

Analyse de performance d'un solveur direct hiérarchique

Abel Calluaud^{1,2}, Mathieu Faverge², David Lugato¹, et Pierre Ramet²

¹CEA CESTA, 15 Avenue des Sablières, 33405 Le Barp, France

²Univ. Bordeaux, CNRS, Bordeaux INP, INRIA, LaBRI, UMR 5800, 33400 Talence, France

Résumé

La recherche de performance pour la résolution de systèmes linéaires de grande taille est la clé pour de nombreuses applications de simulation numérique. Ainsi, le code de simulation ARLENE du CEA s'intéresse à la résolution des équations de Maxwell en fréquence pour étudier les problèmes de furtivité électromagnétique. Il repose sur une méthode de factorisation directe extrêmement gourmande en mémoire et en temps de calcul, limitant la taille des problèmes qui peuvent être étudiés. Les besoins industriels pour la résolution de systèmes toujours plus grands ont motivé l'utilisation de techniques de compression hiérarchique pour réduire les besoins en termes de mémoire et de puissance de calcul. Les approximations \mathcal{H} -matrices ont rendu atteignable en temps raisonnable la résolution sur de très grands systèmes en ramenant la complexité de la factorisation à $O(n \log(n))$ contre $O(n^3)$ pour un algorithme dense. L'introduction de \mathcal{H} -matrices introduit toutefois des difficultés pour une implémentation efficace de par la grande combinatoire de noyaux de calculs qui varient en taille et en intensité arithmétique. Les algorithmes sur ce format hiérarchique s'expriment naturellement à l'aide d'un paradigme de tâches dont l'ordonnancement est délégué à un support d'exécution dédié. Basé sur le modèle Séquentiel Task Flow (STF), celui-ci bénéficie d'extensions permettant d'exploiter les particularités du domaine, notamment la prise en compte des dépendances hiérarchiques de données à grain fin afin d'éviter des synchronisations excessives. Cette approche a permis de factoriser efficacement des problèmes qui comportent jusqu'à des dizaines de millions d'inconnues en quelques heures sur quelques dizaines de milliers de cœurs. Cependant, l'irrégularité des calculs laisse une question de recherche ouverte sur l'adaptation de la granularité des tâches à l'architecture matérielle. Les petits noyaux peuvent d'une part souffrir d'inefficacités et les larges tâches gagneraient à être parallélisées. Dans ce contexte, nous proposons une méthodologie et des métriques afin d'évaluer la performance des noyaux par rapport aux limites matérielles. Nous présentons les résultats de performance des noyaux et en faisons une analyse détaillée. Nous mettons en évidence les problèmes de granularité et faisons une estimation du gain potentiel d'une stratégie de fusion pour le grain faible et de parallélisation pour le gros grain.
