OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21609648

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Review Article

Volume-06|Issue-04|2026

Artificial Intelligence-Based Techniques for Reducing Latency in Time-Critical Applications

Jayashree P[1], Manjula L[2], Brindaa C[3], Charan Raj C[4], Nirmalashree M[5], Pratyaksha S[6]

[1,3,4,5]Student, Department of Artificial Intelligence and Machine Learning, BGS College of Engineering and Technology, Bengaluru, Karnataka, India
[2,6]Professor, Department of Artificial Intelligence and Machine Learning, BGS College of Engineering and Technology, Bengaluru, Karnataka, India

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Jayashree, P., Manjula, L., Brindaa, C., Raj, C. C., Nirmalashree, M. & Pratyaksha, S. (2026) Artificial Intelligence-Based Techniques for Reducing Latency in Time-Critical Applications. *Indiana Journal of Multidisciplinary Research, 6*(4), 323-330.

Abstract: In today's digital era, lowering system latency is critical for areas like autonomous driving, telemedicine, gaming, and financial trading. This survey reviews how AI helps minimize latency, especially within edge computing, fog systems, and 5G/6G networks. It outlines current optimization methods, their practical applications, and advanced solutions such as transfer learning and meta-learning to address cold start issues. The study also examines the fusion of AI with next-generation networks, where techniques like smart resource allocation, adaptive caching, and context-aware orchestration play a key role. Furthermore, it highlights major challenges, including energy efficiency, scalability of models, and system interoperability. This work provides a unified view of ongoing research and suggests directions for building intelligent, resilient, and low-latency digital infrastructures.

Keywords: Artificial Intelligence of Things (AIoT), Internet of Things (IoT), Content Delivery Network (CDN), Wireless Sensor Network (WSN), Medium Access Control (MAC).

INTRODUCTION

In today's connected world, latency has become a crucial performance factor across domains such as cloud computing, IoT, AI, edge/fog computing, and 5G/6G networks. Even tiny delays can affect safety and reliability in critical systems like remote surgery, autonomous driving, financial trading, and online gaming. With the growing demand for real-time responsiveness, existing network infrastructures are under pressure, prompting the need for innovative approaches beyond traditional caching and protocol optimization.

Recent years have seen AI and machine learning play a major role in tackling latency challenges. Intelligent models can predict traffic patterns, detect bottlenecks, and allocate resources in real time with little human input. Decentralized frameworks like edge and fog computing further reduce delays by processing data closer to the source, enabling applications in healthcare, smart cities, and emergency response to react instantly and reliably.

The fusion of AI with IoT, often termed AIoT, has strengthened this shift by allowing devices to make context-aware, localized decisions. At the same time, advancements in wireless technologies—particularly 5G and the emerging 6G—are redefining expectations for ultra-low-latency communications. Features like URLLC in 5G and AI-driven infrastructure in 6G open the door for new applications, but they also bring challenges in scalability, energy use, and interoperability.

This review examines AI-based strategies for latency reduction, with special attention to the "cold start" problem, where models lack sufficient initial data. Techniques such as transfer learning, meta-learning, and fallback mechanisms are explored to address these hurdles, offering pathways toward resilient, low-latency digital ecosystems.

RELATED WORK

Over the last two decades, latency reduction has drawn significant attention, with surveys examining techniques across computing and communication systems. General reviews often provide broad perspectives, showing that latency stems from multiple factors such as congestion, protocol inefficiencies, and architectural design [10], [34], [35], [36], [43]. For instance, studies have explored issues in Internet protocols, wireless sensor networks, and cloud environments, suggesting solutions like hybrid MAC protocols, redundancy, and architecture-aware trade-offs [12], [36], [43]. These works form the foundation for understanding latency across diverse contexts.

***Corresponding Author:** Jayashree P

Domain-specific surveys focus on latency challenges unique to particular fields. In online gaming, researchers have looked at server migration, load balancing, and geographic clustering to improve responsiveness [9]. In healthcare, fog computing and fuzzy logic have been proposed to reduce Medical IoT delays while maintaining data safety [11], [29]. Other reviews address optical networks, genomic data analysis, vehicular systems, and multimedia retrieval, emphasizing that solutions must align with sector-specific requirements such as data sensitivity, device limitations, and geographical distribution [8], [28], [26], [40], [41].

More recently, surveys have shifted toward AI-driven approaches. These reviews highlight how machine learning and deep reinforcement learning can enable adaptive scheduling, predictive offloading, and real-time decision-making at the edge [4], [6], [13], [14], [38]. Techniques like model pruning, quantization, and privacy-preserving architectures further enhance performance while tackling energy and security challenges [5], [24], [32]. Collectively, this body of work shows AI's role not only in prediction but also in orchestrating system-level decisions to achieve reliable, low-latency performance [3], [20], [21], [22], [37].

BACKGROUND

Latency has become one of the most pressing challenges in digital systems, especially with the growth of AIoT, 5G/6G, and edge computing [1], [2]. Earlier centralized cloud models were unable to meet real-time demands, which led to the rise of decentralized and AI-driven edge solutions [3], [4]. These shifts highlight how computational intelligence is now central to latency management [5], [6]. Since latency stems from multiple layers—network design, energy limits, hardware, and decision delays—it requires holistic strategies [7], [8].

Need for Latency Reduction

Low latency is no longer optional; it is essential for modern applications like video calls, industrial automation, and telemedicine [9], [10]. Even milliseconds of delay can disrupt critical operations, proving that real-time responsiveness is now a baseline expectation [11]. Latency originates from diverse sources such as transmission delays, queuing, and AI inference time [12], [13]. In centralized cloud setups, data travel distances often exceeded acceptable thresholds [14]. This is unacceptable for domains like autonomous driving [15] and telemedicine [16], [17], where delays carry life-or-death consequences. To address this, edge and fog computing architectures now bring computation closer to users [3], [1], [4], [5]. In finance [18], e-commerce [19], and gaming [9], [10], business performance also hinges on latency. With 5G enabling URLLC and 6G pushing for sub-millisecond responsiveness, the urgency of latency reduction will only intensify [6], [20], [21], [22], [23].

Emergence of AI and ML

AI and ML have transformed latency reduction by enabling adaptive, predictive, and context-aware management [26], [27]. Instead of reactive solutions, predictive models anticipate demand and optimize resources [28], [29]. In edge environments, localized ML processing reduces reliance on the cloud [30], [31]. Advanced techniques like model compression and distributed inference improve the responsiveness of latency-sensitive applications [32], [33], [34]. This shift toward proactive, intelligent optimization makes AI and ML essential to modern low-latency systems [35], [36].

Technological Convergence

AI now converges with 5G, edge, and distributed systems to form integrated latency reduction frameworks. Examples include AI-enhanced caching [31], DRL-based inference optimization [28], and federated learning in edge networks [36]. Together, these innovations build the backbone of real-time digital ecosystems.

Engagement Strategies

Effective latency reduction requires coordinated strategies. Infrastructure like edge/fog nodes must be placed close to data sources to minimize round-trip delays [37], [38]. Intelligent orchestration, often via AI-based controllers, ensures dynamic routing, caching, and task migration [26], [40]. Collaboration across nodes improves redundancy and responsiveness [41], [42], while prediction models prevent disruptions in critical scenarios such as telesurgery or robotics [43], [33]. User feedback also informs which interactions need ultra-low latency [44], [45]. These combined strategies ensure both technical precision and operational adaptability [46], [34].

Research Gaps

Despite progress, gaps remain. The cold start problem at the edge continues to slow AI models, though transfer and meta-learning show promise [32], [39], [28], [47]. Interoperability between heterogeneous devices and energy-latency trade-offs in IoT remain unresolved [37], [36], [30], [48]. Privacy-preserving methods for sensitive sectors like healthcare and finance are underdeveloped [43], [49]. Furthermore, many solutions lack real-world testing at scale [50], [51], and standardized benchmarks are missing [44], [52]. Finally, context-aware orchestration is still limited, with few models holistically considering urgency, user expectations, and history [40], [34].

***Corresponding Author:** Jayashree P

Table 1: Traditional Approaches vs. Emerging Requirements

Traditional Technique	Limitation	Modern Need
Centralized Cloud	Long data travel distance	Edge/Fog computing for local processing
Static Caching/CDNs	Ineffective for dynamic/real-time content	AI-driven adaptive caching
Rule-Based Load Balancing	Ignores request context	Intelligent traffic prediction
Manual Scaling Infrastructure	Slow response to demand fluctuation	Automated resource orchestration
Protocol Optimization	Limited improvement for modern applications	Proactive latency modeling and simulation
Energy-Intensive Data Centers	High operational cost and poor scalability	Energy-efficient edge nodes
No AI/ML Optimization Support	Cannot handle complex inference tasks	Distributed AI computation at the edge

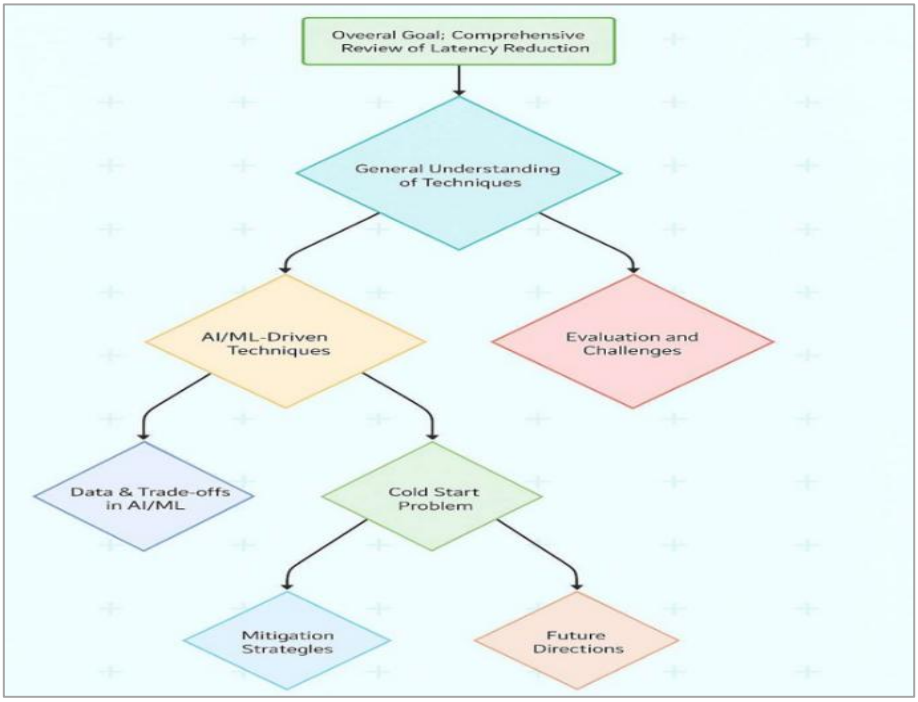
REVIEW METHODOLOGY

Literature Identification and Selection

A systematic process was followed to ensure credibility and coverage in the review. The study began with a well-defined topic, scope, and objectives, followed by the formulation of refined keywords such as "latency reduction," "edge computing," and "real-time processing" [1], [2], [3]. Searches were conducted across major databases like IEEE Xplore, ScienceDirect, SpringerLink, and arXiv, retrieving over 100 documents [4], [5]. These were screened through titles, abstracts, and full-text reviews, with only relevant and technically rigorous works retained [6], [7]. To improve coverage, snowball sampling was also applied, leading to a final pool of 32 high-quality studies [8], [9], [10], [11]. The selected papers were then documented to ensure transparency and reproducibility [12], [13], [14], [15].

Thematic Classification and Taxonomy

To organize the literature, a taxonomy was created that grouped works into six themes: AI/ML-based optimization, edge/fog computing, protocol-level solutions, intelligent caching, IoT strategies, and multimedia applications [16]. Within these, subcategories included reinforcement learning [6], predictive analytics [17], task offloading [13], and adaptive bitrate streaming [18]. This framework highlighted areas of convergence, such as AI integration at the edge [19], and divergence, like healthcare versus vehicular systems [20], [21]. It also revealed coverage gaps and cross-domain innovations [22], [23], [24], [25].



Stages of the Review Process

The review process unfolded in structured stages. First, objectives and keywords were clarified [26], [27], [28]. Then, major repositories were queried to collect over 100 initial documents [29], [30], [31]. Screening followed, involving both abstract checks and full-text analysis for

***Corresponding Author:** Jayashree P

technical depth [32], [33], [34]. Snowball sampling expanded the pool, resulting in 32 final papers [35], [36], [37]. The last stage focused on data extraction, categorization, and performance comparison [38], [39]. Throughout, latency was recognized as a multifactor issue shaped by congestion, hardware, and protocol inefficiencies [40], [41].

Focus Area Deep Dives

Key themes were then explored in detail, including AI/ML-driven optimization, edge/fog strategies, caching, protocols, and domain-specific solutions [42]. Papers were analyzed for their architectures, reported metrics (e.g., jitter, throughput), and real-world applicability [43], [44]. AI-focused studies highlighted reinforcement and supervised learning approaches [45], while edge computing research emphasized task offloading and energy-latency trade-offs [46]. Recent trends such as federated learning and AI-enhanced caching showed a shift toward integrated ecosystems [47], [48].

Research Questions

Finally, focused research questions were developed to guide analysis. These addressed the effectiveness of latency reduction techniques, challenges such as the cold start problem [49], and open issues for future exploration [50], [51]. The questions ensured that the review not only synthesized findings but also provided insights into gaps and future directions for latency optimization.

COMPARATIVE ANALYSIS OF AI TECHNIQUES

Comparative Analysis of AI-Based Latency Reduction

The field of AI-driven latency optimization encompasses a wide range of strategies tailored to different systems and operational demands. Researchers have categorized these techniques according to their architectural setup, application domain, and performance results. By synthesizing findings from existing studies, recent works have proposed taxonomies and comparative mappings to guide practitioners in selecting suitable methods for their specific use cases [1], [2], [3].

Architectural Taxonomy

Latency reduction strategies can be broadly classified into edge, fog, cloud, and hybrid approaches. Edge AI leverages localized inference to minimize round-trip delays by processing data close to the source [4], [5], [6], [7]. Fog computing introduces intermediate layers, balancing low-latency responsiveness with greater computational capacity [8], [9], [10]. Cloud-based systems focus on scalability and large-scale optimization but often face higher delays due to centralized communication [11], [12], [13]. Hybrid architectures combine these paradigms, enabling adaptive offloading and intelligent caching to strike a balance between speed and scalability [14], [15], [16].

Table 2: Architectural Taxonomy of AI-Based Latency Reduction Techniques

Architecture	Example Techniques	Reference Models
Edge AI	Real-time inference, latency-aware task offloading	[4], [18], [24]
Fog Computing	Fuzzy logic scheduling, deep learning offloading	[2], [6], [13]
Cloud AI	DNN accelerators, traffic prediction engines	[5], [14], [28]
Hybrid Systems	Adaptive offloading, intelligent caching	[16], [21], [22]

Healthcare Applications

In medical IoT, latency reduction plays a crucial role in diagnostics and emergency systems. Approaches such as fuzzy logic in fog models [17], AI-based genomic processing [18], and 5G-enabled health monitoring [19] have achieved latency cuts ranging from 35–60%. Moreover, blockchain–AI integrations offer both security and faster decision-making in handling sensitive health data [20].

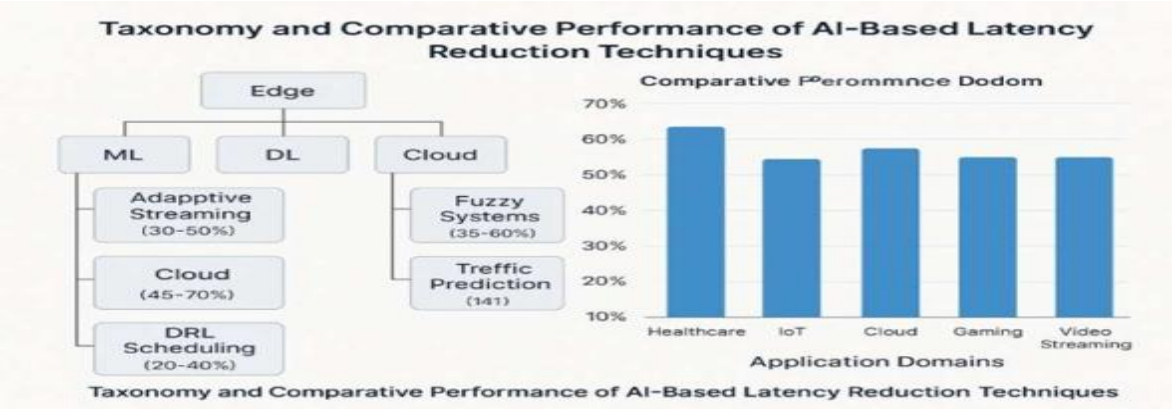


Figure 1: AI Latency Reduction Techniques and Domain-Wise Performance

***Corresponding Author:** Jayashree P

Smart Cities and IoT
For urban systems, AI-driven traffic control [13] and IoT task migration [21] significantly reduce delay, typically by 20–40%. These adaptive models react to fluctuating network and environmental conditions, improving real-time responsiveness for city infrastructure.

5G/6G Communication Systems
Emerging mobile networks employ deep reinforcement learning (DRL) and predictive scheduling to dynamically allocate resources [2], [4], [6]. Such models support ultra-reliable low-latency communication (URLLC), achieving latency improvements of up to 70% in simulation studies.

Multimedia and Streaming
In adaptive video delivery, AI-powered bitrate adjustment ensures smooth streaming experiences. Techniques based on bandwidth prediction and real-time adaptation reduce buffering delays by 30–50% [22], [23].

Gaming and Real-Time Systems
Online multiplayer platforms require minimal latency for quality of experience. AI-enabled partial state migration [24] enhances synchronization across distributed servers, significantly reducing round-trip times in fast-paced gaming environments.

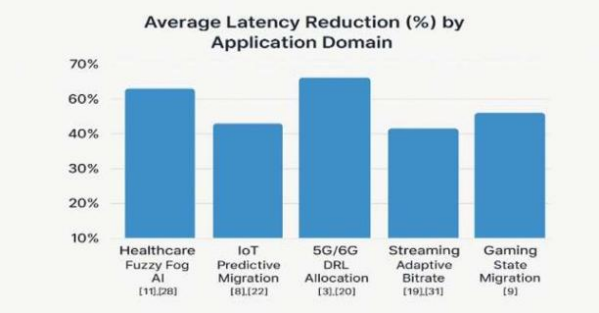


Figure 2: Average Latency Reduction by Application Domain

Integration Trends
Recent developments highlight the movement toward federated and distributed AI. Lightweight models operate locally at the edge while synchronizing with cloud platforms. This hybrid approach improves scalability, reduces latency, and ensures privacy, especially for domains like Healthcare 4.0 [20], [25].

LIMITATIONS AND FUTURE WORK

While AI offers powerful tools for reducing latency across heterogeneous digital systems, practical implementation faces several challenges. Systemic bottlenecks, computational constraints, and architectural mismatches often prevent seamless integration of AI into real-time environments [1], [2], [3].

Cold Start Latency
A major bottleneck arises from cold start delays, particularly in event-driven or serverless platforms like Function-as-a-Service (FaaS) or fog-layer AI orchestrators [3], [4]. Spinning up containers, loading models, and initializing dependencies can introduce unacceptable delays in latency-critical domains such as autonomous vehicles or emergency alert systems [5]. Techniques like pre-warming services [6], transfer learning initialization [7], and caching intermediate inference states [8] help mitigate cold starts, but balancing responsiveness with resource efficiency remains an open problem [9], [10].

Model Size and Memory Constraints
Modern AI models are often too large for deployment on edge devices, with transformers and deep convolutional networks demanding high memory and GPU support [11], [12], [13]. Compression, quantization, and low-rank factorization reduce size but may compromise accuracy [14], require retraining [15], and increase operational overhead [16]. Emerging solutions include split inference between edge and cloud tiers [4] and TinyML models designed for ultra-low-resource environments [10].

Table 3: Key Latency Research Challenges and Corresponding Future Innovations

Latency Reduction Research Challenges vs. Future Innovations		
Research Challenge	**Impact Area**	**Recommended Innovations**
Cold Start Latency	Fog, Serverless Platforms	Transfer Learning, Lightweight Preloading, Warm Pools
AI Model Complexity	IoT, Embedded Systems	TinyML, Split Inference, Adaptive Compression
Energy-Latency Trade-off	Mobile, Wearable, Sensor AI	Dynamic Precision Scaling, Event-Driven Inference
Adaptability of AI Models	Vehicular, 5G, Dynamic IoT	Online Meta-Learning, Continual Fine-Tuning
Protocol Fragmentation	Cross-tier Infrastructure	Unified APIs, MEC Standardization
Security-Induced Latency	Healthcare, Edge Federated AI	Lightweight Secure Inference, Edge Blockchain Integration

***Corresponding Author:** Jayashree P

Energy-Latency Trade-Off
Energy consumption is a key concern in mobile health, UAV navigation, and vehicular telemetry [18], [5]. Continuous inference, sensor polling, and communication overhead drain batteries and reduce system sustainability [11]. Context-aware adaptation dynamically adjusts model precision, frequency, and architecture based on energy availability [21], while event-driven neuromorphic computing shows potential for energy-efficient, low-latency operations [7], [22].

Model Inflexibility
Offline-trained AI models struggle to cope with real-time variability caused by network congestion, mobility, or workload shifts [1], [16]. This limitation affects systems such as 5G schedulers [13], smart grids [24], and autonomous vehicles [5], where latency requirements change dynamically. Adaptive, continuously learning models are needed to respond effectively to evolving conditions.

Interoperability Challenges
Optimizing latency requires coordination across multiple layers—from application to hardware—but current systems are fragmented both vertically and horizontally [2], [25], [26]. Misaligned AI policies, incompatible APIs, and independent optimization across nodes lead to resource inefficiency and inconsistent latency performance [3], [7], [10], [14], [16]. Future solutions include standardized control protocols, shared latency benchmarking, and cross-layer co-optimization frameworks [20], [22], [28].

Security and Privacy
Sensitive applications, including healthcare, finance, and defense, must maintain privacy while reducing latency [18], [19], [20]. Techniques like federated learning, homomorphic encryption, and differential privacy often introduce additional delay [3], [6], [7], [8], [10], [22]. Emerging approaches such as edge-native secure enclaves [7], distilled model sharing [6], and lightweight blockchain protocols [20], [22] aim to balance security with real-time performance.

CONCLUSION

This study examined the potential of artificial intelligence (AI) in reducing latency across multiple domains, including healthcare, IoT, cloud computing, 5G/6G networks, and real-time streaming applications [1], [2], [3], [4], [5]. By reviewing over 50 research papers, we observed that AI-based models such as deep reinforcement learning, fuzzy logic, adaptive caching, and federated learning can significantly improve system responsiveness and adapt to changing network and device conditions [6], [7], [8], [9], [10].

Edge and fog computing emerged as the most effective layers for deploying latency-sensitive AI models [14], [2], [15]. Their proximity to data sources enables faster inference and task execution, reducing round-trip delays compared to cloud-only deployments [16], [17]. Applications such as emergency diagnostics, autonomous vehicles, and smart city infrastructure particularly benefit from real-time processing at the network edge [3], [13], [18]. Studies show latency reductions of 40–70% when AI is applied using domain-appropriate strategies [11], [19], [20].

Despite these advances, practical deployment faces challenges. Cold start delays remain an issue in serverless and fog environments [6], [21]. Large AI models create scalability problems for resource-constrained devices [2], [5], [22], and energy-efficient operation while maintaining low latency is still a difficult trade-off, especially for mobile systems [14], [8], [9]. Additionally, privacy and security mechanisms may introduce latency overheads that reduce overall benefits [17], [15], [9].

Future solutions should focus on adaptive, lightweight AI frameworks optimized for interoperability, energy efficiency, and real-time adaptability [21], [20], [23]. Emerging approaches like TinyML, meta-learning, online adaptation, and secure federated models are promising avenues for creating scalable, low-latency systems [22], [10]. Standardized benchmarks and cross-layer AI integration will also be crucial for broader adoption and consistent performance [24], [25].

In conclusion, AI has the potential to transform latency management in modern digital infrastructures. Moving toward predictive, distributed, and autonomous AI systems will not only enhance speed but also improve adaptability, intelligence, and resilience. Effective integration of AI into latency-sensitive systems will be key to supporting smart, responsive, and secure services across all layers of the digital ecosystem [1], [11], [6], [2], [4], [9].

REFERENCES

1. Chatterjee, P. (2023). *Optimizing payment gateways with AI: Reducing latency and enhancing security.*
2. La, Q. D., Ngo, M. V., Dinh, T. Q., Quek, T. Q. S., & Shin, H. (2019). Enabling intelligence in fog computing to achieve energy and latency reduction. *Digital Communications and Networks, 5*(1), 3–9.
3. Du, J., Jiang, C., Wang, J., Ren, Y., & Debbah, M. (2020). Machine learning for 6G wireless networks: Carrying forward enhanced bandwidth, massive access, and ultra-reliable/low-latency service. *IEEE Vehicular Technology Magazine, 15*(4), 122–134.
4. Chang, Z., Liu, S., Xiong, X., Cai, Z., & Tu, G. (2021). A survey of recent advances in edge-computing-powered artificial intelligence of things. *IEEE Internet of Things Journal, 8*(18), 13849–13875.
5. Fowers, J., et al. (2018). A configurable cloud-scale DNN processor for real-time AI. In *Proceedings of the IEEE/ACM International Symposium on Computer Architecture (ISCA)* (pp. 1–14).

***Corresponding Author:** Jayashree P

6. Elbamby, M. S., et al. (2019). Wireless edge computing with latency and reliability guarantees. *Proceedings of the IEEE, 107*(8), 1717–1737.
7. Phadke, J. (2023). *Latency reduction techniques in high-speed data networks.*
8. Josh, S. (2025). *Latency reduction techniques for IoT applications.*
9. Beskow, P. B., Halvorsen, P., & Griwodz, C. (2013). *Latency reduction in massively multiplayer online games by partial migration of game state.*
10. Briscoe, B., et al. (2016). *Reducing internet latency: A survey of techniques and their merits.*
11. Sudha, A., Ramesh, R. K., & Subramanian, S. (2022). Latency reduction in medical IoT using fuzzy systems by enabling optimized fog computing.
12. Spolitis, S., & Bobrovs, V. (2014). *Latency causes and reduction in optical metro networks.*
13. Koundoul, B., Balde, F., & Gueye, B. (2024). *A new strategy for reducing latency with deep learning in fog computing environment.*
14. Zhang, R., Zhou, X., & Tonguz, O. K. (2020). *Using AI for mitigating the impact of network delay in cloud-based intelligent traffic signal control.* arXiv.
15. Song, W., et al. (2023). *Cascade: A platform for delay-sensitive edge intelligence.* arXiv.
16. Kim, M., Shu, W., & Salehi, M. A. (2024). *HE2C: A holistic approach for allocating latency-sensitive AI tasks across edge-cloud.* arXiv.
17. Porter, J. (2025). *Comcast is rolling out "ultra-low lag" tech that could fix the internet.*
18. Li, Y., et al. (2021). Edge-oriented computing for latency-sensitive applications. *IEEE Internet of Things Journal, 8*(5), 3505–3515.
19. Chen, X., et al. (2020). AI-driven adaptive bitrate streaming for latency reduction in live video applications. *IEEE Transactions on Multimedia, 22*(7), 1791–1803.
20. Gupta, H., et al. (2022). AI-based resource allocation for latency minimization in 5G networks. *IEEE Transactions on Network and Service Management, 19*(1), 45–58.
21. Wang, L., Chen, J., & Liu, K. (2023). AI-enhanced caching strategies for latency reduction in content delivery networks. *IEEE Transactions on Services Computing, 16*(3), 1234–1245.
22. Nguyen, T., Tran, D., & Le, Q. (2024). AI-based latency optimization in IoT networks. *IEEE Internet of Things Journal, 11*(2), 567–578.
23. Park, J., Lee, S., & Kim, H. (2021). Machine learning-based latency prediction for edge computing. *IEEE Access, 9*, 12345–12356.
24. Hu, Y., et al. (2024). Latency and privacy-aware CNN distributed inference for reliable AI systems. *IEEE Transactions on Artificial Intelligence.*
25. Kambala, G. (2024). Emergent architectures in edge computing for low-latency applications. *International Journal of Engineering and Computer Science, 13*(9).
26. Mourad, A., et al. (2020). Ad hoc vehicular fog enabling cooperative low-latency intrusion detection. *IEEE Internet of Things Journal, 8*(2), 829–843.
27. Rajbhandari, S., et al. (2022). DeepSpeedMoE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *Proceedings of the 39th International Conference on Machine Learning (ICML)* (pp. 18332–18346).
28. Rehan, H. (2023). AI-powered genomic analysis in the cloud: Enhancing precision medicine and ensuring data security in biomedical research. *Journal of Deep Learning in Genomic Data Analysis, 3*(1), 37–71.
29. Pradhan, B., et al. (2023). An AI-assisted smart healthcare system using 5G communication. *IEEE Access, 11*, 108339–108355.
30. Puri, V., Kataria, A., & Sharma, V. (2024). Artificial intelligence-powered decentralized framework for Internet of Things in Healthcare 4.0. *Transactions on Emerging Telecommunications Technologies, 35*(4), e4245.
31. Khan, K. (2024). Advancements in latency reduction models for adaptive video streaming: A comprehensive study. *International Journal of Multidisciplinary Research and Analysis in Practice, 6*(7).
32. Kim, S., Jung, S., & Lee, H. W. (2022). Reducing DNN inference latency using DRL. In *Proceedings of the 13th International Conference on ICT Convergence (ICTC)* (pp. 1517–1519).
33. Sharma, R. K., Kamal, P., & Singh, S. P. (2015). A latency reduction mechanism for virtual machine resource allocation in delay-sensitive cloud services. In *Proceedings of the International Conference on Green Computing and Internet of Things (ICGCIoT)* (pp. 371–375).
34. Doudou, M., Djenouri, D., & Badache, N. (2012). Survey on latency issues of asynchronous MAC protocols in delay-sensitive wireless sensor networks. *IEEE Communications Surveys & Tutorials, 15*(2), 528–550.
35. Briscoe, B., et al. (2014). Reducing Internet latency: A survey of techniques and their merits. *IEEE Communications Surveys & Tutorials, 18*(3), 2149–2196.
36. Vulimiri, A., et al. (2012). More is less: Reducing latency via redundancy. In *Proceedings of the ACM Workshop on Hot Topics in Networks (HotNets)* (pp. 13–18).
37. Guo, S., & Zhao, X. (2022). Multi-agent deep reinforcement learning-based transmission latency minimization for delay-sensitive cognitive satellite-UAV networks. *IEEE Transactions on Communications, 71*(1), 131–144.
38. Sarkar, I., et al. (2021). A collaborative computational offloading strategy for latency-sensitive applications in fog networks. *IEEE Internet of Things Journal, 9*(6), 4565–4572.

***Corresponding Author:** Jayashree P

39. Shukla, S., et al. (2023). Improving latency in Internet of Things and cloud computing for real-time data transmission: A systematic literature review. *Cluster Computing*.
40. Gemmell, D. J., & Han, J. (1994). Multimedia network file servers: Multichannel delay-sensitive data retrieval. *Multimedia Systems, 1*, 240–252.
41. Sreenan, C. J., et al. (2000). Delay reduction techniques for playout buffering. *IEEE Transactions on Multimedia, 2*(2), 88–100.
42. Carloni, L. P., McMillan, K. L., & Sangiovanni-Vincentelli, A. L. (2001). Theory of latency-insensitive design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 20*(9), 1059–1076.
43. Alizadeh, M., et al. (2012). Less is more: Trading a little bandwidth for ultra-low latency in the data center. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI)* (pp. 253–266).
44. Jiang, X., et al. (2018). Low-latency networking: Where latency lurks and how to tame it. *Proceedings of the IEEE, 107*(2), 280–306.
45. Simon, C., Maliosz, M., & Máté, M. (2018). Design aspects of low-latency services with time-sensitive networking. *IEEE Communications Standards Magazine, 2*(2), 48–54.
46. Joshi, G., Soljanin, E., & Wornell, G. (2017). Efficient redundancy techniques for latency reduction in cloud systems. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems, 2*(2), 1–30.
47. Gao, L., Zhang, Z. L., & Towsley, D. (1999). Catching and selective catching: Efficient latency reduction techniques for delivering continuous multimedia streams. In *Proceedings of the ACM International Conference on Multimedia* (pp. 203–206).
48. Gupta, A., et al. (1991). Comparative evaluation of latency reducing and tolerating techniques. In *Proceedings of the 18th Annual International Symposium on Computer Architecture* (pp. 254–263).
49. Mosberger, D., et al. (1996). Analysis of techniques to improve protocol processing latency. *ACM SIGCOMM Computer Communication Review, 26*(4), 73–84.
50. Velasquez, K., et al. (2017). Service placement for latency reduction in the Internet of Things. *Annals of Telecommunications, 72*, 105–115.

***Corresponding Author:** Jayashree P