# QoE Ready to Respond: A QoE-aware MEC Selection Scheme for DASH-based Adaptive Video Streaming to Mobile Users

Wanxin Shi*[†], Qing Li[†‡], Ruishan Zhang*, Gengbiao Shen*
Yong Jiang*[†], Zhenhui Yuan[+], Gabriel-Miro Muntean[§]
* International Graduate School, Tsinghua University, Shenzhen, China
[†] Peng Cheng Laboratory, Shenzhen, China
[‡] Southern University of Science and Technology, Shenzhen, China
[+] Department of Computer and Information Sciences, Northumbria University, UK
[§] School of Electronic Engineering, Dublin City University, Galsnevin Campus, Dublin 9, Ireland
shiwx17@mails.tsinghua.edu.cn,liq@pcl.ac.cn,zrs20@mails.tsinghua.edu.cn
sgb16@mails.tsinghua.edu.cn,jiangy@sz.tsinghua.edu.cn,zhenhui.yuan@northumbria.ac.uk,gabriel.muntean@dcu.ie

## ABSTRACT

The Multi-access Edge Computing (MEC) paradigm offers cloud-computing support to rich media applications, including Dynamic Adaptive Streaming over HTTP (DASH)-based ones at the edge of the network, close to mobile users. MEC servers, typically deployed at base stations (BS), help reduce latency and improve quality of experience (QoE) of video streaming. Unfortunately the communications involving mobile users require handovers between BSs and these influence both transmission efficiency because of the relative position of the MEC servers and transit cost. At the same time, serving MEC for a mobile user should not necessarily be changed when handover occurs. This paper introduces **QoE Ready to Respond (QoE-R2R)**, a QoE-aware MEC Selection scheme for DASH-based mobile adaptive video streaming for optimizing video transmission in a MEC-supported network environment. Simulation-based testing shows that the proposed (**QoE-R2R**) scheme outperforms some traditional alternative solutions. Compared to hit rate and delay-based schemes, QoE-R2R reduces by 27.6% transmission time and improves with 6.2% QoE.

## CCS CONCEPTS

• **Networks** → *Mobile networks*; *In-network processing*; • **Information systems** → **Multimedia streaming**.

## KEYWORDS

DASH, mobile users, MEC selection, handover

**ACM Reference Format:**
Wanxin Shi*[†], Qing Li[†‡], Ruishan Zhang*, Gengbiao Shen*, Yong Jiang*[†], Zhenhui Yuan[+], Gabriel-Miro Muntean[§]. 2021. QoE Ready to Respond: A QoE-aware MEC Selection Scheme for DASH-based Adaptive Video Streaming to Mobile Users. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475325

## 1 INTRODUCTION

The latest rapid increase in the number of smart mobile device users and large diversity of video services put under pressure video transmissions at high quality. Fuelled by this large number of video transmissions, the global video streaming market is expected to reach USD 184.2 billion by 2027 [22]. However, supporting high viewer quality of experience (QoE) for these video services is very challenging and therefore there is a need for optimization of content delivery. 5G network solutions already support video transmissions with large bandwidth requirements, but the latest use cases which include ultra-high definition (UHD) videos, virtual/augmented reality (VR/AR) and interactive/panoramic live videos require additional support. Solutions employ Dynamic Adaptive Streaming over HTTP (DASH)-based adaptation, which enables dynamic adjustment of the video content to suit network delivery conditions and Multi-access Edge Computing (MEC), which brings computation capabilities to the network edge and therefore closer to users.

Unfortunately, the existing solutions do not fully solve the problems, especially related to mobility, mostly due to the need for handover (HO) between base stations (BS) and lack of consideration of MEC states. Some of these aspects are discussed next. **First**, the expected capacity gain offered by 5G network densification is achieved at the expense of increased HO rates. This may affect network operation and degrade user QoE. Flexible HO methods for different applications may be needed [13, 25]. For example, a dense deployment of BSs may cause a ping-pong effect during handover. Although some research [12, 21] focused on overcoming the ping-pong effect, they did not find a solution for sequential and chunk-based DASH videos or consideration of MEC cache states. **Secondly**, the use of MEC brings content closer to mobile users, but it also introduces the problem of MEC selection [11]. The serving MEC may not be the one deployed on the serving BS. The handover will make the user communicate with a target BS, but it will not bring the connection to a neighbor MEC. Therefore, the content providers or network operators need to find an appropriate MEC for each mobile user. There is a need for a MEC selection process which has to be performed in a timely fashion, in order to support high service quality.

Corresponding author: Qing Li (liq@pcl.ac.cn).

*MM '21, October 20–24, 2021, Virtual Event, China*
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475325

In this context, a novel MEC selection scheme is designed which considers HO execution and edge cache states in order to achieve high QoE for DASH-based adaptive video services. The scheme is named **QoE Ready to Respond (QoE-R2R)**. QoE-R2 considers that the response efficiency is the best when the selected MEC is mounted on the serving BS. It is also assumed that MEC caches the requested content. We design hit rate and delay based methods for MEC selection and find that hit rate and delay may not be the most important metrics. Then a QoE-aware method is proposed to directly optimize QoE. So the proposed policies are executed based on cache hit, delay and QoE with the goal of utilizing the cached content in the MEC mounted on the serving BS as much as possible.

The core of mobile network, i.e., Evolved Packet Core (EPC) or 5G Core (5GC) includes the function of mobility management. To be compatible with MEC selection, the mobility management should try its best to avoid triggering HO events before finishing delivering a video chunk. This is as DASH is chunk-based and the playback of a given chunk cannot begin until the download of the entire chunk is finished. If the HO is triggered during the download, the HO delay will influence the playback of the chunk. Besides, the MEC cache is updated according to the requests received by the BS instead of the MEC itself. This is as the ideal result of the proposed scheme is to make the MEC cache host of the most frequently-requested content for the users served by its BS. Then, by employing the proposed scheme, the user can be served by the MEC mounted on the serving BS in the near future. The proposed solution enables to finish delivering a whole video chunk before executing HO and utilize the MEC mounted on the serving BS as much as possible.

The following are the contributions of this paper.

- It investigates user request and MEC response pattern and formulates the response problem in the context of a MEC-based edge framework for DASH-based video delivery.
- A new QoE-aware MEC selection scheme (QoE-R2R) which considers user request-motivated HO decision and edge cache states is designed and described.
- The paper validates the proposed scheme using simulations, showing how QoE-R2R outperforms existing solutions in terms of responsiveness and estimated user QoE.

The proposed scheme QoE-R2R was implemented and tested in the discrete-event network simulator NS-3 [18]. The experiments are performed in two major scenarios, involving a single user and multiple users, respectively. The former one tests the user's request mode and the server's response pattern. The latter one confirms that the QoE-aware selection method takes effect and outperforms the other approaches in terms of transmission duration and QoE.

This paper is organized as follows. Related works are discussed in Section 2. The network model is introduced in Section 3. The proposed MEC selection method is presented in Section 4. In Section 5 includes simulations-based testing and results. Section 6 concludes the paper.

## 2 RELATED WORKS

DASH video streaming accounts for a large part of traffic in mobile networks. DASH viewers request video content dynamically, chunk by chunk according to network capacity. Various researchers perform research which focuses on client-side Adaptive Bit Rate algorithms (ABR) [14, 26] and edge-side optimization like transcoding [24], super-resolution [27] and prefetching [8]. Highly relevant to this paper, next we discuss some research works on server selection and handover in a mobile video delivery context.

### 2.1 Server Selection

Various server selection schemes, including remote-side and edge-based ones are proposed [4, 16, 29]. The authors of [30] introduce a Multi-user Edge server Selection method based on Particle swarm optimization (MESP). The MESP method selects a mobile edge server in advance of its use within a polynomial time. An edge server selection method based on a genetic algorithm and a simulated annealing algorithm is devised in [31], which minimizes the overhead of the user. The authors of [29] proposed a new server-selection policy for multiple servers based on the defined unified cost metric. This policy takes the network, latency, and media distortion into account. Some learning-based methods are proposed in [10, 28] for the MEC server selection and focus on computation offloading. Although these methods are intelligent, they are not dedicated to DASH streaming and lack consideration of MEC states.

Related to DASH-based video streaming, sequential and chunk-based characteristics may be important for MEC selection. As a key metric, QoE should also be considered in the MEC selection process. With MEC servers deployed at BSs, content providers or network operators should find a proper MEC to serve user equipments (UE), which should take into consideration multiple edge-related information e.g., cache state, workload and link capacity.

### 2.2 Handover

The traditional handover algorithms in Long Term Evolution (LTE) are mainly based on signal strength, i.e., Reference Signal Receiving Quality (RSRQ) and Reference Signal Receiving Power (RSRP). Apart from signal strength, different other metrics could be considered in HO such as throughput and SINR, which may depend on precise prediction. To overcome the ping-pong effect during the traditional signal-based handover, Leu et al. introduce a class of fast handover algorithms that removes only the fast fading component from the received signal strength [12]. It is helpful to reduce the impact of corner effects. The handover algorithm introduced by [21] employs local averaging, a drop timer, and hysteresis to eliminate ping-pong. Based on the popular machine learning, there are some works which optimize HO by using learning methods. Colin et al. [7] execute threshold selection for handover parameters in accordance with the reward configuration from throughput. The HO is optimized by solving the contextual bandit problems. Alkhateeb et al. [3] predict channel blockage to assure reliability and reduce latency, which avoids the sudden blockage of the line-of-sight link. However, none of these works addresses the DASH-based video delivery supported by MEC, while also considering the user and edge states. Therefore, there is a need to consider both the features of DASH video streaming and MEC paradigm in our work.

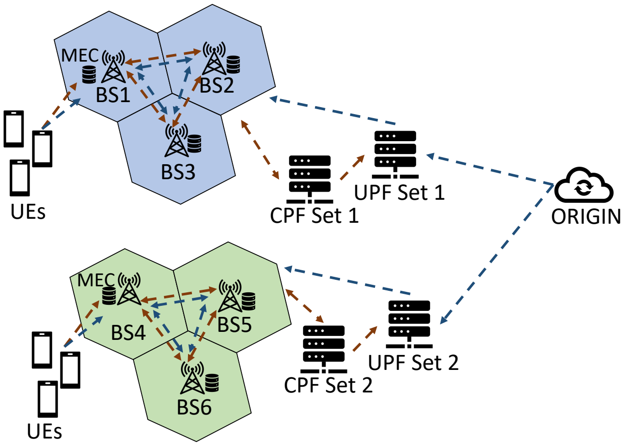


**Figure 1: Network model**

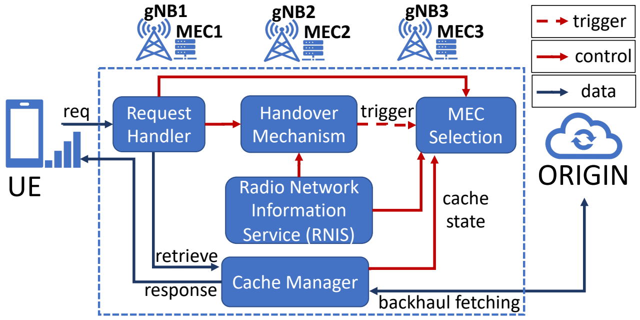


**Figure 2: The components in the system**

# 3 SYSTEM MODEL

## 3.1 Architecture

Fig. 1 illustrates a simplified 5G network architecture [1], in which each UE accesses the cellular network via a nearby BS. In order to address the issues related to high traffic load in the backhaul paths, content localization is employed. Therefore, the content is brought closer to its consumers by deploying MECs. Assuming cooperation between content providers and network operators, each BS may have associated a MEC server to cache videos and provide better viewing service for end users. MEC servers can be flexibly deployed in different locations (i.e., from the local BS to EPC/5GC). EPC/5GC refers to the core of mobile network 4G/5G, acting as a traffic administrator for mobile users. As the connection anchor between 5G network and MEC, all core network data must be forwarded by the User Plane Function (UPF) [2] before flowing to the external network.

Fig. 2 presents the major system components of the proposed solution. These components include the *Request Handler*, *MEC Selection* module and *Cache Manager* - deployed at the level of MEC and *Handover Mechanism (HO)* module and *Radio Network Information Service (RNIS)* module, located at the level of EPC/5GC. When a UE requests a video chunk, the *Request Handler* block is responsible to resolve the request and provide related information to *HO* and *MEC Selection* modules. The information provided includes the ID of the requested content and UE state data (e.g., buffer occupancy). *RNIS* also assists the *HO* module with periodical reports about the available BSs and their signal strengths. *HO* module makes HO decisions in accordance with the information provided by the *Request Handler* and *RNIS*. If a HO is triggered, *MEC Selection* module is in charge with selecting an appropriate MEC for responding to the user. *Cache Manager* module acts as the decision-maker for content placement/replacement and retrieves content to be sent back to the user. It takes into account content popularity and storage space for cache placement/replacement. *Cache Manager* also provides cache state data to the *MEC Selection* module after cache replacement.

**Table 1: Summary of mathematical symbols**

| Symbols | Definition |
|---|---|
| $B_t$ | $b_i \in B_t$, $i = 1, \dots, n$, the accessible BSs for $U_e$ |
| $S$ | $s_j \in S$, $i = 1, \dots, n'$, the available MEC video server for $U_e$ |
| $U_e$ | the current user equipment |
| $D(U_e, b_i, t)$ | the distance between $U_e$ and $b_i$ at the moment $t$ |
| $w(t, b_i)$ | the workload of $b_i$ at the moment $t$ |
| $w'(t, s_j)$ | the workload of $s_j$ at the moment $t$ |
| $GA(U_e, b_i, t)$ | the channel gain between $U_e$ and $b_i$ at the moment $t$ |
| $PW(b_i)$ | the transmission power of $b_i$ |
| $R(U_e, b_i, t)$ | the transmission rate from $b_i$ to $U_e$ at the moment $t$ |
| $R(s_r, s_{r'}, t)$ | the transmission rate between $s_r$ and $s_{r'}$ at the moment $t$ |
| $\mathcal{N}_{b_i}$ | the neighbor BS set of $b_i$ |
| $\mathcal{N}_{s_j}$ | the nearby MEC set of $s_j$ |
| $V_m^k$ | the $k$th chunk of the video $V$ with the representation $m$ |
| $f_{i,o}$ | the number of requests for $o$ per minute from $U_e$ to BS $b_i$ |

## 3.2 Rate Definitions for Different Transmission Situations

Table. 1 lists the mathematical symbols used in this paper. Figure. 3 illustrates the mobile data transmission process from origin to UE, which assumes MECs deployed at BSs. With deployment of MECs, the UE may be in two states. **State 1** considers that the serving MEC is mounted on the serving BS as illustrated in Figure. 3. In such a situation, the MEC responds to the UE directly, as shown in case $(a)$ or avails from the support of a nearby MEC as pictured in case $(b)$. If there is no MEC caching the requested content, the content will be fetched from the origin server, as indicated in case $(c)$. **State 2** considers that the serving MEC is not mounted on the serving BS. In this situation, there are also three cases $(d)$, $(e)$ and $(f)$, depending on whether the content is located at the serving MEC, at a nearby MEC or at the origin server. Case $(g)$ which indicates that there is no MEC service is out of the scope of this work. The transmission rates in different situations are detailed as follows.

*Rate from BS to UE:* We define the base station set $B(t) = \{b_i\}$ as containing the accessible BSs for $U_e$. At moment $t$, the distance between $U_e$ and $b_i$ is $D(t, U_e, b_i)$. The workload of $b_i$ is defined as $w(t, b_i)$ and we assume the available bandwidth between $U_e$ and $b_i$ is $R[t, D(t, U_e, b_i), w(t, b_i)]$. In this context, the transmission rate from $b_i$ to $U_e$ is related to distance and BS workload. To better illustrate the available bandwidth in a cellular network scenario, $R[t, D(t, U_e, b_i), w(t, b_i)]$ should be modified as $R(U_e, b_i, t)$ which is actually pertinent to transmission power and channel gain. We

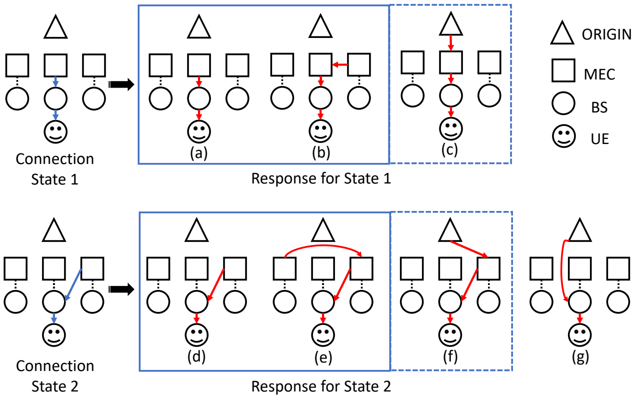


**Figure 3: UE connection states and MEC response modes**

can use typical values of transmission power introduced in [23], e.g., $PW(b_i) = 20W$. Here, the channel gain between $U_e$ and $b_i$ at moment $t$ can be modeled as $GA(U_e, b_i, t)$. Different BSs may have different propagation and loss values [15, 19]. The channel gain in [5] is defined as follows:

$$\begin{aligned} &GA(U_e, b_i, t) \\ &= 40 \cdot (1 - 4 \times 10^{-3} \times \mathcal{HT}) \cdot log_{10}[D(t, U_e, b_i)] \\ &- 18 \cdot log_{10}(\mathcal{HT}) + 21 \cdot log_{10}(\mathcal{FR}) + 80 + \mathcal{FD} \end{aligned} \tag{1}$$

where $\mathcal{HT}$, $\mathcal{FR}$ and $\mathcal{FD}$ are antenna height in meters, carrier frequency in MHz and shadowing channel fading in dB, respectively. We define noise power as $\sigma^2$. The Signal to Interference plus Noise Ratio (SINR) can be computed as follows:

$$SINR(U_e, b_i, t) = \frac{GA(U_e, b_i, t) \cdot PW(b_i)}{\sigma^2} \tag{2}$$

According to the Shannon formula, the instantaneous transmission rate from $b_i$ to $U_e$ is given by:

$$R(U_e, b_i, t) = BW(U_e, b_i, t) \cdot log_2[1 + SINR(U_e, b_i, t)] \tag{3}$$

where $U_e$ associated with $b_i$ is allocated a fraction of bandwidth $BW(U_e, b_i, t)$ that in general is a fixed value. $BW(U_e, b_i, t)$ is channel bandwidth referring to the frequency range (Hz) in which the signal can be transmitted with appropriate fidelity. It is channel inherent and it is not related to the signal carried.

*Rate from MEC to MEC:* We define the MEC set $S = \{s_j\}$ as the available MEC video servers for $U_e$. The workload of $s_j$ influences the transmission rate between these MECs because the links carry traffic. We define the workload as $w'(t, s_j)$. The transmission rate between two MEC servers $s_r$ and $s_{r'}$ at moment $t$ is $R(w'(t, s_r), w'(t, s_{r'})) \Leftrightarrow R(s_r, s_{r'}, t)$.

## 3.3 Transmission Formulation

Due to user mobility, it is difficult for a user to download the whole video within the coverage of a single BS, especially when the user is at high speed or the size of file is large. Note that the content may be retrieved from the serving MEC, a neighboring MEC or the origin server. Therefore, the transmission formulation is based on video chunks, according to the chunk-based characteristics of DASH video transmissions.

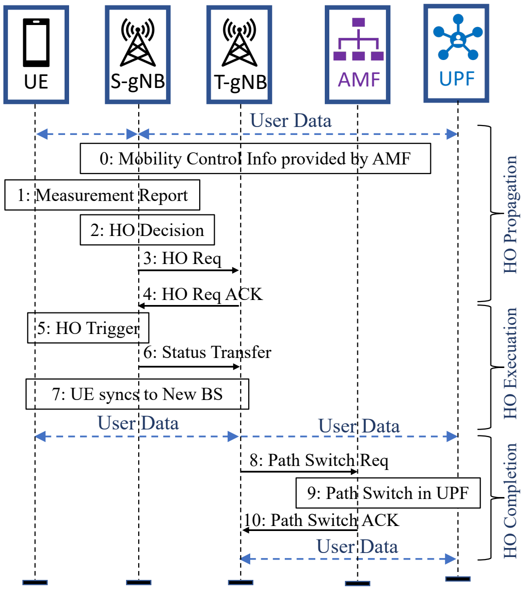


**Figure 4: Handover process**

We assume that $U_e$ is currently attached to BS $a(t) \in B(t)$ and served by MEC $c(t) \in S$. Note that MEC $c(t)$ may not be mounted on BS $a(t)$ due to BS-MEC asynchronization. Assume $U_e$ current requested video chunk is $V_m^k$ of size $SIZ(V_m^k)$. Here, $R(U_e, a(t), t)$ is simply mapped to $R(U_e, a(k), k)$ to represent the average throughput during transmission of the $k^{th}$ chunk. The transmission time for sending $V_m^k$ from the MEC server mounted on $a(t)$ to $U_e$ is expressed as:

$$\tau^{a(t) \to U_e}(V_m^k) = \frac{SIZ(V_m^k)}{R[U_e, a(t), t]} = \frac{SIZ(V_m^k)}{R[U_e, a(k), k]} \tag{4}$$

However, during sending of chunk $V_m^k$, $a(k)$ may not be invariant due to handovers. Here we assume that the BS tries not to handover until finishes sending a video chunk. This problem will be discussed and solved later.

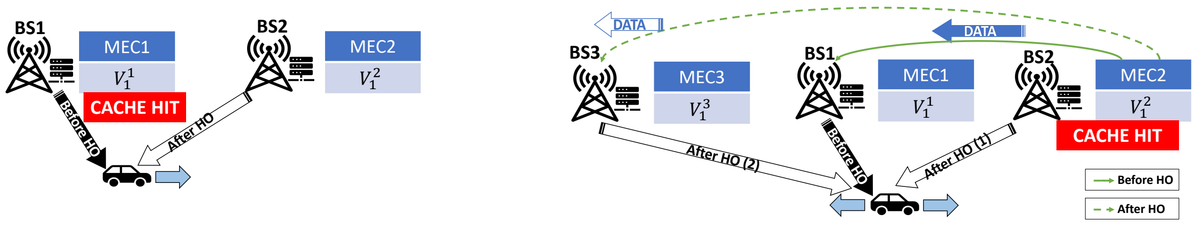
When the $U_e$ requested chunk is not cached at $c(t)$, the request will be forwarded to a nearby MEC $s_r \in \mathcal{N}_{c(t)} \in S$. First, $\mathcal{N}_{b_i}$ is defined as the neighbor BS set of $b_i$. Similarly, the set of **neighbor** BSs of $a(t)$ is $\mathcal{N}_{a(t)}$. Finally, $\mathcal{N}_{c(t)}$ is the set of nearby MECs of $c(t)$. The MEC(s) in $\mathcal{N}_{c(t)}$ with the requested content is(are) mounted on the **nearby** BSs that are not exactly $\mathcal{N}_{a(t)}$. The transmission time between $s_r$ and $c(t)$ is denoted as $\tau^{s_r \to c(t)}(V_m^k)$ when delivering $V_m^k$ and is given as follows.

$$\tau^{s_r \to c(t)}(V_m^k) = \frac{SIZ(V_m^k)}{R[c(t), s_r, t]} = \frac{SIZ(V_m^k)}{R[c(k), s_r, k]} \tag{5}$$

where $R[c(k), s_r, k]$ is link capacity between $c(k)$ and $s_r$.

When there is no MEC caching the requested content, $c(t)$ forwards the request to the origin server through the backhaul path. We define $\tau^{c(t) \to OS}$ as the transmission time between $c(t)$ and the origin server, which is given by:

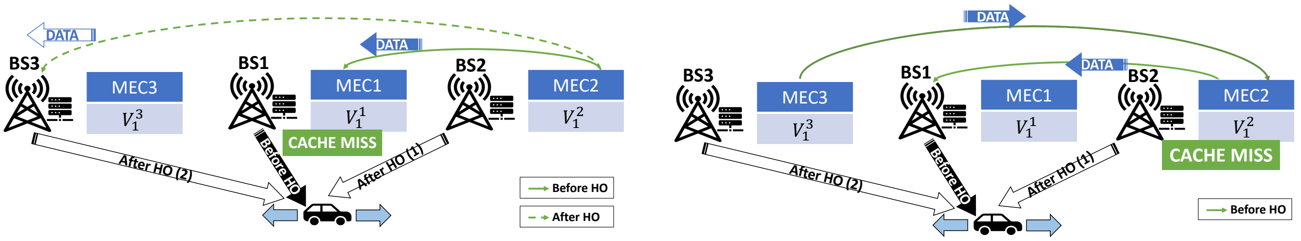
$$\tau^{OS \to c(t)}(V_m^k) = \frac{SIZ(V_m^k)}{R[c(t), OS, t]} = \frac{SIZ(V_m^k)}{R[c(k), OS, k]} \tag{6}$$



**(a) Current MEC is mounted on the serving BS.**

**(b) Current MEC is not mounted on the serving BS.**

**Figure 5: How handover occurs if the current MEC caches the requested content.**



**(a) Current MEC is mounted on the serving BS.**

**(b) Current MEC is not mounted on the serving BS.**

**Figure 6: How handover occurs if the current MEC does not cache the requested content.**

where $R[c(k), OS, k]$ is the capacity of backhaul path when transmitting the $k$th chunk.

If $c(t)$ is not mounted on $a(t)$, there will be the delay from the serving MEC to the BS access point. We define this time as $\tau^{c(t)\rightarrow a(t)}(V_m^k)$. Besides, if there is no MEC service, there will be the transmission delay between the serving BS and the origin server which is defined as $\tau^{OS\rightarrow a(t)}$. So the total delay in sending back the requested video chunk to the client is given by:

$$\begin{aligned}\tau_{total}(U_e, V_m^k) =& \tau^{a(k)\rightarrow U_e}(V_m^k) + \alpha_1 \tau^{c(k)\rightarrow a(k)}(V_m^k) \\ &+ \alpha_2 \tau^{s_r\rightarrow c(k)}(V_m^k) + \alpha_3 \tau^{OS\rightarrow c(k)}(V_m^k) \\ &+ \alpha_4 \tau^{OS\rightarrow a(k)}(V_m^k)\end{aligned} \tag{7}$$

In the formula above, $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are factors which indicate whether the corresponding transmission time exists or not, respectively.

## 4 QOE-R2R DESIGN

### 4.1 Cache-based HO Control for MEC selection

Before diving into the details of the MEC selection and its related HO-based trigger mechanism, we first illustrate the handover process in a 5G context in Fig. 4. In such a 5G scenario, when the Access and Mobility Management Function (AMF) receives a session establishment request from the UE, it will communicate with the Session Management Function (SMF) to finish session establishment. With the UPF (mentioned in Section 3.1), SMF is able to manage session context while transmitting data. Meanwhile, AMF is responsible to provide mobility control for UEs. The UE periodically submits measurement reports, e.g., RSRP and RSRQ, to the serving BS for HO decisions. Before the handover is triggered, the serving and target BSs make mutual confirmations first. Then UE synchronizes to the new BS. After the HO execution, the target BS finishes the path switch with AMF and UPF.

To be compatible with the MEC selection, the concrete HO principle is to try best to let the serving MEC be with the serving BS like case **(a)** in Fig. 3. That is, if the serving MEC is mounted on the serving BS and it caches the requested content, the HO should be triggered late. Otherwise, the HO decision should be made in advance in case of missing the MEC selection. To better illustrate the HO principle, Fig. 5 and Fig. 6 show the HO details. In Fig. 5a, the serving BS for $U_e$ is $b_1$. We assume that the MEC in $b_1$ caches $V_1^1$ while that in $b_2$ caches $V_1^2$. If the user requests $V_1^1$, the handover should be triggered as **late** as possible to avoid the unnecessary cost. Apart from the nearly 2$s$ delay from the control plane [2], frequent handovers cause the overhead of context maintenance. Besides, it is likely that the MEC in $b_2$ will be the optimal one for the next chunk(s). If the current chunk $V_1^1$ is not delivered completely by the MEC in $b_1$, it will cause the transit cost of $\tau^{s_r\rightarrow c(t)}(\cdot)$.

However, if the user requests $V_1^2$ that is not cached in $s_1$ as in Fig. 6a, the HO decision should be made as **soon** as possible. Because the MEC selection is triggered by the HO execution. For example, if the target BS is $b_2$, the MEC $s_2$ may be the optimal one due to its cached content. Even though the UE moves backwards and is handed over to $b_3$, the transmission efficiency will not be damaged. Because the response mode is similar to that before HO like the case ($b$) in Fig. 3. Similarly, if the serving MEC caches the requested content $V_1^2$ but is not mounted on the serving BS like Fig. 5b, handover decision should also be made as **soon** as possible. When the target BS is $b_2$, the response mode will change to the case ($a$) in Fig. 3, which is optimal. Even though the target BS is $b_3$, without considering the origin server, the transmission mode will be similar to that before HO like the case ($d$) in Fig. 3.

As Fig. 6b shows, when the UE served by $b_1$ requests $V_1^3$ and its serving MEC is $s_2$, the HO should be triggered as **soon** as possible. Because the serving MEC $s_2$ does not cache $V_1^3$. If the UE is handed

**Algorithm 1** Hit Rate-based MEC Selection Algorithm

```
Inputs: $k, U_e, V_m^k, a(k+1), c(k), S, \mathcal{HO}(U_e), \mathcal{SCK}$,
    $\mathcal{CS}_{s_{a(k+1)}}(V^{k+1\rightarrow k+\mathcal{SCK}}), \mathcal{CS}_{s\in S\backslash s_{a(k+1)}}(V^{k+1\rightarrow k+\mathcal{SCK}})$.
Outputs: $c(k+1 \rightarrow k+\mathcal{SCK})$.
 1: $\mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}}) = -1$;
 2: if $k == 0$ then
 3:     $\mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}}) = s_{a(0)}$;
 4: else
 5:     if $\mathcal{HO}(U_e) == 1$ then
 6:         if $\mathcal{CS}_{s_{a(k+1)}}(V^{k+1\rightarrow k+\mathcal{SCK}}) == \mathcal{SCK}$
            & $c(k) \neq s_{a(k+1)}$ then
 7:             $\mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}}) = s_{a(k+1)}$;
 8:         else if $\mathcal{CS}_{s_{a(k+1)}}(V^{k+1\rightarrow k+\mathcal{SCK}}) == \mathcal{SCK}$
            & $c(k) == s_{a(k+1)}$ then
 9:             $\mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}}) = c(k)$;
10:         else
11:             for $s \in S\backslash s_{a(k+1)}$ do
12:                 if $\mathcal{CS}_s(V^{k+1\rightarrow k+\mathcal{SCK}}) == \mathcal{SCK}$
                    & $c(k) == s$ then
13:                     $\mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}}) = c(k)$
14:                     break;
15:                 else if $\mathcal{CS}_s(V^{k+1\rightarrow k+\mathcal{SCK}}) == \mathcal{SCK}$
                    & $c(k) \neq s$ then
16:                     $\mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}}) = s$
17:                 else
18:                     $\mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}}) = ORIGIN$
19:                 end if
20:             end for
21:         end if
22:     end if
23: end if
24: $c(k+1 \rightarrow k+\mathcal{SCK}) = \mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}})$;
25: return $c(k+1 \rightarrow k+\mathcal{SCK})$
```

over to $b_2$ and the serving MEC is not changed, the response mode will change from case $(e)$ to case $(b)$, as illustrated in Fig. 3. The transmission efficiency will be improved. Even if the serving MEC is changed to another nearby MEC without the cached content, the response mode will be similar to that before HO. If the UE is handed over to $b_3$, the serving MEC may be changed to $s_3$ after HO (i.e., case $(a)$ in Fig. 3). Even if the serving MEC is not changed or just changed to another nearby MEC such as $s_1$, the transmission efficiency will not be damaged.

## 4.2 QoE-aware MEC Selection

The process of MEC selection involves finding the best MEC in geographical terms for the UE.

*4.2.1 Hit Rate-based MEC Selection.* A hit rate-based method which considers the transmission overhead formulated in Section 3 is described in Algorithm. 1 If some MECs cache the same contents, the one mounted on the serving BS is the best choice for MEC selection. This is presented in lines 5-7 and lines 8-9 of the algorithm. The case covered in lines 8-9 is the situation that the originally-serving MEC is the one mounted on the serving BS. If the MEC mounted on the serving BS does not cache the requested content, the originally-serving MEC is a good choice for MEC selection, as presented in lines 12-14. An inferior choice is to pick an alternative MEC to the originally-serving MEC and the one mounted on the serving BS as shown in lines 15-16. The worst case is to fetch the content from the origin server as performed in lines 17-18.

*4.2.2 Delay-based MEC Selection.* Due to mobility, a hit rate-based scheme may bring frequent MEC changes and reduce overall efficiency. Therefore, a delay-based MEC selection scheme is considered based on the transmission mode formulated in Section 3. In general, the transmission duration of fetching from origin server is high due to the shared backhaul path. This happens in cases $(c)$, $(f)$ and $(g)$. If the serving MEC is deployed in the serving BS, transmission efficiency will be the best, like in case $(a)$. This case avoids any delivery overheads between servers or communication overheads between BSs. In case $(b)$ and case $(d)$, the use of MECs attached to nearby BSs within short ranges also avoids backhaul traffic and reduces transmission time. However, the optimal solution remains case $(a)$. If all components of the transmission time are considered, the optimal MEC for $U_e$ to fetch $\mathcal{SCK}$ chunks are as follows.

$$\begin{aligned} &\mathcal{SVR}^*(U_e, V^{k+1\rightarrow k+\mathcal{SCK}}) \\ &= \arg\min \sum_{k'=k+1}^{k+\mathcal{SCK}} [\tau^{a(k+1)\rightarrow U_e}(V^{k'}) \\ &+ \alpha_1 \tau^{s\rightarrow a(k+1)}(V^{k'}) + \alpha_2 \tau^{s\rightarrow c(k)}(V^{k'}) \\ &+ \alpha_3 \tau^{OS\rightarrow s}(V^{k'}) + \alpha_4 \tau^{OS\rightarrow a(k+1)}(V^{k'})]. \\ s.t. \quad &\alpha_2 + \alpha_3 \leq 1 \\ &\text{if } \alpha_4 == 1, \alpha_1, \alpha_2, \alpha_3 == 0 \end{aligned} \tag{8}$$

Due to the sequential and chunk-based characteristics of DASH video streaming, the MEC selection based on the minimal transmission time may not be optimal. This is as multiple nearby-located MECs may result in similar delays and provide similar service. Additionally, DASH viewers may be more interested in video quality, smoothness and stalling. An optimal MEC should provide the shortest latency to ensure best QoE.

*4.2.3 QoE-based MEC Selection.* When assessing user QoE when viewing DASH videos, notable is that QoE is mainly related to video quality, rebuffering and bitrate switching [14, 26]. Although QoE is non-discrete, it is important to be estimated based on video chunks. Unfortunately, the sequential and chunk-based characteristics of DASH streaming do not allow the video chunk to be played until the transmission of the chunk is finished. Here $QoE(t)$ represents the QoE for serving the last chunk, which can be mapped to $QoE(k)$.

$$\begin{aligned} QoE(U_e, t) &= QoE(U_e, V^k) \\ &= \beta_1 \cdot q(V^k) \\ &- \beta_2 \cdot q(V_{min})[\tau_{total}(U_e, V^k) - \mathcal{BUF}(U_e, V^k)]_+ \\ &- \beta_3 \cdot |q(V^k) - q(V^{k-1})| \end{aligned} \tag{9}$$

where $\beta_1, \beta_2, \beta_3$ indicate the weight factors associated with bitrate, rebuffering and smoothness, respectively. In this work, $q(\cdot)$ indicates the video quality measured using the structural similarity (SSIM) index, which reflects better the subjective user experience [6, 20].

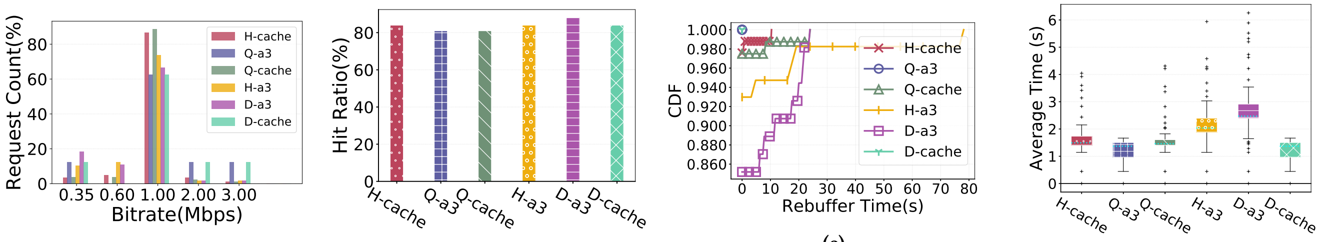


**Figure 7: Bitrate, hit ratio, rebuffering and transmission time for a high-speed user**

For a short-term target, the optimal server should provide maximum QoE possible, as follows.

$$\mathcal{SVR}^*(U_e, V^k) = \arg\max_{\mathcal{SVR}} QoE(U_e, V^k) \tag{10}$$

It can be argued that the goal above is not reasonable enough because MEC change may guide users to request many video chunks in the near future. This MEC server selection strategy is only pertinent to transmission time, which is common with the original delay-based scheme. However, the future bitrate selection will be influenced by the previous bitrate selection. Then the future QoE will also be influenced. So the optimization goal is to maximize QoE for a moment $\tau$ corresponding to $K$ chunks.

$$\begin{aligned}
&\mathcal{SVR}^*(U_e, V^{k+1\to k+K}) \\
&= \arg\max_{\mathcal{SVR}} QoE(U_e, V^{k+1\to k+K}) = \arg\max_{\mathcal{SVR}} \sum_{k'=k+1}^{k+K} QoE(U_e, V^{k'}) \\
&= \sum_{k'=k+1}^{k+K} [\beta_1 \cdot q(V^{k'}) \\
&- \beta_2 \cdot q(V_{min})[\tau_{total}(U_e, V^{k'}) - \mathcal{BUF}(U_e, V^{k'})]_+ \\
&- \beta_3 \cdot |q(V^{k'}) - q(V^{k'-1})|]
\end{aligned} \tag{11}$$

By following this formula, the selected MEC should improve video quality, avoid rebuffering and maintain smoothness for the next several chunks. However, the future UE requested content is unknown. Besides, the MEC selection will further influence users' decisions. Therefore, finding an optimal server is an unsolvable problem. A heuristic approach can be employed instead, involving five discrete bitrate levels and associated to five different chunk sizes. Simple linear regression can be used to predict future throughput which is useful to find the possible requested bitrate of users.

## 4.3 BS-request-based Cache

The MEC server at the mobile network edge can capture radio access network (RAN) conditions through its intrinsic Radio Network Information Service (RNIS) function to achieve better intelligent video adaptation [9]. With assistance of RNIS, if certain MECs are able to cache enough video chunks for some clients, server reselection may not be triggered frequently. Besides, the cache-based handover will also be more compatible with user requests.

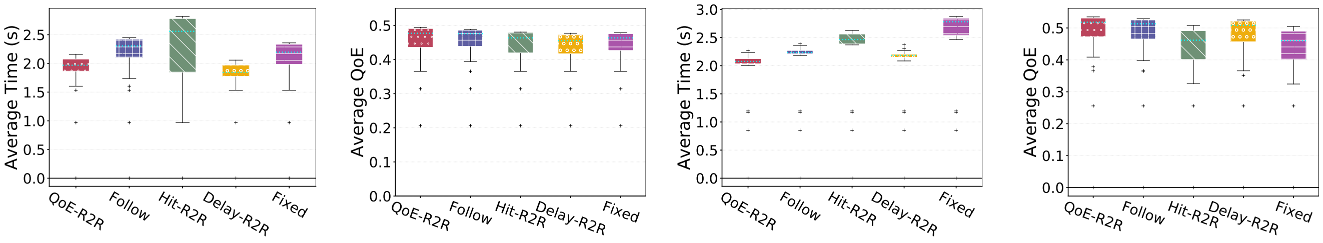
Each BS should maintain a request statistic list for its internal MEC, including the requests sent via itself when working as a serving BS. We define the cache update period as $T_c$. The number of requests for $V_m^k$ is $\mathcal{REQ}_{b_i}(V_m^k)$ while those for $V(o)$ is $\mathcal{REQ}_{b_i}[V(o)]$ that includes all the requests for chunks in $V(o)$. Each MEC also needs to keep a list to collect the requests received by itself. We define the number of requests received by the MEC as $\mathcal{REQ}_{s_i}(V_m^k)$. The number of requests received from $b_i$ and sent to $s_i$ is as $\mathcal{RREQ}_{s_i}(V_m^k)$. If $\mathcal{REQ}_{s_i}(V_m^k) \leq \mathcal{REQ}_{b_i}(V_m^k)$, some requests for $V_m^k$ are served by other MECs. If $\mathcal{REQ}_{s_i}(V_m^k) \geq \mathcal{REQ}_{b_i}(V_m^k)$, it means that $s_i$ provides $V_m^k$ to other BSs or MECs. The requests in $b_i$ and $s_i$ should be listed in the descending order in accordance with $\Delta\mathcal{REQ}_{s_i}(V_m^k) = \mathcal{REQ}_{b_i}(V_m^k) + \mathcal{REQ}_{s_i}(V_m^k) - \mathcal{RREQ}_{s_i}(V_m^k)$. Then the cache in $s_i$ can be periodically updated according to $\Delta\mathcal{REQ}_{s_i}(\cdot)$ to maximize the cache hit possibility of responding UEs with the MEC cache mounted on its serving BS.

# 5 SIMULATION-BASED EVALUATION

The experiments are performed in Network Simulator NS-3 [18]. The network topology is as illustrated in Fig. 2 and involves micro BSs deployed on the straight road. There are 50 users served by 5 MEC servers that are all equipped with 5GB storage space. These MECs are deployed on the BSs individually and share the same origin server. The UE and BS transmission powers are 10dBm and 46dBm, respectively. The path loss model employed is the MmWavePropagationLossModel [17].

## 5.1 Single User Response Pattern

The single-user experiment tests the response pattern to different requests when different flavours of the proposed MEC selection method, based on hit rate (H), delay (D) and QoE (Q), are employed. These methods are combined with the traditional signal-based A3 HO algorithm [18] and proposed cache-based HO control, resulting in H-a3, H-cache, D-a3, D-cache , Q-a3 and Q-cache, respectively. The request pattern of bitrate representations is relatively stable because of the setting of network capacity as in Figure. 7a. Except from the schemes **Q-cache** and **Q-a3**, the other schemes obtain a higher hit ratio as shown in Figure. 7b. However, **Q-cache** and **Q-a3** are associated with less rebuffering time as presented in Fig. 7c. Besides, as Fig. 7c shows, the schemes with the proposed cache-based HO control outperform the ones based on the traditional A3-based HO. In Fig. 7d can also be noted how the QoE-based schemes **Q-a3** and **Q-cache** generally perform better in terms of average transmission time than the others. In conclusion, it can be said that the MEC selection is improved by the addition of the cache-based HO control.



(a)

(b)

(c)

(d)

Figure 8: The comparison of transmission time and QoE with different cache replacement

Table 2: The comparison of different schemes

| Cache Policy | Metrics | Hit-R2R | Delay-R2R | QoE-R2R | Fixed | Follow |
|---|---|---|---|---|---|---|
| **LRU** | Bitrate | 0.731 | 0.902 | **0.885** | 0.834 | 0.817 |
| | Rebuffer | 0.201 | 0.145 | **0.100** | 0.205 | 0.105 |
| | Variation | 0.211 | 0.249 | **0.161** | 0.234 | 0.183 |
| | Hit Ratio | 0.526 | 0.478 | **0.478** | 0.478 | 0.591 |
| | Delay | 2.818 | 2.057 | **2.160** | 2.358 | 2.447 |
| | QoE | 0.480 | 0.477 | **0.494** | 0.478 | 0.488 |
| **BS-based** | Bitrate | 1.009 | 1.024 | **1.062** | 0.997 | 1.049 |
| | Rebuffer | 0.192 | 0.123 | **0.082** | 0.198 | 0.105 |
| | Variation | 0.196 | 0.224 | **0.222** | 0.239 | 0.225 |
| | Hit Ratio | 0.552 | 0.574 | **0.587** | 0.527 | 0.572 |
| | Delay | 2.628 | 2.201 | **2.078** | 2.872 | 2.233 |
| | QoE | 0.508 | 0.525 | **0.535** | 0.504 | 0.529 |

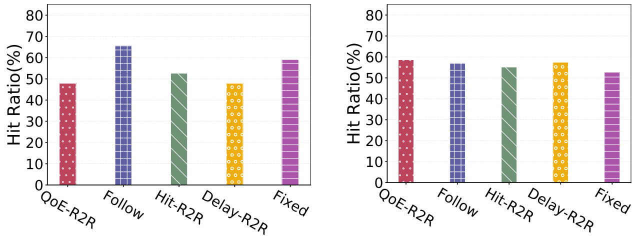


Figure 9: The comparison of hit ratio

## 5.2 Multiple User Response Results

The multi-user experiment's goal is to evaluate the proposed QoE-aware MEC selection method in comparison with alternative approaches. The following five MEC selection methods are tested, methods which differ in the manner they interact with the cache-based HO control: **Hit-R2R** (employs the hit rate-based MEC selection), **Delay-R2R** (uses the delay-based MEC selection), **QoE-R2R** (is based on the QoE-based MEC selection), **follow** (always chooses the MEC mounted on the serving BS) and **fixed** (never changes the serving MEC). Here we regard the former three schemes as R2R-related. The experiment is divided into two groups, employing different cache policies: the first group uses the Least Recently Used (LRU) policy and the other group - the BS-request Motivated Cache Replacement policy.

The concrete statistical testing results in terms of hit ratio are listed in Table. 2. It can be concluded that the user request pattern will change because of the BS-request Motivated Cache Replacement, with at least 16.3% bitrate improved. There is also less rebuffering for the schemes with BS-request Motivated Cache Replacement. The scheme **QoE-R2R** reduces about 22% rebuffering time on average. This also explains why the QoE in the second group is relatively better. From Fig. 9, it can be concluded that the BS-request motivated cache replacement helps R2R-related schemes attain a high hit ratio. Because relying on BS request statistics, the cached contents of a certain MEC will follow the request mode of the users covered by the BS. But this cache replacement may not work with the other two simple MEC selection methods, i.e., **Fixed** and **Follow**. Fig.8a and Fig.8c show the transmission time for all the video chunks requested by users. The average transmission time of **Hit-R2R** is longer, 2.818 and 2.628 seconds in the two groups of experiment, respectively. This is as the solution focuses on maximizing hit ratio instead of the response speed. Although the proposed scheme **QoE-R2R** is associated with a slightly higher delay than the delay optimisation solution **Delay-R2R**, it achieves the best QoE in both groups of the experiment, as shown by Fig. 8b and Fig. 8d.

# 6 CONCLUSIONS AND FUTURE WORKS

This paper formulates the problem of response in the context of a MEC-based edge framework for DASH-based video delivery. We design a new QoE-aware MEC selection scheme (QoE-R2R) which considers user request-motivated HO decision and edge cache states. Simulations show that the proposed QoE-R2R scheme outperforms other strategies by achieving reduced transmission time and improved QoE. Future work will focus on employing learning-based methods in a 5G scenario to improve MEC-based DASH video streaming.

# 7 ACKNOWLEDGEMENTS

This work is supported by the Key-Area Research and Development Program of Guangdong Province under grant No. 2020B0101130006, National Natural Science Foundation of China under grant No. 61972189, Guangdong Province Key Area R&D Program under grant No. 2018B010113001, the Shenzhen Key Lab of Software Defined Networking under grant No. ZDSYS20140509172959989 and Science Foundation Ireland (SFI) via grant number 12/RC/2289_P2.

# REFERENCES

[1] ETSI TS 123 501-2021. 2021. 5G; System architecture for the 5G System (5GS) (V16.7.0; 3GPP TS 23.501 version 16.7.0 Release 16). (2021).

[2] Mukhtiar Ahmad, Syed Usman Jafri, Azam Ikram, Wasiq Noor Ahmad Qasmi, Muhammad Ali Nawazish, Zartash Afzal Uzmi, and Zafar Ayyub Qazi. 2020. A Low Latency and Consistent Cellular Control Plane. In *ACM SIGCOMM*.

[3] Ahmed Alkhateeb, Iz Beltagy, and Sam Alex. 2018. Machine Learning for Reliable mmWave Systems: Blockage Prediction and Proactive Handoff. In *IEEE GlobalSIP*.

[4] Niklas Carlsson and Derek L Eager. 2010. Server Selection in Large-Scale Video-on-Demand Systems. *ACM TOMM* 6, 1 (2010), 1–26.

[5] Jiayin Chen, Huaqing Wu, Peng Yang, Feng Lyu, and Xuemin Shen. 2020. Cooperative Edge Caching With Location-Based and Popular Contents for Vehicular Networks. *IEEE Transactions on Vehicular Technology* 69, 9 (2020), 10291–10305.

[6] Federico Chiariotti, Stefano D'Aronco, Laura Toni, and Pascal Frossard. 2016. Online Learning Adaptation Strategy for DASH Clients. In *ACM MMSys*.

[7] Igor Colin, Albert Thomas, and Moez Draief. 2018. Parallel Contextual Bandits in Wireless Handover Optimization. In *IEEE ICDMW*.

[8] Chang Ge, Ning Wang, Gerry Foster, and Mick Wilson. 2017. Toward QoE-Assured 4K Video-on-Demand Delivery through Mobile Edge Virtualization with Adaptive Prefetching. *IEEE TMM* 19, 10 (2017), 2222–2237.

[9] Chang Ge, Ning Wang, Severin Skillman, Gerry Foster, and Yue Cao. 2016. QoE-driven DASH Video Caching and Adaptation at 5G Mobile Edge. In *ACM ICN*.

[10] Tai Manh Ho and Kim-Khoa Nguyen. 2020. Joint Server Selection, Cooperative Offloading and Handover in Multi-access Edge Computing Wireless Network: A Deep Reinforcement Learning Approach. *IEEE Transactions on Mobile Computing* (2020).

[11] Sami Kekki, Walter Featherstone, Yonggang Fang, Pekka Kuure, Alice Li, Anurag Ranjan, Debashish Purkayastha, Feng Jiangping, Danny Frydman, Gianluca Verin, et al. 2018. MEC in 5G networks. *ETSI White Paper* 28 (2018), 1–28.

[12] Alexe E Leu and Brian L Mark. 2003. An Efficient Timer-based Hard Handoff Algorithm for Cellular Networks. In *IEEE WCNC*.

[13] Yuanjie Li, Qianru Li, Zhehui Zhang, Ghufran Baig, Lili Qiu, and Songwu Lu. 2020. Beyond 5G: Reliable Extreme Mobility Management. In *ACM SIGCOMM*.

[14] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *ACM SIGCOMM*.

[15] J Meredith. 2016. Study on Channel Model for Frequency Spectrum above 6 GHz. *3GPP TR 38.900, Jun, Tech. Rep.* (2016).

[16] Brian Meskill, Alan Davy, and Brendan Jennings. 2011. Server Selection and Admission Control for IP-based Video on Demand Using Available Bandwidth Estimation. In *IEEE LCN*.

[17] Marco Mezzavilla, Menglei Zhang, Michele Polese, Russell Ford, Sourjya Dutta, Sundeep Rangan, and Michele Zorzi. 2018. End-to-End Simulation of 5G mmWave Networks. *IEEE Communications Surveys & Tutorials* 20, 3 (2018), 2237–2263.

[18] George F Riley and Thomas R Henderson. 2010. The NS-3 Network Simulator. In *Modeling and tools for network simulation*. Springer, 15–34.

[19] Indranil Sen and David W Matolak. 2008. Vehicle–Vehicle Channel Models for the 5-GHz Band. *IEEE TITS* 9, 2 (2008), 235–245.

[20] Wanxin Shi, Chao Wang, Yong Jiang, Qing Li, Gengbiao Shen, and Gabriel-Miro Muntean. 2021. CoLEAP: Cooperative Learning-Based Edge Scheme with Caching and Prefetching for DASH Video Delivery. *IEEE TMM* (2021).

[21] Wu Shih-Jung and KC Lo Steven. 2011. Handover Scheme in LTE-based Networks with Hybrid Access Mode. *JCIT* 6, 7 (2011), 68–78.

[22] Graphene Market Size. 2020. Video Streaming Market Size, Share & Trends Analysis Report By Streaming Type, By Solution, By Platform, By Service, By Revenue Model, By Deployment Type, By User And Segment Forecasts, 2020 - 2027. *Grand View Research* (2020).

[23] Kyuho Son, Eunsung Oh, and Bhaskar Krishnamachari. 2011. Energy-Aware Hierarchical Cell Configuration: from Deployment to Operation. In *IEEE INFOCOM Workshops*.

[24] Zhi Wang, Lifeng Sun, Chuan Wu, Wenwu Zhu, and Shiqiang Yang. 2014. Joint Online Transcoding and Geo-Distributed Delivery for Dynamic Adaptive Streaming. In *IEEE INFOCOM*.

[25] Dongzhu Xu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu, Congkai An, Yiming Shi, Liang Liu, and Huadong Ma. 2020. Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption. In *ACM SIGCOMM*.

[26] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. 2015. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. In *ACM SIGCOMM*.

[27] Aoyang Zhang, Qing Li, Ying Chen, Xiaoteng Ma, Longhao Zou, Yong Jiang, Zhimin Xu, and Gabriel-Miro Muntean. 2021. Video Super-Resolution and Caching– An Edge-Assisted Adaptive Video Streaming Solution. *IEEE Transactions on Broadcasting* (2021).

[28] Ke Zhang, Yongxu Zhu, Supeng Leng, Yejun He, Sabita Maharjan, and Yan Zhang. 2019. Deep Learning Empowered Task Offloading for Mobile Edge Computing in Urban Informatics. *IEEE IoT-J* 6, 5 (2019), 7635–7647.

[29] Qian Zhang, Zhe Xiang, Wenwu Zhu, and Lixin Gao. 2004. Cost-based Cache Replacement and Server Selection for Multimedia Proxy Across Wireless Internet. *IEEE TMM* 6, 4 (2004), 587–598.

[30] Wenming Zhang, Yiwen Zhang, Qilin Wu, and Kai Peng. 2019. Mobility-Enabled Edge Server Selection for Multi-User Composite Services. *Future Internet* 11, 9 (2019), 184.

[31] Yi-wen Zhang, Wen-ming Zhang, Kai Peng, Deng-cheng Yan, and Qi-lin Wu. 2021. A Novel Edge Server Selection Method based on Combined Genetic Algorithm and Simulated Annealing Algorithm. *Automatika* 62, 1 (2021), 32–43.