IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 June 2023, accepted 17 July 2023, date of publication 24 July 2023, date of current version 31 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298556

RESEARCH ARTICLE

QoE-Aware Analysis and Management of Multimedia Services in 5G and Beyond Heterogeneous Networks

MOHAMAD T. SULTAN AND HESHAM EL SAYED
College of Information Technology, United Arab Emirates University (UAEU), Abu Dhabi, United Arab Emirates
Emirates Center for Mobility Research (ECMR), United Arab Emirates University, Abu Dhabi, United Arab Emirates

Corresponding author: Mohamad T. Sultan (202190237@uaeu.ac.ae)

This work was supported in part by the ASPIRE Award for Research Excellence under Project AARE20-368.

ABSTRACT The explosion of mobile applications and phenomenal adoption of mobile connectivity by end-users has generated an increasing amount of mobile data traffic. Application posing stringent network requirements of high bandwidth and low latency (e.g., immersive videos) and the substantial amount of data traffic has put tremendous pressure on existing network infrastructures. Cognizant of the need of increasing network capacity, the simultaneous use of heterogeneous network technologies (HetNet) has been proposed to address this imperative problem. 5G is expected to further drive the concept of HetNet by allowing the use of huge available bandwidth at milli-meter wave frequencies. While HetNet concentrates on improving network capacity from the data transmission perspective, it overlooks the importance of enhancing the support of user's Quality of Experience (QoE) for the evolving new services. There is a wide range of factors influencing user's QoE, such as network performance including delay, jitter and throughput, contextual influence such as personalized content delivery, mobility aware content caching and dissemination, and human impact such as human roles and demographic attributes. This paper initially presents a QoE-centric analysis, and evaluation of multimedia services over 5G networks. Then, to address the shortcomings of existing mobile networks, we propose a framework to enhance the support of QoE, to enable smooth delivery of personalized immersive video environment and personalized interaction with an immersive video, anywhere, anytime and on any device. Finally, we propose new solutions to achieve practically feasible spectrum allocation and personalized content caching and dissemination, to provide uninterrupted multimedia services to end-users.

INDEX TERMS Quality of experience, network slicing, multimedia services, HAS, 5G, SDN.

I. INTRODUCTION

The consumption of multimedia services has significantly increased since the introduction of 4G/LTE networks. The use of mobile video streaming services like YouTube and Mobile TV on smart devices is projected to keep growing as future networks like 5G emerge and develop [1]. There is a growing demand from end users for higher-quality services from service providers, which has led to a shift towards network management that prioritizes the Quality of Experience (QoE). This involves effectively using network resources to meet user expectations. However, existing network technologies face challenges in adapting to varying network conditions and have limitations in terms of available resources. As a result, service providers are facing difficulties in delivering QoE-aware multimedia services [1]. Operators keep evolving their capacity planning, deployment, and optimization strategies to meet such stringent application requirements with ultimate goal of improving users' quality of experience, which is widely perceived as a measure of user satisfaction. Despite these efforts improving users' quality of experience remains a challenging task due to many factors, which include the variability of network resources, the variety of networks technologies (e.g., fixed and mobile), the emergence of

The associate editor coordinating the review of this manuscript and approving it for publication was Yogendra Kumar Prajapati.

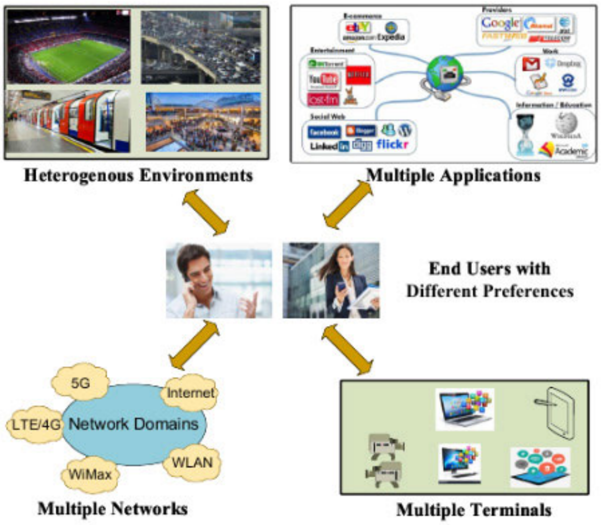

FIGURE 1. Quality of experience challenges.

new immersive services like video streaming, video gaming, virtual and augmented reality, and heterogeneity of end-user devices with different capabilities [2]. Prevailing QoE approaches are mostly operator-oriented, while users play somewhat limited roles. Required is support for QoE as perceived subjectively by end-users. How to achieve this anytime, anywhere and on any device is critical to the immersive videos. It is also challenging to accurately quantify video QoE in near real time and at scale [2]. Communication bit-pipes connecting a video server and an end-user contribute heavily to QoE, which mostly comprises of a series of heterogeneous links connecting Content Delivery Network (CDN), and Internet Service Provider (ISP). It is therefore difficult to isolate and gauge the impact of individual links and their interconnections on video QoE. To meet end-users' QoE requirements and expectations, several QoE issues such as QoE monitoring, control, and management present challenges to service providers. Addressing these challenges requires the integration of advanced intelligent architectures in the current and beyond 5G future networks with the development of smart QoE-centric adaptation techniques [3]. Therefore, to improve end-users' quality of experience, several QoE-enabled functions in the future networks, such as QoE-centric resource allocation, server selection, routing, admission, and control mechanisms should be adaptive to changes in network conditions. Figure 1 presents the QoE management challenges in the current and future networks.

The objective of this research is to provide an initial analysis and evaluation of QoE-centric video streaming services and to explore new approaches to achieve practically feasible spectrum allocation and personalized content caching and dissemination, to provide uninterrupted multimedia services to mobile users, while maximizing network resource utilization through predictive models. It provides a systematic framework to enable personalized immersive multimedia system, for smooth video streaming delivery and personalized interaction with an immersive video, anywhere, anytime and on any device. The framework takes advantage of user mobility prediction, user preference inference, and fine-grained resource reservation, to improve overall user's QoE via personalized data caching and dissemination in HetNet. The advantage of this framework is two-fold. First, network traffic can be reduced by delivering only relevant information to target users. Second, enhanced support of QoE can be achieved by synergistic content caching and resource reservation that differentiate applications and users. To achieve these goals, the proposed framework integrates edge computing [4], software defined network (SDN), and ETSI MANO [5] smart orchestrator to support user mobility management, personalized content caching and dissemination among edge servers, QoE-based multimedia flow routing and QoE-aware resource allocation on communication infrastructures. Initially, we present an analysis and evaluation of multimedia streaming services over 5G and future networks based on system-level simulations, real-world datasets, and experiments in a prototype of the proposed system. then we propose a QoE-aware solutions that focus on three coherent research thrusts to investigate (1) mobility-aware SDN-enabled caching and dissemination of personalized contents, (2) Smart orchestrator for QoE-aware efficient resource allocation in HetNets and (3) predictive mobility model for adaptive timed-QoE guarantees in HetNets.

II. INTERNET MULTIMEDIA STREAMING SERVICES

With the growth of streaming video traffic over 5G networks, online video streaming applications like video on demand (VoD) have advanced significantly. While future networks like 5G and beyond promise to support higher throughputs and lower end to end delay, the management and adaptation of these networks to the rapid growth and increase of video streaming applications and maintaining a smooth quality of experience (QoE) to the end users is still a challenging task [6]. One of the key factors that impacts the end user quality of experience in multimedia streaming services is the underlying quality of service (QoS) network level performance metrics like throughput, delay, RTT, jitter, and packet lose. Variation in these QoS metrics contribute highly to determining the end users QoE satisfaction. Numerous techniques have been proposed to meet end-users' quality of experience, including the development of HTTP Adaptive Streaming (HAS) which is as a de facto streaming technology used by popular video streaming platforms like Netflix and YouTube that allows the adaptation of downloaded video quality to current network conditions [7]. The advantages of HAS include that it can run over HTTP, cache infrastructure reuse capability and delivering reliable transmission sessions. The connection between a client and a standard HTTP web server in HAS architecture is shown in Figure 2.

In HAS, adaptation techniques are used to provide a smooth streaming session by switching between various qualities of the video segments to avoid stalling streamed frames

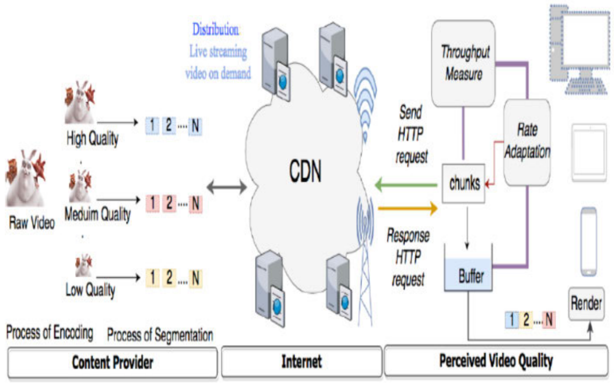

FIGURE 2. Architecture of HAS.

in a highly congested user's network. Moreover, another key factor besides the QoS metrics which plays a major role in end user satisfaction is the selection of the underlying Adaptive Bitrate Streaming (ABS) algorithm on the client side. Thus, it is important to understand the behavior of different proposed ABS algorithms and the corresponding impact of the relationship between quality of service and quality of experience key performance indicators (KPIs). It worth noting that network level quality of service KPIs like throughput, delay, packet lose has a direct impact on the application-level quality of experience KPI like bitrate, stall and resolution which eventually influence the end user QoE subjective feedback. Implementing ABS algorithms to estimate QoE in real life scenarios of 5G networks is a complex task due to cost, time, and decisions [7]. Thus, in the following section we present an initial virtualized network testbed for performance evaluation of two different ABS algorithms in Dynamic Adaptive Streaming over HTTP (DASH) using real 5G network traces.

A. PERFORMANCE EVALUATION AND RELATED LITERATURE

The video streaming file in DASH is split into small segments or chunks of same duration during the process of segmentation where these smaller chunks are further encoded with various resolutions and bitrates that are stored in a representation MPD file hosted on the server side as shown in Figure 2. However, at the client side, where a video player is hosted, the corresponding ABS algorithm is responsible of reading the MPD file from the server side and detecting and requesting the most appropriate video segments taking into consideration the current network condition to ensure a smooth video streaming experience for the end user. Different ABS algorithms have been proposed for adaptive video streaming [8]. Traditionally, these algorithms fall into different categories which are either rate-based or buffer-based algorithms. However there have been proposed a hybrid algorithm the combines estimation techniques from both categories. The rate-based algorithms (e.g., exponential, and conventional) work by estimating the bandwidth to determine which video streaming segment to download [9]. On other hand, the buffer-based ABS algorithms (e.g., BBA [10] and Logistic [11]) utilize the buffer state to estimate the appropriate video rate based on the relationship between the different levels of video quality and network buffer utilization. Numerous studies exist in the literature that deal with ABS performance analysis and evaluation. The authors in [12] have proposed a method for resolution identification based on HTTP/2 features for video segments analysis. Their method is called H2CI and can monitor QoE to infer the adaptation behavior of encrypted video streaming. Similarly, the authors in [13] have proposed a quality of experience DASH emulator called QoE-DASH that can be utilized in multi-access edge computing. Their emulator is developed based on popular goDASH framework where two joint caching techniques are used for the purpose of performance evaluation for QoE metrics taking into account user preferences and network properties. Moreover, the authors in [14] conducted a performance analysis of video streaming over constrained application protocol using different setups experiments by adjusting the CoAP transmission parameters. The researcher in [15] conducted an analysis for adaptive point cloud streaming by utilizing different objective and subjective QoE metrics. They investigated the impact of bandwidth, viewport prediction, user mobility and adaptation rate in their analysis. Moreover, the authors in [16] proposed a methodology for measuring quality of experience in encrypted traffic of video streaming services. Their methodology proposes various levels of QoE degradation due to different affecting factors like quality variations, stalling and average video quality. For the purpose of demonstration of video quality adaptation, we selected two states of the art buffer based: Logistic and BBA ABS algorithms to assess its QoE performance. To this end, we considered building a QoE-aware network using DASH video application in a realistic emulation environment. This performance evaluation provides QoS to QoE mapping and analysis using real 5G cellular network traces.

B. EXPERIMENTAL SETUP

In this research a virtualized network testbed is presented for the purpose of performing performance evaluation of two different buffer-based ABS algorithms to evaluate the QoE of end-users. This presents QoS to QoE metrics mapping and analysis. To support implementing experiments of adaptive video streaming the proposed system setup is presented in Figure 3. Generally, this topology is used in the experiments for data acquisition that is eventually being processed and analyzed for the purpose of ABS algorithms performance evaluation and comparison. The system design contains different components such as the network emulator Mininet [17], Caddy cache server for DASH video content and headless DASH video player GoDash [18] on the client side where the ABS algorithms reside. The content of video streaming segments is delivered from the video streaming cache server to the GoDash clients through different network access switches to assess the impact of simultaneous

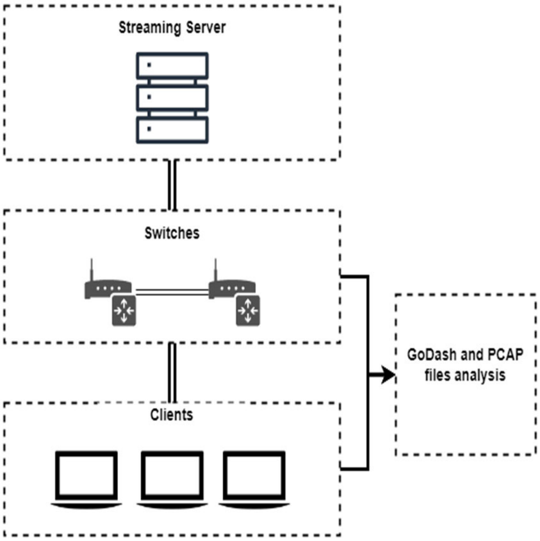

FIGURE 3. Proposed system design setup.

TABLE 1. Sample output logs generated by GoDash.

Seg.	Arr. time	Del. time	Stall Dur.	Del. Rate	Bye Size	Buff Level
1	133	133	0	4955	82382	4000
2	2353	1371	0	25074	4297176	8000
3	4043	869	0	23466	2549074	10310

TABLE 2. Sample QoE model output by GoDash.

Seg.	Algorithm	Codec	P.1203	Claey	Duanmu
1	Logistic	h264	1.871	0.000	46.465
2	BBA	h264	2.543	0.163	40.823
3	Logistic	h264	3.288	0.097	48.888

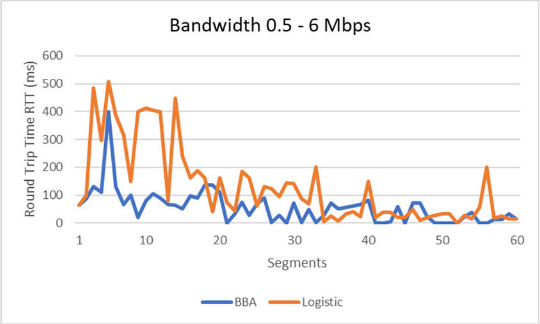

FIGURE 4. Round trip time for BBA and Logistic ABS algorithms.

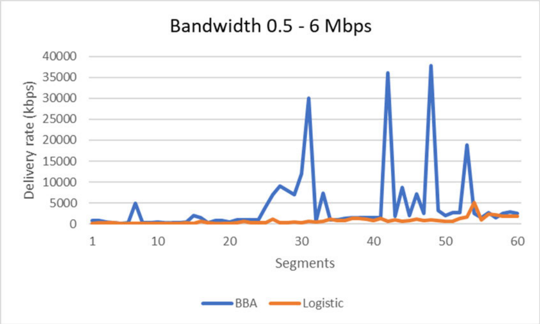

FIGURE 5. Delivery rate parameter for BBA and Logistic ABS algorithms.

video streaming on end-users. The clients compete for video streaming services from the same cache server. Two switches are used, and the link bandwidth utilizes 5G trace parameters to sample values. The 5G trace data is an open source freely available dataset that's collected from an Irish mobile operator [19]. It contains QoS data of client-side cellular key performance indicators (KPIs) such as throughput information, and two different patterns of mobility generated which are static and dynamic over two application patterns (file download and video streaming). On the other hand, the GoDash client-side video player hosts the ABS algorithms, and the generated GoDash log files from these algorithms contains useful information such as per segment objective quality of experience KPIs (e.g., bitrate, stall, segment arrival time etc.) as shown in Table 1. and per segment output using five well-known QoE evaluation models including the International Telecommunication Union (ITU) P.1203 standard, Yin, Duanmu, Yu, and Clae as shown in Table 2 [18]. All the tools and data used in this research are open source and publicly available.

C. PERFORMANCE EVALUATION RESULTS

The performance results of evaluating the buffer-based BBA and Logistic adaptive bitrate algorithms using static 5G network traces are presented in this section. The generated client side GoDash logs and corresponding PCAP log files are processed and used for performance analysis. In every experimental run, we collect per-segment QoS KPIs, generated by the PCAP logs. The running time is 120 seconds for a total of 60 segments, and we used the ITU P.1203 standard metric as a QoE evaluation model that predicts a mean opinion score based on QoS network level KPIs. The performance analysis results for 60 segments of video streaming against different performance parameters such as round-trip time (RTT), stall, delivery rate and P.1203 standard are presented in Figure. 4, Figure 5, Figure 6, and Figure 7 respectively. The network bandwidth values are estimated from the 5G trace parameters, and it is exchanged between the network switches.

In our evaluation and analysis, the per-segment QoS results and KPIs are collected for each experiment. It can be noticed that both algorithms experienced higher RTT at the beginning of the video streaming as shown in Figure 4. The fluctuations in RTT values are due to the nature of the buffer-based algorithms as those algorithms require defining the initial number of segments to be downloaded before the stream begins. However, the RTT for both algorithms started to gradually decrease after segment 20. Noticeably, the logistic algorithm

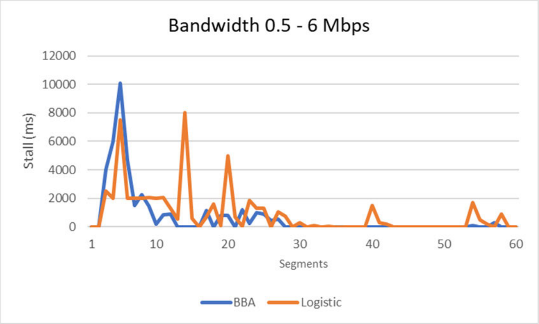

FIGURE 6. Stall duration parameter for BBA and Logistic ABS algorithms.

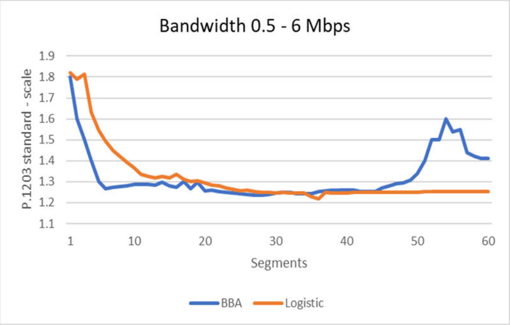

FIGURE 7. P.1203 QoE standard for BBA and Logistic ABS algorithms.

experienced higher RTT than the BBA algorithm for this scenario. Similarly, the results in Figure 5, indicate that BBA has a higher delivery rate than the ABS logistic algorithm. Moreover, in terms of stall or video freezing, both algorithms experienced higher stall at the beginning of the streaming due to selecting the wrong segments that do not match the network conditions and resources. Wrong segment choices can happen in these algorithms due to fluctuations in the bandwidth. However, stall values started to level down at the subsequent segments. It is worth noting that BBA experienced less stall issues in comparison to Logistic algorithm as shown in Figure 6. Finally, the results in Figure 7, present the well-known QoE estimation model P.1203 where the values of this QoE metric are estimated from the previous QoS results. The results show a quite similar trend for the P.1203 score over the experiment time with an overall slight advantage to the BBA algorithm. The Logistic algorithm experienced higher RTT and higher number of stalls and eventually lower P.1203 score.

The experimental results obtained from this performance evaluation using realistic emulation environment aid in understanding the behavior, challenges, and issues in the current video streaming technologies such as the HTTP Adaptive Streaming (HAS) which is the dominant video streaming technology for multimedia services over cellular 5G and future networks. This highlights the need for the proposal of new techniques, approaches and methods that help in providing better QoE for end-users, which will be discussed in the next sections.

III. QOE PROPOSED FRAMEWORK AND SOLUTIONS

The objective of this section is to explore the development of new approaches in terms of QoE control and management to achieve practically feasible spectrum allocation and personalized content caching and dissemination, to provide uninterrupted video streaming and multimedia services, while maximizing network resource utilization through virtualized networks and predictive models. Although different traffic modes can be considered in this research, one particular application domain which is the vehicular network, will be used to illustrate the proposed framework. As the number of driverless cars increases, so will the demand for various forms of content including road state, traffic condition, and entertainments, e.g., video streaming services, 360 videos, movies, and real-time news. Instead of downloading contents from original servers, the proposed framework allows a user to obtain them from nearby edge servers that already cached the content before the user arrives. To achieve this goal, enabling technologies like edge computing and SDN are the integral techniques that enable the proposed framework, as illustrated in Figure 8. We adopt the ETSI reference MANO architecture [5] and include modules in different layers to enhance support of user's QoE.

The very top layer represents the application plane where applications and services are located. In this layer, we introduce three important databases, namely the traffic profile database (e.g., traffic congestion and road situation data), user mobility profile database (e.g., user most visited route and hierarchy), and spectrum database (e.g., spectrum availability and registration information). They provide useful information to mobility management and resource reservation modules to achieve accurate location prediction and optimized resource allocation. The second layer is network service and abstraction layer that provides a global view of the network, e.g., current network status information, content availability information, and buffer capacity information on edge servers. The newly introduced content availability information is extremely important because the optimal content distribution can only be realized if such information is available. To support user's QoE in different applications and various user hierarchies, we also introduce the service model module within this layer. In this module, the time guaranteed period, cluster reservation threshold, and bandwidth reservation threshold are configured for individual applications and services.

The innovation of the proposed solution lies in the incorporation of the proposed frameworks into SDN architecture and its integration with management platforms of ETSI reference architecture, which supports dynamic network slicing and permits the exploitation of Network Slice Selection Function (NSSF) for 5G (refer to 3GPP Rel. 18 [19]) to further boost 5G performance. The content delivery, QoE-aware resource reservation, and predictive mobility management modules are introduced to guarantee users receive satisfactory services in HetNets. The management plane, located at the third layer, is considered the most important place where all proposed

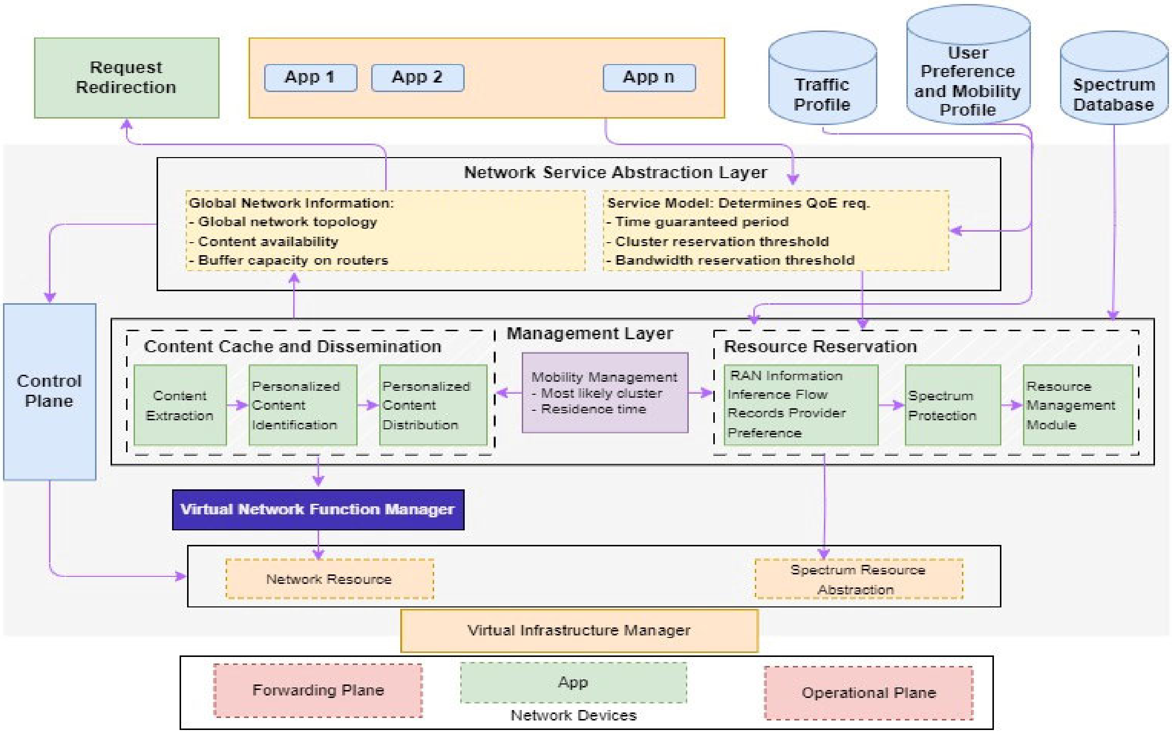
Request Redirection
App 1
App 2
App n
Traffic Profile
User Preference and Mobility Profile
Spectrum Database
Network Service Abstraction Layer
Global Network Information:
- Global network topology
- Content availability
- Buffer capacity on routers
Service Model: Determines QoE req.
- Time guaranteed period
- Cluster reservation threshold
- Bandwidth reservation threshold
Control Plane
Management Layer
Content Cache and Dissemination
Content Extraction
Personalized Content Identification
Personalized Content Distribution
Mobility Management
- Most likely cluster
- Residence time
Resource Reservation
RAN Information Inference Flow Records Provider Preference
Spectrum Protection
Resource Management Module
Virtual Network Function Manager
Network Resource
Spectrum Resource Abstraction
Virtual Infrastructure Manager
Forwarding Plane
App
Network Devices
Operational Plane

FIGURE 8. Proposed framewok.

intelligent models and algorithms are implemented. Specifically, the mobility management module keeps track of user's locations and predicts the Most Likely Clusters (MLC) of servers that a user will likely visit in the future.

Based on speed information, the prediction will also offer the user's earliest arrival time, latest arrival time, and latest departure time to the MLC. This module synthetically considers user mobility profile and traffic profile to accurately predict user's future locations. The prediction results set the foundation for personalized content distribution and resource reservation as contents will flow to potential mobile recipients and resources are reserved along the way. The resource reservation module considers the application of QoE requirements and spectrum availability to provide seamless communications to users by differentiating their hierarchy and mobility. Moreover, the content caching and dissemination module caches personalized popular contents on servers that will likely serve the corresponding users, based on user's mobility prediction and user's preference inference. It also notifies the network service and application layer to update the global content availability information. The actual data transmission in the core network is implemented by the control plane that efficiently forwards contents from an edge server to another. Instructions on how to move contents and network configuration parameters for radio access are then sent to the device and resource abstraction layer. Network devices layer, lowest layer in architecture, contains both physical and virtual devices that physically forwards contents. In the following sections we will further investigate the following coherent research thrusts to achieve an optimized end-user QoE based on the proposed framework in emerging and future networks.

- **Thrust I: Mobility-aware SDN enabled personalized content caching and dissemination.** This thrust will exploit the capabilities of SDN to obtain real-time global network state information and achieve an optimal personalized content caching and dissemination.
- **Thrust II: QoE-aware network slicing in HetNets.** The goal of this thrust is to an achieve efficient and practically feasible bandwidth allocation mechanism to provide continuous service to mobile users, while maximizing the network resources utilization.
- **Thrust III: Predictive mobility model for adaptive timed-QoE guarantees in HetNets.** The focus of this thrust is to develop a novel predictive model that supports different modes of user mobility, varying from human mobility to urban mobility.

A. MOBILITY-AWARE SDN-ENABLED CACHING FOR PERSONALIZED CONTENTS

The characteristics of SDN, including programmability, adaptability, and cost effectiveness, make it well-suited and a viable option for bandwidth-intensive multimedia applications like live video streaming apps. Thus, the objective of this research thrust is to leverage edge computing and SDN technology to explore a personalized content caching

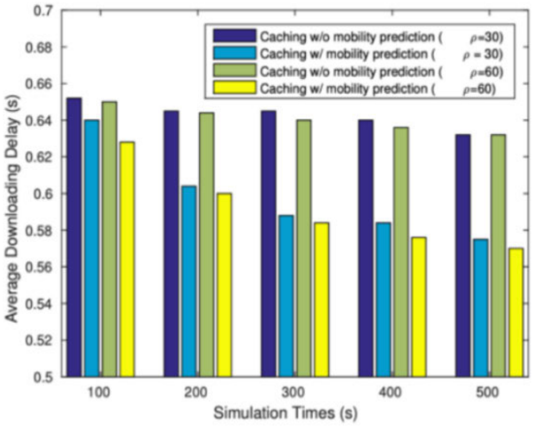

FIGURE 9. **Average downloading delay.**

and dissemination approach, which harnesses the synergy between mobile networks and recommendation systems. Existing techniques for content caching in mobile networks mainly focus on storing the most popular contents on servers that have the greatest demand for the contents. Assuming content popularity follows the Zipf distribution, the researchers in [20] utilized different vehicle density (ρ) and demonstrated that a user's average downloading delay can be significantly reduced, if the most popular contents are cached on the servers that will serve the users in future, as shown in Figure 9.

To improve the cache hitting rate as well as user's QoE, it is critical to design a personalized content caching mechanism that considers user's preference to predict the most popular contents for every individual user. Personalized content caching and dissemination in HetNets involves solving two major technical problems: (1) identifying popular contents for each user, (2) determining when and on which server(s) to cache the contents. Popular content for a user could be their frequently accessed content. Identifying frequently accessed contents requires a complete history of the contents accessed by every user, which is infeasible in a HetNet, where users frequently disconnect from the network for various reasons. With a partial content-accessing history, it is important to recover the missed ones and cache them on edge servers. Identifying the most-likely-accessed content implies predicting contents that a user has not yet accessed but might be interested in in future. To achieve the objectives of this thrust the following task could be performed.

Task 1: Personalized Content Identification (PCI): Focusing on user's preference rather than content's popularity, we can obtain a relatively stable user preference to precisely predict the contents that a user may be interested in, and thus provide a fine-grained personalized content caching and dissemination service. Previous works on content caching in wireless networks assume the existence of a popularity profile for every content and perform content caching on a per website basis, per cluster-of-objects basis or per community-of-objects basis [21]. This assumption, however, ignores the importance of user preference in accessing contents, the variability of content popularity, and the emergence of dynamic contents that suddenly become popular. We argue that it is critical to consider user preference in identifying personalized popular contents so that contents are cached on edge servers, on a per user basis.

This problem can be formulated as follows: Given $\boldsymbol{n}$ users $\boldsymbol{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{\boldsymbol{n}}\}$ and $\boldsymbol{m}$ contents $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{\boldsymbol{m}}\}$, we use accessing-frequency matrix $\boldsymbol{M}_{\boldsymbol{n}\times\boldsymbol{m}}= \{\boldsymbol{f}_{\boldsymbol{i},\boldsymbol{j}}\}$ to record the number of times user $\boldsymbol{u}_{\boldsymbol{i}}$ accessed content $\boldsymbol{c}_{\boldsymbol{j}}$, where $\boldsymbol{i} = 1, 2, \ldots, \boldsymbol{n}$. Similarly, $\boldsymbol{j} = 1, 2, \ldots, \boldsymbol{m}$. The problem is to identify and predict the top $\boldsymbol{K}$ popular content accessed by every user in the network. Although collaborative filtering (CF) was widely used in recommendation systems e.g., suggesting movies to a Netflix user, it is yet an unexplored problem whether it can accurately identify personalized popular contents of HetNet users. To solve this problem, we need to conquer the following technical challenges. The first technical challenge is to obtain an accurate estimate of the user similarity matrix $\boldsymbol{S} = \{\boldsymbol{S}_{\boldsymbol{ik}}\}$ where $\boldsymbol{S}_{\boldsymbol{ik}}$ denotes the similarity between users $\boldsymbol{i}$ and $\boldsymbol{k}$, where $\boldsymbol{i}, \boldsymbol{k} \in \boldsymbol{(1, 2, \ldots, n)}$. This is a challenge because $\boldsymbol{S}$ can only be computed from matrix $\boldsymbol{M}$ that may contain partial or inaccurate content accessing data. To address this challenge, we propose to design a genetic algorithm that starts from the $\boldsymbol{S}$ that is directly computed from $\boldsymbol{M}$, and then evolutionarily updates $\boldsymbol{S}$ until it offers the most accurate prediction results.

The second technical challenge is to address the data sparsity issue in $\boldsymbol{M}$, a well-known problem in CF [22]. This is a challenge because the performance of CF highly depends on the completeness of $\boldsymbol{M}$. To address this challenge, we plan to develop an innovative CF approach that iteratively updates $\boldsymbol{M}$ based on the estimated $\boldsymbol{S}$ until $\boldsymbol{M}$ converges to the correct one.

The third technical challenge is to minimize the overhead of collecting data to construct $\boldsymbol{M}$, which requires continuous monitoring of which user is accessing what contents. To address this issue, the monitoring process will be started to update $\boldsymbol{M}$, only if certain conditions are satisfied. Specifically, if user similarity matrix $\boldsymbol{S}$ derived from updated $\boldsymbol{M}$ is significantly different from current, the monitoring process will continue for a period of time; otherwise, it stops collecting any data.

An important item to consider when developing a solution is the timing constraint to achieve QoE. It was discovered that traffic characteristics always follow a clear daily pattern in traditional cellular networks [23], which will also be the case in a HetNet. A viable solution to address this problem is to deconstruct $\boldsymbol{M}$ into several smaller matrices, based on the time periods during which they are generated. Decoupling $\boldsymbol{M}$ into smaller matrices will again cause data sparsity issues that will degrade CF's performance. Additionally, we could pre-process $\boldsymbol{M}$, by grouping correlated contents together, to reduce the number of columns in it. With the pro-processed $\boldsymbol{M}$, all contents in the same category need to be cached on edge servers, leading to a huge networking

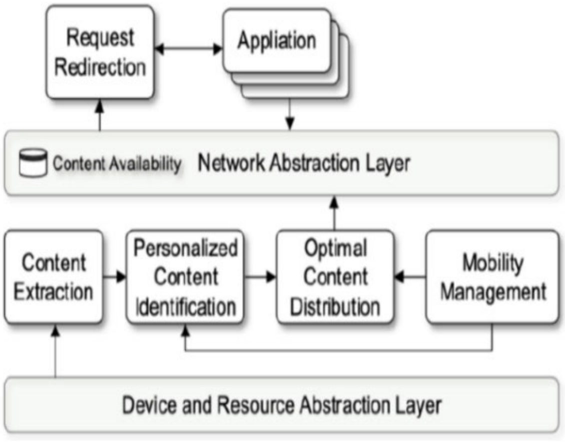
Request Redirection
Appliation
Content Availability
Network Abstraction Layer
Content Extraction
Personalized Content Identification
Optimal Content Distribution
Mobility Management
Device and Resource Abstraction Layer

FIGURE 10. Proposed architecture for optimal content caching.

overhead. To mitigate this issue, we propose to divide matrix $\boldsymbol{M}$ into a set of sub-matrices $\{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots \boldsymbol{etc}.\}$ wherein each sub-matrix $\boldsymbol{M}_l$ only contains users with similar preferences. A genetic algorithm based iterative collaborative filtering (GAICF) algorithm will then be applied to each $\boldsymbol{M}_l$ to infer most popular contents, not a group of contents, for every user.

Task 2: SDN-Enabled Optimal Content Caching and Dissemination: Optimal content caching and dissemination (OCCD) in HetNets deals with the problem of caching personalized popular contents to edge based on predicted user locations, aiming at minimizing the overhead in caching and disseminating the contents. To solve the OCCD problem, it is critical to have a global view of the current network status that is impossible in traditional networks but feasible in an SDN based network. Figure 10 provides an architecture for optimal content caching system wherein the key component, SDN controller, consists of three sub-layers: network abstraction layer, network management layer, and device and resource abstraction layer. Because the current SDN technology focuses on flows rather than packets, we propose a "content extraction" module to support deep packet inspections so that contents transferred in each flow can be extracted. The "personalized content identification" module is introduced to identify the most-frequently and most-likely-accessed contents for each user. Based on the predicted user's locations and most popular contents, "optimal content distribution" module computes a solution to OCCD problem.

Solving the OCCD problem requests a priori knowledge about future network demands, i.e., the optimal solution to OCCD cannot be directly applied in practice. In particular, the servers that serve a user change over time and space, i.e., the spatial location of a user at a given time determines the server to which the request should be addressed. To this end, we propose a family of heuristic algorithms to determine what contents to replicate on which server to minimize the overall networking overhead, based on the predictions of user's future locations and his most-likely-accessed contents. The proposed online algorithm requires that the SDN controller keeps track of the number of requests received on edge servers and predicts future demand on every edge server. Based on the predicted demands and the current network status.

B. QOE-AWARE NETWORK SLICING IN HETNETS

The focus of this research thrust is to develop a predictive and adaptive mechanism to increase the likelihood of hand-off success, while achieving weighted-fair resource allocation among end users with different priorities and traffic types. HetNets densification has potential to increase effective bandwidth available to users, but also increases hand-off frequency, due to small coverage areas. The need to evolve future mobile networks toward supporting, reliably and efficiently, a wider range of telecommunication and multimedia content distribution applications is a critical design requirement of next generation wireless networks. This section focuses on three technical problems: (1) identifying best base station(s) to associate for each user, (2) optimizing spectrum reservation strategy to balance user's QoE and resource utilization efficiency, (3) designing algorithms to handle congestion.

(1) Predictive base station association: With overlapping coverages supported by small cells of 5G and heterogeneous radio access technologies, the key question for an end user is selecting a cluster of the best base station(s) to associate along the predicted trajectory. Both local and global optimization exist in this process. In detail, base station(s) should optimize a user's QoE, meanwhile, balance the workload among all base stations. Therefore, two sub-questions exist here. The first one is to identify base station(s) that support QoE for end users while considering users' hardware limitations, mobility, application requirements, and potential interference level. The second one is about load balance among all base station(s) to avoid potential contention. To identify a cluster of base station(s) that will provide the best service, a predictive model for future base station associations should be established.

(2) Predictive and adaptive spectrum Pre-Allocation: There is a need for new approaches to achieve efficient and practically feasible bandwidth allocation to guarantee uninterpreted service for end users, after hand-offs occur, while maximizing network resource utilization. We argue that it is critical to design a resource reservation scheme in 5G HetNets to enhance the support of QoE. Spectrum reservation for mobile users can be achieved by edge computing, service based architectural components of 5G (3GPP release 18 [19]) and SDN. On the contrary to classical approaches, wireless virtualization on edge computing aims at centralizing the control plane for both radio access network and spectrum to support wireless services and higher QoE requirements [24]. By leveraging local radio signal processing, cooperative radio resource management, and distributed storage capabilities in edge devices, this approach maximizes the benefits while minimizing the strain on front haul and eliminates the need for centralized base band units to perform extensive radio signal processing on a large scale.

(3) Adaptive spectrum allocation strategy to handle network congestion: Developing an adaptive spectrum allocation framework that adaptively handles network congestion while considering the heterogeneity of the mobile devices. When network congestion occurs, the adaptive spectrum allocation strategy adapts the bandwidth allocated to the applications in a fair fashion, according to their priorities. Moreover, factors such as spectrum utilization efficiency, mobile units, QoE requirements and priority levels, moving characteristics such as speed (impacts hand-offs), and user mobility classes should be considered in the adaptive spectrum allocation framework.

C. PREDICTIVE MOBILITY MODEL FOR ADAPTIVE TIMED-QOE GUARANTEES IN HETNETS

The goal here is to design a mobility prediction model to facilitate personalized content caching and dissemination as well as resource reservation. It is important as personalized content caching and resource allocation are feasible only if accurate knowledge of end users' paths and their arrival and departure are available for the duration of wireless communications. Inaccurate predictions will lead to not only waste of resources but also low QoE or even service termination. However, acquiring such knowledge is challenging in a wireless environment as it is characterized by a high scale of uncertainty both in user mobility and resource availability.

Mobility prediction for end users in HetNets requires both temporal and spatial information, which is different from existing mobility research that focuses on either spatial or temporal domain. In the spatial domain, three main factors considered which are destination recommendation, path prediction, and prediction on next cell. In the temporal domain, total travel, and sojourn times, as well as remaining time in a cell can be predicted [25]. In this research our proposed framework supports all mobility modes, including both human and urban mobility, in temporal-spatial domain. As illustrated in Figure 8, two databases serve as input for the mobility prediction, namely traffic profile and user mobility profile. In traffic profile database, traffic conditions such as congestion, accidents, road condition, are stored. User mobility profile database contains two types of information: mobility profile and user preference profile. Mobility profile has input such as transportation types, mobile unit's position, direction, speed, while user preference profile has input such as most traveled routes, most visited designations, driving patterns, etc. To achieve this goal, the most likely cluster (MLC) which forms a collection of contiguous coverages that are expected to be visited by the target mobile user need to be predicted to pre-allocate spectrum and cache contents, based on the data collected on edge servers. Here, coverage is used, instead of cells or base stations, to capture the geographic area that will likely be visited by the user. From the predicted coverage information, the cell, base stations, and edge servers to be visited can be derived easily.

IV. CONCLUSION

In recent years, the consumption trend of multimedia services which run enhanced applications such as live video streaming, videos on demand, immersive videos, and video gaming have put tremendous pressure on existing mobile network infrastructures in terms of management of multimedia services. This trend is expected to last even over beyond-5G future networks as end-users become adapted to more resource demanding services with higher quality. While existing solutions concentrate on improving network capacity from the data transmission perspective, it overlooks the importance of enhancing the support of user's quality of experience for the evolving new services. In this sense, the optimization of the quality of experience is increasingly receiving higher attention as service providers are converging towards the future softwarized networks. The present research paper is focused on the evaluation of various multimedia services and applications that target optimizing the QoE. It explores how the QoE of current internet multimedia streaming services is impacted by factors such as network infrastructure, streaming technology, and user terminals. In the first phase of this research, a test methodology is designed for assessing QoE of live video streaming over 5G cellular networks using buffer-based adaptive video streaming algorithms. Then we proposed a framework to enhance the support for QoE and explored new approaches that leverage existing and new cutting-edge technologies such as SDN, predictive mobility models, network slicing and edge computing to achieve practically feasible spectrum allocation and personalized content caching and dissemination, to provide uninterrupted multimedia services to end users. We have provided an investigation of three coherent research thrusts to support our proposed QoE-aware framework which are: mobility-aware SDN-enabled caching and dissemination of personalized contents, smart orchestrator for QoE-aware efficient resource allocation in heterogenous networks, and predictive mobility model for adaptive timed-QoE guarantees in in heterogenous networks. For future works we will provide an implementation of our proposed framework, validation and verification based on system-level simulations, real-world datasets, and experiments in a prototype of the proposed system.

REFERENCES

[1] H. Chen, H. Sarieddeen, T. Ballal, H. Wymeersch, M.-S. Alouini, and T. Y. Al-Naffouri, "A tutorial on terahertz-band localization for 6G communication systems," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 3, pp. 1780–1815, 3rd Quart., 2022, doi: 10.1109/COMST.2022.3178209.

[2] P. Pérez, D. Corregidor, E. Garrido, I. Benito, E. González-Sosa, J. Cabrera, D. Berjón, C. Díaz, F. Morán, N. García, J. Igual, and J. Ruiz, "Live free-viewpoint video in immersive media production over 5G networks," *IEEE Trans. Broadcast.*, vol. 68, no. 2, pp. 439–450, Jun. 2022, doi: 10.1109/TBC.2022.3154612.

[3] H.-W. Kao and E. H. Wu, "QoE sustainability on 5G and beyond 5G networks," *IEEE Wireless Commun.*, vol. 30, no. 1, pp. 118–125, Feb. 2023, doi: 10.1109/MWC.007.2200260.

[4] K. Cao, Y. Liu, G. Meng, and Q. Sun, "An overview on edge computing research," *IEEE Access*, vol. 8, pp. 85714–85728, 2020, doi: 10.1109/ACCESS.2020.2991734.

[5] *Multi-Access Edge Computing (MEC); Framework and Reference Architecture*, document ETSI GS MEC, Jan. 2019.

[6] J. Chen, Z. Luo, Z. Wang, M. Hu, and D. Wu, "Live360: Viewport-aware transmission optimization in live 360-degree video streaming," *IEEE Trans. Broadcast.*, vol. 69, no. 1, pp. 85–96, Mar. 2023, doi: 10.1109/TBC.2023.3234405.

[7] M. Liubogoshchev, D. Zudin, A. Krasilov, A. Krotov, and E. Khorov, "DeSlice: An architecture for QoE-aware and isolated RAN slicing," *Sensors*, vol. 23, no. 9, p. 4351, Apr. 2023, doi: 10.3390/s23094351.

[8] A. Bentaleb, B. Taani, A. C. Begen, C. Timmerer, and R. Zimmermann, "A survey on bitrate adaptation schemes for streaming media over HTTP," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 562–585, 1st Quart., 2019, doi: 10.1109/COMST.2018.2862938.

[9] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, "Probe and adapt: Rate adaptation for HTTP video streaming at scale," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 719–733, Apr. 2014, doi: 10.1109/JSAC.2014.140405.

[10] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," in *Proc. ACM Conf. SIGCOMM*. Chicago, IL, USA: ACM, Aug. 2014, pp. 187–198.

[11] Y. Sani, A. Mauthe, and C. Edwards, "Modelling video rate evolution in adaptive bitrate selection," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Miami, FL, USA, Dec. 2015, pp. 89–94, doi: 10.1109/ISM.2015.65.

[12] H. Wu, X. Li, G. Wang, G. Cheng, and X. Hu, "Resolution identification of encrypted video streaming based on HTTP/2 features," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 2, pp. 1–23, May 2023, doi: 10.1145/3551891.

[13] J. P. Esper, A. C. B. L. Moncao, K. B. C. Rodrigues, C. B. Both, S. L. Correa, and K. V. Cardoso, "QoE-DASH: DASH QoE performance evaluation tool for edge-cache and recommendation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2022, pp. 757–762, doi: 10.1109/ICC45855.2022.9839234.

[14] W. U. Rahman, Y.-S. Choi, and K. Chung, "Performance evaluation of video streaming application over CoAP in IoT," *IEEE Access*, vol. 7, pp. 39852–39861, 2019, doi: 10.1109/ACCESS.2019.2907157.

[15] J. Van Der Hooft, M. T. Vega, C. Timmerer, A. C. Begen, F. De Turck, and R. Schatz, "Objective and subjective QoE evaluation for adaptive point cloud streaming," in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Athlone, Ireland, May 2020, pp. 1–6, doi: 10.1109/QoMEX48832.2020.9123081.

[16] G. Dimopoulos, I. Leontiadis, P. Barlet-Ros, and K. Papagiannaki, "Measuring video QoE from encrypted traffic," in *Proc. Internet Meas. Conf.* Santa Monica, CA, USA: ACM, Nov. 2016, pp. 513–526, doi: 10.1145/2987443.2987459.

[17] D. Raca, M. Salian, and A. H. Zahran, "Enabling scalable emulation of differentiated services in Mininet," in *Proc. 13th ACM Multimedia Syst. Conf.* Athlone Ireland: ACM, Jun. 2022, pp. 240–245, doi: 10.1145/3524273.3532893.

[18] J. O'Sullivan, D. Raca, and J. J. Quinlan, "Godash 2.0—The next evolution of HAS evaluation," in *Proc. IEEE 21st Int. Symp. 'A World Wireless, Mobile Multimedia Networks' (WoWMoM)*, Aug. 2020, pp. 185–187, doi: 10.1109/WoWMoM49955.2020.00043.

[19] X. Lin, "An overview of 5G advanced evolution in 3GPP release 18," *IEEE Commun. Standards Mag.*, vol. 6, no. 3, pp. 77–83, Sep. 2022, doi: 10.1109/MCOMSTD.0001.2200001.

[20] Z. Su, Y. Hui, and Q. Yang, "The next generation vehicular networks: A content-centric framework," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 60–66, Feb. 2017, doi: 10.1109/MWC.2017.1600195WC.

[21] B. Zolfaghari, G. Srivastava, S. Roy, H. R. Nemati, F. Afghah, T. Koshiba, A. Razi, K. Bibak, P. Mitra, and B. K. Rai, "Content delivery networks: State of the art, trends, and future roadmap," *ACM Comput. Surveys*, vol. 53, no. 2, pp. 1–34, Mar. 2021, doi: 10.1145/3380613.

[22] S. Wang, L. Cao, Y. Wang, Q. Z. Sheng, M. A. Orgun, and D. Lian, "A survey on session-based recommender systems," *ACM Comput. Surv.*, vol. 54, no. 7, pp. 1–38, Sep. 2022, doi: 10.1145/3465401.

[23] G. Soos, D. Ficzere, and P. Varga, "Towards traffic identification and modeling for 5G application use-cases," *Electronics*, vol. 9, no. 4, p. 640, Apr. 2020, doi: 10.3390/electronics9040640.

[24] H. R. D. Filgueiras, E. S. Lima, M. S. B. Cunha, C. H. D. S. Lopes, L. C. De Souza, R. M. Borges, L. A. M. Pereira, T. H. Brandão, T. P. V. Andrade, L. C. Alexandre, G. Neto, A. Linhares, L. L. Mendes, M. A. Romero, and A. Cerqueira, "Wireless and optical convergent access technologies toward 6G," *IEEE Access*, vol. 11, pp. 9232–9259, 2023, doi: 10.1109/ACCESS.2023.3239807.

[25] Q. Cui, X. Hu, W. Ni, X. Tao, P. Zhang, T. Chen, K.-C. Chen, and M. Haenggi, "Vehicular mobility patterns and their applications to Internet-of-Vehicles: A comprehensive survey," *Sci. China Inf. Sci.*, vol. 65, no. 11, Nov. 2022, Art. no. 211301, doi: 10.1007/s11432-021-3487-x.

MOHAMAD T. SULTAN received the B.Sc. degree in computer engineering and the master's degree in information technology from Universiti Tenaga Nasional (UNITEN), Kuala Lumpur, Malaysia, in 2009 and 2012, respectively. He is currently a Computer Engineer and a Researcher. He is also with the College of Information Technology, United Arab Emirates University, United Arab Emirates. His current research interests include computer networks, machine learning, connected and autonomous vehicles, and quality of experience.

HESHAM EL SAYED is currently a Professor with the College of Information Technology, United Arab Emirates University (UAE University), United Arab Emirates. Before joining UAE University, he held several industrial positions with Nortel Networks, Wind River Systems (acquired by Intel Corporation), and Paragon Networks (acquired by Carrier Access Corporation, then Turin Networks, Force10 Networks, and Dell Inc.) During his tenure in the industry, he led several projects focused on the performance optimization of network protocols and architectures. His current research interests include the Internet of Things, trust management, intelligent transportation systems, and vehicular ad-hoc networks. His contributions in these areas had a tangible impact on the performance of real commercial products, which led to the receiving of several prestigious awards, including the Nortel Networks President's Award of Excellence in Leadership and the Paragon Networks CEO Award of Excellence.