International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol 13 Issue 02 | 2026

Load Balancing Mechanism in Fog Computing

Dr. Sachin A Vyawhare
Assistant Professor, [1]Sanmati Engineering College, Washim, Maharashtra

DOI: 10.5281/zenodo.20606296

ABSTRACT

Recent years have seen the emergence of fog computing as a contemporary distributed paradigm that works in conjunction with cloud computing through the provision of services. As a result of the fog system's ability to improve storage and compute at the network's edge, it is able to successfully handle the issues of service computing for delay-sensitive applications while also supporting location awareness and mobility assistance. Load balancing is an essential component of fog networks that helps to prevent the occurrence of fog nodes that are either underutilized or with an excessive amount of workload. In order to improve Quality of Service (QoS) characteristics, load balancing can be utilized. These characteristics include resource utilization, throughput, cost, response time, performance, and energy consumption. Within the past few years, there has been research carried out on various load balancing strategies for fog networks; nevertheless, there is a lack of a comprehensive study that consolidates these results. A comprehensive analysis of load balancing algorithms in fog computing is presented in this article. These algorithms are categorized into four distinct categories: approximation, exact, fundamental, and hybrid techniques. The analysis is based on articles that were published between 2013 and August 2020. In this article, load balancing measurements are discussed, along with the benefits and drawbacks that are linked with various load balancing algorithms that are utilized in fog networks. The procedures and instruments of evaluation that were applied for each of the studies that were analyzed are also investigated. Furthermore, the important open difficulties that these processes face as well as the potential tendencies within them are investigated.

Keywords: Fog computing, load balancing, Quality of Service, Internet of things, systematic review

1. INTRODUCTION

Fog computing is an emerging paradigm that extends traditional cloud computing by distributing computational resources closer to end-users and Internet of Things devices. Unlike centralized cloud systems, fog computing deploys geographically distributed nodes—known as fog nodes—at the network edge. These nodes provide local processing, storage, and networking capabilities, enabling faster data handling and real-time decision-making. This proximity to devices significantly enhances system responsiveness and supports delay-sensitive applications. In traditional cloud-only architectures, large volumes of data generated by IoT devices are transmitted to centralized cloud servers for processing, analysis, and storage. This approach often leads to increased latency, bandwidth consumption, and potential security risks. Additionally, applications that require real-time responses—such as smart healthcare, autonomous vehicles, and industrial automation—cannot tolerate such delays. Fog computing addresses these challenges by processing data closer to the source, thereby reducing latency, improving reliability, and enhancing mobility support. The fog layer works in close coordination with Cloud Computing, forming a hybrid architecture that leverages the strengths of both paradigms. While fog nodes handle time-sensitive tasks locally, the cloud is responsible for large-scale data storage, complex analytics, and long-term processing. This collaboration enables the development of next-generation applications that demand both real-time responsiveness and high computational power. One of the critical challenges in fog computing environments is maintaining Quality of Service (QoS), especially due to the dynamic and heterogeneous nature of fog networks. Users increasingly expect applications to respond instantly and operate efficiently. To meet these expectations, load balancing plays a vital role. Load balancing refers to the efficient distribution of incoming tasks across multiple fog nodes or between fog and cloud resources. Proper load balancing ensures that no single node becomes overloaded while others remain underutilized.

Unlike cloud environments, where resources are relatively stable and centralized, fog networks consist of distributed and diverse nodes with varying capabilities. Therefore, traditional cloud-based load balancing techniques cannot be directly applied to fog environments. Instead, adaptive and intelligent load balancing mechanisms are required to handle dynamic workloads and network conditions effectively.

The primary objective of load balancing in fog computing is to optimize resource utilization while improving system performance. An efficient load balancing strategy helps in maximizing throughput and minimizing response time, operational cost, and energy consumption. By distributing workloads intelligently, the system can maintain stability, enhance user experience, and ensure seamless operation of latency-sensitive IoT applications. Fog computing plays a crucial role in addressing the limitations of cloud computing by bringing

computation closer to the data source. When combined with effective load balancing strategies, it significantly improves performance, scalability, and efficiency in modern IoT-driven systems.

2. LITERATURE SURVEY

With the rapid growth of the Internet of Things, it is estimated that more than 50 billion devices are connected to the Internet, generating an unprecedented volume of data. Handling such massive data using traditional computing models like distributed systems or Cloud Computing has become increasingly challenging. These models often suffer from limitations such as high communication delays, increased network traffic, and privacy concerns. As IoT applications like smart cities, eHealth systems, industrial automation, and smart transportation continue to expand, these challenges significantly impact the overall performance and efficiency of cloud-based systems.

To address these limitations and bring computational capabilities closer to end devices (such as sensors, mobile phones, vehicles, and embedded systems), the concept of fog computing has been introduced. Fog computing acts as an intermediate layer between cloud data centers and end devices, enabling faster processing and localized decision-making. The term “fog” was initially introduced by the Cisco, inspired by the natural phenomenon where fog exists closer to the ground while clouds remain higher in the sky. Similarly, fog computing places computational resources closer to users, reducing the distance that data must travel.

Fog computing is considered a distributed and virtualized platform that provides computing, storage, and networking services between end devices and cloud data centers. According to early definitions by researchers such as Flavio Bonomi, fog computing is not limited to the network edge but extends across multiple layers of the network, enabling seamless interaction between devices and centralized systems. This multi-layer architecture supports real-time processing, reduces latency, and improves overall system responsiveness.

A common misconception is that fog computing is the same as edge computing; however, there are important differences between the two. Fog computing operates in a hierarchical, multi-layered architecture where processing can occur at various intermediate nodes between the cloud and end devices. It supports dynamic reconfiguration of both hardware and software components, allowing flexibility for diverse applications. In contrast, edge computing typically performs computations directly on or near specific devices in a fixed location, focusing on localized processing without extensive hierarchical coordination.

Furthermore, fog computing goes beyond simple data processing by incorporating additional functionalities such as data storage, network control, and acceleration of processing tasks. It enables intelligent decision-making closer to the data source while still maintaining coordination with cloud systems for large-scale analytics and long-term storage. This makes fog computing particularly suitable for latency-sensitive and location-aware applications.

To effectively utilize fog computing services, IoT devices or clients may exhibit several characteristics, such as context awareness, low-latency communication, mobility support, and real-time data processing capabilities. These features collectively distinguish fog computing from other computing paradigms and highlight its importance in supporting next-generation IoT applications.

In conclusion, fog computing provides a powerful solution to the challenges posed by large-scale IoT environments. By bridging the gap between cloud infrastructure and end devices, it enhances performance, reduces latency, and enables efficient handling of massive data generated in modern smart systems.

3. FOG ARCHITECTURE

The reference model for fog computing architecture is an important piece of information that should be investigated intellectually. The majority of fog computing designs are developed from a framework that is composed of three layers [11], [16]. Recent times have witnessed the emergence of a wide range of fog computing architectures. To expand cloud services to the outside of the network, the fog network suggests the creation of a fog layer between the cloud and user devices. This layer would be placed between the cloud and the user devices. The fog architecture is depicted in a hierarchical organization, and it is composed of three layers, which are as follows:

Cloud layer: The cloud layer is the layer of cloud computing that refers to the collection of high-performance computers and a variety of storage devices. It is accountable for the development of a variety of services and apps over the course of its existence. Because of its high computing and storage capabilities, it is able to support the long-term storage of a considerable amount of information, as well as extended processing and analysis. Additionally, it can support the storage of information for multiple years. However, it is essential to keep in mind that not all processing and storage tasks can be finished in the cloud, which is not the same as the architecture of traditional cloud computing [19]. This is something that should be taken into consideration.

Fog layer: The fog layer is located at the periphery of the network and is composed of a few fog nodes. These fog nodes include access points, routers, switches, gateways, and other devices that are comparable to these. It is possible to find them in both the cloud and on end devices where they are distributed. It is possible for end devices to easily connect with fog nodes in order to get services. Their powers include computing, storing, and

transmitting the information that they have gathered from their senses. Real-time analysis and the execution of latency-sensitive applications are both possible in the fog layer, which provides an environment that is ideal for both operations. Additionally, we are able to make reference to the connection that is established between the cloud data center and the fog nodes through the use of the IP core network technology. Fog nodes are assigned with the responsibility of interfacing and working together with the cloud in order to gain more powerful computing and storage capabilities. This is done in order to acquire more powerful capabilities.

User device: The layer of a user device is located in close proximity to the end-user and the actual environment. There are many different Internet of Things devices that are included in this layer. Some examples of these devices are sensors, telephones, smart cars, cards, and readers. In spite of the fact that wireless mobile phones and intelligent automobiles are both outfitted with computer capabilities, the fundamental purpose of these gadgets is that of intelligent sensing devices. They are accountable for the sensing of feature data that is associated with events or physical objects, and they are responsible for transmitting this data to higher layers so that it can be processed and stored. They can be found all over the place, both in terms of geography and in general.

Through the utilization of technologies that allow for the establishment of wired or wireless connections, this architecture ensures that all end devices or smart objects are linked to fog nodes simultaneously. 3G, 4G, wireless local area networks (LAN), ZigBee, Bluetooth, and Wi-Fi are all examples of these technologies. Bluetooth and Bluetooth are two further examples. The technologies that permit wireless or wired communication communication can make it easier for fog nodes to connect with one another and communicate with one another. Through the utilization of the Internet Protocol core network, fog nodes are able to establish connections to the cloud [10].

4. LOAD BALANCING

In a fog computing environment, load balancing plays a crucial role in ensuring efficient distribution of workloads across available resources. The primary objective is to maintain continuous service availability, even in the event of component failures, while optimizing resource utilization. By dynamically provisioning and de-provisioning application instances, load balancing helps maintain system stability and performance. Since fog environments consist of heterogeneous nodes with varying capabilities and traffic patterns, an effective load balancing mechanism is essential to enhance application performance and network efficiency.

Load balancing distributes workloads across multiple nodes to prevent situations of overload or underutilization. This distribution can be implemented either through hardware-based solutions or software-based algorithms. The key goals of load balancing include maximizing throughput, minimizing response time, optimizing network traffic, and improving scalability in distributed systems. Additionally, it aims to reduce server-side resource consumption and minimize task processing delays, thereby enhancing overall system efficiency.

In fog computing, load balancing strategies can be broadly categorized into static and dynamic methods. Static load balancing relies on predefined rules and prior knowledge of the system. However, due to the unpredictable nature of user behavior and network conditions, static methods often fail to adapt effectively in real-time environments. In contrast, dynamic load balancing methods are more efficient as they consider the current state of the system and distribute workloads accordingly. These methods continuously monitor system performance and make adaptive decisions, making them more suitable for fog networks.

Dynamic load balancing mechanisms operate based on several key policies that guide task distribution and resource management:

Transfer Policy

This policy determines whether a task should be transferred from one node to another. It evaluates incoming tasks and decides, based on workload conditions, whether to process them locally or offload them to another node. It plays a central role in task migration and rescheduling decisions.

Selection Policy

The selection policy identifies which tasks are suitable for transfer. It considers factors such as migration overhead, execution time, and system calls involved. This ensures that only tasks that benefit from migration are selected, avoiding unnecessary overhead.

Location Policy

Once a task is selected for transfer, the location policy identifies suitable under-loaded nodes where the task can be executed. It ensures that the selected node has the required resources and services to handle the task efficiently.

Information Policy

This policy is responsible for collecting and maintaining up-to-date information about the system state, including node workloads and resource availability. It determines when and how system information should be gathered and shared with other policies to support informed decision-making.

These policies work together in a coordinated manner. Initially, the transfer policy evaluates incoming tasks and decides whether they should be migrated. If migration is required, the location policy identifies an appropriate

destination node. If no suitable node is found, the task is processed locally. The selection policy ensures that only appropriate tasks are chosen for transfer, while the information policy continuously updates the system state to support all decision-making processes.

In summary, load balancing in fog computing is a critical mechanism for achieving high performance, reliability, and scalability. By intelligently distributing workloads and adapting to real-time conditions, dynamic load balancing techniques significantly enhance the efficiency of Fog Computing environments and support the growing demands of modern IoT applications.

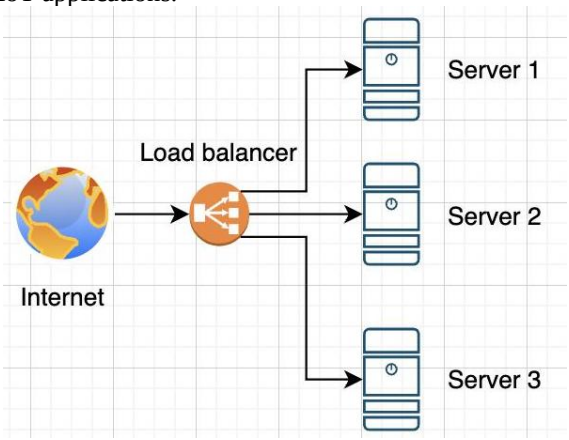

Fig. 1 Load Balancer Architecture

5. CHALLENGES IN FOG COMPUTING

In order to address issues and deficiencies that are associated with the Internet of Things (IoT) at the network edge, fog computing is considered to be an evolved extension of the cloud computing system. On the other hand, processing nodes in fog computing are separated from one another and are diverse. In addition to this, the services that are based on fog technology are required to be compatible with a variety of features of the limited environment. In addition, the assurance of security is the most important aspect of fog computing. Consequently, the following is a summary of the issues that can be discovered in fog computing from the perspectives of service-oriented computing, structural computing, and security related to this technology: Service-oriented resources do not enrich every fog node in the network. As a result, the enhancement of applications on a comprehensive scale in nodes with limited resources is not natural when compared to typical data centers. Since this is the case, it is necessary to implement dispersed application development demands for potential programming platforms in Fog. Furthermore, it is necessary for a fog administrator to specify the policies in order to allocate the necessary jobs among the sensors and Internet of Things devices, as well as the fog infrastructure.

Various components from both the core and the outskirts of networks make up the infrastructure of fog computing. This infrastructure presents a number of structural challenges. Although they are equipped with a different calculation, these types of components are not suited for universal computing due to their design. Redesigning or modifying the compute unit for the component is therefore an incredibly tough aspect of the process of setting up the system. Additionally, the selection of the appropriate device, locations of deployment, and resource configurations that match to those locations are essential in fog computing. This is because the selection of these factors is based on the execution activities and operational needs. In fog computing, computational devices are dispersed over the borders of the network and there is the possibility of sharing or virtualizing them. In this particular scenario, it is essential to establish appropriate indicators, techniques for inter-nodal collaboration, and effective resource supply. Aspects Relating to Security Because fog computing is dependent on traditional networking components, it is extremely vulnerable to security threats. A globally spread paradigm, such as fog computing, presents a number of challenges when it comes to ensuring the protection of users' privacy and providing them with authenticated access to computing and storage facilities. Therefore, maintaining quality of service during the installation of security is problematic, especially in situations when the integrity of the data center is adequate, which makes the topic of security in fog computing challenging.

6. PROPOSED SYSTEM

Cloud load balancing refers to the distribution of workload and computational resources within cloud computing environments. It enables enterprises to manage workload or application requirements by distributing resources among several PCs, networks, or servers. Cloud load balancing involves the management of workload distribution and requirements across the Internet. Internet traffic is rapidly increasing, currently accounting for

approximately 100% of annual traffic. As a result, the server workload is increasing swiftly, leading to server overload, especially for the most frequented web servers. Two primary strategies exist to mitigate the problem of server overloading.

The primary alternative entails upgrading to a more advanced single-server setup. However, the new server may soon encounter overload, requiring an additional update. Moreover, the improvement process is arduous and expensive.

The second is a multi-server strategy that creates a scalable service framework on a cluster of servers. Thus, developing a server cluster system for network services is more cost-effective and scalable.

Cloud-based servers can achieve improved scalability and availability by employing farm server load balancing. Load balancing is beneficial for almost all service categories, including HTTP, SMTP, DNS, FTP, and POP/IMAP.

7. CONCLUSION

Load balancing is a fundamental component in Fog Computing environments, playing a critical role in ensuring efficient resource utilization, improved performance, and reliable service delivery. Due to the heterogeneous and dynamic nature of fog networks, traditional cloud-based load balancing approaches are not sufficient. Instead, adaptive and intelligent mechanisms are required to manage workload distribution effectively across fog nodes and cloud resources. Dynamic load balancing techniques, supported by policies such as transfer, selection, location, and information, provide a flexible and efficient way to respond to real-time changes in system conditions. These techniques help prevent overload and underutilization of resources, thereby enhancing throughput, reducing response time, and optimizing energy consumption. The coordination among different policies ensures that tasks are allocated to the most suitable nodes, maintaining system stability and performance. Overall, effective load balancing in fog computing enables scalable, responsive, and high-performance systems, which are essential for supporting modern Internet of Things applications. As IoT continues to expand, the development of more intelligent and context-aware load balancing strategies will be crucial for meeting the increasing demands of next-generation distributed computing environments.

REFERENCES

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, 2012: ACM, pp. 13-16.

[2] A. Yousefpour *et al.*, "All one needs to know about fog computing and related edge computing paradigms: A complete survey," *Journal of Systems Architecture,* vol. 98, pp. 289-330, 2019/09/01/ 2019.

[3] N. Auluck, A. Azim, and K. Fizza, "Improving the Schedulability of Real-Time Tasks using Fog Computing," *IEEE Transactions on Services Computing,* pp. 1-1, 2019.

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

[5] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *Journal of Systems and Software,* vol. 80, no. 4, pp. 571-583, 2007/04/01/ 2007.

[6] P. Jamshidi, A. Ahmad, and C. Pahl, "Cloud Migration Research: A Systematic Review," *IEEE Transactions on Cloud Computing,* vol. 1, no. 2, pp. 142-157, 2013.

[7] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Information and Software Technology,* vol. 51, no. 1, pp. 7-15, 2009/01/01/ 2009.

[8] C. Jatoth, G. R. Gangadharan, and R. Buyya, "Computational Intelligence Based QoS-Aware Web Service Composition: A Systematic Literature Review," *IEEE Transactions on Services Computing,* vol. 10, no. 3, pp. 475-492, 2017.

[9] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: a proposal and a discussion," *Requirements engineering,* vol. 11, no. 1, pp. 102–107, 2005.

[10] M. Haghi Kashani, A. M. Rahmani, and N. Jafari Navimipour, "Quality of service-aware approaches in fog computing," *International Journal of Communication Systems,* 2020/01/10 2020.

[11] M. Rahimi, M. Songhorabadi, and M. H. Kashani, "Fog-based smart homes: A systematic review," *Journal of Network and Computer Applications,* vol. 153, p. 102531, 2020/03/01/ 2020.

[12] S. Bazzaz Abkenar, M. Haghi Kashani, E. Mahdipour, and S. M. Jameii, "Big data analytics meets social media: A systematic review of techniques, open issues, and future directions," *Telematics and Informatics,* p. 101517, 2020/10/14/ 2020.

[13] P. Hu, S. Dhelim, H. Ning, and T. Qiu, "Survey on fog computing: architecture, key technologies, applications and open issues," *Journal of network and computer applications*, vol. 98, pp. 27-42, 2017.

[14] P. Asghari, A. M. Rahmani, and H. H. S. Javadi, "Internet of Things applications: A systematic review," *Computer Networks* vol. 148, pp. 241-261, 2019/01/15/ 2019.

[15] E. Marín-Tordera, X. Masip-Bruin, J. García-Almiñana, A. Jukan, G.-J. Ren, and J. Zhu, "Do we all really know what a fog node is? Current trends towards an open definition," *Computer Communications*, vol. 109, pp. 117-130, 2017.

[16] O. C. A. W. Group, "OpenFog reference architecture for fog computing," *OPFRA001*, vol. 20817, p. 162, 2017.

[17] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 416-464, 2017.

[18] M. Iorga, L. Feldman, R. Barton, M. Martin, N. Goren, and C. Mahmoudi, "Fog computing conceptual model, recommendations of the National Institute of Standards and Technology," *NIST Special Publication*, pp. 500-325, 2018.