Multi-Tier Edge-to-Cloud Architecture for Adaptive Video Delivery

Roger Immich, Leandro Villas, Luiz Bittencourt, Edmundo Madeira
Institute of Computing
University of Campinas
Campinas, Brazil
{roger, leandro, bit, edmundo}@ic.unicamp.br

***Abstract*—In the last few years, there has been a rapid proliferation of a wide range of real-time video services and applications. These technologies flood the wireless systems with video content on a daily basis. As a result of this sharp increase in video traffic, the prospect of errors due to network interference and congestion rises. Incidentally, the adoption of the 5th generation of wireless systems (5G) will allow this growth to be even greater due to its high bandwidth capacity and low latency. However, even with these improvements on the wireless capabilities, a reliable and high-quality video transmission still imposes several challenges, such as how to handle a large number of heterogeneous devices and how to better use the resource-richer Edge, Fog, and Cloud computing sources to meet the user's requirements. To overcome these issues, this work proposes a multi-tier video delivery architecture relying upon several technologies such as Multi-access Edge computing (MEC), 5G slices, and microservice placement/chaining. Furthermore, to assess the proposed idea an experimental proof-of-concept testbed of the multi-tier architecture was designed, implemented, and evaluated using real-world tools and actual video sequences. The results obtained supported our claim that a multi-tier video delivery system is feasible and can greatly benefit the end-users.**

***Keywords*-Multi-tier Video Delivery, Edge computing, Microservice chaining**

I. Introduction

Owing to the ever-growing number of wireless-capable devices and the broader amount of communication technologies that can be adopted, optimization techniques for video delivery are required [1]–[3]. These mechanisms will have to accommodate heterogeneous networks and devices allowing them to coexist while providing satisfactory levels of Quality of Experience (QoE). The integration of these networks with the Cloud partially solves some of the scalability, availability, and interoperability issues, but at the same time, introduces new ones (e.g., higher latency) [4], [5]. To solve the new issues while increasing the resilience, Fog and Edge computing can be used along with 5G slices.

In addition to that, according to Cisco, the global Internet traffic will be 127 times higher by 2021 in comparison to 2005. Furthermore, the video-related IP services are expected to represent over 82% of all the Internet traffic [6]. To make the matter worse, it is also foreseen a massive increase in the mobile data traffic as new applications start to get momentum such as immersive videos or 360-degree, virtual and augmented reality, as well as ultra-high definition video. This tendency is especially relevant in environments with a high user's concentration such as in Smart Cities. These applications require ultra-low latency and sustained bandwidth to guarantee an acceptable performance, these requirements cannot be fulfilled by Cloud computing alone.

There are several works that try to solve a number of issues regarding the combined use of Cloud, Fog, and Edge computing. Generally, the existing works are based on architecture design and deployment issues [7]–[10]. They try to improve virtual machine migration, mobility adversities or provide smart caching for a specific type of content as well as to enhance video delivery with adaptive streaming and Fog nodes. However, the solutions presented tend to be general and do not take into consideration video-specific needs.

Taking into account the aforementioned scenarios and new ones yet to be revealed in the next few years there is a growing need for a multi-tier video delivery architecture. It could take advantage, at the same time, of the elastic resource pool that Cloud computing provides in association with the low-latency and high-throughput offered by the Edge computing. This paper aims to advance the idea of multi-tier video delivery using off-the-shelf open-source tools and real video sequences. The main goal is to prove that it is possible to build a real multi-tier environment to improve video delivery quality.

The remainder of this paper is organized as follows. The related work is given in Section II. Section III describes the proposed multi-tier video delivery architecture. Following this, Section IV shows a proof-of-concept with real open source tools, where actual video sequences assessments are presented supporting the idea of the proposed architecture. In the end, conclusions and future work are summarized in Section V.

II. Related Work

There are several proposals in the last few years to improve the video delivery over a variety of network architectures.

The FogRoute [11] presented a hybrid data dissemination in Fog computing. The authors applied Delay-Tolerant Network (DTN) and Software-Defined Network (SDN) to exploit the offload capabilities of these networks. In doing that, several distributed Fog servers are employed to disseminate data using DTN techniques and let the Cloud as a control plane to organize the flows and process the content update. This solution helps to alleviate the network traffic by precaching content on the Fog nodes and then distributing to the end-users. This proposal provided a good solution for data-dissemination and other contents that are latency-agnostic, however, it does not perform well with low latency content such as live video or video-on-demand.

The ICN-Fog architecture [12] proposes a horizontal Fog-to-Fog abstraction layer supported by Information-Centric Networking (ICN). The proposed architecture enables horizontal data transfer between nodes allocated in the Fog layer. It also grants the capability of distributing the processing tasks among these nodes. Because of that, the ICN-Fog provides loose dependence between Fog and Cloud applications, it improves the usage of heterogeneous devices (even constrained devices are able to participate in the network) and enhances mobility support through connectionless name-based data communication. One of the disadvantages of this architecture is the high cost of orchestrating all the communication and dataflows that span through multiple networks domains between the Fog and Cloud nodes.

Another work proposes several strategies for preemptive migration in ICN caches and Edge nodes of mobile networks [13]. The main idea is to keep the data, that users are interested in, always close to them. In doing that, the authors expected to improve delivery performance while supporting a higher number of users. In their proposal, the user mobility is predicted by a controller which decides whether to migrate contents and which is the best Edge node to receive this migration. The video content is categorized and migrated based on popularity. The assessment results indicated that the proposed work is able to both increase the cache hit rate at Edge nodes and reduce the download latency. Using this technique, the author provided a global cache optimization that benefits a group of users that consume popular videos, on the other hand, they penalize users that are not interested in the available popular content. To yield a holistic optimization other factors related to the users' profile should be also considered such as the users' personal preferences.

III. Towards the design of a Multi-tier Video Delivery architecture

The convergence of Cloud, Fog, and Edge computing requires a proactive resource orchestration scheme [14]. It is worth mentioning that one of the major challenges in Edge-centric computing is how to achieve the best tradeoff between node computing, communication, and Edge processing. Since heterogeneous devices are expected to exist in the network, several details must be taken into consideration, such as the limited battery of mobile devices, different communication capabilities as well as processing power [15]. An optimal solution should also deal with the user's expectation of QoE and its impact on the system as a whole.

Besides offloading the Cloud environment and offering low latency communications, the Fog/Edge computing also provides location awareness since data can be collected and processed based on geographic location and Radio Access Network (RAN) details, such as link quality and cell load. Because of that, it is important to have the support of multi-tier microservices placement mechanisms. This allows placing microservices in the best locations of the Edge computing system [16]. This further improves latency and also provides both high availability and resilience. Intelligent microservice placement mechanisms should constantly assess the network conditions through the RAN details, the user status, and profile as well as the popularity of the content to steer the placement decision. In addition, the mechanisms should be able to seamless migrate the microservices, to and from any tier of the Edge computing, whenever necessary.

Furthermore, the provision of a high QoE is important from the point-of-view of the end-users [17]. One way to do that is through the creation, configuration, and optimization of 5G slices, which are also expected to be one of the key resources in 5G networks [18]. They will provide a holistic end-to-end virtual network for a given user, so-called tenants. This means that the physical mobile network will have its resources partitioned and customized according to the system needs [19]. The combination of two technologies is envisioned to offer support to the network slices, namely Network Functions Virtualization (NFV) and SDN [20]. The former allows instantiating on-demand specific network functions and services, which enables load balance, failure recovery, and better hardware utilization. The latter decouples the control and data planes, making it possible to abstract the underlying network infrastructure. The outcome is a more flexible, reliable, scalable, and secure network. Using these technologies, in many situations, the networks will be able to reconfigure slices within seconds to quickly respond to local demands, such as an unexpected gathering of people or to prioritize emergency systems. On the other hand, it is also possible to program a long-term lease, for example, to an electrical utility company to accommodate its smart grid components such as meters, sensors, controllers, and other IoT devices. A short-term lease is also feasible, for example, when a public venue or a concert promoter wants to have a dedicated slice for a weekend-long festival and optimize it for streaming high-quality video and music data.

The main goal of this work is to design, implement, and

assess a reliable and high-quality multi-tier video delivery architecture to be used in Smart City environments. The proposed scheme will take advantage of several network-related technologies such as Cloud, Fog, and Edge computing, as well as intelligent microservice placement and chaining. Fig. 1 depicts a multi-tier network architecture. The first layer represents a data center with hundreds of servers and the ability to process a high volume of requests through the use of elastic resource virtualization. The second layer aims to bring computational capacity near the users and devices at the edge of the network. Following the same idea as the Cloud tier, this layer can also provide the elasticity of resources required, however in a more restricted way than the previous layer. The Fog layer can be composed of several tiers, ranging from the access networks to the regional data centers. The Cyber-Physical layer is composed of a heterogeneous set of devices and applications using the resources. Based on these premises, a multitude of parameters should be assessed to define which microservices need to be deployed as well as the most suitable tier to deploy each one of them. In addition, network slices also should be defined, created, configured, or re-arranged based on these parameters. The assessed parameters include, but are not limited to, the user's profile, the load of the local cell, the link quality, the type of the video application (eHealth, entertainment, augmented or virtual reality, 360° videos), the motion complexity of the videos, and also smart city details such as the location and the traced route in case of users with mobility.

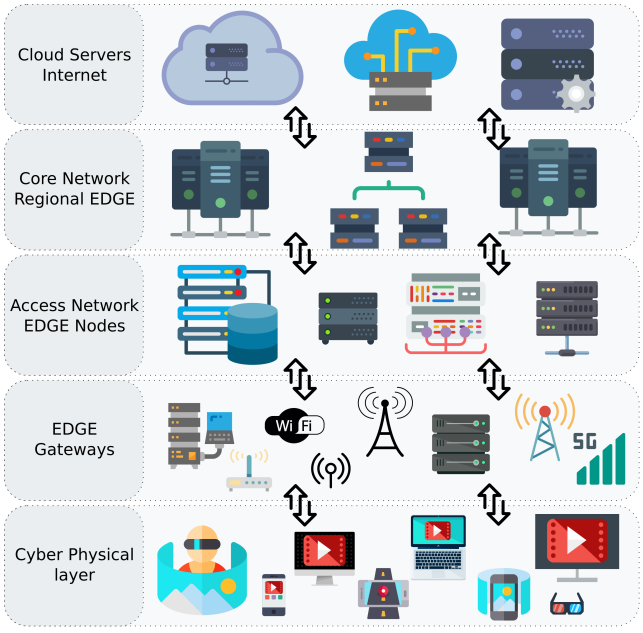

Fig. 1. Multi-tier architecture

The Smart City scenario is complex because it relies on a large number of heterogeneous communication technologies, such as radio-frequency, cellular, wireless, wired, and hybrid. This means that the communication protocols have to handle distinct underlying networks and different addressing schemas. There is also the need to take into consideration the node's mobility. Some of them can be stationary, but others can range from low to high mobility patterns. In order to take into consideration all these details, the proposed methods will have to fetch information from the routing protocols as well as the cell and subscriber radio interface information, such as location, cell load, and link quality. In order to advance in this field, the proposed work presents a proof-of-concept testbed of a multi-tier video delivery architecture with microservice chaining and network slices. The testbed was implemented and assessed with real of-the-shelf tools and actual video sequences.

IV. Testbed Assessment and Results

The section describes the proposed testbed architecture. The experimental multi-tier environment is composed of four nodes, which are one client and three servers. Each server represents a different layer, for example, the 3rd-tier is the Cloud server node, the 2nd-tier is regional Fog node, and the 1st-tier is the Edge node. Because of that, three setups were created, i.e., with the cache microservice at each one of the tiers chained to the others that work as a proxy microservice. The cache microservice works as a simple temporary data storage so that users can access the information faster, it also helps to reduce the core network bandwidth demand. The proxy microservice just forward the information to the next node. Each one of these setups was assessed with and without the aid of network slices.

A. Experiment settings

All the experiments were performed using real open-source tools and an actual Big Buck Bunny video stream. A set of tools were used to create a multi-tier Edge computing environment as follows. The first tool used is Docker. It allows to create, configure, and run containers, which is a lightweight way to provide isolation, automation, compatibility, and integration of services. The second tool is the Kubernetes. This is a container-orchestration platform to automate several container-related activities, such as automatic deployment, management, scaling, and operations between different clusters and hosts. Another tool is JUJU. It is a modeling tool to perform quick deploys of container-based applications in several private and public Cloud services. It also helps to configure, interact, and easily perform several other operational tasks.

The containers were loaded with either an NFV cache or proxy microservice. The tool used for the microservices was NGINX. This tool is known for its stability, low resource consumption, large feature set, and high performance. In each scenario, there was one cache server and two proxy

nodes. The microservices were manually chained through the configuration files. The network slices were configured as Virtual Private Network (VPN), with resource reservation and isolation. Additionally, the control plane and the data plane were also isolated. In the future, SDN and NFV functions will be used to make these configurations dynamically allocable and adjustable.

Furthermore, the MPEG-DASH Adaptive Bitrate Streaming (ABR) technique was adopted. This is an HTTP-based streaming largely used nowadays. The client node was equipped with a Google Chrome browser and the DASH Industry Forum Player. The only modification to the player was the addition of a function to log the assessed metrics. The player was configured to not "Schedule While Paused", to not "Allow Local Storage", and to not use the "Low Latency Mode". On the other hand, it was set to "Jump Small Gaps", to allow "Fast Switching ABR", and to use the "Strategy: Dynamic ABR", which means that it dynamically switch between the "BOLA ABR" and the "Throughput" strategy.

The Big Buck Bunny video was encoded with MPEG-DASH and resulted in the Media Presentation Description (MPD) parameters described below. This document is an eXtensible Markup Language (XML) file which holds important metadata about the several different media segments and how they relate to each other. This allows the video player client to choose between them to provide the most appropriate configuration to the end-users. The minimum buffer time was set to two (3) seconds. There Digital Rights Management (DRM) was not set. It uses the MIME type "video/mp4" ("m4v" file type) with 30 frames per second (fps), "16:9" aspect ratio, and progressive scan. Additionally, the video was encoded with ten (10) distinct resolutions, ranging from "320x180" up to "3840x2160", each one with a specific bitrate as shown in Table I. The audio was encoded with MPEG4 audio file "m4a" (MIME type "audio/mp4") using the Advanced Audio Coding (AAC) codec with lossy compression and a single sampling rate of "48000". A constant User Datagram Protocol (UDP) background traffic was generated with iperf version 2.0.10 during the experiments. The assessment was performed in the first 60 seconds of the video stream. Table II summarizes the testbed parameters.

TABLE I
VIDEO RESOLUTION AND BITRATE

RESOLUTION	BITRATE
320x180 pixels	200 Kbps
320x180 pixels	400 Kbps
480x270 pixels	600 Kbps
640x360 pixels	800 Kbps
640x360 pixels	1000 Kbps
768x432 pixels	1500 Kbps
1024x576 pixels	2500 Kbps
1280x720 pixels	4000 Kbps
1920x1080 pixels	8000 Kbps
3840x2160 pixels	12000 Kbps

TABLE II
TESTBED PARAMETERS

PARAMETERS	VALUE
Display sizes	320x180 up to 3840x2160
Frame rate	30 fps
Aspect ratio	16:9
Video mimeType	video/mp4
MPEG4 video file	m4v
Audio mimeType	audio/mp4
audioSamplingRate	48 kHz
MPEG4 audio file	m4a
Dash Player	Reference Client 2.9.0
Dash Schedule While Paused	Not selected
Dash Allow Local Storage	Not selected
Dash Low Latency Mode	Not selected
Dash Jump Small Gap	Selected
Dash Fast Switching ABR	Selected
Dash Fast Switching Strategy	Dynamic ABR
Segment Size	$\approx$ 2 seconds
scanType	progressive
minBufferTime	PT3.00S
JUJU version	2.4.3
Kubernetes version	1.11/stable
Docker version	18.06.1-ce
NGINX version	1.15.4
Google Chrome version	68.0.3440.106 (64-bit)
Client to 3rd-tier (Cloud) link delay	200ms
Client to 2nd-tier (Fog) link delay	70ms
Client to 1st-tier (Edge) link delay	22ms

B. Bitrate assessment

Fig. 2 shows the results of the bitrate assessment of scenario without slices and Fig. 3 of the scenario with slices. This metric depicts the visual quality of the video stream as perceived by the end-users. There is no standard deviation in this result because the experiments were conducted in a controlled environment and even with background traffic, there was no significant variation of this metric. There was a negligible variation due to the ABR operation, where the perceived quality remained the same. In the scenario with no network slices (Fig. 2), the video resolution with the cache microservice at the Cloud level stayed at "480x270 600 Kbps". This was expected because the video streaming had to go through several layers before being delivered to the end-users as well as compete with the background traffic for resource allocation. Owing to the ABR protective procedures there was no stalls and re-buffering during the experiments. Additionally,

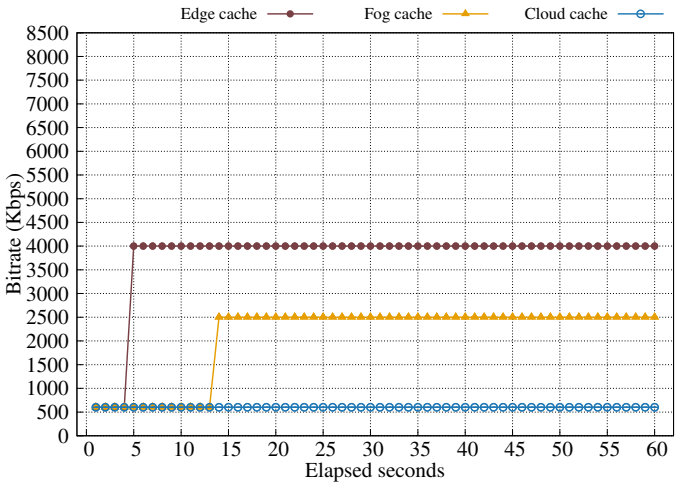

Fig. 2. Delivered bitrate without network slices

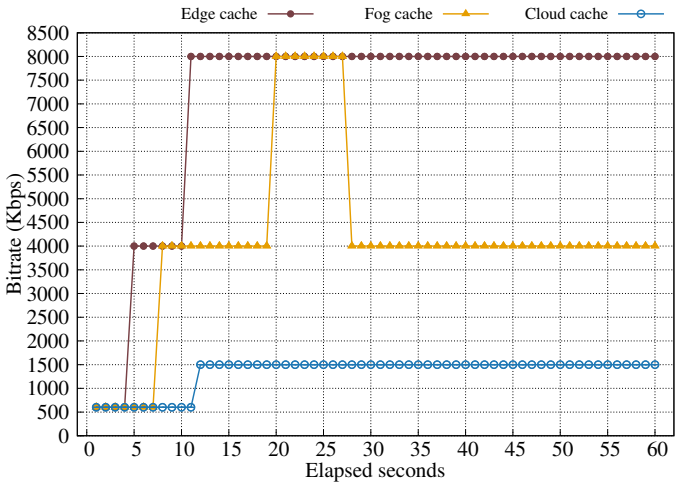

Fig. 3. Delivered bitrate with network slices

when the cache microservice was hosted at the Fog level there was an increase in the video resolution up to "1024x576 2500 Kbps". After several seconds with a lower resolution, the ABR protocol was able to change the video quality according to improved bandwidth delivered by a cache closer to the end-user. The best results of this metric, in the scenario without network slices, were obtained with the cache microservice at the Edge layer. In this experiment, the video resolution reached the resolution "1280x720 4000 Kbps" after a very short period of adaptation. This result was expected due to the proximity of the cache microservice to the end-user. It is important to notice that after the negotiation phase the MPEG-DASH is able to maintain the same video quality during the remaining of the stream time because there is little fluctuation in the network. As mentioned before, the experiments were conducted using several real tools and services in a controlled environment. This means that the network tends to keep the same features.

In the scenario with network slices (Fig. 3), the delivery bitrate was improved, as expected. At first, with the cache microservice at the Cloud level, the video resolution started at "480x270 600 Kbps" up to 10 seconds, and then increase the bitrate to 1500 Kbps with the resolution of "768x432" pixels. This increase in the resolution was perceived in the scenario without network slices and it reaches more than twice of the previous bitrate, which is translated as a higher video quality for the end-users. Moreover, when the cache service was placed at the Fog layer, the video resolution reached "1280x720" with 4000 Kbps of bitrate after a 7 seconds adaptation period. This result displays an increase of 60% in the bitrate. After 19 seconds of video reproduction, the ABR protocol increased again the video resolution to "1920x1080", however, it was only able to keep this resolution for 7 seconds, returning to the previous one. This can be explained because there is a considerable gap, in the bitrate size, between the "1280x720" and "1920x1080" resolutions. In this case, the bitrate doubles from one to another. This experiment can indicate that if there was an intermediate bitrate between these two resolutions, the ABR protocol would probably be able to keep the intermediate resolution, providing better video quality. Finally, when the cache service was hosted at the Edge level and with the help of network slices, the video resolution climbed to "1920x1080" with 8000 Kbps of bitrate after a short warm-up period. On the basis of the results referred to above, it is possible to conclude that the adoption of network slices, in this particular scenario, yield results similar to bringing the cache microservices one tier closer to the end user. The best situation, of course, is to allocate the resources closer to end users and also implement network slices.

C. *Buffer assessment*

Another assessed metric is the buffer occupancy as shown in Fig. 4 and Fig. 5, for the scenarios without and with network slices, respectively. This evaluation is important because a healthy buffer will be able to hide the delay/latency/jitter variation, ensuring a more pleasant experience for the end-users. It also prevents stalls and can guarantee video quality even when the network is facing time-varying errors and bandwidth fluctuation. As aforementioned, here again, there is no standard deviation shown because the values were negligible. In the scenario without network slices, it is clear that the optimal buffer size is around 21 seconds worth of video. When the cache microservice is hosted at the Cloud layer, the time needed to reach the optimal buffer size is of 44 seconds. This may not be an issue for VoD services with a long duration, however, for short videos, this can have an impact on the perceived quality. This holds true because, if there is a fluctuation in the network conditions, the buffer will not be able to properly compensate for the problem. On

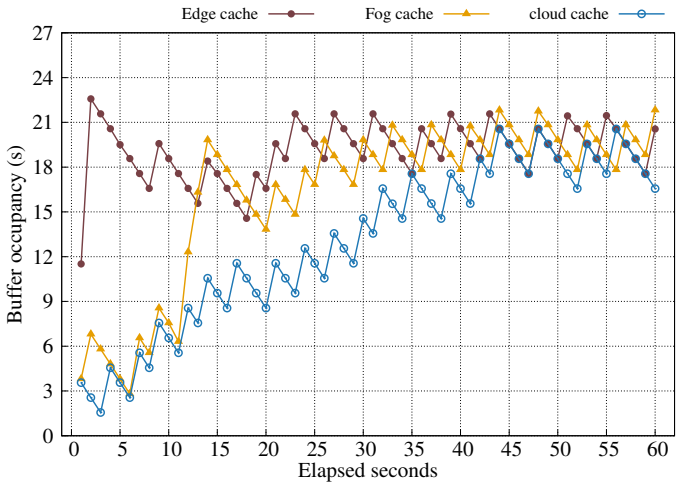

Fig. 4. Buffer occupancy without network slices

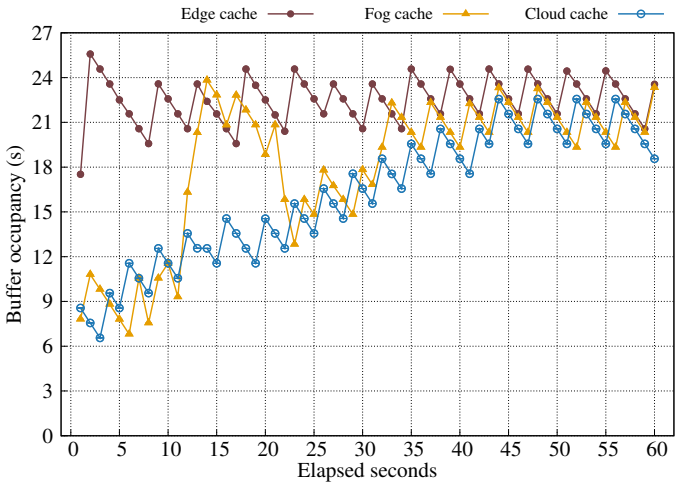

Fig. 5. Buffer occupancy with network slices

the other hand, with a Fog cache scenario without network slices, it took less time to reach the optimal buffer level. In this scenario, only 14 seconds were needed. After the initial negotiation stage, with the duration of 11 seconds, in this case, the buffer quickly reached a satisfactory occupancy level and a few seconds later the optimal level. It is important to notice that the increase in the buffer occupancy is also aligned with the bitrate enhancement, both have similar negotiation stage time in this tier at 11 and 13 seconds, respectively. Additionally, as expected in the scenario without slices, the faster buffer occupancy level was with the Edge cache. In this scenario, it only took 2 seconds to reach a satisfactory level and an additional few seconds to reach the optimal level. The same alignment between the negotiation stage time of the buffer occupancy and the bitrate level can be noticed here. With 2 seconds the buffer occupancy quickly arose as almost with the same time (4 seconds) as the bitrate boost.

In the scenario with network slices (Fig. 5), the results obtained had a quite similar pattern. Here the optimal buffer size was around 24 seconds and the players, considering the cache microservices in all the tiers, reached the optimal level earlier than in the scenario without network slices. For example, when the cache service was at the Cloud tier, the video player fills the buffer in 39 seconds, which means 5 seconds earlier than the without slices scenario. This is a reasonable improvement when considering short duration videos. It is also important to notice that even before filling the buffer, the scenario with network slices always provided a higher buffer occupancy, which means a higher chance to recover from errors in case of a network failure. Additionally, only 14 seconds were necessary to reach a good buffer size with the cache service at the Fog level. With the rapid increase in the buffer occupancy, the ABR protocol increased again the video bitrate, going from 4000 Kbps to 8000 Kbps. However, the available network resources were not enough to support this increase in the bitrate and the buffer occupancy started to drop very quickly. With 23 seconds the buffer reached a level not acceptable for this bitrate and the ABR mechanism was forced to decrease the video resolution. As mentioned before, if there was an intermediate bitrate between 4000 Kbps and 8000 Kbps, the ABR probably would select it, providing a more tailored video delivery according to the user resources. This means that the analysis of this type of architecture can also help to decide which video resolutions and bitrates can be adopted included in the video codification to better suit the user's needs. Moreover, when the cache service is hosted by the Edge tier, there is only a small difference in the scenarios with and without network slices. That being said, it is important to notice that when the slices are in place, the buffer filling and consumption is steadier. This lack of fluctuation indicates that even with microservices placed closer to the users, the adoption of slices can always further improve the video delivery.

D. Latency assessment

Fig. 6 and Fig. 7 depicts the latency in the three experiments set-up, without and with the adoption of network slices, respectively. The results here refer to the latency in the transmission of one MPEG-DASH segment. The size of the segment may vary according to the video bitrate and generally, it represents 2 seconds worth of video. The results in the scenario without network slices have a high variation of the standard deviation, even in a controlled network environment with low background traffic. One possible explanation for this is the Transmission Control Protocol (TCP) congestion-avoidance algorithm scheme called slow start. Because of this scheme, the transmissions started with a small amount of data and gradually increase the transmission size. It is possible to

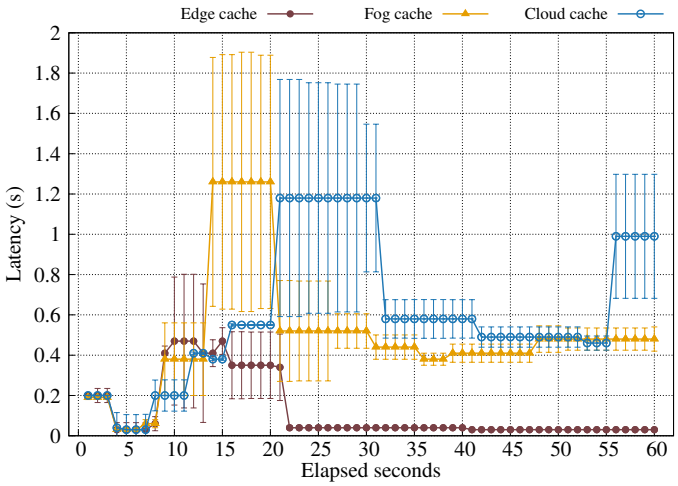

Fig. 6. Segment latency without network slices

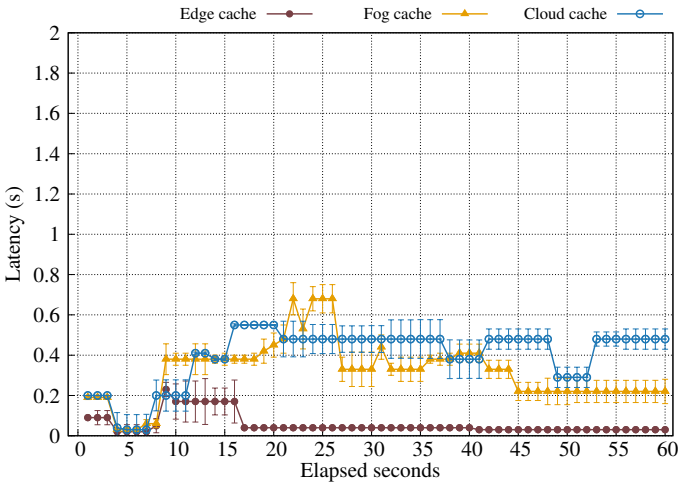

Fig. 7. Segment latency with network slices

notice in Fig. 6 that, for a period of time, there is a large variation in the latency. After that, it stabilizes and produces less variation. Besides that, as expected, the scenario with the cache microservice at the Edge tier presented the lowest latency. It is important to know that, owing to the video resolution and bitrate, the size of this segment in bytes is the second highest of all the experiments. It only stays behind when the microservice is placed at the Edge tier and with network slices in place. When the Fog and Cloud tier holds the cache service, there was a bigger variation on the results. This can be attributed to the additional hops that the packets had to go through from the source to the destination. As in the previous metric, the latency is also aligned with the buffer occupancy levels. For example, in the Edge tier environment, there was an increase in the latency from the 9th to the 21st second. The result of this was a large than expected decrease on the buffer occupancy levels that spread from the 8th to the 22nd second. The same situation is also noticeable in the Fog cache scenario: the latency increased from the 9th to the 20th second and the decrease in buffer occupancy levels is visible from 14th to the 20th second. Following the same pattern, the Cloud level experiment was also affected by this situation. The highest increase in the latency happened from the 21st to the 31st second, and as expected, the buffer occupancy levels were below the expectation between the 21st and the 31st second.

The latency assessment with the network slices in place gave quite different results. First of all, the standard deviation is several times less than in the scenario without slices. This is a foretold advantage of using this network feature. Since the resources are reserved and isolated from external interference it is expected very little variation on the network conditions, this enables better control over the resources as well as delivering the information in a timely manner. In the Edge setup, it is possible to notice that the latency drops earlier in the slice-enabled network, at 17 seconds against 23 seconds in comparison to the scenario without network slices. Additionally, when analyzing the Fog results, it is worth mentioning that when the DASH increased the resolution from "1280x720" at 4000 Kbps to "1920x1080" pixels at 8000 Kbps there is a spike in the segment latency. This happens around the 20th second. As mentioned before, this is the segment latency and segments of different resolutions have distinct sizes. In this case, it is possible to observe that the segment latency rapidly increased because the segment increased in size. In the Cloud setup, the latency had a discrete reduction, however, it provided less variation in the value and standard deviation, this means that the network connections were more stable and reliable. These results confirm the importance of using network slices together with elastic resources close to the end-users.

V. Conclusion and Future Works

Taking into consideration the ever-growing video transmission needs and the increasing number of available Edge devices to store, process, and manage information is evident the need for network architectures that exploit all this potential. This work proposes a multi-tier video delivery architecture to make use of this structure with the aid of Edge computing, microservices chaining, and network slices. The proof-of-concept testbed presented here, using a set of real tools and actual video, has demonstrated that such architecture can be used to improve the video delivery to the end-users. As future work, an additional number of microservices are going to be integrated into the proposed architecture and other network-related parameters are going to be assessed such as modification in real-time of the slices, the assessment of RAN parameters and video details, as well as smart city details. In addition, an SDN- and NFV-based microservice orchestration is going to be proposed, implemented, and evaluated.

ACKNOWLEDGMENT

This work was partially supported by grants #2018/02204-6 and #2015/24494-8, São Paulo Research Foundation (FAPESP), and it is part of the INCT project called the Future Internet for Smart Cities (CNPq 465446/2014-0, CAPES 88887.136422/2017-00 and FAPESP 2014/50937-1). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] R. Immich, E. Cerqueira, and M. Curado, "Efficient high-resolution video delivery over vanets," *Wireless Networks*, Feb 2018.

[2] 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)*, pp. 13–14, Dec 2018.

[3] C. Quadros, E. Cerqueira, A. Neto, A. Pescap, A. Riker, R. Immich, and M. Curado, "A quality of experience handover system for heterogeneous multimedia wireless networks," in *2013 International Conference on Computing, Networking and Communications (ICNC)*, pp. 1064–1068, Jan 2013.

[4] M. Curado, H. Madeira, P. R. da Cunha, B. Cabral, D. P. Abreu, J. Barata, L. Roque, and R. Immich, *Internet of Things*, pp. 381–401. Springer International Publishing, 2018.

[5] R. Immich, E. Cerqueira, and M. Curado, "Shielding video streaming against packet losses over vanets," *Wireless Networks*, vol. 22, pp. 2563–2577, Nov 2016.

[6] Cisco, "White paper: Cisco VNI forecast and methodology, 2016-2021," tech. rep., Cisco, September 2017.

[7] C. C. Byers, "Architectural imperatives for fog computing: Use cases, requirements, and architectural techniques for fog-enabled iot networks," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 14–20, 2017.

[8] F. Song, Z.-Y. Ai, J.-J. Li, G. Pau, M. Collotta, I. You, and H.-K. Zhang, "Smart collaborative caching for information-centric iot in fog computing," *Sensors*, vol. 17, no. 11, 2017.

[9] F. van Lingen, M. Yannuzzi, A. Jain, R. Irons-Mclean, O. Lluch, D. Carrera, J. L. Perez, A. Gutierrez, D. Montero, J. Marti, R. Maso, and a. J. P. Rodriguez, "The unavoidable convergence of nfv, 5g, and fog: A model-driven approach to bridge cloud and edge," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 28–35, 2017.

[10] D. Rosario, M. Schimuneck, J. Camargo, J. Nobre, C. Both, J. Rochol, and M. Gerla, "Service migration from cloud to multi-tier fog nodes for multimedia dissemination with qoe support," *Sensors*, vol. 18, no. 2, 2018.

[11] L. Gao, T. H. Luan, S. Yu, W. Zhou, and B. Liu, "Fogroute: Dtn-based data dissemination model in fog computing," *IEEE Internet of Things Journal*, vol. 4, pp. 225–235, Feb 2017.

[12] D. Nguyen, Z. Shen, J. Jin, and A. Tagami, "Icn-fog: An information-centric fog-to-fog architecture for data communications," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–6, Dec 2017.

[13] A. Gomes, T. Braun, and E. Monteiro, "Enhanced caching strategies at the edge of lte mobile networks," in *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, pp. 341–349, May 2016.

[14] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, *Fog Computing: A Platform for Internet of Things and Analytics*, pp. 169–186. Cham: Springer International Publishing, 2014.

[15] L. Bittencourt, R. Immich, R. Sakellariou, N. Fonseca, E. Madeira, M. Curado, L. Villas, L. da Silva, C. Lee, and O. Rana, "The internet of things, fog and cloud continuum: Integration and challenges," *Internet of Things*, 2018.

[16] T. Taleb, S. Dutta, A. Ksentini, M. Iqbal, and H. Flinck, "Mobile edge computing potential in making cities smarter," *Comm. Mag.*, vol. 55, pp. 38–43, Mar. 2017.

[17] R. Immich, E. Cerqueira, and M. Curado, "Towards a qoe-driven mechanism for improved h.265 video delivery," in *2016 Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, pp. 1–8, June 2016.

[18] H. Zhang, N. Liu, X. Chu, K. Long, A. H. Aghvami, and V. C. M. Leung, "Network slicing based 5g and future mobile networks: Mobility, resource management, and challenges," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, 2017.

[19] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5g network slice broker," *IEEE Communications Magazine*, vol. 54, pp. 32–39, July 2016.

[20] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN - key technology enablers for 5G networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 2468–2478, Nov 2017.