International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume-11, Issue-3, June 2024
https://doi.org/10.55524/ijirem.2024.11.3.10
Article ID IJIR2548, Pages 65-70
www.ijirem.org

Quality of Service Based on Scheduling in Cloud/Fog Environment

Lakshya Mall[1], Priyanka Vashisht[2], and Ashima Narang[3]

[1,2,3] Amity School of Engineering and Technology (ASET), Amity University, Gurugram, Haryana, India

Correspondence should be addressed to Lakshya Mall lakshyaofficial123@gmail.com

Received: 4 May 2024 Revised: 18 May 2024 Accepted: 1 June 2024

ABSTRACT- In the area of fog and cloud computing, efficient work scheduling is critical for optimising resource allocation and raising Quality of Service (QOS) levels. Research in this area has resulted in a variety of ways for dealing with this problem. One line of research focuses on lowering data latency and improving QOS in fog-cloud systems by precise job scheduling. This requires intelligently dividing jobs across fog and cloud resources to reduce data processing and delivery delays, ultimately improving the overall user experience. Furthermore, the investigation encompasses novel methodologies such as hybrid evolutionary algorithms and heuristic strategies. These approaches seek to create a compromise between competing objectives, such as cost and energy efficiency, while also achieving QOS criteria. Researchers aim to develop scheduling systems that optimise resource utilisation and reduce operating costs without sacrificing service quality by applying evolutionary concepts or combining heuristic methods. Furthermore, ongoing discussions focus on improving resource scheduling approaches to achieve optimal QOS results. This entails continual evaluation and adaption of scheduling algorithms to changing environmental conditions, workload fluctuations, and user requests. Furthermore, developing methodologies such as the CODA approach provide intriguing avenues for efficient job scheduling in fog computing settings, with the potential to revolutionise future resource allocation and QOS enhancement efforts.

KEYWORDS- Quality of Service (QOS) levels, CODA, Fog Computing, Cloud Computing

I. INTRODUCTION

Ensuring good Quality of Service (QOS) has become a major concern for both service providers and companies in the quickly evolving cloud and fog computing ecosystem. Because QoS-based scheduling assigns resources and prioritizes tasks based on user expectations and performance goals, it is essential in resolving this issue. Historically, cloud computing has been depended upon to provide computer services over the internet because of its scalable architecture and centralized data centres. Fog computing, on the other hand, brings new possibilities and problems for QOS optimization as it spreads cloud capabilities to the edge of the network.

Numerous facets of QOS-based scheduling in cloud and fog settings have been clarified by recent research. Examining resource scheduling methods in fog computing settings offers insights into various QOS management strategies [1]. It has been suggested that hybrid heuristic algorithms improve QOS by minimizing data latency and maximizing work scheduling [2]. Furthermore, the significance of cost- and energy-efficient solutions is shown by the effect of job scheduling on QOS and financial cost in collaborative fog-cloud systems [3]. Examining fog computing job scheduling algorithms further emphasises how important node availability and processing power are for effective resource allocation [4]. Finally, it has been determined that one way to raise QOS in fog computing systems is through effective workflow scheduling in cloud settings [5]. These studies collectively underscore the importance of QOS-based scheduling in optimizing resource utilization and delivering satisfactory user experiences in cloud and fog environments. As shown in Figure 1.

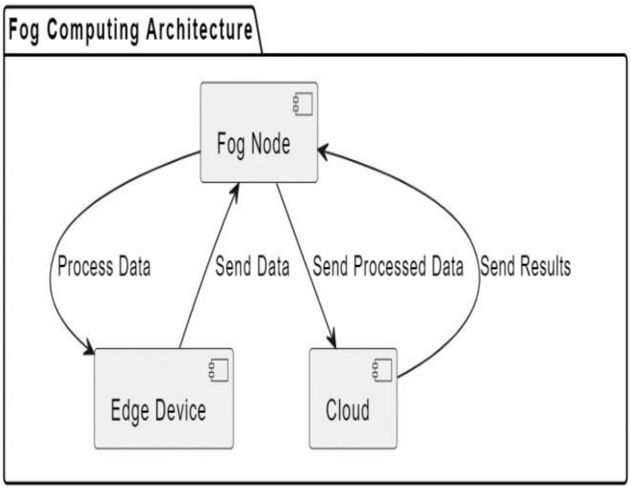

Figure 1: Fog Computing Architecture

Table 1: Analysis of the Related Works

Paper Title	Methodology	Key Findings and Contributions
Resource Scheduling Techniques for Optimal Quality of Service in Fog Computing Environment: A Review	Systematic Literature Review	Conducted a systematic review of existing literature on resource scheduling techniques in fog computing environments. Identified and analysed various approaches for managing Quality of Service (QoS) in fog environments, including heuristic-based, optimization-based, and machine learning-based techniques. Provided insights into trends, challenges, and best practices in QoS-based scheduling [1].
Hybrid heuristic algorithm for cost-efficient QoS aware task scheduling.	Algorithm Development and Evaluation	Proposed a novel hybrid heuristic algorithm for task scheduling in fog-cloud environments, aiming to optimize resource allocation while ensuring Quality of Service (QoS) requirements are met. Combined multiple heuristic techniques, such as genetic algorithms and simulated annealing, to enhance scheduling efficiency and effectiveness. Evaluated the algorithm through simulations or experiments, demonstrating its capability to improve QoS and reduce data delay [2].
Cost and Energy Efficient Task Scheduling in a Cloud-Fog Computing Environment.	Quantitative Research	Conducted a quantitative analysis to examine the impact of task scheduling on both Quality of Service (QoS) and financial cost in collaborative fog-cloud systems. Gathered empirical data on resource usage, QoS metrics, and operational expenses from real-world fog-cloud deployments. Employed statistical analysis techniques, such as regression analysis or cost-benefit analysis, to identify factors influencing cost and QoS. Identified cost and energy-efficient scheduling strategies essential for optimizing resource utilization while minimizing operational expenses [3].
An Effective analysis on various task scheduling algorithms in fog computing.	Comparative Analysis	Conducted a comprehensive comparative analysis of different task scheduling algorithms in fog computing environments. Implemented multiple scheduling algorithms and evaluated their performance using simulation or experimentation. Compared the effectiveness of algorithms in terms of task completion time, resource utilization, and Quality of Service (QoS) satisfaction. Provided insights into the suitability of different algorithms under various scenarios and workload conditions [4].
Improving Quality of Services of Fog Computing Through Efficient Workflow Scheduling.	Qualitative Research	Employed qualitative research methods, such as interviews or surveys, to explore the role of workflow scheduling in improving Quality of Service (QoS) in fog computing systems. Investigated current workflow management practices, challenges, and opportunities for improvement. Identified key factors influencing QoS in fog environments and proposed recommendations or new workflow scheduling techniques to enhance system performance and user satisfaction [5].

II. THE SIGNIFICANCE OF QOS-BASED SCHEDULING

Scheduling algorithms are critical in maintaining Quality of Service (QoS) provision in computing systems because they efficiently manage resources and orchestrate jobs to satisfy performance goals and user expectations. These algorithms allocate computational resources such as CPU, memory, and network bandwidth to competing activities or applications depending on predetermined criteria such as priority, deadline, or resource restrictions. Scheduling algorithms help maintain target QoS levels by optimizing resource utilisation and minimizing resource contention. Different scheduling methods, such as First-Come-First-Served (FCFS), Round Robin, and Earliest Deadline First (EDF), are intended to properly priorities and schedule tasks based on their characteristics and application-specific QoS needs [6].

QoS has a substantial impact on user experience and application performance since it influences many aspects of system behaviour and responsiveness. For example, low latency promotes faster response times and shorter wait times for user interactions, resulting in a more seamless and responsive user experience. High throughput facilitates efficient data processing and transfer, allowing for the timely delivery of services and content to users. Reliability provides continuous performance and service availability while minimizing disruptions and downtime, which can reduce user happiness and trust in the system. Furthermore, QoS has a direct impact on application performance and functionality, influencing parameters such as reaction time, data consistency, and overall system efficiency. Prioritizing QoS considerations in system design and administration can boost user pleasure, improve application performance, and achieve business objectives more effectively [9].

QoS metrics serve as quantitative measures to assess and evaluate the performance and reliability of computing systems and services. Examples of QoS metrics include:

A. Latency

The time delay between the initiation and completion of a task or communication process, representing the responsiveness of the system. Low latency is essential for real-time applications, such as video streaming, online gaming, and telecommunication services.

B. Throughput

The rate at which data is processed, transmitted, or delivered within a given timeframe, indicating the system's capacity to handle workload demands. High throughput ensures efficient data transfer and processing, supporting scalable and high-performance applications.

C. Reliability:

The probability that a system or service will perform its intended functions without failures or errors over a specified period, reflecting the system's stability and resilience. Reliable systems minimize service disruptions and data loss, ensuring consistent performance and user satisfaction.

D. Availability

The percentage of time that a system or service is operational and accessible to users, accounting for planned maintenance, unplanned downtime, and system failures. High availability ensures continuous access to services, maximizing uptime and productivity for users and organizations. Other QoS metrics may include security, scalability, response time, jitter, and resource utilization, depending on the specific requirements and objectives of the system or application. By monitoring and optimizing these QoS metrics, organizations can ensure the delivery of high-quality services, meet user expectations, and maintain a competitive edge in the market. As shown in Fig 2, QoS is used to guarantee the performance of the network and make sure the most critical traffic is prioritized in using the bandwidth. It usually gives a low latency and favorable user experience an essential status

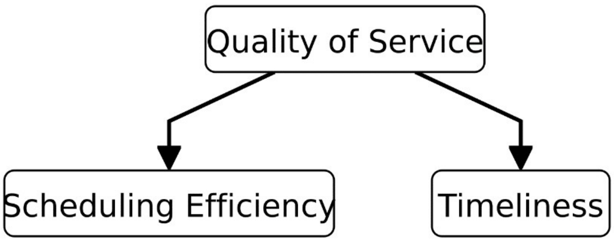

Figure 2: Importance of QOS

III. NAVIGATING CHALLENGES AND KEY CONSIDERATIONS

Addressing scalability issues in QoS-based scheduling involves designing algorithms and mechanisms that can efficiently handle increasing workloads and growing system complexity without compromising performance or QoS guarantees. Scalability challenges arise due to the dynamic nature of cloud and fog environments, where the number of tasks, users, and resources may vary unpredictably over time [10].

One approach to address scalability is to design scheduling algorithms that can dynamically adjust resource allocation and task prioritization based on system load and demand fluctuations. For example, adaptive scheduling algorithms may automatically scale resources up or down in response to changes in workload intensity, ensuring that QoS objectives are met while minimizing resource wastage.

Another strategy involves leveraging distributed and parallel processing techniques to distribute scheduling decisions across multiple nodes or components in the system. By decentralizing scheduling logic and workload management, distributed scheduling architectures can improve scalability by distributing the computational burden and reducing bottlenecks.

Furthermore, the use of efficient data structures, caching mechanisms, and optimization techniques can help improve the efficiency and scalability of QoS-based scheduling algorithms by reducing computational overhead and latency associated with task scheduling and resource management operations.

Overall, addressing scalability issues in QoS-based scheduling requires a combination of adaptive algorithms, distributed architectures, and optimization techniques tailored to the specific characteristics and requirements of cloud and fog computing environments.

The trade-offs between QoS and resource utilization arise from the inherent tension between providing high-quality services and maximizing resource efficiency. While QoS aims to ensure optimal performance and user satisfaction by meeting specified performance targets, resource utilization focuses on maximizing the efficient use of available resources to minimize costs and improve system efficiency [13].

One trade-off involves allocating resources to prioritize QoS-sensitive tasks or applications at the expense of resource utilization. For example, dedicating excess resources to critical applications may lead to underutilization and increased operational costs, but it ensures that QoS requirements are met and performance objectives are achieved.

Conversely, optimizing resource utilization by consolidating workloads and sharing resources among multiple applications can lead to improved efficiency and cost savings but may result in degraded QoS for latency-sensitive or high-priority tasks. Balancing QoS and resource utilization requires careful consideration of workload characteristics, system requirements, and performance objectives to determine the optimal resource allocation strategy.

Moreover, trade-offs may arise from competing objectives and constraints, such as cost, energy consumption, and scalability. For example, implementing resource-intensive QoS guarantees may require additional hardware resources or specialized infrastructure, leading to increased costs and complexity [11].

Ultimately, achieving a balance between QoS and resource utilization involves trade-offs that require careful consideration of system requirements, workload characteristics, and performance objectives to optimize resource allocation and ensure satisfactory user experiences while maximizing resource efficiency.

Heterogeneity and dynamism in cloud and fog environments pose significant challenges to QoS management, affecting resource allocation, task scheduling, and system performance. Heterogeneity refers to the diversity of resources, devices, and infrastructure components in the system, including differences in processing capabilities, memory capacities, and network bandwidth. Dynamism refers to the variability and unpredictability of system conditions, such as changes in workload intensity, network conditions, and resource availability over time [12].

The impact of heterogeneity and dynamism on QoS management includes:

- **Resource Variability:** Heterogeneous environments may have varying levels of resource availability and performance characteristics, making it challenging to allocate resources effectively and meet QoS requirements consistently across different devices and platforms.
- **Load Balancing:** Dynamically balancing workload distribution and resource utilization becomes more challenging in heterogeneous and dynamic environments, requiring adaptive scheduling algorithms that can adjust resource allocation based on real-time system conditions and workload demands.
- **Interoperability:** Integrating diverse devices, protocols, and technologies in fog environments requires interoperability mechanisms to ensure seamless communication and coordination between heterogeneous components, facilitating QoS-aware service delivery and management.
- **Scalability:** Heterogeneity and dynamism can impact system scalability by introducing complexity and overhead in resource management and scheduling operations, requiring scalable architectures and algorithms to handle increasing workloads and growing system complexity. As shown in Fig 3, Heterogeneity and dynamism in networks can lead to complex interactions and unpredictable performance. These factors require adaptive strategies to manage diverse device capabilities and varying traffic patterns effectively.

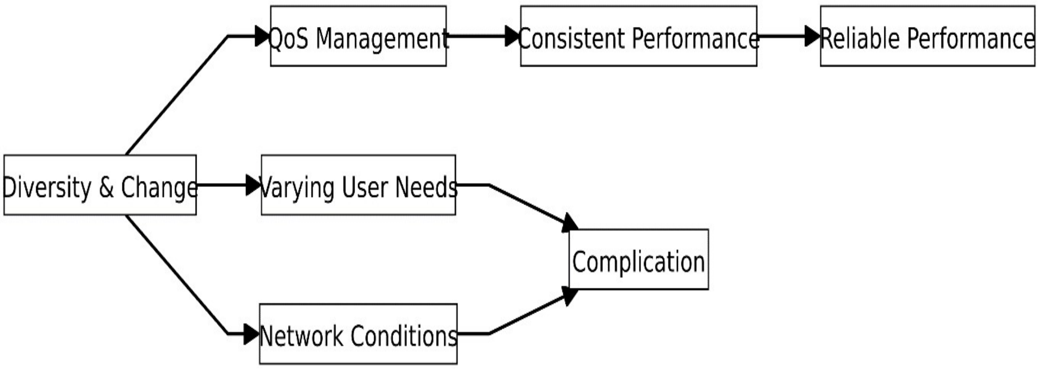

Figure 3: Effects of Heterogeneity and Dynamism

Addressing the impact of heterogeneity and dynamism on QoS management necessitates adaptive and flexible techniques that can dynamically adjust to changing system conditions and demands. This could include using machine learning algorithms, predictive analytics, and autonomic computing approaches to anticipate and respond to changes in workload, resource availability, and network circumstances, resulting in optimal QOS supply in heterogeneous and dynamic cloud and fog settings.

IV. SURVEY AND ANALYSIS

Quality of service (QoS)-based scheduling is critical for optimizing resource utilisation and providing good user experiences in cloud and fog computing environments. With the widespread adoption of distributed computing paradigms, efficient resource allocation while achieving QoS requirements has become critical. This essay examines the significance of QoS-based scheduling in cloud and fog settings, focusing on recent research in the subject.

Cloud computing, defined as the supply of computing services via the internet, has transformed how organizations and consumers access and manage data and applications. However, as data volume and complexity expand, traditional cloud designs confront issues such as latency, bandwidth limits, and data processing at the network's edge. Fog computing, an outgrowth of cloud computing, tackles these issues by decentralising computer resources and relocating data processing closer to the source. This proximity lowers latency, increases data privacy, and boosts overall system performance.

Effective task scheduling is crucial for optimizing resource utilization and meeting QoS requirements in cloud and fog environments. A review of resource scheduling techniques for optimal QoS in fog computing environments provides valuable insights into various approaches for QoS management [1]. Hybrid heuristic algorithms have been proposed to address the challenges of data delay reduction and enhancing QoS in fog-cloud environments [2]. These algorithms leverage heuristic techniques to efficiently allocate tasks based on QoS metrics such as latency, throughput, and reliability.

Furthermore, the effect of job scheduling on QoS and financial costs in collaborative fog-cloud systems has been investigated. Cost and energy-efficient task scheduling solutions are critical for maximizing resource utilisation while reducing operating costs [3]. Analyzing various job scheduling techniques in fog computing throws insight on aspects such as node availability and processing power, which are critical for efficient resource allocation [4].

Efficient workflow scheduling in cloud environments has also been recognized as a strategy for improving QoS in fog computing systems. Cloud-based apps that optimize workflow execution can improve overall system performance and user satisfaction in fog situations [5].

V. FUTURE DIRECTIONS AND RESEARCH OPPORTUNITIES

Identifying emerging trends in QoS-based scheduling entails recognizing new advances and directions that are influencing the sector. One rising trend is to include machine learning and artificial intelligence approaches into scheduling algorithms. Algorithms that use AI may adaptively learn from past data, analyze real-time system conditions, and make intelligent judgments to optimize resource allocation and improve QoS. Another emerging trend is the usage of edge computing and fog computing paradigms, which bring computer resources closer to end users and devices. QoS-based scheduling in edge and fog environments necessitates unique ways for successfully managing resource limits, latency-sensitive applications, and changing workload characteristics.

Proposing areas for further research and development in QoS-based scheduling involves identifying gaps in current knowledge and exploring opportunities for innovation and improvement. One area for research is the development of decentralized and self-adaptive scheduling algorithms that can operate autonomously in distributed environments without centralized coordination. Additionally, investigating the impact of emerging technologies such as quantum computing, blockchain, and 5G networks on QoS-based scheduling can uncover new challenges and opportunities for advancement. Furthermore, exploring interdisciplinary approaches that combine insights from fields like operations research, control theory, and human-computer interaction can lead to novel solutions for QoS management in complex and dynamic computing environments.

Discussing the potential impact of emerging technologies such as edge computing and AI on QoS-based scheduling involves considering how these innovations can reshape the landscape of resource management and service delivery. Edge computing enables low-latency processing and real-time analytics at the network edge, allowing for QoS-sensitive applications to be executed closer to end-users. This can lead to improved responsiveness, reduced network congestion, and enhanced user experiences. Similarly, AI techniques such as reinforcement learning and deep learning can optimize scheduling decisions by learning patterns from data, predicting future workload trends, and dynamically adapting resource allocations to meet QoS objectives. By harnessing the power of emerging technologies, QoS-based scheduling can become more adaptive, efficient, and responsive to the evolving needs of modern computing environments.

VI. CONCLUSION

Summarizing key findings from the review involves highlighting the main insights and conclusions drawn from the examination of QoS-based scheduling in cloud and fog environments. Key findings may include:

- The critical role of QoS-based scheduling algorithms in optimizing resource utilization, improving system performance, and ensuring satisfactory user experiences in cloud and fog computing.
- The effectiveness of various scheduling techniques, such as priority-based scheduling, deadline-driven scheduling, and adaptive resource allocation, in meeting QoS requirements and addressing scalability challenges.
- The impact of heterogeneity, dynamism, and resource constraints on QoS management in cloud and fog environments, highlighting the need for adaptive and flexible scheduling strategies.
- The emerging trends in QoS research, such as the integration of machine learning, edge computing, and AI techniques into scheduling algorithms, and their potential to reshape the future of QoS provisioning in distributed computing environments.

Reinforcing the importance of QoS-based scheduling in cloud and fog environments emphasizes the critical role of efficient resource management in ensuring optimal performance, reliability, and user satisfaction. QoS-based scheduling enables organizations to:

- Meet service-level agreements (SLAs) and performance targets by dynamically allocating resources and prioritizing critical tasks based on their QoS requirements.
- Enhance system scalability and flexibility by adapting resource allocations to changing workload patterns and environmental conditions.
- Improve resource utilization and efficiency by minimizing resource wastage and maximizing the utilization of available computing resources.
- Support emerging applications and technologies, such as IoT, edge computing, and real-time analytics, by providing low-latency, high-throughput, and reliable services at the network edge.
- Stay competitive in the marketplace by delivering high quality services, meeting user expectations, and differentiating themselves from competitors.

Providing closing remarks on the future of QoS research in the field offers insights into the direction and potential advancements in QoS-based scheduling. The future of QoS research may involve:

- Continued exploration of emerging technologies, such as edge computing, AI, and blockchain, and their impact on QoS provisioning in distributed computing environments.
- Development of decentralized and self-adaptive scheduling algorithms that can operate autonomously in dynamic and heterogeneous environments without centralized coordination.
- Integration of predictive analytics and proactive management techniques to anticipate and prevent QoS violations before they occur, improving system reliability and performance.
- Collaboration between academia and industry to address practical challenges and real-world applications of QoS-based scheduling in diverse domains, such as healthcare, finance, and smart cities.
- Emphasis on interdisciplinary research that combines insights from fields like computer science, operations research, and network engineering to develop holistic approaches to QoS management and optimization.

Overall, the future of QoS research holds promise for addressing the evolving needs and challenges of modern computing environments, driving innovation, and advancing the state-of-the-art in resource management and service delivery.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest between them and with any third party.

REFERENCES

1) G. Goel and Dr. R. Tiwari, "Resource Scheduling Techniques for Optimal Quality of Service in Fog Computing Environment: A Review," Wireless Personal Communications, vol. 131, pp. 1-24, 2023, doi: 10.1007/s11277-023-10421-4.
2) S. M. Hussain and G. R. Begh, "Hybrid heuristic algorithm for cost-efficient QoS aware task scheduling in fog–cloud environment," Journal of Computational Science, vol. 64, 2022, Art no. 101828, ISSN 1877-7503.
3) M. S. Kumar and G. R. Karri, "EEOA: Cost and Energy Efficient Task Scheduling in a Cloud-Fog Framework," Sensors, vol. 23, no. 5, p. 2445, 2023. [Online]. Available: https://doi.org/10.3390/s23052445.
4) P. Choppara and S. Mangalampalli, "An Effective analysis on various task scheduling algorithms in Fog computing," EAI Endorsed Transactions on Internet of Things, vol. 10, Dec. 2023.
5) G. Prasanth, M. Kumar, R. Al-Jawry, A. Ali, H. Sabah, and M. Al-Tahee, "Improving Quality of Services of Fog Computing Through Efficient Work Flow Scheduling," in 2023 International Conference on Advances in Computing, Communication and Information Technology (ICACITE), 2023, pp. 43-47, doi: 10.1109/ICACITE57410.2023.10183104.
6) P. Vashisht, R. Kumar, and A. Sharma, "Efficient dynamic replication algorithm using agent for data grid," The Scientific World Journal, vol. 2014, 2014.
7) P. Vashisht, A. Sharma, and R. Kumar, "Strategies for replica consistency in data grid–a comprehensive survey," Concurrency and Computation: Practice and Experience, vol. 29, no. 4, p. e3907, 2017.
8) P. Vashisht, V. Kumar, R. Kumar, and A. Sharma, "Optimization of replica consistency and conflict resolution in data grid environment," International Journal of Mathematical, Engineering and Management Sciences, vol. 4, no. 6, pp. 1420-1431, 2019.
9) P. Vashisht, V. Kumar, R. Kumar, and A. Sharma, "Optimizing replica creation using agents in data grids," in 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, UAE, 2019, pp. 542-547, doi: 10.1109/AICAI.2019.8701395.
10) P. Vashisht and V. Kumar, "Agent based optimized replica management in data grids," Investigación Operacional, vol. 41, no. 2, pp. 232-249, 2020.
11) P. Vashisht and V. Kumar, "A Cost Effective and Energy Efficient Algorithm for Cloud Computing," International Journal of Mathematical, Engineering and Management Sciences, vol. 7, pp. 681-696, 2022.
12) S. D. Vispute and P. Vashisht, "Optimized Energy Efficient Task Scheduling in Fog Computing," in International Conference on Innovations in Computational Intelligence and Computer Vision, Singapore, 2022, pp. 735-746, Springer Nature Singapore.
13) S. D. Vispute and P. Vashisht, "Energy-efficient task scheduling in fog computing based on particle swarm optimization," SN Computer Science, vol. 4, no. 4, p. 391, 2023.