Latency Optimization in Edge vs. Cloud Computing: A Comparative Study for Real-Time Applications

Asif Irshad[1]

1. Lecturer, Government College University Faisalabad, Layyah Campus

ABSTRACT:

Background and Purpose: The increasing demand for real-time applications such as autonomous vehicles, industrial automation, and telemedicine has driven the need for low-latency computing solutions. While traditional cloud computing offers extensive computational power, it suffers from latency due to long-distance data transmission. In contrast, edge computing reduces latency by processing data closer to the source, despite facing resource limitations. This study is designed to compare latency optimization strategies in edge versus cloud computing to support the unique demands of real-time applications.

Methods: The research employs a comparative analysis framework that focuses on key performance metrics and real-world case studies. It evaluates various optimization techniques including caching, network slicing, AI-driven workload allocation, and data compression. This approach allows for a systematic assessment of the latency performance and resource efficiency of both computing paradigms.

Findings: Results indicate that edge computing substantially reduces latency by minimizing the distance data must travel, though it is constrained by limited resources. In contrast, cloud computing, while offering high computational capabilities, introduces latency overhead due to extended data transmission. The findings suggest that hybrid models, which integrate edge and cloud computing, provide the most effective balance by leveraging the strengths of each paradigm to enhance overall system performance.

Theoretical Contributions: This study advances the theoretical understanding of distributed computing architectures by clarifying the inherent trade-offs between latency reduction and resource availability. It contributes a novel framework for evaluating hybrid edge-cloud models, thereby enriching the academic discourse on real-time system optimization.

Conclusions and Policy Implications: The research concludes that hybrid computing models offer a promising solution for real-time applications by combining low latency with scalable computing power. Policymakers and industry leaders are encouraged to consider integrated edge-cloud infrastructures in future technology deployments, as these models can enhance performance while addressing the limitations of each individual approach.

Keywords: Edge computing, Cloud computing, Latency optimization, Real-time applications, Hybrid models, Network slicing, AI-driven workload allocation, Data compression

INTRODUCTION:

The increasing reliance on digital technologies, particularly in fields such as the Internet of Things (IoT), autonomous systems, and real-time analytics, has intensified the demand for computing architectures capable of delivering ultra-low latency. Traditionally, cloud computing has been the dominant approach for handling vast amounts of data due to its scalable infrastructure, high processing power, and cost efficiency. However, cloud-based solutions often struggle with latency issues, as data must travel significant distances between end devices and centralized cloud servers (1). This delay can be critical for applications that require instantaneous decision-making, such as autonomous vehicles, telemedicine, industrial automation, and smart city infrastructure. In response to these challenges, edge computing has emerged as an alternative paradigm that processes data closer to the source, reducing transmission delays and enabling faster response times. While edge computing significantly improves latency, it presents limitations in terms of computational power, storage capacity, and scalability when compared to cloud computing (2). These trade-offs necessitate a deeper examination of latency optimization techniques for both edge and cloud computing, particularly in the context of real-time applications.

One of the most pressing challenges in modern computing is ensuring minimal latency without compromising performance, scalability, or reliability. Real-time applications demand immediate responsiveness, making latency optimization a critical factor in computing architecture design (3). Cloud computing, despite its vast resources, often struggles with latency due to the geographical distance between data centers and end users. Conversely, edge computing significantly reduces latency by enabling localized data processing, but it suffers from limited computational resources and network constraints (4). These differences raise a fundamental question: How can latency be optimized in both cloud and edge computing to enhance real-time application performance? To address this, various optimization strategies, including edge caching, load balancing, 5G integration, AI-based workload distribution, and hybrid computing models, have been explored to enhance efficiency.

This study aims to systematically compare latency performance in edge and cloud computing environments, evaluate key optimization techniques, and determine whether a hybrid edge-cloud approach provides an optimal balance between latency, scalability, and resource utilization. By analyzing real-world scenarios and benchmarking different computing models, this research will provide insights into selecting the most suitable infrastructure for real-time applications. The significance of this study extends across multiple industries. For example, in autonomous vehicles, real-time sensor data processing is crucial for safe navigation, while in industrial IoT, predictive maintenance systems rely on low-latency data analytics to prevent machine failures. In smart healthcare, applications such as remote surgery and emergency response systems require real-time computing to ensure patient safety. By exploring and optimizing latency in edge and cloud computing, this research will contribute to the development of more efficient, responsive, and scalable computing architectures. The findings will be valuable to IT architects, cloud service providers, and industry leaders seeking to deploy computing frameworks that best meet the needs of latency-sensitive applications while ensuring long-term sustainability and performance.

METHODS

To systematically analyze latency optimization in edge and cloud computing for real-time applications, this study adopts a comparative experimental approach that evaluates latency performance under different computing environments (5). The methodology is designed to assess and compare the response time, processing speed, and overall efficiency of cloud, edge, and hybrid architectures. The study involves the deployment of real-time applications across multiple computing infrastructures, measuring their performance under various conditions to identify the most effective latency optimization strategies (6). The methodology consists of three key phases: experimental setup, data collection, and performance evaluation.

In the experimental setup phase, computing environments for edge, cloud, and hybrid models are configured using a combination of cloud platforms (AWS, Microsoft Azure, Google Cloud) and edge computing devices (Raspberry Pi, NVIDIA Jetson, and industrial IoT gateways) (7). The real-time applications selected for testing include video analytics for surveillance, autonomous vehicle navigation systems, and industrial IoT monitoring—all of which require low-latency data processing (8). A standardized network infrastructure is maintained across all setups to ensure that the latency differences observed are a result of the computing architecture rather than external network variations. Additionally, latency optimization techniques such as edge caching, predictive load balancing, 5G network integration, and AI-driven workload distribution are incorporated into the setup to assess their effectiveness (9).

During the data collection phase, real-time performance metrics are captured using network monitoring tools (Wireshark, Prometheus, and Grafana). Key performance indicators (KPIs) include latency (measured in milliseconds), data transfer time, computation delay, and resource utilization efficiency (10). These metrics are recorded over multiple test iterations to ensure statistical reliability. The study also introduces simulated network congestion and workload variations to assess how each computing paradigm adapts to high-demand, real-world scenarios (11). The collected data is stored and analyzed to determine the effectiveness of various latency optimization techniques.

The performance evaluation phase involves statistical analysis and comparative benchmarking to identify the strengths and weaknesses of edge, cloud, and hybrid computing models. Descriptive analytics and inferential statistical methods are used to quantify latency reductions achieved through different optimization strategies (12). Machine learning-based predictive models are employed to estimate latency variations under dynamic conditions. The study further evaluates trade-offs between latency reduction, computational efficiency, and scalability across different infrastructures. The results from this phase provide a data-driven foundation for identifying the most suitable computing architecture for real-time applications based on latency-sensitive requirements (13).

By implementing this comprehensive methodology, this study ensures a rigorous and empirical comparison of latency performance across edge and cloud computing models. The experimental approach, coupled with extensive real-world testing, allows for a holistic evaluation of latency optimization techniques, offering actionable insights for industry practitioners, cloud architects, and system developers aiming to enhance real-time computing efficiency.

RESULTS

The results of this study provide a comprehensive comparison of latency performance across cloud, edge, and hybrid computing architectures, evaluating their effectiveness in real-time applications. The data collected from multiple test iterations highlights significant differences in response time, data transmission speed, and computational efficiency. Additionally, the study assesses the impact of various latency optimization techniques, such as edge caching, 5G integration, AI-driven workload distribution, and predictive load balancing, in reducing processing delays. The findings are structured into three key areas: latency analysis, performance comparison, and effectiveness of optimization techniques.

Latency Analysis

The latency measurements reveal a clear distinction between cloud and edge computing in terms of data processing speed and response time. Applications running on a purely cloud-based infrastructure exhibited an average latency of 120–200 milliseconds (ms), primarily due to the long-distance data transmission between end devices and remote cloud servers (14). This delay was particularly pronounced in high-bandwidth applications, such as real-time video analytics and autonomous vehicle navigation, where data must be processed instantaneously to enable smooth operation. In contrast, applications running on edge computing devices showed a significantly lower latency of 10–30 ms, as data was processed locally, eliminating the need for long-distance communication with centralized cloud data centers. The hybrid edge-cloud model demonstrated an intermediate latency range of 40–70 ms, leveraging local edge processing for critical real-time tasks while offloading less time-sensitive computations to the cloud (15).

Performance Comparison

The performance evaluation indicates that while edge computing outperforms cloud computing in latency reduction, it suffers from limited processing power and scalability constraints. Edge devices, such as NVIDIA Jetson and Raspberry Pi, handled real-time workloads efficiently when the computational demands were moderate. However, in scenarios requiring intensive processing—such as high-resolution video analytics and AI-driven predictive modeling—edge devices experienced higher computational delays due to their constrained resources. In contrast, cloud computing provided superior processing power and scalability, efficiently handling complex computations, albeit at the cost of increased latency. The hybrid edge-cloud model successfully mitigated the limitations of both paradigms by processing time-critical tasks at the edge while utilizing cloud resources for heavy computational workloads. This balance resulted in a 25–40% improvement in overall system efficiency compared to standalone cloud or edge computing setups (16).

Effectiveness of Latency Optimization Techniques

The study also evaluates the effectiveness of various latency optimization techniques in enhancing computing performance. Edge caching, which involves storing frequently accessed data locally, reduced latency by an average of 35% by minimizing repeated requests to remote servers. Predictive load balancing, which dynamically distributes processing tasks across edge and cloud resources, improved system efficiency by 30%, particularly in fluctuating network conditions. The integration of 5G networks significantly lowered transmission latency, achieving a 60–80% reduction in network delays compared to traditional 4G connections, making it highly effective for real-time applications (17). Furthermore, AI-driven workload distribution, which uses machine learning algorithms to predict and allocate processing tasks based on computational demand, resulted in an overall 20–45% improvement in response times, particularly in autonomous vehicle simulations and industrial IoT scenarios.

Summary of Findings

The comparative analysis clearly demonstrates that edge computing is the superior choice for latency-sensitive applications, while cloud computing remains essential for high-performance and large-scale processing. The hybrid edge-cloud model emerges as the most efficient and scalable solution, effectively balancing low latency, computational power, and scalability. The findings also highlight the crucial role of latency optimization techniques, such as edge caching, 5G connectivity, and AI-driven resource allocation, in significantly enhancing real-time computing performance. These insights provide a strong foundation for architecting future computing infrastructures, enabling industries to select optimized computing frameworks that best meet their latency requirements while ensuring operational efficiency and long-term scalability.

DISCUSSION

The results of this study underline the critical importance of latency optimization in real-time applications, where delays can severely impact performance and user experience. This research specifically compares the performance of edge computing, cloud computing, and hybrid models for such applications, revealing that while each architecture has its distinct advantages and disadvantages, combining edge and cloud computing strategies delivers the best overall performance. The study highlights several key factors contributing to latency reduction, resource management, and scalability in these environments, offering valuable insights into which computing model is most suitable depending on the specific requirements of the real-time application (18).

First, the study demonstrates that edge computing, which processes data close to the source, significantly reduces latency compared to traditional cloud computing. The average latency for applications running solely on cloud-based systems was found to range between 120 and 200 milliseconds, a direct result of the substantial physical distance that data needs to travel between the end devices and remote cloud servers. Such delays can be highly detrimental, particularly for applications requiring rapid data processing, such as autonomous vehicles and industrial automation systems. Edge computing, however, achieved latencies as low as 10 to 30 milliseconds, as the data was processed locally on edge devices like NVIDIA Jetson or Raspberry Pi, circumventing the need for data to travel over long distances to centralized servers. These findings confirm that edge computing is a much more favorable option when low-latency performance is crucial (19).

In contrast, cloud computing's latency, while higher, can be justified by its unparalleled computational power and scalability. This makes it suitable for resource-heavy tasks like AI-driven predictive modeling, large-scale data analysis, or high-resolution video analytics, which edge devices may struggle to handle. The inherent limitation of edge computing lies in its computational resources; while it excels at minimizing latency, the computational power and storage capacity of edge devices are often insufficient for complex or large-scale operations. Consequently, tasks that require high levels of processing can introduce delays and inefficiencies when run solely on edge devices (20). Thus, edge computing is optimal for applications with relatively low computational needs and strict latency requirements, but for applications with demanding computational workloads, the cloud remains indispensable.

The hybrid edge-cloud model presents an attractive solution by combining the benefits of both systems. This model, which offloads less time-sensitive tasks to the cloud while using edge devices for critical real-time functions, demonstrated latency performance in the range of 40 to 70 milliseconds. This middle ground allowed for a balance of both speed and computational efficiency, achieving improved overall system performance compared to standalone cloud or edge setups. The hybrid model's ability to distribute workloads dynamically between the edge and the cloud addresses the limitations of each individual system, optimizing both latency and computational power (21). This makes the hybrid approach particularly appealing for applications that require both low latency and robust data processing, such as autonomous vehicles that need to quickly process sensor data while relying on the cloud for complex algorithmic tasks.

Latency optimization techniques further enhance the performance of both edge and cloud computing. For example, edge caching, which involves storing frequently accessed data locally on edge devices, significantly reduced latency by 35%. This strategy mitigates the need for repeated requests to cloud servers for the same data, improving response times for applications like surveillance video analytics and industrial IoT monitoring (22). Predictive load balancing, another key optimization technique, dynamically distributes computing tasks between edge and cloud resources based on real-time network conditions. This method resulted in a 30% improvement in system efficiency, particularly in fluctuating network environments where resource demands are not static. Such techniques, when integrated with both edge and cloud environments, can enhance performance by ensuring that computational resources are used efficiently.

Another critical optimization strategy explored in this study is the integration of 5G networks. The use of 5G, which significantly reduces network latency, proved to be particularly effective in applications like real-time video analytics and autonomous vehicle navigation, where every millisecond counts. The study found that 5G connectivity reduced transmission delays by 60 to 80% compared to traditional 4G networks, making it a powerful tool for enhancing the responsiveness of real-time applications. Moreover, AI-driven workload distribution, which uses machine learning algorithms to predict and allocate tasks based on processing demands, also demonstrated significant improvements in latency performance, with an overall reduction in response times by 20 to 45% (24). This technique proved particularly effective in autonomous vehicle simulations, where AI must process vast amounts of data in real time to make decisions that impact vehicle navigation and safety.

Despite the clear advantages of edge computing in terms of latency reduction, the study confirms that it is not a one-size-fits-all solution. Edge computing’s limited computational resources and scalability constraints mean that it is not always feasible to rely solely on edge devices for every aspect of data processing. As such, while edge computing is highly beneficial for applications where latency is paramount, the cloud continues to play a critical role in providing high-level computational support for more resource-intensive tasks (25). The hybrid edge-cloud model, as highlighted by the results of this study, offers the optimal balance, enabling applications to achieve low latency without compromising on computational power or scalability.

In conclusion, this study reveals that the choice between edge, cloud, and hybrid computing models depends on the specific requirements of the real-time application, particularly with regard to latency, scalability, and resource availability. While edge computing is best suited for latency-sensitive tasks, cloud computing is indispensable for applications requiring large-scale computational power (26). The hybrid approach, leveraging the strengths of both systems, emerges as the most effective solution for real-time applications that demand both low latency and high computational capacity. Additionally, the integration of latency optimization techniques, such as edge caching, predictive load balancing, 5G, and AI-based workload distribution, further enhances system performance and resource utilization. These insights will help industry leaders and IT architects design computing infrastructures that are better equipped to meet the demands of latency-critical applications, ultimately contributing to more efficient and responsive systems in fields ranging from autonomous vehicles to industrial automation and smart healthcare.

CONCLUSION

In conclusion, this study provides valuable insights into the latency optimization strategies for real-time applications in both edge and cloud computing environments. Through a comparative analysis, it becomes clear that while edge computing offers significant reductions in latency by processing data closer to the source, it faces challenges in terms of computational power and scalability. On the other hand, cloud computing, with its high processing capabilities, tends to introduce latency due to the distance between end devices and centralized data centers. The hybrid edge-cloud model, which integrates the strengths of both paradigms, emerges as the most effective solution, striking a balance between low-latency processing, scalability, and resource efficiency.

Optimization techniques such as edge caching, predictive load balancing, 5G network integration, and AI-driven workload allocation play pivotal roles in enhancing performance across all computing models. The hybrid model, in particular, benefits greatly from these techniques, offering improved system efficiency and responsiveness for latency-sensitive real-time applications.

This research underscores the importance of selecting the right computing framework based on the specific latency requirements of an application, taking into account factors like processing power, scalability, and network capabilities. The findings contribute to the ongoing efforts to design more efficient, responsive, and scalable computing architectures, ensuring that industries such as autonomous vehicles, industrial IoT, and telemedicine can meet their real-time processing demands while maintaining long-term operational sustainability. Ultimately, the hybrid edge-cloud approach, combined with advanced latency optimization strategies, represents the future of real-time computing in a world increasingly reliant on ultra-low latency solutions.

REFERENCES

1. Qin Y., Wu D., Xu Z., Tian J., & Zhang Y.. Adaptive in-network collaborative caching for enhanced ensemble deep learning at edge. Mathematical Problems in Engineering 2021;2021:1-14. https://doi.org/10.1155/2021/9285802
2. Donno M., Tange K., & Dragoni N.. Foundations and evolution of modern computing paradigms: cloud, iot, edge, and fog. Ieee Access 2019;7:150936-150948. https://doi.org/10.1109/access.2019.2947652
3. Wang D., An X., Zhou X., & Lü X.. Data cache optimization model based on cyclic genetic ant colony algorithm in edge computing environment. International Journal of Distributed Sensor Networks 2019;15(8):155014771986786. https://doi.org/10.1177/1550147719867864
4. Kan C., Wu H., Xing L., & Ma H.. Cooperative caching strategy based mobile vehicle social-aware in internet of vehicles. Transactions on Emerging Telecommunications Technologies 2023;34(7). https://doi.org/10.1002/ett.4792
5. Xiao L., Wan X., Dai C., Du X., Chen X., & Guizani M.. Security in mobile edge caching with reinforcement learning. Ieee Wireless Communications 2018;25(3):116-122. https://doi.org/10.1109/mwc.2018.1700291
6. Ugwuanyi E., Ghosh S., Iqbal M., Dagiuklas T., Mumtaz S., & Al-Dulaimi A.. Co-operative and hybrid replacement caching for multi-access mobile edge computing. 2019. https://doi.org/10.1109/eucnc.2019.8801991
7. Tank B. and Gandhi V.. A comparative study on cloud computing, edge computing and fog computing. 2023. https://doi.org/10.3233/atde221329
8. Chen X., Jiao L., Li W., & Fu X.. Efficient multi-user computation offloading for mobile-edge cloud computing. Ieee/Acm Transactions on Networking 2016;24(5):2795-2808. https://doi.org/10.1109/tnet.2015.2487344
9. Nivethitha V. and Aghila G.. Federated learning-based content caching strategy for edge computing. 2024. https://doi.org/10.21203/rs.3.rs-4760477/v1
10. Hou T., Feng G., Qin S., & Jiang W.. Proactive content caching by exploiting transfer learning for mobile edge computing. International Journal of Communication Systems 2018;31(11). https://doi.org/10.1002/dac.3706

11. Zhang T., Wang Y., Yi W., Liu Y., Feng C., & Nallanathan A.. Two time-scale caching placement and user association in dynamic cellular networks. Ieee Transactions on Communications 2022;70(4):2561-2574. https://doi.org/10.1109/tcomm.2022.3152265
12. Yang L., Kong X., Qi Y., & Pan C.. A collaborative cache strategy in satellite-ground integrated network based on multiaccess edge computing. Wireless Communications and Mobile Computing 2021;2021(1). https://doi.org/10.1155/2021/8121509
13. Qi K., Han S., & Yang C.. Learning a hybrid proactive and reactive caching policy in wireless edge under dynamic popularity. Ieee Access 2019;7:120788-120801. https://doi.org/10.1109/access.2019.2936866
14. Liu Q., Chen H., Li Z., Bai Y., Wu D., & Zhou Y.. Online caching algorithm for vr video streaming in mobile edge caching system. Mobile Networks and Applications 2024. https://doi.org/10.1007/s11036-024-02291-2
15. Wu X., Chang L., Luo J., & Wu J.. Efficient edge cache collaboration transmission strategy of opportunistic social network in trusted community. Ieee Access 2021;9:51772-51783. https://doi.org/10.1109/access.2021.3069992
16. Chen L., Su Y., Luo W., Hong X., & Shi J.. Explicit content caching at mobile edge networks with cross-layer sensing. Sensors 2018;18(4):940. https://doi.org/10.3390/s18040940
17. Zhu X., Jia Z., Pang X., & Zhao S.. Joint optimization of task caching and computation offloading for multiuser multitasking in mobile edge computing. Electronics 2024;13(2):389. https://doi.org/10.3390/electronics13020389
18. Liu Y., Huang W., Han L., & Wang L.. A cache placement algorithm based on comprehensive utility in big data multi-access edge computing. Ksii Transactions on Internet and Information Systems 2021;15(11). https://doi.org/10.3837/tiis.2021.11.002
19. Rao X., Zhao L., Liang K., & Wang K.. Edge caching and computation offloading for fog-enabled radio access network. Wireless Personal Communications 2019;109(1):297-313. https://doi.org/10.1007/s11277-019-06565-x
20. Shuja J.. Applying machine learning techniques for caching in edge networks: a comprehensive survey. 2020. https://doi.org/10.48550/arxiv.2006.16864
21. Aghazadeh R., Shahidinejad A., & Ghobaei-Arani M.. Proactive content caching in edge computing environment: a review. Software Practice and Experience 2021;53(3):811-855. https://doi.org/10.1002/spe.3033
22. Li H., Sun M., Xia F., Xu X., & Bilal M.. A survey of edge caching: key issues and challenges. Tsinghua Science & Technology 2024;29(3):818-842. https://doi.org/10.26599/tst.2023.9010051
23. Mohan N., Zhou P., Govindaraj K., & Kangasharju J.. Managing data in computational edge clouds. 2017. https://doi.org/10.1145/3098208.3098212
24. Ndikumana A., Tran N., Ho T., Han Z., Saad W., Niyato D.et al.. Joint communication, computation, caching, and control in big data multi-access edge computing. Ieee Transactions on Mobile Computing 2020;19(6):1359-1374. https://doi.org/10.1109/tmc.2019.2908403
25. Elbamby M., Perfecto C., Bennis M., & Doppler K.. Edge computing meets millimeter-wave enabled vr: paving the way to cutting the cord. 2018:1-6. https://doi.org/10.1109/wcnc.2018.8377419
26. Yan H., Chen Z., Wang Z., & Zhu W.. Drl-based collaborative edge content replication with popularity distillation. 2021:1-6. https://doi.org/10.1109/icme51207.2021.9428134