An Interference-aware Virtual Machine Placement Strategy for Small-scale HPC Applications in Clouds

Maicon Alves[1], Lucia Drummond[1]

[1]Universidade Federal Fluminense

Abstract. The cross-interference problem may occur when applications are executed in virtual machines placed in a same physical machine. Although many previous works have proposed several different strategies for Virtual Machine Placement, neither of them have employed a suitable method for predicting cross-interference nor have considered the minimization of the number of used physical machines at the same time. In this thesis, we define the Interference-aware Virtual Machine Placement Problem for small-scale HPC applications in Clouds (IVMPP) that tackles both problems by minimizing, at the same time, the cross-interference of small-scale HPC applications, that can share physical machines, and the number of physical machines used to allocate them. We propose a mathematical formulation and a strategy based on the Iterated Local Search framework to solve this problem. Moreover, we also propose a quantitative and multivariate model to predict interference for a set of applications allocated to the same physical machine. Experiments executed in a real scenario, by using applications from the oil and gas industry and the HPCC benchmark suite, showed that our method outperforms several heuristics from the related literature in terms of interference, while using the same number of physical machines.

Resumo. Em um ambiente de nuvem computacional, aplicacões de alto desempenho podem sofrer interferência ao serem executadas em máquinas virtuais que estejam alocadas em uma mesma máquina física. Embora alguns trabalhos tenham proposto estratégias de alocação de máquinas virtuais cientes deste problema, nenhuma dessas estratégias empregou um método adequado para predizer a interferência nem considerou, ao mesmo tempo, tanto a minimização da interferência quanto do número de máquinas físicas ativas na nuvem. Nesta tese, define-se o Problema de Alocação de Máquinas Virtuais ciente da Interferência para Aplicações de Alto Desempenho de Baixa Escalabilidade, um problema que tem a finalidade de minimizar, simultaneamente, (i) a interferência sofrida por aplicações de alto desempenho que estejam sendo executadas em uma mesma máquina física e (ii)o número de máquinas físicas necessárias para alocar essas aplicações na nuvem. Este trabalho apresenta uma formulação matemática para o problema, além de propor uma estratégia baseada na metaheurística Busca Local Iterada para resolvê-lo. Para predizer a interferência, esta estratégia utiliza um modelo quantitativo e multivariado que leva em conta a quantidade e similaridade de acesso aos recursos compartilhados e o número de aplicações co-alocadas. Uma análise experimental, utilizando aplicações reais da área de petróleo e o benchmark HPCC, mostraram que o

método proposto foi capaz de superar,em termos de redução de interferência, várias heurísticas da literatura. Os resultados revelaram que, mesmo usando o número de máquinas físicas indicados por tais heurísticas, a estratégia proposta conseguiu reduzir o nível de interferência sofrido pelas aplicações alocadas na nuvem.