Contents

Literature Review

The world is experiencing a notable transformation in the manner people are consuming content for entertainment or learning purposes. This change can be considered quite substantial for the different industries content is part of. Broadcast television, for instance, is being phased out for online media services thanks to the growth of streaming solutions that have taken advantage of the robust internet connections in many markets (Fisher, 2015). One aspect of the development can be viewed as advantageous because it places consumers in total control of their viewing experiences. Put differently; customers have the ability to choose what they should consume on their media players, what time they can play videos, and the ability to watch anywhere and what device to watch the content on. However, the ability to choose from a variety of options comes at a subtle and real cost. Specifically, a variety of networks, content, devices and places implies that streams that reach consumers at any given time cut through unchartered channels.

In some cases, the quality of experience (QoE) is not the same for all users. Some have the best experience while others continue to report deteriorated service. Therefore, it is vital to examine what metrics define Q.O.E. for applications such as live streaming live actions such as sports.

A thorough examination of QoE for live streaming services should include the analysis of critical factors such as number points over a complicated and diverse value chain that has several handoffs. The sheer number of handoffs that are involved in transmitting content makes accountability unclear. Nevertheless, in modern times, Q.O.E. is sourced from an ecosystem that encompasses crucial elements. One of them is content, which entails resolution and quality of live streams from a broadcaster and the encoding process used. Secondly, content distribution networks (C.D.N.) that involve performance and optimization of video cache is key to the assessment process (Fisher, 2015).

A (C.D.N.) content distribution network or content distributed network is an extensive geographically distributed network setup consisting of location-based servers which are placed in various pop-sites and their data centres, which provide both high performance as well as high availability by serving out data from a particular server based on the end-users location. (Al-Abbasi, Aggarwal and Ra, 2019).

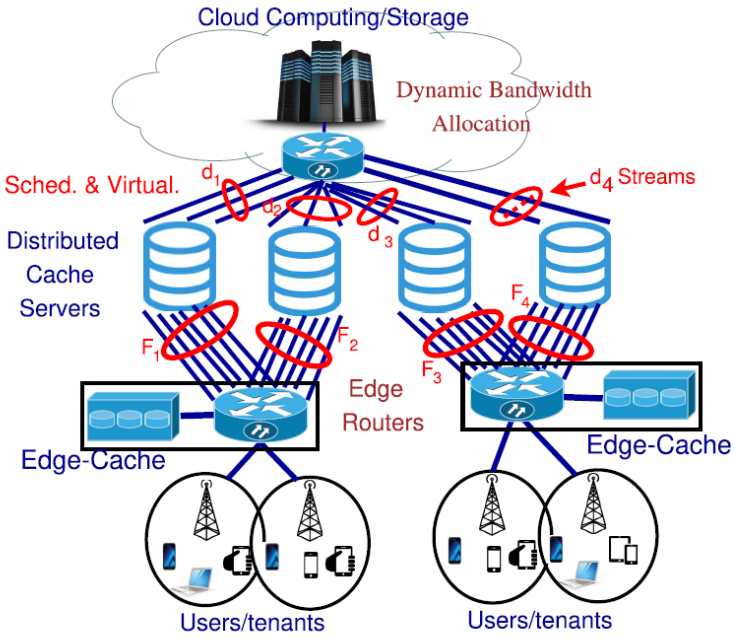

Figure 1: Diagram showing a Content Delivery Network model for video distribution, consisting of a single Datacenter, four caching servers and two edge routers with multiple parallel connections between the datacenter and caching servers as well as between the cache server and the edge router. (Al-Abbasi, Aggarwal and Ra, 2019).

Devices used, as well as their application further affect the quality of viewing live streams. User behaviour plays a crucial role based on consumer interaction with client applications of devices. Furthermore, end to end, watching should be simplified to address some technical problems and push for a small set of metrics that can be leveraged for reporting purposes. To this end, live streaming services can be summed up based on three critical parameters that mainly affect QoE. They are: start time, which is the elapsed period from when a user starts viewing live content and when content starts streaming; rebuffer rate, which is the number of times rebuffering process takes place during a live event; and lastly, average bit rate, which is the average rate of live streams that are presented in megabits per second (Mbps).

Cache replacement modules

Osuga, Asakura, and Taniguchi (2013, pp. 1151) support that live streaming events have been growing in popularity. The authors argue that up until now, caching systems that were leveraged to deliver live events from cache servers located near their users have cut traffic between the cache and origin servers. The development has continued to provide improved quality. Efficiencies such as Least Recently Used (L.R.U.) and Least Frequently Used (L.F.U.) have been deployed to improve caching and storing content with higher access frequency at cache servers. The high number of people live-streaming events has also forced service providers to use high-performance hardware such as fast RAM sticks and solid-state drives. However, processes that are used to replace caches have been subject to critical issues, especially when access frequency varies. The issue implies that while cached content is being pushed out of storage because of lower access frequency, videos cannot be removed because it is locked until delivery is concluded. Ultimately, the cache hit ratio, which is a quantitative measurement used to express the percentage of successful and missed cache hits over a given time frame, would deteriorate. The problem is commonplace in high-traffic applications. The problem must also be addressed for high-performance cache severs because the ejection content tends to have a higher chance of working status in the event of high traffic.

Using multi-tier fog architecture

Santos (2018, pp. 46) states that some approaches have been deployed to address the discussed concern. For instance, it has been proposed that a multi-tier fog architecture can be used to optimize a live streaming application for both wired and wireless connections that are used mainly for mobile devices.

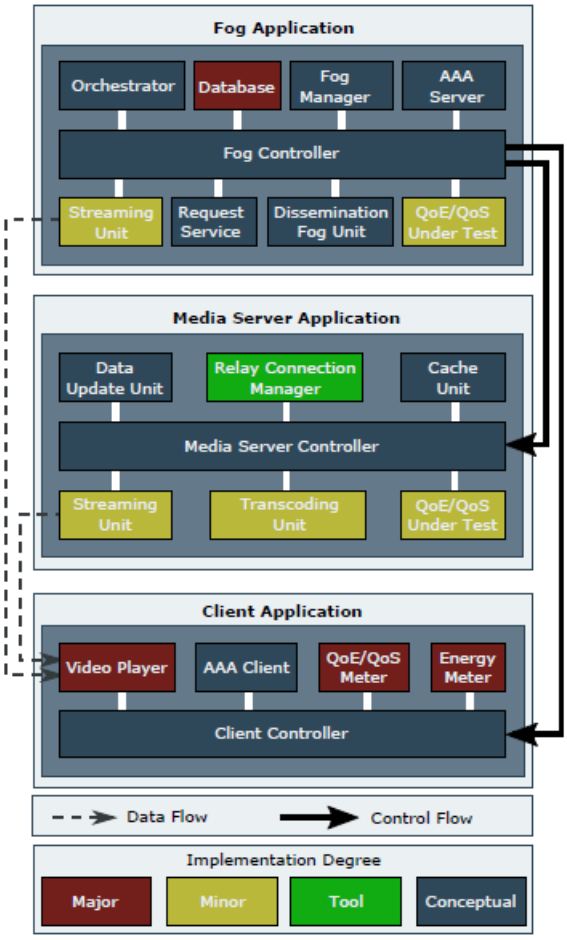

Figure 2: Diagram showing a Multi-tier fog architecture (Santos, 2018, pp. 35)

The elements of architecture seek to improve live streaming from the cloud and network operators, to mention a few. Santos further explains that open-source tools can be used for several tests as well as validation testing by using code intervention processes. The element implementation aspect based on code intervention in free and open-source software with custom

configuration can also be used to gain access to a suitable experimental environment with easy control over the quality accessed by customers. Furthermore, the implementation degree can be classified as a conceptual, tool, minor and significant. Conceptual degrees can be deployed for analyzing attributes in an element. The tool is used in situations where software is used to create an experimental setup. Minor is used for simplified configuration settings, and lastly, the principal plays an instrumental role in detailing creations or interventions. In addition, it is advised to examine some forms of unique architectures for multimedia distributions, and how their applications can improve user experiences for consumers streaming live events. The deployment should also go past operation usage of a network core and consumer devices to enforce an enhanced experience in a collaborative system.

The use of the three-tier fragmentation system

Li, Wang, L., Cui, and Zheng (2018, pp. 87) examined a new fragmentation system for video of HTTP live streaming. Features such as HTTP live streaming (H.L.S.) that was proposed by Apple have enabled live streams to be consumed on mobile devices, both for Android and iOS. Segmented HTTP demand has also been used for live streams and video-on-demand applications. Moreover, the authors state that the key to meeting robust video on demand requests is segmenting media files. The same approach can also be used for live events. A three-tire fragmentation strategy for transmission can also be used for communication of live events in P2P-based IPTV. The layers deployed in such a scenario are sub piece, chunk and piece. Chunk is the necessary element of a reference point for information exchange between peer nodes. The piece is leveraged as a data transmission end in a common working space, whereas sub piece entails a data transmission system in a degraded network space. It also serves as the starting point of data transmission. Moreover, performance testing must be conducted to establish the reliability of the fragmentation strategy. Simply put, the assessment verifies the superiority of the configuration by comparing it to a two-tier approach in the same amount of time. Some of the elements examined include percentage padding data and time of vide reception and split.

Common issues observed during HTTP adaptive streaming

Bentaleb et al. (2018, pp. 567) mention some of the most experienced issues for HTTP adaptive streaming (H.A.S.). H.A.S. is a technology widely adopted by various video distributors as well as content delivery platforms which use this technology to adapt the video being transmitted to the various conditions present on the network. This enables providers to improve the quality of the stream as well as the Quality of Experience (QoE) by stacking information from different network layers in order to adapt and deliver video in the best possible quality. Therefore, this allows for the network to take into account available quality levels of video, capabilities of end-user devices, server usage as well as current network conditions.

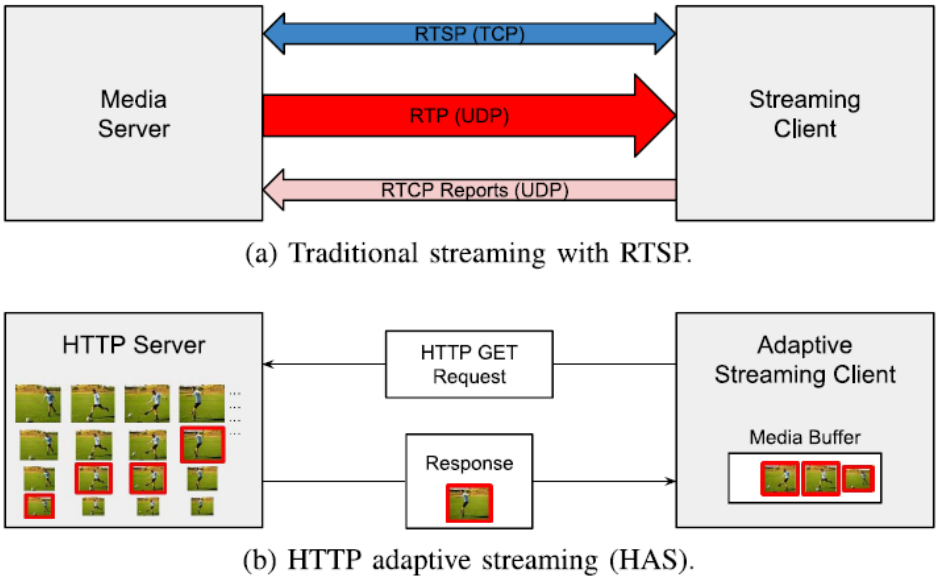

Figure 3: Diagram showing differences between traditional RTSP (image a) and H.A.S. streaming (image b). (Bentaleb et al. 2018, pp. 563)

The concerns are based on the heterogeneous nature of networks, a high number of users and the popularity of high-quality live streams. The issues include multi-client rivalry that has led to stability lapses, consistent-quality live streams, QoE measurement and optimization and inter-destination content synchronization. For competition and stability issues, H.A.S. has leverage stability for clients who are advised to skip frequent bit-rate switching that causes live streams to

stall. Fairness is also introduced for several H.A.S. clients in order to free bandwidth for wide pool users. High utilization helps in deploying resources efficiently in events that are preceded attempts to make streams stable and accessible to all users. Consistent quality streaming has been supported by assessments that have established a non-linearity between video bit-rate and perceptual quality. While bandwidth can be availed, the delivered live streams can have varying conditions for end-users. Nevertheless, it is suggested or preferred to stream content with reliable quality than at a consistent bit-rate. As discussed previously, Q.O.E. optimization is subject to issues due to the shifts observed in best-effort environments. One primary concern is associated with limitations in a unified and quantitative method of measuring QoE. Lastly, inter-destination content synchronization has been subject to issues linked to H.A.S. because clients adaptively live stream events based on their existing network conditions. Thus, dedicated Q.O.E. models need to be created to improve visual quality while taking synchronization accuracy into account.

Bit-rate adaptation

Sobhani, Yassine, and Shirmohammadi (2017, pp. 3) suggest a series of contributions for addressing high traffic and improving QoE. One of their models is based on a fuzzy logic controller (FLC) for H.A.S. that should dynamically manage the rate of a requested live stream or video-on-demand content and offers a superior decision-making process for downloading the stream or video content. The author's approach is said to address issues of defining absolute buffer limits by examining fuzzy (uncertain) elements of buffer thresholds. The method is further reported to continuously download segments provided there is a video bit-rate that surpasses the estimated and available bandwidth. Generally speaking, the approach is critical because it cuts the ON-OFF traffic models in affected scenarios, which in the end reduced stability issues as well as unfairness for competing service providers. Ultimately, QoE is improved. Furthermore, the method allows devices such as congestion controllers to remain active over an extended period than ordinary H.A.S. Hence, solutions that can deploy the method stand a chance to share available bandwidth for in a fair manner to their customers compared to other H.A.S. processes. Lastly, the approach is supported by a Grey-model predictor that supports FLC, which, as a result, deploys more informed decisions on a predicted buffer level.

Cache partitioning

Moreover, authors Li, Sharief, Fayed, and Hassanein (2018, pp. 408) discuss bit-rate adaptions and caching partitioning techniques that have been leveraged for streaming services via information networks. Their investigations argue that cache placement should be based on adaptive streaming because bit-rate adaptive approaches do not match with generic caching processes. Further assessments look into the problem of oscillation dynamics that are preceded by interplay that is observed in bit-rate adaptation management and in-network caching. The entire issue presents a fundamental approach to caching and can be argued as a novel premise that can be leveraged for safeguarding cache partitions, particularly those that need elevated bit-rates. Ultimately, a system that allows ideal caching placement is created, and serves as a reliable tool for adaptive streaming content. Additionally, the safeguarding systems proposed by the authors support bit-rate partitioning of cache volumes. The process allows the stabilization of bandwidth that usually goes up and down. In fact, the system is projected to perform admirably because a network of caches is visible throughout all forwarding channels such as from consumers to the network core. The concept of the proposed model is further supplemented by its potential gain, which is validated and implemented via a caching scheme, and illustrates how partitioning would improve QoE based on video quality and a substantial cut in bit-rate oscillation.

The outcome in bit-rate adaptation

According to Nguyen, Jin, and Tagami (2016, pp. 74), adaptions that are cache-friendly can be deployed to address congestion issues in information-centric networks (ICN) in a manner similar to Li, Sharief, Fayed, and Hassanein. Their studies examine utility-fair bit-rate adaptation. In that case, utility proportional fairness (U.P.F.) is leveraged to quantify the available bandwidth. By design, the process avoids cases of network instability and unfairness. The process is different from other commonplace bit-rate adaptations because the approaches deploy bandwidth to users differently. Generally speaking, utility-fair bit-rate adaption calculates bandwidth through congestion feedback that then enforces fairness for users who may live streaming and event or accessing V.O.D. services via different platforms. Furthermore, as one data packet is received, another separate interest packet is created and deployed to a network after a defined delay that is

based on playback buffer level. The authors further give an example where a user supposedly downloads 2 G.B. of content in a congested network haring 24Mbit/s bottleneck among eight other users. To this end, the requested bit-rate of the users can be summarized, as shown in Figure 4.

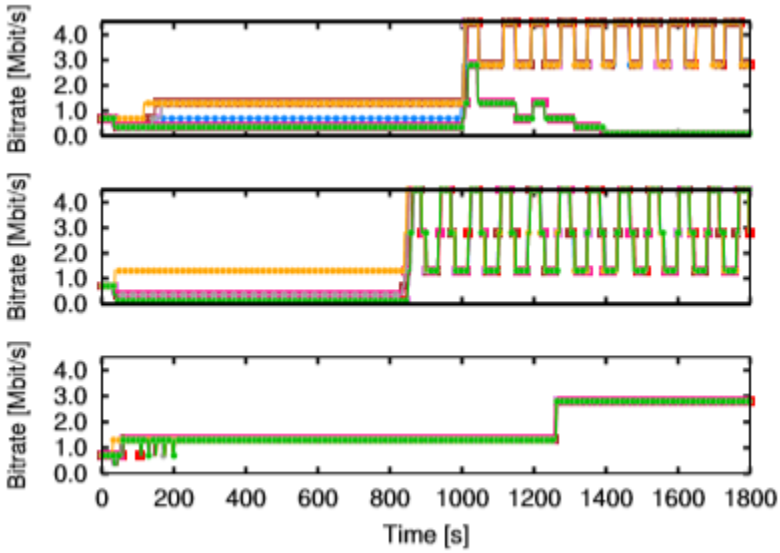

Figure 4: Chart showing eight streaming session alongside one downloading session (Jin, and Tagami, 2016, pp. 75)

The first half of the simulation demonstrates a situation when an elastic user is downloading materials, and each bit-rate behaves differently based on elastic background traffic. Thus, utility-fair adaptation is used to ensure fairness for all members using the network prior to and after the conclusion of the elastic traffic.

Replacing cache based on distance

Further work has been done by Kamiyama, Nakano, and Shiomoto (2016, pp. 849), who have based their work on cache placement on the distance to primary servers. The proposal suggested by the researchers is based on a cache-replacement policy that is based on the need to cut the link load that arises from delivering online materials from cache servers. The method suggests that storage hardware of cache servers should be separated into virtual caches, and content elements

should be controlled by each virtual cache separately based on hop distance to origin servers. In addition, it is advised to allocate space optimally to each virtual cache at each nodal point in order to optimize the mean reduction of flow hop length. Generally speaking, the contributions of the authors can be summed up with a few statements. First, to exhaustively improve the cutting effect on link load that is sourced from delivering materials from cache servers without engaging in challenging operations at each stage, the group recommends a cache replacement system that is built with several virtual caches that are associated to their proximity to origin servers. The second element discussed a mode that can be used to reduce flow hop length of content channels by using a procedure that optimally sets aside storage for each virtual cache. Lastly, the proposed system can cut the average link load substantially based on the evaluation of network topologies in the U.S.A.

Use of HTTP adaptive streaming

Other researchers such as Lee, Dovrolis, and Begen (2014, pp. 33) have discussed caching in HTTP adaptive streaming as either a development or a setback in events that require quality live streaming. The authors present a system referred to as Visit Cache, which analyzes and shapes intelligent cache. It monitors bandwidth metrics as one of the inputs that affect the development of algorithms and moving content from origin servers to the customers' devices. While there are several other processes that can be used to transferring bandwidth, the researchers chose to use a different element of cache servers for measurements that were established for their tests. The shaping of traffic from cache is equally critical because it was noted that oscillations led to inaccurate bandwidth readings for channels that were managed from the cache. The issue was addressed by a proposal that defined a cache-based model for traffic shaping that also ensured that end-users did not request parts that surpassed the path bandwidth. The assessment of representation bit-rates was also essential for the cache servers to redeploy client requests for channels that were marked with defined representation bit-rates as well as shaping impactful transfers. A visci implementation constituted different models for gaining access to content, based on a given scenario. For instance, the content could be read directly if the provider employed a plaintext M.D.P. file distributed over unencrypted paths. The other and last aspect was duplicating a M.D.P. for the requirement for the cache server to impart more logic to pinpoint an event.

References

Al-Abbasi, A., Aggarwal, V. and Ra, M. (2019). Multi-Tier Caching Analysis in CDN-Based Over-the-Top Video Streaming Systems. IEEE/ACM Transactions on Networking, 27(2), pp.835-847.

Bentaleb, A., Taani, B., Begen, A.C., Timmerer, C. and Zimmermann, R., 2018. A survey on bit-rate adaptation schemes for streaming media over HTTP. *IEEE Communications Surveys & Tutorials*, *21*(1), pp.562-585.

Fisher, M. (2015). Making Sense of Streaming Video Quality of Experience - Qwilt. [online] Qwilt. Available at: https://qwilt.com/making-sense-of-streaming-video-quality-of-experience/ [Accessed 10 Jan. 2020].

Kamiyama, N., Nakano, Y. and Shiomoto, K., 2016. Cache replacement based on the distance to origin servers. *IEEE Transactions on Network and Service Management*, *13*(4), pp.848-859.

Lee, D.H., Dovrolis, C. and Begen, A.C., 2014, March. Caching in HTTP adaptive streaming: Friend or foe? In *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop* (p. 31). A.C.M.

Li, W., Sharief, O., Fayed, M. and Hassanein, H.S., 2018, October. Bit-rate adaptation-aware cache partitioning for video streaming over Information-centric Networks. In *2018 IEEE 43rd Conference on Local Computer Networks (L.C. N)* (pp. 401-408). IEEE.

Li, X., Wang, L., Cui, J. and Zheng, B., 2016, December. A New Fragmentation Strategy for Video of HTTP Live Streaming. In *2016 12th International Conference on Mobile Ad-Hoc and Sensor Networks (M.S.N.)* (pp. 86-89). IEEE.

Nguyen, D., Jin, J. and Tagami, A., 2016, September. Cache-friendly streaming bit-rate adaptation by congestion feedback in icn. In *Proceedings of the 3rd A.C.M. Conference on Information-Centric Networking* (pp. 71-76). A.C.M.

Osuga, T., Asakura, T. and Taniguchi, K., 2013, March. A Cache Replacement Method for Crowded Streaming Cache Servers Responding to Rapidly Changing Access Patterns.

In *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (A.I.N.A.)* (pp. 1150-1156). IEEE.

Santos, H.L.M.D., 2018. A Multi-tier fog architecture for video-on-demand streaming.

Sobhani, A., Yassine, A. and Shirmohammadi, S., 2017. A video bit-rate adaptation and prediction mechanism for HTTP adaptive streaming. *A.C.M. Transactions on Multimedia Computing, Communications, and Applications (T.O.M.M.)*, *13*(2), p.18.