Cost-Saving Streaming: Unlocking the Potential of Alternative Edge Node Resources

Yu Tian
tianyu21b@ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China

Zhenyu Li
zyli@ict.ac.cn
ICT, CAS
UCAS
Purple Mountain Laboratories
Beijing, China

Matthew Yang Liu
matthewliu58@google.com
Bilibili Inc.
Beijing, China

Jian Mao
maojian@bilibili.com
Bilibili Inc.
Beijing, China

Gareth Tyson
gtyson@ust.hk
Hong Kong University of Science and Technology (GZ)
Guangzhou, China

Gaogang Xie
xie@cnic.cn
CNIC, CAS
UCAS
Beijing, China

ABSTRACT

As the demand for online video content drives up bandwidth costs for content providers (CPs), there have been efforts to integrate cost-effective techniques to mitigate their bandwidth expenditure (*e.g.* using set-top boxes to share content). However, the use of such resources requires considerable effort to balance cost vs. user-perceived quality of service. This paper serves as a first step to quantify this trade-off. We collect and analyze data from a major CP that serves millions of users per day using both traditional CDN resources and alternative cheaper resources. Our analysis reveals that introducing cheaper alternative resources does *not* always yield anticipated cost savings and may lead to a reduction in quality of experience for users. We provide insights into the reasons behind these issues and propose strategies for better utilization of alternative network resources. We work with a major CP to deploy our proposals, and offer insights on how to better leverage different kinds of bandwidth resources for improved cost-efficiency and streaming delivery.

CCS CONCEPTS

• **Networks** → **Network measurement**; **Network performance analysis**; *Network services.*

KEYWORDS

Edge Node Resources, Cost-effective CDN, live streaming

IMC '24, November 4–6, 2024, Madrid, Spain

ACM ISBN 979-8-4007-0592-2/24/11
https://doi.org/10.1145/3646547.3689025

ACM Reference Format:
Yu Tian, Zhenyu Li, Matthew Yang Liu, Jian Mao, Gareth Tyson, and Gaogang Xie. 2024. Cost-Saving Streaming: Unlocking the Potential of Alternative Edge Node Resources. In *Proceedings of the 2024 ACM Internet Measurement Conference (IMC '24), November 4–6, 2024, Madrid, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3646547.3689025

1 INTRODUCTION

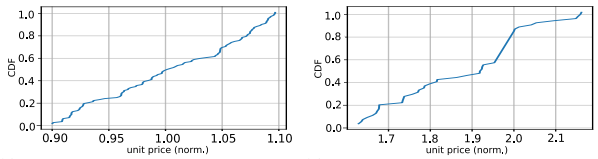
As the demand for streaming services continues to grow, optimizing delivery strategies and managing financial costs have become pivotal for content providers (CPs). To achieve this, they rely on Content Delivery Networks (CDNs), comprising geographically distributed edge nodes. These CDN edge nodes primarily originate from two sources: self-built infrastructure [1, 28, 32] and third-party cloud vendors [4, 7]. While these offer advantages (such as large bandwidth capacity and high stability), they are often associated with high bandwidth unit prices [4, 7]. As a result, bandwidth planning, which aims to generate cost-effective configurations for the use of edge resources, has emerged as a common practice for cost reduction [2, 21, 24, 33, 42, 43].

These high bandwidth costs have also been driving CPs to leverage "alternative" edge resources to further reduce bandwidth costs. These resources, such as set-top boxes with residential bandwidth leased by users, offer significantly lower unit prices, yet unpredictable availability [35, 38]. To date, there is a lack of research examining the practical deployment of these resources and their impact on cost savings and user-perceived quality.

To address this gap, we have gathered data from a large-scale live streaming CP that serves millions of users per day. The CP leverages hundreds of geo-distributed edge nodes, sourced from multiple providers, for bandwidth allocation in streaming delivery. Exploiting this data, we show that introducing lower-cost resources does not always yield expected cost savings. This unexpected outcome is primarily due to the discrepancy between the actual and planned percentile bandwidth usage, which is often caused by long persistent connections. Additionally, while the alternative resources demonstrate comparable first frame latency, they may result in higher buffering rates. We discovered that the primary cause of

Table 1: The overview of bandwidth capacity.

	MIN	MAX	AVG
traditional resources	14 Gbps	450 Gbps	155 Gbps
alternative resources	3 Gbps	17 Gbps	7.75 Gbps



(a) Distribution of anonymous unit price of *alternative resources*. (b) Distribution of anonymous unit price of *traditional resources*.

Figure 1: Distribution of anonymous unit price of *alternative resources* and *traditional resources*.

increased buffering rates is the significant mismatch between the advertised and the actual available bandwidth capacity. To address these challenges, we conduct improved deployment with strategies that take into account both connection time and available bandwidth capacity, showing notable reductions in both bandwidth costs (8.8% to 16.7%) and buffering rates (26.89%).

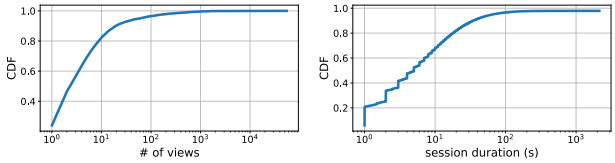
In summary, our contributions are:

- We characterize the use of alternative edge node resources, by gathering data from a large-scale live streaming CP, encompassing 100 million viewing sessions watched by approximately 10 million unique viewers daily.
- We evaluate the impact of introducing cheaper alternative resources on cost savings and user-perceived quality, identifying the reasons behind the unexpected observations. These insights offer a deeper understanding of the challenges in deploying alternative resources in streaming services and strategies for their optimal utilization.
- We propose strategies to enhance performance based on our observations, and the results of our re-deployment validate these strategies, offering CPs valuable insights on leveraging network assets for improved cost-efficiency and service delivery.

2 BACKGROUND

2.1 Edge Node Resources.

Video CPs strive to offer seamless video experiences to their users, relying on CDNs with geographically distributed edge nodes [1, 32]. Traditionally, these nodes originate from self-built CDNs [1, 28, 32] or cloud vendors [4, 7, 8], designed for substantial capacity and high redundancy, leading to elevated bandwidth unit prices, termed as *traditional resources* in this paper. Billing models for bandwidth usage typically follow two prevalent methods: (*i*) monthly percentile billing charged on the 95th-percentile bandwidth usage [33, 42, 43], and (*ii*) fixed billing based on a pre-set limit, regardless of actual usage [8, 31]. Recently, CPs have begun exploring cheaper resources from third-party vendors to augment the pricier CDN-hosted resources. These vendors aggregate idling resources [29, 31, 38], such as user-leased set-top boxes, into virtual edge nodes, known as *alternative resources*. These alternatives offer significantly lower unit prices but come with reduced stability.



(a) Viewer distribution. (b) Session duration distribution.

Figure 2: Distribution of viewers and session duration.

2.2 The Examined Streaming CP

The examined streaming CP, serving millions daily, employs a diverse edge system for content delivery. This system encompasses hundreds of edge nodes, including *traditional resources* from its self-built CDN and *alternative resources* rented from various third-party vendors. The ratio of *traditional resources* to *alternative resources* is roughly 3:7. Table 1 presents the statistics of bandwidth capacity among edge nodes. Over 90% of *alternative resources* exhibit a bandwidth capacity below 10Gbps, while more than half of *traditional resources* have a capacity exceeding 100Gbps. Furthermore, there is a considerable price disparity between *alternative resources* and *traditional resources* (see Figure 1), with an average ratio of 1:2. Meanwhile, *alternative resources* are billed based on daily percentile bandwidth usage [13] while *traditional resources* are billed based on monthly usage.

Most CDNs employ a centralized streaming center and a set of geo-distributed nodes to facilitate content distribution [23, 27]. The streaming center manages media processing and request scheduling. The CDN nodes are typically organized into a hierarchical structure, consisting of edge nodes that directly serve users and reflector nodes that reside between connect edges and the streaming center. This paper's focus (*i.e. alternative resources* and *traditional resources*) is on edge nodes. Note that the examined live streaming service offers fixed-rate streaming for viewers, where live channels with different bitrates are treated as distinct streams. Consequently, when a user switches resolution on the same channel, it is deemed as a request for a different stream. As to the mapping of individual viewing requests to edge nodes, rule-based schedulers map requests to the nearest or cheapest edge node, while planning-based schedulers solve optimization problems to map requests [12]. The examined CP'CDN uses a planning-based approach, optimizing bandwidth costs by configuring the billable bandwidth of edge nodes as in [33]. The chances being assigned new requests to an edge node is proportional to the gap between the billable bandwidth and the current bandwidth usage. The scheduler will no longer assign new requests to a node if its bandwidth usage exceeds its configured billable bandwidth. The alternative resources were initially incorporated into the examined CDN by simply taking them as traditional edge nodes (referred to as INIT deployment).

3 MEASUREMENTS & ANALYSIS

In this section, we gather one-month production data from a large streaming CP to characterize *alternative resources* and *traditional resources*, attempting to answer the following questions: (*i*) Do *alternative resources* reduce bandwidth cost? If so, to what extent? (*ii*) Do they have any impact on user-perceived quality? If so, what kind of impact do they have?

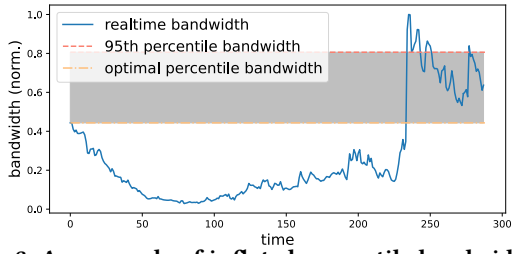


Figure 3: An example of inflated percentile bandwidth utilization.

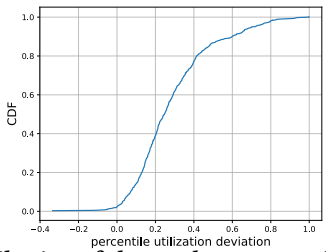


Figure 4: Distribution of the gap between optimal and actual 95th bandwidth utilization.

3.1 Dataset Collection

The production data consists of two datasets, gathered separately from the node-level and the viewing session-level.

Edge Node Dataset: This dataset encompasses detailed records from hundreds of edge nodes operated by the examined CP. *Traditional resources* span 34 cities, while *alternative resources* cover 24 cities. For each node, we document its type (*traditional* or *alternative resources*), the bandwidth capacity as stated by resource providers, the billing method along with unit price, and real-time bandwidth usage recorded every 5 minutes over the month.

Streaming Session Dataset: We also gather statistics at the viewing session-level. On average, our dataset captures over **100 million viewing sessions** per day, watched by approximately **10 million unique viewers**. To provide a clearer profile of streams and viewing sessions, we present a snapshot of the distribution of viewers and session duration in Figure 2. Figure 2(a) shows the Cumulative Distribution Function (CDF) of the number of viewers per stream at a given moment, while Figure 2(b) presents the CDF of viewing session durations for streams with at least one viewer. Notably, more than 80% of streams have fewer than 10 viewers, and over 20% of views last less than one second. We concentrate on three key metrics: (*i*) First frame latency, which gauges the time between the client initiating a content request and the arrival of the first frame; (*ii*) Buffering rate, indicating the frequency of streaming interruptions. Specifically, we compute the buffering rate for a viewing session as $\frac{\#buffering}{session\ duration}$. (*iii*) Serving edge node, identifying the specific node that facilitated the streaming session.

3.2 Analysis of Cost Savings

In this section, we investigate the degree to which the integration of *alternative resources* actually lowers bandwidth costs and whether these savings align with our projections.

Methodology. We first assess the cost savings achieved through the INIT deployment described in Section 2.2 when compared to a deployment solely relying on traditional nodes. This analysis aims to determine whether the introduction of *alternative resources* indeed leads to cost reductions. Then, to investigate whether these savings align with our projections , we solve an optimization problem (detailed in Appendix A), using the demand data from the entire month and the attributes of *alternative resources* and *traditional resources* as input, to calculate the theoretical minimum achievable bandwidth cost, denoted as *Oracle*. By comparing with Oracle, we can understand the gap between the cost of INIT and the minimum cost, as well as the space for further cost savings. Given that cross-regional traffic transmission can introduce additional latency, we limit traffic allocation within a single network region[1]. The comparison results across various network regions are presented in Table 2 where negative values indicate that INIT incurs lower costs and positive values indicate higher costs.

Table 2: The cost difference of INIT compared to the *Oracle* and the traditional-nodes-only deployment (tra-only) in various network regions.

Regions	R1	R2	R3	R4	R5	R6	R7
tra-only	-11.4%	-12.1%	-13.6%	-15.4%	-16.8%	-18.2%	-19.4%
Oracle	+36.1%	+46.3%	+45.3%	+28.9%	+28.1%	+37.9%	+32.9%

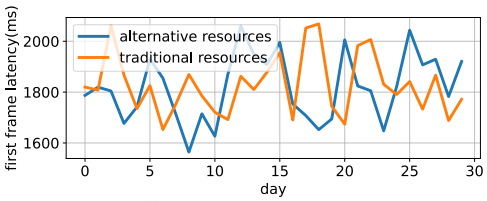
Comparison with Traditional-Nodes-Only Deployment. The cost savings achieved by INIT, when compared to a deployment solely relying on traditional nodes, vary from 11.4% to 19.4% across different network regions. These savings are primarily due to: (*i*) Lower unit costs: The presence of *alternative resources* decreases the overall bandwidth needed from *traditional resources*, and their lower unit price further reduces the total bandwidth cost. (*ii*) Decreased billable bandwidth: Node utilization during 5% of the time intervals does not affect the 95th percentile cost calculation, effectively making these intervals "free" in terms of cost. As the number of nodes increases, the collective "free" intervals also increase, allowing content providers to keep the 95th percentile bandwidth low and reduce the total billable bandwidth.

***Comparison with* Oracle.** We next compare the initial deployment against the theoretical optimum. The *Oracle* saves 28.1% to 46.3% more bandwidth cost across different network regions, indicating that there is still significant room for improvement before reaching our expectations.

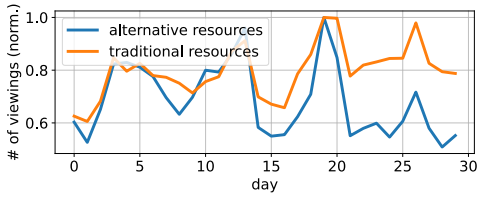
To understand the gap from *Oracle*, we illustrate the inflation of 95th percentile bandwidth utilization in Figure 3. The blue line shows real-time bandwidth, while the red and yellow dashed lines indicate the actual and optimal 95th percentile bandwidth, respectively. The shaded area highlights the gap between the actual and optimal values, arising from misalignment between the node's actual usage and theoretical configuration. We calculate the gap and present the cumulative distribution function (CDF) plot in Figure 4. Among various *alternative resources*, the gap ranges from -0.4 to 1.0, with only 1.3% of nodes having a negative gap and approximately 50% having a gap higher than 0.25, significantly increasing costs compared to *Oracle*.

This observation can be attributed to two main factors: (*i*) *Long Persistent Connections:* Live streaming connections, once established,

[1]In our study, users within the same AS (Autonomous System) and metropolitan area are grouped into one user group, which corresponds to a specific network region. This grouping is based on their IP prefix.



(a) Tail first frame latency.



(b) Number of viewing sessions.

Figure 5: First frame latency comparison.

tend to remain open to ensure low-latency streaming. These connections continuously consume bandwidth until released. The INIT scheduler, unaware of connection duration, causes bandwidth utilization to deviate from its optimal configuration due to these long-lasting connections. While *traditional resources*, billed monthly, can adjust utilization over subsequent days to mitigate percentile value impacts, *alternative resources*, billed daily, must bear unexpectedly high daily percentile values since bandwidth utilization on different days does not affect each other. A connection time-aware request scheduler that prevents *alternative resources* from suffering long consistent connections could reduce the gap issues, unlocking further potential for cost savings. (*ii*) *Inaccuracy of Traffic Prediction:* Offline planning relies on historical data to forecast demand for each time interval in the coming month. However, such predictions often lack precision, leading to gaps from expected node bandwidth. Research on better prediction of high dynamic traffic will be left for future work.

3.3 Analysis of Quality of Experience (QoE)

We next examine the QoE attained by *alternative resources* in terms of first frame latency and buffering rate.

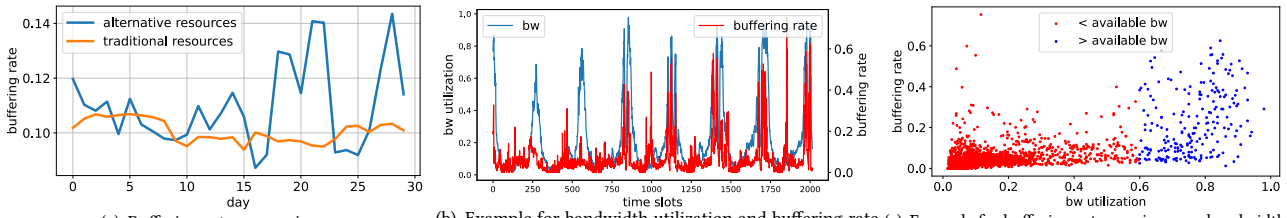
First frame latency. In streaming services, the first frame latency is a crucial performance indicator. Figures 5(a) and 5(b) present a comparative analysis of the 90th percentile tail latency and the normalized serving sessions for both *traditional resources* and *alternative resources* throughout a one-month period. The normalization of serving sessions is based on the maximum daily sessions for each node type during that month.

We see that both node types exhibit comparable fluctuations in serving requests over time. Notably, the Spearman's rank correlation coefficient between first frame latency and the volume of serving sessions is calculated as -0.02 (p-value=0.91) for *alternative resources* and -0.07 (p-value=0.71) for *traditional resources*, indicating that there is no discernible connection between the first frame latency and the volume of serving sessions. The average 90th percentile latency for *alternative resources* and *traditional resources* during the monitored month amounts to 1,816ms and 1,827ms, respectively. Remarkably, even with a comparatively smaller bandwidth capacity, *alternative resources* exhibit latency equivalent to that of *traditional resources*. This observation can primarily be ascribed to the closer physical proximity of *alternative resources* to the end-users and implies that, despite limited bandwidth, deploying *alternative resources* can yield competitive first frame latency.

Buffering rate. To further investigate the performance of *alternative resources*, we compare the buffering rate between *alternative resources* and *traditional resources*. This is because the buffering rate is one of the factors that has the greatest impact on user engagement [14]. The buffering rate is calculated as the total number of buffering events divided by the total length of the viewing sessions served by the corresponding edge resources (or specific edge node). Therefore, the buffering rate is normalized by the total viewing duration, minimizing the impact of the distribution of individual viewing session duration. Figure 6(a) presents the average buffering rate of *alternative resources* and *traditional resources* over a month and we can see disparities in both the variance and mean values of buffering rates between *alternative resources* and *traditional resources*. Notably, while *alternative resources* exhibit a slightly higher buffering rate (0.1097) compared to *traditional resources* (0.1006), the difference in average buffering rates is statistically insignificant, as indicated by a t-test (t-statistic=1.42, p-value=0.15). However, a concerning aspect is the notably higher standard deviation observed in *alternative resources* (0.0149), indicating intermittent spikes or surges in buffering, potentially compromising the streaming quality of experience. Conversely, *traditional resources* demonstrate a far more stable performance with a significantly lower standard deviation (0.0040).

We explore factors contributing to fluctuating buffering rates in *alternative resources*. A clear trend emerges: a correlation between bandwidth utilization and buffering rates. Figure 6(b) visualizes this relationship over a week for one node, showing that buffering rates tend to increase during peak traffic hours. To further support this finding, Figure 6(c) plots buffering rates against bandwidth utilization of one node. When bandwidth utilization exceeds a threshold (0.6 in this case), the buffering rate jumps significantly, with increased fluctuations. At higher bandwidth utilization, the average buffering rate is 0.2213 (standard deviation 0.1566), much higher than the 0.0499 rate (standard deviation 0.0522) at lower utilization. The Pearson correlation coefficient is 0.32 at higher bandwidth utilization and 0.28 at lower, indicating a weak positive correlation between buffering rate and bandwidth utilization in their respective intervals.

To generalize our findings, we next group *alternative resources* based on their advertised bandwidth capacities and randomly selected nodes from each group. We then plot the variation of average buffering rates across different bandwidth utilization bins in Figure 7, where the bin size is 0.05 and the group ID increases with bandwidth capacity. Each line in the graph exhibits an elbow point, indicating the practical bandwidth utilization limit of each edge node. This finding underscores a key disparity between the advertised and actual available bandwidth capacities of *alternative resources*. Besides, we don't see a significant fluctuation in buffering rates across various *traditional resources*, nor have we observed such a disparity between the advertised and actual available bandwidth capacities of *traditional resources* (as shown in Figure 9). It urges content providers to identify these limits for *alternative resources*,



(a) Buffering rate comparison.

(b) Example for bandwidth utilization and buffering rate varying over time.

(c) Example for buffering rate varying over bandwidth utilization.

Figure 6: The buffering rate of *alternative resources*.

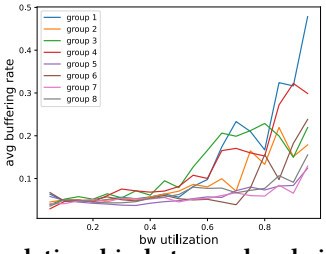


Figure 7: The relationship between bandwidth utilization and buffering rate.

perhaps using the elbow method, and take proactive steps to maintain bandwidth usage below the available bandwidth capacities to ensure optimal streaming performance.

3.4 Summary & Takehomes.

Our findings reveal that the introduction of cheaper *alternative resources* does *not* always yield anticipated cost savings and can lead to increased buffering rates. Furthermore, we offer the following two insights to improve cost savings and enhance the performance: (*i*) Prolonged persistent connections can lead to significant gaps from the optimal node bandwidth configuration. Implementing a connection time-aware request scheduling strategy can mitigate these gaps, unlocking greater potential for cost savings. (*ii*) There is often a discrepancy between the advertised bandwidth capacity and the actual available bandwidth. Identifying the true bandwidth capacity of alternatives and scheduling requests to stay within each node's practical bandwidth limits can significantly reduce buffering rates and enhance stability.

4 IMPROVED STRATEGIES

With the above analysis in-mind, we propose some practical improvements for *alternative resources*. We focus our efforts on simple and heuristic approaches that can easily scale-up in production environments.

4.1 Key strategies

Connection time-aware request scheduling. Our previous findings underscore the influence of connection duration on bandwidth cost savings as the connection continuously consumes the bandwidth until actively terminated by the client. In the context of live streaming, CPs predominantly use the HTTP-FLV (Flash Video over HTTP) [19] and HLS (HTTP Live Streaming) [30] protocols to facilitate video streaming. Since HLS relies on short HTTP connections, it is more resilient to short-lived and intermittent connections, making it a suitable candidate for *alternative resources*. On the other hand, HTTP-FLV, with its persistent connections, requires a more stable environment to maintain a smooth streaming experience. Therefore, we propose to prioritize redirecting FLV requests to *traditional resources*, while the remaining HLS requests can be prioritized for *alternative resources* instead. This approach aims to prevent the serious deviation of the real-time bandwidth utilization of *alternative resources* from its optimal configuration, minimizing bandwidth costs as much as possible.

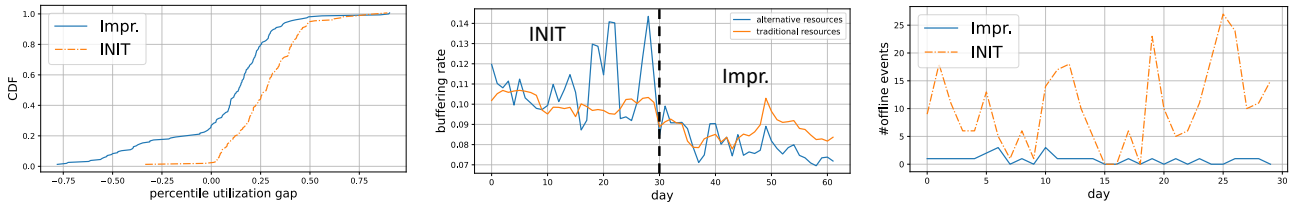
Note that we acknowledge that other factors, such as historical viewing patterns and video content characteristics, can also inform connection duration. While gathering such data is more time-consuming, it holds promise for further optimization. However, even without these additional data, our protocol-based approach represents a substantial step forward.

***Identify available bandwidth utilization for* alternative resources.** Given that there is often a discrepancy between the advertised bandwidth capacity and the practical available bandwidth, coupled with the increasing buffering rates when the bandwidth surpasses the available limit, we need to identify the practically available bandwidth for *alternative resources*. To ascertain the practically bandwidth capacity of each alternative node, we employ the elbow method. Specifically, we construct a curve, as exemplified in Figure 7, and visually pinpoint the elbow point where the curve starts to rise steeply. This elbow point is interpreted as the node's available bandwidth utilization. Subsequently, we enforce bandwidth utilization limits for *alternative resources*, ensuring they do not exceed their respective capacities.

Additionally, we recognize that streaming traffic patterns exhibit distinct diurnal and weekly patterns, with peak evening hours witnessing significantly higher volumes. To address the bandwidth pressure on *alternative resources* during peak periods, we recommend implementing dynamic bandwidth utilization thresholds for peak and non-peak hours. Specifically, during peak hours, we will lower the available bandwidth utilization thresholds to prevent *alternative resources* from approaching saturation too quickly under the high volume of concurrent requests. This strategy aims to maintain a consistent and reliable streaming experience for users while optimizing cost-efficiency.

4.2 Results

To test the efficacy of the innovations mentioned, we perform a redeployment that incorporates the improved strategies, utilizing the same set of resources in the initial deployment. Over a one-month period, we gather data including *edge node dataset* and *streaming session dataset* as described in Section 3.1. The key differences between the improved deployment and INIT are as follows: (*i*) In the planning phase, a similar optimization model is used, but with additional constraints to ensure that the output bandwidth limit for



(a) Percentile BW utilization distribution. (b) Buffering rate reduction. (c) The number of offline events of *alternative resources*.

Figure 8: The results of improved deployment.

Table 3: The cost savings in various network regions of improved deployment (Impr.) compared with INIT.

Regions	R1	R2	R3	R4	R5	R6	R7
Impr.	13.2%	12.1%	10.6%	8.8%	9.5%	16.7%	12.4%

alternative resources does not exceed their available bandwidth. (*ii*) In the request mapping phase, the scheduler outputs edge nodes for FLV and HLS requests separately. It's worth noting that there was no significant difference in the volume of serving requests between the two months, as confirmed by t-test for *alternative resources* (t-statistic=0.33, p-value=0.75) and *traditional resources* (t-statistic=0.89, p-value=0.38). We briefly detail our findings below.

The savings are detailed in Table 3. We attain substantial cost savings ranging from 8.8% to 16.7% across various network regions when compared to the initial deployment.

We then illustrate the distribution of the gap between the percentile bandwidth utilization and planned limit in the improved deployment, as shown in Figure 8(a). Approximately 62.5% of *alternative resources* exhibit an absolute value of percentile bandwidth deviation within 0.25. This finding underscores that our improved deployment aligns more closely with *Oracle*, thereby improving cost savings.

Figure 8(b) plots the daily buffering rates for both *alternative resources* and *traditional resources*. Compared to the initial deployment, a notable decline (26.89%) in the buffering rate of *alternative resources* is evident. This reduction highlights the effectiveness of our approach in managing buffering rates through the identification of available bandwidth capacity. Additionally, we utilize the 3-sigma principle, classifying sessions with buffering rates exceeding $\bar{b} + 3 * \sigma_b$ as abnormal. We find that sessions experiencing abnormally elevated buffering rates have decreased by 17.1%.

Furthermore, during the INIT, we observed several offline events involving *alternative resources* as edge resources are forcibly taken offline when their buffering rates exceed a predefined threshold. Figure 8(c) presents the comparison of daily offline events between the initial and improved deployments, showing a substantial drop in forced offline occurrences from over 10 incidents per day to no more than 3. This significant improvement can further underscore the efficacy of our strategy in managing the buffering rate, which is attributed to the identification of available bandwidth capacity.

4.3 Discussion

While this paper highlights the cost savings and QoE benefits of *alternative resources*, several challenges may limit their broader adoption. *Traditional resources*, sourced from the CP's own CDN, offer better management and stability. In contrast, *alternative resources* often face frequent updates and replacements, leading to less consistent performance and reduced resilience against unexpected traffic surges. Future research should explore solutions to these issues to improve practical deployment.

5 RELATED WORK

Bandwidth cost optimization. With the surge in bandwidth demand, optimizing bandwidth cost has become crucial for CPs. Research has extensively used MILP frameworks to model cost-saving strategies, focusing on various billing methods[10, 16, 20, 22, 25, 36, 37, 40]. While linear billing methods have been well-explored[2, 24, 26, 42], recent studies have addressed cost-saving opportunities under the monthly 95th-percentile billing in WANs [21, 34] and cloud networks [33, 43]. Additionally, cheaper edge node resources, such as P2P systems used for content delivery [9, 18, 39, 41], offer potential for cost reduction. However, optimization for these alternative resources remain unexplored.

Measuring video delivery systems. There is extensive literature on measuring video delivery systems. Some works analyze the impact of the video quality on user engagement [14, 15] while some studies examine how various factors affect performance and user experience [11, 17, 32]. There are also efforts in performance enhancement through multi-dimensional routing strategies [3, 5, 6, 32]. Prior research lacks a focus on billing and performance of alternative edge resources versus traditional CDN nodes in production. Our work identifies opportunities and provides insights into how to achieve a balance between performance and cost savings by selectively leveraging alternative edge node resources.

6 CONCLUSION

This paper has presented a characterization of the use of *alternative resources* in a video streaming context. We have identified the potential for improved cost-effectiveness and discussed the issue of increased buffering rates. Our findings reveal that the gap between the percentile bandwidth of *alternative resources* from the planned one caused by serving long persistent connections leads to the limited cost savings. We identify that the significant gap between claimed bandwidth capacity and practically available bandwidth capacity is the main reason for the inflated buffering rates of *alternative resources*. Accordingly, we propose a connection time-aware and available bandwidth capacity-aware strategy to improve the deployment of *alternative resources*.

7 ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. This work was supported in part by the National Key R&D Program of China (2022YFB2901800), Natural Science Foundation of China (U20A20180, 62072437). Corresponding Author: Zhenyu Li.

REFERENCES

[1] Vijay Kumar Adhikari, Yang Guo, Fang Hao, Matteo Varvello, Volker Hilt, Moritz Steiner, and Zhi-Li Zhang. 2012. Unreeling netflix: Understanding and improving multi-cdn movie delivery. In *2012 Proceedings IEEE Infocom*. IEEE, 1620–1628.

[2] Micah Adler, Ramesh K Sitaraman, and Harish Venkataramani. 2011. Algorithms for optimizing the bandwidth cost of content delivery. *Computer Networks* 55, 18 (2011), 4007–4020.

[3] Adnan Ahmed, Zubair Shafiq, Harkeerat Bedi, and Amir Khakpour. 2017. Peering vs. transit: Performance comparison of peering and transit interconnections. In *2017 IEEE 25th International Conference on Network Protocols (ICNP)*. IEEE, 1–10.

[4] Akamai. 2024. *Akamai Cloud CDN*. https://www.akamai.com/solutions/content-delivery-network

[5] Aditya Akella, Bruce Maggs, Srinivasan Seshan, and Anees Shaikh. 2008. On the performance benefits of multihoming route control. *IEEE/ACM Transactions on Networking* 16, 1 (2008), 91–104.

[6] Aditya Akella, Bruce Maggs, Srinivasan Seshan, Anees Shaikh, and Ramesh Sitaraman. 2003. A measurement-based analysis of multihoming. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*. 353–364.

[7] Alibaba. 2024. *Alibaba Cloud CDN*. https://www.alibabacloud.com/zh/product/content-delivery-network/pricing

[8] Amazon. 2024. *Amazon Cloud Services Pricing*. https://aws.amazon.com/cn/cloudfront/pricing

[9] Nasreen Anjum, Dmytro Karamshuk, Mohammad Shikh-Bahaei, and Nishanth Sastry. 2017. Survey on peer-assisted content delivery networks. *Computer Networks* 116 (2017), 79–95.

[10] Jeremy Bogle, Nikhil Bhatia, Manya Ghobadi, Ishai Menache, Nikolaj Bjørner, Asaf Valadarsky, and Michael Schapira. 2019. TEAVAR: striking the right utilization-availability balance in WAN traffic engineering. In *Proceedings of the ACM Special Interest Group on Data Communication*. 29–43.

[11] Hyunseok Chang, Sugih Jamin, and Wenjie Wang. 2009. Live streaming performance of the Zattoo network. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. 417–429.

[12] Huan Chen, Huiyou Zhan, Haisheng Tan, Huang Xu, Weihua Shan, Shiteng Chen, and Xiang-Yang Li. 2022. Online Traffic Allocation Based on Percentile Charging for Practical CDNs. In *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*. IEEE, 1–10.

[13] JD Cloud. 2024. *JD Cloud Services*. https://m-console-buy.jdcloud.com/init?product=Z

[14] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. 2011. Understanding the impact of video quality on user engagement. *ACM SIGCOMM computer communication review* 41, 4 (2011), 362–373.

[15] Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. 2007. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 15–28.

[16] Kathrin Hanauer, Monika Henzinger, Lara Ost, and Stefan Schmid. 2023. Dynamic Demand-Aware Link Scheduling for Reconfigurable Datacenters. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*.

[17] Xiaojun Hei, Chao Liang, Jian Liang, Yong Liu, and Keith W Ross. 2007. A measurement study of a large-scale P2P IPTV system. *IEEE transactions on multimedia* 9, 8 (2007), 1672–1687.

[18] Yan Huang, Tom ZJ Fu, Dah-Ming Chiu, John CS Lui, and Cheng Huang. 2008. Challenges, design and analysis of a large-scale p2p-vod system. *ACM SIGCOMM computer communication review* 38, 4 (2008), 375–388.

[19] Adobe Systems Incorporated. 2010. *Adobe Flash Video File Format Specification*. https://download.macromedia.com/f4v/video_file_format_spec_v10_1.pdf

[20] Paras Jain, Sam Kumar, Sarah Wooders, Shishir G Patil, Joseph E Gonzalez, and Ion Stoica. 2023. Skyplane: Optimizing Transfer Cost and Throughput Using {Cloud-Aware} Overlays. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 1375–1389.

[21] Virajith Jalaparti, Ivan Bliznets, Srikanth Kandula, Brendan Lucier, and Ishai Menache. 2016. Dynamic pricing and traffic engineering for timely inter-datacenter transfers. In *Proceedings of the 2016 ACM SIGCOMM Conference*. 73–86.

[22] Virajith Jalaparti, Peter Bodik, Ishai Menache, Sriram Rao, Konstantin Makarychev, and Matthew Caesar. 2015. Network-aware scheduling for data-parallel jobs: Plan when you can. *ACM SIGCOMM Computer Communication Review* 45, 4 (2015), 407–420.

[23] Jinyang Li, Zhenyu Li, Ri Lu, Kai Xiao, Songlin Li, Jufeng Chen, Jingyu Yang, Chunli Zong, Aiyun Chen, Qinghua Wu, Chen Sun, Gareth Tyson, and Hongqiang Harry Liu. 2022. LiveNet: a low-latency video transport network for large-scale live streaming. In *Proceedings of the ACM SIGCOMM 2022 Conference* (Amsterdam, Netherlands) *(SIGCOMM '22)*. Association for Computing Machinery, New York, NY, USA, 812–825. https://doi.org/10.1145/3544216.3544236

[24] Wenxin Li, Keqiu Li, Deke Guo, Geyong Min, Heng Qi, and Jianhui Zhang. 2016. Cost-minimizing bandwidth guarantee for inter-datacenter traffic. *IEEE Transactions on Cloud Computing* 7, 2 (2016), 483–494.

[25] Yingling Mao, Xiaojun Shang, and Yuanyuan Yang. 2022. Joint resource management and flow scheduling for sfc deployment in hybrid edge-and-cloud network. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 170–179.

[26] Murtaza Motiwala, Amogh Dhamdhere, Nick Feamster, and Anukool Lakhina. 2012. Towards a cost model for network traffic. *ACM SIGCOMM Computer Communication Review* 42, 1 (2012), 54–60.

[27] Matthew K. Mukerjee, David Naylor, Junchen Jiang, Dongsu Han, Srinivasan Seshan, and Hui Zhang. 2015. Practical, Real-time Centralized Control for CDN-based Live Video Delivery. *SIGCOMM Comput. Commun. Rev.* 45, 4 (aug 2015), 311–324. https://doi.org/10.1145/2829988.2787475

[28] Netflix. 2021. *A cooperative approach to content delivery*. https://openconnect.netflix.com/Open-Connect-Briefing-Paper.pdf

[29] Onething. 2024. *Onething Services*. https://www.onething.net

[30] Roger Pantos and William May. 2017. HTTP Live Streaming. RFC 8216. https://doi.org/10.17487/RFC8216

[31] Qiniu. 2024. *Qiniu CDN Pricing*. https://www.qiniu.com/prices/qcdn

[32] Brandon Schlinker, Italo Cunha, Yi-Ching Chiu, Srikanth Sundaresan, and Ethan Katz-Bassett. 2019. Internet performance from facebook's edge. In *Proceedings of the Internet Measurement Conference*. 179–194.

[33] Rachee Singh, Sharad Agarwal, Matt Calder, and Paramvir Bahl. 2021. Cost-effective cloud edge traffic engineering with cascara. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. 201–216.

[34] Rade Stanojevic, Nikolaos Laoutaris, and Pablo Rodriguez. 2010. On economic heavy hitters: Shapley value analysis of 95th-percentile pricing. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 75–80.

[35] StellarCloud. 2024. *A introduction for StellarCloud*. https://www.xycloud.com

[36] Yu Sun, Chi Lin, Jiankang Ren, Pengfei Wang, Lei Wang, Guowei Wu, and Qiang Zhang. 2022. Subset selection for hybrid task scheduling with general cost constraints. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 790–799.

[37] Midhul Vuppalapati, Giannis Fikioris, Rachit Agarwal, Asaf Cidon, Anurag Khandelwal, and Eva Tardos. 2023. Karma: Resource Allocation for Dynamic Demands. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*.

[38] Dehui Wei, Jiao Zhang, Haozhe Li, Zhichen Xue, Yajie Peng, Xiaofei Pang, Rui Han, Yan Ma, and Jialin Li. 2024. Swarm: Cost-Efficient Video Content Distribution with a Peer-to-Peer System. arXiv:2401.15839 [cs.NI]

[39] Hao Yin, Xuening Liu, Tongyu Zhan, Vyas Sekar, Feng Qiu, Chuang Lin, Hui Zhang, and Bo Li. 2010. Livesky: Enhancing cdn with p2p. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 6, 3 (2010), 1–19.

[40] Menglu Yu, Ye Tian, Bo Ji, Chuan Wu, Hridesh Rajan, and Jia Liu. 2022. Gadget: Online resource optimization for scheduling ring-all-reduce learning jobs. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1569–1578.

[41] Ge Zhang, Wei Liu, Xiaojun Hei, and Wenqing Cheng. 2014. Unreeling Xunlei Kankan: Understanding hybrid CDN-P2P video-on-demand streaming. *IEEE Transactions on Multimedia* 17, 2 (2014), 229–242.

[42] Zheng Zhang, Ming Zhang, Albert G Greenberg, Y Charlie Hu, Ratul Mahajan, and Blaine Christian. 2010. Optimizing Cost and Performance in Online Service Provider Networks.. In *NSDI*. 33–48.

[43] Gongming Zhao, Jingzhou Wang, Hongli Xu, and Zhuolong Yu3 Chunming Qiao. 2023. COIN: Cost-Efficient Traffic Engineering with Various Pricing Schemes in Clouds. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE.

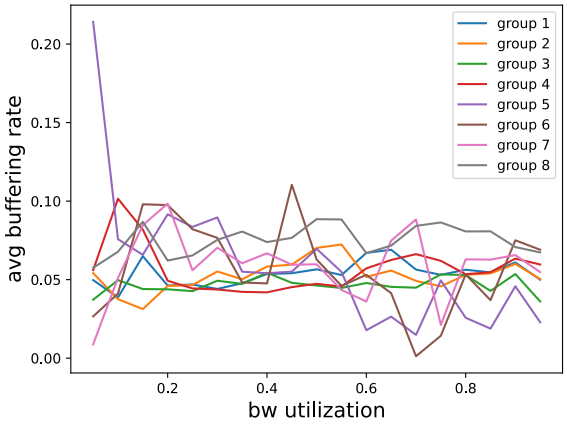


Figure 9: Generalization for the rela- tionship between bandwidth utilization and buffering rate of *traditional resources*.

A FORMULATION FOR BANDWIDTH PLANNING MODEL

A.1 Bandwidth cost formulation

We initiate by modeling the monthly bandwidth cost of each type of node. Formally, the bandwidth cost of node k is:

$$c_k = p_k \cdot u_k$$

, where where u_k represents the billing bandwidth during this billing cycle and p_k represents the unit price of the node.

We denote δ_{kt} as the bandwidth utilized by node k during time interval t and T as the total number of time intervals. For monthly percentile billing, the billing bandwidth of node k is:

$$u_k = 95^{th}\text{-}percentile(\delta_{k1}, \ldots, \delta_{kT}) \tag{1}$$

Furthermore, we add the following constraints to represent Eq. (1), enforcing that only 5% of the time slots δ_{kt} exceed u_k. Specifically, we have:

$$u_k \geq \delta_{kt} - M \cdot \lambda_{kt}, \quad \forall t$$
$$\sum_t \lambda_{kt} = \left\lfloor \frac{T}{20} \right\rfloor$$

. Here, M is a sufficiently large integer, and λ_{kt} is a binary variable that indicates whether the bandwidth usage in a given time slot exceeds the 95th percentile threshold.

For daily percentile billing, we assume there are N days in a month, with T_{day} intervals per day. We have:

$$u_k = \mathsf{avg}(z_1, z_2, \ldots, z_N)$$

Where z_i represents the 95th-percentile bandwidth for the ith day, which is calculated as:

$$z_i = 95^{th}\text{-}\mathsf{percentile}(\delta_{ki_1}, \ldots, \delta_{ki_{T_{day}}}) \tag{2}$$

Here, $i_j = (i-1) * T_{day} + j - 1$ serves as the index for the intervals within each day. To further illustrate the calculation of z_i in Eq. (2), we can introduce the following constraints:

$$z_i \geq \delta_{ki_j} - M \cdot \lambda_{ki_j}, \qquad \forall i \in \{1, \ldots, N\}, j \in \{1, \ldots, T_{day}\}$$
$$\sum_{i_1 \leq t \leq i_{T_{day}}} \lambda_{kt} = \left\lfloor \frac{T_{day}}{20} \right\rfloor, \quad \forall i \in \{1, \ldots, N\}$$

For fixed billing, the billing bandwidth equals to the predefined fixed capacity C_k.

A.2 Bandwidth cost optimization

The problem of bandwidth cost optimization considering both daily and monthly bandwidth-charged nodes (*alternative resources* and *traditional resources*) can be formulated as a mixed integer linear programming (MILP) problem:

$$\min \sum_k p_k \cdot u_k$$
$$s.t. \quad \sum_k \delta_{kt} \geq \sum d_t, \qquad \forall t \tag{3}$$
$$\delta_{kt} \leq C_k, \qquad \forall t, k \tag{4}$$

Inputs. On the supply side, each node k possesses a bandwidth capacity C_k and a unit price p_k. On the demand side, d_t represents the demand in the region at time interval t.

Outputs. The outputs include: (*i*) The billing bandwidth of each node, which is denoted as u_k. (*ii*) The arrangement of free intervals for each node, denoted as λ_{kt} (see Section A.1 for details).

Constraints. Constraint (3) constrains the total outbound bandwidth of all nodes, ensuring it exceeds the demand at any given time slot; Constraint (4) restricts the outbound bandwidth of each node to not surpass its capacity throughout the entire duration. Additionally, the constraints in Section A.1 should also be introduced.

B COMPARISON RESULT

Similar to Figure 7, we categorize *traditional resources* based on their advertised bandwidth capacities and randomly select nodes from each group. We then plot the variation in average buffering rates across different bandwidth utilization bins in Figure 9, with a bin size of 0.05 and group IDs increasing with bandwidth capacity, ranging from small to large. There is no significant elbow point in each line, and buffering rates show minimal fluctuation across various *traditional resources*, indicating that the actual available bandwidth capacities align closely with the advertised values.