Um estudo sobre a medição do tempo de resposta de Servidores Web instanciados em uma plataforma global de Computação em Nuvem

Walber José Adriano Silva
Departamento de Engenharias e Tecnologia
Universidade Federal Rural do Semi-Árido - Pau dos Ferros, Brasil
walber.silva@ufersa.edu.br

***Resumo*—Servidores Web são tecnologias onipresentes na Internet. O modelo de Computação em Nuvem possibilita a criação de instâncias com o mínimo de esforço gerencial e em localidades diversas no globo. Como o tempo de resposta é uma das métricas para se definir Acordos de Nível de Serviço, este estudo investiga, através de experimentos no lado do cliente (em uma arquitetura Cliente-Servidor), a variação desta métrica em acessos a múltiplas instâncias em uma plataforma global de Computação em Nuvem. Aspectos como quantidade de informações acessadas (tamanhos de páginas Web estáticas) e a localidade dos acessos aos Servidores Web foram consideradas (e.g., servidores Web em São Paulo, Frankfurt, Sydney e Tóquio). Verificou-se que quantidades pequenas de dados em páginas Web, e.g., 1KB, permite que usuários não experimentem lentidão acima de um máximo aceitável (acima de 1 segundo). Contudo, quando a quantidade de dados se eleva, maior quantidade de bytes na página Web, os resultados deste trabalho indicam que ter os acessos acontecendo o mais próximo possível da localidade onde a página foi solicitada é a melhor decisão em termos de redução da métrica tempo de resposta.**

***Palavras-chave*—Plataforma de Computação em Nuvem, Tempo de resposta, Servidor Web.**

I. Introdução

O crescimento da Internet oferece problemas na perspectiva da arquitetura e dos serviços da rede [1]. Um deles é a variação do tempo de resposta durante os acessos aos Servidores Web [2]. Essa variação de tempo entre o início de uma solicitação e o devido processamento no lado do cliente pode ter várias causas, tais como sinais que viajam através de múltiplos enlaces de comunicação, tempo de processamento, sincronização, entre outros. Nesta pesquisa apresenta-se um estudo da medição do tempo de resposta de Servidores Web instanciados em uma plataforma global de Computação em Nuvem.

Computação em Nuvem é um modelo que habilita o acesso ubíquo, conveniente, sob demanda via rede a um agrupamento de recursos computacionais configuráveis que podem ser rapidamente provisionado e liberado com um esforço gerencial mínimo [3]. Uma das tecnologias habilitadas pela Computação em Nuvem é o Servidor Web.

Servidor Web é uma tecnologia onipresente na Internet. Por meio deles, os sites de diversas organizações oferecem serviços e produtos para clientes ou usuários em todo o mundo. O domínio Internet Live Stats [4], que produz estatísticas sobre o estado atual da Internet, relatou que o número de sites disponíveis na Internet está perto de 2 bilhões na metade de 2021. Além disso, o tempo de resposta é uma métrica preciosa para definir o consumo de um serviço na Internet por parte dos usuários [5]. O desempenho de sites varia em relação ao tipo de serviço que o usuário irá consumir e isso pode afetar as expectativas do usuário. Por exemplo, uma pessoa navegando na Internet será mais paciente esperando o carregamento de vídeo do que aguardando pelo resultado de uma consulta de um texto em um site de buscas na Internet por causa da quantidade de dados envolvida em cada um dos tipos de solicitação.

É importante destacar que desde meados da década de 90, estudos sobre o desempenho de Servidores Web com a métrica de tempo de resposta vêm sendo realizados [6], [7]. Desde então, muitas melhorias tecnológicas foram feitas tanto no software quanto no hardware para reduzir o tempo de resposta dos Servidores Web. Contudo, a relevância e importância de se investigar o tempo de resposta de aplicações em Servidores Web em um ambiente de Computação em Nuvem está no fato de tal métrica ser fundamental na definição de Acordos de Nível de Serviço (ou *Service Level Agreement* - SLA) [8].

Além de servir para definição de SLAs, o tempo de resposta pode definir o sucesso e satisfação da interação entre usuários, clientes, e sites Web, providos por Servidores Web [5]. Medir tal métrica levando em consideração as localidades das partes que realizam essa interação e a quantidade de dados envolvida culmina na justificativa desta investigação experimental. Portanto, o objetivo deste estudo é realizar experimentos para medir o tempo de respostas de Servidores Web instanciados em diversas localidades em uma plataforma global de Computação em Nuvem quando for utilizado diferentes tamanhos de páginas Web.

II. Revisão da literatura

Um estudo clássico de um modelo de desempenho de Servidor Web [7], que se usa a Teoria das Filas na modelagem do sistema, onde se captura a essência em sistemas de

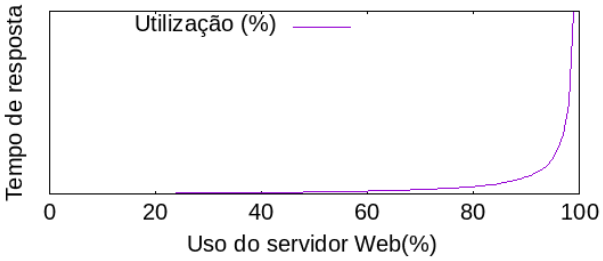

Figura 1: O comportamento do tempo de resposta e a utilização de uma máquina de serviço Web seguindo o modelo simples de fila do tipo $M/M/1$.

servidor único. O trabalho define um limite superior teórico na capacidade de serviço de um Servidor Web. As métricas usadas, nesse estudo, foram os tempos de resposta para uma solicitação de tamanho de arquivo fixo, velocidade do servidor e largura de banda da rede. Destaca-se que o tempo de resposta aumenta para o infinito ao se aproximar da utilização total (vide Figura 1). Esse limite é particularmente sensível ao tamanho médio dos arquivos servidos pelos servidores Web. Os autores encontraram que utilizar múltiplos servidores em paralelo mitiga o tempo de respostas para este tipo de solicitação (aumento da capacidade), e que quando a velocidade da rede é o gargalo, aumentar a velocidade da rede gera um desempenho melhor do que a ampliação da capacidade do sistema.

Em plataformas de Computação em Nuvem, os recursos computacionais podem ser provisionados e reconfigurados a tal ponto que ajustes ao desempenho, por exemplo um Servidor Web, possam ser sanados. Contudo, a localidade onde estes Servidores Web estão e a quantidade de dados a serem transmitidos podem causar um impacto no tempo de resposta percebido pelo cliente durante acesso aos serviços destes servidores. Trabalhos na literatura definem limiares de valores para a métrica tempo de resposta quando a interação Cliente-Servidor é realizada pelo usuário [5], [9]. Desta forma, é possível definir objetivamente o que um usuário irá perceber durante o acesso a um site Web (provido via Servidor Web). Diferente de trabalhos que adotam modelos e simulações, este trabalho realiza experimentos em uma plataforma de Computação em Nuvem que têm recursos computacionais disponíveis em diversas localidades do planeta.

III. Medindo o tempo de resposta

A experimentação da coleta das medições do tempo de resposta de múltiplos acessos aos Servidores Web foi desenvolvida como base na arquitetura Cliente-Servidor descrita na Figura 2.

No lado do Cliente, o serviço de monitoramento Zabbix[1] foi instalado em uma máquina com imagem do Sistema Operacional Ubuntu 20.04. Este cliente dispara as solicitações de acessos às páginas Web nos Servidores Web instanciados

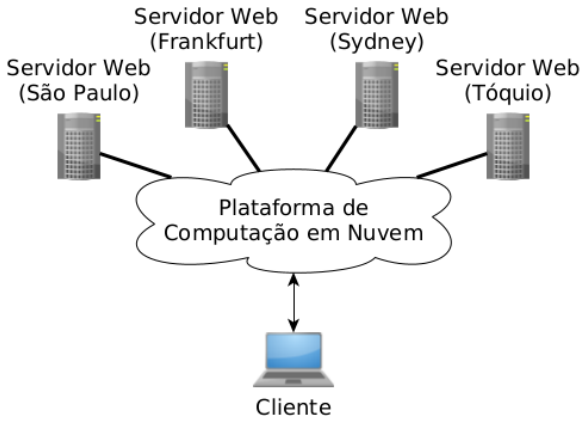

Figura 2: Arquitetura utilizada na realização das medições no lado do cliente.

na plataforma de Computação em Nuvem disponíveis em diferentes localidades do planeta. Este procedimento simula o acesso às páginas Web como se fosse executado por um usuário humano e, dessa forma, o Zabbix executando no Cliente realiza a coleta dos dados relacionados ao tempo de resposta de acesso aos servidores espalhados pelo mundo. Em termos de localidade, o Cliente executa em uma máquina localizada em uma cidade do nordeste brasileiro.

Também foi coletado dados sobre o atraso no tempo de viagem ida-e-volta (*Round-Trip Time*) nas solicitações do Cliente aos Servidores Web. Esta medição indica o tempo total de viagem ida-e-volta de pacotes saindo do Cliente até os Servidores em cada cidade selecionada neste estudo. A coletada usou comando o utilitário *ping* (disponível largamente em diversos Sistemas Operacionais) com o protocolo Internet Control Message Protocol (ICMP). Com essa métrica, é possível verificar o impacto do distanciamento geográfico entre o Cliente e os Servidores Web.

A plataforma global de Computação em Nuvem usada foi a Amazon AWS[2]. Apesar desta plataforma incluir múltiplas cidades ao redor do mundo onde instâncias de Servidores Web podem ser executadas, optou-se por utilizar apenas quatro cidades neste estudo: São Paulo, no Brasil; Frankfurt, na Alemanha; Sydney, na Austrália; e Tóquio, no Japão. Essas cidades estão destacadas na Figura 3 como um meio de referenciar as localidades consideradas.

No lado dos Servidores Web executou-se o Apache2 versão 2.4.41. Em cada servidor, os arquivos no formato de texto de tamanho fixo de 1KB, 100KB e 1000KB estavam disponíveis para serem acessados pelo Cliente (vide Tabela I). Assim, tais arquivos estáticos são os alvos do Cliente para realizar a medição dos tempos de respostas dos Servidores Web. Optou-se por realizar uma consulta de 500 amostras para cada Cliente e arquivo.

[1]O software do Zabbix está disponível em (último acesso: 26 de setembro de 2021): https://www.zabbix.com/

[2]Plataforma disponível em (último acesso: 26 de setembro de 2021): https://aws.amazon.com

Figura 3: Cidades selecionadas para realizar a medição de instâncias.

Tabela I: Tamanho dos arquivos, a cidade onde a instância foi executada, a quantidade de amostras utilizadas no cálculo do tempo de resposta médio e quantidade de amostras descartadas por erro de medição.

Tamanho	Cidade	Número de amostras utilizadas	Número de amostras descartadas
1 KB	São Paulo	500	0
1 KB	Frankfurt	478	22
1 KB	Sydney	476	24
1 KB	Tóquio	479	21
100 KB	São Paulo	500	0
100 KB	Frankfurt	480	20
100 KB	Sydney	478	22
100 KB	Tóquio	481	19
1000 KB	São Paulo	500	0
1000 KB	Frankfurt	483	17
1000 KB	Sydney	476	24
1000 KB	Tóquio	481	19

Sobre a configuração das instâncias que foram utilizadas no experimento, adotou-se o Sistema Operacional Ubuntu Server versão 20.04 LTS de arquitetura x86 de 64-bit e sobre uma plataforma de Virtualização Assistida por *Hardware* (*Hardware Virtual Machine* - HVM). Com relação ao sistema de armazenamento foram utilizados volumes do tipo Disco de Estado Sólido (*Solid-State Drive* - SSD). Além disto, cada instância possuía uma Unidade de Processamento Virtual (vCPU), com 1GB de Memória Principal (RAM) e disco com capacidade de 8GB e velocidade de Operações de Entrada e Saída por Segundo (IOPS) de 100/3000.

O algoritmo descrito no Algoritmo 1 é executado no lado do Cliente pelo serviço Zabbix. O algoritmo recebe os endereços URLs a serem acessados e o número de amostras a serem atingidas para o experimento e, como resultado, gera as medições sobre o tempo de resposta. A lista *D* produzida pelo algoritmo contém os dados a serem analisados.

IV. Resultados e discussão

Após a execução dos experimentos, algumas amostras precisaram ser descartadas da análise final do tempo de resposta

Algoritmo 1 Algoritmo de medição de Servidores Web utilizado no lado do cliente

Input: URLs a serem acessadas *URLS*, número de amostras *N*, tempo entre cada amostra (T)

Output: Dicionário contendo informações sobre o tempo de respostas para cada acesso realizado pela aplicação no lado do cliente (D).

Inicialização: Lista de amostras vazias para o tempo de respostas de cada acesso (L).

1: **for** cada URL em *URLS* **do**
2: $contador = 0$
3: **while** $contador < N$ **do**
4: Aguarda o tempo *T*
5: Medir o tempo de resposta do acesso ao Servidor Web na URL.
6: $contador+ = 1$
7: Adicionar o tempo de resposta em L.
8: **end while**
9: Calcular a média ($media$), o desvio padrão ($desvio$) e o erro padrão da medição ($erro$) com 95% de intervalo de confiança nas amostras em L.
10: $D[URL] = L$
11: **end for**
12: **return** D

Tabela II: Medição da média do tempo de ida-e-volta aos Servidores Web utilizando pacotes ICMP medidos em milissegundos.

Cidade	Tempo (*Round-trip time*)	Intervalo de confiança de 95%
São Paulo	63.18	±1.18
Frankfurt	204.71	±1.06
Tóquio	325.98	±1.05
Sydney	374.83	±3.19

(*outliers*). O critério adotado para exclusão da amostra foi possuir o dobro do valor da média de todas as amostras, que é um critério coerente para tratamaneto de outliers indicado pela literatura [10]. A Tabela I indica o número de amostras descartadas com base neste critério.

A Tabela II indica os tempos do ida-e-volta (*Round-Trip Time*) experimentado pelo Cliente durante a execução dos experimentos. Como é intuitivo, as cidades mais próximas da cidade do Cliente têm tempos de ida-e-volta menores. Contudo, a importância dessas medições está no fato de ter evidências de quanto essa distância está impactando no tempos de respostas aos Servidores Web. Apesar de haver elementos influenciando as medições dos tempos de ida-e-volta, como tempo de processamento nos roteadores entre o caminho Cliente-Servidor, alterações de rotas, roteamento assimétrico, entre outros, a quantidade de amostras e a aplicação de um intervalo de confiança de 95% permite, estatisticamente, medir e ter confiança nos valores finais.

Com base em trabalhos da literatura [5], [9], definiu-se um limiar mínimo e máximo para o tempo de resposta com

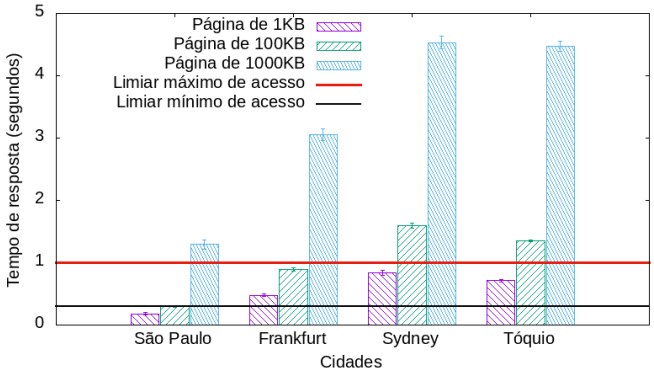

Figura 4: Tempo de resposta de servidores Web para diferentes Datacenters em algumas cidades espalhadas pelo globo.

valores 0,3 s e 1s, respectivamente. Dentro deste intervalo, um usuário humano irá considerar o tempo de acesso aceitável, e.g. definido em um SLA. Abaixo do intervalo a percepção é que a página é carregada de maneira instantânea e acima do intervalo é considerado um acesso lento. Com estes critérios, o tempo de resposta pode ser analisado.

A Figura 4 indica os resultados das medições do tempo de resposta para cada tamanho de arquivo e para as instâncias executando nas cidades alvo deste estudo. Um intervalo de confiança de 95% também foi usado no cálculo do erros das medições (indicado na Figura 4 em cima de cada barra do gráfico). Verificou-se que quantidades pequenas de dados em páginas Web, e.g., 1KB, permite que usuários não experimentem lentidão acima do máximo aceitável (acima de 1 segundo), indicado pela linha "Limiar máximo de acesso". Contudo, quando a quantidade de dados se eleva, maior quantidade de KB na página Web, os resultados indicam que ter os acessos acontecendo o mais próximo possível da localidade que a página foi solicitada é a melhor decisão em termos de redução da métrica tempo de resposta.

A. *Limitações e trabalhos futuros*

Os tempos de resposta de Servidores Web podem ser influenciados por diferentes fatores: rede utilizada pelo cliente, otimização de páginas Web via JavaScript, parâmetros do TCP/IP, configurações no navegador (e.g., Proxies), sistemas de controle de acesso entre o servidor e o cliente (e.g., firewall, VPNs, etc), entre outros. Estudos mais detalhados podem esmiuçar tais elementos influenciadores para essa métrica.

Apesar do estudo focar na métrica tempo de resposta, outras métricas podem ser adotadas em trabalhos futuros, por exemplo as métricas instantâneas como, *document completion time (onLoad)*, *Above-the-fold*, métricas integrais como, tempo de indexação do Google (*SpeedIndex*), e métricas compostas, como *PageSpeed* [11].

Ademais, uma forma de acelerar a entrega de conteúdo (reduzir o tempo de resposta de páginas Web), principalmente se tal conteúdo for estático, é utilizar uma rede de entrega de conteúdo (Content Delivery Network - CDN) [12]. Desta maneira, a interação Cliente-Servidor pode acontecer através de locais mais próximos providenciados pela CDN o que, consequentemente, auxilia na redução da métrica tempo de resposta, que foi adotada neste estudo.

V. Considerações finais

Este estudo realizou experimentos sobre a medição do tempo de resposta de Servidores Web que foram instanciados em uma plataforma de Computação em Nuvem atual e global. O tempo de resposta de Servidores Web é fator essencial na interação Cliente-Servidor e, aqui, os aspectos de localidade e tamanho de páginas Web foram considerados. Verificou-se que a combinação da proximidade dessa interação com a pouca quantidade de dados de uma página permite aos usuários receberem respostas dos Servidores Web dentro de intervalo de tempo adequado.

Como resultado, trabalhos mais complexos que investiguem a experiência do usuário poderão se basear neste estudo para conduzir os experimentos mais elaborados sobre a qualidade da experiência de acesso às aplicações Web [13], bem como, formas alternativas de acessar a entrega de conteúdos de Servidores Web podem também serem extensões deste estudo.

Referências

[1] R. Khan, S. U. Khan, R. Zaheer, and S. Khan, "Future internet: The internet of things architecture, possible applications and key challenges," in *Proceedings - 10th International Conference on Frontiers of Information Technology, FIT 2012*, 2012.

[2] M. Villamizar, O. Garcés, L. Ochoa, H. Castro, L. Salamanca, M. Verano, R. Casallas, S. Gil, C. Valencia, A. Zambrano, and M. Lang, "Cost comparison of running web applications in the cloud using monolithic, microservice, and AWS Lambda architectures," *Service Oriented Computing and Applications*, 2017.

[3] P. Mell and T. Grance, "The NIST Definition of Cloud Computing - Recommendations of the National Institute of Standards and Technology," *NIST*, vol. 1, pp. 1–7, 2011.

[4] Internet Live Stats, "Total number of Websites," 2014. [Online]. Available: http://www.internetlivestats.com/total-number-of-websites/

[5] S. Lohr, "For Impatient Web Users, an Eye Blink Is Just Too Long to Wait," 2012. [Online]. Available: http://www.nytimes.com/2012/03/01/technology/impatient-web-users-flee-slow-loading-sites.html?_r=0

[6] F. Prefect, L. Doan, S. Gold, T. Wicki, and W. Wilcke, "Performance limiting factors in http (web) server operations," *Digest of Papers - COMPCON - IEEE Computer Society International Conference*, pp. 267–272, 1996.

[7] L. P. Slothouber, "A Model of Web Server Performance," *Proceedings of the 5th International World wide web Conference*, vol. 1, no. June 1995, pp. 1–15, 1995.

[8] A. Anand, M. Dhingra, J. Lakshmi, and S. K. Nandy, "Resource usage monitoring for KVM based virtual machines," *2012 18th Annual International Conference on Advanced Computing and Communications, ADCOM 2012*, 2012.

[9] J. Nielsen, "Response Times: The 3 Important Limits," 1993. [Online]. Available: https://www.nngroup.com/articles/response-times-3-important-limits/

[10] R. Jain, *Art of Computer Systems Performance Analysis Techniques For Experimental Design Measurements Simulation And Modeling*. Wiley Computer Publishing, 1991.

[11] E. Bocchi, L. De Cicco, and D. Rossi, "Measuring the quality of experience of web users," *Computer Communication Review*, vol. 46, no. 4, pp. 8–13, 2016.

[12] V. K. Adhikari, Y. Guo, F. Hao, V. Hilt, Z. L. Zhang, M. Varvello, and M. Steiner, "Measurement Study of Netflix, Hulu, and a Tale of Three CDNs," *IEEE/ACM Transactions on Networking*, pp. 1–14, 2014.

[13] H. Z. Jahromi, D. T. Delaney, and A. Hines, "Beyond First Impressions: Estimating Quality of Experience for Interactive Web Applications," *IEEE Access*, vol. 8, pp. 47 741–47 755, 2020.