Passage à l'échelle des bibliothèques de communication sur grappes de noeuds massivement multi-coeurs

Responsable: Alexandre DENIS (Alexandre.Denis@inria.fr),
Emmanuel JEANNOT (Emmanuel.Jeannot@inria.fr)

Lieu: INRIA/LaBRI, équipe Satanas, projet Runtime.

Mots-clés: communications réseau, calcul haute-performance, multi-coeur, MPI.

L'émergence des processeurs multi-coeurs a radicalement changé la physionomie des machines courantes et des grappes de calcul. Désormais les noeuds de calcul sont dotés courament de plus de dix coeurs. Cette tendance à l'augmentation du nombre de coeurs par noeud est nette et semble durable, avec l'annonce récente d'Intel de la prochaine génération de Xeon Phi *Knights Landing* dotée de 72 coeurs et qui sera utilisable en tant que processeur principal. À l'avenir, les machines de calcul seront *massivement multi-coeur*.

Cette révolution du matériel a d'énormes répercussions sur le logiciel, et en particulier sur la gestion des communications. En effet, même si certains noeuds sont désormais équipés de plusieurs cartes réseaux, elles sont en tout état de cause en nombre bien plus restreint que le nombre de coeurs. L'accès au réseau est donc source de contention, qui va aller en s'aggravant à mesure que le nombre de coeurs augmente.

Nous avions proposé la bibliothèque NewMadeleine, capable de tirer profit des flux de communications parallèles issus de différents threads, en appliquant une stratégie d'optimisation à la volée sur les séquences de paquets. En s'appuyant sur la bibliothèque d'ordonnancement de tâches d'E/S PIOMan, elle est par ailleurs capable de paralléliser la progression des communications. En revanche, aussi bien la stratégie d'optimisation que l'ordonnancement des tâches reposent sur la connaissance d'un *état*

global de tous les paquets et toutes les tâches ordonnancées ou en cours. Cette approche globale, donc centralisée, est efficace sur un petit nombre de coeurs, mais atteint désormais ses limites avec le grand nombre de coeurs que l'on trouve dans les nouvelles architectures.

Dans le domaine des cartes réseau haute-performance, le modèle de communication a évolué. Le modèle dominant est longtemps resté la réception explicite, dite également *send/recv*. Avec l'émergence et maintenant la domination de la technologie InfiniBand, le modèle le plus répandu est désormais le RDMA, à savoir l'accès mémoire à distance. Dans ce modèle, il est possible d'accéder directement en lecture ou en écriture à la mémoire d'un noeud distant, sans action de sa part. Ce changement de paradigme réseau a également de nombreuses répercussions sur la gestion des communication. En effet, s'il est actuellement courant d'écrire des pilotes qui simulent un mode *send/recv* sur un réseau nativement RDMA, ceci a un impact sur les performances. Il nous semble opportun d'envisager une bibliothèque de communications nativement RDMA de bout en bout, pour tirer pleinenement profit des performances et capacités du réseau.

L'objet de ce sujet de thèse est de s'intéresser aux problématiques posées à bas niveau dans la gestion des communications consécutivement aux évolutions des architectures vers un grand nombre de coeurs et un modèle réseau RDMA. En particulier, il s'agira pour l'étudiant de s'intéresser aux aspects suivants :

- Étudier le passage à l'échelle en fonction du nombre de coeurs des mécanismes impliqués dans la stratégie d'optimisation des paquets et d'ordonnancement de tâches.

- Étudier en quoi le passage au RDMA contribue à la réduction de la contention du côté du récepteur, et ce qu'il implique sur les stratégies d'optimisation.

- Proposer des mécanismes ayant de meilleurs propriétés de passage à l'échelle, éventuellement en renonçant à la connaissance de l'état global. Il semble pertinent de s'orienter vers une approche hierarchique tenant compte de la topologie de la machine.

- Implémenter les mécanismes proposés et évaluer leur comportement sur des micro-benchmarks et applications.

Références bibliographiques

- François Trahay. *De l'interaction des communications et de l'ordonnancement de threads au sein des grappes de machines multi-coeurs.* PhD thesis, Université Bordeaux 1, November 2009.

- François Trahay and Alexandre Denis. *A scalable and generic task scheduling system for communication libraries.* In Proceedings of the IEEE International Conference on Cluster Computing, New Orleans, LA, September 2009. IEEE Computer Society Press.

- François Trahay, Élisabeth Brunet, Alexandre Denis, and Raymond Namyst. *A multithreaded communication engine for multicore architectures.* In CAC 2008: Workshop on Communication Architecture for Clusters, held in conjunction with IPDPS 2008, Miami, FL, April 2008. IEEE Computer Society Press.

- Elisabeth Brunet, François Trahay, Alexandre Denis, and Raymond Namyst. *A sampling-based approach for communication libraries auto-tuning.* In International Conference on Cluster Computing (IEEE Cluster), Austin, Texas, pages 299-307, September 2011. IEEE Computer Society Press.