International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

Survey of Load Balancing Strategies in Fog-Cloud Architectures for IoT Integration

Dr. Puneet Garg

Associate Professor, Department of CSE-AI, KIET Group of Institutions, Delhi NCR, Ghaziabad, India

puneet.garg@kiet.edu

Publication History: Received: 18.03.2026; Revised: 10.04.2026; Accepted: 13.04. 2026; Published: 18.04.2026.

ABSTRACT: The Internet of Things (IoT) innovations have experienced high growth and, therefore, generate a lot of real-time data, which requires processing frameworks with a considerable efficiency and scalability low latency system. Traditional cloud computing has many benefits, but it might not be able to meet the bandwidth and latency requirements of time-dependent Internet of Things devices. To solve these shortcomings, fog computing has come to redefine a related computing paradigm, which pushes computation and storage resources to network endpoints. Fog-cloud architecture, which combines fog with cloud, offers a hybrid capability that improves system scalability, responsiveness, and resource efficiency. Devices running on heterogeneous platforms gain greatly from IoT interaction with the cloud. Applications based on the Internet of Things produce vast amounts of data from various sensors. Decisions are made by analyzing this data. Nevertheless, use cases of IoT environments are dynamic and heterogeneous posing challenging issues in the distribution of work load in such a layered infrastructure. The provided study includes a thorough analysis of fog-cloud system load balancing solutions, including their categorization, performance metrics, and applicability. It also discusses many vital issues, including latency constraints, resource heterogeneity, energy efficiency, and provides future research meets the objectives of intelligent, adaptive, and context-sensitive load balancing in the next-generation IoT systems.

KEYWORDS: Fog Computing, CC, Fog-Cloud Architecture, Load Balancing, IoT, Real-Time Processing, Resource Allocation, Edge Computing, Latency Reduction, Scalability, QoS.

I. INTRODUCTION

Consumer interest in data-intensive services has skyrocketed thanks in large part to cloud computing's spectacular ascent. Both academics and businesses have recently taken an interest in cloud service portfolios [1]. Cloud computing platforms offer the optimal computational paradigm for meeting user needs simultaneously by merging various resources. With a multi-provider multi-service system design, users' requests are typically satisfied by simulating the service execution process on clouds. To ensure that a cloud-based solution lives up to customer service expectations, it is possible to integrate many clouds and services from various suppliers into an interconnected set of services [2][3]. Despite being adept at compute and storage, traditional cloud computing infrastructures are unable to provide the ultra-low latency, location awareness, and real-time responsiveness required by contemporary IoT applications [4]. Communication delays and bandwidth constraints are frequently encountered when data is sent between geographically dispersed edge devices and centralized cloud data centers [5]. Particularly for latency-sensitive applications like emergency response systems or smart traffic management, these limitations may worsen QoS, increase energy consumption, and decrease performance [6].

To get around these problems, fog computing has emerged as a supplemental paradigm to cloud computing. Fog computing decentralizes processing, storage, and networking resources by relocating them to the edge, near the actual IoT devices [7]. This improves real-time processing capabilities, lowers latency, and eases network congestion [8]. A fog-cloud hybrid architecture, which combines fog and cloud systems, offers a scalable and adaptable computing continuum. Workloads are dynamically split between fog nodes located nearer data sources and centralized cloud servers in such an architecture, guaranteeing effective resource use and flexible handling of IoT workloads. One crucial mechanism is load balancing [9]. To maintain high system availability, avoid node overloading, and guarantee appropriate job allocation, effective load balancing techniques are crucial. Because the workloads generated by the Internet of Things are inherently varied and dynamic, load balancing techniques are crucial for achieving energy economy, low latency, fault tolerance, and adherence to Service Level Agreements (SLAs) [10]. The main difficulty is distributing computing loads across several fog and cloud nodes while taking mobile devices and changing network circumstances into account [11]. As a result, a wide range of

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

load balancing techniques, from sophisticated AI-driven methods to rule-based heuristics, have been created to meet the particular requirements of IoT-powered Fog-Cloud integrated settings.

A. Structured of the paper

This paper is organized as follows: The foundations of fog-cloud architecture are reviewed in Section II. The topic of load balancing in fog-cloud environments is covered in Section III. Section IV illustrates IoT in a fog-cloud context. In Section V, recent literature are reviewed with their research gaps, Section VI concludes the study and future directions are discussed.

II. FUNDAMENTALS OF FOG-CLOUD ARCHITECTURES

A centralized computing paradigm known as cloud computing uses the internet to provide on-demand services including storage, processing power, and software programs. It makes use of multi-tenant architecture and virtualization to provide scalability, resource sharing, and cost effectiveness. In the context of IoT, cloud platforms are responsible for long-term data storage, intensive analytics, and global service orchestration [12]. However, latency-sensitive and bandwidth-intensive IoT applications face limitations due to the physical remoteness of cloud data centers.

A decentralized computing model called fog computing moves cloud functions to the network's periphery, moving towards the sources of information, i.e., IoT devices and sensors. Fog computing, introduced by Cisco, seeks to lower latency and augment location awareness [13][14], as well as provide real-time processing capabilities. In essence, fog computing is a cloud extension that is more akin to devices that handle Internet of Things data [15]. Fog computing, as seen in Fig. 1, serves as a bridge between the cloud and endpoints, bringing networking, processing, and storage functions closer to the endpoints. We refer to these gadgets as fog nodes. Anywhere there is a network connection, they may be set up [16]. Fog nodes can be any device that has computation, storage, and network connection, including embedded servers, switches, routers, industrial controls, and security cameras.

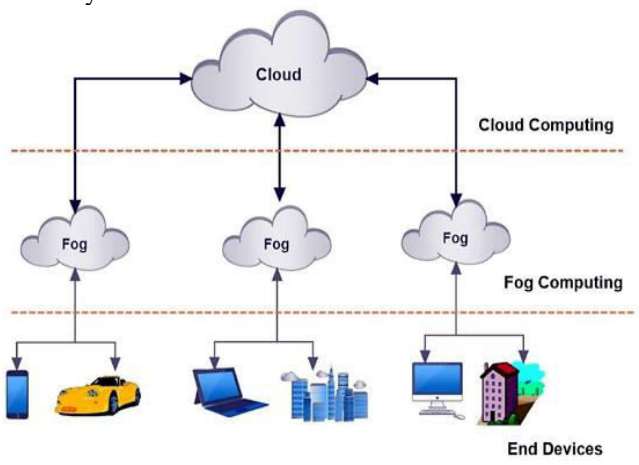

Fig. 1. Hierarchical Fog-Cloud Architecture for IoT Data Processing

Fog nodes carry out intermediary processing and data storage activities at the site of generation or the closest locations to the source, in contrast to typical cloud computing, which decentralizes processing and data storage in remote data centers [17]. The following are some essential features of fog computing:

- **Nearness to paying customers:** Eliminates delays and provides better responsiveness.
- **Mobility support:** Enables easy access to data in moving conditions.
- **Location awareness:** provides the ability to contextual applications.
- **Low latency:** Essential to real-time applications like autonomy vehicles as well as industrial automation.
- **Distributed architecture:** It increases fault tolerance and scalability[18].

B. Fog-Cloud Synergy for IoT Applications

In IoT ecosystems, cloud and fog play complementary roles. The cloud serves as the backbone for large-scale analytics, system-wide updates, and global coordination. Fog nodes, on the other hand, serve as middlemen to lessen data transfer to the cloud, facilitating quicker decision-making and easing network congestion [19][20]. The fog layer also offers enhanced

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

support for mobility, location-awareness, and context-aware services, making it ideal for distributed and dynamic environments. Fog and cloud combining, or less formally fog-cloud architecture, is a tactical integration of decentralized and centralized data processing paradigms, set to accommodate the requirements of real-time, data-sensitive IoT networks. The constraints of cloud-only based infrastructures have also been solved by fog computing so that computation [16], Data storage and analysis Decision effects are concentrated at the network's edge, while the cloud offers deep analytics, elastic scalability, and global visibility [21]. This hybrid system improves the efficiency of the wholesome system, which smartly off-loads tasks depending on the predilection of the computation, latency, and the resource. Table I shows a fog vs cloud computing with the role of the IOT integrations are as follows:

TABLE I. COMPARISON OF FOG AND CLOUD COMPUTING

Aspect	Fog Computing	Cloud Computing
Architecture Type	Decentralized, distributed across edge devices	Centralized, located in data centers.
Proximity to Devices	Closer to IoT devices (edge level)	Distant from IoT devices
Latency	Very low (supports real-time responses)	Higher latency due to network transmission delays
Data Processing	Local, real-time, and context-aware processing	Batch processing and large-scale data analytics
Bandwidth Usage	Reduced, as only filtered/processed data is sent to cloud	High, as raw data is transmitted from devices
Storage Capacity	Limited, local storage	Virtually unlimited cloud storage
Scalability	Moderate scalability, localized scaling	High scalability, elastic resources
Reliability in IoT	High, ensures continuous operation during connectivity loss	Dependent on stable internet connectivity
Security & Privacy	Better for local data handling and privacy-sensitive applications	Centralized controls with possible compliance concerns
Use Case Examples	Real-time health monitoring, traffic control, smart manufacturing	Historical trend analysis, AI model training, data warehousing
Primary Role in IoT	Handles immediate, real-time, and location-aware IoT requirements	Supports global insight, orchestration, and deep analytics

C. Architectural Layers and Communication Flow

As seen in Fig. 2 of the hierarchy, the fog-cloud architecture is often organized into three levels: the edge layer, the fog layer, and the cloud layer.

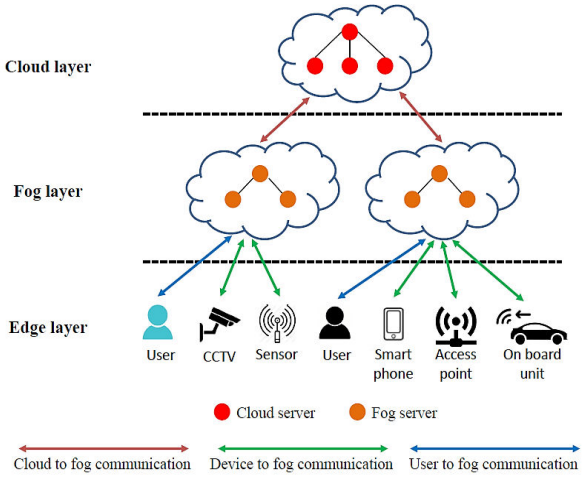

Fig. 2. Three-layer architecture of fog computing

The fog computing architecture consists of the following three primary layers:

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

- **Edge Layer:** The layer includes data creation and initial transmission and is made up of IoT devices, sensors, actuators, and gateways. It encourages minimal processing and performs the calculations using nearby fog nodes.
- **Fog Layer:** It is an intermediate controller that hosts fog nodes or edge servers that play a role in data filtering, local analytics, real-time response, and temporary storage. Due to their physical dispersion and proximity to the data source, these nodes minimize latency.
- **Cloud Layer:** There is a centralized data center in the top layer which delivers high performance computing, worldwide data agglomeration, deep learning[22], and long-term storage.

There is usually a two-way flow of communication in this architecture:

- **Upstream**: The sensor-generated data will go from the edge to the fog nodes and maybe to the cloud environment for in-depth analysis.
- **Downstream**: Information, data, or management messages are transmitted upward to the devices by the cloud/fog nodes.

The multi-layer approach makes task offloading efficient, response time low latency and workload adaptive and that is needed in scalable and responsive IoT systems.

III. LOAD BALANCING IN FOG-CLOUD ARCHITECTURE

The goal of load balancing is to ensure that nodes in a distributed system are not overworked or idle and that all nodes are able to respond quickly to jobs by distributing the overall system workload evenly among all of their components, such as disc drivers, network links, and central processing units. Elastic scalability and cloud computing both rely on load balancing [13]. Load balancing is a common tool for preventing system failure by regulating input traffic and removing work from resources that are too busy or unresponsive. In fog-cloud architectures, the favorable distribution of the computational workload is a critical factor in preserving the entire system performance, especially regarding IoT, where the creation of data never stops and sometimes may need real-time or even the near-real-time processing needs [23]. With load balancing, the load can be distributed based on available resources effectively hence one node is not overloaded and another one is not underutilized be it fog nodes, cloud servers or edge devices [24]. Proper load balancing contributes to the usage of resources to the best, delay in response, prevention of system bottlenecks, and enhanced Quality of Service (QoS) [25]. With the heterogenous and changing environment of IoT applications and the diverse types and capabilities of devices that operate under such network environment, it is necessary to distribute the load intelligently to achieve sustainable performance and guarantee reliability through it.

A. Importance of Load Balancing in IoT Ecosystems

An ecosystem of the IoT contains millions of interlinked sensors, actuators and devices that produce huge amounts of data constantly. When load balancing is not performed properly, fog or cloud-based infrastructure can be overloaded, which results in service degradation, higher latency or crashes [26][27].Major reasons load balancing is important in fog-cloud environments driven by IoTs are as follows:

- **Maintaining Low Latency**: Applications such as health monitoring, emergency response systems and traffic control are real-time applications and they need immediate processing. The task assignment mechanism of load balancing works to assign a task of this nature to the closest or least loaded fog nodes to reduce delay.
- **Efficient Resource Utilization**: Resource Utilization Efficiency: Distributed computing systems usually have nodes that have different capabilities. Load balancing does not allow overloading of some of the nodes and underutilization of others to enable maximum utilization of the processing power as well as storage.
- **Improved Fault Tolerance and Resilience**: A typical IoT deployment at a large-scale experience's failure and disconnection of nodes. Load balancing dynamically reallocates work whenever a failure occurs, preserving the continuity in the system and decreasing downtimes.
- **Adaptability to Fluctuating Workloads**: Response to Unstable Business Demands: The data generation through IoT is usually uncertain and even driven. Load balancing algorithms are able to react to any change in workloads, and thus, they are able to help keep a certain level of performance.
- **Energy Efficiency**: Energy Efficiency: A lot of fog gadgets are resource-limited and power-conscious. The load balancing will be able to transfer energy-high-demand activities to nodes with more resources or that consume less power, etc., minimizing the overall consumption of power and lengthening the lifespan of the devices.
- **Scalability and Growth Support**: Scalability and Growth Support: The IoT network is expanding and, simultaneously, the number of devices and data grows. A properly developed load balancing system will have the ability to scale with ease since the workloads can be distributed without altering the architecture significantly.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

B. Key Metrics of Load Balancing

Load balancing strategies should be evaluated based on the quantitative measure of performance in order to determine their effectiveness in fog-cloud architectures related to IoT [28][29]. Such metrics give an understanding of whether or not a system is well performing in terms of QoS needs, especially in time-sensitive and resource-limited settings[30]. The three popular measurements are latency, throughput and energy consumption:

- **Latency**: Latency is a time delay between an assignment of a task and the processing of the result of the task. When used in IoT-based applications, like telemedicine or self-driving cars, a delay may lead to disastrous effect, even minor [31][32]. The set-up of load balancing algorithm can reduce the latency through an optimal distribution of the job to a node which is geographically closer to it or less congested or with more processing resources available. This commonly implies using fog nodes to handle real-time operations in fog-cloud systems, and moving non-time sensitive tasks to the cloud.
- **Throughput**: The quantity of data or tasks that a system can complete in a given length of time is measured by a metric called throughput. High throughput implies that one can push a lot of work to the system, and it can handle the tasks efficiently without causing a bottleneck. Fog-cloud environments: A throughput offers an aggregate efficiency of both fog and cloud layers. Its goal is to maximize throughput and still be fair among different nodes through an appropriate load-balancing strategy.
- **Energy Consumption**: Energy efficiency is obligatory value, especially in the case of fog and edge devices which can be battery-powered or have limited power storage abilities. Energy-aware load balancing allocates tasks in such a way that the total power consumed is as minimal as possible, and latency, throughput requirement is still satisfied. This has as a trade-off between energy savings and performance, and often employs methods like dynamic voltage scaling, task offloading, or workload consolidation.

The other relevant metrics include: task Completion Time, Resource Utilization, Scalability, and Reliability. These metrics are used to evaluate how well a load-balancing technique meets the applications' quality-of-service requirements.

C. Classification of Load Balancing Strategies

Fog-cloud architecture load balancing strategies may be divided into a few broad categories in terms of their methodology and model of implementation:

- **Static Load Balancing**: The division of work is made in advance with the help of pre-designed rules or prior data. Although it is basic and low-overhead, it is not adaptable to the real-time changes.
- **Dynamic Load Balancing**: On-demand choices apply to the real time status of the system (e.g., node load and network traffic). It is flexible, but can bring in computational overhead.
- **Centralized Load Balancing**: All the decision of load balancing is done by a central controller. It allows worldwide visibility, however, can become decisive or a collapse point.
- **Distributed Load Balancing**: Nodes have equal input on the decision-making process of tasks allocation. It scales and is resilient.
- **AI/ML-Based Load Balancing**: Utilizes reinforcement learning, predictive models [33], or heuristics in dynamically scheduling the workloads depending on the patterns and the past data.

Such strategies are frequently unique, adapted or hybridized to fit to the particular IoT situations and infrastructure limitations.

IV. INTERNET OF THINGS (IOT) IN FOG-CLOUD CONTEXT

In order to overcome the particular difficulties presented by extensive, real-time IoT ecosystems, the fog-cloud IoT combines centralized cloud capabilities with decentralized edge intelligence in a synergistic manner. Traditional cloud computing alone often falls short in meeting the low-latency, high-bandwidth, and location-aware requirements of modern IoT applications [34]. Fog computing bridges this gap by placing processing, storage, and decision-making closer to the data source, reducing response time and alleviating network congestion. By dynamically allocating workloads, this hierarchical architecture facilitates the effective management of heterogeneous data streams from IoT devices, including sensors, actuators, and wearables. Long-term or computationally demanding analytics are offloaded to the cloud, while latency-sensitive tasks are managed at the fog layer. Such a paradigm is critical in real-world scenarios like smart cities, industrial automation, e-health, and precision agriculture, where intelligent, context-aware load balancing between fog and cloud nodes ensures scalability, reliability, and optimal QoS in IoT-enabled systems.

A. IoT Architecture and Components

The IoT is a huge network of physically linked objects that gather, send, and exchange data, including sensors, actuators, RFID tags, and embedded systems [35]. The standard IoT architecture is typically organized into three layers:

- **Perception Layer:** This layer includes sensors and actuators that capture real-world phenomena and transmit raw data.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

- **Network Layer:** Responsible for data transmission through wireless or wired communication protocols such as Wi-Fi, LTE, ZigBee, or 5 G. This layer ensures secure routing and connectivity between edge and core components.
- **Application Layer:** This layer delivers services to end-users by processing data in sectors like healthcare, transportation, smart cities, and industrial automation.

For intermediate data processing and protocol translation, edge devices, gateways, and middleware are also essential.

B. Role of Fog and Cloud in Handling IoT Workloads

Fog and cloud computing integration offers a hierarchical paradigm to satisfy IoT system requirements:

- **Fog Computing:** Fog nodes, which are positioned closer to the data source, manage response, preprocessing, and real-time analytics. This facilitates location-aware services, lowers latency, and eases congestion in the core network.
- **Cloud Computing:** Cloud data centers offer vast computational resources for batch processing, deep analytics, and historical data storage. They are instrumental for training machine learning models, long-term insights, and centralised control.
- **Hybrid Approach:** Dynamic load distribution is made possible via a fog-cloud continuum. While computationally demanding or delay-tolerant jobs are assigned to the cloud, time-sensitive tasks are offloaded to fog nodes.

In IoT networks, this dual-layer strategy greatly enhances scalability, energy efficiency, and performance.

C. Real-World IoT Scenarios Requiring Load Balancing

In practical IoT deployments, maintaining performance, dependability, and responsiveness requires efficient load balancing across fog and cloud infrastructures. Context-aware task allocation is essential in a variety of application fields, as demonstrated by the following illustrative scenarios:

- **Smart Traffic Management:** Real-time data from surveillance cameras and road sensors is essential to IoT-enabled traffic solutions. Fog nodes located near intersections or roadside units process data locally to trigger immediate control actions such as traffic signal adjustments, congestion alerts, or accident notifications.
- **Industrial IoT:** High-frequency telemetry data is continually produced by industrial sensors and machines in smart manufacturing environments. Real-time defect prevention is made possible by fog computing, which enables localized processing for time-sensitive activities like anomaly detection and equipment monitoring. Conversely, cloud systems are used for predictive maintenance scheduling and thorough analytics [36], and production optimization using trends in long-term data.
- **Remote Health Monitoring:** Medical devices that are worn or implanted transmit physiological data to adjacent fog gateways for edge-based diagnostics, warning generating, and instant filtering. This guarantees quick action in urgent circumstances. Large-scale patient data storage, AI-powered health trend analysis, and model training for illness prediction and individualized treatment planning are all handled concurrently by the cloud.
- **Smart Agriculture:** In precision agriculture, environmental sensors distributed across large farmlands collect data on soil moisture, temperature, and humidity. Fog nodes enable localized decision-making for irrigation scheduling, pesticide application, and microclimate control.

These scenarios emphasize the need for dynamic and adaptive load balancing mechanisms that account for latency sensitivity, data volume, energy efficiency, and bandwidth availability. Optimal workload distribution across the fog-cloud continuum enhances system performance, supports scalability, and ensures QoS (Quality of Service) in increasingly complex and data-intensive IoT ecosystems.

Key Challenges in Load Balancing for Fog-Cloud IoT Environments

These are the major issues in load balancing of Fog-Cloud architectures that exist in an IoT context and should be listed in point form:

- **Heterogeneity of Devices and Resources:** The uniform distribution of the load is made easier by the heterogeneity of fog and IoT devices, which results from variations in compute capacity, energy limits, and protocol.
- **Unpredictable and Dynamic Workloads:** Workloads involving IoT data flow may be highly unpredictable and subjected to substantial variance in volume/frequency, which is hard to balance at real time**.**
- **Latency and Real-Time Constraints:** Latency of time-sensitive applications is also a key challenge, and one which becomes increasingly difficult as application deployment gets more mobile or distributed.
- **Resource Constraints at the Edge:** The energy, memory, and computing power of fog and edge devices are not necessarily set up to provide a large loading capacity.
- **Scalability Problem:** The larger an IoT network [37], the more challenging it is to ensure that the load is distributed in a balanced manner without the underlying bottlenecks.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

- **Fault Tolerance and Reliability:** Effective systems must be able to deal with nodes failing, going off grid/connection or getting overloaded, and being dynamically reliable.
- **Security and Privacy Issues:** It is possible that decisions made in regards to load balancing contain sensitive data, thus concerns on safe data processing and privacy are pronounced.
- **Phasing of Fog and Clouds Layer:** At distributed layers, coordination and various types of decisions are difficult to make [33], owing to varying performance qualities and objectives.

V. LITERATURE REVIEW

With a focus on key architectural elements, strategies, domain-specific implementations, assessment frameworks, and technological developments, this section reviews the literature on load balancing tactics in fog-cloud architectures for IoT integration.

B and Dhas (2025) explores various models and their applications in high-performance cloud environments, emphasizing the advantages of combining deep learning techniques with cloud computing resources for load balancing. This article offers a critical assessment of the methods now employed to enhance cloud computing efficiency and handle load-balancing problems. The optimization methods for adjusting network parameters to enhance cloud computing performance and control load-balancing are also examined, in addition to deep learning techniques. This paper reviews a number of performance criteria that were employed in previous studies to assess the effectiveness of the model. This dual-layer approach significantly improves performance, scalability, and energy efficiency in IoT networks [38]. Chowdhury and Katangur (2025) presents a load balancing method designed to guarantee a fair allocation of workloads across nodes, based on threshold concepts. The algorithm's primary goal is to prevent virtual machines (VMs) in the cloud from becoming overloaded with work or from becoming idle as a result of insufficient task allocation while there are active activities running. The algorithm's threshold settings ensure that virtual machines are deployed sparingly, preventing both job overload and idle states brought on by inadequate work allocation. According to simulation results, our threshold-based algorithm significantly improves task/request response times and data processing times in datacenters, surpassing existing algorithms like Round Robin, First Come First Serve, and the Equally Spread Current Execution Load Balancing algorithm [39].

Van Anh and Nguyen (2025) proposed approach introduces a dynamic priority assignment mechanism that leverages real-time patient data for swift processing of critical events and an adaptive resource allocation strategy that optimizes performance under varying workloads. The framework's superiority is demonstrated through simulations and real-world case studies, which show a 25% improvement in resource usage and a 30% decrease in average reaction time for key events when compared to state-of-the-art techniques. Contributions include: a novel M/M/C/K priority queue model integrated with edge-fog-cloud architecture; dynamic priority assignment and adaptive resource allocation strategies; and comprehensive evaluation through simulations and case studies [40]. Nayak et al. (2024) A PSO based approach to load balancing in cloud-fog systems is introduced, addressing resource management challenges in dynamic and dispersed environments. Effective load balancing techniques are required due to the complex interactions between fog and cloud computing in order to maximize resource usage and improve system performance. The proposed PSO-based load balancing architecture uses the swarm intelligence concept to dynamically distribute tasks across fog nodes and the cloud to minimize processing delays and improve overall system performance. Extensive simulations and performance evaluations demonstrate the PSO algorithm's effectiveness in achieving load balance and emphasize its adaptability to fluctuating workloads. Comparing the PSO-based load balancing strategy to conventional methods, the results show a considerable improvement in response times and resource utilization [41].

Kumar, Agrawal and Tapaswi (2024) delves into existing partitioning concepts and load balancing models within the realm of public clouds. It also advances the subject of load balancing in cloud computing by introducing a thorough generalized model that is intended to manage a range of network load conditions with ease. In order to effectively distribute workloads among cloud infrastructure nodes and maximize overall performance, load balancing techniques are essential. By ensuring that load factors are managed effectively, these techniques significantly enhance system efficiency. This is particularly relevant in data centers, which serve as fundamental elements within cloud computing frameworks. Effective utilization of data centers can lead to marked improvements in system performance [42]. Vachhani et al. (2024) to achieve optimal resource efficiency and scalability quickly by distributing loads effectively among virtual machines. There is a strong desire for the exploration and development of new algorithms in this field to advance technology and make progress in resource allocation applications in cloud computing. This study investigates innovative techniques for handling resources in cloud computing, specifically emphasizing enhancing efficiency by utilizing load balancing and dynamic predictive

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

resource allocation. This study involves using machine learning models like LSTM networks and ARIMA within a framework to allocate resources and predict workloads dynamically [43]. Key studies on load balancing for fog-cloud architecture in IOT settings are included in Table II, which also includes highlights of the study's methods, primary conclusions, current issues, and potential future research areas.

TABLE II. SUMMARY OF RELATED LITERATURE ON FOG-CLOUD ARCHITECTURES AND LOAD BALANCING IN IOT ENVIRONMENTS

Reference	Study Focus	Methods/ Approaches	Key Findings	Limitations / Future Work
B et.al. (2025)	Analytical review of DL techniques and optimization strategies for load balancing in cloud computing	Review of various deep learning (DL) models and optimization approaches for network parameter tuning	Emphasizes advantages of integrating DL with cloud resources; surveys multiple performance metrics and existing methods	Lacks empirical evaluation; future work suggested in bridging research gaps and exploring new DL-based solutions
Chowdhury, et.al. (2025)	Development of threshold-based load balancing algorithm for equitable VM task allocation	Threshold-based algorithm compared against FCFS, Round Robin, and ESCE	Improved response time and processing efficiency within data centers; avoids VM overload and idle states	Specific thresholds may need tuning for varying workloads; future work can focus on adaptive threshold models
Van Anh, et.al. (2025)	Load balancing for real-time healthcare systems using edge-fog-cloud and dynamic priorities	M/M/C/K priority queue model; dynamic priority assignment; adaptive resource allocation	Achieved 30% faster response time and 25% higher resource utilization; strong performance in critical healthcare use-cases	Limited to healthcare scenarios; future work could generalize to other real-time applications
Nayak et al. (2024)	PSO-based dynamic load balancing in cloud-fog hybrid systems	Particle Swarm Optimization (PSO) algorithm for task distribution	Demonstrates improved response time and resource utilization over traditional methods	Requires tuning of PSO parameters; future research could explore hybrid optimization with AI/ML integration
Kumar, et.al. (2024)	Generalized model for load balancing in public cloud systems	Conceptual framework incorporating existing partitioning and load balancing models	Highlights importance of efficient data center load handling; proposes scalable generalized approach	Conceptual model without simulation; future work includes implementation and validation on real cloud datasets
Vachhani et al. (2024)	Predictive resource allocation using ML for efficient cloud load balancing	Machine learning models: LSTM and ARIMA used for workload prediction and resource management	Demonstrates effectiveness of predictive models in dynamic load distribution and resource optimization	Lacks comparative performance evaluation with state-of-the-art; future work may explore ensemble or real-time learning methods

VI. CONCLUSION AND FUTURE WORK

Fog and cloud computing are an emerging architectural model that appears to meet the scalability, low latency and high availability needs of contemporary IoT ecosystems. In this paper, the principles of fog-cloud architectures have been surveyed, paying special attention to their synergistic potential in distributed data processing, outlined load balance as an important aspect of resource optimization, resilience and response time of a system. An extensive load balancing strategy taxonomy based on static/dynamic as well as topology-based approaches, AI techniques was provided, and the considerations related to important evaluation metrics and factors are brought up. The combative and diverse IoT environments still create major challenges though a number of strategies display efficiency in particular settings. Future studies in fog-cloud load balancing ought to concentrate on coming up with versatile, situational conscious algorithms that could work in the exceedingly dynamic and constrained assets at IOT levels. The level of AI integration and ML methodology is highly beneficial in the prediction of the workload patterns and real-time task optimization. Moreover, in subsequent research, the efforts are to be made to cover security-aware load balancing, particularly those applications that are sensitive, such as healthcare and smart infrastructure. It requires some developments to improve interoperability, minimize energy use, and mobility, as well as Vault Tolerance, in various IoT solutions. It will be necessary to make sure

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

that the actual effectiveness of offered solutions is confirmed by using standardized benchmarking philosophies and testing them in real large-scale-scenarios.

REFERENCES

[1] C. Patel, "A Review of Multi-Channel CRM Strategies Using Big Data and Cloud Integration," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 8, no. 1, pp. 577–588, 2022.
[2] A. Parupalli and H. Kali, "An In-Depth Review of Cost Optimization Tactics in Multi-Cloud Frameworks," Int. J. Adv. Res. Sci. Commun. Technol., vol. 3, no. 5, pp. 1043–1052, Jun. 2023, doi: 10.48175/IJARSCT-11937Q.
[3] S. K. Chintagunta and S. Amrale, "Enhancing Cloud Database Security Through Intelligent Threat Detection and Risk Mitigation," TIJER--Int. Res. J, vol. 9, no. 10, pp. 49–55, 2022.
[4] L. Golightly, V. Chang, Q. A. Xu, X. Gao, and B. S. C. Liu, "Adoption of cloud computing as innovation in the organization," Int. J. Eng. Bus. Manag., vol. 14, Nov. 2022, doi: 10.1177/18479790221093992.
[5] A. Goyal, "Enhancing Engineering Project Efficiency through Cross-Functional Collaboration and IoT Integration," Int. J. Res. Anal. Rev., vol. 8, no. 4, pp. 396–402, 2021.
[6] A. R. Toorpu, S. K. Vududala, A. Nerella, and B. P. Madupati, "Hybrid AI Models for Privacy-Preserving Big Data Analytics in Distributed Environments," in 2025 Global Conference in Emerging Technology (GINOTECH), IEEE, May 2025, pp. 1–8. doi: 10.1109/GINOTECH63460.2025.11076666.
[7] G. Maddali, "An Efficient Bio-Inspired Optimization Framework for Scalable Task Scheduling in Cloud Computing Environments," Int. J. Curr. Eng. Technol., vol. 15, no. 3, pp. 229–238, 2025.
[8] H. Rashid Abdulqadir et al., "A Study of Moving from Cloud Computing to Fog Computing," Qubahan Acad. J., vol. 1, no. 2, pp. 60–70, Apr. 2021, doi: 10.48161/qaj.v1n2a49.
[9] S. Singamsetty, "An Intelligent Framework for Secure and Fair Cloud Resource Distribution," in 2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA), IEEE, Oct. 2025, pp. 686–690. doi: 10.1109/ICIDCA66325.2025.11280502.
[10] A. Saoud and A. Recioui, "Hybrid algorithm for cloud-fog system based load balancing in smart grids," Bull. Electr. Eng. Informatics, vol. 11, no. 1, pp. 477–487, Feb. 2022, doi: 10.11591/eei.v11i1.3450.
[11] K. Shukla, N. Patel, and H. Mistry, "Securing the Cloud: Strategies and Innovations in Network Security for Modern Computing Environments," Int. Res. J. Eng. Technol., vol. 11, no. 04, 2024.
[12] S. P. Bheri and G. Modalavalasa, "Advancements in Cloud Computing for Scalable Web Development: Security Challenges and Performance Optimization," JCT Publ., vol. 13, no. 12, pp. 01–07, 2024.
[13] C. Martín, D. Garrido, L. Llopis, B. Rubio, and M. Díaz, "Facilitating the monitoring and management of structural health in civil infrastructures with an Edge/Fog/Cloud architecture," Comput. Stand. Interfaces, vol. 81, p. 103600, Apr. 2022, doi: 10.1016/j.csi.2021.103600.
[14] B. Y. Akhil Reddy Bairi, Venkatesha Prabhu Rambabu, "AI-Enhanced Test Prioritization in Continuous Integration for SaaS Platforms," Am. J. Auton. Syst. Robot. Eng., vol. 2, pp. 110–145, 2022.
[15] K. M. R. Seetharaman, "Internet of Things (IoT) Applications in SAP: A Survey of Trends, Challenges, and Opportunities," Int. J. Adv. Res. Sci. Commun. Technol., vol. 3, no. 2, pp. 499–508, Mar. 2021, doi: 10.48175/IJARSCT-6268B.
[16] H. F. Atlam, R. J. Walters, and G. B. Wills, "Fog Computing and the Internet of Things: A Review," Big Data Cogn. Comput., vol. 2, no. 2, p. 10, Apr. 2018, doi: 10.3390/bdcc2020010.
[17] T. Shah, "Cloud-Based Data Warehousing for Marketing Agility: Lessons from FinTech Migrations to Snowflake and AWS," Int. J. Adv. Res. Sci. Commun. Technol., vol. 4, no. 4, pp. 1–11, 2024.
[18] P. Hu, S. Dhelim, H. Ning, and T. Qiu, "Survey on fog computing: architecture, key technologies, applications and open issues," J. Netw. Comput. Appl., vol. 98, pp. 27–42, Nov. 2017, doi: 10.1016/j.jnca.2017.09.002.
[19] A. Gupta, "What Is The Right Security Posture? A Perspective on Cloud Computing Security Threats and Risk Assessment," Int. J. Emerg. Res. Eng. Technol., vol. 4, no. 4, 2023, doi: 10.63282/3050-922X.IJERET-V4I4P112.
[20] N. R. Barot, "Transparency-Driven Operational Intelligence: A New Data Governance Model for High-Risk Industrial Automation," J. Inf. Syst. Eng. Manag., vol. 10, no. 63s, pp. 1019–1028, Dec. 2025, doi: 10.52783/jisem.v10i63s.13975.
[21] V. Prajapati, "Role of Identity and Access Management in Zero Trust Architecture for Cloud Security: Challenges and Solutions," Int. J. Adv. Res. Sci. Commun. Technol., vol. 5, no. 3, pp. 6–18, Mar. 2025, doi: 10.48175/IJARSCT-23902.
[22] C.-Y. Weng, C.-T. Li, C.-L. Chen, C.-C. Lee, and Y.-Y. Deng, "A Lightweight Anonymous Authentication and Secure Communication Scheme for Fog Computing Services," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3123234.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 2, March - April 2026||

DOI:10.15662/IJRPETM.2026.0902015

[23] B. Madupati, M. M. Mohammed, L. Upadhyay, D. P. Guda, K. Kaushik, and M. Soni, "Integrating Artificial Intelligence with Cybersecurity for Resilient Wireless Communication Against Advanced Threats," in 2025 International Conference on Artificial Intelligence and Machine Vision (AIMV), IEEE, Aug. 2025, pp. 1–5. doi: 10.1109/AIMV66517.2025.11203666.
[24] M. Vijarania, S. Gupta, A. Agrawal, M. O. Adigun, S. A. Ajagbe, and J. B. Awotunde, "Energy Efficient Load-Balancing Mechanism in Integrated IoT–Fog–Cloud Environment," Electronics, vol. 12, no. 11, p. 2543, Jun. 2023, doi: 10.3390/electronics12112543.
[25] S. Kabade and A. Sharma, "Intelligent Automation in Pension Service Purchases with AI and Cloud Integration for Operational Excellence," Int. J. Adv. Res. Sci. Commun. Technol., pp. 725–735, Dec. 2024, doi: 10.48175/IJARSCT-14100J.
[26] B. Pourghebleh and V. Hayyolalam, "A comprehensive and systematic review of the load balancing mechanisms in the Internet of Things," Cluster Comput., vol. 23, no. 2, pp. 641–661, Jun. 2020, doi: 10.1007/s10586-019-02950-0.
[27] S. Bhat, S. R. Sirikonda, V. Katoch, and R. Jain, "Carbon-Kube: A Kubernetes-Native Framework for Multi-Objective Carbon-Aware Scheduling of Big Data Pipelines," in 2026 9th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), IEEE, Feb. 2026, pp. 1–6. doi: 10.1109/IEMENTech202669403.2026.11434192.
[28] A. Kushwaha, P. Pathak, and S. Gupta, "Review of optimize load balancing algorithms in cloud," Int. J. Distrib. Cloud Comput., vol. 4, no. 2, pp. 1–9, 2016.
[29] H. Ravilla, J. Yarra, and S. Dilip, "Role of SOQL and Database Optimization in Large-Scale Salesforce Implementations," Int. J. Eng. Archit., vol. 3, no. 1, pp. 13–31, Feb. 2026, doi: 10.58425/ijea.v3i1.481.
[30] M. A. Shahid, N. Islam, M. M. Alam, M. M. Su'ud, and S. Musa, "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach," IEEE Access, vol. 8, pp. 130500–130526, 2020, doi: 10.1109/ACCESS.2020.3009184.
[31] H. S. Chandu, "A Review of IoT-Based Home Security Solutions: Focusing on Arduino Applications," TIJER – Int. Res. J., vol. 11, no. 10, pp. a391–a396, 2024.
[32] H. B. Dama, "A Survey of MySQL Database Administration Techniques and Best Practices," ESP J. Eng. Technol. Adv., vol. 6, no. 1, pp. 89–98, 2026.
[33] R. Gao, X. Xie, and Q. Guo, "K-TAHP: A Kubernetes Load Balancing Strategy Base on TOPSIS+AHP," IEEE Access, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3313643.
[34] S. Garg, "Next-Gen Smart City Operations with AIOps & IoT : A Comprehensive look at Optimizing Urban Infrastructure," J. Adv. Dev. Res., vol. 12, no. 1, 2021.
[35] H. S. Chandu, "Enhancing Manufacturing Efficiency: Predictive Maintenance Models Utilizing IoT Sensor Data," IJSART, vol. 10, no. 9, 2024.
[36] S. Garg, "AI/ML Driven Proactive Performance Monitoring, Resource Allocation and Effective Cost Management in SAAS Operations," Int. J. Core Eng. Manag., vol. 6, no. 6, pp. 263–273, 2019.
[37] J. Thomas, "The Effect and Challenges of the Internet of Things (IoT) on the Management of Supply Chains," Int. J. Res. Anal. Rev., vol. 8, no. 3, pp. 874–878, 2021.
[38] D. B and A. S. Dhas, "A Systematic Literature Review on Dynamic Load Balancing Techniques in Cloud Computing Environment: Techniques, Challenges, and Future Prospects," in 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS), 2025, pp. 1440–1446. doi: 10.1109/ICMLAS64557.2025.10968023.
[39] S. Chowdhury and A. Katangur, "Optimizing Cloud Computing Performance Through Integration of a Threshold-Based Load Balancing Algorithm With Multiple Service Broker Policies," IEEE Trans. Cloud Comput., vol. 13, no. 2, Apr. 2025, doi: 10.1109/TCC.2025.3563848.
[40] D. Van Anh and V.-H. Nguyen, "Leveraging Priority Queueing in IoT-Edge-Fog-Cloud Infrastructures for Efficient Healthcare Monitoring," IEEE Access, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3565679.
[41] A. Nayak, S. S. Tripathy, S. Bebortta, and B. Tripathy, "An Intelligent Study Towards Nature-Inspired Load Balancing Framework for Fog-Cloud Environments," in 2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE), IEEE, Feb. 2024. doi: 10.1109/ICWITE59797.2024.10503194.
[42] R. Kumar, N. Agrawal, and S. Tapaswi, "Improved Load Balancing and Partitioning Model for Cloud Networks," in 2024 OITS International Conference on Information Technology (OCIT), IEEE, Dec. 2024. doi: 10.1109/OCIT65031.2024.00144.
[43] M. Vachhani, Z. Patel, D. Garg, K. Patel, and M. Patel, "Enhancing Cloud Computing Efficiency: Dynamic and Predictive Resource Allocation and Load Balancing Strategies," in 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), IEEE, Dec. 2024. doi: 10.1109/ICICNIS64247.2024.10823380.