Dossier de soutenance de thèse

Loris Marchal

Communications collectives et ordonnancement en régime permanent sur plates-formes hétérogènes

Sommaire

Curriculum vitæ

État civil

Loris Marchal

• Date de naissance :	22 février 1980
• Lieu de naissance :	Lyon 6eme (69)
• Nationalité :	française
• Situation familiale :	célibataire
• Adresse professionnelle :	LIP, ENS-Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07
• Téléphone professionnel :	04 72 72 85 04
• Adresse personnelle :	23 rue René Leynaud, 69001 LYON.
• Téléphone :	06 64 51 58 95
• Adresse électronique :	Loris.Marchal@ens-lyon.fr
• Page personnelle :	http://perso.ens-lyon.fr/loris.marchal/

Cursus

2003–2006 : **Thèse**

Titre : **Communications collectives et ordonnancement en régime permanent sur plates-formes hétérogènes**

Directeurs : Olivier Beaumont et Yves Robert

Lieu : École normale supérieure de Lyon
LIP, Laboratoire de l'informatique et du parallélisme

Situation : Moniteur allocataire de recherche (allocation couplée)

2000–2003 : **Magistère d'Informatique et Modélisation** de l'ENS Lyon

2002–2003 : **DEA d'informatique fondamentale** à l'ENS Lyon

2001–2002 : **Maîtrise d'Informatique** à l'ENS Lyon.

1998–2000 : **Licence d'Informatique** à l'ENS Lyon.

1998–2000 : **Classes préparatoires** au lycée "Aux Lazaristes" (Lyon 5eme).

1997–1998 : **Baccalauréat** série **S** option mathématiques.

Fonctions occupées

2003–2006 : Allocataire de recherche au sein du projet Graal du LIP, ENS Lyon. Moniteur à l'ENS Lyon.

2000–2003 : Élève normalien de l'ENS Lyon.

Formation

Stages de recherches

2003 : Stage de DEA de cinq mois au Laboratoire de l'Informatique du Parallélisme de l'ENS Lyon, *Sujet de stage : communications collectives pipelinées sur plate-forme hétérogène*, encadré pas Yves Robert.

2002 : Stage de dix semaines à l'université de Californie à San Diego : *A Network Model for Simulation of Grid Application*, encadré par Henri Casanova.

2001 : Stage de six semaines au Laboratoire d'Informatique de Marseille, *Automates cellulaires number-conserving : dynamique et décidabilité*, encadré par Bruno Durand et Enrico Formenti.

Formation complémentaire

Stages CIES : Évaluation d'un enseigné et d'un enseignement, Initiation à la zététique : comment enseigner l'esprit critique ?

Atelier CIES : Approche interdisciplinaire des mathématiques pour les sciences de la vie.

Module d'insertion professionnelle : "Histoire et philosophie des sciences".
Participation à deux Workshops organisés dans le cadre le l'équipe associé INRIA I-ARTUHR, en 28-30 septembre 2004 et 12-14 novembre 2005.

Participation aux Journées de l'ACI PairAPair, 24-25 mars 2005.
Participation à la réunion de l'ANR Alpage, 3 et 4 juin 2006.

Enseignement

2006–2007 : Moniteur ENS-Lyon : Architecture, Réseaux et Système (Licence 3) et Algorithmique des Réseaux et des Télécoms (Master 1)

2005–2006 : Moniteur ENS-Lyon : Algorithmique Parallèle (Master 1) et Algorithmique des Réseaux et des Télécoms (Master 1)

2004–2005 : Moniteur ENS-Lyon : Algorithmique Parallèle (Master 1) et Bases de données et Algorithmique, à l'INSA (première année de classe préparatoire)

Doctorat

Titre	Communications collectives et ordonnancement en régime permanent sur plates-formes hétérogènes
Directeurs	Olivier Beaumont et Yves Robert
mots clés	ordonnancement, régime permanent, communications collectives, optimisation combinatoire, programmation linéaire.
Date de soutenance prévue	16 octobre 2006

Contributions et résumé de la thèse

Contexte de la thèse

L'objet de ce travail est d'étudier diverses techniques d'ordonnancement pour les plates-formes distribuées à grande échelle. Ce type de plates-formes, rassemblant des ressources de calculs distribuées à l'échelle d'un pays, ou d'un continent, voit son intérêt grandir, en particulier parce qu'il offre une alternative relativement peu coûteuse aux super-calculateurs monolithiques comme les récents «BlueGene» ou «Earth Simulator». De nombreux projets tentent de rassembler des ressources distribuées afin de créer des «grilles de calcul», comme le projet français Grid5000. La contrepartie du coût limité de ces plates-formes par rapport à des super-calculateurs est leur irrégularité et leur hétérogénéité, présente à tous les niveaux :

- au niveau matériel, les processeurs sont différents, et le réseaux d'interconnexion très complexe,
- au niveau logiciel, ces machines utilisent des systèmes d'exploitation, des librairies de calcul, et des intergiciels différents,
- et même au niveau administratif l'accès aux différentes ressources n'est souvent pas centralisé.

Notre objectif est d'étudier comment utiliser efficacement de telles plates-formes. Une première constatation s'impose : l'utilisation d'une plate-forme distribuée, puissante mais complexe, n'est justifiée que pour l'exécution d'une application nécessitant beaucoup de calculs. De plus, vu l'hétérogénéité de la plate-forme, on ne peut espérer y exécuter un code fortement couplé. Notre étude s'oriente donc vers des applications nécessitant beaucoup de calculs, mais relativement simples. On peut par exemple penser au modèle de tâches indépendantes, pour lequel l'application consiste en un grand nombre de tâches similaires et indépendantes. Ce modèle convient par exemple au projets de grilles participatives, utilisant les ordinateurs de personnes volontaires pendant leur absence ; on peut par exemple citer le projet BOINC, ou encore les très classiques SETI@home ou Einstein@home.

Nous voulons également étudier des applications qui ne sont pas aussi simples que des tâches indépendantes, quoique présentant une certaine régularité. Pour ceci, on s'intéresse aux communications impliquées par l'exécution d'une application distribuée. Ces communications

peuvent souvent être rassemblées sous la forme de primitives de communications collectives, comme par exemple la diffusion de données : une des machines transmet à toutes les autres une copie d'une donnée qu'elle possède. Pour qu'une application mérite d'être exécutée sur une plate-forme à grande échelle, il est probable que le volume de données à communiquer soit important, on peut par exemple imaginer qu'une base de données de grande taille soit nécessaire à l'application, et doive ainsi être diffusée à toutes les machines avant le début du calcul. Pour effectuer ces transferts d'importants volumes de données, nous les découpons en une série d'un grand nombre de communications de taille plus modeste : la diffusion d'une donnée de grande taille sera alors transformée en une série d'un grand nombre de diffusion de messages de taille plus modeste, que l'on cherche à effectuer de façon pipelinée.

Régime permanent

Nous allons profiter de la régularité des applications pour optimiser leur temps d'exécution. Nous supposons que ces applications consistent en un grand nombre d'actions répétitives : soit une série de tâches indépendantes, soit une série de communications collectives, soit même une série de graphes de tâches similaires. Plutôt que de chercher à minimiser le temps total d'exécution de l'application, nous tentons de maximiser le débit d'opérations effectuées pendant la phase de régime permanent. Ceci a un triple avantage. D'abord, cette approche *simplifie* le problème d'optimisation : en négligeant les phases d'initiation et de terminaison des calculs, nous nous occupons uniquement des quantités moyennes de calcul et de communication allouées à chaque machine. Ensuite, cette approche est *efficace* : nous construisons des solutions optimales pour le régime permanent sous la forme d'ordonnancements périodiques, qui sont décrits de façon compacte et peuvent ainsi être implantés facilement. Enfin, cette approche a l'avantage de l'*adaptation* à la dynamicité naturelle des plates-formes à grande échelle : comme nos solutions sont périodiques, on peut mesurer les caractéristiques de la plate-forme pendant l'exécution d'une période et mettre à jour la solution en conséquence pour la période suivante, ce qui permet de réagir rapidement aux changements de la plate-forme.

Contributions

Communications collectives

Nous avons tout d'abord étudié l'optimisation du régime permanent pour les communications collectives. Nous proposons un cadre théorique d'étude de différentes primitives de communications collectives, sous l'angle de la maximisation du débit. Pour certaines d'entre elles, la diffusion, la distribution et la réduction de données, nous proposons des algorithmes efficaces qui construisent des ordonnancements périodiques de débit optimal. D'autres primitives de communications se révèlent plus difficiles : nous montrons en particulier que pour la diffusion partielle et le calcul des préfixes, la recherche du débit optimal est un problème NP-complet. D'un point de vue expérimental, nous validons notre approche en la comparant aux méthodes existantes par simulation, et nous proposons également des heuristiques pour résoudre les problèmes d'optimisation difficiles. Nous proposons également, pour le problème de la diffusion, une implantation réelle, testée sur la grille de recherche française Grid5000.

Applications multiples

Dans un deuxième temps, nous étendons notre étude du régime permanent à l'ordonnancement d'applications sur les grilles de calcul. Nous prenons en compte le fait que plusieurs applications concurrentes se partagent les ressources disponibles. Nous étudions en particulier comment ordonnancer des applications consistant chacune en un grand nombre de tâches

indépendantes, sur une plate-forme hiérarchique, en forme d'arbre. Nous montrons que des stratégies non coopératives peuvent conduire à des situations de famine (une application est complètement défavorisée par rapport aux autres), et nous évaluons des stratégies décentralisées, en les comparant à une stratégie optimale centralisée.

D'autre part, nous nous intéressons également à la modélisation fine du réseau d'interconnexion des plates-formes distribuées. Dans ce modèle, nous montrons que l'exécution équitable de débit optimal d'un ensemble d'applications divisibles est un problème NP-complet, et nous proposons des algorithmes heuristiques pour résoudre ce problème. En collaboration avec l'équipe RESO du LIP, nous nous intéressons également à la réservation de ressources dans les réseaux d'interconnexion des grilles de calcul, en particulier pour les problèmes liés à la contention des connexions au niveau de points d'entrée et de sortie du cœur du réseau. Nous montrons que le problème d'optimisation associé est NP-complet, et nous proposons des stratégies heuristiques pour le résoudre.

Autres problèmes d'ordonnancement

Au cours de cette thèse, nous avons également été amenés à nous intéresser à d'autres problèmes d'ordonnancement en marge du thème principal de l'optimisation du régime permanent. Nous nous sommes intéressés à l'ordonnancement de tâches indépendantes en présence de mémoire limitée : dans les autres travaux, nous supposons avoir à notre disposition sur chaque processeur une mémoire pouvant contenir une nombre illimité de tâches à traiter. Nous montrons que si cette hypothèse n'est pas vérifiée, alors un grand nombre de problèmes d'ordonnancement que l'on savait résoudre, en particulier l'optimisation du débit, deviennent NP-complets. Par contre, nous présentons des résultats de simulation montrant que lorsqu'un seuil dans la taille de la mémoire est franchi, alors on peut espérer atteindre le débit optimal calculé sans contrainte de mémoire. Nous avons également apporté une contribution à la théorie des tâches divisibles, en étudiant l'ordonnancement de telles tâches avec messages de retour. En effet, on ne considère souvent que les données en entrée des calculs, tout en négligeant les résultats. Nous montrons que prendre en compte les messages de retour contenant les résultats rend les problèmes d'ordonnancement significativement plus difficiles. En particulier, nous exhibons le premier cas (à notre connaissance) d'ordonnancement optimal sous un modèle sans latence où tous les processeurs ne prennent pas part au calcul.

Publications de Loris MARCHAL

Revues internationales avec comité de lecture

[1] O. Beaumont, L. Marchal et Y. Robert. «Complexity results for collective communications on heterogeneous platforms». *Int. Journal of High Performance Computing Applications* (2006, à paraître).

[2] L. Marchal, Y. Yang, H. Casanova et Y. Robert. «Steady-state scheduling of multiple divisible load applications on wide-area distributed computing platforms». *Int. Journal of High Performance Computing Applications* (2006, à paraître).

[3] A. Legrand, L. Marchal et Y. Robert. «Optimizing the steady-state throughput of scatter and reduce operations on heterogeneous platforms». *Journal of Parallel and Distributed Computing* **65**, numéro 12 (2005), 1497–1514.

[4] O. Beaumont, A. Legrand, L. Marchal et Y. Robert. «Steady-state scheduling on heterogeneous clusters». *Int. J. of Foundations of Computer Science* **16**, numéro 2 (april 2005).

[5] O. Beaumont, A. Legrand, L. Marchal et Y. Robert. «Pipelining broadcasts on heterogeneous platforms». *IEEE Trans. Parallel Distributed Systems* **16**, numéro 4 (2005).

[6] O. Beaumont, A. Legrand, L. Marchal et Y. Robert. «Scheduling strategies for mixed data and task parallelism on heterogeneous clusters». *Parallel Processing Letters* **13**, numéro 2 (2003).

Conférences internationales avec comité de lecture

[7] O. Beaumont, L. Carter, J. Ferrante, A. Legrand, L. Marchal et Y. Robert. «Centralized versus distributed schedulers for multiple bag-of-task applications». Dans *International Parallel and Distributed Processing Symposium IPDPS'2006* (2006), IEEE Computer Society Press.

[8] O. Beaumont, L. Marchal, V. Rehn et Y. Robert. «FIFO scheduling of divisible loads with return messages under the one-port model». Dans *HCW'2006, the 15th Heterogeneous Computing Workshop* (2006), IEEE Computer Society Press.

[9] O. Beaumont, L. Marchal et Y. Robert. «Scheduling divisible loads with return messages on heterogeneous master-worker platforms». Dans *International Conference on High Performance Computing HiPC'2005* (2005, to appear), LNCS, Springer Verlag.

[10] L. Marchal, P. V.-B. Primet, Y. Robert et J. Zeng. «Optimizing Network Resource Sharing in Grids». Dans *IEEE Global Telecommunications Conference (Gloebcom'2005)* (2005, to appear), IEEE Computer Society Press.

[11] L. Marchal, Y. Yang, H. Casanova et Y. Robert. «A realistic network/application model for scheduling divisible loads on large-scale platforms». Dans *International Parallel and Distributed Processing Symposium IPDPS'2005* (2005), IEEE Computer Society Press.

[12] O. Beaumont, L. Marchal et Y. Robert. «Broadcast Trees for Heterogeneous Platforms». Dans *International Parallel and Distributed Processing Symposium IPDPS'2005* (2005), IEEE Computer Society Press.

[13] O. Beaumont, A. Legrand, L. Marchal et Y. Robert. «Independent and Divisible Tasks Scheduling on Heterogeneous Star-schaped Platforms with Limited Memory». Dans *13th Euromicro Conference on Parallel, Distributed and Network-based Processing PDP'2005* (2005), IEEE Computer Society Press, pp. 179–186.

[14] O. Beaumont, A. Legrand, L. Marchal et Y. Robert. «Pipelining broadcasts on heterogeneous platforms». Dans *International Parallel and Distributed Processing Symposium IPDPS'2004* (2004), IEEE Computer Society Press.

[15] A. Legrand, L. Marchal et Y. Robert. «Optimizing the steady-state throughput of scatter and reduce operations on heterogeneous platforms». Dans *APDCM'2004, 6th Workshop on Advances in Parallel and Distributed Computational Models* (2004), IEEE Computer Society Press.

[16] O. Beaumont, A. Legrand, L. Marchal et Y. Robert. «Complexity results and heuristics for pipelined multicast operations on heterogeneous platforms». Dans *Proceedings of the 33rd International Conference on Parallel Processing (ICPP'04)* (2004), IEEE Computer Society Press.

[17] O. Beaumont, A. Legrand, L. Marchal et Y. Robert. «Steady-state scheduling on heterogeneous clusters : why and how ?». Dans *6th Workshop on Advances in Parallel and Distributed Computational Models APDCM 2004* (2004), IEEE Computer Society Press.

[18] H. Casanova, A. Legrand et L. Marchal. «Scheduling Distributed Applications : the SimGrid Simulation Framework». Dans *Proceedings of the third IEEE International Symposium on Cluster Computing and the Grid (CCGrid'03)* (may 2003).

[19] O. Beaumont, A. Legrand, L. Marchal et Y. Robert. «Assessing the impact and limits of steady-state scheduling for mixed task and data parallelism on heterogeneous platforms». Dans *HeteroPar'2004 : International Conference on Heterogeneous Computing, jointly published with ISPDC'2004 : International Symposium on Parallel and Distributed Computing* (2004), IEEE Computer Society Press.

Conférences nationales avec comité de lecture

[20] O. Beaumont, A.-M. Kermarrec, L. Marchal et Étienne Rivière. «Voronet, un réseau objet-à-objet sur le modèle petit-monde». Dans *CFSE'5 : Conférence Française sur les Systèmes d'Exploitation* (2006).