QoE-SDN APP: A Rate-guided QoE-aware SDN-APP for HTTP Adaptive Video Streaming

Eirini Liotou, Konstantinos Samdanis, Emmanouil Pateromichelakis, Nikos Passas, Lazaros Merakos

***Abstract*—While video streaming has dominated the Internet traffic, Video Service Providers (VSPs) compete on how to assure the best Quality of Experience (QoE) to their customers. HTTP Adaptive Streaming (HAS) has become the de facto way that helps VSPs work-around potential network bottlenecks that inevitably cause stallings. However, HAS-alone cannot guarantee a seamless viewing experience, since this highly relies on the Mobile Network Operators' (MNOs) infrastructure and evolving network conditions. Software-Defined Networking (SDN) has brought new perspectives to this traditional paradigm where VSPs and MNOs are isolated, allowing the latter to open their network for more flexible, service-oriented programmability. This paper takes advantage of recent standardization trends in SDN and proposes a programmable QoE-SDN APP, enabling network exposure feedback from MNOs to VSPs towards network-aware video segment selection and caching, in the context of HAS. The video selection problem is formulated using Knapsack optimization and relaxed to partial sub-problems that provide segment encodings that can mitigate stallings. Furthermore, a mobility prediction mechanism based on the SLAW model is introduced, towards proactive segment caching. A number of use cases, enabled by the QoE-SDN APP, are designed to evaluate the proposed scheme, revealing QoE benefits for VSPs and bandwidth savings for MNOs.**

***Index Terms*—HTTP Adaptive Streaming, Software-Defined Networking, Quality of Experience, Mobile Network Operator, Video Service Provider, video streaming**

I. INTRODUCTION

The emerging 5G networks are expected to enable a service ecosystem that facilitates new business opportunities, supporting also market players that do not necessarily own a network infrastructure, such as verticals and service/application providers. Such a 5G paradigm will scale-up further traffic volumes due to the mass adoption of content-rich multimedia applications and cloud services, introducing stringent service requirements in dense areas and on the move [1]. Alongside the launch of new 5G services including massive Internet of Things (mIoT), vehicular, and critical communications, etc., 5G networks will diversify the desired performance requirements in terms of throughput, latency, jitter, etc. This plethora of 5G services creates pressure for Mobile Network Operators (MNOs) who cannot simply react by overprovisioning the network infrastructure, since the service race for the same set of resources is endless and the associated infrastructure cost is tremendous. Instead, to assure that the best experience is always assigned and follows a user, i.e., irrespective of location and network conditions, enhanced intelligent Quality of Experience (QoE) mechanisms are needed considering the service type specifics and network conditions. For instance, novel paradigms such as the Follow-Me Cloud (FMC) have emerged, allowing a service (and not just the content) to follow a mobile user, ensuring service continuity [2]. Regardless of this immense potential, MNOs continue to offer only a "communication pipe", while being in search for new business models to allow them to enter the service/application provider market.

As multimedia services are dominating the mobile economy, an ever-increasing number of Video Service Providers (VSPs) such as Netflix, Amazon, YouTube, etc. is expected to contribute towards a threefold grow of IP video traffic by 2021 [3]. New opportunities for video-related services still arise, especially with 5G, e.g. augmented and virtual reality video, but also outside the entertainment business with various verticals dependent on video such as e-health, security, safety, etc. Currently, VSPs offer Over-The-Top (OTT) services considering the underlying infrastructure as a "black box" supporting best-effort services. HTTP Adaptive Streaming (HAS) has appeared as a work-around solution of VSPs to confront network bottlenecks by dynamically controlling the rate at which video is offered, with the ultimate goal to avoid stalling events, i.e., video freezing, which constitutes the most crucial QoE degrading factor [4]. Despite the success in mitigating stallings, HAS may lead to an inevitably sub-optimal solution, since: a) quality adjustments are done re-actively after the service has already degraded, b) HAS tries to overcome a network problem without having any network control, and c) it relies on the subjective and isolated user perception regarding bandwidth availability.

The high competitiveness in the VSP market as well as the large business potential encourage service providers to find new means to offer higher QoE to their customers. The World Economic Forum recognizes that MNOs need to launch new business models, where they partner directly with various vertical markets (e.g. VSPs), in the direction of transforming their networks into more flexible, open, and customized infrastructures, as well as providing differentiation in a software-based way [5]. MNOs can therefore exploit their exclusively owned assets and capabilities, namely a) user

This work has been partially funded by European Union under Grant Agreement 645393 (H2020-MSCA-RISE CASPER) and under Grant Agreement 762057 (H2020 5G-PICTURE).

E. Liotou (eliotou@di.uoa.gr), N. Passas (passas@di.uoa.gr) and L. Merakos (merakos@di.uoa.gr) are with the National & Kapodistrian University of Athens - Department of Informatics & Telecommunications, Athens, Greece.

K. Samdanis (konstantinos.samdanis@huawei.com) and E. Pateromichelakis (emmanouil.pateromichelakis@huawei.com) are with Huawei Technologies, Munich, Germany.

Article last revised: February 28, 2018

information, b) network conditions, and c) technological options relative to their infrastructure, to create and offer additional services. Leveraging the benefit of such information and by opening their networks for collaboration, MNOs can form new business models considering network, user and service intelligence (e.g. regarding congestion and location, big data related to users, etc.) as well as open Application Programming Interfaces (APIs), enhancing the VSPs' capabilities beyond just application-level parameter control [6]. For instance, application-oriented bearer elasticity may be introduced in order to guarantee or augment the QoE of specific VSPs or specific VSPs' customers [7], or multi-path routing combined with a precise bandwidth allocation may be applied in a Software Defined Networking (SDN) environment [8].

Currently, SDN [9]-[10] facilitates programmability and openness, enabling VSPs to interact with the network layer via open APIs, which allows MNOs and VSPs to build a close collaboration with a positive value for both stakeholders. In particular, the benefits of such a collaboration paradigm via the means of SDN are: a) VSP customers are served with better QoE, enabled by the direct interaction among VSPs and MNOs, b) application/service-awareness allows MNOs to manage network resources more efficiently, and c) MNOs can get into the revenue loop of the APP market, offering big data and QoE-related information through their open APIs to third parties. A SWOT analysis from the MNOs' perspective is provided in Figure 1, elaborating on the Weaknesses and Threats in the current Telecom status quo (where MNOs and VSPs are isolated), but also on the Strengths and Opportunities that arise from eliminating such an isolation.

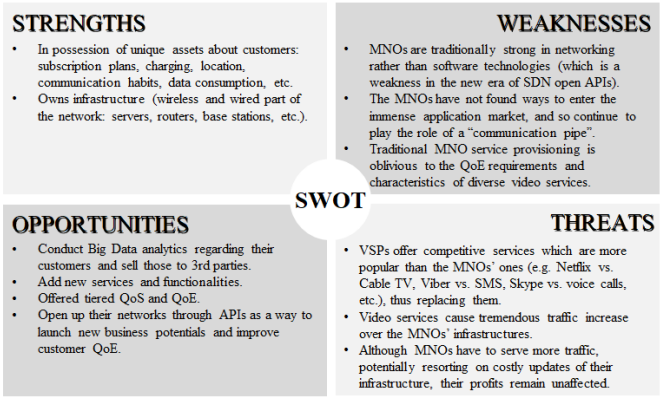

Fig. 1. SWOT analysis from the MNO's perspective.

This paper incentivizes and provides a technologically feasible realization of an MNO-VSP collaboration, where feedback from the MNO is enabled and application-awareness is enforced. A novel QoE-SDN APP is proposed, which can be flexibly programmed and customized to assure the desired QoE for verticals, VSPs and OTT providers, relying on the specifications of the SDN paradigm. The analysis considered in this paper focuses on the case of video-on-demand with the objective to enhance the HAS paradigm. In particular, in our approach a feedback mechanism is facilitated from the MNO to the VSP, in order to enhance end user QoE. This QoE enhancement is achieved through proactive video selection and encoding, which accounts for the user movement and the potential network conditions in the process of assigning the required video encoding rate that reduces stalling probability. We complementary explore the use of Multi-access Edge Computing (MEC) [11], which can cache HAS segments in advance based on forecasted user mobility in order to enhance QoE, while allowing MNOs to utilize the network resources more efficiently. We formulate an optimization problem with the objective of improving the user QoE and model the solution using a Discrete Time Markov Chain model combined with a video segment-to-quality mapping problem that optimizes the video encoding selection. Moreover, we propose three novel use cases in the context of HAS, unlocked by the proposed framework, which incorporate mobility and rate guidance towards a better video encoding selection and a more efficient video segment caching. A set of simulations in a realistic and challenging mobile cellular environment demonstrate the added value of the proposed scheme, in terms of QoE amelioration of VSPs' customers and network resource savings for the benefit of the MNOs.

The remainder of this paper is summarized as follows. In Section II, we review the related state-of-the-art in the areas of QoE provisioning in SDN-based environments. Section III describes the HAS paradigm in a mobile cellular environment, the proposed QoE-SDN APP, and the supporting SDN-based architecture, including required APIs, components and operations. Section IV models the system and formulates an optimization problem of video encoding towards improving end user QoE. Section V then approaches this optimization problem by relaxing it to partial sub-problems, and presents the mobility forecasting and rate estimation logic. Section VI describes three novel use cases in the context of HAS that are activated by the QoE-SDN APP, presents the evaluation environment and respective QoE indicators, as well as the evaluation results. Finally, Section VII concludes the paper.

II. Related Work

The notion of QoE has emerged as a subjective performance measure from the user's perspective with respect to an application or service, and can assist MNOs and application/service providers to understand the overall quality of their services [12]. QoE can be obtained directly from end users or be derived from empirical estimation models that link various performance measures like delay, jitter, loss, stalling, etc., to user experiences, commonly in the form of Mean Opinion Scores (MOS) [13]. SDN, via the means of open APIs, can offer programmability that enables service providers to obtain QoE measures regarding the offered applications as well as the capability to interact with the network, introducing adjustments on the networking resources considering also the application requirements.

Preliminary SDN-based solutions considering QoE concentrate on the core and transport networks taking advantage of the global network view to perform dynamic traffic steering and optimal Content Distribution Network (CDN) selection. In [14], a jointly optimized path assignment and service utility decision for multimedia flows is performed by OpenFlow considering the resource requirements of competing services. Similarly, [15] improves the QoE of video streaming applications using an SDN controller that monitors video QoE metrics at the client side and dynamically selects

TABLE 1
COMPARISON OF SDN-BASED HAS SOLUTIONS.

Solution	Approach	Network	Prediction	HAS strategy	Asset	Weakness	SDN add-on
A. Bentaleb et al. [17]-[18]	Hybrid	Fixed	No	Upper bounded bit rate recommendation & buffer level	Optimized QoE per user, User heterogeneity support	A new user communication interface is required	Internal and external SDN-based resource management components
J. W. Kleinrouweler et al. [19]	Hybrid	Fixed	No	Target bit rate pushed to each user	Explicit adaptation assistance with fairness criteria	Users have to cooperate with the Service Manager	HAS-aware Service Manager
D. Bhat et al. [20]	Hybrid	Fixed	Short-term prediction (ARIMA)	User assisted with information about cache location and link bandwidth	Video segment decision remains at the user's control (scalable)	Overhead due to both bandwidth and cache occupancy monitoring	SABR module
P. Georgopoulos et al. [21]	Hybrid	Fixed	No	Optimum bit rates that ensure fairness pushed to users	Optimized QoE, Heterogeneity support, Fairness	Utility functions need to be pre-calculated and stored for all video content at each resolution	Orchestrating OpenFlow module
QoE-SDN APP	Bitrate Guidance	Mobile	Longer-term (cluster based)	Rate-guided, prediction-based	Network exposure feedback enabled, No change needed at HAS clients	Assumes VSP-MNO collaboration	QoE-SDN APP

delivery nodes via the means of traffic engineering. In the context of HAS, [16] investigates three different network-assisted video streaming approaches: a) Bandwidth Reservation, where optimal bandwidth slices are assigned to video flows, b) Bitrate Guidance, where optimal video bit rates are estimated centrally and then enforced to the users, and c) hybrid approaches, that combine both. Such hybrid solutions are explored in [17]-[21]. SDNDASH [17] relies on an SDN-based management and resource allocation architecture with the goal to maximize the QoE per user considering heterogeneous QoE requirements. Each user's adaptation logic is then based on a combination of optimal bit rate recommendations and buffer levels. As an extension to this work, [18] proposes a more scalable architecture, called SDNHAS, which estimates optimal QoE policies for groups of users and requests a bandwidth constraint slice allocation, while providing encoding recommendations to HAS players. Furthermore, [19] proposes a network application controller, called Service Manager, which oversees video traffic and fairly allocates network resources among competing HAS flows, while enforcing QoS guarantees. A target bit rate is assigned to each client, which can be used as a reference in their adaptation logic regarding the maximum encoding they should request. Then, [20] considers caching, and proposes an SDN-based Adaptive Bit Rate (SABR) architecture, where video users are informed regarding each cache's content as well as get a short-term prediction of the bottleneck bandwidth to reach each cache, so that their adaptation decisions are better. In parallel, OpenFlow guides routing between clients and selected caches. Finally, [21] proposes an OpenFlow-assisted QoE Fairness Framework (QFF), with the objective to fairly optimize QoE among HAS clients with heterogeneous device requirements, expressed via bitrate-to-QoE utility functions. Our QoE-SDN APP adopts joint network and application programmability via the means of open APIs, but in contrary to all previous approaches, we concentrate our efforts on *mobile* networks, which require a higher flexibility due to constantly evolving network dynamics. Moreover, our approach guides HAS-related decisions considering also longer-term forecasted information regarding user mobility and network load. A point-by-point comparison with aforementioned SDN-based HAS solutions is presented at Table 1.

For Radio Access Networks (RANs), the notion of flexibility and programmability goes beyond routing and forwarding, due to mobility, load and radio conditions and, hence, the role of SDN is crucial for auguring QoE. One of the earliest proposals for softwarizing the access network (and not just the core) has been elaborated in [22], where the "SoftRAN" vision is described. The SoftRAN architecture describes a software-defined controller that abstracts physical base stations, while it conducts radio access mechanisms such as load and interference management in a logically centralized manner. Other examples in the direction of "Software Defined Mobile Networks (SDMN)" are described in [23], where the technical- and business-added value of such schemes is thoroughly analyzed. A flexible 5G RAN architecture based on software-defined control is proposed in [24], where a QoE/QoS mapping and monitoring function dictates the way in which the radio or core networks are (re)configured with respect to the decomposition and allocation of Virtual Network Functions (VNFs). However, the use of SDN in these proposals focuses on MNOs' efficient resource management considering the requirements of the application but not actively interacting with third parties (e.g. VSPs), nor leveraging of VSPs' capabilities.

Assuring a desired QoE in mobile networks may also involve admission control and policy provision, where new connections will be restricted or existing ones will be handed over, based on QoE criteria. Such mechanisms are explored considering femtocell networks in [25], where a "QoS/QoE mapper" creates a statistical profile of relevant QoS metrics

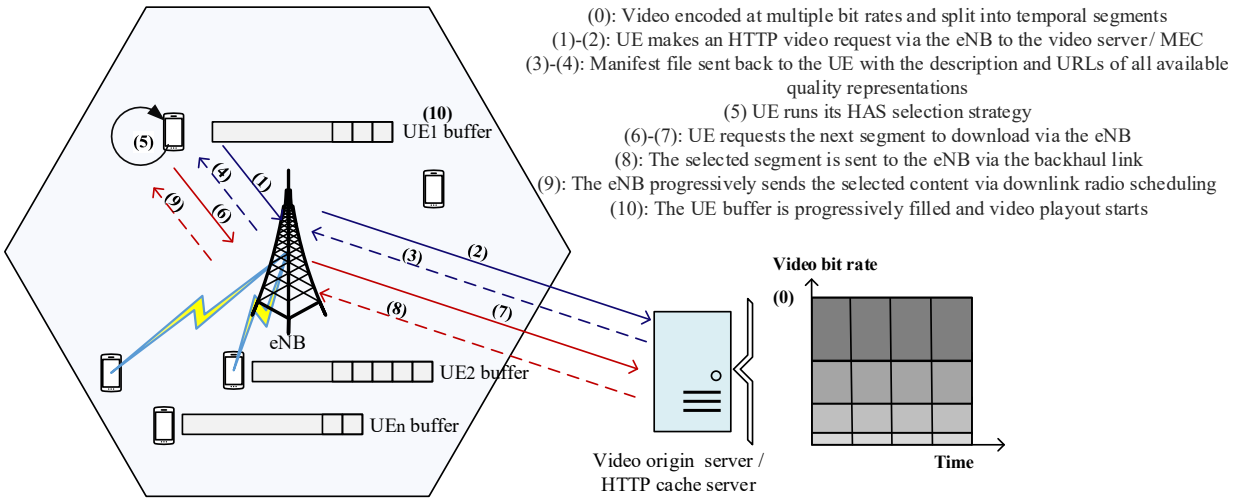

Fig. 2. HAS paradigm in cellular networks.

(e.g. bandwidth availability) and maps this to user satisfaction, defining a QoE-based admission control policy. Moreover, in the context of HAS, [26] describes a novel mobile edge function for transcoding video segments on-the-fly, in the case that this requirement is triggered by a QoE assessor, while [27] introduces an SDN-enabled resource allocation mechanism, called UFair, to fairly orchestrate resources among competing HAS flows.

The adoption of SDN logic in a network can also serve the purposes of application awareness and data analytics. For instance, [28] envisions an architecture relying on a "Video Quality Application", which queries information regarding video content, client information, and network data in order to help the operators better understand their network (e.g. congestion points) through QoE analytics. QoE analytics may also result in a user recommendation engine, as proposed in the case of the "u-map" system [29], where user collected subjective and objective quality metrics are uploaded in the u-map server, followed by feedback to the users regarding the performance of provided services in a specific region. Furthermore, [30] describes an intelligent system that collects real-time user QoE feedback in order to enforce elasticity by scaling up-/down the corresponding cloud resources. In this paper, we build-up on our previous work in [31], introducing a QoE-SDN APP that allows VSPs to program and control the desired QoE with the assistance of the MNO.

A collaboration model between OTT parties and Internet Service Providers (ISPs) is also described in [32], but from a revenue perspective, thus, proving the concept, viability and mutual benefit of such collaboration paradigms. Also, [33], explores the MEC paradigm, proposing a reference architecture for orchestration and management, where Channel State Information (CSI) is sampled to enforce service-level management.

The proposed QoE-SDN APP allows MNOs to dynamically provide network capability exposure feedback to the corresponding VSP based on mobility and rate forecasting mechanisms, proactively guiding in this way the video segment distribution towards particular edge caches as well as the video encoding, in order to avoid stalling events.

III. QoE-SDN APP: HAS Concepts, QoE Functions & SDN Supporting Architecture

A. *HTTP Adaptive Streaming in a cellular network*

HAS is an adaptive streaming technique in which each video is encoded at the server side in multiple versions, called representations or quality layers, which result in a distinct video bit rate and video resolution. Each version is divided into segments of around 2-10 seconds each. A manifest file reporting the availability of different representations and segments is sent to each user upon a video view request. Each user independently requests the next segment in order to maximize the video bit rate, while diminishing the probability of stalling. This decision is taken at the user side based on the manifest file, the user's buffer status, and the user's subjective perception of network congestion.

The HAS strategy followed by typical users is based on the weighted perceived downlink data rate of previously downloaded segments. In a dynamic mobile environment, the achieved data rate is a result of: a) the scheduling algorithm combined with the Modulation and Coding Scheme (MCS), b) the user location in the cell, and c) the momentary load in each cell sector, as a result of competing flows' requests for bandwidth. The HAS operations in a cellular network environment are illustrated in Figure 2, describing step by step the end-to-end logic of video streaming, starting from the user request for watching a video, up to the point that video playout starts at the user side. (The Long-Term Evolution (LTE) notation is adopted in this figure, i.e., evolved eNB – eNB, is the LTE base station, and User Equipment – UE, is the end user).

B. *QoE-SDN APP functions and architecture*

The QoE-SDN APP relies on the SDN architecture [9]-[10], allowing the SDN controller to maintain a corresponding APP template. Such template offers VSPs the opportunity to program their QoE requirements and QoE assessment logic once subscribed. VSPs can then use the QoE-SDN APP to enhance their video segment encoding and distribution procedures by getting network feedback exposed by the

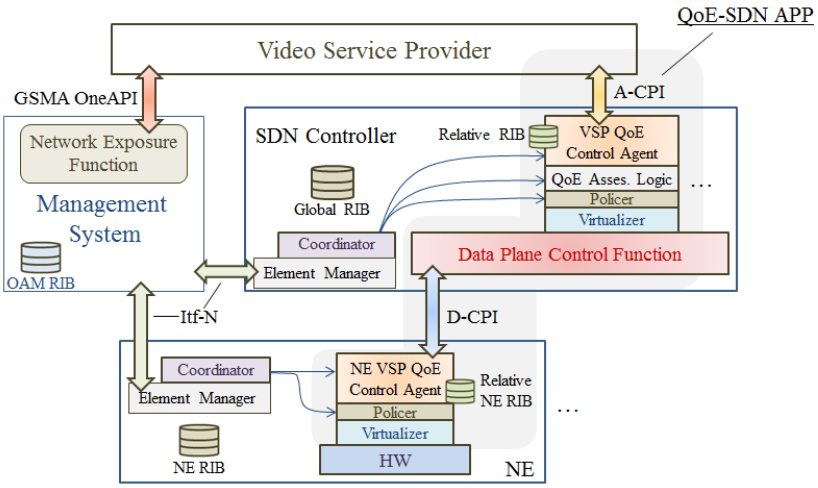

Fig. 3. QoE-SDN APP functions and architecture.

MNOs. The VSP can contact the MNO to request the setup of the QoE-SDN APP via conventional 3GPP management system means, i.e., through the Network Exposure Function (NEF) [34] and an open API, such as GSMA OneAPI [35]. The NEF provides authentication and secure access for VSPs, charging, as well as the means for requesting the QoE-SDN APP. Once a VSP QoE-SDN APP request is authorized, the network management system installs the corresponding SDN-related functions within the SDN controller and within the corresponding Network Elements (NE), e.g. base stations, via the Coordinator function. The Coordinator is contacted through the conventional Itf-N interface and Element Manager. The basic functions of the QoE-SDN APP within the SDN controller are the following:

- ***VSP QoE Control Agent*** is a function that allows VSPs to collaborate with the underlying MNO's infrastructure and resides within the SDN controller. It facilitates the communication and control between the two parties, i.e., providing feedback to the VSP regarding required encoding rates and control capabilities related to the data plane within the MNO infrastructure. The QoE control agent uses a relative global view of the underlying network, i.e., relative RAN Information Base (RIB), considering the abstracted resources allocated to the particular VSP via a virtualizer component.
- ***QoE Assessment Logic*** is the core of the QoE-SDN APP, which can be programmed by the VSP according to the application characteristics and requirements. In particular, the VSP can provide the QoE estimation model and associated parameters, the desired monitoring metrics as well as the policy for retrieving such metrics, e.g. monitoring periodicity, circumstance, etc. The QoE assessment logic is responsible for: a) determining the QoE per application using the MOS scale or appropriate application-specific Key Performance Indicators (KPIs), e.g. stalling events in case of video streaming, b) instructing the Data Plane Control Function to introduce alternations into the allocated network resources with the purpose of maximizing the perceived QoE, and c) determining guidance decisions for the VSP regarding the encoding rate and caching strategy that should be adopted considering future user mobility and network load. The QoE assessment logic relies on feedback collected by the Data Plane Control Function from NE agents or from the MNO management system. Another significant process of the QoE assessment logic is user mobility forecasting that determines future user positions considering the current location, duration of a session and gravity points, i.e., areas with higher user concentration. Based on such forecasted users' locations, the traffic load can be determined at particular RAN points with respect to time, which can be used to guide the encoding rate of VSP content and the video segment distribution, considering also potential re-configurations of the network resources.
- ***Policer*** defines the policy applied to the allocated resources of the VSP and corresponding QoE-SDN APP.

The Data Plane Control Function operates on the allocated resources carrying out all QoE-SDN APP processes related with data acquisition, video segment distribution and potentially network resources' programmability. The data acquisition process takes place periodically or optionally on-demand and can also adjust the input type of collected QoE data including its nature, i.e., real-time measurements or statistics, which is retrieved via agents of specified NEs located in the RAN and in the core network that can capture service-related parameters. These agents can be dynamically configured considering topology changes, e.g. upon a user movement. The QoE-SDN APP functionalities within each NE include a NE VSP QoE Control Agent and Policer, which are responsible for carrying out QoE monitoring and policy processes on the allocated resources, i.e., relative NE RIB, within the NE. An overview of the QoE-SDN APP architecture is illustrated in Figure 3.

Moreover, the SDN controller can communicate with the 3GPP network management system in order to collect the conventional network monitoring information such as interference, load and other KPIs, which can be stored in a RIB creating a global network view.

The Application-Controller Plane Interface (A-CPI) can facilitate programmability for the VSPs in order to program the QoE assessment logic, while the Data-Controller Plane Interface (D-CPI) offers the interaction means between the

SDN controller and the corresponding NE of the MNO, carrying out QoE monitoring as well as resource and policy re-configuration instructions. As for the location of the SDN Controller in the network, it may be elastically distributed as proposed in [36].

IV. System Model & Problem Formulation

A. System model

The system under study is considered as a downlink multi-cell OFDMA cellular network that consists of a ring topology of tri-sector Base Stations (BSs). The entire network comprises $m = 1,2,..,M$ BSs, each co-located with a MEC server, used for caching of video segments. Nevertheless, the notion of a BS may vary especially in 5G, i.e., not restricted to the distributed LTE eNB. In a cloud-RAN environment a BS architecture contains the elements of remote radio head as well as the baseband unit (in 5G the baseband unit is further split into the so called Centralized Unit (CU) and Decentralized Unit (DU)) introducing the notion of a virtual BS. In such a case, the MEC platform can be collocated at the baseband unit (or in a 5G scenario at the CU or DU). We have a set of users U, which can be served by a different BS for a period of time; hence $u(m,t)$ refers to a set of users that is served by BS m at time t. It is assumed that all users have ongoing sessions in time window T. Users follow the Self Similar Least-Action Walk (SLAW) mobility model [37], derived by empirical studies of real-life human-walk traces. One main property of SLAW is the existence of "gravity points" or so-called "clusters", i.e., of popular areas where users tend to accumulate with certain probability. SLAW provides a realistic outlook in terms of the network traffic per square meter, as compared to random mobility models, and can have a realistic application for instance in the case of a mall or train station.

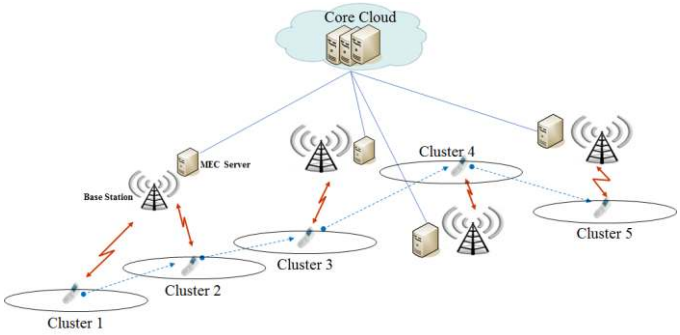

Fig. 4. Exemplary system model.

We divide the total area into $i = 1,2,..I$ clusters based on the SLAW clustering logic. Each cluster can be served by one BS and each BS may include one or more clusters. For each cluster $i \in I, I_m \subseteq I$ is defined as the set of clusters belonging to BS m. The spectrum which is allocated to each BS for wireless access is shared, meaning that each of the $n = 1,..,N$ sub-channels (or Resource Block, RB) is re-used by each base station. An exemplary system overview is illustrated in Figure 4.

B. Generic problem formulation

Initially, we formulate the per-user segment selection strategy of HAS logic as a Knapsack optimization problem, using [38] as a basis. We consider a video split into $s = 1..S$ video segments, while each segment is available in $l = 1..L$ quality layers. Moreover, as mentioned before, there are $u = 1..U$ mobile users in the system and $m = 1..M$ BS (and equal MEC platforms). In this Knapsack problem, the ***value*** which quantifies the level of importance associated with each decision is the quality layer. The higher the index of the quality layer, the more valuable the solution. On the other hand, the ***cost*** of each decision is the size of the video segment needed to transfer to satisfy it. The basic parameters of this problem are represented as:

v_{sl} = the ***value*** associated with segment s of quality l (here: quality is the quality layer index);

c_{sl} = the ***cost*** associated with segment s of quality l (here: the size of segment s of quality l);

$V(t)$ = the total data downloaded until moment t;

R_u = the achieved data rate per user u (in bps);

D_k = the deadline of segment k, meaning that segment k needs to be downloaded by that moment, otherwise a stalling will occur.

In order to estimate $V(t)$, the information about the R_u is required, so:

$$V(t) = R_u * t \tag{1}$$

Moreover, the deadline can be found as follows:

$$D_k = T_0 + k\tau, \forall\, k = 1..S \tag{2}$$

where T_0 is the video start-up delay (initial delay), and τ is the segment duration. The unknown optimization variable in this problem is x_{musl}, which represents the selection of a segment with index number s of quality l that is destined for user u from the BS/MEC m. It is a binary variable, namely a segment with index number s of quality l is either selected or not. Using the above notation, the optimization problem of segment selection is formulated as follows:

$$maximize \sum_{m=1}^{M}\sum_{u=1}^{U}\sum_{s=1}^{S}\sum_{l=1}^{L} v_{sl}x_{musl} \tag{3}$$

subject to:

$$x_{musl} \in \{0,1\} \tag{4}$$

$$\sum_{l=1}^{L}\sum_{m=1}^{M} x_{musl} = 1, \ \ \forall\, u = 1..U, \forall\, s = 1..S \tag{5}$$

$$\sum_{m=1}^{M}\sum_{u=1}^{U}\sum_{s=1}^{k}\sum_{l=1}^{L} c_{sl}x_{musl} \le V(D_k), \ \ \forall\, k = 1..S \tag{6}$$

Equation (3) expresses the optimization goal of maximizing the quality layers of the segments selected, as those will bring higher video bit rates to the users. In terms of the constraints imposed, equation (4) expresses the binary nature of the unknown variable x_{musl}, while equation (5) mandates that each user can request a segment at only one quality layer and from only one MEC platform. Finally, the last constraint (6) expresses the requirement that all segments need to be downloaded before their deadline (on the right-hand side of (6) $V(D_k)$ expresses the maximum amount of data that can be downloaded until the deadline of k, so as to prevent a stalling).

TABLE 2
SUMMARY OF NOTATIONS IN PROBLEM ANALYSIS.

Symbol	Description	Symbol	Description	Symbol	Description
M	Number of BSs / MECs	L	Number of quality layers	t_u	Number of segments per user u
m	BS / MEC index	l	Quality layer index	v_u	User velocity
U	Number of users	S	Number of video segments	$\xi_{m,u}$	Fraction of time user u resides at cell m
u	User index	s	Video segment index	p_{m,K_u}	Probability that user u has an ongoing session at cell m after K_u changes
I	Number of clusters	K_u	Cluster changes for user u	$N_{seg}(m,u)$	Segments per user u stored at m cell
i	Cluster index	$S_{i,j}$	State of a user residing at cluster i after j number of cluster changes	$R_{u(m,t),n}$	Data rate of user u
N	Number of sub-channels	H_u	Handover sequence for user u	$P_{m,n,t}$	BS m transmit power in the sub-channel n at time frame t
n	Sub-channel index	N_u	Number of BS in handover sequence of user u	SR	Segment rate

This optimization problem restricts the existence of any stalling events, due to constraint (6). Therefore, if a stalling event is inevitable, then the optimization problem will be infeasible, namely it will not be solved by an optimizer (e.g. GUROBI). However, if we relax this problem and allow the existence of stalling events, then we can bypass constraint (6), and re-write formula (2) as $D_k = T_0 + k\tau + dur_{stalling}$, $\forall\, k = 1..S$, where $dur_{stalling}$ is the total stalling duration prior to segment k's playout.

Solving such an optimization problem requires a priori perfect knowledge of R_u for all users, and for the whole duration of the video streaming session (namely until all segments S are downloaded), which is impossible in real networks and hence, we carry out solutions that would simplify it.

V. PROBLEM ANALYSIS & SOLUTIONS

A. Problem analysis

Due to the aforementioned challenges, the generic problem can be relaxed by using a predictive mobility model and by de-coupling it into two sub-problems as follows. The related methodology is illustrated in Figure 5 and includes the steps described next. The most important parameters used in the problem analysis are summarized at Table 2, for convenience. At first, we use a Discrete Time Markov Chain (DTMC) model to derive the stationary probabilities of a mobile user residing at each candidate cluster after a given number of cluster changes (denoted as state), while having an ongoing video session. Here, we initially assume that the total area is divided to a grid-type of candidate clusters, which are uniformly distributed (since the information on the actual clusters as defined by the SLAW model cannot be a-priori known). A location state aggregation method similar to [39] is used to model the user mobility. A user can be either idle or can turn to idle after a certain number of cluster changes. With this step we can find the probability of a user following a route, comprising a sequence of cluster "handovers", while streaming a video session.

In the second step, a stationary probability of each state is used (which represents the probability of a user residing in a candidate cluster while having an ongoing video session). We incorporate the outcome of the SLAW model that provides the actual clusters (gravity points), which may be a subset of the candidate clusters (based on the users' traffic) and the route of each user. By using this information and the time of a video session, we extract the fraction of time that each user resides in a cluster, while streaming the video session. This solves the problem of distributing video segments associated with each user to a particular BS (i.e., by adding up all the gravity points/clusters that belong to each BS). Hence, by this step we estimate how many segments per user need to be cached per BS (assuming that a MEC server is co-located with a BS).

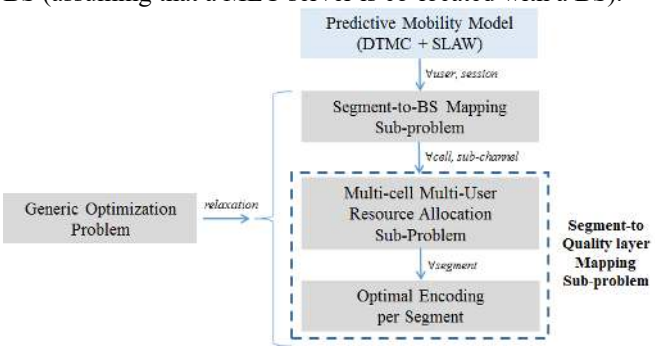

Fig: 5. Optimization problem interpretation.

In the last step, with the knowledge of the segments per user per BS and the connection map of all the users (based on their traces) for a time window, we derive the optimal mapping of segments to quality layers, taking into account the access and queuing constraints. In particular, we first perform conventional multi-cell multi-user interference-aware scheduling, to allow high spectral efficiency for all users within a time window. To enable fairness, we also include multi-channel Proportional Fair (PF) scheduling. The outcome of this problem will provide the optimal power and RB allocation per user to maximize the system's performance. The result will be the optimal user rate per segment, by taking into account the constraint of each user buffer at each time instance to be non-empty, i.e., avoiding stalling events. Since the general objective of this problem is to find the optimal encoding per segment, we map the optimal segment rate to a respective encoding, since the encoding level is closely

coupled with the achievable rate per user and segment, for instance the highest the rate, the highest the encoding of the segment.

1) Discrete Time Markov Chain model

Initially a user is at state S_0, receiving a video session. When a user "handovers" in the next cluster, a "logical chain" also defined as "forwarding chain" increases by 1, which means that the video session of the user will be transmitted over 2 clusters. Hence, the user will receive traffic while residing in these two clusters. If a user, while receiving a video streaming, changes state from S_0 to S_K this means that the video will be transmitted over $i = K$ clusters (where K is set as the maximum number of clusters that the user trespasses). In the transition diagram, shown in Figure 6, $S_{i,j}$ refers to the state of a user residing at cluster i after j number of cluster changes, while having an ongoing video session.

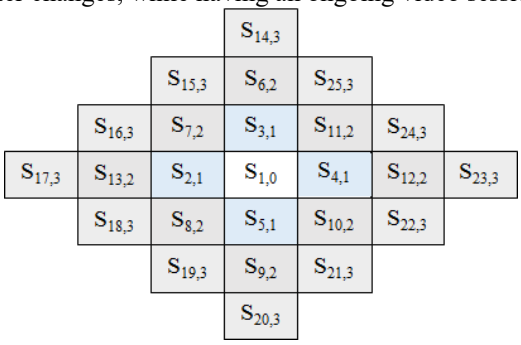

Fig. 6. Grid cluster deployment.

The transition probabilities strongly depend on the cluster layouts, which in our case is a grid. When a user moves out of the coverage area of a cluster regardless the direction of the movement, the cluster logical chain will be increased by 1 (given the transition probabilities). Hence, some states can be aggregated considering this logic. Here, $i^{(j)}$ can be defined as the aggregated state j while increasing the chain by i. For example, $S_{2,1}$ $S_{3,1}$ $S_{4,1}$ $S_{5,1}$ can be aggregated to $1^{(1)}$, the states with $j = 2$ $S_{6,2}$ $S_{9,2}$ $S_{12,2}$ $S_{13,2}$ are aggregated to $2^{(1)}$ due to the fact that the user at these states has a three-fourths probability of increasing its forwarding chain and a one-fourth probability of decreasing it.

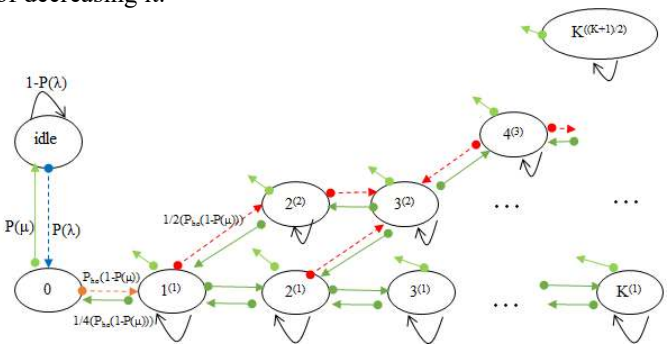

Fig. 7. State transition diagram.

On the other hand, the remaining states of $j = 2$ are aggregated to $2^{(2)}$ since the user has a 50% probability of increasing its forwarding chain and a 50% probability of decreasing it. Figure 7 illustrates the state transition diagram, where we have $i^{(j)}$ aggregated states plus the state for being idle, where K is the maximum number of cluster changes. In this, the following probabilities can be defined: $P(\lambda) = \lambda t$ is the session arrival probability (follows Poisson distribution), $P(\mu) = \mu t$ is the session departing probability and $P_{ho} = Mr \cdot t$ is the cluster HO probability based on the user mobility rate Mr. When a user is at the idle state, the probability of initiating a session during a time slot t is $P(\lambda)$. Thus, the current cluster becomes the local point, and the user moves to the state $0^{(1)}$ with this probability and stays at the idle state with the probability $1 - P(\lambda)$ at the end of a time slot. When a session is initiated, the probability of a session terminating during a time slot t is $P(\mu)$. Let π_{idle} and $\pi_{i,j}$ be the probability of being at state S_{idle} and $i^{(j)}$ respectively. Based on the transition diagram, the stationary probabilities can be found by solving the set of equations below:

$$
\begin{aligned}
\pi_{idle} &= \big(1 - P(\lambda)\big)\pi_{idle} + \mathrm{P}(\mu)\sum_{i=0}^{K}\sum_{j=1}^{(i+1)/2}\pi_{i,j} \\
\pi_{0,1} &= (1 - P(\mu) - (1 - P(\mu)P_{ho})\pi_{0,1} + P(\lambda)\pi_{idle} \\
&\quad + \frac{1}{4}\big(1 - \mathrm{P}(\mu)\big)\mathrm{P_{ho}}\pi_{1,1} \\
\pi_{1,1} &= (1 - P(\mu) - (1 - P(\mu)P_{ho})\pi_{1,1} \\
&\quad + \big(1 - \mathrm{P}(\mu)\big)\mathrm{P_{ho}}(\pi_{0,1} + \frac{1}{4}\pi_{2,1} \\
&\quad + \frac{1}{2}\pi_{2,2}) \\
\pi_{\iota,1} &= (1 - P(\mu) - (1 - P(\mu)P_{ho})\pi_{\iota,1} \\
&\quad + \frac{1}{4}\big(1 - \mathrm{P}(\mu)\big)\mathrm{P_{ho}}(\pi_{\iota-1,1} + \pi_{\iota+1,1} \\
&\quad + \pi_{\iota+1,2}) \\
\pi_{\mathrm{K},1} &= (1 - P(\mu) - (1 - P(\mu)P_{ho})\pi_{\mathrm{K},1} \\
&\quad + \frac{1}{4}\big(1 - \mathrm{P}(\mu)\big)\mathrm{P_{ho}}\pi_{\mathrm{K}-1,1} \\
\pi_{2,2} &= (1 - P(\mu) - (1 - P(\mu)P_{ho})\pi_{2,2} \\
&\quad + \big(1 - \mathrm{P}(\mu)\big)\mathrm{P_{ho}}(\frac{1}{2}\pi_{1,1} + \frac{1}{4}\pi_{3,2}) \\
\pi_{3,2} &= (1 - P(\mu) - (1 - P(\mu)P_{ho})\pi_{3,2} \\
&\quad + \big(1 - \mathrm{P}(\mu)\big)\mathrm{P_{ho}}(\frac{1}{2}\pi_{2,1} + \frac{1}{2}\pi_{2,2} \\
&\quad + \frac{1}{4}\pi_{4,2} + \frac{1}{2}\pi_{4,3}) \\
\pi_{idle} &+ \textstyle\sum_{i=0}^{K}\sum_{j=1}^{(i+1)/2}\pi_{i,j} = 1
\end{aligned}
\tag{7}
$$

2) Segment-to-BS mapping sub-problem

By solving the stationary probabilities, the probability of a user starting a video session at a certain cluster and terminating this session at cluster K after j cluster changes can be calculated. However, there is information that this model cannot capture, including:

- The actual/estimated trace of each user. Each user may start from a distinct point and may have different direction. The cell handover probability will require information on which clusters are involved.
- The number of changes required between the establishment and termination of the video session. For

each user, this would be different based on the time of the video and the mobility of the user.

Furthermore, each of the clusters may belong to different BSs or to the same BS. So, in order to find the segments that need to be handled by one BS (local MEC cache) we need to find the proportion of time the user stays at the BS coverage so as to transmit the requested segments in advance. It is assumed that both aforementioned points can be derived by the SLAW model and by the knowledge of the video duration. In particular, let the handover sequence for user u be: $H_u =< h_1, h_2, \dots, h_m >$ and the number of BSs in this sequence is N_u. Both vectors are assumed to be known using the mobility prediction model.

Also, let the number of segments per user based on the video content be t_u, and T_{max} the maximum number of segments for a video transmission. For a given probability $p_{m,j} = \sum_{i \in I_m} \pi_{i,j}, \forall j \in K$, and by taking into account H_u, and user velocity v_u, we can translate the probability of a user moving between clusters to a function corresponding to the fraction of time that each user resides at each cell:

$$\xi_{m,u} = f\left(p_{m,K_u}, H_u, v_u\right) \tag{8}$$

Note that K_u is the number of cluster changes from $u(1,1)$ till $u(m, t_u)$ and p_{m,K_u} denotes the probability that user u has an ongoing session at cell m after K_u cluster changes. The number of segments per cell and user can be easily derived by the following equation:

$$N_{seg}(m,u) \triangleq \left\lceil \xi_{m,u} t_u \right\rceil \tag{9}$$

where $N_{seg}(m,u)$ shows how many segments per user u are going to be stored at m cell (or m cache). The sub-problems remaining to be investigated based on this analysis are: a) how to allocate resources to maximize system's throughput while providing fairness to all the users, and subsequently b) how to map the estimated performance to encoding levels.

3) Segment-to-Quality layer mapping sub-problem

As soon as the number of segments per BS per user is known from the previous step, the segments can be cached in the respective queues at the local caches, e.g. MECs. However, there are two challenges which need to be solved:

- How to allocate resources to all users so as to achieve the maximum quality per segment and per user, while keeping the intra/inter-cell interference low, and assuming dynamically changing wireless channel conditions and availability.
- How to ensure that all users at each time instance have non-empty buffers. In conventional schedulers, where the aggregated rate is optimized by multi-user diversity, users with favorable channel conditions might be preferred instead of users with low channel quality. This might lead to resource starvation and empty buffers for some users with low channel qualities in some time instances.

Taking into consideration the aforementioned challenges, the problem formulated focuses on how to maximize QoE (i.e., quality layers) for all users and all segments, taking into account the user fairness requirements and the expected load per cell (by the expected segment provisioning of the previous step). Afterwards, it is straightforward to extract the encoding level per segment based on the optimal rates.

a) Multi-cell multi-user resource allocation sub-problem

The problem of network optimization can be translated to a weighted sum rate maximization problem, where the weighting factors can be tuned accordingly to maintain fairness. The problem is to find the optimal resource allocation per user for all cells, assuming the locations of the users in t timeslots and the demand of the users in terms of segments per cell. Initially, we define the binary variable $b(u,m,t)$ which is 1 if the user u is served by cell m at time slot t; and 0 otherwise. Based on this, the Signal to Interference plus Noise Ratio (SINR) can be formulated as:

$$SINR_{u(m,t),n} = \frac{\sum_m b(u,m,t) P_{m,n,t} G_{m,u(m,t),n}}{\left(\sum_{m' \neq m \in M} P_{m',n,t} G_{u(m,t),m',n} + \eta\right)} \tag{10}$$

Here, $P_{m,n,t}$ is the serving BS transmit power at time slot t and $G_{m,u(m,t),n}$ is the channel gain between BS m and user u in the sub-channel n at time frame t. Moreover, η is the power of the thermal noise and m' accounts for the interferer BS in a specific sub-channel n. Also, $R_{u(m,t),n}$ accounts for the achievable user's data rate in terms of spectral efficiency on each sub-channel (using the truncated Shannon capacity formula) multiplied by the unit bandwidth (BW) per resource chunk, and is represented as:

$$R_{u(m,t),n} = BW log_2(1 + \rho SINR_{u(m,t),n}) \tag{11}$$

where $\rho = -1.5/\ln(5 \cdot BER)$ is the SNR gap to data-rate, linked to a particular target Bit Error Rate (BER).

The optimization problem is to find the optimal resource allocation (subcarrier and power control) in order to maximize the weighted sum-rate:

$$\max_{A,P_{m,n,t}} \sum_{t \in T} \sum_{m=1}^{M} b(u,m,t) \sum_{u(m,t)=1}^{U_{m,t}} \sum_{n=1}^{N} (w_{u(m,t),n} R_{u(m,t),n,t} \, a_{u(m,t),n,t}) \tag{12}$$

$subject\ to$:

$$a_{u(m,t),n,t} \in \{0,1\}, \forall m \in M, n \in N, \forall t \tag{13}$$

$$\sum_{n=1}^{N} P_{m,n,t} \leq P_{m,max}, \forall t \tag{14}$$

$$\sum_{u(m,t) \in U_{m,t}} a_{u(m,t),n,t} \leq 1, \forall\ m \in M, n \in N, \forall t \tag{15}$$

$$\sum_t \sum_n R_{u(m,t),n,t} \geq N_{seg}(u,m) R_{min,u} \tag{16}$$

$$R_{max,u} \geq \sum_n R_{u(m,t),n,t} \geq R_{min,u}, \forall m, t \tag{17}$$

where $A = \{a_{u(m,t),n,t} | a_{u(m,t),n,t} \in \{0,1\}\}$ is the binary variable corresponding to the allocation decision for the sub-channel n to user u that belongs to cell m at instance t, i.e., $a_{u(m,t),n,t} = 1$ if user u which resides in cell m *at* t is allocated sub-channel n, where $N = \{n | \forall n \in 1,2,..,N\}$ is the set of sub-channels. Moreover, $P_{m,n,t}$ accounts for the cell transmit power per sub-channel. Hence, the optimization problem is a weighted sum-rate maximization over the network in presence of inter-cell interference subject to power constraint of $P_{m,max}$ per node m as in (14) and orthogonal allocation at intra-cell as in (15). Furthermore, according to the expected positioning of users in different time instances we can provision how many segments are going to be served by cell m to user u. So, we have constraint (16) as minimum

demand rate, which is proportional to the number of segments per cell-user link. $R_{min,u}$ corresponds to the minimum rate with the lowest quality. Finally, constraint (17) accounts for the requirement to have at every time instance all the users to be served with at least the minimum rate $R_{min,u}$, so as to avoid non-empty user buffer, i.e., stalling events, while not exceeding $R_{max,u}$ which denotes the maximum rate in order to avoid buffer overflow.

Concerning the intra-cell scheduling, PF scheduling is used for a multi-channel system in each cell to provide a fair allocation of resources between multiple users. Each user feedbacks the achievable data rate to its BS per sub-channel $n = 1,2,\dots,N$ per timeslot and the BS calculates the ratio of the achievable spectral efficiency $R_{u(m,t),n,t}$ to the average spectral efficiency $\bar{R}_{u(m,t),t}$ for each user and time-slot $t = 1,2,..,T$. Thereafter, each BS forms a matrix consisting of the ratios of the achievable rate to the average spectral efficiency for the allocated users, corresponding to their individual weighted rates:

$$w_{u(m,t),n}R_{u(m,t),n,t} = \frac{R_{u(m,t),n,t}}{\bar{R}_{u(m,t),t}}, \forall u \in U, \forall n \in N, \forall t \in T, \forall m \quad (18)$$

b) Derivation of optimal encoding per segment

The solution of the previous sub-problem will provide the optimal resource allocation and power vectors per user, segment and timeslot. Assuming that power can be fixed per RB (for Reuse-1 cases, this is straightforward) as $P_{m,n,t} = \frac{P_{max}}{N}, \forall m,n,t$, the output is $a^*_{u(m,t),n,t}, \forall u,m,n,t$. So, the optimal estimated rate per user can be written as:

$$Rate^*_{u,t} = \sum_{n=1}^{N} R_{u(m,t),n,t}a^*_{u(m,t),n,t} \quad (19)$$

Assuming that each segment may last one or more timeslots, the per segment achievable rate (defined as *SR*) accounts for the summation of the rates for all timeslots that each segment s will be transmitted to the user (set of timeslots at segment *s* of user u is defined as $T_{u,s}$). So, the rate can be written as:

$$SR^*_{u,s} = \sum_{t \in T_{u,s}} Rate^*_{u,t}, \forall \quad (20)$$

Considering that different encoding levels require a certain achievable rate range, it can be easily shown that from the per segment rate we can derive the encoding per segment. Let $l = 1,2,..,L$ encoding levels and the lower and upper rate bounds per encoding level: $SR_{l,min}$ and $SR_{l,max}$. The per segment rate can be now translated to encoding level per segment of index *s* and user u (defined as $seg(u,s)$) if it lies within the aforementioned thresholds. In particular:

$$\forall s,l: seg(u,s) \in l \rightarrow SR^*_{u,s} \in [SR_{l,min}, SR_{l,max}] \quad (21)$$

B. Mobility prediction & rate adaptation solutions

After the relaxation of the original problem into less complex sub-problems, this section provides the solutions to the aforementioned sub-problems. In particular, a mobility prediction solution is provided as well as a rate adaptation algorithm for the segment-to-BS and segment-to-quality layer mapping problems, respectively. For the first solution, we implemented a mobility prediction algorithm based on the SLAW mobility model [37] taking advantage of the "gravity points" or "clusters", which tend to accumulate users with certain "self-similar waypoints".

Algorithm 1: SLAW-based mobility prediction

- Set of all waypoints based on SLAW pattern: $w = \{w_i, i = 1..W\}$
- Set of all clusters $c = \{c_k, k = 1..C\}$, where $c_k = \{w_m, \dots, w_l\}$, so that $d(w_i, w_j) < clustering\ range$ for all $w_{i,j} \in c_k$
- Set of visited clusters: v'
- Starting user waypoint: $s \in v'$
- Present user waypoint: $p = (x_p, y_p) \in v'$
- $p \leftarrow s$
- Identify to which cluster c_k the waypoint p belongs, $v' \leftarrow c_k$
- Set $cluster\_ratio$ (percentage of clusters to visit), $velocity$, $prediction\_interval$

for $cluster = 1:C$

- Calculate each cluster's popularity as the number of waypoints per cluster over the total clusters available: $P = \frac{|c_k|}{C}$
- Order the first $\frac{C}{cluster\_ratio}$ number of clusters in descending popularity → set v of clusters to visit
- Calculate cluster centers $\bar{c}$: $\bar{c_k} = mean(w_m, \dots, w_l)$

end for

while v is not empty ***do***

- Calculate distances from p to the center $\bar{c}$ of all unvisited clusters v: $d(p,v) = \|p - \bar{c}\|^2$, for all $v \neq v'$
- Order clusters in increasing distance omitting the one with the least distance (which is the current cluster)
- The next movement prediction is towards cluster c_k which is the first element of the previous vector
- Future predicted position: $(x_f, y_f) = (x_p + velocity * prediction\_interval * cos\varphi, y_p + velocity * prediction\_interval * sin\varphi)$ where $\varphi = \tan^{-1}\frac{\bar{y_{c_k}} - y_p}{\bar{x_{c_k}} - x_p}$
- $v' \leftarrow v' \cup c_k$
- $v \leftarrow v - \{c_k\}$

end while

In the proposed QoE-SDN APP, mobility prediction is introduced to guide QoE control decisions at the VSP and network layer. The mobility prediction algorithm adopted is based on SLAW and runs per user relying on information that the MNO has at its disposal, i.e., the popularity of visited locations (available from statistics kept at the MNO) and the user current positions. Such information can be fed to the QoE assessment logic via the SDN controller, which communicates with the network management system via the Itf-N interface. In detail, the mobility prediction algorithm uses as input the set w of visit-able waypoints and the set c of clusters, with the objective to find the next visited cluster per user, based on the user's current position p. All clusters that a user can potentially visit are sorted by popularity, with the logic that more waypoints will be accumulated in the most popular clusters. Each user is going to visit a total of v clusters, subject to the trace generation duration. Then for each user, the algorithm estimates the distances $d(p,v)$ between the user's position p and the center of each yet unvisited cluster, c_k, ordering them in increasing distance from p. The predicted next movement will be towards the cluster center at the smallest distance out of this list, while the exact position for the next prediction interval will be a function of the user's velocity and direction. The operation of the SLAW-based mobility prediction is illustrated in Algorithm 1.

Based on the user mobility prediction we then estimate the corresponding data rate in order to identify and proactively

handle congestion conditions in the RAN, considering bandwidth conditions on a cluster-basis, as elaborated in Algorithm 2. As mentioned, Algorithm 2 uses as input the mobility prediction estimations, which reveal the set of clusters that each user can potentially visit during a pre-defined future time window. Based on such information, it can approximate the rate for each user as the mean data rate of the cluster that it will reach. In this way, when a user moves from a low-congested to a higher-congested cluster, the estimated data rate will be conservative (ensuring no stalling events), i.e., it may be predicted lower compared to what each user would subjectively perceive, since this prediction will be based on the mean data rate experience of the to-be-visited cluster. A similar idea may be found in [40], where a mobility-prediction-aware bandwidth reservation (MPBR) scheme is proposed. This scheme predicts when a user will perform handovers along his movement path, while a rate estimation scheme calculates the available bandwidth along this path in order to drive call admission control with QoS guarantees for ongoing calls.

Algorithm 2: Congestion-aware proactive rate estimation

- Set of future predicted positions per user $f = (x_f, y_f)$
- Set of cluster centers $\bar{c}$
for $user = 1: all$
  for $step = current\_tti: current\_tti + prediction\_interval$
    - Read the next predicted position of the UE: $f = (x_f, y_f)$
    - Find cluster k closest to this position: $arg_k \min\{\|f - \bar{c_k}\|^2\}$
    - Identify other users belonging to the same cluster k
    - Estimate the mean data rate from all users in the cluster, r, during the latest second
    - The predicted rate for this user for this step is equal to r
  end for
end for

Such rate forecasting estimates can then help the QoE assessment logic to guide VSPs to take proactive service provisioning decisions, as will be shown later by the evaluation use cases in Section VI. The MNO in turn is aware of the achieved throughput per user, as each user positively or negatively acknowledges the scheduled packets per Transmission Time Interval (TTI) to the serving base station.

VI. Simulation Setup & Evaluation Analysis

A. Simulation setup

The performance evaluation is carried out using the Vienna simulator, a 3GPP-compliant LTE system-level simulator [41], in which we have developed the proposed QoE-SDN APP introducing the QoE assessment logic that contains the mobility prediction and rate estimation algorithms, as well as the corresponding SDN programmability functionalities for providing feedback to VSPs regarding the HAS encoding rate and segment distribution. For the purposes of the simulations, the complete end-to-end HAS logic (i.e., video file encodings at different rates, streaming logic, user HAS strategies, user buffers with a maximum buffer size and a minimum playout threshold, etc.) is adopted considering also caching logic within BSs that represent MEC platforms, while the user distribution and mobility are implemented using the SLAW model. For ensuring fairness, PF scheduling is used. The simulation specific parameters are summarized at Table 3.

TABLE 3
SIMULATION PARAMETERS.

Parameter	Value
SLAW parameters (their meaning is explained in [37])	
Number of waypoints	1000
Hurst parameter	1
Alpha, Beta	3, 1
Pause time	0 s
Clustering range	40m
Trace generation time	Set to simulation time of 1 min
Maximum area size	Set to simulation area
User speed	1.38 m/s
Network parameters	
Bandwidth available	20MHz
Radio scheduler	Proportional fair
Network geometry	7 cells with 3 sectors each
Number of eNBs, M	21
Inter-eNodeB distance	500 m
Number of mobile users	105 users
Initial user positions	SLAW-based
Prediction interval	4 s
Traffic distribution	FTP: 20%, HTTP: 20%, VoIP: 20%, Video streaming: 20%, Gaming: 20%
Application parameters	
Max buffer size	64 s
Min buffer playout threshold	2.5 s
Segment duration	2 s
Available video bit rates (representations)	235, 375, 560, 750, 1050, 1750, 2350, 3000, 3850, 4300 kbps
First segment selection	At lowest quality layer
QoE model	Equation (22)
Video utility model	Equation (23)
Simulation parameters	
Number of SLAW topologies tested per use case	4 randomly created SLAW topologies

We concentrate our evaluation on HAS, considering both user and network KPIs. The former include QoE-related metrics that the end user perceives, while the latter focus on overall network performance metrics. For the end users' experience, we use QoE insights extracted via subjective experiments, which have led to the identification of the following main KPIs affecting the video delivery quality [4]. In order of importance, these KPIs are:

- **Stalling events** refer to the interruption of video playback that occurs when the playout buffer runs out, and it is the most significant QoE degradation factor. In the case of YouTube videos, a QoE model has been derived based on the number and duration of stalling events [42]. This model follows the "IQX hypothesis", i.e., a perception-centric QoS-to-QoE mapping [43]. According to the IQX hypothesis, the relationship between QoE and QoS is negative exponential, in the form:
$$MOS = \alpha \cdot e^{-\beta(L)\cdot N} + \gamma \quad (22)$$
where N and L are the number and duration of stalling events, respectively, and α, γ and $\beta(L)$ are coefficients derived from the experimental process. Specifically, $\beta(L)$ has a linear relation with the stalling duration L, which is defined as: $\beta(L) = 0.15 \cdot L + 0.19$ for the case of YouTube. Typical values for α and γ coefficients are $\alpha = 3.5$ and $\gamma = 1.5$.
- **Video characteristics** that shape QoE concentrate on the

resolution and video bit rate, i.e., a higher resolution and video bit rate result in more satisfied users. A video utility model can be used to represent the video quality, using as input the video resolution and mean bit rate [44]. For the cases of 720p videos, the video utility function is as follows:

$$VQ_{720p} = -4.85 * VR^{-0.647} + 1.011 \quad (23)$$

where VR is the video bit rate experienced by the user. Video utility takes values between 0 and 1, where 1 represents the highest quality. Moreover, the percentage of time at each quality layer that the user spent while watching a video is another meaningful KPI, strongly correlated to the resulting video bit rate. Nevertheless, the impact of unexpected stallings is much more severe than a controlled bandwidth reduction on the video bit rate [45]; therefore stallings are the main QoE performance KPI we judge in the evaluation subsection.

- **Adaptation and stability,** consider cases that attempt to improve the overall video viewing quality and avoid stallings, in where parameters such as video resolution and video bit rate are subject to adaptation. As a result, unexpected changes among the selected video encodings may be enforced, as guided by the bandwidth availability, thus causing instability. These changes in the encoding quality are referred to as quality *switches*. The frequency of these switches should be kept at a minimum, allowing a smooth and stable streaming experience with controlled uninterrupted flow [4],[44]. Moreover, the *amplitude*, which describes the depth of the switch (i.e., the gap between two subsequent layers), is another useful stability metric. The smaller the amplitude, the more gradual and smoother the adaptation and thus, the better viewing experience. In order to measure stability, the Switching Impact (SI) is used [44]:

$$SI = |VQ - VQ'| * e^{-0.015*(t-t_i)} \quad (24)$$

where VQ' is the video utility after a switch, and t_i is the instant of the quality switch i. This metric integrates a) the forgiveness effect of a switch, in the sense that the impact of a switch fades out over time (exponential term), as well as b) the impact of amplitude ($|VQ - VQ'|$). Summing up the SI for all switches, we get the *accumulative SI*. Since VQ ranges from 0 to 1, also SI ranges from 0 to 1, and thus the accumulative SI ranges from 0 to switches_number.

The average system throughput is a generic quality indicator typically not sufficient to accurately capture the video streaming experience from the network perspective. For instance, considering the following two extreme cases: a) all users are served with a medium-quality layer, versus b) half users are served with a high-quality layer and half with a low-quality layer, both cases lead to the same average experienced throughput. However, the QoE among users significantly differs. Hence, a useful complementary KPI is fairness in the achieved QoE values (i.e., MOS), which can be estimated using Jain's index as follows, when there are U users in the system:

$$QoE\ fairness = \frac{(\sum_{u=1}^{U} MOS_u)^2}{U * \sum_{u=1}^{U} {MOS_u}^2} \quad (25)$$

For our evaluation analysis, we adopted the following three use cases, considering first the HAS segment selection enforcement problem, then the segment encoding and placement, and finally the proactive segment selection and placement. These use cases fall into the Bitrate Guidance category [16].

In line with the proposed architecture (Section III), the communication flow that realizes the proposed use cases, once the QoE-SDN APP is setup by the VSP, is as follows: (1) The QoE assessment logic requests a periodic estimate of the data rates and positions of users of interest, i.e., VSP customers. An MNO can facilitate this requirement by the Data Plane Control Function via the D-CPI interface. (2) The MNO installs monitoring rules to any involved eNBs in order to collect and provide, in response, this information back to the QoE assessment logic (namely, eNBs serve as Network Elements). (3) The QoE assessment logic then predicts the data rate that each monitored HAS user is expected to achieve, based on per-cluster rate forecasting and mobility prediction (using Algorithms 1 and 2). (4) Finally, the QoE assessment logic enforces the segment selection of each user (use case 1), the segment encoding and placement (use case 2), or the proactive segment selection and placement (use case 3) and passes this information to the VSP side by the QoE control agent via the A-CPI interface. In more detail:

1) Use case 1: Segment selection enforcement demonstrates the potential of assisting users in their HAS segment selection decisions. The information exposed by the MNOs to the VSPs is meant to help users take more informed decisions reflecting how the user perceived rate is expected to evolve. Such a procedure can be useful in cases of unexpected or rapid congestion, i.e., when the conventional segment selection decisions might prove detrimental and lead to stalling events. As explained before, the QoE assessment logic collects the desired KPIs periodically and forecasts the expected rate based on the estimated per-cluster rate and mobility prediction. Such estimated data rate is then used to guide the VSPs either by directly replacing the segment selection of particular users if required, or by indirectly limiting their available options to select (in the case video streaming is about to begin and the manifest file is prepared). Therefore, the suggested segment selection enforcement that takes place serially per user overrides the user's selection and delivers a safer segment alternative. Hence, the goal of this scheme is: a) to reduce stallings by proactively decreasing the quality layer that a user has individually selected based on his current perception of the network, if rate was overestimated, or b) to maximize the quality layer selection if rate was underestimated.

2) Use case 2: Segment encoding and placement considers the network-aware encoding and potential distribution of segments to MECs based on expected network conditions within each BS coverage area. As stated before, HAS requires the encoding of the video content at multiple bit rates (quality layers), which are pre-defined in a network-agnostic manner. To avoid caching video streams at all available quality layers and save backhauling as well as edge cloud resources, there is a need to consider the network resource variation in time and location within the process of distributing video segments, e.g. encode and place low video quality layers to high congested cells and vice versa. In addition, by limiting the available representations based on network resource prediction, users

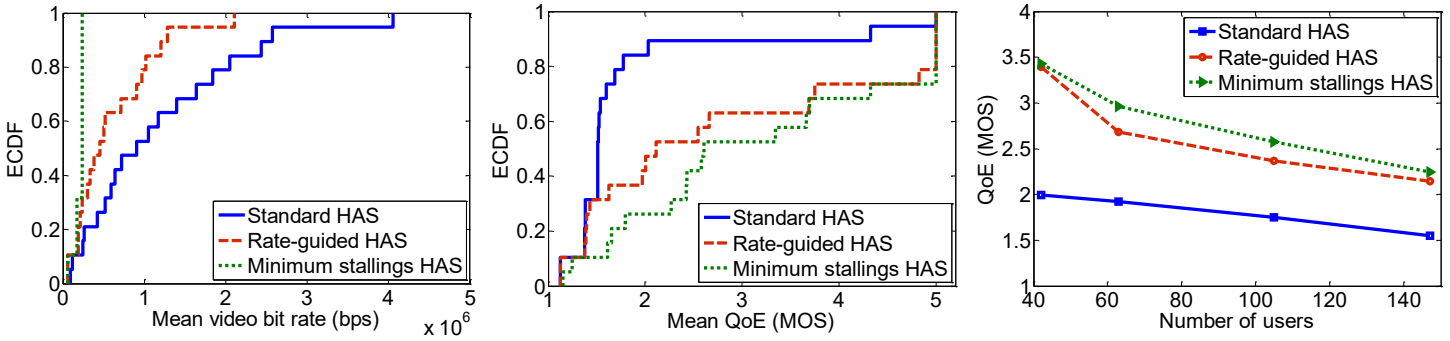

Fig. 8. Use case 1 evaluation results - (a) ECDF of mean video bit rate for all users, (b) ECDF of MOS for all users, (c) QoE for varying load.

will be led to take more accurate HAS decisions reducing stalling events and increasing QoE. Segment encoding and placement decisions will be valid for a next interval, and then the entire process will be repeated. In cases of live video streaming, such a procedure could lead to significant backhaul bandwidth savings.

3) Use case 3: Proactive segment selection and placement: in contrary to use case 2 that periodically performs a massive caching of video segments for all users, the third use case proactively enforces the caching of pre-recorded video segments in advance destined for a user, i.e., before the user requests a segment. The rational is to proactively cache appropriate segment encodings (based on rate estimation) in appropriate edge cloud platforms or MEC locations considering user's mobility prediction, thus avoiding the backhaul delay, while regulating congestion on backhaul links. The appropriate segment is placed on the MEC server closer to the user, considering the user mobility prediction with respect to a predefined prediction window, and will be offered to the user replacing the original segment selection that may lead to stalling events.

B. *Evaluation results*

Simulations were conducted comparing the aforementioned use cases with a standard, i.e., state of the art, version of HAS and with a conservative HAS variation that introduces minimum stalling events. The evaluation process was performed for each use case separately considering measurements in terms of various meaningful KPIs. In these use cases, the partial sub-problems introduced in Section V are solved; specifically, a) the segment-to-BS mapping sub-problem is approached via the mobility prediction algorithm (Algorithm 1) in use cases 2 and 3, determining where appropriate segment encodings will be placed (i.e., in which MEC platforms), and b) the segment-to-quality layer mapping sub-problem is approached via the rate adaptation algorithm (Algorithm 2) in all use cases, determining the appropriate segments to be selected (use cases 1 and 3) or to be encoded (use case 2).

1) Evaluation analysis of use case 1: Segment selection enforcement considers three different HAS variations: a) the *standard HAS*, where always 10 representations are available per segment (this is the baseline strategy), b) the *rate-guided HAS*, where the segment selection of each user is guided by the QoE-SDN APP providing feedback to the VSP based on mobility- and cluster-based rate estimations, and finally c) the *minimum stallings HAS,* where only the lowest bit rate is requested (here 235kbps per segment), leading to the least number of stalling events at the cost of very low video bit rate. The latter case represents a benchmark in terms of stalling events taking into account the specifics of the simulation environment. The evaluation results are illustrated in Figure 8 as well as in Table 4. Figure 8 presents: a) the ECDF for mean video bit rate in the system, b) the ECDF for mean user QoE in the MOS scale, and c) the mean user QoE for increasing traffic load, while Table 4 includes the mean values for various significant KPIs, such as mean MOS, stalling probability, video utility, fairness, switching impact, etc.

As shown in Figure 8(a), the experienced mean video bit rate per user is higher for the standard case, followed by the rate-guided HAS (with the QoE-SDN APP) and the minimum stallings HAS. This is due to the fact that the standard HAS case allows users to select segments with a higher quality layer in contrast with the proposed rate-guided HAS, which follows a more conservative approach, guiding users to select segments with a lower quality, as shown in Table 4 (the mean quality layer downloaded for the standard case is 5.65, while for the rate-guided one it is 2.66). However, the proposed rate-guided HAS as well as the minimum stallings HAS allow more segments (i.e., more playtime) to be buffered, preparing the video player better for imminent congestion and worse channel conditions. Therefore, such higher quality layer selection for standard HAS, is the result of overestimated subjective bandwidth calculations that mislead users to request segments with a higher quality layer, and thus, eventually experience stalling events. This effect is illustrated in Figure

TABLE 4
KPI ESTIMATIONS FOR USE CASE 1.

HAS logic	Mean video bit rate (bps)	Mean quality layer	Mean QoE (MOS)	QoE fairness	Mean video utility	Mean stalling probability	Mean stalling duration (sec)	Average stallings per user	Mean accumulative SI
Standard	1.20E+06	5.65	1.83	0.76	0.82	0.89	21.67	1.87	0.25
Rate-guided	6.13E+05	2.66	2.79	0.75	0.79	0.63	23.39	0.85	0.44
Min stallings	2.04E+05	1.00	3.15	0.78	0.75	0.57	26.52	0.64	0

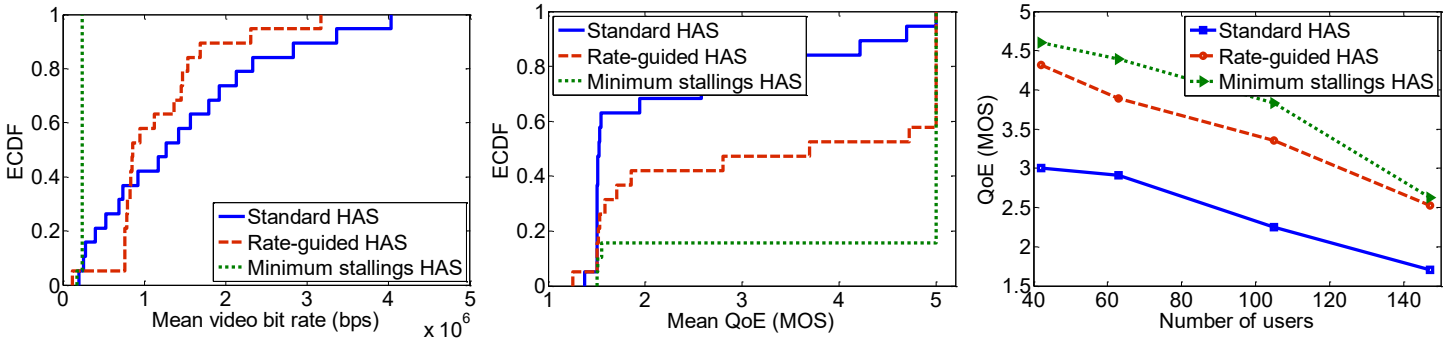

Fig. 9. Use case 2 evaluation results - (a) ECDF of mean video bit rate for all users, (b) ECDF of MOS for all users, (c) QoE for varying load.

8(b), where the QoE model of equation (22) gives an estimation of the MOS as a function of the number and duration of stalling events (their values are shown in Table 4), showing the benefits in terms of QoE for the proposed rate-guided HAS. In fact, for the standard HAS, more than 80% of the users have $MOS < 2$. Since stalling is the most important QoE shaping factor, such an improvement is highly desirable for the users (and therefore, the VSPs). Moreover, even though the minimum stallings HAS leads to the lowest stalling rate (and thus, phenomenally higher QoE), it is not an acceptable solution, since it completely ignores the adaptation logic of HAS providing no video utility improvements even in low congestion scenarios.

In Figure 8(c) we also present results regarding the performance of the QoE-SDN APP when the traffic load (i.e., number of served users) in the system increases. We observe that the rate-guided strategy always yields higher QoE scores compared to the standard case, and that this benefit is higher for lower traffic loads. Also, as it was expected, the higher the load, the lower the QoE for all cases. Finally, as presented at Table 4, all strategies depict very similar fairness and stability levels (note that mean accumulative *SI* is not a percentage value).

2) Evaluation analysis of use case 2: Segment encoding and placement demonstrates high benefits in terms of QoE preserving a high video bit rate, while it can save backhaul capacity. As before, three HAS variations are considered: a) the *standard HAS*, where all 10 quality layers are encoded and cached, b) the *rate-guided HAS*, where the cached amount and video bit rate of the quality layers are driven by per-cluster rate estimation, and c) the *minimum stallings HAS*. Figure 9 and Table 5 illustrate the compared evaluation results.

Similarly to use case 1, as shown in Figure 9(a), the standard HAS provides the highest bit rate, since segments with higher quality layers are selected (5.63 for the standard HAS as opposed to 2.28 for the rate-guided HAS) at the cost of QoE, since MOS is tightly connected to stalling events (Figure 9(b)). In fact, according to Table 5, the mean QoE value is higher by 1.1 MOS value for the rate-guided case. For the same reasons, the minimum stallings HAS assures a better MOS since it always selects segments with the lowest quality layer, which however impacts significantly the user-experienced video bit rates and is not a viable adaptive video streaming logic. It is also observed, that the proposed rate-guided HAS provides a fair trade-off between video bit rates and MOS, i.e., occurrence of stalling events, while it can also result in significant backhaul capacity savings since, as presented at Table 5, on average only 1.09 quality layers instead of all 10 quality layers need to be cached, even leading to better QoE, as the users are prevented from a plethora of stalling-prone segment selections. Hence, the QoE-SDN APP improves significantly the end user QoE, assisting VSPs to have more satisfied customers, and MNOs to use their backhaul resources more efficiently. Finally, in Figure 9(c) we observe the same trends as in the previous use case, and that for a large number of users, the minimum stallings and rate-guided HAS strategies converge, as users tend to request the lowest layer segments as a way to avoid stallings.

3) Evaluation analysis of use case 3: Proactive segment selection and placement studies the impact of proactive HAS segment caching. For the purposes of evaluation, we introduce a simplistic backhaul delay, which depends on the size of the transmitted segment, as: *Backhaul delay = Segment size / Backhaul rate,* in order to demonstrate the impact of the backhaul. The backhaul rate is set to 10Mbps (i.e., the achieved backhaul rate per user on average), so that the access network connectivity is not backhaul restricted (actually, the mean video bit rate is much less, as shown in Table 6). As before, three HAS variations are considered named: a) the *standard HAS*, where there is no proactive caching, b) the *rate-guided HAS*, where the rate estimation is used to enforce the VSP segment selection, with the mobility prediction guiding the *proactive caching* of these selected segments to the appropriate MEC locations, and c) the *minimum stallings HAS* that caches the lowest segment quality layers only. The results obtained are showed in Figure 10 and Table 6.

TABLE 5
KPI ESTIMATIONS FOR USE CASE 2.

HAS logic	Mean video bit rate (bps)	Mean quality layer	Mean QoE (MOS)	QoE fairness	Mean video utility	Mean stalling probability	Mean stalling duration (sec)	Average stallings per user	Mean accumu-lative SI	***Average number of active layers***	***Bandwidth savings (bps)***
Standard	1.46E+06	5.63	2.25	0.75	0.94	0.82	22.33	1.87	0.29	10	-
Rate-guided	5.00E+05	2.28	3.35	0.80	0.88	0.50	23.69	0.67	0.08	1.09	1.77E+07
Min stallings	2.32E+05	1	3.83	0.85	0.86	0.35	26.26	0.40	0	1	-

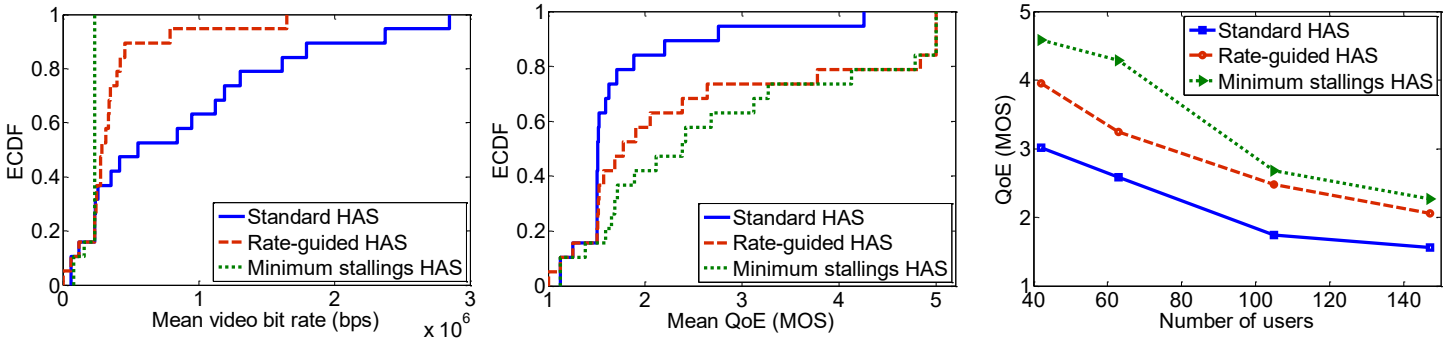

Fig. 10. Use case 3 evaluation results - (a) ECDF of mean video bit rate for all users, (b) ECDF of MOS for all users, (c) QoE for varying load.

TABLE 6
KPI ESTIMATIONS FOR USE CASE 3.

HAS logic	Mean video bit rate (bps)	Mean quality layer	Mean QoE (MOS)	QoE fairness	Mean video utility	Mean stalling probability	Mean stalling duration (sec)	Average stallings per user	Mean accumulative SI
Standard	8.72E+05	4.08	1.74	0.84	0.83	0.94	25.36	1.57	0.22
Rate-guided	3.70E+05	2.07	2.48	0.75	0.79	0.76	18.90	1.01	0.34
Min stallings	2.10E+05	1	2.68	0.76	0.78	0.71	18.54	0.94	0

Similarly to the previous use cases, and as observed from Figures 10(a)-(c) and Table 6, the proposed rate-guided HAS provides a better balance in the achieved video bit rate and MOS compared to the standard and minimum stallings HAS. It is also observed that stalling events are less likely to occur when proactive caching is used (reduced to 76% as opposed to 94% for the baseline approach). The reason for that, additionally to the benefits of the rate-guided segment selection process, is that this scheme reduces the backhaul delay required to fetch a video segment upon request; therefore, the user has more chances of downloading this segment early enough, i.e., before the segment's deadline. Finally, in terms of fairness and stability, the results are similar to the previous use cases. It is also worth noting that the current simulation presents a very challenging scenario in terms of stallings, since due to SLAW, users tend to accumulate to specific clusters; thus, certain eNBs are severely loaded. Thus, in the current setting, each user will experience at least one stalling as shown in Table 6.

VII. CONCLUSIONS AND FURTHER DISCUSSION

In this paper, we have introduced a programmable QoE-SDN APP, based on the openness and flexibility provided by the SDN paradigm. This QoE-SDN APP can serve the customers of VSPs, improving their QoE by reducing the occurrence of the highly undesirable stalling events. Focusing on HAS applications, and by running a mobility forecasting and rate estimation function within the MNO's domain, the proposed scheme manages to significantly improve the QoE of video streaming users. This improvement has been highlighted and quantified through the proposal and evaluation of use cases for video segment encoding, selection and placement that are "unlocked" by the proposed architecture. These techniques take advantage of network feedback information exposed by the MNO related to the positions and data rates of mobile users, in order to trade off stalling events with video bit rates, since the former have a much stronger QoE impact. Based on the simulations conducted, the rate-guided HAS strategies enforced by the QoE-SDN APP also ensure fairness among users and stability in the viewing quality, in parallel to improving QoE.

Apart from the technical novelty of the proposed scheme, added business value is expected. Specifically, the introduction of the QoE-SDN APP has an impact not only on the reputation of various service providers, but also on the revenues of the MNOs, stemming from bandwidth savings and from direct financial benefits through API exposure to service providers. The activation of the QoE-SDN APP can be on-demand, rather than being an "always-on" function and can be programmed according to the particular service needs. For instance, some VSPs already differentiate their customers, based on their subscription type, to gold or standard users; in this case, the QoE-SDN APP can be triggered only for the former type of users. Similarly, the QoE-SDN APP may be designed as an add-on feature, which customers can activate on-demand, and for a limited amount of time, i.e., in the form of time-bounded purchased tokens or pay-as-you schemes. When any of these schemes is recognized, then the QoE-SDN APP and the accompanying QoE management cycle will automatically instantiate the essential monitoring and control actions within the MNO that will boost the customer QoE.

The need to improve the end users' experience together with the emergence of technologies such as SDN, MEC and personalized network slicing [46], which enable such improvements through service/application and user/OTT differentiation, pose a challenge to net neutrality principles. The QoE-SDN APP offers a differentiated and enhanced experience to the users of VSPs that choose to adopt it, in a broad sense. However, it raises none net neutrality concerns, since in the context of the HAS use cases the QoE-SDN APP does not require any special traffic treatment to different traffic flows by the MNO, such as prioritization against other traffic classes; it just enables QoE assessment and network exposure feedback mechanism to VSPs that helps them better handle video streaming. Nevertheless, our view is that such

solutions and architectures will inevitably continue to emerge and finally prevail towards the 5G era, as also enforced by the recent net neutrality repeal order of the Federal Communications Commission (FCC), which questions the future of net neutrality in the name of "Internet freedom" [47].

Future work involves the real implementation of the proposed QoE-SDN APP on an SDN testbed, to showcase the applicability of this scheme for real HAS services and devices. Moreover, scalability issues related to the placement of the QoE-SDN APP need to be investigated (i.e., centralized vs. distributed), closely related to the challenge of optimal placement of SDN controllers [48]-[49]. Finally, even though this study has concentrated on HAS, the benefits for other types of services and verticals remain to be investigated.

REFERENCES

[1] NGMN Alliance, "5G White Paper," 2015.

[2] T. Taleb, A. Ksentini, and P. Frangoudis, "Follow-Me Cloud: When cloud services follow mobile users," IEEE Trans. Cloud Comput., 2016.

[3] Cisco Visual Networking Index: Forecast and Methodology, 2016-2021, Sept. 2017.

[4] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," IEEE Commun. Surv. Tutorials, vol. 17, no. 1, pp. 469–492, Jan. 2015.

[5] World Economic Forum, Digital Transformation Initiative Telecommunications Industry, White Paper, 2017.

[6] GSMA, The Mobile Economy, Report, 2017.

[7] T. Taleb, K. Samdanis, and A. Ksentini, "Towards elastic application-oriented bearer management for enhancing QoE in LTE networks," in 2016 IEEE Wireless Communications and Networking Conference, 2016, pp. 1–6.

[8] D. L. C. Dutra, M. Bagaa, T. Taleb, and K. Samdanis, "Ensuring end-to-end QoS based on multi-paths routing using SDN technology," in 2017 IEEE Global Communications Conference (GLOBECOM), 2017, pp. 1–6.

[9] ONF SDN, SDN Architecture issue 1.0 - Open Networking Foundation, 2014.

[10] ONF SDN, SDN Architecture issue 1.1 - Open Networking Foundation, 2016.

[11] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, D. Sabella "On multi-access edge computing: A survey of the emerging 5G network edge architecture & orchestration," IEEE Communications Surveys & Tutorials, 3rd Quarter 2017.

[12] Qualinet European Network on Quality of Experience in multimedia systems and services, "Definitions of Quality of Experience (QoE) and related concepts," White Paper, 2012.

[13] E. Liotou, D. Tsolkas, N. Passas and L. Merakos, "Quality of Experience management in mobile cellular networks: Key issues and design challenges," IEEE Communications Magazine, Network & Service Management Series, vol. 53, no. 7, pp. 145–153, July 2015.

[14] A. Kassler, L. Skorin-Kapov, O. Dobrijevic, M. Matijasevic, and Peter Dely, "Towards QoE-driven multimedia service negotiation and path optimization with software defined networking," in 20th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2012, pp. 1–5.

[15] H. Nam, K.-H. Kim, J. Y. Kim, and H. Schulzrinne, "Towards QoE-aware video streaming using SDN," in 2014 IEEE Global Communications Conference, 2014, pp. 1317–1322.

[16] G. Cofano, L. De Cicco, T. Zinner, A. Nguyen-Ngoc, P. Tran-Gia, and S. Mascolo, "Design and experimental evaluation of network-assisted strategies for HTTP adaptive streaming," in 7th International Conference on Multimedia Systems - MMSys '16, 2016, pp. 1–12.

[17] A. Bentaleb, A. C. Begen, and R. Zimmermann, "SDNDASH: Improving QoE of HTTP adaptive streaming using software defined networking," in 2016 ACM on Multimedia Conference - MM '16, 2016, pp. 1296–1305.

[18] A. Bentaleb, A. C. Begen, R. Zimmermann, and S. Harous, "SDNHAS: An SDN-enabled architecture to optimize QoE in HTTP Adaptive Streaming," IEEE Trans. Multimed., vol. 19, no. 10, pp. 2136–2151, Oct. 2017.

[19] J. W. Kleinrouweler, S. Cabrero, and P. Cesar, "Delivering stable high-quality video: An SDN architecture with DASH assisting network elements," in 7th International Conference on Multimedia Systems - MMSys '16, 2016, pp. 1–10.

[20] D. Bhat, A. Rizk, M. Zink, and R. Steinmetz, "Network assisted content distribution for adaptive bitrate video streaming," in 8th ACM on Multimedia Systems Conference - MMSys'17, 2017, pp. 62–75.

[21] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, "Towards network-wide QoE fairness using openflow-assisted adaptive video streaming," in ACM SIGCOMM workshop on Future human-centric multimedia networking, 2013, pp. 15–20.

[22] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software Defined Radio Access Network," in Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking - HotSDN '13, 2013, pp. 25–30.

[23] T. Chen, M. Matinmikko, X. Chen, X. Zhou, and P. Ahokangas, "Software defined mobile networks: concept, survey, and research directions," IEEE Commun. Mag., vol. 53, no. 11, pp. 126–133, Nov. 2015.

[24] M. Gramaglia, I. Digon, V. Friderikos, D. von Hugo, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Flexible connectivity and QoE/QoS management for 5G Networks: The 5G NORMA view," in 2016 IEEE International Conference on Communications Workshops (ICC), 2016, pp. 373–379.

[25] T. Taleb and A. Ksentini, "QoS/QoE predictions-based admission control for femto communications," in 2012 IEEE International Conference on Communications (ICC), 2012, pp. 5146–5150.

[26] S. Dutta, T. Taleb, P. A. Frangoudis, and A. Ksentini, "On-the-fly QoE-aware transcoding in the mobile edge," in 2016 IEEE Global Communications Conference (GLOBECOM), 2016, pp. 1–6.

[27] M. Mu, M. Broadbent, A. Farshad, N. Hart, D. Hutchison, Q. Ni, and N. Race, "A scalable user fairness model for adaptive video streaming over SDN-assisted future networks," IEEE J. Sel. Areas Commun., vol. 34, no. 8, pp. 2168–2184, Aug. 2016.

[28] S. Ramakrishnan, and X. Zhu, "An SDN based approach to measuring and optimizing ABR video Quality of Experience", Cisco Systems, Technical Paper, 2014.

[29] M. Katsarakis, G. Fortetsanakis, P. Charonyktakis, A. Kostopoulos, and M. Papadopouli, "On user-centric tools for QoE-based recommendation and real-time analysis of large-scale markets," IEEE Commun. Mag., vol. 52, no. 9, pp. 37–43, Sep. 2014.

[30] S. Dutta, T. Taleb, and A. Ksentini, "QoE-aware elasticity support in cloud-native 5G systems," in 2016 IEEE International Conference on Communications (ICC), 2016, pp. 1–6.

[31] E. Liotou, G. Tseliou, K. Samdanis, D. Tsolkas, F. Adelantado, and C. Verikoukis, "An SDN QoE-service for dynamically enhancing the performance of OTT applications," in 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), 2015, pp. 1–2.

[32] A. Ahmad, A. Floris, and L. Atzori, "QoE-centric service delivery: A collaborative approach among OTTs and ISPs," Comput. Networks, vol. 110, pp. 168–179, Dec. 2016.

[33] S. Peng, J. O. Fajardo, P. S. Khodashenas, B. Blanco, F. Liberal, C. Ruiz, C. Turyagyenda, M. Wilson, and S. Vadgama, "QoE-oriented mobile edge service management leveraging SDN and NFV," Mob. Inf. Syst., vol. 2017, pp. 1–14, 2017.

[34] 3GPP TR 23.708, Architecture enhancement for Service Capability Exposure, Rel.13, Jun. 2015.

[35] GSMA OneAPI, Online: www.gsma.com/oneapi

[36] A. Aissioui, A. Ksentini, A. M. Gueroui, and T. Taleb, "Toward elastic distributed SDN/NFV controller for 5G mobile cloud management systems," IEEE Access, vol. 3, pp. 2055–2064, 2015.

[37] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: Self-Similar Least-Action Human Walk," IEEE/ACM Trans. Netw., vol. 20, no. 2, pp. 515–529, Apr. 2012.

[38] E. Liotou, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas, "Enriching HTTP adaptive streaming with context awareness: A tunnel case study," in 2016 IEEE International Conference on Communications (ICC), 2016, pp. 1–6.

[39] R. Langar, N. Bouabdallah and R. Boutaba, "A comprehensive analysis of mobility management in MPLS-based wireless access networks," in IEEE/ACM Transactions on Networking, vol. 16, no. 4, pp. 918–931, Aug. 2008.

[40] A. Nadembega, A. Hafid, and T. Taleb, "Mobility-prediction-aware bandwidth reservation scheme for mobile networks," IEEE Trans. Veh.

Technol., vol. 64, no. 6, pp. 2561–2576, Jun. 2015.
[41] J. C. Ikuno, M. Wrulich, and M. Rupp, "System level simulation of LTE networks," in 2010 IEEE 71st Vehicular Technology Conference, 2010, pp. 1–5.
[42] T. Hoßfeld, R. Schatz, E. W. Biersack, and L. Plissonneau, "Internet video delivery in YouTube: From traffic measurements to quality of experience," in Data Traffic Monitoring and Analysis, E. Biersack, C. Callegari, and M. Matijasevic, Eds. Springer Berlin Heidelberg, 2013, pp. 264–301.
[43] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," IEEE Network, vol. 24, no. 2, pp. 36–41, Mar-2010.
[44] A. Farshad, P. Georgopoulos, M. Broadbent, M. Mu, and N. Race, "Leveraging SDN to provide an in-network QoE measurement framework," in 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2015, pp. 239–244.
[45] T. Zinner, T. Hoßfeld, T. N. Minash, and M. Fiedler, "Controlled vs. uncontrolled degradations of QoE - The provisioning–delivery hysteresis in case of video," in Proc. 1st Workshop QoEMCS, Tampere, Finland, 2010, pp. 1–3.
[46] T. Taleb, B. Mada, M.-I. Corici, A. Nakao, and H. Flinck, "PERMIT: Network slicing for personalized 5G mobile telecommunications," IEEE Commun. Mag., vol. 55, no. 5, pp. 88–93, May 2017.
[47] FCC 17-166, "Restoring Internet Freedom", Jan. 2018.
[48] A. Ksentini, M. Bagaa, and T. Taleb, "On using SDN in 5G: The controller placement problem," in 2016 IEEE Global Communications Conference (GLOBECOM), 2016, pp. 1–6.
[49] A. Ksentini, M. Bagaa, T. Taleb, and I. Balasingham, "On using bargaining game for optimal placement of SDN controllers," in 2016 IEEE International Conference on Communications (ICC), 2016, pp. 1–6.

Eirini Liotou received the Diploma in Electrical & Computer Engineering from the National Technical University of Athens, the MSc in Communications & Signal Processing from the Imperial College of London and her Ph.D. degree from the Department of Informatics and Telecommunications, University of Athens in 2017. She has worked as a Senior Software Engineer in Siemens Enterprise Communications within the R&D department. She is currently working as a Senior Researcher in the Communication Networks Laboratory of the University of Athens. Her research interests are in the area of Quality of Experience in mobile cellular networks.

Konstantinos Samdanis is a Principal Researcher at Huawei for 5G carrier networks. He is involved in strategy and research for 5G architectures and network slicing for transport networks, while being also active in the 5G stream at BBF in Wireless-Wired Converged Networks Working Group and IETF in the Network and Routing Area Working Groups. Previously, he worked for NEC Europe, Germany, as a Senior Researcher and a Broadband Standardization Specialist, involved in numerous EU projects including 5G-NORMA, iJOIN, BeFemto, and BBF, 3GPP SA2 and SA5. Konstantinos has served as editor on the Network Slicing feature topic at IEEE Communications Magazine in 2017, and has organized the feature topic on Enabling 5G Verticals & Services through Network Softwarization and Slicing at IEEE Communications Standards Magazine. He has arranged and authored a book in Green Communications, and is the author of more than 60 academic publications and 20 patent applications. He received his Ph.D. and M.Sc. degrees from Kings College London.

Emmanouil Pateromichelakis received his Diploma in Information and Communication Systems Engineering from University of the Aegean, Greece. He received his MSc and PhD degree in Mobile Communications from University of Surrey, UK in 2009 and 2013 respectively. Thereafter, for the period 2013-2015 he was working as post-doctoral Research Fellow in the ICS at University of Surrey. Since 2015, he has been working as Senior Researcher in Huawei Technologies, at the German Research Center, focusing on 5G and beyond solutions. Part of this work has been published in several conferences and journal papers and he has also filed numerous patents on 5G related topics.

Nikos Passas received his Diploma (honors) from the Department of Computer Engineering, University of Patras, Greece, and his Ph.D. degree from the Department of Informatics and Telecommunications, University of Athens, Greece, in 1992 and 1997, respectively. Since 1995, he has been with the Communication Networks Laboratory of the University of Athens, working as a Senior Researcher in a number of national and European research projects. His research interests are in the area of mobile network architectures and protocols.

Lazaros Merakos received the Diploma in Electrical and Mechanical Engineering from the National Technical University of Athens, Greece, in 1978, and the M.S. and Ph.D. degrees in Electrical Engineering from the State University of New York, Buffalo, in 1981 and 1984, respectively. From 1983 to 1986, he was on the faculty of the Electrical Engineering and Computer Science Department, University of Connecticut, Storrs. From 1986 to 1994, he was on the faculty of the Electrical and Computer Engineering Department, Northeastern University, Boston, MA. During the period 1993–1994, he served as Director of the Communications and Digital Processing Research Center, Northeastern University. During the summers of 1990 and 1991, he was a Visiting Scientist at the IBM T. J. Watson Research Center, Yorktown Heights, NY. In 1994, he joined the faculty of the University of Athens, Athens, Greece, where he is presently a Professor in the Department of Informatics and Telecommunications, and Scientific Director of the Networks Operations and Management Center.