A Context-aware QoE-driven Strategy for Adaptive Video Streaming in 5G multi-RAT Environments

Bouali, F., Moessner, K. & Fitch, M.

Original citation & hyperlink:
Bouali, F, Moessner, K & Fitch, M 2018, A Context-aware QoE-driven Strategy for Adaptive Video Streaming in 5G multi-RAT Environments. in 20th International Symposium on Wireless Personal Multimedia Communications (WPMC). IEEE, pp. 354-360, 20th International Symposium on Wireless Personal Multimedia Communications, Bali, Indonesia, 17/12/17.
https://dx.doi.org/ 10.1109/WPMC.2017.8301838

DOI 10.1109/WPMC.2017.8301838
ISBN 978-1-5386-2769-3
ISBN 978-1-5386-2768-6

Publisher: IEEE

A Context-aware QoE-driven Strategy for Adaptive Video Streaming in 5G multi-RAT Environments

F. Bouali, K. Moessner
5G Innovation Centre (5GIC)
University of Surrey, UK
Email: {f.bouali, k.moessner}@surrey.ac.uk

M. Fitch
BT Research
Adastral Park, Ipswich IP5 3RE, UK
Email: michael.fitch@bt.com

***Abstract*—This paper extends traditional dynamic adaptive streaming over HTTP (DASH) to efficiently exploit all available bands and licensing regimes in a given context. A novel objective quality-of-experience (QoE) metric is proposed to capture the most relevant factors that impact user perception during streaming sessions. Based on it, a QoE-driven adaptation strategy is devised to jointly select the best radio access technology (RAT) and quality for each video segment depending on the various components of the context. It relies first on fuzzy logic to estimate the QoE provided by each available RAT subject to the uncertainty level associated with DASH clients. Then, a fuzzy multiple attribute decision making (MADM) methodology is developed to combine the QoE estimates with the heterogeneous components of the context to assess the in-context suitability levels. The proposed approach is applied to adapt video streaming across available RATs in dense deployments for a set of *Bronze* and *Gold* subscriptions. The results reveal that the proposed strategy always assigns *Gold* clients to the well-regulated licensed band, while switches *Bronze* clients between licensed and unlicensed bands depending on the operating conditions. It strikes a balance between maximising video quality and reducing playback stalling, which significantly improves the perceived QoE compared to the traditional DASH approach.**

I. Context/Motivation

The traditional quality-of-service (QoS) metrics fall short in meeting the stringent requirements associated with emerging interactive applications (e.g., two-way gaming). As a result, the quality-of-experience (QoE) level perceived by the end-user is expected to become one of the most relevant metrics in the fifth-generation (5G) of wireless networks. In this respect, there has been an increasing trend towards incorporating QoE-driven mechanisms to assist, e.g., network management [1], scheduling [2], and content caching [3].

Most of these mechanisms have been developed to support the emerging dynamic adaptive streaming over HTTP (DASH) standard developed by the moving picture experts group (MPEG) [4]. DASH has become the preferred choice of most video providers (e.g., Youtube and Netflix), and has been recently adopted by the third generation partnership project (3GPP) [5]. The key concept behind DASH is that video sequences are split into segments that are encoded with different qualities (e.g., resolutions and bit-rates) and delivered over Transmission Control Protocol (TCP) transport from conventional web-servers. During streaming sessions, a DASH client requests, for each of these segments, the most appropriate quality depending on its local conditions (e.g., available bandwidth and buffer level). Given that the implementation details of such adaptation were left unspecified by the standard [4], lots of effort has been made to come up with the most efficient adaptation logic. Some proposals have focused on minimizing the likelihood of stalling [6, 7], while others have given more priority to maximize video quality [8, 9]. More interestingly, other strategies have been devised to strike a balance between the conflicting objectives of each of these approaches [10, 11].

A common shortcoming of all these proposals is that they all have been designed to exploit a single RAT, i.e., video quality is down-graded when the bandwidth of the in use RAT is reduced. However, high-quality segments could still have been delivered if all RATs typically available for current devices (e.g., cellular and Wi-fi) were exploited. The second key limitation is that the traditional DASH adaptation does not take into account various relevant components of the context (i.e., remaining credit, velocity, and battery level), which may hurt the QoE of some users, e.g., a user with limited credit would get an "out-of-credit" drop after starting a high-quality streaming session over licensed, but may be able to finish the session with lower quality over free unlicensed access. Finally, each of these proposals has considered different criteria (i.e., buffer level, available bandwidth or combination) to select the best video quality. Having a common QoE metric, based on which the adaptation is performed, would provide a common ground where all algorithms could co-exist.

The above discussion clearly calls for a novel DASH adaptation logic that exploits all available RATs depending on the operating conditions (e.g., interference and contention levels in licensed and unlicensed bands, respectively) and all relevant components of the context (e.g., remaining credit and battery level). To this end, this paper exploits the generic context-aware framework that was previously built in [12] and successfully instantiated to support voice-over-IP (VoIP) [12] and fixed-rate live video [13].

Motivated by the proven usefulness of the proposed framework, this paper extends it to support the more challenging case of adaptive buffered (i.e., DASH-like) video streaming. Specifically, it makes the following contributions:

- A novel objective QoE metric is proposed to jointly capture the most relevant factors that impact user perception during adaptive video streaming sessions,
- Based on this metric, a QoE-driven adaptation strategy is developed to dynamically select the best RAT and video quality in a given context. It relies on fuzzy logic to estimate the QoE provided by each RAT and MADM to efficiently combine the heterogeneous components of the context. To the best of our knowledge, jointly selecting the best RAT and quality to meet the QoE requirements of adaptive streaming has not been considered previously,
- Finally, the effectiveness of the proposed methodology in supporting video streaming is evaluated and benchmarked against the traditional DASH approach.

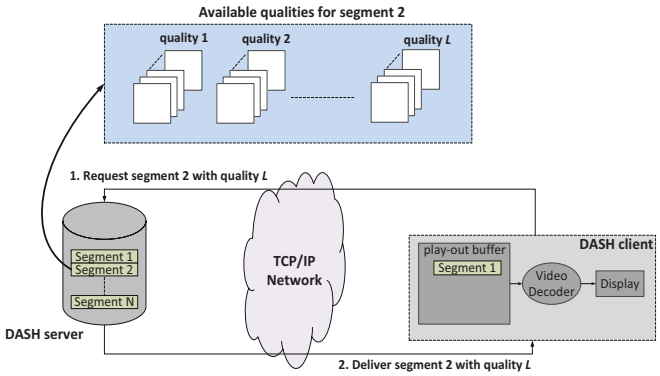


Fig. 1. Architecture for DASH video streaming.

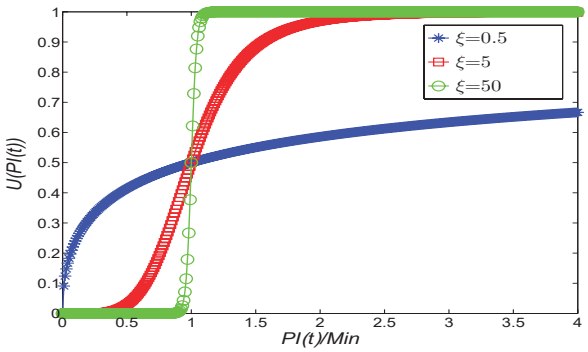


Fig. 2. Behavior of the considered utility function.

The remainder of this paper is organized as follows. A novel objective QoE metric for adaptive video streaming is proposed in Section II together with a functional architecture to facilitate its estimation. Then, a QoE-driven strategy is developed in Section III to perform a context-aware adaptation of video quality across all available RATs. The initial results are presented in Section IV to assess the effectiveness of the proposed methodology in supporting video streaming. The conclusions and future directions are provided in Section V.

II. Context-aware QoE-driven Framework for Adaptive Video Streaming in multi-RAT Environments

An ultra-dense environment is considered, where K available RATs ($\{RAT_k\}_{1\leq k\leq K}$) are exploited by N_{video} clients to establish DASH-like streaming sessions with a remote server. All clients are assumed to belong to one of S subscription profiles ($\{P_s\}_{1\leq s\leq S}$). To capture the constraints associated with various licensing regimes, K=2 RATs are considered, namely LTE and WLAN operating in licensed and unlicensed bands, respectively.

A. DASH model

During streaming sessions, DASH clients download over a TCP transport video sequences that are divided into multiple segments, each of duration T_S. To absorb the delay that may be introduced by TCP retransmissions, the various clients progressively download and buffer these segments during video playback as illustrated in Fig. 1. Furthermore, each segment is encoded into L different qualities defined in bit-rates for the sake of simplicity. Based on the experimental results conducted in [14], L=4 qualities are considered, namely {5,10,15,20} Mbps, where only 20 Mbps meets the requirements of 4K resolution.

In conventional DASH adaptation [6–11], clients dynamically select, for the j-th segment to be requested at a given time t, the most appropriate quality $Q_j(t)$ based on a sub-set of the following metrics:

- $Q(j-1)$: The quality of the previous segment,
- $B(t)$: The current level of the play-out buffer expressed in seconds of playback duration,
- $R_k(t)$: The perceived bit-rate on the in use RAT_k expressed in Mbps.

In contrast, it is proposed to jointly select the best RAT and quality of each video segment based on not only the QoE requirements, but also the various components of the context (e.g., remaining credit and velocity). Note that the video content to be downloaded is assumed to be cached in the neighborhood of each DASH client, and thus any available RAT could be used to send it. This means that switching from one RAT to another does not require to establish a new socket with the remote server.

B. Considered problem

Access to licensed bands (i.e., LTE) is assumed to be paid (i.e., consumes some units of the remaining credit), while access to unlicensed bands (i.e., WLAN) is free of charge. This means that adaptive video streaming needs to maximize usage of unlicensed bands as long as the associated QoE requirements are met. To reduce energy consumption of battery-operated clients, a single-homing setting is considered (i.e., each client can use only one RAT at any time).

To tackle the considered problem, a novel QoE metric is proposed in Section II-C and a functional architecture is presented in Section II-D to facilitate its estimation.

C. Proposed QoE metric

This section proposes a novel objective metric to assess the QoE perceived during adaptive video streaming. To this end, the following components are particularly considered:

- Video quality: the quality in Mbps of the video segments that are delivered during streaming sessions.
- Stalling: also referred to as re-buffering or freezing. It occurs in the middle of playback whenever the available bandwidth is not sufficient to sustain the selected video quality and the play-out buffer runs out of data.

In this respect, the utility function previously proposed in [15] is considered as a building block to assess a given performance indicator *PI* evaluated at a given time *t*:

$$U(PI(t))=\frac{(\frac{PI(t)}{Min})^{\xi}}{1+(\frac{PI(t)}{Min})^{\xi}} \quad (1)$$

where Min denotes the minimum performance requirement and ξ is a shaping parameter that captures different degrees of elasticity in meeting it.

To better analyse the behavior of the considered utility, Fig. 2 plots it as a function of the ratio $PI(t)/Min$ for different values of the shaping parameter ξ. It can be seen that $U(.)$ is a monotonic increasing function of the achievable performance $PI(t)$ that equals 0.5 at $PI(t)$=Min, and tends

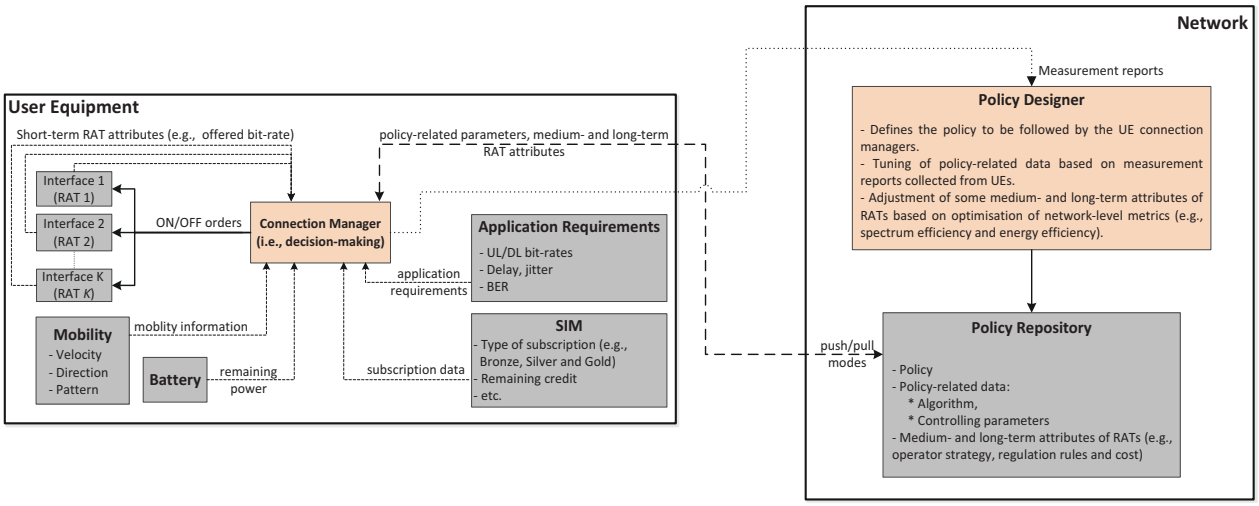


Fig. 3. Functional architecture for context-aware user-driven operation.

asymptotically to 1. The marginal increase of the utility function for large performances $PI(t)$ well above *Min* becomes progressively smaller especially when intermediate values of ξ are used (e.g., ξ=5). Therefore, $U(.)$ provides a measure of the suitability to support the requirement *Min*, with values ranging from 0 (low suitability) to 1 (high suitability).

Based on the above utility, the following metric is proposed to assess the QoE level provided by a given RAT_k at time *t*:

$$QoE_k(t)=U_B(t)\cdot U_R(t,k) \tag{2}$$

where the two multiplicative terms are defined as:

$$U_B(t)=U(B(t))\Big|_{Min=5\ s,\xi=5} \tag{3}$$

$$U_R(t,k)=U(R_k(t))\Big|_{Min=5\ Mbps,\xi=5} \tag{4}$$

According to the above formulations, $U_B(t)$ is penalized when the buffer level enters a risky zone (i.e., $B(t)<5\ s$), where the likelihood of stalling becomes high. On the other hand, $U_R(t,k)$ decreases gradually when the bit-rate is not sufficient to support the maximum quality (i.e., $R_k(t)<20\ Mbps$) and significantly when it does not sustain the minimum quality (i.e., $R_k(t)<5\ Mbps$). Therefore, $QoE_k(t)$ maximises video quality while minimizing video stalling. Note that the proposed metric could be extended to consider other relevant factors (e.g., start-up delay and quality switches), which is left for future consideration.

The evaluation of $QoE_k(t)$ depends on whether RAT_k is active or not at time *t*. When RAT_k is used, $QoE_k(t)$ is obtained by evaluating (2) at $R_k(t)=\overline{R_k}$, where $\overline{R_k}$ denotes the average bit-rate at which the latest N_{avg} segments have been received. In turn, when RAT_k is unused, $R_k(t)$ and $QoE_k(t)$ will be estimated according to the methodology developed in Section III-A.

D. Functional architecture

To enable QoE-driven operation, the generic context-aware framework previously proposed in [12] will be instantiated. According to its functional architecture described in Fig. 3, a connection manager (CM) is introduced at the user equipment (UE) to implement a given decision-making policy (e.g., the proposed strategy in Section III). To this end, it exploits the relevant components of the context available locally (e.g., velocity and battery level) and a radio characterisation of each available RAT in terms of a set of short-term attributes (e.g., current load) obtained through beacons and pilot channels. In particular, it is assumed that each RAT_k broadcasts the maximum bit-rate R_k^{max} that would be offered to a new user as an indicator of its current load. Additionally, the CM collects from the network side some medium- and long-term RAT attributes (e.g., cost and regulation rules) stored in a Policy Repository together with all the policy-related data. The content of the Policy Repository may be retrieved in practice from a local instance following a pull or push mode using e.g., the Open Mobile Alliance-Device Management (OMA-DM) protocol [16]. To offer higher flexibility to the network manager, a Policy Designer entity builds and updates the Policy Repository content based on measurement reports collected from the various UEs and some potential network-level constraints (e.g., operator strategy and regulation rules).

In the next section, the CM will be implemented to solve the problem considered in Section II-B, i.e., perform a context-aware QoE-driven adaptation of video streaming.

III. Connection Manager: Context-aware QoE-driven Adaptation Strategy

This section implements the CM of the functional architecture described in Fig. 3 to perform a context-aware QoE-driven adaptation of video streaming across all available RATs. To this end, the following three-step approach is proposed to select the best RAT and quality for each segment to be requested:

1) Design a fuzzy logic calculator to estimate the QoE level provided by each RAT_k (i.e., QoE_k) based on the local observations of each DASH client.
2) Develop a fuzzy MADM methodology to combine QoE_k with the UE subscription profile (i.e., P_s) and all other components of the context to derive the so-named "in-context" suitability level ($s_k^{ic,s}$).
3) Select the RAT that maximizes the in-context suitability level $s_k^{ic,s}$, and adapt the video quality accordingly.

In what follows, each of these steps will be implemented.

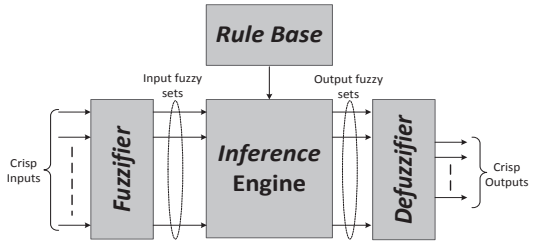


Fig. 4. Block diagram of the fuzzy logic controller.

A. Estimated QoE levels

Given the uncertainty associated with DASH clients, this section relies on fuzzy logic to estimate the QoE level that would be provided by each unused RAT_k (i.e., QoE_k).

The key building block in fuzzy logic reasoning is the fuzzy logic controller (FLC) whose block diagram is described in Fig. 4. It is composed of three main stages, namely the fuzzifier, inference engine and defuzzifier. During fuzzification, crisp (i.e., real) input data are assigned a value between 0 and 1 corresponding to the degree of membership in a given fuzzy set. Then, the inference engine executes a set of *if-then* rules on the input fuzzy sets. These inference rules are maintained in a rule base that is typically built based on previous expert knowledge. Finally, the aggregated output fuzzy sets are converted into crisp outputs using a given defuzzification method.

In the following, a separate FLC is designed to estimate the QoE provided by each of the considered RATs.

1) LTE: The designed FLC and corresponding membership functions are described in Fig. 5. In particular, the following input parameters are considered:

- *RSRQ*: the reference symbol received quality (RSRQ) that captures the radio and interference conditions.
- R_L^{max}: the maximum bit-rate that could be offered to a new user. It is advertised through one of the cell system information blocks (SIBs) as explained in Section II-D. Assuming a fair scheduler (e.g., proportional fair (PF)), the cell calculates it as $R_L^{max}=C_{max}/(N+1)$, where C_{max} and N denote the maximum cell capacity and number of active users, respectively.
- $B(t)$: the buffer level in seconds of playback at time t.

2) WLAN: The proposed FLC and associated membership functions are described in Fig. 6. Specifically, the set of input parameters is designed as follows:

- *SINR*: the signal to interference and noise ratio of the access point (AP) beacon.
- R_W^{max}: the maximum bit-rate offered to a new user. It is estimated based on the number of users, fraction of busy channel, slot duration and minimum contention window [17] and broadcasted on the AP beacon.
- $B(t)$: the buffer level in seconds of playback at time t.

For both FLCs, the inferences rules have been designed based on a sensitivity analysis to the various combinations of the input parameters, which is omitted for the sake of brevity. Finally, the defuzzification process is based on the commonly used centroid method for its accuracy [18].

B. In-context suitability levels

In this section, the previous QoE estimates are combined with the context at hand to assess the in-context suitability.

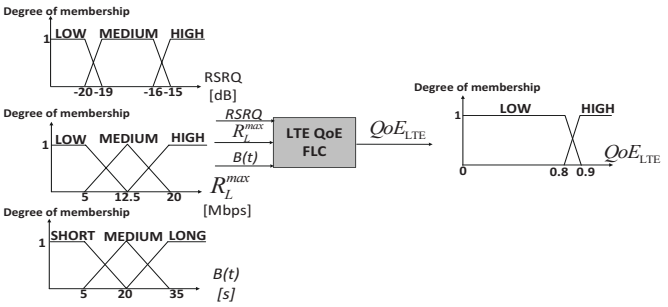


Fig. 5. FLC for estimating QoE_{LTE}.

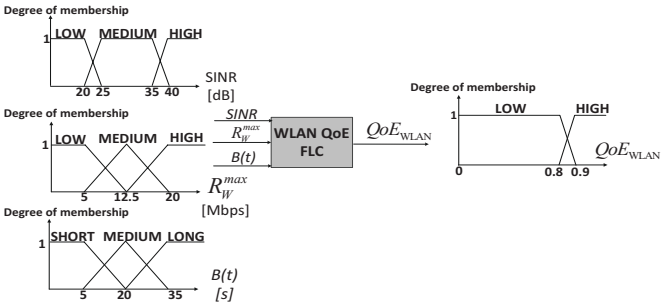


Fig. 6. FLC for estimating QoE_{WLAN}.

To cope with the heterogeneity of the context components, a methodology is developed based on MADM [19].

In this respect, for each $k\in\{1,\ldots,K\}$, RAT_k is characterized in terms of the following $M=4$ attributes:

- QoE_k: the QoE level provided by RAT_k defined in (2). Recall that, if RAT_k is used, it is directly evaluated based on the perceived bit-rate $\overline{R_k}$. Otherwise, it is estimated by the FLCs developed in the previous sub-section.
- $cost_k$: the monetary cost of RAT_k.
- $power_k$: the power consumption level when using RAT_k.
- $range_k$: an assessment of the range to reflect the appropriateness from the UE velocity perspective.

Therefore, the RATs can be fully characterized in terms of a $K\times M$ decision matrix $\mathbf{D}$ whose element $d_{k,m}$ denotes the performance of RAT_k in terms of the m-th attribute:

$$\mathbf{D}=\begin{matrix}LTE\\WLAN\end{matrix}\begin{bmatrix}\overset{QoE}{QoE_{LTE}} & \overset{cost}{HIGH} & \overset{power}{HIGH} & \overset{range}{LARGE}\\ QoE_{WLAN} & LOW & MEDIUM & SMALL\end{bmatrix} \quad (5)$$

Note that, compared to LTE, WLAN is qualified as cheaper, less power-consuming, and smaller.

To adjust the relative importance of the various attributes, a vector $\mathbf{w}^s$ of M weights ($\{w_m^s\}_{1\leq m\leq M}$) is introduced for each s-th subscription profile:

$$\mathbf{w^B}=\begin{bmatrix}HIGH\\HIGH\\LOW\\LOW\end{bmatrix} \quad (6)$$

$$\mathbf{w^G}=\begin{bmatrix}HIGH\\LOW\\LOW\\LOW\end{bmatrix} \quad (7)$$

where B and G stand for the *Bronze* and *Gold* subscription profiles, respectively.

Note that the cost attribute is judged as more relevant for the *Bronze* user (i.e., $w_{cost}^B=HIGH$), while the power and range attributes are not initially considered (i.e., $w_{power}^s=w_{range}^s=LOW$) for the sake of simplicity.

Finally, the vector $\mathbf{s^{ic,s}}$ of in-context suitability levels ($\{s_k^{ic,s}\}_{k\in\{1,\ldots,K\}}$) is obtained as follows:

$$\mathbf{s^{ic,s}} = \begin{bmatrix} s_1^{ic,s} \\ \vdots \\ s_k^{ic,s} \\ \vdots \\ s_K^{ic,s} \end{bmatrix} = \overline{\mathbf{D}} \cdot \mathbf{w^s} \quad (8)$$

where $\overline{\mathbf{D}}$ is the matrix of normalized attributes $\overline{d_{k,m}}$ that are calculated as $\overline{d_{k,m}}=d_{k,m}/\max_k(d_{k,m})$ for benefit attributes (i.e., QoE and range) and $\overline{d_{k,m}}=\min_k(d_{k,m})/d_{k,m}$ for cost attributes (i.e., cost and power).

C. Decision-making

This section proposes a context-aware QoE-driven strategy for adaptive video streaming. To assess its added-value to the existing works, one of the design requirements is to support the adaptation logic of any traditional DASH algorithm.

The proposed strategy is described by the pseudo code of Algorithm 1. Initially, the RAT_{k^*} that maximises the in-context suitability is selected (line 1). Recall that this metric combines the QoE level provided by each RAT with the context at hand. Next, the best video quality is adapted within the selected RAT depending on the local observations. To this end, the bit-rate $R_{k^*}(t)$ is estimated depending on whether RAT_{k^*} is currently used or not. If it is used, $R_{k^*}(t)$ is set to the average perceived bit-rate $\overline{R_{k^*}}$ (line 3). Otherwise, it is set to the maximum offered bit-rate $R_{k^*}^{max}$ advertised by the RAT (line 5). Finally, for the *j*-th segment to be requested at a given time *t*, the quality $Q_j(t)$ is determined based on the quality of the previous segment $Q(j-1)$, buffer level $B(t)$, and estimated bit-rate $R_{k^*}(t)$ (line 7). Note that the generic function $f(.,.,.)$ could implement the adaptation logic of any traditional DASH algorithm.

Algorithm 1 QoE-driven strategy for RAT/quality adaptation

1: Select the best RAT: $k^* = \arg\max_{k\in\{1,\ldots,K\}} \left(s_k^{ic,s}\right)$;
2: **if** RAT_{k^*} is used **then**
3: $\quad R_{k^*}(t)=\overline{R_{k^*}}$;
4: **else**
5: $\quad R_{k^*}(t)=R_{k^*}^{max}$;
6: **end if**
7: Select the video quality for the *j*-th segment:

$$Q_j(t)=f\left(Q(j-1), B(t), R_{k^*}(t)\right); \quad (9)$$

IV. Simulation results

To obtain an insight into the effectiveness of the proposed approach, a set of system-level simulations have been performed using the NS-3 simulator [20].

A. Considered environment

- To model ultra-dense deployments, one single LTE macro-cell operating in licensed mode and overlaid by a set of buildings is considered. Each building is structured according to the dual-stripe layout [21], i.e., as two stripes of rooms with a corridor in-between, which corresponds, e.g., to the set of stores inside a shopping mall.
- Two LTE and WLAN small-cells are dropped randomly inside each room. The LTE small-cells operate in licensed bands according to a co-channel configuration, while the WLAN APs operate in unlicensed bands. As an illustrative example, Fig. 7 describes the SINR map obtained in the licensed band when a building of two 20-room stripes is considered.
- The capacity of both LTE and WLAN is assumed to be well-provisioned to accommodate a large number of high-quality streaming sessions. In this respect, five component carriers (CCs), each of 20 MHz, are aggregated for LTE, while 802.11ac is assumed for WLAN with a maximum bandwidth of 160 MHz.
- To vary the load on WLAN, a set of background traffic flows are established between N_{back} pairs of nodes in an ad-hoc mode on the same channel used by the AP. These flows are transported over TCP for a duration of t_{back}.

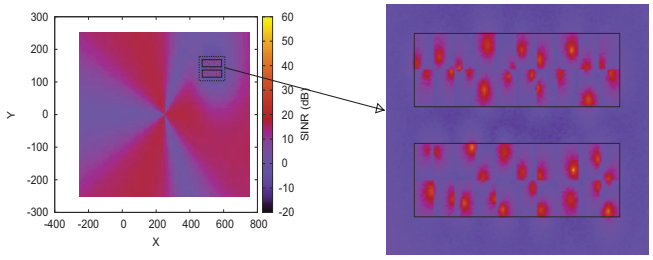


Fig. 7. Illustrative example of SINR map, licensed band.

B. Benchmarking

To benchmark the proposed strategy, the following schemes will be compared:

- *Fixed*: Only 4K resolution (i.e., 20 Mbps) is allowed on WLAN. This could be particularly relevant for clients who do not tolerate any down-grading in video quality.
- *DASH*: This represents the traditional DASH approach on WLAN and will be used as baseline to benchmark the performance of the next scheme. Without loss of generality, the algorithm proposed in [10] is selected.
- *Fuzzy MADM*: This is the strategy developed in Section III, where the adaptation logic function *f(.,.,.)* used in (9) is set to that of the previous scheme.

C. Key performance indicators

The following metrics are considered to assess the QoE level perceived during adaptive video streaming sessions:

- f_{4K}: the fraction of 4K (i.e., 20 Mbps) video segments out of the total number of delivered segments.
- S_{prob}: the probability that the video stalls, i.e., the playout buffer runs out of data. It is calculated as the fraction of stalling duration out of the total playback duration.
- $QoE(t)$: this is the instantaneous QoE level perceived by each DASH client regardless of the in use RAT. It is obtained by evaluating (2) at $R_k(t)=Q(j)$, where $Q(j)$ denotes the quality of the *j*-th segment.

D. Initial assumptions

To provide a proof of concept of the proposed approach, the following assumptions are initially considered:

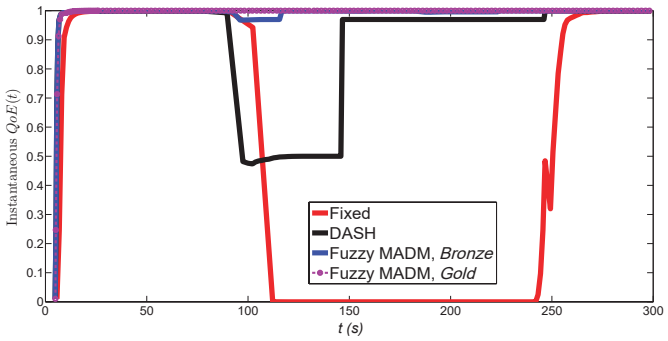


Fig. 8. Evolution of the instantaneous $QoE(t)$, 6^{th} client, t_{back}=3 min.

- A single-room scenario of the dual-stripe layout described in Section IV-A is initially considered.
- N_{video}=10 DASH sessions are established inside the same room together with N_{back}=3 background sessions. The various DASH clients may belong to *Gold* or *Bronze* subscriptions that are associated with an unlimited and limited credit of 4 Gbits, respectively.
- During a simulation time of T_{sim}=300 s, DASH clients continuously request segments, each of T_S=2 s.
- The bit-rate of the in use RAT is estimated over the latest N_{avg}=5 segments.

E. Performance evaluation

This section assesses the effectiveness of the proposed strategy in supporting adaptive video streaming. To this end, the metrics introduced in Section IV-C are evaluated.

Fig. 8 shows the evolution of the instantaneous QoE perceived by an arbitrary (i.e., 6^{th}) DASH client for each of the schemes considered in Section IV-B. Two simulation rounds were initially run where all DASH clients belong to the *Bronze* and *Gold* subscription profiles, respectively. For both rounds, the background sessions are maintained during t_{back}=3 min (i.e., from t=60 s to t=240 s). Note that the performance achieved by *Fixed* and *DASH* is shown only once at it does not depend on the subscription profile.

The observed behavior shows that the proposed strategy (i.e., *Fuzzy MADM*) introduces significant QoE gains. When only 4K resolution is used (i.e., *Fixed*), shortly after background sessions start (i.e., t=60 s), the capacity of the WLAN is no longer sufficient to meet the play-out deadline of the delivered segments. As a result, the play-out buffer runs out of data, and the QoE is strongly degraded due to a null first term in (2) (i.e., $U_B(t)$). In turn, when adaptation is performed within WLAN (i.e., *DASH*), lower qualities are selected for the subsequent segments, which avoids stalling, but degrades the QoE due to the reduction of the second term in (2) (i.e., $U_R(t,k)$). Note that the degradation persists much longer when no adaptation is performed. Unlike these two schemes, *Fuzzy MADM* treats the clients differently depending on the context at hand. For *Gold* clients, it exclusively assigns sessions to the well-regulated licensed band (i.e., LTE) as the cost is not of concern. For *Bronze* clients, it opportunistically exploits the unlicensed band (i.e., WLAN) as long as a good QoE could be sustained, which maintains the highest quality without stalling.

Next, the performance of the *Bronze* clients is further investigated. Fig. 9(a) shows the average QoE perceived by all clients as a function of the duration of background sessions. Fig. 9(b) shows the associated fraction of using WLAN.

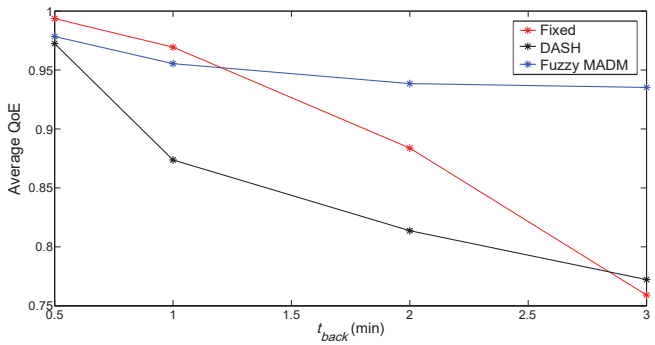
The results show that the QoE gain introduced by the proposed strategy significantly increases as WLAN gets loaded. For low loads (i.e., t_{back}=0.5 min), all schemes exhibit good QoE performance. This is because, at such loads, the play-out buffer is often sufficient to absorb the delay caused by the reduced WLAN bandwidth. As the load increases, all schemes gradually degrade as background sessions tend to occupy WLAN for longer periods. However, the degradation perceived by the proposed strategy remains marginal even at the highest loads (i.e., t_{back}=3 min). It can be seen from Fig. 9(a) and Fig. 9(b) that *Fuzzy MADM* reduces the fraction of using WLAN to the extent that sustains a good QoE level, which efficiently achieves the target behavior of Section II-B. For a better understanding of the contribution of each QoE component, Fig. 10(a) and Fig. 10(b) plot the corresponding fraction of 4K segments and stalling probability, respectively.

The first key observation in Fig. 10(a) is that when quality adaptation is performed within WLAN (i.e., *DASH*), the fraction of 4K segments decreases significantly compared to *Fixed* particularly for high loads. This is because whenever the offered WLAN capacity is reduced, *DASH* reduces the video quality to avoid buffer under-run events. In turn, *Fixed* keeps using the highest quality at the cost of a higher stalling probability (Fig. 10(b)). When both the in use RAT and video quality are adapted (i.e., *Fuzzy MADM*), the fraction of 4K segments significantly increases (Fig. 10(a)) without any degradation in terms of stalling (Fig. 10(b)).

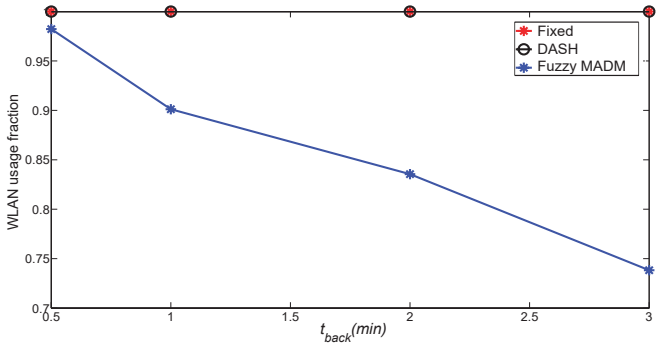
In summary, the proposed fuzzy MADM strategy enables to select the best RAT and video quality depending on the context at hand. For *Gold* users, it exclusively exploits the well-regulated licensed band as the cost is not of concern. For *Bronze* users, it additionally exploits unlicensed bands depending on its existing load. For all users, it strikes a balance between maximising video quality and minimising playback stalling, which significantly improves the overall QoE compared to the considered baselines.

V. Conclusions and Future Directions

This paper extends traditional DASH to efficiently exploit all available bands and licensing regimes in a given context. A novel QoE metric is proposed to maximize video quality subject to minimizing the playback stalling perceived by DASH clients. Based on this metric, a QoE-driven adaptation strategy is devised to jointly select the best RAT and video quality depending on the context at hand. It relies first on fuzzy logic to estimate the QoE level provided by each RAT subject to the lack of information associated with DASH clients. Then, a fuzzy MADM approach is developed to combine these estimates with the heterogeneous components of the context to assess the in-context suitability levels. The proposed strategy is validated in a dense small-cell environment, and its performance is benchmarked against two fixed-quality and DASH baselines. The results reveal that the proposed strategy always assigns *Gold* clients to the well-regulated licensed band, while switches *Bronze* clients between licensed and unlicensed bands depending on the operating conditions of each RAT. It strikes a balance between maximising video quality and minimising stalling, which significantly improves the perceived QoE compared to the considered baselines.

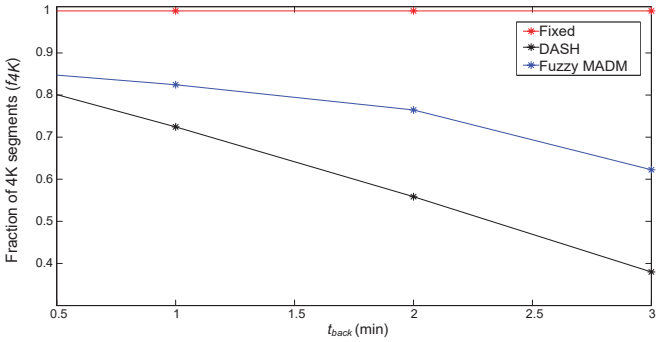


(a) Average QoE.

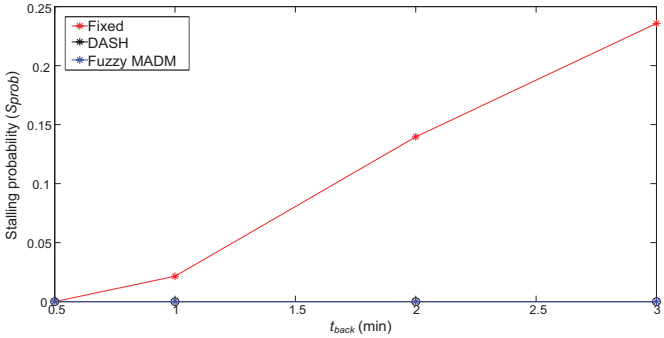


(b) WLAN usage fraction.

Fig. 9. Impact of WLAN load on the performance of *Bronze* clients in terms of



(a) Fraction of 4K segments.



(b) Stalling probability.

Fig. 10. Behavior of the QoE components, *Bronze* clients

As part of future work, it is intended to extend the proposed strategy to consider more QoE factors (e.g., start-up delay and quality switches), and showcase its benefits in a real-word environment (e.g., test-bed).

ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge the support of BT through the University of Surrey 5G Innovation Centre (http://www.surrey.ac.uk/5gic), and the EU funded H2020 5G-PPP project SPEED-5G under the grant agreement no. 671705.

REFERENCES

[1] S. Latré, P. Simoens, B. D. Vleeschauwer, W. V. de Meerssche, F. D. Turck, B. Dhoedt, P. Demeester, S. V. den Berghe, and E. G. de Lumley, "An autonomic architecture for optimizing QoE in multimedia access networks," *Computer Networks*, vol. 53, no. 10, pp. 1587 – 1602, 2009.

[2] J. Navarro-Ortiz, P. Ameigeiras, J. M. Lopez-Soler, J. Lorca-Hernando, Q. Perez-Tarrero, and R. Garcia-Perez, "A QoE-aware scheduler for HTTP progressive video in OFDMA systems," *IEEE Communications Letters*, vol. 17, no. 4, pp. 677–680, April 2013.

[3] C. Ge, N. Wang, S. Skillman, G. Foster, and Y. Cao, "QoE-driven DASH Video Caching and Adaptation at 5G Mobile Edge," in *Proceedings of the 3rd ACM Conference on Information-Centric Networking*, ser. ACM-ICN '16. New York, NY, USA: ACM, 2016, pp. 237–242.

[4] DASH, "MPEG-DASH specification," Tech. Rep. ISO/IEC 23009-1:2014.

[5] 3GPP, "Progressive download and dynamic adaptive streaming over HTTP (3GP-DASH), (Release 14)," Tech. Rep. TS 26.247-V14.1.0, March 2017.

[6] L. De Cicco, S. Mascolo, and V. Palmisano, "Feedback control for adaptive live video streaming," in *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, ser. MMSys '11. New York, NY, USA: ACM, 2011, pp. 145–156.

[7] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 187–198, Aug. 2014.

[8] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate adaptation for adaptive http streaming," in *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, ser. MMSys '11, New York, NY, USA, 2011, pp. 169–174.

[9] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, "Probe and adapt: Rate adaptation for http video streaming at scale," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 4, pp. 719–733, April 2014.

[10] K. Miller, E. Quacchio, G. Gennari, and A. Wolisz, "Adaptation algorithm for adaptive streaming over HTTP," in *2012 19th International Packet Video Workshop (PV)*, May 2012, pp. 173–178.

[11] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over http," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 325–338, Aug. 2015.

[12] F. Bouali, K. Moessner, and M. Fitch, "A context-aware user-driven framework for network selection in 5G multi-RAT environments," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, Sept 2016, pp. 1–7.

[13] ——, "A context-aware user-driven strategy to exploit offloading and sharing in ultra-dense deployments," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–7.

[14] S. H. Bae, J. Kim, M. Kim, S. Cho, and J. S. Choi, "Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services," *IEEE Transactions on Broadcasting*, vol. 59, no. 2, pp. 209–222, June 2013.

[15] F. Bouali, O. Sallent, J. Pérez-Romero, and R. Agustí, "A framework based on a fittingness factor to enable efficient exploitation of spectrum opportunities in cognitive radio networks," in *Wireless Personal Multimedia Communications (WPMC)*, Oct. 2011, pp. 1–5.

[16] Open Mobile Alliance, "OMA Device Management Protocol," Tech. Rep. version 1.2.1, June 2008.

[17] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, March 2000.

[18] T. J. Ross, *Fuzzy logic with engineering applications*. Chichester, U.K. John Wiley, 2010.

[19] C.-L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications A State-of-the-Art Survey*. Springer Berlin Heidelberg, 1981, ch. Methods for Multiple Attribute Decision Making, pp. 58–191.

[20] The network simulator-3 (NS-3). [Online]. Available: https://www.nsnam.org/

[21] 3GPP, "TSG RAN WG4 Meeting 51: Simulation assumptions and parameters for FDD HeNB RF requirements," Tech. Rep. R4-092042, May 2009.