Arquitetura de balaneamento de carga utilizando aws – amazon web services

Load balancing architecture using aws – amazon web services

Renato José da Silva Camões,
Salvador Alves Melo Junior,
Rafael Alves Araújo

Resumo

Computação em nuvem provê acesso a um conjunto de recursos como máquinas virtuais, armazenamento e rede como serviços. Neste contexto, o balanceamento de carga é importante para computação em nuvem, pois, permite a distribuição de carga de trabalho entre vários nós de processamento que podem ser provisionados sob demanda. Este artigo descreve um modelo para o balanceamento de carga web dinâmico com base em uma distribuição e utilização de recursos de forma igualitária e dinâmica. O trabalho demonstrou os benefícios do balanceamento de carga na nuvem que é capaz de lidar com cargas repentinas, entregando recursos sob demanda e mantendo melhor utilização de recursos e infraestrutura.

Palavras-chave: Nuvem, Balanceamento de Carga, Infraestrutura, WEB

Abstract

Cloud computing provides access to a set of resources like virtual machines, storage and networking as services. In this context, load balancing is important for the cloud cloud as it allows for workload distribution across multiple processing nodes that can be provisioned on demand. This article disciplines a model for dynamic web load balancing based on an equal and dynamic distribution and utilization of resources. The work includes the benefits of load balancing in the cloud that is able to handle sudden loads, delivering resources on demand and maintaining better utilization of resources and infrastructure.

Keywords: *Cloud, Load Balancing, Infraestructure, WEB*

INTRODUÇÃO

Computação em nuvem continua aparecendo na tela do radar dos executivos de Tecnologia da Informação. É uma ideia extremamente sedutora: utilizar os recursos ociosos de computadores independentes, sem preocupação com localização física e sem investimento em hardware. Por isso o interesse nesse assunto que é mais um no reflexo de inúmeros artigos técnicos, publicações e eventos sobre o tema. (TAURION, 2009).

Dentro dessas inúmeras publicações é notável que ambas abordam características essenciais em comum como a presença de um link de internet (para o acesso remoto), compartilhamento de recursos (principalmente hardware) e

escalabilidade. Para diferenciar serão colocadas as definições dadas pelas maiores empresas que disponibilizam os serviços em nuvem.

> Para Microsoft a computação em nuvem é o fornecimento de serviços de computação, incluindo servidores, armazenamento, bancos de dados, rede, software, análise e inteligência, pela Internet (“a nuvem”) para oferecer inovações mais rápidas, recursos flexíveis e economias de escala. Normalmente paga-se apenas pelos serviços de nuvem que são utilizados, ajudando a reduzir os custos operacionais, a executar a infraestrutura com mais eficiência e a escalonar conforme as necessidades da empresa. (MICROSOFT AZURE, 2021).
>
> Para Amazon a computação em nuvem é a entrega de recursos de TI sob demanda por meio da Internet com definição de preço de pagamento conforme o uso. Em vez de comprar, ter e manter datacenters e servidores físicos, pode-se acessar serviços de tecnologia, como capacidade computacional, armazenamento e bancos de dados, conforme a necessidade, usando um provedor de nuvem como a Amazon Web Services. (AMAZON, 2021).
>
> Para Cisco a computação em nuvem oferece às empresas modelos práticos para acessar as ofertas de infraestrutura, plataforma e software com pagamento por utilização. Com a computação em nuvem, as empresas estão liberam capital, otimizando a manutenção de TI, modernizando e dimensionando as abordagens de negócios, criando segurança e flexibilidade em serviços e soluções, ajudando os clientes de novas maneiras e ampliando os negócios sob condições de mercado em constante mudança.

COMO FUNCIONA A COMPUTAÇÃO EM NUVEM – CLOUD COMPUTING

Características essências

O trabalho de Peter Mell (2011) com base no National Institute of Standards and Technology - NIST cita alguns elementos principais dentro da computação em nuvem, dentre eles:

- **Autosserviço sob demanda** - Provisionamento de forma autônoma os recursos de computação que será implementado**;**
- **Amplo acesso por rede** - Para acesso aos recursos utilizados, há uma padronização e o acesso pode ser feito em dispositivos com pouco hardware, já que na maioria das vezes o acesso é feito pelo Browser no dispositivo)**;**
- **Agrupamento de recursos** - Os recursos adquiridos em grandes Clouds tem a facilidade e facilidade de ser escolhido a região mais próxima de onde terá mais volume de acesso, dando a liberdade de escolha de região.

- **Elasticidade rápida** - Recursos que quando necessários podem ser disponibilizados de forma autônoma e rápida, nesse caso a disponibilidade depende da demanda a ser solicitada em um determinado momento**;**
- **Serviço mensurado** (Há um controle à risca do que se gasta mensalmente, onde cada serviço e cada recurso tem um valor fixo que é mensurado e informado de forma clara, gerando transparência e economia para o consumidor)**.**

Modalidades de serviço

Existem algumas modalidades de serviços de computação em nuvem que são oferecidas no mercado, dentre as principais estão:

- **Software como Serviço (Software as a Service – SaaS)**: De acordo com Anthony Velte (2011) no SaaS, a aplicação é fornecida ao consumidor por meio de usuário e senha e é comummente usada em Web Browsers, ou Programas de acesso, o cliente nesse caso não tem acesso a infraestrutura e nem ao código fonte da aplicação**;**
- **Plataforma como Serviço (Platform as a Service – PaaS)**: É fornecida ao consumidor uma plataforma onde ele pode instalar a sua aplicação e ter acesso ao código fonte bem como pode realizar atualizações e alterações, o provedor de serviços oferece os aplicativos e uma plataforma segura onde pode ser acessada de qual quer localidade. (AZURE, 2021).
- **Infraestrutura como Serviço (Infrastructure as a Service – IaaS)**: A IaaS, é fornecida ao consumidor o provisionamento de Processamento, Armazenamento e Rede da infraestrutura, dando a liberdade de controle total, podendo instalar sistemas operacionais e gerenciar toda a infraestrutura em nuvem de sua aplicação, mas sem acesso a infraestrutura subjacente (AMAZON. 2021).

Modalidades de Instalação

De acordo com Cezar Taurion (2009) os tipos de Nuvem são:

- **Nuvem privada** - Infraestrutura fornecida de forma dedicada a uma única organização, com recursos de hardware e rede dedicados somente a um cliente, ou unidades de negócio, possui características próprias, segurança de firewall e restrição de acessos somente para pessoas autorizadas na organização;
- **Nuvem comunitária** - A infraestrutura é fornecida e acessada de maneira comunitária, ou seja, de organizações que trabalham com a mesma finalidade ou semelhante);

- **Nuvem pública** - Infraestrutura com acesso público, voltada para o público em geral, geralmente sem custo, como exemplo o Google Drive a nível gratuito;
- **Nuvem híbrida** - Infraestrutura hibrida nesse caso seria a junção das nuvens, privadas, comunitárias e públicas, que contêm diferentes organizações, mas podem ter comunicação entre elas.

A IMPORTÂNCIA DO BALANCEAMENTO DE CARGA EM COMPUTAÇÃO EM NUVEM

Um número alto de acessos pode acabar prejudicando servidores, fazendo com que o desempenho das máquinas diminua, causando a interrupção dos serviços. A saída para evitar esses problemas ou minimizar suas consequências é a utilização do balanceamento, dividindo o trabalho entre mais de uma máquina, no intuito de que ela não fique sobrecarregada facilmente.

Uma outra opção, muito utilizada atualmente é criar um sistema que virtualize o trabalho dos servidores físicos que executam aqueles serviços. Uma definição mais básica é a de equilibrar a carga entre vários servidores físicos, fazendo com que eles pareçam um grande servidor para o mundo externo. Há muitos motivos para fazer isso, mas os principais podem ser resumidos em escalabilidade, alta disponibilidade e previsibilidade.

Dois conceitos são chaves para o balanceamento de carga:

- **Escalabilidade**: De acordo com Ricardo Almeida (2008), escalabilidade e a capacidade de modificação e adaptação fácil e dinâmica ao aumento de carga, sem impacto sobre o desempenho. Se tratando de escalabilidade horizontal, Edmar Ferreira (2010), diz que a escalabilidade horizontal em um sistema distribuído permite que múltiplos servidores sejam adicionados em múltiplas plataformas sem perca do desempenho.
- **Alta Disponibilidade:** É a capacidade de um site manter-se disponível e acessível, mesmo em caso de falha de um ou mais sistemas. De acordo com Christian Engelmann (2005), caracteriza-se por um sistema projetado para ter redundância de componentes e que caso haja falhas, os serviços continuem em funcionamento e de forma transparente para o usuário.

TRABALHOS RELACIONADOS

Balanceamento de carga é um tema de trabalhos de pesquisa que têm como objetivo distribuir a carga de trabalho uniformemente, ou não, entre dois ou mais computadores, enlaces de rede, UCPs, discos rígidos ou outros recursos, a fim de

otimizar a utilização de recursos, maximizar o desempenho, minimizar o tempo de resposta e evitar sobrecarga.

As técnicas de balanceamento de carga em nuvens existentes, consideram vários parâmetros como desempenho, tempo de resposta, escalabilidade, utilização de recursos, tolerância a falhas, tempo de migração, sobrecargas, consumo de energia e emissão de carbono. (KANSAL CHANA, 2012)

Lucas Varela (2021) explora a orquestração da plataforma HPCC Systems *(High Performance Computing Cluster)* em ambientes cloud conteinerizados, com a ferramenta de orquestração *Kubernetes*. O objetivo do trabalho é avaliar as características, benefícios e desafios da implantação da plataforma HPCC Systems nesse paradigma através de diferentes provedores de cloud pública, especificamente *Amazon Web Service* (AWS) e *Microsoft Azure*.

Luciano Gonda (2017) trata sobre o armazenamento em nuvem e uma realidade e cada vez mais os usuários desejam a tranquilidade de ter seus dados mais importantes disponíveis a qualquer momento ou dispositivo. Essa realidade traz desafios para a TI e deve ser avaliada cuidadosamente.

Luis Carreño (2016) propõe a implementar escalabilidade horizontal em servidores web, com base na técnica de balanceamento de carga HTTP; usando um servidor de balanceamento nginx no modo de *proxy* reverso. Este tipo de escalabilidade suporta o tratamento de conexões TCP simultâneas, distribuição de solicitações HTTP com base em balanceamento dinâmico e permite a agregação de novas instâncias de servidor para suportar o crescimento contínuo.

Alfred Sanmiguel (2013) criou um projeto dividido em 2 blocos, são eles: A criação de um cluster de alta disponibilidade para que garanta em caso de falha de um dos servidores físicos a web permanecerá visível ao público. A criação de um cluster de balanceamento de carga que desempenha a função de redirecionar as solicitações para o nó do cluster que está mais ocioso naquele momento.

Luis Pires (2016) desenvolveu estratégias para garantir a disponibilidade oferecida pelos provedores. No presente trabalho, descreve-se uma solução que implementa alta disponibilidade em ambientes *Multi-Cloud*, mediante a distribuição de acesso por DNS e a utilização de proxy reverso. Realizou-se também uma análise financeira, levando-se em conta valores de mercado em serviços de *Cloud Computing*, o que mostrou que a solução proposta pode ser mesmo vantajosa com a relação à solução tradicional.

METODOLOGIA

Neste trabalho é apresentado uma proposta que busca reduzir a sobrecarga em um ambiente legado de aplicações WEB a partir da implementação de um

balanceamento de carga utilizando a AWS. Utilizando uma abordagem centralizada, almeja-se alcançar um balanceamento de carga levando em consideração a redução na utilização de CPU, consumo de memória RAM e tráfego de rede na infraestrutura. Também objetiva-se reduzir os custos de operação e manutenção da infraestrutura atual, com mão de obra para implementação e monitoramento da infraestrutura.

Cenário atual

No cenário atual existe uma infraestrutura legada com servidor alocado fisicamente na empresa, sem escalabilidade, sem tolerância a falhas e sendo compartilhado com outras aplicações. Esse servidor aloca as aplicações WEB, gerenciado por linha de comando e sem software de monitoramento. O servidor opera sem redundância em caso de uma indisponibilidade total do serviço. Destaca-se que o servidor está estruturado em uma plataforma composta por um processador Intel Xeon E5 2689 de 2 núcleos operando a 3.2GHz, 1GB de RAM DDR3, armazenamento em um HD de 500GB SATA 6Gbps a 7200rpm e executa um sistema operacional Ubuntu Linux 20.04.03. Para conexão à rede externa utiliza-se uma rede dedicada instalada em Fibra ótica de 5 Megas de Download e 5 megas de Upload.

Na tabela 1, apresenta a estimativa de custo mensal na infraestrutura atual, totalizando um custo mensal de 1.000 (mil reais) para manter a infraestrutura operando 24x7.

DESCRIÇÃO	RECURSOS GASTOS	TOTAL = R$ 1.000 MENSAL
Alocação do servidor (Local)	Energia Elétrica	R$ 150 reais mensais
Conexões	Link Dedicado	R$ 350 reais mensais
Consultoria	Suporte técnico a infraestrutura atual	R$ 500 reais mensais

Tabela 1. –Estimativa de custos da infraestrutura atual
(AUTOR, 2021).

No cenário atual o servidor WEB conta como servidor de aplicação o Apache2 juntamente com o interpretador de linguagem de programação PHP na versão 7.4. A aplicação hospedada é *OpenSource*, e é usada a aplicação WEB *Dolibarr* na versão 14.0.2 *(Dolibarr* é uma aplicação ERP, CRM voltada para gerenciamento de empresas e negócios, totalmente open source e com código fonte desenvolvido em PHP e Compatível os SGBDs Mysql MariaDB e PostgresSQL). (DOLIBARR, 2021).

Atualmente no ambiente de produção tem-se picos de acesso, durante o período da manhã, chegando a mil usuários simultâneos acessando o sistema. Esse cenário gera muitos incidentes e reclamações de usuários na utilização do sistema, inclusive contabilizando prejuízos por cancelamento de acessos a plataforma.

Na figura 1, é apresentado a arquitetura do cenário atual utilizado em produção para hospedar a aplicação WEB.

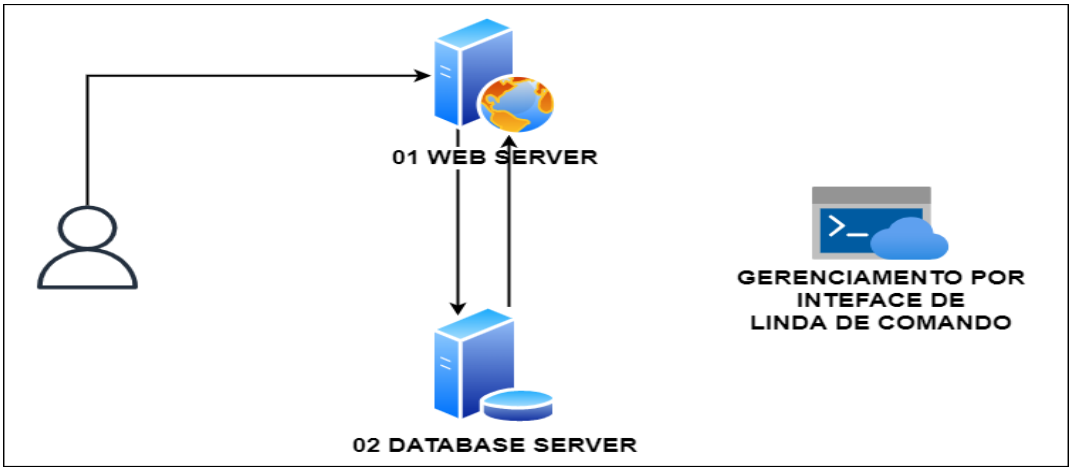

Figura 1. Infraestrutura atual
(Autor, 2021).

Com base nos relatos de lentidão e quedas de acessos relatadas pelos usuários, decidiu-se realizar um teste de sobrecarga e stress no servidor para registrar eventos de sobrecarga na aplicação.

No ambiente de testes proposto, foi utilizado o Apache JMeter realizando uma simulação com 100 usuários, cada um executando 10 interações simultâneas sendo um total de 1000 (mil) solicitações. Nos testes foi possível registrar o uso de CPU, consumo de memória RAM e tráfego de rede.

Na figura 2, verifica-se a utilização de CPU e Memória RAM no servidor. É possível observar que no momento do teste, a CPU chega a 100% de utilização e há o uso de mais de 50% de uso de memória RAM.

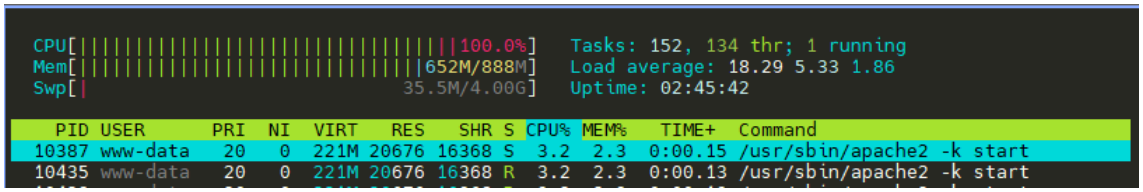

Figura 2. Uso de CPU e Memória RAM.
(Autor, 2021).

Na figura 3, foram coletadas as seguintes informações:

1. Tráfego na Interface onde são recebidas as conexões (ens33).
2. (RX Bytes/Second) - Saída de rede medida em 92,03 Kib por segundo durante o teste;
3. (TX Bytes/Second) - Entrada de rede medida em 540,92 Kib por segundo durante o teste;

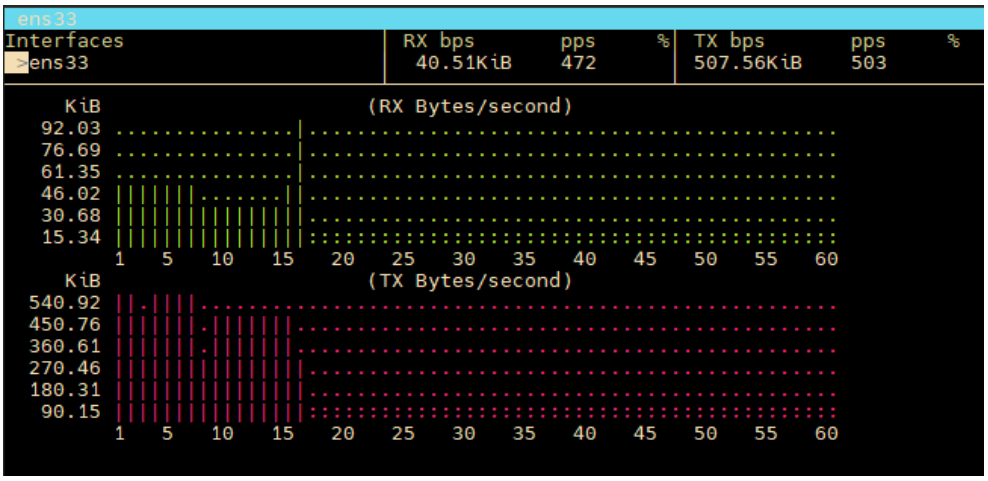

Figura 3. Infraestrutura atua
(Autor, 2021).

Desta forma, o servidor web (ponto de demanda) no recurso de CPU excede a sua capacidade. Esse teste que proporciona a quase realidade de acessos, recebe requisições a uma taxa maior do que pode executar, e por isso, ocorrem falhas no serviço web.

Cenário proposto

Abaixo é apresentado o cenário proposto com implementação de balanceamento de carga em nuvem.

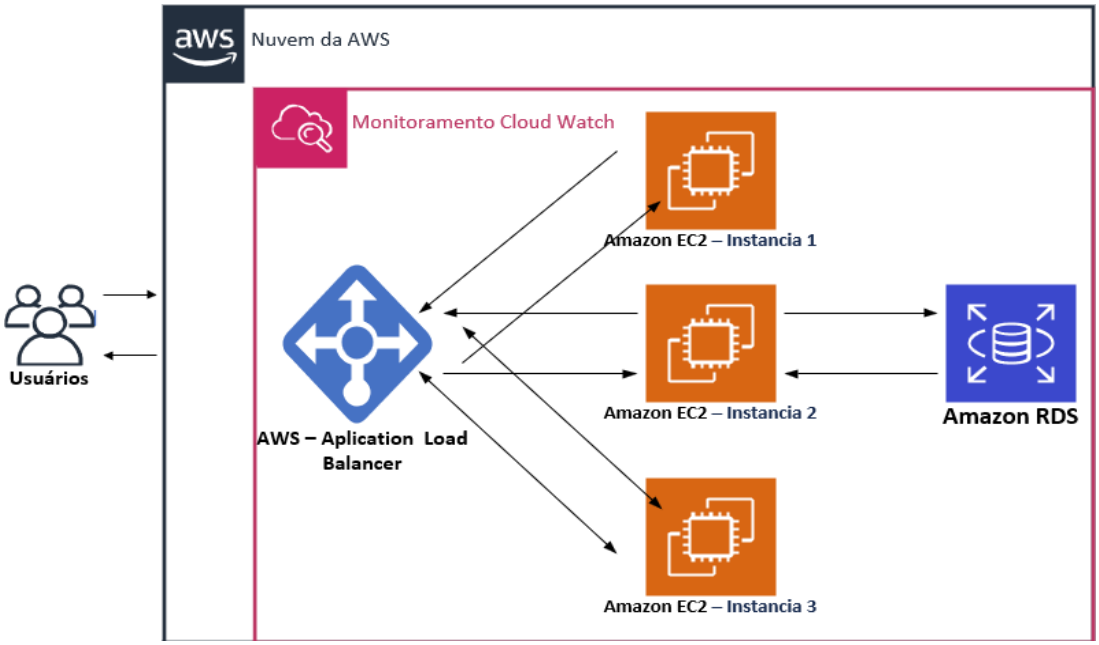

Figura 4. Arquitetura da infraestrutura proposta.
(AUTOR, 2021).

Com a implementação da infraestrutura proposta objetiva-se:

1. A escalabilidade e a alta disponibilidade do serviço.
2. A redução dos problemas de desempenho da aplicação atual, gerando um valor e uma boa visibilidade da aplicação no mercado.
3. A implementação dos melhores recursos de monitoramento e métricas de uso e desempenho da infraestrutura de redes da aplicação,
4. A utilização do *CloudWatch* da AWS para realizar o registro de métricas de desempenho do l*oad Balancing*.

Após a criação do cenário proposto, procurou-se observar e comparar o desempenho com a infraestrutura atual. Assim, pode-se observar-se que o *load balancing* cumpre a função de escalonamento de requisições de forma igualitária trazendo desempenho e alta disponibilidade, sem sobrecarga de uso de CPU, memória RAM e tráfego de rede.

Nas três instâncias implementadas foram instalados os servidores de aplicação Apache2 e o interpretador de linguagem de programação PHP na versão 7.4. A aplicação é a mesma utilizada no cenário atual em sua mesma versão apresentada.

Para o cenário proposto foi implementado o ALB - *Aplication Load Balancing* que faz o roteamento requisições HTTP/HTTPS. O ALB foi escolhido pela compatibilidade e devido ao fato de a aplicação trabalhar em um servidor Apache com requisições HTTP/HTTPS.

O ALB segue representado na figura 3, que representa a requisição HTTP do usuário e o ALB redireciona a aplicação.

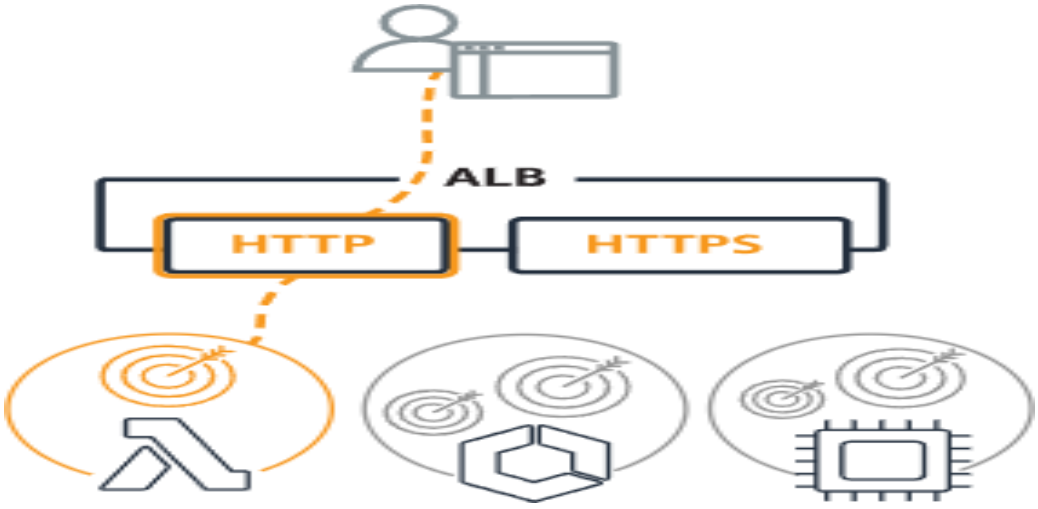

Figura 5. ALB – *Aplication Load Balancing.*
(AWS, 2021).

Baseado na AWS, o cenário foi implementado com os seguintes recursos, na seção de nível gratuito, sendo eles:

RECURSOS	TIPO	DESCRIÇÃO
UBUNTU 18.04	SISTEMA OPERACIONAL	Na amazona EC2, há a disponibilização do sistema operacional Ubuntu para ser instalado nas instâncias, na arquitetura foi instalada a versão 18.04 por se tratar de uma versão que já passou por diversas correções por tanto é mais estável.
EC2 – Elastic Computing Cloud	3 INSTANCIAS T2 MICRO – 1 Core – 1GB RAM – 10GB SSD	Considerado o serviço mais importante de toda AWS. Ele fornece ambientes virtualizados de acordo com a necessidade do projeto, podendo variar de servidores com capacidade mínima utilizados em desenvolvimento e testes, até máquinas com alto poder computacional para suportar aplicações em produção. Instância é o nome dado para cada ambiente virtualizado no EC2. Tecnologias e métodos variam de acordo com a necessidade de cada aplicação.
RDS - Amazon Relational Database Service	1 INSTANCIA RDS (Banco de dados MySQl) – 1 Core – 1GB RAM – 10GB SSD	está disponível em vários tipos de instância de banco de dados – com otimização para memória, performance ou E/S – e oferece seis mecanismos de bancos de dados comuns, incluindo Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle Database e SQL Server. Você pode usar o AWS Database Migration Service para migrar ou replicar facilmente bancos de dados existentes para o Amazon RDS.
ALB – Aplication Load Balancing	1 APLICATION LOAD BALANCER (HTTP/HTTPS)	Distribui automaticamente o tráfego de entrada de aplicações entre diversos destinos, como instâncias do Amazon EC2, contêineres, endereços IP, funções do Lambda e dispositivos virtuais. O serviço pode lidar com a carga variável de tráfego das aplicações em uma única zona de disponibilidade ou em diversas zonas de disponibilidade
Amazon Cloud Watch	Monitoramento de recursos	Além desses serviços é necessário citar o AWS - *CloudWatch*. O *CloudWatch* é um serviço de monitoramento que fornece dados e métricas em que é possível monitorar aplicações, realizar ações a partir de métricas de desempenho ou erros nas mesmas, como por exemplo envio de alertas. Ele coleta dados em forma de logs, métricas e eventos e exibe todos esses dados de forma unificada e simples. Também é possível criar regras para executar ações em determinado período, é possível configurar em quais horários do dia há a execução, os dias da semana e até do mês.

Tabela 2. – Definição de recursos utilizado na infraestrutura do projeto
FONTE: (AWS, 2021)

Baseado na AWS, foi utilizada a calculadora de estimativa de preço, que proporcionou o cálculo os recursos utilizados para cenário proposto (AWS 2021), sendo eles:

DESCRIÇÃO	RECURSOS	TOTAL = 80,89 USD MENSAL
EC2 – *Elastic Computing Cloud*	3 INSTANCIAS T2 MICRO – 1 Core – 1GB RAM – 10GB SSD	3 x 8,13 = 24,39 USD
RDS - *Amazon Relational Database Service*	1 INSTANCIA RDS (Banco de dados MySQl) – 1 Core – 1GB RAM – 10GB SSD	31,72 USD
ALB – *Aplication Load Balancing*	1 APLICATION LOAD BALANCER (HTTP/HTTPS)	30,45 USD
Amazon Cloud Watch	Monitoramento de recursos	0,00 USD

Tabela 3. – Estimativa de custos para o cenário proposto no projeto
FONTE: (AUTOR, 2021)

As definições de custos deram um total de mensal de 86,56 USD sendo 1.038,72 USD por ano, em conversão direta na data desenvolvimento desse trabalho. O gasto mensal ficaria em torno de 482,75 reais mensais, uma economia de mais de 50% em relação aos gastos com a infraestrutura atual.

Análise de Resultados

Durante o experimento foi possível coletar duas análises de resultados. A primeira analisando a utilização de hardware (CPU e memoria) e a segunda analisando a utilização do *load balancing* por meio do *Cloud Watch*, para colher as métricas de forma mais exata da utilização da aplicação após. Dessa forma, a comparação será realizada com mais exatidão nos dois cenários predispõem demonstrar os benefícios em todos os requisitos.

Em ambos os cenários foram feitos testes de requisição HTTP com o Apache JMeter. O Apache JMeter é uma ferramenta desenvolvida na linguagem Java e sob uma arquitetura open source, tem como objetivo facilitar a criação de testes de carga, stress, desempenho e longevidade. A ferramenta conta com vários componentes para a realização dos testes, dentre eles plano de teste, grupo de usuário, ouvinte, configuração do ambiente e testador (JMETER 2021).

Para o experimento do cenário proposto, foi realizada uma simulação com 60 usuários realizando 60 interações, totalizando 3600 solicitações durante o teste, o triplo da capacidade na infraestrutura atual.

Análise da utilização de Hardware

As figuras 7, 8 e 9 representam o consumo de CPU e memoria nas instâncias 1, 2 e 3. Observa-se que nos testes as três instancias apresentaram uso moderado de CPU ficando sempre abaixo de 70% e o uso de memória permaneceu baixo em comparação ao cenário atual.

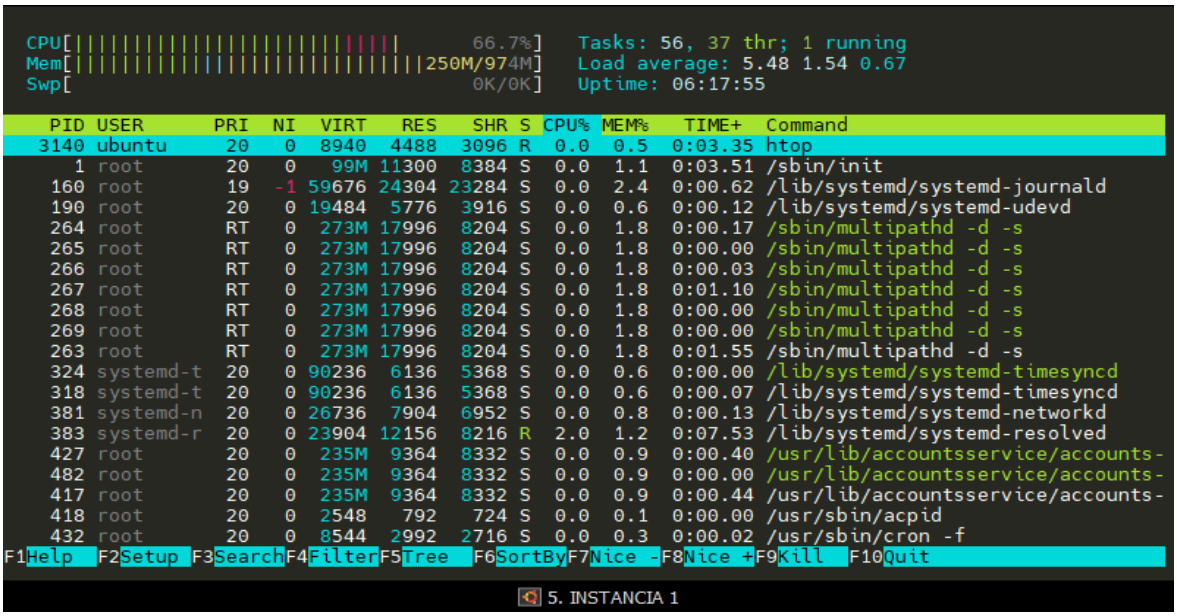

Figura 6. Uso de CPU e Memória RAM – Instância 1
(AUTOR, 2021)

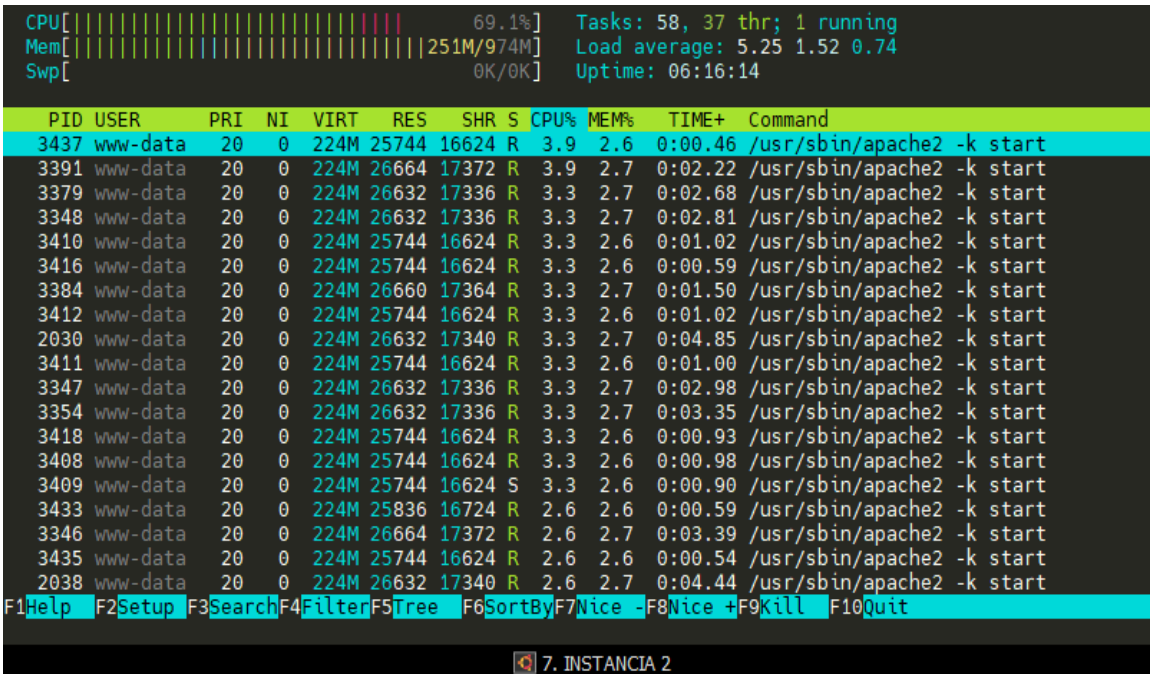

Figura 7. Uso de CPU e Memória RAM – Instância 2.
(AUTOR, 2021)

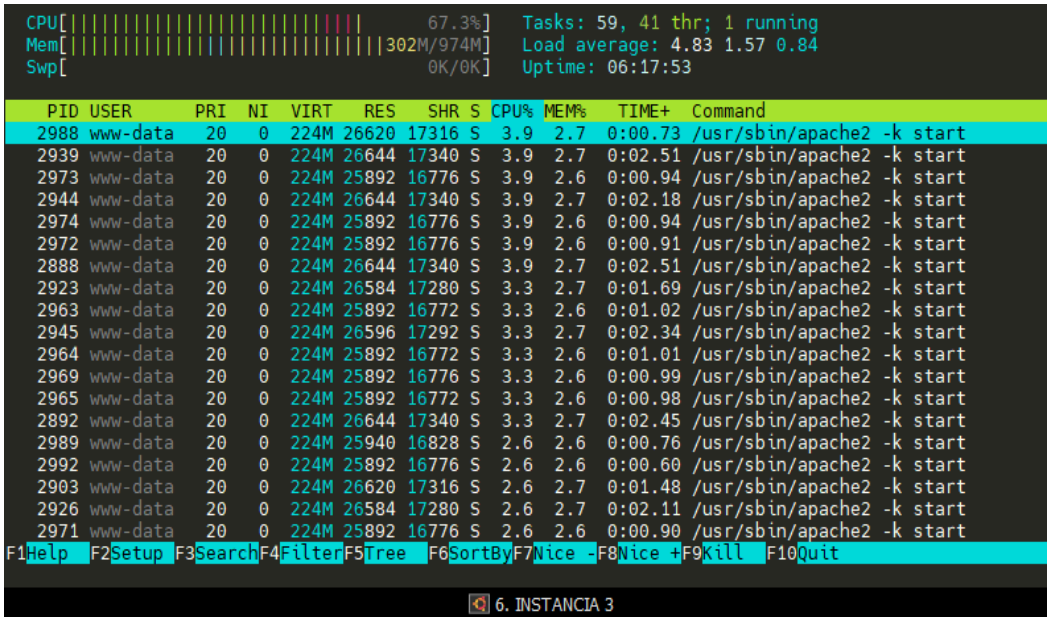

Figura 8. Uso de CPU e Memória RAM – Instância 3.
(AUTOR, 2021)

ANÁLISE DO MONITORAMENTO DO *LOAD BALANCING*

Em sequência foram coletadas as métricas de utilização do *Load Balancing.* Os resultados referentes ao uso do *Aplication Loda Balancig,* serão ilustrados por meio das imagens do painel da AWS. Como mencionado essa análise busca detalhes da utilização da infraestrutura proposta em comparação ao cenário atual. Verifica-se que a escalabilidade e alta disponibilidade atuam em alto desempenho no cenário proposto. Mesmo com número de acessos maior nota-se que tanto no balanceamento quanto nas entradas de redes, o desempenho fica bastante satisfatório para o cenário atual.

Na figura 9, é demonstrada a métrica de contagem total de solicitações 3,6k (três mil e seiscentas).

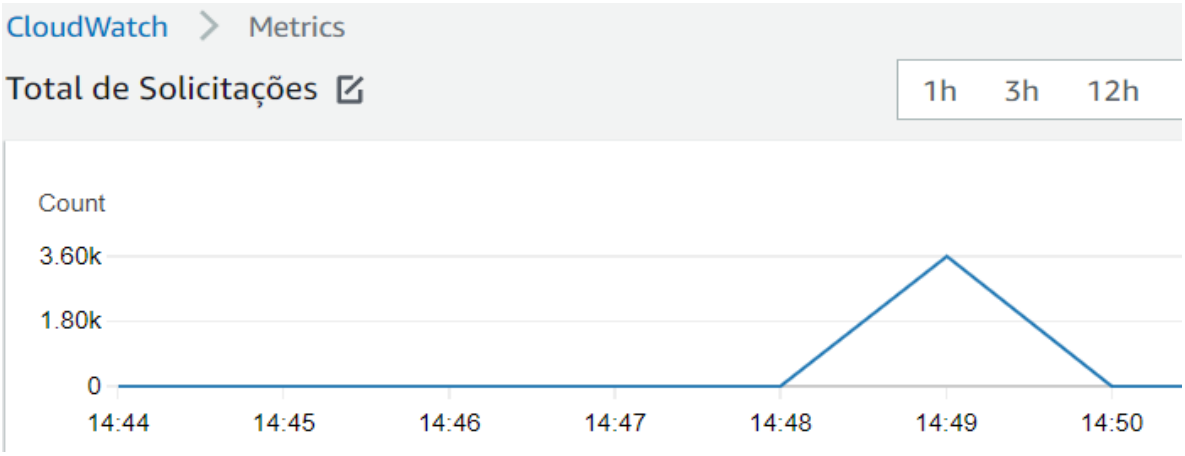

Figura 9. Total de Solicitações.
(AUTOR, 2021).

A figura 10, mostra a métrica de solicitações por instancias que compõem o *load balacing*, com total de 1,2k (mil e duzentas) solicitações por instância, totalizando 3,6k (Três mil e seiscentas) solicitações.

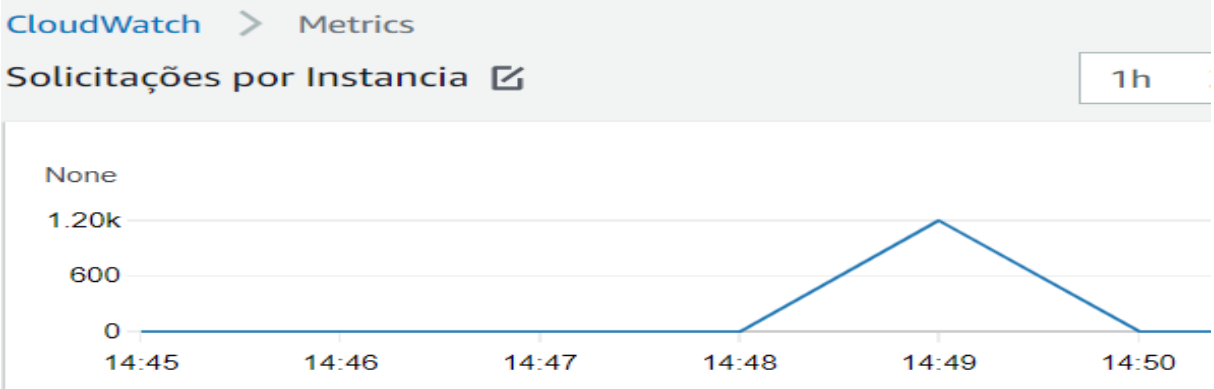

Figura 10. Total de Solicitações por instancia.
(AUTOR, 2021)

Na figura 11, ressalta a métrica da contagem de saída de rede totalizando 2,07 Mb por instancia, observando que as linhas no gráfico se mantem iguais na contagem durante o teste

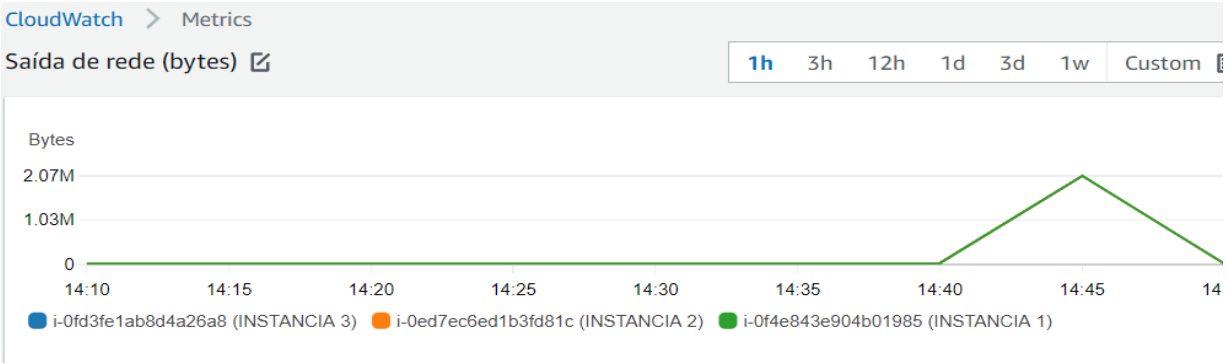

Figura 11. Saída de rede por instancia.
(AUTOR, 2021)

Por último, a figura abaixo simboliza a métrica da contagem de entrada de rede totalizando 1,30 Mb por instancia, observando que as linhas no gráfico se mantem iguais na contagem durante o teste

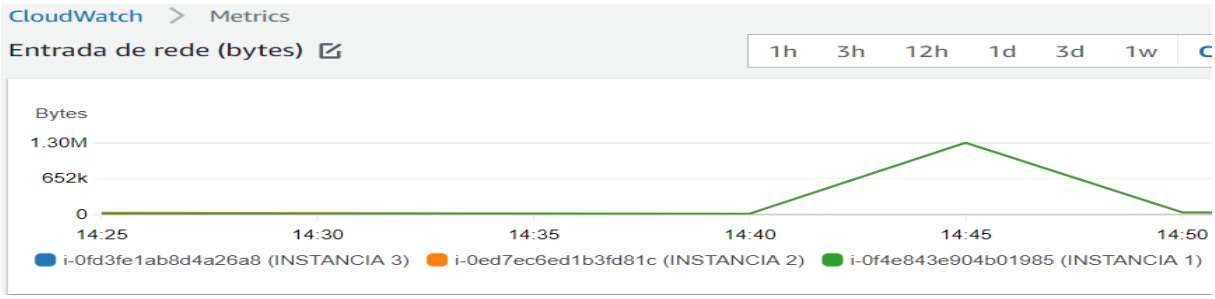

Figura 12. Entrada de rede por instancia.
(AUTOR, 2021).

CONCLUSÃO

Este trabalho apresenta uma arquitetura de balanceamento de carga web para uma nuvem privada AWS, que provê lata disponibilidade sob demanda baseada no estado das instâncias.

Foram apresentados os resultados da nuvem em função do estado de suas métricas: quantidade de requisições, escalonamento de conexões e utilização de CPU. O trabalho demonstrou alguns benefícios das plataformas de computação em nuvem, que são capazes de lidar com cargas repentinas, entregando recursos sob demanda para os usuários e mantendo maior utilização de recursos e infraestrutura, reduzindo assim, custos de gestão. Além disso, de maneira sucinta, foi abordado questões financeiras e a redução de gastos na implementação do projeto de serviços WEB.

A maior contribuição do trabalho foi apresentar uma proposta de balanceamento de carga dinâmica que pode evitar problemas de sobrecarga de recursos, e possíveis problemas de sobrecarga de sistemas e provisionamento para cargas de pico, o que poderia gerar travamentos e indisponibilidade de serviços no ambiente de infraestrutura.

O *Aplication Load Balacing* (ALB), permite configurar um ambiente escalável de alta disponibilidade e processamento com consumo baixo de CPU e tráfego de rede moderado. Nos experimentos, a escalabilidade servida conseguiu provisionar os recursos necessários ao tratamento da carga de trabalho evitando o sobrecarga de recursos.

Desta forma, é possível confirmar que o balanceamento de carga é uma solução favorável para momentos de pico de tráfego de acesso. O balanceador aplicado nesse trabalho cuidou da distribuição de tarefas entre os processadores disponíveis para que as instâncias não entrassem em situações inconvenientes em sua experiência de transações. Com balanceamento de carga adequado, os serviços evitam percalços financeiros e dissabores na relação com os usuários como nas situações hipotéticas descritas acima.

Como proposta futura, é favorável a aplicação de técnicas melhoradas de *auto-scaling* e testes de desempenho com outras soluções de balanceamento de carga.

REFERÊNCIAS

ALMEIDA, Ricardo. **Escalabilidade! = Performance.** 2008. Disponível em: http:// manifestonaweb.wordpress.com/2008/06/18/ escalabilidade-performance. Acesso em 10 de fevereiro de 2021.

CARREÑO ORTIZ, Luis Giovanny**. Escalabilidad y optimización en aplicaciones web utilizando técnicas de balanceo de carga** 2016. Disponível em http://repositorio.utb.edu.co/handle/20.500.12585/1590. Acesso em 28 de janeiro de 2021.

Definição de computação em nuvem segundo o NIST. 2011 Disponivel em: Definição de computação em nuvem segundo o NIST | Plataforma Nuvem (wordpress.com)

CARREÑO ORTIZ, Luis Giovanny**. Escalabilidad y optimización en aplicaciones web utilizando técnicas de balanceo de carga** 2016. Disponível em http://repositorio.utb.edu.co/handle/20.500.12585/1590. Acesso em 28 de janeiro de 2021.

CISCO, **O que é computação em nuvem?** Disponivel em https://www.cisco.com/c/pt_br/solutions/cloud/what-is-cloud-computing.html

ENGELMANN, Christian et al. **Concepts for high availability in scientific high-end computing.** In: Proceedings of High Availability and Performance Workshop (HAPCW). 2005.

FERREIRA, Edmar. **Escolhendo entre escalabilidade horizontal e escalabilidade vertical.** 2010. Disponível em: < http://escalabilidade.com/2010/09/21/ escolhendo-entre-escalabilidade-horizontal-eescalabilidade-vertical/>.

SOFTWARES **para Criação de diagramas de projetos**. Disponível em: https://cloud.google.com/what-is-cloud-computing?hl=pt

GONDA, Luciano; JUNIOR, B. A. **Nuvem de Dados Corporativa-Un Caso de Sucesso**. 2017. Disponível em: http://www. xiwticifes. ufba. br/modulos/submissao/Upload-353/86118.pdf. Acesso em 15 de maio de 2021.

GUTIÉRREZ SANMIGUEL, Alfred et al. **Clúster de alta disponibilidad y balanceo de carga sobre un servidor web**. 2013.

INFOGRÁFICO **Tipos de Hospedagem**. Disponível em: https://brasilcloud.com.br

JMETER, Apache JMeter. 2021, Disponível em: http://jmeter.apache.or

KANSAL, Nidhi Jain; CHANA, Inderveer. **Cloud load balancing techniques: A step towards green computing**. IJCSI International Journal of Computer Science Issues, v. 9, n. 1, p. 238-246, 2012.

MAGALHÃES, Leandro. **Hospedagem de Sites: conheça mais sobre.** 2020. Disponível em https://blog.brasilcloud.com.br/saiba-mais-sobre-hospedagem-de-sites/ Acesso em: 12 de março de 2021

MELL, Peter et al. **A definição NIST de computação em nuvem**. 2011

MUKHERJEE, Sourav. **Benefits of AWS in modern Cloud**. Available at SSRN 3415956, 2019.

MODELOS **de cloud computing**. 2021 Disponível em: https://aws.amazon.com/pt/types-of-cloud-computing. Acesso em: 12 de março de 2021

NEVES, Ricardo; BERNARDINO, Jorge. Performance and scalability of voldemort nosql. In: **2015 10th Iberian Conference on Information Systems and Technologies (CISTI)**. IEEE, 2015. p. 1-6.

O QUE **é computação em nuvem**. Microsoft. 2021. Disponível em: https://azure.microsoft.com/pt-br/overview/what-is-cloud-computing/#benefits.

O QUE **é PaaS? Plataforma como Serviço**. 2021 Disponivel https://azure.microsoft.com/pt-br/overview/what-is-paas/

PIRES, Luis Paulo Gonçalves et al. **Alta disponibilidade: uma abordagem com DNS e Proxy Reverso em Multi-Cloud**. 2016.

TAURION, Cezar. **Cloud computing-computação em nuvem**. Brasport, 2009.

VARELLA, Lucas et al. **Orquestração de Aplicações de Computação de Alta Performance em Ambientes Cloud Conteinerizados**. In: Anais da XXI Escola Regional de Alto Desempenho da Região Sul. SBC, 2021.

VELTE, A; VELTE, T. J.; ELSENPETER, R. **Computação em Nuvem: Uma abordagem prática**.1 ed. Rio de Janeiro: Alta Books, 2011