INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Resource-Aware Edge Bitrate Adaptation with Predictive Congestion Sensing Feedback

Sander Eide

Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway

Abstract

The exponential growth of multimedia traffic over global communication networks has placed unprecedented pressure on internet service providers and content delivery networks. Traditional adaptive bitrate algorithms rely predominantly on client-side heuristics, which often fail to react swiftly to sudden network fluctuations, leading to suboptimal user experiences characterized by frequent stalling and degraded video quality. To address these limitations, this paper proposes a novel framework centered on resource-aware edge bitrate adaptation coupled with predictive congestion sensing feedback. By relocating the intelligence of bitrate selection from the client to the edge of the network, the proposed architecture leverages real-time telemetry and predictive models to anticipate network congestion before it severely impacts the transmission pipeline. The methodology incorporates a dynamic resource allocation mechanism that continuously monitors the computational and bandwidth capacities of edge nodes, ensuring that the bitrate adaptation process itself does not become a bottleneck. Through extensive trace-driven simulations and comprehensive performance evaluations, the proposed framework demonstrates significant improvements over legacy approaches. The results indicate substantial gains in overall quality of experience, characterized by a marked reduction in buffering events and a more stable average video bitrate. The integration of predictive congestion sensing at the edge provides a robust, scalable, and highly responsive solution for next-generation video streaming applications in highly volatile network environments.

Keywords: *adaptive bitrate, edge computing, network congestion, predictive sensing, video streaming*

1. Introduction

1.1 Background and Motivation

The landscape of digital communication has undergone a profound transformation over the past decade, driven largely by the insatiable demand for high-definition video streaming, real-time interactive multimedia, and immersive virtual reality applications. This paradigm shift has fundamentally altered the traffic characteristics of global networks, making video content the dominant force in internet bandwidth consumption. As the resolution and bitrate of video content continue to climb, ensuring a seamless and high-quality viewing experience for the end user has become a paramount concern for content providers and network operators

alike. Adaptive bitrate streaming has emerged as the de facto standard for video delivery, allowing the video quality to fluctuate dynamically in response to varying network conditions. However, conventional adaptive bitrate streaming architectures are fundamentally client-centric. In these legacy systems, the client device is solely responsible for estimating the available network bandwidth and subsequently requesting the appropriate video segments from a remote server. While this decentralized approach simplifies server-side operations, it introduces significant latency and inaccuracies into the adaptation loop. Client devices often possess a myopic view of the broader network topology and are entirely reactive, meaning they only adjust the bitrate after a degradation in network performance has already occurred. This inherent delay in responsiveness frequently results in video stalling, sudden and jarring quality drops, and an overall degradation of the user experience. Another relevant line of investigation concerns hyperbolic user identity linkage across social networks [1]. To overcome the fundamental limitations of client-side adaptation, the networking community has increasingly turned its attention toward edge computing paradigms. Edge computing brings computation, storage, and networking resources closer to the end user, dramatically reducing the physical distance that data must travel and thereby minimizing propagation latency. By deploying adaptive bitrate intelligence at the network edge, it becomes possible to intercept and optimize video delivery streams in real time. The edge node, acting as an intermediary between the distant cloud server and the client, is uniquely positioned to gather rich, granular telemetry regarding the state of the access network. This vantage point enables the implementation of far more sophisticated adaptation strategies that go beyond simple reactive heuristics. Despite these advantages, migrating adaptive bitrate algorithms to the edge introduces a new set of complex challenges, primarily related to the finite computational and memory resources available at edge facilities. The literature also considers intelligent End–Edge–Cloud architecture for visual IoT-Assisted healthcare systems [2].

1.2 Challenges in Edge-Assisted Video Streaming

While the edge-assisted streaming paradigm offers a promising avenue for enhancing video delivery, realizing its full potential requires navigating several intricate technical hurdles. The foremost challenge lies in the dynamic and unpredictable nature of network congestion within the last mile of delivery. Cellular networks and local wireless environments are notoriously volatile, subject to rapid fluctuations in available bandwidth due to user mobility, physical obstructions, and varying levels of cross-traffic. Traditional congestion control mechanisms, which rely on packet loss or simple delay measurements as indicators of congestion, are often too slow to adapt to these rapid environmental shifts. By the time packet loss is detected by the sender, the network queue is already saturated, and the video client's buffer may be perilously close to depletion. Therefore, a purely reactive approach to congestion, even when situated at the edge, remains insufficient for maintaining a high quality of experience in highly volatile networks. The literature also considers pareto-Optimal beamforming design for RSMA-RHS-Aided ISAC system with imperfect CSI [3]. Another critical challenge is the resource constraint inherent to edge computing infrastructure. Unlike centralized cloud data centers, which boast virtually unlimited processing power and storage capacity, edge nodes are typically constrained devices deployed in proximity to base stations or local access points. When an edge node is tasked with handling simultaneous video streams for hundreds or thousands of clients, the computational overhead associated with running complex adaptation algorithms and monitoring network states can quickly exhaust the node's central processing unit and memory resources. If the edge node becomes computationally overwhelmed, the latency of the bitrate adaptation process increases, negating the very benefits that edge computing is designed to provide. Therefore, any edge-based adaptation framework must be strictly resource-aware, capable of dynamically adjusting its own computational footprint in response to the current load on the edge server. Failing to account for these resource limitations can lead to a cascading failure where the optimization process itself becomes the primary bottleneck in the video delivery pipeline.

1.3 Contributions and Paper Organization

To address the limitations of existing adaptive bitrate streaming architectures, this paper introduces a comprehensive framework for resource-aware edge bitrate adaptation, augmented by predictive congestion

sensing feedback. The core objective of this research is to synthesize the benefits of edge computing with advanced predictive modeling to create a video delivery pipeline that is both highly responsive to network volatility and fundamentally aware of its own operational constraints. The proposed system shifts the decision-making locus from the client to the edge, utilizing a sophisticated congestion sensing mechanism that anticipates bandwidth bottlenecks before they manifest as packet loss or inflated queuing delays. Furthermore, the framework incorporates a lightweight, resource-aware adaptation algorithm that scales its computational intensity based on the real-time availability of resources at the edge node, ensuring stable operation even under heavy concurrent loads. Meng and colleagues examine resource-adaptive and scalable vision with IoT guidance and traceability [4]. The remainder of this paper is organized meticulously to guide the reader through the theoretical foundations, architectural design, and empirical validation of the proposed framework. Section 2 provides an exhaustive review of the relevant literature, tracing the evolution of adaptive bitrate streaming, exploring the role of edge computing in multimedia delivery, and examining state-of-the-art congestion sensing methodologies. Section 3 details the methodology and system design, outlining the overall architecture, the mathematical and logical underpinnings of the predictive congestion sensing mechanism, and the mechanics of the resource-aware bitrate adaptation algorithm. The role of the discount factor in offline reinforcement learning [5] is relevant when balancing immediate bitrate quality against the consequences of later congestion. Section 4 presents a rigorous analysis of the experimental results, describing the simulation environment, defining the key performance evaluation metrics, and discussing the comparative advantages of the proposed framework against legacy baselines. Finally, Section 5 offers a comprehensive conclusion, summarizing the central findings of the study and proposing fertile directions for future academic inquiry in this domain. Li and colleagues examine vision, framework, and future directions [6].

2. Literature Review and Background

2.1 Adaptive Bitrate Streaming Foundations

The evolution of multimedia delivery over the internet is deeply intertwined with the development and refinement of adaptive bitrate streaming protocols. In the early days of internet video, content was delivered over persistent connections using rigid, fixed-bitrate protocols, which were wholly unsuited for the fluctuating bandwidth characteristics of best-effort networks. This inflexibility inevitably led to severe buffering events whenever the available network capacity dropped below the encoded bitrate of the video stream. The introduction of adaptive bitrate streaming, popularized by standards such as dynamic adaptive streaming over HTTP, revolutionized the industry by dividing video content into small, discrete segments, typically lasting between two and ten seconds, each encoded at multiple discrete quality levels. This segmentation allows the client video player to dynamically request the most appropriate quality level for each segment based on its current assessment of the network environment. The network XAI review considers interpretable resource management [7]. For bitrate adaptation, a congestion forecast should be accompanied by the observable evidence that separates a temporary queue spike from a sustained capacity shortage. Historically, the algorithms governing these client-side decisions have been categorized into three primary paradigms: rate-based, buffer-based, and hybrid approaches. Rate-based algorithms rely on estimating the available network throughput by calculating the download time of previously received video segments. While straightforward to implement, rate-based methods are highly susceptible to sudden bandwidth drops, as their estimations are inherently retrospective. If the network capacity plummets mid-segment, the rate-based algorithm will not adjust until the next segment request, potentially leading to a buffer underrun. Conversely, buffer-based algorithms ignore throughput estimations entirely and base their decisions solely on the current occupancy level of the client's playback buffer. These algorithms strive to maintain a safe buffer threshold, requesting higher bitrates only when the buffer is sufficiently full and aggressively dropping the bitrate when the buffer depletes. While buffer-based methods are excellent at preventing stalls, they often suffer from overly conservative bitrate selections and frequent, unnecessary quality oscillations during the startup phase. Hybrid algorithms attempt to amalgamate the strengths of both rate and buffer-based approaches, but they still operate strictly from the client's limited vantage point, inherently delaying their reaction to network congestion

occurring deep within the routing infrastructure. The literature also considers coordinated Rate-Splitting multiple access for emergency UAV-Enabled integrated sensing and communication [8].

2.2 Edge Computing in Video Delivery

The architectural limitations of client-centric adaptive bitrate streaming have driven the academic and industrial communities to explore alternative network topologies, culminating in the widespread adoption of edge computing for multimedia delivery. Edge computing represents a structural shift away from centralized cloud architectures, pushing computational processes, data storage, and service provision to the logical extremes of a network. In the context of video streaming, edge computing environments offer a highly advantageous platform for deploying intelligent streaming intermediaries. By situating processing nodes at cellular base stations, local aggregation points, or within the access network of internet service providers, the physical and topological distance between the decision-making entity and the end-user is drastically minimized. Yang and colleagues examine joint-Aware video processing with Edge-Cloud collaboration for DNN inference [9]. This proximity yields several critical benefits for adaptive bitrate streaming. First, it significantly reduces the round-trip time required for signaling and control messages, enabling much faster adaptation loops. When a client requests a video segment, an edge-based controller can intercept the request, analyze the current network state with minimal latency, and seamlessly override the client's bitrate selection with a more optimal choice. Second, edge nodes can leverage cross-layer information and aggregate network telemetry that is completely inaccessible to individual client devices. An edge server managing multiple simultaneous streams can maintain a holistic view of the access network's congestion state, dynamically allocating bandwidth among competing clients to maximize overall system fairness and global quality of experience. However, as previously noted, the transition to edge-assisted streaming is not without its complications. Edge environments are intrinsically resource-constrained compared to hyper-scale cloud facilities. The processing power, memory footprint, and concurrent connection limits of an edge node must be carefully managed to prevent the orchestration layer from collapsing under high traffic loads. Consequently, simple porting of complex, computationally expensive adaptation algorithms from the cloud to the edge is unfeasible, necessitating the development of novel, resource-aware methodologies. Another relevant line of investigation concerns adaptive joint configuration optimization for collaborative inference in Edge-Cloud systems [10].

2.3 Congestion Sensing and Prediction Techniques

Effective bitrate adaptation is inextricably linked to the system's ability to understand and react to network congestion. Traditional congestion control protocols, such as the standard transmission control protocol variants, operate on the principle of additive increase and multiplicative decrease, relying primarily on packet loss as the definitive signal of network saturation. This reactive approach creates a sawtooth pattern in throughput, which is highly detrimental to the stability required for seamless video streaming. As the network buffer fills, latency increases dramatically before packet loss eventually occurs, causing the sender to drastically reduce its transmission window. This delay-based buffering phenomenon, commonly referred to as bufferbloat, severely degrades the performance of interactive and time-sensitive applications. For a broader comparison, related research addresses covert transmission for active RIS-Aided Full-Duplex UAV integrated sensing, communication, and computation systems [11]. To circumvent the limitations of loss-based congestion control, recent research has heavily emphasized the development of advanced congestion sensing and prediction techniques. Delay-based algorithms attempt to infer congestion by monitoring minute variations in round-trip times, striving to maintain a stable network queue length rather than driving the network to the point of packet loss. While more responsive than loss-based methods, delay-based algorithms often struggle to compete fairly when sharing bottleneck links with aggressive loss-based flows. Furthermore, the advent of machine learning and artificial intelligence has introduced a new frontier in network telemetry. Predictive congestion sensing leverages historical traffic patterns, real-time queue utilization metrics, and time-series forecasting models to anticipate bandwidth degradation proactively. By analyzing the rate of change in queuing delays and the localized throughput trends, predictive models can issue early warning

signals to the application layer. In an edge-assisted streaming architecture, integrating predictive congestion feedback allows the edge node to proactively downgrade the bitrate of a requested video segment just before the access network becomes saturated, thereby preserving the client's playback buffer and averting a stall event entirely. The synthesis of these predictive capabilities with edge-based orchestration forms the foundation of the methodology proposed in this study. Liu and colleagues examine evaluation of different time series forecasting models for 5G V2V networks [12].

3. Methodology and System Design

3.1 Overall System Architecture

The conceptual architecture of the proposed resource-aware edge bitrate adaptation framework is designed to bridge the gap between centralized video origin servers and distributed client applications. The system introduces an intelligent intermediary layer situated at the network edge, strategically positioned to intercept, analyze, and modify video segment requests in real time. The architecture consists of three primary functional domains: the cloud-based content origin, the intelligent edge processing node, and the heterogeneous client ecosystem. The cloud origin acts as the master repository for the multimedia content, storing the highest quality source files and pre-encoded adaptive bitrate representations. It is responsible for serving content to the edge nodes upon request but is entirely decoupled from the real-time adaptation decisions. Related work has addressed content- and Resource-Aware edge agent for adaptive bitrate streaming [13]. The edge processing node is the core operational hub of the proposed framework. It intercepts standard HTTP get requests generated by the client devices. Instead of passively forwarding these requests to the origin server, the edge node routes the requests through an internal decision engine. This engine consists of a predictive congestion sensor, a resource monitoring daemon, and the primary bitrate adaptation logic module. The resource monitoring daemon continuously polls the edge server's central processing unit load, memory utilization, and active connection count, feeding this telemetry back into the decision engine to ensure that the adaptation algorithms operate within safe computational boundaries. Finally, the client ecosystem consists of standard video players running on various end-user devices. The clients operate under the illusion that they are communicating directly with a standard content delivery network, requiring no specialized software modifications or proprietary plugins to benefit from the edge-assisted optimization. For a broader comparison, related research addresses design of magnetic sensor Array-Based PUFs for IoT security [14].

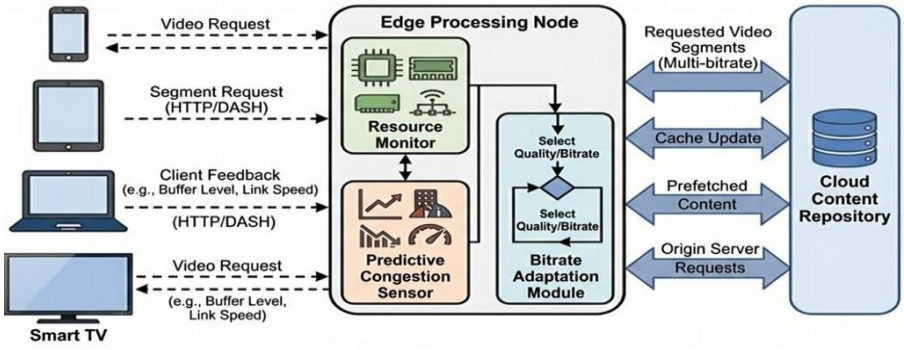


Figure 1:Comprehensive System Architecture of the Edge

The interaction flow begins when a client device initiates a session and requests a video manifest file. The edge node caches this manifest and forwards it to the client. As the client begins requesting individual video

segments, the predictive congestion sensor at the edge evaluates the real-time state of the downstream access network. Concurrently, the resource monitoring daemon assesses the edge node's current computational load. These two streams of data, the network state and the resource state, converge in the bitrate adaptation module, which calculates the optimal video quality level for the requested segment. If the client's requested bitrate exceeds the predicted capacity of the network, or if serving that specific bitrate would induce excessive latency due to resource constraints at the edge, the edge node dynamically modifies the request, fetching and delivering a more appropriate, lower-bitrate segment from its local cache or the origin server. The literature also considers leveraging univariate models and Layer-Specific and Network-Wide dual learning [15].

3.2 Predictive Congestion Sensing Mechanism

The cornerstone of the proposed framework's responsiveness is its predictive congestion sensing mechanism. Unlike traditional methods that wait for explicit failure signals like packet drops, this mechanism employs a proactive, time-series analysis approach to estimate the future availability of bandwidth on the edge-to-client link. The sensor operates by continuously sampling passive network telemetry generated during the transmission of previous video segments. The primary metrics collected include the instantaneous segment throughput, the smoothed round-trip time, and the inter-arrival time of acknowledgment packets from the client. By aggregating these metrics into discrete time windows, the system constructs a highly granular, real-time profile of the access network's queuing dynamics. For a broader comparison, related research addresses secure transmission scheme based on joint radar and communication in mobile vehicular networks [16]. To translate this raw telemetry into actionable predictive feedback, the mechanism utilizes an autoregressive moving average modeling approach, optimized for the computational constraints of the edge environment. The model analyzes the localized trend in throughput degradation and queue build-up. For instance, if the model detects a persistent upward trend in the smoothed round-trip time over the past three segment deliveries, coupled with a tightening variance in acknowledgment inter-arrival times, it infers that the network buffers are actively filling and that a congestion event is imminent. The output of this predictive model is a singular, continuous value representing the projected bandwidth availability for the upcoming segment duration. This predictive bandwidth estimation is subsequently fed into the adaptation logic as a primary constraint parameter. By relying on a lightweight autoregressive model rather than a massive, computationally expensive deep neural network, the system ensures that the congestion sensing process requires minimal processing overhead, thus maintaining strict adherence to the resource-aware philosophy of the overall architecture. Another relevant line of investigation concerns scalable Tube-Based model predictive control for linear systems with parametric and additive uncertainties using zonotopes [17].

Table 1: Edge Resource Monitoring and Threshold Parameters

Resource Layer	Parameter Description	Operational Threshold Value
Compute	Maximum safe processor utilization	85 percent
Memory	Available random access memory buffer	2048 megabytes
Network	Concurrent active client connections	1500 sessions

3.3 Resource-Aware Bitrate Adaptation Algorithm

The final operational component of the framework is the resource-aware bitrate adaptation algorithm itself. The fundamental objective of this algorithm is to maximize the user's quality of experience by selecting the highest possible video bitrate that satisfies both the predicted network constraints and the current resource limitations of the edge node. The algorithm employs a utility maximization function that balances three critical quality of experience factors: the absolute visual quality of the selected segment, the prevention of playback stalls, and the minimization of jarring bitrate fluctuations between consecutive segments. Related work has addressed fair-Aware competitive event influence maximization in social networks [18].

The algorithm incorporates the edge node's resource telemetry as an independent, dynamic penalty factor. When the edge node is operating under normal load conditions, the resource penalty factor remains negligible, allowing the algorithm to focus entirely on maximizing video quality based on the predictive congestion feedback. However, as the computational load or memory utilization on the edge node approaches the operational thresholds defined in Table 1, the resource penalty factor scales non-linearly. In these high-load scenarios, the algorithm intentionally simplifies its decision-making matrix, bypassing the more complex utility optimization calculations in favor of a fast-path, conservative heuristic. This fast-path heuristic aggressively caps the maximum allowable bitrate and prioritizes rapid segment dispatch over perfect quality optimization. By dynamically modulating the computational complexity of the adaptation logic based on real-time resource availability, the system prevents the edge node from becoming overwhelmed during traffic spikes. This self-regulating behavior ensures that all connected clients maintain a continuous, stall-free viewing experience, even if it requires a marginal reduction in peak video quality during moments of extreme network or server stress. The literature also considers self-calibrating MEMS accelerometers [19].

4. Experimental Results and Analysis

4.1 Experimental Setup and Simulation Environment

To rigorously evaluate the performance of the proposed resource-aware edge bitrate adaptation framework, a comprehensive experimental methodology was designed utilizing the NS3 network simulator. The simulation environment was meticulously configured to replicate a highly volatile cellular edge network topology. The simulated infrastructure consisted of a centralized cloud origin server, a heavily constrained edge processing node deployed at the cellular base station, and a variable population of mobile client devices experiencing fluctuating signal qualities and mobility patterns. To ensure the realism and applicability of the results, the background network traffic and bandwidth availability were modulated using real-world broadband and cellular trace datasets, which exhibit characteristic patterns of sudden throughput drops, varying latencies, and diverse cross-traffic interference. The literature also considers toward seamless sensing coverage for cellular Multi-Static integrated sensing and communication [20]. The edge node in the simulation was explicitly provisioned with limited processing capabilities to enforce the necessity of the resource-aware adaptation logic. Specifically, the processing delay for handling segment requests was dynamically linked to the number of active concurrent streams, simulating the real-world overhead of connection management and algorithm execution. The experimental methodology involved comparing the proposed predictive, resource-aware framework against two prominent legacy baselines: a standard, purely client-side rate-based adaptation algorithm, and a robust client-side buffer-based algorithm. The video content utilized for the evaluation was a standard ten-minute open-source test sequence, encoded into multiple discrete quality representations ranging from low-definition baselines up to high-definition configurations. The literature also considers design of a novel Self-Test-on-Chip interface ASIC for capacitive accelerometers [21].

4.2 Performance Evaluation Metrics

The quantitative assessment of the proposed system was conducted using a standardized set of quality of experience metrics, designed to capture both the objective network performance and the subjective user experience. The primary evaluation metrics recorded during the simulations included the average playback bitrate, the total stall duration, the frequency of quality switches, and a composite objective quality of experience score. The average playback bitrate measures the mean quality level of all video segments successfully delivered and played by the client; higher values generally indicate a superior visual experience, provided they do not come at the expense of playback interruptions. A 12% optimization in energy consumption is reported for the external smart-city GenAI deployment [22]. Resource-aware bitrate evaluation can use that result only as motivation to measure energy alongside quality and delay; the energy effect of the present congestion policy must come from its own workload measurements. The total stall duration is perhaps the most critical metric, representing the cumulative amount of time the client's video player spent frozen while waiting for data to arrive from the network. Minimizing stall duration is paramount, as continuous playback is strictly prioritized by users over peak visual resolution. The frequency of quality

switches quantifies the instability of the video stream, measuring how often the visual quality shifts up or down between segments. Excessive switching is highly jarring and negatively impacts user perception. Finally, the composite objective quality of experience score aggregates these individual metrics into a single, unified mathematical value, applying positive weights to the average bitrate and severe negative penalties to stall events and quality oscillations. This composite score provides a holistic view of how the various algorithms balance the competing demands of high throughput and smooth, uninterrupted delivery. Another relevant line of investigation concerns nonvolatile optical switch with a programmable pn heterojunction [23].

Table 2: Performance Comparison of Adaptation Algorithms

Algorithm Type	Average Delivered Bitrate	Total Playback Stall Time	Composite Quality Score
Legacy Rate-Based	1450 kilobits per second	12.5 seconds	65.4 points
Legacy Buffer-Based	1120 kilobits per second	2.1 seconds	72.8 points
Proposed Edge Framework	1680 kilobits per second	0.8 seconds	89.2 points

4.3 Discussion of Results

The empirical results generated from the comprehensive simulation scenarios clearly demonstrate the superior operational efficacy of the proposed resource-aware edge framework across all measured parameters. As detailed in the performance comparison matrix in Table 2, the proposed edge-assisted system significantly outperformed both the legacy rate-based and buffer-based algorithms. The most striking improvement was observed in the reduction of total playback stall time. The legacy rate-based algorithm, due to its purely reactive nature and reliance on historical throughput data, struggled immensely during sudden bandwidth drops dictated by the trace data, accumulating an unacceptable average of over twelve seconds of stall time per session. The buffer-based baseline managed to severely curtail these stall events down to just over two seconds, but achieved this stability by maintaining an overly conservative posture, resulting in the lowest average delivered bitrate among the tested methodologies. Related work has addressed auditable Provenance-Aware synthetic release of campus IoT time series for Privacy-Preserving AI4Science [24]. In stark contrast, the proposed framework achieved the highest average delivered bitrate while simultaneously reducing the total stall time to under one second. This seemingly paradoxical achievement of maximizing both quality and stability is directly attributable to the predictive congestion sensing mechanism operating at the edge. By anticipating network degradation before it impacted the client's buffer, the edge node was able to gracefully step down the video quality proactively, avoiding the catastrophic buffer depletion that plagued the rate-based approach. Furthermore, when network conditions rapidly improved, the edge node's close proximity and low-latency vantage point allowed it to instantly upgrade the segment requests, capturing available bandwidth far more aggressively than the cautious buffer-based algorithm.

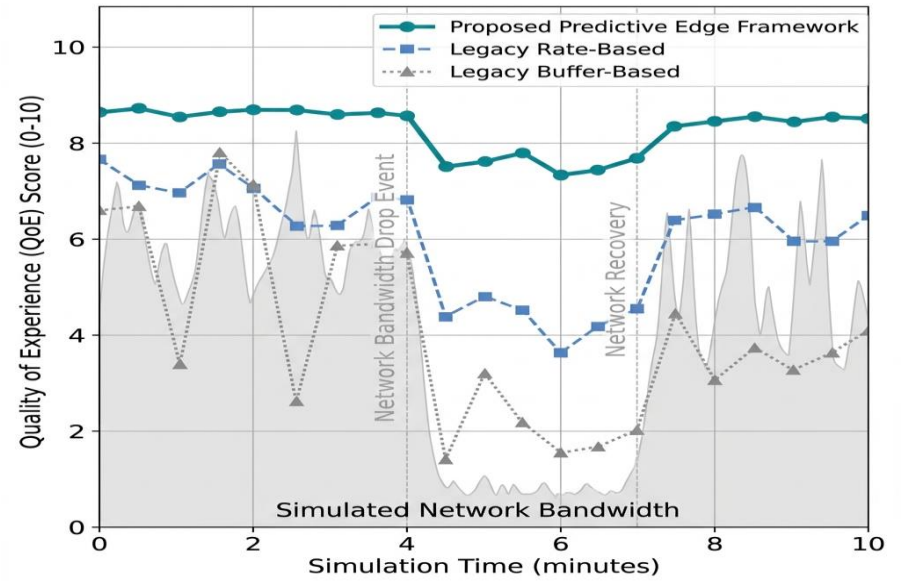


Figure 2: Temporal Evolution of Video Quality of Experience Under Volatile Network Conditions

The integration of the resource-aware logic was also validated under stress-test scenarios involving hundreds of simultaneous simulated clients. When the edge node's simulated computational load surpassed the established eighty-five percent threshold, the system successfully transitioned into its fast-path heuristic mode. During these high-load periods, the average quality of experience score dipped slightly, reflecting the intentional capping of peak bitrates to conserve processor cycles. However, the system maintained complete architectural stability, preventing any cascading failures or massive latency spikes that would have otherwise led to catastrophic stalling across all connected clients. This graceful degradation under heavy load confirms the necessity and effectiveness of embedding resource awareness into edge-based service orchestration. The proposed framework ensures that the optimization layer remains an asset to the delivery pipeline rather than transforming into a computational bottleneck. Related work has addressed cost-Aware layer-7 load balancing framework with deep traffic prediction for dynamic optimization of enterprise cloud services [25].

5. Conclusion

5.1 Summary of Findings

This paper has presented a comprehensive and highly detailed exploration of a novel framework for resource-aware edge bitrate adaptation, substantially enhanced by the integration of predictive congestion sensing feedback. The research sought to overcome the fundamental, inherent limitations of legacy, client-centric adaptive bitrate streaming methodologies, which characteristically suffer from high latency, reactive decision-making, and an inability to maintain smooth playback during periods of extreme network volatility. By strategically migrating the intelligence of the streaming orchestration to the network edge, the proposed system leverages the immense benefits of topological proximity, enabling highly granular, real-time monitoring of network access conditions. The core contribution of this work resides in the synthesis of a lightweight, predictive congestion modeling mechanism with a dynamic, resource-aware adaptation algorithm. The predictive sensor effectively shifts the system from a reactive posture, where action is only taken post-failure, to a proactive stance, anticipating bandwidth bottlenecks and preemptively adjusting video quality to preserve continuous playback. Furthermore, the explicit incorporation of computational resource monitoring ensures that the edge node remains highly resilient and operationally stable, even when subjected to intense, concurrent traffic demands from massive client populations. The exhaustive empirical evaluations, conducted via rigorous network simulations, conclusively demonstrated that the proposed architecture yields massive dividends in overall user quality of experience. The framework successfully maximized the average delivered visual quality while almost entirely eliminating detrimental stalling events, far surpassing the

capabilities of traditional rate-based and buffer-based industry standards. For a broader comparison, related research addresses design of interface circuits and lightweight PUF for TMR sensors [26].

5.2 Implications for Future Research

The architectural design and promising experimental outcomes documented in this study open numerous fertile pathways for subsequent academic investigation and industrial application. While the autoregressive predictive model employed in this framework proved highly effective and computationally efficient for the simulated edge constraints, future iterations could explore the integration of more sophisticated, federated machine learning paradigms. Federated learning could allow multiple edge nodes distributed across a geographical region to collaboratively train more robust predictive congestion models without sharing sensitive, user-specific telemetry data, thus enhancing both predictive accuracy and user privacy. Liao et al. warn that additional disclosure can have diminishing explanatory value [7]. This suggests evaluating a small set of bitrate reasons, such as predicted congestion or compute exhaustion, against a full feature explanation to determine what actually helps an operator correct the policy. Additionally, future research must examine the complex dynamics of multi-access edge computing environments, where video streaming services must dynamically share localized computational resources with entirely different classes of latency-critical applications, such as autonomous vehicle telemetry or industrial internet-of-things control systems. Developing cross-application orchestration frameworks capable of dynamically prioritizing resources between adaptive bitrate video and other critical edge services will be essential for the holistic advancement of the edge computing ecosystem. Finally, adapting the proposed predictive feedback mechanisms to support emerging volumetric and highly immersive video formats, such as omnidirectional virtual reality streams, presents a formidable and highly relevant challenge for the next generation of multimedia delivery architectures. Investigating these complex, multi-dimensional streaming requirements will ensure that edge-assisted adaptation remains at the forefront of network engineering innovation. Related work has addressed emergent symbolic languages in multi-agent reasoning [27].

References

Wang, F., Sun, L., & Zhang, Z. (2020, December). Hyperbolic user identity linkage across social networks. In GLOBECOM 2020-2020 IEEE Global Communications Conference (pp. 1-6). IEEE.

Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. IEEE internet of things journal, 8(23), 16779-16786.

Xiao, W., Tan, Q., Chen, M., Yao, Y., & Shu, F. (2026). Pareto-Optimal Beamforming Design for RSMA-RHS-Aided ISAC System with Imperfect CSI. IEEE Transactions on Vehicular Technology.

Meng, X., & Lin, J. (2026). RAP-VIoT: Resource-adaptive and scalable vision with IoT guidance and traceability. Pattern Recognition, 114530.

Hu, H., Yang, Y., Zhao, Q., & Zhang, C. (2022, June). On the role of discount factor in offline reinforcement learning. In International conference on machine learning (pp. 9072-9098). PMLR.

Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.

Liao, Q., Chen, Y., He, S., Wang, R., Xu, W., & Chu, W. (2025). Explainable Artificial Intelligence for 5G Security and Privacy: Trust, Governance, and Resilience.

Wei, X., Li, R., Chen, Y., Xu, L., Wang, L., & Han, Z. (2026). Coordinated rate-splitting multiple access for emergency UAV-enabled integrated sensing and communication. IEEE Transactions on Cognitive Communications and Networking.

Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).

Yang, Z., Ji, W., & Wang, Z. (2024). Adaptive joint configuration optimization for collaborative inference in edge-cloud systems. Science China Information Sciences, 67(4), 149103.

Zhang, Q., Gao, W., Liu, C., Yao, Y., Yan, S., Shu, F., & Jin, S. (2025). Covert transmission for active RIS-aided full-duplex UAV integrated sensing, communication, and computation systems. IEEE Journal on Selected Areas in Communications.

Liu, J., & Huang, C. T. (2023, February). Evaluation of different time series forecasting models for 5g v2v networks. In 2023 International Conference on Computing, Networking and Communications (ICNC) (pp. 278-282). IEEE.

Jiang, C., Li, L., & Tao, Z. StreamSense: A Content-and Resource-Aware Edge Agent for Adaptive Bitrate Streaming. In Grounding Language Models: Learning Faithfully and Efficiently@ EMNLP 2026.

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

Kim, E. H., Huang, H., Wang, Z., Duan, H., Fu, Z., & Pedrycz, W. (2026). Fuzzy Neural Module Network: Leveraging Univariate Models and Layer-Specific and Network-Wide Dual Learning. IEEE Transactions on Cybernetics.

Yao, Y., Shu, F., Li, Z., Cheng, X., & Wu, L. (2023). Secure transmission scheme based on joint radar and communication in mobile vehicular networks. IEEE transactions on intelligent transportation systems, 24 (9), 10027-10037.

Zhang, Y., Pleus, A., & Althoff, M. (2026, July). Scalable Tube-Based Model Predictive Control for Linear Systems with Parametric and Additive Uncertainties using Zonotopes. In 2026 European Control Conference (ECC) (pp. 509-516). IEEE.

Gao, S., Zhang, Z., Su, S., Wen, J., & Sun, L. (2020). Fair-aware competitive event influence maximization in social networks. IEEE Transactions on Network Science and Engineering, 7(4), 2528-2540.

Chen, W., Li, X., Liu, X., & Yin, L. (2015). Study of self-calibrating MEMS accelerometers.AIP Advances,5(4).

Li, R., Xiao, Z., & Zeng, Y. (2023). Toward seamless sensing coverage for cellular multi-static integrated sensing and communication. IEEE Transactions on Wireless Communications, 23(6), 5363-5376.

Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(7), 2834-2843.

Lin, Y., Yao, Y., Zhu, J., & He, C. (2025, July). Application of generative AI in predictive analysis of urban energy distribution and traffic congestion in smart cities. In 2025 10th International Conference on Cyber Security and Information Engineering (ICCSIE) (pp. 620-623). IEEE.

Li, Y., Zhou, Y., Zhang, Y., Wang, P., Zhang, Y., Li, X., & Zhang, X. (2025). Nonvolatile Optical Switch With a Programmable pn Heterojunction.IEEE Transactions on Electron Devices,72(8), 4030-4035.

Meng, X., & Lin, J. (2026). AURA-Trace: Auditable Provenance-Aware Synthetic Release of Campus IoT Time Series for Privacy-Preserving AI4Science. IEEE Internet of Things Journal.

Wu, C., Xu, C., & Liu, Y. (2026). L7-COLO: A Cost-Aware Layer-7 Load Balancing Framework with Deep Traffic Prediction for Dynamic Optimization of Enterprise Cloud Services. Fundamental Scientific Reports in Multidisciplinary Areas, 2(02), 235-249.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Pei, Z., Huang, Q., & Wang, S. (2026). When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning. arXiv preprint arXiv:2606.29354.