Dossier de qualification à la fonction de Maître de Conférences

27ème section

Camille Coti

Iowa State University of Science and Technology
Department of Mathematics
396 Carver Hall
50011 Ames, Ia, USA
Email : coti@iastate.edu
Site Web : http://coti.public.iastate.edu
Téléphone : +1 515 294 8687

Présentation

Dans ce dossier, je résume les activités de recherche et d'enseignement que j'ai menées au cours des stages effectués lors de mon cursus, des trois années de ma thèse et du début de séjour post-doctoral.

J'ai effectué ma thèse sous la direction de Franck Cappello, Directeur de Recherches à l'INRIA Saclay-Île de France, et de Thomas Hérault, Maître de Conférences à l'Université Paris Sud-XI. J'étais intégrée au projet INRIA Grand Large de l'INRIA Saclay-Île de France et l'équipe Parallélisme du Laboratoire de Recherche en Informatique (LRI) de l'Université Paris Sud-XI.

J'ai également effectué durant ma thèse trois séjours longs à l'Université du Tennessee d'une durée de six mois chacun.

J'ai bénéficié pour cette thèse d'un contrat de recherche doctorale dont le financement provenait du projet Européen QosCosGrid, qui s'est déroulé de septembre 2006 à mai 2009.

Mes activités d'enseignement ont été réalisées en tant que vacataire à l'IUT d'Orsay et à l'école d'ingénieurs Polytech'Paris-Sud (anciennement IFIPS).

Mon sujet de thèse porte sur les environnements d'exécution d'applications parallèles, en particulier suivant le paradigme de passage de messages normalisé par MPI, sur des machines de grande taille. La grande échelle soulève des problématiques que j'ai examinées selon troix axes : le passage à l'échelle de l'environnement d'exécution lui-même et de ses fonctionnalités vis-à-vis de l'application, la tolérance aux défaillances, inhérente aux systèmes à grande échelle, et un type particulier de systèmes à grande échelle avec les grilles de calcul.

J'effectue actuellement un séjour post-doctoral à l'Université d'État de l'Iowa où je travaille d'une part sur l'aide à la sûreté de programmation d'applications parallèles et l'évaluation d'implémentations de bibliothèques de communication, et d'autre part sur l'évaluation des performances de machines à architecture émergente.

J'ai auparavant effectué un stage d'une durée de deux mois au King's College, London dans le cadre de mes études d'ingénieur à Télécom INT. J'ai travaillé sur les réseaux de neurones et les processus gaussiens en graphes aléatoires finis sous la direction de Peter Sollich.

Mots-clés : Calcul à hautes performances, grilles de calcul, calcul à grande échelle, tolérance aux pannes, outils pour le développement et l'exécution d'applications parallèles.

Résumé du dossier

Recherche

Articles publiés :
- 8 articles sont publiés en conférences internationales
- 1 article est publié dans une revue internationale
- 1 article est en cours de soumission dans une revue internationale

Rapports de recherche :
- 1 rapport de recherche INRIA
- 7 rapports de livrables de projets Européen (QosCosGrid) et national (ANR SPADES)

Exposés :
- 3 communications à des conférences internationales
- 6 communications à des groupes de travail (projets QosCosGrid, Grid'5000 et HARNESS)
- 5 communications à des séminaires

Enseignement

95H équivalent TD :
- 46H équivalent TD à l'IFIPS
- 49H équivalent TD à l'IUT d'Orsay

Table des matières

1 Curriculum vitae

1.1 Identité

Nom	Coti
Prénom	Camille
Nationalité	Française
Date de naissance	17 novembre 1984
Lieu de naissance	Paris 12ème
Adresse personnelle	2101 Oakwood Rd, Apt 327 50014 Ames, Ia, USA
Téléphone personnel	+1 865 201 3065
Adresse professionnelle	Iowa State University of Science and Technology, Department of Mathematics 396 Carver Hall 50011 Ames, Ia, USA
Téléphone professionnel	+1 515 294 8687
Email	coti@iastate.edu
Site Web	http://coti.public.iastate.edu

1.2 Titres universitaires

2009 **Thèse de doctorat** en informatique de l'Université Paris Sud - XI effectuée au Laboratoire de Recherche en Informatique (LRI) et soutenue le 10 novembre 2009, mention très honorable, titre : "Environnements d'exécution pour applications parallèles communiquant par passage de messages pour les systèmes à grande échelle et les grilles de calcul"

- Sous la direction de Franck Cappello (DR, INRIA Saclay) et co-encadrée par Thomas Hérault (maître de conférences, LRI)
- Rapporteurs : Jean-François Méhaut (ENSIMAG) et Raymond Namyst (Université de Bordeaux I)
- Président de jury : Yannis Manoussakis (LRI)
- Examinateurs : George Bosilca (University of Tennessee), Franck Cappello (INRIA Saclay), Thomas Hérault (LRI)

2006 **Diplôme d'ingénieur** en télécommunications à Télécom INT. Option de troisième année : Architecte de Services en Réseaux.

- Stage de fin d'études réalisé à l'INRIA Saclay sous la direction de Thomas Hérault, intitulé "Conception et évaluation d'un algorithme de retour arrière sur points de reprise pour MPICH2" et soutenu devant le jury composé de Thomas Hérault (directeur de stage), Éric Renault (suivi administratif) et François Meunier (examinateur). Stage récompensé par le prix de la Fondation Louis Leprince-Ringuet récompensant les 3 meilleurs stages sur l'ensemble des étudiants du Groupement des Écoles de Télécommunications.
- Projet de fin d'études intitulé "Réalisation d'un service de grilles pour Globus 4" réalisé sous la direction de Daniel Millot et soutenu devant le jury composé de Daniel Millot (encadrant), Guy Bernard (examinateur) et Nigel Barnett (examinateur)

1.3 Fonctions exercées

Depuis novembre 2009 : post-doc à Iowa State University, département de Mathématiques, groupe High Performance Computing.

Octobre 2006 à novembre 2009 : doctorante à l'INRIA Saclay-Île de France, projet Grand Large. Membre de l'équipe Parallélisme / Grand Large commune à l'INRIA Saclay-Île de France et au Laboratoire de Recherche en Informatique (LRI).

Avril 2006 à septembre 2006 : stage de fin d'études à l'INRIA Saclay-Île de France, projet Grand Large.

Juillet 2005 à août 2005 : stage au King's College, London, département de mathématiques.

2 Activités d'enseignement

Mes activités d'enseignement ont été réalisées dans le cadre de vacations à l'IUT d'Orsay et à l'IFIPS. J'ai effectué des TD et TP à des étudiants en formation initiale à l'IUT d'Orsay en 2007-08 et à des étudiants en apprentissage à l'IFIPS 2008-09 puis on m'a confié la responsabilité d'un cours pour des étudiants en apprentisage à l'IFIPS en 2009-10.

2.1 Résumé

Le volume total d'enseignement est de 95H équivalent TD, proche du maximum de 96H autorisé par mon statut de doctorante INRIA. Ce total est composé de 16H de cours magistral et 71H de TD et TP.

Module	Niveau	Année	Formation	Volume
Programmation objet avec Java	L2	2007-08	formation initiale IUT	49H
Programmation parallèle	M2	2008-09	apprentissage IFIPS	10H
Réseaux avancés	M2	2009-10	apprentissage IFIPS	36H

Le contenu des modules où je suis intervenue est détaillé dans la suite de cette section.

2.2 Programmation objet avec Java

J'ai pris en charge pendant toute la durée du cours, c'est-à-dire un semestre, un groupe de TD de deuxième année de l'IUT d'Orsay pour le cours d'initiation à la programmation orientée objet avec Java dirigé par Sylvie Delaët.

Ce cours a pour but d'introduire et d'approfondir les concepts de programmation orientée objet en s'intéressant particulièrement à leur application aux méthodes de programmation d'applications. Ainsi, le cours contient des références au cours d'UML pour illustrer et mettre en œuvre les notions d'héritage, d'encapsulation et d'interface. Il est également fortement lié au cours de bases de données et donne lieu à un projet commun à ces deux matières. Une partie du module est consacrée à la programmation d'interfaces graphiques avec l'introduction de la représentation modèle-vue-contrôleur.

Le module est organisé comme suit : un cours magistral hebdomadaire d'une durée de 1H30 présente des nouveaux concepts qui seront mis en œuvre lors d'une scéance de TD/TP en classe entière de 2H et une scéance de TP de 1H30 en demi-groupe en alternance. Chaque sujet de TP peut être réalisé en une ou plusieurs scéances suivant sa complexité (croissante au fur et à mesure de l'avancement du module) et est à rendre afin de constituer la note de TP, comptant pour la moitié de la note de contrôle continu. Par ailleurs, deux examens de contrôle continu sur tâble ont lieu pendant le module, en plus d'une note de projet et d'un contrôle final sur table. Les sujets de contrôle continu sont conçus par les chargés de groupe en binôme : ainsi, le premier examen a été conçu par Sylvie Delaët et posé à nos deux groupes et j'ai conçu le deuxième examen, qui a été posé à nos deux groupes. Le contrôle final est corrigé collégialement par tous les enseignants du module, chacun ne corrigeant pas de copies de son groupe. Les moyennes de chaque groupe permettent de vérifier l'homogénéité des connaissances acquises par les étudiants d'un groupe, et surtout d'un enseignant, à l'autre. Le groupe dont j'ai eu la charge a obtenu une moyenne égale à celle des quatre autres groupes.

2.3 Programmation parallèle avec MPI et OpenMP

En 2008-09 j'ai été chargée de TD pour ce cours de l'IFIPS, aujourd'hui Polytech' Paris-Sud. Les étudiants dont j'ai eu la charge étaient en dernière année de leur cursus d'ingénieur, spécialité informatique, et avaient la particularité de suivre une formation en apprentissage. Ainsi, ils disposaient d'une expérience significative de la programmation dans des langages variés et parfois éloignés des langages adaptés à ce cours. Cependant, leur formation technique étant avancée et mise en pratique, la plupart des étudiants de cette filière disposent d'une maturité particulière vis-à-vis de leur formation et de l'attente qu'ils en ont (rendant l'enseignement de cette matière spécialisée plus facile avec certains, plus difficile avec d'autres, moins réceptifs).

Le cours a été conçu par Mathieu Jan qui en assurait également les cours magistraux. Les TP ont été assurés par un autre vacataire. Mathieu Jan étant également vacataire, nous étions sous la responsabilité de Claude Barras, permanent de l'IFIPS (maître de conférences) et cadre de la filière d'apprentisage.

Le programme a couvert l'introduction du parallélisme et des techiques de programmation pour machines à mémoire distribuée en utilisant le paradigme de programmation par passage de messages avec MPI et pour machines à mémoire partagée en utilisant plusieurs threads avec OpenMP.

MPI a été introduit en premier : les étudiants ont vu la nécessité de déplacer les données explicitement dans leur application. Les deux premiers cours magistraux, les six premières heures de TD et le premier TP leur ont présenté les fonctionnalités les plus utilisées de la norme MPI : les différentes routines de communication point-à-point, les communications collectives, le concept et les manipulations de communicateurs, les topologies et les coordonnées, les types de données. Chaque notion a été illustrée en TD par un exemple de calcul, introduisant souvent en même temps différentes méthodes de découpage des données et de parallélisation du calcul.

Le dernier cours et les quatre dernières heures de TD ont été consacrée à OpenMP. Les mécanismes de gestion de la concurrence d'accès à une mémoire partagée et de découpage du travail, ainsi que le modèle fork-join ont été présentés et illustrés d'exemples à la manière de la présentation de MPI.

2.4 Réseaux avancés

Au début de l'année scolaire 2009-10, on m'a confié la responsabilité du cours de réseaux avancés de dernière année de la formation en apprentissage de l'IFIPS. Ce cours a été découpé en quatre partie d'environ 8 heures chacune (4H de cours et 4H de TD) et la dernière partie, sur la sécurité, a été l'objet de deux scéances de TP de 4H chacune.

Dans la première partie de ce module, j'ai présenté les architectures des réseaux globaux à haute capacité et à haut débit pour les infrastructures de communication et les problématiques spécifiques de routage, d'interconnexion et de contrôle de congestion qui leur sont associées ainsi que les algorithmes et les mécanismes utilisés pour y répondre. Des exemples de technologies ont été approfondis avec ATM et IPv6.

La deuxième partie de ce module a porté sur les réseaux locaux, notamment à haut débit et basse latence et les réseaux à temps réel. Les réseaux Ethernet et le protocole ARP ont été vus en détail. Des applications de réseaux rapides comme Infiniband ont été présentées avec les clusters et les systèmes de stockage NAS et SAN. Enfin, les principes des réseaux embarqués à temps réel comme le multiplexage par TDMA et des applications avec les réseau embarqués CAN et FlexRay et le réseau téléphonique commuté ont été

présentés en détail.

La troisième grande partie de ce cours est la plus orientée vers les télécommunications avec la présentation des infrastructures réseau pour la téléphonie mobile et leur évolution depuis la première génération, RadioComm 2000 jusqu'à l'UMTS et l'EDGE et les perspectives à venir. Des notions périphériques ont été présentées comme les problèmes de coloration de graphe appliqués à l'attribution de fréquences et les techniques de codage de canal utilisées pour obtenir des communications robustes à haut débit en multiplexant sur une fréquence.

Enfin, le dernier volet du module a porté sur la cryptographie avec la présentation des principes de base et les trois fonctions de la cryptographie : le cryptage-décryptage de messages, la signature et l'authentification. Les principes mathématiques sur lesquels repose RSA ont été présentés ainsi que l'importance du générateur de nombres aléatoires. Enfin, des failles célèbres ont été présentées pour montrer la faiblesse de certains systèmes de sécurité.

J'ai bénéficié d'une grande liberté dans la conception de ce cours, et c'est volontairement que j'ai choisi de présenter un large spectre d'applications des réseaux de communication, du calcul aux télécommunications. Cette variété a généré un certain enthousiasme chez les étudiants qui ont, à ce que j'ai pu constater au vu de leur participation en cours et en TD, apprécié le contenu du module.

2.5 Enseignements au cours de mon séjour post-doctoral

Bien que ces activités d'enseignement n'aient pas commencé à l'heure de cette candidature, mon séjour post-doctoral inclut des enseignements à l'Université d'État de l'Iowa. Je suis, à partir du semestre de printemps 2010, instructrice chargée du cours de calcul différentiel et intégral (Calculus) de niveau Bachelor. Le volume horaire est de 4 heures par semaine.

Mes responsabilités incluront la conception du cours à partir d'un programme global, la conception et la correction des devoirs à réaliser chaque semaine par les étudiants ainsi que la conception et la correction de deux examens de contrôle continu et la correction de l'examen final (conçu par un autre enseignant du département).

Il s'agit d'un exercice d'enseignement relativement différent des activités menées jusqu'alors, avec un niveau technique peu élevé (correspondant à un programme de première et terminale scientifique en France) mais des attentes importantes sur le plan pédagogique et au niveau de la façon de présenter et d'amener des connaissances nouvelles à des étudiants débutant leur cursus scientifique et disposant le plus souvent de bases minces dans la matière étudiée. En ce sens, on peut le rapprocher des enseignements en programmation que j'ai réalisés à l'IUT d'Orsay.

2.6 Projet d'enseignement

Les cours auxquels j'ai participé m'ont beaucoup intéressée pour des raisons différentes. J'ai trouvé très intéressant d'enseigner une technique de programmation à des étudiants ayant relativement peu d'expérience de la programmation (deuxième année d'IUT) pour leur présenter, en plus d'une technique et d'un ensemble de concepts, une façon de programmer et un ensemble de "bonnes pratiques" qui, je l'espère, leur sera utile quel que soit le langage ou l'environnement de programmation dans lequel ils évolueront ensuite.

J'ai également trouvé intéressant d'intervenir face à des étudiants ayant déjà une certaine expérience de l'informatique en général et de la programmation en particulier puis-

qu'en fin d'études d'ingénieur, qui plus est en alternance. Lors du cours de programmation parallèle les étudiants avaient tous une solide pratique de la programmation séquentielle, utilisant des langages et des méthodes de conception variés, et mon rôle était de leur présenter une nouvelle technique de programmation. Lors du cours de réseaux avancés certains étudiants avaient une certaine pratique de l'administration de réseaux locaux IP. Il s'agissait d'un cours apportant des notions plutôt avancées à des étudiants dont les connaissances étaient hétérogènes suivant l'expérience dont ils disposaient dans ce domaine (certains avaient un thème d'apprentissage en entreprise lié aux réseaux).

Je suis particulièrement intéressée par deux aspects de l'enseignement. D'une part, j'apprécie d'enseigner à un groupe d'étudiants "novices", peu expérimentés et en début de cursus, pour leur apporter les premières bases techniques. Il s'agit d'une situation où l'aspect pédagogique de l'enseignement est fortement accentué puisqu'il faut présenter des notions de niveau relativement peu élevé de façon claire, cohérente et facilement assimilable. L'encadrement plus serré des étudiants du fait du nombre important de TD et de TP en petits groupes permet d'apporter un suivi plus personnalisé des étudiants et d'ajuster sa réponse aux besoins de chacun, plus lents ou plus rapides. C'est pourquoi je suis prête à participer à des cours d'initiation à la programmation et à l'algorithmique de niveau L1 et L2, notamment dans la construction du programme et la conception des supports de cours.

Le deuxième aspect qui m'intéresse particulièrement concerne les cours de niveau technique avancé apportant des notions pointues dans un domaine particulier. S'adresser à des étudiants plus avancés dans leur cursus et plus à l'aise avec l'assimilation de notions technique permet d'aborder des aspects techniquement pointus. En ce sens, je suis disposée à m'investir dans des cours de niveau M1 et M2 sur des thèmes connexes à mes domaines de recherche et où je dispose d'une expertise technique plus importante, comme l'algorithmique distribuée, la programmation parallèle, les concepts de systèmes d'exploitation, l'architecture des machines et les réseaux, ou d'introduction aux systèmes d'exploitation de niveau L3.

3 Activités de recherche

3.1 Stage d'école d'ingénieurs

Au cours de mes études d'ingénieur à Télécom INT, j'ai effectué un stage de deux mois au département de mathématiques du King's College, London, sous la direction de Peter Sollich. Outre l'expérience internationale requise par mon cursus, cette expérience m'a donné un premier contact avec la recherche, la démarche scientifique et un laboratoire universitaire.

J'ai travaillé dans le domaine des réseaux de neurones sur les processus gaussiens dans des graphes aléatoires de taille finie. Ces travaux m'ont permis de mettre en œuvre mes connaissances en statistiques, théorie des graphes, algèbre linéaire et surtout programmation de calculs numériques pour des simulations.

J'ai étudié les marches aléatoires dans un graphe aléatoire régulier fini, et plus précisément les effets de cette finitude sur la marche aléatoire. On néglige souvent ces effets en considérant le graphe comme "de grande taille", c'est-à-dire d'un diamètre grand devant le nombre d'étapes de la marche aléatoire.

Or, nous avons constaté que lorsque le nombre d'étapes multiplé par la probabilité de se déplacer sur un nœud voisin atteint le même ordre de grandeur que le diamètre du graphe, la marche aléatoire repasse à des endroits par lesquels elle est déjà passée (création de boucles). Les matrices d'adjacence correspondantes deviennent alors semblables : la distribution de la marche aléatoire est stationnaire.

Dans un contexte d'apprentissage automatique par régression par processus Gaussiens, le noyau atteint alors une forme spécifique donnant une erreur de généralisation qui correspond à l'erreur Bayésienne.

Les résultats de ce stage ont servi d'étude préliminaire à une thèse de doctorat sur le sujet effectuée actuellement par Matthew Urry, étudiant au King's College. Ils ont été publiés dans une conférence internationale majeure du domaine des réseaux de neurones (NIPS).

3.2 Stage de fin d'études

J'ai choisi d'effectuer un stage de fin d'études d'ingénieur orienté recherche, similaire à un stage de DEA. Je l'ai effectué à l'INRIA Saclay, au sein de l'équipe Grand Large dont les axes sont l'étude de systèmes à grande échelle, sous la direction de Thomas Hérault. J'ai travaillé dans le projet MPICH-V qui avait pour but de concevoir, implémenter et évaluer des protocoles de tolérance aux pannes par retour arrière sur points de reprise pour applications MPI.

Dans un premier temps, j'ai participé à l'implémentation d'un protocole de retour arrière coordonné bloquant dans MPICH2 et à sa comparaison avec un algorithme non bloquant. Les performances de ces deux algorithmes ont été comparées sur trois plateformes d'exécution typiques : un cluster à réseau rapide, un cluster de grande taille et une grille de calcul à grande échelle.

Les résultats expérimentaux ont montré que l'algorithme bloquant, plus simple et moins intrusif dans le chemin critique des communications, était plus efficace dans un contexte de communications rapides mais passait moins bien à l'échelle du fait de son caractère bloquant et de la concurrence engendrée vis-à-vis du support d'exécution (enregistrement des points de reprise).

La deuxième partie de mon stage a été consacrée à la conception d'un protocole causal

de journalisation de messages hiérarchique, destiné aux grilles de calcul et aux systèmes à très grande échelle. J'ai réalisé une implémentation prototypaire de ce protocole afin d'en évaluer les performances.

Ce protocole étant basé sur la distribution d'un composant jusqu'alors central, il passe mieux à l'échelle en nombre de processus de l'application. De plus, il rend possible la mise en place d'une affinité des composants de l'environnement d'exécution par rapport aux processus de l'application, permettant alors d'être efficace sur la grille.

Les résultats de ce stage ont donné lieu à une publication dans une conférence internationale (SC|06) et ont été récompensés par le prix de la Fondation Louis Leprince-Ringuet, décerné par un jury d'industriels majeurs du domaine des télécommunications et distinguant chaque année trois stages considérés comme les meilleurs parmi l'ensemble des stages de fin d'études réalisés par des étudiants du Groupement des Écoles de Télécommunications (aujourd'hui nommé Institut Télécom, environ 1 200 étudiants en dernière année).

3.3 Thèse de doctorat en informatique

Les travaux de recherche effectués durant ma thèse concernent les supports exécutifs et les problématiques liées à la grande échelle pour les applications parallèles communiquant par passage de messages (en particulier, suivant la norme MPI). J'ai travaillé en particulier sur les environnements d'exécutions pour applications MPI.

Le rôle d'un environnement d'exécution vis-à-vis de l'application parallèle qu'il supporte est d'assurer son lancement sur les ressources disponibles, la mise en relation des processus de l'application les uns avec les autres, la surveillance des processus (mort ou terminaison normale) et le comportement à tenir en cas de changement d'état, et de transmettre les entrées-sorties et les signaux.

J'ai pour cela suivi trois axes de recherche : le passage à l'échelle de l'environnement d'exécution, la tolérance aux pannes, qui sont inévitables dans les systèmes à grande échelle, et la programmation et l'exécution de programmes en MPI sur grilles de calcul.

3.3.1 Passage à l'échelle

Un premier axe de recherche concerne le passage à l'échelle de l'environnement d'exécution lui-même, c'est-à-dire de ses fonctionnalités internes. J'ai effectué trois séjours de six mois chacun dans l'équipe MPI du laboratoire Innovative Computing Laboratory, à l'Université du Tennessee à Knoxville. J'y ai travaillé sur le développement d'OpenMPI, une des principales implémentations open-source de la norme MPI, notamment sur les questions de passage à l'échelle.

Les communications internes de l'environnement d'exécution (dites *out-of-band* car non-utiles directement à l'application) sont un élément déterminant des performances. J'ai travaillé sur la topologie des communications point-à-point, et sur le routage des signaux et des communications. Les communications collectives sont également un élément prépondérant, car elles synchronisent les processus de l'application entre eux (lancement, signaux). J'ai travaillé sur l'optimisation de ces opérations dans OpenMPI.

Le lancement peut être délégué à une application tiers (souvent intégrée à un système de réservation de ressources) et appelée par l'environnement d'exécution, ou être pris en charge par l'environnement d'exécution lui-même. J'ai amélioré le système de lancement utilisé par OpenMPI.

Enfin, j'ai travaillé sur le système de composants sur lequel repose la modularité d'OpenMPI afin de diminuer l'utilisation de la mémoire quand on augmente la taille du

système.

J'ai présenté mes résultats dans le cadre du projet HARNESS (De-AC05-00OR22725) qui vise à fournir des outils pour le calcul à hautes performances.

3.3.2 Tolérance aux pannes

Un axe de recherche lié à ces problématiques est la tolérance aux pannes. Plus le nombre de composants d'un système est important, plus la probabilité d'être victime d'une panne est importante. Le comportement par défaut en cas de panne des environnements de programmation MPI est de terminer l'application.

Une approche de la tolérance aux pannes est de sauvegarder l'état de l'application pour pouvoir reprendre ultérieurement, par exemple si l'exécution a été interrompue par une panne. La sauvegarde de l'état d'une application distribuée ne se fait pas aussi simplement que pour un seul processus, et il est nécessaire de s'assurer de la cohérence de l'état de l'application après récupération de son état. Pour cela, on dispose de plusieurs approches présentant des caractéristiques différentes en termes de surcoût en l'absence de panne et de temps de récupération après une panne.

J'ai travaillé à la conception, l'implémentation et l'évaluation de plusieurs algorithmes de tolérance aux pannes pour les applications communiquant par passage de messages, basés sur une prise d'état coordonnée ou non coordonnée, avec ou sans journalisation de messages. Les implémentations ont été faites dans les bibliothèques MPI MPICH2 puis OpenMPI. Les évaluations de performances comparées ont été faites sur des clusters à réseaux rapides, à grande échelle et sur grille de calcul sur la plate-forme Grid'5000.

Les algorithmes de tolérance aux pannes non coordonnés nécessitent un support plus important de l'environnement d'exécution. Ainsi, j'ai conçu un protocole de restauration de l'état d'un environnement d'exécution sous les contraintes liées au contexte de l'environnement d'exécution, et l'ai implémenté dans OpenMPI. La réparation de l'environnement d'exécution tout en continuant à fournir ses services aux autres processus de l'application permet non seulement de mettre en place des protocoles non-coordonnés et supposés comme passant mieux à l'échelle, mais est également plus rapide et passe mieux à l'échelle qu'un redéploiement total de l'environnement d'exécution et de l'application.

Une autre approche de la tolérance aux pannes dans une infrastructure de communication telle que les environnements d'exécution est apportée par l'auto-stabilisation. J'ai donc participé à la conception d'un algorithme auto-stabilisant de construction et de maintien de l'infrastructure de communications.

Certains de mes travaux s'inscrivent dans le cadre du réseau d'excellence CoreGrid (EU FP6-004265), en particulier dans la tâche 4.4 : "Fault-tolerance and robustness".

3.3.3 Grilles de calcul

Les grilles sont un type particulier de système à grande échelle, posant, outre les problèmes liés à tout système de grande taille, des problèmes particuliers dûs à son hétérogénéité et à sa nature hiérarchique.

Tout au long de ma thèse, j'ai travaillé pour le projet Européen QosCosGrid (*Quasi-Opportunistic Supercomputing for Complex Systems in Grid environments*, EU FP6 IST-2005-033883), dont le but est de fournir une plate-forme de calcul composée de plusieurs clusters (en considérant une grille comme une fédération de clusters) utilisés ensemble ponctuellement (quasi-opportunisme) à des applications nécessitant d'importantes ressources de calcul et de mémoire (les systèmes complexes).

J'ai d'abord participé à la conception globale de la pile logicielle du système QosCosGrid, après avoir examiné les besoins des applications cible. Le système QosCosGrid permet de choisir les ressources nécessaires selon un schéma d'ordonnancement déterminé par le meta-scheduler, d'effectuer des réservations coordonnées sur les ressources éventuellement localisées dans différents clusters, et enfin de déployer et exécuter l'application.

Mes travaux se sont principalement portés sur la bibliothèque de communication et de programmation parallèle permettant d'exécuter des applications MPI sur une grille de calcul, appelée QCG-OMPI. J'ai conçu et implémenté une extension de l'environnement d'exécution d'OpenMPI basée sur un ensemble de services de grille et permettant de résoudre les problèmes de connectivité inhérents aux fédérations de clusters (pare-feux complètement fermés, adressage privé).

Une autre fonctionnalité de cet extension est de pouvoir s'interfacer avec les couches basses de la pile logicielle (système de réservation) pour pouvoir obtenir des informations de topologie. J'ai travaillé sur des modèles de calcul et de communication afin d'exploiter ces informations pour organiser les communications de l'application afin de communiquer efficacement sur la grille. Enfin, j'ai proposé un ensemble de communications collectives tirant parti de ces informations et optimisées pour les grilles.

J'ai également travaillé au niveau du système de réservation afin de mettre en place un outil de réservations coordonnées sur la grille.

J'ai participé à la mise en place d'une plate-forme de développement internationale dédiée à QosCosGrid et disposant de machines en France, Irlande du Nord, Pologne, Hongrie, Espagne, Pays Bas et Australie.

Enfin, j'ai collaboré avec les utilisateurs du middleware QosCosGrid, des biologistes, astronomes et physiciens, afin de les aider à adapter et optimiser leurs applications pour exploiter au mieux les possibilités de la bibliothèque de communication QCG-OMPI.

Par ailleurs, l'articulation des fonctionnalités spécifiques aux grilles de QCG-OMPI avec des nouveaux algorithmes d'algèbre linéaire à évitement de communications a permis de montrer pour la première fois qu'il était possible d'obtenir des bonnes performances avec ce type d'applications sur une grille.

Les performances de la bibliothèque QCG-OMPI sur des applications tests typiques puis sur des applications adaptées ont été réalisées sur la plate-forme Grid'5000, qui présente bien les caractéristiques topologiques d'une fédération de clusters, et sur la plate-forme dédiée QosCosGrid, qui présente une grande diversité de restrictions d'accès allant de l'ouverture complète des ports à un pare-feu complètement fermé, en passant par un pare-feu offrant un intervalle de ports ouverts.

3.4 Séjour post-doctoral

J'effectue actuellement un séjour post-doctoral à l'Université d'État de l'Iowa, dans le groupe de calcul à hautes performances, sous la direction de Glenn Luecke. Je travaille d'une part sur l'aide à la sûreté de programmation d'applications parallèles et l'évaluation d'implémentations de bibliothèques de communication, et d'autre part sur l'évaluation des performances de machines à architecture émergente.

La sûreté de programmation d'applications consiste à rechercher dans les programmes des erreurs qui ne sont pas détectées par la vérification syntaxique du compilateur. Deux types d'erreurs sont à distinguer : les erreurs détectables à la compilation et les erreurs détectables lors de l'exécution.

Je travaille principalement sur des systèmes à mémoire partagée (éventuellement distribuée et donc virtuellement partagée) avec UPC et SHMEM. Les erreurs peuvent être liées

à la gestion de la mémoire (fuites, utilisation de zones non initialisées, accès hors limites de tableaux) ou aux fonctionnalités du langage (erreurs d'arguments, interblocages).

Par ailleurs, je poursuis des travaux débutés sur la fin de ma thèse concernant les applications parallèles sur les grilles de calcul, en particulier les applications d'algèbre linéaire, réputées inadaptées aux grilles de calcul. L'utilisation d'une nouvelle famille d'algorithmes dits "à évitement de communication" permet de confiner les communications au sein des clusters et d'effectuer un nombre minimal de communications entre les clusters. Les premiers travaux dans ce sens, effectués sur une factorisation QR, montrent un passage à l'échelle parfait sur la grille (performance d'un cluster multipliée par deux en utilisant deux clusters, par quatre en utilisant quatre clusters).

3.5 Expérience internationale et séminaires

3.5.1 Présentations

10/09 Séminaire de l'équipe Parallélisme / Grand Large, Orsay, France
"QCG-OMPI : executing MPI applications on grids"

08/09 Présentation à la conférence EuroPar'09, Delft, Pays-Bas
"MPI Applications on Grids : A Topology-Aware Approach"

08/09 Séminaire du laboratoire ICL, Knoxville, Tn, USA
"Runtime Environment Support for Fault-Tolerant MPI Applications"

08/09 Retraite annuelle du laboratoire ICL, Townsend, Tn, USA
"Run-Time Environment Scalability for MPI"

02/09 Séminaire du laboratoire ICL, Knoxville, Tn, USA
"MPI on the Grid"

06/08 Réunion de travail du projet HARNESS, Oak Ridge, Tn, USA
"FT-ORTE : Toward a self-healing, self-tuning RTE"

05/08 Présentation à la conférence CCGRID'08, Lyon, France
"Grid Services for MPI"

05/08 Réunion de travail du projet QosCosGrid, Haifa, Israël
"Introduction to QCG-OMPI"

05/08 Réunion de travail du projet HARNESS, Atlanta, Ga, USA
"Scalability Issues of Runtime Environments for Distributed Applications"

10/07 Présentation courte à la conférence EuroPVM/MPI'07, Paris, France
"Grid Services for MPI"

08/07 Retraite annuelle du laboratoire ICL, Townsend, Tn, USA
"Scalability issues in OpenMPI"

06/07 Réunion de travail du projet HARNESS, Atlanta, Ga, USA
"Runtime environment for Large-Scale and Grid Computing"

03/07 Cérémonie de remise des prix de la Fondation Louis Leprince-Ringuet, siège de Bouygues Télécom, Boulogne, France
"Conception et évaluaton d'un algorithme de tolérance aux fautes par points de reprise coordonnés pour MPICH2"

10/06 Réunion des utilisateurs de GdX, Orsay, France
"MPICH-Vcl vs MPICH-Pcl"

3.5.2 Visites à l'étranger

Ma thèse s'étant déroulée dans le cadre d'un projet Européen impliquant 8 partenaires dans 7 pays différents, j'ai été amenée à voyager pour me rendre à des réunions techniques

et visiter des équipes de recherche. Ces visites ont permis d'établir des relations avec des chercheurs dont les domaines sont liés à mes recherches. C'est le cas de l'Observatoire d'Astrophysique de Leiden, aux Pays-Bas, avec qui une collaboration a été mise en place. J'ai été invitée deux fois à me rendre en visite aux Pays-Bas pour travailler sur des projets communs. C'est également le cas du laboratoire de systèmes distribués de l'Institut Technique d'Haïfa (Technion), en Israël, où j'ai effectué une visite d'une semaine et avec qui j'ai collaboré.

Enfin, j'ai effectué trois séjours de six mois chacun à l'Université du Tennessee à Knoxville, dans le labratoire Innovative Computing Laboratory, dans le cadre d'une collaboration avec mon équipe. Lors de ces trois séjours j'ai travaillé avec les chercheurs locaux afin d'intégrer les développements effectués dans le cadre de ma thèse dans OpenMPI. J'ai également travaillé avec le groupe d'algèbre linéaire sur des applications pour les grilles de calcul.

3.6 Projet de recherche

Les conclusions tirées dans ma thèse ouvrent des perspectives à court et moyen terme suivant les trois axes que j'ai suivis. Je souhaiterais en ce sens approfondir les recherches effectuées dans ma thèse.

Les questions de passage à l'échelle et les solutions apportées doivent maintenant être adaptables dans un environnement plus générique. Cela passe par une contribution plus étroite des systèmes de lancement d'applications parallèles. De plus, certains choix techniques effectués jusqu'à maintenant sous certaines hypothèses sont remises en cause dans des environnements à très grande échelle, ces hypothèses n'étant plus toujours valides.

L'apport du support d'exécution à la tolérance aux pannes des applications permet d'explorer les protocoles de tolérance aux pannes orientés par l'application. Cette approche est prometteuse en terme de passage à l'échelle et est en cours de normalisation dans le cadre du standard MPI 3.0.

Enfin, l'approche hiérarchique a montré son efficacité sur des applications et mérite d'être poursuivie pour les grilles de calcul, dont elle permet d'exploiter efficacement les ressources de calcul. Le modèle de quasi-opportunisme développé par QosCosGrid, conssistant à mettre ponctuellement des ressources en commun pour obtenir un système à grande échelle, est alors tout à fait justifié en regard du fait que l'on dispose désormais de techniques de programmation permettant d'exécuter des applications passant à l'échelle sur de telles plate-formes. On peut également se tourner vers cette approche pour les systèmes traditionnels (clusters) à très grande échelle, dont la taille ne permet plus de suivre une approche "à plat" et nécessite pour passer à l'échelle la mise en place d'une hiérarchie interne de l'application.

Par ailleurs, la complexité de la programmation d'applications parallèles la rendant difficile à débugger, les outils de détection d'erreur peuvent aider à améliorer la justesse des programmes et à faciliter le cycle de développement et de correction des erreurs.

4 Activités administratives et collectives

J'ai été assistante à l'organisation des conférences HPDC'06, EuroPVM/MPI'07 et SC|08 en tant qu'étudiante bénévole (volunteer student). À SC|08 j'ai en particulier participé à la mise en place et au démontage du réseau SCinet, dit "le réseau le plus rapide du monde".

Je participe à l'évaluation des soumissions en tant que relectrice d'article pour la revue FGCS et j'ai été relectrice externe pour les conférences PCGrid'07, AlgoTel'07, EuroPar'07, PDP'08, EuroPVM/MPI'08, ISPDC'08 et CCGRID'09. J'ai en outre assisté le directeur du comité de programme de CCGRID'09 (Franck Cappello) dans le processus de distribution des articles à relire, de sélection des articles et d'établissement du programme de la conférence.

J'ai co-organisé le séminaire commun de l'équipe Parallélisme du LRI et du projet Grand Large du 2006 à 2008.

5 Liste des publications

5.1 Actes de conférences internationales

1. Emmanuel Agullo, Camille Coti, Jack Dongarra, Thomas Herault et Julien Langou : QR Factorization of Tall and Skinny Matrices in a Grid Computing Environment, à paraître dans *Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS'10)*, Atlanta, Georgia, USA, avril 2010.
2. Pavel Bar, Camille Coti, Derek Groen, Thomas Herault, Valentin Kravtsov, Assaf Schuster et Martin Swain : Running parallel applications with topology-aware grid middleware, à paraître dans *Proceedings of the 5th IEEE International Conference on e-Science (eScience 2009)*, Oxford, UK, décembre 2009.
3. Peter Sollich, Matthew Urry et Camille Coti : Kernels and learning curves for Gaussian process regression on random graphs, dans *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, Vancouver, Canada, décembre 2009.
4. George Bosilca, Camille Coti, Thomas Herault, Pierre Lemarinier et Jack Dongarra : Constructing Resiliant Communication Infrastructure for Runtime Environments, dans *International Conference in Parallel Computing (ParCo2009)*, Lyon, France, septembre 2009.
5. Camille Coti, Thomas Herault et Franck Cappello : MPI Applications on Grids : a Topology-Aware Approach, dans *Proceedings of the 15th European Conference on Parallel and Distributed Computing (EuroPar'09)*, Delft, Pays-Bas, LNCS volume5704, pages 477-477, août 2009,
6. Camille Coti, Thomas Herault, Sylvain Peyronnet, Ala Rezmerita et Franck Cappello : Grid Services For MPI, dans *Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid'08)*, pages 417–424, Lyon, France, mai 2008.
7. Camille Coti, Ala Rezmerita, Thomas Herault et Franck Cappello : Grid Services For MPI, dans *Proceedings of the 14th European PVM/MPI Users' Group Meeting (EuroPVM/MPI)*, Paris, France, pages 393–394, LNCS volume 4757, octobre 2007.
8. Camille Coti, Thomas Herault, Pierre Lemarinier, Laurence Pilard, Ala Rezmerita, Eric Rodriguez et Franck Cappello : Blocking vs. Non-Blocking Coordinated Checkpointing for Large-Scale Fault Tolerant MPI, dans *Proceedings of the Int. Conf. for High Performance Networking Computing, Networking, Storage and Analysis (SC|06)*, ACM press, Tampa, Fl, USA, novembre 2006.

5.2 Revue internationale

9. Darius Buntinas, Camille Coti, Thomas Herault, Pierre Lemarinier, Laurence Pilard, Ala Rezmerita, Eric Rodriguez et Franck Cappello : Blocking vs. Non-Blocking Coordinated Checkpointing for Large-Scale Fault Tolerant MPI, dans *Future Generation Computer Systems*, volume 24, numéro 1, pages 73–84, 2008.

5.3 Workshop et posters

10. Peter Sollich et Camille Coti : Covariance functions and Bayes errors for GP, dans *Bayesian Research Kitchen (BaRK'08)*, workshop du réseau d'excellence EU FP7 PASCAL II, septembre 2008, Ambleside, Lake District, UK.

11. Camille Coti, Ala Rezmerita, Thomas Herault et Franck Cappello : Grid Services for MPI, Poster à *EuroPVM/MPI'07*, Paris, Fra, octobre 2007..
12. Camille Coti, Thomas Herault, Pierre Lemarinier, Laurence Pilard, Ala Rezmerita, Eric Rodriguez et Franck Cappello : MPICH-Pcl vs MPICH-Vcl, Poster à *PariSTIC*, Nancy, Fra, 22-24 novembre 2006.
13. MPICH-V, MPI Implementation for volatile resources, Poster sur le stand INRIA à *SC|06*, Tampa, Fl, 11-17 novembre 2006.

5.4 Autres publications

14. Elisabeth Brunet Franck Cappello, Camille Coti, Thomas Herault et Sylvain Peyronnet : Supports d'exécution pour environnements petascales : État de l'art, rapport de projet ANR SPADES ANR 08-ANR-SEGI-025, avril 2010, 19 pages.
15. Camille Coti, Thomas Herault, Derek Groen et Mariusz Mamonski : D1.2c : Adapted version of the OpenMPI Communication Library, rapport de projet UE QosCosGrid FP6-IST-2005-033883, juin 2009, 28 pages.
16. Camille Coti, Thomas Herault et Ala Rezmerita : D1.2b : Adapted version of the OpenMPI Communication Library, rapport de projet UE QosCosGrid FP6-IST-2005-033883, octobre 2008, 45 pages.
17. Krzysztof Kurowski, Mariusz Mamonski, Piotr Grabowski, Yannick Langlois, Guillaume Mecheneau, Thomas Herault, Camille Coti et Mark Ragan : D1.4 : Second Prototype and Integration of Grid Services Together with QoS-Aware Grid MW Providers, rapport de projet UE QosCosGrid FP6-IST-2005-033883, octobre 2008, 28 pages
18. Camille Coti, Thomas Herault et Franck Cappello : MPI Applications on Grids : A Topology-Aware Approach, rapport de recherche INRIA #6633, septembre 2008, 21 pages.
19. Camille Coti, Thomas Herault, Pierre Lemarinier, Sylvain Peyronnet et Ala Rezmerita : D1.2a : OpenMPI Communication Library, rapport de projet UE QosCosGrid FP6-IST-2005-033883, octobre 2007, 37 pages.
20. David Carmeli, Valentin Kravtsov, Benny Yoshpa, Aassaf Schuster, Krzysztof Kurowski, Camille Coti et Thomas Herault : D2.1 Part 1 : Grid Services for Quasi Opportunistic Super Computing, rapport de projet UE QosCosGrid FP6-IST-2005-033883, avril 2007, 75 pages.
21. Camille Coti, Thomas Herault, Krzysztof Kurowski, Pierre Lemarinier et Guillaume Mecheneau : D1.1 : State of the art/Gap analysis of Existing Grid Middleware Services for CS modeling, rapport de projet UE QosCosGrid FP6-IST-2005-033883, avril 2007, 32 pages.

5.5 Article en soumission

22. Emmanuel Agullo, Camille Coti, Thomas Herault, Sylvain Peyronnet, Ala Rezmerita, Franck Cappello et Jack Dongarra : QCG-OMPI : MPI Applications on Grids, en soumission à *Future Generation Computer Systems*.

6 Pièces jointes

6.1 Photocopie d'une pièce d'identité avec photographie

6.2 Attestation provisoire en remplacement d'une copie du diplôme de doctorat

6.3 Documents de soutenance

- copie du rapport de soutenance ;
- copies des deux rapports de manuscrit de thèse.

6.4 Documents concernant mon statut et mon cursus

- une copie de mon diplôme d'ingénieur ;
- une copie de mon attestation de réussie au diplôme d'ingénieur ;
- une copie de mon contrat de doctorante à l'INRIA Saclay (avec les différents avenants émis).

6.5 Lettre de recommandation pour la recherche

- Franck Cappello, directeur de recherches à l'INRIA Saclay-Île de France, directeur de thèse.

6.6 Lettre de recommandation pour l'enseignement

- Joël Falcou, maître de conférences à l'IFIPS, responsable de la filière d'apprentissage.

6.7 Exemplaires de documents relatifs à mes travaux

- un exemplaire de la thèse ;
- un exemplaire de trois articles publiés en conférence ou revue internationale : FGCS, CCGRID'08 et ParCo'09.