Aalborg Universitet

Minimizing Latency to Support VR Social Interactions over Wireless Cellular Systems via Bandwidth Allocation

Jihong Park, [†]Petar Popovski, and [‡]Osvaldo Simeone

Abstract—Immersive social interactions of mobile users are soon to be enabled within a virtual space, by means of virtual reality (VR) technologies and wireless cellular systems. In a VR mobile social network, the states of all interacting users should be updated synchronously and with low latency via two-way communications with edge computing servers. The resulting end-to-end latency depends on the relationship between the virtual and physical locations of the wireless VR users and of the edge servers. In this work, the problem of analyzing and optimizing the end-to-end latency is investigated for a simple network topology, yielding important insights into the interplay between physical and virtual geometries.

Index Terms—Virtual reality (VR), social network, latency, resource management.

I. INTRODUCTION

Virtual reality (VR) is a key use case for 5G [1]–[4]. Its emergence is powered by the recent advances in computing, which enable immersive real-time interactions with virtual objects. As announced by Facebook [5] and Microsoft [6], users will soon be able to interact with each other within virtual communities using VR technologies. In this paper, we consider the problem of supporting *VR-based mobile social networks* over cellular systems by means of edge computing [1]–[4].

A key new element of this challenging problem is the discrepancy between virtual and physical locations of the participating users. In fact, traffic is generated by VR communities in a virtual space, but the supporting network resource for communication and computation are located within the physical network infrastructure. Therefore, the users in the same VR community may not always be co-located in the physical space. For example in Fig. 1, user C belonging to VR community 1 is close in the physical space to user D affiliated to VR community 2, but far from users A and B in VR community 1. This spatial difference between virtual and physical topologies affects the operation of resource allocation and transmission techniques over both Radio Access Network (RAN) and backhaul.

To elaborate, consider VR mobile users that interact in a virtual space. In order for these interactions to be perceived as natural, the network needs to guarantee low latency of e.g. 10 ms for tactile interactions [1]. At the same time, all user states should be properly synchronized in the shared virtual environment. Each user's end-to-end latency is thus dominated by the user from the VR community that experiences the worst latency, accrued due to communication and processing. This, in turn, depends on the physical distribution of users belonging to the same VR community and on the spatial availability of communication and computation resources within the physical network infrastructure.

J. Park was with Aalborg University, and now he is with Centre for Wireless Communications, University of Oulu, Finland (email: jihong.park@oulu.fi).

[†]P. Popovski is with Department of Electronic Systems, Aalborg University, Denmark (email: petarp@es.aau.dk).

[‡]O. Simeone is with Department of Informatics, King's College London, United Kingdom (email: osvaldo.simeone@kcl.ac.uk).
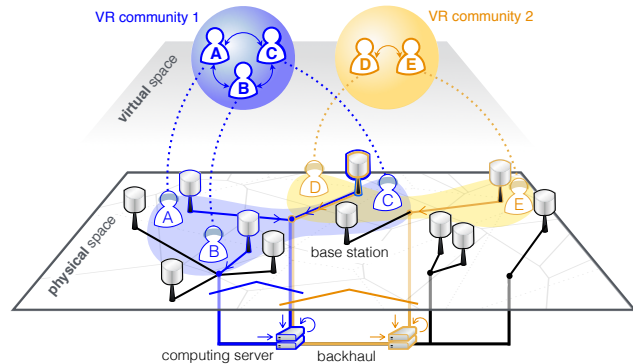
Fig. 1. Illustration of VR mobile social network, where the traffic is generated by virtual-space user interactions and supported by a physical cellular network.

In this work, we study the problem of supporting a VR mobile social network over a multi-cell wireless cellular system with the goal of minimizing the end-to-end latency. Specifically, we focus on the problem of minimizing the end-to-end latency via the bandwidth allocation of the uplink and downlink channels used for communication between users and computing servers. To this end, we formulate a simple model based on a linear cellular topology that captures the interplay between the social interactions within the VR mobile social network and the location of the computation and communication resources within the physical network. The average end-to-end latency is evaluated by accounting for the contributions of uplink, downlink, and backhaul transmissions, as well as for processing times at the servers. The resulting latency is minimized through a stochastic optimization technique.

Related Works – Current VR headsets provide wireless connections via WiFi and/or WiGig (60 GHz) technologies using unlicensed frequency bands [7]. The resulting short-range barrier can be overcome by enabling 5G wireless connections. For such 5G-enabled VR headsets, computing tasks will be conceivably offloaded to edge-cloud servers, in order to overcome the restrictions brought by the limited computing capability and battery capacity of mobile devices.

The required wireless capacity needed to support immersive VR experiences has recently been investigated in [2]. To minimize the VR traffic volume, a caching approach has been proposed in [3]. In a VR theater scenario, a multicast design has been studied in [4]. These works [2]–[4] focus only on optimizing the downlink operations. For augmented reality (AR) applications, the optimization of both uplink and downlink transmissions in terms of end-to-end latency has been studied in [8] for a single-cell scenario. Due to the focus on AR, end-to-end latency model of reference [8] does not take into account virtual social interactions. Finally, virtual social interactions underlie Massively Multiplayer Online game (MMO) applications such as Second Life [9]. Within more restricted virtual spaces, immersive VR social interactions have been recently provisioned by Facebook and Microsoft [5], [6].
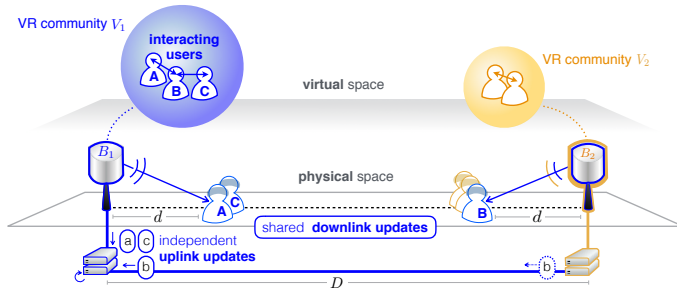
Fig. 2. An illustration of a virtual space traffic flows in a one-dimensional physical network model. In the virtual space, VR users A, B, and C interact with each other within VR community V_1. Their uplink updates a, b, and c are sent to the cloud computing server at B_1 via *unicast* transmissions. The resulting downlink update needs to be sent to all three users for synchronous interactions via *multicast* transmissions

II. SYSTEM MODEL

This section describes the physical network infrastructure, including RAN, backhaul, and computation resources, as well as the VR data traffic model. To focus on the key ideas, we consider a one-dimensional physical network model with two VR communities, as well as two base stations (BSs) in the physical space, as illustrated in Fig. 2.

We use the subscript $i \in \{1, 2\}$ to indicate VR communities 1 or 2. The subscript $j \in \{1, 2\}$ identifies the two BSs. The superscript X $\in \{\text{up}, \text{dn}\}$ describes uplink (up) or downlink (dn) operations at a BS.

A. Physical Network and Channel Model

The network under study comprises a set of users $\mathcal{K} = \{1, 2, \cdots, N\}$ and two BSs B_1 and B_2. Each BS is equipped with a computing server that supports a single VR community. The computing server for VR community V_i is located at BS B_i, unless otherwise specified. In the virtual space, each VR community V_i includes a subset of users $\mathcal{V}_i = \{1, 2, \cdots, N_{V_i}\} \subseteq \mathcal{K}$, representing a fraction $p_{V_i} = N_{V_i}/N$ of the set of users, with $p_{V_1} + p_{V_2} = 1$. Furthermore, a subset of users $\mathcal{B}_j = \{1, 2, \cdots, N_{B_j}\} \subseteq \mathcal{K}$ associates with BS B_j for both uplink and downlink transmissions in the physical space, representing a fraction $p_{B_j} = N_{B_j}/N$, with $p_{B_1} + p_{B_2} = 1$. There are four types of possible user assignments in the virtual and physical spaces, partitioning the set of users into four subsets that are defined as $\mathcal{N}_{ij} = \mathcal{V}_i \cap \mathcal{B}_j$ with $i \in \{1, 2\}$ and $j \in \{1, 2\}$. Each type $\mathcal{N}_{ij}$ includes $N_{ij} \leq \min\{N_{V_i}, N_{B_j}\}$ users, representing a fraction $p_{ij} = N_{ij}/N$. Note that we have the equalities $p_{i1} + p_{i2} = p_{V_i}$, and $p_{1j} + p_{2j} = p_{B_j}$.

The two BSs are located at the edges of a one-dimensional physical space with length D, and are connected by a wired backhaul. The BSs use disjoint spectrum bands, hence not interfering with each other. BS B_j assigns orthogonal bands W^{up} to uplink and to W^{dn} downlink following Frequency Division Duplex (FDD). In the uplink, BS B_j serves each of the N_{B_j} associated users via Frequency Division Multiple Access (FDMA) using *unicast* transmissions. In the downlink, instead, each BS B_j uses orthogonal *multicast* transmissions in order to update the users in the two VR communities. For a given user configuration $\mathbf{N} = \{N_{11}, N_{12}, N_{21}, N_{22}\}$, we denote by w_{ij}^{up} the bandwidth allocated in the uplink to *each* user from the subset $\mathcal{N}_{ij}$ and by w_{ij}^{dn} the bandwidth allocated for multicasting to *all users* in $\mathcal{N}_{ij}$. The users associated with any BS are located at a distance $d \leq D/2$ as illustrated in Fig. 2. For the purpose of obtaining worst-case performance

results, all users are assumed to lie at the maximum distance d. Extensions of the analysis to the more general scenario of arbitrary user-BS distances that are upper bounded by d are possible, but call for more cumbersome notation.

Users in the subsets $\mathcal{N}_{ij}$ with $j \neq i$, denoted as *cross-type* users, are in VR community V_i, but are associated with BS B_j. The uplink data of these cross-type users must be forwarded through a wired backhaul to BS B_i in order to be processed by its attached server. As defined below, each backhaul transmission entails a random delay with the average value proportional to the distance D and the data size.

For the given physical distance d between a BS and the assigned user, the signal-to-noise ratio (SNR) is determined by path loss attenuation $d^{-\alpha}$ for $\alpha \geq 2$ and by independent Rayleigh fading. Therefore, the SNR in uplink or downlink for a given user is

$$\mathsf{SNR}^X = P^X d^{-\alpha} g/\sigma^2, \tag{1}$$

where P^X denotes transmit power and σ^2 indicates the noise variance. The term g represents a small-scale fading coefficient that follows an exponential distribution with unitary mean. These coefficients are independently and identically distributed (i.i.d.) across users in uplink and downlink. We assume the use of type-I Hybrid Automatic Repeat reQuest (HARQ), while the instantaneous SNR information is unknown at the BSs.

B. Virtual Space Traffic

In the virtual space, all the N_i VR users in community V_i are assumed to interact with each other, directly or indirectly, as seen in Fig. 2. In order to enable these virtual interactions, each user uploads its uplink state update message with size b^{up} at regular time intervals, and all the N_i users download the common downlink update message with size b^{dn}. We hereafter consider a fixed users' allocation in physical and virtual spaces given by $\mathbf{N} = \{N_{ij}\}$. With this given user configuration $\mathbf{N}$, we focus on a reference user o. This user o is uniformly randomly selected in the set of N users, and thus has a type $\mathcal{N}_{ij}$ with probability $p_{ij} = N_{ij}/N$. For a single state update, the VR traffic of user o is characterized by the following phases.

- **Step 1 (Upload)** – The user $o \in \mathcal{N}_{ij}$ uploads its update message with size b^{up} bits to the associated BS B_j. If user o is of cross-type, i.e. $j \neq i$, the uplink data is forwarded to the desired computing server at B_i through the inter-BS wired backhaul;
- **Step 2 (Compute)** – The computing server at BS B_i collects all input data from user o as well as from its interacting $N_{V_i} - 1$ users, and then produces their synchronous output states;
- **Step 3 (Download)** – The output states are updated with a common message of size b^{dn} bits to all N_{V_i} users through wireless and, for cross-type users, backhaul links.

Note that, in order to carry out Step 2, the computing server needs to collect data from all N_{V_i} users in VR community V_i. For this reason, the delay prior to computing is limited by the user with *the worst uploading delay in VR community* V_i, as described next.

C. Physical Space Delay

In this section, we fix the user configuration $\mathbf{N}$ and the spectrum allocation $\mathbf{w} = \{\mathbf{w}^{\text{up}}, \mathbf{w}^{\text{dn}}\}$ with $\mathbf{w}^{\text{up}} = \{w_{ij}^{\text{up}}\}$ and $\mathbf{w}^{\text{dn}} = \{w_{ij}^{\text{dn}}\}$ for $i, j \in \{1, 2\}$, and analyze the latency of a reference user o with a fixed type $\mathcal{N}_{ij}$. According to the described VR input/output data flow, conditioned on $\mathbf{N}$, $\mathbf{w}$, and the reference user's type, the average end-to-end latency

$T_{ij}(\mathbf{N}, \mathbf{w})$ of user $o \in \mathcal{N}_{ij}$ consists of the average uploading delay $T_i^{\mathrm{up}}(\mathbf{N}, \mathbf{w}^{\mathrm{up}})$, computing delay $T_i^{\mathrm{c}}(\mathbf{N})$, and downloading delay $T_{ij}^{\mathrm{dn}}(\mathbf{w}^{\mathrm{dn}})$ as in

$$T_{ij}(\mathbf{N}, \mathbf{w}) = T_i^{\mathrm{up}}(\mathbf{N}, \mathbf{w}^{\mathrm{up}}) + T_i^{\mathrm{c}}(\mathbf{N}) + T_{ij}^{\mathrm{dn}}(\mathbf{w}^{\mathrm{dn}}). \qquad (2)$$

We now discuss the three terms in (2). First, the average uploading delay $T_i^{\mathrm{up}}(\mathbf{N}, \mathbf{w}^{\mathrm{up}})$ is, as discussed, the worst user's uploading delay for the N_{V_i} users in VR community V_i. Denoting by $D_\ell^{\mathrm{w.up}}(\mathbf{w}^{\mathrm{up}})$ and $D_\ell^{\mathrm{b.up}}$ the instantaneous uplink wireless and backhaul delays for any user $\ell \in \mathcal{V}_i$, the average uploading delay is given as

$$(\textbf{Upload}) \quad T_i^{\mathrm{up}}(\mathbf{N}, \mathbf{w}^{\mathrm{up}}) = \mathsf{E}\big[\max_{\ell \in \mathcal{V}_i} \big\{ D_\ell^{\mathrm{w.up}}(\mathbf{w}^{\mathrm{up}}) + D_\ell^{\mathrm{b.up}} \big\} \big]. \quad (3)$$

In (3), the expectation is taken over the random number of transmission time slots required by the HARQ process as well as the random backhaul delay, as detailed next.

The uplink wireless delay $D_\ell^{\mathrm{w.up}}(\mathbf{w}^{\mathrm{up}})$ in (3) depends on the instantaneous uplink SNRs, which are random due to the small-scale fading coefficients in (1). Specifically, if the instantaneous $\mathrm{SNR}^{\mathrm{up}}$ is no smaller than a target threshold θ^{up}, the received signal is successfully decoded; otherwise, a retransmission occurs. For a target success probability η, such that $\Pr(\mathrm{SNR}^{\mathrm{up}} \geq \theta^{\mathrm{up}}) = \eta$, the threshold is given as

$$\theta^{\mathrm{up}} = P^{\mathrm{up}} \log_2\left(\eta^{-1} \right) / (d^\alpha \sigma^2). \qquad (4)$$

As a result, the number M_ℓ^{up} of transmission attempts by the ℓ-th user follows a geometric distribution with mean $1/\eta$. Measuring the achievable rate via Shannon capacity, each transmission lasts for $b^{\mathrm{up}}/[w_\ell^{\mathrm{up}} \log_2(1 + \theta^{\mathrm{up}})]$ seconds, where the uplink spectrum allocation equals $w_\ell^{\mathrm{up}} = w_{ij}^{\mathrm{up}}$ if user $\ell \in \mathcal{N}_{ij}$. The total uplink wireless transmission delay of the user ℓ is hence given as

$$D_\ell^{\mathrm{w.up}}(\mathbf{w}^{\mathrm{up}}) = M_\ell^{\mathrm{up}} b^{\mathrm{up}} / [w_\ell^{\mathrm{up}} \log_2(1 + \theta^{\mathrm{up}})], \qquad (5)$$

where $M_\ell^{\mathrm{up}} \sim \mathsf{Geometric}\,(\eta)$.

The uplink backhaul delay $D_\ell^{\mathrm{b.up}}$ in (3) is non-zero only for a cross-type user $\ell \in \mathcal{N}_{ij}$, with $j \neq i$. We define the indicator δ_ℓ such that $\delta_\ell = 1$ if user ℓ is of cross-type, and otherwise we have $\delta_\ell = 0$. Following [10], we model the random backhaul delay for cross-type users to follow a Gamma distribution

$$D_\ell^{\mathrm{b.up}} = \delta_\ell B^{\mathrm{up}}, \qquad (6)$$

where $B^{\mathrm{up}} \sim \mathsf{Gamma}(D, c_b b^{\mathrm{up}})$. The constant $c_b > 0$ represents the propagation delay per bit and unit backhaul length, e.g., $c_b = 10^{-8}$ seconds [10]. Note that the average backhaul delay of cross-type users is $c_b D b^{\mathrm{up}}$, and is hence proportional to D and b^{up}.

Second, the average computing delay $T_i^{\mathrm{c}}(\mathbf{N})$ of user $o \in \mathcal{N}_{ij}$ in (2) is the server processing time required to update the VR model based on input data from the N_{V_i} users in community V_i. With the clock speed f_s, the processing of the VR input data of size b^{up} per user requires b^{up}/f_s seconds. The average computing delay is thereby given as:

$$(\textbf{Compute}) \quad T_i^{\mathrm{c}}(\mathbf{N}) = b^{\mathrm{up}} N_i / f_s. \qquad (7)$$

Finally, the average downloading delay $T_{ij}^{\mathrm{dn}}(\mathbf{w}^{\mathrm{dn}})$ of user o in (2) comprises the average downlink wireless transmission delay induced by the HARQ process and the wired backhaul delay. Denoting as $D^{\mathrm{w.dn}}(\mathbf{w}^{\mathrm{dn}})$ and $D^{\mathrm{b.dn}}$ the instantaneous downlink wireless transmission and backhaul delays, the average downloading delay for user o is computed as

$$(\textbf{Download}) \quad T_{ij}^{\mathrm{dn}}(\mathbf{w}^{\mathrm{dn}}) = \mathsf{E}[D^{\mathrm{w.dn}}(\mathbf{w}^{\mathrm{dn}})] + \mathsf{E}[D^{\mathrm{b.dn}}], \quad (8)$$

where the two terms are discussed next.

For downlink transmissions, BS B_j keeps multicasting to the associated N_{ij} users until all of them successfully decode the output update message. Nevertheless, the latency for user $o \in \mathcal{N}_{ij}$ depends solely on its own decoding process. Following the same reasoning as for (5), the number M^{dn} of transmission attempts follows a Geometric distribution, resulting in the average downlink wireless transmission delay

$$\mathsf{E}[D^{\mathrm{w.dn}}(\mathbf{w}^{\mathrm{dn}})] = b^{\mathrm{dn}} / [w_{ij}^{\mathrm{dn}} \eta \log_2(1 + \theta_{ij}^{\mathrm{dn}})]. \quad (9)$$

The downlink SNR threshold $\theta_{ij}^{\mathrm{dn}}$ in (9) is set such that all N_{ij} users can successfully decode their received multicast signals with probability η. Denoting SNR_ℓ as user ℓ's SNR, this implies the condition $\Pr(\min_{\ell \in \mathcal{N}_{ij}} \{\mathsf{SNR}_\ell\} \geq \theta_{ij}^{\mathrm{dn}}) = \eta$, which yields

$$\theta_{ij}^{\mathrm{dn}} = P^{\mathrm{dn}} \log_2\left(\eta^{-1} \right) / (d^\alpha \sigma^2 N_{ij}). \qquad (10)$$

For the backhaul term, as in (6), a backhaul transfer occurs when the user o is of cross-type. This entails a random delay following $\mathsf{Gamma}(D, c_b b^{\mathrm{dn}})$ distribution, yielding the average downlink backhaul delay

$$\mathsf{E}[D^{\mathrm{b.dn}}] = \delta_o c_b b^{\mathrm{dn}} D, \qquad (11)$$

where $\delta_o = 1$ if user o is of cross-type, and $\delta_o = 0$ otherwise.

III. END-TO-END LATENCY MINIMIZATION

The objective of this section is to tackle the minimization of the average end-to-end latency studied in the previous section with respect to the spectrum allocation $\mathbf{w}$ for a given user configuration $\mathbf{N}$. Besides the average taken in (2) over HARQ and backhaul delays, here we further consider the expectation over the choice of a reference user o. Assuming a uniformly distributed selection, user o belongs to the type $\mathcal{N}_{ij}$ with probability $p_{ij} = N_{ij}/N$, in which case it experiences the average end-to-end latency $T_{ij}(\mathbf{N}, \mathbf{w})$ in (2). The optimal spectrum allocation $\mathbf{w}^* = \{\mathbf{w}^{\mathrm{dn}^*}, \mathbf{w}^{\mathrm{up}^*}\}$ minimizing the average end-to-end latency of user o is then given as

$$\mathbf{w}^* = \underset{\{\mathbf{w}^{\mathrm{dn}}, \mathbf{w}^{\mathrm{up}}\}}{\arg\min} \sum_{j=1}^{2} \sum_{i=1}^{2} p_{ij} T_{ij}(\mathbf{N}, \mathbf{w}) \qquad (12)$$

$$\text{s.t.} \quad \sum_{i=1}^{2} w_{ij}^{\mathrm{dn}} = W^{\mathrm{dn}}, \ w_{ij}^{\mathrm{dn}} \geq 0 \quad \forall j \in \{1, 2\} \qquad (13)$$

$$\sum_{i=1}^{2} w_{ij}^{\mathrm{up}} N_{ij} = W^{\mathrm{up}}, \ w_{ij}^{\mathrm{up}} \geq 0 \quad \forall j \in \{1, 2\}. \qquad (14)$$

The constraints (13) and (14) impose that each BS entirely allocates its available bandwidth for downlink multicast and uplink unicast transmissions, respectively.

Due to the additive form of the end-to-end latency (2) and to the distinct constraints (13) and (14), the optimal allocation $\mathbf{w}^*$ can be achieved by minimizing (12) separately with respect to the downlink allocation $\mathbf{w}^{\mathrm{dn}}$ and the uplink allocation $\mathbf{w}^{\mathrm{up}}$.

For the downlink, using (8), one needs to equivalently minimize the objective function $N_{1j}/w_{1j}^{\mathrm{dn}} + N_{2j}/w_{2j}^{\mathrm{dn}}$ under the constraint (13). This convex problem can be solved by applying the Karush-Kuhn-Tucker (KKT) conditions, yielding the optimal downlink allocation

$$w_{ij}^{\mathrm{dn}^*} = W^{\mathrm{dn}} \sqrt{N_{ij}} / (\sqrt{N_{1j}} + \sqrt{N_{2j}}). \qquad (15)$$

For uplink spectrum allocation, using (3), the said problem is equivalent to minimizing the objective function $\sum_{i=1}^{2} N_{V_i} \mathsf{E}[\max_{\ell \in \mathcal{V}_i} \{a M_\ell^{\mathrm{up}}/w_\ell^{\mathrm{up}} + D_\ell^{\mathrm{b.up}}\}]$ under the constraint (14), with $a = b^{\mathrm{up}}/\log_2(1 + \theta^{\mathrm{up}})$. This is a convex but non-differentiable stochastic problem. It can be tackled by

means of the stochastic approximation method, whereby the objective function is estimated via its empirical mean with a number T of independent samples of the relevant random variables [11], as further detailed below. The resulting problem can be solved by applying the projected subgradient method with some number K of iterations [12]. Convergence to the global optimum $\mathbf{w}^{\mathrm{up}^*}$ is guaranteed for sufficiently large T and K under suitable technical conditions [11], [12].

To elaborate, we draw T independent samples of the transmission attempt numbers $\{M_{\ell,t}^{\mathrm{up}}\}$ and of the backhaul delays $\{D_{\ell,t}^{\mathrm{b.up}}\}$, and consider the empirical objective function $1/T \sum_{t=1}^{T} \sum_{i=1}^{2} N_{V_i} \max_{\ell \in \mathcal{V}_i}\{aM_{\ell,t}^{\mathrm{up}}/w_\ell^{\mathrm{up}} + D_{\ell,t}^{\mathrm{b.up}}\}$. For a given spectrum allocation $w_\ell^{\mathrm{up}(k)} \in \{w_{ij}^{\mathrm{up}(k)}\}$ at the k-th iteration of the subgradient method, we define the index $\ell_{i,t}^{(k)}$ of the user inducing the worst uploading delay in VR community V_i for the t-th sample as $\ell_{i,t}^{(k)} = \arg\max_{\ell \in \mathcal{V}_i}\{aM_{\ell,t}^{\mathrm{up}}/w_\ell^{\mathrm{up}(k)} + D_{\ell,t}^{\mathrm{b.up}}\}$. The subgradient $g_{ij}^{(k)}$ and the uplink allocation $\tilde{w}_{ij}^{\mathrm{up}(k+1)}$ for the next $(k + 1)$-th iteration, prior to the enforcement of the constraint (14), are then given as

$$g_{ij}^{(k)} = -\frac{aN_{V_i}}{T(w_{ij}^{\mathrm{up}(k)})^2} \sum_{t=1}^{T} \delta_{ij}^{(k)} M_{\ell_{i,t}^{(k)},t}^{\mathrm{up}} \quad \text{and} \quad (16)$$

$$\tilde{w}_{ij}^{\mathrm{up}(k+1)} = w_{ij}^{\mathrm{up}(k)} - \beta g^{(k)}, \quad (17)$$

respectively with a step size $\beta > 0$. The indicator function equals $\delta_{ij}^{(k)} = 1$ if $\ell_{i,t}^{(k)} \in \mathcal{N}_{ij}$, and $\delta_{ij}^{(k)} = 0$, otherwise. The next-iterate uplink allocation $\{w_{ij}^{\mathrm{up}(k+1)}\}$ is finally obtained by projecting $\{\tilde{w}_{1j}^{\mathrm{up}(k+1)}, \tilde{w}_{2j}^{\mathrm{up}(k+1)}\}$ in (17) onto the segment (14). This can be easily seen to yield as follows: (i) if $\tilde{w}_{1j}^{\mathrm{up}(k+1)} < (\tilde{w}_{2j}^{\mathrm{up}(k+1)} - W^{\mathrm{up}}/N_{2j})N_{1j}/N_{2j}$, then $w_{1j}^{\mathrm{up}(k+1)} = 0$ and $w_{2j}^{\mathrm{up}(k+1)} = W^{\mathrm{up}}/N_{2j}$; (ii) if $\tilde{w}_{2j}^{\mathrm{up}(k+1)} < (\tilde{w}_{1j}^{\mathrm{up}(k+1)} - W^{\mathrm{up}}/N_{1j})N_{2j}/N_{1j}$, then $w_{1j}^{\mathrm{up}(k+1)} = W^{\mathrm{up}}/N_{1j}$ and $w_{2j}^{\mathrm{up}(k+1)} = 0$; and (iii) otherwise,

$$w_{1j}^{\mathrm{up}(k+1)} = \frac{N_{1j}W^{\mathrm{up}} - N_{1j}N_{2j}\tilde{w}_{2j}^{\mathrm{up}(k+1)} + (N_{2j})^2\tilde{w}_{1j}^{\mathrm{up}(k+1)}}{(N_{1j})^2 + (N_{2j})^2}, \quad (18)$$

$$w_{2j}^{\mathrm{up}(k+1)} = \frac{N_{2j}W^{\mathrm{up}} - N_{1j}N_{2j}\tilde{w}_{1j}^{\mathrm{up}(k+1)} + (N_{1j})^2\tilde{w}_{2j}^{\mathrm{up}(k+1)}}{(N_{1j})^2 + (N_{2j})^2}. \quad (19)$$

In the next section, the optimal allocation scheme is compared with the equal allocation baselines for downlink multicast and uplink unicast, which are given as

$$(\text{Equal Dn}) \ w_{ij}^{\mathrm{dn}} = W^{\mathrm{dn}}/2 \ \text{and} \ (\text{Equal Up}) \ w_{ij}^{\mathrm{up}} = W^{\mathrm{up}}/N_{B_j}. \quad (20)$$

IV. NUMERICAL RESULTS AND DISCUSSION

In this section, we evaluate the end-to-end latency of a VR mobile social network under different user configurations in virtual and physical spaces. The virtual and physical geometries determine the number of cross-type users. To capture this important parameter, we define $0 \leq \rho_c \leq 1$ as the ratio of cross-type users. Specifically, we consider the symmetric setting with equal loads of the two BSs and with an equal fraction of cross-type users for each BS, i.e. $p_{12} = p_{21} = \rho_c/2$ and $p_{11} = p_{22} = (1-\rho_c)/2$, which satisfies $p_{V_i} = p_{B_j} = 0.5$. Other simulation parameters are given as: $N = 50$, $b_i^{\mathrm{X}} = 1$ kbit, $c_b = 10^{-8}$ seconds, $f_s = 2$ GHz, $W^{\mathrm{X}} = 1$ GHz, $\eta = 0.7$, $P^{\mathrm{up}}/\sigma^2 = 20$ dB, $P^{\mathrm{dn}}/\sigma^2 = 30$ dB, $\alpha = 3$, $D = 500$ m, and $d = 15$ m. For the proposed stochastic optimization method, we set $T = K = 50$.

Fig. 3 shows that the average end-to-end latency versus the cross-type user ratio ρ_c. The latency is seen to increase monotonically with ρ_c, owing to the increasing backhaul delays
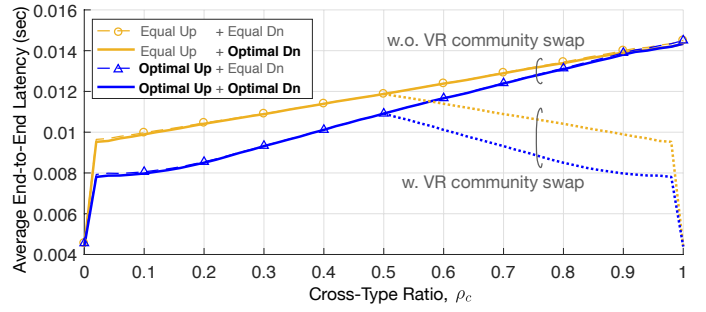

Fig. 3. Average end-to-end latency of a uniformly randomly selected user o as cross-type ratio ρ_c increases, where cross-type users connect to the BS that does not run the corresponding VR community.

of cross-type users. The proposed optimal uplink spectrum allocation is able to partially compensate for the backhaul delays by providing more spectrum to cross-type users. As seen in the figure, this yields up to 25.1% latency reduction as compared to the equal allocation in the uplink for the intermediate value of ρ_c. Note that optimal uplink spectrum allocation cannot improve the latency performance when we have $\rho_c = 0$, i.e., no cross-type users, or $\rho_c = 1$, i.e., all cross-type users. For such cases, all uploading user delays are identically distributed, and it is thus not possible to prioritize spectrum allocation to any group of users.

We also observe that the mentioned gains are achieved by and large even with equal downlink bandwidth allocation, since the end-to-end latency is mostly dictated by the worst uploading delay. This is due to the fact that, thanks to multicasting and to the larger transmission power of the BSs, downlink transmission delays are typically shorter than unicast uplink transmission delays.

Finally, for $\rho_c > 0.5$, it is beneficial to support VR community V_i on the computing server at BS B_j with $j \neq i$, as seen via the dotted curves in Fig. 3. This is obtained by swapping the two VR communities V_1 and V_2. The said observation emphasizes the importance of the computing server locations, which is an interesting topic for further research. Another possible extension of this work is to incorporate stochastic VR interactions that depend on the virtual-space user locations.

REFERENCES

[1] ABI Research and Qualcomm, "Augmented and Virtual Reality: The First Wave of 5G Killer Apps," *White Paper*, Feb. 2017.
[2] E. Baştuğ, M. Bennis, M. Médard, and M. Debbah, "Towards Interconnected Virtual Reality: Opportunities, Challenges and Enablers," *IEEE Commun. Mag.*, vol. 55, pp. 110–117, Jun. 2017.
[3] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Edge Computing Meets Millimeter-wave Enabled VR: Paving the Way to Cutting the Cord," *to appear in Proc. IEEE WCNC 2018*.
[4] Athul Prasad, Mikko A. Uusitalo, David Navrátil, and Mikko Säily, "Challenges for Enabling Virtual Reality Broadcast Using 5G Small Cell Network," *to appear in Proc. IEEE WCNC Wksp. CmMmW5G 2018*.
[5] Facebook Spaces, *Webpage [Online]. URL: http://facebook.com/spaces*.
[6] AltspaceVR, "AltspaceVR joins Microsoft," *[Online]. URL: http://altvr.com/joining-microsoft*.
[7] Displaylink, *Webpage [Online]. URL: http://www.displaylink.com/vr*.
[8] A. AL-Shuwaili and O. Simeone, "Energy-Efficient Resource Allocation for Mobile Edge Computing-Based Augmented Reality Applications," *IEEE Wireless Commun. Lett.*, vol. 6, pp. 398–401, Apr. 2017.
[9] Second Life, *Webpage [Online]. URL: http://secondlife.com*.
[10] G. Zhang, T. Q. S. Quek, M. Kountouris, A. Huang, and H. Shan, "Fundamentals of Heterogeneous Backhaul Design Analysis and Optimization," *IEEE Trans. Commun.*, vol. 64, pp. 876–889, Feb. 2016.
[11] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Stochastic Convex Optimization," *Proc. COLT*, 2009.
[12] S. Boyd, "Subgradient Methods," *notes for EE394b, Stanford University*, Spring 2013-2014.