Reliability Enhancement for VR Delivery in Mobile-Edge Empowered Dual-Connectivity Sub-6 GHz and mmWave HetNets

Zhuojia Gu, Hancheng Lu, Peilin Hong, and Yongdong Zhang

Abstract

The reliability of current virtual reality (VR) delivery is low due to the limited resources on VR head-mounted displays (HMDs) and the transmission rate bottleneck of sub-6 GHz networks. In this paper, we propose a dual-connectivity sub-6 GHz and mmWave heterogeneous network architecture empowered by mobile edge capability. The core idea of the proposed architecture is to utilize the complementary advantages of sub-6 GHz links and mmWave links to conduct a collaborative edge resource design, which aims to improve the reliability of VR delivery. From the perspective of stochastic geometry, we analyze the reliability of VR delivery and theoretically demonstrate that sub-6 GHz links can be used to enhance the reliability of VR delivery despite the large mmWave bandwidth. Based on our analytical work, we formulate a joint caching and computing optimization problem with the goal to maximize the reliability of VR delivery. By analyzing the coupling caching and computing strategies at HMDs, sub-6 GHz and mmWave base stations (BSs), we further transform the problem into a multiple-choice multi-dimension knapsack problem. A best-first branch and bound algorithm and a difference of convex programming algorithm are proposed to obtain the optimal and sub-optimal solution, respectively. Numerical simulations demonstrate the performance improvement using the proposed algorithms, and reveal that caching more monocular videos at sub-6 GHz BSs and more stereoscopic videos at mmWave BSs can improve the VR delivery reliability efficiently.

Index Terms

Virtual reality, sub-6 GHz and mmWave heterogeneous networks, reliability enhancement, mobile edge computing, stochastic geometry.

Zhuojia Gu, Hancheng Lu, Peilin Hong are with CAS Key Laboratory of Wireless-Optical Communications, University of Science and Technology of China, Hefei 230027, China (email: guzj@mail.ustc.edu.cn; hclu@ustc.edu.cn; plhong@ustc.edu.cn).

Yongdong Zhang is with the Department of Electronic Engineering and Information Science (EEIS), University of Science and Technology of China, Hefei 230027, China (email: zhyd73@ustc.edu.cn).

I. INTRODUCTION

In recent years, the interest of virtual reality (VR) in academia and industry has been unprecedented [1]. To achieve the immersive experience, the most important task is to increase the resolution of VR applications to the resolution of human eyes [2]. It is impractical to cache all rendered VR videos locally at VR head-mounted displays (HMDs) in advance, especially in the scenarios where user interactions are required. Moreover, users prefer to experience VR videos anytime and anywhere, compared with being tied down by wired cables. Therefore, VR videos are expected to be delivered real-time over wireless networks. However, the bandwidth required for delivering VR videos is usually 4-5 times the bandwidth required for delivering conventional high-definition videos [3], which puts tremendous pressure on the network bandwidth. On the other hand, VR applications are typical ultra-reliable low-latency communication (URLLC) applications, and the end-to-end delay exceeding 20 ms can cause dizziness of users [4]. Thus, it is essential to ensure the high reliability of VR delivery networks, which means that ensuring more data packets delivered to HMDs within the latency requirement of VR applications for a satisfactory VR viewing experience. In this regard, two fundamental problems for VR delivery should be addressed: *1) How to enhance the reliability of VR delivery when delivering a large amount of video data over wireless networks,* and *2) how to enhance the reliability of VR delivery when performing time-consuming projecting and rendering of raw VR viewpoints?*

For the first problem, using the current sub-6 GHz network is limited by the bottleneck of wireless bandwidth and cannot meet the URLLC requirements of VR applications. The sufficient spectrum resources brought by millimeter wave (mmWave) communication are considered to be the key enablers of 5G applications. Sufficient bandwidth resources make it possible to transmit large-capacity VR videos in real time. However, mmWave signals are prone to be blocked and suffer severe fading. This poses a great challenge to VR applications, because users may frequently experience blockage caused by buildings, human bodies, and environmental facilities. To tackle the problem, dual-connectivity (DC) sub-6 GHz and mmWave heterogeneous networks (HetNets) are promising for enhancing the reliability of VR delivery. DC is an appealing technique for performance enhancement in wireless networks, and it allows the user to have two simultaneous connections and utilize both radio resources [5]. Inherently, DC sub-6 GHz and mmWave HetNets utilize the complementary advantages of the wide signal coverage in sub-6 GHz networks and sufficient spectrum resources in mmWave networks, which is suitable for

reliability enhancement of VR delivery.

For the second problem, computation offloading is seen as a key enabler to provide the required video projection rendering capabilities for VR delivery. Edge computing servers are suitable for performing high CPU- and GPU-intensive computing tasks. By providing efficient computing resources close to users, mobile edge computing (MEC) strikes a balance between communication delay and computing delay. Specifically, the HMD can upload the tracking information (e.g., game actions or viewpoint preferences) to the MEC server to offload computing tasks. In this case, the MEC server uses the computing resources to project monocular videos (MVs) into stereoscopic videos (SVs), and sends the downlink SVs to the HMD. In addition, the caching capability of MEC servers can help save the projection and rendering time.

Based on the above discussion, in this paper, we propose a mobile-edge empowered DC sub-6 GHz and mmWave HetNet architecture, and aim to conduct a collaborative design of edge network resources to enhance the reliability of VR delivery.

A. Related Work

1) Reliability Enhancement for Wireless Transmission: Ultra-reliable communication has become the vital support of mission-critical 5G applications. Mei *et al.* [6] proposed a reliability guaranteed resource allocation scheme in vehicle-to-vehicle (V2V) networks. Guo *et al.* [7] extended the research to perform a reliability-aware resource allocation in V2V networks based only on large-scale fading channel information. Popovski *et al.* [8] analyzed the fundamental tradeoffs between ultra-reliability and some other metrics (i.e., latency, bandwidth occupancy and energy consumption) for wireless access. Recently, the joint application of sub-6 GHz and mmWave in HetNets has been proposed as an attractive solution to improve the network performance [9]. Semiari *et al.* [10], [11] introduced DC mode into sub-6 GHz and mmWave HetNets to ensure data transmission reliability while considering the user association and scheduling. DC implements carrier aggregation between sites through base station coordination to realize diversified association, which can reduce the handover failure and wireless link failure, thus improving the transmission reliability. DC mode has been extended to multi-connectivity (MC) mode in [12], [13], which enables simultaneous connections of multiple air interfaces such as cellular, device-to-device (D2D) and WiFi links to further support ultra-reliable applications.

2) DC-assisted mobile edge computing: Some prior works proposed to use DC technology for mobile edge computing to achieve energy-efficient computation offloading performance. Guo *et*

al. [14] studied the problem of computation offloading for multi-access mobile edge computing in ultra-dense networks to reduce the overall energy consumption. Guo *et al.* [15] also extended the scenario in [14] to the energy-constrained Internet of Things (IoT) devices with DC capability to tackle the conflict between resource-hungry mobile applications and energy-constrained IoT devices. Authors in [16], [17] investigated DC-assisted computation offloading scheme in non-orthogonal multiple access system to minimize the total energy consumption. Note that these works mainly focused on utilizing DC-assisted mobile edge computing to achieve green-oriented computation offloading for data services. However, the benefits of both edge caching and the complementary sub-6 GHz and mmWave bands to the delay-sensitive VR application in the DC-assisted mobile edge computing scenario have not been investigated.

3) Mobile Edge Computing for VR Delivery: Authors in [2], [18], [19] inspired the use of MEC resources for VR delivery to obtain the potential performance gain, but they did not establish theoretical formulation and provided efficient algorithms. Some efficient task offloading algorithms in wireless cellular networks with MEC were proposed in [20]–[24]. The MEC task offloading technology was introduced for the delivery of VR video in [25]–[27]. Sun *et al.* [25] focused on balancing the local and edge computing resources to maximize the average bandwidth usage for VR delivery. Dang *et al.* [26] extended the scenario in [25] to an edge fog computing network to achieve an economical computing offloading scheme that minimized the average delay for VR delivery. Chakareski *et al.* [27] proposed an edge server cooperation framework to efficiently offload the local computing tasks. It is worth noting that [25]–[27] focused on improving the average bandwidth usage or delivery latency in MEC empowered VR delivery networks. Nevertheless, little work has focused on improving the reliability of VR delivery.

B. Motivation and Contributions

As presented in existing related works, wireless channels and connections have a significant impact on the reliability of data transmission. In the context of VR delivery, these wireless factors have not been well studied. First, the transmission delay and the reliability of wireless VR delivery will be randomly affected by the path loss, fading, and signal blockage of wireless channels. However, the impact of wireless channel fluctuation on the performance of VR delivery is largely overlooked in the existing literatures. For example, the authors in [26] simplified the characteristics of wireless channels as a deterministic value of the minimum allowable transmission rate of each VR viewpoint, without considering the randomness of wireless channels.

To this end, some probabilistic analysis is required to more accurately characterize the impact of wireless channel fluctuation on the transmission delay and the reliability of wireless VR delivery. Second, although the dual-connectivity sub-6 GHz and mmWave technology has been proved to improve the reliability of data transmission, existing works on VR delivery mostly assumed single-connectivity (i.e., one wireless interface) for a HMD (e.g., see [20]–[27]). These works mainly focused on the task offloading of VR video projection and rendering in the MEC scenario. Nevertheless, DC is considered as a promising technology for realizing URLLC applications, while the resource coordination for VR delivery in DC sub-6 GHz and mmWave HetNets with edge caching and computing capability is challenging when the interface diversity is implemented. On the one hand, it is essential to decide whether to deliver the viewpoint over sub-6 GHz links or mmWave links when the requested viewpoint is not cached locally. This should not only consider the influence of the wireless channel conditions, but also consider whether the requested viewpoint is cached at sub-6 GHz BSs (μBSs) or mmWave BSs (mBSs), as well as the impact of the limited computation resources of HMDs, μBSs and mBSs on the computation delay. On the other hand, a proper resource coordination between sub-6 GHz networks and mmWave networks is vital for adapting the computation-intensive VR videos. Specifically, local cache can save the transmission delay of viewpoints at the cost of occupying limited local caching capacity. Caching simultaneously at μBSs and mBSs can reduce the transmission delay of some popular viewpoints with a higher probability, but at the cost of occupying more caching capacity of BSs compared with caching at either μBSs or mBSs. In addition, the caching and computing strategies in DC sub-6 GHz and mmWave HetNets are coupled. If MVs are cached at μBSs or mBSs, the caching capacity can be saved, but the computation time for projection is increased. While caching SVs save the computation time at the cost of more caching capacity.

To address the aforementioned issues, in this paper, we utilize tools from stochastic geometry to model the DC sub-6 GHz and mmWave channels and theoretically analyze the reliability of VR delivery by comprehensively considering the influence of communication, caching, and computation (3C). We then perform the resource coordination and jointly optimize the caching and computing strategy, aiming to ensure higher reliability of VR delivery. The main contributions of this paper are summarized as follows:

- We present a DC sub-6 GHz and mmWave HetNet architecture empowered by edge caching and computing capability, aiming to meet the requirement of ultra reliability of VR delivery. Based on this architecture, we conduct a collaborative design of caching and computing

resource utilization that adapts to the complementary advantages of sub-6 GHz and mmWave links by utilizing a stochastic geometry analytical framework.

- We derive closed-form expressions for the reliability of VR delivery in the DC sub-6 GHz and mmWave HetNet. We propose a link selection strategy based on the minimum-delay delivery, and derive the VR delivery probability over sub-6 GHz links and mmWave links. The relationship between the delivery probability and the difference of caching and computing strategy in sub-6 GHz and mmWave tiers is provided. We theoretically demonstrate that sub-6 GHz links can be used to enhance the reliability of VR delivery despite the large mmWave bandwidth.

- Based on the analytical work, the coupling 3C resource allocation can be formulated as a joint caching and computing optimization problem, which is designed for the feasibility of practical implementation. By analyzing the coupling relationship between caching and computing strategies at HMDs, μBSs and mBSs, we transform the problem into a multiple-choice multi-dimension knapsack problem (MMKP). We propose a best-first branch and bound algorithm (BFBB) by calculating the upper and lower bounds of the MMKP to obtain the optimal solution. To further reduce the complex of the algorithm, the problem is transformed into a continuous optimization problem, and the difference of convex programming technique is utilized to obtain a sub-optimal solution.

- We conduct numerical evaluation to validate the theoretical analysis of reliability with respect to several key parameters, such as blockage density, CPU cycles, and cache size of μBSs/mBSs/HMDs. Numerical simulations show great promise of the proposed DC sub-6 GHz and mmWave HetNet architecture to enhance the reliability of VR delivery. The results also reveal that it is more advantageous to cache MVs at μBSs and cache SVs at mBSs for enhancing the reliability efficiently.

The rest of this paper is organized as follows. We first introduce the system model in Section II. We then analyze the reliability of VR delivery over DC sub-6 GHz and mmWave HetNets and provide the link selection strategy to formulate the joint caching and computing problem in Section III. In Section IV, we first provide the BFBB algorithm based on the computation of upper and lower bounds of the optimal solution, and a difference of convex programming algorithm with lower complexity is proposed to maximize the reliability of VR delivery. Numerical simulations are provided in Section V, and Section VI concludes this paper with summary.
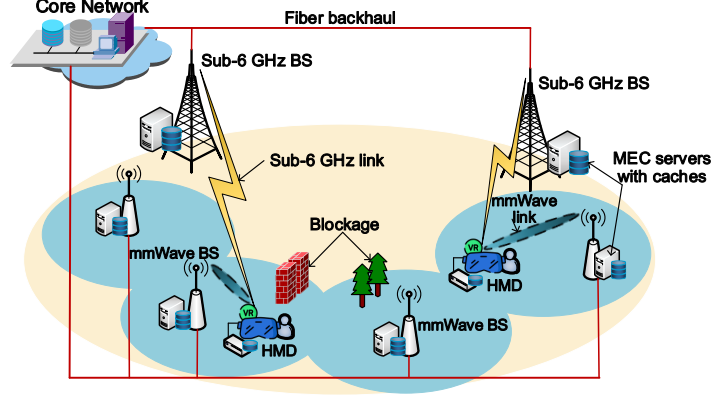
Fig. 1. Dual-connectivity sub-6 GHz and mmWave heterogeneous network architecture empowered by mobile edge capability.

II. SYSTEM MODEL

A. Network Model

As illustrated in Fig. 1, we consider the VR delivery in a two-tier dual-connectivity (DC) sub-6 GHz and mmWave HetNet. Similar to [28]–[30], the locations of sub-6 GHz BSs (μBSs), mmWave BSs (mBSs) and HMDs are modeled as three independent and homogeneous Poisson Point Processes (PPPs)[1], which are denoted by Φ_μ, Φ_m, Φ_H of densities λ_μ, λ_m, λ_H, respectively. DC is implemented at HMDs by performing packet data convergence protocol (PDCP) layer integration [32], so a HMD can be simultaneously connected to a μBS and a mBS. Instead of delivering $360°$ VR videos, μBSs and mBSs only deliver the requested MVs and SVs. As mentioned above, MVs are projected and rendered using the computing resources at BSs or HMDs to create SVs. The set of all viewpoints is denoted by $\mathcal{J} = \{1, 2, \cdots, J\}$. Each viewpoint $j \in \mathcal{J}$ corresponds to an MV and an SV. The data sizes of the j-th MV and SV are denoted by d_j^M and d_j^S, respectively. Note that d_j^S is at least twice larger than d_j^M, i.e., $d_j^S/d_j^M \geq 2$, due to the creation of stereoscopic video.

In the sub-6 GHz tier, all channels undergo independent identically distributed (i.i.d.) Rayleigh fading. Without loss of generality, when a HMD located at the origin o requests the j-th viewpoint from the associated μBS, its received signal-to-interference-plus-noise ratio (SINR) is given by,

$$\Upsilon_j^\mu = \frac{P_\mu h_j^\mu r^{-\alpha_\mu}}{I_j^\mu + \sigma_\mu^2},$$

(1)

---

[1]PPPs are generally used for modeling the scenarios that do not consider hot-spot areas (e.g., in rural areas). If considering hot-spot areas, the Poisson cluster process (PCP) can be a more realistic model which captures the clustering nature of base stations and users in hot-spot areas [31]. Modeling the locations of base stations as PCPs will have an impact on the outage and coverage performance of the HetNet, which will be considered in our future work.

where P_μ is the transmit power of μBS, h_j^μ is the Rayleigh channel gain between the HMD and its serving μBS which follows the exponential distribution. $r^{-\alpha_\mu}$ is the path loss with the distance r, where α_μ is the path loss exponent. $I_j^\mu = \sum_{n \in \Phi_\mu \setminus b_\mu} P_\mu h_j^{\mu,n} r^{-\alpha_\mu,n}$ denotes the inter-cell interference, where b_μ denotes the associated μBS. σ_μ^2 is the noise power of a sub-6 GHz link.

In the mmWave tier, unlike the conventional sub-6 GHz counterpart, mmWave transmissions are highly sensitive to blockage. We adopt a two-state statistical blockage model for each mmWave link as in [33], such that the probability of the link to be LOS or NLOS is a function of the distance between the HMD and its serving mBS. Assume that the distance between them is r, then the probability that a link of length r is LOS or NLOS can be modeled as

$$\rho_{\mathrm{L}}(r) = \mathrm{e}^{-\kappa r}, \ \rho_{\mathrm{N}}(r) = 1 - \mathrm{e}^{-\kappa r}, \tag{2}$$

respectively, where κ is a parameter determined by the density and the average size of the blockages [34]. Under dual-connectivity mode, we consider the blockage effects for the mmWave tier by using the defined LOS/NLOS probability function. Assume that the antenna arrays at mBSs and HMDs perform directional beamforming with the main lobe directed towards the dominant propagation path and having less radiant energy in other directions. For tractability in the analysis, we adopt a sectorial antenna pattern [35]. Denote θ as the main lobe beamwidth, and M and m as the directivity gain of main and side lobes, respectively. Then the random antenna gain/interference G between the mBS and the HMD has 3 patterns with different probabilities, which is given as

$$G = \begin{cases} M^2, & \text{with prob. } (\dfrac{\theta}{2\pi})^2, \\ Mm, & \text{with prob. } \dfrac{2\theta(2\pi - \theta)}{(2\pi)^2}, \\ m^2, & \text{with prob. } (\dfrac{2\pi - \theta}{2\pi})^2. \end{cases} \tag{3}$$

Independent Nakagami fading is assumed for each link. Parameters of Nakagami fading N_{L} and N_{N} are assumed for LOS and NLOS links, respectively. Therefore, when the HMD requests the j-th viewpoint from its associated mBS, the received SINR is given by

$$\Upsilon_j^m = \frac{P_m h_j^m G r^{-\alpha_m}}{I_j^m + \sigma_m^2}, \tag{4}$$

where P_m is the transmit power of the mBS, h_j^m is the Nakagami channel fading which follows Gamma distribution. The path loss exponent $\alpha_m = \alpha_{\mathrm{L}}$ when it is a LOS link and $\alpha_m = \alpha_{\mathrm{N}}$ when

it is an NLOS link. $I_j^m = \sum_{n \in \Phi_m \backslash b_m} P_m h_j^{m,n} r^{-\alpha_{m,n}}$, where b_m denotes the associated mBS. σ_m^2 is the noise power of a mmWave link. The rate of sub-6 GHz and mmWave links are given by the Shannon's formula as $R_j^l = B_l \log_2(1 + \Upsilon_j^l), l \in \{\mu, m\}$, where B_l denotes the subchannel bandwidth of sub-6 GHz or mmWave links.

B. Caching and Computing Model

The probability of the service request of the HMD for the j-th viewpoint is $p_j, j \in \mathcal{J}$. Considering all J viewpoints, we have $\sum_{j=1}^{J} p_j = 1$. μBSs, mBSs, and HMDs all have caching and computing capabilities. The cache size at μBSs, mBSs, and HMDs is C^μ, C^m, and C^H, respectively. The caching decision for the MV and SV of the j-th viewpoint is denoted by $y_j^{q,M} \in \{0,1\}$ and $y_j^{q,S} \in \{0,1\}, q \in \{\mu, m, H\}$. $y_j^{q,\omega} = 1, \omega \in \{M, S\}$ indicates that the MV or SV of the j-th viewpoint is cached at device q, otherwise $y_j^{q,\omega} = 0$. With the cache size constraint, we have $\sum_{j=1}^{J} d_j^M y_j^{q,M} + d_j^S y_j^{q,S} \leq C^q$.

The computing decisions at μBSs, mBSs, and HMDs are considered to process the projection and rendering. The CPU-cycle frequency of μBSs, mBSs, and HMDs is denoted by $f_q, q \in \{\mu, m, H\}$. The average energy consumption constraint of μBSs, mBSs, and HMDs is denoted by $E_q, q \in \{\mu, m, H\}$. Define ε as the number of computation cycles required to process the projection and rendering of one bit input.

According to [36]–[38], the energy consumption of a CPU cycle can be expressed as $k_q = \eta_q f_q^2$, where f_q is the CPU-cycle frequency of device q, and η_q is a constant related to the hardware architecture of device q. The computing decision for the j-th viewpoint is denoted by $z_j^q \in \{0,1\}$, where $z_j^q = 1$ indicates that the projection is computed at device q, otherwise $z_j^q = 0$. Then, the average energy consumption of a device for processing a computing task of a viewpoint can be calculated by averaging all J viewpoints, which can be expressed as $\sum_{j=1}^{J} p_j \varepsilon k_q d_j^M z_j^q$. Considering that the HMD is energy-constrained, the average energy consumption of the device for processing a computing task of a viewpoint should be limited, otherwise the onboard battery will be depleted quickly, and the device will heat up due to the overload of the CPU, which will degrade the user experience. Therefore, a constraint of the average energy consumption of the device $\sum_{j=1}^{J} p_j \varepsilon k_q d_j^M z_j^q \leq E^q$ should be satisfied, which can ensure that the CPU of the device will not be overloaded.

In the case where the requested viewpoint cannot be found in HMD cache, μBS cache or mBS cache, the requested viewpoint is retrieved from the core network through the fiber backhaul links with an extra backhaul retrieving delay τ_j^r. We assume that the backhaul transmission rate

TABLE I
SUMMARY OF NOTATIONS

Notation	Description	Notation	Description
Φ_μ / Φ_m / Φ_H	PPP of μBSs / mBSs / HMDs	λ_μ / λ_m / λ_H	density of μBSs / mBSs / HMDs
P_μ / P_m	Transmit power of μBSs / mBSs	B_μ / B_m	Bandwidth for each user at sub-6 GHz / mmWave
α_L / α_N	Path loss exponent of LOS and NLOS	N_L / N_N	Nakagami fading parameter for LOS / NLOS link
h_j^μ / h_j^m	Channel fading of sub-6 GHz / mmWave links	M / m	Mainlobe antenna gain / sidelobe antenna gain
Υ_j^μ / Υ_j^m	Received SINR at the HMD from μBSs / mBSs	R_j^μ / R_j^m	Data rate of sub-6 GHz / mmWave links
θ	Mainlobe beamwidth	κ	Blockage density
$\mathcal{J}$	Set of viewpoints	J	The number of viewpoints
C^μ / C^m / C^H	The cache size at μBSs / mBSs / HMDs	d_j^M / d_j^S	Size of MVs / SVs
p_j	Request probability of the j-th viewpoint	δ	Skewness of the viewpoint popularity
f_μ / f_m / f_H	CPU cycle of μBSs / mBSs / HMDs	η_μ / η_m / η_H	Energy efficiency coefficient of HMDs / μBSs / mBSs
T_j	End-to-end delay threshold of the j-th viewpoint	τ_j^μ / τ_j^m	End-to-end latency of the j-th viewpoint over the sub-6 GHz / mmWave link
A_j^μ / A_j^m	Probability of the sub-6 GHz / mmWave link being selected to deliver the j-th viewpoint	$\rho_L(r)$ / $\rho_N(r)$	LOS / NLOS probability of mmWave links with length r
ε	The number of computation cycles required for 1 bit input	x_{jk}	Joint caching and computing decision of the j-th viewpoint
$y_j^{q,M}$ / $y_j^{q,S}$	Caching decision for the MV / SV of the j-th viewpoint at device q	z_j^q	Computing decision of the j-th viewpoint at device q
$\mathcal{R}_j^l$	Reliability of delivering the j-th viewpoint over sub-6 GHz / mmWave links	$\mathcal{R}_j$	Reliability of delivering the j-th viewpoint in DC sub-6 GHz and mmWave HetNets
$\tau_j^{l,t}$ / $\tau_j^{l,c}$ / $\tau_j^{l,b}$, $l \in \{\mu, m\}$	Transmission / computing / backhaul retrieving delay of the j-th viewpoint over the sub-6 GHz / mmWave link	ξ_{jk}^q / ζ_{jk}^q / φ_{jk}	Caching occupancy / computing energy consumption / backhaul cost of the j-th viewpoint for the k-th strategy

is R_j^b, and the backhaul capacity constraint is B^b. The notations used in Sec. II to Sec. IV are summarized in Table I.

III. RELIABILITY ANALYSIS OF VR DELIVERY AND PROBLEM FORMULATION

In order to investigate the reliability performance of VR delivery in DC sub-6 GHz and mmWave HetNets, we refer to the reliability defined by 3GPP [39], which is the probability of experiencing an end-to-end latency below the threshold required by the targeted service. Accordingly, using the law of total probability, the reliability of VR delivery can be expressed as,

$$\mathcal{R} = \sum_{j=1}^{J} p_j \mathbb{P}[\tau_j^\mu < T_j \cup \tau_j^m < T_j], \qquad (5)$$

where T_j is the end-to-end latency threshold[2] required by the j-th viewpoint, $\tau_j^l, l \in \{\mu, m\}$ denotes the end-to-end latency when the j-th viewpoint is retrieved over a sub-6 GHz link or mmWave link. Delay contributions to the end-to-end VR delivery latency include the over-the-air transmission delay $\tau_j^{l,t}$, the computing delay $\tau_j^{l,c}$, the sensor sampling delay τ_j^s, the display refresh delay τ_j^d and the backhaul retrieving delay $\tau_j^{l,b}$ if the requested viewpoint is not cached

---

[2]Instead of assuming a pre-determined minimum allowable transmission rate as in [26], we assume a pre-determined end-to-end delay threshold for VR viewpoints, which is a more recognized and practical indicator to ensure the user experience of VR videos, typically no more than 20 ms [4].

[2]. Thus, when the j-th viewpoint is delivered over a sub-6 GHz link or mmWave link, the end-to-end delay is calculated as

$$\tau_j^l = \tau_j^{l,t} + \tau_j^{l,c} + \tau_j^{l,b} + \tau_j^s + \tau_j^d, j \in \mathcal{J}, l \in \{\mu, m\}. \tag{6}$$

Note that $\tau_j^{l,t}$ is a random variable which is affected by the channel uncertainty of wireless environments and the data size of the delivered viewpoint. The value of $\tau_j^{l,c}$ can be calculated when the caching and computing strategy is determined. Whether the end-to-end delay contains $\tau_j^{l,b}$ depends on whether the requested viewpoint is cached. τ_j^s and τ_j^d are assumed to be constants.

A. Reliability Analysis over DC sub-6 GHz and mmWave HetNets

The reliability of VR delivery is mainly affected by the channel uncertainty of wireless environments. In wireless environments where temporary outages are common due to impairments in SINR, VR's non-elastic traffic behavior poses additional difficulty. In this subsection, we utilize the statistical wireless channel state information for sub-6 GHz links and mmWave links described in Sec. II-A to analyze the reliability of VR delivery. Utilizing tools from stochastic geometry, tractable expressions can be obtained to characterize the reliability of VR delivery.

The computing delay $\tau_j^{l,c} = \frac{\varepsilon d_j^M}{f_q}, q \in \{\mu, m, H\}$ when the j-th MV is calculated into SV at device q. Considering the caching and computing strategy for the j-th viewpoint, the computing delay $\tau_j^{l,c}$ can be expressed as

$$\tau_j^{l,c} = \frac{\varepsilon d_j^M (y_j^{H,M} \| y_j^{l,M})(z_j^H \| z_j^l)}{z_j^{l,M} f_l + (1 - z_j^{l,M}) f_H}, \tag{7}$$

where $\cdot \| \cdot$ is the logical OR operator. And the backhaul retrieving delay $\tau_j^{l,b}$ can be expressed as

$$\tau_j^{l,b} = (1 - y_j^{l,M} \| y_j^{l,S}) \tau_j^r. \tag{8}$$

Note that the transmission delay over sub-6 GHz links or mmWave links is affected by the wireless channel fluctuation. Thus, the transmission delay is a random variable related to the distance between communicating nodes, the channel fading and the blockage probability for mmWave links. When the computing delay $\tau_j^{l,c}$ and the backhaul retrieving delay $\tau_j^{l,b}$ are obtained under a given caching and computing strategy, we can obtain the transmission delay threshold $T_j^{l,t}$ for the j-th viewpoint, which is defined as

$$T_j^{l,t} = T_j - \tau_j^{l,c} - \tau_j^{l,b}. \tag{9}$$

Then the calculation of the reliability of VR delivery can be transformed into the calculation of the probability that the transmission delay is less than the threshold $T_j^{l,t}$.

The data size of the j-th delivered viewpoint over wireless links also depends on the caching and computing strategy, which can be calculated as

$$D_j^l = y_j^{l,M} z_j^l d_j^S + y_j^{l,M}(1 - z_j^l)d_j^M + y_j^{l,S} d_j^S + (1 - y_j^{l,M})(1 - y_j^{l,S})d_j^S, \tag{10}$$

Then, the reliability of delivering the j-th viewpoint over sub-6 GHz or mmWave links is a function of D_j^l and $T_j^{l,t}$, which can be expressed as

$$\mathcal{R}_j^l(D_j^l, T_j^{l,t}) = \mathbb{P}[\tau_j^l < T_j] = \mathbb{P}[\tau_j^{l,t} < T_j^{l,t}] \overset{(a)}{=} \mathbb{P}[R_j^l > D_j^\omega/T_j^{l,t}], l \in \{\mu, m\}, \tag{11}$$

where (a) follows from $\tau_j^{l,t} = D_j^l/R_j^{l,t}$, $D_j^l \in \{d_j^M, d_j^S\}$.

Proposition 1. *When the j-th viewpoint is delivered to the HMD over sub-6 GHz links, the reliability of VR delivery is given as,*

$$\mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t}) = \sum_{i=1}^q w_i \mathrm{e}^{r_i + \beta_\mu(\nu_j^\mu, r_i)} f_\mu(r_i), \tag{12}$$

where $w_i = \frac{r_i}{(q+1)^2[L_{q+1}(r_i)]^2}$, r_i is the i-th zero of $L_q(r)$, $L_q(r)$ denotes the Laguerre polynomials, and q is a parameter balancing the accuracy and complexity. $\nu_j^\mu = 2^{\frac{D_j^\mu}{T_j^{\mu,t} B_\mu}} - 1$, $f_\mu(r) = 2\pi\lambda_\mu r \mathrm{e}^{-\pi\lambda_\mu r^2}$, $\beta_\mu(\nu_j^\mu, r) = -\nu_j^\mu r^{\alpha_\mu}\sigma_\mu^2 - \pi\lambda_\mu r^2 H_\delta(\nu_j^\mu) + \pi\lambda_\mu s^\delta\Gamma(1 + \delta)\Gamma(1 - \delta)$, $s = \nu_j^\mu r^{\alpha_\mu} P_\mu^{-1}$, $\delta = 2/\alpha_\mu$, $\Gamma(\cdot)$ is the Gamma function, and $H_\delta(x) \triangleq {}_2F_1(1, \delta; 1 + \delta; -x)$ is the Gauss hypergeometric function.

Proof. Please refer to Appendix A. □

The delivery reliability (12) is in the form of the complementary cumulative distribution function (CCDF) of SINR over the sub-6 GHz tier. The reliability is monotonically decreasing with the SINR threshold ν_j^μ, which indicates that the reliability increases with the increase of $T_j^{\mu,t}$, while decreases with the increase of D_j^μ.

Proposition 2. *When the j-th viewpoint is delivered to the HMD over mmWave links, the reliability of VR delivery is given as,*

$$\mathcal{R}_j^m(D_j^m, T_j^{m,t}) = \sum_{i=1}^q w_i \mathrm{e}^{r_i} \beta_m(\nu_j^m, r_i) f_m(r_i), \tag{13}$$

where $\beta_m(\nu_j^m, r_i) = \sum_{\ell \in \{L,N\}} \rho_\ell(r_i) \sum_{k=1}^{N_\ell} (-1)^{k+1} \binom{N_\ell}{k} e^{-\frac{k\eta_\ell \nu_j^m r_i^{\alpha_\ell} \sigma_m^2}{P_m G}} \mathcal{L}(r_i)$, $\nu_j^m = 2^{\frac{D_j^m}{T_j^{m,t} B_m}} - 1$, $f_m(r) = 2\pi \lambda_m r e^{-\pi \lambda_m r^2}$, $\eta_\ell = N_\ell(N_\ell!)^{-\frac{1}{N_\ell}}$, *and*

$$\mathcal{L}(r) = \prod_{n \in \{L,N\}} \prod_G \exp\left[-2\pi \lambda_m p_G \sum_{u=1}^{N_n} \binom{N_n}{u} \frac{r^{-\frac{1}{\alpha_n}\left(u - \frac{2}{\alpha_n}\right)}}{u\alpha_n - 2} {}_2F_1\left(N_n, u - \frac{2}{\alpha_n}; 1 + u - \frac{2}{\alpha_n}; -sr^{-\frac{1}{\alpha_n}}\right) \right], \quad (14)$$

where p_G *is the probability of the random antenna gain defined in (3), and* $s = \frac{k\eta_n G \nu_j^m r^{\alpha_n} \sigma_m^2}{M^2 N_n}$.

Proof. Please refer to Appendix B. □

Taking into consideration the LOS or NLOS channel state, the delivery reliability (13) is the CCDF of SINR over the mmWave tier. Then some remarks can be concluded from Proposition 2.

Remark 1: The reliability of VR delivery over the mmWave tier (13) indicates that the Laplace transform of the interference $\mathcal{L}(r)$ in (14) is independent of the transmit power P_m.

Remark 2: (13) is a monotonically increasing function with respect to $T_j^{m,t}$, and a monotonically decreasing function with respect to D_j^m.

Remark 3: Considering a special case when the viewpoints are only delivered through LOS links, then (13) is a decreasing and convex function with respect to the blockage density κ.

For DC sub-6 GHz and mmWave network, according to the addition rule for probability, the delivery reliability for viewpoint j is calculated as

$$\mathcal{R}_j^{DC} = \mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t}) + \mathcal{R}_j^m(D_j^m, T_j^{m,t}) - \mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t})\mathcal{R}_j^m(D_j^m, T_j^{m,t}), \quad (15)$$

Note that $\mathcal{R}_j^{DC}$ is the expression of the reliability when the j-th viewpoint is delivered through wireless links. On the other hand, if the MV or SV of the j-th viewpoint is cached in the HMD, we assume that the end-to-end delay is less than the delay threshold with probability 1, i.e., the reliability is set to 1. Therefore, the complete expression of the reliability can be written as

$$\mathcal{R}_j = \mathbf{1}(y_j^{H,M} \| y_j^{H,S} = 1) + \mathbf{1}(y_j^{H,M} + y_j^{H,S} = 0)\mathcal{R}_j^{DC}, \quad (16)$$

where $\mathbf{1}(\cdot)$ is the indicator function.

B. Link Selection Strategy and Problem Formulation

Since the DC mode is adopted at HMD, it is essential to give the criteria for delivery the viewpoints over the sub-6 GHz link or the mmWave link. In this paper, since the end-to-end

delay is the key factor affecting the reliability of VR delivery, we propose the minimum-delay delivery criteria in DC mode as the link selection strategy, which can be written as

$$l_0 = \arg \min_{l \in \{\mu, m\}} \tau_j^l. \tag{17}$$

Proposition 3. *When the j-th viewpoint is not locally cached by the HMD, based on the minimum-delay delivery criteria, the probability that the mmWave link is selected to deliver the viewpoint is given as,*

$$A_j^m = \int_0^\infty p_{\mathcal{R}_j^\mu(D_j^\mu, t)} \mathcal{R}_j^m(D_j^m, (t - \tau_0)^+) \mathrm{d}t, \tag{18}$$

where $p_{\mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t})} = \frac{\partial \mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t})}{\partial T_j^{\mu,t}}$, $\tau_0 = \tau_j^{m,c} + \tau_j^{m,b} - \tau_j^{\mu,c} - \tau_j^{\mu,b}$, and $(\cdot)^+ = \max(\cdot, 0)$.

Proof. According to (17), the delivery probability over mmWave links can be expressed as

$$A_j^m = \mathbb{P}[\tau_j^m < \tau_j^\mu] = \mathbb{P}[\tau_j^{m,t} < \tau_j^{\mu,t} - \tau_0], \tag{19}$$

where $\tau_0 = \tau_j^{m,c} + \tau_j^{m,b} - \tau_j^{\mu,c} - \tau_j^{\mu,b}$. Using the results in Propositions 1 and 2, the cumulative distribution function of $\tau_j^{m,t}$ is directly obtained in (13), and denote the probability density function of (12) with respect to $T_j^{\mu,t}$ as $p_{\mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t})} \triangleq \frac{\partial \mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t})}{\partial T_j^{\mu,t}}$, we have

$$A_j^m = \int_0^\infty p_{\mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t})} \mathcal{R}_j^m(D_j^m, (T_j^{\mu,t} - \tau_0)^+) \mathrm{d}T_j^{\mu,t}. \tag{20}$$

$\square$

Accordingly, the probability that the mmWave link is selected to deliver the viewpoint is given as $A_j^m = 1 - A_j^\mu$. Proposition 3 can be used to calculate the expected amount of transmitted data over the sub-6 GHz link and the mmWave link. Thus, the expected number of MVs computed at the μBS, mBS, and the HMD can be obtained, respectively.

Note that τ_0 represents the difference of the delays (except for the transmission delays) between mmWave links and sub-6 GHz links, where $\tau_j^{l,c}, \tau_j^{l,b}, l \in \{\mu, m\}$ can be calculated by (7) and (8), respectively. From Remark 2, we can deduce that A_j^m is decreasing with τ_0, which indicates that the caching and computing strategy will have an impact on the delivery probability over mmWave and sub-6 GHz links. For example, if the blockage density in the mmWave tier is low, we can adjust the caching and computing strategy to get a smaller value of τ_0, thus increasing the delivery probability over mmWave links to improve the reliability of VR delivery.

Remark 4: From Remark 3 and the delivery probability over mmWave links (18), we conclude

that when the viewpoints are only delivered through LOS links, A_j^m is a decreasing and convex function with respect to the blockage density κ. In addition, when $\kappa \to \infty$, we have $A_j^m \to 0$, and when $\kappa = 0$, $A_j^m < 1$. This indicates that mmWave links cannot be utilized when the blockage density is too high, while sub-6 GHz links can be utilized even if there is no blockage for mmWave links. In other words, sub-6 GHz links can be utilized to enhanced the reliability of VR delivery despite the large mmWave bandwidth.

Next, we can formulate the problem of maximizing the reliability of VR delivery in DC sub-6 GHz and mmWave HetNets by jointly optimizing the caching and computing decision at the μBS, mBS, and HMD as follow,

$$\textbf{P1:} \quad \max_{\{y_j^{q,M}, y_j^{q,S}, z_j^q\}} \quad \sum_{j=1}^{J} p_j \mathcal{R}_j \tag{21a}$$

$$\text{s.t.} \quad \sum_{j=1}^{J} d_j^M y_j^{q,M} + d_j^S y_j^{q,S} \leq C^q, q \in \{H, \mu, m\}, \tag{21b}$$

$$\sum_{j=1}^{J} p_j \varepsilon k_H d_j^M (y_j^{H,M} + A_j^{\mu,M} y_j^{\mu,M} + A_j^{m,M} y_j^{m,M}) z_j^H \leq E^H \tag{21c}$$

$$\sum_{j=1}^{J} p_j \varepsilon A_j^l k_l d_j^M z_j^l \leq E^l, l \in \{\mu, m\}, \tag{21d}$$

$$\sum_{j=1}^{J} \sum_l A_j^l d_j^S (1 - y_j^{H,M} \| y_j^{H,S} \| y_j^{l,M} \| y_j^{l,S}) \leq B^b, j \in \mathcal{J}, l \in \{\mu, m\}, \tag{21e}$$

$$y_j^{q,M} \in \{0,1\}, y_j^{q,S} \in \{0,1\}, z_j^q \in \{0,1\}, j \in \mathcal{J}, q \in \{H, \mu, m\}, \tag{21f}$$

where (21a) is the reliability of VR delivery according to (5). (21b) is the cache capacity constraint. (21c) is the average computing energy consumption at the HMD. (21d) is the average computing energy consumption at the μBS or mBS. (21e) is the backhaul bandwidth constraint, and (21f) ensures the binary decision of the caching and computing decision.

Remark 5: Note that the coupling 3C resources are jointly considered in MEC-enabled networks for performance improvement, though the communication part is not added as an optimization variable in Problem **P1**. Specifically, the communication strategy A_j^l is affected by the caching and computing strategies reflected by τ_0 as shown in Eq. (18), while the caching and computing strategies in the formulated Problem **P1** are affected by the communication strategy A_j^l involved in constraints (21c)–(21e). From the perspective of practical implementation, the caching placement phase is prior to the viewpoint delivery phase (e.g., caching placement is performed during off-peak time), so it is impractical to perform a joint communication, caching and computing optimization. Instead, the caching and computing strategies are optimized based

on the probabilistic communication decision derived in Proposition 3, then the caching and computing decisions can be determined by solving Problem **P1**, which is a practical solution for 3C resource optimization.

C. Problem Reformulation

Problem **P1** is difficult to solve because the objective function (21a) and the constraint (21e) contain logical operations. Fortunately, we can explore the coupling relationship between the caching and computing strategies to greatly reduce the solution space.

Lemma 1. *For any viewpoint $j \in \mathcal{J}$, the following caching and computing relationships hold,*
1) $(y_j^{H,M} \| y_j^{H,S}) + (y_j^{\mu,M} \| y_j^{\mu,S} \| y_j^{m,M} \| y_j^{m,S}) \leq 1$,
2) $y_j^{q,M} + y_j^{q,S} \leq 1$, $q \in \{\mu, m, H\}$,
3) *if $y_j^{l,M} = 1$, then $z_j^l + z_j^H = 1$ holds, $l \in \{\mu, m\}$.*

Proof. 1) means that the HMD and the MEC server (at the μBS or mBS) should not cache the j-th viewpoint simultaneously. This is because when the j-th viewpoint is cached at the HMD, the reliability is set to 1, thus there is no need to cache it at the MEC server.

2) means that there is no need to cache the MV and the SV of the j-th viewpoint simultaneously at the HMD or the MEC server. This is because the caching and the computing strategy is determined before delivering the j-th viewpoint, and caching MV or SV is two different strategies, which should not exist simultaneously.

3) holds because when the MV of the j-th viewpoint is cached at the μBS or mBS, it should be computed either at the MEC server or at the HMD. On the other hand, the computing strategy over the sub-6 GHz link can be different from that over the mmWave link. $\qquad\square$

According to Lemma 1, the caching and computing strategies are coupled, which can be used to reduce the solution space to 18 joint caching and computing strategies for the j-th viewpoint, which are listed in Table II. In Table II, we rewrite the joint caching and computing strategies of the j-th viewpoint in the form of 9-tuple as $\left[y_j^{H,M} \ y_j^{H,S} \ z_j^H, y_j^{\mu,M} \ y^{\mu,S} \ z^\mu, y^{m,M} \ y^{m,S} \ z^m \right]$. To distinguish whether the MVs computed at the HMD is cached locally, or cached at μBSs or mBSs, we further rewrite z_j^H as $z_j^H \triangleq (z_j^{H(H)} z_j^{H(\mu)} z_j^{H(m)})$.

The 18 joint caching and computing strategies can be divided into 4 types based on different caching strategies, i.e., local caching, caching at μBSs and mBSs simultaneously, caching at μBSs or mBSs, and backhaul retrieving. In each type, different computing strategies are included. Detailed descriptions are as follows,

- **Local caching (Strategy 1, 2):** In this type, the viewpoint j is cached locally at the HMD, i.e. $y_j^{H,M} = 1$, or $y_j^{H,S} = 1$. Since over-the-air transmission and backhaul retrieving are not required in this case, the viewpoint j can be guaranteed to be obtained in time, so the delivery reliability is set to 1. When $y_j^{H,M} = 1$, the j-th MV is cached locally, at the cost of d_j^M cache size and $k_H d_j^M \varepsilon$ computing energy consumption. When $y_j^{H,S} = 1$, the j-th SV is directly obtained locally, at the cost of d_j^S cache size.

- **Caching at μBSs and mBSs simultaneously:** In this type, depending on whether the MV or SV of the j-th viewpoint is cached at μBSs and mBSs, 4 cases can be generated,

  - **(Strategy 3 – 6)** The j-th MV is cached at both μBSs and mBSs. The delivery reliability over the sub-6 GHz link is $\mathcal{R}_j^\mu(d_j^S, T_j - \frac{\varepsilon d_j^M}{f_\mu})$ when the MV is projected into SV at the μBS, and $\mathcal{R}_j^\mu(d_j^M, T_j - \frac{\varepsilon d_j^M}{f_H})$ when the MV is projected into SV at the HMD. Likewise, the delivery reliability over the mmWave link is $\mathcal{R}_j^m(d_j^S, T_j - \frac{\varepsilon d_j^M}{f_m})$ when the MV is projected into SV at the mBS, and $\mathcal{R}_j^m(d_j^M, T_j - \frac{\varepsilon d_j^M}{f_H})$ when the MV is projected into SV at the HMD. The computing energy consumption is $A_{q_1} k_{q_2} d_j^M \varepsilon, q_1 \in \{\mu, m\}, q_2 \in \{\mu, m, H\}$ at the μBS, mBS or HMD for strategy 3, 4, 5, and $k_v d_j^M \varepsilon$ at the HMD for strategy 6.

  - **(Strategy 7, 8)** The MV is cached at μBSs, and the SV is cached at mBSs. The delivery reliability over the sub-6 GHz link depends on the computing strategy similar to the case in strategy 3 – 6. The delivery reliability over the mmWave link is $\mathcal{R}_j^m(d_j^S, T_j)$. The computing energy consumption is $A_\mu k_q d_j^M \varepsilon, q \in \{\mu, v\}$ at the μBS or HMD.

  - **(Strategy 9, 10)** The SV is cached at μBSs, and the MV is cached at mBSs. The delivery reliability over the sub-6 GHz link is $\mathcal{R}_j^\mu(d_j^S, T_j)$. The delivery reliability over the mmWave link depends on the computing strategy similar to the case in strategy 3 – 6. The computing energy consumption is $A_m k_q d_j^M \varepsilon, q \in \{m, v\}$ at the mBS or HMD.

  - **(Strategy 11)** The SV is cached at both μBSs and mBSs. The delivery reliability over the sub-6 GHz link is $\mathcal{R}_j^\mu(d_j^S, T_j)$, and that over the mmWave link is $\mathcal{R}_j^m(d_j^S, T_j)$. The cost of cache size is d_j^S at both μBSs and mBSs, without computing energy consumption.

- **Caching at μBSs or mBSs:** In this type, depending on whether the MV or SV is cached at μBSs or mBSs, 4 cases can be generated,

  - **(Strategy 12, 13)** The MV is cached at μBSs. The delivery reliability over the sub-6 GHz link depends on the computing strategy similar to the case in strategy 3 – 6.

TABLE II
JOINT CACHING AND COMPUTING STRATEGY IN DC SUB-6 GHZ AND MMWAVE HETNETS

Type	Strategy Index k	Joint Decision	sub-6 GHz Link		mmWave Link		Caching Occupancy ξ_{jk}^q			Computing Energy Consumption ζ_{jk}^q			Backhaul Cost φ_{jk}
			D_j^μ	$\tau_j^{\mu,c}+\tau_j^{\mu,b}$	D_j^m	$\tau_j^{m,c}+\tau_j^{m,b}$	μBS	mBS	HMD	μBS	mBS	HMD	
Local Caching	1	[10(100),000,000]	–	–	–	–	0	0	d_j^M	0	0	$k_H d_j^M \varepsilon$	0
	2	[01(000),000,000]	–	–	–	–	0	0	d_j^S	0	0	0	0
Caching at μBSs and mBSs simultaneously	3	[00(000),101,101]	d_j^S	$\frac{\varepsilon d_j^M}{f_\mu}$	d_j^S	$\frac{\varepsilon d_j^M}{f_m}$	d_j^M	d_j^M	0	$A_\mu k_\mu d_j^M \varepsilon$	$A_m k_m d_j^M \varepsilon$	0	0
	4	[00(001),101,100]	d_j^S	$\frac{\varepsilon d_j^M}{f_\mu}$	d_j^M	$\frac{\varepsilon d_j^M}{f_H}$	d_j^M	d_j^M	0	$A_\mu k_\mu d_j^M \varepsilon$	0	$A_m k_H d_j^M \varepsilon$	0
	5	[00(010),100,101]	d_j^M	$\frac{\varepsilon d_j^M}{f_H}$	d_j^S	$\frac{\varepsilon d_j^M}{f_m}$	d_j^M	d_j^M	0	0	$A_m k_m d_j^M \varepsilon$	$A_\mu k_H d_j^M \varepsilon$	0
	6	[00(011),100,101]	d_j^M	$\frac{\varepsilon d_j^M}{f_H}$	d_j^M	$\frac{\varepsilon d_j^M}{f_H}$	d_j^M	d_j^M	0	0	0	$k_H d_j^M \varepsilon$	0
	7	[00(000),101,010]	d_j^S	$\frac{\varepsilon d_j^M}{f_\mu}$	d_j^S	0	d_j^M	d_j^S	0	$A_\mu k_\mu d_j^M \varepsilon$	0	0	0
	8	[00(010),100,010]	d_j^M	$\frac{\varepsilon d_j^M}{f_H}$	d_j^S	0	d_j^M	d_j^S	0	0	0	$A_\mu k_H d_j^M \varepsilon$	0
	9	[00(000),010,101]	d_j^S	0	d_j^S	$\frac{\varepsilon d_j^M}{f_m}$	d_j^S	d_j^M	0	0	$A_m k_m d_j^M \varepsilon$	0	0
	10	[00(001),010,100]	d_j^S	0	d_j^M	$\frac{\varepsilon d_j^M}{f_H}$	d_j^S	d_j^M	0	0	0	$A_m k_H d_j^M \varepsilon$	0
	11	[00(000),010,010]	d_j^S	0	d_j^S	0	d_j^S	d_j^S	0	0	0	0	0
Caching at μBSs or mBSs	12	[00(000),101,000]	d_j^S	$\frac{\varepsilon d_j^M}{f_\mu}$	d_j^S	τ_j^r	d_j^M	0	0	$A_\mu k_\mu d_j^M \varepsilon$	0	0	$A_m d_j^S$
	13	[00(010),100,000]	d_j^M	$\frac{\varepsilon d_j^M}{f_H}$	d_j^S	τ_j^r	d_j^M	0	0	0	0	$A_\mu k_H d_j^M \varepsilon$	$A_m d_j^S$
	14	[00(000),010,000]	d_j^S	0	d_j^S	τ_j^r	d_j^S	0	0	0	0	0	$A_m d_j^S$
	15	[00(001),000,100]	d_j^S	τ_j^r	d_j^M	$\frac{\varepsilon d_j^M}{f_H}$	0	d_j^M	0	0	0	$A_m k_H d_j^M \varepsilon$	$A_\mu d_j^S$
	16	[00(000),000,101]	d_j^S	τ_j^r	d_j^S	$\frac{\varepsilon d_j^M}{f_m}$	0	d_j^M	0	0	$A_m k_m d_j^M \varepsilon$	0	$A_\mu d_j^S$
	17	[00(000),000,010]	d_j^S	τ_j^r	d_j^S	0	0	d_j^S	0	0	0	0	$A_\mu d_j^S$
Backhaul Retrieving	18	[00(000),000,000]	d_j^S	τ_j^r	d_j^S	τ_j^r	0	0	0	0	0	0	d_j^S

Considering the backhaul retrieve delay, the delivery reliability over the mmWave link is $\mathcal{R}_j^m(d_j^S, T_j - \tau_j^r)$. The computing energy consumption is $A_\mu k_q d_j^M \varepsilon, q \in \{\mu, \upsilon\}$ at the μBS or HMD, and the backhaul cost is $A_m d_j^S$.

– **(Strategy 14)** The SV is cached at μBSs. The delivery reliability over the sub-6 GHz link is $\mathcal{R}_j^\mu(d_j^S, T_j)$, and that over the mmWave link is $\mathcal{R}_j^m(d_j^S, \tau_j^r)$. No computing delay or computing energy consumption is generated, and the backhaul cost is $A_m d_j^S$.

– **(Strategy 15, 16)** The MV is cached at mBSs. Considering the backhaul delay, the delivery reliability over the sub-6 GHz link is $\mathcal{R}_j^\mu(d_j^S, T_j - \tau_j^r)$. The delivery reliability over the mmWave link depends on the computing strategy similar to the case in strategy 3 – 6. The computing energy consumption is $A_m k_q d_j^M \varepsilon, q \in \{m, \upsilon\}$ at the mBS or HMD, and the backhaul cost is $A_\mu d_j^S$.

– **(Strategy 17)** The SV is cached at mBSs. The delivery reliability over the sub-6 GHz link is $\mathcal{R}_j^\mu(d_j^S, T_j - \tau_j^r)$, and that over the mmWave link is $\mathcal{R}_j^m(d_j^S, T_j)$. No computing delay or computing energy consumption is generated, and the backhaul cost is $A_\mu d_j^S$.

• **Backhaul Retrieve (Strategy 18):** The delivery reliability over the sub-6 GHz link is $\mathcal{R}_j^\mu(d_j^S, T_j - \tau_j^r)$, and that over the mmWave link is $\mathcal{R}_j^m(d_j^S, T_j - \tau_j^r)$. The viewpoint j is retrieved via the backhaul at the cost of d_j^S.

Denote k as the strategy index of the 18 strategies listed in Table II, and $x_{jk} \in \{0, 1\}$

as the binary decision variable for the strategy of the j-th viewpoint. For the k-th strategy, the corresponding caching occupancy, computing energy consumption and the backhaul cost is denoted as ξ_{jk}^q, ζ_{jk}^q, $q \in \{\mu, m, H\}$, and φ_{jk}, respectively. Then problem **P1** can be reformulated as a multiple-choice multi-dimension knapsack problem (MMKP) as follows,

$$\textbf{P2:} \quad \max_{\{x_{jk}\}, j \in \mathcal{J}, k \in \{1,2,\cdots,18\}} \quad \sum_{j=1}^{J} \sum_{k=1}^{18} p_j \mathcal{R}_{jk} x_{jk} \tag{22a}$$

$$\text{s.t.} \quad \sum_{j=1}^{J} \sum_{k=1}^{18} \xi_{jk}^q x_{jk} \leq C^q, q \in \{H, \mu, m\}, \tag{22b}$$

$$\sum_{j=1}^{J} \sum_{k=1}^{18} p_j \zeta_{jk}^q x_{jk} \leq E^q, q \in \{H, \mu, m\}, \tag{22c}$$

$$\sum_{j=1}^{J} \sum_{k=1}^{18} \varphi_{jk} x_{jk} \leq B^b, \tag{22d}$$

$$\sum_{k=1}^{18} x_{jk} = 1, j \in \mathcal{J}, \tag{22e}$$

$$x_{jk} \in \{0,1\}, j \in \mathcal{J}, k \in \{1, 2, \cdots, 18\}. \tag{22f}$$

Problem **P2** is a 18-choice 7-dimensional MMKP problem, which is proved to be NP-hard. Specifically, J viewpoints are considered as J classes, while the 18 joint caching and computing strategies belonging to each class j are considered as 18 items. The problem is to choose one item from each class such that the profit sum is maximized while satisfying the capacity and energy consumption constraints.

IV. OPTIMIZATION OF JOINT CACHING AND COMPUTING

A. Optimal solution using BFBB algorithm

We propose a best-first branch and bound (BFBB) algorithm to solve the MMKP problem **P2**. The key idea of the BFBB algorithm can be described as follows: 1) Compute upper bounds and lower bounds of the optimal solution and discard branches that cannot produce a better solution than the best one found so far by the algorithm; 2) Utilize a priority queue to select the node with the highest priority as the expanded node of the search tree.

1) Computation of upper and lower bound: First, to compute the upper bound of problem **P2**, we construct an auxiliary problem **P3**, which relaxes the constraints into the sum of constraints

(22b)–(22d) as follows,

$$\textbf{P3:} \quad \max_{\{x_{jk}\}} \quad \sum_{j=1}^{J}\sum_{k\in\mathcal{K}} p_j \mathcal{R}_{jk} x_{jk} \tag{23a}$$

$$\text{s.t.} \quad \sum_{j=1}^{J}\sum_{k\in\mathcal{K}} \varpi_{jk} x_{jk} \le C_0, \tag{23b}$$

$$(22e), (22f),$$

where $\varpi_{jk} = \left(\varphi_{jk} + \sum_{q\in\{H,\mu,m\}} (\xi_{jk}^q + p_j \zeta_{jk}^q) \right)$ and $C_0 = B^b + \sum_{q\in\{H,\mu,m\}} (C^q + E^q)$.

For each viewpoint j, select the joint caching and computing decision k which maximizes $\frac{\mathcal{R}_{jk}}{\varpi_{jk}}$, and denote the corresponding decision as $k_{\max}$ for each viewpoint j. Let $\mathcal{R}_{\text{UB}}$ denote the upper bound of problem **P3**, and let $\varpi_0 = \sum_{j\in\mathcal{J}} \varpi_{jk_{\max}}$, $\mathcal{R}_0 = \sum_{j\in\mathcal{J}} \mathcal{R}_{jk_{\max}}$, then there exist the following two cases,

- $\varpi_0 > C_0$. In this case, the solution of problem **P3** is upper bounded by

$$\mathcal{R}_{\text{UB}} = \sum_{j\in\mathcal{J}} \mathcal{R}_{jk_{\max}} \times \left(\frac{C_0}{\sum_{j\in\mathcal{J}} \varpi_{jk_{\max}}} \right). \tag{24}$$

- $\varpi_0 < C_0$. In this case, the constraint (23b) is not violated when decision $x_{jk_{\max}}$ is selected for all $j \in \mathcal{J}$. The remaining items are merged into the same class L, with items indexed by $\ell = 1, \cdots, N_f$. The remaining items are sorted in descending order of $\frac{\mathcal{R}_\ell}{\varpi_\ell}$. Then, the problem is converted into choosing the remaining items with the remaining capacity constraint equals to $C_0 - \varpi_0$. Adopting greedy algorithm, we select the items in descending order of $\frac{\mathcal{R}_\ell}{\varpi_\ell}$ to fill the remaining capacity until the capacity constraint is violated. In particular, denote $\hat{\ell} \in [1, \ell]$ the index of the item that violates the remaining capacity constraint $C_0 - \varpi_0$, which can be defined as,

$$\hat{\ell} = \min \left\{ \hat{\ell} : \sum_{\ell=1}^{\hat{\ell}-1} \varpi_\ell \le C_0 - \varpi_0 < \sum_{\ell=1}^{\hat{\ell}} \varpi_\ell \right\}. \tag{25}$$

Then the solution of problem **P3** is upper bounded by

$$\mathcal{R}_{\text{UB}} = \mathcal{R}_0 + \sum_{\ell=1}^{\hat{\ell}-1} \mathcal{R}_\ell + \left(\frac{(C_0 - \varpi_0) - \sum_{\ell=1}^{\hat{\ell}-1} \varpi_\ell}{\varpi_q} \right) \times \mathcal{R}_{\hat{\ell}}. \tag{26}$$

Lemma 2. $\mathcal{R}_{\text{UB}}$ *is an upper bound for the auxiliary problem **P3** as well as the problem **P2**.*

Proof. The proof can be shown by contradiction for the first case ($\varpi_0 > C_0$) of this lemma. Suppose that there exists a solution $\hat{X} = (\hat{x}_{1k_1}, \cdots, \hat{x}_{jk_j}, \cdots, \hat{x}_{Jk_J})$ satisfying problem **P3** with

corresponding objective value $\hat{\mathcal{R}} = \sum_{j \in \mathcal{J}} \mathcal{R}_{jk_j}$ such that $\mathcal{R}_0 \times (\frac{C_0}{\varpi_0}) < \hat{\mathcal{R}}$. Then we have $\frac{\mathcal{R}_0}{\varpi_0} < \frac{\hat{\mathcal{R}}}{C_0}$. Since $\hat{\varpi} = \sum_{j \in \mathcal{J}} \varpi_{jk_j} \leq C_0$, we have $\frac{\hat{\mathcal{R}}}{C_0} \leq \frac{\hat{\mathcal{R}}}{\hat{\varpi}}$. Thus $\frac{\mathcal{R}_0}{\varpi_0} < \frac{\hat{\mathcal{R}}}{\hat{\varpi}}$.

According to the descending order of $\frac{\mathcal{R}_{jk}}{\varpi_{jk}}$, the inequality $\frac{\mathcal{R}_{jk_{\max}}}{\varpi_{jk_{\max}}} \geq \frac{\mathcal{R}_{jk_j}}{\varpi_{jk_j}}, \forall j \in \mathcal{J}$ holds, thus we have

$$\frac{\sum_{j \in \mathcal{J}} \mathcal{R}_{jk_{\max}}}{\sum_{j \in \mathcal{J}} \varpi_{jk_{\max}}} \geq \frac{\sum_{j \in \mathcal{J}} \mathcal{R}_{jk_j}}{\sum_{j \in \mathcal{J}} \varpi_{jk_j}}, \tag{27}$$

which is contradictory to $\frac{\mathcal{R}_{\max}}{\varpi_{\max}} < \frac{\hat{\mathcal{R}}}{\hat{\varpi}}$.

For the second case ($\varpi_0 \leq C_0$), the remaining capacity is filled with the remaining items as a knapsack problem, so the upper bound can be obtained also by the greedy algorithm.

Since problem **P3** is a relaxed problem of the problem **P2**, it is obvious that UB is an upper bound for **P2**.

$\square$

To determine an initial feasible solution that can be used as the starting lower bound, we develop a heuristic algorithm for obtaining the lower bound of problem **P2** as described in Algorithm 1. The algorithm is based on the conception of aggregate resource saving [40] considering the multiple resource constraints in **P2**. Specifically, the solutions for each viewpoint that achieve lowest reliability is selected as initial solution. The solution is then upgraded by choosing a new solution for a viewpoint which has the maximum aggregate resource saving while increasing the objective function of total reliability. The aggregate resource saving is defined as

$$\Delta \varpi_{jk} = \varpi_{jk_j} - \varpi_{jk}. \tag{28}$$

Note that there might be some solutions that violate the constraints but achieve higher reliability. To obtain a tighter lower bound, we downgrade some of the solutions to get a feasible solution, which may increase the total value of reliability.

2) Best-first principle of priority: The candidate nodes to be expanded are stored in the priority queue, and assume that the level of the search tree where a candidate node is located is denoted as j_0. Define the best-first principle of priority as the maximum upper bounds of the candidate nodes, which can be calculated as

$$\mathcal{P} = \sum_{j=1}^{j_0} p_j \mathcal{R}_{jk_j} x_{jk_j} + \widetilde{\mathcal{R}}_{\mathrm{UB}}, \tag{29}$$

where $\widetilde{\mathcal{R}}_{\mathrm{UB}}$ is the upper bound of the remaining capacity constraint problem of a candidate node

---

Algorithm 1: Heuristic Lower Bound Computation Algorithm

Input: $\mathcal{P}_{\text{s}}$, F, C, ϵ;
Output: Optimal solution $\mathbf{p}^* = (p_i^*)_{i \in \mathcal{F}}$;

1 initialization: ;

2 Select the lowest-valued solution for each viewpoint.

3 Replace a strategy of a viewpoint which has the highest positive value of $\Delta \varpi_{jk}$ and subject to the resource constraints (22b)–(22d) . If no such strategy is found, then a strategy with the highest $\Delta \mathcal{R}_{jk} / \Delta \varpi_{jk}$;

4 if *no such solution is found in step 3* **then**

5   go to step 8;

6 else

7   look for another solution in step 3;

8 if *there exist higher-valued solutions than the selected solution for any viewpoint* **then**

9   select a highest-valued strategy $\mathcal{R}_{jk}$ with the minimum aggregate resource consumption ϖ_{jk};

10 Replace a lower-valued strategy $\mathcal{R}_{jk}$ that consumes the maximum aggregate resource ϖ_{jk} with the strategy selected in step 9;

11 if *a solution is found in step 10* **then**

12   **if** *the solution satisfies the constraints (22b)–(22d)* **then**

13   look for a better solution in step 3;

14   **else**

15   go to step 10 for another downgrade.

16 else

17   revive the solution obtained at the end of step 3 and terminate.

---

as follows,

$$\textbf{P4:} \max_{\{x_{jk}\}} \quad \sum_{j=j_0+1}^{J} \sum_{k \in \mathcal{K}} p_j \mathcal{R}_{jk} x_{jk} \tag{30a}$$

$$\text{s.t.} \quad \sum_{j=j_0+1}^{J} \sum_{k \in \mathcal{K}} \varpi_{jk} x_{jk} \leq C_0 - \sum_{j=1}^{j_0} \sum_{k \in \mathcal{K}} \varpi_{jk_j} x_{jk_j}, \tag{30b}$$

$$(22e), (22f).$$

The best-first principle of priority $\mathcal{P}$ is the sum of the current reliability value of a candidate node and its remaining possible reliability value, i.e., the upper bound of problem **P4**. Note that the form of problem **P4** is the same as that of problem **P3**, thus the upper bound of problem **P4** can be obtained using the similar method described in Lemma 2.

The advantage of applying the best-first principle of priority is that the first obtained feasible solution using the branch and bound searching method achieves the maximum value of reliability

of VR delivery, which accelerates the searching process to find the optimal solution faster.

3) The proposed BFBB algorithm: In this subsection, we propose the optimal joint caching and computing solution based on BFBB algorithm to solve problem **P2**, as described in Algorithm 2. The BFBB algorithm starts by initializing 18 strategies of the first viewpoint into the priority queue, and using Algorithm 1 to obtain the lower bound of the solution. The next node to be expanded is selected according to the best-first principle of priority (29), and the lower bound is then updated by directly computing the solution of $\mathbf{x}_{j_0 k_{j_0}}$. To perform the branch and bound operation, the feasibility of the expanded child node is verified, and the upper bound of the extended child node is computed using (29). To accelerate the searching process, the lower bound $\mathcal{R}_{\mathrm{LB}}$ is updated using Algorithm 1 in step 13, thus more branches of the search tree can be pruned in step 11. The optimality of the BFBB algorithm is guaranteed by the best-first principle. Since the nodes added into the PQ all correspond to feasible solutions, when $j_0 = J$, i.e., at step 5, the algorithm obtains the biggest value such that $\mathbf{x}_{J k_J}$ is feasible. Then $\mathbf{x}_{J k_J}$ is the optimal solution for problem **P2**.

Algorithm 2: The Optimal Joint Caching and Computing Solution based on BFBB

Input: $p_j, \mathcal{R}_{jk}, \xi^q_{jk}, \zeta^q_{jk}, \varphi^q_{jk}, C^q, E^q, B^b, j \in \mathcal{J}, k \in \mathcal{K}, q \in \{H, \mu, m\}$;
Output: Optimal solution $\{x^*_{jk}\}, j \in \mathcal{J}, k \in \mathcal{K}$;

1 **initialization:** Using Algorithm 1 to obtain the lower bound of the solution $\mathcal{R}_{\mathrm{LB}}$; denote the priority queue as PQ;
2 **while** PQ $\neq \varnothing$ **do**
3   Select a node $\mathbf{x}_{j_0 k_{j_0}} \triangleq \{x_{1k_1}, x_{2k_2}, \cdots, x_{j_0 k_{j_0}}\}$ from PQ according to the best-first principle of priority (29), and remove $\mathbf{x}_{j_0 k_{j_0}}$ from PQ;
4   Update the lower bound of the solution $\mathcal{R}_{\mathrm{LB}}$ using $\mathbf{x}_{j_0 k_{j_0}}$;
5   **if** $j_0 = J$ **then**
6   Exit the loop with the optimal solution $\mathbf{x}_{J k_J}$;
7   Expand the node $\mathbf{x}_{j_0 k_{j_0}}$ to obtain its child nodes $\mathbb{C}$;
8   **while** $\mathbb{C} \neq \varnothing$ **do**
9   **if** $\mathbf{x}_{(j_0+1)k}$ *corresponds to a feasible solution* **then**
10   Calculate the upper bound of the solution $\mathcal{R}_{\mathrm{UB}}$ containing $\mathbf{x}_{(j_0+1)k}$ using (29);
11   **if** $\mathcal{R}_{\mathrm{UB}} > \mathcal{R}_{\mathrm{LB}}$ **then**
12   add $\mathbf{x}_{(j_0+1)k}$ into PQ;
13   Update the current maximum value of $\mathcal{R}_{\mathrm{LB}}$ using Algorithm 1;

B. Suboptimal solution based on DCP

In this subsection, a suboptimal solution for solving **P2** is proposed by using difference of convex programming (DCP) with lower computation complexity. First, we equivalently transform

P2 into a continuous optimization problem, which can be written as,

$$\textbf{P5:} \quad \min_{\{x_{jk}\}} \quad \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} -p_j \mathcal{R}_{jk} x_{jk} \tag{31a}$$

$$\text{s.t.} \quad 0 \le x_{jk} \le 1, j \in \mathcal{J}, k \in \mathcal{K}, \tag{31b}$$

$$\sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} x_{jk}(1 - x_{jk}) \le 0, \tag{31c}$$

$$(22b), (22c), (22d), (22e),$$

where the binary variable constraint (22f) in **P2** is substituted by continuous variable constraints (31b) and (31c). Thus, by solving the continuous optimization problem **P5**, the computation complexity is greatly reduced compared with that by directly solving **P2**. Nevertheless, the constraint (31c) is a concave function instead of a convex function, which becomes the main difficulty for solving **P5**.

Algorithm 3: Suboptimal solution based on DCP

Input: $p_j, \mathcal{R}_{jk}, \xi_{jk}^q, \zeta_{jk}^q, \varphi_{jk}^q, C^q, E^q, B^b, j \in \mathcal{J}, k \in \mathcal{K}, q \in \{H, \mu, m\}$;
Output: Suboptimal solution of $\{x_{jk}\}$;

1 **Initialize:** counter $i = 0$, obtain an initial feasible solution $\{x_{jk}^0\}$ using Algorithm 1, threshold $\epsilon = 10^{-5}$;

2 **repeat**

3   **Compute:** $\hat{f}(\mathbf{x}; \mathbf{x}^i) \triangleq f(\mathbf{x}^i) + \nabla f(\mathbf{x}^i)^T(\mathbf{x} - \mathbf{x}^i)$.

4   **Solve:** Set the value of $\{x_{jk}^{i+1}\}$ to be a solution of the following convex problem:

$$\min_{\{x_{jk}\}} \quad f_0(\{x_{jk}\}) - \phi \hat{f}(\{x_{jk}\}; \{x_{jk}^i\})$$

$$\text{s.t.} \quad (22b), (22c), (22d), (22e), (31b).$$

5   where $f_0(\{x_{jk}\}) = \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} -p_j \mathcal{R}_{jk} x_{jk}$.

6   **Update:** $i \leftarrow i + 1, \{x_{jk}^i\} \leftarrow \{x_{jk}^{i-1}\}$;

7 **until** $(f_0(\boldsymbol{x}^{i-1}) - \phi f(\boldsymbol{x}^{i-1})) - (f_0(\boldsymbol{x}^i) - \phi f(\boldsymbol{x}^i)) \le \epsilon$;

To tackle this difficulty, we utilize a penalty function to bring the concave constraint into the objective function, which can be written as,

$$\textbf{P6:} \quad \min_{\{x_{jk}\}} \quad \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} -p_j \mathcal{R}_{jk} x_{jk} - \phi f(\mathbf{x}) \tag{32a}$$

$$\text{s.t.} \quad (22b), (22c), (22d), (22e), (31b),$$

where ϕ is a penalty parameter and $\phi > 0$, $f(\mathbf{x}) = \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} x_{jk}(x_{jk} - 1)$. To ensure the equivalence between **P5** and **P6**, according to the exact penalty property of DCP in [41], ϕ

should satisfy,

$$\phi > \frac{-p_j \mathcal{R}_{jk} x_{jk}^0 - g(0)}{\phi_0}, \tag{33}$$

where x_{jk}^0 denotes any feasible solution, $g(\phi)$ denotes the optimal objective value for **P6**, and $\phi_0 = \min_{\mathbf{x}}\{\sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} x_{jk}(1 - x_{jk}) : (22b), (22c), (22d), (22e), (31b)\}$.

Note that the objective function in **P6** is in the form of a difference of two convex functions. Therefore, we adopt DCP to solve this problem as outlined in Algorithm 3. Specifically, the objective function is convexified by the affine minorization function of the second term $f(\mathbf{x})$ in step 3. The complexity of Algorithm 3 lies in solving a series of convex problems in steps 3-6. These convex problems can be solved using the polynomial interior point method [42], which requires a third degree polynomial complexity in terms of the number of variables. Let L be the total number of iterations, then the complexity of Algorithm 3 is $\mathcal{O}(L(JK)^3)$, where J is the number of viewpoints, and K is the number of strategies for the j-th viewpoint.

Remark 6: It is worth mentioning that the authors in [26] proposed to use the Lagrangian dual decomposition (LDD) approach to obtain a suboptimal solution of the MMKP problem. The Lagrangian relaxation method was adopted in [26] to relax all the resource constraints into the objective function, without considering the specific structure of the objective function and the constraints, so it is hard to obtain a high-quality solution close to the optimal solution by using the LDD approach. Besides, the solution of the Lagrangian relaxed problem is not necessarily a feasible solution to the original problem, so the initial condition need to be changed multiple times to obtain a feasible solution of the original problem. In comparison, we first equivalently transform the MMKP problem into a continuous non-convex optimization problem. To deal with the concave function in the constraints, we exploit the specific structure of the problem and bring the concave constraint into the objective function. Note that this is also an equivalent problem transformation owing to the use of exact penalty property of the DCP problem. Further, the DCP problem is solved by approximating the penalty function as its affine minorization function, which can obtain a high-quality feasible solution effectively, and will be validated in Sec. V.

Remark 7 (Practical Implication): The proposed BFBB and DCP algorithms are implemented based on the pre-known statistical channel state information of sub-6 GHz and mmWave links, as well as the pre-known caching, computing and bandwidth resource information, thus they are offline algorithms which can obtain the optimal and sub-optimal joint caching and computing strategy, respectively. The overall decision is get made in two steps in the practical implementation. First, during the off-peak time, the HMD and the MEC servers obtain the joint caching and computing strategy by using the proposed BFBB or DCP algorithm, and perform caching

TABLE III
PARAMETER VALUES

Parameters	Physical meaning	Values
P_μ / P_m	Transmit power of μBSs / mBSs	30 / 30 dBm
B_μ / B_m	Bandwidth assigned to each user at sub-6 GHz / mmWave	100 / 500 MHz
α_L / α_N	Path loss exponent of LOS and NLOS	2.5 / 4
θ	Mainlobe beamwidth	30°
M / m	Mainlobe antenna gain / sidelobe antenna gain	10 / -10 dB
κ	Blockage density	6×10^{-4} (Unless otherwise stated)
N_L / N_N	Nakagami fading parameter for LOS and NLOS channel	3 / 2
λ_μ / λ_m / λ_H	density of μBSs / mBSs / HMDs	10^{-5} / 3×10^{-5} / 10^{-4} nodes/m^2
T_j	End-to-end delay threshold	20 ms (Unless otherwise stated)
δ	Skewness of the viewpoint popularity	0.8
C^μ / C^m / C^H	The cache size at μBSs / mBSs / HMDs	30 / 30 / 10 Mb
J	The number of viewpoints	20
d_j^M / d_j^S	Size of MVs / SVs	1 / 3 Mb
ε	The number of computation cycles required for 1 bit input	10
f_H / f_μ / f_m	CPU cycle of HMDs / μBSs / mBSs	10^9 / 3×10^9 / 3×10^9
η_H / η_μ / η_m	Energy efficiency coefficient of HMDs / μBSs / mBSs	10^{-25} / 10^{-26} / 10^{-26}
τ_j^r	Backhaul retrieving delay	10 ms

placement according to the caching strategy. Second, in the viewpoint delivery phase, the HMD calculates the end-to-end delay of sub-6 GHz links and mmWave links based on the channel state information and the joint caching and computing strategy. Then the HMD selects the link to receive viewpoint data according to the minimum-delay delivery criteria, thus the communication strategy can be determined.

V. PERFORMANCE EVALUATION

In this section, the performances of the proposed BFBB and DCP algorithms are evaluated through numerical simulations. The default parameter values of network environment, mobile edge resources and VR videos are listed in Table III. The μBSs, mBSs, and HMDs are scattered in a square area of 1km × 1km based on three independent PPPs with densities listed in Table III, and the mobility of the HMDs follows the random-walk model [43]. For the sake of comparison, five benchmark algorithms are provided including the LDD algorithm used in [26], and the other four benchmark algorithms listed as follows,

- Local SV caching (LSC): First, the SVs are cached locally at the HMD according to the descending order of the popularity of the viewpoints until the caching capacity is full. Then, the remaining SVs are cached at μBSs and mBSs according to the descending order of the popularity until the caching capacities are full. Then the remaining viewpoints are delivered via the backhaul.

- Local computing with MV caching preferentially (LC-MCP): The MVs are cached at the HMD according to the descending order of the popularity of the viewpoints until the local computing resources are fully utilized. If there remains caching capacity, the SVs are cached until the caching capacity is full. The remaining MVs are cached at μBSs and mBSs until the edge computing resources or caching capacities are fully utilized.

- μBS-only: Only μBSs are deployed in the VR delivery network, i.e., let $\lambda_m = 0$, and use the proposed BFBB algorithm to make the joint caching and computing decision.

- mBS-only: Only mBSs are deployed in the VR delivery network, i.e., let $\lambda_\mu = 0$, and use the proposed BFBB algorithm to make the joint caching and computing decision.

A. Reliability Performance of the Proposed Algorithms

We first evaluate the reliability of VR delivery with various network parameters as shown in Fig. 2(a) and Fig. 2(b). It is observed that the BFBB algorithm achieves the highest reliability, and the DCP algorithm achieves reliability close to that of the BFBB algorithm, which is 97.8% of that of the BFBB algorithm on average. This is because the proposed algorithms comprehensively consider the characteristics of the sub-6 GHz link and the mmWave link to determine a joint caching and computing strategy. Note that the LDD algorithm is not able to achieve the same reliability performance as the DCP algorithm, which is mainly due to the operation of relaxing all the constraints into the objective function without exploiting the specific structure of the MMKP problem as the DCP algorithm does.

The reliability of VR delivery with various blockage densities is shown in Fig. 2(a). The blockage model is in accordance with the NLOS probability function defined in Sec. II-A, and the blockages are generated according to the blockage density parameter κ and the NLOS probability function. We assume independent LOS probability between different links, i.e., potential correlations of blockage effects between links are ignored, which causes negligible loss of accuracy for performance evaluation [33], [34]. It is observed that the performance of LC-MCP is better than that of LSC. This is because caching MVs in the HMD can save more caching capacity and more viewpoints can be cached. Using the local computing capability of the HMD can help more viewpoints to be played in time. The mBS-only can achieve performance close to the proposed algorithms when the blockage density is low, while the performance is significantly reduced when the blockage density is increased. This is because the blockage effect in the mmWave tier causes the link rate to drop rapidly, and the viewpoints cannot be transmitted through the relatively stable sub-6 GHz link, making the reliability of VR delivery very low. In contrast, the reliability of VR delivery using the μBS-only strategy remains unchanged with
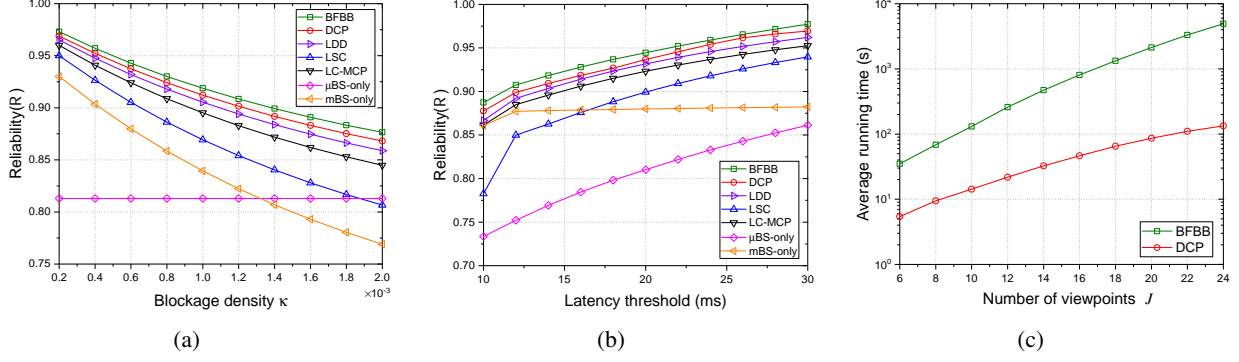
Fig. 2. (a) Reliability of VR delivery versus blockage density κ, and (b) reliability of VR delivery versus latency threshold, (c) average running time of the proposed algorithms with various numbers of viewpoints.

various blockage densities. This is because the wireless channel fluctuation of the sub-6 GHz tier is not affected by the blockages, which makes the reliability independent of blockage factors.

Fig. 2(b) shows the reliability of VR delivery with various delay thresholds. It can be seen that the proposed BFBB and DCP algorithms also achieve higher reliability. The performance of LC-MCP is better than that of LSC, especially when the delay threshold is low. This is because LC-MCP can cache more viewpoints locally or at BSs, eliminating the backhaul retrieve delay, which is more important when the delay threshold is low. Observing the μBS-only and mBS-only algorithms, it is seen that mBS-only basically does not change with the increase of the delay threshold, while μBS-only increases much with the increase of the delay threshold. This indicates that the CDFs of the coverage probabilities of the two channels have different characteristics, and it is necessary to combine the complementary advantages of the two channels to achieve higher reliability.

Fig. 2(c) shows the average running time of the proposed algorithms with various numbers of viewpoints, which is tested on the computer with an Intel Core CPU i5 with a clock rate of 2.30 GHz, and an RAM with 8 G. It is observed that the DCP algorithm runs much faster than the BFBB algorithm, especially when the number of viewpoint becomes larger. This indicates that the DCP algorithm is a more practical algorithm than the BFBB algorithm when applied to the scenario where the number of viewpoints is large. On the other hand, the BFBB algorithm is also meaningful, because it can be used to obtain the optimal solution of the MMKP problem, which can be viewed as the upper bound of the reliability performance in the considered scenario. Based on this optimal solution, the distance between the results obtained by the suboptimal algorithm and the optimal algorithm can be calculated, which can be used to quantitatively measure the performance of the suboptimal algorithm.
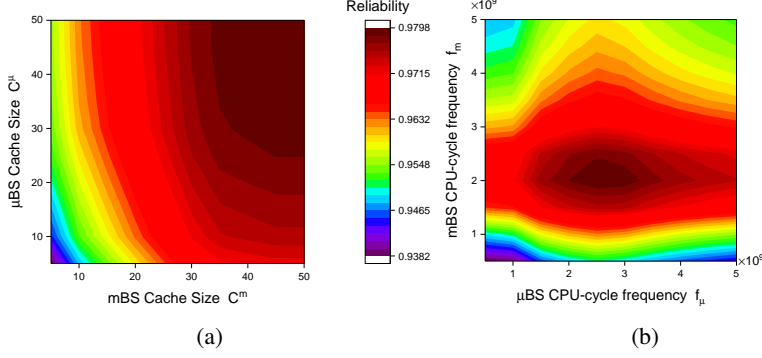
Fig. 3. (a) Reliability of VR delivery for various cache size of μBSs and mBSs, and (b) reliability of VR delivery for various CPU-cycle frequency of μBSs and mBSs.
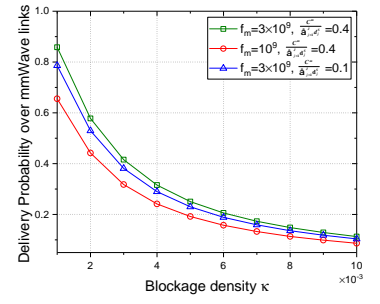
Fig. 4. Delivery probability over mmWave links for various mmWave network parameters.

The impact of caching capacities of μBSs and mBSs on the reliability of VR delivery is shown in Fig. 3(a). In general, the reliability is higher with a larger caching capacity. It is observed that the caching capacity has a greater gain in the reliability of VR delivery over mmWave links. For example, the reliability is about 0.97 when $C^\mu = 10, C^m = 50$, while the reliability is about 0.95 when $C^\mu = 50, C^m = 10$. This is because the SVs cached at mBSs can be delivered over large-bandwidth mmWave links, so that the SVs can be delivered to the HMD in time.

Fig. 3(b) shows the impact of CPU-cycle frequency reliability of μBSs and mBSs on the reliability of VR delivery. It is observed that, it is not that the higher the CPU-cycle frequency, the higher the reliability can be obtained. The reason is analyzed in conjunction with the energy consumption limitation in the following subsection. Note that the reliability of VR delivery is more sensitive to f_m compared with f_μ. In other words, the impact on the reliability is greater when f_m is changed. This is because the mBSs needs to consume more computing resources to project and render the MVs into SVs, and deliver them over the large-bandwidth mmWave links. In contrast, the number of MVs computed into SVs at μBSs is relatively small, which results in less impact on the reliability.

B. Delivery Probability

The delivery probability over mmWave links is validated with various network parameters as shown in Fig. 4. It can be observed that the delivery probability over mBSs decreases with the increase of the blockage density due to the deteriorating channel conditions, which reflects the importance of sub-6 GHz links to enhance the reliability of VR delivery despite the large mmWave bandwidth. In addition, the delivery probability over mBSs increases when the CPU-cycle frequency or the caching capacity of mBSs increases. This is because the increase in
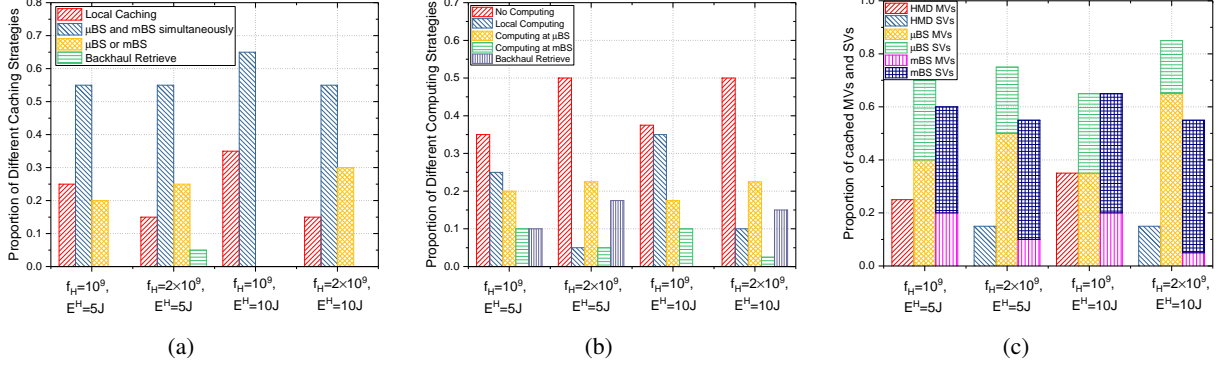
Fig. 5. (a) Proportion of different caching strategies under various f_H and E^H, (b) proportion of different computing strategies under various f_H and E^H, and (c) proportion of cached MVs and SVs under various f_H and E^H.

caching and computing resources reduces the average end-to-end delay, thus more viewpoints are delivered over mmWave links according to the minimum delay criterion.

C. Caching and Computing Strategies

We now evaluate the caching and computing strategies to obtain design insights. Fig. 5(a) shows the proportion of different caching strategies under various f_H and E^H. It can be seen that the locally cached viewpoints decrease when the local CPU-cycle frequency f_H increases. This is due to the local energy consumption limitation. Higher computing frequency leads to higher energy consumption, as shown in Fig. 5(b). Therefore, the proportion of local computing should be reduced, otherwise it will cause HMD to heat up and reduce the user experience. Thus, more SVs are cached in the HMD. On the other hand, when f_H is unchanged and E^H is increased, the proportion of local caching or simultaneously cached in μBS and mBS increases. This is because the local energy consumption limitation is loosen, and more MVs can be cached locally or at BSs. Then more MVs are delivered to the HMD over sub-6 GHz links or mmWave links and computed into SVs locally, as shown in the proportion of local computing in Fig. 5(b). It is worth noting that when $f_H = 10^9$, the proportion of cached at μBS or mBS decreases with the increase of E^H. This is because the energy consumption limitation is strict at HMD when E^H is small, thus more SVs should be cached in the limited caching capacity at μBSs and mBSs to avoid the backhaul retrieve delay. When E^H increases, more MVs can be cached, so more viewpoints can be cached simultaneously at μBSs and mBSs.

The proportion of cached MVs and SVs under various f_H and E^H is shown in Fig. 5(c). It can be seen that SVs are mainly cached at mBSs, while MVs are more cached at μBS. This indicates that it is beneficial to cache more SVs at mBSs and deliver them over mmWave links,

while cache more MVs at μBSs and deliver them over sub-6 GHz links. This is reasonable, because delivering more SVs over mmWave links will not increase the transmission delay much due to the large mmWave bandwidth, and the computing delay is saved. On the contrary, more MVs can be delivered over sub-6 GHz links due to the relatively low bandwidth.

VI. CONCLUSION

In this paper, we propose a dual-connectivity sub-6 GHz and mmWave HetNet architecture empowered by mobile edge capability, with the aim of improving the reliability of VR delivery. Based on the differentiated characteristics of sub-6 GHz links and mmWave links, we utilize their complementary advantages to conduct a collaborative design to improve the reliability of VR delivery. From the perspective of stochastic geometry, we first derive closed-form expressions for the reliability of VR delivery. We propose a link selection strategy based on the minimum-delay delivery, and derive the VR delivery probability over sub-6 GHz links and mmWave links. We theoretically show the necessity of sub-6 GHz links to improve the reliability despite the large mmWave bandwidth. We then formulate a joint caching and computing optimization problem to maximize the reliability of VR delivery. By analyzing the coupling caching and computing strategies, we further transform the problem into a MMKP and propose a BFBB algorithm to obtain the optimal solution. To further reduce the complexity of the algorithm, we leverage DCP algorithm to obtain a sub-optimal solution. Numerical simulations demonstrate the performance improvement using the proposed algorithms, and shows great promise to improve the VR delivery reliability in mobile-edge empowered DC sub-6 GHz and mmWave HetNets.

In future work, a prospective direction is to improve the quality of experience (QoE) for VR users in DC sub-6 GHz and mmWave HetNets. The QoE of VR video transmission is influenced by many factors such as the video rendering, quality of viewpoints, end-to-end delay, and delivery reliability, which is more challenging for modeling and optimization. A QoE-driven cross-layer design framework for mobile-edge empowered DC sub-6 GHz and mmWave HetNets is anticipated, in which resource coordination that dynamically adapts to network conditions can be designed to achieve QoE enhancement.

APPENDIX A

PROOF OF PROPOSITION 1

According to (1) and (11), the reliability of VR delivery over sub-6 GHz links can be derived as

$$\mathcal{R}_j^\mu(D_j^\mu, T_j^{\mu,t}) = \mathbb{P}\left[\frac{P_\mu h_j^\mu r^{-\alpha_\mu}}{I_j^\mu + \sigma_\mu^2} > \nu_j^\mu\right] = \int_0^\infty \mathbb{P}\left[\frac{P_\mu h_j^\mu r^{-\alpha_\mu}}{I_j^\mu + \sigma_\mu^2} > \nu_j^\mu \Big| r\right] f_r(r)\mathrm{d}r,$$

$$\stackrel{(a)}{=} \int_0^\infty e^{-\pi\lambda_\mu r^2} e^{-\nu_j^\mu r^{\alpha_\mu} P_\mu^{-1}\sigma_\mu^2} \mathcal{L}_{I_r}(\nu_j^\mu r^{\alpha_\mu} P_\mu^{-1}) \cdot 2\pi\lambda_\mu r\mathrm{d}r, \tag{34}$$

where (a) follows by using the fact that h_j^μ obeys exponential distribution, and $\mathcal{L}_{I_j^\mu}(\nu_j^\mu r^{\alpha_\mu} P_\mu^{-1})$ is the Laplace transform of random variable I_r. Let $s = \nu_j^\mu r_\mu^\alpha P_\mu^{-1}$, we have

$$\mathcal{L}_{I_j^\mu}(s) = \mathbb{E}_{\Phi_\mu, h}[e^{-sI_j^\mu}] = \mathbb{E}_{\Phi_\mu}\left[\prod_{x\in\Phi_\mu\backslash b(o,r)} \frac{s\ell(x)}{1+s\ell(x)}\right] \stackrel{(b)}{=} \exp\left(-2\pi\lambda_\mu\int_r^\infty \frac{sx}{s+x^{\alpha_\mu}}\mathrm{d}x\right),$$

$$\stackrel{(c)}{=} \exp(\pi\lambda_\mu s^{\delta_\mu}\Gamma(1+\delta_\mu)\Gamma(1-\delta_\mu) - \pi\lambda_\mu r^2 H_{\delta_\mu}(r^{\alpha_\mu}/s)),$$

where $\ell(x) = x^{-\alpha_\mu}$, $\delta_\mu = 2/\alpha_\mu$. (b) follows from the probability generating functional (PGFL) of the PPP [44], and (c) follows from the use of gamma function or Gauss hypergeometric function for integration, where $H_\delta(x) \triangleq {}_2F_1(1, \delta; 1+\delta; -x)$ is the Gauss hypergeometric function. By applying the Gauss-Laguerre Quadrature [45], the desired proof is obtained.

APPENDIX B

PROOF OF PROPOSITION 2

According to (4) and (11), the reliability of VR delivery over mmWave links is derived as

$$\mathcal{R}_j^m(D_j^m, T_j^{m,t}) = \mathbb{P}\left[\frac{P_m h_j^m G r^{-\alpha_m}}{I_j^m + \sigma_m^2} > \nu_j^m\right],$$

$$\stackrel{(d)}{=} \int_0^\infty \sum_{i\in\{L,N\}} \rho_i(r)\left\{1 - \mathbb{E}_{I_j^m}\left[\left(1 - \exp\left(-\frac{\eta_i\nu_j^m r^{\alpha_i}(I_j^m + \sigma_m^2)}{P_m G}\right)\right)^{N_i}\right]\right\} f_r(r)\mathrm{d}r,$$

$$\stackrel{(e)}{=} \int_0^\infty \sum_{i\in\{L,N\}} \rho_i(r)\left\{\sum_{k=1}^{N_i}(-1)^{k+1}\binom{N_i}{k}e^{-\frac{k\eta_i\nu_j^m r^{\alpha_i}\sigma_m^2}{P_m G}}\mathcal{L}_{I_j^m}\left(\frac{k\eta_i\nu_j^m r^{\alpha_i}\sigma_m^2}{P_m G}\right)\right\} f_r(r)\mathrm{d}r,$$

where (d) follows from the Alzer's approximation of a gamma random variable [46]. (e) follows by using Binomial theorem and the assumption that N_i is an integer.

$$\mathcal{L}_{I_j^m}(s_i) \stackrel{(f)}{=} \mathbb{E}_{I_j^m}\left[\prod_{\ell\in\Phi_m\backslash b_m} \mathbb{E}_h\left[\exp\left(-s_i h_\ell P_m G r^{-\alpha_i}\right)\right]\right],$$

$$\stackrel{(g)}{=} \exp\left[-2\pi\lambda_m p_G\int_r^\infty\left(1 - \mathbb{E}_h\left[e^{-s_i h_\ell P_m G t^{-\alpha_i}}\right]\right) t\mathrm{d}t\right],$$

$$\stackrel{(h)}{=} \exp\left[-2\pi\lambda_m p_G\int_r^\infty\left(1 - \frac{1}{(1 + s_i P_m G t^{-\alpha_i}/N_i)^{N_i}}\right) t\mathrm{d}t\right], \tag{35}$$

where (f) follows from the i.i.d. distribution of h and the independence of the PPP. (g) follows by computing the PGFL of the PPP. (h) follows by computing the moment generating function of the gamma random variable h. Applying the integral formula of powers of t and powers of binomials [47], (35) can be written in the form of Gauss hypergeometric function, then the desired result is obtained.

REFERENCES

[1] F. Hu, Y. Deng, W. Saad, M. Bennis, and A. H. Aghvami, "Cellular-connected wireless virtual reality: Requirements, challenges, and solutions," *IEEE Communications Magazine*, vol. 58, no. 5, pp. 105–111, 2020.

[2] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Network*, vol. 32, no. 2, pp. 78–84, 2018.

[3] A. TaghaviNasrabadi, A. Mahzari, J. D. Beshay, and R. Prakash, "Adaptive 360-degree video streaming using layered video coding," in *2017 IEEE Virtual Reality (VR)*. IEEE, 2017, pp. 347–348.

[4] B. Han, "Mobile immersive computing: Research challenges and the road ahead," *IEEE Communications Magazine*, vol. 57, no. 10, pp. 112–118, 2019.

[5] M. G. Kibria, K. Nguyen, G. P. Villardi, W.-S. Liao, K. Ishizu, and F. Kojima, "A stochastic geometry analysis of multiconnectivity in heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9734–9746, 2018.

[6] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, "A latency and reliability guaranteed resource allocation scheme for lte v2v communication systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3850–3860, 2018.

[7] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for vehicular communications with low latency and high reliability," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3887–3902, 2019.

[8] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (urllc)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[9] H. Elshaer, M. N. Kulkarni, F. Boccardi, J. G. Andrews, and M. Dohler, "Downlink and uplink cell association with traditional macrocells and millimeter wave small cells," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6244–6258, 2016.

[10] O. Semiari, W. Saad, and M. Bennis, "Joint millimeter wave and microwave resources allocation in cellular networks with dual-mode base stations," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4802–4816, 2017.

[11] Y. Shi, H. Qu, J. Zhao, and G. Ren, "Downlink dual connectivity approach in mmwave-aided hetnets with minimum rate requirements," *IEEE Communications Letters*, vol. 22, no. 7, pp. 1470–1473, 2018.

[12] C. She, Z. Chen, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Improving network availability of ultra-reliable and low-latency communications with multi-connectivity," *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5482–5496, 2018.

[13] A. Wolf, P. Schulz, M. Dörpinghaus, J. C. S. Santos Filho, and G. Fettweis, "How reliable and capable is multi-connectivity?" *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1506–1520, 2018.

[14] H. Guo, J. Liu, and J. Zhang, "Computation offloading for multi-access mobile edge computing in ultra-dense networks," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 14–19, 2018.

[15] H. Guo, J. Liu, J. Zhang, W. Sun, and N. Kato, "Mobile-edge computation offloading for ultradense iot networks," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4977–4988, 2018.

[16] Y. Wu, L. P. Qian, J. Zheng, H. Zhou, and X. S. Shen, "Green-oriented traffic offloading through dual connectivity in future heterogeneous small cell networks," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 140–147, 2018.

[17] C. Li, H. Wang, and R. Song, "Intelligent offloading for noma-assisted mec via dual connectivity," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2802–2813, 2021.

[18] E. Bastug, M. Bennis, M. Médard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 110–117, 2017.

[19] S. Mangiante, G. Klas, A. Navon, Z. GuanHua, J. Ran, and M. D. Silva, "Vr is on the edge: How to deliver 360 videos in mobile networks," in *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, 2017, pp. 30–35.

[20] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, 2016.

[21] T. Q. Dinh, J. Tang, Q. D. La, and T. Q. Quek, "Offloading in mobile edge computing: Task allocation and computational frequency scaling," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3571–3584, 2017.

[22] Y. Mao, J. Zhang, S. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5994–6009, 2017.

[23] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 4924–4938, 2017.

[24] S. Bi, L. Huang, and Y.-J. A. Zhang, "Joint optimization of service caching placement and computation offloading in mobile edge computing systems," *IEEE Transactions on Wireless Communications*, 2020.

[25] Y. Sun, Z. Chen, M. Tao, and H. Liu, "Communications, caching, and computing for mobile virtual reality: Modeling and tradeoff," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7573–7586, 2019.

[26] T. Dang and M. Peng, "Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 7, pp. 1594–1607, 2019.

[27] J. Chakareski, "Viewport-adaptive scalable multi-user virtual reality mobile-edge streaming," *IEEE Transactions on Image Processing*, vol. 29, pp. 6330–6342, 2020.

[28] W. Yi, Y. Liu, and A. Nallanathan, "Cache-enabled hetnets with millimeter wave small cells," *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5497–5511, 2018.

[29] S. Biswas, T. Zhang, K. Singh, S. Vuppala, and T. Ratnarajah, "An Analysis on Caching Placement for Millimeter-Micro-Wave Hybrid Networks," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1645–1662, 2019.

[30] S. Kuang, X. Liu, and N. Liu, "Analysis and optimization of random caching in k -tier multi-antenna multi-user hetnets," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5721–5735, 2019.

[31] Y. J. Chun, M. O. Hasna, and A. Ghrayeb, "Modeling Heterogeneous Cellular Networks Interference Using Poisson Cluster Processes," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2182–2195, 2015.

[32] O. Semiari, W. Saad, M. Bennis, and M. Debbah, "Integrated millimeter wave and sub-6 ghz wireless networks: A roadmap for joint mobile broadband and ultra-reliable low-latency communications," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 109–115, 2019.

[33] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, 2014.

[34] T. Bai, R. Vaze, and R. W. Heath, "Analysis of blockage effects on urban cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 5070–5083, 2014.

[35] M. Di Renzo, "Stochastic geometry modeling and analysis of multi-tier millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 5038–5057, Sep. 2015.

[36] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[37] W. Yuan and K. Nahrstedt, "Energy-efficient cpu scheduling for multimedia applications," *ACM Transactions on Computer Systems*, vol. 24, no. 3, pp. 292–331, 2006.

[38] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4569–4581, 2013.

[39] 3GPP, "5G; Service requirements for next generation new services and markets," 3rd Generation Partnership Project, Rel 15, Tech. Rep., 2018.

[40] R. Parra-Hernandez and N. J. Dimopoulos, "A new heuristic for solving the multichoice multidimensional knapsack problem," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 35, no. 5, pp. 708–717, 2005.

[41] H. A. Le Thi, T. P. Dinh, and H. Van Ngai, "Exact penalty and error bounds in dc programming," *Journal of Global Optimization*, vol. 52, no. 3, pp. 509–535, 2012.

[42] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[43] K.-H. Chiang and N. Shenoy, "A 2-d random-walk mobility model for location-management studies in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 2, pp. 413–424, 2004.

[44] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[45] G. H. Golub, "Calculation of gauss quadrature rules," *Mathematics of computation*, vol. 23, no. 106, pp. 221–230, 1969.

[46] H. Alzer, "On some inequalities for the incomplete gamma function," *Mathematics of Computation*, vol. 66, no. 218, pp. 771–778, 1997.

[47] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.