Análise do Impacto da Elasticidade em Nuvens Computacionais Híbridas

Emanuel F. Coutinho[1], Danielo G. Gomes[2], Maurício M. Neto[3]
Leonardo O. Moreira[1], José Neuman de Souza[3]

[1]Instituto Universidade Virtual (UFC VIRTUAL)
[2]Departamento de Engenharia em Teleinformática (DETI)
[3]Mestrado e Doutorado em Ciência da Computação (MDCC)
Universidade Federal do Ceará – Fortaleza – CE – Brasil

emanuel@virtual.ufc.br, danielo@ufc.br, maumneto@alu.ufc.br

leoomoreira@virtual.ufc.br, neuman@ufc.br

Abstract. *Hybrid computational clouds can be an attractive option in the use of computational resources with elasticity. However, adequate elasticity management is necessary because computational resources can be inefficiently provisioned, resulting in waste or idleness. This paper proposes an analysis of the elasticity effects on a hybrid computational cloud. As a main result, it was identified that elasticity maintains the quality level defined for the attendance of applications executed in a hybrid cloud in face of varied workloads.*

Resumo. *Nuvens computacionais híbridas podem ser uma atrativa opção na utilização de recursos computacionais com elasticidade. Porém é necessária uma gestão adequada da elasticidade, pois recursos computacionais podem ser provisionados de maneira ineficiente, resultando em desperdício ou ociosidade. Este artigo propõe uma análise dos efeitos da elasticidade em uma nuvem computacional híbrida. Como principal resultado, identificou-se que a elasticidade mantém o nível de qualidade definido para o atendimento das aplicações executadas em uma nuvem híbrida diante de cargas de trabalho variadas.*

1. Introdução

O aumento do acesso aos ambientes de nuvens computacionais, devido à facilidade de utilização e pagamento por uso, provoca um acréscimo na quantidade de usuários e suas respectivas cargas de trabalho. Assim, provedores de serviços devem gerenciar melhor seus recursos para garantir o nível de qualidade acordado (*Service Level Agreement* - SLA). Caso contrário, ocorrendo violações no SLA, multas podem ser aplicadas. Devido a esse crescimento na utilização dos recursos computacionais, uma das características principais da computação em nuvem tem se tornado bastante atrativa: a elasticidade.

[Herbst et al. 2013] definiram elasticidade como o quanto um sistema é capaz de se adaptar a variações na carga de trabalho pelo provisionamento e desprovisionamento de recursos de maneira autonômica, de modo que em cada instante no tempo os recursos computacionais disponíveis combinem com a demanda da carga de trabalho o mais próximo possível. Atualmente diversos provedores (e.g. Amazon EC2, Microsoft Azure e HP Cloud Services) disponibilizam serviços de elasticidade aos usuários.

Em geral, para avaliar o desempenho de sistemas computacionais são utilizadas métricas de alguma característica do ambiente, como CPU, memória e rede, pois muitas vezes a elasticidade está associada a algum recurso do provedor. Além disso, existe uma grande quantidade de tecnologias e estratégias para o provimento da elasticidade, aplicadas aos diferentes tipos de nuvens computacionais.

Nuvens híbridas surgem como uma atraente opção em termos de custo/benefício. Entretanto, sem uma gestão adequada da elasticidade, recursos computacionais podem ser provisionados de maneira ineficiente, resultando em desperdício de recursos ou ociosidade. A nuvem híbrida possibilita aos usuários que, possuam acesso a sua própria nuvem privada, possam redimensionar recursos de computação para as nuvens públicas ocasionalmente [Imai et al. 2013]. Nesse contexto, podemos analisar a seguinte hipótese: a elasticidade mantém o nível de qualidade definido para o atendimento das aplicações executadas em uma nuvem computacional híbrida.

O SLA definido para o atendimento das aplicações executadas em uma nuvem híbrida é um fator importante, pois diferentes ambientes computacionais são mais complexos de se integrarem entre si, e de manter a qualidade no atendimento aos serviços dos usuários. Além disso, diferentes provedores possuem diferentes tecnologias e políticas as quais os usuários se submetem. Espera-se que haja uma associação com a elasticidade, pois esta capacidade se propõe a ajustar os recursos do ambiente de maneira que estes se adequem às necessidades das cargas de trabalho impostas. Também se espera uma manutenção do SLA devido à adequação dos recursos provida pela elasticidade. Nesse contexto, o objetivo deste trabalho é avaliar os efeitos da elasticidade sobre uma nuvem computacional híbrida diante de cargas de trabalho aplicadas ao ambiente.

2. Material e Métodos

Os experimentos têm como objetivo avaliar a elasticidade de uma nuvem híbrida, onde inicialmente são utilizados recursos de uma nuvem privada, e conforme a necessidade por mais recursos, estes são adicionados a partir de uma nuvem pública. Para apoiar a análise de desempenho, utilizou-se um *framework* conceitual específico para elasticidade em nuvem, apresentado em [Coutinho 2014]. Ele é composto por três macroatividades, relacionadas ao planejamento, inicialização de serviços e análise.

Os experimentos utilizaram dois ambientes diferentes de nuvens computacionais: uma nuvem privada e uma nuvem pública. Para a nuvem privada, o OpenNebula 3.8, com cada máquina virtual foi criada com 1 VCPU, 1 GB de memória RAM e sistema operacional Linux Ubuntu Server 12.04 64 bits. Para a nuvem pública, foi utilizada a plataforma da Microsoft Azure, com instâncias criadas do tipo A1 padrão (1 núcleo e 1.75 GB de memória RAM) e sistema operacional Linux Ubuntu Server 14.04 64 bits. Utilizou-se como servidor *web* o Apache Tomcat, balanceador de carga o NGINX, e gerador de cargas de trabalho HTTPERF. Para a instanciação das atividades do *framework* utilizou-se Java e *shell script*. A Figura 1 exibe o *testbed* utilizado, baseado na arquitetura proposta por [Coutinho et al. 2016b].

Arquivos texto registram o *log* das operações de coleta das máquinas virtuais e das demais informações consolidadas. O *log* gerado para cada máquina virtual em arquivo texto contém a data da coleta, valores de utilização de CPU, memória, disco, rede e tempo de resposta das requisições, assim como a média de utilização de CPU, alocação

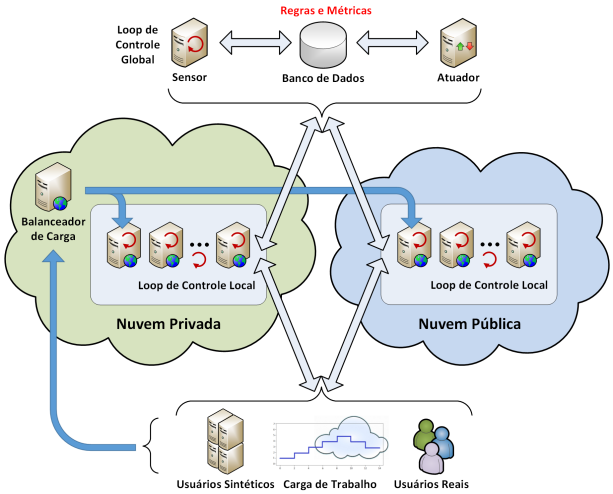

Figura 1. Arquitetura e ambiente experimental [Coutinho et al. 2016b]

de recursos e dados sobre a elasticidade. Os dados coletados e consolidados são apresentados sob a forma de gráficos de linha. O intervalo de coletas definido foi de 1 segundo adicionado do custo da coleta e análise dos resultados.

A geração de cargas de trabalho para os experimentos ocorreu de duas maneiras distintas: (i) requisições encaminhadas diretamente na máquina virtual do balanceador de carga, geradas pelo HTTPERF e por navegadores *web*, distribuídas entre as demais máquinas virtuais alocadas; e (ii) requisições executadas diretamente nas máquinas virtuais da infraestrutura. Dessa maneira é possível emular a concorrência pelos recursos em um ambiente de computação em nuvem.

Para a elasticidade, um mecanismo baseado na arquitetura proposta em [Coutinho et al. 2016b] foi implementado. Uma estratégia de elasticidade horizontal foi utilizada, onde sempre que recursos são necessários, novas máquinas virtuais são adicionadas ou removidas do balanceador de carga conforme a necessidade. Para disparar ações de elasticidade, foi utilizada a média do percentual de utilização de CPU das máquinas virtuais.

Algumas métricas específicas para a elasticidade foram utilizadas neste trabalho. As métricas propostas por [Coutinho et al. 2016a] são baseadas nos conceitos de elasticidade da Física e da Microeconomia, e avaliam a alocação e ajuste de recursos de uma nuvem computacional, indicando a necessidade por mais recursos ou não. Tais métricas possibilitam uma avaliação da elasticidade por meio de uma análise numérica e gráficos. Adicionalmente, as métricas de Elasticidade de *Scaling Up* e Elasticidade de *Scaling Down* [Herbst et al. 2013] foram utilizadas para comparação.

Os limiares utilizados para as ações de elasticidade foram: acima de 70% (aloca uma nova máquina virtual), abaixo de 60% (desaloca uma máquina virtual), e entre 60% e 70% (mantém alocação). Esse valor foi calculado como a média das 10 últimas coletas de utilização de CPU nas máquinas virtuais. Como mecanismo de predição foi utilizado regressão multilinear sobre valores de utilização de CPU, memória, disco e rede. Para o provisionamento dos recursos, a estratégia de balanceamento de carga foi utilizada, onde máquinas virtuais são adicionadas conforme a necessidade.

3. Resultados

Para o experimento, apenas uma máquina virtual na nuvem privada e na nuvem pública foram utilizadas. A Figura 2 exibe a média de utilização de CPU consolidada em todas as máquinas virtuais, a alocação das máquinas virtuais, as métricas de elasticidade baseadas em conceitos da Física e da Microeconomia [Coutinho et al. 2016a] e o tempo de resposta das requisições. Este experimento teve duração de 36min10s. Dessa maneira foi possível verificar se a infraestrutura proveria recursos a partir das duas nuvens, constituindo uma nuvem híbrida, conforme a necessidade gerada pela carga de trabalho aplicada.

A média do percentual de utilização de CPU foi bastante diversificada, chegando em alguns pontos próximo a 100% de utilização. Isto ocorreu devido a muitas requisições e poucas máquinas virtuais para o atendimento. Entretanto, sempre que o limiar de 70% estabelecido como limite superior era ultrapassado, a máquina virtual da nuvem pública era adicionada, reduzindo o valor do percentual médio de utilização de CPU. Violações ocorreram por causa da utilização de apenas duas máquinas virtuais no experimento (poucos recursos), e assim que a utilização de CPU superava o limite e ambas máquinas virtuais estavam alocadas, não existia mais a possibilidade de se alocar mais recursos. Momentos de alocação das duas máquinas virtuais em geral coincidem com momentos de alta utilização de CPU.

As métricas para elasticidade possuíram valores baixos até cerca da metade do

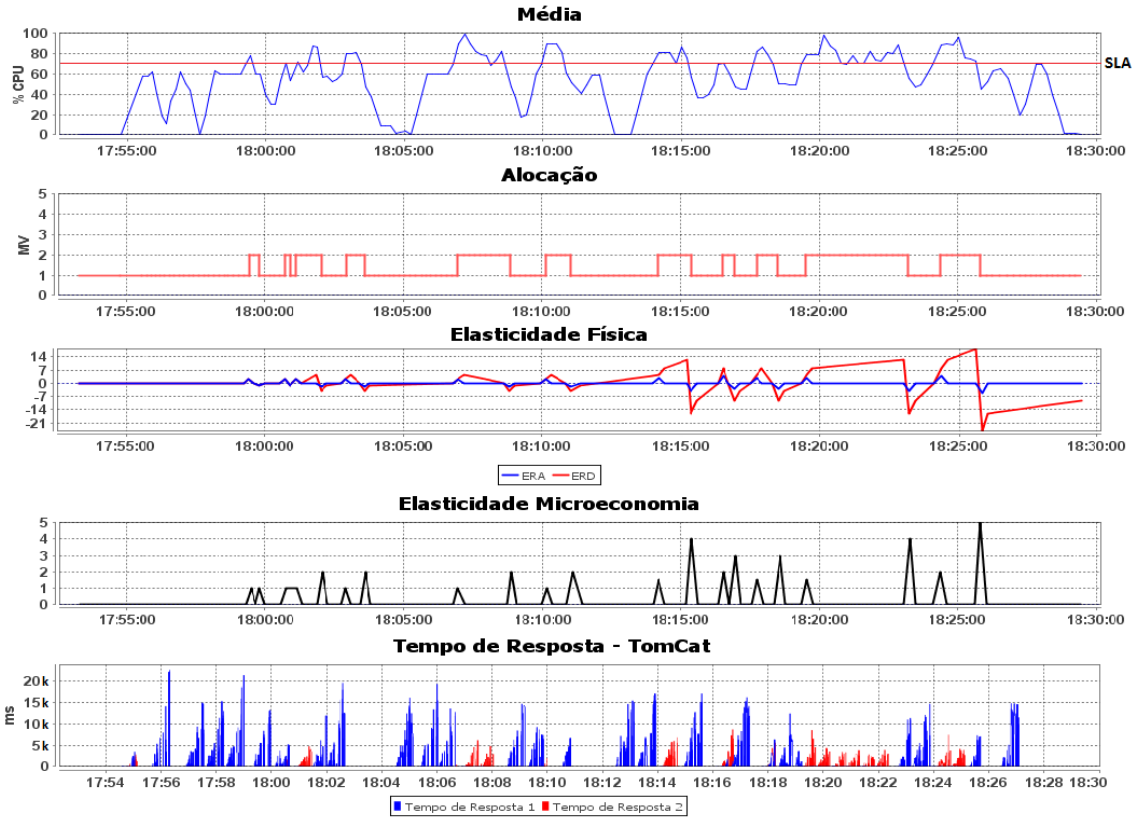

Figura 2. Média de utilização de CPU, alocação das máquinas virtuais, métricas de elasticidade baseadas em conceitos da Física e da Microeconomia e tempo de resposta das requisições.

experimento. A partir da metade do tempo do experimento os valores começaram a aumentar, e a utilização média de CPU em geral permaneceu elevada, indicando uma necessidade maior de recursos. Na maioria dos casos, momentos de alocação e desalocação coincidiram com os picos nas métricas de elasticidade. O que provocou essa variação foi o SLA definido e a velocidade na qual recursos são alocados e desalocados.

O tempo de resposta das requisições possuiu os maiores valores no início do experimento, e à medida em que o experimento avançava, estes tempos foram reduzindo. Como só haviam duas máquinas virtuais disponíveis para o experimento, assim que as requisições eram distribuídas pelo balanceador de carga, o tempo de resposta diminuía. Porém isso não implicava no pleno atendimento do SLA em todo o período do experimento, visivelmente identificado no gráfico de utilização de CPU.

3.1. Discussão dos Resultados

A Tabela 1 exibe métricas de elasticidade dos experimentos e da literatura. A Elasticidade de *Scaling Up* e Elasticidade de *Scaling Up* indicaram a velocidade de alocação e desalocação de recursos. Isso se deve ao fato de apenas existir duas máquinas virtuais envolvidas no experimento, e rapidamente se alocava e desalocava recursos.

A duração das requisições em vários pontos dos experimentos resultou em altos valores. Em geral, uma requisição duraria 2ms, e algumas duraram cerca de 20.000ms, sendo muito tempo para uma requisição. O tempo de resposta não foi considerado nas regras para ações de elasticidade, os três experimentos foram prejudicados no atendimento das requisições. O ideal é que métricas de aplicação, como o tempo de resposta e vazão, sejam incluídas nas regras para a execução de ações de elasticidade.

A utilização de máquinas virtuais em nuvens públicas não teve um alto impacto nos resultados dos experimentos do ponto de vista de infraestrutura. Entretanto, utilizar o percentual médio de utilização de CPU como métrica para acionar as ações de elasticidade poderia ser mais efetiva, pois facilmente se atingiria o limite definido, gerando violações no SLA e altos tempos de resposta das requisições. A latência da rede pode ter sido um fator que influenciou no elevado tempo de resposta, podendo também ter sido utilizada como uma regra para elasticidade.

A Figura 2 exibiu EDi picos com uma diferença visual em relação à largura dos picos, sendo relativamente estreitos, indicando uma não estabilidade crescente nas alocações dos experimentos. Para $ERAi$ e $ERDi$, a partir da metade do tempo dos experimentos, ocorreu uma maior variação nos valores, indicando uma alocação mais intensa. Entretanto, a alta variação entre extremos indicou uma necessidade de recursos maior do que o ambiente possui.

Foi observado pelos experimentos realizados que na maior parte do tempo a elasticidade manteve o SLA definido para o atendimento das requisições na nuvem híbrida.

Tabela 1. Métricas de elasticidade

Métricas	Experimento
Elasticidade Média da Física ($\overline{E_{RAi}}$)	0.00
Elasticidade Média da Física ($\overline{E_{RDi}}$)	-0.08
Elasticidade Média da Microeconomia ($\overline{E_{Di}}$)	0.70
Elasticidade de *Scaling Up*	0.04
Elasticidade de *Scaling Down*	0.09

Entretanto, estabilidade não necessariamente implica em atendimento ao SLA. O experimento possuiu maior estabilidade nos estados de alocação, porém necessitou de mais recursos e mais tempo passou alocando e desalocando recursos, obtendo a maior quantidade de violações. Em relação à quantidade de recursos alocados, o experimento, mesmo tendo apenas 2 máquinas virtuais envolvidas, realizou várias alocações e desalocações.

Para a melhoria do ambiente, sugere-se a utilização de novas regras para a execução de ações de elasticidade, sendo a mais indicada o tempo de resposta das requisições e a latência da rede. É possível utilizar estratégias de alocação e predição mais efetivas, para evitar momentos longos de subprovisionamento e sobreprovisionamento de recursos, evitando violações no SLA, ociosidade e desperdício.

4. Conclusão

O mecanismo de elasticidade aplicado manteve o nível de qualidade definido para o atendimento das requisições na nuvem híbrida, confirmando a hipótese proposta. As principais conclusões deste trabalho são: (i) as métricas de elasticidade aplicadas possuem alta relação com o projeto de carga de trabalho; (ii) a elasticidade se comportou bem em relação à utilização de CPU, mas não foi eficiente em relação ao tempo de resposta das requisições; (iii) métricas de aplicação, como tempo de resposta, devem ser utilizadas em ações de elasticidade e em regras para melhorar o desempenho do ambiente.

Apenas métricas de infraestrutura foram utilizadas como critério de qualidade e ações de elasticidade, o que prejudicou a avaliação do ambiente com métricas de aplicação como o tempo de resposta. O ideal era ter mais experimentos para comparar tais valores, e ter uma análise mais completa, com subsídio das métricas calculadas. Como trabalhos futuros pretende-se investigar os efeitos da elasticidade em nuvens híbridas com diferentes mecanismos de elasticidade, como elasticidade vertical, replicação e estratégias combinadas, utilizando métricas de aplicação.

Referências

Coutinho, E. F. (2014). *FOLE: Um Framework Conceitual para Avaliação de Desempenho da Elasticidade em Ambientes de Computação em Nuvem*. Doutorado, Mestrado e Doutorado em Ciência da Computação (MDCC), Universidade Federal do Ceará (UFC), Fortaleza.

Coutinho, E. F., Rego, P. A., Gomes, D. G., and de Souza, J. N. (2016a). Physics and microeconomics-based metrics for evaluating cloud computing elasticity. *J. Netw. Comput. Appl.*, 63(C):159–172.

Coutinho, E. F., Rego, P. A. L., Gomes, D. G., and de Souza, J. N. (2016b). An architecture for providing elasticity based on autonomic computing concepts. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16.

Herbst, N. R., Kounev, S., and Reussner, R. (2013). Elasticity in cloud computing: What it is, and what it is not. In *Proceedings of the 10th International Conference on Autonomic Computing(ICAC 2013), San Jose, CA*, pages 23–27. USENIX.

Imai, S., Chestna, T., and Varela, C. (2013). Accurate resource prediction for hybrid iaas clouds using workload-tailored elastic compute units. In *Utility and Cloud Computing (UCC), 2013 IEEE/ACM 6th International Conference on*, pages 171–178.