Exploração do Espaço de Projetos para Alocação de Recursos em Nuvem

Danillo Christi A. Arigoni, Ricardo Ribeiro dos Santos, Liana Dessandre Duenha

[1]Faculdade de Computação (FACOM)
Universidade Federal de Mato Grosso do Sul - UFMS

danillo.arigoni@ufms.br, ricardo@facom.ufms.br, lianaduenha@facom.ufms.br

Resumo. *A computação em nuvem oferece uma enorme gama de recursos computacionais disponíveis sob demanda. Contudo, encontrar a melhor configuração que reduza custos e atenda as exigências do usuário tornou-se um grande desafio. Este desafio compartilha características essenciais com um problema da área de arquitetura de computadores, a exploração de espaço de projetos (Design Space Exploration - DSE). Em DSE, o foco é escolher, dentre uma grande quantidade de soluções arquiteturais, qual a mais indicada para uma determinada demanda, buscando atender objetivos e cumprindo as restrições de projeto. Diante disso, este trabalho propõe a aplicação de técnicas de exploração de espaço de projeto como potencial solução para o problema de alocação de recursos em nuvem. Este trabalho fará uso de técnicas já disponíveis junto ao fluxo da ferramenta de DSE denominada MultiExplorer, resultando, assim, em uma extensão dessa ferramenta para também atuar no problema de alocação de recursos em nuvem.*

1. Introdução

À medida que aumenta a quantidade e a carga de trabalho das aplicações para execução em nuvem, torna-se essencial a otimização dos recursos e serviços que serão disponibilizados nesse ambiente. A heterogeneidade das aplicações gera demandas por recursos computacionais também heterogêneos; por exemplo, alguns *workloads* podem necessitar uso intenso de CPU (são definidos como *CPU-intensive*), enquanto outros podem demandar uso intenso de entrada e saída (*I/O-intensive*), ou ainda necessitar de GPUs ou outros aceleradores. Desta forma, a nuvem deve ter recursos heterogêneos com diferentes capacidades, custos e desempenho [Lee and Katz 2011].

Uma vez conhecida a aplicação é necessário alocar uma combinação de recursos computacionais que melhor atenda a sua necessidade. Deve-se encontrar a máquina que atenda a múltiplos objetivos, por exemplo: maximizar o desempenho da aplicação, minimizar custo, minimizar tempo de execução, minimizar a quantidade de máquinas do *cluster*, minimizar tempo de comunicação entre máquinas do *cluster*, etc [Buyya et al. 2009, Calheiros et al. 2011]. Mesmo com a disponibilidade de ferramentas e propostas de soluções para indicar a configuração de máquina virtual mediante demandas de aplicações, percebe-se que essas soluções ainda apresentam algum tipo de limitação seja pelo tipo de aplicação que conseguem analisar ou mesmo pela limitação do espaço de busca de soluções de máquinas virtuais.

Na área de sistemas de computação, mais especificamente de arquitetura de computadores, exploração do espaço de projeto (*Design Space Exploration* - DSE) é conhecido como a aplicação de técnicas, que podem abranger desde modelos de otimização até algoritmos de aprendizado de máquina, para escolher, dentre uma grande quantidade de parâmetros arquiteturais (o espaço de projeto), soluções arquiteturais para um projeto em particular. As técnicas de DSE, de forma geral, visam atender múltiplos objetivos (como maximizar desempenho ou minimizar consumo energético), obedecendo a diversas restrições de projeto (como custo, área, dissipação de potência, entre outros).

Os dois problemas (alocação de recursos em nuvem e exploração do espaço de projetos) parecem compartilhar características essenciais. Enquanto um demanda configurações adequadas, o outro oferece metodologias e técnicas para explorar alternativas de configurações para atender essas demandas. Diante do exposto, esta proposta de trabalho, em nível de mestrado, tem como objetivo propor uma solução para o problema de alocação de recursos em nuvem utilizando técnicas de exploração do espaço de projetos, aproveitando a heterogeneidade das aplicações e dos recursos disponíveis. Especificamente, vislumbra-se utilizar uma infraestrutura de exploração de espaço de projetos de sistemas computacionais, denominada MultiExplorer [Devigo et al. 2015], e estendê-la para resolver o problema de alocação de recursos em nuvem computacional.

2. Estágio Atual de Desenvolvimento do Trabalho

A ferramenta MultiExplorer, cujo o fluxo de execução é mostrado na Figura 1, preenche uma lacuna experimental e de exploração de projeto de sistemas. Conhecida a carga de trabalho ou mesmo a demanda computacional de uma aplicação, é necessário alocar uma combinação de recursos computacionais que melhor atende essa demanda. MultiExplorer utiliza um algoritmo de exploração de espaço de projetos a partir de uma estratégia genética para evolução de uma solução (arquitetura) inicial [Santos et al. 2018].

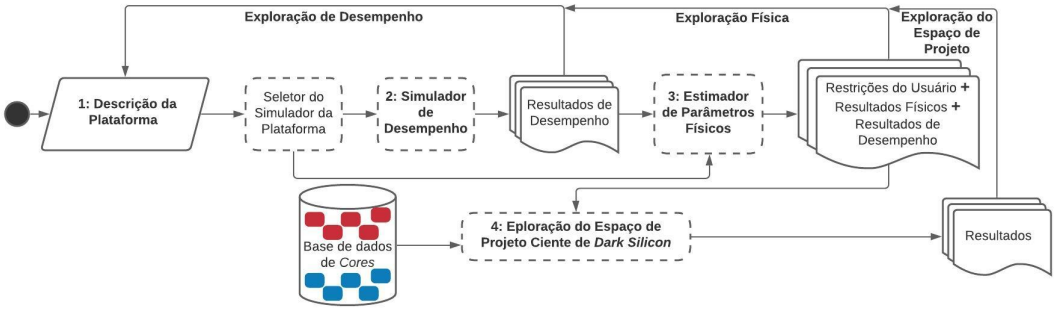

Figure 1. Fluxo de execução do MultiExplorer.

Em se tratando de alocação de recursos utilizando ambiente de nuvem computacional, observa-se constantemente a necessidade de determinar uma ou mais máquinas que atendam a múltiplos objetivos. Diante desse contexto, entende-se que o atendimento do problema de alocação de recursos em nuvem pode ser obtido utilizando os algoritmos e técnicas existentes no MultiExplorer [Santos et al. 2016, Santos et al. 2018]. Essa associação é possível uma vez que a exploração pode contemplar tanto a utilização de plataformas heterogêneas (máquinas virtuais em nuvem) disponíveis quanto a elasticidade dessas plataformas para o atendimento das demandas das aplicações.

O desenvolvimento atual deste trabalho está focado na etapa 4 da Figura 1, em que o módulo de exploração do espaço de projeto é executado visando apresentar alternativas arquiteturais (configurações de recursos computacionais em nuvem) que atendem requisitos e demandas estipulados. Para tanto, esse módulo deve se basear em um banco de dados de cores de exploração que foi previamente povoado com 27 configurações de máquinas virtuais[1] e é a fonte para a busca dessas alternativas arquiteturais. Outra parte relevante do trabalho é o desenvolvimento dos preditores de tempo e custo de configurações de máquinas virtuais. Preditores de tempo e custo são utilizados para estimar a performance oferecida por uma configuração de máquina virtual e o custo associado à utilização dessa configuração. Resultados preliminares sobre o desenvolvimento desses preditores são apresentados na Seção 3.

3. Resultados Preliminares

O desenvolvimento de preditor de tempo (performance) para máquinas virtuais utiliza o simulador de nuvem CloudSim [Goyal et al. 2012] com aplicações do *benchmark* NPB [Bailey et al. 1995]. Esse simulador estima o desempenho e custo de configurações de máquinas virtuais mediante determinadas cargas de trabalho, funcionando assim como um gabarito para um preditor de desempenho. Os resultados gerados pelo CloudSim junto com 27 configurações de máquinas virtuais AWS, geraram um *dataset* com 82862 saídas de simulação com configurações homogêneas e heterogêneas. A partir desses dados, algoritmos de aprendizado de máquina[2] foram avaliados com objetivo de gerar preditores eficientes de tempo e custo de máquinas virtuais em nuvem, que serão utilizados no *framework* de exploração de espaço de projeto MultiExplorer. As Tabelas 1 e 2 apresentam resultados obtidos com diferentes preditores de tempo e custo gerados a partir de técnicas de aprendizado de máquina. Para fins de escolha do preditor, este trabalho considerou a combinação entre o *score* de teste (R^2), o tempo de execução do preditor e o MAPE (*Mean Absolute Percentage Error*).

Nossos resultados preliminares apresentados nas Tabelas 1 e 2 mostram que o modelo baseado em *Decision Tree* apresenta alto score de teste e baixo tempo de execução tanto para os preditores de tempo quanto custo, sendo assim a técnica inicialmente escolhida neste trabalho. Preditores baseados nessa técnica apresentaram o menor tempo de execução entre todos os modelos testados, um alto *score* de teste (Coeficiente de determinação R^2) acima de 99% e um erro percentual médio absoluto (MAPE) de 6,4% e 0,7% para os preditores de tempo e custo, respectivamente.

4. Conclusão

Com a necessidade de encontrar a configuração de máquina virtual adequada de acordo com as demandas de desempenho e custo das cargas de trabalho, este trabalho propõe soluções para o problema de alocação de recursos em nuvem, utilizando técnicas de DSE.

A extensão da ferramenta MultiExplorer para lidar com o problema de alocação de recursos em nuvem, exige ações como a criação de preditores de tempo e custo para

[1] As configurações são baseadas nas maquinas virtuais disponíveis no ambiente *Amazon Web Services - AWS.*

[2] Para todos os métodos e parâmetros foram usados os valores padrões fornecidos pelo Scikit-Learn[Pedregosa et al. 2011].

Modelo	MAPE (%)	Score de Teste (%)	Tempo (s)
KNN n=1	1.241	99.866	0.295
KNN n=2	2.834	99.589	0.302
KNN n=3	4.646	99.411	0.367
RandomForest	5.781	99.998	0.328
DecisionTree	6.378	99.998	0.004
KNN n=4	6.823	99.245	0.378
KNN n=5	9.913	99.153	0.380
KNN n=6	14.751	99.103	0.395
KNN n=7	21.195	99.107	0.407
KNN n=8	31.783	99.064	0.454
KNN n=9	47.877	98.878	0.452

Table 1. Avaliação dos modelos para preditor de tempo.

máquinas virtuais e o povoamento do banco de dados com configurações computacionais em nuvem. Os resultados preliminares mostram que os preditores baseados na técnica de *Decision Tree* destacam-se entre os outros modelos devido ao menor tempo de *score* de teste.

A continuidade de atividades no âmbito deste trabalho objetivam aprofundar o estudo e experimentação com a ferramenta MultiExplorer, com enfoque no algoritmo genético NSGA-II [Cordeiro and Silva-Filho 2010]. Adicionalmente, ênfase será dada no estudo teórico da fronteira de Pareto aplicado à exploração de projetos.

Agradecimentos

Os autores agradecem à Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul (FUNDECT) pelo apoio dado a este trabalho de mestrado através da concessão de bolsa com número de processo 71/700.151/2020. Agradecem também às agências CAPES, CNPq e à UFMS pelo suporte dado às pesquisas desenvolvidas no Laboratório de Sistemas Computacionais de Alto Desempenho (LSCAD).

References

Bailey, D., Harris, T., Saphir, W., Van Der Wijngaart, R., Woo, A., and Yarrow, M. (1995). The nas parallel benchmarks 2.0. Technical report, Technical Report NAS-95-020, NASA Ames Research Center.

Buyya, R., Ranjan, R., and Calheiros, R. N. (2009). Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities. In *2009 international conference on high performance computing & simulation*, pages 1–11. IEEE.

Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A., and Buyya, R. (2011). Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience*, 41(1):23–50.

Modelo	MAPE (%)	Score de Teste (%)	Tempo (s)
DecisionTree	0.722	99.643	0.004
RandomForest	0.855	99.793	0.340
KNN n=1	2.762	98.620	0.269
KNN n=2	3.829	98.706	0.309
KNN n=3	4.416	98.722	0.370
KNN n=4	4.892	98.606	0.413
KNN n=5	5.456	98.374	0.482
KNN n=6	5.897	98.328	0.483
KNN n=7	6.369	98.352	0.529
KNN n=8	6.834	98.300	0.536
KNN n=9	7.343	98.148	0.605
Polynomial5	13.455	99.986	0.107
Polynomial4	21.243	99.921	0.059
Polynomial3	36.157	99.573	0.022
Polynomial2	80.120	97.357	0.008

Table 2. Avaliação dos modelos para preditor de custo.

Cordeiro, F. and Silva-Filho, A. (2010). Nsgaii applied to unified second level cache memory hierarchy tuning aiming energy and performance optimization. In *2010 11th Symposium on Computing Systems*, pages 64–71. IEEE.

Devigo, R., Duenha, L., Azevedo, R., and Santos, R. (2015). Multiexplorer: A tool set for multicore system-on-chip design exploration. In *26th International Conference on Application-specific Systems, Architectures and Processors*, pages 160–161. IEEE.

Goyal, T., Singh, A., and Agrawal, A. (2012). Cloudsim: simulator for cloud computing infrastructure and modeling. *Procedia Engineering*, 38:3566–3572.

Lee, G. and Katz, R. H. (2011). Heterogeneity-aware resource allocation and scheduling in the cloud. *HotCloud*, 11:4–8.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Santos, R., Duenha, L., Silva, A. C., Sousa, M., Tedesco, L. A., Melgarejo, J. C., Santos, T., Azevedo, R., and Moreno, E. (2018). Dark-silicon aware design space exploration. *Journal of Parallel and Distributed Computing*, 120:295–306.

Santos, T., Silva, A., Duenha, L., Santos, R., Moreno, E., and Azevedo, R. (2016). On the dark silicon automatic evaluation on multicore processors. In *28th International Symposium on Computer Architecture and High Performance Computing*, pages 166–173. IEEE.