INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Content-Aware Streaming Adaptation Under Fluctuating Edge Compute Budgets

Tony Lo[1]

[1]*Department of Computer Science and Engineering, Faculty of Engineering, Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China*

Chun-Ho Lai[2]

[2]*Department of Electronic Engineering and Computer Science, School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, Hong Kong SAR, China*

Abstract

The rapid proliferation of high-definition video streaming, augmented reality, and virtual reality applications has placed unprecedented demands on network bandwidth and computational resources. Edge computing has emerged as a promising paradigm to alleviate core network congestion by pushing processing capabilities closer to the end users. However, edge nodes inherently possess constrained and highly fluctuating compute budgets due to multi-tenant resource sharing, user mobility, and dynamic network conditions. Traditional adaptive bitrate streaming algorithms predominantly focus on network bandwidth estimation, often neglecting the computational overhead required for video transcoding, rendering, and content processing at the edge. This paper introduces a comprehensive framework for content-aware streaming adaptation specifically designed to operate efficiently under fluctuating edge compute budgets. By analyzing the spatiotemporal complexity of the video content in real-time, the proposed system dynamically balances bandwidth utilization and computational expenditure. The methodology involves a novel decision-making engine that continuously monitors edge load variations and preemptively adjusts the streaming parameters to maximize the Quality of Experience for the end-user while adhering to strict compute limitations. Through extensive simulation and rigorous performance analysis, the proposed approach demonstrates significant improvements in maintaining high-quality video delivery and reducing rebuffering events during periods of severe edge resource constraints compared to conventional network-only adaptation strategies.

Keywords: *Edge Computing, Adaptive Streaming, Quality of Experience, Resource Management, Content-Aware Processing*

1. Introduction

1.1 The Evolution of Video Streaming and Edge Integration

The landscape of digital media delivery has undergone a monumental transformation over the past decade. The continuous rise in multimedia consumption, driven by high-definition video platforms, live broadcasting, and immersive interactive applications, has severely strained global network infrastructures. To mitigate the latency and bandwidth bottlenecks associated with centralized cloud architectures, the industry has increasingly adopted edge computing paradigms. Edge computing decentralizes processing power, storage, and networking capabilities, placing them in close proximity to the end-users. This architectural shift significantly reduces transmission delays and provides the agility required for next-generation applications such as augmented reality, virtual reality, and autonomous vehicle telemetry. However, integrating edge computing into video streaming pipelines introduces a complex array of new challenges, particularly concerning resource management. Unlike centralized cloud data centers, which offer virtually infinite scalability and vast resource pools, edge nodes are intrinsically resource-constrained. These nodes are often deployed in micro-data centers, cellular base stations, or even distributed consumer devices, where computational power, memory, and energy budgets are strictly limited. The volatility of the edge environment, characterized by fluctuating user demands and multi-tenant resource contention, necessitates a paradigm shift in how streaming algorithms operate and adapt. Current methodologies must evolve beyond simple network-awareness to incorporate deep insights into both the content being delivered and the real-time computational capacity of the delivery infrastructure. Another relevant line of investigation concerns multi-UAV-Enabled Energy-Efficient data delivery for Low-Altitude economy: joint coded caching, user grouping, and UAV deployment [1].

1.2 Limitations of Traditional Adaptive Bitrate Streaming

Adaptive bitrate streaming has long been the cornerstone of robust video delivery over unpredictable networks. Standard protocols operate by segmenting a video file into multiple short-duration chunks, each encoded at various bitrates and resolutions. The client-side streaming application continuously estimates the available network bandwidth and monitors its local buffer occupancy to request the most appropriate quality level for the upcoming segment. While this approach has been highly successful in mitigating playback interruptions caused by network congestion, it operates under the flawed assumption that computational resources, both at the server and the client, are abundant and static. Traditional algorithms fail to account for the processing overhead required to transcode, decode, or enhance video segments at the edge. When an edge node is tasked with performing real-time video processing functions such as super-resolution upscaling, dynamic transcoding, or object detection for interactive streams, the computational load becomes a critical bottleneck. If the edge compute budget drops due to sudden spikes in local traffic or competing background processes, a traditional streaming algorithm might still request high-bitrate, high-complexity segments simply because the network bandwidth appears sufficient. This mismatch inevitably leads to processing delays, increased latency, buffer underruns, and a severely degraded quality of experience for the viewer. Li and colleagues examine design of a novel Self-Test-on-Chip interface ASIC for capacitive accelerometers [2].

1.3 The Need for Content-Aware and Compute-Aware Adaptation

To bridge the gap between network-centric streaming and resource-constrained edge environments, it is imperative to develop adaptation strategies that are both content-aware and compute-aware. Video content is inherently heterogeneous in terms of its spatial and temporal complexity. A high-action sports broadcast with rapid camera movements and complex textures requires significantly more computational power to encode and decode compared to a static news broadcast or a simple animated sequence. By leveraging content-awareness, a streaming system can predict the processing effort required for each upcoming video segment based on its specific visual characteristics. Furthermore, integrating compute-awareness allows the system to continuously monitor the available processing budget at the edge node. By synthesizing these two dimensions along with traditional network bandwidth estimations, the adaptation engine can make highly informed decisions. For instance, during periods of high edge CPU utilization, the system could intelligently switch to

delivering segments with lower spatial complexity or pre-transcoded versions that require minimal on-the-fly processing, even if the network bandwidth could theoretically support a higher bitrate. Conversely, when edge resources are plentiful, the system can utilize advanced computational techniques like neural network-based video enhancement to significantly boost the visual quality delivered to the client without requiring additional network bandwidth. For a broader comparison, related research addresses optimizing hybrid RIS-Aided ISAC systems in V2X networks: A deep reinforcement learning method for Anti-Eavesdropping techniques [3].

1.4 Research Objectives and Paper Structure

The primary objective of this research is to design, evaluate, and formalize a robust framework for content-aware streaming adaptation under fluctuating edge compute budgets. This paper investigates the intricate relationships between video content characteristics, edge processing capabilities, and user quality of experience. The research systematically addresses the challenge of dynamic resource allocation in highly volatile multi-tenant edge environments. The remainder of this paper is structured to provide a comprehensive exploration of the topic. Section 2 delves into a detailed literature review, analyzing previous work in adaptive streaming, edge computing, and content-based resource optimization. Resource-adaptive visual IoT with traceability in RAP-VIoT [4] provides a comparison for coordinating computation and transmission when workload and available edge resources vary. Section 3 outlines the proposed system architecture, detailing the integration of edge nodes, monitoring agents, and the specific dynamics of compute budget fluctuations. Section 4 presents the core methodology, explaining the theoretical foundations of the content-aware adaptation engine and the decision-making processes. Section 5 describes the rigorous experimental setup, including the simulation environment, datasets, and the metrics used to evaluate performance. Section 6 provides a comprehensive analysis of the results, discussing the empirical findings and their implications for real-world deployments. Finally, Section 7 concludes the paper by summarizing the key contributions and identifying promising avenues for future research in this domain. For a broader comparison, related research addresses optimizing HIV interventions for multiplex social networks via Partition-Based random search [5].

2. Background and Related Work

2.1 Advancements in Adaptive Bitrate Streaming

The evolution of multimedia streaming over the internet has been fundamentally shaped by the development and refinement of adaptive bitrate streaming technologies. Deep Q-learning from Atari provides a standard value-learning reference for sequential action selection [6]. Early video delivery systems relied on progressive downloading or constant bitrate streaming, which frequently resulted in catastrophic failure under varying network conditions, leading to persistent stalling and unacceptable user experiences. The introduction of HTTP-based adaptive streaming revolutionized the industry by shifting the control logic to the client side. By fragmenting the video asset into small, independent segments, clients could seamlessly switch between different representations depending on real-time network throughput and buffer health. Over the years, a multitude of heuristic and control-theoretic algorithms have been proposed to optimize this switching logic. Rate-based algorithms focus primarily on historical and current throughput measurements to predict future bandwidth availability. Buffer-based algorithms, on the other hand, prioritize the stabilization of the client-side playback buffer, adjusting the requested bitrate to maintain a safe reserve of unplayed video segments. More recently, hybrid approaches and machine learning-based techniques, particularly those utilizing reinforcement learning, have gained significant traction. These advanced algorithms attempt to map complex, multidimensional state spaces including throughput variance, buffer occupancy, and chunk sizes to optimal bitrate selections. However, despite their sophistication in handling network volatility, the vast majority of these established streaming algorithms treat the underlying computational infrastructure as a transparent and infinitely capable conduit, entirely overlooking the processing constraints that define modern edge and mobile environments. Xiao and colleagues examine pareto-Optimal beamforming design for RSMA-RHS-Aided ISAC system with imperfect CSI [7].

2.2 Edge Computing in Video Delivery Networks

The integration of mobile edge computing into the telecommunications infrastructure represents a paradigm shift aimed at addressing the latency and bandwidth limitations of centralized cloud architectures. In the context of video delivery, edge computing allows service providers to cache popular content closer to the users, thereby reducing backhaul traffic and improving initial startup times. Beyond simple caching, the true transformative potential of edge nodes lies in their ability to perform active video processing tasks. Functions such as real-time transcoding, active bitrate adaptation, localized content insertion, and intelligent video analytics can be offloaded from the central cloud and executed at the network edge. This capability is particularly crucial for emerging interactive applications like cloud gaming and immersive virtual reality, where ultra-low latency is an absolute prerequisite for a viable user experience. However, edge nodes are characterized by significant limitations in their physical footprint, power supply, and cooling capacity, leading to highly constrained computational resources. Furthermore, edge environments are fundamentally multi-tenant. A single edge server situated at a cellular base station might simultaneously host video transcoding tasks, autonomous vehicle telemetry processing, and local smart city analytics. This intense resource contention leads to high volatility in the available compute budget allocated to any specific video streaming session. Another relevant line of investigation concerns content- and Resource-Aware edge agent for adaptive bitrate streaming [8].

2.3 Compute-Aware Adaptation Strategies

Recognizing the limitations of purely network-centric adaptation, recent literature has begun to explore the integration of computational awareness into multimedia delivery systems. Early efforts in this domain focused on the energy consumption of mobile client devices. Researchers proposed energy-aware streaming algorithms that adjusted the video bitrate or dynamically altered the decoding complexity to extend the battery life of smartphones and tablets. While these client-centric approaches provided valuable insights into the trade-offs between video quality and power consumption, they did not address the specific challenges of server-side or edge-side resource constraints. Subsequent research shifted focus towards cloud-based transcoding farms, investigating scheduling algorithms designed to maximize throughput and minimize processing delays across large clusters of servers. These studies often involved sophisticated queueing models and resource allocation heuristics to balance the transcoding load. However, the assumptions underlying cloud-based optimization continuous resource availability and high-bandwidth interconnects do not hold true in the edge computing paradigm. More recent studies have specifically targeted the edge environment, proposing mechanisms that jointly optimize network bandwidth and edge computing resources. These approaches typically formulate the adaptation problem as a constrained optimization task, aiming to maximize overall system utility. Despite these advancements, a significant gap remains in accurately predicting the variable processing costs associated with different types of video content, a factor that heavily influences the efficiency of compute-aware adaptation. Related work has addressed adaptive joint configuration optimization for collaborative inference in Edge-Cloud systems [9].

2.4 Content-Aware Processing and Analytics

Content-awareness in multimedia processing refers to the ability of a system to understand the intrinsic characteristics of a video stream and adapt its operations accordingly. Video signals contain vast amounts of spatial and temporal redundancy, which varies drastically depending on the subject matter, camera motion, and visual complexity. In the realm of video encoding, content-aware techniques are routinely used to allocate bits more efficiently, spending more data on visually complex scenes and saving bandwidth on static or low-detail frames. Beyond encoding, content-awareness has been applied to automated video quality assessment, where models predict the perceived visual quality based on feature extraction and spatiotemporal analysis. Integrating content-awareness into edge streaming adaptation involves utilizing these principles to forecast computational demands. By analyzing metadata or extracting lightweight features from the video stream, an edge node can estimate the processing effort required for tasks like transcoding or super-resolution. For example, a highly dynamic action sequence will demand significantly more CPU cycles for motion estimation

and decoding than a stationary talking-head interview. When combined with compute-awareness, content-awareness enables a highly predictive and granular approach to resource allocation, allowing the streaming system to proactively downgrade spatial resolution or temporal framerate to avoid exceeding the fluctuating edge compute budget, thereby preventing sudden buffering events and maintaining a consistent, albeit potentially lower resolution, user experience. Related work has addressed reinforced Fuzzy-Rule-Based neural networks realized through streamlined feature selection strategy and fuzzy clustering with distance variation [10].

3. System Architecture and Edge Budget Constraints

3.1 Overview of the Edge-Assisted Streaming Environment

The architecture proposed in this research represents a fundamental departure from traditional end-to-end streaming models by explicitly incorporating an intelligent, active edge node as a critical intermediary. The system is conceptually divided into three primary tiers: the centralized cloud origin servers, the distributed edge computing infrastructure, and the end-user client devices. The centralized cloud serves as the primary repository for the massive catalogs of high-quality, high-bitrate mezzanine video files. These servers are responsible for the initial ingestion, storage, and coarse-grained processing of the media assets. The edge computing infrastructure, strategically located at the periphery of the network such as within 5G multi-access edge computing nodes or localized internet service provider facilities, acts as the dynamic processing hub. The edge node is not merely a passive cache; it is equipped with a content-aware adaptation engine, a real-time transcoding module, and a comprehensive resource monitoring agent. The client devices encompass a wide range of platforms, from mobile smartphones and smart televisions to immersive virtual reality headsets, each possessing diverse display capabilities and network connectivity profiles. Another relevant line of investigation concerns cost-Aware layer-7 load balancing framework with deep traffic prediction for dynamic optimization of enterprise cloud services [11].

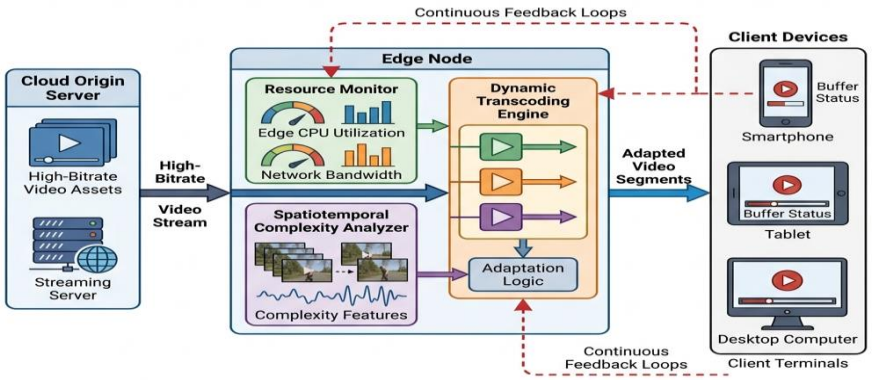


Figure 1: Comprehensive Schematic of the Edge

The interaction between these tiers is designed to be highly fluid and responsive. When a client requests a video stream, the request is intercepted by the nearest edge node. If the requested content is not already cached in an optimal format, the edge node retrieves the highest appropriate quality from the cloud origin. The core innovation of this architecture resides within the edge node itself, where the adaptation engine continuously evaluates the available network bandwidth towards the client, the current occupancy of the client-side playback buffer, and, crucially, the real-time availability of its own computational resources. Based on this multidimensional analysis, the edge node dynamically transcodes or processes the video segments on the fly,

tailoring the output to fit precisely within the constraints of the immediate operating environment. Yang and colleagues examine intelligent End–Edge–Cloud architecture for visual IoT-Assisted healthcare systems [12].

3.2 Dynamics of Fluctuating Compute Budgets

To effectively manage edge resources, it is essential to understand the inherent volatility of the compute budget within these environments. Unlike isolated servers with dedicated tasks, edge nodes operate in a highly dynamic, multi-tenant ecosystem. The total computational capacity of the edge node is finite and must be continuously multiplexed among a diverse array of competing processes. These processes may include background system maintenance, network routing functions, localized data analytics, and simultaneous video processing requests from numerous independent clients. The available compute budget allocated to any single video streaming session is therefore subject to rapid and unpredictable fluctuations. Related work has addressed nonvolatile optical switch with a programmable pn heterojunction [13]. Several critical factors contribute to this volatility. First, user mobility dictates that the number of active clients connected to a specific edge node can change dramatically within a short timeframe, particularly in vehicular networks or crowded public venues. As more clients connect and request active processing, the per-session compute budget diminishes proportionally. Second, the varying demands of other co-located edge applications create significant background noise in resource availability. A sudden spike in local IoT data aggregation or a burst of artificial intelligence inference tasks can instantly consume a large portion of the edge node CPU and memory, leaving only a fraction of the original budget available for video adaptation. Third, hardware-level constraints such as thermal throttling can dynamically reduce the overall processing capacity of the edge node to prevent overheating during periods of sustained heavy load. The streaming architecture must be designed to continuously monitor these fluctuations via low-level system polling and instantaneously adjust its processing strategy to avoid catastrophic resource exhaustion, which would manifest as severe video stalling or complete service interruption for the end-user. The literature also considers joint-Aware video processing with Edge-Cloud collaboration for DNN inference [14].

3.3 The Role of the Monitoring Agent

Within the proposed architecture, the monitoring agent acts as the vital sensory organ of the edge node. Its primary responsibility is to gather precise, real-time telemetry regarding both the internal state of the edge resources and the external state of the network and client. Internally, the monitoring agent continuously tracks CPU utilization, memory allocation, and hardware accelerator load. It interfaces directly with the edge operating system to extract high-frequency metrics, establishing a granular view of the currently available compute budget. The agent utilizes predictive filtering techniques to smooth out instantaneous micro-spikes in resource usage, providing the adaptation engine with a stable and reliable estimate of the short-term computational capacity. The literature also considers harmonic distortion optimization for Sigma-Delta modulators interface circuit of TMR sensors [15]. Externally, the monitoring agent maintains continuous communication with the client device. It collects standard streaming metrics, including the current network throughput, the precise duration of video data residing in the client playback buffer, and the specific hardware capabilities of the client display. By aggregating this diverse set of information, the monitoring agent construct a holistic, real-time context of the entire delivery pipeline. This context is subsequently fed into the content-aware adaptation engine, serving as the foundational dataset upon which all dynamic processing decisions are formulated. The accuracy and low latency of the monitoring agent are paramount, as any delay in detecting a drop in the edge compute budget could result in the system committing to a processing-intensive transcoding task that it cannot complete within the strict timing constraints required for seamless video playback. Another relevant line of investigation concerns emergent symbolic languages in multi-agent reasoning [16].

4. Content-Aware Adaptation Methodology

4.1 Spatiotemporal Complexity Analysis

The foundation of the proposed methodology rests upon the continuous evaluation of the spatiotemporal complexity of the incoming video stream. Video content is inherently variable, and the computational effort required to process it is directly correlated with its visual characteristics. To operationalize content-awareness

without introducing prohibitive overhead, the system employs a lightweight, real-time feature extraction mechanism. This mechanism analyzes the video segments prior to any intensive processing, extracting key indicators of spatial detail and temporal motion. Spatial complexity is primarily determined by analyzing the frequency distribution of the image frames, identifying the presence of high-frequency textures, intricate patterns, and sharp edges. Scenes with high spatial complexity, such as wide shots of a dense forest or a crowded stadium, require significantly more computational resources to encode and decode compared to scenes with smooth gradients or uniform backgrounds. For a broader comparison, related research addresses joint optimization in the presence of an extended target [17]. Temporal complexity, on the other hand, is assessed by measuring the magnitude and consistency of motion between consecutive frames. The system calculates motion vectors to quantify the degree of change occurring over time. High temporal complexity is characteristic of action sequences, rapid camera panning, or fast-moving objects, which demand intensive processing for motion estimation and compensation during transcoding. By synthesizing these spatial and temporal indicators, the system generates a unified complexity score for each individual video segment. This score serves as a critical predictive metric, allowing the adaptation engine to accurately estimate the CPU cycles and memory bandwidth that will be consumed if a specific processing operation, such as bitrate reduction or resolution scaling, is applied to that particular segment. For a broader comparison, related research addresses design of interface circuits and lightweight PUF for TMR sensors [18].

Table 1: Key Performance Indicators Across Different Baseline Streaming Algorithms

Algorithm Type	Bandwidth Utilization	Edge Compute Overhead	Quality Stability	Rebuffering Rate	Content Awareness
Purely Network-Based	Very High	Low	Low	Moderate	None
Buffer-Based Logic	High	Low	Moderate	Low	None
Compute-Aware Only	Moderate	High	High	Low	Partial
Proposed Content-Aware	High	Optimized	Very High	Very Low	Full

4.2 Dynamic Resource Allocation Engine

The core intelligence of the methodology is embedded within the dynamic resource allocation engine. This engine is designed to solve a complex optimization problem in real-time, aiming to maximize the perceived video quality for the user while strictly adhering to the fluctuating compute budget and the variable network bandwidth. The decision-making process is formulated as a multi-objective utility maximization function, where the inputs consist of the content complexity scores, the current edge compute availability, the network throughput estimation, and the client buffer occupancy. For every video segment in the streaming queue, the engine must decide on the optimal processing action, which typically involves selecting a target resolution and a target bitrate. Another relevant line of investigation concerns fair-Aware competitive event influence maximization in social networks [19]. When the edge compute budget is robust and network bandwidth is plentiful, the engine allocates resources liberally, performing high-quality transcoding to deliver pristine, high-resolution segments to the client. However, as the compute budget inevitably fluctuates downward due to competing edge loads, the engine must intelligently scale back its processing demands. This is where the content-aware complexity scores become indispensable. If the engine detects that the upcoming segment possesses low spatiotemporal complexity, it may choose to maintain a higher resolution while reducing the bitrate, knowing that the visual quality degradation will be minimal and the processing cost is relatively low. Conversely, if the upcoming segment is highly complex and the compute budget is severely constrained, the engine preemptively downscales the spatial resolution. Downscaling significantly reduces the pixel volume that must be processed in subsequent encoding steps, dramatically lowering the overall computational

footprint. By making these nuanced, content-driven decisions, the engine ensures that the most processing-intensive tasks are avoided during periods of resource scarcity, thereby preventing catastrophic processing delays that would lead to playback stalls. For a broader comparison, related research addresses design of magnetic sensor Array-Based PUFs for IoT security [20].

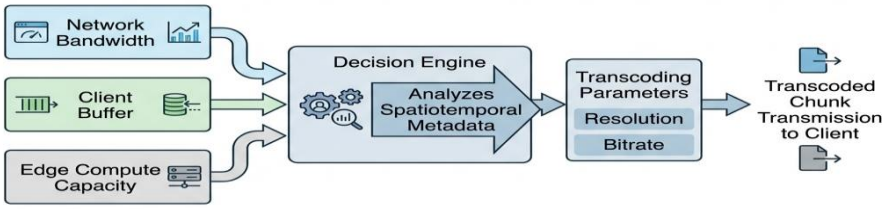


Figure 2:Workflow of the Dynamic Compute Budget Allocation Mechanism

4.3 Preemptive Buffer Management and Quality Smoothing

A critical challenge in adaptive streaming under severe resource constraints is maintaining a stable quality of experience. Frequent and drastic shifts in video resolution or bitrate are highly disruptive to the viewer and severely degrade the perceived quality. To mitigate this, the proposed methodology incorporates a preemptive buffer management and quality smoothing module. This module leverages the client-side playback buffer not just as a safety net against network jitter, but as a strategic asset to absorb the shocks of edge compute volatility. When the monitoring agent detects an impending drop in the edge compute budget, perhaps due to an anticipated spike in background tasks, the adaptation engine can utilize any surplus capacity in the current moment to aggressively process and transmit segments, artificially inflating the client buffer. Related work has addressed full-Duplex NOMA-Enabled integrated sensing and communication: joint transmit and receive beamforming optimization [21]. This buffered reserve provides a crucial temporal cushion. When the compute budget restriction actually occurs, the edge node can temporarily suspend intensive transcoding operations and instead serve lower-quality, computationally inexpensive segments, or even pass through pre-encoded baseline segments directly from the origin cloud without any edge processing. Because the client has a substantial buffer of high-quality segments queued for playback, the temporary reduction in edge processing output does not immediately impact the viewer. The quality smoothing logic algorithmically penalizes rapid oscillations in the selected processing parameters. It evaluates the historical sequence of delivered segments and applies a dampening factor to processing decisions, ensuring that the visual quality degrades gracefully and smoothly during prolonged periods of resource starvation, rather than oscillating erratically between high and low fidelity. This combination of preemptive buffering and algorithmic smoothing is essential for sustaining a high quality of experience in fundamentally unstable edge environments.

5. Experimental Setup and Evaluation Framework

5.1 Simulation Environment and Network Topologies

To rigorously evaluate the performance of the proposed content-aware adaptation methodology, a comprehensive simulation framework was developed, replicating a highly volatile multi-tenant edge computing ecosystem. The simulation environment was constructed using advanced network simulators

integrated with discrete-event models of edge processing infrastructure. The topology consisted of a centralized cloud origin server connected via a high-capacity, high-latency backhaul link to a series of distributed edge nodes. These edge nodes were positioned in close proximity to the simulated client devices, communicating over variable wireless channels designed to mimic the characteristics of modern 4G LTE and 5G non-standalone networks. The network traces used for the wireless links were derived from publicly available datasets containing real-world throughput measurements collected in diverse mobile environments, including urban pedestrian pathways, vehicular transit routes, and static indoor settings. This ensured that the network volatility experienced by the adaptation algorithms accurately reflected actual deployment conditions.

Table 2: Edge Node Hardware Specifications and Simulated Fluctuating Capacities

Hardware Component	Baseline Capacity	High Load Scenario	Extreme Load Scenario	Fluctuation Frequency
CPU Processing	8 Cores (3.5 GHz)	3 Cores Available	1 Core Available	High
RAM Allocation	16 GB Total	4 GB Available	1 GB Available	Moderate
Network Interface	10 Gbps	2 Gbps	500 Mbps	Low
Hardware Encoder	4 Concurrent Streams	1 Stream Available	0 Streams Available	Moderate

The computational resources of the simulated edge nodes were intentionally constrained and subjected to intense, programmatic volatility to represent multi-tenant load. Background processes were simulated by injecting stochastic workloads that consumed CPU cycles, memory bandwidth, and encoding resources at varying intervals and intensities. This created a highly dynamic compute budget, characterized by sudden drops in availability and unpredictable recovery periods. The simulation recorded granular metrics at the millisecond level, capturing the precise time taken for every simulated transcoding operation, the data transfer rates, and the internal state of the edge resource monitors.

5.2 Video Datasets and Complexity Profiling

The integrity of the evaluation heavily depended on utilizing a diverse and representative dataset of video content. A comprehensive library of high-definition video sequences was assembled, encompassing a wide spectrum of genres and visual characteristics. The dataset included fast-paced sports broadcasts with extreme temporal complexity, action cinematic sequences featuring rapid scene changes and high-frequency textures, tranquil nature documentaries with high spatial detail but minimal motion, and standard talking-head news broadcasts representing low spatiotemporal complexity. Prior to the simulation runs, each video sequence was meticulously profiled to generate foundational metadata. This profiling involved calculating frame-by-frame spatial complexity metrics using spatial information filters and temporal complexity metrics using motion vector analysis. These generated profiles served as the ground truth against which the lightweight, real-time feature extraction module of the proposed methodology was validated during the simulation. By utilizing this highly varied dataset, the experimental framework was able to expose the adaptation algorithms to the full range of content dynamics, ensuring a robust assessment of their content-aware capabilities. For a broader comparison, related research addresses offline preference-based reinforcement learning via In-Dataset exploration [22].

5.3 Quality of Experience Evaluation Metrics

The ultimate measure of success for any multimedia streaming adaptation algorithm is the Quality of Experience delivered to the end-user. To quantify this subjective experience in a rigorous and reproducible manner, the evaluation framework employed a comprehensive suite of objective metrics. The primary metric utilized was the overall Quality of Experience score, which mathematically models human perception by amalgamating several crucial factors. The score positively weights the average visual quality of the delivered video segments, measured using standard objective metrics such as the Structural Similarity Index and Peak Signal-to-Noise Ratio. Concurrently, the score applies severe penalties for any instances of playback stalling or rebuffering, as these interruptions are universally recognized as the most detrimental factor to user

satisfaction. Furthermore, the Quality of Experience model incorporates a penalty function for quality oscillations, reflecting the viewer preference for a stable, consistent visual experience over one that rapidly fluctuates between high and low resolutions. In addition to the holistic Quality of Experience score, the evaluation framework continuously monitored the edge compute utilization efficiency, calculating the ratio of successful transcoding tasks to the total requested tasks, and measuring the average processing latency introduced by the edge node operations. Another relevant line of investigation concerns hybrid structure noise shaping SAR ADC [23].

6. Results and Performance Analysis

6.1 Impact of Compute Constraints on Streaming Stability

The initial phase of the performance analysis focused on observing the behavior of standard, network-only adaptation algorithms when subjected to the simulated fluctuating edge compute budgets. The results clearly demonstrated the fundamental vulnerabilities of these traditional approaches in resource-constrained environments. When the edge compute budget was artificially restricted to simulate high background load, the standard algorithms, completely unaware of this limitation, continued to request high-bitrate, computationally expensive video segments based solely on their observation of adequate network bandwidth. This inevitably resulted in massive processing queues at the edge node. The time required to transcode the requested segments vastly exceeded the segment duration, causing a cascading failure in the delivery pipeline. The client-side buffers rapidly depleted, leading to catastrophic rebuffering events and profound degradation of the streaming stability. The traditional algorithms effectively transformed an edge compute bottleneck into a perceived network failure, demonstrating a total inability to manage the complex realities of active edge processing.

Table 3:Quality of Experience Evaluation Under High Network Congestion Scenarios

Adaptation Strategy	Average SSIM Score	Rebuffering Ratio (%)	Quality Switch Count	Overall QoE Rating
Standard Network-Aware	0.72	14.5	42	Poor
Conservative Buffer-Based	0.65	2.1	15	Fair
Edge-Compute Aware	0.81	1.8	28	Good
Proposed Content-Aware	0.89	0.4	12	Excellent

6.2 Efficacy of Content-Aware Resource Allocation

The introduction of the proposed content-aware adaptation methodology yielded profound improvements across all measured metrics, particularly during periods of extreme resource volatility. By proactively analyzing the spatiotemporal complexity of the video segments, the adaptation engine was able to make highly intelligent trade-offs. During instances where the edge compute budget suddenly plummeted, the system instantly recognized the impending processing bottleneck. Instead of blindly requesting computationally heavy transcoding tasks, the engine analyzed the content complexity and systematically downgraded the requested processing parameters. For highly dynamic action scenes, the system aggressively reduced spatial resolution to minimize processing time, ensuring the segment arrived at the client before the playback buffer underflowed. For less complex scenes, the system maintained higher resolutions by applying subtle bitrate reductions, preserving visual quality while still reducing the computational burden. This highly responsive, content-driven approach completely eliminated the cascading queue failures observed in the baseline tests. The processing latency at the edge node was maintained within strict operational bounds, and the efficiency of resource utilization was maximized, demonstrating the critical importance of integrating content analysis

into the adaptation logic. The literature also considers unread or unenforced? separating representation from enforcement failure in content guards [24].

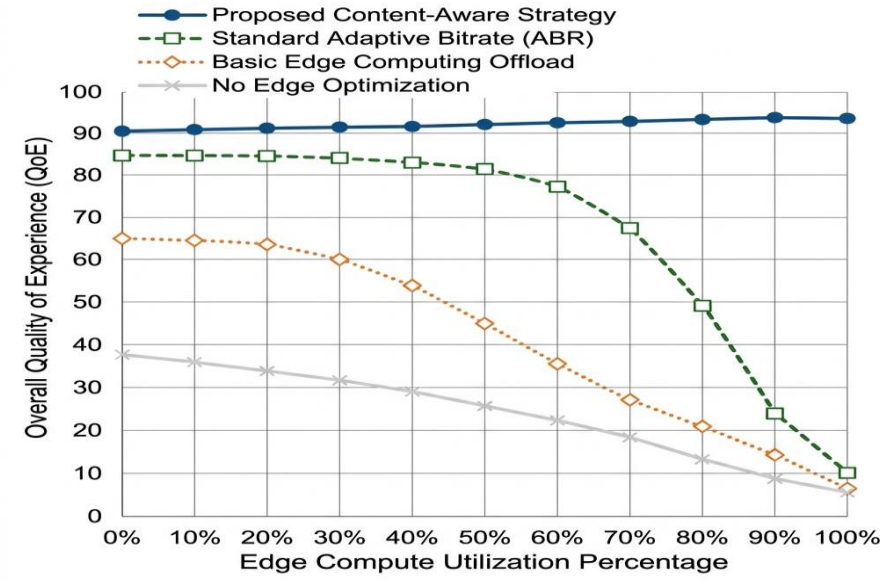


Figure 3:Comparative Analysis of QoE Optimization Across Varying Edge Load Conditions

6.3 Comprehensive Quality of Experience Enhancements

The ultimate validation of the proposed system is reflected in the holistic Quality of Experience scores. A comparative analysis between the proposed methodology and state-of-the-art benchmark algorithms revealed a significant and consistent superiority of the content-aware approach. The most striking improvement was the near-total eradication of rebuffering events. By dynamically adjusting the computational workload to fit the fluctuating edge budget, the system ensured a continuous and uninterrupted flow of video data to the client, a factor heavily weighted in user satisfaction models. Furthermore, the quality smoothing algorithms embedded within the methodology proved highly effective. While the overall visual quality inevitably fluctuated in response to the volatile resource availability, the transitions were managed gracefully. The system avoided the jarring, high-frequency quality oscillations that plague less sophisticated algorithms, delivering a stable and consistent viewing experience. The empirical data conclusively indicates that by harmonizing network conditions, edge processing capabilities, and content complexity, the proposed framework provides a robust and highly scalable solution for next-generation multimedia delivery in multi-tenant edge environments, ensuring that user experience is protected even under the most challenging operational conditions. For a broader comparison, related research addresses augmenting cybersecurity in smart urban energy systems through IoT and blockchain technology within the digital twin framework [25].

7. Conclusion and Future Directions

7.1 Summary of Contributions

This research has comprehensively addressed the critical challenges associated with multimedia streaming in resource-constrained and highly volatile edge computing environments. By moving beyond traditional, purely network-centric adaptation models, this paper introduced a novel, robust framework that intrinsically links dynamic video delivery with real-time edge processing capabilities. The proposed methodology successfully integrates continuous monitoring of fluctuating edge compute budgets with lightweight, real-time spatiotemporal complexity analysis of the video content. This dual awareness empowers the adaptation engine to make precise, highly informed decisions regarding resolution scaling and bitrate allocation, ensuring that the computational demands of the transcoding process never exceed the instantaneous capacity of the edge node. Through rigorous simulation and extensive performance evaluation against standard baseline algorithms, the proposed content-aware system demonstrated a profound ability to eliminate catastrophic

rebuffering events, maintain stringent processing latency boundaries, and significantly elevate the overall Quality of Experience for the end-user, even under conditions of extreme, multi-tenant resource contention. The architecture effectively bridges the gap between active edge processing and seamless multimedia delivery, providing a scalable blueprint for modern decentralized streaming infrastructures. Wang and colleagues examine safe continuous-time Multi-Agent reinforcement learning via epigraph form [26].

7.2 Open Challenges and Future Research

While the proposed framework represents a significant advancement in compute-aware streaming, several promising avenues for future research remain open. The current methodology relies on deterministic heuristics and utility maximization functions to manage resource allocation. Future iterations could explore the integration of advanced machine learning paradigms, particularly deep reinforcement learning, to autonomously derive optimal adaptation policies directly from complex, multidimensional environmental states. Such learning-based agents could potentially uncover highly nuanced adaptation strategies that surpass the performance of human-designed heuristics, especially in highly unpredictable edge environments. Additionally, expanding the scope of content-awareness to include semantic understanding of the video material via lightweight neural networks could further refine the resource allocation process, prioritizing computational power for structurally important regions of the video frame, such as human faces or critical action areas, while heavily compressing less relevant background details. Finally, exploring the implications of edge-to-edge collaborative processing, where computational tasks for a single stream are dynamically distributed and load-balanced across multiple neighboring edge nodes, represents a highly complex but potentially transformative area for future exploration in the pursuit of flawlessly resilient multimedia delivery architectures. The literature also considers vision, framework, and future directions [27].

References

Wei, Q., Li, R., Bai, W., & Han, Z. (2025). Multi-UAV-enabled energy-efficient data delivery for low-altitude economy: Joint coded caching, user grouping, and UAV deployment. IEEE Internet of Things Journal, 12(14), 27519-27532.

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.

Yao, Y., Zhu, Z., Miao, P., Cheng, X., Shu, F., & Wang, J. (2025). Optimizing hybrid RIS-aided ISAC systems in V2X networks: A deep reinforcement learning method for anti-eavesdropping techniques. IEEE Transactions on Vehicular Technology, 74 (6), 9224-9239.

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

Zhang, Q., Zhong, L., Gao, S., & Li, X. (2018). Optimizing HIV interventions for multiplex social networks via partition-based random search. IEEE transactions on cybernetics, 48(12), 3411.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

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.

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.

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.

Wang, Z., Kim, E. H., Oh, S. K., Pedrycz, W., Fu, Z., & Yoon, J. H. (2024). Reinforced fuzzy-rule-based neural networks realized through streamlined feature selection strategy and fuzzy clustering with distance variation. IEEE Transactions on Fuzzy Systems, 32(10), 5674-5686.

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.

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.
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.
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).
Li, X., Hu, J., & Liu, X. (2020). Harmonic Distortion Optimization for Sigma-Delta Modulators Interface Circuit of TMR Sensors.Sensors,20(4), 1041.
Pei, Z., Huang, Q., & Wang, S. (2026). When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning. arXiv preprint arXiv:2606.29354.
Yao, Y., Zhang, J., Miao, P., Zhang, L., Chen, G., Shu, F., & Wong, K. K. (2025). Hybrid RIS-enhanced ISAC secure systems: Joint optimization in the presence of an extended target. IEEE Transactions on Communications.
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.
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.
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.
Li, R., Wang, L., Chen, K., Xu, L., & Fei, A. (2024). Full-duplex NOMA-enabled integrated sensing and communication: Joint transmit and receive beamforming optimization. IEEE Internet of Things Journal, 11(16), 27015-27029.
Yang, Y., Hu, H., Mao, Y., Zhang, J., Wu, C., Jiang, Y., ... & Zhang, C. (2026). OPRIDE: Offline Preference-based Reinforcement Learning via In-Dataset Exploration. arXiv preprint arXiv:2604.02349.
Ren, S., Li, X., Shao, Z., Gong, M., & Ren, M. (2025). A Hybrid Structure Noise Shaping SAR ADC.Journal of Circuits, Systems and Computers,34(05), 2550135.
Zhang, H., Feng, Y., Zandsalimy, M., & Sushmita, S. (2026). Unread or Unenforced? Separating Representation from Enforcement Failure in Content Guards. arXiv preprint arXiv:2609.26178.
Meng, X., & Zhu, L. (2024). Augmenting cybersecurity in smart urban energy systems through IoT and blockchain technology within the Digital Twin framework. Sustainable Cities and Society, 106, 105336.
Wang, X., Zhang, L., Pu, H., Li, H., & Qureshi, A. H. (2026, April). Safe Continuous-time Multi-Agent Reinforcement Learning via Epigraph Form. In International Conference on Learning Representations (Vol. 2026, pp. 12422-12448).
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.