International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 11 - Issue 09 || September 2025 || PP. 05-12

Optimization of Video Content Delivery in Hybrid Cloud Architectures

Avinash Kumar
Software Development Engineer, Google Seattle, USA

Abstract: The article presents an analysis of architectural, network, and algorithmic factors that determine the quality of experience in hybrid cloud–edge video delivery systems, with a focus on their impact on delivery stability and QoE parameters. The study is conducted within an interdisciplinary paradigm combining distributed computing theory, network engineering, multimedia processing, and resource management in wireless networks. The methodological basis is a content analysis of peer-reviewed publications on low-latency bitrate adaptation, software-defined intercloud backbones with points of presence, GPU-accelerated encoding, and SDN control using the "Lyapunov drift-plus-penalty" function. Three interrelated groups of optimization mechanisms are identified and systematized: transport and routing (PoP-Overlay, jitter reduction, latency predictability), computational (multi-level caching, serverless GPU processing), and algorithmic (CMAF with chunked transfer, low-latency ABR strategies, content-aware scheduling). Three analytical tables are presented, including an intercontinental backbone comparison, HEVC accelerated encoding metrics, and comparative results of adaptive streaming algorithms. It is shown that the coordinated integration of transport, computational, and algorithmic solutions forms a predictable delivery loop that minimizes playback stalls and abrupt quality switches. The findings substantiate the need to design cloud–edge video systems based on QoE-oriented management models. The article will be of interest to researchers in multimedia networking, cloud and edge platform engineers, adaptive video streaming developers, and specialists in distributed architecture optimization.

Keywords: hybrid cloud–edge architecture, video content delivery, quality of experience, QoE-oriented management, low-latency bitrate adaptation, software-defined networking, points of presence, CMAF, edge caching, GPU-accelerated encoding.

Introduction

Current video delivery practice shows a steady rise in quality expectations alongside pressure to reduce latency within distributed, hybrid cloud infrastructures. The mix of public clouds, private data centers, and edge computing has become standard for video services, streaming platforms, and enterprise media. The need for optimization is driven by traffic volumes and the diversity of scenarios (on-demand, live, interactive), as well as the direct impact of network metrics on user experience and business outcomes.

Field observations indicate that perceptual degradation is "zoned" across the stack layers. On the client side, it appears as spikes in rebuffering and frequent quality shifts. In transport, it manifests as jumps in latency and jitter. At the edge, it involves radio-resource shortages and misaligned caching. Such heterogeneity of conditions and content profiles makes unified tuning of adaptation algorithms and network policies difficult, even with mature HTTP adaptive streaming technologies.

Interest in the topic is further fueled by the growing share of latency-sensitive services: low-latency streaming, event-driven live, and cloud interactive experiences. Every additional few tens of milliseconds here convert into lower engagement and revenue. Optimization ceases to be a local task and becomes an interdisciplinary project at the intersection of broadcast protocols, intercontinental connectivity, and software-defined edge control, requiring coordination of decisions along the entire delivery path.

The scientific literature highlights several key avenues of improvement. Low-latency bitrate adaptation paired with fragmented delivery reduces time to first frame and stabilizes quality, although the effect depends on segment duration, network profile, and content type [3]. Software-defined global network architectures with points of presence co-located with cloud data centers have demonstrated reductions in both latency and jitter and give operators end-to-end path control by the criteria of latency/reliability/security. QoE-oriented cloud–edge coordination models account for resource asymmetry and justify dynamic task and traffic placement, including the use of learning methods for prediction and adaptation. At the wireless edge, integrating SDN, caching, and scalable video coding with Lyapunov drift-plus-penalty control stabilizes buffers and reduces stalls without a priori channel statistics [10].

The aim of the study is to analyze the determinants of optimizing video delivery in hybrid cloud architectures, classify cross-layer factors, and synthesize principles of coordinated control based on results in low-latency quality adaptation, software-defined inter-cloud networks with points of presence, QoE-oriented cloud–edge streaming, and SDN-based control.

Practical cases from major platforms confirm the relevance of the topic. For instance, at Amazon, the transition from a monolithic architecture to distributed microservices eliminated bottlenecks in the shipment generation system, thereby enhancing the scalability and resilience of the global order fulfillment network.

Materials and Methods

This is a theoretical study based on content analysis of peer-reviewed publications on hybrid cloud–edge architectures for video delivery. The analytical framework was formed using the view of next-generation edge computing in M. Ergen et al. [1] and the survey of QoE-oriented adaptation in cloud–edge interaction by W. Wang et al. [9]. No experimental measurements or modeling were performed.

The literature corpus covers recent scientific publications. Alongside academic sources, the study takes into account the practical experience of implementing large-scale distributed solutions, including Prime Video's transition to the Startover Playback feature and the redesign of the subtitle pipeline into the JAB (just-after-broadcast) pipeline, as well as the re-architecture of Amazon's shipment generation platform. Architectures for caching and processing video blocks in hierarchical wireless networks were considered following D. Kim et al. [2], and cloud encoding stages and typical bottlenecks were taken from W. Moina-Rivera et al. [3]. The methodological device of tracing delivery paths and decision points relied on the concept of multi-hop accountability in J. Pennekamp et al. [4]. The relationship between edge function placement and quality-of-experience management was described by W. u. Rahman and E. N. Huh [5], while acceleration of encoding on serverless edge platforms with GPUs was drawn from A. Salcedo-Navarro et al. [6]. The comparison of traditional and low-latency adaptation algorithms followed S. Uddin et al. [7], and issues of organizing a hybrid backbone between distributed data centers followed J. Wang et al. [8]. Consolidated QoE terminology and metric classification were taken from W. Wang et al. [9], and aspects of centralizable control of video distribution at the wireless edge from T. Zhao et al. [10].

Synthesis was performed in three interrelated planes: architectural (placement of compute and caches); network (path and access control mechanisms); algorithmic (adaptive streaming strategies). For each plane, definitions, assumptions, and limitations were recorded as stated by the source authors, with mandatory citation. Redefinition of terms was not permitted. This strategy ensured comparability of conclusions without extrapolation beyond reported data and tied the solutions described in the literature to a unified conceptual apparatus.

Results

The transport layer of a hybrid cloud architecture governs the predictability of video delivery, the stability of buffering, and the controllability of the adaptive codec. M. Ergen [1] and W. Moina-Rivera [3] emphasize that variability of latency and loss in the backbone becomes a hidden limiting factor for QoE. Against this background, J. Wang [8] provides indicative results comparing classic On-Premises-Overlay modes with an On-POP-Overlay architecture that elevates route control to points of presence (PoPs). Table I presents key metric comparisons on intercontinental routes.

Table I: Intercontinental path metrics comparison (Source: [8])

Architecture / Link	**RTT avg (ms)**	**Jitter (ms)**	**Loss (%)**	**MOS (5)**	**Hops**	**Distance (km)**	**Packets**
On-Premises / IPsec VPN	224.2	1.5	0.1	4.07	6	17,000	55,197
On-Premises / Internet	222.1	1.4	0.7	4.03	21	17,000	50,036
On-Premises / MPLS VPN	213.6	2.0	0.1	4.10	16	17,000	109,189
On-POP-Overlay / Internet	192.09	0.53	0.823	4.13	17	20,000	110,325

In the traditional On-Premises-Overlay configuration, the lowest average latency is observed on the dedicated line (213.6 ms), yet jitter there is the highest (2.0 ms), exceeding both best-effort Internet and the IPsec tunnel. For streaming, this implies more frequent quality-level changes and a higher risk of micro-stalls. S. Uddin [7] demonstrated the decisive role of transport stability for low-latency algorithms. The modest advantage in subjective score (4.10) at minimal loss is offset by unstable segment delivery times, as reflected in J. Wang's intercontinental analysis [8].

Figure 1 illustrates the comparative performance of different transport architectures for intercontinental video delivery. The visualization shows that a PoP-controlled backbone provides more predictable latency and jitter, resulting in smoother playback and higher perceived quality compared to traditional on-premises or Internet-based approaches.

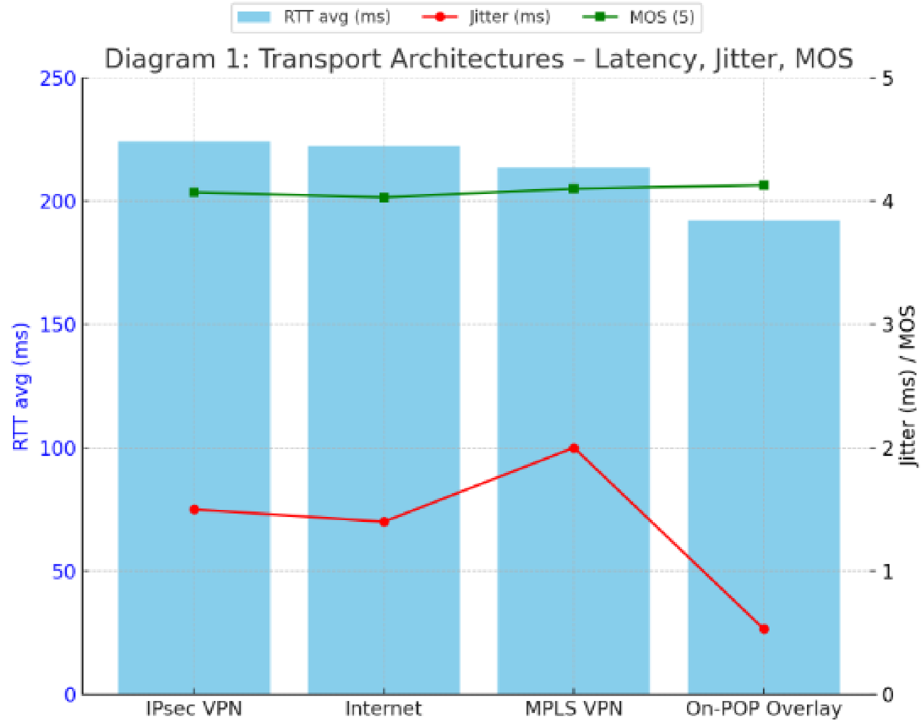

Figure 1 – Transport Architectures – Latency, Jitter, MOS (Compiled by the author based on source [8])

Comparing best-effort Internet and IPsec yields similar average latencies (222.1 vs. 224.2 ms) with substantially higher loss on the open Internet (0.7% vs. 0.1%). The score reduction to 4.03 aligns with content-aware optimization, where loss accelerates buffer drain and triggers quality downgrades, as shown by W. u. Rahman [5]. The slight latency overhead in IPsec is explained by processing costs, also noted by J. Wang [8].

The On-POP-Overlay architecture exhibits the key effect: despite the longer route, average latency drops to 192.09 ms and jitter to 0.53 ms owing to PoP-based routing and queue reduction along the path [4]. Stable segment arrival reduces both the frequency and amplitude of quality switches. Improved edge sampling efficiency for chunks is confirmed by D. Kim [2], and software-defined resource coordination supports uninterrupted playback under a variable radio channel, as shown by T. Zhao [10].

The aggregate MOS of 4.13 in the On-POP-Overlay case is explained by predictability of latency: for video services its contribution outweighs small differences in loss, as emphasized by W. Wang [9]. Added path predictability creates conditions for accelerated encoding and serverless edge pipelines, as A. Salcedo-Navarro shows for HEVC [6].

From a hybrid design standpoint, the conclusion is unambiguous: a PoP-controlled backbone yields more deterministic long-haul delivery and sets the envelope for caching decisions, edge compute placement, and encoding profile selection. This systemic view is reinforced in M. Ergen's survey [1]. The practical implementation of such principles is exemplified by the introduction of the Startover Playback feature in Prime Video. This technology enabled users to begin watching a live broadcast from the very start, even if they joined later, thus creating a DVR-like experience in live streaming. The global rollout of the feature increased audience engagement and retention, particularly during sports and entertainment events. A graphical analysis of the platform's internal metrics demonstrates an increase in viewing time per user and a reduction in session drop-off rates.

GPU-serverless effects for content preparation per A. Salcedo-Navarro [6]. Table II summarizes confirmed metrics from experiments on a local Knative functions platform for HEVC encoding with two implementations: hevc_nvenc (GPU-accelerated) and libx265 (CPU).

Table II: HEVC encoding outcomes in serverless-edge (Source: [6])

Metric / Scenario	**Value**	**Comment**
HEVC NVENC vs x265 (1 repr) speed	8.33× faster	Scenario 5 vs CPU baseline
Multi-resolution (5 repr) GPU vs CPU	12.43× faster	Mean per-segment gain
Best layout	1 fat VM + 4 slim replicas (Scen.5)	Encoding time improvement +15.2% vs other GPU scenarios
x265: best CPU layout	1 fat VM + 4 slim (Scen.2)	+8.42% faster among CPU scenarios
Cold start when requesting GPU	+10.47% / +17.44% / +36.50%	Scenario pairs (1–3 vs 4–6)
RAM reduction (GPU vs CPU)	−77.7% / −71.2% / −76.7%	(4 vs 1), (5 vs 2), (6 vs 3)
Mean GPU utilization	~10.5%	hevc_nvenc

The advantage of accelerated encoding is twofold. A. Salcedo-Navarro [6] shows that, for a single representation, preparation time shrinks by 8.33×, and for simultaneous generation of five representations it improves by 12.43×. This speeds construction of the adaptive "ladder" and reduces preparation latency—critical for perceived quality in cloud–edge adaptive delivery per W. Wang [9]. The same work captures the importance of placement: the "one fat VM plus four slim replicas" configuration minimizes encoding time for both GPU (+15.2% vs. other GPU scenarios) and CPU (+8.42% among CPU scenarios) [6]. This effect aligns with the benefits of peripheral caching and fragment distribution in D. Kim [2]. This effect aligns with the advantages of peripheral caching and fragment distribution demonstrated in the work of D. Kim [2]. Figure 2 illustrates the comparative performance of GPU-based (hevc_nvenc) and CPU-based (libx265) encoding in serverless-edge scenarios. As shown, GPU processing provides significant acceleration, achieves additional gains through optimal layout configuration, and markedly reduces memory consumption compared to CPU implementations.

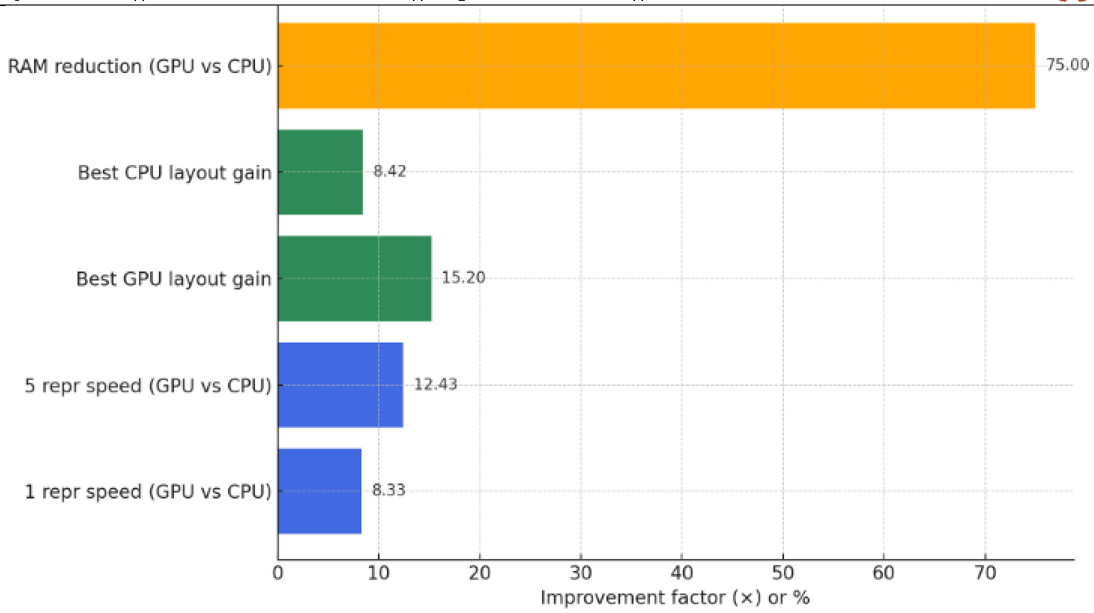

Figure 2 – GPU vs CPU performance in HEVC encoding under serverless-edge scenarios (Compiled by the author based on source [6])

Another area of optimization was the design and implementation of the subtitle processing system within Prime Video's JAB pipeline. The solution was targeted at devices with limited memory, such as legacy smart TVs. The segmentation of subtitles into short time blocks and the elimination of duplicate entries, formalized in patent US 10,893,331 B1, reduced resource usage and improved playback stability. Experiments demonstrated a significant reduction in memory consumption and the elimination of failures associated with incorrect subtitle rendering.

T. Zhao [10] shows that combining SDN, scalable video coding, and a Lyapunov drift-plus-penalty construction enables per-frame decisions on quality level and power allocation without a priori channel statistics. In simulation, user buffers stabilize around the target of 200 seconds by roughly the 80th time slot. The contrast with baselines is clear. Under "Max Arrive," the average buffer drifts toward ~250 seconds due to persistent delivery of low quality; under "Max Quality," buffers frequently hit zero, causing stalls. The proposed algorithm, by building the buffer, steadily raises quality and yields a higher overall perception score. These effects are documented via buffer trajectories, QoE dynamics, and comparisons against both benchmarks [10].

Embedding perceptual criteria in the edge control loop aligns with W. Wang's survey [9], where QoE components include quality level, the frequency and amplitude of switches, and rebuffering; these very terms enter the edge objective and penalty functions. Content-aware coupling of network loss, buffer depletion, and forced profile downgrades is emphasized by W. u. Rahman [5], in which loss and stalls directly map into QoE penalties and trigger client-side adaptations.

Discussion

Low-latency client-side ABR algorithms have become a keystone of hybrid delivery, because perceived quality depends on throughput, segment-arrival predictability, and buffer stability. Experimental studies show that combining short segments, CMAF, and chunked transfer reduces startup and inter-segment delays, regularizes the arrival cadence, and lowers stall probability [7]. These mechanisms are particularly effective when underpinned by stabilized transport and near-user caching: software-defined optimization of the backbone via PoPs reduces jitter to the benefit of adaptive clients [4], while multi-level caching and edge chunk fetching shorten time to first byte and cushion short-term channel dips [2]. Survey work underscores that such a coordinated "transport–edge–client" loop is foundational for QoE-oriented delivery in cloud–edge systems [9], and content-aware control penalizes rebuffering and sharp quality switches more heavily than minor fluctuations in average bitrate [5]. Given compute and energy limits at the edge, the need for robust strategies that are economical in buffer and startup time is emphasized in analyses of future wireless networks [1]. Table III consolidates key facts from comparative tests of low-latency client algorithms in DASH/CMAF, which inform the subsequent analytic interpretation.

Table III: Key facts on ABR/QoE in DASH/CMAF (Source: [7])

Scenario / Metric	**Result**
Mean bitrate and stability (all experiments)	L2A-LL: 2149.47 kbps, CV=0.6390; LOL+: 2054.21 kbps, CV=0.6462
Profile 2 (8→0.5→8 Mbps, 2 s)	LOL+: minimum rebuffering (0–1 events), with similar average quality; Throughput/Dynamic — more and longer rebuffering
10 s segments, complex profile (NP4)	L2A-LL > Throughput: p=0.0053, Holm 0.0529, BF10=8.147, g=0.917; L2A-LL > Dynamic: p=0.0076, Holm 0.0686, BF10=6.122, g=0.874
CMAF + Chunked Transfer	No playback interruptions for all algorithms across tests; faster startup
Interpretation	In a hybrid architecture, combining CMAF/CTE with low-latency ABR is beneficial for robust QoE under variable channels.

The obtained values show a consistent advantage of low-latency algorithms over simple throughput-based schemes. In S. Uddin's study, L2A-LL achieves a higher average bitrate with comparable variability, while LOL+ minimizes rebuffering in a stress profile with sharp channel swings [7]. For a complex scenario with 10-second segments, L2A-LL outperforms Throughput and Dynamic both statistically and practically, as confirmed by p-values, Holm corrections, Bayes factors, and Hedges' g [7]. Combined with CMAF and chunked transfer, all algorithms exhibited no playback interruptions and reduced startup time, which lowers buffer pressure and decreases the frequency of quality switches [3].

The linkage to transport and edge components is direct. Reducing jitter and dispersion of latency via PoP-based routing raises the predictability of segment arrival times and supports the functioning of low-latency strategies [4]. Edge placement and chunk fetching shorten the data path and limit the amplitude of brief troughs [2]. In QoE-oriented models, the ultimate gains are expressed not only in average bitrate but, above all, in fewer stalls and smoother quality profiles, as highlighted in content-aware perception optimization [5] and the cloud–edge adaptive streaming survey [9]. In the hybrid context, this yields a practical recommendation: jointly design the representation format (CMAF with chunked transfer), a transport with controlled pathing, and low-latency client adaptation, while accounting for the constraints of peripheral nodes noted in studies of future wireless networks [1]. At the corporate level, similar principles are confirmed by Amazon's experience. The transition from a monolithic shipment generation system to a distributed microservice architecture with workflow orchestration demonstrated that eliminating architectural bottlenecks and increasing throughput directly affect the resilience of operations and the manageability of complex interdependent processes. These results resonate with the patterns identified in QoE-oriented management of hybrid video systems.

In a hybrid architecture, content delivery rests on transport and media preparation and on an edge control loop where decisions are made about radio-resource allocation and quality level. The cloud–edge adaptive streaming survey underscores that practical QoE improvements arise when QoE metrics are explicitly embedded into decision algorithms and when client adaptation is closely coupled with resource planning at the edge [4]. This sets a methodological guidepost: criteria for quality, switching, and rebuffering should become part of the edge controller's objective rather than external network constraints.

T. Zhao [10] proposes an edge-control configuration combining SDN with caching and SVC, in which the SDN controller abstracts radio resources into a common pool and, at each discrete step, selects a "power–quality level" pair for each user online. The key mechanism is minimizing a Lyapunov drift-plus-penalty function, which stabilizes virtual buffer queues without a priori traffic or channel statistics. The problem decomposes into two subproblems—power allocation and quality selection—where the optimal allocation follows from Karush–Kuhn–Tucker conditions, and resource-block assignment follows a maximizing rule. Empirically, with such a regulator, user buffers converge to a steady regime by about slot 80, trending toward the upper bound of the target reserve, and integral perception scores exceed two indicative benchmarks: "Max Quality" (prioritizing maximal quality regardless of channel) and "Max Arrive" (prioritizing the number of delivered low-quality blocks) [10]. Visualized queue dynamics show a "hand-off" of power between users: once one client's buffer reaches the target, resources shift to the others, preventing buffer underflow and reducing the frequency of forced quality switches.

Coupling to W. Wang's theoretical frame [9] suggests a controller evolution: the penalty term can be parameterized directly by a QoE model that separately accounts for contributions of average quality, switching frequency/amplitude, and stall duration. This enables a shift from proxy metrics to direct optimization of perception—raising the weight of buffer-depletion events, tempering switching aggressiveness, and, where needed, introducing contextual coefficients by device and network profile. In practice, such a controller becomes a "conciliation node" between client adaptation and edge planning. The client receives a time-predictable set of chunks, while the edge, by managing power and selecting SVC layers, keeps buffers in a stable region. This "SDN controller with online optimization → stable queues → rare switches" loop is system-forming for hybrid delivery. It minimizes structural causes of QoE degradation and makes client algorithms resilient to channel variability, in line with recommendations for QoE-oriented design of cloud–edge systems [9].

In addition to the theoretical analysis, it is relevant to highlight a number of applied projects that demonstrate the industrial significance of the identified architectural principles.

Re-architecture of Amazon's shipment generation platform. The transition from a monolithic system to a distributed microservice architecture with workflow orchestration eliminated architectural bottlenecks and increased throughput. As a result, the system was able to handle significantly larger order volumes while maintaining the stability and fault tolerance of the global order fulfillment network.

Startover Playback feature in Prime Video. The development of a mechanism enabling users to start watching a live broadcast from the beginning provided a DVR-like streaming experience. The feature was rolled out globally and used for major sports and entertainment events, which enhanced audience engagement and retention.

Subtitle optimization in Prime Video's JAB pipeline. The creation of a segmentation and optimization system for subtitles on resource-constrained devices, formalized in US Patent 10,893,331 B1, reduced memory consumption and improved playback stability on legacy smart TVs.

These examples illustrate that the architectural, network, and algorithmic approaches identified in the study are validated by industrial practice and reinforce the article's conclusions on the necessity of comprehensive optimization in hybrid cloud systems.

Conclusion

The study identified key architectural, network, and algorithmic determinants that govern perceived video quality in hybrid cloud–edge systems and classified cross-layer mechanisms for their optimization. Based on the source analysis, transport parameters (latency, jitter, loss), adaptive streaming algorithms (segment duration, bitrate-selection strategy, representation format), and edge control (power allocation, caching, SVC encoding) were systematized, and their interrelations in the "transport–edge–client" loop were recorded. It was shown that stabilizing routes via PoPs and reducing jitter create conditions for the robust operation of low-latency algorithms, while integrating GPU-accelerated edge encoding cuts segment preparation time by orders of magnitude while reducing resource consumption.

Classifying edge-control methods by functional domains—transport/routing, compute, and algorithmic—confirmed that the greatest perceptual gains arise when they are applied in concert within QoE-oriented models. In particular, the use of software-defined backbones with PoP routing, multi-level caching, and low-latency ABR in CMAF with chunked transfer reduces rebuffering and smooths the quality profile. At the same time, an SDN controller optimizing a Lyapunov drift-plus-penalty criterion can stabilize buffer queues without a priori channel statistics and redistribute resources toward clients with low reserves, minimizing the probability of buffer depletion.

Comparative delivery analysis showed that the On-POP-Overlay architecture, despite longer routes, provides the lowest latency and jitter—critical for latency-sensitive services. GPU serverless models for HEVC encoding yielded multi-fold acceleration in building adaptive profiles and a 71–78% reduction in memory consumption, expanding the capabilities of resource-constrained edge nodes. These data confirm that holistic optimization spanning transport, content preparation, and edge control creates a resilient delivery loop with predictable QoE parameters.

Thus, optimizing video delivery in hybrid cloud architectures requires an integrated approach that unites architectural, network, and algorithmic solutions into a single QoE-oriented control system. The proposed factor classification and rationale for system-forming mechanisms provide methodological guideposts for designing next-generation cloud–edge video systems. Future research should focus on adaptive models that directly optimize subjective perception and on integrating traffic-prediction methods into the edge-planning loop.

References

[1]. Ergen, M., Saoud, B., Shayea, I., El-Saleh, A. A., Ergen, O., Inan, F., & Tuysuz, M. F. (2024). Edge computing in future wireless networks: A comprehensive evaluation and vision for 6G and beyond. ICT Express, 10(5), 1151–1173. https://doi.org/10.1016/j.icte.2024.08.007

[2]. Kim, D., Seo, D.-W., & Choi, M. (2024). Edge caching and computing of video chunks in multi-tier wireless networks. Journal of Network and Computer Applications, 226, 103889. https://doi.org/10.1016/j.jnca.2024.103889

[3]. Moina-Rivera, W., Garcia-Pineda, M., Gutiérrez-Aguado, J., et al. (2024). Cloud media video encoding: Review and challenges. Multimedia Tools and Applications, 83, 81231–81278. https://doi.org/10.1007/s11042-024-18763-2

[4]. Pennekamp, J., Henze, M., Wehrle, K., Niemietz, P., & Hiller, J. (2020). Private multi-hop accountability for supply chains. 2020 IEEE International Conference on Communications Workshops (ICC Workshops), 1–7. https://doi.org/10.1109/ICCWorkshops49005.2020.9145100

[5]. Rahman, W. u., & Huh, E. N. (2023). Content-aware QoE optimization in MEC-assisted mobile video streaming. Multimedia Tools and Applications, 82, 42053–42085. https://doi.org/10.1007/s11042-023-15163-w

[6]. Salcedo-Navarro, A., Peña-Ortiz, R., Claver, J. M., et al. (2025). Towards GPU-enabled serverless cloud edge platforms for accelerating HEVC video coding. Cluster Computing, 28, 68. https://doi.org/10.1007/s10586-024-04692-0

[7]. Uddin, S., Grega, M., Leszczuk, M., & Rahman, W. u. (2025). Evaluating HAS and low-latency streaming algorithms for enhanced QoE. Electronics, 14(13), 2587. https://doi.org/10.3390/electronics14132587

[8]. Wang, J., Bewong, M., & Zheng, L. (2024). SD-WAN: Hybrid edge cloud network between multi-site SDDC. Computer Networks, 250, 110509. https://doi.org/10.1016/j.comnet.2024.110509

[9]. Wang, W., Wei, X., Tao, W., Zhou, M., & Ji, C. (2025). Quality of experience-oriented cloud-edge dynamic adaptive streaming: Recent advances, challenges, and opportunities. Symmetry, 17(2), 194. https://doi.org/10.3390/sym17020194

[10]. Zhao, T., Jia, Y., Zhou, J., Liu, X., & Guo, Z. (2025). Video distribution strategy based on software defined network at the wireless edge. Digital Communications and Networks. Advance online publication. https://doi.org/10.1016/j.dcan.2025.06.004