Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Joint uplink and downlink resource allocation for low-latency mobile virtual reality delivery in fog radio access networks*

Tian DANG, Chenxi LIU‡, Xiqing LIU, Shi YAN

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China

E-mail: tiandang@bupt.edu.cn; chenxi.liu@bupt.edu.cn; liuxiqing@bupt.edu.cn; yanshi01@bupt.edu.cn

Received June 30, 2021; Revision accepted Sept. 14, 2021; Crosschecked Nov. 8, 2021

Abstract: Fog radio access networks (F-RANs), in which the fog access points are equipped with communication, caching, and computing functionalities, have been anticipated as a promising architecture for enabling virtual reality (VR) applications in wireless networks. Although extensive research efforts have been devoted to designing efficient resource allocation strategies for realizing successful mobile VR delivery in downlink, the equally important resource allocation problem of mobile VR delivery in uplink has so far drawn little attention. In this work, we investigate a mobile VR F-RAN delivery framework, where both the uplink and downlink transmissions are considered. We first characterize the round-trip latency of the system, which reveals its dependence on the communication, caching, and computation resource allocations. Based on this information, we propose a simple yet efficient algorithm to minimize the round-trip latency, while satisfying the practical constraints on caching, computation capability, and transmission capacity in the uplink and downlink. Numerical results show that our proposed algorithm can effectively reduce the round-trip latency compared with various baselines, and the impacts of communication, caching, and computing resources on latency performance are illustrated.

Key words: Virtual reality delivery; Fog radio access network (F-RAN); Round-trip latency; Resource allocation
https://doi.org/10.1631/FITEE.2100308 **CLC number:** TN914

1 Introduction

Fifth generation and beyond (5G and beyond) wireless networks have been an indispensable part of our daily life, enabling many novel and "intellicise" applications. As stated in Zhang P et al. (2022), "intellicise" means that 5G and beyond wireless networks can provide these novel applications with endogenous intelligence and primitive conciseness. Among these applications, virtual reality (VR) is regarded as a transformative application, since it can provide immersive experiences for users, and therefore receives significant attention from both academia and industry (Bastug et al., 2017; Hu et al., 2020). However, VR applications require real-time broadband communications. As anticipated by Cisco, VR traffic will be 254.4 petabytes per month by 2022 (Cisco System, 2019), which creates many challenges for current and future wireless networks, such as high data rate and low latency. Against this backdrop, increasing research attention has been

‡ Corresponding author
* Project supported by the Beijing Natural Science Foundation, China (No. JQ18016), the National Key R&D Program of China (No. 2020YFB1806703), the National Natural Science Foundation of China (Nos. 62001047, 61901315, and 61901044), the National Program for Special Support of Eminent Professionals, China, the Young Elite Scientist Sponsorship Program by China Institute of Communications, and the Project of China Railway Corporation (No. P2020G004)
ORCID: Tian DANG, https://orcid.org/0000-0002-9589-6967; Chenxi LIU, https://orcid.org/0000-0002-9134-1235

paid to delivering immersive VR experiences in wireless networks.

One of the main research problems in mobile VR delivery is the resource allocation problem. Specifically, in scenarios where fog and mobile edge computing (MEC) capabilities are deployed to render VR videos, Yoshihara and Fujita (2019) offloaded the rendering of the background view from the cloud server to the fog nodes and achieved a low-latency VR game. In Zhang Y et al. (2019), the rendering modules were dynamically placed at the distributed MEC servers to reduce the operational cost and communication delay. In Huang et al. (2018), VR videos were designed to be pre-cached at the user destination to reduce the transmission latency and optimize the access probability. In Dai et al. (2020), a transmission latency minimization problem was studied by hierarchically and adaptively caching VR videos at the cloud server and the remote radio heads (RRHs). Moreover, in Du et al. (2020), the offloaded rendering and transmit power control were jointly optimized to minimize energy consumption with a deep reinforced learning approach. An MEC-based VR delivery framework was presented in Sun et al. (2019), in which the caching and computation resources were jointly optimized and the tradeoffs among communications, caching, and computing were revealed. In Zhou et al. (2021), a joint bandwidth, caching, and computation resource allocation was designed to minimize the maximum communication and computing latency.

Recently, fog radio access networks (F-RANs) have been regarded as a promising network architecture for mobile VR delivery, because they deploy caching and computing functionalities near user equipment, thereby greatly alleviating the fronthaul traffic load and reducing the latency (Peng et al., 2016; You et al., 2019). Chiu et al. (2019) examined how to formulate fog node groups according to various communication and computation resources to achieve low-latency VR service. In Park SH et al. (2016), cloud and edge processing were jointly designed to maximize the delivery rate with enhanced RRHs equipped with local caching and baseband processing capabilities. Liu et al. (2018) developed a strategy involving a joint offloading decision, computing, power, and bandwidth allocation that minimized latency and energy consumption. In our previous work (Dang and Peng, 2019), we considered the impact of downlink transmissions only on VR delivery in F-RANs and formulated a tolerant latency maximization problem. We proposed a Lagrangian dual decomposition based approach that jointly optimizes the content placement and task offloading to solve the formulated problem, and illustrated the impact of caching and computing resources on the average tolerant delay with numerical results.

The aforementioned works focused mainly on the resource allocation problem in downlink transmissions. However, the limited resources in uplink transmissions can have a significant impact on the performance of mobile VR delivery. In Park J et al. (2018), uploading delay in the uplink was regarded as the bottleneck in the downlink delay; the proposed uplink spectrum allocation yielded up to 25.1% average end-to-end latency reduction compared to the equal uplink allocation. In Chen et al. (2020), tracking information was transmitted in the uplink over a sub-6 GHz frequency band to provide reliable communication with limited bandwidth, emphasizing that the uplink transmission latency cannot be ignored. Given the significance of the uplink transmissions in mobile VR delivery, it is therefore essential to address the impact of both the uplink and downlink in designing the corresponding resource allocation strategies.

Motivated by the above observations, in this work, we investigate how to effectively and efficiently deliver mobile VR services in F-RANs. Both the uplink and downlink transmissions are considered. Specifically, in the uplink, the tracking information is collected from VR users. In the downlink, VR videos are delivered to the VR users, and the processing tasks are offloaded to either the fog access points (F-APs) or the VR users. In this work, a round-trip latency minimization problem is formulated to allocate the caching, computing, and communication resources. Our contributions are as follows:

1. We characterize the round-trip latency of mobile VR delivery in F-RANs, revealing its dependence on communication resource allocation in both uplink and downlink, as well as caching and computation resource allocations between F-APs and VR users.

2. We formulate a round-trip latency minimization problem, subject to practical constraints on the uplink and downlink transmission capacities, the F-AP caching size, and the computation capabilities

of the F-APs and VR users. To solve this NP-hard problem, we decompose it into two subproblems, the mode selection problem and the joint communication and computation resource allocation problem, which are tackled with the branch-and-bound method and convex optimization, respectively. Finally, a simple yet effective algorithm is proposed to iteratively solve these subproblems.

3. We examine the impact of key system parameters on round-trip latency. Compared to various baselines, our proposed algorithm can effectively improve mobile VR delivery performance, especially when the downlink transmission capacity is relatively large.

Unless specified, the notations used throughout this paper are as summarized in Table 1.

2 System model

As illustrated in Fig. 1, the considered F-RAN-based mobile VR delivery system consists of one cloud server, one high-power node (HPN), M F-APs, N VR users, and multiple RRHs. The cloud server includes a baseband unit (BBU) pool. The F-APs and VR users are denoted by $\mathcal{M}=\{1,2,\ldots,M\}$ and $\mathcal{N}=\{1,2,\ldots,N\}$, respectively. In this system, the cloud server extracts 360° VR videos into monocular videos (MVs) and stereoscopic videos (SVs). The F-APs are equipped with certain caching and computation capabilities and can judiciously cache VR videos and process computing tasks. The VR users are equipped with limited computational capabilities, and access either the RRHs or the F-APs. The RRHs communicate with the VR users through centralized cooperative transmission. The resource management decision is transmitted through the HPN in the control link. In the uplink, the tracking information is uploaded. In the downlink, the requested VR videos are delivered to the VR users.

We now describe the VR delivery process, including the following:

1. Tracking. The VR users upload the tracking information to the associated access points, i.e., F-APs or RRHs, through the uplink.

2. Extraction. Using the tracking information, the F-APs and the cloud server extract the VR videos (i.e., SVs or MVs).

3. Projection. The F-APs and VR users project the MVs into the SVs when the computing tasks are offloaded.

Table 1 Notations used in this paper

Notation	Description
M, N	Numbers of F-APs and VR users, respectively
K	Number of VR services
s_k	Tracking data size of VR service k
$v_{n,k}$	Request indicator of VR user n for VR service k
D_k^{S}, D_k^{M}	Sizes of SV k and MV k, respectively
α_k	Ratio of D_k^{S} to D_k^{M}
P_k	Request probability of VR service k
C_m^{A}	Cache size of F-AP m
w_k	Number of cycles per bit for VR service k
E_n^{V}, E_m^{A}	Maximum energy consumptions of VR user n and F-AP m, respectively
$\varepsilon_n^{\mathrm{V}}$, $\varepsilon_m^{\mathrm{A}}$	Power efficiencies of VR user n and F-AP m, respectively
L_n^{V}, L_m^{A}	Processing frequencies of VR user n and F-AP m, respectively
R^{UF}, R_m^{Umax}	Uplink transmission capacities of the fronthaul and F-AP m, respectively
R^{F}, $R_m^{\max}$	Downlink transmission capacities of the fronthaul and F-AP m, respectively
$a_{m,n}$	User association between F-AP m and VR user n
$c_{m,k}^{\mathrm{AM}}, c_{m,k}^{\mathrm{AS}}$	Content placements of MV k and SV k at F-AP m, respectively
$d_{m,n}^{\mathrm{A}}, d_n^{\mathrm{V}}$	Task offloading decisions of VR user n to F-AP m and to itself, respectively
$l_{m,n}^{\mathrm{A}}$	Processing frequency allocation from F-AP m to VR user n
b_n^{U}	Uplink data rate of VR user n
$b_{m,n}^{\mathrm{A}}, b_n^{\mathrm{R}}$	Downlink data rates of VR user n from F-AP m and the RRHs, respectively
$q_{n,i}$	Service mode selection of VR user n in service mode i
τ_n^{UL}	Uplink transmission latency of VR user n
τ_n^{DL}	Downlink transmission latency of VR user n
τ_n^{CP}	Processing latency of VR user n
$\tau_{n,i}$	Round-trip latency of VR user n in service mode i

4. Rendering. The VR users render the SVs into 360° videos and present them. Note that the projection exists if and only if the computing tasks are offloaded to the F-APs or the VR users.

The service modes considered in this work are summarized into five categories (Fig. 2). Note that there are $3M+2$ service modes, consisting of two modes through the RRHs and three modes through each F-AP. Let $q_{n,i} \in \{0,1\}$ ($n \in \mathcal{N}$, $i \in \{1,2,\ldots,3M+2\}$) denote the service mode selection variable of VR user n, in which $q_{n,i}=1$ indicates that the VR video is delivered to VR user n in service mode i, and $q_{n,i}=0$ otherwise. The service modes in Fig. 2 are as follows:

1. $q_{n,1}=1$. In service mode 1, VR user n accesses the RRHs and the MV is delivered to the VR user from the cloud server. Meanwhile, the computing task is processed at VR user n.

2. $q_{n,2}=1$. In service mode 2, VR user n accesses the RRHs and the SV is delivered to the VR user from the cloud server.

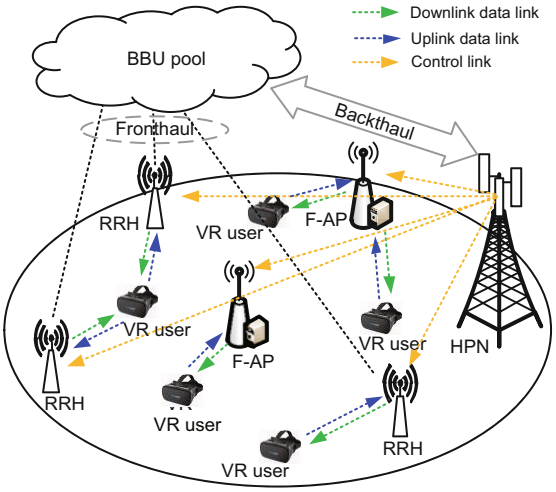

Fig. 1 VR delivery in F-RANs

BBU: baseband unit; RRH: remote radio head; VR: virtual reality; HPN: high-power node; F-AP: fog access point

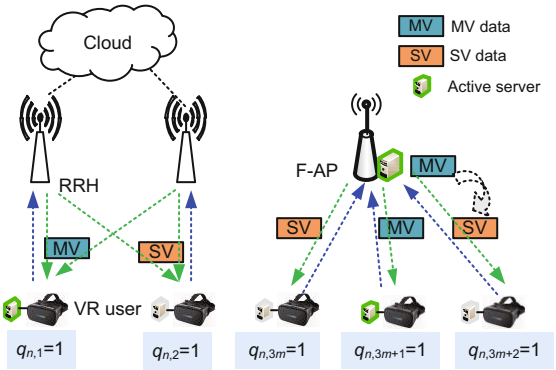

Fig. 2 Service modes for mobile VR delivery

3. $q_{n,3m}=1, m \in \mathcal{M}$. In service mode 3, VR user n accesses F-AP m and the SV is delivered to the VR user from F-AP m. The SV must be pre-cached at F-AP m.

4. $q_{n,3m+2}=1, m \in \mathcal{M}$. In service mode $3m+2$, VR user n accesses F-AP m with the requested MV cached. The computing task is offloaded to F-AP m to project the MV to SV. Then, SV is transmitted to VR user n.

2.1 Cache and computation model

In our system, K VR services exist and they are indexed by $\mathcal{K}=\{1,2,\ldots,K\}$. Let $v_{n,k} \in \{0,1\}$ denote the request indicator of VR user n for VR service k. In particular, $v_{n,k}=1$ denotes that VR user n requests the k^{th} viewpoint, and $v_{n,k}=0$ otherwise. Accordingly, the subset of the VR users that request VR service k can be represented by $\mathcal{N}_k=\{1,2,\ldots,N_k\}$, which satisfies $v_{n,k}=1\ \forall n \in \mathcal{N}_k$. Suppose that only one VR service is requested by each VR user. In that case, $\mathcal{N}=\cup_{k\in\mathcal{K}}\mathcal{N}_k$ and $N=\sum_{k\in\mathcal{K}}N_k$.

For each viewpoint k, there is one SV k and one MV k. Let D_k^{S} and D_k^{M}, in bits, denote the data size of the SV and MV, respectively. Then, the ratio of D_k^{S} to D_k^{M} for viewpoint k is given by $\alpha_k=D_k^{\mathrm{S}}/D_k^{\mathrm{M}}$, which satisfies $\alpha_k \geq 2$, because the visual images need to be different for two eyes to create a stereoscopic vision. The request probability of each viewpoint k is denoted by P_k, which satisfies $\sum_{k\in\mathcal{K}}P_k=1$. The volume of the tracking information for each viewpoint k is denoted by s_k in bits.

The cache size of F-AP m is C_m^{A} in Gbit (Gb for short). F-APs and VR users use the virtualization technology for adaptive allocation of computation resources, e.g., central processing unit (CPU) frequency. Specifically, the CPU-cycle frequency capacities of F-AP m and VR user n are L_m^{A} and L_n^{V} in GHz, respectively. The computing energy stored at F-AP m and VR user n are E_m^{A} and E_n^{V} in J, respectively. For each VR user, all the L_n^{V} are used when the VR user processes a computation task. For each F-AP, let $l_{m,n}^{\mathrm{A}}$ in GHz denote the CPU-cycle frequency allocated to VR user n. The power efficiencies of the CPU server at the F-APs and VR user are $\varepsilon_m^{\mathrm{A}}$ and $\varepsilon_n^{\mathrm{V}}$, respectively. Moreover, for each viewpoint k, the number of computation cycles required

for processing the projection of one bit input is denoted by w_k, in cycle/bit.

Denote $a_{m,n} \in \{0,1\}$ $(m \in \{0\} \cup \mathcal{M})$ as the user association variable, in which $a_{0,n} = 1$ denotes that VR user n accesses the RRHs, and $a_{m,n} = 1$ $(m \in \mathcal{M})$ denotes that VR user n accesses F-AP m. Denote $c^{\mathrm{AM}}_{m,k},\ c^{\mathrm{AS}}_{m,k} \in \{0,1\}$ as the caching variables at the F-APs. In particular, $c^{\mathrm{AM}}_{m,k}$ $(c^{\mathrm{AS}}_{m,k}) = 1$ indicates that MV(SV) k is cached at F-AP m; otherwise, $c^{\mathrm{AM}}_{m,k}$ $(c^{\mathrm{AS}}_{m,k}) = 0$. Then, the cache size constraint at the F-APs is expressed as

$$\sum_{k=1}^{K} c^{\mathrm{AM}}_{m,k} D^{\mathrm{M}}_k + c^{\mathrm{AS}}_{m,k} D^{\mathrm{S}}_k \le C^{\mathrm{A}}_m. \tag{1}$$

Denote $d^{\mathrm{A}}_{m,n},\ d^{\mathrm{V}}_n \in \{0,1\}$ as the offloading variables of the projection tasks at the F-APs and VR users, respectively. Particularly, when the computation task requested by VR user n is processed at F-AP m (VR user n), $d^{\mathrm{A}}_{m,n}$ $(d^{\mathrm{V}}_n) = 1$; otherwise, $d^{\mathrm{A}}_{m,n}$ $(d^{\mathrm{V}}_n) = 0$. Then, the constraint on the CPU frequency capacity at the F-APs is given by

$$\sum_{n=1}^{N} l^{\mathrm{A}}_{m,n} d^{\mathrm{A}}_{m,n} \le L^{\mathrm{A}}_m. \tag{2}$$

In addition, the computing energies consumed by the projection at the F-APs and VR users are calculated as $\varepsilon^{\mathrm{A}}_m (l^{\mathrm{A}}_{m,n})^2$ and $\varepsilon^{\mathrm{V}}_n (L^{\mathrm{V}}_n)^2$ per cycle, respectively. The energy consumption constraints can be expressed as

$$\varepsilon^{\mathrm{A}}_m \sum_{k=1}^{K} w_k D^{\mathrm{M}}_k \sum_{n \in \mathcal{N}_k} \left(l^{\mathrm{A}}_{m,n}\right)^2 d^{\mathrm{A}}_{m,n} \le E^{\mathrm{A}}_m, \tag{3}$$

$$\varepsilon^{\mathrm{V}}_n w_k D^{\mathrm{M}}_k \left(L^{\mathrm{V}}_n\right)^2 d^{\mathrm{V}}_n \le E^{\mathrm{V}}_n,\ \forall n \in \mathcal{N}_k. \tag{4}$$

Accordingly, the service mode selection of VR user n can be represented by different user association, caching, and offloading decisions, as summarized in Table 2. Particularly, $q_{n,1} = 1$ indicates that VR user n accesses the RRHs to acquire the MV and that the computing task is offloaded to the user; as such, $a_{0,n} = 1$ and $d^{\mathrm{V}}_n = 1$. When $q_{n,2} = 1$, VR user n accesses the RRHs to acquire the SV, and in addition, $a_{0,n} = 1$. $q_{n,3m} = 1$ $(m \in \mathcal{M})$ implies that VR user n accesses F-AP m to acquire the SV, so $a_{m,n} = 1$ and $c^{\mathrm{AS}}_{m,k} = 1$. Similarly, $q_{n,3m+1} = 1$ $(m \in \mathcal{M})$ indicates that VR user n accesses F-AP m to acquire the MV and that the computing task is offloaded to the user, so $a_{m,n} = 1$, $c^{\mathrm{AM}}_{m,k} = 1$, and $d^{\mathrm{V}}_n = 1$. $q_{n,3m+2} = 1$ $(m \in \mathcal{M})$ indicates that VR user n accesses F-AP m to acquire the MV and that the computing task is offloaded to the F-AP, so $a_{m,n} = 1$, $c^{\mathrm{AM}}_{m,k} = 1$, and $d^{\mathrm{A}}_{m,n} = 1$. For VR user n, the user association and offloading decisions must satisfy $\sum_{m=0}^{M} a_{m,n} = 1$ and $\sum_{m=1}^{M} d^{\mathrm{A}}_{m,n} + d^{\mathrm{V}}_n = 1$.

Table 2 Summary of service modes for VR users

Service mode	Association, caching, offloading
$q_{n,1} = 1$	$a_{0,n} = 1,\ d^{\mathrm{V}}_n = 1$
$q_{n,2} = 1$	$a_{0,n} = 1$
$q_{n,3m} = 1$	$a_{m,n} = 1,\ c^{\mathrm{AS}}_{m,k} = 1$
$q_{n,3m+1} = 1$	$a_{m,n} = 1,\ c^{\mathrm{AM}}_{m,k} = 1,\ d^{\mathrm{V}}_n = 1$
$q_{n,3m+2} = 1$	$a_{m,n} = 1,\ c^{\mathrm{AM}}_{m,k} = 1,\ d^{\mathrm{A}}_{m,n} = 1$

2.2 Transmission model

In the mobile VR delivery system, both the uplink and downlink data transmissions are considered. Specifically, the uplink transmissions are used to upload the tracking information, while the downlink transmissions are used to deliver the VR videos. Denote the volume of the tracking information at VR user $n \in \mathcal{N}_k$ as $S_n = s_k$. In the uplink transmission, the data rate to upload the tracking information from VR user n to the F-APs or RRHs is denoted by b^{U}_n in Gb/s. In the downlink transmission, the data rate for the VR video delivery from F-AP m to VR user n is given by $b^{\mathrm{A}}_{m,n}$ in Gb/s, and the data rate from the RRHs to VR user n is b^{R}_n in Gb/s. The data volume of the downlink VR video for VR user n is denoted by S^{DL}_n. Then, the transmission capacity constraints on both the uplink and downlink are given by

$$\sum_{n=1}^{N} a_{0,n} b^{\mathrm{U}}_n \le R^{\mathrm{UF}}, \tag{5}$$

$$\sum_{n=1}^{N} a_{m,n} b^{\mathrm{U}}_n \le R^{\mathrm{Umax}}_m, \tag{6}$$

$$\sum_{n=1}^{N} a_{m,n} b^{\mathrm{A}}_{m,n} \le R^{\mathrm{max}}_m, \tag{7}$$

$$\sum_{n=1}^{N} a_{0,n} b^{\mathrm{R}}_n \le R^{\mathrm{F}}, \tag{8}$$

where R^{UF} denotes the maximum uplink transmission capacity at the RRHs, R^{Umax}_m and R^{max}_m denote

the maximum uplink and downlink transmission capacities at F-AP m respectively, and R^{F} denotes the maximum downlink fronthaul capacity.

2.3 Round-trip latency

To evaluate the performance of mobile VR delivery, we adopt round-trip latency as the metric, given by

$$\tau=\sum_{n=1}^{N}(\tau_n^{\mathrm{UL}}+\tau_n^{\mathrm{DL}}+\tau_n^{\mathrm{CP}}), \tag{9}$$

where τ_n^{UL}, τ_n^{DL}, and τ_n^{CP} denote the uplink transmission, the downlink transmission, and the processing latency of VR user n, respectively. Specifically, the computation capacity of the cloud server is considered powerful enough and the processing latency in the cloud server is ignored in the round-trip latency. Therefore, when $q_{n,2}=1$ or $q_{n,3m}=1$, the processing latency is reduced to zero. For VR user $n\in\mathcal{N}_k$, τ_n^{UL}, τ_n^{DL}, and τ_n^{CP} are given by

$$\tau_n^{\mathrm{UL}}=\frac{S_n}{b_n^{\mathrm{U}}}, \tag{10}$$

$$\tau_n^{\mathrm{DL}}=\begin{cases}\dfrac{S_n^{\mathrm{DL}}}{b_n^{\mathrm{R}}}, & i=1,2,\\ \dfrac{S_n^{\mathrm{DL}}}{b_{m,n}^{\mathrm{A}}}, & i=3m,3m+1,3m+2,\end{cases} \tag{11}$$

$$\tau_n^{\mathrm{CP}}=\begin{cases}\dfrac{C_n}{L_n^{\mathrm{V}}}, & i=1,3m+1,\\ \dfrac{C_n}{l_{m,n}^{\mathrm{A}}}, & i=3m+2,\\ 0, & i=2,3m,\end{cases} \tag{12}$$

where C_n denotes the computing workload. Substituting Eqs. (10)–(12) into Eq. (9), the round-trip latency of VR user n is re-expressed as

$$\tau_{n,i}(\boldsymbol{b},\boldsymbol{l})=\begin{cases}\dfrac{S_n}{b_n^{\mathrm{U}}}+\dfrac{S_n^{\mathrm{DL}}}{b_n^{\mathrm{R}}}+\dfrac{C_n}{L_n^{\mathrm{V}}}, & i=1,\\ \dfrac{S_n}{b_n^{\mathrm{U}}}+\dfrac{S_n^{\mathrm{DL}}}{b_n^{\mathrm{R}}}, & i=2,\\ \dfrac{S_n}{b_n^{\mathrm{U}}}+\dfrac{S_n^{\mathrm{DL}}}{b_{m,n}^{\mathrm{A}}}, & i=3m,\\ \dfrac{S_n}{b_n^{\mathrm{U}}}+\dfrac{S_n^{\mathrm{DL}}}{b_{m,n}^{\mathrm{A}}}+\dfrac{C_n}{L_n^{\mathrm{V}}}, & i=3m+1,\\ \dfrac{S_n}{b_n^{\mathrm{U}}}+\dfrac{S_n^{\mathrm{DL}}}{b_{m,n}^{\mathrm{A}}}+\dfrac{C_n}{l_{m,n}^{\mathrm{A}}}, & i=3m+2,\end{cases} \tag{13}$$

where $\boldsymbol{b}$ and $\boldsymbol{l}$ are the tuples of the data rate and processing frequency allocations, respectively. Note that the round-trip latency $\tau_{n,i}(\boldsymbol{b},\boldsymbol{l})$ is the convex function of $\boldsymbol{b}$ and $\boldsymbol{l}$.

3 Problem formulation and resource allocation scheme

In this section, we first formulate a mixed-integer nonlinear programming (MINLP) problem to minimize the round-trip latency by jointly optimizing the communication, caching, and computation resource allocation. To solve this NP-hard problem, we decompose it into two subproblems: the mode selection problem and the joint communication and computation resource allocation problem. The joint communication and computation resource allocation problem is convex and can be solved with the convex optimization method. The mode selection problem is a 0-1 linear programming problem, which is solved with the proposed branch-and-bound-based algorithm, in which an optimal mode selection can be obtained within finite iterations.

3.1 Problem formulation

The key goal of our work is to jointly optimize the communication, caching, and computing resources to minimize the round-trip latency τ, satisfying the caching capacity, processing frequency, energy consumption, and transmission capacity constraints. Mathematically, the optimization problem can be formulated as

$$\min_{\boldsymbol{a},\boldsymbol{c},\boldsymbol{d},\boldsymbol{b},\boldsymbol{l}}:\tau\quad\text{s.t. Eqs. (1)}-\text{(8)}, \tag{14}$$

where $\boldsymbol{a}$, $\boldsymbol{c}$, and $\boldsymbol{d}$ denote the tuples of the user association, caching, and offloading decision, respectively. It is observed that problem (14) is an NP-hard MINLP problem. To tackle this problem, we replace the user association, caching, and offloading decisions, i.e., $\boldsymbol{a}$, $\boldsymbol{c}$, and $\boldsymbol{d}$, with the mode selection $q_{n,i}$. Then problem (14) can be transformed to

$$\min_{\boldsymbol{b},\boldsymbol{l},\boldsymbol{q}}:\sum_{n=1}^{N}\sum_{i=1}^{3M+2}\tau_{n,i}(\boldsymbol{b},\boldsymbol{l})\,q_{n,i}$$

$$\text{s.t.}\quad \text{C1:}\sum_{k=1}^{K}\sum_{n\in\mathcal{N}_k}\sum_{i=1}^{3M+2}p_{n,i}^{(1,m)}q_{n,i}\le C_m^{\mathrm{A}},$$

$$
\begin{aligned}
&\text{C2:} \sum_{k=1}^{K} \sum_{n \in \mathcal{N}_k} \sum_{i=1}^{3M+2} p_{n,i}^{(2,m)}(\boldsymbol{l})\, q_{n,i} \le E_m^{\mathrm{A}},\\
&\text{C3:} \sum_{n=1}^{N} \sum_{i=1}^{3M+2} p_{n,i}^{(3,m)}(\boldsymbol{l})\, q_{n,i} \le L_m^{\mathrm{A}},\\
&\text{C4:} \sum_{i=1}^{3M+2} p_{n,i}^{(4)} q_{n,i} \le E_n^{\mathrm{V}},\ \forall n \in \mathcal{N}_k,\\
&\text{C5:} \sum_{n=1}^{N} \sum_{i=1}^{3M+2} p_{n,i}^{(5)}(\boldsymbol{b})\, q_{n,i} \le R^{\mathrm{UF}},\\
&\text{C6:} \sum_{n=1}^{N} \sum_{i=1}^{3M+2} p_{n,i}^{(6,m)}(\boldsymbol{b})\, q_{n,i} \le R_m^{\mathrm{Umax}},\\
&\text{C7:} \sum_{n=1}^{N} \sum_{i=1}^{3M+2} p_{n,i}^{(7,m)}(\boldsymbol{b})\, q_{n,i} \le R_m^{\max},\\
&\text{C8:} \sum_{n=1}^{N} \sum_{i=1}^{3M+2} p_{n,i}^{(8)}(\boldsymbol{b})\, q_{n,i} \le R^{\mathrm{F}},\\
&\text{C9:} \sum_{i=1}^{3M+2} q_{n,i} = 1,\\
&\text{C10:}\ q_{n,i} \in \{0,1\},
\end{aligned}
\tag{15}
$$

where $\boldsymbol{q}$ denotes the tuple of $q_{n,i}$, and C1–C8 are obtained from Eqs. (1)–(8) by replacing $\boldsymbol{a}$, $\boldsymbol{c}$, and $\boldsymbol{d}$ with $\boldsymbol{q}$. C9 and C10 ensure that only one service mode can be chosen for each VR user.

In problem (15), constraint C1 means that the data volume of the VR videos cached at F-AP m must be less than the caching capacity of F-AP m, in which $p_{n,i}^{(1,m)}$ denotes the data volume of the VR videos when VR user $n \in \mathcal{N}_k$ is served in mode i, given by

$$
p_{n,i}^{(1,m)} = \begin{cases} D_k^{\mathrm{S}}, & i = 3m, \\ D_k^{\mathrm{M}}, & i = 3m+1, 3m+2, \\ 0, & \text{otherwise.} \end{cases} \tag{16}
$$

C2 and C4 indicate that the energy consumption for processing is no more than the total energy stored at the F-APs and VR users, respectively. $p_{n,i}^{(2,m)}(\boldsymbol{l})$ and $p_{n,i}^{(4)}$ denote the computing energy consumption at F-AP m and VR user n, respectively, when VR user $n \in \mathcal{N}_k$ is served in mode i, expressed as

$$
p_{n,i}^{(2,m)}(\boldsymbol{l}) = \begin{cases} \varepsilon_m^A \left(l_{m,n}^{\mathrm{A}}\right)^2 D_k^{\mathrm{M}} w_k, & i = 3m+2, \\ 0, & \text{otherwise,} \end{cases} \tag{17}
$$

$$
p_{n,i}^{(4)} = \begin{cases} \varepsilon_n^V \left(L_n^{\mathrm{V}}\right)^2 D_k^{\mathrm{M}} w_k, & i = 1, 3m+1, \\ 0, & \text{otherwise.} \end{cases} \tag{18}
$$

C3 ensures that the total computing frequency allocated to the VR users from the F-APs is less than L_m^{A}, where $p_{n,i}^{(3,m)}(\boldsymbol{l})$ denotes the frequency for VR user n at F-AP m when VR user n is served in mode i, given by

$$
p_{n,i}^{(3,m)}(\boldsymbol{l}) = \begin{cases} l_{m,n}^{\mathrm{A}}, & i = 3m+2, \\ 0, & \text{otherwise.} \end{cases} \tag{19}
$$

C5–C8 indicate that the total bandwidth for the VR users is less than the transmission capacity in both the uplink and downlink, where $p_{n,i}^{(5)}(\boldsymbol{b})$, $p_{n,i}^{(6,m)}(\boldsymbol{b})$, $p_{n,i}^{(7,m)}(\boldsymbol{b})$, and $p_{n,i}^{(8)}(\boldsymbol{b})$ denote the data rates of VR user n served in mode i, expressed as

$$
p_{n,i}^{(5)}(\boldsymbol{b}) = \begin{cases} b_n^{\mathrm{U}}, & i = 1, 2, \\ 0, & \text{otherwise,} \end{cases} \tag{20}
$$

$$
p_{n,i}^{(6,m)}(\boldsymbol{b}) = \begin{cases} b_n^{\mathrm{U}}, & i = 3m, 3m+1, 3m+2, \\ 0, & \text{otherwise,} \end{cases} \tag{21}
$$

$$
p_{n,i}^{(7,m)}(\boldsymbol{b}) = \begin{cases} b_{m,n}^{\mathrm{A}}, & i = 3m, 3m+1, 3m+2, \\ 0, & \text{otherwise,} \end{cases} \tag{22}
$$

$$
p_{n,i}^{(8)}(\boldsymbol{b}) = \begin{cases} b_n^{\mathrm{R}}, & i = 1, 2, \\ 0, & \text{otherwise.} \end{cases} \tag{23}
$$

3.2 Problem decomposition

It is observed that reformulated problem (15) is a non-convex MINLP problem, which is challenging to solve. To make it more tractable, we first decompose problem (15) into two subproblems as follows:

$$
\min_{\boldsymbol{q}} : \sum_{n=1}^{N} \sum_{i=1}^{3M+2} \tau_{n,i} q_{n,i} \quad \text{s.t. C1–C10,} \tag{24}
$$

$$
\min_{\boldsymbol{b},\boldsymbol{l}} : \sum_{n=1}^{N} \sum_{i=1}^{3M+2} \tau_{n,i}(\boldsymbol{b},\boldsymbol{l})\, q_{n,i} \quad \text{s.t. C2, C3, C5–C8.} \tag{25}
$$

Particularly, subproblem (24) is yielded with given $\boldsymbol{b}$ and $\boldsymbol{l}$, and subproblem (25) is obtained with given $\boldsymbol{q}$. Note that there is a mutual dependence among $\boldsymbol{b}$, $\boldsymbol{l}$, and $\boldsymbol{q}$ in problem (15); this decomposition implies that the objective values of problems (24) and (25) are the upper bounds for problem (15). To make the upper bounds approach the optimal round-trip latency, we propose Algorithm 1, an iterative algorithm, where $\delta_{\mathrm{out}} > 0$ is the termination parameter with a small value. Note that

Algorithm 1 is guaranteed to converge (Zhou et al., 2021).

Note that the round-trip latency achieved by Algorithm 1 is sensitive to the initial resource allocation decisions. Therefore, we perform Algorithm 1 repeatedly and select the resource allocation with the lowest round-trip latency.

Algorithm 1 The iterative algorithm for solving problem (15)

1: **Initialize** a feasible resource allocation $\boldsymbol{q}_0$, $\boldsymbol{b}_0$, and $\boldsymbol{l}_0$ and iteration index $j=1$. The round-trip latency is $\hat{\tau}_0$.
2: **repeat**
3: With the given $\boldsymbol{q}_{j-1}$, problem (25) is a convex optimization problem. Then, $\boldsymbol{b}_j$ and $\boldsymbol{l}_j$ are achieved by solving problem (25) with the traditional convex optimization approach.
4: With the given $\boldsymbol{b}_j$ and $\boldsymbol{l}_j$, problem (24) is a 0-1 linear programming problem. Then, $\boldsymbol{q}_j$ can be obtained by solving problem (24) with the branch-and-bound approach.
5: The corresponding round-trip latency can be calculated with $\boldsymbol{q}_j, \boldsymbol{b}_j$, and $\boldsymbol{l}_j$ and denoted as $\hat{\tau}_j$.
6: Set $j=j+1$.
7: **until** $|\hat{\tau}_j-\hat{\tau}_{j-1}|\leq\delta_{\text{out}}$

3.3 Service mode selection

Subproblem (24) is a 0-1 linear programming problem, which can be traditionally solved with an exhaustive search at the cost of a high computational complexity. To reduce the complexity, we resort to the branch-and-bound method (the branch-and-bound method can solve the formulated problem with less computational complexity compared to the traditional exhaustive search method) (Boyd and Mattingley, 2018). Denote τ^{opt} as the optimal objective value of problem (24). Relax the binary variable $\boldsymbol{q}$ into the continuous variable $0\leq q_{n,i}\leq 1$ to yield the following problem:

$$\begin{aligned}&\min_{\boldsymbol{q}}: \sum_{n=1}^{N}\sum_{i=1}^{3M+2}\tau_{n,i}q_{n,i}\\ \text{s.t.}\quad &\text{C1–C9}\\ &\text{C10':}\quad q_{n,i}\in[0,1].\end{aligned}\tag{26}$$

It is observed that problem (26) is a linear programming problem, which can be tackled with the traditional simplex method (Nelder and Mead, 1965). Let L_1 denote the objective latency of problem (26), which is a lower bound for τ^{opt}. If $L_1\to\infty$, the problem is surely infeasible. Otherwise, by rounding the relaxed variables and substituting them into the objective of problem (24), an upper bound for τ^{opt}, denoted by U_1, can be obtained. τ^{opt}, L_1, and U_1 satisfy $L_1\leq\tau^{\text{opt}}\leq U_1$.

Denote $\delta_{\text{in}}>0$ as the termination parameter with a small value. If the difference between L_1 and U_1 is small enough, which means that $U_1-L_1\leq\delta_{\text{in}}$, the required tolerance is satisfied, the algorithm terminates, and we have $\tau^{\text{opt}}=L_1$. Otherwise, we focus on branching. Pick any index (n,i) and form two subproblems from problem (24) by making $q_{n,i}=1$ and 2:

$$\begin{aligned}&\min_{\boldsymbol{q}}: \sum_{n=1}^{N}\sum_{i=1}^{3M+2}\tau_{n,i}q_{n,i}\\ \text{s.t.}\quad &\text{C1–C10}\\ &\text{C11:}\quad q_{n,i}=0,\end{aligned}\tag{27}$$

$$\begin{aligned}&\min_{\boldsymbol{q}}: \sum_{n=1}^{N}\sum_{i=1}^{3M+2}\tau_{n,i}q_{n,i}\\ \text{s.t.}\quad &\text{C1–C10}\\ &\text{C11:}\quad q_{n,i}=1.\end{aligned}\tag{28}$$

In this case, problem (24) is called the parent problem of problems (27) and (28). Then, relax problems (27) and (28), and solve them to obtain the lower and upper bounds for the optimal objective value of each subproblem. Denote these bounds as L_{s1} and U_{s1} for $q_{n,i}=0$ and L_{s2} and U_{s2} for $q_{n,i}=1$. L_{s1} and L_{s2} must be at least as large as L_1, i.e., $\min\{L_{s1},L_{s2}\}\geq L_1$. Similarly, the upper bounds satisfy $\min\{U_{s1},U_{s2}\}\leq U_1$. Then, τ^{opt} satisfies $L_2=\min\{L_{s1},L_{s2}\}\leq\tau^{\text{opt}}\leq U_2=\min\{U_{s1},U_{s2}\}$. We also observe that $U_2-L_2\leq U_1-L_1$.

If $U_2-L_2\leq\delta_{\text{in}}$, the algorithm terminates, and we have $\tau^{\text{opt}}=L_2$. Otherwise, consider problems (27) and (28) as the parent problems and split them on a new index (n',i'), which has not been used before. Similarly, the subproblems can be tackled in the same way as shown above, and a set of lower bounds and upper bounds is obtained, the minima of which give a lower bound and an upper bound for τ^{opt}. Denote the lower bound and upper bound as L_j and U_j, respectively. It can be observed that L_j is non-decreasing and that U_j is non-increasing. Thus, these bounds satisfy $U_{j+1}-L_{j+1}\leq U_j-L_j$. The iteration must quit when $U_j-L_j\leq\delta_{\text{in}}$.

Finally, the branch-and-bound-based algorithm for service mode selection is summarized in Algorithm 2. The convergence of this algorithm is guaranteed, because the algorithm must terminate in fewer than $2^{(3M+2)N}$ iterations, with $U_j = L_j$.

Algorithm 2 The branch-and-bound-based algorithm for solving problem (24)

1: **Initialize** termination precise δ_{in} and iteration index $j = 1$.
2: Solve the relaxation of problem (24) and obtain a lower bound L_1. If $L_1 \to \infty$, the problem is infeasible, and the algorithm terminates. Otherwise, calculate an upper bound U_1.
3: **repeat**
4: Determine the split indexes n' and i' and make them different from the values that have been used to split the parent problems.
5: Split the parent problems into several subproblems on $q_{n',i'} = 0$ and $q_{n',i'} = 1$.
6: Solve the relaxations of these subproblems and obtain the set of lower bounds.
7: If it is infeasible, prune this branch. Otherwise, calculate and obtain the set of upper bounds.
8: Choose the minimum lower bound as L_{j+1} and the minimum upper bound as U_{j+1}.
9: Set $j = j + 1$.
10: **until** $U_j - L_j \le \delta_{\text{in}}$

4 Numerical results

In this section, numerical results are presented to demonstrate the effectiveness of the proposed algorithm and examine the impact of the communication, caching, and computation resources on the round-trip latency performance. Consider one F-AP, 10 VR users, and 6 VR services, i.e., $M = 1$, $N = 10$, and $K = 6$ in this system. For each VR service, the value of D_k^{M} follows the uniform distribution $U(10, 15)$, the expectation of which is $E(D_k^{\text{M}}) = 12.5$ Mbit (Mb for short). The ratio of D_k^{S} to D_k^{M} follows the uniform distribution, given by $\alpha_k \sim U(2, 3)$, the data volume of the tracking information is $s_k \in [0, 0.7]$ Mb, the request probability follows Zipf distribution $P_k \propto 1/k^{0.8}$ (Zipf, 1929), and the number of cycles to process one bit is $w_k = 10$ cycles/bit. For VR user n, the request variable is $v_{n,k} = 1$ when VR service k is stochastically requested according to the probability. Unless specified, the power efficiency of the CPU, processing frequency, and energy of VR user n are $\varepsilon_n^{\text{V}} = 10^{-25}$, $L_n^{\text{V}} = 2$ GHz, and $E_n^{\text{V}} = 50$ J, respectively. For F-AP m, the power efficiency of the CPU, processing frequency, and energy are $\varepsilon_m^{\text{A}} = 10^{-26}$, $L_m^{\text{A}} = 10$ GHz, and $E_m^{\text{A}} = 100$ J, respectively. The caching capacity is $C_m^{\text{A}} = 0.2$ Gb. In addition, the uplink transmission capacities are given by $R^{\text{UF}} = 0.2$ Gb/s and $R_m^{\text{Umax}} = 0.1$ Gb/s, and the downlink transmission capacities are $R^{\text{F}} = 6$ Gb/s and $R_m^{\max} = 3$ Gb/s.

4.1 Performance comparison

First, we compare the proposed Algorithm 1 with the centralized-modes-only (CMO), distributed-modes-only (DMO), and half-centralized-half-distributed (HCHD) schemes. Specifically, in the CMO scheme, all VR users access the RRHs, where the mode selection and resource allocation are achieved with Algorithm 1. Similarly, in the DMO scheme, all VR users access the F-APs, where the mode selection and resource allocation are achieved with Algorithm 1. In the HCHD scheme, half of the VR users access the RRHs and the remaining half access the F-APs, randomly. The joint bandwidth and computing frequency allocation is achieved by solving convex problem (25).

Fig. 3 illustrates the round-trip latency per VR user and the number of VR users that access the RRHs vs. the downlink transmission capacity of the F-APs. In Fig. 3a, it is observed that the round-trip latency achieved by Algorithm 1 decreases as the downlink transmission capacity of the F-APs increases, and always outperforms the three other baselines. This situation occurs because the VR users can adaptively select the service mode under the joint bandwidth and computing frequency allocation. Specifically, it is observed that our proposed algorithm outperforms these baselines and can achieve up to 53% smaller round-trip latency for $R_m^{\max} = 8$ Gb/s. Fig. 3b shows the number of VR users that access the RRHs vs. the downlink transmission capacity of the F-APs for different schemes; the number of VR users is fixed at 10, 0, and 5 for the CMO, DMO, and HCHD schemes, respectively. As shown in Fig. 3b, the number of VR users that access the RRHs in Algorithm 1 decreases with the increase of the downlink transmission capacity of the F-APs. This situation occurs because with a higher $R_m^{\max}$, more VR users tend to access the F-APs to achieve a higher transmission rate. Furthermore, it is noted

that the HCHD scheme and Algorithm 1 achieve almost the same round-trip latency performance when the corresponding numbers of VR users that access the RRHs are the same.

In Fig. 4, we compare the round-trip latency per VR user achieved by our proposed algorithm with that achieved by the optimal solution obtained with an exhaustive search for different values of average D_k^{M}. It is observed that our proposed scheme achieves almost the same performance as the optimal solution, especially when D_k^{M} is relatively small.

4.2 Impact of caching capacity

The impact of the caching capacity on the round-trip latency is studied in Fig. 5. It is observed that the round-trip latency per VR user can be reduced by increasing the caching capacity of the F-APs. For instance, when $E(D_k^{\mathrm{M}}) = 16$ Mb, the round-trip latency decreases from 51.55 to 46.76 ms when the caching capacity is increased from 0.05 to 0.15 Gb. This situation occurs because more VR videos can be pre-cached at the F-APs with a larger caching capacity, which means that more VR users can be served by the F-APs for lower round-trip latency. This result suggests that the F-APs should be equipped with a larger caching capacity to achieve lower round-trip latency. Furthermore, a larger MV data volume leads to a higher round-trip latency.

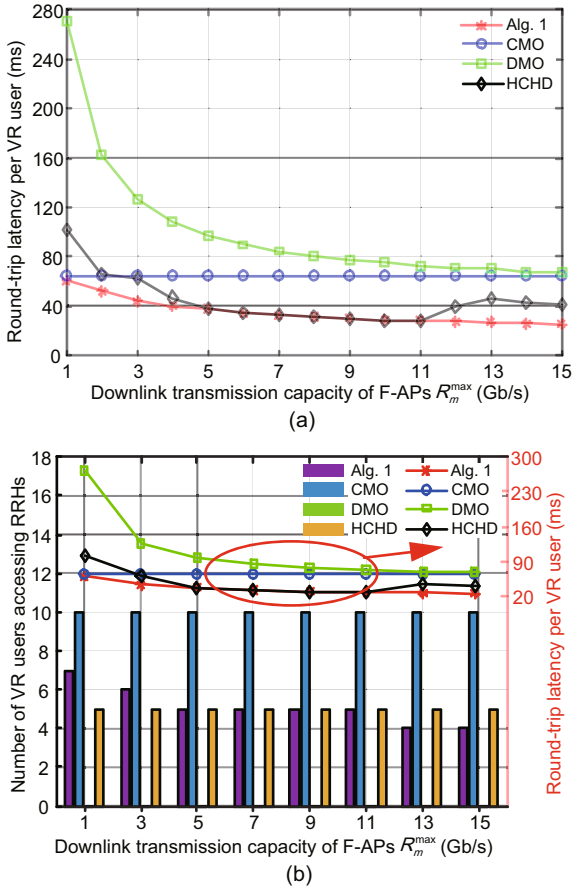

Fig. 3 Round-trip latency per VR user vs. downlink transmission capacity of F-APs (a) and number of VR users accessing RRHs vs. downlink transmission capacity of F-APs (b) with different resource allocation schemes

4.3 Impact of computing frequency

Fig. 6 shows the round-trip latency per VR user and the computing energy consumption vs. the computing frequency of the F-APs. The round-trip latency per VR user decreases based on the computing frequency of the F-APs, because higher computing frequency leads to lower processing latency according to Eq. (13). We also find that the computing energy consumption of the F-APs increases as the

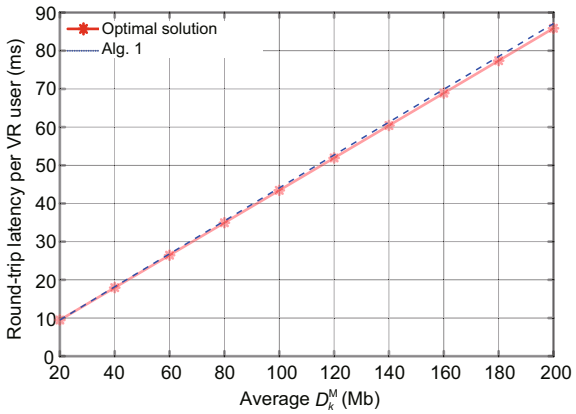

Fig. 4 Round-trip latency per VR user vs. average D_k^{M}

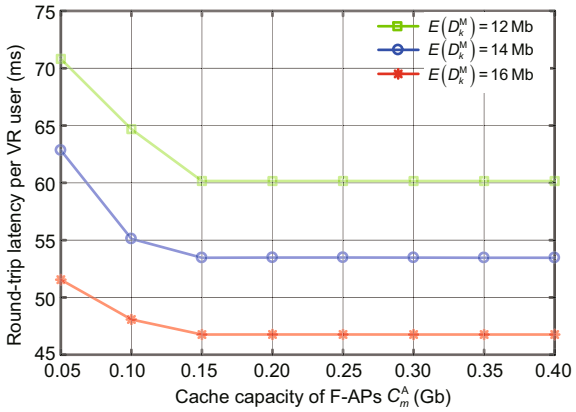

Fig. 5 Round-trip latency per VR user vs. caching capacity of F-APs C_m^{A} ($N = 10$ and $K = 10$)

computing frequency increases. Furthermore, the round-trip latency per VR user remains unchanged for different values of E_m^{A} for both small and large computing frequencies, because no computing task is offloaded to the F-APs when the computing frequency is low and the computing energy consumption of the F-APs meets the constraints when the computing frequency is high.

4.4 Impact of transmission capacity

Finally, we evaluate the latency performance with different uplink and downlink transmission capacities. Fig. 7 shows the round-trip latency performance as the uplink transmission capacity of the F-APs varies. It is observed that the round-trip latency per VR user decreases when the uplink transmission capacity increases. For instance, in the scenario of one F-AP and $E(s_k) = 0.4$ Mb, the round-trip latency per VR user decreases from 50.92 to 46.48 ms when the uplink transmission capacity R_m^{Umax} increases from 0.1 to 0.3 Gb/s. We also find that deploying more F-APs could reduce the latency, and that the larger data volume of the tracking information could lead to higher latency. In Fig. 8, we illustrate the round-trip latency per VR user vs. the uplink fronthaul capacity. It can be seen that the round-trip latency per VR user decreases as the uplink fronthaul capacity increases. For example, when the number of VR users is 10 and $E(s_k) = 0.4$ Mb, the round-trip latency per VR user decreases from 58.11 to 48.53 ms as the uplink fronthaul capacity increases from 0.1 to 0.3 Gb/s. We also find that more VR users and the larger data volume of the tracking information could lead to higher round-trip latency.

Fig. 9 shows the round-trip latency per VR user vs. the downlink transmission capacity of the F-APs for 10 VR users and 6 VR services. It is observed that

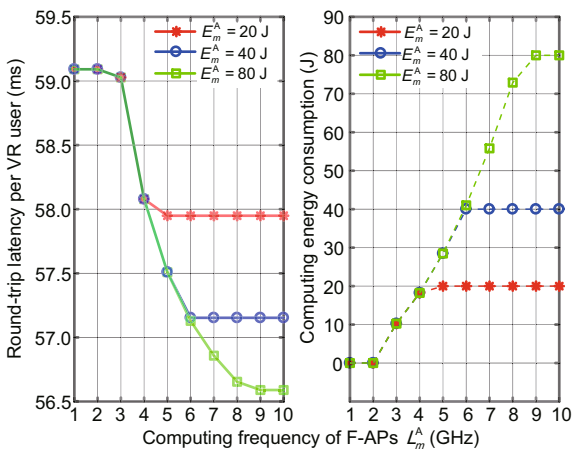

Fig. 6 Round-trip latency per VR user and computing energy consumption vs. computing frequency of F-APs ($K = 1$, $L_n^{\mathrm{V}} = 1.5$ GHz, $E_n^{\mathrm{V}} = 30$ J, $C_m^{\mathrm{A}} = 0.1$ Gb, and $R^{\mathrm{F}} = 3.5$ Gb/s)

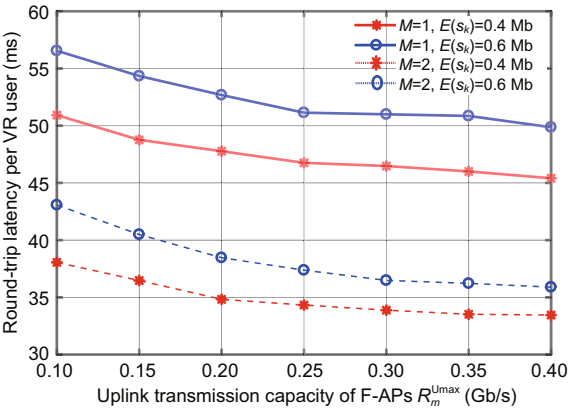

Fig. 7 Round-trip latency per VR user vs. uplink transmission capacity of F-APs R_m^{Umax}

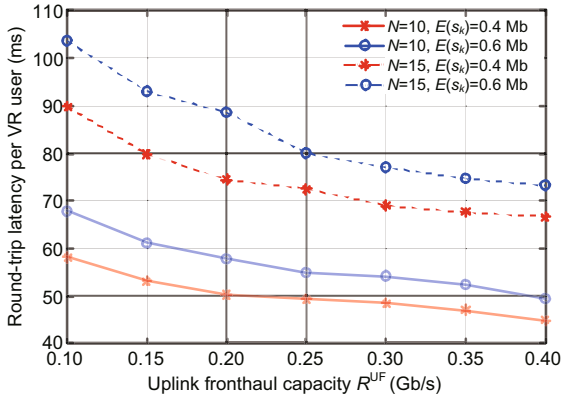

Fig. 8 Round-trip latency per VR user vs. uplink fronthaul capacity R^{UF}

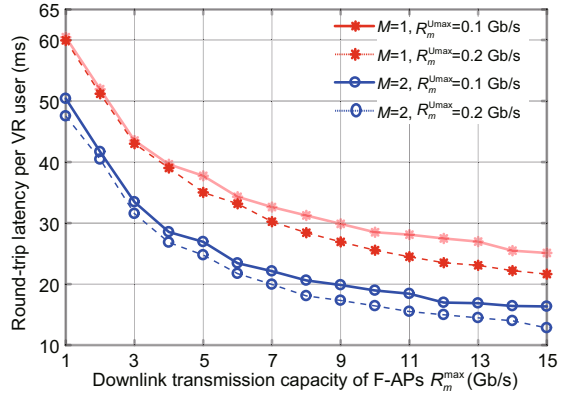

Fig. 9 Round-trip latency per VR user vs. downlink transmission capacity of F-APs R_m^{max} (10 VR users, 6 VR services)

the round-trip latency per VR user decreases with increased F-AP downlink transmission capacity. For example, in the case with one F-AP (i.e., $M = 1$) and the uplink transmission capacity of the F-APs of $R_m^{\mathrm{Umax}} = 0.1$ Gb/s, the round-trip latency per VR user decreases from 37.74 to 28.50 ms when the downlink transmission capacity $R_m^{\max}$ increases from 5 to 10 Gb/s. This situation occurs because the large downlink transmission capacity leads to lower transmission latency for the VR users that access the F-APs. Fig. 9 also shows that adding more F-APs and increasing the uplink transmission capacity help reduce the round-trip latency.

Fig. 10 illustrates the round-trip latency per VR user vs. the downlink fronthaul capacity, with one F-AP and 6 VR services, i.e., $M = 1$ and $K = 6$. It is observed that the round-trip latency per VR user decreases with increased downlink fronthaul capacity. With 10 VR users, up to 29.1% latency reduction is obtained when the downlink fronthaul capacity R^{F} increases from 5 to 10 Gb/s. This situation occurs because large downlink fronthaul capacity helps reduce the downlink transmission latency for the VR users that access the RRHs. We also find that more VR users could cause higher round-trip latency due to the constrained resource.

Finally, we evaluate the uplink transmission latency as the average s_k varies, as shown in Fig. 11. In Fig. 11, we consider a benchmark scheme, namely "equal uplink bandwidth," in which the uplink transmission capacity is equally allocated to the VR users. It is observed that the uplink transmission latency increases as the average s_k increases. Considering the case of 10 VR users, up to 32.4% latency increase is created with the average s_k increasing from 0.14 to 0.18 Mb. We also find that the proposed Algorithm 1 always outperforms the equal uplink bandwidth scheme, which demonstrates the effectiveness of our proposed Algorithm 1.

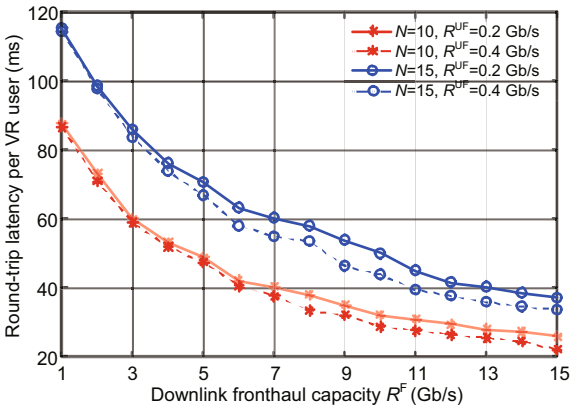

Fig. 10 Round-trip latency per VR user vs. downlink fronthaul capacity R^{F} for $R_m^{\mathrm{Umax}} = 0.2$ Gb/s ($M = 1$, $K = 6$)

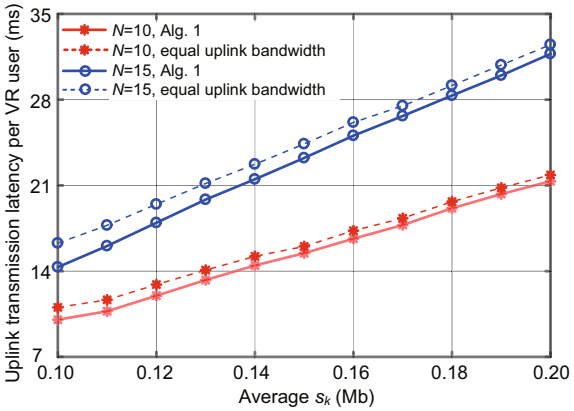

Fig. 11 Uplink transmission latency per VR user vs. average s_k ($C_m^{\mathrm{A}} = 0.1$ Gb, $R^{\mathrm{UF}} = 0.06$ Gb/s, $R_m^{\mathrm{Umax}} = 0.03$ Gb/s)

5 Conclusions

In this study, a mobile virtual reality (VR) delivery framework in fog radio access networks was investigated, in which both the uplink and downlink transmissions were considered. Specifically, we characterized the round-trip latency and showed how it was determined by the communication, caching, and computation resource allocations. Then, a simple yet efficient algorithm was proposed to minimize the round-trip latency, while satisfying the practical constraints on caching, computation capabilities, and transmission capacity. Numerical results were provided to verify the effectiveness of our proposed algorithm.

Contributors

Tian DANG carried out the analysis of the resource allocation optimization and drafted the paper. Chenxi LIU modeled the system and designed the algorithms. Xiqing LIU performed the simulations and helped organize the paper. Shi YAN helped with the optimization analysis and simulations. Tian DAN and Chenxi LIU revised and finalized the paper.

Compliance with ethics guidelines

Tian DANG, Chenxi LIU, Xiqing LIU, and Shi YAN declare that they have no conflict of interest.

References

Bastug E, Bennis M, Medard M, et al., 2017. Toward interconnected virtual reality: opportunities, challenges, and enablers. *IEEE Commun Mag*, 55(6):110-117. https://doi.org/10.1109/MCOM.2017.1601089

Boyd S, Mattingley J, 2018. Branch and Bound Methods. Stanford University, Stanford, USA.

Chen MZ, Semiari O, Saad W, et al., 2020. Federated echo state learning for minimizing breaks in presence in wireless virtual reality networks. *IEEE Trans Wirel Commun*, 19(1):177-191. https://doi.org/10.1109/TWC.2019.2942929

Chiu TC, Pang AC, Chung WH, et al., 2019. Latency-driven fog cooperation approach in fog radio access networks. *IEEE Trans Serv Comput*, 12(5):698-711. https://doi.org/10.1109/TSC.2018.2858253

Cisco System, 2019. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2020. White Paper.

Dai JM, Zhang ZL, Mao SW, et al., 2020. A view synthesis-based 360° VR caching system over MEC-enabled C-RAN. *IEEE Trans Circ Syst Video Technol*, 30(10):3843-3855. https://doi.org/10.1109/TCSVT.2019.2946755

Dang T, Peng MG, 2019. Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks. *IEEE J Sel Areas Commun*, 37(7):1594-1607. https://doi.org/10.1109/JSAC.2019.2916486

Du JB, Yu FR, Lu GY, et al., 2020. MEC-assisted immersive VR video streaming over terahertz wireless networks: a deep reinforcement learning approach. *IEEE Int Things J*, 7(10):9517-9529. https://doi.org/10.1109/JIOT.2020.3003449

Hu FH, Deng YS, Saad W, et al., 2020. Cellular-connected wireless virtual reality: requirements, challenges, and solutions. *IEEE Commun Mag*, 58(5):105-111. https://doi.org/10.1109/MCOM.001.1900511

Huang HC, Liu B, Chen L, et al., 2018. D2D-assisted VR video pre-caching strategy. *IEEE Access*, 6:61886-61895. https://doi.org/10.1109/ACCESS.2018.2868766

Liu YM, Yu FR, Li X, et al., 2018. Distributed resource allocation and computation offloading in fog and cloud networks with non-orthogonal multiple access. *IEEE Trans Veh Technol*, 67(12):12137-12151. https://doi.org/10.1109/TVT.2018.2872912

Nelder JA, Mead R, 1965. A simplex method for function minimization. *Comput J*, 7(4):308-313. https://doi.org/10.1093/comjnl/7.4.308

Park J, Popovski P, Simeone O, 2018. Minimizing latency to support VR social interactions over wireless cellular systems via bandwidth allocation. *IEEE Wirel Commun Lett*, 7(5):776-779. https://doi.org/10.1109/LWC.2018.2823761

Park SH, Simeone O, Shitz SS, 2016. Joint optimization of cloud and edge processing for fog radio access networks. *IEEE Trans Wirel Commun*, 15(11):7621-7632. https://doi.org/10.1109/TWC.2016.2605104

Peng MG, Yan S, Zhang KC, et al., 2016. Fog-computing-based radio access networks: issues and challenges. *IEEE Netw*, 30(4):46-53. https://doi.org/10.1109/MNET.2016.7513863

Sun YP, Chen ZY, Tao MX, et al., 2019. Communications, caching, and computing for mobile virtual reality: modeling and tradeoff. *IEEE Trans Commun*, 67(11):7573-7586. https://doi.org/10.1109/TCOMM.2019.2920594

Yoshihara T, Fujita S, 2019. Fog-assisted virtual reality MMOG with ultra low latency. 7th Int Symp on Computing and Networking, p.121-129. https://doi.org/10.1109/CANDAR.2019.00022

You D, Doan TV, Torre R, et al., 2019. Fog computing as an enabler for immersive media: service scenarios and research opportunities. *IEEE Access*, 7:65797-65810. https://doi.org/10.1109/ACCESS.2019.2917291

Zhang P, Peng MG, Cui SG, et al., 2022. Theory and techniques for "intellicise" wireless networks. *Front Inform Technol Electron Eng*, 23(1):1-4. https://doi.org/10.1631/FITEE.2210000

Zhang Y, Jiao L, Yan JY, et al., 2019. Dynamic service placement for virtual reality group gaming on mobile edge cloudlets. *IEEE J Sel Areas Commun*, 37(8):1881-1897. https://doi.org/10.1109/JSAC.2019.2927071

Zhou Y, Pan CH, Yeoh PL, et al., 2021. Communication-and-computing latency minimization for UAV-enabled virtual reality delivery systems. *IEEE Trans Commun*, 69(3):1723-1735. https://doi.org/10.1109/TCOMM.2020.3040283

Zipf GK, 1929. Relative Frequency as a Determinant of Phonetic Change. PhD Thesis, Harvard University, Cambridge, USA.