Grenze International Journal of Engineering and Technology, June Issue

Technical Comparison between Load Balancing Algorithm and Scheduling Algorithm over Cloud

Ajay Kumar Sahu[1], Shivani Dubey[2], Vikas Singhal[3], Nitish Kumar Kushwaha[4], Akash Kumar[5], Kajal Jaiswal[6], Abhishek Kumar[7] and Abhinav Kumar Rai[8]
[1-8]Greater Noida Institute of Technology/ Information Technology, Greater Noida, India
Email:ajay4989@gmail.com,{dubey.shivani,vikassinghal75,kushwahaniti850,ak9744741,kj741806,abhishekgandhar141, abhinavkumarroy002}@gmail.com

***Abstract*— The meteoric rise of cloud computing stands as a hallmark in computational technology, bolstered by the burgeoning strength of cloud-based virtual machines empowered by dynamic resource provisioning. This pioneering featureallows on-the-fly allocation of resources from a dynamic pool, a watershed moment vital for fortifying the resilienceof mission-critical web applications within cloudcomputing infrastructure. These mission-critical applications, typified by the indomitable titans such as Facebook, Google, YouTube, and Amazon, operate under the unyielding mandate of zero downtime despite the capricious influx of incoming traffic The bedrock of success for these mission-critical web applications liessquarely on the shoulders of efficient load balancing techniques and algorithms. These mechanisms, orchestrating the judicious distribution of HTTP/HTTPS traffic among diverse nodes within server clusters, emergeas the linchpin for optimized resource utilization. Yet, the quest for implementing an optimal load balancer that impeccably shares the load among server nodes in an optimized fashion remains a formidable frontier. In this paper, a concise overview of cloud computing technology is presented, unraveling both the affirmative and negative facets of prevalent load.**

***Index Terms*— Virtual Machine (VM) Load balancing, load balancing techniques, cloud computing, fault tolerance design patterns (FTD)**

I. INTRODUCTION

In recent years, the technological landscape has witnessed a paradigm shift with the advent of cloud computing a transformative computing model that has redefined thedynamics of resource provisioning and accessibility. As elucidated by the US National Institute of Standards and Technology (NIST), cloud computing stands as a revolutionary framework, characterized by its ubiquitous presence, unparalleled convenience, and on-demand accessibility to a myriad of virtual machine resources. At the core of this computing model lies the premise of instantaneous Access configurable shared pool and diverse computing resources[1]. Ranging from the fundamental building blocks of computation such as RAMand CPU to the expansive realms of storage and webservice-related applications, cloud computing encapsulates. NIST's definition underscores the essence of this technological marvel, portraying cloud computing as an omnipresent and user-centric ecosystem that transcends theconstraints of traditional computing architectures. The ubiquity and convenience it offers coupled with the dynamic allocation of resources from a communal pool, lay the groundwork for a transformative approach to computing—one that

Grenze ID: 01.GIJET.10.2.549

empowers users with agility, scalability, and unparalleled flexibility. This paper delves deeper into the multifaceted realm of cloud computing, exploring its underpinnings, the mechanisms driving its functionalities, and the transformative impact it has exerted on the technological landscape, paving the way for an era defined by accessibility, efficiency, and innovation.

II. Literature Review

A. Essential characteristic of Cloud Computing

On demand self service

The concept of "on-demand self-service" stands as a cornerstone within the framework of cloud computing, reflecting the intrinsic flexibility and user-centric nature ofthis paradigm. In the context of this research, "on-demand self-service" emerges as a pivotal component that embodies the agility and autonomy afforded to cloud service consumers, especially in the provisioning and managementof virtual machine resources. This attribute facilitates the dynamic adjustment of critical computing elements—suchas RAM, CPU, and storage based on fluctuating demands stemming from varying traffic loads and the nature of the services offered The paper acknowledges this fundamental tenet as a core element driving the efficiency and responsiveness of cloud computing infrastructures[2].

Broad network access

The deployment of web applications within a cloud infrastructure heralds a new era of accessibility, transcending the limitations imposed by device heterogeneity. Cloud-hosted web applications exemplify a paradigm where global accessibility becomes an inherent trait, catering seamlessly to an expansive array of client devices, whether thin or thick, spanning mobiles, tablets, laptops, and desktop computers[3].

Measured Service

The tenet of "Measured Service" stands as a fundamental principle in cloud computing, imbuing the utilization of virtual machine (VM) resources with transparency and control. Within this paradigm, VM resources are akin to metered commodities, wherein cloud consumers wield the capability to meticulously allocate and utilize only the requisite resources from the available pool. This metering capability, operating at an appropriate level of abstraction commensurate with the service type (be it storage, processing, bandwidth, or active user accounts), empowers virtual machines to autonomously govern and optimize resource utilization. This autonomous control mechanism ensures that resources are allocated judiciously, aligning precisely with the service requirements, thus optimizing efficiency.

B) Common Characteristics of Cloud Computing

The facet of "Massive Scale" within cloud computing is intrinsically tied to the dynamic scalability of resources, both horizontally and vertically, to cater to the ever- changing landscape of HTTP/HTTPS load traffic and the intricacies of varied web application types. This scalability serves as a cornerstone for ensuring enhanced service delivery and uninterrupted 24/7 availability[4].

Virtualization: Virtualization within the realm of cloud computing encapsulates the process of generating virtualized instances or representations of servers, operating systems, hardware, and software components.

Service Orientation: Cloud platforms, through their service-oriented paradigms, empower users with the agility to request, utilize, and release computing services based on real-time demands or pre-agreed performance benchmarks.

C) Cloud Services

Software as a service (SaaS)

Absolutely, Software as a Service (SaaS) encapsulates applications designed, developed, and hosted by application providers within a cloud environment. These applications are accessible from various devices, featuring coded functionalities, data definitions, and remote management by providers. Examples of SaaS include Google Spreadsheet, YouTube, and Facebook[5].

Infrastructure as a service (IaaS)

IaaS serves as the bedrock of cloud computing, providing users with virtualized resources that include but are not limited to RAM, storage, networking capabilities, virtual machines, hypervisors, and raw hardware elements. This

layer empowers users with the flexibility to dynamically provision, configure, and manage these resources remotely, catering to their specific computational needs without the burden of physical infrastructure maintenance.

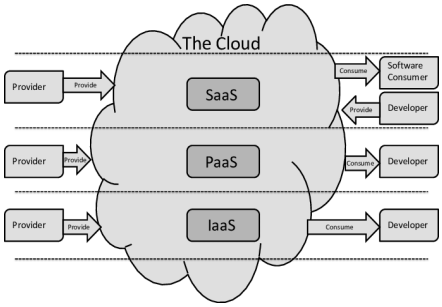

Figure 1: Cloud computing Architecture

F) Load Balancing Metrics

Response Time: Response time measures the time it takes for the server to process the request and send the response to the client. Load balancing aims to minimize response times by distributing the workload efficiently, ensuring that each server handles an optimal amount of traffic to expedite request processing.
Throughput: Throughput refers to the rate at which a server or the entire system processes incoming requests within a given timeframe. Load balancing strategies aim to maximize throughput by evenly distributing the workload among servers[6].
Server Utilization: Server utilization gauges the percentage of computing resources (CPU, memory, disk I/O) utilized by each server within the cluster. Balancing the workload aims to evenly distribute utilization across servers to prevent overloading or underutilization, optimizing resource usage.

III. METHODOLOGY

Here's an outline for a methodology that could be employed for a unique and comprehensive comparison between load balancing and scheduling algorithms in cloud environments:

A) Objective Definition:

Clearly define the specific objectives of the comparison. Determine the metrics for evaluation (performance, scalabilit, resource utilization, etc.) and the scope of algorithms to be compared.

1) Selection of Algorithms

Identify a set of representative load balancing and scheduling algorithms commonly used in cloud environments. Consider algorithms like Weighted Round Robin, Least Connections for load balancing, and Round Robin, First Come First Serve, Shortest Job First for scheduling ,task scheduling algorithms such as Round Robin[7].

2) Performance Metrics

Response Time: Measure the time taken by each algorithm to process requests/tasks.
Throughput: Evaluate the number of requests/tasks processed per unit time.
Resource Utilization: Analyze the efficiency of resource usage by algorithms under different workloads.
Scalability: Test the algorithms' ability to handle an

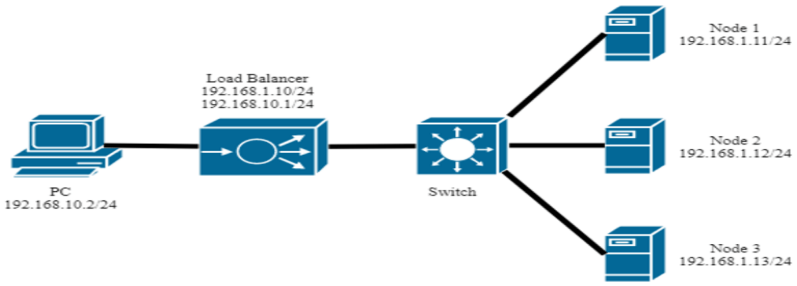

Figure 2 cluster server

Discussion and Conclusion

Summarize the findings, highlighting the strengths and weaknesses of each algorithm. Discuss the implications of the results in real-world cloud deployments, considering factors like efficiency, adaptability, and suitability for different applications[8].

A) How DNS works

When a user enters a URL (Uniform Resource Locator) like "www.xyz.com" into a web browser, the browser initiates a request to access that particular website The DNS resolver or server processes the request by first checking its cache to see if it already has the IP address associated with the domain name. If found in the cache (cached DNS record), it returns the IP address directly to the browser.

Load Balancing using DNS server

DNS servers are configured to associate multiple IP addresses with a single domain name (e.g., "www.xyz.com"). Each IP address corresponds to a different server within the server cluster hosting the web application. DNS-based load balancing employs a simple round-robin mechanism, where the DNS server rotates through the list of IP addresses in a sequential manner for each new request it receives. For instance[9].
First request: The DNS server returns the IP address of Node 1 (e.g., 192.168.1.11/24).
Second request: The DNS server returns the IP address of Node 2 (e.g., 192.168.1.12/24).
Third request: The DNS server returns the IP address of Node 3 (e.g., 192.168.1.13/24).
Subsequent requests follow this sequence, evenly Distribute traffic between servers in clusters. This technique ensures that incoming requests are equally distributed among all servers in the cluster. Each server receives its share of traffic, allowing for better resource utilization and preventing overload on any single server[10].

IV. FINDING

Nature of Functionality: Load Balancing: Focuses on distributing incoming network traffic across multiple servers to ensure optimal utilization of resources and minimize response time. It deals with real-time adjustments based on traffic pattern[11].
Scheduling: Concentrates on task allocation and resource management, ensuring efficient utilization of available resources by allocating tasks to appropriate nodes or virtual machines. It handles long-term resource allocation decisions.
Dynamic Adaptability: Load Balancing: Primarily reacts to immediate changes in network traffic, adjusting resource allocation on-the-fly to hand le fluctuations in demand or traffic patterns.

1) *Resource Optimization*: Load Balancing: Aims to evenly distribute incoming requests across multiple servers to avoid overloading specific nodes, thereby preventing bottlenecks and maximizing throughput[12].
2) *Latency and Response Time Impact*: Load Balancing: Directly influences response time by directing traffic to the least loaded server, reducing latency and enhancing user experience.
3) *Impact on Workload Management*: Load Balancing: Plays a crucial role in handling varying workloads by dynamically distributing incoming requests, ensuring balanced loads across servers.
4) *Algorithmic Variations*: Load Balancing: Algorithms vary from simple Round Robin to more complex Weighted Round Robin, Least Connections, or Dynamic algorithms adjusting to real-time conditions.

This unique comparison outlines how load balancing and scheduling algorithms operate differently within cloud environments, showcasing their distinctive roles, adaptability, and impact on resource utilization and performance optimization[13].

V. CONCLUSION

In conclusion, the comparison between load balancing and scheduling algorithms in cloud environments underscores their unique yet complementary roles in optimizing resource utilization and enhancing system performance.Load balancing algorithms, with their real-time adaptability, excel in dynamically distributing incoming traffic across servers, minimizing latency, and ensuring optimal response times. They are pivotal in handling immediate demands and maintaining system stability during varying traffic patterns.
Both types of algorithms exhibit a diverse range of variations tailored to specific scenarios and workload characteristics, ensuring optimal resource allocation and performance optimization within cloud infrastructures. In essence, the synergy between load balancing and scheduling algorithms is indispensable for the seamless operation of cloud environments. Their distinct functionalities, adaptabilities, and impacts collectively contribute to the

efficiency, scalability, and reliability of distributed systems in the cloud, forming the backbone of modern computing infrastructures.

DISCUSSION

Load Balancing and Scheduling: Balancing Real-Time Demands and Resource Allocation
Load balancing and scheduling are pivotal elements in the realm of cloud computing, each addressing distinct facets of resource optimization and performance enhancement. Load balancing primarily deals with the immediate distribution of incoming traffic to ensure optimal utilization of resources and minimal latency. On the other hand, scheduling algorithms focus on the efficient allocation of resources for tasks or jobs based on predefined criteria, aiming for long-term resource utilization efficiency.

A) Dynamic Adaptability vs. Planned Resource Allocation

The essence of load balancing lies in its real-time adaptability. It dynamically responds to fluctuating traffic patterns, ensuring that no single server gets overloaded, thus optimizing response times and maintaining system stability. Contrastingly, scheduling algorithms are more planned in their approach, making decisions based on predefined policies or algorithms. They aim for optimized resource allocation over a more extended period, considering workload characteristics and resource demands.

B) Immediate Impact vs. Long-term Efficiency

Load balancing directly impacts immediate response times and throughput by distributing incoming requests among servers in real-time. Its goal is to minimize latency and enhance user experience by routing traffic to the most available and responsive server.

REFERENCES

[1] S Afzal, G Kavitha, “Load balancing in cloud computing A hierarchical taxonomical classification,” Journal of Cloud Computing, Vol.8, Issue.1, pp.386-398, 2019.

[2] Peter M. Mell, Timothy Grance, “The NIST definition of cloud computing,” National Institute of Standards and Technology, Vol.1, Issue.1, pp.1-7, 2011.

[3] Sahu A. K. & Kishore A. (2011). “A Throttle Based Approach to Mitigate Distributed Denial of Service (DDOS) Attacks”, International Journal of Applied Science and Technology (IJAST); Vol. 1(1), pp. 9-18. ISSN: 2231-3842.

[4] Shadab Siddiqui, Manuj Darbari, Diwakar Yagyasen, “Evaluating Load Balancing Algorithms for Performance Optimization in Cloud Computing,” International Journal of Innovative Technologyand Exploring Engineering (IJITEE),Vol.8, Issue.12, pp.2217- 2221, 2019.

[5] K. Suresh Kumar, Vinay Kumar Nassa, Dipesh Uike, Ashima kalra, Ajay Kumar Sahu, Vijay Anant Athavale, V. Saravanan (2022), "A Comparative Analysis of Blockchain in Enhancing the Drug Traceability in Edible Foods Using Multiple Regression Analysis", Journal of Food Quality, vol. 2022, 6 pages. https://doi.org/10.1155/2022/1689913. Hindawi, ISSN: 1687-5265, Impact Factor. 3.200, https://www.hindawi.com/journals/cin/

[6] Hyame Almandine, Sara Ayoubi, “An Efficient Survivable Design with Bandwidth Guarantees for Multi-Tenant Cloud Networks,”IEEE Transaction on network and service management, Vol.14, Issue.2, pp.357-372, 2017.

[7] Jula, Amin, Elan kovan Sundarajan and Zalinda Othman, “Cloud Computing Service Composition: A systematic review,” Journal of information systems and communications, Vol.3, Issue.1, pp.87-91, 2012.

[8] Sahu, A.K., & Kumar, A. (2019), “SPAS: An Authentication Scheme to Prevent Unauthorized Access of Information from Smart Card”, Pertanika Journal of Science and Technology; Vol. 27(1): pp. 175-192. ISSN: 0128-7680. http://www.pertanika.upm.edu.my/

[9] Talwana Jonathan Charity, Gu Chun Hua ,”Resource Reliability using Fault Tolerance in Cloud Computing,” In the proceedings of the 2016 International Conference on Next Generation Computing Technologies-IEEE , India , 2016.

[10] Sahu A. K., & Kumar, A. (2016), “An Improved Remote User Password Authentication Scheme Using Smart Card with Session Key Agreement”, International Journal of Control Theory and Applications; Vol. 9(17), pp. 8445-8454

[11] Fatima Shakeel, Seema Sharma, “Green Cloud Computing: A review on Efficiency of Data Centres and Virtualization of Servers,” In the proceeding of the 2017 International ConferenceonComputing, Communication and Automation–IEEE, India 2017.

[12] Sahu, A. K., & Kumar, A. (2021), “A Novel Verification Protocol to Restrict Unconstitutional Access of Information From Smart Card”, International Journal of Digital Crime and Forensics (IJDCF); Vol. 13(1): pp. 65-78. doi:10.4018/IJDCF.2021010104.

[13] Ibrahim, D. Kliazovich, P. Bouvry, “On Service Level Agreement Assurance in Cloud Computing Data Centers,” In the proceedings of the 2016 International Conference on Cloud Computing-IEEE, India , 2016.