© April 2026 | IJIRT | Volume 12 Issue 11 | ISSN: 2349-6002

Performance Analysis of Edge Computing versus Cloud Computing for Real-Time IoT Applications

Dr. G. J. Navale[1], Dr. Kishor Wagh[2], Mohit Khairnar[3], Shubham Jadhav[4], Deep Kale[5]
[1,2,3,4,5] *Computer Engineering, AISSMS Institute of Information Technology* Pune,India

Abstract—**The resulting proliferation of Internet of Things (IoT) devices has created massive data streams, thus frustrating traditional cloud-based processing models. This paper analyzes performance trade-offs of the edge and cloud computing in real-time IoT applications. Latency, bandwidth utilisation, scalability, energy consumption and cost are the key metrics that are discussed and an explanation is given on how each paradigm solves these aspects. As an example, edge computing provides processing of data adjacent to the source, and is associated with truncating round-trip latencies and network traffic, whereas cloud computing resources are virtually unlimited and are centrally managed. The practical use (e.g. smart healthcare, autonomous vehicles, industrial IoT) is analyzed to show how these paradigms are applied; the problem of resource limits and security are addressed, and the perspectives of the future are presented (e.g. TinyML, 5G, and hybrid models). The results point to the fact that edge and cloud are not always the best options; instead, a combination of these strategies that more or less combine the advantages of the two slowly prevails.**

I. INTRODUCTION

The Internet of Things (IoT) is a network of billions of sensors and devices, which allows billions of applications, including smart cities and industrial automation. The current IoT implementations generate mind-boggling amounts of data; the current networks are capable of producing more data daily than the total throughput of the entire data centres ten years ago. Conventionally, the data in IoT is sent to remote cloud computing servers where it is processed to provide immense computational and storage capabilities. However, cloud-only solutions are faced with significant drawbacks on real-time and latency-sensitive activities. According to empirical research, cloud processing frequently experiences end-to-end latencies ranging from tens to hundreds of milliseconds. This is unacceptable for applications that require response times in the sub-millisecond range, such as virtual reality or remote surgery.

Edge computing is the solution to this gap, as it brings the calculation to the places of data. Processing near the IoT sensors is done by edge nodes, such as a gateway, local server, or even embedded device, and thus minimizes the distance that data needs to go. This decentralisation technique minimises the communication delay and bandwidth consumption significantly. In essence, edge computing is the cloud's complement; heavier analytics and long-term storage are handled on the cloud, while real-time, low-level processing is done on the edge. Based on recent empirical research and case studies, the current paper offers a thorough performance evaluation of these two paradigms within the context of the Internet of Things. In addition to discussing practical applications, problems, and upcoming trends, it compares edge and cloud in terms of latency, bandwidth, scalability, energy, and cost.

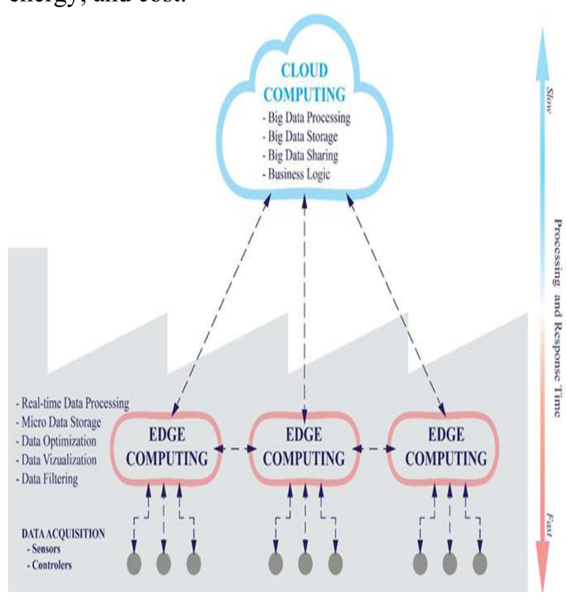

Fig. 1. The framework of differences between Edge Computing and Cloud Computing

II. BACKGROUND AND RELATED WORK

Cloud computing is a paradigm that has been well established, and provides on-demand computing, storage and services via the Internet. Central databases and enormous scalability are offered by large cloud solutions (AWS, Azure, Google Cloud, etc.). The cloud, when applied in the context of the Internet of Things (IoT), allows aggregating sensor data and more complex analytics in a central place, using the elastic resources and advanced AI/ML capabilities that would be practically impossible on the limited devices. The pay-as-you-go model of pricing also offers flexibility of costs, and thus allows users to dynamically scale resources. Nevertheless, the centralization nature of the cloud requires every data to pass through the network to the remote data centers. With the spread of IoT devices, the so-called all-to-cloud model creates bottlenecks, such as high latency, bandwidth overload, and possible reliability problems in the case of poor network conditions. This is also associated with security and privacy, in which sensitive IoT data can be delivered to third-party servers.

A portion of processing is moved to the network edge, which is physically near devices, by edge computing, which is often used interchangeably with fog computing. Edge resources can include servers within the premises, gateways or even the end devices. Edge computing helps eliminate or lessen network load, communication delays, by moving computations nearer to the data source(s). An example is that the reaction to sensor input (e.g. stopping an erroneous machine) can be immediate in an edge node without the round trip to a distant data centre. Edge systems have the ability to filter, aggregate or pre-process data and only send upstream the information that is relevant. This edge and cloud hybrid model has emerged as a point of current research and practice in the industry as it attempts to realize the benefits of both paradigms. There have been many surveys and experiments related to these approaches that are recorded in the literature. To give an example, a massive measurement experiment has shown that a significantly large fraction of end-users, about 58 percent, were able to contact an edge server in less than 10 ms, which is a well-known 29 percent in case of a cloud data center. This result validates that edge compute can significantly decrease latency to most users, which is a huge advantage of edge compute to time-sensitive IoT applications.

III. PERFORMANCE ANALYSIS

Edge and cloud computing assessment in this paper is across five critical performance dimensions, namely latency, bandwidth consumption, scalability, energy consumption, and cost. All of these dimensions can have a strong impact on the suitability of paradigm to real-time IoT workloads. Most of the time, edge computing is faster, in terms of latency and bandwidth, and cloud computing is faster, in terms of raw scalability; both have trade-offs in terms of energy use and cost. The comparative traits of both metrics are as follows:

Because edge computing processes data close to the source, end-to-end latency is greatly reduced. In reality, cloud round-trip times are frequently more than tens of milliseconds, and edge node response times are within a few milliseconds or occasionally less. As an example, a study has noted that the total latency of cloud deployments is normally above 80ms, whilst edge processing may meet strict requirements of about 1ms in terms of interactive IoT services. In real-time applications (like autonomous robotics and augmented reality), where decisions must be made in an instant, this low latency cannot be tolerated. On the other hand, the cloud would most probably violate these latency limits. This means that edge computing is typically better suited to tasks with a high level of latency sensitivity, while cloud computing (although slower) can be used to process batches that are non-urgent.

Total IoT Response Latency

$$\mathcal{L}_{tot} = \mathcal{L}_{tx} + \mathcal{L}_{proc} + \mathcal{L}_{rx} \quad (1)$$

Where:

- $\mathcal{L}_{tx}$= data transmission delay
- $\mathcal{L}_{proc}$= computation/processing delay
- $\mathcal{L}_{rx}$= response propagation delay

Edge vs Cloud Latency

$$\mathcal{L}_{edge} < \mathcal{L}_{cloud} \text{ such that } \mathcal{L}_{cloud} = \mathcal{L}_{edge} + \Delta_{wan} \quad (2)$$

Where Δ_{wan} represents additional wide-area network delay incurred when using cloud servers.

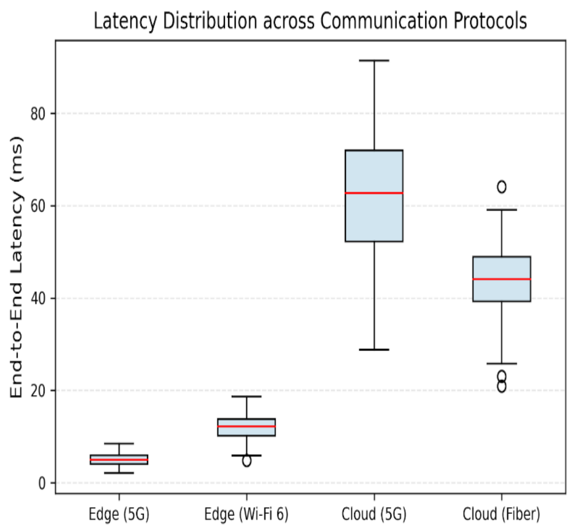

Fig.2. End-to-End Delay in Edge vs Cloud Communication

Edge computing greatly reduces the amount of data sent over wide-area networks. Rather than streaming raw sensor feeds to the cloud, edge nodes can pre-process or filter data locally, sending only essential information or alerts upstream. This conserves network bandwidth and avoids congestion. In high-bandwidth IoT contexts (e.g. video streams, LIDAR scans), edge analytics can cut backhaul traffic by orders of magnitude. For instance, processing surveillance video at the edge and transmitting only events or summaries can yield ~90%+ reduction in uplink data. Cloud-centric IoT systems, on the other hand, may saturate links by pushing all raw data to data centers, incurring heavy bandwidth costs and potential delays. In summary, edge computing optimizes bandwidth by short-circuiting high-volume data flows, whereas cloud computing requires more robust network capacity.

Network Bandwidth Requirement

$$\mathcal{B}_{net} = \rho_d \times \sigma_r \qquad (3)$$

Where:

- ρ_d= data generation rate (MB/s)
- σ_r= ratio of transmitted data

For edge computing:

Edge Bandwidth Reduction

$$\sigma_r^{edge} \ll \sigma_r^{cloud} \Rightarrow \mathcal{B}_{edge} \ll \mathcal{B}_{cloud} \qquad (4)$$

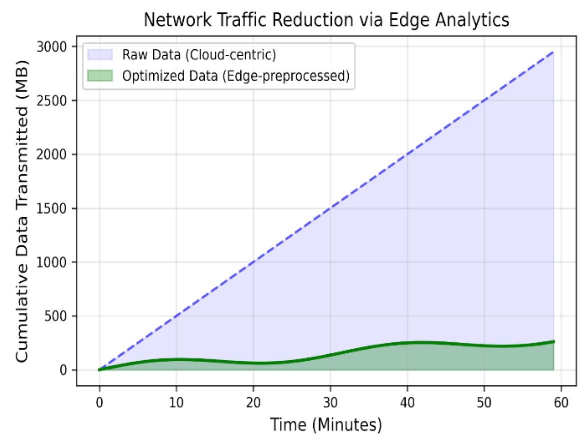

Fig.3. Comparison of Data Transmission With and Without Edge Processing

Cloud applications are by definition scalable indefinitely. They have large resource pools that are capable of scaling down and down to support increased or reduced Internet-of-Things (IoT) workloads. Cloud system is able to accommodate any sudden increase in demand by simply creating more servers making it especially ideal when a large-scale deployment is required. In contrast, edge computing has limited capacity due to the limited capacity of each node on an edge. The edge systems can be scaled by the distribution of workloads to a variety of local nodes including more gateways, micro-data centres or devices. Edge architectures can scale using this distributed paradigm but every node is constrained by a limited amount of CPU, memory and storage. Moreover, an edge network has to control heterogeneity and inter-node synchronization. In real-world use, most systems don't rely only on the cloud or only on edge devices. Instead, they use a mix of both. Tasks that need quick responses are handled by nearby edge servers, while heavier processing and data collection are sent to the cloud whenever required.Therefore, cloud computing provides unmatched centralized scalability, and edge systems will scalability is accomplished with the deployment of extra local resources.

Processing Capacity Limit

$$\mathcal{C}_i = \sum_{k=1}^{n} \psi_k \leq \Omega_i \qquad (5)$$

Where:

- ψ_k= workload of IoT task k

- Ω_i= maximum compute capacity of node i

For cloud systems:

$$\Omega_{cloud} \to \infty \qquad (6)$$

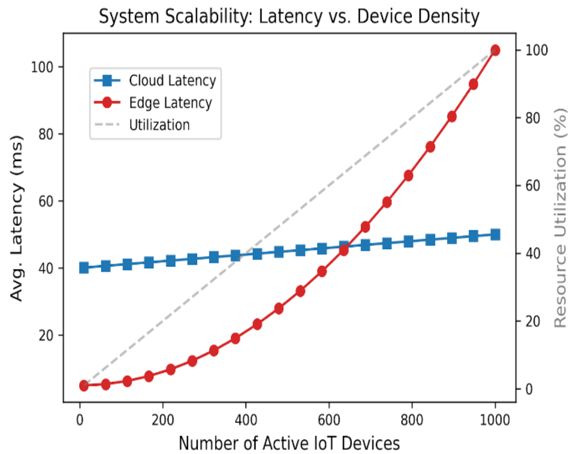

Fig.4. Latency and Scalability of Edge vs Cloud Systems

The implementation of data processing in the network edge can be used to improve the energy efficiency of Internet of Things (IoT) systems. Wireless transfers often use up a large percentage of the power budget of mobile and battery sensors, limiting the overall performance and lifespan of the system. Extensive computation on localized devices combined with fewer data transfers allows edge computing to reduce the energy used by the computing device itself as well as the network infrastructure associated with it. These benefits are consistent with the available empirical evidence: transferring the parts of the computationally intensive load off of IoT nodes to proximal edge servers was reported to increase battery lifespan and accelerate speed by a significant factor. In one study, execution speed was reported to have increased approximately 20-fold with a 5% energy consumption decrease with utilizing edge-assisted of processing as opposed to cloud-only structures. Another way the cloud is able to help with saving on energy is through workload aggregation and centralized optimization. As a result, the reduction of long-range communication, which is made possible through edge processing, is likely to reduce the overall energy usage which is specifically important in large-scale IoT systems where power demands are severe.

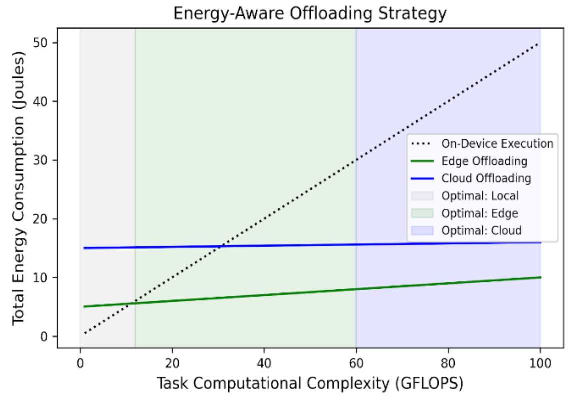

Fig.5. Energy Usage Across Local, Edge, and Cloud Processing

Energy per IoT Task

$$\mathcal{E}_{task} = \mathcal{E}_{comp} + \mathcal{E}_{comm} \qquad (7)$$

Where:

- $\mathcal{E}_{comp}$= computation energy
- $\mathcal{E}_{comm}$= communication energy

Edge optimization condition:

$$\mathcal{E}_{comm}^{edge} < \mathcal{E}_{comm}^{cloud} \qquad (8)$$

Modeling economies of edge and cloud computing are very different. Cloud computing is usually based on a pay-as-you-go model of operation and an operational expenditure (OpEX) paradigm. The users thus end up eliminating initial investment on infrastructure, and they pay a price per compute time, storage and, most importantly, data egress. The fees on data egress and bandwidth might be important in large-scale Internet of Things (IoT) systems. The more expensive element of edge computing is higher capital expenditure (CapEx) to install the devices and gateways; however, it may decrease the current cost through limiting the usage of clouds. Through local data processing, organisations use less cloud bandwidth and pay less infrastructure; in one case, sending only discrete events, as opposed to raw streams of data to the cloud would reduce costly data transfer significantly. One lesson learned is that the total cost of ownership (TCO) in edge deployments is workload-dependent: applications with high volumes of data or data processing neediness tend to justify edge deployments, while compute-intensive analytics with small data input may not be cost-effective to do in the cloud. Analysts suggest a total TCO approach as edge devices and maintenance have their fixed costs, whereas the overall costs of clouds, including data transfer and computation are often the major cost factors in high-scale IoT applications.

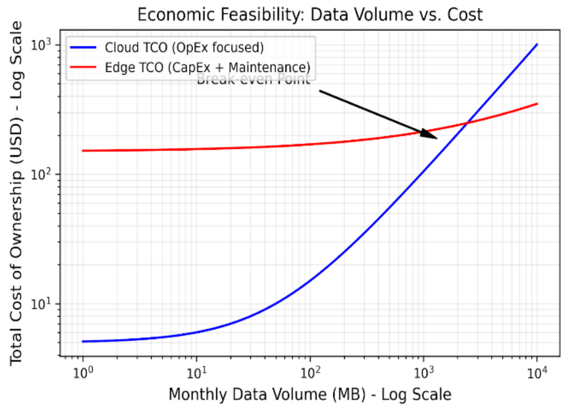

Fig.6. When Edge Computing Becomes More Cost-Effective Than the Cloud

Total Cost of Ownership (TCO)

$$\mathcal{T}_{sys} = \mathcal{C}_{infra} + \mathcal{C}_{comp} + \mathcal{C}_{data} \quad (9)$$

Where:

- $\mathcal{C}_{infra}$= infrastructure cost
- $\mathcal{C}_{comp}$= computation cost
- $\mathcal{C}_{data}$= data transfer/storage cost

Hybrid advantage:

$$\mathcal{T}_{hybrid} < \mathcal{T}_{cloud} \text{ for high-volume IoT workloads} \quad (10)$$

IV. REAL-WORLD USE CASES

In practice, many IoT applications leverage edge and cloud computing in tandem. Below are representative scenarios illustrating how each paradigm supports specific use cases:

A. Smart Healthcare:

The use of edge computing in healthcare monitoring and emergency response is increasing. For instance, the elderly fall detection system uses either cameras or wearable devices at the edge of their home that can perform real-time analytics. A local edge server with a deep learning model can identify a video frame as either normal or falling and trigger an alert in milliseconds. This is in line with the demand for a critical response time (typically <7 minutes to post-fall support) and avoids the network delay of cloud computing. Moreover, the use of edge computing in health data addresses the issues of privacy and compliance. It has been shown that the use of edge computing in geriatric monitoring results in a substantial improvement in the safety of geriatric patients with immediate notifications. Cloud-only systems, on the other hand, are prone to delayed notification in emergency situations. Wearable devices for vital sign monitoring are also included in Smart-health IoT, and local anomaly detection (on-device or gateway) can provide immediate notifications of an emergency situation without uploading the streams to the cloud.

B. Autonomous Vehicles:

Self-driving vehicles and drones are ideal real-time IoT systems that are based on edge computing. Every vehicle is in effect an edge node on wheels, with sensor (cameras, LIDAR, radar) input being processed on-board. The automotive system is an example where local compute is used to detect objects, avoid collisions, and navigate vehicles in real-time without having to wait before clouds verify them. Benchmarks have shown that autonomous vehicles will have to react in tens of milliseconds, which is much lower than the round-trip times of a cloud. This can be made possible by edge computing, which can be run on edge-capable hardware (such as automotive GPUs or special chips) and only connected to the cloud to perform non-critical functions (such as map updates). Practically, cars are frequently transmitting only reduced telemetry to the cloud to learn about the fleet and perform diagnostics, reducing bandwidth. Independent vehicles would not be safe nor achieve latency without edge processing.

C. Industrial IoT:

Edge and cloud computing are often integrated in manufacturing and industrial applications. In the factory, industrial IoT relies on sensors and cameras to track equipment in real-time. Edge gateways can be installed close to equipment to immediately analyze and control the equipment. For example, one industrial IoT application involves vision-based edge computing to sort items and optimize manufacturing processes in real-time. Edge nodes can also be used for predictive maintenance. Vibration or acoustic sensors on equipment transmit data to a local edge server, which identifies anomalies and sends notifications, even with intermittent network connections. (For example, in an oil rig application, edge inference identifies key outlier patterns and only transmits urgent notifications, significantly reducing the need for data transmission.)

The cloud provides overall optimization and long-term analysis by aggregating data from multiple locations.

V. CHALLENGES AND CONSIDERATIONS

In spite of the numerous beneficial properties that the edge computing offers, the edge computing also has a multiplicity of challenges and problems. Resource Constraints Edge devices and gateway: Edge devices and gateways are significantly weaker than their cloud counterparts and are less powerful in terms of CPU, memory, and storage. Particularly IoT sensors often are being subjected to severe battery or power budgets. The management of the effective offloading of tasks between devices and edge nodes is still a topic of ongoing research. This limitation of the edge nodes creates the need to carefully plan the time and assign resources to avoid congestion at one point. In adaptive IoT systems, the traditional methods of static allocation are unsatisfied, and scientific interest has turned to the use of AI-based scheduling solutions, e.g. reinforcement learning, to optimally assign workloads between the edge and cloud.

IoT-edge ecosystem is highly heterogeneous as devices may use a variety of hardware platforms (CPUs, GPUs, ASICs), operating systems and network interfaces. It is intrinsically difficult to come up with applications that will work well under this kind of diversity. The variety of platform variations often have to be addressed by the programmers in a way that makes their development and maintenance more difficult. The new solutions are either on standardizing the runtime environments (e.g., containerization) or on offering abstraction mechanisms to have a wide variety of edge node types where the code would execute.

Edge computing calculability redefines security. Although the risk of exposure could be reduced with local storage of sensitive data in comparison with the Internet, each edge device also turns into a possible point of attack. Getting end-to-end security in a system of thousands of remote nodes is challenging. Idleness of usage patterns, including power usage or camera views, can be used to predict personal knowledge, thus requiring the implementation of the concept of privacy by design in the edge architecture. Field devices also necessitate a high level of authentication, encryption and over-the-air update measures in terms of protecting edge deployments. Weaknesses in one of the nodes can affect the integrity of the whole system.

The coordination of a distributed edge network is a more difficult task when compared to the coordination of a centralized cloud. Providering edge devices, deploying software updates, health monitoring, and field failure are among the tasks. Most organizations do not have immature automated management tools on massive sets of edge nodes. Any hardware repairs can also be an issue with devices such as hardware that has been widely dispersed. Such managerial overheads come with costs and a cost engineer will have to factor this overhead in any IoT deployment program. • Consistency and Integration: The data and the state between the edge and the cloud cannot be maintained consistently and easily. IoT systems might be required to coordinate models or datasets across layers. Integrating hardware and cloud services of different vendors may be a problem. There also exist some regulated settings, like healthcare, finance, etc., which have a higher data processing requirements that make edge-cloud partitioning even more complex.

VI. FUTURE PROSPECTS

Edge-cloud computing is also occurring faster with the development of new technologies, which bring intelligence to the point of access to the data sources. The main driving force behind this change is the spread of the TinyML and lightweight edge artificial intelligence, which was enabled by the advances made to model compression and energy-efficient neural architecture. These innovations can support the most basic inference tasks (i.e., keyword spotting, anomaly detection and simple decision-making) running directly on low-power devices, lowering latency and eliminating the need to rely on centralized edge or cloud infrastructure, third, on-the-edge intelligence will form the basis of the future Internet of Things (IoT) deployments. Increased communication networks also support edge-computing. With the 5G roll-out and the continuous development of 6G, ultra-low latency, high-bandwidth, and ultra-reliable low-latency communication (URLLC) have become available that cannot be done without in real-time applications. These characteristics would support hybrid execution models in which workloads that need wide processing power can be offloaded to nearby

multi-access edge computing (MEC) servers and at the same time meet high response-time demands, especially in autonomous systems and industrial automation and the extended reality industry.
Cloud-native concepts are gradually infiltrating edge computing systems via serverless code and containerization. Applications deployed on dispersed edge nodes are uniformly deployed, scaled and managed using platforms like AWS IoT Greengrass and Azure IoT Edge. The frameworks achieve high scalability by providing a significant reduction in overhead operations due to the abstraction of infrastructure complexity. Even Distributed analytics and digital twins are expected to form the core of future IoT architectural designs. The physical systems can be optimistically monitored and simulated locally because an edge processor can continually update virtual images of real physical dependencies on sensory data, and because predictive monitoring is possible. This approach enhances responsiveness, manages network congestion and is as well. capable of facilitating deep-sea analysis on the cloud.

Finally, the AI-based orchestration solutions are emerging to address the increasing heterogeneity of the IoT environment. It is believed to dynamically allocate resources among devices, edge, and cloud infrastructure through techniques such as deep reinforcement learning. The self-adaptive architectures of the latter are designed to provide fault tolerance, focusing on resource efficiency, and plug-and-play simplicity, thus driving the IoT networks towards the autonomous era.

VII. CONCLUSION

In conclusion, edge and cloud computing are essential in real-time IoT. Edge computing involves moving computation closer to the sources of data, thus significantly reducing latency and network usage, which is required in real-time systems. Cloud computing provides massive centralized resources that perform massive analytics and massive storage. We note that edge has the lowest latency and bandwidth and limited local capacity, while cloud has the highest scalability but high latency and fixed data costs. A hybrid model would normally be the best: leverage the benefit to offer local and immediate processing and the cloud to offer in-depth analytics and coordination. This multi-layered strategy is becoming increasingly popular as IoT systems continue to grow in size. Finally, edge vs. cloud choices would depend on the applications required (real-time processing, volume of data, and privacy). This survey has outlined the current trade-offs in performance and pointed to future trends/directions, such as TinyML, 5G, and better orchestration that will shape the future convergence of edge and cloud in IoT.

REFERENCES

[1] R. K. Maddumala, “Comparison of Edge vs. Cloud Computing Architectures in IoT-Based Smart Agriculture Systems,” Global Journal of Engineering Innovations and Interdisciplinary Research, vol. 5, no. 6, pp. 121–132, 2025.
[2] A. Alkilany Dallaf, “Edge Computing in IoT Networks: Enhancing Efficiency, Reducing Latency, and Improving Scalability,” International Journal of Advanced Network Monitoring and Controls, vol. 10, no. 1, pp. 103–119, 2025.
[3] S. Alamouti, “Quantifying Energy and Cost Benefits of Hybrid Edge Cloud: Analysis of Traditional and Agentic Workloads,” arXiv preprint arXiv:2501.14823, 2024.
[4] K. Reddy and R. Kothinti, “Benchmarking Edge Computing vs. Cloud Computing for Real-Time Data Processing,” IRE Transactions on Electron Devices, vol. 7, no. 11, pp. 750–760, 2024.
[5]] F. C. Andriulo et al., “Edge Computing and Cloud Computing for Internet of Things: A Review,” Informatics, vol. 11, no. 4, art. 71, 2024.