Multi-timescale QoE Provisioning for Adaptive Video Streaming in Heterogeneous Deployments

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

***Abstract*—This paper considers an optimization problem that maximizes an aggregate utility, formulated as the weighted geometric mean of the "in-context" suitability of a set of radio access technologies (RATs) to support adaptive video streaming, subject to the existence of legacy data transfers. Motivated by the unfeasibility of solving the formulated problem centrally when the various RATs are loosely integrated (i.e., at core network (CN) level), a hybrid (i.e., network-assisted user-driven) strategy is devised to approximate its optimum solution. Unlike previous hybrid approaches, the proposed methodology exploits network assistance to ensure a friendly co-existence between adaptive video streaming clients and legacy users. It operates on different timescales, where the fastest timescale operation is performed on the video clients according to a policy that is tuned by the network on slower timescales. A user tuning on the fastest timescale (i.e., tens of ms) enables to adapt video streaming depending on the perceived quality-of-experience (QoE) and local components of the context (e.g., remaining credit and battery level). A small-cell tuning on a slower timescale (i.e., hundreds of ms) enables to preempt the resources used by legacy users based on the operating conditions (e.g., load and type of scheduler). Finally, a tuning performed by the network on the slowest timescale (i.e., few seconds) offloads legacy data transfers to unlicensed bands whenever the amount of interference on licensed bands reaches critical levels, which helps to sustain good QoE for all video clients. A cost-benefit analysis reveals that the proposed methodology performs closely to its centralized counterpart with much less control overhead on the radio interface.**

I. Introduction

HTTP-based adaptive video streaming has emerged as the technology of choice for most video providers (e.g., YouTube and Netflix). The associated standard, known as dynamic adaptive streaming over HTTP (DASH), was initially published in 2012 by the moving picture experts group (MPEG) [1], and since then has been adopted by many standardization and industry bodies, including the 3rd generation partnership project (3GPP) [2], digital video broadcasting (DVB) [3] and Hybrid broadcast broadband TV (HbbTV) [4]. Compared to traditional streaming solutions relying on dedicated media servers (e.g., real-time streaming protocol (RTSP) and real-time messaging protocol (RTMP)), DASH offers fundamental benefits, such as content being stored on regular HTTP servers, firewall pass-through and high scalability. At the same time, it introduces a set of challenges that still need to be overcome.

F. Bouali is with the 5G Innovation Centre (5GIC), University of Surrey, UK, E-mail: f.bouali@surrey.ac.uk.

K. Moessner is with the Professorship for Communications Engineering, Faculty for Electrical Engineering and Information Technology, Chemnitz University of Technology, Germany and the Institute for Communication Systems and 5G Innovation Centre (5GIC), at the University of Surrey, UK.

M. Fitch is with BT Research, Adastral Park, Ipswich IP5 3RE, UK.

A. Background information

The DASH standard is designed for client-pull based deployments primarily with HTTP protocol for media delivery. Video sequences are split into segments that are encoded with different qualities (e.g., resolutions and bit-rates) and delivered from conventional web-servers over a transmission control protocol (TCP) transport. During a DASH streaming session, a client first retrieves a manifest file, known as media presentation description (MPD), and then retrieves and renders content segments based on that metadata. For each of the retrieved segments, video quality adaptation is performed by selecting the most appropriate quality depending on the client local conditions (e.g., available bandwidth and buffer level).

Given that the implementation details of the DASH video quality adaptation were left unspecified by the standard [1], a significant effort has been made to come up with the most efficient adaptation logic assuming a single radio access technology (RAT) [5–7]. This means that, whenever the bandwidth of the in-use RAT is overloaded, low-quality segments are delivered to prevent buffer underruns and offer the end-user a continuous media experience. However, high-quality segments could still be delivered if all RATs, typically available for current devices (e.g., long-term evolution (LTE) and wireless local area network (WLAN)), are jointly used. The exploitation of different RATs with different characteristics (e.g., low-cost and limited-footprint for WLAN versus high-cost and extended-coverage for LTE) would bring flexibility and diversity in sustaining higher quality-of-experience (QoE) during adaptive video streaming sessions. As such, there has been an increasing interest in extending DASH with multi-RAT support.

B. Related work

To extend DASH with multi-RAT support, three directions can be explored, namely centralised, distributed and hybrid (i.e., network-assisted user-driven) approaches.

Multi-RAT inter-working has been traditionally achieved through centralised architectures interconnecting available cellular (e.g., wideband code division multiple access (WCDMA) and LTE) and non-cellular (e.g., WLANs) access networks [8–10]. In particular, the 3GPP has focused on a radio-level integration of the various RATs at different layers of the protocol stack e.g., medium access control (MAC) and packet data convergence protocol (PDCP) for licensed-assisted access (LAA) [11] and LTE-WLAN aggregation (LWA) [12], respectively. These works considered a common radio resource management (cRRM) based on a tight coupling between the various RATs, which is generally cumbersome in terms of

multi-RAT measurements and inter-working [13], or even unfeasible in many practical scenarios e.g., user-deployed WLAN access points (APs) and LTE femto-cells.

To overcome the above shortcomings, the various user equipments (UEs), where information about the various RATs is readily available (i.e., through beacons and pilot channels), could be exploited as a cost-effective distributed solution to achieve inter-working. As such, a user-driven decision-making would be much more suitable particularly when the performance strongly depends on the actual perception of the end-users (e.g., QoE for video traffic) and should be optimized rapidly (e.g., upon sudden degradation of radio conditions). To cope with the limited information typically available to UEs, various stochastic proposals that combine exploitation (i.e., using identified RATs) and exploration (i.e., trying unknown ones) have been developed [14–16]. These distributed schemes have been shown to achieve good individual performances, but cannot achieve network-wide efficiency due to their lack of global knowledge.

To strike a balance between the aforementioned extreme (i.e., fully centralised and distributed) options, various hybrid approaches have been proposed [17–19]. Most of these proposals focused on studying the theoretical aspects (e.g., convergence, equilibria and fairness) of their methodologies assuming simple traffic models (e.g., full buffer and file transfer). As such, they cannot capture the specificities associated with the emerging adaptive video streaming traffic. More recently, few hybrid schemes have been specifically designed for DASH-like traffic with a focus on assisting video adaptation across various video clients. In [20], a heuristic-based algorithm is proposed to jointly maximize the QoE and proportional fairness across mobile video streaming clients in multi-access edge computing (MEC) environments. The proposed algorithm performs better compared to purely client-based DASH heuristics, particularly when the achievable throughput is high or the wireless link characteristics of the mobile clients are comparable. In [21], an optimal slot-based resource allocation policy is proposed to minimize the probability of stalling of video streaming due to buffer underflows for a group of clients. The optimality of the proposed multi-user policy is established, and its usefulness is verified under realistic network conditions. In [22], a network-assisted adaptive video streaming methodology is proposed to enable simultaneous clients to select their bitrates in a coordinated fashion, which maximizes their QoE levels, achieve proportional fairness, and balance the load among video servers. These hybrid schemes have been proven to eliminate the fluctuations traditionally experienced by client-based adaptation. However, they all assumed a controlled setup of video clients without any legacy operation.

In practice, the legacy operation will be always around and the lack of coordination between DASH and legacy users may significantly impact their performance. As a matter of fact, the fundamental decentralised and client-driven nature of DASH traffic may be conflicting with the traditional network-driven traffic types (e.g., file transfer) and mechanisms (e.g., resource allocation and admission control). To cite a few illustrative scenarios, DASH clients may request high-quality segments based on the available bandwidth, but eventually get only a small fraction of it due to the network prioritisation of some legacy users. A given RAT may also be initially selected by various DASH clients due to its abundant bandwidth, but get shortly saturated by a high number of legacy users if no proper admission control is in place. Finally, service providers may not be able to guarantee a consistent or a premium quality of service (QoS) to some legacy users due to their lack of control over the individual behaviour of DASH clients.

The above discussion clearly calls for a form of network assistance to exploit the diversity offered by all available RATs to support adaptive video streaming, while regulating the behaviour of DASH clients to achieve a peaceful co-existence with all legacy users.

C. Contributions

As a first step in this direction, we have constructed in [23] a generic network-assisted user-driven framework, where the decision-making process is delegated to the end-user to select the best RAT depending on its local conditions. The methodology has been instantiated to support a variety of applications, including voice-over-IP (VoIP) [23], real-time video [24] and adaptive buffered video streaming [25]. It has been shown that, when the constructed framework is applied in a controlled small-scale setup, significant improvement can be achieved in the performance of all users.

The proposed user-driven operation may not be needed across larger-scale deployments and could even be incompatible with some use cases, where a centralized management would be more appropriate (e.g., interference reduction via coordinated multi-point (CoMP)). As such, the proposed framework would be better integrated into existing networks as an optional feature that is activated only when needed. This would result in a heterogenous deployment, where the uncontrollable legacy operation may hurt the performance of the proposed user-driven operation.

To cope with these challenges, this paper makes the following contributions:

- Formulate an optimization problem to maximize an aggregate utility that captures the "in-context" suitability of available RATs to support adaptive video streaming subject to the existence of a set of legacy data transfers. To the best of the authors' knowledge, no previous work has considered the practical problem of co-existence between adaptive video streaming clients and legacy users.
- Extend and instantiate the generic framework we previously constructed in [23] to solve the formulated problem when only a loose (i.e., core network (CN)) integration is available between the various RATs. First, a set of mechanisms are introduced to allow a multi-timescale operation, where the user-driven behaviour is regulated based on the network-level strategy and constraints. Then, the extended framework is mapped onto the 3GPP 5G architecture and instantiated as an add-on feature that can be activated only for the relevant use cases.

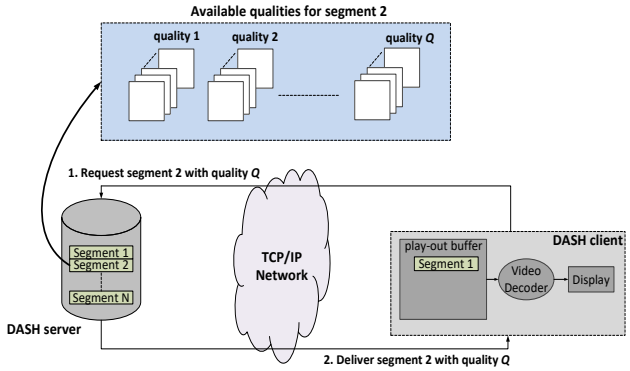

Fig. 1. Architecture for DASH video streaming.

- Based on the instantiated framework, devise a hierarchical QoE-driven methodology to adapt video streaming across all available RATs. It operates on different timescales, where the fastest timescale operation is performed on the video clients according to a policy that is tuned by the network on slower timescales. To the best of the authors' knowledge, multi-timescale operation has been previously considered only to split the functionalities on the network side [26–28]. In contrast, this paper introduces a specific timescale associated with the video clients to capture all the relevant factors affecting their perceived QoE levels, while ensuring fair co-existence with the legacy users.
- Conduct a cost-benefit analysis of the proposed approach versus its fully centralized and distributed counterparts. The performance is evaluated in terms of the individual and aggregated performances, while the cost is assessed in terms of the control overhead on the radio interface.

The remainder of this paper is organized as follows. The system model is described in Section. II, including a mixture of adaptive video streaming and legacy data transfer. An optimization problem is formulated in Section. III to maximize an aggregate utility that jointly captures the performances achieved by all users. A hierarchical network-assisted user-driven strategy is devised in Section. IV to efficiently solve the considered problem. The obtained results are presented in Section. V to benchmark the effectiveness of the proposed methodology against other reference schemes. The conclusions and future directions are provided in Section. VI.

II. System Model

A heterogeneous environment is considered, where a set of *K* available RATs ($\{RAT_k\}_{1\leq k\leq K}$) are exploited by a set of N_v video streaming clients subject to existence of a set of N_d legacy data transfers. The video streaming sessions are optimized based on the context (e.g., user velocity and battery level), while legacy data transfers are managed from a pure radio perspective. The various RATs are assumed to be loosely integrated (i.e., at CN-level) in cases where a tight integration would be too costly (e.g., user-deployed APs and femto-cells).

In what follows, each of the considered applications will be modeled together with their associated performances.

A. Adaptive video streaming

To capture the latest progress in adaptive video streaming, the emerging DASH standard [1] is considered.

1) Model: The considered video sequences are split into multiple segments, each of duration T_S, that are encoded with different qualities (e.g., resolutions and bit-rates). As illustrated in Fig. 1, DASH clients progressively download these segments over a TCP transport from conventional web-servers. The downloaded segments are buffered during video playback to absorb the delay that may be introduced by TCP retransmissions. Based on the experimental study conducted in [29], each segment is encoded into Q=4 qualities, namely $\{5,10,15,20\,\mathrm{Mbps}\}$, where only $20\,\mathrm{Mbps}$ can meet the requirements associated with 4K resolution.

In conventional DASH [5–7], the video quality is adapted within a given RAT. As such, each *n*-th video client selects, for the *j*-th segment to be requested at time *t*, the best quality $Q_n(j(t))$ based on a sub-set of the following metrics:

- $Q_n(j-1)$: The quality of the previous segment,
- $B_{v,n}(t)$: The current level of the play-out buffer expressed in seconds of playback duration,
- $R_{v,n,k}(t)$: The achievable bit-rate on the in-use RAT_k in Mbps.

In contrast, it is proposed to jointly select the best RAT and quality of each video segment based on the various components of the context (e.g., QoE level, remaining credit and velocity). Note that the video content is assumed to be cached in the neighborhood of the DASH client, and thus could be delivered by any available RAT without the need to establish a new socket with the remote server.

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

- Stalling: also referred to as re-buffering. It occurs whenever the offered bandwidth cannot sustain the selected video quality and the play-out buffer runs out of data.
- Video quality: the quality in Mbps of the video segments that are delivered during streaming sessions.

In this respect, the following utility is considered as a building block to assess the degree of fulfillment of a given requirement [30]:

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

where *PI(t)* denotes the achieved performance at a given time *t* and Min is the associated minimum requirement. The controllable shaping parameter ξ allows to capture different degrees of elasticity in meeting the considered requirement.

To better analyze the behavior of the considered utility, Fig. 2 plots it as a function of the ratio $PI(t)/Min$ for different values of ξ. It can be seen that the proposed formulation is a monotonic increasing function that equals 0.5 at $PI(t)$=Min, and tends asymptotically to 1. Its marginal increase for large performances $PI(t)$ well above the requirement *Min* becomes progressively smaller, especially when

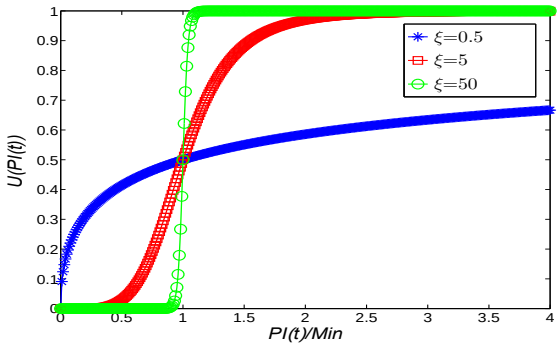

Fig. 2. Behavior of the considered utility function.

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).

It is worth pointing out that, for a given application, the controlling parameter ξ could be set in practice by the service provider based on some well-known traffic classes. For instance, in Fig. 2, ξ=50 could provide a good approximation of the step function associated with inelastic traffic (e.g., VoIP), ξ=5 could be used for slightly elastic traffic (e.g., adaptive video streaming) and ξ=0.5 would better characterize fully elastic traffic (e.g., data transfer).

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

$$QoE_{v,n,k}(t)=U^B_{v,n}(t)\cdot U^R_{v,n,k}(t) \tag{2}$$

where the two multiplicative terms are defined as

$$\begin{aligned} U^B_{v,n}(t) &= U(B_{v,n}(t))\Big|_{Min=B_v^{min}=5\,\mathrm{s},\xi=5} \\ &= \frac{\left(\frac{B_{v,n}(t)}{B_v^{min}}\right)^5}{1+\left(\frac{B_{v,n}(t)}{B_v^{min}}\right)^5} \end{aligned} \tag{3}$$

$$\begin{aligned} U^R_{v,n,k}(t) &= U(R_{v,n,k}(t))\Big|_{Min=R_v^{min}=5\,\mathrm{Mbps},\xi=5} \\ &= \frac{\left(\frac{R_{v,n,k}(t)}{R_v^{min}}\right)^5}{1+\left(\frac{R_{v,n,k}(t)}{R_v^{min}}\right)^5} \end{aligned} \tag{4}$$

The formulations in (3) and (4) assess the degree of fulfillment of the minimum buffer (i.e., B_v^{min}) and video quality (i.e., R_v^{min}) requirements. The shaping parameters have been set to ξ=5 to capture the partial elasticity associated with adaptive video streaming.

Based on these considerations, the first and second terms in (2) are penalized when the stalling likelihood is high (i.e., $B_{v,n}(t)<B_v^{min}$) and the minimum quality cannot be sustained (i.e., $R_{v,n,k}(t)<R_v^{min}$), respectively. Therefore, $QoE_{v,n,k}(t)$ maximizes video quality while minimizing stalling.

3) "In-context" suitability: This section proposes to combine the QoE level with the various components of the context.

To this end, L is introduced to denote the number of attributes that characterize the achievable performance by the *n*-th video client on each available RAT_k. Without loss of generality, the following attributes are considered (i.e., L=4):

- $QoE_{v,n,k}(t)$: the achievable QoE level given by (2).
- $C_{v,n,k}(t)$: the consumable credit in the next ΔT if RAT_k is selected:

$$C_{v,n,k}(t)=R_{v,n,k}(t)\cdot\Delta T \tag{5}$$

- $D_{v,n,k}(t)$: the dwell likelihood that the client will remain in the coverage area of RAT_k. It depends on the position of the client, its velocity $v_n(t)$ and range of RAT_k.
- $E_{v,n,k}(t)$: the consumable energy in the next ΔT if RAT_k is selected:

$$E_{v,n,k}(t)=E_{off-on}+P_{v,n,k}(t)\cdot\Delta T \tag{6}$$

where E_{off-on} is a state-dependent variable that models the energy overhead needed to turn on the RAT radio (i.e., set to zero when the RAT radio is already on at time *t*).

The above expression is composed of two terms. The first evaluates the energy loss to turn on the RAT radio. The second assesses the consumable energy on RAT_k in the next ΔT, where $P_{v,n,k}(t)$ denotes the associated power (W) estimated as a linear fit of the achievable bit-rate [31].

$$P_{v,n,k}(t)=\alpha_k\cdot R_{v,n,k}(t)+\beta_k \tag{7}$$

where α_k and β_k are the associated linear fit coefficients. Note that only the communication component is considered in (6). The other components (e.g., screen) are assumed to be independent from the in-use RAT.

Next, the so-named "in-context" suitability level is defined as a weighted sum of the various attributes:

$$\begin{aligned} s^{ic}_{v,n,k}(t)=&w_{v,n,Q}(t)\cdot QoE_{v,n,k}(t)+w_{v,n,C}(t)\cdot C_{v,n,k}(t) \\ &+w_{v,n,D}(t)\cdot D_{v,n,k}(t)+w_{v,n,E}(t)\cdot E_{v,n,k}(t) \end{aligned} \tag{8}$$

where $w_{v,n,Q}(t)$, $w_{v,n,C}(t)$, $w_{v,n,D}(t)$ and $w_{v,n,E}(t)$ denote the weights associated with the QoE, credit, dwell and energy attributes, respectively.

B. Legacy data transfer

Legacy data transfers are associated with a loose bit-rate requirement of R_d^{min}=1 Kbps. Their performance is assessed in terms of the "out-of-context" suitability level defined as

$$\begin{aligned} s^{oc}_{d,n,k}(t) &= U(R_{d,n,k}(t))\Big|_{Min=R_d^{min},\xi=0.5} \\ &= \frac{\left(\frac{R_{d,n,k}(t)}{R_d^{min}}\right)^{0.5}}{1+\left(\frac{R_{d,n,k}(t)}{R_d^{min}}\right)^{0.5}} \end{aligned} \tag{9}$$

where $R_{d,n,k}(t)$ denotes the achievable bit-rate by the *n*-th legacy data transfer using RAT_k expressed in Mbps.

The above formulation assesses the degree of fulfillment of the requirement R_d^{min}. The shaping parameter has been set to ξ=0.5 to reflect the fully elasticity (i.e., loose bitrate requirement) of the associated traffic.

III. Optimization problem

This section considers to adapt video streaming across all available RATs subject to the legacy activity.

A. Problem formulation

To reflect the assignments that may be performed at a given time *t*, let $\boldsymbol{x(t)}=\{x_{m,n,k}(t)\}$ denote the vector of assignment indicators of all users, where for each $m\in\{d,v\}$ and $n\in\{1,\cdots,N_m\}$, $x_{m,n,k}(t)$ is a binary indicator that takes the value 1 if RAT_k is assigned to the *n*-th user, and 0 otherwise.

Based on the above considerations, the optimum RAT assignment can be achieved by solving the following problem:

$$\begin{aligned}
\underset{\boldsymbol{x(t)}}{\text{maximize}} \quad & f\left(\boldsymbol{x(t)}\right) \\
\text{subject to} \quad & \sum_{m\in\{d,v\}}\sum_{n=1}^{N_m}\left(x_{m,n,k}(t)\cdot R_{m,n,k}(t)\right)\leq R_k^{tot},\ \forall k. \\
& \sum_{k=1}^{K}\left(x_{v,n,k}(t)\cdot C_{v,n,k}(t)\right)\leq C_{v,n}^{av}(t),\ \forall n\in\{1,\cdots,N_v\}. \\
& \sum_{k=1}^{K}\left(x_{v,n,k}(t)\cdot E_{v,n,k}(t)\right)\leq E_{v,n}^{av}(t),\ \forall n\in\{1,\cdots,N_v\}. \\
& \sum_{k=1}^{K} x_{m,n,k}(t)\leq 1,\ \forall m\in\{d,v\},\ \forall n\in\{1,\cdots,N_m\}. \\
& x_{m,n,k}(t)\in\{0,1\},\forall m,\ \forall n,\ \forall k.
\end{aligned} \tag{10}$$

where R_k^{tot} is the total capacity of RAT_k, while $C_{v,n}^{av}(t)$ and $E_{v,n}^{av}(t)$ denote the available credit and energy level of the *n*-th video client at time *t*, respectively.

The first constraint in (10) ensures that the aggregated bit-rate of video clients and legacy data transfers does not exceed the total RAT capacity (i.e, R_k^{tot}). The second and third constraints ensure that the data volume to be downloaded and energy to be consumed by the *n*-th video client do not exceed the available credit and energy level, respectively. The forth constraint reflects a single-homing configuration (i.e., each user can use only one RAT at any time). This is particularly relevant to reduce the energy consumption of battery-powered video clients.

The objective function $f(\cdot)$ is defined in (11) as the weighted geometric mean of all user performances, where the first and second double-product operators assess the performance associated with video clients and legacy data transfers, respectively. It maximizes proportional fairness (i.e., product of all user contributions) and inherently gives priority to the most-demanding users with a possible further tuning of the various priorities via the controlling weights $\{\delta_{m,n}\}$.

Note that the considered problem assumes that the scheduling strategies within each RAT are already enforced and cannot be influenced, which is particularly relevant for the considered scenario of loosely-integrated RATs (e.g. user-deployed APs and femto-cells).

B. Problem analysis

As it is formulated, the problem in (10) is complex to solve due to the following challenges:

- Due to the loose (i.e., CN-level) integration of the various RATs, the centralized approach that jointly determines and enforces the best RAT assignment for the various users is not feasible and is therefore ruled out.
- Due to the existence of legacy users, a pure distributed approach where each video client selects its best RAT would result in a sub-optimum performance as traditionally experienced by fully distributed schemes [14–16].
- The various attributes and their associated weights in (8) are not easily quantifiable, which calls for a reliable strategy to estimate them.

Enlightened by the above analysis, a network-assisted user-driven methodology will be devised to efficiently solve the considered problem. The effectiveness of the proposed solution will be later benchmarked in Section. V against its fully centralized and distributed counterparts.

IV. Multi-timescale QoE Provisioning for Adaptive Video Streaming in Heterogeneous Deployments

To enable network-assisted user-driven operation, a functional split between the user and network domains should be first made to identify the logical entities on each side.

A. Functional architecture

The functional architecture described in Fig. 3 is considered [23]. Specifically, a connection manager (CM) is introduced at the UE to implement a given decision-making policy (e.g., strategy of Section. IV-C). The CM exploits the relevant components of the context available locally (e.g., velocity and battery level) and a short-term characterization of each available RAT obtained through 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 indicator of its load. As the target devices are battery-powered, it is assumed that only one RAT interface is kept active, while all the others are kept in a power-efficient state (e.g., discontinuous reception (DRX) [32]). The inactive RAT interfaces could be activated on an event-triggered basis (e.g., upon performance degradation). Additionally, the CM collects from the network a set of medium- and long-term RAT attributes (e.g., cost) stored in a policy repository (PR) together with all the policy-related data. The content of the PR 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 [33]. To offer higher flexibility, a policy designer (PD) builds and updates the content of the PR based on call detail records (CDRs) collected from the various UEs.

B. Estimation strategy

This section proposes to estimate the various "in-context" suitability levels based on the following methodology:

1) Design a fuzzy logic controller (FLC) to estimate the QoE level provided by each RAT_k (i.e., $\widehat{QoE}_{v,n,k}(t)$).
2) Develop a fuzzy multiple attribute decision making (MADM) methodology to estimate the "in-context"

$$f\left(\boldsymbol{x(t)}\right)=\left(\prod_{n=1}^{N_v}\prod_{k=1}^{K}\left(s^{ic}_{v,n,k}(t)\right)^{\delta_{v,n}\cdot x_{v,n,k}(t)}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\left(s^{oc}_{d,n,k}(t)\right)^{\delta_{d,n}\cdot x_{d,n,k}(t)}\right)^{\frac{1}{\sum_{m\in\{d,v\}}\sum_{n=1}^{N_m}\delta_{m,n}}} \tag{11}$$

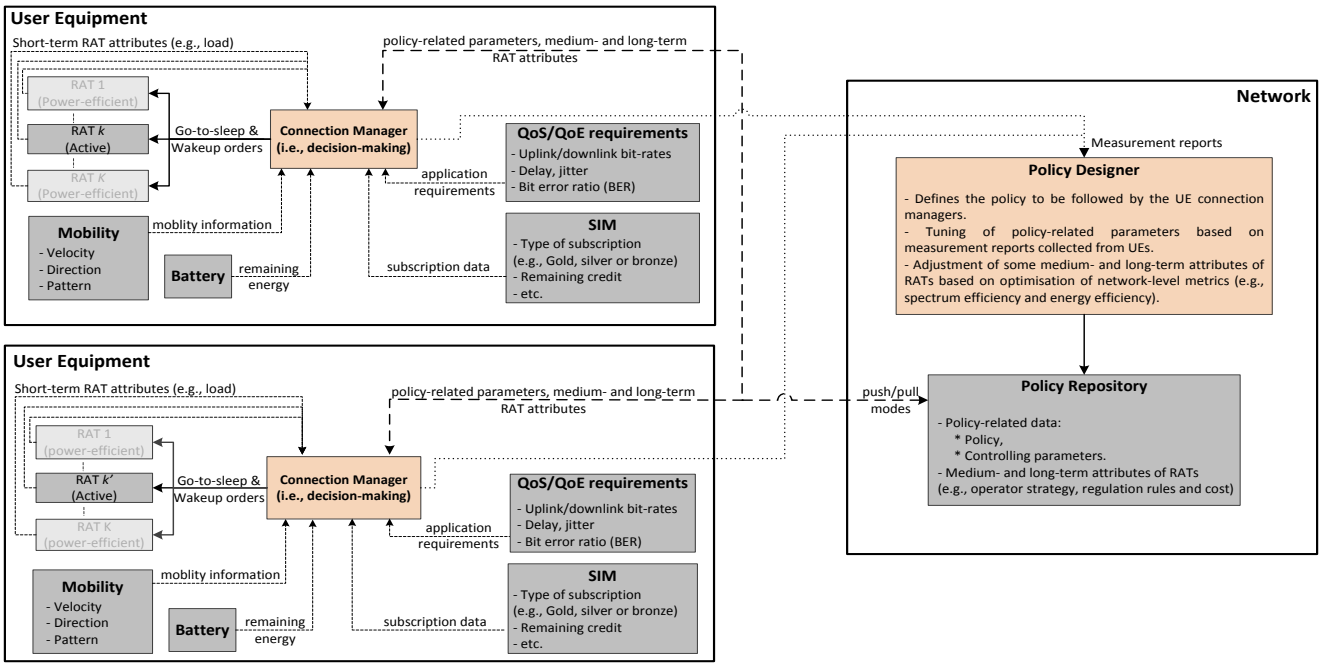

Fig. 3. Functional architecture of the proposed network-assisted user-driven framework.

suitability levels (i.e., $\{\widehat{s}^{ic}_{v,n,k}(t)\}_{1\leq k\leq K}$). This is achieved by defining and combining a set of linguistic characterizations for each of the attributes and weights in (8).

The proposed methodology combines two relevant tools, namely fuzzy logic to cope with the uncertainty level associated with DASH clients and MADM to efficiently combine the heterogeneous components of the context. The reader is referred to our previous works in [23] and [25] for a detailed description of the considered fuzzy MADM estimation methodology and its specific application to adaptive video streaming, respectively.

C. Network-assisted QoE-driven adaptation strategy

This section exploits the estimates of the previous section to adapt video streaming across all available RATs.

In this respect, the CM of Fig. 3 is implemented based on the pseudo code of Algorithm 1. Initially, the best RAT that maximizes the "in-context" suitability (i.e., $RAT_{k^*(n)}$) is selected. Next, the video quality is adapted within the selected RAT based on the local observations. To this end, the achievable bit-rate is estimated depending on whether the selected RAT is in-use or not. If it is used, it is set to the average bit-rate $\overline{R_{v,n,k^*(n)}}$ perceived over the latest N_{avg} segments (line 3). Otherwise, it is set to the maximum bit-rate advertised by the RAT (line 5). Finally, the quality of the next segment is determined based on the quality of the previous segment, buffer level, and estimated bit-rate (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: $k^*(n)=\arg\max_{k\in\{1,\ldots,K\}}\left(\widehat{s}^{ic}_{v,n,k}(t)\right)$;
2: **if** $RAT_{k^*(n)}$ is used **then**
3: $\quad R_{v,n,k^*(n)}(t)=\overline{R_{v,n,k^*(n)}}$;
4: **else**
5: $\quad R_{v,n,k^*(n)}(t)=R^{max}_{k^*(n)}$;
6: **end if**
7: Select the video quality for the j-th segment:

$$Q_n(j(t))=f\left(Q_n(j-1),B_{v,n}(t),R_{v,n,k^*(n)}(t)\right); \tag{12}$$

D. User tuning

This section proposes to tune the controlling parameters of the proposed strategy based on the local user conditions.

In this respect, the linguistic characterizations of the MADM weights used in Section. IV-B are dynamically tuned based on the strategy described in Table I. For instance, if the user is moving at low speed (i.e., $v_n(t)\leq D^1_{thr}$), the likelihood that it stays in the same cell is quite high, so the dwell likelihood attribute is given less importance by setting $w_{v,n,D}(t)=LOW$. However, if the user is moving at high speed (i.e., $v_n(t)\geq D^2_{thr}$), the likelihood that it stays in the same cell is quite low, and therefore the dwell likelihood attribute is given higher importance (i.e., $w_{v,n,D}(t)=HIGH$). In turn, $w_{v,n,C}(t)$ and $w_{v,n,E}(t)$ are set to HIGH whenever the available credit and battery level fall below C^1_{thr} and E^1_{thr}, respectively. Note that the QoE weight (i.e., $w_{v,n,Q}(t)$) is always set to HIGH to meet the video streaming requirements at all times.

TABLE I
ADJUSTMENT OF THE CONTROLLING WEIGHTS

Weight	Relevant parameter	Unit	Range	Value
$w_{v,n,C}(t)$	Available credit ($C^{av}_{v,n}(t)$)	Bits	$\leq C^1_{thr}$ $\in \left]C^1_{thr}, \cdots, C^2_{thr}\right[$ $\geq C^2_{thr}$	HIGH MEDIUM LOW
$w_{v,n,D}(t)$	Velocity ($v_n(t)$)	Km/h	$\leq D^1_{thr}$ $\in \left]D^1_{thr}, \cdots, D^2_{thr}\right[$ $\geq D^2_{thr}$	LOW MEDIUM HIGH
$w_{v,n,E}(t)$	Available energy ($E^{av}_{v,n}(t)$)	J	$\leq E^1_{thr}$ $\in \left]E^1_{thr}, \cdots, E^2_{thr}\right[$ $\geq E^2_{thr}$	HIGH MEDIUM LOW

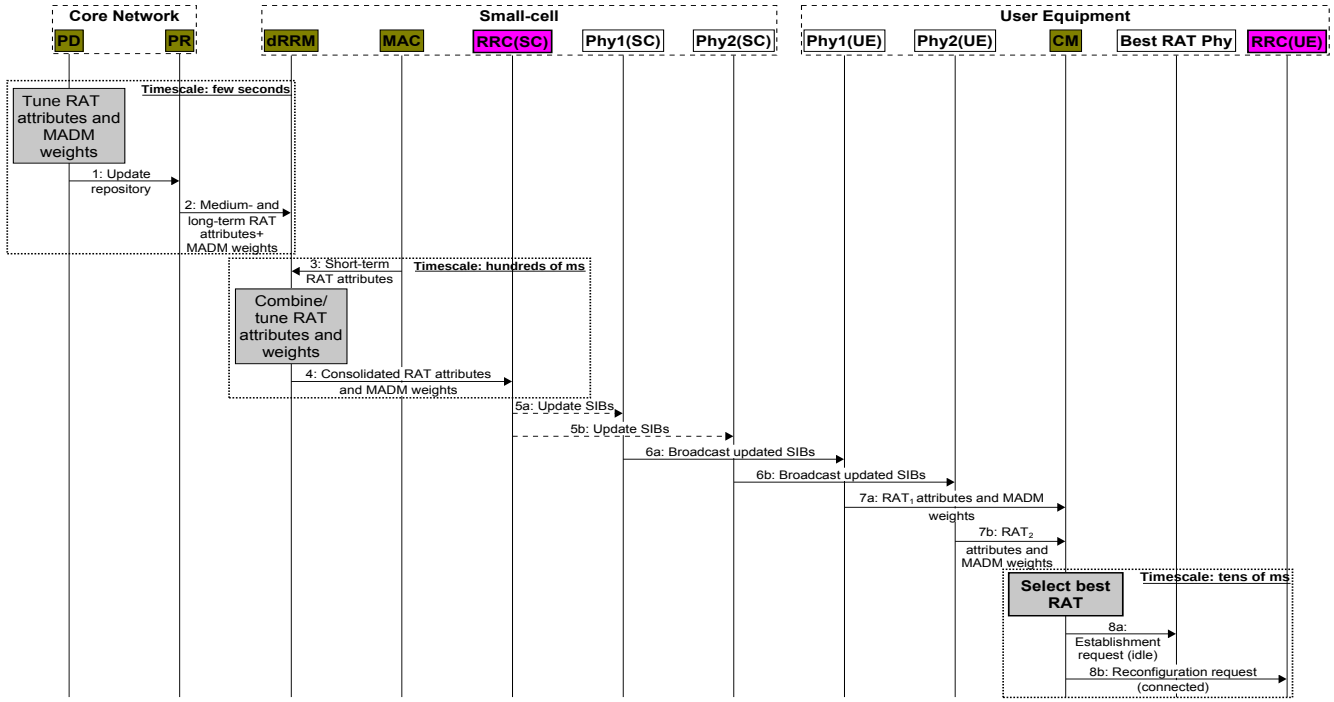

Fig. 4. Exchanged signaling messages with their associated timescales.

E. Multi-timescale operation

This section regulates the proposed QoE-driven operation on slower timescales based on the legacy activity.

Fig. 4 describes the signaling messages exchanged between the entities of Fig. 3 with their associated timescales.

To achieve network-wide efficiency, the PD may adjust some of the medium- and long-term RAT attributes (e.g., credit and energy) together with their associated weights on a relatively long timescale (i.e., few seconds) based on the legacy activity. The adjusted parameters are then passed to the distributed radio resource management (dRRM) agents of the various small-cells (SCs). Each agent combines the received information with a set of short-term RAT attributes (e.g. load) obtained from the MAC entity with a possible tuning on a slower timescale (i.e., hundreds of milliseconds) depending on the local conditions (e.g., load and type of scheduler). The consolidated set of RAT attributes and MADM weights is then communicated to the radio resource control (RRC) entity that accommodates it in the broadcasted system information blocks (SIBs). Finally, the CM extracts the broadcasted information from each of the physical (Phy) interfaces and combines it with the local context components to determine the best RAT on the fastest timescale (i.e., tens of ms).

F. Mapping onto the 5G architecture

This section integrates the proposed network-assisted user-driven framework into the 5G architecture as an add-on feature activated only for the relevant use cases.

Fig. 5 maps the functional modules of Fig. 3 onto the latest 3GPP architecture [34]. Specifically, the network-side functional entities (i.e., PD and PR) are mapped to the relevant CN modules, while the decision-making entity (i.e., CM) is placed in the UE. Compared to the default 3GPP setting, this mode of operation bypasses the SCs in terms of intelligence and delegates the final decision to the UE. The controlling parameters of the user-driven policy are adjusted by the policy control function (PCF) and session management function (SMF) and communicated to the SCs via the access management function (AMF). It is worth pointing out that, from the signalling perspective, the proposed extension would require the exchange of the messages shown in the message sequence chart of Fig. 4 on top of the standard 3GPP message flow.

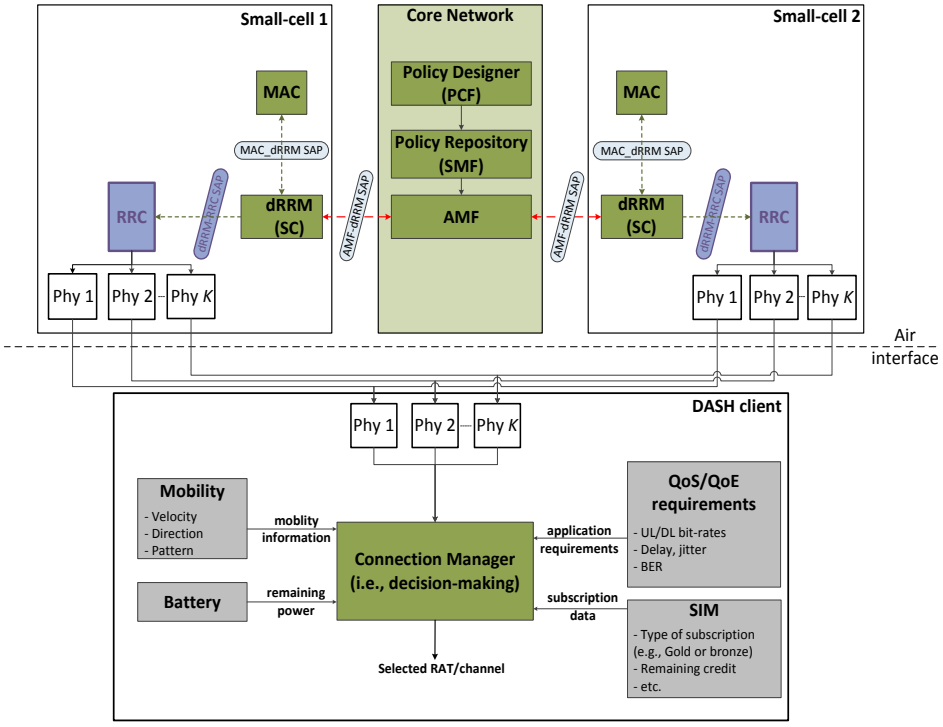

Fig. 5. Mapping onto the 3GPP 5G architecture.

It is worth pointing out that there is an on-going 3GPP work item (WI) on access traffic steering, switch and splitting (ATSSS) between 3GPP and non-3GPP accesses based on a policy framework, where various ATSSS agents are introduced on the UE and CN sides to assist decision-making on the network side [35]. A contribution to this WI based on our framework is being prepared to support user-driven operation.

Finally, the proposed add-on feature (i.e., network-assisted user-driven operation) could be particularly facilitated by the network slicing paradigm, which creates a set of logical network instances (i.e., slices) on top of the same physical infrastructure. Based on its softwarisation enablers (e.g., software-defined networking (SDN) and network function virtualisation or (NFV)), the proposed CM entity could be virtualised and implemented as a virtual network function (VNF) just for the slices where it is mostly needed, which is left for future consideration.

V. Simulation Results

To obtain an insight into the effectiveness of the proposed approach, a set of extensive system-level simulations have been carried out using the NS-3 simulator [36].

A. Considered environment

- An hexagonal setting of LTE macro-cells (MCs) overlaid by a set of buildings is considered. Each building is structured according to the dual-stripe layout [37], i.e., as two stripes of rooms with a corridor in-between. The various propagation losses in the presence of buildings are modeled using the hybrid building model [38].
- Two LTE and WLAN SCs are dropped randomly inside each room (i.e., K=2). The LTE cells (i.e., MCs and SCs) operate in licensed bands according to a co-channel configuration, while the WLAN APs operate in unlicensed bands. As an illustrative example, Fig. 6 describes the signal-to-interference-and-noise-ratio (SINR) map of the licensed band, where a building of two 20-room stripes is dropped on top of an hexagonal layout of 27 LTE MCs.
- The legacy activity inside each room is modeled as two sets of N_d^L and N_d^W data transfers established on licensed and unlicensed bands, respectively. The licensed users are managed by LTE is a traditional way, while the unlicensed sessions are established between pairs of

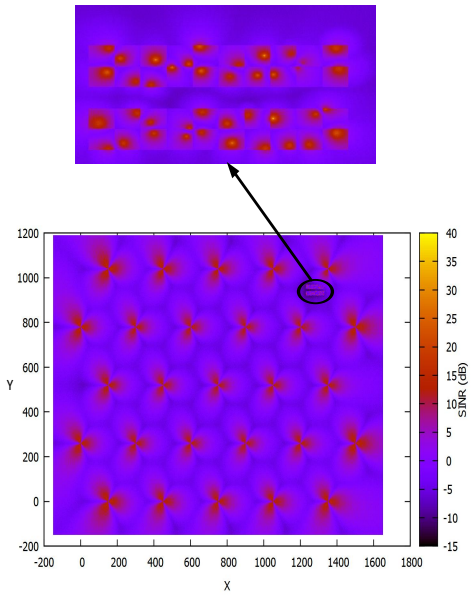

Fig. 6. Illustrative example of SINR map, licensed band.

nodes in an ad-hoc mode on the same channel used by the WLAN AP.
- The capacity of both LTE and WLAN is well-provisioned to accommodate a large number of streaming sessions. In this respect, five component carriers (CCs), each of 20 MHz, are aggregated for LTE assuming proportional fair (PF) MAC scheduling, while 802.11ac is assumed for WLAN with a maximum bandwidth of 160 MHz.
- Access to licensed bands (i.e., LTE) is assumed to be paid (i.e., consumes part of the available credit), while access to unlicensed bands (i.e., WLAN) is free-of-charge.

B. Benchmarking

To benchmark the proposed strategy, the traditional DASH approach (i.e., adapting video quality inside a given RAT) will be first used as baseline:

- *DASH(RAT_k)*: This represents the legacy DASH applied on a single RAT (i.e., RAT_k). Without loss of generality, one of the most cited algorithms [7] is selected.

To extend DASH with a multi-RAT capability, the following schemes will be assessed and compared:

- *UE*: A distributed scheme that combines exploitation and exploration without any assistance from the network. When the serving RAT (i.e., $RAT_{k^*(n)}$) is suitable (i.e., $s^{ic}_{v,n,k^*(n)} \geq Thr$), it exploits (i.e., keeps using) it. In turn, when the serving RAT becomes unsuitable (i.e., $s^{ic}_{v,n,k^*(n)} < Thr$), it performs a trial-and-error exploration of the other RATs with equal probabilities.
- *MEC*: This is a centralized scheme based on multi-access edge computing (MEC) [39] that solves the formulated problem in (10) at the edge of the radio access network each ΔT. To this end, it collects the relevant components of the context from the various users each T_{rep}. Note that, due to the loose integration between the various RATs, this scheme is not feasible, but would be used to benchmark the performance of the other schemes.
- *NQA*: The Network-assisted QoE-driven Adaptation (NQA) strategy described in Section IV-C, where the adaptation logic function *f(.,.,.)* of Algorithm 1 is set to that of *DASH(RAT_k)* [7].

Finally, the following variants will be compared to assess the effectiveness of the proposed multi-timescale operation:

- *NQA+SC*: It is equal to the *NQA* strategy with a dynamic adjustment of its controlling parameters on a slow timescale (i.e., hundreds of ms) performed by the various SCs as described in Fig. 4.
- *NQA+SC+PD*: In addition to the tuning performed by the various SCs, the PD of Fig. 3 may alter some of the MADM controlling parameters on the slowest timescale (i.e., few seconds) based on the network conditions (e.g., interference amount in a given neighborhood). Note that only a loose integration between the various RATs is needed for this scheme due to its slow operation.

C. Key performance indicators

The reliability of the proposed strategy is assessed based on the following metrics:

- $\overline{f(\cdot)}$: The average objective function collectively achieved by all DASH and legacy users. It is computed by averaging the values of $f(\cdot)$ in (11).
- $\overline{s^{ic}_{v,n,k^*}(t)}$: The average "in-context" suitability achieved by DASH clients evaluated in (8).
- $\overline{s^{oc}_{d,n,k^*}(t)}$: The average "out-of-context" suitability perceived by legacy data transfers evaluated in (9).
- $\overline{Fairness}$: The average Jain's fairness index [40] defined at a given time *t* as

$$Fairness(t)=\frac{\left(\sum_{n=1}^{N_v} s^{ic}_{v,n,k^*}(t)+\sum_{n=1}^{N_d} s^{oc}_{d,n,k^*}(t)\right)^2}{N_v\cdot\sum_{n=1}^{N_v}(s^{ic}_{v,n,k^*}(t))^2+N_d\cdot\sum_{n=1}^{N_d}(s^{oc}_{d,n,k^*}(t))^2} \quad (13)$$

  Note that $Fairness(t)$ assesses whether the DASH and legacy users are achieving comparable performances.

To have a deeper look at the actual perception of DASH clients, the following metrics are considered:

- $\overline{QoE(t)}$: this is the average QoE level achieved by all DASH clients, where each instantaneous QoE is evaluated in (2) at $R_{v,n,k}(t)=Q_n(j(t))$.
- 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 play-out buffer runs out of data. It is calculated as the fraction of stalling duration out of the total playback duration.
- D_{prob}: the probability that the video streaming session drops due to credit depletion or full battery discharge.

Finally, the generated control overhead is evaluated based on the following indicators:

- $Overhead$: The total signaling overhead generated on the radio interface given by

$$Overhead=(N_{estab}\cdot\overline{estab})+(N_{accept}\cdot\overline{accept})+(N_{rep}\cdot\overline{rep})+(N_{switch}\cdot\overline{switch}) \quad (14)$$

  where N_{estab}, N_{accept}, N_{rep} and N_{switch} denote the number of establishment requests, accepted requests, measurement reports and RAT switches, respectively. The corresponding costs are $\overline{estab}$=266, $\overline{accept}$=64, $\overline{rep}$=43, and $\overline{switch}$=167 Bytes, respectively [41].
- $Reporting$: The amount of over-the-air reporting evaluated in terms of the number of reported bits per active session. Recall that the centralized scheme (i.e., *MEC*) relies on a periodic reporting of all achievable performances, while the proposed methodology (i.e., *NQA*) generates only one CDR to guide future adjustments on the network side as explained in Section IV-A.
- N_{switch}/s: The number of RAT switches per second.

D. Initial assumptions

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

- A single-room scenario is considered, where $N_v \in \{3,\cdots,15\}$ active video clients are considered together with N_v^{idle} idle users such that $N_v+N_v^{idle}=15$.
- Inside the same room, a set of $N_d^W=10$ unlicensed legacy data transfers are initially established and maintained for

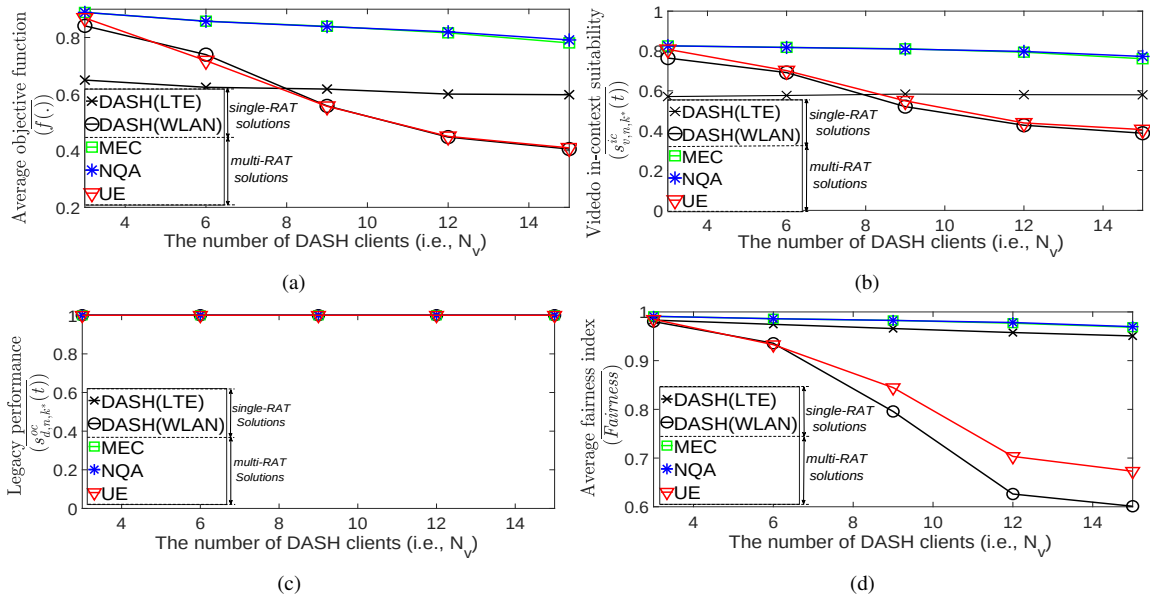

Fig. 7. Performance evaluation in terms of (a) Objective function, (b) Video "in-context" suitability, (c) Legacy performance and (d) Fairness index.

a duration of t_d^W=3 min (i.e., from t=60 s to t=240 s). A set of licensed legacy users will be additionally considered in Section. V-G.

- Video clients have a limited credit of $C_{v,n}^{av}(t=0)$=4 Gbits, full battery capacity of $E_{v,n}^{av}(t=0)$=37 800 J (2100 mAh at a voltage of 5 V - A typical smartphone battery capacity) and move inside the same room at $v_n(t)$=0.3 Km/h.
- The weights associated with video and legacy users in (11) are set to $\delta_{v,n}$=5 and $\delta_{d,n}$=1, respectively.
- The consumed energy by the RAT radios is calculated based on the state of the Phy layer (e.g., Idle, Tx and Rx) and its associated current draw in Ampere [42, 43]. Additionally, based on the study performed in [44], the power level associated with power-efficient states is set to 10.01 mW, while the energy and time needed to switch between power-efficient and active states are set to E_{act}=4.4 µJ and T_{act}=20 µs.
- The proposed methodology (i.e., *NQA*) estimates the perceived bit-rate over the latest N_{avg}=5 segments (line 3 of Algorithm 1), the centralized scheme (i.e., *MEC*) solves (10) each ΔT=250 ms based on a periodic reporting each T_{rep}=120 ms, while the distributed scheme (i.e., *UE*) uses a satisfaction threshold of Thr=0.9.
- During a simulation time of T_{sim}=300 s, each DASH client continuously request segments, each of T_S=2 s.

E. Cost-benefit analysis

1) Performance evaluation: This section assesses the effectiveness of the proposed strategy in supporting adaptive video streaming subject to the existence of legacy data transfers. To this end, the performances of the schemes considered in Section V-B are compared based on the metrics of Section V-C.

Fig. 7(a) plots the objective function (i.e., $\overline{f(\cdot)}$) achieved by each scheme as a function of the number of DASH clients. The individual performances of video (i.e., $\overline{s_{v,n,k^*}^{ic}(t)}$) and legacy (i.e., $\overline{s_{d,n,k^*}^{oc}(t)}$) users are separately shown in Figs. 7(b) and 7(c), respectively. Finally, the associated fairness index (i.e., $\overline{Fairness}$) is plot in Fig. 7(d).

The results show that the proposed methodology (i.e., *NQA*) and its centralized counterpart (i.e., *MEC*) significantly outperform all other schemes (Fig. 7(a)). An analysis of the individual performances reveals that, for the considered traffic mixture (i.e., adaptive video streaming and legacy data transfer), the observed overall improvement in Fig. 7(a) is mainly due to a better satisfaction of the most demanding application (i.e., adaptive video streaming) in Fig. 7(b). The the loose bit-rate requirement of the less demanding legacy data transfers (i.e., R_d^{min}=1 Kbps) is easily satisfied by any of the assigned radio access technologies (RATs), and thus by any of the considered schemes (i.e., *DASH (LTE)*, *DASH (WLAN)*, *MEC*, *NQA* and *UE*) as can observed in Fig. 7(c). On the one hand, adapting video quality within one single RAT cannot sustain a good performance. When only LTE is considered (i.e., *DASH(LTE)*), the unloaded licensed band helps to achieve the best QoE levels, but is too costly given the limited credit of DASH clients. In turn, when only WLAN is exploited (i.e., *DASH(WLAN)*), the contention level created by legacy users degrades the performance. On the other hand, the fully distributed scheme (i.e., *UE*) exploits all RATs, but its lack of global knowledge leads to sub-optimum performance. When comparing the best performing schemes (i.e., *NQA* and *MEC*), it can be seen that they perform equally at low loads, but *MEC* gets slightly degraded at higher loads (i.e., $N_v \geq 12$). This is because *MEC* relies on an uplink signaling that becomes a bottleneck at the highest loads.

Next, the perception of DASH clients is further analyzed.

Fig. 8(a) plots the average QoE level achieved by all clients (i.e., $\overline{QoE(t)}$). To assess its various components, Figs. 8(b), 8(c) and 8(d) show the associated video quality (i.e., f_{4K}), stalling (i.e., S_{prob}) and dropping (i.e., D_{prob}) probabilities, respectively. For better analysis, Fig. 9 shows the

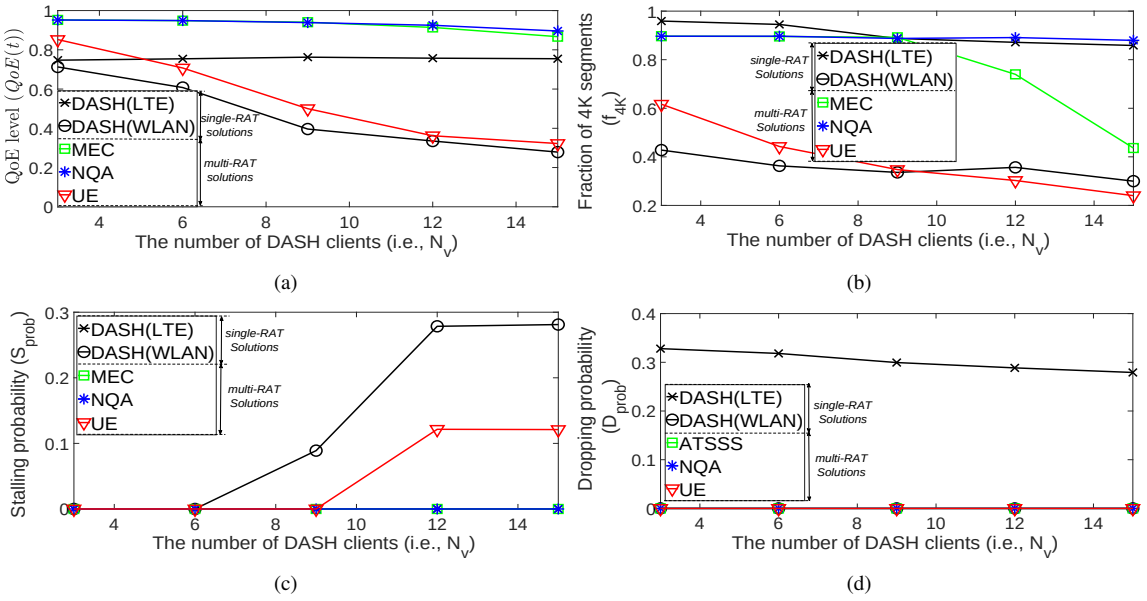

Fig. 8. Video performance in terms of (a) Average QoE, (b) Video quality, (c) Stalling probability and (d) Dropping probability.

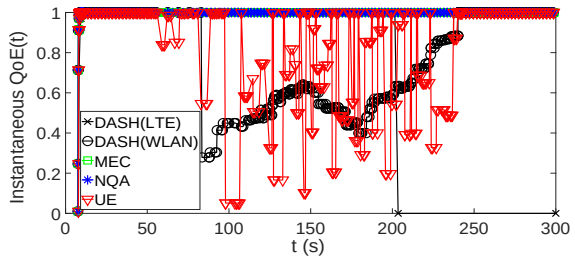

Fig. 9. Evolution of the instantaneous $QoE(t)$, N_v=9, 6$^{\text{th}}$ client.

instantaneous QoE of an arbitrary (i.e., 6$^{\text{th}}$) client for N_v=9.

The first observation in Fig. 8(a) is that the observed performance exhibits a similar behaviour to Figs. 7(a) and 7(b). When adaptation is performed within LTE (i.e., *DASH(LTE)*), a high fraction of 4K segments (i.e., f_{4K}) can be delivered on the unloaded licensed band (Fig. 8(b)). However, the limited credit units are exhausted at about t=205 s as can be observed in Fig. 9, which drops the session (Fig. 8(d)). In turn, when adaptation is performed within WLAN (i.e., *DASH(WLAN)*), shortly after the start of legacy sessions (i.e., t=60 s), the remaining WLAN capacity is no longer sufficient to sustain the highest quality (Fig. 8(b)), which degrades the QoE due to the reduction of the second term (i.e., $U^R_{v,n,k}(t)$) in (2). As the number of video sessions increases (i.e.. $N_v \geq 9$), the offered capacity cannot even sustain the lowest quality (i.e., 5 Mbps), and thus the video stalls (Fig. 8(c)). When both RATs are exploited without any assistance from the network (i.e., *UE*), the performance is slightly improved, but the continuous exploration of better RATs results in an instable behaviour as reflected by the frequent oscillations observed in Fig. 9.

2) Cost analysis: This section conducts an analysis of the control overhead generated on the radio interface (i.e., $Overhead$) and its components defined in Section V-C.

Fig. 10(a) plots the signaling overhead introduced by all schemes as a function of the number of active DASH clients. Fig. 10(b) shows the corresponding amount of measurements reports (i.e., *Reporting*) generated by *MEC* and *NQA*. The other schemes are not shown as they do not generate any reporting. Fig. 10(c) plots the associated number of RAT switches (i.e., N_{switch}/s) performed by multi-RAT schemes (i.e., *MEC*, *NQA* and *UE*). Note that the first two metrics are plot in logarithmic scale for improved visualization.

The results show that the control overhead of *NQA* is much smaller compared to its fully centralized and distributed counterparts (Fig. 10(a)). On the one hand, the observed improvement with respect to *MEC* is mainly due to the reporting component as can be seen in Fig. 10(b). This is because *MEC* requires to continuously report all achievable bit-rates ($\{R_{m,n,k}(t)\}$) even when RATs are not used, which increases the amount of signaling per session, particularly when the number of active users is low. On the other hand, the overhead reduction in reference to *UE* is mainly due to the RAT switch component as illustrated in Fig. 10(c).

In summary, at low loads, the proposed methodology performs closely to its centralized counterpart with much less control overhead. It scales up well, while the centralized approach gets overwhelmed by its uplink signaling.

F. Impact of controlling parameters

This section performs a sensitivity analysis to the dynamic adjustment of the various weights on the user side. Without loss of generality, the credit and energy attributes are particularly considered. To isolate the mutual effects between the various clients, N_v=1 video session is considered.

1) Credit: This section conducts an analysis of the sensitivity to the credit weight. To this end, the following variants of the strategy described in Table I are considered:

- *Quality First*: The credit weight $w_{v,n,C}(t)$ is set to LOW regardless of the available credit (i.e., C^2_{thr}=0).
- *Credit First*: The credit weight $w_{v,n,C}(t)$ is always set to HIGH (i.e., C^1_{thr}=$C^{av}_{v,n}(t)$).

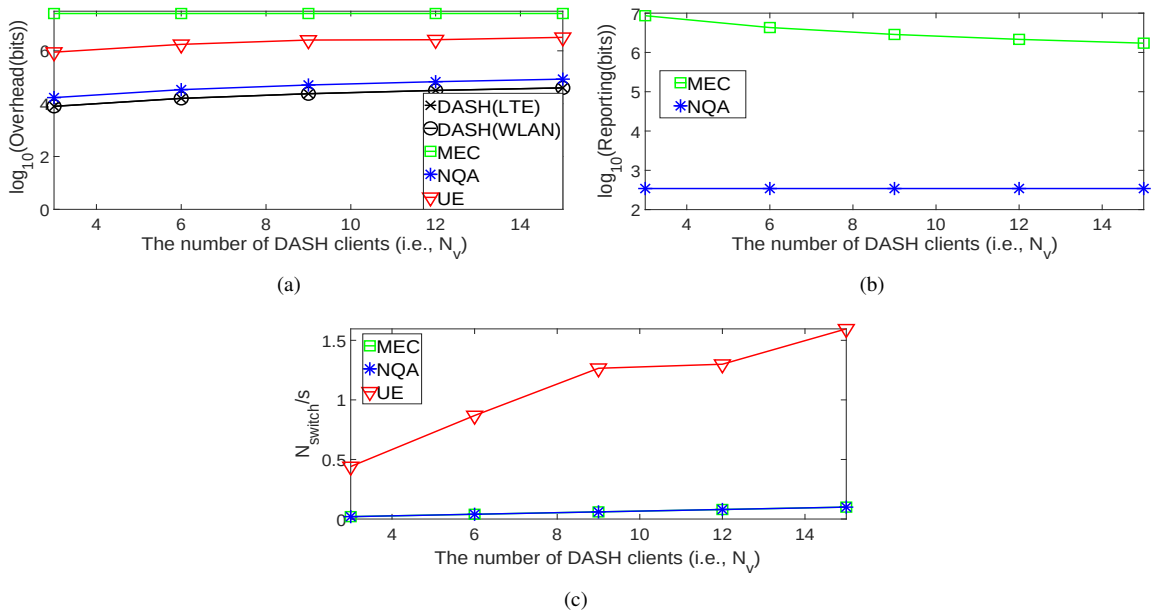

Fig. 10. Analysis of the control overhead in terms of (a) *Overhead*, (b) *Reporting* and (c) N_{switch}/s.

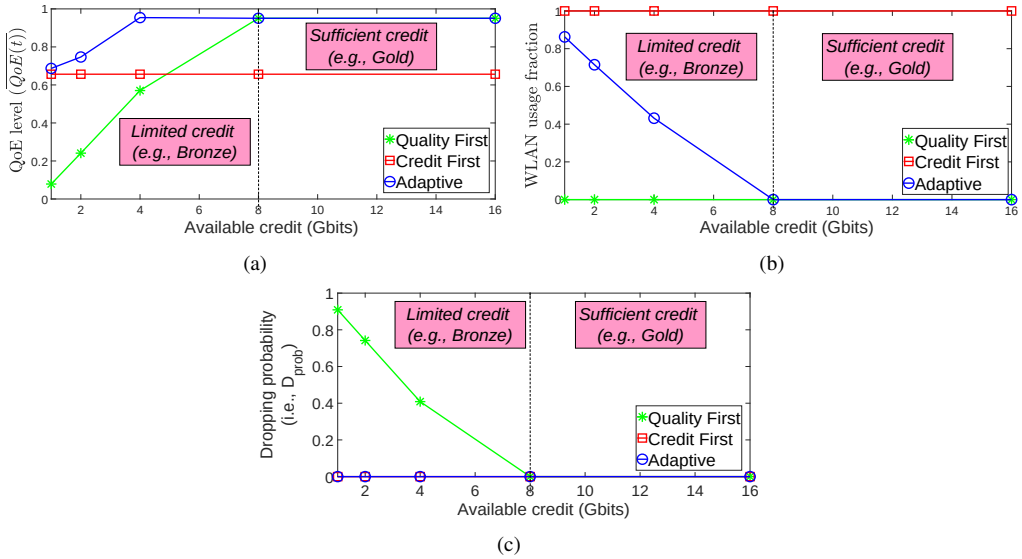

Fig. 11. Sensitivity analysis to the credit weight in terms of (a) Average QoE level, (b) WLAN usage fraction and (c) Dropping probability.

- *Adaptive*: this variant dynamically adjusts $w_{v,n,C}(t)$ to maximize video quality subject to avoiding credit depletion. This is achieved by the following settings:

$$C^1_{thr}{=}100\,\text{Mbps},\ C^2_{thr}{=}T_{rem}{\cdot}\overline{C} \tag{15}$$

  where T_{rem} and $\overline{C}$ denote the remaining session duration and credit consumption rate, respectively.

Fig. 11(a) plots the QoE level achieved by the various variants as a function of the available credit. To isolate the impact of the energy attribute, a full battery capacity of $E^{av}_{v,n}(t{=}0){=}37\,800\,\text{J}$ is assumed. The associated objective function and "in-context" suitability are exhibiting similar behaviour, and are thus not shown. Figs. 11(b) and 11(c) show the associated fraction of using WLAN and dropping probability that may result from credit depletion, respectively.

The results show that the dynamic adjustment of weights (i.e., *Adaptive*) maintains the best performance at all levels of available credit. On the one hand, when the quality is blindly prioritised (i.e., *Quality First*), the costly licensed band is exclusively selected to achieve the highest QoE levels (Fig. 11(b)). If the available credit is abundant (i.e., $C^{av}_{v,n}(t){\geq}8\,\text{Gbits}$), the good performance could be maintained till the end of the session. Otherwise, the credit is depleted in the middle of video streaming, which results in an out-of-credit drop (Fig. 11(c)). On the other hand, when saving the credit units is favoured (i.e., *Credit First*), the affordable WLAN is exclusively used (Fig. 11(b)), which results in a performance degradation due to the contention level experienced on the unlicensed band (Fig. 11(a)). In contrast, it can be seen from Figs. 11(b) and 11(c) that *Adaptive* uses LTE exclusively when the credit is sufficient (e.g., Gold subscription), and

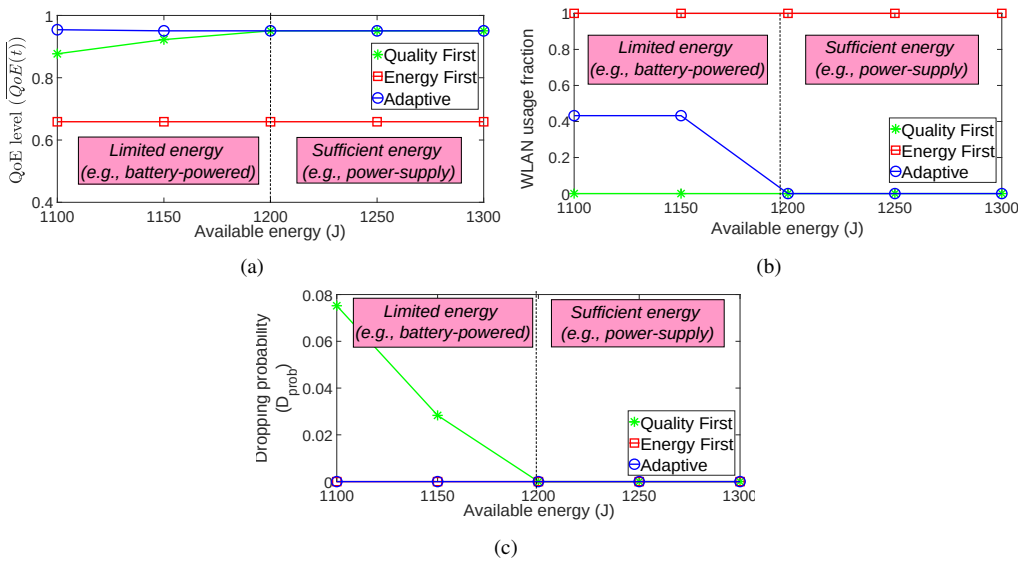

Fig. 12. Sensitivity analysis to the energy weight in terms of (a) Average QoE level, (b) WLAN usage fraction and (c) Dropping probability.

WLAN opportunistically when the credit becomes limited (e.g., Bronze subscription) to the extent that avoids dropping.

2) Energy: This section conducts an analysis of the sensitivity to the energy weight. In this respect, the following variants of the strategy described in Table I are considered:

- *Quality First*: The energy weight $w_{v,n,E}(t)$ is set to LOW regardless of the available battery level (i.e., E^2_{thr}=0).
- *Energy First*: The energy weight $w_{v,n,E}(t)$ is always set to HIGH (i.e., E^1_{thr}=$E^{av}_{v,n}(t)$).
- *Adaptive*: this variant dynamically adjusts $w_{v,n,E}(t)$ to maximize video quality subject to avoiding battery discharge. This is achieved by the following settings:

$$E^1_{thr}=T_{rem}\cdot\overline{P_{min}},\ E^2_{thr}=T_{rem}\cdot\overline{P_{max}} \qquad (16)$$

where $\overline{P_{min}}$ and $\overline{P_{max}}$ denote the power levels associated with the least and most power-consuming RATs.

Fig. 12(a) plots the QoE achieved by the various variants as a function of the available energy for $C^{av}_{v,n}(t{=}0)$=16 Gbits. For a realistic assessment, the power consumed by the screen display has been set to P_{screen}=2 W assuming a maximum brightness level [45]. Figs. 12(b) and 12(c) show the associated fraction of using WLAN and dropping probability that may result from full battery discharge, respectively.

The results show that the dynamic adjustment of weights (i.e., *Adaptive*) achieves the best performance at all levels of available energy. On the one hand, when the quality is prioritised (i.e., *Quality First*), the energy-consuming licensed band is exclusively selected (Fig. 12(b)). If the available energy is sufficient (i.e., $E^{av}_{v,n}(t)\geq$1200 J), the good performance can be maintained till the end. Otherwise, the battery is fully discharged in the middle of video streaming and the session drops (Fig. 12(c)). On the other hand, when saving energy is favoured (i.e., *Energy First*), WLAN is exclusively selected (Fig. 12(b)), which degrades the performance due to the contention level generated by legacy users (Fig. 12(a)). In contrast, it can be seen from Figs. 12(b) and 12(c) that *Adaptive* uses LTE exclusively when the available energy is abundant (e.g., power-supply), and WLAN opportunistically when the energy budget is limited (e.g., battery-powered devices) to the extent that avoids dropping.

G. Multi-time scale operation

This section assesses the effectiveness of the proposed multi-timescale operation to cope with the legacy activity. To this end, N_v=15 DASH clients, each associated with an abundant credit (i.e., 16 Gbits) and full battery level (i.e., 37 800 J), are placed in an arbitrary room of the layout of Fig. 6 with a variable density of legacy users around. The legacy density is varied by considering an increasing number of legacy rooms (N^{room}_d), where each room is associated with a set of N^L_d=5 and N^W_d=10 legacy data transfers established on licensed and unlicensed bands, respectively.

Fig. 13(a) plots the objective function achieved by the proposed variants in Section. V-B as a function of the legacy density. Figs. 13(b) and 13(c) show the individual performances achieved by the video and legacy users, respectively.

The observed behaviour shows that, when only video sessions are established (i.e., N^{room}_d=0), the user-driven operation (i.e., *NQA*) is sufficient to achieve good performance. Recall that the licensed band is exclusively used in this case due to the abundant credit and full battery level. When legacy data transfers are established inside the same room (i.e., N^{room}_d=1), the increased contention level overwhelm the video sessions as the considered MAC scheduling (i.e., PF) is not QoS-aware. In turn, when the controlling parameters are tuned at the cell level (i.e., *NQA+SC*), the resources used by legacy users are preempted and assigned to the more demanding video sessions, which maintains the good performance. When more legacy data transfers are established in the surrounding area (i.e., $N^{room}_d\geq$4), the tuning performed by *NQA+SC* cannot cope with the increasing amount of interference in licensed bands, which degrades the individual video performance (Fig. 13(b)) and the overall efficiency in

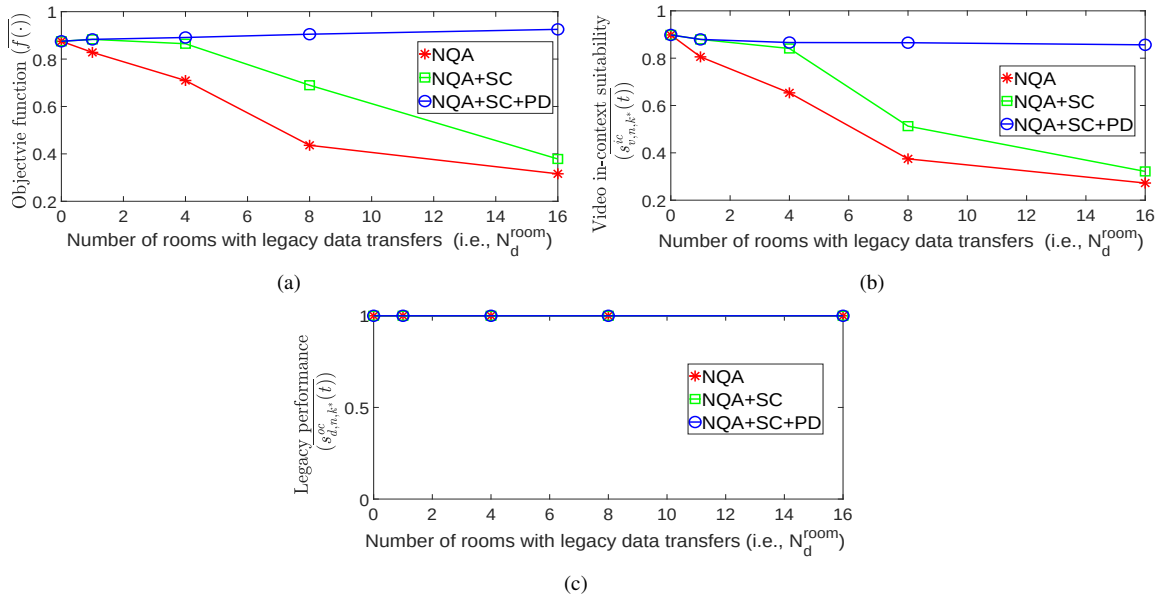

Fig. 13. Impact of the legacy activity in terms of (a) Objective function, (b) Video "in-context" suitability and (c) Legacy performance.

consequence (Fig. 13(a)). Finally, when the user-driven operation is regulated at the network-level (i.e., *NQA+SC+PD*), the legacy users are offloaded to unlicensed bands in the whole neighborhood, which significantly decreases the amount of interference in licensed bands and helps to sustain good performance for all video clients. Note that, in this case, the objective function increases as a function of the legacy density due to a higher contribution of legacy data transfers in (11).

VI. Conclusions and Future Directions

This paper formulates an optimization problem to maximize an aggregate utility that captures the "in-context" suitability of available radio access technologies (RATs) to support adaptive video streaming subject to the existence of a set of legacy data transfers. To efficiently solve the formulated problem when the various RATs are loosely integrated (i.e., at core network level), a generic hybrid (i.e., network-assisted user-driven) framework, constructed in a prior work, is extended and instantiated. First, a set of mechanisms are introduced to allow a multi-timescale operation, where the user-driven behaviour is regulated based on the network-level strategy and constraints. Then, the extended framework is mapped onto the 3GPP 5G architecture and instantiated as an add-on feature that can be activated only for the relevant use cases. Based on it, a hierarchical quality-of-experience (QoE)-driven methodology is devised to adapt video streaming across all available RATs, while ensuring fair co-existence with the legacy users. It operates on different timescales, where the fastest timescale operation is performed on the end-users according to a policy tuned by the network on slower timescales. A cost-benefit analysis reveals that the proposed strategy performs closely to its centralized counterpart with much less control overhead on the radio interface. A user tuning on the fastest timescale (i.e., tens of ms) enables to adapt video streaming depending on the perceived QoE and local components of the context (e.g., remaining credit and battery level). A small-cell tuning on a slower timescale (i.e., hundreds of ms) enables to preempt the resources used by legacy users based on the operating conditions (e.g., load and type of scheduler). Finally, a tuning performed by the network on the slowest timescale (i.e., few seconds) offloads legacy data transfers to unlicensed bands whenever the amount of interference on licensed bands reaches critical levels, which helps to sustain good QoE for all video clients.

As part of future work, it is intended to assess the implications of the proposed methodology on the processing load and energy consumption of battery-powered devices and implement it based on network slicing.

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 European Commission in the framework of the H2020-ICT-19-2019 project 5G-HEART (Grant agreement no. 857034).

References

[1] "Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats," ISO/IEC, Switzerland, Tech. Rep. ISO/IEC 23009-1:2019, August 2019.

[2] 3GPP, "Technical Specification Group Services and System Aspects; Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH), (Release 16)," Tech. Rep. TS 26.247-V16.2.0, December 2019.

[3] "Digital Video Broadcasting (DVB); MPEG-DASH Profile for Transport of ISO BMFF Based DVB Services over IP Based Networks," ETSI, Tech. Rep. TS 103 285 V1.3.1, February 2020.

[4] "Hybrid Broadcast Broadband TV," ETSI, Tech. Rep. TS 102 796 V1.2.1, November 2012.

[5] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman, "BOLA: Near-optimal bitrate adaptation for online videos," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, April 2016, pp. 1–9.

[6] 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.

[7] 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.

[8] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2069–2084, Sep. 2017.

[9] V. Sagar, R. Chandramouli, and K. P. Subbalakshmi, "Software defined access for HetNets," *IEEE Communications Magazine*, vol. 54, no. 1, pp. 84–89, January 2016.

[10] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, "5G multi-RAT LTE-WiFi ultra-dense small cells: Performance dynamics, architecture, and trends," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1224–1240, June 2015.

[11] 3GPP, "New Work Item on enhanced LAA for LTE," Tech. Rep. RP-152272, December 2015.

[12] ——, "New Work Item on enhanced LWA," Tech. Rep. RP-160600, March 2016.

[13] "5G Whitepaper: The Flat Distributed Cloud (FDC) 5G Architecture Revolution," 5G Innovation Centre, University of Surrey, Tech. Rep., January 2016.

[14] M. Yan, G. Feng, J. Zhou, and S. Qin, "Smart multi-RAT access based on multiagent reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4539–4551, May 2018.

[15] A. Keshavarz-Haddad, E. Aryafar, M. Wang, and M. Chiang, "Hetnets selection by clients: Convergence, efficiency, and practicality," *IEEE/ACM Transactions on Networking*, vol. 25, no. 1, pp. 406–419, Feb 2017.

[16] K. Zhu, D. Niyato, and P. Wang, "Network selection in heterogeneous wireless networks: Evolution with incomplete information," in *2010 IEEE Wireless Communication and Networking Conference*, April 2010, pp. 1–6.

[17] D. D. Nguyen, H. X. Nguyen, and L. B. White, "Reinforcement learning with network-assisted feedback for heterogeneous RAT selection," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 6062–6076, 2017.

[18] B. H. Jung, N. Song, and D. K. Sung, "A network-assisted user-centric WiFi-offloading model for maximizing per-user throughput in a heterogeneous network," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 4, pp. 1940–1945, 2014.

[19] S. Maghsudi and S. Stańczak, "Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1309–1322, 2015.

[20] A. Mehrabi, M. Siekkinen, and A. Ylä-Jääski, "Edge computing assisted adaptive mobile video streaming," *IEEE Transactions on Mobile Computing*, vol. 18, no. 4, pp. 787–800, 2019.

[21] E. Ozfatura, O. Ercetin, and H. Inaltekin, "Optimal network-assisted multiuser DASH video streaming," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 247–265, 2018.

[22] A. Mehrabi, M. Siekkinen, and A. Ylä-Jääski, "Joint optimization of QoE and fairness through network assisted adaptive mobile video streaming," in *2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2017, pp. 1–8.

[23] 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.

[24] ——, "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.

[25] ——, "A Context-aware QoE-driven Strategy for Adaptive Video Streaming in 5G multi-RAT environments," in *2017 20th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Dec 2017, pp. 354–360.

[26] J. Kwak, J. Moon, H. Lee, and L. B. Le, "Dynamic network slicing and resource allocation for heterogeneous wireless services," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–5.

[27] F. Teng and D. Guo, "Resource management in 5G: A tale of two timescales," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, Nov 2015, pp. 1041–1045.

[28] G. Dimitrakopoulos *et al.*, "Adaptive resource management platform for reconfigurable networks," *Mobile Networks and Applications*, vol. 11, no. 6, pp. 799–811, Dec 2006.

[29] 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.

[30] 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), 2011 14th International Symposium on*, Oct. 2011, pp. 1–5.

[31] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4G LTE networks," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '12. New York, NY, USA: ACM, 2012, pp. 225–238.

[32] R. Vassoudevan and P. Samundiswary, "Performance analysis of DRX power saving technique for LTE based UE under bursty web traffic," in *2015 International Conference on Communications and Signal Processing (ICCSP)*, April 2015, pp. 0924–0928.

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

[34] 3GPP, "3GPP technical specification group services and system aspects; system architecture for the 5G system; stage 2," Tech. Rep. 23.501 V16.4.0, March 2020.

[35] ——, "3GPP Technical Specification Group Services and System Aspects; Study on Access Traffic Steering, Switch and Splitting support in the 5G system architecture Phase 2 (Release 17)," Tech. Rep. TR 23.700-93 V1.0.0, September 2020.

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

[37] 3GPP, "TSG RAN WG4 Meeting 51: Simulation assumptions and parameters for FDD HeNB RF requirements," Tech. Rep. R4-092042, May 2009.

[38] "NS-3 Model Library," Tech. Rep. Release NS-3.30, August 2019. [Online]. Available: https://www.nsnam.org/docs/release/3.30/models/ns-3-model-library.pdf

[39] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1657–1681, thirdquarter 2017.

[40] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Digital Equipment Corporation, Tech. Rep., Sep 1984.

[41] F. Bouali, O. Sallent, J. Pérez-Romero, and R. Agustí, "A cogni-

tive management framework for spectrum selection," *Computer Networks*, vol. 57, no. 14, pp. 2752 – 2765, 2013.

[42] D. Halperin, B. Greenstein, A. Sheth, and D. Wetherall, "Demystifying 802.11n power consumption," in *Proceedings of the 2010 International Conference on Power Aware Computing and Systems*, ser. HotPower'10. Berkeley, CA, USA: USENIX Association, 2010.

[43] A. R. Jensen, M. Lauridsen, P. Mogensen, T. B. Sørensen, and P. Jensen, "LTE UE power consumption model: For system level energy and performance optimization," in *2012 IEEE Vehicular Technology Conference (VTC Fall)*, Sept 2012, pp. 1–5.

[44] M. Lauridsen, G. Berardinelli, F. M. L. Tavares, F. Frederiksen, and P. Mogensen, "Sleep modes for enhanced battery life of 5G mobile terminals," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–6.

[45] M. Kennedy, H. Venkataraman, and G.-M. Muntean, *Energy Consumption Analysis and Adaptive Energy Saving Solutions for Mobile Device Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 173–189.

Faouzi Bouali is a Research Fellow in the 5G Innovation Centre (5GIC) of the University of Surrey, United Kingdom. Since 2009, he has been contributing to various research projects funded by the European Commission Horizon 2020 Framework Programme (i.e., 5G-HEART, 5GENESIS and Speed-5G), 7th Framework Programme for Research and Technological Development (i.e., FARAMIR, OneFIT and HELP) and BT Research, United Kingdom. Additionally, he has been working for over a decade as radio optimization engineer and consultant in various public and private projects. His expertise and research interests span the field of mobile radio-communication systems with a specific focus on innovative architectures for next-generation networks, advanced radio resource management (RRM), vehicular communications, spectrum sharing and quality of service/experience (QoS/QoE) provisioning in heterogeneous and cognitive radio networks. At current, he is the Technical Manager (TM) of the Automotive vertical of the Horizon 2020 5G-HEART project (2019-2022) investigating and trialling the capability of 5G systems to support future Automotive use cases. Dr. Bouali is full member of the Engineering and Physical Sciences Research Council (EPSRC) Peer Review College and British Standards Institution (BSI). He is an Associate Editor of IEEE Access and Fellow of the British Higher Education Academy (HEA).

Klaus Moessner is Professor for Communications Engineering at the University of Technology Chemnitz, and Professor in Cognitive Networks at the Institute for Communication Systems and the 5G Innovation Centre, at the University of Surrey. Klaus was involved in a large number of projects in the Cognitive Communications, Service provision and IoT areas. He was responsible for the work on cognitive decision making mechanisms in the CR project ORACLE, and led the work on radio awareness in the ICT FP7 project QoSMOS, led the H2020 Speed5G project. In the past, Klaus was the founding chair of the IEEE DYSPAN Working Group (WG6) on Sensing Interfaces for future and cognitive communication systems. His research interests include cognitive networks, IoT deployments and sensor data based knowledge generation, as well as reconfiguration and resource management; he is senior member of the IEEE. At current he does lead the EU-Taiwan project Clear5G investigating the extensions 5G systems need to serve the particular requirements of the Factories of the Future.

Michael Fitch works for BT in the Research and Innovation department, providing technical leadership to a small research team specialising in physical and systems aspects of wireless communications. He is currently working on a number of projects on emerging wireless technologies such as small cells, radio resource management and 5G. In addition he provides engineering consultancy to other parts of BT on LTE, WiFi and other wireless topics. Previous experience is with modelling, trials and deployments of Satellite, WiMAX, 3G and LTE systems. Michael has a first degree in maths and physics, a PhD in satellite communications and he is a member of the IET.