EXPLORING SIMPLE LOAD BALANCING TECHNIQUES FOR IMPROVED CLOUD PERFORMANCE

Manisha Mittal, Rashmi Dutta, Shilpa Seth, Pooja
E-Mail Id: manishamittal22@gmail.com, rdutta.csc@mlncollegeynr.ac.in, sseth.csc@mlncollegeynr.ac.in, pooja.csc@mlncollegeynr.ac.in
Department of Computer Science & Application, Mukand Lal National College, Yamuna Nagar, Haryana, India

Abstract- The paradigm of cloud computing is one in which users receive resources like computation power warehousing and applications as a service over the internet. It provides common assets, information, software and other measures based on the requirement of the client at a particular time. Service needs to be improved as cloud computing expands rapidly and required computing attracts more users. Good load balancing techniques are required for efficient resource management. Load adjusting in distributed computing is a basic module for conveying approaching organization traffic or computational responsibilities across different servers or assets to guarantee ideal asset usage, expand throughput, limit reaction time and keep up with framework soundness. It is essential for enhancing cloud-based applications and services performance, scalability and dependability.
Keywords: Cloud Computing, Computation power, Load Balancing, Distributed Computing, Scalability, Warehousing, Network Traffic.

1. INTRODUCTION

A cloud represents a unique IT environment designed specifically to deliver extensible and metered services remotely. It constitutes a data processing approach where resources are shared rather than confined to individual or local appliances. Applications can be centrally managed on cloud servers, distinguishing it from standard personal server systems. In cloud computing, the term "cloud" metaphorically refers to the internet, defining it as an internet-based computing environment wherein organizations access various services, servers, applications, and storage through internet-connected tools.When comparing cloud computing with usual ownership and usage models, the need to purchase and sustain infrastructure diminishes. Users can utilize IT resources in real-time, aligning with their immediate needs. Consequently, cloud computing capabilities swift and convenient access to a shared pool of computing resources—including applications, networks, storage, and services—on a pay-as-you-go basis.

1.1 Cloud Computing Architecture

Cloud computing is experiencing significant development in today's dynamic environment with discussions revolving around the facts of cloud technology the services it offers and its various deployment models. [1]

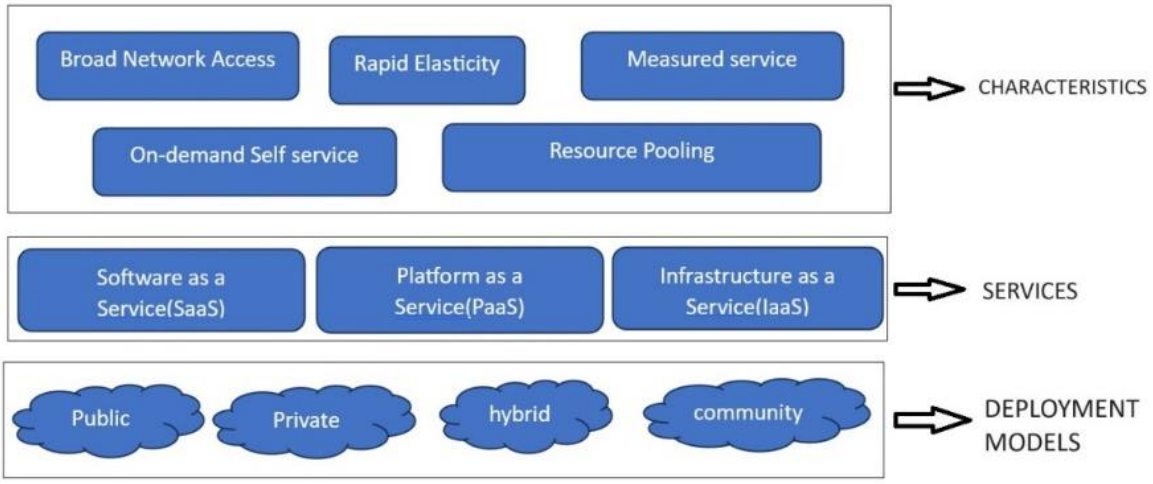

Fig. 1.1 Illustrating the basic characteristics, three service model and deployment models.[1]

2. CLOUD VIRTUALIZATION

The concept of virtualization is highly valuable when it comes to cloud computing. Although virtualization offers all the amenities of the actual world, it is akin to "something that is not real." This is a computer software that emulates a real machine and allows for the execution of many programs. Since users can access many cloud services, virtualization is a component of cloud computing. Remote data centers with full or partial virtualization

All Right Reserved @ IJTRS

provide the end user with all these varied services. The two types of virtualizations that are now accessible are explained here.

2.1 Complete Virtualization

When using full virtualization, one system's installation is completed entirely on another machine. This makes it feasible to emulate hardware that is placed on separate systems and to share a computer system among several users. All of the software that is present on the actual server will also be available in the virtual system.

2.2 Virtualization Para

Para virtualization is introduced to reduce the limitation of device emulation. To achieve better performance, a guest operating system or device drivers is modified to support VMM interfaces and to speed up I/O operations. The performance of paravirtualization is better than the pure emulation approach, but it is significantly slower as compared to the direct access to the device.[2]

3. LOAD BALANCING

Load balancing involves effectively distributing incoming network traffic among a collection of backend servers, which are often referred to as a server farm or server pool.

A load balancer acts as a central coordinator, directing incoming client requests across all available servers to optimize speed and resource usage while preventing server overload. It also ensures continuous service availability by redirecting traffic away from offline servers to those that are operational. Moreover, when new servers are introduced, the load balancer automatically incorporates them into the server pool, maintaining efficient distribution of requests. Essentially, a load balancer efficiently distributes network load, maintains high availability, and accommodates dynamic changes in server capacity.

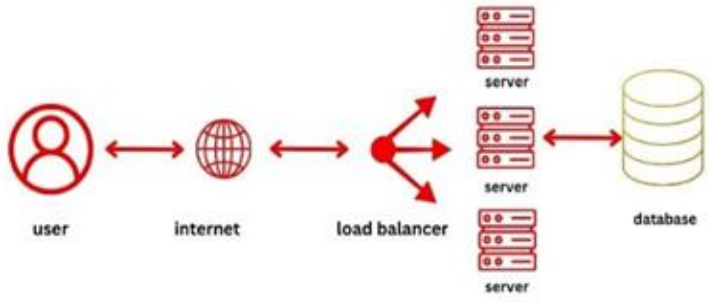

Fig. 3.1 Diagram shows how load balancing works. Load balancing is accomplished through a hardware device or software

3.1 Software Load Balancer

Software load balancers, which operate as software components on grade servers, offer versatility across different setups. They function by having the application select the first server from a designated list to retrieve data. In case of repeated failures beyond a set retry threshold, rendering the server inaccessible, it is removed from the rotation, and an alternative server from the list is chosen to fulfill the demand. This method represents a cost-effective approach to implementing load balancing.

3.2 Hardware Load Balancer

Hardware load balancers, also referred to as Layer 4-7 Routers, utilize physical appliances to distribute traffic among a cluster of network servers. They possess the capability to manage various types of network traffic, including HTTP, HTTPS, TCP, and UDP.

Unlike software load balancers, which operate on general-purpose servers, hardware load balancers are purpose-built devices dedicated solely to traffic management. They can efficiently handle high volumes of traffic but come with a substantial cost and limited flexibility.

If any server fails to deliver the expected response, traffic is immediately diverted away from it by the load balancer. Although hardware load balancers serve as the initial entry point for user requests due to their robust capabilities, many organizations opt for internal software load balancers for subsequent traffic management within their infrastructure due to the high cost and complexity associated with hardware load balancers.

3.3 Virtual load balancer

A virtual load balancer is a key module for distributing coming network traffic on all parts of various servers or resources in a virtualized environment, like data centers employing virtualization platforms such as VMware, Hyper-V, or KVM. Its primary function is to optimize resource utilization, amplify response times, and prevent server overloads by evenly spreading the loads.

All Right Reserved @ IJTRS

3.4 Load Balancing Algorithms

The collection of guidelines that a load balancer uses to choose the optimal server for each type of client request is known as a load balancing algorithm. There are two major types of load balancing algorithms.[3]

3.4.1 Balanced Static Load

Static load balancing methods don't depend on the state of the server at any one time and adhere to set rules. A static cloud load balancing algorithm is a fixed way of distributing incoming requests across multiple servers or resources. It does not consider real-time conditions but relies on predefined rules. For example, Round Robin and IP Hash. Although static algorithms are simple and predictable, they may not optimize resource usage or adapt to changing workloads. Here are some static load balancing examples.

3.4.2 Round-robin Technique

IP addresses on servers indicate to clients where to submit requests. The IP address is a lengthy, hard-to-remember number. A Domain Name System links website names to servers to make things simple. Your browser sends a request to our name server, which then delivers our IP address to your browser, when you type aws.amazon.com. Rather than using specialized hardware or software for load balancing, the round-robin method uses an authoritative name server. The IP addresses of each server in the server farm are returned by the name server either in a round-robin or turn-by-turn manner.

3.4.3 IP hashing Technique

The load balancer uses the client IP address to carry out a mathematical operation known as hashing when using the IP hash method. The IP address of the client is transformed into a number, which is subsequently assigned to other servers.

3.4.4 Balanced Dynamic Load

Before distributing traffic, dynamic load balancing algorithms check the condition of the servers. A few instances of dynamic load balancing algorithms are as follows.

3.4.5 Least-Connection Technique

An open channel of communication between a client and a server is called a connection. The client and server authenticate and establish an active connection when the client sends the first request to the server. The load balancer provides traffic to the servers with the fewest active connections when using the least connection approach. According to this approach, every connection needs the same amount of processing power from every server.

3.4.6 Weighted least-connection Technique

Certain servers are assumed to be able to manage more active connections than others via weighted least connection methods. As a result, you can give each server a different weight or capacity. The load balancer then routes new client requests to the server with the fewest connections based on capacity.

3.5 Limitation of Existing Techniques

Load balancing in cloud computing comes with several limitations, which includes :

3.5.1 Dynamic Workload

Workloads in the cloud can be highly dynamic, with traffic fluctuations throughout the day. Load balancers must adapt quickly to changing conditions to efficiently distribute traffic.

3.5.2 Scalability

Cloud environments often span multiple regions and availability zones, requiring load balancers to scale horizontally to handle increasing traffic loads while maintaining low latency.

3.5.3 Heterogeneous Environments

Cloud environments may consist of diverse resources, including virtual machines, containers, and server less functions. Load balancers must support various types of infrastructure while ensuring consistent performance and reliability.

3.5.4 Health Monitoring

Load balancers must continuously monitor the health of backend servers or instances to detect failures or performance degradation. Effective health checks are crucial for maintaining high availability and reliability.

3.5.5 Security

Load balancers are critical components for securing applications in the cloud. They must support encryption protocols, handle SSL/TLS termination securely, and integrate with other security mechanisms such as web application firewalls (WAFs) to protect against cyber threats.

All Right Reserved @ IJTRS

3.5.6 Cost Management

Load balancing services may incur costs based on factors like the volume of traffic, number of instances, or features utilized. Optimizing costs while ensuring performance and reliability requires careful planning and resource allocation.

3.5.7 Complexity

Managing load balancers in large-scale cloud environments can be complex, especially when dealing with multiple services, configurations, and integration points. Automation and orchestration tools are essential for streamlining operations and reducing manual overhead.

3.5.8 Vendor Lock-In

Cloud providers offer proprietary load balancing solutions, which may result in vendor lock-in. Organizations must weigh the benefits of using native cloud load balancers against the risks of limited interoperability and dependency on a single provider. Addressing these challenges requires a combination of advanced technologies, best practices, and continuous optimization efforts to ensure that load balancing in the cloud remains effective, scalable, and resilient.

4. PROPOSED SYSTEM

According to author the current system is not able to resolve the challenges like scalability, network management. The proposed system tries to resolve the cost and the scalability of the load balancing technique of cloud with Content Delivery Network method.

4.1 Content Delivery Network (CDN)

It's a system of distributed servers that deliver web content to users based on their geographic location and the origin of the webpage. CDNs are designed to improve website performance by reducing latency and increasing load times. They work by caching copies of website content on servers located in various data centers around the world. When a user requests a webpage, the CDN routes the request to the nearest server, which then delivers the content to the user. This helps to minimize the distance the data needs to travel, resulting in faster loading times and a better user experience. CDNs are commonly used for delivering web pages, images, videos, scripts, and other static or dynamic content.

4.1.1 Geographical Distribution

CDNs consist of multiple servers distributed across different geographical locations. When a user requests content, such as a webpage or a video, the CDN routes that request to the server that's geographically closest to the user. This minimizes latency and improves the overall user experience.

4.1.2 Caching Content

CDNs cache static content, like images, CSS files, JavaScript, and even entire web pages, across their network of servers. When a user requests this content, the CDN delivers it from the server closest to the user, reducing the load on the origin server. This caching mechanism helps improve website performance and reduces the load on the origin server.

4.1.3 Load Balancing Algorithms

Within the CDN network, load balancing algorithms determine which server should handle each incoming request. These algorithms consider factors such as server load, network proximity to the user, and current network conditions. By distributing requests evenly across servers, load balancing ensures optimal performance and prevents any single server from becoming overwhelmed.

4.1.4 Failover and Redundancy

CDNs often include failover mechanisms to ensure high availability and reliability. If one server becomes unavailable due to maintenance, hardware failure, or other issues, the CDN can automatically reroute traffic to other healthy servers. This redundancy minimizes downtime and ensures continuous service availability.

4.1.5 Dynamic Content Delivery

While CDNs are typically used for caching static content, they can also handle dynamic content delivery. In this scenario, requests for dynamic content are routed to the origin server, which generates the content dynamically based on user requests. CDNs can still assist in load balancing by directing these requests to the most suitable origin server based on factors such as server load and proximity. Overall, CDNs enhance load balancing by optimizing content delivery, reducing latency, improving scalability, and ensuring high availability for websites and web applications.

CONCLUSION AND FUTURE SCOPE

The paper provides an overview of different load balancing methods tailored for cloud computing. Load balancing aims to meet customer demands by dynamically allocating workload among nodes, optimizing resource usage by

All Right Reserved @ IJTRS

redistributing total workload to individual nodes. This strategy ensures efficient and equitable distribution of resources, leading to enhanced system performance. Load balancing in cloud computing is poised for remarkable advancements in the future. Expect to see more intelligent load balancers leveraging AI and machine learning to dynamically manage workloads. These load balancers will seamlessly distribute workloads across multi-cloud and hybrid environments, including edge computing locations, while being more application-aware to optimize performance. Additionally, the paper delves into cloud virtualization and outlines the qualitative criteria needed for effective load balancing.

REFERENCES

[1] Desai, Tushar, and Jignesh Prajapati. "A survey of various load balancing techniques and challenges in cloud computing." International Journal of Scientific & Technology Research 2.11 (2013): 158-161.

[2] Phaphoom, N., Wang, X., & Abrahamsson, P. (2013, February 7). Foundations and Technological Landscape of Cloud Computing. ISRN Software Engineering. https://doi.org/10.1155/2013/782174

[3] Sagar, S., Ahmed, M., Husain, M.Y. (2022). Fuzzy Randomized Load Balancing for Cloud Computing. In: Barolli, L. (eds) Advances on P2P, Parallel, Grid, Cloud and Internet Computing. 3PGCIC 2021. Lecture Notes in Networks and Systems, vol 343. Springer, Cham. https://doi.org/10.1007/978-3-030-89899-1_3

[4] S. Aslam and M. A. Shah, "Load balancing algorithms in cloud computing: A survey of modern techniques," 2015 National Software Engineering Conference (NSEC), Rawalpindi, Pakistan, 2015, pp. 30-35, doi: 10.1109/NSEC.2015.7396341.

[5] J. M. Shah, K. Kotecha, S. Pandya, D. B. Choksi and N. Joshi, "Load balancing in cloud computing: Methodological survey on different types of algorithms," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, India, 2017, pp. 100-107, doi: 10.1109/ICOEI.2017.8300865.

[6] Dalia Abdulkareem Shafiq, N.Z. Jhanjhi, Azween Abdullah,"Load balancing techniques in cloud computing environment: A review",Journal of King Saud University - Computer and Information Sciences,Volume 34, Issue 7,2022,Pages 3910-3933, ISSN 1319-1578.

[7] A. Jain and R. Kumar, "A multistage load balancing technique for cloud environment," 2016 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, 2016, pp. 1-7, doi: 10.1109/ICICES.2016.7518921.

[8] V. R. Kanakala, V. K. Reddy and K. Karthik, "Performance analysis of load balancing techniques in cloud computing environment," 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2015, pp. 1-6, doi: 10.1109/ICECCT.2015.7226052.

[9] Sujit Kumar et al 2021. Strategies to Enhance Solar Energy Utility in Agricultural Area of Rajasthan State, India. J. Phys.: Conf. Ser. 1854 012013. DOI 10.1088/1742-6596/1854/1/012013

[10] Vyas, Megha & Yadav, Vinod & Vyas, Shripati & Joshi, R. (2021). Voltage Sag Mitigation Using Distribution Static Compensator. 10.1007/978-981-15-8586-9_24.

[11] R. Jangid; et al., "Development of Advance Energy Management Strategy for Standalone Hybrid Wind & PV System Considering Rural Application", IEEE 2nd International Conference on Smart Systems and Inventive Technology, Organized by Francis Xavier Engineering College during November 27-29, 2019 at Tirunelveli, India.

[12] Tirole, R., Joshi, R.R., Yadav, V.K., Maherchandani, J.K. and Vyas, S. (2022). Intelligent Control Technique for Reduction of Converter Generated EMI in DG Environment. In Intelligent Renewable Energy Systems (eds N. Priyadarshi, A.K. Bhoi, S. Padmanaban, S. Balamurugan and J.B. Holm-Nielsen). https://doi.org/10.1002/9781119786306.ch4.

[13] K. Garala, N. Goswami and P. D. Maheta, "A performance analysis of load Balancing algorithms in Cloud environment," 2015 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2015, pp. 1-6, doi: 10.1109/ICCCI.2015.7218063.

All Right Reserved @ IJTRS