu dans les hôpitaux, universités ou administrations. D'un autre coté plusieurs centaines voir milliers d'ordinateurs sont dans les bureaux ou étages et seulement rarement utilisés pour des tâches nécessitant un calcul intensif. L'utilisation de telles ressources peut aider à résoudre des problèmes de calcul dans la recherche.

3 Matériel et méthodes

Cette article va présenter l'infrastructure informatique des hôpitaux universitaires de Genève en se concentrant sur les aspects de calcul et de stockage. Plus particulièrement, la structure pour les bases de données clinique et celles pour la recherche sont présentées. En seconde partie, les contraintes pour une implémentation et les applications sont mises en place et enfin les idées du projet KnowARC pour résoudre les problèmes identifiés auparavant sont exposées.

4 Résultats

4.1 L'infrastructure d'ordinateurs dans les Hôpitaux Universitaires de Genève

Les Hôpitaux Universitaires de Genève incluent six hôpitaux dans le territoire du canton de Genève et une à proximité du canton. En total 2'200 lits sont servis par 10'000 employés. Une architecture complexe a été mise en place pour servir le dossier patient électronique et l'archive d'images (PACS – Picture Archival and Communication System). De grands serveurs existent pour ces tâches de stockage de grands jeux de données.

Au total, il y a environ 6'000 ordinateurs à l'hôpital, dont la plus grand partie sont des machines de bureau situées dans les bureaux personnels ou dans les étages à l'hôpital. La plupart de ces machines sont utilisées pour l'affichage des données patient ou pour le traitement de texte seulement. Une partie croissante mais encore petite de ces machines sont des ordinateurs portables. Les machines ont une installation standard et sont gérées à distance pour ce qui concerne la mise à jour. Il y a également plusieurs centaines de PDAs gérés par le service central de l'hôpital. Les logiciels sont distribués de façon centralisée, permettant une rapide distribution. La sécurité du réseau est la partie la plus importante, vu que la majorité des machines contiennent des données patient sensibles. En général, les contraintes de l'utilisation de l'infrastructure sont strictes y compris la configuration des

[6] http://www.swissbiogrid.org/
[7] http://www.cscs.ch/

machines sur lesquelles les utilisateurs ne sont en général pas administrateurs. Le firewall est également configuré de façon stricte, interdisant par exemple les accès extérieurs par voie cryptée (ssh).

Un grand majorité des systèmes de bureau fonctionnent sous Windows et les serveurs en grand partie sous Solaris. Dans la partie radiologie, les ordinateurs Apple sont fréquents et quelques chercheurs travaillent également sous linux.

4.2 Contraintes d'utilisation de cette infrastructure des grilles

Pour utiliser et installer une infrastructure telle que les grilles de calcul, il faut prendre en compte un grand nombre de contraintes qui sont nettement plus strictes que celles rencontrées dans un environnement universitaire. Quelques premières contraintes ont été identifiées pour l'écriture du projet knowARC afin d'y trouver des solutions techniques aux problèmes.

- L'infrastructure existante n'est pas toujours fiable pour une utilisation dans un grid car les ordinateurs peuvent être en utilisation, éteint, des PC portables peuvent être retirés du réseaux, etc.
- L'infrastructure est extrêmement hétérogène car les PCs sont souvent renouvelés tous les 5 ans, seulement, et on trouve toutes les architectures intermédiaires.
- L'objectif principal est le fonctionnement de la clinique et non pas la recherche, aussi au niveau de l'infrastructure réseau etc. La configuration du Firewall peut être extrêmement stricte pour toute communication à l'intérieur et avec l'extérieur.
- La sécurité des données et des utilisateurs est la plus importante, donc l'accès aux données sur un ordinateur d'un processus du grid est inimaginable et doit être évité de façon technique.
- Il n'y a pas de personnel pour maintenir une infrastructure complexe pour la recherche seulement, qui ne fait pas partie de l'utilisation standard de l'institution. L'installation et la maintenance du système doit être assez simple et créée par les chercheurs eux-mêmes.
- Les ordinateurs ne sont pas dédiés au calcul, mais sont la pour d'autres fonctions, donc le middleware ne peut pas disposer de la totalité des ressources de l'ordinateur.
- La plupart des infrastructures des hôpitaux sont basées sur Windows tandis que les projets grid sont souvent sous linux ou autre unix.

Ces problèmes ont été identifiés pour les Hôpitaux Universitaires de Genève, mais ne sont pas propres à cette institution. Les universités et administrations publiques ont souvent une infrastructure comparable et une grande partie de solutions pour ces problèmes peuvent aussi aider à l'application des grilles dans des autres domaines.

4.3 Solutions envisagées par le projet KnowARC

Le but du projet KnowARC est de réunir divers partenaires dans la recherche, l'industrie et l'administration pour faciliter l'application des grids dans ces domaines. Il est important pour le développement d'un middleware d'avoir le feedback des utilisateurs directement inclus au travail des développeurs d'un middleware. Pour une partie des problèmes d'acceptance des grids déjà identifiés, des solutions ont été pensées ;

- L'utilisation d'un middleware à partir de composants qui sont déjà utilisés pour avoir des ressources à disposition directement (ARC du Nordugrid).
- Une virtualisation de la partie grid sur les machines hôtes en prenant en compte que le middleware tourne sur linux et que les machines sont installées en

majorité sous linux ; Xen et d'autres méthodes de virtualisation permettent ceci, ainsi que de gérer le problème de la sécurité, car le processus exécutant des tâches n'a pas d'accès aux autres processus de la machine.

- Un processus d'installation et de maintenance simple avec deux clics similaires à d'autres composants windows. Ceci permet de tester le logiciel et de changer rapidement une infrastructure.
- Un outil de soumission de tâches simple avec une interface graphique permettant rapidement de « griddifier » un logiciel et voir le succès de son exécution.
- Un middleware qui occupe seulement une partie des ressources permettant l'utilisation des ordinateurs pour les tâches primaires.
- La création d'un load balancing system pour gérer une infrastructure très hétérogène.

Ceci ne sont que les premières idées pour des solutions techniques destinées à aider l'application des grilles de calcul dans les domaines biomédicaux. Ceci devrait augmenter l'adoption des grilles et résoudre les problèmes de calcul parfois présent dans la recherche biomédicale.

L'objectif final de cette infrastructure est de contribuer à l'amélioration les applications biomédicales comme la recherche d'information visuelle [6] ou textuelle [7]. Ces deux domaines profitent en ce moment des petits serveurs qui permettent un indexation traditionnelle avec des attentes, ainsi des ressources plus importantes peuvent fortement aider à développer des nouvelles approches qui sont très conséquentes en terme de calcul [8].

5 Discussion - Conclusion

Cet article a présenté le projet KnowARC de l'Union Européene (FP6) qui a comme but de développer une infrastructure (middleware) grid facilement applicable et utilisable pour des non-experts. Pour ceci, une étroite coopération entre les divers partenaires est nécessaire. La première étape était déjà de définir les problèmes et les buts communs pour le développement d'un middleware, aussi décrit dans cet article. L'accent a été mis sur les partenaires dans le domaine biomédical.

Les grilles de calculs ont le potentiel de résoudre plusieurs problèmes présents dans la recherche biomédicale pour mieux utiliser une infrastructure existante et fournir en même temps une force de calcul pour la recherche, souvent négligée dans les institutions. Plusieurs points qui bloquent encore l'application des grilles dans le domaine biomédical ont été identifiés et des solutions pour les résoudre sont proposées.

Il y a encore beaucoup de travail avant d'implémenter une telle infrastructure dans un hôpital ou une université, surtout sur le plan politique. Ceci inclut une éducation des personnes qui peuvent mettre à disposition leurs machines pour la grille. Ceci a bien fonctionné dans le projet global Seti@home. Un autre objectif est politique car les problèmes rencontrés ne sont pas purement techniques.

Il est prévu comme prochaine étape d'effectuer un sondage qui explique aussi la technologie afin de sensibiliser les dirigeants. Ceci devrait améliorer la compréhension de la technique dans notre domaine et en même temps nous donner des idées concrètes pour une implémentation d'un middleware capable de surmonter les problèmes tant que possible.

Ceci devrait nous aider a créer une meilleure structure pour la recherche dans notre domaine, à savoir notamment la recherche et l'analyse de l'information textuelle et visuelle.

Remerciements

Une partie de cette recherche a été financée par le fonds national suisse de la recherche scientifique (FNRS) avec le numéro 205321-109304/1. Une autre partie a été financé par le projet UE FP6 KnowARC (IST 032691).

Références

[1] Litzkow M, Livny M, Mutka M, Condor – a hunter of idle workstations, in proceedings of the 8th international conference of distribute computing systems, 1988, pp. 104-111.

[2] Romberg M, The Unicore architecture: Seamless access to distributed resources, in proceeddings of the 8th IEEE International Symposium on High Performance Distributed Computing (HPDC), 1999, pp. 287-293.

[3] Foster I, Kesselmann C, Globus: A metacomputing infrastructure toolkit, International journal of supercomputer applications 11(2), 1997, pp. 115-128.

[4] Ellert M, Grønager M, Konstantinov A, Kónia B, Lindemann J, Livenson I, Nielsen JL, Niinimäki M, Smnirnova O, Wäänänen A, Advanced resource connector middleware for lightweight computational grids, Future generation computer systems, 2006 – to appear.

[5] Müller H, Garcia A, Vallée JP, Geissbuhler A, Grid computing at the University Hospitals of Geneva, proceedings of the 1st international healthgrid conference, 2003.

[6] Müller H, Michoux N, Bandon D, Geissbuhler A, A review of content-based image retrieval systems in medicine - clinical benefits and future directions, *International Journal of Medical Informatics*, 73, 2004, pp. 1-23.

[7] Ruch P, Automatic assignment of biomedical categories: toward a generic approach, Bioinformatics 22(6), 2006, pp. 658-664.

[8] Hegerath A, Deselaers T, Ney H, Patch-based Object Recognition Using Discriminatively Trained Gaussian Mixtures. In 17th British Machine Vision Conference (BMVC06), Edinburgh, UK, September 2006 – to appear.

Adresse de correspondance

Dr. Henning Müller

Service d'informatique médicale

Université et hôpitaus universitaires de Genève

24, Rue Micheli-du-Crest

1211 Genève 14, Suisse

Tel +41 22 372 6175 Fax +41 22 372 8680

Henning.mueller@sim.hcuge.ch