

A REVIEW OF COMPARATIVE STUDY OF EDGE COMPUTING NETWORKS Vs. CLOUD COMPUTING NETWORKS FOR LATENCY-SENSITIVE APPLICATIONS

Kajal Singh¹, Mrs.Sahreen Hijab²

¹Master of Technology, Computer Science and Engineering, Sagar Institute of Technology and Management, Barabanki, India

²Assistant Professor, Department of Computer Science and Engineering, Sagar Institute of Technology and Management, Barabanki, India

Abstract - The rapid growth of latency-sensitive applications such as autonomous vehicles, augmented and virtual reality, industrial automation, and remote healthcare has exposed the limitations of traditional cloud computing networks in meeting stringent delay and reliability requirements. This review presents a comprehensive comparative study of cloud computing and edge computing networks with a focus on their suitability for latency-sensitive applications. The paper systematically analyzes fundamental concepts, architectural differences, performance metrics, and application-specific requirements based on existing literature. The review highlights that while cloud computing offers scalability and centralized resource management, it suffers from high latency and bandwidth constraints. Edge computing, by contrast, significantly reduces latency by enabling localized data processing and real-time decision-making. However, challenges related to security, scalability, and resource management persist. The study concludes that hybrid edge-cloud architectures provide an effective balance between low-latency performance and scalable computation, making them a promising solution for next-generation networked systems.

Key Words: Edge computing, Cloud computing, Latency-sensitive applications, Quality of Service, Internet of Things, 5G networks.

1. INTRODUCTION

1.1 Background

The rapid proliferation of data-intensive and real-time digital services has fundamentally transformed modern computing paradigms. Cloud computing has emerged as a dominant model by offering centralized, scalable, and on-demand computational resources over the Internet (Mell and Grance, 2011). However, the exponential growth of Internet of Things (IoT) devices, mobile users, and real-time applications has exposed critical limitations of centralized cloud infrastructures, particularly in terms of latency, bandwidth congestion, and reliability. These challenges have motivated the exploration of alternative computing paradigms that can support stringent performance requirements closer to end users. As a result, edge

computing has gained significant attention as a complementary approach to cloud computing, enabling computation and data processing at the network edge to improve responsiveness and service quality (Shi et al., 2016).

1.2 Importance of Latency-Sensitive Applications

Latency-sensitive applications are systems in which even minimal communication or processing delays can significantly degrade functionality, safety, or user experience. Such applications demand ultra-low latency, high reliability, and real-time data processing to operate effectively. With the advancement of 5G and emerging 6G networks, the number of applications requiring millisecond-level latency has increased substantially, making latency a critical performance metric in modern networked systems (Cisco, 2020).

1.2.1 Characteristics of Latency-Sensitive Applications

Latency-sensitive applications are characterized by strict timing constraints, continuous data exchange, and often mission-critical operations. These applications typically involve real-time decision-making, where delays in data transmission or processing may result in system instability, safety hazards, or poor quality of service (QoS). Unlike traditional batch-processing workloads, latency-sensitive systems require predictable and deterministic response times, making centralized cloud processing less suitable in many scenarios (Satyanarayanan, 2017).

Prominent examples include augmented and virtual reality (AR/VR), where motion-to-photon latency must be minimized to prevent motion sickness and ensure immersive user experiences (Shi et al., 2016). Autonomous vehicles rely on real-time sensor data processing and vehicle-to-everything (V2X) communication, where delays may lead to catastrophic outcomes (Taleb et al., 2017). In industrial IoT environments, real-time control and monitoring systems require immediate feedback to maintain operational efficiency and safety. Similarly, telemedicine and remote surgery demand ultra-reliable and low-latency

communication to ensure precision and patient safety (Khan et al., 2020).

1.3 Traditional Cloud Computing Challenges

Despite its scalability and cost efficiency, traditional cloud computing faces inherent challenges when supporting latency-sensitive applications. Centralized data centers are often geographically distant from end users, resulting in increased round-trip delays and network congestion. Additionally, reliance on wide-area networks (WANs) introduces unpredictable latency and packet loss, which can adversely affect real-time services. Security and privacy concerns also arise due to the transmission of sensitive data over public networks. These limitations have prompted researchers to reconsider the suitability of cloud-only architectures for emerging real-time and mission-critical applications (Satyanarayanan, 2017).

1.4 Emergence of Edge Computing

Edge computing has emerged as a promising paradigm to address the shortcomings of traditional cloud computing by bringing computation, storage, and intelligence closer to data sources and end users. By deploying edge nodes such as base stations, gateways, and micro data centers near the network edge, edge computing significantly reduces latency and alleviates backbone network congestion (Shi et al., 2016). Furthermore, edge computing enhances context awareness, improves data privacy, and supports localized decision-making. This paradigm is increasingly viewed as an essential component of next-generation networks, particularly in conjunction with 5G and IoT ecosystems (Mach and Becvar, 2017).

1.5 Objective and Scope of the Review

The primary objective of this review paper is to present a comprehensive comparative study of edge computing networks and cloud computing networks with a specific focus on latency-sensitive applications. This study systematically analyzes architectural differences, performance metrics, and application suitability of both paradigms. The scope of the review includes an examination of existing literature, identification of research gaps, and discussion of challenges and future research directions.

2. FUNDAMENTALS

This section presents the fundamental concepts required to understand and compare cloud computing networks and edge computing networks. It introduces formal definitions, architectural distinctions, and key performance metrics relevant to latency-sensitive applications.

2.1 Definitions

Computing paradigms have evolved to address the growing demand for scalable, efficient, and real-time data processing. Among these paradigms, cloud computing and edge computing represent two distinct yet complementary approaches. Understanding their definitions is essential to establish a foundation for comparative analysis.

2.1.1 Cloud Computing Networks

Cloud computing networks refer to a centralized computing model in which computational resources such as processing power, storage, and networking are delivered as services over the Internet. According to the National Institute of Standards and Technology (NIST), cloud computing enables ubiquitous, convenient, and on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort (Mell and Grance, 2011). In cloud computing networks, data is typically transmitted from end devices to large-scale, geographically distributed data centers where processing and decision-making occur. This model is well suited for applications requiring high scalability and massive storage but is often constrained by network latency for real-time use cases (Armbrust et al., 2010).

2.1.2 Edge Computing Networks

Edge computing networks represent a decentralized computing paradigm that brings computation, storage, and intelligence closer to data sources and end users. Instead of relying solely on centralized cloud data centers, edge computing deploys processing capabilities at the network edge, such as base stations, gateways, routers, and micro data centers (Shi et al., 2016).

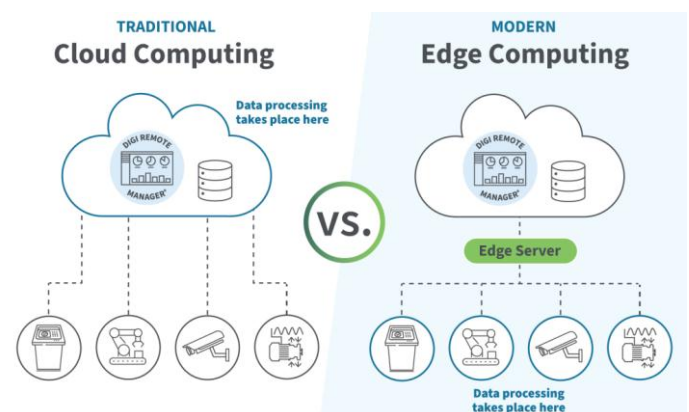


Figure 1: Cloud vs Edge computing architectural comparison showing local and remote processing points.

2.2 Architectural Differences

The architectural distinction between cloud computing and edge computing lies primarily in the location of computational resources and data processing. Cloud computing architectures are centralized, relying on remote data centers connected via wide-area networks (WANs). In contrast, edge computing architectures are distributed, with multiple edge nodes positioned closer to end devices. While cloud architectures emphasize resource pooling and elastic scalability, edge architectures prioritize proximity, low latency, and localized processing. In practice, modern systems often adopt a hybrid architecture, where edge computing handles time-critical tasks and the cloud manages large-scale analytics and long-term storage (Mach and Becvar, 2017).

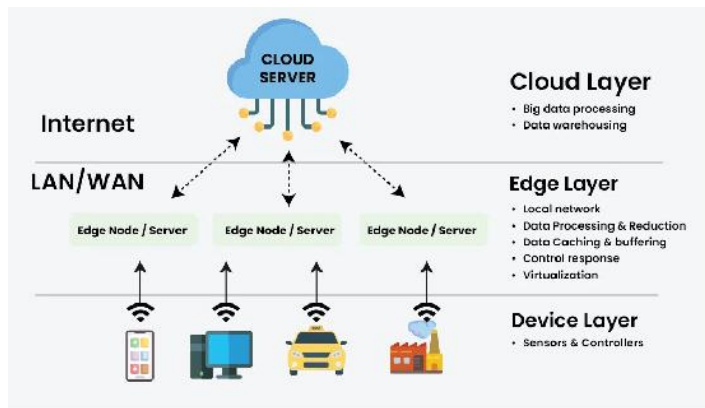


Figure 2: Edge-cloud hybrid architecture demonstrating distributed processing and storage layers.

2.3 Key Performance Metrics

Evaluating cloud and edge computing networks for latency-sensitive applications requires a set of well-defined performance metrics. These metrics enable objective comparison and help determine application suitability.

2.3.1 Latency

Latency refers to the time delay between data generation at the source and the receipt of processed results. It is one of the most critical metrics for real-time and latency-sensitive applications. In cloud computing networks, latency is influenced by the physical distance between users and data centers, as well as network congestion. Edge computing significantly reduces latency by performing computation closer to the data source, thereby improving responsiveness and quality of service (QoS) (Satyanarayanan, 2017).

2.3.2 Bandwidth Utilization

Bandwidth utilization measures the amount of network capacity consumed during data transmission. Cloud-centric

architectures often require continuous transmission of large volumes of raw data to centralized servers, leading to increased bandwidth consumption and potential network bottlenecks. Edge computing mitigates this issue by processing and filtering data locally, transmitting only relevant or aggregated information to the cloud. This approach reduces backbone network traffic and improves overall network efficiency (Shi et al., 2016).

2.3.3 Energy Efficiency

Energy efficiency is a crucial metric, particularly for mobile and IoT devices with limited power resources. Cloud computing may increase energy consumption due to frequent long-distance data transmission. Edge computing can improve energy efficiency by reducing communication overhead and enabling localized processing. However, the deployment of numerous edge nodes introduces new energy management challenges, making efficient resource orchestration essential (Deng et al., 2016).

2.3.4 Scalability

Scalability refers to a system's ability to handle increasing workloads without performance degradation. Cloud computing excels in scalability due to its centralized resource pooling and virtualization capabilities. Edge computing, while beneficial for latency reduction, faces scalability challenges arising from heterogeneous devices, limited resources at edge nodes, and dynamic network conditions. Effective coordination between cloud and edge layers is necessary to achieve scalable and reliable systems (Mach and Becvar, 2017).

2.3.5 Security and Privacy Concerns

Security and privacy are critical concerns in both cloud and edge computing environments. Cloud computing faces risks related to data breaches, unauthorized access, and data sovereignty due to centralized data storage. Edge computing improves privacy by processing sensitive data locally; however, it introduces new security challenges such as physical node vulnerability, distributed attack surfaces, and trust management among heterogeneous edge devices. Addressing these concerns requires robust encryption, authentication, and access control mechanisms across both paradigms (Roman, Lopez and Mambo, 2018).

3. LITERATURE REVIEW

This section systematically reviews existing research related to cloud computing and edge computing networks, with a particular emphasis on their applicability to latency-sensitive applications. The review highlights key findings, comparative insights, and research gaps identified in prior studies.

3.1 Overview of Reviewed Works

A comprehensive review of the literature was conducted to understand the evolution, capabilities, and limitations of cloud and edge computing paradigms. The reviewed works primarily focus on architectural designs, performance evaluation, and application-specific implementations related to latency-sensitive systems. Both survey papers and experimental studies were considered to ensure balanced coverage of theoretical and practical perspectives.

3.1.1 Selection Criteria

To ensure relevance and quality, a structured selection methodology was adopted for identifying research articles. The review emphasizes peer-reviewed journal papers, conference proceedings, and authoritative technical reports.

3.2 Key Findings from Cloud Computing Research

Cloud computing has been extensively studied as a scalable and cost-effective solution for data-intensive applications. However, its suitability for latency-sensitive applications has been a subject of ongoing investigation.

3.2.1 Latency Characterization in Cloud Environments

Several studies have analyzed latency behavior in cloud environments, identifying network distance, congestion, and virtualization overhead as major contributing factors. Armbrust et al. (2010) demonstrated that centralized cloud architectures introduce unpredictable latency due to shared resources and wide-area network dependencies. Subsequent studies highlighted that while cloud data centers offer high computational power, the round-trip delay significantly impacts real-time application performance, particularly for mobile and IoT-based systems (Li et al., 2018).

3.2.2 Cloud-Based Solutions for Real-Time Applications

Researchers have proposed various cloud-based optimization techniques to support real-time applications, including resource provisioning, workload scheduling, and virtualization enhancements. Techniques such as priority-aware scheduling and latency-aware resource allocation have shown moderate improvements in response time (Zhang et al., 2015). However, these approaches often rely on complex orchestration mechanisms and are limited by the inherent distance between cloud servers and end devices.

3.2.3 Limitations Identified in Literature

Despite ongoing optimizations, existing literature consistently reports that cloud computing struggles to meet ultra-low latency requirements. Key limitations include dependency on stable network connectivity, limited context

awareness, and privacy concerns due to centralized data processing. These challenges make cloud-only solutions unsuitable for mission-critical applications such as autonomous driving and remote healthcare (Satyanarayanan, 2017).

3.3 Key Findings from Edge Computing Research

Edge computing research has rapidly expanded as a response to the limitations of cloud-centric architectures. The literature emphasizes its potential to support real-time and latency-sensitive applications.

3.3.1 Latency Reduction Techniques

Multiple studies demonstrate that edge computing significantly reduces latency by offloading computation from centralized clouds to nearby edge nodes. Techniques such as task offloading, computation caching, and data pre-processing at the edge have been shown to reduce end-to-end delay by up to several milliseconds (Shi et al., 2016). These latency reduction strategies are particularly effective in dynamic environments such as vehicular networks and smart cities.

3.3.2 Edge Network Architectures and Protocols

Research has explored various edge network architectures, including fog computing, mobile edge computing (MEC), and multi-access edge computing. Mach and Becvar (2017) highlighted that hierarchical and distributed edge architectures improve scalability and fault tolerance. Protocols designed for edge environments emphasize lightweight communication, local orchestration, and seamless integration with cloud backends to support hybrid deployment models.

3.3.3 Edge-Based Solutions for IoT, 5G, and 6G

Edge computing plays a critical role in enabling next-generation IoT and 5G/6G networks. Studies indicate that edge-assisted IoT frameworks enhance real-time analytics, reduce network congestion, and improve energy efficiency (Khan et al., 2020). In 5G and emerging 6G systems, edge computing supports ultra-reliable low-latency communication (URLLC), making it essential for applications such as augmented reality and autonomous systems (Taleb et al., 2017).

3.4 Comparative Studies in Prior Research

Comparative analyses between cloud and edge computing provide valuable insights into their relative strengths and limitations.

3.4.1 Studies Directly Comparing Edge vs. Cloud

Several studies have explicitly compared cloud and edge computing architectures in terms of latency, bandwidth usage, and resource efficiency. Satyanarayanan (2017) reported that edge-based systems consistently outperform cloud-based systems for latency-sensitive workloads. Hybrid models combining cloud scalability with edge responsiveness were found to offer balanced performance in diverse scenarios.

3.4.2 Findings on Latency and Quality of Service

Comparative evaluations indicate that edge computing significantly improves Quality of Service (QoS) by reducing response time and jitter. Experimental results in vehicular and industrial IoT environments show that edge-based processing meets stringent latency constraints that cloud-based approaches fail to satisfy (Deng et al., 2016). However, cloud computing remains advantageous for compute-intensive and non-real-time workloads.

3.4.3 Identified Gaps in Comparative Analysis

Despite existing comparative studies, the literature reveals a lack of standardized evaluation metrics and benchmark scenarios. Many studies focus on specific applications or controlled environments, limiting generalizability. Additionally, few works comprehensively analyze security, scalability, and energy efficiency alongside latency, highlighting the need for holistic comparative frameworks.

4. COMPARATIVE ANALYSIS

This section presents a detailed comparative analysis of edge computing networks and cloud computing networks with respect to key performance aspects relevant to latency-sensitive applications. The comparison highlights theoretical foundations, empirical findings, and practical implications.

4.1 Latency in Edge vs. Cloud Networks

Latency is the most critical metric for evaluating the suitability of computing paradigms for real-time and mission-critical applications. Cloud and edge computing differ fundamentally in how latency is introduced and managed due to their architectural designs.

4.1.1 Theoretical Insights

From a theoretical perspective, latency in cloud computing networks is primarily influenced by physical distance, routing complexity, and shared network infrastructure. Centralized cloud data centers are often located far from end users, resulting in increased propagation delay and queuing latency. In contrast, edge computing minimizes these delays by placing computation and storage resources closer to data sources. Queueing theory and network models consistently

show that reducing hop count and transmission distance significantly lowers end-to-end latency, which explains the inherent advantage of edge-based architectures for latency-sensitive workloads (Satyanarayanan, 2017; Shi et al., 2016).

4.1.2 Experimental Evidence

Experimental studies further validate the theoretical advantages of edge computing. Empirical evaluations in vehicular networks, smart city deployments, and industrial IoT environments demonstrate that edge computing can reduce latency by 30–70% compared to cloud-only solutions (Deng et al., 2016). Taleb et al. (2017) reported that multi-access edge computing (MEC) significantly improves response times in 5G networks, enabling ultra-reliable low-latency communication (URLLC). These results confirm that edge computing consistently outperforms cloud computing in scenarios with stringent latency constraints.

4.2 Bandwidth and Network Traffic

Bandwidth consumption and network traffic patterns differ substantially between cloud and edge computing networks. Cloud-centric models require continuous transmission of large volumes of raw data to centralized data centers, which increases backbone network traffic and may cause congestion during peak loads. Edge computing alleviates this issue by performing data filtering, aggregation, and preprocessing locally. As a result, only essential or summarized data is transmitted to the cloud, leading to reduced bandwidth usage and improved network efficiency (Shi et al., 2016). This characteristic is particularly beneficial for large-scale IoT deployments generating massive data streams.

4.3 Processing and Computation Distribution

In cloud computing networks, computation is predominantly centralized, with end devices acting mainly as data producers and consumers. This centralized processing model simplifies management but introduces latency and scalability challenges. Edge computing adopts a distributed computation model, where processing tasks are dynamically offloaded between end devices, edge nodes, and cloud servers. Such hierarchical computation distribution improves responsiveness and fault tolerance while enabling localized decision-making. Hybrid edge–cloud frameworks are increasingly favored, as they combine the real-time benefits of edge processing with the computational power of centralized clouds (Mach and Becvar, 2017).

4.4 Cost and Resource Utilization

Cost efficiency and resource utilization are important considerations when comparing cloud and edge computing. Cloud computing benefits from economies of scale, offering cost-effective resource provisioning through virtualization and pay-as-you-go models. However, frequent data

transmission and latency penalties may increase operational costs for real-time applications. Edge computing reduces communication overhead and latency-related costs but introduces additional expenses related to deploying, managing, and maintaining distributed edge infrastructure. Studies suggest that hybrid architectures can optimize cost and performance by allocating time-critical tasks to the edge and compute-intensive tasks to the cloud (Deng et al., 2016).

4.5 Security and Privacy Implications

Security and privacy implications differ significantly between cloud and edge computing networks. Centralized cloud environments are attractive targets for large-scale cyberattacks and raise concerns about data sovereignty and regulatory compliance. Edge computing enhances privacy by enabling local data processing, reducing the need to transmit sensitive information over public networks. However, the distributed nature of edge nodes increases the attack surface and introduces challenges related to physical security, trust management, and authentication. Consequently, both paradigms require robust, multi-layered security mechanisms tailored to their architectural characteristics (Roman, Lopez and Mambo, 2018).

4.6 Scalability and Flexibility for Future Networks

Scalability and flexibility are critical for supporting future networks characterized by massive device connectivity and dynamic workloads. Cloud computing offers high scalability through centralized resource pooling but may struggle to maintain performance under stringent latency requirements. Edge computing provides flexibility by enabling localized scaling and adaptive resource allocation but faces challenges due to heterogeneous hardware and limited edge resources. Future network architectures are expected to rely on seamless integration of cloud and edge computing, supported by AI-driven orchestration and software-defined networking, to achieve scalable, flexible, and low-latency services (Taleb et al., 2017).

5. LATENCY-SENSITIVE APPLICATIONS

Latency-sensitive applications require immediate data processing and response to function correctly. In such applications, delays of even a few milliseconds can degrade system performance, compromise safety, or negatively impact user experience. This section categorizes major latency-sensitive applications, discusses their performance requirements, and compares the suitability of cloud and edge computing paradigms for each application domain.

5.1 Application Categories

Latency-sensitive applications span multiple domains, including transportation, multimedia, industrial systems, and healthcare. These applications typically involve real-time data acquisition, processing, and actuation,

necessitating ultra-low latency, high reliability, and continuous availability.

5.1.1 Autonomous Vehicles

Autonomous vehicles rely on real-time processing of sensor data collected from cameras, LiDAR, radar, and vehicle-to-everything (V2X) communications. Decision-making tasks such as obstacle detection, path planning, and collision avoidance must be performed within strict latency constraints to ensure passenger safety. Studies indicate that end-to-end latency requirements for autonomous driving applications can be as low as 1–10 milliseconds (Taleb et al., 2017). Edge computing enables rapid local processing and real-time coordination among nearby vehicles, whereas cloud-based solutions alone are insufficient due to communication delays and reliability concerns.

5.1.2 Augmented and Virtual Reality (AR/VR)

Augmented and virtual reality applications demand extremely low latency to provide immersive and seamless user experiences. Motion-to-photon latency, which measures the delay between user movement and visual feedback, must typically be below 20 milliseconds to prevent motion sickness and disorientation (Shi et al., 2016). Edge computing supports AR/VR by offloading computationally intensive rendering and tracking tasks to nearby edge servers, significantly reducing response time compared to centralized cloud processing (Satyanarayanan, 2017).

5.1.3 Industrial Automation and Robotics

Industrial automation systems and collaborative robots operate in time-critical environments where real-time control and monitoring are essential. Applications such as predictive maintenance, process optimization, and robotic control require deterministic latency and high reliability. Centralized cloud architectures may introduce unpredictable delays, making them unsuitable for closed-loop control systems. Edge computing facilitates local data processing and real-time feedback, enabling faster response and improved operational efficiency in smart manufacturing environments (Mach and Becvar, 2017).

5.1.4 Healthcare and Remote Surgery

Healthcare applications, particularly telemedicine and remote surgery, impose stringent latency and reliability requirements to ensure patient safety. Remote surgical procedures require end-to-end latency of less than 10 milliseconds and near-zero packet loss to enable precise control of surgical instruments (Khan et al., 2020). Edge computing enhances healthcare systems by enabling real-time patient monitoring, local data analysis, and reduced transmission delays, while cloud computing supports long-term data storage and large-scale medical analytics.

5.2 Requirements and Performance Metrics

Latency-sensitive applications share common performance requirements, including ultra-low latency, high reliability, minimal jitter, and consistent quality of service. Additional metrics such as bandwidth efficiency, energy consumption, and fault tolerance are also critical, particularly in mobile and IoT-based environments. For applications operating over 5G and emerging 6G networks, support for ultra-reliable low-latency communication (URLLC) is essential. These requirements necessitate computing architectures capable of adaptive resource management and real-time processing (Taleb et al., 2017).

5.3 Cloud and Edge Compare per Application

Comparative studies consistently show that edge computing outperforms cloud computing for latency-sensitive applications by reducing response time and improving reliability. For autonomous vehicles and industrial automation, edge-based processing enables immediate decision-making and localized coordination. In AR/VR applications, edge computing significantly improves user experience by minimizing motion-to-photon latency. In healthcare, edge computing enhances real-time monitoring and emergency response, while cloud computing remains valuable for non-time-critical tasks such as data archiving and large-scale analytics. Consequently, hybrid edge-cloud architectures are widely regarded as the most effective solution, combining the low-latency benefits of edge computing with the scalability and computational power of cloud platforms (Satya narayanan, 2017; Deng et al., 2016).

6. CHALLENGES AND RESEARCH ISSUES

Despite the significant advantages offered by edge computing over traditional cloud computing for latency-sensitive applications, several challenges and open research issues remain. These challenges span technical, architectural, and operational domains and must be addressed to enable large-scale, reliable, and secure deployment of edge-cloud systems.

6.1 Technical Challenges

The technical challenges associated with edge and cloud computing primarily arises from the distributed and heterogeneous nature of modern networked systems. Efficient coordination between end devices, edge nodes, and centralized cloud servers remains a complex task, particularly for real-time applications with strict latency constraints.

6.1.1 Resource Management

Resource management is a critical challenge in edge computing environments due to limited computational capacity, storage, and energy availability at edge nodes.

Unlike cloud data centers, which offer virtually unlimited resources through virtualization, edge nodes must dynamically allocate resources among competing applications. Efficient task scheduling, workload offloading, and resource orchestration mechanisms are required to meet latency and quality-of-service requirements while minimizing overhead (Deng et al., 2016). Poor resource management can lead to performance degradation and service instability.

6.1.2 Heterogeneous Network Support

Edge computing environments are inherently heterogeneous, consisting of diverse devices, communication protocols, and network technologies. Supporting seamless operation across heterogeneous networks such as Wi-Fi, cellular, 5G, and future 6G systems poses significant challenges. Differences in hardware capabilities, operating systems, and network conditions complicate application deployment and management. Research efforts emphasize the need for adaptive middleware and abstraction layers to ensure interoperability and consistent performance across heterogeneous environments (Mach and Becvar, 2017).

6.2 Standardization and Interoperability

The lack of unified standards remains a major barrier to the widespread adoption of edge computing. Unlike cloud computing, which benefits from mature standardization frameworks, edge computing involves multiple stakeholders, vendors, and deployment models. This fragmentation hinders interoperability between edge platforms and cloud services. Standardization efforts by organizations such as ETSI and IEEE aim to define reference architectures and interfaces; however, achieving global consensus remains an open research issue (Taleb et al., 2017).

6.3 Security, Privacy and Trust Management

Security, privacy, and trust management are among the most critical challenges in edge-cloud environments. While edge computing improves privacy by enabling local data processing, it also introduces vulnerabilities due to the distributed nature of edge nodes. Physical exposure of edge devices increases the risk of tampering, while decentralized trust management complicates authentication and authorization. Ensuring secure data exchange, maintaining user privacy, and establishing trust among heterogeneous entities require advanced cryptographic techniques, secure hardware, and distributed trust frameworks (Roman, Lopez and Mambo, 2018).

6.4 Energy Efficiency and Sustainability

Energy efficiency is a key concern, particularly for edge nodes deployed in resource-constrained environments. Continuous computation and communication at the edge can

increase energy consumption, affecting system sustainability and operational costs. Research highlights the importance of energy-aware task scheduling, adaptive resource allocation, and energy-efficient hardware design to reduce power consumption. Sustainable edge-cloud systems are essential for supporting long-term deployment of IoT and smart city applications (Shi et al., 2016).

6.5 Scalability for Massive IoT

The rapid growth of IoT devices presents significant scalability challenges for both cloud and edge computing systems. Massive IoT deployments generate large volumes of data and require simultaneous connectivity for millions of devices. While edge computing reduces network congestion by localizing data processing, managing and orchestrating a large number of edge nodes remains complex. Scalability issues related to device mobility, dynamic workload distribution, and fault tolerance remain open research problems, particularly in ultra-dense network environments (Khan et al., 2020).

7. CONCLUSION

This review presented a comprehensive comparative analysis of cloud computing networks and edge computing networks with a specific focus on latency-sensitive applications. The study highlighted that while cloud computing remains highly effective for scalable, compute-intensive, and non-real-time workloads, its centralized architecture introduces inherent latency, bandwidth, and reliability constraints that limit its suitability for time-critical applications. In contrast, edge computing addresses these limitations by decentralizing computation and bringing processing capabilities closer to data sources and end users. Through analysis of existing literature, it is evident that edge computing significantly improves latency performance, bandwidth efficiency, and quality of service for applications such as autonomous vehicles, AR/VR, industrial automation, and healthcare systems. However, the review also indicates that neither paradigm is sufficient in isolation. Hybrid edge-cloud architectures emerge as the most promising solution, combining low-latency responsiveness at the edge with the scalability and computational power of the cloud. Overall, this review provides valuable insights for researchers and practitioners designing next-generation networks to support real-time and mission-critical applications.

8. LIMITATIONS OF THE REVIEW

Despite its comprehensive scope, this review has several limitations. First, the analysis is primarily based on existing literature and does not include original experimental validation or real-world deployment results. Second, the reviewed studies often use different evaluation metrics and experimental setups, which may affect the direct comparability of results. Third, emerging technologies such as 6G, AI-driven orchestration, and federated learning are

still in early research stages, limiting the availability of mature performance data. Finally, economic and regulatory considerations were discussed only at a high level, leaving room for deeper investigation in future work.

REFERENCES

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I. and Zaharia, M. (2010) 'A view of cloud computing', *Communications of the ACM*, 53(4), pp. 50–58.
2. Cisco (2020) Cisco annual Internet report (2018–2023). Cisco Systems.
3. Deng, R., Lu, R., Lai, C., Luan, T.H. and Liang, H. (2016) 'Optimal workload allocation in fog-cloud computing', *IEEE Transactions on Vehicular Technology*, 65(10), pp. 8783–8793.
4. Khan, L.U., Yaqoob, I., Tran, N.H., Kazmi, S.M.A., Dang, T.N. and Hong, C.S. (2020) 'Edge computing enabled smart cities', *IEEE Internet of Things Journal*, 7(10), pp. 10200–10232.
5. Li, W., Chou, W. and Zhou, Y. (2018) 'Performance evaluation of cloud-based real-time applications', *Future Generation Computer Systems*, 80, pp. 245–256.
6. Mach, P. and Becvar, Z. (2017) 'Mobile edge computing: A survey', *IEEE Communications Surveys & Tutorials*, 19(3), pp. 1628–1656.
7. Mell, P. and Grance, T. (2011) The NIST definition of cloud computing. NIST Special Publication 800-145. National Institute of Standards and Technology.
8. Roman, R., Lopez, J. and Mambo, M. (2018) 'Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges', *Future Generation Computer Systems*, 78, pp. 680–698.
9. Satyanarayanan, M. (2017) 'The emergence of edge computing', *Computer*, 50(1), pp. 30–39.
10. Shi, W., Cao, J., Zhang, Q., Li, Y. and Xu, L. (2016) 'Edge computing: Vision and challenges', *IEEE Internet of Things Journal*, 3(5), pp. 637–646.
11. Taleb, T., Samdanis, K., Mada, B., Flinck, H., Dutta, S. and Sabella, D. (2017) 'On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration', *IEEE Communications Surveys & Tutorials*, 19(3), pp. 1657–1681.
12. Zhang, Q., Chen, M., Li, L. and Li, S. (2015) 'Dynamic resource allocation for real-time cloud services', *Future Generation Computer Systems*, 52, pp. 83–94.

13. Bonomi, F., Milito, R., Zhu, J. and Addepalli, S. (2012) 'Fog computing and its role in the Internet of Things', Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, pp. 13–16.
14. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J. and Brandic, I. (2009) 'Cloud computing and emerging IT platforms', Future Generation Computer Systems, 25(6), pp. 599–616.
15. Chiang, M. and Zhang, T. (2016) 'Fog and IoT: An overview of research opportunities', IEEE Internet of Things Journal, 3(6), pp. 854–864.
16. Dinh, H.T., Lee, C., Niyato, D. and Wang, P. (2013) 'A survey of mobile cloud computing', Wireless Communications and Mobile Computing, 13(18), pp. 1587–1611.
17. ETSI (2019) Multi-access edge computing (MEC); Framework and reference architecture. ETSI GS MEC 003.
18. Gupta, H., Vahid Dastjerdi, A., Ghosh, S.K. and Buyya, R. (2017) 'iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things', Software: Practice and Experience, 47(9), pp. 1275–1296.
19. Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E. and Chiroma, H. (2016) 'The role of big data in smart city', International Journal of Information Management, 36(5), pp. 748–758.
20. Hou, X., Li, Y., Chen, M., Wu, D., Jin, D. and Chen, S. (2016) 'Vehicular fog computing', IEEE Communications Magazine, 54(11), pp. 87–93.
21. Hu, Y.C., Patel, M., Sabella, D., Sprecher, N. and Young, V. (2015) 'Mobile edge computing—A key technology towards 5G', ETSI White Paper, 11(11), pp. 1–16.
22. ITU-R (2020) Minimum requirements related to technical performance for IMT-2020 radio interface(s). International Telecommunication Union.
23. Li, X., Dang, Y., Zhang, M. and Chen, J. (2019) 'Latency-aware workload offloading in mobile edge computing', IEEE Access, 7, pp. 125398–125409.
24. Mao, Y., You, C., Zhang, J., Huang, K. and Letaief, K.B. (2017) 'A survey on mobile edge computing', IEEE Communications Surveys & Tutorials, 19(4), pp. 2322–2358.
25. Ni, L., Zhang, J., Jiang, C., Yan, C. and Yu, K. (2019) 'Resource allocation in mobile edge computing', IEEE Communications Surveys & Tutorials, 21(3), pp. 2291–2316.
26. OpenFog Consortium (2017) OpenFog reference architecture for fog computing. OpenFog Consortium.
27. Park, S., Kim, Y., Kim, S. and Jung, J. (2018) 'Latency-aware resource management for edge computing', IEEE Transactions on Network and Service Management, 15(1), pp. 295–308.
28. Perera, C., Liu, C.H., Jayawardena, S. and Chen, M. (2014) 'A survey on Internet of Things from industrial market perspective', IEEE Access, 2, pp. 1660–1679.
29. Rahmani, A.M., Liljeberg, P., Preden, J.S. and Jantsch, A. (2018) 'Fog computing in the Internet of Things', Springer International Publishing.
30. Wang, S., Zhang, X., Zhang, Y., Wang, L., Yang, J. and Wang, W. (2017) 'A survey on mobile edge networks', IEEE Access, 5, pp. 25505–25530.
31. Xu, X., Chen, S., Li, J., Xu, Y. and Xu, Y. (2018) 'Latency and energy efficient computation offloading', IEEE Transactions on Mobile Computing, 17(10), pp. 2463–2476.
32. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K. and Zhang, J. (2019) 'Edge intelligence', Proceedings of the IEEE, 107(8), pp. 1655–1674.