Escalabilidade Dinâmica em Nuvens Construídas a partir de Recursos Computacionais Compartilhados

Raphael de Aquino Gomes[1,2], **Fábio Moreira Costa**[1], **Luciana Nishi**[3]

[1]Instituto de Informática – Universidade Federal de Goiás
Câmpus Samambaia, Caixa Postal 131, Goiânia - GO, Brasil

[2]Instituto Federal de Goiás - Campus Goiânia
Rua 75, n. 46, Centro. CEP: 74055-110 - Goiânia - GO - Brasil

[3]Centro Universitário de Anápolis - UniEVANGÉLICA
Av. Universitária Km. 3,5 - Cidade Universitária - Anápolis - GO - Brasil

raphael@ifg.edu.br, fmc@inf.ufg.br, luciana.nishi@unievangelica.edu.br

Abstract. *The specification of the non-functional requirements of an application and the definition of the required resources are activities the are commonly performed in ad hoc way, which can cause quality of service lost due inefficent use of resources. Cloud Computing is an alternative to provision these resources, which can be done using the same infrastructure that is already allocated to run the provider's local applications, which further complicates resource management due to competition with local applications. In this paper we present a set of proposals for resource management in clouds and for provisioning resources in a more efficient way and more easily by the end user.*

1. Introdução

Empresas pioneiras no mercado de computação em nuvem [Mell e Grance 2009] logo perceberam a vantagem de oferecer sua capacidade computacional excedente na forma de serviços de nuvem. Com isto, o gerenciamento de recursos assume caráter duplo: manter os acordos de nível de serviço estabelecidos com os clientes, ao mesmo tempo em que são garantidos recursos suficientes para atender a demanda das aplicações locais. Essas garantias devem ser oferecidas de forma dinâmica, considerando que tanto a demanda das aplicações locais quanto a demanda gerada pelos clientes de computação em nuvem variam com o tempo.

Dessa forma, é crucial mover algumas atividades executadas na fase de projeto para a fase de execução do sistema [Baresi e Ghezzi 2010]. Uma estratégia para conseguir isso é o uso de modelos em tempo de execução [Bencomo et al. 2010]. O uso de modelos possibilita ao cliente expressar seus requisitos de recursos com base nas necessidades das aplicações que esses recursos deverão hospedar; e permite um gerenciamento mais preciso da capacidade computacional disponível.

Diante disso, neste trabalho apresentamos propostas de soluções eficientes para os problemas de gerenciamento dinâmico de recursos e especificação de requisitos em plataformas compartilhadas de computação em nuvem. O objetivo geral consiste em propor um conjunto de mecanismos para o gerenciamento dinâmico de recursos em plataformas de computação em nuvem, considerando o uso, primariamente, de infraestruturas computacionais compartilhadas com aplicações de uso local. O objetivo desses mecanismos,

por sua vez, consiste em tornar mais eficiente a utilização dos recursos da infraestrutura computacional, permitindo o suporte a um número maior de clientes dos serviços de computação em nuvem ao mesmo tempo em que são respeitados os acordos de nível de serviço. As propostas apresentadas neste artigo constituem um trabalho em andamento.

2. Abordagem de Solução

Em nossa proposta, a especificação dos requisitos da aplicação é definida pelo usuário na forma de um modelo. Além dos requisitos, o modelo também pode incluir restrições impostas pelo próprio cliente ou por outro *stakeholder* envolvido no sistema. Exemplos de restrições incluem características físicas como quantidade de memória e capacidade de processamento, além de características orçamentárias, como o limite de gastos aceitável para a manutenção do recurso.

A especificação dos recursos de hardware necessários na fase de implantação é derivada em outra etapa, que recebe como entrada o modelo dos requisitos das aplicações e gera, como saída, um outro modelo, o qual representa os recursos necessários para satisfazer os requisitos. O modelo de recursos gerado representa a infraestrutura ótima específica para a aplicação.

A alocação de recursos é feita de acordo com o modelo de recursos gerado. Inicialmente, a nuvem formada pela infraestrutura computacional da organização, é analisada para verificar a possibilidade de admitir a nova aplicação, dado o modelo de recursos estabelecido para ela. Se for possível a alocação dos recursos, a aplicação é alocada neste ambiente, por meio da criação de uma ou mais máquinas virtuais conforme a especificação de recursos derivada anteriormente. Caso contrário, outros provedores são considerados, sendo, portanto, realizada a descoberta de recursos nos provedores disponíveis. Neste caso é utilizada informação sobre os tipos de recursos disponíveis na infraestrutura. Essas informações são mantidas na forma de um repositório que é atualizado sempre que há alguma mudança nos tipos de recurso disponíveis na infraestrutura. Como os tipos disponíveis não representam necessariamente os recursos estabelecidos para a aplicação, a seleção é feita através de aproximação para os tipos mais adequados.

Somente após a alocação e o escalonamento do recurso é feita a criação da *appliance* virtual correspondente à aplicação e o início de sua execução. Neste contexto, um recurso toma a forma de uma máquina virtual com configuração compatível com o modelo de recursos ou com o tipo selecionado. Durante a execução da aplicação é realizado o monitoramento contínuo para avaliar possíveis mudanças na definição de requisitos da aplicação e/ou na disponibilidade de recursos, o que pode causar a geração de um novo modelo e tarefas adicionais de gerenciamento.

Referências

Baresi, L. e Ghezzi, C. (2010). The disappearing boundary between development-time and run-time. In *Proceedings of the FSE/SDP workshop on Future of software engineering research*, pages 17–22. ACM.

Bencomo, N., Whittle, J., Sawyer, P., Finkelstein, A., e Letier, E. (2010). Requirements reflection: requirements as runtime entities. In *Software Engineering, 2010 ACM/IEEE 32nd International Conference on*, volume 2, pages 199–202. IEEE.

Mell, P. e Grance, T. (2009). Draft NIST working definition of cloud computing.