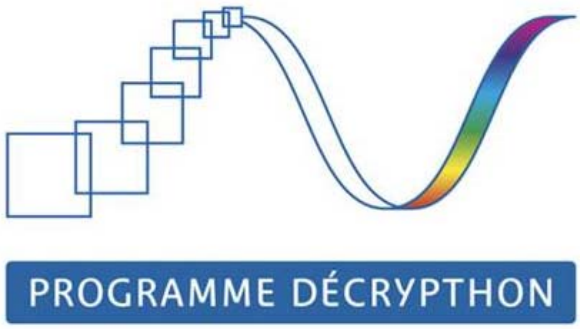
Le projet Décrypthon

Philippe d'Anfray*
CEA, Délégation ANR-CI
Centre de Saclay
91191 Gif-sur-Yvette Cedex
Philippe.d-Anfray@cea.fr

Frédéric Desprez
INRIA Rhône-Alpes, ENS-Lyon
46 allée d'Italie 69364 Lyon cedex 07
Frederic.Desprez@ens-Lyon.fr

Raphäel Bolze
ENS-Lyon
46 allée d'Italie 69364 Lyon cedex 07
Raphael.Bolze@ens-lyon.fr



Résumé

Le projet Décrypthon est le fruit d'une collaboration entre le CNRS, l'Association Française contre les Myopathies (AFM) et la société IBM. Il vise à mettre à la disposition des équipes de recherche en bio-informatique des ressources informatiques de calcul et de stockage. Ces ressources placées dans les universités constituent une plate-forme de type grille comprenant six sites -Bordeaux, Jussieu, Lille, Lyon, Orsay, Rouen- reliés par le réseau RENATER. Le Décrypthon met en oeuvre les moyens nécessaires à l'exploitation de la grille; il finance des équipes de recherche selectionnées sur appels d'offres et les accompagne pour les aspects informatiques des projets (modélisation, portage des applications sur la grille, gestion des données, etc...). Le programme Décrypthon fait aussi appel à une grille d'internautes via le WCG d'IBM. Nous présentons brièvement les caractéristiques de la grille Décrypthon basée sur l'intergiciel DIET ainsi que les applications qui y sont actuellement déployées. Ces applications utilisent de façon intensive la ressource réseau (mouvement de données, répartition des calculs, etc...).

Mots clefs

grille, applications bioinformatiques, middleware

1 Introduction

Le projet actuel « programme Décrypthon », lancé en 2004 fait suite au « Décrypthon » lancé en 2001 par l'AFM [1] avec le soutien des sociétés IBM et Genomining. Ce premier programme avait permis de mobiliser jusqu'à 75 000 internautes pour aboutir à la comparaison de 559 275 séquences protéiques. Une base de données Décrypthon était alors mise à la disposition de la communauté scientifique (Infobiogen) c'était la « *première comparaison exhaustive de toutes les protéines ou fragments de protéines identifiées chez les êtres vivants (animal, végétal, humain), disponibles à la date de janvier 2002 et provenant de 76 protéomes dont la séquence est entièrement connue* ». Le programme Décrypthon fait toujours appel aux internautes *via* le WCG [2] (World Computing Grid) d'IBM mais vise à mettre en place un ensemble de ressources propres au projet et les moyens nécessaires à leur exploitation. Ces ressources placées dans les universités constituent une plate-forme de type grille comprenant six sites -Bordeaux, Jussieu, Lille, Lyon, Orsay, Rouen- reliés par le réseau RENATER utilisant l'intergiciel DIET [3] développé au Laboratoire de l'Informatique du Parallélisme à Lyon.

2 Ressources

2.1 Ressources matérielles

Les ressources de calcul et de stockage sont constituées de serveurs installés par IBM dans les universités dans le cadre de bourses SUR (*Shared University Research)* et représentent:

- 64 processeurs Power4 et Power5 ;
- environ 480 Gflops de puissance de calcul ;
- 4+8 To de stockage.

Ces ressources sont intégrées aux centres de calcul des universités en complément des machines existantes et sont affectées en priorité au traitement des travaux selectionnés dans le cadre du programme Décrypthon. Ponctuellement, si nécessaire, la puissance déjà installée dans l'université peut s'ajouter aux ressources propres du projet.

Le détail des ressources est le suivant:

- Bordeaux 1 : 2 serveurs P575 (soit 8 procs Power5 AIX);
- Lille (USTL) : 2 serveurs P575 (soit 8 procs Power5 AIX);
- Paris 6 (Jussieu) : 1 serveur P655 (8 procs Power4 AIX) ;

*GIP RENATER jusqu'en Octobre 2007

- Paris Sud (Orsay) : 2 serveurs P575 (soit 8 procs Power5 AIX);
- Lyon (ENS) : 1 serveur P550 (4 procs Power5 AIX) et une baie de stockage de 4 To;
- Rouen (CRIHAN) : 1 serveur P550 (4 procs Power5 AIX), une baie de stockage de 8 To.

A tout cela s'ajoutent des machines d'administration (gestion de la grille, serveur d'application) :

- Paris Sud (Orsay) : 2 machines ZPRO (XEON/Linux) bi-processeur.

2.2 Ressources logicielles

Le fonctionnement de la grille était basé initialement sur le logiciel GridMP de United Devices [4]. Mais la montée en puissance, courant 2006, de la configuration matérielle (calcul et stockage) a nécessité l'étude d'autres solutions Open Source : DIET [3] et g-Lite [5].

L'intergiciel DIET [3] développé dans le projet GRAAL [6] a été retenu pour assurer la continuité du support de la grille Décrypthon. Il assure la répartition du travail sur l'ensemble des 6 centres de calcul universitaires au travers du réseau RENATER.

Parallèlement aux besoins en calcul des projets scientifiques. Un système de fédération de données biologiques hétérogènes publiques ou privées a été mis en place sur un serveur à Lyon (ENS) par l'équipe strasbourgeoise d'Olivier Poch. Ce système baptisé BIRD (Biological Integration and Retrieval Data) est utilisé pour extraire des connaissances biologiques dans les applications s'exécutant sur la grille et également utilisé pour répertorier les résultats provenant des différents projets scientifiques.

3 La grille Décrypthon

3.1 Présentation de DIET

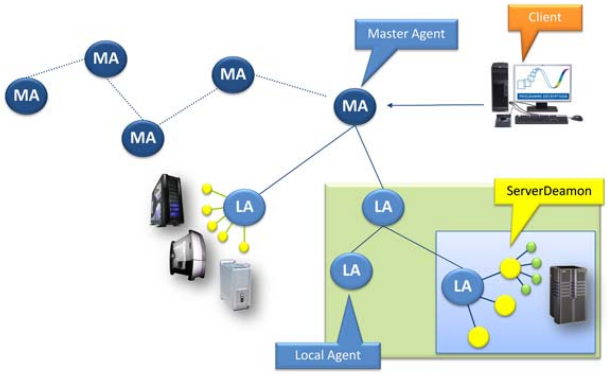


Figure1 *: Intergiciel DIET*

DIET (*distributed interative engineering toolbox*) est un intergiciel développé par l'équipe GRAAL[1] à l'ENS-Lyon. Il permet le déploiement d'une application de type Client-Serveur sur grand nombre de machines. DIET se décompose en plusieurs couches :

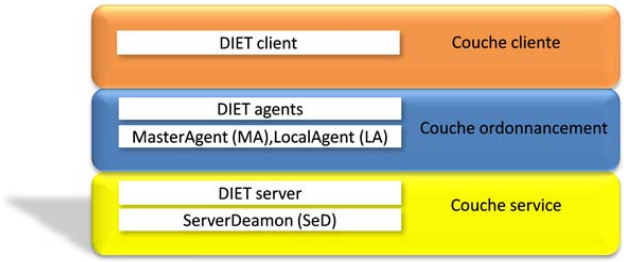


Figure 2 *: Couches de l'intergiciel DIET*

La couche service : le *Server Deamon* (*SeD*) est chargé de l'exécution du service demandé par un client. Le *SeD* évalue les performances, effectue le transfert des données et exécute les services qu'il est capable d'accomplir.
La couche ordonnancement et gestion des ressources : comprend les agents (*Master Agent* et *Local Agent*); ils sont responsables de la recherche et de l'optimisation des services gérés par la plate-forme.
La couche cliente : le client est le programme qui fait appel à DIET via une API de type gridRPC [7] pour demander l'exécution d'un service.

Le but de DIET est de rendre l'accès à des machines distantes transparent pour un utilisateur désirant l'exécution d'un service. Autour de DIET, plusieurs logiciels sont développés pour faciliter son déploiement et son utilisation. GoDIET, VizDIET, LogService et DietDashboard [3].

3.2 Fonctionnement de la grille

La figure ci-dessous montre l'architecture actuelle de la grille Décrypthon.

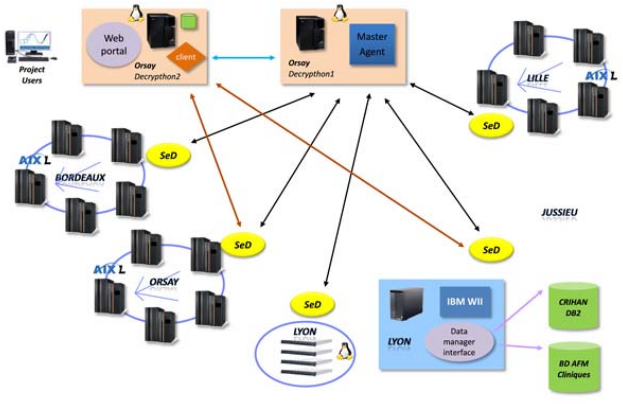


Figure 3 : *La grille Décrypthon.*

[1]projet INRIA – CNRS – ENS-Lyon – UCBL.

La grille Décrypthon s'articule autour de plusieurs éléments séparés : l'intergiciel de grille DIET, un portail web pour l'accès aux ressources de la grille, des gestionnaires locaux de ressources propres à chaque centre de calcul.

Le portail Web se trouve sur une machine dédiée à Orsay. Il contient une application web spécifique à chaque projet scientifique permettant la soumission des calculs sur l'ensemble des ressources du Décrypthon. Le portail s'appuie ensuite sur l'intergiciel DIET pour soumettre les calculs sur les ressources adaptées aux besoins de chaque application.

L'intergiciel DIET déployé pour la grille Décrypthon se compose d'un *MasterAgent* hébergé sur une machine à Orsay, et d'un *ServerDeamon* (*SeD*) lancé sur chaque frontale d'accès aux ressources de calcul des centres de calcul universitaire. Les *SeDs* sont connectés au *MasterAgent*. Ils sont destiné à collecter les informations de disponibilité et de performances des serveurs de calcul et à soumettre les travaux aux gestionnaires locaux des centres de calcul (Loadleveler, PBS, OAR ...). Il assure de plus la migration des données nécessaires pour les calculs et le stockage des résultats vers les serveurs dédiés à cet usage.

Le gestionnaire local de ressources (*batch scheduler*) : c'est le système propre à chaque centre universitaire qui assure au niveau local la répartition et le bon déroulement des calculs sur les machines scientifiques dédiées. Chaque site définit une politique d'utilisation de ses ressources, les machines dédiées au Décrypthon s'intégrant dans le parc des machines scientifiques de l'université. De cette manière toute les machines d'un centre de calcul sont partagées par les utilisateurs locaux et les utilisateurs des projets Décrypthon.

4 La grille d'internautes

Le programme Décrypthon a décidé de faire appel à la grille d'internautes *World Community Grid* pour les projets scientifiques éligibles à ce type de grille.

Ainsi lorsqu'un module de calcul d'un projet scientifique remplit les conditions nécessaires pour pouvoir être lancé sur un PC d'internaute, nous proposons l'utilisation de la grille WCG. Le projet devient alors un projet de recherche pour le WCG, et bénéficie de ce fait de l'infrastructure mise en place par l'équipe WCG pour être exécuté sur les machines des volontaires.

C'est ainsi que le 23 janvier 2007 après une phase de test d'un mois, le projet HCMD « *Help Cure Musculare Dystrophy* » a bénéficié de la puissance répartie de quelques 600 000 membres du WCG pour l'étude de la modélisation d'interactions protéine-protéine. Ce projet s'est terminé le 13 juin 2007, il a demandé plus de 80 siècles de calcul et a généré 123 Go de données, pour l'analyse d'une base de 168 protéines.

L'ensemble des résultats est actuellement en cours d'analyse afin de déterminer leur pertinence et la continuité du projet vers une base contenant plus de protéines.

5 Les projets utilisateurs

5.1 Les appels d'offre

Le troisième appel d'offre du Décrypthon a eu lieu début 2007. Le support aux projets retenus comporte :

- l'accès aux ressources (grille Décrypthon et infrastructure WCG);
- le support technologique de la part d'IBM et d'intervenants CNRS pour la validation, le portage et l'adaptation à la grille de l'application;
- un support financier de la part de l'AFM, typiquement sous la forme d'un contrat de type « postdoc » pour la durée du projet (18mois).

5.2 Les projets actuels

Actuellement cinq projets utilisent les ressources du programme Décrypthon.

MS2PH « Mutations structurales avec les conséquences sur le phénotype des pathologies humaines », projet coordonné par Olivier Poch (Illkirch, IGBMC / CNRS / Inserm) et Gilbert Deléage (Lyon, IBCP/CNRS). Il s'agit de comprendre les mécanismes qui contrôlent la fonction des protéines, de mettre au point une grille d'analyse descriptive des protéines avec des mutations connues et d'élaborer un outil prédictif pour faciliter la compréhension de ces mutations dans les maladies humaines.

« Défauts d'épissage et maladies génétiques », projet coordonné par Christiane Branlant, Fabrice Leclerc (Nancy, Laboratoire de Maturation des ARN et Enzymologie Moléculaire / CNRS), et Yann Guermeur (Nancy, LORIA / CNRS / INP Lorraine / INRIA). Ce projet propose d'analyser les liens qui existent entre les défauts d'épissage et les maladies génétiques. L'analyse des mutations dans les gènes chez des personnes atteintes de maladies génétiques et les conséquences de l'épissage devraient apporter des données fondamentales pour la compréhension de ces maladies.

« Help Cure Muscular Dystrophy », projet coordonné par Alessandra Carbone de l'université Pierre et Marie Curie, Paris (Inserm U511-Immunologie cellulaire et moléculaire des infections parasitaires-Génomique analytique). Ce projet s'attache à mettre au point des outils informatiques pour repérer à la surface des protéines des sites d'interactions, qui leur permettent d'interagir avec des protéines, de l'ADN ou des ligands, selon une méthode appelée *docking moléculaire*. Pour des centaines, voire des milliers de protéines, les calculs peuvent se compter « en siècles ».

« SpikeOMatic », projet coordonné par Christophe Pouzat de l'université René Descartes, Paris V (CNRS UMR 8118-Laboratoire de physiologie cérébrale). Afin de déceler les dysfonctionnements des neurones dans le cerveau ou des motoneurones qui commandent les fibres musculaires, on enregistre leur activité électrique sous forme de potentiels d'action. Ce projet propose, en se basant sur les probabilités, d'automatiser le tri de ces potentiels d'action en mesurant leur amplitude et en analysant leur forme.

« Gènes véritablement importants dans les processus neuromusculaires normaux et anormaux », projet coordonné par Marc Robinson-Rechavi de la faculté de Biologie et de Médecine de l'université de Lausanne (Département d'écologie et évolution). Ce projet permettra d'identifier précisément quels sont les gènes qui devraient s'exprimer ou qui s'expriment à tort dans les cellules musculaires: informations capitales pour comprendre les pathologies neuromusculaires. En pratique, ce travail revient à croiser, entre autres, les expressions de plusieurs dizaines de milliers de gènes.

6 Conclusion

Né de la synergie entre différents acteurs du monde de la recherche et de l'industrie, le programme Décrypthon répond aux besoins des scientifiques dans un cadre définit par un appel d'offre. Son dynamisme et sa force proviennent sans aucun doute du fait qu'il a su mettre en commun les atouts d'une collaboration multi-disciplinaire d'équipes de recherche à la pointe de leur domaine.

Bibliographie

[1] AFM : Association Française contre les Myopathie http://www.afm-france.org

[2] WCG: World Community Grid http://www.worldcommunitygrid.org

[3] Abelkader Amar, Raphaël Bolze, Yves Caniou, Eddy Caron, Benjamin Depardon, Jean-Sébastien Gay, Gaël Le Mahec, and David Loureiro. **Tunable Scheduling in a GridRPC Framework,** *Concurrency & Computation: Practice & Experience*, 2007.

[4] United Device : http://www.ud.com

[5] Intergiciel g-Lite http://glite.web.cern.ch/glite

[6] Equipe Graal http://graal.ens-lyon.fr

[7] Keith Seymour, Hidemoto Nakada, Satoshi Matsuoka, Jack Dongarra, Craig Lee et Henri Casanova, **Overview of GridRPC: A Remote Procedure Call API for Grid Computing,** *Proceedings of the Third International Workshop on Grid Computing***,** 274-278, 2002.