Video Streaming Analysis in Multi-tier Edge-Cloud Networks

Eduardo S. Gama[†], Lucas Otávio N. de Araújo[†], Roger Immich[§], and Luiz F. Bittencourt[†]
[†]Institute of Computing - State University of Campinas (UNICAMP), Brazil
[§]Federal University of Rio Grande do Norte (UFRN), Brazil
eduardogama@lrc.ic.unicamp.br, l240106@dac.unicamp.br,
roger@imd.ufrn.br, and bit@ic.unicamp.br

***Abstract*—Video streaming services represent most internet traffic, and according to Cisco forecasts, in 2022, 82% of all internet traffic will be dominated by video streaming. This includes current video services as well as innovative services such as Real-Time video Streaming and future cloud gaming, whereas, for mobile devices, this estimate represents 78% of all mobile data traffic. A good cloud architecture partially solves some issues related to the live stream and Video on Demand (VoD) services to accommodate video traffic. However, a centralized cloud service introduces some issues such as higher latency and core network congestion. Therefore, to improve video services, it is paramount to distribute video streams according to their requirements properly: a real-time video streaming infrastructure is an interactive service that needs reduced delays (a few milliseconds). At the same time, a non-interactive VoD delivery can tolerate higher delays without impairing the quality of experience. This work discusses and gives evidence for the need for proper management and orchestration of video delivery over the Internet as it is core to the smooth coexistence of video services in multi-tier edge/cloud environments. The results assessment corroborate that well-defined video management can considerably increase the end-user QoE.**

***Index Terms*—Edge computing, Cloud computing, Quality of Experience, Video Streaming, Video-on-Demand, VoD**

I. Introduction

Over the years, Internet traffic has grown exponentially around the world, mainly due to multimedia content streaming, which currently represents 70% of the whole traffic [1], [2]. The Over-the-top (OTT) provider can use the network edge to cache and transmit the video traffic with an uninterrupted streaming experience, as smooth as possible, to accommodate this growing video demand. This includes standard video services as well as innovative services such as real-time video streams, and future gaming platforms using cloud infrastructures (e.g., AWS Wavelength and Google Stadia) [3]. This trend imposes new challenges in video provisioning to satisfy the Quality of Experience (QoE) guarantees for a wide range of subscribers [4]. As it was originally designed to consider the best-effort internet model for data transmission [5]–[7].

When the content provider saturates available bandwidth or a particular set of service metrics across the network, the cloud server (OTT provider) can reroute active connections to one or more edge caches to service the required demand. Such a model can be represented by a network organized hierarchically in multi-tiers to maintain a distributed, balanced traffic load [8].

Using the edge of the network with cache and replica capabilities helps improve user satisfaction, but it can also bring negative impacts. When many users start to request video segments at the edge nodes, these surrogated nodes consider static users and just one edge node tiers. These nodes are not ready to deal with the constant changes in the number of users and switching between different Access Points (AP) due to user mobility. Several works of literature highlight edge/cloud computing to deal with the new video traffic demands, but there are neglected aspects related to dynamic load solutions.

Figure 1 depicts a multi-tier network architecture, which is composed of a heterogeneous set of devices and applications using distributed computing resources through multi-access communication technology. Below the cloud layer, the network edge is divided into three layers organized hierarchically. Core Network Regional Edge can manage coordination across the distributed infrastructure in this multi-tier ecosystem, for example, in a smart city, followed by the Access Network Edge, which supports a few dozen to maybe a few hundred local nodes at the middle layer of the edge computing infrastructure. The Edge gateways can be distributed on local edge nodes, where the node re-transmits video content employing wired or wireless communication.

Motivated by the characteristics mentioned above in multi-tier edge/cloud scenarios to improve users' satisfaction and accommodate the growing video traffic, this paper discusses the need for an orchestrator to provide video streaming content in dynamic scenarios. Our goal is to provide designers and operators a performance analysis of a hierarchical edge network and its impacts on the users' QoE. In this paper, the models we present are evaluated by a series of simulations using a controlled environment, suggesting that dynamic resource allocation mechanisms are necessary to cope with video

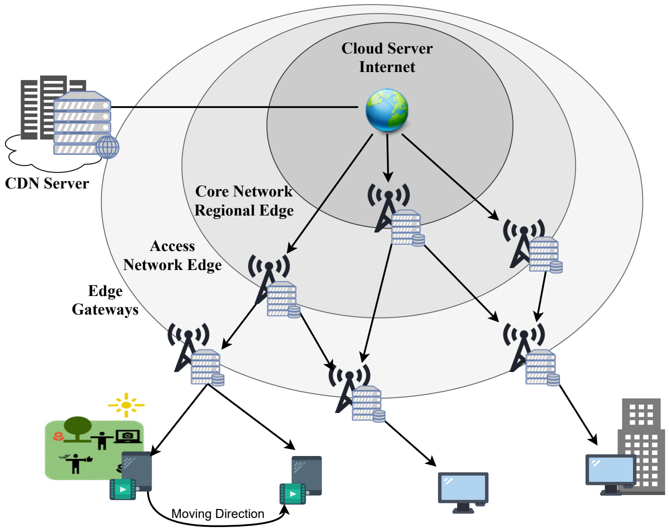

Fig. 1: A General Overview of the multi-tier network environment.

demands, such as when users are mobile.

This paper is organized as follows. Section II presents the related work on the impact of edge/cloud networks on video streaming services. A brief analysis of the proposed multi-tier edge/cloud network and a description of some opportunities of using such architectures are given in Section III. Furthermore, Section IV shows the preliminary results on the impact of the network performance for video streaming services and motivates further research on the topic, and Section V concludes the paper.

II. Motivation and Related Work

This section describes the related work in edge/cloud computing for video streaming. Here, some representative works in QoE are summarized regarding the edge network topologies and their impacts on video provisioning.

Guan *et al.* [9] demonstrate the performance of the two-tier edge caching network, presenting an algorithm with a 15% hit rate in multimedia content consuming 20% less memory. When a user requests a video, the address IP mapping redirects the request to the nearest cache. If the cache node does not hit the video, the edge node forwards it to the upper tier, which forwards the request to the upper-tier until it reaches the source. Otherwise, on any node storing the video in a cache, it will return immediately without forwarding the request to an upper tier.

Rosario *et al.* [8] present an architecture for virtual machine migration in real-time. During the migration, the video provisioning is moved forward over a multi-tier network. The architecture is based on the SDN paradigm for video distribution with QoE support. The work divides the edge into three tiers to ensure storage, upload, and download capacity. The cloud distributes the video content to the different edge levels with the multimedia service in the experimental scenario.

Shen *et al.* [10] work with a set of cache proxy services to analyze the cache miss occurrences. This work implements a reactive approach where cache proxies download chunks of multimedia content when requested. The cache services use probability theory to improve the efficiency of transferring corresponding video segments in the cloud. In this way, they demonstrated an improvement in users' QoE.

In Zhang *et al.* [11], the network consists of a cloud server connected to a pool of baseband units and a set of remote radio heads as cache nodes. The environment is organized hierarchically in multiple layers, where the proposed heuristic is formulated to consider the download rate between the nodes in the cache. The solution has two objectives, first to minimize the amount of backhaul traffic, and second, to improve the hit rate in VoD systems.

Bentaleb *et al.* [12] develop a Game Theory Algorithm, a new customer-oriented scheme within the client-side that seeks to select the best bitrate. Unlike most works in Multimedia Systems, in which users strive to maximize the QoE of the viewer without considering other entities on the network, this solution allows efficient collaboration between different players. The multimedia scheme improves the users' QoE with emphasis on the perceptual quality of the viewer, without explicit communication overload, respecting the decision requirements of the existing players. In addition, this work considers the cross-traffic in different network conditions.

The approaches above could decrease the traffic load and improve QoE. However, more issues arise in such dynamic scenarios: user mobility, collaborative cache services over multi-edge, users' number during flash crowds, and interactive streaming requirements are not fully considered. Proper management and orchestration of video delivery over the Internet is core to the smooth coexistence of heterogeneous video services. This work focuses on edge/cloud hierarchy to show the impacts of video streaming services in a multi-tier environment and discusses the need for real-time video streaming orchestration.

III. Analysis and Opportunities for a Video Streaming Architecture

Designing a cache hierarchy on vertically organized edge nodes with an arbitrary number of tiers can present improvements in users' QoE [13]. The architecture mentioned above works toward such advantages by serving the requested content as close as possible to the end-user, efficiently forwarding requests between parent edge nodes within the hierarchy, and balancing the video traffic considering hop counts and users attended. In addition, the network core congestion is reduced since it represents an operational overhead for the content provider. As a preliminary outcome achieved by a multi-tier network experiment, the QoE impact over a video streaming service is assessed. After that, we describe some results about the QoE characteristics and insights on the opportunities of caching multimedia content in edge nodes of multi-tier networks.

A. Impact of Fog Multi-tier Network Approach

To illustrate the differences in users' performance requesting a video from cache nodes in different tiers, consider two users requesting the same multimedia content from different layers. Then, an analysis of the impacts over the network is performed. Figure 2 illustrates a two-tier network model. The graphs show the results of bitrate, interruptions (stalls), video buffer, and representations switch, respectively, from left to right. Although a user requests a cache edge node video, it can be located in layers L1 or L2. Each layer level has different network factors, e.g., load, latency, so on. In Figure 1, an L1 layer can be interpreted as a personal computer, Access Point or Base Station, and this layer transmits the video content through wired or wireless communication channels, whereas in the L2 layer, a specific Edge gateway can be distributed on local edge nodes.

Apart from the initial interruption before the start of the video, both users did not experience the same issue again during the video execution. However, the user who received the video from the nearest edge layer had a higher bit rate than the user receiving the video from the upmost layer. Note that the user receiving the video from the closest layer has filled the buffer faster, and thus he/she managed to keep the video playing at the best possible resolution. On the other hand, the user who received the video from a more distant layer worked with the buffer at the limit and constantly switched resolutions so that there was no interruption during the video execution. Even without any interruptions in both video playbacks, the transmission from different tiers directly impacts the users' QoE.

B. Multi-tier Edge-Cloud Network Opportunities

Below are discussed the opportunities for resource management in multi-tier edge/cloud networks to provide video transmission. The advantages of nodes closer to end-users can improve the functioning of the network as a whole. We discuss some insights that can be used in favor of network and provider admin in the infrastructure for video transmission.

1) Improving the User's QoE: In a multi-tier environment composed of more than one-tier resource availability, the resources can host the video content near the end-users. This allows reducing latency and mitigating the load on the network core. The edge nodes are composed of specific resources combined to carry out the video transmission to integrate video streaming services in such communication environments. Within this context, integrating QoE feedback models inside the player raises an opportunity to improve the users' satisfaction. The results reported in Figure 2 suggest that it is possible to improve the users' satisfaction using the edge multi-tier network. Depending on the video service allocation level, it is possible to provide a smooth video playback even in an overloaded network.

2) Potential Bandwidth Saving: Videos streamed in higher quality increase the network bandwidth use. Consequently, provisioning from the Cloud will incur high communication expenses in the core network. The process of delivering part of the video along the network can significantly save bandwidth instead of sending the entire video to an edge server or by lowering the encoding quality of uninteresting portions of the video. Different delivery approaches can have different performances to reduce the uplink bandwidth use. Because of that, it is possible to contemplate an end-to-end design of video streaming, wherein the edge server adapts the video streams based on uplink and downlink bandwidth capacities. Additionally, new forms of video content are being generated today and may present opportunities for bandwidth saving and video services orchestration in edge-cloud infrastructures.

3) Cacheability: Nodes located at the edge are responsible for providing resources to VoD providers to allocate their caches. The allocated caches make multi-hops within the edge itself to serve the end-user. With this characteristic, different possibilities in the multi-tier network still have to be studied, such as cache allocation, placement, replacement, and selection caches, usually making decisions in real-time. These problems can offer a better video streaming service. In addition, an orchestrator has to consider the user's mobility and/or the possibility of predicting the direction of movement. In this way, as the user changes their trajectory, new caching mechanisms can use this information to optimize streaming video services.

IV. Performance Evaluation

This section describes the experimental evaluation of the proposed multi-tier video delivery architecture, including initial evaluation scenarios, metrics, methodology, and outcomes.

A. Experimental Setup

We use Adaptive Multimedia Streaming (AMuSt) [14] to implement a video streaming server, and the server is implemented through Dynamic Adaptive Streaming over HTTP (DASH), with users that allow adaptive video streaming. The AMuSt framework provides a set of applications for producing and consuming adaptable videos based on the DASH standard. DASH functionality is provided by the libdash library [15], an open source library that provides

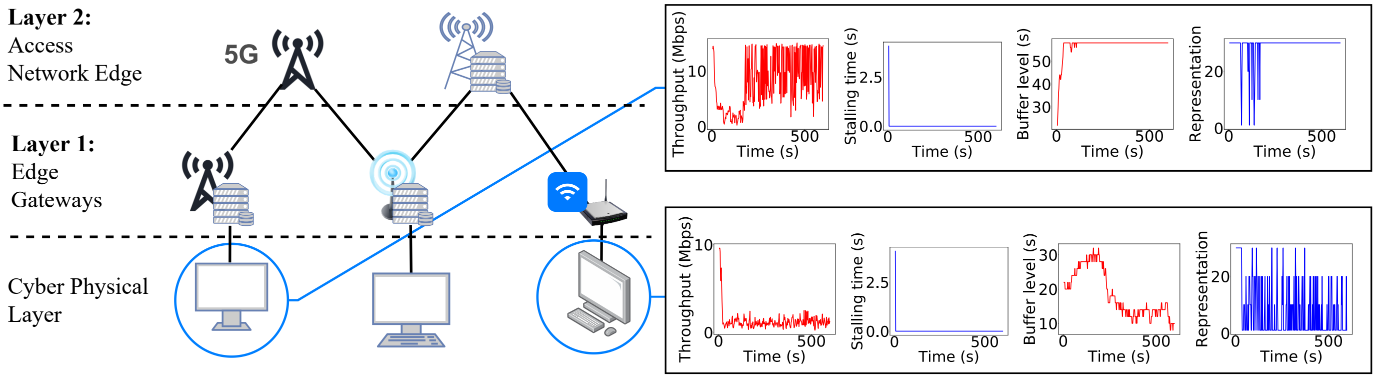

Fig. 2: Bitrate switches, stalls, buffer size, the startup delay in seconds of a DASH player was requesting a video with 10 bitrate levels varying from 50 to 4,500Kbps and from nodes in different tiers.

an interface to the DASH standard. Currently, libdash is the official reference software for the DASH standard. We consider that users are interested in an available video with ten different bit rate representations {235kbps, 375kbps, 560kbps, 560kbps, 750kbps, 1050kbps, 1750kbps, 2350kbps, 3000kbps, 4300kbps, 5800kbps}, which are used by Netflix subsets [16]. Each representation is divided into 2-second segments. All the experiments are executed once with a video of 1600 seconds (800 segments). For simplicity, the multimedia content used in the simulation is deployed beforehand in the edge peering nodes.

We simulate the scenario in a binary tree topology with seven nodes and a Cloud Provider connected to the root node. The last four nodes are Access Points (AP), and the others are edge peering points. Figure 3 illustrates this scenario.

The AP nodes are implemented on wireless devices that communicate via IEEE 802.11g in 2.4GHz. The APs have wired connections to the edge peering points, while the end-users are wireless. Each user connected to the AP is located precisely 8 meters away from the AP. The Bandwidth available is 20Mbs on links 0-1 and 0-2; and 30Mbs on links 1-3, 1-4, 2-5, and 2-6.

We present three approaches to address the impacts identified in Section III into the edge/cloud multi-tier network, namely *cloud-only*, *edge cache*, and mobile-based scenarios. The cloud-only scenario uses only the Cloud Provider node to deliver the video content. On the other hand, the *edge cache* approach uses nodes 1 and 2 as auxiliary nodes to deliver the video. The simulation starts with the users requesting the video from the cloud. When a congested link is detected, the edge cache below the link is turned on. Thereafter, the users receiving the video over the congested link are redirected to the edge nodes 1 and 2.

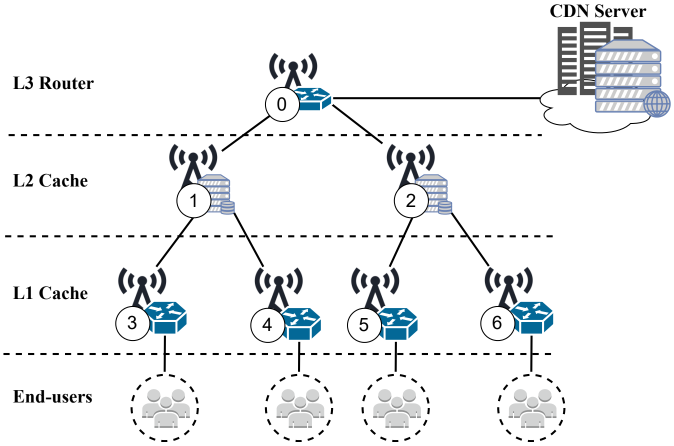

Fig. 3: Overview of the multi-tier network environment.

The number of end-user devices communicating through each wireless AP may change over time due to the mobility of the end-user devices. In practice, the number of end-user devices connecting to the wireless AP changes frequently. To show the problems that can arise in such dynamic load scenarios, a connection change between the APs occurs. We then reran the second experiment, maintaining connections between users and edge nodes 1 and 2, but changing the connection between users and APs, emulating users' mobility. In this mobility scenario, half the users from each AP move. Let i the AP index ($3 \leq i \leq 6$) such that $i < 5$, users from AP_i go directly to AP_{i+2}, otherwise they go to AP_{i-2}.

B. QoE Assessment

There are many QoE models in the literature. We describe how QoE metrics can be used to score user satisfaction. Firstly, each video quality chunk is computed by a logarithmic law over bitrates [17], where a video quality model for DASH is proposed as shown in Equation 1. Each video has N segments and is encoded with L bitrate levels. r_i represents a specific bitrate level. At each step i, the quality of segment i is defined as:

$$q(r_i) = a_1 * log(a_2 * (r_i / r_{|L|})) \quad (1)$$

It is required a flexible QoE model that includes the most influential metrics to quantify long-term users' QoE. We consider Equation 2 from [12], which consists of four metrics: (a) the average chunk perceptual quality, (b) the average number of quality oscillations, (c) the average number of stall events and their duration, and (d) the startup delay. K represents the total segments of the video, S_i is the stall duration, and ST_i is the startup delay of user i.

$$QoE_i = \frac{1}{K}\sum_{k=1}^{K} q(r_k) - \frac{1}{K-1}\sum_{k=1}^{K-1} |q(r_{k+1}) - q(r_k)| - \frac{1}{K}\sum_{k=1}^{K} S_k - ST_i \quad (2)$$

The QoE_i for each user i can range from 1 to 5, where 1 = bad, 2 = poor, 3 = fair, 4 = good, and 5 = excellent.

C. Results and Discussion

The experiments illustrated in Figures 4, 5, and 6 show the average QoE calculated by Equation 2 when there are 15, 20, and 25 users per access point requesting videos in the simulated infrastructure. Each boxplot represents the Cloud-only, Edge cache and Mobility scenarios. The overall average performance of the Edge cache is better than the Cloud-only and Mobility scenarios, mainly due to the choices of nodes at the edge peering nodes for serving the requests from users. For instance, in the Edge cache experiment, when congestion occurs in intermediate links $e_{0,1}$ e $e_{0,2}$, the edge peering nodes are activated to serve the end-users below those nodes. In this way, the traffic passing through the upward link will now be smoothed out, so the users can have their QoE improved to an excellent level. Since the users request segments from the closer nodes and with no congestion link in scenarios with Edge cache, as expected, we observe a QoE increase. The performance difference from the Cloud-only and mobility scenarios to the Edge cache experiment is near one level of satisfaction for 15 and 20 users and approximately two satisfaction levels for 25 users.

It is important to note that the QoE results of scenarios after mobility are similar to the results of Cloud-only, even with the edge peering nodes using the QoE. This is due to the lack of a rerouting mechanism in real-time when users switch to another AP. The connections between the edge-peering nodes and the users remain unchanged so that the paths through which the packets pass are longer, negatively impacting the performance of the network as a whole.

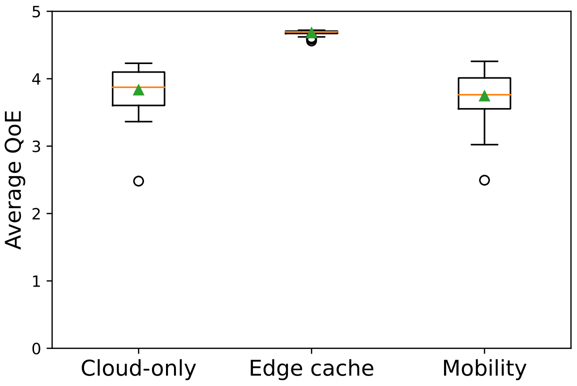

Fig. 4: Average QoE results (15 users per AP).

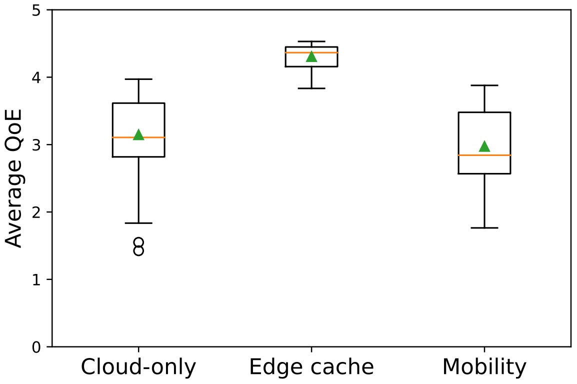

Fig. 5: Average QoE results (20 users per AP).

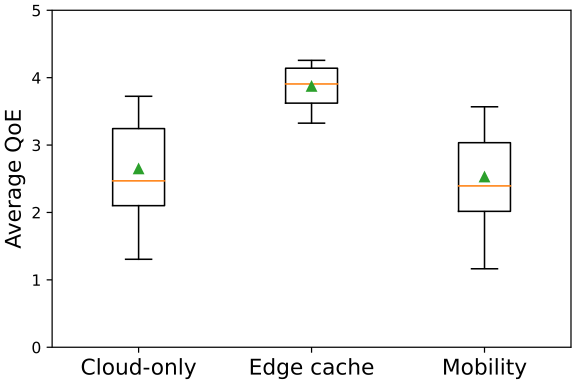

Fig. 6: Average QoE results (25 users per AP).

Figures 7, 8, and 9 show the final QoE of each user per start time (time to start the segment requests). Each blue tick represents a user, and the y-axis QoE means the final user satisfaction for a fully watched video. Here, we can see the final QoE degradation for each user's entrance during the simulation execution. The red plot represents the number of active users simultaneously. As users reach the final QoE in this experiment, it tends to be slightly lower than the previous one. When looking at the final QoE delta between users u_i and u_{i+1}, this seems to be irrelevant, but as we increase the delta, the QoE starts to become considerably different. We can also confirm this behavior by showing that as the number of users increases, the average QoE decreases, and the variation between users increases. Also, note that in the simulation with 25 users per AP, the system already shows a degradation in the quality, negatively perceived by the end-user.

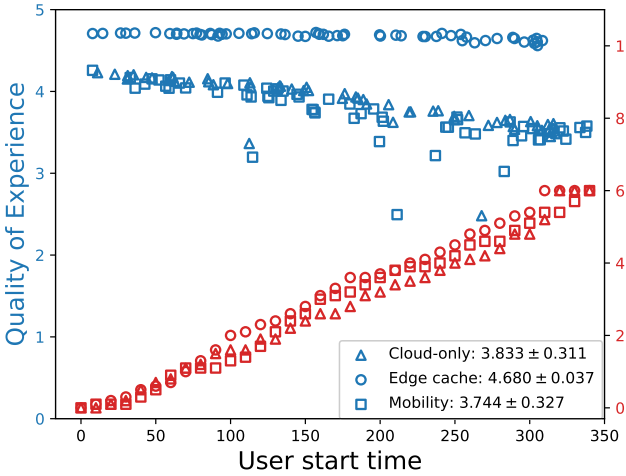

Fig. 7: Average QoE for each user (15 users per AP).

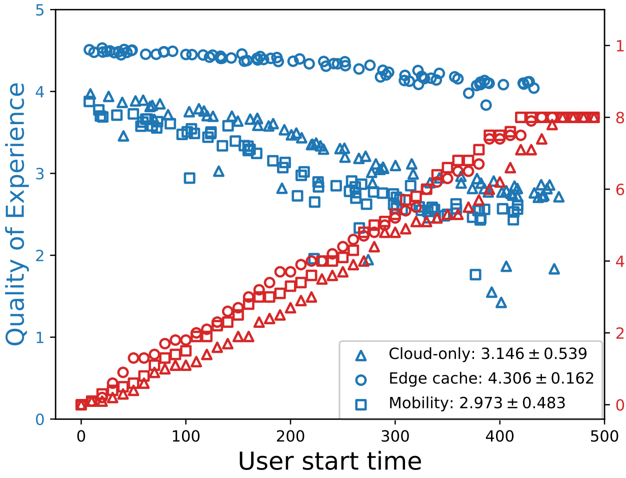

Fig. 8: Average QoE for each user (20 users per AP).

An interesting discussion takes place when the QoE per user is analyzed separately. According to the numerical results, the final QoE tends to worsen as the number of active users increases. However, it is not entirely true for the Edge cache scenario, where the final QoE for each user remains close. The standard deviation of 0.037, 0.162, and 0.273, respectively, for scenarios with 15, 20, and 25 users per AP, indicates a closer QoE between the users. Only in the scenario for 25 users per AP, the average QoE showed a drop with a satisfaction close to regular. In contrast to the other two scenarios presenting a users' satisfaction from good to excellent. Whereas for cloud-only and mobility scenarios, the network operates with a high standard deviation. As the number of users increases, some outliers start to appear with the worst level of satisfaction, being between poor and inadequate.

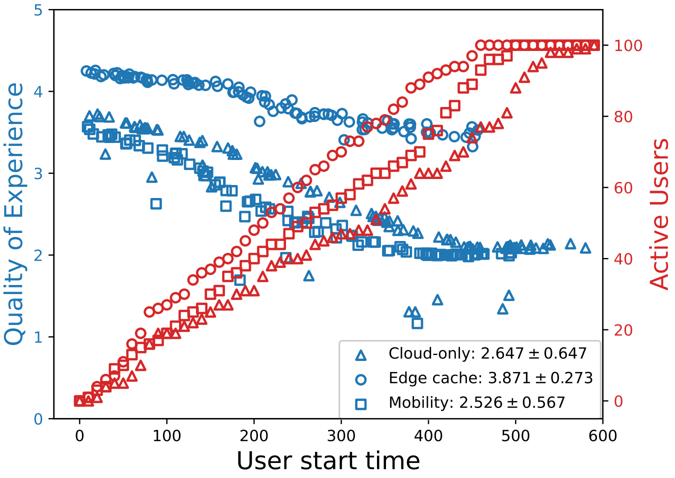

Fig. 9: Average QoE for each user (25 users per AP).

Based on these observations, a simple strategy of moving the video to the edge can significantly improve the user's QoE. In this way, the video transmission system can provide user satisfaction qualities and keep them watching the video up to the end. However, if there is no correct management of connections in real-time and a dynamic mechanism to tackle with a varying load coming, for example, from the mobility of users, we can conclude that the impact introduced by the AP changes can significantly decrease their QoE. The user experience can end up getting worse even using the edge of the network. Proper management of multimedia content can be done in which the VoD focuses on providing a better QoE for the users' connection changes. A content migration mechanism at the edge and upper tiers can mitigate the problem. Also, performing a rerouting between the active users' connections and the server nodes may help in improving and balance user's QoE.

V. Conclusion

This paper investigates the characteristics of multi-tier edge/cloud scenarios with a VoD service. Numerical results for a binary tree network suggest that the correct video management can substantially improve the users' QoE. However, introducing simple connection changes between the APs and users, with the lack of an adequate orchestration of the connections, can negatively impact user satisfaction.

As future work, we intend to implement mechanisms capable of orchestrating the users' connections in real-time and improving video streaming in multi-tier edge/cloud environments. Another improvement is to assess how service behaves to provide a delicate balance between cost and customer satisfaction in terms of QoE.

Acknowledgment

This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, FAPESP procs. 14/50937-1, 15/24485-9, and 2020/16763-7.

References

[1] Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022, 2017. ; accessed 30-may-2021.

[2] Roger Immich, Leandro Villas, Luiz Bittencourt, and Edmundo Madeira. Multi-tier edge-to-cloud architecture for adaptive video delivery. In *2019 7th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 23–30, 2019.

[3] Aws wavelength forneça aplicações de latência ultrabaixa para dispositivos 5g, 2021. ; accessed 30-may-2021.

[4] Carlos Quadros, Eduardo Cerqueira, Augusto Neto, Andre Riker, Roger Immich, and Marilia Curado. A mobile qoe architecture for heterogeneous multimedia wireless networks. In *2012 IEEE Globecom Workshops*, pages 1057–1061, 2012.

[5] Huan Wang, Guoming Tang, Kui Wu, and Jianping Wang. PLVER: joint stable allocation and content replication for edge-assisted live video delivery. *CoRR*, abs/2006.07505, 2020.

[6] Z. Ye, F. D. Pellegrini, R. El-Azouzi, L. Maggi, and T. Jimenez. Quality-aware dash video caching schemes at mobile edge. In *2017 29th International Teletraffic Congress (ITC 29)*, volume 1, pages 205–213, Sept 2017.

[7] E. S. Gama, R. Immich, and L. F. Bittencourt. Towards a multi-tier fog/cloud architecture for video streaming. In *2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion)*, pages 13–14, Dec 2018.

[8] Denis Rosário, Matias Schimuneck, João Camargo, Jéferson Nobre, Cristiano Both, Juergen Rochol, and Mario Gerla. Service migration from cloud to multi-tier fog nodes for multimedia dissemination with qoe support. *Sensors*, 18(2), 2018.

[9] Yu Guan, Xinggong Zhang, and Zongming Guo. Caca: Learning-based content-aware cache admission for video content in edge caching. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 456–464, New York, NY, USA, 2019. ACM.

[10] Gengbiao Shen, Qing Li, Yong Jiang, Richard Sinnott, Dong Lin, Zehua Guo, and Yi Wang. Chunk-level request-grant-transfer mode for qoe-sensitive video delivery in cdn. In *Proceedings of the International Symposium on Quality of Service*, IWQoS '19, pages 12:1–12:10, New York, NY, USA, 2019. ACM.

[11] Zhilong Zhang, Danpu Liu, and Yaxiong Yuan. Layered hierarchical caching for svc-based http adaptive streaming over c-ran. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2017.

[12] Abdelhak Bentaleb, Ali C. Begen, Saad Harous, and Roger Zimmermann. Want to play dash?: A game theoretic approach for adaptive streaming over http. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, pages 13–26, New York, NY, USA, 2018. ACM.

[13] Omer Rana, Manjerhussain Shaikh, Muhammad Ali, Ashiq Anjum, and Luiz Bittencourt. Vertical workflows: Service orchestration across cloud & edge resources. In *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 355–362. IEEE, 2018.

[14] Christian Kreuzberger, Daniel Posch, and Hermann Hellwagner. Amust framework - adaptive multimedia streaming simulation framework for ns-3 and ndnsim, 2016.

[15] C. Mueller, S. Lederer, J. Poecher, and C. Timmerer. Demo paper: Libdash - an open source software library for the mpeg-dash standard. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–2, July 2013.

[16] Per-title encode optimization, 2015. ; accessed 20-novembro-2019.

[17] Weiwen Zhang, Yonggang Wen, Zhenzhong Chen, and Ashish Khisti. Qoe-driven cache management for http adaptive bit rate streaming over wireless networks. *IEEE Transactions on Multimedia*, 15(6):1431–1445, 2013.