Communication-and-Computing Latency Minimization for UAV-Enabled Virtual Reality Delivery Systems

Yi Zhou, Cunhua Pan, Phee Lep Yeoh, Kezhi Wang, Maged Elkashlan,
Branka Vucetic, *Fellow, IEEE*, and Yonghui Li, *Fellow, IEEE*

***Abstract*—In this paper, we propose a low-latency virtual reality (VR) delivery system where an unmanned aerial vehicle (UAV) base station (U-BS) is deployed to deliver VR content from a cloud server to multiple ground VR users. Each VR input data requested by the VR users can be either projected at the U-BS before transmission or processed locally at each user. Popular VR input data is cached at the U-BS to further reduce backhaul latency from the cloud server. For this system, we design a low-complexity iterative algorithm to minimize the maximum communications and computing latency among all VR users subject to the computing, caching and transmit power constraints, which is guaranteed to converge. Numerical results indicate that our proposed algorithm can achieve a lower latency compared to other benchmark schemes. Moreover, we observe that the maximum latency mainly comes from communication latency when the bandwidth resource is limited, while it is dominated by computing latency when computing capacity is low. In addition, we find that caching is helpful to reduce latency.**

***Index Terms*—UAV communication, computing, caching, latency minimization, joint optimization.**

I. INTRODUCTION

THE demand for virtual reality (VR) applications that can create high-definition ultra-immersive VR environments for mobile users has increased significantly in 5G and beyond wireless networks [1, 2]. However, due to the low computing capability and finite battery lifetime of VR users, it is extremely challenging for wireless networks to support these computing-intensive and latency-sensitive VR applications. To alleviate computing resource constraints and reduce latency, mobile edge computing (MEC) has emerged as a promising enabler for VR delivery by equipping high-capacity computing resources at the network edge [3, 4]. In [5], a task scheduling strategy was proposed to solve a transmission data consumption minimization problem with delay constraint in MEC-enabled VR systems. In [6], by optimizing the bandwidth allocation of the uplink and downlink channels, the authors solved an end-to-end delay minimization problem in a VR mobile social edge network. In [7], the authors proposed an efficient algorithm to minimize the offloading energy consumption under latency and power constraints for augmented reality applications. In [8], an energy consumption minimization framework was developed for a two-tier computing offloading MEC network. In [9], the authors implemented and developed a low-latency management framework for distributed service function chains enabling tactile internet with MEC. In [10], by jointly coordinating the task assignment, computing, and transmission resources among edge devices, multi-layer MEC servers and cloud center, the authors proposed an efficient algorithm that aimed at minimizing the system latency including total computing and transmission time in heterogeneous multi-layer MEC networks. In [11], by jointly optimizing the users' transmit power, computing capacity allocation, and user association, a latency minimization problem of an MEC system was formulated.

To further reduce latency consumption, caching has been considered for MEC servers to pre-cache popular data files from the cloud servers during off-peak periods [12–14]. By doing so, the backhaul latency for requesting data from the cloud server can be minimized during peak periods. In [12], the authors formulated a joint radio communication, caching and computing decision problem to maximize the average tolerant delay with a given transmission rate constraint in a fog radio access network. In [13], joint caching and computing optimization was proposed to minimize the average transmission rate in MEC-based VR delivery systems. The authors in [14] jointly optimized the computation offloading, content caching and resource allocation such that the total latency consumption is minimized.

Due to its mobility and flexibility, unmanned aerial vehicle (UAV) is an ideal platform to provide high-quality and low-latency transmissions by deploying the UAV in close proximity to serve ground users [15–18]. Different from a ground base station which suffers from highly scattered Rayleigh fading channels, the UAV can exploit a strong line-of-sight (LoS) channel when it is above a certain altitude and the propagation conditions between the UAV and ground users can be approximated as free space. Furthermore, the UAV can be

The work of P. L. Yeoh was supported in part by ARC under Grant DP190100770. The work of Y. Li was supported by ARC under Grant DP190101988 and DP210103410. The work of B. Vucetic was supported in part by ARC Laureate Fellowship under Grant FL160100032. *(Corresponding author: Yonghui Li.)*

Y. Zhou, P. L. Yeoh, B. Vucetic, and Y. Li are with the School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia (e-mail: yi.zhou@sydney.edu.au; phee.yeoh@sydney.edu.au; branka.vucetic@sydney.edu.au; yonghui.li@sydney.edu.au).

C. Pan and M. Elkashlan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: c.pan@qmul.ac.uk; maged.elkashlan@qmul.ac.uk).

K. Wang is with the Department of Computer and Information Sciences, Northumbria University, Newcastle NE2 1XE, U.K. (e-mail: kezhi.wang@northumbria.ac.uk).

optimally deployed between the ground users and cloud server to further reduce the transmission and backhaul latency, which is perfectly suitable for latency-sensitive applications. Several papers have addressed the performance of UAV-enabled MEC systems with computing resource constraints [19–21]. In [19], a security maximization UAV-enabled MEC framework was proposed by jointly optimizing the UAV location, users' transmit power, UAV jamming power, offloading ratio, UAV computing capacity, and offloading user association. The authors in [20] developed a low-complexity power minimization algorithm by jointly optimizing user association, power control, computation capacity allocation and location planning in a MEC network with multiple UAVs. In [21], the UAV trajectory, user association and user offloading ratio were jointly optimized to minimize the maximum latency in UAV-MEC networks. The performance of a UAV-enabled caching system was investigated in [22–24]. In [22], a secure transmission scheme was proposed for a UAV-enabled caching system based on interference alignment. In [23], the UAV location, content caching decision and user association were jointly optimized to maximize the users' quality-of-experience. In [24], the file caching policy, UAV trajectory and file transmission scheduling were jointly optimized in a UAV-enabled network with proactive caching. Notably, no prior works have jointly considered the communication, computing, and caching (3C) performance of UAV systems which is critical for successful low-latency VR delivery, thus motivating this work.

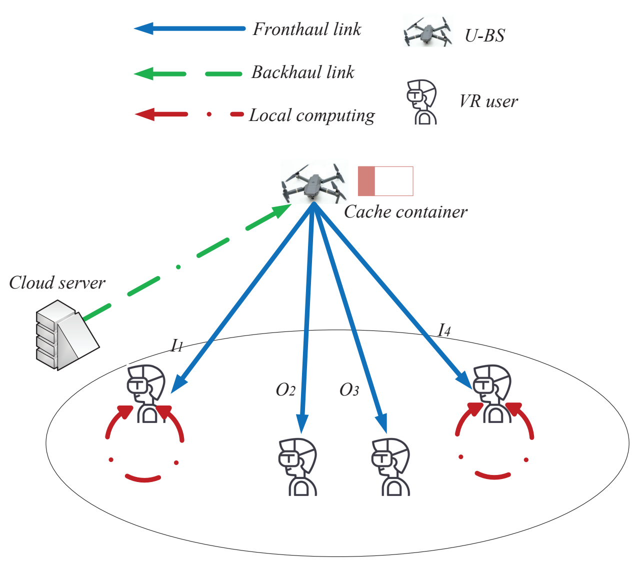


Fig. 1. UAV-enabled communication, computing and caching VR delivery system.

In this paper, we present a novel framework with the aim of minimizing the maximum latency of a UAV-enabled communication, computing and caching VR delivery system as shown in Fig. 1, which consists of one cloud server, one UAV aerial base station (U-BS) equipped with both caching and computing capabilities, and multiple ground VR users with local computing resources. To reduce the traffic burden on the backhaul link and backhaul communication latency, the U-BS caches the most popular input data in its cache container and the data which has not been cached at the U-BS needs to be transmitted from the cloud server via a wireless backhaul link. Moreover, to further reduce latency, the U-BS may choose to process the input data with its computing resource and transmit the projected output data to VR users for display, or send the input data to VR users directly for local computing. We note that compared to [5–7] where a VR delivery system was proposed for ground communications, our work exploits the advantages of UAV communications where the latency consumption can be further reduced by optimizing the UAV location, UAV computing capacity allocation and UAV caching policy. In addition, our work which jointly considers the computing and caching capabilities of UAV-enabled systems is different from other research on UAV communications such as [19–21] and [22–24] which solely focused on either computing or caching capabilities, respectively.

The main contributions of this paper are summarized as follows.

- We formulate a maximum latency minimization problem of a UAV-enabled VR delivery system by jointly optimizing the U-BS location, fronthaul and backhaul bandwidth allocation, computing capacity allocation, data caching policy and computing offloading policy subject to computing, caching and power constraints.
- To solve the non-convex optimization problem, we first apply the block coordinate descent (BCD) method to decouple the original optimization problem into six subproblems and propose a low-complexity algorithm to solve each subproblem alteratively. We solve the U-BS location subproblem by applying a successive convex approximation (SCA) on the U-BS data rate. Then, we apply Lagrangian dual decomposition method to efficiently solve the bandwidth and computing capacity allocation subproblems. Finally, we obtain efficient closed-form solutions for the caching and computing policy subproblems.
- Simulation results show that our proposed algorithm achieves a lower latency compared to benchmark strategies and highlight a tradeoff between latency and the primary resource requirements of communication, computing and caching.

The rest of this paper is organized as follows. Section II introduces the UAV-enabled communication, computing and caching VR delivery system model and formulates the joint optimization problem. In Section III, we propose an efficient iterative algorithm to minimize the maximum latency consumption. The effectiveness of our proposed solution is shown through simulation results in Section IV. Finally, we conclude the paper in Section V.

II. System Model

Fig. 1 depicts our proposed UAV-enabled communication, computing and caching VR delivery system with N ground VR users, one U-BS and one ground cloud server, where the set of VR users is denoted as $\mathcal{N} = \{1, 2, ..., N\}$. We

TABLE I
TABLE OF NOTATIONS

Notation	Description
$\mathcal{N}$	Set of VR users
θ_i	Fraction of fronthaul bandwidth allocated to the i-th VR user
η_i	Fraction of backhaul bandwidth allocated for transmitting I_i
f_i	Computing capacity of U-BS assigned to the i-th VR user
c_i	Caching policy variable
a_i	Computing policy variable
p_u, p_c	Transmit power for the U-BS and cloud server
σ^2	Noise spectral density
β_0	Reference channel power gain
$\mathbf{y}, \mathbf{w}_i, \mathbf{v}$	Horizontal location of the U-BS, the i-th VR user and cloud server
H_u	Altitude of U-BS
r_i	Fronthaul data rate at the i-th VR user
r_i^b	Backhaul data rate for transmitting I_i
I_i, O_i	Input and output data size of the i-th VR user
α	Ratio of size between O_i and I_i
F_i	CPU cycles for computing data I_i
f_i^{local}	Local computing capacity at the i-th VR user

consider that the U-BS has caching and computing capabilities which enable it to cache the data requested by each VR user from the cloud server via wireless backhaul and compute the data, respectively. Each VR user with computing capability is able to compute the data locally. We assume that all devices are equipped with a single antenna for transmitting or receiving. Due to the long distance and blockages from the cloud server to the VR users, the direct links between them are not applicable.

A. Computing Model

We assume that the i-th VR user has the computationally intensive task U_i to be executed as follows [13]

$$U_i = (I_i, O_i, F_i), \forall i \in \mathcal{N}, \tag{1}$$

where I_i is the input data in bits of the VR video required by the i-th user which is available in the remote cloud server, and may or may not be cached at the U-BS, $O_i = \alpha I_i$ is the output data in bits after being processed at the U-BS or locally with $\alpha \geq 2$ as the ratio of size between O_i and I_i [13], and F_i is the number of CPU cycles for computing one bit of input data I_i.

We consider VR projection and rendering in our system and define $a_i = \{0, 1\}, \forall i \in \mathcal{N}$, as the computing policy variable where $a_i = 1$ indicates that the input data I_i required by the i-th VR user will be projected at the U-BS. Thus, the U-BS processes the input data and transmits the output data O_i to VR users for display. On the other hand, $a_i = 0$ indicates that the i-th VR user decides to compute its data locally, but this user also needs to receive the input data, I_i, from the U-BS for calculation. Thus, the fronthaul transmission latency between the U-BS and each VR user is jointly decided by the computing policy, data size, and transmission rate, which is given by

$$t_i^{tr} = a_i \cdot \frac{O_i}{r_i} + (1 - a_i) \cdot \frac{I_i}{r_i}, \forall i \in \mathcal{N}, \tag{2}$$

where r_i is the transmission rate between the U-BS and the i-th VR user which is shown in (10). The first term in the right-hand-side (RHS) of (2) shows that if the data is computed at the U-BS, the output data O_i after being processed will be transmitted from U-BS to the VR user and the second term means that if the data is computed locally, the U-BS transmits the input data I_i to the VR user for calculation.

The computing latency which depends on computing policy, data size, computing capacity, and the required CPU cycles of the computing data, is given by

$$t_i^c = a_i \cdot \frac{I_i \cdot F_i}{f_i} + (1 - a_i) \cdot \frac{I_i \cdot F_i}{f_i^{local}}, \forall i \in \mathcal{N}, \tag{3}$$

where f_i^{local} is the local computing capacity at the i-th VR user and f_i is the computing capacity of the U-BS assigned to compute the data requested by the i-th VR user, which is constrained by a maximum computing capacity given by

$$\sum_{i=1}^{N} a_i f_i \leq f_{max}. \tag{4}$$

We note that if the i-th VR user decides to locally compute its data and $a_i = 0$, the U-BS will not allocate any computing capacity to this VR user and $f_i = 0$. We set the first term in the RHS of (3) to zero when $a_i = 0$ and $f_i = 0$.

We model the power consumption at the U-BS for computing the input data requested by the i-th VR user as [19]

$$p_i^c = \kappa f_i^3, \forall i \in \mathcal{N}, \tag{5}$$

where κ is the effective switched capacitance on the chip.

The total power consumption at the U-BS which consists of transmit power, p_u, and computing power should be limited by a maximum budget p_{max}, which is given by

$$p_u + \sum_{i=1}^{N} a_i \kappa f_i^3 \leq p_{max}. \tag{6}$$

B. Caching model

We define $c_i = \{0, 1\}, \forall i \in \mathcal{N}$ as the caching policy variable where $c_i = 1$ represents that the input data requested by the i-th VR user has been cached in the U-BS and $c_i = 0$ otherwise. We note that if the U-BS has cached the input data requested by the i-th VR user, it can apply the data directly from its cache container. Otherwise, the input data has to be transmitted from the cloud server to the U-BS and the corresponding backhaul latency is given by

$$t_i^b = (1 - c_i) \cdot \frac{I_i}{r_i^b}, \forall i \in \mathcal{N}, \tag{7}$$

where r_i^b representing the backhaul rate for transmitting I_i is given in (12).

Since different portions of the VR video are viewed by different VR users based on their geographical locations, we assume that the input data required by each VR user is different from each other. Note that the caching storage at the U-BS should be bounded by a maximum budget c_{max}, which is given by

$$\sum_{i=1}^{N} c_i I_i \leq c_{max}. \tag{8}$$

C. Communication Model

Assume that the coordinate of the i-th VR user is denoted by $\mathbf{w}_i = (x_i, y_i)^T \in \mathbb{R}^{2\times 1}, \forall i \in \mathcal{N}$. The U-BS is fixed at altitude H_u, which is the minimum altitude required by regulations to avoid building obstacles, and its horizontal location is denoted by $\mathbf{y} = (x_u, y_u)^T \in \mathbb{R}^{2\times 1}$. For the air-to-ground channel, we adopt a simple channel model where the channel power gains are dominated by the LoS links. Then, the channel power gain between the U-BS and the i-th VR user is given as [25, 26]

$$h_i = \frac{\beta_0}{||\mathbf{y} - \mathbf{w}_i||^2 + H_u^2}, \forall i \in \mathcal{N}, \tag{9}$$

where β_0 denotes the channel power gain at the reference distance of one meter.

Define $\theta_i \geq 0, \forall i \in \mathcal{N}$ as the fronthaul bandwidth allocation factor which represents the fraction of fronthaul bandwidth that the U-BS allocates to the i-th VR user. The achievable data rate at the i-th VR user is denoted by r_i in bits/second (bps), which is expressed as

$$r_i = \theta_i B \log_2 \left(1 + \frac{p_u h_i}{\theta_i B \sigma^2}\right), \forall i \in \mathcal{N}, \tag{10}$$

where p_u is the transmit power at the U-BS, B is the total fronthaul bandwidth, and σ^2 is the noise spectral density.

Assume that the coordinate of the cloud server is denoted by $\mathbf{v} = (x_c, y_c)^T \in \mathbb{R}^{2\times 1}$. Then, the channel power gain between the cloud server and the U-BS is given as [25, 26]

$$h_b = \frac{\beta_0}{||\mathbf{y} - \mathbf{v}||^2 + H_u^2}. \tag{11}$$

We define $\mathcal{N}_{uncached} = \{i | c_i = 0, \forall i \in \mathcal{N}\}$ as the set of VR users whose input data has not been cached in the U-BS and $\eta_i \geq 0, \forall i \in \mathcal{N}_{uncached}$ as the backhaul bandwidth allocation factor which represents the fraction of backhaul bandwidth that the cloud allocates to transmit the input data I_i which has not been cached in the U-BS. We note that for the i-th VR user whose requested input data I_i has been cached in the U-BS, i.e., $c_i = 1$, the cloud will not allocate any backhaul bandwidth for transmitting I_i and $\eta_i = 0, \forall i \in \mathcal{N}/\mathcal{N}_{uncached}$. The achievable backhaul data rate for transmitting I_i is denoted by r_i^b in bits/second (bps), which is expressed as

$$r_i^b = \eta_i B_{back} \log_2 \left(1 + \frac{p_c h_b}{\eta_i B_{back} \sigma^2}\right), \forall i \in \mathcal{N}_{uncached}, \tag{12}$$

where p_c is the transmit power at the cloud server and B_{back} is the total backhaul bandwidth.

According to (2), (3), (7), (10), and (12), the total latency to complete the task at each VR user is given by

$$\begin{aligned} t_i &= t_i^{tr} + t_i^c + t_i^b \\ &= a_i \left(\frac{O_i}{r_i} + \frac{I_i F_i}{f_i}\right) + (1 - a_i)\left(\frac{I_i F_i}{f_i^{local}} + \frac{I_i}{r_i}\right) + \frac{(1 - c_i) I_i}{r_i^b}, \\ &\qquad \forall i \in \mathcal{N}. \end{aligned} \tag{13}$$

D. Problem Formulation

We note that the satisfaction of VR experience among all users is dominated by the user who experiences the worst latency. To achieve the fairness among all VR users, we formulate an optimization problem aimed at minimizing the maximum latency among all VR users subject to computing capacity, caching storage and total power constraints. We jointly optimize the U-BS location $\mathbf{y} = \{(x_u, y_u)^T\}$, fronthaul bandwidth allocation $\boldsymbol{\theta} = \{\theta_i, \forall i \in \mathcal{N}\}$, backhaul bandwidth allocation $\boldsymbol{\eta} = \{\eta_i, \forall i \in \mathcal{N}_{uncached}\}$, computing capacity allocation $\mathbf{f} = \{f_i, \forall i \in \mathcal{N}\}$, data caching policy $\mathbf{c} = \{c_i, \forall i \in \mathcal{N}\}$, and computing policy $\mathbf{a} = \{a_i, \forall i \in \mathcal{N}\}$. The optimization problem can be formulated as

$$\min_{\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{f}, \mathbf{c}, \mathbf{a}} \quad \max_{i \in \mathcal{N}} t_i \tag{14a}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} a_i f_i \leq f_{max} \tag{14b}$$

$$\sum_{i=1}^{N} c_i I_i \leq c_{max} \tag{14c}$$

$$p_u + \sum_{i=1}^{N} a_i \kappa f_i^3 \leq p_{max} \tag{14d}$$

$$a_i = \{0, 1\}, \forall i \in \mathcal{N} \tag{14e}$$

$$c_i = \{0, 1\}, \forall i \in \mathcal{N} \tag{14f}$$

$$\sum_{i=1}^{N} \theta_i \leq 1 \tag{14g}$$

$$\sum_{i=1}^{N} \eta_i \leq 1. \tag{14h}$$

Define an auxiliary variable $T \triangleq \max_{i \in \mathcal{N}} t_i$ as the maximum latency among all VR users, we can reformulate the original optimization problem as

$$\min_{\mathbf{y},\boldsymbol{\theta},\boldsymbol{\eta},\mathbf{f},\mathbf{c},\mathbf{a},T} \quad T \tag{15a}$$

$$\text{s.t.} \quad a_i\left(\frac{O_i}{r_i}+\frac{I_iF_i}{f_i}\right)+(1-a_i)\left(\frac{I_iF_i}{f_i^{local}}+\frac{I_i}{r_i}\right)+\frac{(1-c_i)I_i}{r_i^b} \leq T, \forall i \in \mathcal{N} \tag{15b}$$

$$(14b)-(14h),$$

where the newly defined constraint (15b) is based on the intrinsic limitation that the latency consumption of each user should be less than its maximum value. Although reformulated, Problem (15) is still a non-convex optimization problem due to the following reasons. First, the optimizing variables for computing policy **a** and data caching policy **c** are binary integers. Second, even with given **a** and **c**, (15b) is still a non-convex constraint with respect to U-BS location **y**. Therefore, the main challenge that we will address in the following section is to develop an efficient algorithm to solve the latency optimization problem in (15).

III. Proposed Latency Minimization Algorithm

In this section, we detail our proposed latency minimization algorithm for UAV-enabled VR delivery systems. To solve Problem (15), we apply the BCD method which alternately optimizes one block of optimization variable in each iteration while keeping other blocks of optimization variables fixed to obtain a high-quality suboptimal solution [15]. Therefore, we can decouple the original optimization problem into six subproblems to solve the U-BS location **y**, fronthaul bandwidth allocation $\boldsymbol{\theta}$, backhaul bandwidth allocation $\boldsymbol{\eta}$, computing capacity allocation **f**, data caching policy **c**, and computing policy **a** in an iterative manner.

A. U-BS Location Subproblem

For any given $\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{f}, \mathbf{c}$, and **a**, the U-BS location of Problem (15) can be optimized by solving the following problem

$$\min_{\mathbf{y},T} \quad T \tag{16a}$$

$$\text{s.t.} \quad \frac{a_iO_i+(1-a_i)I_i}{\theta_i B\log_2\left(1+\frac{\zeta_i}{||\mathbf{y}-\mathbf{w}_i||^2+H_u^2}\right)}+\frac{(1-c_i)I_i}{\eta_i B_{back}\log_2\left(1+\frac{\gamma_i}{||\mathbf{y}-\mathbf{v}||^2+H_u^2}\right)} \leq T-\rho_i, \forall i \in \mathcal{N}, \tag{16b}$$

where the constraint (16b) corresponds to (15b), and all the other constraints in (15) are not applicable. In (16), we define $\zeta_i = \frac{p_u\beta_0}{\theta_i B\sigma^2}, \gamma_i = \frac{p_c\beta_0}{\eta_i B_{back}\sigma^2}$, and $\rho_i = a_i\frac{I_iF_i}{f_i}+(1-a_i)\frac{I_iF_i}{f_i^{local}}$. Note that (16) is a non-convex optimization problem and the non-convexity arises from the logarithm terms. In the following, we first introduce slack variables $\boldsymbol{\epsilon} \triangleq \{\epsilon_i, \forall i \in \mathcal{N}\}$ and $\boldsymbol{\omega} \triangleq \{\omega_i, \forall i \in \mathcal{N}_{uncached}\}$, and reformulate the U-BS location subproblem as

$$\min_{\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\omega},T} \quad T \tag{17a}$$

$$\text{s.t.} \quad \frac{a_iO_i+(1-a_i)I_i}{\epsilon_i}+\frac{(1-c_i)I_i}{\omega_i} \leq T-\rho_i, \forall i \in \mathcal{N} \tag{17b}$$

$$\underbrace{\theta_i B\log_2\left(1+\frac{\zeta_i}{||\mathbf{y}-\mathbf{w}_i||^2+H_u^2}\right)}_{\mathcal{I}_i} \geq \epsilon_i, \forall i \in \mathcal{N} \tag{17c}$$

$$\underbrace{\eta_i B_{back}\log_2\left(1+\frac{\gamma_i}{||\mathbf{y}-\mathbf{v}||^2+H_u^2}\right)}_{\mathcal{Z}_i} \geq \omega_i, \forall i \in \mathcal{N}_{uncached}. \tag{17d}$$

We note that the constraint (17b) is convex now and the non-convexity of Problem (17) arises from constraints (17c) and (17d). Next, we adopt the SCA technique, where the original function can be approximated by a more tractable expression at a given local point in each iteration [15][19]. We note that $\mathcal{I}_i$ is convex with respect to $||\mathbf{y}-\mathbf{w}_i||^2$, thus, a concave lower bound expression $\mathcal{I}_i^{lb}$ can be derived by applying the first-order Taylor expansion with given U-BS location $\mathbf{y}[m]$ in the m-th iteration, which is given by

$$\mathcal{I}_i^{lb} = \theta_i B\log_2\left(1+\frac{\zeta_i}{||\mathbf{y}[m]-\mathbf{w}_i||^2+H_u^2}\right) - \frac{\theta_i B\zeta_i(||\mathbf{y}-\mathbf{w}_i||^2-||\mathbf{y}[m]-\mathbf{w}_i||^2)}{(||\mathbf{y}[m]-\mathbf{w}_i||^2+H_u^2+\zeta_i)(||\mathbf{y}[m]-\mathbf{w}_i||^2+H_u^2)\ln 2}. \tag{18}$$

We apply a similar approach on $\mathcal{Z}_i$ and the corresponding concave lower bound $\mathcal{Z}_i^{lb}$ is given by

$$\mathcal{Z}_i^{lb} = \eta_i B_{back}\log_2\left(1+\frac{\gamma_i}{||\mathbf{y}[m]-\mathbf{v}||^2+H_u^2}\right) - \frac{\eta_i B_{back}\gamma_i(||\mathbf{y}-\mathbf{v}||^2-||\mathbf{y}[m]-\mathbf{v}||^2)}{(||\mathbf{y}[m]-\mathbf{v}||^2+H_u^2+\gamma_i)(||\mathbf{y}[m]-\mathbf{v}||^2+H_u^2)\ln 2}. \tag{19}$$

With given U-BS location $\mathbf{y}[m]$ and the lower bound expressions in (18) and (19), the U-BS location subproblem can be solved as

$$\min_{\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\omega},T} \quad T \tag{20a}$$

$$\text{s.t.} \quad \frac{a_iO_i+(1-a_i)I_i}{\epsilon_i}+\frac{(1-c_i)I_i}{\omega_i} \leq T-\rho_i, \forall i \in \mathcal{N} \tag{20b}$$

$$\mathcal{I}_i^{lb} \geq \epsilon_i, \forall i \in \mathcal{N} \tag{20c}$$

$$\mathcal{Z}_i^{lb} \geq \omega_i, \forall i \in \mathcal{N}_{uncached}. \tag{20d}$$

We note that Problem (20) is a convex optimization problem and it can be efficiently solved by utilizing mathematical optimization software with the polynomial complexity [27].

B. Fronthaul Bandwidth Allocation Subproblem

For any given $\mathbf{y}, \boldsymbol{\eta}, \mathbf{f}, \mathbf{c}$, and **a**, the fronthaul bandwidth allocation of Problem (15) can be optimized by solving the following problem

$$
\begin{aligned}
&\min_{\boldsymbol{\theta},T} \quad T && \text{(21a)}\\
&\text{s.t.} \quad \frac{\chi_i}{T-\nu_i} \le \theta_i B \log_2\left(1+\frac{p_u h_i}{\theta_i B \sigma^2}\right), \forall i \in \mathcal{N} && \text{(21b)}\\
&\sum_{i=1}^{N} \theta_i \le 1. && \text{(21c)}\\
&\theta_i \ge 0, \forall i \in \mathcal{N} && \text{(21d)}\\
&T \ge \nu_i, \forall i \in \mathcal{N}, && \text{(21e)}
\end{aligned}
$$

where $\nu_i = a_i \frac{I_i F_i}{f_i} + (1-a_i)\frac{I_i F_i}{f_i^{local}} + \frac{(1-c_i) I_i}{r_i^b}$ and $\chi_i = a_i O_i + (1-a_i) I_i$. We define $\theta_i B \log_2\left(1+\frac{p_u h_i}{\theta_i B \sigma^2}\right) \triangleq 0$ when $\theta_i = 0, \forall i \in \mathcal{N}$, such that the RHS of (21b) is continuous with respect to θ_i over the whole domain. We analyze the convexity of Problem (21) in the following lemma.

Lemma 1. *Problem* (21) *is a convex problem.*

Proof. It can be easily noted that (21a), (21c), (21d), and (21e) are convex terms due to their linearity. Therefore, proving Lemma 1 is equivalent to proving that the constraint (21b) is convex. To show this, we define $g(x) = x \log_2\left(1+\frac{1}{x}\right), x > 0$, and we have

$$
\frac{\partial^2 g}{\partial x^2} = -\frac{1}{x(x+1)^2 \ln 2} < 0, \ \forall x > 0, \tag{22}
$$

which indicates that $g(x)$ is a concave function. Thus, we can conclude that the RHS of (21b) is a concave term with respect to θ_i. Moreover, We note that the left-hand-side (LHS) of (21b) is a convex term with respect to T. Therefore, we show that the constraint (21b) is convex and prove that Problem (21) is a convex problem. □

Next, we apply the Lagrangian dual decomposition method to solve this convex problem. It can be verified that the Slater's condition is satisfied for Problem (21), which indicates that the duality gap between (21) and its dual problem is zero [27]. The partial Lagrangian function of Problem (21) is given by

$$
\begin{aligned}
\mathcal{L}(T, \boldsymbol{\theta}, \boldsymbol{\mu}, \iota) = T &+ \sum_{i=1}^{N} \frac{\mu_i \chi_i}{T-\nu_i} \\
&+ \sum_{i=1}^{N} \left[\iota \theta_i - \mu_i \theta_i B \log_2\left(1+\frac{p_u h_i}{\theta_i B \sigma^2}\right)\right] - \iota,
\end{aligned} \tag{23}
$$

where $\boldsymbol{\mu} = \{\mu_i, \forall i \in \mathcal{N}\}$ and ι are Lagrangian multipliers associated with constraints (21b) and (21c), respectively. The boundary constraints (21d) and (21e) will be absorbed into the optimal solution in the following. The dual function is given by

$$
\begin{aligned}
&f(\boldsymbol{\mu}, \iota) = \min_{T, \boldsymbol{\theta}} \mathcal{L}(T, \boldsymbol{\theta}, \boldsymbol{\mu}, \iota) && \text{(24a)}\\
&s.t. \quad \theta_i \ge 0, T \ge \nu_i, \forall i \in \mathcal{N}, && \text{(24b)}
\end{aligned}
$$

and the dual problem of (21) is given by

$$
\begin{aligned}
&\max_{\boldsymbol{\mu}, \iota} \ f(\boldsymbol{\mu}, \iota) && \text{(25a)}\\
&s.t.\ \boldsymbol{\mu} \succeq 0, \iota \ge 0. && \text{(25b)}
\end{aligned}
$$

To derive the primal optimal solution of Problem (21), we apply the Lagrange duality and derive $f(\boldsymbol{\mu}, \iota)$ by solving Problem (24). We note that with given dual variables $\boldsymbol{\mu}$ and ι, Problem (24) can be decomposed into $N+1$ independent subproblems where one subproblem is for optimizing T and the other N subproblems are for optimizing $\theta_i, \forall i \in \mathcal{N}$. The subproblem for optimizing T can be formulated as

$$
\begin{aligned}
&\min_{T} \ T + \sum_{i=1}^{N} \frac{\mu_i \chi_i}{T-\nu_i} && \text{(26a)}\\
&s.t.\ T \ge \nu_i, \forall i \in \mathcal{N}. && \text{(26b)}
\end{aligned}
$$

By setting the first-order derivative of (26a) with respect to T to zero, we observe that the optimal T should satisfy

$$
T = \left\{ T \,|\, \sum_{i=1}^{N} \frac{\mu_i \chi_i}{(T-\nu_i)^2} = 1, T \ge \nu_i \right\}, \tag{27}
$$

which can be found by applying the bisection method.

Moreover, the subproblem for optimizing $\theta_i, \forall i \in \mathcal{N}$ can be formulated as

$$
\begin{aligned}
&\min_{\theta_i} \ \iota \theta_i - \mu_i \theta_i B \log_2\left(1+\frac{p_u h_i}{\theta_i B \sigma^2}\right) && \text{(28a)}\\
&s.t.\ \theta_i \ge 0, \forall i \in \mathcal{N}. && \text{(28b)}
\end{aligned}
$$

By setting the first-order derivative of (28a) with respect to θ_i to zero, we obtain the closed-form expression of the optimal bandwidth allocation as

$$
\theta_i = \left[\frac{p_u h_i}{B \sigma^2} \left(-\frac{1}{\mathcal{W}\left(-\frac{1}{\exp\left(1+\frac{\iota \ln 2}{\mu_i B}\right)}\right)} - 1 \right)^{-1} \right]^{+}, \tag{29}
$$

where $[x]^{+} = \max\{x, 0\}$ and $\mathcal{W}(x)$ is the Lambert function, which is defined as the inverse function of $f(x) = x \exp(x)$.

The value of dual variables $\boldsymbol{\mu}$ and ι can be determined by the sub-gradient method. The updating procedure can be given by

$$
\mu_i = \left[\mu_i + \phi \left(\frac{\chi_i}{T-\nu_i} - \theta_i B \log_2\left(1+\frac{p_u h_i}{\theta_i B \sigma^2}\right) \right) \right]^{+}, \forall i \in \mathcal{N} \tag{30a}
$$

$$
\iota = \left[\iota + \phi \left(\sum_{i=1}^{N} \theta_i - 1 \right) \right]^{+}, \tag{30b}
$$

where $\phi > 0$ is a dynamic step-size sequence, which can be selected by using the typical self-adaptive scheme [18].

We note that in the primal problem of (21), the optimal T and $\boldsymbol{\theta}$ can be derived by solving (27) and (29), respectively. Moreover, in the dual problem of (21), the optimal dual variables $\boldsymbol{\mu}$ and ι can be found by solving (30a) and (30b), respectively. The details for obtaining the optimal solution to Problem (21) are summarized in Algorithm 1. We note that Problem (24) has been decomposed into $N+1$ subproblems. To solve the subproblem for optimizing T, the complexity of solving (27) via the bisection method is $\mathcal{O}(\log_2(1/\epsilon))$ with ϵ being the iterative accuracy. To solve each of the N subproblems, since the closed-from solution has been derived in (29), the complexity for these N subproblems is $O(N)$. Moreover, the complexity of updating dual variables is $\mathcal{O}(N)$ according to (30a) and (30b). As a result, the total complexity of Algorithm 1 is $\mathcal{O}(L_1 L_2 N^2 \log_2(1/\epsilon))$, where L_1 is the number of iterations for outer layer in Algorithm 1 and L_2 is the number of iterations via the dual method of solving Problem (21).

Algorithm 1 Fronthaul Bandwidth Allocation Algorithm for Solving Problem (21).

1: Initialize $\boldsymbol{\mu}$ and ι.
2: **repeat**
3: Obtain the optimal T and $\boldsymbol{\theta}$ by solving (27) and (29), respectively;
4: Update the Lagrangian multipliers $\boldsymbol{\mu}$ and ι by solving (30a) and (30b), respectively;
5: **until** The objective function in (21a) converges.

C. Backhaul Bandwidth Allocation Subproblem

For any given $\mathbf{y}, \boldsymbol{\theta}, \mathbf{f}, \mathbf{c}$, and $\mathbf{a}$, the backhaul bandwidth allocation of Problem (15) can be optimized by solving the following problem

$$\min_{\boldsymbol{\eta}, T} \quad T \tag{31a}$$

$$\text{s.t.} \quad \frac{I_i}{T - z_i} \leq \eta_i B_{back} \log_2\left(1 + \frac{p_c h_b}{\eta_i B_{back} \sigma^2}\right), \forall i \in \mathcal{N}_{uncached} \tag{31b}$$

$$\sum_{i=1}^{N_{uncached}} \eta_i \leq 1 \tag{31c}$$

$$\eta_i \geq 0, \forall i \in \mathcal{N}_{uncached} \tag{31d}$$

$$T \geq z_i, \forall i \in \mathcal{N}_{uncached}, \tag{31e}$$

where $z_i = a_i \left(\frac{O_i}{r_i} + \frac{I_i F_i}{f_i}\right) + (1 - a_i)\left(\frac{I_i F_i}{f_i^{local}} + \frac{I_i}{r_i}\right)$. We set $\eta_i B_{back} \log_2\left(1 + \frac{p_c h_b}{\eta_i B_{back} \sigma^2}\right) \triangleq 0$ when $\eta_i = 0, \forall i \in \mathcal{N}_{uncached}$, such that the RHS of (31b) is continuous with respect to η_i over the whole domain. We note that Problem (31) is a convex problem and the proof is similar to that of (21) in Subsection III-B.

As such, we can apply the Lagrangian dual decomposition method to solve Problem (31). We denote $\boldsymbol{\lambda} = \{\lambda_i, \forall i \in \mathcal{N}_{uncached}\}$ and ς as Lagrangian multipliers associated with constraints (31b) and (31c), respectively. The boundary constraints (31d) and (31e) will be absorbed into the optimal solution in the following. The partial Lagrangian function of Problem (31) is given by

$$\mathcal{L}(T, \boldsymbol{\eta}, \boldsymbol{\lambda}, \varsigma) = T + \sum_{i=1}^{N_{uncached}} \frac{\lambda_i I_i}{T - z_i} + \sum_{i=1}^{N_{uncached}} \left[\varsigma \eta_i - \lambda_i \eta_i B_{back} \log_2\left(1 + \frac{p_c h_b}{\eta_i B_{back} \sigma^2}\right)\right] - \varsigma. \tag{32}$$

The dual function is given by

$$f(\boldsymbol{\lambda}, \varsigma) = \min_{T, \boldsymbol{\eta}} \mathcal{L}(T, \boldsymbol{\eta}, \boldsymbol{\lambda}, \varsigma) \tag{33a}$$

$$s.t. \quad \eta_i \geq 0, T \geq z_i, \forall i \in \mathcal{N}_{uncached}, \tag{33b}$$

and the dual problem of (31) is given by

$$\max_{\boldsymbol{\lambda}, \varsigma} f(\boldsymbol{\lambda}, \varsigma) \tag{34a}$$

$$s.t. \, \boldsymbol{\lambda} \succeq 0, \varsigma \geq 0. \tag{34b}$$

To derive the primal optimal solution of Problem (31), we apply the Lagrange duality method and derive $f(\boldsymbol{\lambda}, \varsigma)$ by solving Problem (33). We note that with given dual variables $\boldsymbol{\lambda}$ and ς, Problem (33) can be decomposed into $N_{uncached} + 1$ independent subproblems where one subproblem is for optimizing T and the other $N_{uncached}$ subproblems are for optimizing $\eta_i, \forall i \in \mathcal{N}_{uncached}$. The subproblem for optimizing T can be formulated as

$$\min_{T} \; T + \sum_{i=1}^{N_{uncached}} \frac{\lambda_i I_i}{T - z_i} \tag{35a}$$

$$s.t. \; T \geq z_i, \forall i \in \mathcal{N}_{uncached}. \tag{35b}$$

By setting the first-order derivative of (35a) with respect to T to zero, we find that the optimal T should satisfy

$$T = \left\{T \middle| \sum_{i=1}^{N_{uncached}} \frac{\lambda_i I_i}{(T - z_i)^2} = 1, T \geq z_i\right\}, \tag{36}$$

which can be solved by applying the bisection search method.

Moreover, the subproblem for optimizing $\eta_i, \forall i \in \mathcal{N}_{uncached}$ can be formulated as

$$\min_{\eta_i} \; \varsigma \eta_i - \lambda_i \eta_i B_{back} \log_2\left(1 + \frac{p_c h_b}{\eta_i B_{back} \sigma^2}\right) \tag{37a}$$

$$s.t. \; \eta_i \geq 0, \forall i \in \mathcal{N}_{uncached}. \tag{37b}$$

By setting the first-order derivative of (37a) with respect to η_i to zero, we obtain the closed-form expression of the optimal backhaul bandwidth allocation as

$$\eta_i = \left[\frac{p_c h_b}{B_{back} \sigma^2}\left(-\frac{1}{\mathcal{W}\left(-\frac{1}{\exp\left(1 + \frac{\varsigma \ln 2}{\lambda_i B_{back}}\right)}\right)} - 1\right)^{-1}\right]^+. \tag{38}$$

The value of dual variables $\boldsymbol{\lambda}$ and ς can be determined by the sub-gradient method. The updating procedure is given by

$$\lambda_i = \left[\lambda_i + \phi\left(\frac{I_i}{T - z_i} - \eta_i B_{back} \log_2\left(1 + \frac{p_c h_b}{\eta_i B_{back} \sigma^2}\right)\right)\right]^+ \quad \forall i \in \mathcal{N}_{uncached} \tag{39a}$$

$$\varsigma = \left[\varsigma + \phi\left(\sum_{i=1}^{N_{uncached}} \eta_i - 1\right)\right]^+. \tag{39b}$$

The procedures for obtaining the optimal solution to Problem (31) is summarized in Algorithm 2. Similar to the complexity analysis in Subsection III-B, we note that the total complexity of Algorithm 2 is $\mathcal{O}(L_3 L_4 N^2 \log_2(1/\epsilon))$, where L_3 is the number of iterations for outer layer in Algorithm 2 and L_4 is the number of iterations via the dual method of solving Problem (31).

D. Computing Capacity Allocation Subproblem

We define $\mathcal{N}_{as} = \{i | a_i = 1, \forall i \in \mathcal{N}\}$ as the set of VR users who choose to compute their input data at the U-BS. Note that for the VR users who choose to self-execute their tasks, the U-BS will not allocate computing capacity to them and $f_i = 0, \forall i \in \mathcal{N}/\mathcal{N}_{as}$. For any given $\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{c}$, and $\mathbf{a}$, the computing capacity allocation of Problem (15) can be optimized by solving the following problem

Algorithm 2 Backhaul Bandwidth Allocation Algorithm for Solving Problem (31).

1: Initialize $\boldsymbol{\lambda}$ and ς.
2: **repeat**
3: Obtain the optimal T and $\boldsymbol{\eta}$ by solving (36) and (38), respectively;
4: Update the Lagrangian multipliers $\boldsymbol{\lambda}$ and ς by solving (39a) and (39b), respectively;
5: **until** The objective function in (31a) converges.

$$\min_{\mathbf{f},T} \quad T \tag{40a}$$
$$\text{s.t.} \quad \frac{I_i F_i}{T-\varpi_i} \le f_i, \forall i \in \mathcal{N}_{as} \tag{40b}$$
$$\sum_{i=1}^{N_{as}} \kappa f_i^3 \le p_{max} - p_u \tag{40c}$$
$$\sum_{i=1}^{N_{as}} f_i \le f_{max} \tag{40d}$$
$$f_i \ge 0, \forall i \in \mathcal{N}_{as} \tag{40e}$$
$$T \ge \varpi_i, \forall i \in \mathcal{N}_{as}, \tag{40f}$$

where $\varpi_i = a_i \frac{O_i}{r_i} + (1-a_i)(\frac{I_i F_i}{f_i^{local}} + \frac{I_i}{r_i}) + \frac{(1-c_i)I_i}{r_i^b}$. We note that Problem (40) is a convex optimization problem since the objective function and all constraints are convex and it can be effectively solved via the Lagrangian dual decomposition method.

We denote $\boldsymbol{\tau} = \{\tau_i, \forall i \in \mathcal{N}_{as}\}$, δ, and ξ as the Lagrangian multipliers associated with constraints (40b), (40c) and (40d), respectively. The boundary constraints (40e) and (40f) will be absorbed into the optimal solution in the following. The partial Lagrangian function of Problem (40) is given by

$$\begin{aligned}\mathcal{L}(T, \mathbf{f}, \boldsymbol{\tau}, \delta, \xi) =& T + \sum_{i=1}^{N_{as}} \frac{\tau_i I_i F_i}{T-\varpi_i} + \sum_{i=1}^{N_{as}} \left(\delta\kappa f_i^3 + \xi f_i - \tau_i f_i\right) \\ &+ \delta p_u - \delta p_{max} - \xi f_{max}.\end{aligned} \tag{41}$$

The dual function is given by

$$f(\boldsymbol{\tau}, \delta, \xi) = \min_{T,\mathbf{f}} \mathcal{L}(T, \mathbf{f}, \boldsymbol{\tau}, \delta, \xi) \tag{42a}$$
$$s.t. \quad f_i \ge 0, T \ge \varpi_i, \forall i \in \mathcal{N}_{as}, \tag{42b}$$

and the dual problem of (40) is given by

$$\max_{\boldsymbol{\tau},\delta,\xi} \ f(\boldsymbol{\tau}, \delta, \xi) \tag{43a}$$
$$s.t. \boldsymbol{\tau} \succeq 0, \delta \ge 0, \xi \ge 0. \tag{43b}$$

To derive the primal optimal solution of Problem (40), we apply the Lagrange duality method and derive $f(\boldsymbol{\tau}, \delta, \xi)$ by solving Problem (42). We note that with given dual variables $\boldsymbol{\tau}$, δ, and ξ, Problem (42) can be decomposed into $N_{as}+1$ independent subproblems where one subproblem is for optimizing T and the other N_{as} subproblems are for optimizing $f_i, \forall i \in \mathcal{N}_{as}$. The subproblem for optimizing T can be formulated as

$$\min_{T} \ T + \sum_{i=1}^{N_{as}} \frac{\tau_i I_i F_i}{T-\varpi_i} \tag{44a}$$
$$s.t. \ T \ge \varpi_i, \forall i \in \mathcal{N}_{as}. \tag{44b}$$

By setting the first-order derivative of (44a) with respect to T to zero, we observe that the optimal T should satisfy

$$T = \left\{ T | \sum_{i=1}^{N_{as}} \frac{\tau_i I_i F_i}{(T-\varpi_i)^2} = 1, T \ge \varpi_i \right\}, \tag{45}$$

which can be solved by applying the bisection search method.

Moreover, the subproblem for optimizing $f_i, \forall i \in \mathcal{N}_{as}$ can be formulated as

$$\min_{f_i} \ \delta\kappa f_i^3 + \xi f_i - \tau_i f_i \tag{46a}$$
$$s.t. \ f_i \ge 0, \forall i \in \mathcal{N}_{as}. \tag{46b}$$

By setting the first-order derivative of (46a) with respect to f_i to zero, we obtain the closed-form expression of the optimal computing capacity allocation as

$$f_i = \left[\sqrt{\frac{\tau_i - \xi}{3\delta\kappa}}\right]^+. \tag{47}$$

The value of dual variables $\boldsymbol{\tau}$, δ, and ξ can be determined by the sub-gradient method. The updating procedure can be given by

$$\tau_i = \left[\tau_i + \phi\left(\frac{I_i F_i}{T-\varpi_i} - f_i\right)\right]^+, \forall i \in \mathcal{N}_{as} \tag{48a}$$
$$\delta = \left[\delta + \phi\left(\sum_{i=1}^{N_{as}} \kappa f_i^3 - p_{max} + p_u\right)\right]^+ \tag{48b}$$
$$\xi = \left[\xi + \phi\left(\sum_{i=1}^{N_{as}} f_i - f_{max}\right)\right]^+, \tag{48c}$$

where $\phi > 0$ is the step-size in each iteration.

We summarize the procedures for obtaining the optimal solution to Problem (40) in Algorithm 3. Similar to the complexity analysis in Subsection III-B, we note that the total complexity of Algorithm 3 is $\mathcal{O}(L_5 L_6 N_{as}^2 \log_2(1/\epsilon))$, where L_5 is the number of iterations for outer layer in Algorithm 3 and L_6 is the number of iterations via the dual method of solving Problem (40).

Algorithm 3 Computing Capacity Allocation Algorithm for Solving Problem (40).

1: Initialize $\boldsymbol{\tau}$, δ, and ξ.
2: **repeat**
3: Obtain the optimal T and $\boldsymbol{f}$ by solving (45) and (47), respectively;
4: Update the Lagrangian multipliers $\boldsymbol{\tau}$, δ, and ξ by solving (48a), (48b) and (48c), respectively;
5: **until** The objective function in (40a) converges.

E. Caching Policy Subproblem

For any given $\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}$, $\mathbf{f}$, and $\mathbf{a}$, the caching policy of Problem (15) can be optimized by solving the following problem

$$\min_{\mathbf{c},T} \quad T \tag{49a}$$

$$\text{s.t.} \quad \varrho_i + \frac{(1-c_i)I_i}{r_i^b} \le T, \forall i \in \mathcal{N} \tag{49b}$$

$$\sum_{i=1}^{N} c_i I_i \le c_{max} \tag{49c}$$

$$c_i = \{0,1\}, \forall i \in \mathcal{N}, \tag{49d}$$

where $\varrho_i = a_i \left(\frac{O_i}{r_i} + \frac{I_i F_i}{f_i} \right) + (1-a_i) \left(\frac{I_i F_i}{f_i^{local}} + \frac{I_i}{r_i} \right)$. Due to the linearity of the objective function and all constraints, we note that Problem (49) is a binary linear programming.

We first analyze the ideal scenario where the U-BS has a sufficiently large storage capability, i.e., $c_{max} \ge \sum_{i=1}^{N} I_i$. In this case, since a lower maximum latency T might be achieved with a higher c_i according to (49b), we can easily obtain that the optimal solution for Problem (49) is $c_i = 1, \forall i \in \mathcal{N}$, i.e., the U-BS has cached the input data requested by all VR users and all requested input data can be directly obtained from the cache container of the U-BS without the backhaul transmissions, which significantly reduces the latency by eliminating the backhaul latency for all VR users.

For the general scenario where $c_{max} \le \sum_{i=1}^{N} I_i$, due to the limited storage capability of the U-BS, only specific data which is requested by the VR users with high latency consumption will be pre-cached so that the maximum latency can be reduced via caching. Thus, to minimize the maximum latency T, we first sort the users based on the descending order in terms of $\varrho_i + \frac{I_i}{r_i^b}$. Next, we consider the input data required by the user with a higher $\varrho_i + \frac{I_i}{r_i^b}$ will be cached at the U-BS with higher priority until the caching constraint cannot be satisfied. To derive the closed-form solution, we define a new indicator set $\mathcal{S} \triangleq \{s_1, s_2, \cdots, s_N\}$ which is sorted in a descending order in terms of $\varrho_i + \frac{I_i}{r_i^b}$, i.e., $s_1 = arg\max_{\{\forall i \in \mathcal{N}\}} \varrho_i + \frac{I_i}{r_i^b}$ and $s_N = arg\min_{\{\forall i \in \mathcal{N}\}} \varrho_i + \frac{I_i}{r_i^b}$. We further define the set $\mathcal{S}_0 \triangleq \{s_1, s_2, \cdot \cdot \cdot, s_{m-1}\}, m = \min\{m: \sum_{i=1}^{m} I_{s_i} > c_{max}\}$. By following [28], a closed-form expression for the optimal solution of (49) is given as

$$c_i = \begin{cases} 0, \text{if } c_{max} \le \sum_{i=1}^{N} I_i \text{ and } i \notin \mathcal{S}_0 \\ 1, \text{otherwise.} \end{cases} \tag{50}$$

We note that the complexity for the caching policy subproblem is upper bounded by $\mathcal{O}(N)$ and the actual complexity may be much smaller than this upper bound since the proposed approach may terminate when the caching storage constraint cannot be satisfied.

F. Computing Policy Subproblem

For any given $\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}$, $\mathbf{f}$, and $\mathbf{c}$, the computing policy Problem (15) can be optimized by solving the following problem

$$\min_{\mathbf{a},T} \quad T \tag{51a}$$

$$\text{s.t.} \quad a_i \upsilon_i + o_i \le T, \forall i \in \mathcal{N} \tag{51b}$$

$$\sum_{i=1}^{N} a_i f_i \le f_{max} \tag{51c}$$

$$p_u + \sum_{i=1}^{N} a_i \kappa f_i^3 \le p_{max} \tag{51d}$$

$$a_i = \{0,1\}, \forall i \in \mathcal{N}, \tag{51e}$$

where $\upsilon_i = \frac{O_i}{r_i} + \frac{I_i F_i}{f_i} - \frac{I_i F_i}{f_i^{local}} - \frac{I_i}{r_i}$ and $o_i = \frac{I_i F_i}{f_i^{local}} + \frac{I_i}{r_i} + \frac{(1-c_i)I_i}{r_i^b}$. We note that Problem (51) is a binary linear programming since the objective function and all constraints are linear.

To solve Problem (51), we first analyze the scenario when $\upsilon_i \ge 0$. In this case, we observe that a larger a_i might result in a higher T according to (51b). Thus, to minimize T, we can easily derive that $a_i = 0$ when $\upsilon_i \ge 0$. This corresponds to the scenario that when the transmission and computing latencies at the i-th user of local computing is less than that of U-BS processing, the VR user chooses to self-execute its input data to reduce latency.

When $\upsilon_i \le 0$, a lower maximum latency consumption might be achieved with a larger a_i according to (51b). However, due to the computing capacity and power constraints, only specific input data which is requested by the VR users with higher latency consumption will be processed at the U-BS so that the maximum latency can be minimized benefiting from the higher computing capacity at the U-BS. Thus, to minimize the maximum latency T, we first sort the users based on the descending order in terms of o_i. Next, we consider the input data of the user with a higher o_i will be processed at the U-BS with higher priority until the computing capacity constraint or the power constraint cannot be satisfied. To derive the closed-form solution, we define a new indicator set $\mathcal{K} \triangleq \{k_1, k_2, \cdot \cdot \cdot, k_N\}$ which is sorted in a descending order in terms of o_i, i.e., $k_1 = arg\max_{\{\forall i \in \mathcal{N}\}} o_i$ and $k_N = arg\min_{\{\forall i \in \mathcal{N}\}} o_i$. We further define the set $\mathcal{K}_0 \triangleq \{k_1, k_2, \cdot \cdot \cdot, k_{l-1}\}, l = \min\{l_1, l_2\}$ where $l_1 = \min\{l_1 : \sum_{i=1}^{l_1} f_{k_i} > f_{max}\}$ and $l_2 = \min\{l_2 : p_u + \sum_{i=1}^{l_2} \kappa f_{k_i}^3 > p_{max}\}$. Similar to the caching policy subproblem, a closed-form optimal solution of Problem (51) with complexity of $\mathcal{O}(N)$ is given as

$$a_i = \begin{cases} 1, \text{if } \upsilon_i \le 0 \text{ and } i \in \mathcal{K}_0 \\ 0, \text{otherwise.} \end{cases} \tag{52}$$

G. Proposed Iterative Algorithm

The iterative procedure for solving Problem (15) is summarized in Algorithm 4, where the U-BS location, fronthaul and backhaul bandwidth allocation, computing capacity allocation, data caching policy and computing policy are successively optimized while keeping the other variables fixed until convergence, and the suboptimal solutions to Problem (15) can be obtained. In addition, the derived solution in each iteration will be applied as the input for the next iteration. We note that for U-BS location subproblem, since we only solve the

Algorithm 4 Proposed Iterative Optimization for Problem (15).

1: Initialize $m=0$, $\mathbf{y}[m]$, $\boldsymbol{\theta}[m]$,$\boldsymbol{\eta}[m]$, $\mathbf{f}[m]$, $\mathbf{c}[m]$, $\mathbf{a}[m]$.
2: **repeat**
3: Given $\{\boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]\}$, find the optimal U-BS location $\mathbf{y}[m+1]$ by solving (20);
4: Given $\{\mathbf{y}[m+1], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]\}$, find the optimal fronthaul bandwidth allocation $\boldsymbol{\theta}[m+1]$ according to Algorithm 1;
5: Given $\{\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]\}$, find the optimal backhaul bandwidth allocation $\boldsymbol{\eta}[m+1]$ according to Algorithm 2;
6: Given $\{\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m+1], \mathbf{c}[m], \mathbf{a}[m]\}$, find the optimal computing capacity allocation $\mathbf{f}[m+1]$ according to Algorithm 3;
7: Given $\{\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m+1], \mathbf{f}[m+1], \mathbf{a}[m]\}$, find the optimal caching policy $\mathbf{c}[m+1]$ according to (50);
8: Given $\{\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m+1], \mathbf{f}[m+1], \mathbf{c}[m+1]\}$, find the optimal computing policy $\mathbf{a}[m+1]$ according to (52);
9: Update $m=m+1$;
10: **until** convergence.

approximated subproblem optimally, the convergence analysis for this subproblem should be studied.

Denote $T_{UBS}^{ub}(\mathbf{y}[m], \boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m])$ as the objective values of (16). First, in step 3 of Algorithm 4, since the first-order Taylor expansions in (18) and (19) are tight bounds at given local point $\mathbf{y}[m]$ for the original U-BS location subproblem (16), we have

$$\begin{aligned} &T(\mathbf{y}[m], \boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]) \\ =&T_{UBS}^{ub}(\mathbf{y}[m], \boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]). \end{aligned} \tag{53}$$

Notice that the U-BS location solution $\mathbf{y}[m+1]$ for Problem (20) is optimal with other variables fixed, then it follows that

$$\begin{aligned} &T_{UBS}^{ub}(\mathbf{y}[m], \boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]) \\ \geq&T_{UBS}^{ub}(\mathbf{y}[m+1], \boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]) \\ \geq&T(\mathbf{y}[m+1], \boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]), \end{aligned} \tag{54}$$

where the last inequality holds due to the fact that the objective value of Problem (20) is the upper bound of that of its original problem (16). Next, in step 4-8, since we solve the fronthaul and backhaul bandwidth allocation, computing capacity allocation, data caching policy and computing policy optimally, we have

$$\begin{aligned} &T(\mathbf{y}[m+1], \boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]) \\ \geq&T(\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]) \\ \geq&T(\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m+1], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]) \\ \geq&T(\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m+1], \mathbf{f}[m+1], \mathbf{c}[m], \mathbf{a}[m]) \\ \geq&T(\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m+1], \mathbf{f}[m+1], \mathbf{c}[m+1], \mathbf{a}[m]) \\ \geq&T(\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m+1], \mathbf{f}[m+1], \mathbf{c}[m+1], \mathbf{a}[m+1]). \end{aligned} \tag{55}$$

According to (53)-(55), we can conclude that

$$\begin{aligned} &T(\mathbf{y}[m], \boldsymbol{\theta}[m], \boldsymbol{\eta}[m], \mathbf{f}[m], \mathbf{c}[m], \mathbf{a}[m]) \\ \geq&T(\mathbf{y}[m+1], \boldsymbol{\theta}[m+1], \boldsymbol{\eta}[m+1], \mathbf{f}[m+1], \mathbf{c}[m+1], \mathbf{a}[m+1]), \end{aligned} \tag{56}$$

which shows that the algorithm yields a non-increasing sequence of the objective value. In addition, the objective value is lower bounded by zero. Hence, our proposed algorithm is guaranteed to converge. Although the obtained solution is generally suboptimal, we validate the effectiveness of our proposed Algorithm 4 in reducing the latency consumption via simulation results by comparing it with other benchmark strategies in Section IV. We note that the complexity of Algorithm 4 is the addition of the complexity in each step [20]. According to the aforementioned complexity analysis of each subproblem, we obtain that the overall complexity of Algorithm 4 is $\mathcal{O}(L_1 L_2 N^2 \log_2(1/\epsilon) + L_3 L_4 N^2 \log_2(1/\epsilon) + L_5 L_6 N_{as}^2 \log_2(1/\epsilon) + 2N)$, which shows that the complexity of Algorithm 4 is polynomial in the worst scenario.

IV. Simulation Results

In this section, numerical results are presented to evaluate the performance of our proposed algorithm. We consider $N=6$ VR users that are randomly and uniformly distributed within a 400m $\times$ 400m square area. We set the altitude of U-BS as $H_u = 150$ m. The channel power gain is set as $\beta_0 = 10^{-5}$. We set the effective switched capacitance at the UAV as $\kappa = 10^{-27}$ [21]. The noise spectral density is $\sigma^2 = -169$ dBm/Hz. The transmit powers at the U-BS and the cloud server are $p_u = p_c = 0.5$ W. We consider the input data size I_i follows a uniform distribution with $I_i \sim U[10, 15]$ KB, the ratio between O_i and I_i is set as $\alpha = 2$, and the required number of CPU cycles per bit is distributed as $F_i \sim U[500, 800]$ cycles/bit. The computing capacity of VR users follows a distribution of $f_i^{local} \sim U[0.5, 1]$ GHz. The maximum computing capacity, power budgets and caching storage of U-BS are set as $f_{max} = 5$ GHz, $P_{max} = 4$ W, and $c_{max} = 60$ KB, respectively. The fronthaul and backhaul bandwidth are $B = B_{back} = 1$ MHz.

Fig. 2 shows the convergence behavior of Algorithm 4 with different U-BS altitude H_u. This figure shows that our proposed algorithm quickly converges within 8 iterations. Moreover, we observe that compared to its initial value, the maximum latency reduces by 63.7% from 47.6 ms to 17.3 ms when $H_u = 150$ m, which verifies the effectiveness of our proposed solution.

Fig. 3 shows the initial latency and optimized latency of each VR user where the initial latency is generated based on one set of random realization of input data, computing capacity of each VR user and required number of CPU cycles per bit. It can be seen that our proposed joint optimization solution significantly reduces the maximum latency consumption among all VR users by comparing the initial latency and optimized latency. Moreover, we can see that the optimized latency of each VR user is almost equal, which shows that minimizing the maximum latency among all VR users is equivalent to guaranteeing the fairness among all VR users, so that the minimum quality-of-service can be improved.

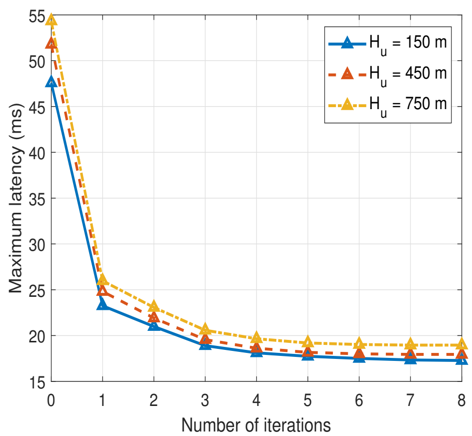


Fig. 2. Maximum latency versus number of iterations with different U-BS altitude H_u.

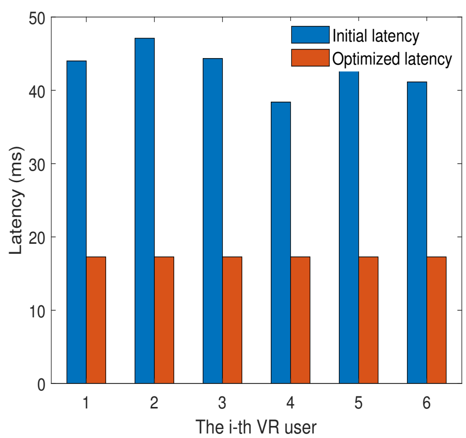


Fig. 3. The initial latency and optimized latency of each VR user.

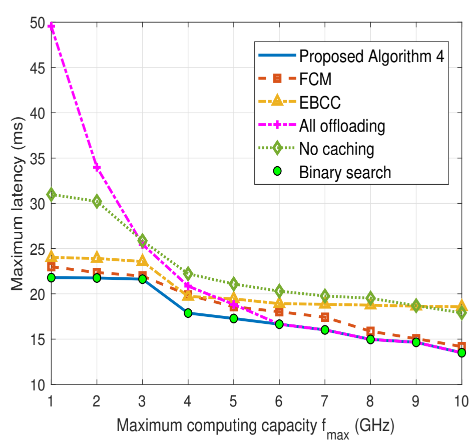


Fig. 4. Maximum latency as a function of maximum computing capacity f_{max}.

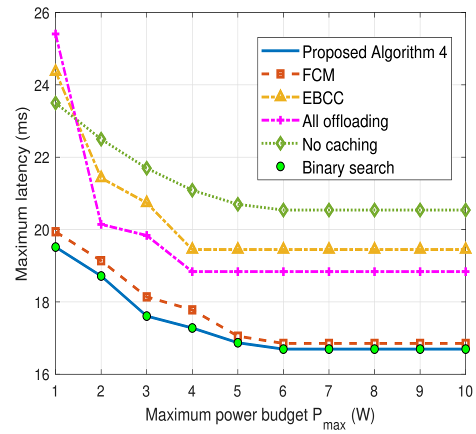


Fig. 5. Maximum latency as a function of maximum power budget P_{max}.

In Fig. 4, we plot the maximum latency as a function of maximum computing capacity f_{max}. We compare our proposed Algorithm 4 with the following five benchmark schemes: 1) Fuzzy C-Means Clustering algorithm (FCM) [20]: the U-BS location is optimized based on FCM algorithm and all the other variables are optimized by using Algorithm 4; 2) Equal bandwidth and computing capacity (EBCC): We set $\theta_i = 1/N, f_i = a(i) * \min\left(f_{max}/N_{as}, ((P_{max} - p_u)/N_{as}/\kappa)^{1/3}\right), \eta_i = (1 - c(i))/N_{uncached}, \forall i \in \mathcal{N}$ and all the other variables are optimized by using Algorithm 4; 3) No caching: We set $c_{max} = 0$ KB and all the other variables are optimized by using Algorithm 4; 4) All offloading: We set $a_i = 1, \forall i \in \mathcal{N}$ and all the other variables are optimized by using Algorithm 4; 5) Binary search: We solve the caching policy and computing policy subproblems by applying the binary search method and all the other variables are optimized by using Algorithm 4. It can be seen that compared to the "Binary search" scheme which has an exponential complexity of $\mathcal{O}(L_1L_2N^2\log_2(1/\epsilon) + L_3L_4N^2\log_2(1/\epsilon) + L_5L_6N_{as}^2\log_2(1/\epsilon) + 2^{N+1})$, our proposed Algorithm 4 with polynomial complexity achieves the same latency performance, which indicates that our proposed algorithm is stable and computationally efficient. Moreover, Fig. 4 shows that our proposed Algorithm 4 achieves a lower latency compared to other benchmark schemes except the "Binary search" scheme. Interestingly, we find that the performance gap between Algorithm 4 and "All offloading" scheme is significant when f_{max} is low, while it reduces to 0 when f_{max} is greater than 6 GHz. This is because when f_{max} is limited, e.g., $f_{max} = 1$ GHz, each VR user chooses to project the input data locally to reduce latency and the maximum latency is dominated by the computing latency which occupies 87%. While when $f_{max} \geq 6$ GHz, the input data requested by all VR users will be processed at the U-BS, resulting in the same performance as "All offloading" scheme.

Fig. 5 shows the maximum latency as a function of maximum power budget P_{max}. It can be seen that our proposed Algorithm 4 achieves the same latency performance compared to the "Binary search" scheme and outperforms all the other baseline solutions. Moreover, we find that the maximum latency first decreases then keeps unchanged with an increasing P_{max}. This is because when P_{max} is limited, increasing P_{max} increases the computing resource allocated to process the input data requested by offloading VR users, resulting in a lower computing latency. However, when P_{max} is sufficient, the computing resource allocation is bounded by the maximum computing capacity, which makes the latency unchanged. In addition, we observe that caching is helpful to reduce the latency.

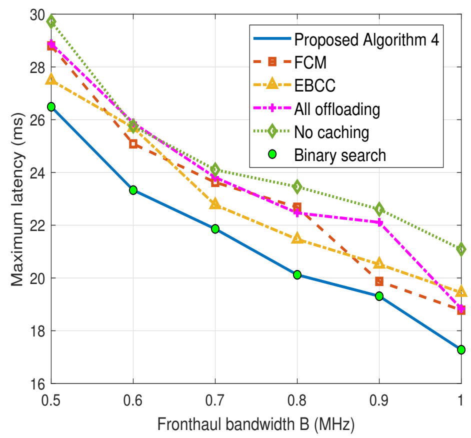


Fig. 6. Maximum latency as a function of fronthaul bandwidth B.

In Fig. 6, we plot the maximum latency as a function of fronthaul bandwidth B. We observe that our proposed Algorithm 4 achieves the same latency performance compared to the "Binary search" scheme and outperforms all the other baseline. Moreover, we observe that when B is limited, e.g., $B = 0.5$ MHz, the maximum latency can be up to 26.5 ms and it mainly comes from the transmission latency, which occupies 57%. While when $B = 1$ MHz, the maximum latency is 17.3 ms and the portion of transmission latency reduces to 27%.

V. Conclusions

In this paper, we have presented the maximum latency minimization problem for a UAV-enabled communication, computing and caching VR delivery network. Specifically, we have jointly optimized the the U-BS location, fronthaul and backhaul bandwidth allocation, computing capacity allocation, caching and computing policies. To solve this nonconvex optimization problem, we have proposed an efficient iterative algorithm by applying the block coordinate descent method, the successive convex approximation technique and Lagrangian dual decomposition method. Simulation results demonstrated that our proposed algorithm significantly reduces the latency compared to benchmark schemes. Moreover, it can be seen that the maximum latency is mainly due to the transmission latency when the bandwidth is limited, whereas it is dominated by the computing latency when the computing resource is low. In addition, we showed that caching is helpful to reduce latency. We note that our work can be extended to consider that each VR user will execute multiple computation tasks to address the impact of queueing delay and consider the use of multiple UAVs to address limited battery capacity. Moreover, the extension to a more practical scenario that different VR users may require the same input data would be an interesting future research direction which results in a more-complex optimization problem with multiple possible computing and caching strategies.

References

[1] M. Chen, W. Saad, C. Yin and M. Debbah, "Data Correlation-Aware Resource Management in Wireless Virtual Reality (VR): An Echo State Transfer Learning Approach," in *IEEE Trans. Commun.,* vol. 67, no. 6, pp. 4267-4280, June 2019.

[2] M. Chen, W. Saad and C. Yin, "Virtual Reality Over Wireless Networks: Quality-of-Service Model and Learning-Based Resource Management," in *IEEE Trans. Commun.,* vol. 66, no. 11, pp. 5621-5635, Nov. 2018.

[3] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward Low-Latency and Ultra-Reliable Virtual Reality," in *IEEE Network,* vol. 32, no. 2, pp. 78-84, March-April 2018.

[4] E. Bastug, M. Bennis, M. Medard, and M. Debbah, "Toward Interconnected Virtual Reality: Opportunities, Challenges, and Enablers," in *IEEE Commun. Mag.,* vol. 55, no. 6, pp. 110-117, June 2017.

[5] X. Yang, Z. Chen, K. Li, Y. Sun, N. Liu, W. Xie, and Y. Zhao, "Communication-Constrained Mobile Edge Computing Systems for Wireless Virtual Reality: Scheduling and Tradeoff," in *IEEE Access,* vol. 6, pp. 16665-16677, 2018.

[6] J. Park, P. Popovski, and O. Simeone, "Minimizing Latency to Support VR Social Interactions Over Wireless Cellular Systems via Bandwidth Allocation," in *IEEE Wireless Commun. Lett.* vol. 7, no. 5, pp. 776-779, Oct. 2018.

[7] A. Al-Shuwaili and O. Simeone, "Energy-Efficient Resource Allocation for Mobile Edge Computing-Based Augmented Reality Applications," in *IEEE Wireless Commun. Lett.,* vol. 6, no. 3, pp. 398-401, June 2017.

[8] Y. Dai, D. Xu, S. Maharjan, and Y. Zhang, "Joint computing offloading and user association in multi-task mobile edge computing," in *IEEE Trans. Veh. Technol.,* vol. 67, no. 12, pp. 12313–12325, Dec 2018.

[9] Z. Xiang, F. Gabriel, E. Urbano, G. T. Nguyen, M. Reisslein and F. H. P. Fitzek, "Reducing Latency in Virtual Machines: Enabling Tactile Internet for Human-Machine Co-Working," in *IEEE J. Sel. Areas Commun.,* vol. 37, no. 5, pp. 1098-1116, May 2019.

[10] P. Wang, Z. Zheng, B. Di and L. Song, "HetMEC: Latency-Optimal Task Assignment and Resource Allocation for Heterogeneous Multi-Layer Mobile Edge Computing," in *IEEE Trans. Wireless Commun.,* vol. 18, no. 10, pp. 4942-4956, Oct. 2019.

[11] Y. Zhou, P. L. Yeoh, C. Pan, K. Wang, M. Elkashlan, Z. Wang, B. Vucetic, and Y. Li, "Offloading Optimization for Low-Latency Secure Mobile Edge Computing Systems," in *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 480-484, April 2020.

[12] T. Dang and M. Peng, "Joint Radio Communication, Caching, and Computing Design for Mobile Virtual Reality Delivery in Fog Radio Access Networks," in *IEEE J. Sel. Areas Commun.,* vol. 37, no. 7, pp. 1594-1607, July 2019.

[13] Y. Sun, Z. Chen, M. Tao and H. Liu, "Communications, Caching, and Computing for Mobile Virtual Reality: Modeling and Tradeoff," in *IEEE Trans. Commun.,* vol. 67, no. 11, pp. 7573-7586, Nov. 2019.

[14] J. Zhang, X. Hu, Z. Ning, E. C. -H. Ngai, L. Zhou, J. Wei, J. Cheng, B. Hu, and V. C. M. Leung, "Joint Resource Allocation for Latency-Sensitive Services Over Mobile Edge Computing Networks With Caching," in *IEEE Internet Things J.,* vol. 6, no. 3, pp. 4283-4294, June 2019.

[15] Q. Wu, Y. Zeng, and R. Zhang, "Joint Trajectory and Communication Design for Multi-UAV Enabled Wireless Networks," in *IEEE Trans. Wireless Commun.,* vol. 17, no. 3, pp. 2109-2121, March 2018.

[16] Y. Zhou, P. L. Yeoh, H. Chen, Y. Li, R. Schober, L. Zhuo, and B. Vucetic, "Improving physical layer security via a UAV friendly jammer for unknown eavesdropper location," in *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11280-11284, Nov 2018.

[17] C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint blocklength and location optimization for URLLC-enabled UAV relay systems," in *IEEE Commun. Lett.,* vol. 23, no. 3, pp. 498-501, March 2019.

[18] Z. Yang, C. Pan, M. Shikh-Bahaei, W. Xu, M. Chen, M. Elkashlan, and A. Nallanathan, "Joint altitude, beamwidth, location, and bandwidth optimization for UAV-enabled communications," in *IEEE Commun. Lett.,* vol. 22, no. 8, pp. 1716-1719, Aug 2018.

[19] Y. Zhou, C. Pan, P. L. Yeoh, K. Wang, M. Elkashlan, B. Vucetic, and Y. Li, “Secure Communications for UAV-Enabled Mobile Edge Computing Systems,” in *IEEE Trans. Commun.,* vol. 68, no. 1, pp. 376-388, Jan. 2020.

[20] Z. Yang, C. Pan, K. Wang, and M. Shikh-Bahaei, “Energy Efficient Resource Allocation in UAV-Enabled Mobile Edge Computing Networks,” in *IEEE Trans. Wireless Commun.,* vol. 18, no. 9, pp. 4576-4589, Sept. 2019.

[21] Q. Hu, Y. Cai, G. Yu, Z. Qin, M. Zhao, and G. Y. Li, “Joint offloading and trajectory design for UAV-enabled mobile edge computing systems,” in *IEEE Internet Things J.,* vol. 6, no. 2, pp. 1879-1892, April 2019.

[22] N. Zhao, F. Cheng, F. R. Yu, J. Tang, Y. Chen, G. Gui, and H. Sari, “Caching UAV Assisted Secure Transmission in Hyper-Dense Networks Based on Interference Alignment,” in *IEEE Trans. Commun.,* vol. 66, no. 5, pp. 2281-2294, May 2018.

[23] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, “Caching in the Sky: Proactive Deployment of Cache-Enabled Unmanned Aerial Vehicles for Optimized Quality-of-Experience,” in *IEEE J. Sel. Areas Commun.,* vol. 35, no. 5, pp. 1046-1061, May 2017.

[24] X. Xu, Y. Zeng, Y. L. Guan, and R. Zhang, “Overcoming Endurance Issue: UAV-Enabled Communications With Proactive Caching,” in *IEEE J. Sel. Areas Commun.,* vol. 36, no. 6, pp. 1231-1244, June 2018.

[25] Q. Wu and R. Zhang, “Common Throughput Maximization in UAV-Enabled OFDMA Systems With Delay Consideration,” in *IEEE Trans. Commun.,* vol. 66, no. 12, pp. 6614-6627, Dec. 2018.

[26] Y. Zeng, X. Xu, and R. Zhang, “Trajectory Design for Completion Time Minimization in UAV-Enabled Multicasting,” in *IEEE Trans. Wireless Commun.,* vol. 17, no. 4, pp. 2233-2246, April 2018.

[27] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[28] J. Dai, Z. Hu, B. Li, J. Liu, and B. Li, “Collaborative hierarchical caching with dynamic request routing for massive content distribution,” in *Proc. IEEE INFOCOM,* pp. 2444-2452, 2012.

Yi Zhou received her Ph.D. degree from School of Engineering and Information Technologies, University of Sydney, Australia, in 2020. Her research interests include physical layer security, UAV communications, and 5G related communications. She was a recipient of the Postgraduate Scholarship and the Norman I. Price Scholarship from the Center of Excellence in Telecommunications, School of Electrical and Information Engineering, University of Sydney.

Cunhua Pan received the B.S. and Ph.D. degrees from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2010 and 2015, respectively. From 2015 to 2016, he was a Research Associate at the University of Kent, U.K. He held a post-doctoral position at Queen Mary University of London, U.K., from 2016 and 2019, where he is currently a Lecturer.

His research interests mainly include intelligent reflection surface (IRS), ultra-reliable low latency communication (URLLC) , machine learning, UAV, Internet of Things, and mobile edge computing. He serves as a TPC member for numerous conferences, such as ICC and GLOBECOM, and the Student Travel Grant Chair for ICC 2019. He also serves as an Editor of IEEE Wireless Communication Letters, IEEE Communications Letters and IEEE ACCESS.

Phee Lep Yeoh (S'08-M'12) received the B.E. degree with University Medal and the Ph.D. degree from the University of Sydney (USYD), Australia, in 2004 and 2012, respectively. From 2008 to 2012, he was with the Telecommunications Laboratory at the University of Sydney and the Wireless and Networking Technologies Laboratory at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. From 2012 to 2016, he was with the Department of Electrical and Electronic Engineering at the University of Melbourne, Australia. Since 2016, he has been a Senior Lecturer in the School of Electrical and Information Engineering at USYD.

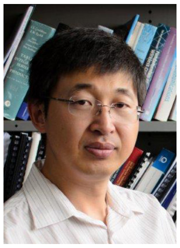
Dr. Yeoh is a recipient of the 2020 USYD Robinson Fellowship, the 2017 Alexander von Humboldt Research Fellowship for Experienced Researchers, and the 2014 Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA). He has served as TPC chair for the 2016 Australian Communications Theory Workshop (AusCTW) and TPC member for IEEE GLOBECOM, ICC, and VTC conferences. He has received best paper awards at IEEE ICC 2014 and IEEE VTC-Spring 2013, and best student paper awards at AusCTW 2013 and 2019. His current research interests include secure wireless communications, ultra-reliable and low-latency communications (URLLC), ultra-dense networks, and multiscale molecular communications.

Kezhi Wang received his B.E. and M.E. degrees in School of Automation from Chongqing University, China, in 2008 and 2011, respectively. He received his Ph.D. degree in Engineering from the University of Warwick, U.K. in 2015. He was a Senior Research Officer in University of Essex, U.K. Currently he is a Senior Lecturer in Department of Computer and Information Sciences at Northumbria University, U.K. His research interests include wireless communication, mobile edge computing, UAV communication and machine learning.

Maged Elkashlan (M'06-SM'20) received the Ph.D. degree in Electrical Engineering from the University of British Columbia, Canada, 2006. From 2007 to 2011, he was with the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. During this time, he held visiting appointments at University of New South Wales and University of Technology Sydney. In 2011, he joined the School of Electronic Engineering and Computer Science at Queen Mary University of London, UK. His research interests fall into the broad areas of communication theory and statistical signal processing.

Dr. Elkashlan is an Editor of the IEEE Transactions on Vehicular Technology and the IEEE Transactions on Molecular, Biological and Multi-scale Communications. He received the Best Paper Awards at the IEEE International Conference on Communications (ICC) in 2016 and 2014, the International Conference on Communications and Networking in China (CHINACOM) in 2014, and the IEEE Vehicular Technology Conference (VTC-Spring) in 2013.

Branka Vucetic (SM'00-F'03) is an ARC Laureate Fellow and Director of the Centre of Excellence for IoT and Telecommunications at the University of Sydney. Her current work is in the areas of wireless networks and Internet of Things. In the area of wireless networks, she works on communication system design for millimetre wave (mmWave) frequency bands. In the area of the Internet of things, she works on providing wireless connectivity for mission critical applications. She is a Fellow of the Australian Academy of Science, IEEE, the Australian Academy of Technological Sciences and Engineering and the Engineers Australia.



Yonghui Li (M'04-SM'09-F'19) received his PhD degree in November 2002 from Beijing University of Aeronautics and Astronautics. From 1999 to 2003, he was affiliated with Linkair Communication Inc, where he held a position of project manager with responsibility for the design of physical layer solutions for the LAS-CDMA system. Since 2003, he has been with the Centre of Excellence in Telecommunications, the University of Sydney, Australia. He is now a Professor in School of Electrical and Information Engineering, University of Sydney. He is the recipient of the Australian Queen Elizabeth II Fellowship in 2008 and the Australian Future Fellowship in 2012.

His current research interests are in the area of wireless communications, with a particular focus on MIMO, millimeter wave communications, machine to machine communications, coding techniques and cooperative communications. He holds a number of patents granted and pending in these fields. He is now an editor for IEEE Transactions on Communications and IEEE Transactions on Vehicular Technology. He was also the guest editor for IEEE JSAC special issue on Millimeter Wave Communications for Future Mobile Networks. He received the best paper awards from IEEE International Conference on Communications (ICC) 2014, IEEE PIMRC 2017 and IEEE Wireless Days Conferences (WD) 2014. He is Fellow of IEEE.