Compas'2022 : Parallélisme / Architecture/ Système
MIS/UPJV - Amiens France, 5-8 juillet 2022

Interactions entre calculs et communications au sein des systèmes HPC distribués : évaluation et modélisation.

Philippe SWARTVAGHER *

Inria Bordeaux – Sud-Ouest,
200 Avenue de la Vieille Tour
33405 Talence - France
philippe.swartvagher@inria.fr

Résumé
Les supports d'exécution parallèles distribués permettent généralement de recouvrir les communications réseau par des calculs, pour amortir le coût des communications, et ainsi espérer optimiser les performances des applications HPC. Le recouvrement signifie exécuter simultanément les communications et les calculs.
Nous avons commencé par étudier les interactions entre les calculs et les communications lorsqu'ils sont exécutés en parallèle. Nous avons découvert que des interactions existent, et qu'elles peuvent dégrader les performances. En plus des variations de fréquences et du surcoût causé par le support d'exécution, la cause principale de cet impact est la contention mémoire. Cette contention peut être due au trafic au sein du système gérant la mémoire, qui est généré par les accès mémoire émis simultanément par les communications réseaux et les cœurs de calcul. Nous avons montré que cette contention peut fortement pénaliser les performances du réseau, mais également, dans une moindre mesure, perturber les calculs. Nous nous sommes intéressés de plus aux différents facteurs qui font varier la contention mémoire et donc l'impact sur les performances : la contention dégrade plus les performances lorsque les calculs sont limités par les accès mémoire et la taille des messages échangés sur le réseau est importante.
Afin de décrire et prédire le débit mémoire accordé respectivement aux calculs et aux communications lorsqu'ils sont exécutés en parallèle, nous avons proposé un modèle prenant en compte la contention mémoire, ainsi que le placement des données utilisées par les communications et les calculs. L'élaboration de ce modèle nous a permis de comprendre que, parmi les composants du système mémoire, les contrôleurs des nœuds NUMA sont plus sujets à contention que les liens inter-processeurs. De même qu'en cas de contention, le système préfère d'abord réduire la bande-passante mémoire accordée aux communications, pour pouvoir satisfaire le plus longtemps possible les besoins des cœurs exécutant des calculs. Cependant, une bande-passante mémoire minimale est toujours assurée pour les communications. Le modèle a été évalué sur plusieurs machines avec différentes caractéristiques et ses prédictions ont une erreur en moyenne inférieure à 4 %.

Mots-clés : MPI, Contention mémoire, NUMA, Bande-passante, Modèle

---

*. Encadrants : Alexandre DENIS et Emmanuel JEANNOT