Proceedings of CECNet 2021
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210441

Ultra-HD Video Streaming in 5G Fixed Wireless Access Bottlenecks

Koffka KHAN [1], Wayne GOODRIDGE
Department of Computing and Information Technology (DCIT), The University of the West Indies, St. Augustine, Trinidad and Tobago, West Indies

Abstract. 5G Fixed Wireless Access (FWA) is an enabling technology in intelligent systems (IS) that may provide Ultra-HD (UHD) video streaming services with high Quality of Experience (QoE) in a small business use case setting. However, UHD streaming over 5G FWA is difficult in terms of latency and dependability due to numerous network factors. Due to this there may be multiple video players competing for network resources when streaming a UHD video. To date there has been very little work of 5G 'last mile access' streaming over bottleneck FWA Local Area Networks (LANs) under congested network conditions. The bottleneck link is the 5G FWA gateway. In these networks viewers may get sub-optimal QoE. Adaptive bitrate (ABR) algorithms are used to select the near optimal bitrates during a streaming session. To obtain the QoE of viewers in 5G FWA bottleneck networks we study the performance of four DASH-based adaptive video streaming algorithms (MPC, BOLA, Oboe and Pensive). BOLA performs the best and Pensive the worst. However, BOLA's overall performance is sub-optimal. This work supports the need for developing new ABR algorithms for the 5G FWA environment.

Keywords. 5G, FWA, Ultra-HD, video, streaming, QoE, last mile access, gateway, bottleneck

1. Introduction

Fixed Wireless Access (FWA) is one of the first planned technologies that could transform the digital environment during the early stages of 5G deployment. FWA delivers Internet services to end consumers with lower infrastructure costs by using both dedicated fixed networks and shared mobile networks (e.g., 4G and 5G networks), see Figure 1. FWA will offer innovative and more flexible wireless broadband options for homes and businesses. Many systems are considering bands at 'higher' millimeter wave frequencies due to the need for high bandwidth and new spectrum. Current 5G FWA devices can attain 150 Mbps on 5G New Radio (5G-NR) bands below 7 GHz [1]. Future 5G-NR deployments in the millimeter wave (mmWave) spectrum will allow the FWA to attain multi-gigabit per second (Gbps) rates, equivalent to super fast fiber networks. However, a clear LoS (Line of Sight) is required from the base station to consumer premises equipment (CPE) [2]. Practically this has its challenges. To address this challenge, a number

[1]Corresponding Author: Lecturer in Computer Science, Department of Computing and Information Technology (DCIT), The University of the West Indies, St. Augustine, Trinidad and Tobago, West Indies; E-mail: koffka.khan@sta.uwi.edu

of options are being considered. This includes the placement of the base station and CPE in a direct LoS which gives the best coverage. Multiple input, multiple output (MIMO) and beamforming are used to overcome the high data loss incurred when the transmitter and receiver are not in a direct LoS [3].

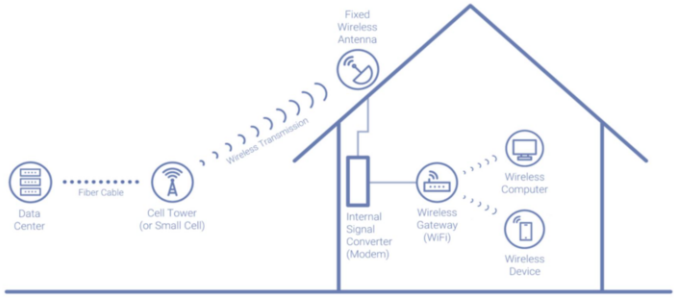

Figure 1. Fixed Wireless Access (Source: cbinsights.com).

HTTP Adaptive Streaming (HAS) is the de-facto solution for Internet video distribution and is used by prominent providers such as YouTube and Netflix. DASH's main idea is that video sequences are broken into chunks that are encoded with varied qualities (e.g., resolutions and bit rates) and provided through Hyper Text Transfer Protocol (HTTP) and Transmission Control Protocol (TCP) from web servers [4,5]. During streaming sessions, a DASH client requests the best quality for each of these chunks based on its local conditions (e.g., available bandwidth and buffer level). Given that the DASH standard does not specify the implementation details of the adaptation logic, a great deal of effort has gone into developing effective solutions. This logic can be implemented at the client [6,5], in-network [7] or server [7].

UHD has a resolution of 3840 x 2160 pixels. UHD video streaming is difficult for a variety of reasons [8]. Your connection speed must be at least 25Mbps in order to watch a UHD video stream a small business [9]. Thus, our first reason is UHD video streams consume a lot of bandwidth. Our second reason is DASH-based video streaming systems often utilize bitrate adaptation, buffer control, Quality of Experience (QoE) [10] inference in their optimization strategies which are not suited to the 5G environment [8]. Our third reason is TCP's [11] transitory behavior via its congestion control mechanism significantly underutilizes network bandwidth. Finally, mmWave 5G experiences [12] dramatic and frequent performance variations which results in highly varying network speeds [8], see Figure 2. This might possibly confound network and application layer logic, such as ABR video streaming, and result in underutilization of the carrier's channel capacity and resources.

The contributions of this paper are: (1) First time the performance of BOLA [13], MPC [14], Pensieve [15] and Oboe [16] DASH ABR algorithms are studied under a UHD DASH video streaming testbed mimicking 5G FWA. (2) Subjective QoE metrics have never been used to evaluate the performance of these streaming algorithms in this environment. (3) Experimental work involving varying number of UHD DASH players are performed and analyzed. We present our work in five Sections. MPC, BOLA, Oboe and Pensive DASH algorithms are described in Section 2. These algorithms had superb performance in 4G video streaming. Our experimental 5G FWA Testbed is described in

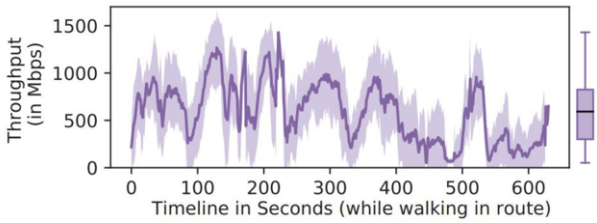

Figure 2. Variation in 5G Throughput [8].

Section 3. Objective and subjective QoE findings are given in Section 4 with a discussion on how sharing 5G DASH UHD video using ABR algorithms in a FWA environment impacts viewer QoE. Finally, we emphasize the need to develop new 5G ABR algorithms in our conclusion, Section 5.

2. DASH approaches

2.1. MPC: Model Predictive Control

MPC [14] selects bitrates that use the best QoE parameters across the projected segment range using both throughput estimations and buffer occupancy data. The characteristics that are generally evaluated include average video quality, average video quality fluctuations, rebuffer or startup delay. However, because viewers may have various priorities regarding which of these factors is more important, the weighted total of these four metrics is used to define the QoE of the video segment. In this paper we assume equal weightings for the four metrics. The MPC method makes appropriate use of both bandwidth estimation and buffer levels. MPC is resistant to prediction error because at any step it reduces the inaccuracy in prediction, see Figure 3 for the MPC adaptation workflow algorithm and [14] for accompanying details.

1: Initialize
2: **for** $k = 1$ to K **do**
3: **if** player is in startup phase **then**
4: $\hat{C}_{[t_k,t_{k+N}]}$= ThroughputPred($C_{[t_1,t_k]}$)
5: $[R_k, T_s] = f^{st}_{mpc}\left(R_{k-1}, B_k, \hat{C}_{[t_k,t_{k+N}]}\right)$
6: Start playback after T_s seconds
7: **else if** playback has started **then**
8: $\hat{C}_{[t_k,t_{k+N}]}$= ThroughputPred($C_{[t_1,t_k]}$)
9: $R_k = f_{mpc}\left(R_{k-1}, B_k, \hat{C}_{[t_k,t_{k+N}]}\right)$
10: **end if**
11: Download chunk k with bitrate R_k, wait till finished
12: **end for**

Figure 3. Video adaptation workflow using MPC [14].

2.2. BOLA: Buffer Occupancy based Lyapunov Algorithm

The BOLA [13] design is purely buffer-based. The DASH session is modeled as a stochastic optimization problem with a time-average objective over a finite horizon. Dynamic programming (DP) is used to solve it. BOLA maximizes the playback's utility as determined by a weighted sum of the average quality bitrate and average stalling time. BOLA solves the chunk quality choice problem with a Lyapunov optimization approach for each chunk request, that is whether to avoid downloading a new chunk or to download a new utility-maximizing chunk. BOLA avoids the overheads of more complicated bandwidth prediction ABR schemes and is more reliable in the face of bandwidth variations, see Figure 4 for the BOLA algorithm and [13] for accompanying details.

for n in $[1, N]$ **do**
  $t \leftarrow \min$[playtime from begin, playtime to end]
  $t' \leftarrow \max[t/2, 3p]$
  $Q^{\mathrm{D}}_{\max} \leftarrow \min[Q_{\max}, t'/p]$
  $V^{\mathrm{D}} \leftarrow (Q^{\mathrm{D}}_{\max} - 1)/(\upsilon_M + \gamma p)$
  $m^*[n] \leftarrow \arg\max_m (V^{\mathrm{D}} \upsilon_m + V^{\mathrm{D}} \gamma p - Q)/S_m$
  if $m^*[n] > m^*[n-1]$ **then**
    $r \leftarrow$ bandwidth measured when downloading segment $(n-1)$
    $m' \leftarrow \max m$ such that $S_m/p \leq \max[r, S_1/p]$
    if $m' \geq m^*[n]$ **then**
      $m' \leftarrow m^*[n]$
    else if $m' < m^*[n-1]$ **then**
      $m' \leftarrow m^*[n-1]$
    else if some utility sacrificed for fewer oscillations **then**
      pause until $(V^{\mathrm{D}} \upsilon_{m'} + V^{\mathrm{D}} \gamma p - Q)/S_{m'} \geq (V^{\mathrm{D}} \upsilon_{m'+1} + V^{\mathrm{D}} \gamma p - Q)/S_{m'+1}$ ▷ BOLA-O
    else
      $m' \leftarrow m' + 1$ ▷ BOLA-U
    end if
    $m^*[n] \leftarrow m'$
  end if
  pause for $\max[p \cdot (Q - Q^{\mathrm{D}}_{\max} + 1), 0]$
  download segment n at bitrate index $m^*[n]$, possibly abandoning
end for

Figure 4. The BOLA Algorithm [13].

2.3. Oboe

Oboe [16] calculates the best potential parameters for a particular ABR algorithm under various network situations. At runtime, it proactively updates the settings for real-world network scenarios. Oboe takes advantage of piecewise-stationary network connections. It addresses the problem of network configuration sensitivity to changing network environments. It begins with an offline stage in which the best configuration for each network state is calculated in advance. Then, it employs an online stage. This stage monitors changes in network status throughout a session and sets the optimal pre-calculated configuration to the present (stationary) state. Oboe can also include other parameters such as session type (live and video on demand (VOD)), streaming rate thresholds and QoE measures (e.g., choice between frequent video stalls vs. UHD quality level). Oboe tweaks robustMPC [14], an MPC version developed to employ a 5-segment horizon, in our 5G trials. It employs a conservative throughput estimate, which normalizes the de-

fault throughput prediction by taking the highest estimation error over the preceding five chunks. Robust MPC efficiently optimizes the worst-case QoE. Oboe's offline and online pipeline is shown in Figure 5 with [16] for accompanying details.

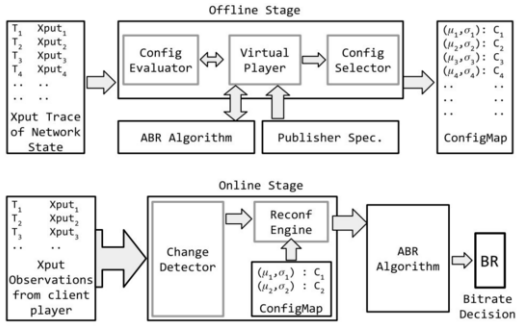

Figure 5. Logical diagram of Oboe's offline and online pipeline [16].

2.4. Pensieve

Reinforcement learning (RL) is used by Pensieve [15] to solve difficulties that video players have during DASH session. It creates a neural network model that uses receiver's video player input to pick bitrates for potential video chunks. Only through analyzing the outcomes of previous decisions can Pensieve develop the capacity to make adaptive bitrate (ABR) decisions. Pensieve constructs its policy with the actor-critical algorithm (see Figure 6 and [15] for accompanying details), a policy gradient method. By looking at the execution paths, policy gradient techniques can quantify the gradient of the projected cumulative reward. Once Pensieve has created an ABR algorithm using our 5G simulations, the model's concepts must be applied to real DASH sessions. When a client requests for individual chunks arrive at the video provider, Pensieve feeds the incoming data into its neural network model, which responds to the video client with the bitrate level to be used for the next chunk update.

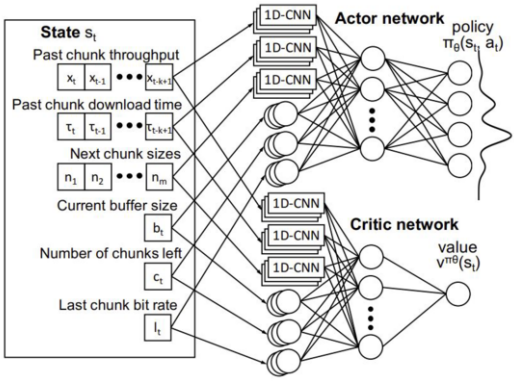

Figure 6. The Actor-Critic algorithm that Pensieve uses to generate ABR policies [15].

3. Empirical 5G FWA Testbed

3.1. 5G FWA Business viewer environments

3.1.1. 5G FWA Business Environment 1: Worker added when viewing video

All 10 workers of the business begin watching their (separate) video at the same time until it ends.

3.1.2. 5G FWA Business Environment 2: Worker removed when viewing videos

One of the workers begin to watch their video. In addition, each worker begins watching their (separate) video at 60 second intervals. Thus, all 10 workers will be watching a video during the last minute.

3.2. 5G FWA small business network

Two-second Group of Pictures (GOP) are used to create a two-minute and twenty-two-second clip in 3840x2160 (4K) with thirteen bitrates ranging from 1.8Mbps to 18Mbps and on demand DASH profiles [17]. There are three 4K bitrates, two Full HD bitrates, two HD bitrates, and six lower resolution bitrates. The two second GOP is repeated five times with different segment lengths (2, 4, 6, 10 and 20 seconds). We used 2 second chunks in experiments. The video clips (Big Buck Bunny (BBB), Sintel, and Tears of Steel (TOS)), each had a duration of more than 10 minutes.

One laptop serves as the UHD video streamer, while the remainder of the system (including wireless gateway) is hosted on a high-end PC. The testbed is set up to be able to stream UHD video. The Streamer machine runs Ubuntu 18.04.5 LTS 64-bit and comes with 32 GB of RAM, an AMD Ryzen 9 5900X 12-core, 24-Thread CPU and a 2.0 TB hard drive. In our FWA environment the laptop acts as the base station. The high-end PC runs Ubuntu 16.04.7 LTS 64-bit and comes with 64 GB of RAM, a Intel(R) Core(TM) i9-9980HK CPU @ 2.40GHz processor, and a 3.0 TB hard drive. The wireless gateway (CPE in our FWA environment) is hosted on a FreeBSD dummynet virtual machine on the PC. It is connected to the UHD players (business workers in our FWA environment) who are also hosted on the PC. Network parameters were set to follow the SPEED-5G [9] specifications.

We set the wireless gateway at 120Mbps. There are 10 workers streaming UHD video. Each video requires 18Mbps download speeds to be viewed at the highest quality. Thus, when the sixth worker joins in experiment 2's 5G FWA Business Environment 1 (see Table 1) or the entire experiment 1's 5G FWA Business Environment 1 the gateway will act as a bottleneck link. UHD players will have to compete for network resources in order to serve high quality to their users. We explore the results of this competition in our experiments, see Section 4.

Time (s)	60	120	180	240	300	360	420	480	540	600
Add Worker	1	2	3	4	5	6	7	8	9	10
UHD Bandwidth Usage (Mbps)	18	43	68	93	118	143	168	193	218	243

Table 1. Bottleneck: Bandwidth usage exceeds 120Mbps

4. Experimental results

4.1. 5G FWA Business Environment 1

Except for BOLA, all other ABR algorithms suffer from more than 2 video stalls while operating over 5G FWA, see Table 2. Video stalls are the most serious issue with video streaming over 5G. BOLA performs best across four objective QoE metrics (switches, stalls, stall duration and average utilization) followed by MPC, Oboe and Pensieve.

ABR	Switches	Stalls	Avg. Stall Duration (s)	Avg. Utilization
BOLA	75	2	0.5	23.27
MPC	87	5	1.5	18.59
Oboe	108	11	3	14.38
Pensieve	116	16	4	12.59

Table 2. QoE Objective metric performance of DASH ABR in 5G FWA Business Environment 1

4.2. 5G FWA Business Environment 1

Video stalls are still numerous in this scanario but are less than the experiment in Section 4.1, see Table 3. All players compete for bandwidth throughout the experiment in Section 4.1. However, bottleneck conditions only occur after 360 s in this experiment so there is more overall bandwidth available for players initially and up to the last minute where all workers are using the network resources. Again, BOLA performs best across the objective QoE metrics followed by MPC, Oboe and Pensieve.

ABR	Switches	Stalls	Avg. Stall Duration	Avg. Utilization
BOLA	37	1	0.3	24.13
MPC	48	3	1.2	19.83
Oboe	64	8	2.5	15.26
Pensieve	25	10	3.3	13.79

Table 3. QoE Objective metric performance of DASH ABR in 5G FWA Business Environment 2

4.3. Subjective QoE tests

The MPC, BOLA, Oboe, and Pensieve ABR algorithms were evaluated against the identical 5G bandwidth traces that we obtained during our study (see Section 4.1). This enabled fair comparisons to be made. The video chunks were then merged into a single UHD file. For each ABR algorithm, this procedure was repeated. We recruited 76 people to take part in the subjective tests in our research. All of the volunteers were not color-blind and had normal vision acuity, and they had no prior awareness of the DASH ABR algorithm utilized in the experiment. Twenty sequences were assessed by the volunteers. They were only required to watch five sequences of a specific UHD video to retain their focus on the evaluation. As a result, 19 people assessed every test sequence. This was

judged sufficient to guarantee that the results were not skewed by the presence of a few volunteers. The subjective assessments took place in a lab with controlled lighting. The sequences were presented on a 31.5 inches monitor. A resolution of 3840 x 2160 and aspect ratio of 16:9 were used.

BOLA promotes lower number of switches, fewer stalls, small stall duration times and better bandwidth utilization among multiple UHD players, thereby leading to a better subjective user experience, as justified by its MOS rating [18] in all evaluation environments, see Figure 7 and Figure 8.

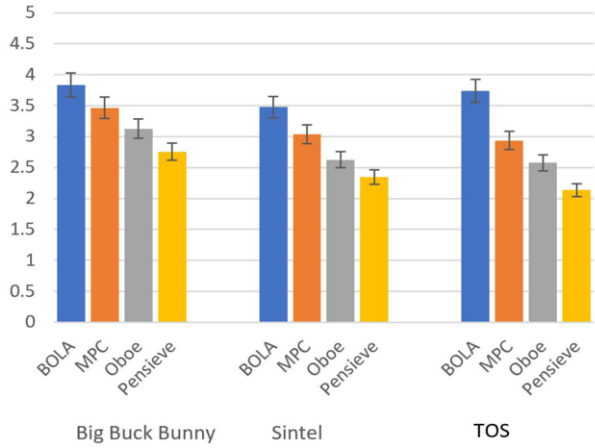

Figure 7. 5G FWA Business Environment 1: Subjective visual quality comparison of UHD DASH - an average MOS score ratings.

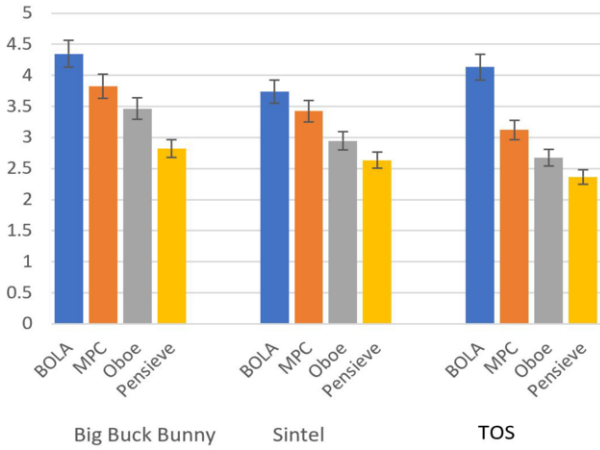

Figure 8. 5G FWA Business Environment 1: Subjective visual quality comparison of UHD DASH - an average MOS score ratings.

4.4. Discussion

BOLA utilizes buffer information to make its next segment streaming decisions. This metric proves more successful than the metrics used by the other ABR DASH-based algorithms explored in this paper. MPC utilized incoming bitrate and buffer metrics in its adaptation strategy. However, the incoming rates are difficult to predict due to highly varying nature of 5G traffic. This fact interferes with the MPC players ability to make proper inferences on the incoming bitrates. Thus, the MPC player performance is lower than the fully buffer-based algorithm, BOLA. Oboe utilizes throughput estimate over the previous five chunks. Its performance suffers and again, we see sub-optimal performance with an ABR using incoming rate estimates as part of its adaptation strategy. Pensieve performs the worst. A possible explanation is that, in order to train the model to understand 5G FWA specific features and make better judgments, a larger dataset is required. However, this merits more investigation. In addition, coping with the high bandwidth variations in 5G FWA may cause the player to sometimes run out of data chunks due to for example a prolonged spell of low bandwidth which could result in more frequent video stalls.

5. Conclusion

There has not been enough research conducted on video streaming in 5G FWA networks. Most researchers focus on mobile 5G video streaming. To bridge this research gap we conducted DASH-based video streaming experiments mimicking a 5G FWA network with a bottleneck gateway. We used the MPC, BOLA, Oboe and Pensive ABR DASH algorithms. Experiments consisted of all small business workers viewing UHD videos or joining the streaming session. Most algorithms perform poorly in the 5G FWA experiments, except BOLA. Many ABR algorithms use a throughput predictor to factor network throughput into their decisions, and their success is highly reliant on predictive accuracy. ABR 5G throughput prediction is poor due to high 5G bandwidth variability. BOLA is a strictly buffer-based ABR algorithm thus does not suffer from inaccurate predictions as the others (MPC, Oboe) or lack of bigger training dataset in the case of Pensieve. Current 5G covers several bands and has a wide range of network performance. Thus, developing improved throughput prediction methods is critical not just for making ABRs function well over 5G FWA, but also for making ABRs work well in general. New adaptive techniques will have to be developed to cope with the ever expanding 5G FWA environment.

References

[1] Asplund H, Astely D, von Butovitsch P, Chapman T, Frenne M, Ghasemzadeh F, et al. Advanced Antenna Systems for 5G Network Deployments: Bridging the Gap Between Theory and Practice. Academic Press; 2020.

[2] Topyan K, Ulema M. Architectural and Financial Considerations for Deploying 5G Based Fixed Wireless Access. In: 2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom). IEEE; 2020. p. 1-6.

[3] Bechta K, Ziółkowski C, Kelner JM, Nowosielski L. Modeling of downlink interference in massive MIMO 5G macro-cell. Sensors. 2021;21(2):597.

[4] Tuysuz MF, Aydin ME. QoE-based mobility-aware collaborative video streaming on the edge of 5G. IEEE Transactions on Industrial Informatics. 2020;16(11):7115-25.
[5] Khan K, Goodridge W. S-MDP: Streaming with markov decision processes. IEEE Transactions on Multimedia. 2019;21(8):2012-25.
[6] Khan K, Goodridge W. B-DASH: broadcast-based dynamic adaptive streaming over HTTP. International Journal of Autonomous and Adaptive Communications Systems. 2019;12(1):50-74.
[7] Khan K, Goodridge W. Server-based and network-assisted solutions for adaptive video streaming. International Journal of Advanced Networking and Applications. 2017;9(3):3432-42.
[8] Ramadan E, Narayanan A, Dayalan UK, Fezeu RA, Qian F, Zhang ZL. Case for 5G-aware video streaming applications. In: Proceedings of the 1st Workshop on 5G Measurements, Modeling, and Use Cases; 2021. p. 27-34.
[9] Navarro-Ortiz J, Romero-Diaz P, Sendra S, Ameigeiras P, Ramos-Munoz JJ, Lopez-Soler JM. A survey on 5G usage scenarios and traffic models. IEEE Communications Surveys & Tutorials. 2020;22(2):905-29.
[10] Kimura T, Kimura T, Matsumoto A, Yamagishi K. Balancing quality of experience and traffic volume in adaptive bitrate streaming. IEEE Access. 2021;9:15530-47.
[11] Le HD, Nguyen CT, Mai VV, Pham AT. On the throughput performance of tcp cubic in millimeter-wave cellular networks. IEEE Access. 2019;7:178618-30.
[12] Sim MS, Lim YG, Park SH, Dai L, Chae CB. Deep learning-based mmWave beam selection for 5G NR/6G with sub-6 GHz channel information: Algorithms and prototype validation. IEEE Access. 2020;8:51634-46.
[13] Spiteri K, Urgaonkar R, Sitaraman RK. BOLA: Near-optimal bitrate adaptation for online videos. IEEE/ACM Transactions on Networking. 2020;28(4):1698-711.
[14] Yin X, Jindal A, Sekar V, Sinopoli B. A control-theoretic approach for dynamic adaptive video streaming over HTTP. In: Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication; 2015. p. 325-38.
[15] Mao H, Netravali R, Alizadeh M. Neural adaptive video streaming with pensieve. In: Proceedings of the Conference of the ACM Special Interest Group on Data Communication; 2017. p. 197-210.
[16] Akhtar Z, Nam YS, Govindan R, Rao S, Chen J, Katz-Bassett E, et al. Oboe: Auto-tuning video ABR algorithms to network conditions. In: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication; 2018. p. 44-58.
[17] Quinlan JJ, Sreenan CJ. Multi-profile ultra high definition (UHD) AVC and HEVC 4K DASH datasets. In: Proceedings of the 9th ACM Multimedia Systems Conference; 2018. p. 375-80.
[18] Kaipio A, Ponomarenko M, Egiazarian K. Merging of MOS of Large Image Databases for No-reference Image Visual Quality Assessment. In: 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP). IEEE; 2020. p. 1-6.