Etablissement d'accueil : **Université Bordeaux 1**

Nom du laboratoire d'accueil **: LaBRI**

Nom du tuteur : **François Pellegrini** et **Jean Roman** (HdR)

francois.pellegrini@labri.fr

Présentation scientifique du projet

Contexte scientifique

Ce sujet s'inscrit dans le cadre du projet ScAlApplix, commun au LaBRI et à l'INRIA Futurs, et qui est une des deux composantes de l'équipe Satanas (Supports et AlgoriThmes pour les Applications Numériques hAutes performanceS) du LaBRI. L'objectif de ce projet est la mise en oeuvre de compétences scientifiques complémentaires pour une recherche pluridisciplinaire dans les domaines de l'informatique hautes performances et de la modélisation numérique, et ce dans le but d'analyser et de résoudre efficacement des problèmes de calcul scientifique provenant d'applications complexes nécessitant des puissances de calcul et manipulant des tailles de données téraflopiques.

Un des points essentiels à la résolution efficace de ces problèmes est que la charge de calcul soit équitablement répartie sur les processeurs, et ce durant toute l'exécution des simulations, ce qui nécessite éventuellement une répartition dynamique de la charge au cours du temps. La création d'outils efficaces de repartitionnement et leur mise à la disposition des scientifiques concepteurs de codes numériques est donc un enjeu scientifique majeur par ses retombées indirectes. Les premiers bénéficiaires en seront les autres membres du projet ScAlApplix (scientifiques de l'IMB et de l'INRIA collaborant autour de la plate-forme régionale mutualisée de calcul PlaFRIM portée par l'INRIA en concertation avec l'IMB et le LaBRI) ainsi que leurs partenaires industriels (tels que le CEA et EDF, déjà utilisateurs du partitionneur Scotch développé au sein du projet ScAlApplix).

Problématique

Du fait des quantités considérables de données et de calculs mises en jeu, la résolution des simulations numériques en vraie grandeur imaginées aujourd'hui par les chercheurs n'est réalisable qu'en ayant recours au calcul parallèle haute performance au sens large. Pour cela, il est nécessaire de pouvoir décomposer efficacement et distribuer de façon équilibrée dans le temps et sur l'ensemble des processeurs les données et les calculs associés à traiter. Une approche très efficace pour résoudre ce problème consiste à utiliser des techniques de partitionnement de graphes, ces graphes modélisant finement l'application à paralléliser. Les graphes utilisés dans ce cas sont valués, les poids entiers portés par leurs sommets modélisant la quantité de calcul dévolue à chacun d'eux.

Lorsque les graphes à partitionner ne dépassent pas quelques millions de sommets, il est possible d'utiliser des techniques séquentielles fournissant en moyenne de très bons résultats. Ces techniques sont basées sur l'utilisation d'une approche multi-niveaux, dans laquelle les

graphes de grandes tailles sont successivement contractés en graphes de même structure topologique mais de tailles à chaque fois plus petites. Il est alors possible de calculer une partition de bonne qualité sur le plus petit graphe, au moyen d'un algorithme global qui serait trop coûteux à utiliser sur le graphe initial, puis à répercuter la partition ainsi calculée, de proche en proche, sur les graphes de tailles croissantes, en utilisant à chaque fois un optimiseur local pour adapter la structure de la partition au niveau de détail du graphe courant.

La parallélisation de l'approche multi-niveaux a déjà donné lieu à plusieurs réalisations parallèles. La plus ancienne, mise en oeuvre au sein du logiciel parMeTiS [1], présentait des problèmes de performance et de stabilité, et c'est pourquoi un travail de thèse fut entrepris il y a trois ans au LaBRI par Cédric Chevalier [2]. Ce travail déboucha en particulier sur de nouveaux algorithmes parallèles d'appariement, permettant au logiciel PT-Scotch de fournir des graphes contractés de qualité équivalente à celle des meilleurs algorithmes séquentiels. Le travail réalisé sur PT-Scotch dans le cadre de la renumérotation de matrices creuses est en cours d'extension au partitionnement de graphes, et un outil parallèle efficace de partitionnement de graphes sera donc bientôt disponible.

Cependant, le repartitionnement parallèle de graphes, nécessaire à l'équilibrage dynamique de la charge lorsque le coût des calculs associés aux sommets évolue dans le temps, est encore mal traité. parMeTiS propose des algorithmes diffusifs pour calculer des repartitionnements minimisant le nombre de sommets à déplacer, mais ces algorithmes, toujours basés sur les mêmes algorithmes multi-niveaux, ne sont pas efficaces [3].

Objectif du sujet

L'objet du travail proposé est la conception et la mise en oeuvre au sein de PT-Scotch d'algorithmes parallèles de repartitionnement de graphes, destinés à l'équilibrage dynamique de la charge d'applications massivement parallèles.

Les algorithmes de type Fiduccia-Mattheyses censés opérer sur les graphes distribués se parallélisant mal, il faut trouver à les remplacer par des algorithmes beaucoup plus scalables. La méthode qui sera étudiée en premier lieu dans le cadre de cette thèse sera basée sur le bipartitionnement récursif au moyen d'une méthode multi-niveaux, et qui remplacera les algorithmes d'optimisation locale intrinsèquement parallèles par des algorithmes de diffusion appliqués aux graphes bandes issus des frontières entre sous-domaines. Ces algorithmes de diffusion seront biaisés afin de prendre en compte le partitionnement initial sur lequel devra s'appuyer les méthodes de repartitionnement afin de minimiser le nombre de sommets déplacés entre sous-domaines.

Un autre bon candidat possible pour cette tâche est la classe des algorithmes génétiques. En opérant une fois encore sur les graphes bandes construits au voisinage du séparateur à raffiner, on peut représenter par un vecteur {0,1} le fait que les sommets appartiennent à l'une des deux parties, et une solution globale peut donc être construite en appariant les sous-vecteurs portés par chaque processeur. Cette classe de méthodes, même si elle est moins efficace que le Fiduccia-Mattheyses, devrait cependant permettre de franchir des maxima locaux lorsque le séparateur ne peut plus être maintenu sur un seul processeur.

Bien sûr, toute autre idée est la bienvenue. Les pistes énumérées ci-dessus sont seulement indicatives, et donnent juste un aperçu du travail à réaliser. Les goulots d'étranglement induits par la très grande latence des communications distantes, par rapport au coût d'accès en mémoire, conduiront dans tous les cas à privilégier l'étude d'algorithmes multi-threadés et spéculatifs, et faisant usage de la mémoire partagée pour les machines à noeuds multi-processeurs.

La réalisation d'une plate-forme parallèle de tests d'algorithmes, à l'image de ce que propose déjà PT-Scotch dans le cadre de la renumérotation parallèle de matrices creuses, est un objectif prioritaire.

Il s'agira également d'étudier comment implémenter des fonctions de coût multi-critères dans les différents algorithmes d'optimisation. Les fonctions de coût multi-critères sont un sujet difficile, car le choix entre plusieurs partitions différentes en fonction de critères multiples (équilibrage des tailles des parties, minimisation de la taille des séparateurs, mais aussi obtention de partitions disposant d'un bon aspect ratio, etc.) suppose une hiérarchisation, ou une combinaison, linéaire ou non, entre les critères individuels évalués, qui dépend du problème étudié, et suppose une grande flexibilité dans l'interface de la bibliothèque.

Références

1. G. Karypis et V. Kumar. Parallel Multilevel k-way Partitioning Scheme for Irregular Graphs. SIAM Review, Vol. 41, No. 2, pp. 278 - 300, 1999.
2. C. Chevalier. Conception et mise en oeuvre d'outils efficaces pour le partitionnement et la distribution parallèles de problèmes numériques de très grande taille. Thèse de doctorat, LaBRI, Université Bordeaux 1, septembre 2007.
3. K. Schloegel, G. Karypis, and V. Kumar. Dynamic Repartitioning of Adaptively Refined Meshes. Supercomputing, 1998.