Optimisation de l'énergie et de la performance d'applications sur des micro-serveurs hétérogènes

Massinissa Ait aba[1], Lilia Zaourar[1], Alix Munier Kordon[2]

[1] CEA, LIST, Laboratoire Calcul et Environnement de conception
91191 GIF SUR YVETTE CEDEX, FRANCE.
{massinissa.aitaba, lilia.zaourar}@cea.fr
[2] UPMC, Laboratoire d'informatique de Paris 6
4 PLACE JUSSIEU, 75005 PARIS, FRANCE.
alix.munier@lip6.fr

Mots-clés : *ordonnancement, optimisation, énergie, makespan.*

1 Introduction

Notre vie quotidienne nécessite des calculs massifs importants sur les ordinateurs de calcul à haute performance (calculs physiques, données médicales, météo...) et les centres de données (recherches Google, Facebook...). Afin de traiter ces charges de travail évolutives, les micro-serveurs sont un format émergeant conçus pour traiter des applications de type *scale out*. Pour répondre au mieux aux exigences des marchés, l'hétérogénéité de ces systèmes devient une tendance croissante pour satisfaire les contraintes d'énergie et gérer la puissance de calcul dans la conception et la gestion des micro-serveurs. Le rapport performance par watt est donc devenu un critère d'optimisation important.
Face à la complexité des applications et des architectures, il devient de plus en plus difficile de distribuer les tâches d'une application parallèle de manière efficace. Plus qu'un simple problème d'équilibrage de charge, l'hétérogénéité conduit à reconsidérer les techniques d'ordonnancement pour tenir compte des spécificités des différentes ressources de calcul.
L'objectif de ce travail est de déterminer un ordonnancement des tâches d'une application parallèle sur l'ensemble des ressources hétérogènes du système. Nous cherchons de minimiser à la fois le temps d'exécution total (makespan) et l'énergie totale du système.

2 Modélisation du problème

2.1 Données et notations

Nous avons une plateforme micro-serveur hétérogène dans laquelle M est un ensemble de m unités de traitement hétérogènes (GPU, CPU, FPGA...) notées PE (Processing Element). Chaque élément PE_k de M est caractérisé par sa fréquence d'exécution $f_k \geq 1$. Un coût de communication fixe entre chaque paire PE_k et PE_l est noté Cm_{kl}.
Une application A de n tâches est représentée par un graphe DAG (Directed Acyclic Graph) orienté pondéré $G(V, E, w)$. Chaque sommet $v_i \in V$ représente une tâche t_i qui est caractérisée par son poids w_i, $i \in \{1, ..., n\}$. Chaque arc $e_{i,j}$ représente une contrainte de précédence entre deux tâches t_i et t_j. $e_{i,j}$ est pondéré par un poids $Ct_{i,j}$ qui représente le volume de communications entre t_i et t_j si elles sont exécutées sur deux machines différentes. Une tâche t_i peut s'exécuter seulement après l'exécution complète de tous ses prédécesseurs. Nous ne permettons pas de duplication des tâches, ni de préemption. Une tâche peut être exécutée par toutes les unités de traitement. L'exécution de la tâche t_i sur PE_k engendre un temps d'exécution égal à $execut_{i,k} = \frac{w_i}{f_k}$ et une puissance égale à $P_{i,k} = w_i * f_k^2$ [1].

Dans cette étude, nous avons imposé une borne sur la consommation d'énergie de la plateforme afin de minimiser l'énergie globale du système. On note par D la quantité d'énergie autorisée durant l'exécution. Cette borne doit être au moins égale à la consommation d'énergie engendrée par l'exécution de toutes les tâches sur le PE avec la plus petite fréquence f_{min} (l'exécution de toutes les tâches sur ce PE donne une solution réalisable, cette dernière engendre une consommation d'énergie minimale).

$$D \geq \sum_{i}^{n} w_i * f_{min}^2$$

L'objectif est de minimiser le makespan tout en respectant la borne D.

3 Résolution

Plusieurs méthodes de résolution existent dans la littérature pour le problème d'ordonnancement sur une plateforme hétérogène avec des paramètres et caractéristiques différentes. La plupart des travaux portent sur des heuristiques et des algorithmes génétiques [2,3] ainsi que l'application de la théorie des jeux [4,5].
L'objectif de ce travail est de proposer des algorithmes d'approximation. Dans ce cadre, nous avons traité le cas particulier d'une chaîne de tâches en proposant un algorithme approché avec un rapport d'approximation atteint. Le principe est de trouver la solution optimale pour le problème avec préemption. Nous avons prouvé que la solution obtenue utilise deux PE seulement (PE_k et PE_{k+1}, $k \in \{1..m-1\}$) quand les temps de communication sont importants. On note par S^* la solution optimale de l'ordonnancement. La valeur de la solution approchée est au plus égale à $\frac{f_{k+1}}{f_k} S^*$.
Nous avons ensuite traité le problème d'ordonnancement pour un DAG sans coût de communication et sans contrainte d'énergie. Nous avons montré que la solution obtenue par l'algorithme de liste est de durée au plus égale à $(2 - \frac{1}{m}) \frac{f_{max}}{f_{min}} S^*$. f_{max} représente la plus grande fréquence. Pour $f_{max} = f_{min}$, on retrouve le rapport de Graham.
Enfin, en introduisant une contrainte d'énergie, nous avons proposé un algorithme de réparation d'ordonnancement obtenu sans tenir compte de la contrainte d'énergie pour les petites instances (<150 tâches) et un algorithme de répartition de tâches pour les grandes instances.

Références

[1] Aupy, Guillaume, et al. "Reclaiming the energy of a schedule : models and algorithms." Concurrency and Computation : Practice and Experience 25.11 (2013) : 1505-1523.

[2] Zhang, Longxin, et al. "Bi-objective workflow scheduling of the energy consumption and reliability in heterogeneous computing systems." Information Sciences (2016).

[3] Sheikh, Hafiz Fahad, and Ishfaq Ahmad. "Efficient heuristics for joint optimization of performance, energy, and temperature in allocating tasks to multi-core processors." Green Computing Conference (IGCC), 2014 International. IEEE, 2014.

[4] Perez, Oscar Carlos Vasquez. "Ordonnancement de tâches pour concilier la minimisation de la consommation d'énergie avec la qualité de service : optimisation et théorie des jeux." PhD diss., Université Pierre et Marie Curie-Paris VI, 2014.

[5] Tarplee, Kyle M., et al. "Energy and makespan tradeoffs in heterogeneous computing systems using efficient linear programming techniques." IEEE Transactions on Parallel and Distributed Systems 27.6 (2016) : 1633-1646.