Ultra-Reliable and Low-Latency Communications Using Proactive Multi-cell Association

Chun-Hung Liu, Di-Chun Liang, Kwang-Cheng Chen and Rung-Hung Gau

Abstract

To attain reliable communications traditionally relies on a closed-loop methodology but inevitably incurs a good amount of networking latency thanks to complicated feedback mechanism. Such a closed-loop methodology thus shackles the current cellular network with a tradeoff between high reliability and low latency. To completely avoid the latency induced by closed-loop communications, this paper aims to study how to jointly employ open-loop communications and multi-cell association in a heterogeneous network (HetNet) so as to achieve ultra-reliable and low-latency communications (URLLC). We first introduce how URLLC mobile users in a large-scale HetNet adopt the proposed proactive multi-cell association (PMCA) scheme to form their virtual cell that consists of multiple access points (APs) and then analyze the communication reliability and latency performances. We show that the communication reliability can be significantly improved by the PMCA scheme and maximized by optimizing the densities of the users and the APs. The analyses of the uplink and downlink delays are also accomplished, which show that extremely low latency can be fulfilled in the virtual cell of a user if the APs are deployed sufficiently for a given user density and the radio resources of each AP are appropriately allocated.

Index Terms

Ultra reliable and low latency communications, coverage, open-loop communications, cell association, heterogeneous network, stochastic geometry.

I. INTRODUCTION

The international telecommunication union has identified ultra-reliable low-latency communication (URLLC), machine-type communication (mMTC) and enhanced mobile broadband (eMBB) as the three main services in the fifth generation (5G) mobile communication that aims to provide good connectivity for many various communication applications [1]–[3]. Among these three services, URLLC remains the most challenging technology due to the need of completely new system design in order to achieve the extremely high system reliability and low latency in 5G cellular systems. Existing mobile communication systems, such as long-term evolution (LTE) systems and its predecessors, were prominently designed to achieve the goal of high throughput
in mobile communications, yet they can also achieve highly reliable communications in the physical layer at the expense of complicated *closed-loop* protocol stack to inevitably result in large networking latency of tens to hundreds of milliseconds (ms). This indicates that there exists a tradeoff between high reliability and low latency in system network architecture and subsequent communication protocols of mobile communication networks. Such a reliability-latency tradeoff problem intrinsically impedes the existing cellular systems to extend their services in mission-critical communication contexts with ultra high reliability and low latency constraints, such as wireless control and automation in industrial environments, vehicle-to-vehicle communications for safety and efficiency improvements, the tactile internet which allows controlling both real and virtual objects with real-time haptic feedback [4], [5].

The message transmission time for mission-critical applications needs to be on the order of tens of microseconds because the human reaction time is on the order of tens of milliseconds [6] or less toward 1 msec for fully autonomous application scenarios. The end-to-end latency of the LTE systems is usually in the range of $30 \sim 100$ ms, which cannot be further reduced in that the backbone network of the LTE systems typically uses the best effort delivery mechanism that is not optimized for latency-sensitive services. To reduce the end-to-end latency in the LTE systems, it is necessary to fundamentally change the system architecture relying on the closed-loop communications and backbone links. The latency of the backbone link can be significantly reduced by appropriate communication architecture and implementation of network protocols to construct the dedicated connection for URLLC services [7]. To reduce the latency in the physical layer, transmission overhead needs to be suppressed by streamlining the grant-free transmission mechanism of the physical layer access and allocating resources properly [8]. Nonetheless, reducing the latency in the communication and backbone links is still insufficient to effectively perform (ultra) low-latency transmission in the current LTE systems because most of the transmission latency is incurred by the control signaling (e.g., grant and pilot signaling usually takes $0.3 \sim 0.4$ ms per scheduling). Accordingly, the most important and essential means that enables URLLC in 5G heterogeneous cellular networks (HetNets) toward the target latency of one millisecond is to completely redesign the transmission protocols in the physical layer of the HetNets [5], [9].

### A. Motivation and Prior Related Work

To effectively reduce latency in 5G HetNets, the essential approach is to adopt feedback-free *open-loop communications* so that no retransmission is needed and receivers can save time in performing additional processing and protocol. Indeed, this approach seems to be the best we can do on the receiver side since retransmission latency is completely avoided. Open-loop communications has a distinct advantage to significantly reduce control signaling overhead relative to closed-loop communications for power control and channel estimation in the traditional and current cellular systems. As such, in this paper we focus on how to fulfill URLLC through
proactive open-loop communications in a HetNet owing to the fact that extremely reliable open-loop communications is the key to low latency.

All the existing URLLC works in the literature are hardly dedicated to studying the open-loop communications or without retransmission \[10\]–\[15\]. Some of the recent works focus on how to perform URLLC in wireless systems by employing retransmissions, short packet designs and their corresponding estimation algorithms for point-to-point transmission \[10\]–\[13\]. Reference \[10\], for example, studied the energy-latency tradeoff problem in URLLC systems with hybrid automatic repeat request (HARQ), whereas reference \[11\] proposed an efficient receiver design that is able to exploit useful information in the data transmission period so as to improve the reliability of short packet transmission. There are some of the recent works that investigated resource allocation problems in wireless networks under the URLLC constraint. In \[14\], the authors studied how to minimize the required system bandwidth as well as optimize the resource allocation schemes to maximize URLLC loads, whereas the problem of optimizing resource allocation in the short blocklength regime for URLLC was investigated in \[15\]. Furthermore, there are few recent works that looked into the URLLC design from the perspective of physical-layer system interfaces and wireless channel characteristics. The recent work in \[16\], for example, adopted coding to seamlessly distribute coded payload and redundancy data across multiple available communication interfaces to offer URLLC without intervention in the baseband/PHY layer design. The problem of how URLLC is affected by wireless channel dynamics and robustness was thoroughly addressed in \[17\].

B. Contributions

Although these aforementioned works and many others in the literature provide a good study on how to achieve URLLC and use it as a constraint to optimize the single-cell performance by using the closed-loop communications and retransmissions, they cannot reveal a good network-wise perspective on how interferences from other cells and user/cell association schemes impact the URLLC performance of cellular systems. To accurately and thoroughly exploit the URLLC performances in a cellular network, in this paper we consider a large-scale multi-tier heterogeneous network (HetNet) model in which all (mobile) users and all access points (APs) adopt open-loop communications and URLLC messages received by each AP are sent to its nearby anchor node performing edge computing in order to reduce the communication latency. Our main contributions are summarized as follows.

- To enhance the communication reliability between users and APs by taking advantage of coordinated multi-point (CoMP) transmission/reception, we propose the proactive multi-cell association (PMCA) scheme that allows each user to associate with multiple APs in the uplink and downlink. The distribution of the number of the users associating with an AP in each tier is accurately derived, which is first found to the best of our knowledge. Also, it importantly indicates that the void AP phenomenon that was discovered in single-cell
association still exists in the PMCA scheme and needs to be considered in the analysis of ultra-reliable communications.

- According to the PMCA scheme, a user can associate with $K$ APs and it thus forms its own virtual cell consisting itself and the $K$ APs. The uplink non-collision reliability of a user in the cell of an AP is found for the proactive open-loop communications and we therefore characterize the uplink communication reliability of a virtual cell for the non-collaborative and collaborative AP cases in a low-complexity form.

- From the analyses of the communication reliability in the uplink and downlink, we are able to show that the communication reliability is significantly influenced by the densities of the users and the APs and the number of the APs in a virtual cell. The PMCA scheme indeed improves the communication reliability of a user and achieves 99.999% communication reliability by appropriately deploying APs for a given user density.

- The uplink and downlink end-to-end delays between users and their anchor node are modeled and analyzed. We not only clarify the fundamental interplay among the delays, the number of the APs in a virtual cell and the user and AP densities, but also show that achieving the target latency of one millisecond is certainly possible provided the APs are deployed with a sufficient density for a given user density and the radio resources of each AP are properly scheduled and allocated.

C. Paper Organization

The rest of this paper is organized as follows. In Section II, we first specify the system architecture of a HetNet for URLLC and then introduce the open-loop communications and propose the PMCA scheme. Section III models and analyizes the uplink and downlink communication reliabilities for the PMCA scheme and some numerical results are provided to validate the correctness and accuracy of the analytical results. In Section IV, the end-to-end latency problem for the open-loop communications and PMCA scheme is investigated and some numerical results are also presented to evaluate the latency performance of the open-loop communications and PMCA scheme. Finally, Section V summarizes our findings in this paper and points out some future research issues in URLLC using the PMCA scheme.

II. System Model and Assumptions

In this paper, we consider an interference-limited planar HetNet in which there are two tiers of APs and the APs in the same tier are of the same type and performance. In particular, the APs in the $m$th tier form an independent homogeneous Poisson point process (PPP) of density $\lambda_m$ and they can be expressed as set $\Phi_m$ given by

$$\Phi_m \triangleq \{A_{m,i} \in \mathbb{R}^2 : i \in \mathbb{N}\}, \quad m = \{1, 2\},$$

where $A_{m,i}$ denotes AP $i$ in the $m$th tier and its location. Without loss of generality, we assume the first tier consists of the macrocell APs and the second tier consists of the small cell APs.
A macrocell AP has a much larger transmit power than a small cell AP, whereas the density of the macro AP is much smaller than that of the small cell APs. To effectively achieve URLLC in the HetNet, all APs are connected to their nearby anchor nodes that manage several APs and are co-located with the edge/fog computing facilities, and a cloud radio access architecture comprised of a core network and a cloud is also employed in the HetNet. Macrocell APs and anchor nodes are connected to the core network which helps send complex computing tasks to the cloud for further data processing and management. An illustrative example of the system model depicted here is shown in Fig. 1 (a). In addition, all (URLLC) users in the HetNet also form an independent homogeneous PPP of density $\mu$ and they are denoted by set $\mathcal{U}$ as

$$\mathcal{U} \triangleq \{ U_j \in \mathbb{R}^2 : j \in \mathbb{N} \} ,$$

where $U_j$ stands for user $j$ and its location. Open-loop communications is used in the HetNet, i.e., there is no feedback between a AP and a user. All APs and users are assumed to be equipped with a single antenna. Such an assumption is made because transmitters are unable to acquire their channel state information feedbacks from their corresponding receivers so that the multi-antenna transmission gain cannot be exploited. In the following, we elaborate the main idea of how to employ open-loop communications to achieve URLLC in the HetNet.

A. Open-loop Communications and Proactive Multi-cell Association

As mentioned in Section I, closed-loop communications fundamentally incurs more latency than open-loop communications owing to feedback. This point manifests that open-loop communications turns out to be the best solution to reducing latency from the receiver perspective.
because feedback-related communication latency is completely avoided. However, the reliability performance of wireless communications could be seriously weakened due to lack of feedback transmission in that it cannot be improved by using the hybrid automatic repeat request (ARQ), a combination of high-rate forward error-correcting coding and ARQ error control, which is commonly used by closed-loop communications. This phenomenon reveals that there seemingly exists a tradeoff between latency and reliability in wireless communications. However, this tradeoff can be absolutely alleviated or tackled by ultra-reliable open-loop communications since closed-loop feedback hardly further benefits the reliability of a wireless channel with extremely high reliability.

To create an ultra-reliable open-loop communications context for the users in the HetNet, the users are suggested to proactively associate with multiple APs at the same time so that their communication reliability can be improved by spatial channel diversity and even boosted whenever the CoMP transmission technique is performed between the associated multiple APs. This proactive multi-cell association approach leads to the concept of the virtual cell of users, that is, each user seems to form its own virtual cell that encloses all the APs associated with it \[18\] and an illustrative example of the virtual cell is shown in Fig. 1(a). All the radio resources in a virtual cell can be scheduled and allocated by utilizing the cloud computing technology. Thus, letting users form their virtual cell (i.e., associate with multiple APs) has an advantage in largely reducing control signaling for frequent handovers between small cell APs, which certainly means the handover latency can be reduced. However, a user should not associate with too many APs at the same time because the signaling overhead due to multi-AP synchronization could deteriorate the latency performance of its virtual cell. To clarify the fundamental interplay among reliability, latency and multi-cell association, we formally propose the PMCA scheme in the following and then study its related statistical properties.

B. Proactive Multi-cell Association (PMCA) and Its Related Statistics

Assume that each user in the network is able to associate with \(K\) APs by using the following PMCA scheme. Let \(V_k\) be defined as

\[
V_k \triangleq \arg \left\{ \begin{align*}
&\max_{m,i: A_{m,i} \in \Phi} \{ w_m \| A_{m,i} \|^{-\alpha} \}, \quad k = 1 \\
&\max_{m,i: A_{m,i} \in \{ \Phi_{k-1} \}} \{ w_m \| A_{m,i} \|^{-\alpha} \}, \quad k > 1
\end{align*} \right\}, \quad k = 1, 2, \ldots, (3)
\]

where \(\Phi \triangleq \bigcup_{m=1}^{2} \Phi_{m}\), \(\Phi_{k-1} = \Phi \setminus \bigcup_{j=1}^{k-1} V_j\), \(\alpha > 2\) is the path-loss exponent, positive constant \(w_m\) is the tier-\(m\) cell association bias, and \(\| \cdot \|\) denotes the Euclidean distance between nodes \(X\) and \(Y\). For a typical user located at the origin, \(V_k\) is thus the \(k\)th biased nearest AP of the typical user by averaging out the channel fading gain effect on the user side. More specifically, \(V_k\) is the \(k\)th nearest AP of the typical user if \(w_m = 1\) for all \(m \in \{1, 2\}\), whereas \(V_k\) becomes the \(k\)th strongest AP of the typical user if \(w_m = P_m\) where \(P_m\) is the transmit power of the
tier-$m$ APs. The $K$ APs associated with the typical user can be expressed as a set given by

$$\mathcal{V}_K \triangleq \bigcup_{k=1}^{K} V_k,$$

which is called the virtual cell of the typical user.

According to our results in [19] [20], the distribution of the number of the users associating with an AP is found for the single-cell association scheme. The method of deriving it cannot be directly applied to the case of the PMCA scheme because the cells of the APs are no longer disjoint in the multi-cell association case. Nonetheless, the idea behind the method is still fairly helpful for us to derive the distribution of the number of the users within the cell of an AP for the PMCA scheme, as shown in the following lemma.

**Lemma 1.** Suppose each user in the network adopts the PMCA scheme in (3) to associate with $K$ APs in the network. Let $N_m$ denote the number of the users associating with a tier-$m$ AP and its distribution (i.e., $p_{m,n} \triangleq P[N_m = n]$) can be accurately found as

$$p_{m,n} \approx \frac{\Gamma(n + \zeta_{m,K})}{n!\Gamma(\zeta_{m,K})} \left( \frac{K\mu}{\zeta_{m,K}\bar{\lambda}_m} \right)^n \left( 1 + \frac{K\mu}{\zeta_{m,K}\bar{\lambda}_m} \right)^{-(n+\zeta_{m,K})},$$

where $\Gamma(x) \triangleq \int_0^\infty t^{x-1}e^{-t}dt$ for $x > 0$ is the Gamma function, $\zeta_{m,K} > 0$ is constant and $\bar{\lambda}_m \triangleq \sum_{i=1}^{2} w_i^2 \lambda_i / w_m^2$.

**Proof:** See Appendix A.

To validate the correctness and accuracy of $p_{m,n}$ in (5), we adopt the network parameters for
a two-tier HetNet shown in Fig. 2 to numerically simulate $p_{m,n}$ for $K \in \{1, 2, \ldots, 5\}$. As can be seen in Fig. 2, the simulated results of $p_{m,n}$ accurately coincide with its corresponding analytical results of $p_{m,n}$ found in (5). Hence, (5) is correct and very accurate. Since the result in Lemma 1 is very accurate, there are some important implications that can be drawn. First, we can learn that the average number of the users associating with a tier-$m$ AP is $K\mu/\tilde{\lambda}_m$ and this means the average cell size of a tier-$m$ AP is $K/\tilde{\lambda}_m$ [20], [21]. In other words, the average cell size of an AP increases $K$ times as users associate with $K$ APs. Second, for $K = 1$ users only associate with a single AP so that the cells of the APs do not overlap and the entire network area consists of weighted Voronoi-tessellated cells. For $K > 1$, the cells of the APs may overlap in part and the cell sizes of the APs and the numbers of the users associating with the APs are no longer completely independent. Third, the probability that a tier-$m$ AP is not associated with any users, referred to as the tier-$m$ void probability, can be found as

$$p_{m,0} = \left(1 + \frac{K\mu}{\zeta_{m,K}\tilde{\lambda}_m}\right)^{-\zeta_{m,K}}.$$  

(6)

For a dense cellular network with a moderate user density, this void probability may not be small that the void APs could be a considerable amount in the network. For example, we use the network parameters for simulation in Fig. 2 to find the void probabilities $p_{1,0} = 0.03$ and $p_{2,0} = 0.2$ for $K = 3$ and the void probability of small cell APs is actually not small at all (there are 20% of the small cell APs that are void.). Thus, such a void cell phenomenon for the PMCA scheme, as illustrated in Fig. 1 (b), must be considered in the interference model [19], [22] when the user density is not very large if compared with the density of the small cell APs.

C. The Truncated Shot Signal Process in a Virtual Cell

As the PMCA scheme and the virtual cell of a user introduced in Section II-B, we define the $K$th-truncated shot signal process of the virtual cell of the typical user as follows.

$$S_K \triangleq \sum_{k=1}^{K} H_k W_k \|V_k\|^{-\alpha}.$$  

(7)

where $V_k \in \mathcal{V}_K$ is already defined in (3), $W_k \in \{w_1, w_2\}$ is the cell association bias of AP $V_k$, and it is a non-negative random variable (RV) associated with $V_k$. We call $S_K$ the $K$th-truncated shot signal process because it does not capture the cumulative effect at the typical user of the $K$ random shocks from the $K$ different random locations (i.e., $V_1, \ldots, V_K$), and $H_k W_k \|V_k\|^{-\alpha}$ can be viewed as the impulse function of AP $V_k$ that gives the $H_k W_k$-weighted attenuation of the transmit power of $V_k$ in space. Let $\mathcal{L}_Z(s) \triangleq \mathbb{E}[\exp(-sZ)]$ denote the Laplace transform of

$^1$When $K$ goes to infinity, $S_{\infty} \triangleq \lim_{K \to \infty} S_K$ is traditionally referred to as (complete) Poisson shot noise process [23], [24] since it contains weighted signal powers in a Poisson field of transmitters. Since $S_K$ only contains the signals emitted from the first $K$ weighted nearest transmitters in the network, it is called the $K$th-truncated shot signal process.
a non-negative RV $Z$ for $s > 0$ and some statistical results regarding $S_K$ are presented in the following theorem.

**Theorem 1.** Assume all the $H_k$’s of the $K$th-truncated shot signal process in (7) are i.i.d. exponential RVs with unit mean. If we define $S_{-K} \triangleq S_\infty - S_K$ and $S_\infty \triangleq \lim_{K \to \infty} S_K$, then the Laplace transform of $S_{-K}$ can be explicitly found as

$$\mathcal{L}_{S_{-K}}(s) = \frac{\langle \pi \tilde{\lambda} \rangle^K}{(K-1)!} \int_0^\infty y^{K-1} \exp \left\{ -\pi \tilde{\lambda} y \left[ 1 + \ell \left( sy^{-\frac{\alpha}{2}}, \frac{2}{\alpha} \right) \right] \right\} dy, \quad (8)$$

where $\tilde{\lambda} \triangleq \sum_{m=1}^2 \frac{w_m^2}{\alpha} \lambda_m$, $\theta_m \triangleq P[W_k = w_m] = \frac{w_m^2}{\alpha} \lambda_m / \tilde{\lambda}$ is the probability that a user associates with a tier-$m$ AP, $\ell(y, z)$ for $y, z \in \mathbb{R}_+$ is defined as

$$\ell(y, z) = \frac{y^z}{\sin(z)} - \int_0^1 \frac{1}{y + t^z} dt, \quad (9)$$

and $\text{sinc}(x) \triangleq \frac{\sin(\pi x)}{\pi x}$. For the Laplace transform of $S_K$, it can be explicitly found as

$$\mathcal{L}_{S_K}(s) = \exp \left[ -\frac{\pi \tilde{\lambda} s^2}{\sin(2/\alpha)} \right] \frac{\langle \pi \tilde{\lambda} \rangle^K}{(K-1)!} \int_0^\infty y^{K-1} \exp \left\{ \pi \tilde{\lambda} y \left[ 1 + \ell \left( sy^{-\frac{2}{\alpha}}, \frac{2}{\alpha} \right) \right] \right\} dy. \quad (10)$$

In addition, the upper bound on $P[S_K \geq x]$ can be found as

$$P[S_K \geq x] \leq 1 - \prod_{k=1}^K \left[ 1 - \mathcal{L}_{Y_k} \left( \frac{2ky}{K(K+1)} \right) \right], \quad (11)$$

where $Y_k \sim \text{Gamma}(k, \pi \tilde{\lambda})$ is a Gamma RV with shape parameter $k$ and rate parameter $\pi \tilde{\lambda}$.

**Proof:** See Appendix C.

The above results of Laplace transform in Theorem 1 indicate that in general the closed-form results of $\mathcal{L}_{S_{-K}}$ and $\mathcal{L}_{S_K}$ are unable to be obtained except in some special cases. For instance, letting $s = \theta y^{\frac{\alpha}{2}}$ and $\mathcal{L}_{y^{\frac{\alpha}{2}} S_{-K}}(\theta)$ can be shown as

$$\mathcal{L}_{y^{\frac{\alpha}{2}} S_{-K}}(\theta) = \frac{\langle \pi \tilde{\lambda} \rangle^K}{(K-1)!} \int_0^\infty y^{K-1} e^{-\pi \tilde{\lambda} y [1 + \ell(\theta, \frac{\alpha}{2})]} dy = \left[ 1 + \ell \left( \theta, \frac{2}{\alpha} \right) \right]^{-K}. \quad (12)$$

Nonetheless, we can still resort to some numerical techniques to evaluate the Laplace transforms of $S_{-K}$ and $S_K$ and the distributions of $S_{-K}$ and $S_K$ by numerically evaluating the inverse Laplace transform of $S_{-K}$ and $S_K$. In addition, we are still able to understand the distribution behaviors of $S_K$ from the closed-form lower bound on $P[S_K \geq y]$. Theorem 1 plays an important role in the following analyses of the communication reliability that will be defined in the following subsection.

III. COMMUNICATION RELIABILITY ANALYSIS FOR PROACTIVE MULTI-CELL ASSOCIATION

In this section, we would like to exploit the fundamental performances and limits of the communication reliability of users in the uplink and downlink when the PMCA scheme is
employed in the HetNet. We assume that the orthogonal frequency division multiple access (OFDMA) is adopted in the HetNet and the communication reliability analyses are proceeded in accordance with how the radio resource blocks (RB) in the cell of each AP are requested by a user in the uplink and allocated by an AP in the downlink. We will first specify how users access the RBs of an AP and then propose and analyze the uplink communication reliability. Afterwards, we will continue to study the communication reliability in the downlink case.

A. Analysis of Uplink Communication Reliability

According to the PMCA scheme and the virtual cell of a user specified in Section II-B, our interest here is to study how likely a user is able to successfully access available RBs of an AP and then send its message to the $K$ APs in its virtual cell through the open-loop uplink communications. To establish the uplink access from a user to the $K$ APs, we propose the following PMCA-based radio resource allocation scheme for uplink open-loop communications:

- To make a user have good uplink connections, the user forms its virtual cell by associating with its first $K$ nearest APs. Thus, all the cell association biases in (3) are unity, i.e., $w_m = 1$ for all $m \in \{1, 2\}$.

- Each radio RB serves as the basic unit while scheduling radio resources. Multiple radio RBs in a single time slot are mapped to a single (virtual) radio resource unit (RRU) for transmitting a message. Users are allowed to transmit one message in each time slot.

- Due to lacking of channel state information of each AP in the virtual cell, a user proactively allocates the radio resource in a distributed manner, that is, it has to proactively (e.g., randomly) select RRUs for the $K$ APs in its virtual cell.

Since each user has to randomly select the uplink RRUs in its virtual cell without considering how other users select their RRUs, multiple users in the cell of an AP could share the same RRUs, which leads to transmission collisions as indicated in Fig. II(a). Such intra-cell collisions among users associating with the same AP may directly lead to users’ failures in the uplink transmissions. The probability that there is no uplink collision in the virtual cell, referred to as the uplink non-collision reliability, is found in the following lemma.

**Lemma 2.** Suppose a user adopts the PMCA scheme in (3) to form its virtual cell with $K$ APs. If the probability that the user selects any one of the RRUs for each AP in its virtual cell is $\delta \in (0, 1)$, then the uplink non-collision reliability of each AP in its virtual cell is found as

$$\rho_{ul} = \sum_{m=1}^{2} \vartheta_m \sum_{n=1}^{\infty} p_{m,n}(1 - \delta)^{n-1},$$

where $\vartheta_m = \lambda_m / \sum_{i=1}^{2} \lambda_i$ is the probability that an AP in the virtual cell is from $\Phi_m$. Hence, the uplink non-collision reliability of the user in its virtual cell is

$$\rho_{ul}^u = 1 - [1 - \rho_{ul}]^K.$$
Proof: See Appendix B. Lemma 2 reveals that the uplink non-collision reliability of each AP is mainly influenced by $K$ and $\delta$, e.g., it decreases whenever $p_{m,n}$ decreases by increasing $K$ and/or $\delta$ decreases; thereby, lesser APs in the virtual cell and/or more radio resources may significantly improve the uplink non-collision reliability$^2$. Note that the uplink non-collision reliability of a user may not always increase as $K$ increases since $\rho_{ul}^K \approx K\rho_{ul}$ for $\rho_{ul} \ll 1$ and increasing $K$ in this situation may not increase $\rho_{ul}^K$ because $\rho_{ul}$ decreases in this case. In addition to the uplink collision problem happening to APs, whether a user is able to successfully send its messages to at least one AP in its virtual cell also depends upon all the communication link statuses in the virtual cell. Let $\gamma_{ul}^k$ denote the uplink signal-to-interference ratio (SIR) from a typical user located at the origin to the $k$th AP in the virtual cell, and it can be expressed as
\[
\gamma_{ul}^k \triangleq \frac{h_k q_k \|V_k\|^{-\alpha}}{\sum_{j:U_j \in U_a} h_j q_j \|V_k - U_j\|^{-\alpha}},
\] (15)
where $h_k$ denotes the uplink fading channel gain from the typical user to AP $V_k$, $q_k$ is the transmit power used by the typical user for AP $V_k$, $q_j$ is the transmit power of user $U_j$, $h_j$ is the uplink fading channel gain from $V_k$, and $U_a \subseteq U$ represents the set of the actively transmitting users using the same RRU as the typical user. All uplink fading channel gains are assumed to be i.i.d. exponential RVs with unit mean, i.e., $h_k \sim \exp(1)$ for all $k \in \{1, \ldots, K\}$ and $h_j \sim \exp(1)$ for all $j \in \mathbb{N}_+$. According to (15), we can consider two cases of non-collaborative and collaborative APs to define the uplink communication reliability in a virtual cell. The case of non-collaborative APs corresponds to the situation in which all APs in the virtual cell are not perfectly coordinated so that they cannot perfectly complete CoMP transmission and reception, whereas when all APs in the virtual cell are perfectly coordinated so that they are able to collaboratively do CoMP transmission and reception corresponds to the case of collaborative APs. For the case of non-collaborative APs, the uplink communication reliability is defined as the probability that a message sent by a user in a virtual cell is successfully received by at least one non-collision AP in the virtual cell, and it can be expressed as
\[
\eta_{ul}^K \triangleq \mathbb{P}\left[ \max_{k \in \{1, \ldots, K\}} \{\gamma_{ul}^k \mathbbm{1}(V_k \in \mathcal{V}_{nc}^K)\} \geq \theta\right],
\] (16)
where $\mathbbm{1}(\mathcal{A})$ is the indicator function that is unity if event $\mathcal{A}$ is true and zero otherwise, $\theta > 0$ is the SIR threshold for successful decoding, and $\mathcal{V}_{nc}^K \subseteq \mathcal{V}_K$ is the subset of the APs without collision in set $\mathcal{V}_K$. For the case of collaborative APs, the uplink communication reliability is

$^2$In general, $\vartheta_m$ does not have a significant impact on $\rho_{ul}^K$ in that usually $P_2 \gg P_1$ as well as $\lambda_1 \ll \lambda_2$ and these two condition leads to $\vartheta_1 \ll \vartheta_2$ in most of practical cases.
defined as
\[ \eta_{ul}^K \triangleq P \left[ \frac{S_{ul}^K}{\sum_{j: U_j \in U \setminus V_K} h_j q_j \| V_k - U_j \|^{-\alpha}} \geq \theta \right], \]
(17)

where \( S_{ul}^K \triangleq \sum_{k: V_k \in V_K^c} h_k q_k \| V_k \|^{-\alpha} \). Namely, \( \eta_{ul}^K \) in (17) is the probability that the non-collision APs in the virtual cell jointly and successfully decode the uplink message.\(^3\)

The analytical results of (16) and (17) are summarized in the following theorem.

**Theorem 2.** Suppose each user employs the PMCA scheme in (5) to form its virtual cell with \( K \) APs. If all the \( K \) APs in the virtual are unable to collaborate, the uplink communication reliability defined in (16) is accurately found as
\[ \eta_{ul}^K \approx 1 - \prod_{k=1}^{K} \left\{ 1 - \rho_{ul} \left( 1 + \frac{\delta \theta^2 (1 - p_0)}{\sin(2/\alpha)} \right)^{-k} \right\}, \]
(18)

where \( p_0 \triangleq p_{m,0} \) that is given in (6) with \( w_m = 1 \) for \( m \in \{1, 2\} \) and \( \rho_{ul} \) is given in (13). For the case of \( K \to \infty \), \( \eta_{ul}^\infty \triangleq \lim_{K \to \infty} \rho_{ul}^K \) can be approximately found as
\[ \eta_{ul}^\infty \approx 1 - \exp \left[ -\frac{\rho_{ul}}{\delta \theta^2} \sin \left( \frac{2}{\alpha} \right) \right]. \]
(19)

When all the \( K \) APs in the virtual cell are able to collaborate to jointly decode the uplink message, \( \eta_{ul}^K \) in (17) can be upper bounded by
\[ \eta_{ul}^K \leq \rho_{ul} \left\{ 1 - \prod_{k=1}^{K} \left[ 1 - \left( 1 + \frac{\delta (1 - p_0)}{\sin(2/\alpha)} \left( \frac{2k \theta}{K(K + 1)} \right) \right)^{-\frac{2}{\alpha} k} \right]^2 \right\}. \]
(20)

**Proof:** See Appendix D. \( \blacksquare \)

From the results in Theorem 2, the uplink reliability \( \eta_{ul}^K \) in (18) monotonically increases as \( K \) increases even though increasing \( K \) makes \( p_0 \) reduce and it thereupon reduces the number of the void cells and induces more interference. However, \( \eta_{ul}^K \) suffers from the diminishing returns problem as \( K \) increases so that associating with too many APs may not be an efficient means to significantly improve \( \eta_K \) for a user. In particular, (18) can be used to obtain the following result
\[ \frac{1 - \eta_{ul}^K}{1 - \eta_{ul}^{K-1}} = 1 - \rho_{ul} \left( 1 + \frac{\delta \theta^2}{\sin(2/\alpha)} \right)^{-K}, \]
(21)

which indicates \( (1 - \eta_{ul}^K)/(1 - \eta_{ul}^{K-1}) \approx 1 \) as \( K \gg 1 \) and we thus know \( \eta_{ul}^K/\eta_{ul}^{K-1} \approx 1 \) for large \( K \), i.e., the diminishing returns problem occurs. According to (18)-(20), another two efficient approaches to boosting \( \eta_{ul}^K \) are reducing the probability of scheduling each RRU in each cell...

\(^3\)For the sake of analytical tractability, we consider that non-coherent signal combing happens among all the non-collision APs in the virtual cell even though such a combining leads to a suboptimal SIR performance.
and densely deploying APs in the HetNet, that is, we need small $\delta$ and large $\rho^{ul}$ in that small $\delta$ suppresses the magnitude of the interference and large $\rho^{ul}$ gives rise to a small number of the users associating with an AP. For instance, if $\theta = 1$, $\alpha = 4$, then $\eta^{ul}_\infty \approx 68.21\%$ for $\delta = 0.5$, $\rho^{ul} = 0.9$, and $\eta^{ul}_\infty \approx 99.76\%$ for $\delta = 0.1$, $\rho^{ul} = 0.95$. Note that $\eta^{ul}_\infty$ in (19) characterizes the fundamental limit of the uplink communication reliability if all APs cannot collaborate in the uplink and it can be used to evaluate whether the PMCA and resource allocation schemes can achieve some target value of $\eta^{ul}_K$. If $\alpha = 4$ and $\theta = 1$, for example, we require $\delta \leq 5\%$ in order to achieve $\eta^{ul}_\infty \geq 99.999\%$. In other words, the target reliability 99.999% is not able to be achieved by the PMCA scheme if $\delta > 5\%$. Furthermore, $\eta^{ul}_K$ in (17) is certainly larger than that in (16). However, the lower bound on $\eta^{ul}_K$ in (20) may not be larger than the result in (19). These aforementioned observations will be numerically validated in Section III-C.

B. Analysis of Downlink Communication Reliability

In this subsection, we would like to study the downlink communication reliability of users in their virtual cells. To establish the benchmark performance, we assume that the frequency reuse factor in this cellular network is unity (i.e., all APs can share the entire available frequency band) so that we can evaluate the downlink communication reliability in the worst case scenario of interference. We also assume that each downlink RRU in the cell of each AP is only allocated to one user in the cell and users adopt the PMCA scheme to associate with the first $K$ strongest APs (i.e., $w_m = P_m^m$ in (3)). In the virtual cell of the typical user, the SIR of the link from the $k$th strongest AP to the typical user is defined as

$$\gamma^{dl}_{k} \triangleq \frac{H_k Q_k \|V_k\|^{-\alpha}}{\sum_{i \in \mathcal{V}_k \setminus \mathcal{V}_k} H_i Q_i \|V_i\|^{-\alpha} + \sum_{m,i : A_{m,i} \in \Phi \setminus \mathcal{V}_K} O_{m,i} H_{m,i} P_m \|A_{m,i}\|^{-\alpha}},$$

(22)

where $H_k$ denotes the downlink fading channel gain from $V_k$ to the typical user, $H_{m,i}$ is also the downlink fading channel gain from $A_{m,i}$ to the typical user, and $O_{m,i}$ is a Bernoulli RV that is unity if $A_{m,i}$ is not void and zero otherwise. All $H_i$’s and $H_{m,i}$’s are assumed to be i.i.d. exponential RVs with unit mean. Note that $\mathbb{P}[O_{m,i} = 1] = 1 - p_{m,0}$ and it can be found by using (6). The term $\sum_{m,i : A_{m,i} \in \Phi \setminus \mathcal{V}_K} O_{m,i} H_{m,i} P_m \|A_{m,i}\|^{-\alpha}$ in (22) is the interference from all non-void APs that are not in the virtual cell, whereas the term $\sum_{i \in \mathcal{V}_K \setminus \mathcal{V}_k} H_i Q_i \|V_i\|^{-\alpha}$ in (22) is the intra-virtual-cell interference that is from other $K - 1$ APs in the virtual cell if the $K - 1$ APs in the virtual cell are not coordinated to avoid using the same RRU used by the $k$th AP. This represents the worst case of the downlink SIR of the $k$th AP in the virtual cell. In this case, the downlink communication reliability of a virtual cell with $K$ APs is defined as

$$\eta^{dl}_K \triangleq \mathbb{P}\left[ \max_{k \in \{1, \ldots, K\}} \{\gamma^{dl}_{k}\} \geq \theta \right],$$

(23)

which is the probability that there is at least one AP in the virtual cell that can successfully transmit to the user in the virtual cell. When all the $K$ APs in the virtual cell can collaborate
to eliminate the intra-virtual-cell interference in the virtual cell, the downlink communication reliability can be simply written as

$$\eta_{dl}^K = \mathbb{P}\left[\frac{\sum_{k=1}^{K} H_k Q_k \|V_k\|^{-\alpha}}{\sum_{m,i:A_{m,i} \in \Phi \setminus \mathcal{V}_K} O_{m,i} H_{m,i} P_m \|A_{m,i}\|^{-\alpha}} \geq \theta \right].$$

(24)

The explicit results of $\eta_{dl}^K$ defined in (23) and (24) are found in the following theorem.

**Theorem 3.** Suppose all APs in a virtual cell are not coordinated so that there exists the intra-virtual-cell interference in the virtual cell. The downlink communication reliability in the case of non-collaborative APs defined in (23) is explicitly upper bounded by

$$\eta_{dl}^K \leq 1 - \prod_{k=1}^{K} \left\{1 - \left[1 - \sum_{m} \vartheta_m (1 - p_{m,0}) \right]^k\right\},$$

(25)

where $\vartheta_m \triangleq P_m^{2/\alpha} \lambda_m / \bar{\lambda}$ and $\bar{\lambda} = \sum_{m=1}^{2} P_m^{2} \lambda_m$. When $K$ goes to infinity, $\eta_{dl}^K \triangleq \lim_{K \to \infty} \eta_{dl}^K$ can be approximately found in closed form given by

$$\eta_{dl}^\infty \approx 1 - \exp\left[-\frac{-\text{sinc}\left(\frac{2}{\alpha}\right)}{\delta \theta \bar{\lambda}}\right].$$

(26)

For the case of collaborative APs, the upper bound on $\eta_{K}^{dl}$ in (24) can be found as

$$\eta_{K}^{dl} \leq 1 - \left\{1 - \left[1 + \delta \ell \left(\frac{\theta}{K^{2}+1}, \frac{2}{\alpha}\right) \sum_{m=1}^{2} \vartheta_m (1 - p_{m,0})\right]^{K}\right\}^{-K}.$$

(27)

**Proof:** See Appendix E.

From the results in Theorem 3, we realize that increasing $K$ indeed improves $\eta_{K}^{dl}$ even though it reduces the tier-$m$ void probability $p_{m,0}$. Yet it also suffers from the diminishing returns problem, like the uplink communication reliability. The tier-$m$ void probability $p_{m,0}$ also significantly impacts $\eta_{K}^{dl}$ when the number of the APs in a virtual cell is not large so that densely deploying APs may largely improve $\eta_{K}^{dl}$ since it helps increase $p_{m,0}$. Moreover, $\delta$ can be interpreted as the probability that all APs statically allocate their RRUs with equal probability and it has to be small in order to achieve ultra-reliable communications. The result in (26) that does not depend on the densities of the APs and users is the fundamental limit of the downlink communication reliability when all non-collaborative APs use different RRUs to transmit a message to the same user. It reveals whether ultra-reliable communications can be attained by the PMCA and resource allocation schemes. For example, if $\alpha = 4$ and $\theta = 1$, we need $\delta < 0.055$ to achieve $\eta_{K}^{dl} \geq 99.999\%$, i.e., each RRU cannot be scheduled with a probability more than 5.5% in this case. Otherwise, the PMCA scheme cannot successfully achieve the downlink communication reliability of 99.999% no matter how many APs are associated in a virtual cell. Furthermore, $\eta_{K}^{dl}$ in (27) for the case of collaborative APs is certainly higher than that in (25) and it can also
TABLE I
NETWORK PARAMETERS FOR SIMULATION

| Parameter \ AP Type (Tier \( m \)) | Macrocell AP (1) | Small cell AP (2) |
|-----------------------------------|------------------|------------------|
| Transmit Power \( P_m \) (W)      | 20               | 5                |
| User Density \( \mu \) (users/m^2) | \( 1 \times 10^{-4} \) |                  |
| AP Density \( \lambda_m \) (APs/m^2) | \( 5.0 \times 10^{-6} \) | \( [0.1\mu, \mu] \) |
| RRU Selection Probability \( \delta \) |                  | 0.05             |
| SIR Threshold \( \theta \)       |                  | 1                |
| Path-loss Exponent \( \alpha \)  |                  | 4                |
| Tier-\( m \) Association Bias \( w_m \) (Uplink, Downlink) | (1, \( P_m \)) |                  |

provide some insights into how to schedule resources and deploy APs in the HetNet so as to achieve the predesignated target value of \( \eta_{K}^{dl} \). In the following subsection, some numerical results and discussions will be provided to evaluate the performances of the downlink communication reliability for the PMCA scheme.

C. Numerical Results and Discussions

To validate the analytical results obtained in the previous subsections and evaluate the communication reliability performances of open-loop communications and PMCA, some numerical results are provided in this subsection and they are obtained based on the network parameters in Table I. Our objective here is to see whether the open-loop communications and the PMCA scheme can achieve the uplink and downlink communication reliabilities up to the target value of 99.999% that is one of the reliability requirement for URLLC services in a 5G system. We first show the simulation results of the uplink reliability for the non-collaborative scenario in Fig. 3 and the analytical results corresponding to this scenario are found by (18).

In Fig. 3(a), we are able to see that the simulated uplink communication reliability \( \eta_{K}^{ul} \) for \( K = 5 \) is higher than 99.999% when \( \mu/\lambda_2 \) is roughly smaller than 0.2 and its corresponding analytical result is just slightly higher than it. This validates not only that PMCA and open-loop communications are indeed able to achieve the reliability target as long as the APs are sufficiently and densely deployed in the HetNet, but also that the analytical result in (18) is fairly close to its simulated counterpart so that the received uplink SIRs at different APs that are dependent in theory can be assumed to be independent while deriving (18). Note that \( \eta_{K}^{ul} \) decreases as \( \mu/\lambda_2 \) increases. This phenomenon is because more and more interferences are generated as \( \mu/\lambda_2 \) is getting larger and larger so that more and more APs are associated with users and getting active. Fig. 3(b) illustrates how \( \eta_{K}^{ul} \) increases as \( K \) increases. As shown in the figure, \( \eta_{K}^{ul} \) significantly improves as \( K \) increases from 1 to 3 and it seems not to improve much after \( K > 3 \), which demonstrates the diminishing returns problem of \( \eta_{K}^{ul} \) mentioned above. In light of this, users may significantly improve their uplink communication reliability by only
associating with 3 or 4 APs, which certainly can be implemented in practice since associating too many APs may incur excess control signaling overhead and latency. The simulation results of the uplink communication reliability for the case of collaborative APs are shown in Fig. 4 and their corresponding analytical upper bound is calculated by using (20). First, we can observe that the simulated results of the uplink reliability in Fig. 4(a) are all above the target reliability of 99.999% even as \( \mu/\lambda_2 \) increases up to a high value, which is much better than the...
simulated results in the case of non-collaborative APs. Thus, the uplink communication reliability in the case of collaborative APs is much less sensitive to \( \mu / \lambda_2 \), which is a good thing from the perspective of AP deployment since we do not need to deploy many APs to reduce \( \mu_2 / \lambda_2 \) so as to increase \( \eta_{ul}^K \), especially when the user population in the network is very large. Like the case of non-cooperative APs, we can also see that the analytical upper bound on \( \eta_{ul}^K \) is also very much tight, which validates the correctness and tightness of (20). Fig. 4(b) illustrates how \( \eta_{ul}^K \) increases with \( K \) in the case of collaborative APs and it is obvious that \( \eta_{ul}^K \) reaches up to the target value 99.999% after \( K \geq 3 \), which is better than the result in Fig. 3(b) as expected. Thus, this expounds that the PMCA scheme should be implemented together with the CoMP scheme in order to easily to serve the URLLC traffic.

The simulated results of the downlink communication reliability for the cases of non-collaborative and collaborative APs are shown in Figs. 5 and 6 respectively. In general, these results also have the same ascending/descending curve trends as their corresponding results in Figs. 3 and 5 but there are still some subtle discrepancies between them. For example, we can observe that in general the downlink communication reliability is better than its uplink counterpart so that they all reach the target value of 99.999%. This is because users do not get much interference from their first \( K \) strongest APs owing to PMCA in the downlink whereas APs would get strong interference from their nearby users in the uplink. Furthermore, in the downlink APs do not “blindly” allocate their RRUs so that they do not suffer from the reliability problem of accessing available channels. This reveals that the the bottle neck of successfully achieving URLLC requirements by using open-loop communications and PMCA is the uplink issue and
Fig. 6. Numerical results of the downlink communication reliability $\eta_{dl}^{K}$ for the case of non-collaborative APs: (a) $\eta_{dl}^{K}$ versus $\mu/\lambda_2$ for $K = 5$, (b) $\eta_{dl}^{K}$ versus $K$ for $\mu_2/\lambda_2 = 0.1$ and $\lambda_2 = 10^{-3}$.

we need to pay more attention to the uplink system designs.

IV. MODELING AND ANALYSIS OF COMMUNICATION LATENCY

Thanks to facilitating open-loop communications, the latency caused by control signaling is completely eliminated as earlier explanation. The communication latency between an anchor node and a user is mostly contributed by the transmission delay between the anchor node and its associated APs and the communication delay between the APs and the user associated with them.

The major difference between uplink and downlink is that uplink RRU collisions incur additional channel access delay. In the following, we will first develop the modeling and analysis approach to the uplink communication delay for open-loop communications and PCMA and we then apply the similar approach to characterize the downlink communication delay.

A. Uplink Communication Delay

The uplink communication delay mainly consists of the channel access delay $D_{ac}^{ul}$, uplink transmission delay $D_{se}^{ul}$ and uplink backhaul delay $D_{tr}^{ul}$, and it can be expressed as

$$D^{ul} = D_{ac}^{ul} + D_{tr}^{ul} + D_{ba}^{ul}.$$  \hspace{1cm} (28)

Since there are $K$ APs in a virtual cell and a user has to randomly select an RRU for each of the $K$ APs, RRU collisions could happen in the cells of the $K$ APs. When a user starts to contend

\footnote{Note that our focus in this section is to study the communication delays induced by open-loop communications and the PMCA scheme and thus some delays not directly related to open-loop communications and the PMCA scheme, such as signal processing delays on the transmitter and receiver sides are ignored.}
RRUs in its virtual cell, the channel access delay can be defined as the lapse of time needed by
the user to successfully contend at least one RRU from the $K$ APs. Thus, $D_{ac}^{ul}$ can be modeled
as a geometric RV with parameter $\rho_K^{ul}$ where $\rho_K^{ul}$ is the non-collision probability found in \textcolor{blue}{[14]},
and its mean is $\mathbb{E}[D_{ac}^{ul}] = 1/\rho_K^{ul}$. $\mathbb{E}[D_{ac}^{ul}] = 1/\rho_K^{ul}$ represents the average number of times for a
user to successfully contend a channel and then send a message, i.e., its unit is channel access
attempts/message, and it can be transformed to seconds/message if each channel attempt is equal
to how many seconds needed for transmitting a message. The uplink transmission delay is the
lapse of time needed by the user to successfully transmit a message to at least one of the $K$
APs and it can be modeled as an ergodic process. Hence, the mean of the uplink transmission
delay is found as

$$\mathbb{E}[D_{tr}^{ul}] = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau D_{tr}^{ul}(t) dt \triangleq \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{t=0}^\tau \mathbb{1}\left(\max_k \{\gamma_k^{ul}(t)\} \geq \theta\right),$$

(29)

where the unit of $\mathbb{E}[D_{tr}^{ul}]$ is transmitting times/message and it can be transformed to seconds if
the time duration (seconds) of transmitting a message is determined.

The uplink backhaul delay is defined as the minimum transmission time needed for the APs
in a virtual cell to transmit a message to their anchor node so that it can be written as

$$D_{ba}^{ul} \triangleq \min_{k, \forall_k \in V_{nc}^K} \{D_{ba,k}^{ul}\},$$

(30)

where $D_{ba,k}^{ul}$ denotes the uplink backhaul delay from the $k$th AP to the anchor node. We further
assume that the backhaul link statuses between the $k$th AP and the anchor node independently
vary with time, which gives rise to a reasonable assumption that $D_{ba,k}^{ul}$ is well characterized by
an exponential RV with parameter $\beta$ so that the message arrival process can be modeled as an
independent Poisson process. In light of this, the distribution of $D_{ba}^{ul}$ in the non-collaborative AP
case is

$$\mathbb{P}[D_{ba}^{ul} \leq x] = 1 - \mathbb{E}\left[\prod_{k=1}^N \mathbb{P}[D_{ba,k}^{ul} \geq x]\right]
= 1 - \mathbb{E}\left[\exp(-x; \beta N)\right] = 1 - \left(1 - \eta_K^{ul} e^{-\beta x}\right)^K$$

since all $D_{ba,k}^{ul}$’s are independent and $N$ that denotes the number of the non-collision APs in the
virtual cell is a binomial RV with parameters $K$ and $\eta_K^{ul}$. We thus have the mean of $D_{ba}^{ul}$ found
as follows:

$$\mathbb{E}[D_{ba}^{ul}] = \int_0^\infty \mathbb{P}[D_{ba}^{ul} \geq x] \ dx = \int_0^\infty \left(1 - \eta_K^{ul} e^{-\beta x}\right)^K \ dx.$$  

(31)

If $\eta_K^{ul} \approx 1$, we can get $\mathbb{E}[D_{ba}^{ul}] \approx 1/\beta K$. For the case of coordinated APs, $\mathbb{P}[D_{ba}^{ul} \geq x] = \exp(-x; \beta)$ because only one message is sent from the virtual cell to the anchor node. The mean
of $D_{ba}^{ul}$ is $\mathbb{E}[D_{ba}^{ul}] = 1/\beta$ and the unit of $\mathbb{E}[D_{ba}^{ul}]$ can be set as times/RRU. The mean of the uplink communication delay is readily obtained by

$$
\mathbb{E} [D_{ba}^{ul}] = \frac{1}{\rho_{K}^{ul}} + \frac{1}{\eta_{K}^{ul}} + \left\{ \begin{array}{ll}
\int_{0}^{\infty} (1 - \eta_{K}^{ul} + \eta_{K}^{ul} e^{-\beta x})^K dx, & \text{non-collaborative} \\
\frac{1}{\beta}, & \text{collaborative}
\end{array} \right.,
$$

(32)

which can be employed to evaluate the uplink latency performance.

### B. Downlink Communication Delay

In the downlink, since each AP is able to allocate its resource to its received message, the downlink communication delay $D_{dl}$ mostly consists of the downlink backhaul delay $D_{ba}^{dl}$ and transmission delay $D_{tr}^{dl}$, i.e., it can be simply written as

$$
D_{dl} = D_{ba}^{dl} + D_{tr}^{dl}.
$$

(33)

The downlink backhaul delay is defined as the maximum time elapsed from the start time of sending a message from the anchor node to the end time when all $K$ APs in the virtual cell receive the message. Suppose the message arrival process at each AP can be modeled as an independent Poisson process and the downlink backhaul delay can be expressed as

$$
D_{ba,k}^{dl} \sim \exp(\beta) \quad \text{is the downlink backhaul delay from the anchor node to the } k\text{th AP in the virtual cell.}
$$

(34)

where $D_{ba,k}^{dl} \sim \exp(\beta)$ is the downlink backhaul delay from the anchor node to the $k$th AP in the virtual cell. In light of this, the distribution of $D_{ba}^{dl}$ is found as

$$
\mathbb{P}[D_{ba}^{dl} \geq x] = \left\{ \begin{array}{ll}
1 - \mathbb{P} \left[ \max_{k \in \{1, \ldots, K\}} \{D_{ba,k}^{dl}\} \leq x \right], & \text{Collaborative APs} \\
\mathbb{P} \left[ \min_{k \in \{1, \ldots, K\}} \{D_{ba,k}^{dl}\} \geq x \right], & \text{Non-collaborative APs}
\end{array} \right.,
$$

$$
= \left\{ \begin{array}{ll}
1 - \prod_{k \in \{1, \ldots, K\}} \mathbb{P}[D_{ba,k}^{dl} \leq x], & \text{Collaborative APs} \\
\prod_{k \in \{1, \ldots, K\}} \mathbb{P}[D_{ba,k}^{dl} \geq x], & \text{Non-collaborative APs}
\end{array} \right.
$$

(35)

The downlink transmission delay is the time duration in which the $K$ APs in the virtual cell successfully transmit a message to the user and its mean can be characterized by the downlink communication reliability, i.e., $\mathbb{E}[D_{tr}^{dl}] = 1/\eta_{K}^{dl}$ and its bound can be found by using Theorem 3.

Accordingly, the mean of the downlink communication delay is given by

$$
\mathbb{E} [D_{dl}] = \frac{1}{\eta_{K}^{dl}} + \left\{ \begin{array}{ll}
\int_{0}^{\infty} \left[1 - \left(1 - e^{-\beta x}\right)^K\right] dx, & \text{Collaborative APs} \\
\frac{1}{\beta K}, & \text{Non-collaborative APs}
\end{array} \right.,
$$

(35)

Hence, $\mathbb{E} [D_{dl}]$ increases as $K$ increases in the case of collaborative APs, but it decreases as $K$ increases in the case of non-collaborative APs.
C. Numerical Results

In this subsection, we would like to numerically demonstrate how \( \mathbb{E}[D_{ul}] \) and \( \mathbb{E}[D_{dl}] \) vary with the densities of the APs and users and the number of the APs in a virtual cell. Assume the total available bandwidth is 100 MHz and there are 20 subbands (channels) so that each subband has 5 MHz and \( \delta = 1/20 = 0.05 \). By assuming an URLLC message of 512 bits and \( \theta = 1 \), users/APs thus need at most 0.1024 ms to transmit one message because the transmitting rate is at least \( 5 \times 10^6 \times \log_2(1 + \theta) = 5 \) Mbps. The means of the uplink and downlink communication delays found in the previous subsection need to transform their unit to ms by multiplying 0.1024 ms. Fig. 7 shows the simulation results of the sum of \( \mathbb{E}[D_{ul}] \) and \( \mathbb{E}[D_{dl}] \) based on the network parameters in Table I and \( \beta = 5 \) messages/ms. In Fig. 7, we see that the sum of \( \mathbb{E}[D_{ul}] \) and \( \mathbb{E}[D_{dl}] \) does not vary much with \( \mu/\lambda_2 \) and the delay performance in the case of collaborative APs is worse than that in the case of non-collaborative APs. Fig. 7(b) also shows how the delay performances in both cases change along \( K \). As expected, when \( K \) gets larger, \( \mathbb{E}[D_{ul}] + \mathbb{E}[D_{dl}] \) in the cases of collaborative and non-collaborative APs becomes larger and smaller, respectively. For the case of collaborative APs, the mean of the total delays increases as the number of the APs in the virtual cell increases, which is mainly because the downlink backhaul delay dominates the total delays. Although increasing the number of the APs in a virtual cell indeed improves the communication reliability, it degrades the delay performance. Hence, we should be aware of the trade-off problem between the communication reliability and latency when users associate with multiple APs.
V. CONCLUDING REMARKS

This paper provides an alternative effective approach to achieving URLLC in a HetNet by using open-loop communications and multi-cell association. Such a solution stems from the idea that extremely reliable communications is hardly benefited by receiver’s feedback that causes additional latency, which breaks the longstanding concept of a tradeoff between communication reliability and latency from the prevailing closed-loop communications in the current cellular system and its predecessors. From the perspective of latency, ultra-reliable open-loop communications is the key to ultimately fulfilling the goal of ultra-low latency in the network. To analytically demonstrate this point, the users in the HetNet are assumed to proactively associate with multiple APs in their virtual cell so as to significantly improve their link reliability. The communication reliability problems in the uplink and downlink are accurately modeled and analyzed by considering the void cell phenomenon for the multi-cell association case. Their analytical and simulated results indicate that the target reliability of $0.99999\%$ can be accomplished by the PMCA scheme. The latency problems in the uplink and downlink are studied as well and their analytical and simulated results ensure that ultra-low latency of 1 ms can be achieved by the PMCA scheme if the APs are sufficiently deployed and the number of the APs in a virtual cell is properly chosen. In addition, there are a few interesting approaches able to further improve the communication reliability of the PMCA scheme when the latency constraint of one ms is satisfied. For example, we may devise new multi-user detection techniques that can be performed over multiple APs to extract the message at the receiver without channel state information. We may also design new error correction coding schemes for open-loop communications that can be distributively performed over multiple APs to further enhance the communication reliability at receivers.

APPENDIX

PROOFS OF LEMMAS AND THEOREMS

A. Proof of Lemma [7]

According to Theorem 1 in [20], we can obtain the following result:

$$P\left[ \max_{m,i:A_{m,i} \in \Phi} w_m \| A_{m,i} \|^{-\alpha} \leq x^{-\alpha} \right] = P\left[ \min_{m,i:A_{m,i} \in \Phi} w_m^{\frac{1}{\alpha}} \| A_{m,i} \| \geq x \right]$$

$$= P\left[ \| \tilde{A}_1 \| \geq x \right] = \exp \left( -\pi x^2 \sum_{m=1}^{\infty} \frac{w_m^2}{\alpha} \lambda_m \right) ,$$

where $\| \tilde{A}_1 \| \triangleq \min_{m,i:A_{m,i} \in \Phi} w_m^{\frac{1}{\alpha}} \| A_{m,i} \|$ and $\| \tilde{A}_1 \|^2 \sim \exp(\pi \sum_{m=1}^{\infty} w_m^{2/\alpha} \lambda_m)$. Let $\tilde{\Phi} \triangleq \{ \tilde{A}_k \in \mathbb{R}^2 : k \in \mathbb{N}_+ \}$ be a homogeneous PPP of density $\sum_{m=1}^{\infty} w_m^{2/\alpha} \lambda_m$ and $\tilde{A}_k$ is the $k$th nearest point in $\tilde{\Phi}$ to the origin. Also, we define $\tilde{\lambda}_m \triangleq \sum_{k=1}^{\infty} (w_i/w_m)^{2/\alpha} \lambda_i$ based on the result of Theorem 1 in [22]. This means
that \( \tilde{\Phi}_m \) can be viewed as a sole homogeneous PPP equivalent to the superposition of all independent homogeneous PPPs in the network (i.e., \( \bigcup_{i=1}^{2} \Phi_i \)) when all tier-\( i \) APs in \( \Phi_i \) are scaled by \( (w_m/u_i)^{2/\alpha} \). Thus, (3) can be equivalently expressed as

\[
\|V_k\| \overset{d}{=} \|A_{m,k}\|, \quad k = 1, 2, \ldots,
\]

where \( \overset{d}{=} \) stands for the equivalence in distribution. This result also indicates that the typical user can be imaged to equivalently associate with the first \( K \) nearest APs in \( \tilde{\Phi}_m \).

Since the typical user can associate with its first \( K \) weighted nearest APs in the network, the distance between the typical user and its \( K \) th weighted nearest AP is \( \|V_K\| \) and \( \|V_K\|^2 \overset{d}{=} \|A_{m,K}\|^2 \). This result also indicates that the typical user can be imaged to equivalently associate with the first \( K \) nearest APs in \( \tilde{\Phi}_m \). From [22] and [25], we learn that the Lebesgue measure of \( C_{m,i} \), denoted by \( \nu(C_{m,i}) \), for \( K = 1 \) can be accurately described by a Gamma RV with the following pdf for all \( i \in \mathbb{N}_+ \):

\[
f_{\nu(C_m)}(x) \approx \frac{(\zeta x \lambda_m)^{\zeta}}{x!} e^{-\zeta x \lambda_m}, \quad \text{for } K = 1,
\]

where \( \zeta = \frac{1}{2} \). Since the mean of \( \nu(C_{m,i}) \) for \( K = 1 \) is \( 1/\lambda_m \) and it is equal to \( \mathbb{E}[\pi\|V_1\|^2] \), the mean of \( \nu(C_{m,i}) \) for \( K > 1 \) is equal to \( \mathbb{E}[\pi\|V_K\|^2] = K/\lambda_m \). Accordingly, for \( K > 1 \), \( f_{\nu(C_m)} \) must be equal to

\[
f_{\nu(C_m)}(x) \approx \frac{(\zeta x \lambda_m/K)^{\zeta_m}}{x!} e^{-\zeta_m x \lambda_m/K}, \quad \text{for } K > 1.
\]

Note that all \( \nu(C_{m,i})'s \) have the same distribution and they may not be independent. Let \( \tilde{\Phi}_m(C_{m,i}) \) denote the number of users associating with \( \tilde{A}_{m,i} \) so that \( p_{m,n} \) for \( K > 1 \) can be expressed as

\[
p_{m,n} \overset{d}{=} \mathbb{P}[N_m = n] = \mathbb{P}[\tilde{\Phi}_m(C_{m,i}) = n] = \mathbb{E} \left[ \frac{\lambda_m \nu(C_m) \mu^n}{n!} \exp(-\lambda_m \nu(C_m) \mu) \right] = \int_0^\infty \frac{(\lambda_m x \mu)^n}{n!} \exp(-\lambda_m x \mu) f_{\nu(C_m)}(x) dx,
\]

which can be completely carried out by using the above expression of \( f_{\nu(C_m)}(x) \) for \( K > 1 \). Thus, the result in (5) is obtained.

**B. Proof of Lemma 2**

Let \( M_k \) denote the number of the users associating with the \( k \) th AP in the virtual cell. The probability of no collisions happening in the cell of the \( k \) th AP is \( \delta(1-\delta)^{M_k-1} \) if the probability of a user selecting any one of the RRRUs for each AP is \( \delta \). Suppose the radio resource (available bandwidth) of each AP can be divided into \( R \) radio resource units so that we have \( \delta = 1/R \). Thus,
the probability that an AP in the virtual cell does not have collisions, denoted by \( \rho^{ul} \), can be written as

\[
\rho^{ul} = \sum_{r=1}^{R} \delta \mathbb{E} \left[ (1 - \delta)^{M_{k}-1} \right] = \sum_{r=1}^{R} \frac{1}{R} \mathbb{E} \left[ (1 - \delta)^{M_{k}-1} \right] = \mathbb{E} \left[ (1 - \delta)^{M_{k}-1} \right],
\]

where \( \mathbb{E}[(1 - \delta)^{M_{k}-1}] = \sum_{m=1}^{2} \mathbb{P}[V_{k} \in \Phi_{m}] \mathbb{E}[(1 - \delta)^{M_{k}-1} V_{k} \in \Phi_{m}] \) and we thus have

\[
\rho^{ul} = \sum_{m=1}^{2} \vartheta_{m} \mathbb{E} \left[ (1 - \delta)^{N_{m}-1} \right] = \sum_{m=1}^{2} \vartheta_{m} \sum_{n=1}^{\infty} p_{m,n} (1 - \delta)^{n-1},
\]

where \( \vartheta_{m} = \mathbb{P}[V_{k} \in \Phi_{m}] \) and \( N_{m} \) is the number of the users associating with a tier-\( m \) AP. Moreover, we know that for any point in the network the probability that the nearest AP to the user is from the \( m \)th tier is \( \mathbb{P}[\|A_{m,*}\|^{-\alpha} \geq \|A_{i,*}\|^{-\alpha}] = \mathbb{P}[\|A_{m,*}\|^2 \leq \|A_{i,*}\|^2] \) in which \( A_{i,*} \) is the nearest point in \( \Phi \) to the point and \( m \neq i \). Since \( c^{-1}\|A_{m,*}\|^2 \sim \exp(c\pi\lambda_{m}) \) for any \( c > 0 \), we thus have

\[
\vartheta_{m} = \mathbb{P} \left[ \|A_{m,*}\|^2 \leq \|A_{k,*}\|^2 \right] = \frac{\lambda_{m}}{\sum_{m=1}^{\infty} \lambda_{i}}.
\]

Substituting the above result of \( \vartheta_{m} \) into the above expression of \( \rho^{ul} \) yields the result in (14). In addition, the uplink non-collision reliability of the user is the probability that there is at least one non-collision AP in the virtual and it thus can be expressed as \( \rho^{ul} = 1 - (1 - \rho^{ul})^{K} \), which is equal to the result in (14) by substituting the result of \( \rho^{ul} \) in (13) into \( \rho^{ul} \).

\( C. \) Proof of Theorem 1

(i) First, we derive the Laplace transform of \( S_{-K} \). According to the definition of \( S_{-K} \), \( S_{-K} \) can be alternatively written as

\[
S_{-K} = \sum_{i=K+1}^{\infty} H_{i} W_{i} \|V_{i}\|^{-\alpha} \overset{d}{=} \sum_{j:V_{j} \in \tilde{\Phi}} H_{j} (\|\tilde{V}_{j}\|^2 + \|\tilde{V}_{i}\|^2)^{-\frac{\alpha}{2}},
\]

where \( \tilde{\Phi} \triangleq \{ \tilde{V}_{j} \in \mathbb{R}^2 : \tilde{V}_{j} = W_{j} V_{j}, V_{j} \in \Phi, j \in \mathbb{N}_{+} \} \) is a homogeneous PPP of density \( \tilde{\lambda} \), and \( \|\tilde{V}_{j+i}\|^2 = \|\tilde{V}_{j}\|^2 + \|\tilde{V}_{i}\|^2 \) for all \( i,j \in \mathbb{N}_{+} \) and \( i \neq j \) based on the proof of Proposition 1 in [26]. Therefore, the Laplace transform of \( S_{-K} \) can be calculated as shown in the following:

\[
\mathcal{L}_{S_{-K}}(s) = \mathbb{E} \left[ \exp \left( -s\|\tilde{V}_{K}\|^{-\alpha} \sum_{j:V_{j} \in \Phi} H_{j} \left( 1 + \frac{\|\tilde{V}_{j}\|^2}{\|\tilde{V}_{K}\|^2} \right)^{-\frac{\alpha}{2}} \right) \right]
\]

\( \overset{(a)}{=} \mathbb{E}_{Y_{K}} \left[ \exp \left( -\pi \tilde{\lambda} \int_{0}^{\infty} \mathbb{E}_{H} \left[ 1 - e^{HsY_{K}^{-2} (1 + \frac{r}{Y_{K}^{2}})^{-\frac{\alpha}{2}}} \right] dr \right) \right]
\]

\( \overset{(b)}{=} \mathbb{E}_{Y_{K}} \left[ \exp \left( -\pi \tilde{\lambda} Y_{K} \int_{1}^{\infty} \mathbb{P} \left[ Y_{K} r' \leq \left( \frac{rH}{Z} \right)^{\frac{2}{\alpha}} \right] dr' \right) \right] = \mathbb{E}_{Y_{K}} \left\{ \exp \left[ -\pi \tilde{\lambda} Y_{K} \ell \left( \frac{sH}{Z^{2}}, \frac{2}{Y_{K}^{2}}, \frac{\alpha}{2} \right) \right] \right\} ,
\]
where \(^{(a)}\) follows from the probability generating functional (PGFL) of the homogeneous PPP \(\Phi \) \(^{[27]} \) \(^{[28]} \) and \(Y_k \triangleq \|V_k\|^2\), and \((b)\) is obtained by using \(Z \sim \exp(1)\). Thus, \(L_{S_{-K}}(s)\) is equal to the result in \(^{(8)}\) because \(Y_k \sim \text{Gamma}(k, \pi \tilde{\lambda})\).

(ii) Next, we find the Laplace transform of \(S_K\). According to the definition of \(V_k\) in \(^{(3)}\), we know that \(S_\infty \triangleq \lim_{K \to \infty} S_K\) in \(^{(7)}\) can be equivalently expressed as

\[
S_\infty = \sum_{m:i:A_m,i \in \Phi} H_{m,i} \|A_{m,i}\|^{-\alpha},
\]

where \(^{d}\) denotes the equivalence in distribution, all \(H_{m,i}\)'s are i.i.d. RVs with the same distribution as \(W_k\) and \(\Phi \triangleq \bigcup_{m=1}^2 \Phi_m\). Thus, \(L_{S_\infty}(s)\) can be found as follows:

\[
L_{S_\infty}(s) = \mathbb{E} \left[ \exp \left( -s \sum_{m:i:A_m,i \in \Phi} H_{m,i} \|A_{m,i}\|^{-\alpha} \right) \right] = \mathbb{E} \left[ \exp \left( -s \sum_{k:V_k \in \Phi} H_k \|V_k\|^\alpha \right) \right] = \exp \left( -\pi \tilde{\lambda} s^\frac{\alpha}{2} \mathbb{E} \left[ H^\frac{\alpha}{2} \right] \Gamma \left( 1 - \frac{\alpha}{2} \right) \right) = \exp \left( -\frac{\pi \tilde{\lambda} s^\frac{\alpha}{2}}{\sin(\pi/\alpha)} \right),
\]

where \(^{(a)}\) follows from Theorem 1 \(^{[20]}\) and \(^{(b)}\) follows from the PGFL of the homogeneous PPP \(\Phi \) \(^{[27]}\). Next, \(L_{S_{-K}}(s)\) can be derived as shown in the following:

\[
L_{S_{-K}}(s) = \mathbb{E} \left[ \exp \left( -s \sum_{k:V_k \in V_\infty \setminus V_K} \frac{H_k}{\|V_k\|^\alpha} \right) \right] = \mathbb{E} \left[ \exp \left( -s \sum_{k:V_k \in V_\infty \setminus V_K} \frac{H_k}{\|V_k\|^\alpha} \right) \right] = \int_0^\infty \mathbb{E} \left[ \prod_{k:V_k \in V_\infty} \exp \left( -\frac{s x^{-\frac{\alpha}{2}} H_k}{(1 + \|V_k\|^2/x)^{\frac{\alpha}{2}}} \right) \right] f_{\|\tilde{V}_K\|^2}(x) \, dx \]

\[
\equiv \mathbb{E}_{Y_K} \left\{ \exp \left[ -\pi \tilde{\lambda} Y_K \ell \left( s Y_K^{-\frac{\alpha}{2}}, \frac{2}{\alpha} \right) \right] \right\}
\]

where \(^{(c)}\) is obtained by first finding the PGFL of \(\tilde{V}_\infty\) and we then follow the derivation steps in the proof of Proposition 4 in \(^{[26]}\) to derive function \(\ell(\cdot, \cdot)\).
In addition, we can know the following:

\[
\mathcal{L}_{S_K}(s) = \mathbb{E} \left[ e^{-s(S_K + S_{\infty})} \right] = \mathbb{E} \left[ \exp \left( -sS_K \right) \cdot \exp \left( -s \sum_{k: \tilde{V}_k \in \tilde{V}_\infty \setminus \tilde{V}_K} \frac{H_k}{\|\tilde{V}_K\|^\alpha} \right) \right]
\]

\[
= \mathbb{E}_{\|\tilde{V}_K\|^2} \left\{ \exp \left( -sS_K \right) \cdot \mathbb{E} \left[ \exp \left( -s \sum_{k: \tilde{V}_k \in \tilde{V}_\infty \setminus \tilde{V}_K} \frac{H_k}{\|\tilde{V}_K\|^\alpha} \right) \right] \right\}
\]

\[
= \mathbb{E}_{Y_K^2} \left\{ \exp \left( -sS_K \right) \exp \left[ -\pi \tilde{\lambda} Y_K^\ell \left( sY_K^{-\frac{\alpha}{2}}, \frac{2}{\alpha} \right) \right] \right\} = \exp \left[ -\frac{\pi \tilde{\lambda} s^2}{\text{sinc}(2/\alpha)} \right],
\]

which yields

\[
\mathcal{L}_{S_K}(s) = \exp \left[ -\frac{\pi \tilde{\lambda} s^2}{\text{sinc}(2/\alpha)} \right] \cdot \mathbb{E} \left\{ \exp \left[ -\pi \tilde{\lambda} Y_K^\ell \left( sY_K^{-\frac{\alpha}{2}}, \frac{2}{\alpha} \right) \right] \right\}
\]

and it can be expressed as (10) due to \( Y_K \sim \text{Gamma}(K, \pi \tilde{\lambda}) \).

(iii) For the upper bound on \( \mathbb{P}[S_K \geq y] \) for \( y \geq 0 \), we can find it by using the following inequality:

\[
\mathbb{P} \left[ \sum_{k=1}^{K} Z_k \geq y \right] = 1 - \mathbb{P} \left[ \sum_{k=1}^{K} Z_k \leq y \right] \leq 1 - \prod_{k=1}^{K} \mathbb{P} \left[ Z_k \leq z_k y \right],
\]

where all \( Z_k \)'s are non-negative RVs, \( z_k \in [0, 1] \) for all \( k \in \{1, 2, \ldots, K\} \) and \( \sum_{k=1}^{K} z_k = 1 \). By using the above inequality, the lower bound on \( \mathbb{P}[S_K \geq y] \) can be thereupon found as follows:

\[
\mathbb{P}[S_K \geq y] \leq 1 - \prod_{k=1}^{K} \mathbb{P} \left[ H_k W_k \|V_k\|^{-\alpha} \leq z_k y \right] = 1 - \prod_{k=1}^{K} \mathbb{P} \left[ H_k \leq z_k y \|\tilde{V}_k\|^\alpha \right]
\]

\[
= 1 - \prod_{k=1}^{K} \mathbb{E} \left[ 1 - \exp \left( -y z_k Y_k^{-\frac{\alpha}{2}} \right) \right] \stackrel{(d)}{=} 1 - \prod_{k=1}^{K} \left[ 1 - \mathcal{L}_{Y_k^2} \left( y z_k \right) \right].
\]

Since \( Y_{k+1} > Y_k \), we select \( z_k \frac{E[Y_k]}{\sum_{k=1}^{K} E[Y_k]} = \frac{2k}{K(K+1)} \) to get a tighter lower bound. Then substituting the above \( z_k \) into the above result in step (d) yields the upper bound on \( \mathbb{P}[S_K \geq y] \) in (11).

**D. Proof of Theorem 2**

First consider the scenario in which all \( K \) APs in the virtual cell do not collaborate and the transmit power \( q \) of users is equally allocated to the \( K \) APs, i.e., \( q_k = q/K \) and \( q_i/K \) for all
$i \in \mathbb{N}_+$. If all the non-collision uplink SIRs in (16) are independent, we have

$$\eta_{ul}^i = 1 - \mathbb{P} \left[ \max_{k \in \{1, \cdots, K\}} \{ \gamma_{ul}^i (V_k \in \mathcal{V}^{nc}_K) \} \leq \theta \right] \approx 1 - \prod_{k=1}^{K} \mathbb{P} \left[ \gamma_{ul}^i (V_k \in \mathcal{V}^{nc}_K) \leq \theta \right]$$

$$= 1 - \prod_{k=1}^{K} \left( \mathbb{P} \left[ \gamma_{ul}^i \leq \theta \right] \mathbb{P}[1 (V_k \in \mathcal{V}^{nc}_K) = 1] + 1 - \mathbb{P}[1 (V_k \in \mathcal{V}^{nc}_K) = 1] \right)$$

$$(a) = 1 - \prod_{k=1}^{K} \left\{ 1 - \rho_{ul}^i \mathbb{P} \left[ \gamma_{ul}^i \geq \theta \right] \right\},$$

where $(a)$ follows from the result of $\mathbb{P}[1 (V_k \in \mathcal{V}^{nc}_K) = 1] = \sum_{m=1}^{2} \theta_m \sum_{n=1}^{\infty} p_{m,n} (1 - \delta)^{n-1} = \rho_{ul}^i$. Whereas $\mathbb{P} \left[ \gamma_{ul}^i \geq \theta \right]$ can be derived as shown in the following:

$$\mathbb{P} \left[ \gamma_{ul}^i \geq \theta \right] = \mathbb{P} \left[ \frac{\lambda V_k}{\sum_{j \in \Phi} h_j \lambda V_k - U_j} \geq \theta \right] \overset{(b)}{=} \mathbb{E} \left[ \exp \left( -\theta \lambda V_k \sum_{j \in \Phi} h_j \lambda V_k - U_j \right) \right]$$

$$\overset{(c)}{=} \mathbb{E} \left[ \frac{\exp \left( -\pi \theta \lambda V_k \lambda V_k \mu_a \right)}{\sin(2/\alpha)} \right] \overset{(d)}{=} \left( 1 + \frac{\theta \lambda V_k \lambda V_k \mu_a \sin(2/\alpha)}{\sin(2/\alpha)} \right)^{-1},$$

where $(b)$ follows from $h_k \sim \exp(1)$ and the Slivnyak theorem saying that the statistical property of a homogeneous PPP evaluated at $V_k$ is the same as that evaluated at the origin (or any point in the network) [21], [28]. $(c)$ is obtained by first applying the PGFL of a homogeneous PPP to $\mathcal{U}_a$ that is a homogeneous PPP of density $\mu_a = \delta (1 - \delta p_0) \sum_{m=1}^{\infty} \lambda_m = \delta (1 - \delta \bar{\lambda})$, and $(d)$ is due to $\lambda V_k \lambda V_k \sim \Gamma(k, \pi \bar{\lambda})$. Then substituting the result of $\mathbb{P} \left[ \gamma_{ul}^i \geq \theta \right]$ into the result of $\eta_{ul}^i$ found in $(a)$ leads to the result in (18). Next, we consider the case in which a user can associate with all APs in the network, i.e., $K$ goes to infinity in this case, and we would like to find $\rho_{ul}^i \triangleq \lim_{K \to \infty} \rho_{ul}^i$. We use the above results in $(a)$ and $(c)$ to express $\rho_{ul}^i$ as

$$\rho_{ul}^i \approx 1 - \mathbb{E} \left\{ \prod_{V_k \in \Phi} \left( 1 - \rho_{ul}^i \mathbb{P} \left[ \gamma_{ul}^i \geq \theta | V_k \right] \right) \right\} = 1 - \mathbb{E} \left\{ \prod_{V_k \in \Phi} \left[ 1 - \rho_{ul}^i \exp \left( -\pi \theta \lambda V_k \lambda V_k \mu_a \sin(2/\alpha) \right) \right] \right\}$$

$$= 1 - \exp \left\{ -\pi \rho_{ul}^i \bar{\lambda} \int_0^{\infty} \exp \left( -\pi \theta \lambda V_k \lambda V_k \mu_a \sin(2/\alpha) \right) \right\},$$

which is equal to the result in (19) by carrying out the integral in the last equality for $\mu_a = \delta \bar{\lambda}$.

Next we would like to find the bounds on $\eta_{ul}^i$ when all the non-collision APs in the virtual cell are able to collaborate. The lower bound on $\eta_{ul}^i$ in (17) can be thereupon derived as follows:

$$\eta_{ul}^i = \mathbb{P} \left[ \frac{\sum_{k:V_k \in \mathcal{V}_K} h_k q_k \lambda V_k \lambda V_k \mu_a \geq \theta}{\sum_{j \in \Phi} h_j q_j \lambda V_k \lambda V_k - U_j} \right] \overset{(c)}{=} \rho_{ul}^i \mathbb{P} \left[ \sum_{k:V_k \in \mathcal{V}_K} h_k \lambda V_k \lambda V_k \mu_a \geq \theta \sum_{j \in \Phi} h_j \lambda V_k \lambda V_k - U_j \right]$$

$$= 1 - \mathbb{E} \left\{ \prod_{V_k \in \Phi} \left( 1 - \rho_{ul}^i \mathbb{P} \left[ \gamma_{ul}^i \geq \theta | V_k \right] \right) \right\} = 1 - \mathbb{E} \left\{ \prod_{V_k \in \Phi} \left[ 1 - \rho_{ul}^i \exp \left( -\pi \theta \lambda V_k \lambda V_k \mu_a \sin(2/\alpha) \right) \right] \right\}$$

$$= 1 - \exp \left\{ -\pi \rho_{ul}^i \bar{\lambda} \int_0^{\infty} \exp \left( -\pi \theta \lambda V_k \lambda V_k \mu_a \sin(2/\alpha) \right) \right\},$$

which is equal to the result in (19) by carrying out the integral in the last equality for $\mu_a = \delta \bar{\lambda}$. 
which yields the result in (25) because \((e)\) follows from \(P[1(V_k \in \mathcal{V}_K^{nc})] = \rho^{ul}\) and \(q_j = q/K\) for all \(j\), and \((f)\) follows from the result in (11) for \(w_m = 1\). Thence, using \(\mathbb{E}_{Y_k}[\exp(-sY_k)] = (1 + s/\pi\tilde{\lambda})^{-k}\) for \(s > 0\) and substituting \(\mu_a = \delta(1 - p_0)\tilde{\lambda}\) into the above result of \(\eta^{ul}_K\) yields the result in (20).

E. Proof of Theorem 3

(i) Since the intra-virtual-cell interference exits in the virtual cell, we know that \(\gamma^{dl}_k\) in (22) can be equivalently expressed as

\[
\gamma^{dl}_k \triangleq \frac{H_k \|V_k\|^{-\alpha}}{\sum_{i:V_i \in \tilde{\Phi}_k \setminus V_k} H_i \|V_i\|^{-\alpha} + I_K}
\]

where \(\tilde{I}_k \triangleq \sum_{j:V_j \in \tilde{\Phi}_k \setminus V_k} O_{jH_j} \|V_j\|^{-\alpha}, \tilde{\Phi}, \tilde{V}_k,\) and \(\tilde{V}_k\) are already defined in Appendix C and \(O_j\) is the Bernoulli random variable associated with \(\tilde{V}_j\) and it is unity if \(\tilde{V}_j\) is not void and zero otherwise. Note that the inequality comes from the fact that \(\tilde{I}_k\) is smaller than \(\sum_{i:V_i \in \tilde{\Phi}_k \setminus V_k} H_i \|V_i\|^{-\alpha} + I_K\) because it does not contain the interference from the first \(k - 1\) APs. Thus, \(\eta^{dl}_K\) is written as

\[
\eta^{dl}_K = 1 - \mathbb{E}_{\tilde{\Phi}_K} \left\{ \prod_{k=1}^{K} \mathbb{P} \left[ \gamma^{dl}_k \leq \theta | \mathcal{V}_K \right] \right\} \leq 1 - \mathbb{E}_{\tilde{\Phi}_k} \left\{ \prod_{k=1}^{K} \mathbb{P} \left[ \frac{H_k \|V_k\|^{-\alpha}}{\tilde{I}_k} \leq \theta \right] \right\}
\]

\[
(\alpha) \triangleq 1 - \prod_{k=1}^{K} \left\{ 1 - \mathbb{E} \left[ \exp \left( -\theta \|V_k\|^{-\alpha} \sum_{j:V_j \in \tilde{\Phi}} O_{jH_j} \|V_j\|^{-\alpha} \right) \right] \right\},
\]

where \((\alpha)\) is obtained by the assumption that all \(\gamma^{dl}_k\)'s are independent and \(H_k \sim \exp(1).\) According to the result in (12) and the density of the APs in \(\tilde{\Phi}\) that uses the same RRU is \(\delta\tilde{\lambda},\) the expectation in step \((\alpha)\) can be simplified as follows:

\[
\mathbb{E} \left\{ \exp \left[ -\theta \sum_{j:V_j \in \tilde{\Phi}} O_{jH_j} \left( \frac{\|V_j\|^2}{\|V_k\|^2} \right)^{-\frac{2}{\alpha}} \right] \right\} = \left\{ 1 + \ell \left( \theta, \frac{2}{\alpha} \right) \delta \mathbb{E} \left[ O_{\frac{2}{\alpha}} \right] \right\}^{-k}.
\]

Thus, we have the following upper bound:

\[
\eta^{dl}_K \leq 1 - \prod_{k=1}^{K} \left\{ 1 - \left( 1 + \delta \ell \left( \theta, \frac{2}{\alpha} \right) \mathbb{E} \left[ O_{\frac{2}{\alpha}} \right] \right)^{-k} \right\},
\]

which yields the result in (25) because \(\tilde{\lambda} \mathbb{E}[O^{2/\alpha}] = \sum_{m=1}^{2} \lambda_m (1 - p_{0,m})\) and \(\mathbb{E}[O^{2/\alpha}] = \mathbb{P}[O = 1] = \sum_{m=1}^{2} \vartheta_m (1 - p_{0,m})\)
(ii) Now consider the case of $K \to \infty$. In this case, all APs in the network are associated with the typical user and $\gamma_k$ in (22) can be rewritten as

$$\gamma_k^{dl} = \frac{H_k Q_k \| \hat{V}_k \|^{-\alpha}}{\sum_{i} H_i Q_i \| V_i \|^{-\alpha}} = \frac{H_k \| \tilde{V}_k \|^{-\alpha}}{I_k},$$

where $I_k \triangleq \sum_{i} H_i \| \hat{V}_i \|^{-\alpha}$. For a given $\| \tilde{V}_k \|^2 = r$ and $H_k \sim \exp(1)$, we have $\mathbb{P}[\gamma_k^{dl} \geq \theta]$ given by

$$\mathbb{P}[\gamma_k^{dl} \geq \theta] = \mathbb{P}[H \geq \theta r^{\frac{\alpha}{2}} I_k] = \mathcal{L}_{I_k}(\theta r^{\frac{\alpha}{2}}) = \exp\left(-\frac{\pi \lambda \theta^{\frac{\alpha}{2}} r}{\text{sinc}(2/\alpha)}\right).$$

Thus, $\eta_{\infty}^{dl}$ can be expressed and derived as follows:

$$\eta_{\infty}^{dl} = 1 - \mathbb{P}\left[ \max_{k:V_k \in \mathcal{V}_\infty} \{ \gamma_k^{dl} \} \leq \theta \right] = 1 - \mathbb{E}_{\mathcal{V}_\infty} \left[ \prod_{k:V_k \in \mathcal{V}_\infty} \mathbb{P}[\gamma_k^{dl} \leq \theta] \right]
\overset{(b)}{=} 1 - \exp\left(-\pi \tilde{\lambda} \int_0^\infty \exp\left(-\frac{\pi \delta \lambda \theta^{\frac{\alpha}{2}} r}{\text{sinc}(2/\alpha)}\right) dr\right),$$

where (b) follows by assuming all $\gamma_k$’s are independent and then finding the PGFL of $\tilde{V}_\infty$. Carrying out the integral in the step of (b) leads to the result in (26).

(iii) When all the $K$ APs in the virtual cell can collaborate, $\eta_{K}^{dl}$ in (24) can be rewritten as

$$\eta_{K}^{dl} = \mathbb{P}\left[ \frac{\sum_{k=1}^{K} H_k \| \hat{V}_k \|^{-\alpha} \max_{k} H_k \| V_k \|^{-\alpha}}{\sum_{i} O_i H_i \| V_i \|^{-\alpha}} \geq \theta \right]
\overset{(c)}{=} 1 - \mathbb{P}\left[ K \max_{k} H_k \leq \theta \| V_k \|^{\alpha} \sum_{i} O_i H_i \| V_i \|^\alpha \right]
\overset{(d)}{=} 1 - \left\{ 1 - \mathbb{E}\left[ \exp\left(-\frac{\theta}{K^{\alpha/2+1}} \left\| \sum_{i} O_i H_i \| V_i \|^\alpha \right\| \right] \right\}^K
\overset{(e)}{=} 1 - \left\{ 1 - \left[ 1 + \delta \ell \left( \frac{\theta}{K^{\frac{\alpha}{2}+1}}, \frac{2}{\alpha} \right) \sum_{m=1}^{\theta m (1-p_m)} \right]^{K} \right\}^{-K},$$

where (c) is because $(\min_k H_k) \| \tilde{V}_k \|^{-\alpha} \leq H_k \| \tilde{V}_k \|^{-\alpha}$ for all $k \leq K$, (d) is due to $\mathbb{P}[\min_k \{ H_k \} \leq x] = 1 - \exp(-Kx)$ and replacing set $\Phi \setminus \tilde{V}_K$ with set $\Phi \setminus \tilde{V}_1$, and (e) follows from the result in (12) for setting $K$ as unity and $\theta$ as $\frac{\theta}{K}$. Hence, (27) is acquired and the proof is complete.

REFERENCES

[1] I.-R. R. M.2083-0, “IMT vision - framework and overall objectives of the future development of IMT for 2020 and beyond,” Sep. 2015.

[2] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, “Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements,” IEEE Netw., vol. 15, no. 2, pp. 8–15, 2018.

[3] K.-C. Chen, T. Zhang, R. Gitlin, and G. Fettweis, “Ultra-low latency mobile networking,” IEEE Netw., pp. 1–7, Early Access 2019.
[4] M. Simsek et al., “5G-enabled tactile internet,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.
[5] S.-Y. Lien, S.-C. Hung, K.-C. Chen, and Y.-C. Liang, “Ultra-low latency ubiquitous connections in heterogeneous cloud radio access networks,” *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 22–31, mar 2015.
[6] L.-T. W. Report, “Tactile internet,” Aug. 2014.
[7] P. Schulz et al., “Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture,” *IEEE Wireless Commun.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.
[8] C. She, C. Yang, and T. Q. S. Quek, “Cross-layer optimization for ultra-reliable and low-latency radio access networks,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.
[9] S.-C. Hung, H. Hsu, S.-Y. Lien, and K.-C. Chen, “Architecture harmonization between cloud radio access networks and fog networks,” *IEEE Access*, vol. 3, pp. 3019–3034, 2015.
[10] A. Avranas, M. Kountouris, and P. Ciblat, “Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2475–2485, Nov. 2018.
[11] B. Lee, S. Park, D. J. Love, H. Ji, and B. Shim, “Packet structure and receiver design for low latency wireless communications with ultra-short packets,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 796–807, Feb. 2018.
[12] J. P. B. Nadas, O. Onireti, R. D. Souza, and H. Alves, “Performance analysis of hybrid ARQ for ultra-reliable low latency communications,” *IEEE Sensors J.*, 2019.
[13] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, “Reliable transmission of short packets through queues and noisy channels under latency and peak-arrival guarantees,” *IEEE J. Sel. Areas Commun.*, Early Access 2019.
[14] A. Anand and G. de Veciana, “Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2411–2421, Nov. 2018.
[15] C. Sun, C. She et al., “Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402–415, Jan. 2019.
[16] J. J. Nielsen, R. Liu, and P. Popovski, “Ultra-reliable low latency communication using interface diversity,” *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322–1334, Mar. 2018.
[17] V. N. Swamy, P. Rigge, and G. Ranade, “Wireless channel dynamics and robustness for ultra-reliable low-latency communications,” *IEEE J. Sel. Areas Commun.*, Early Access 2019.
[18] S.-C. Hung, H. Hsu, S.-M. Cheng, Q. Cui, and K.-C. Chen, “Delay guaranteed network association for mobile machines in heterogeneous cloud radio access network,” *IEEE Trans. Mobile Comput.*, vol. 17, no. 12, pp. 2744–2760, Dec. 2018.
[19] C.-H. Liu and L.-C. Wang, “Random cell association and void probability in poisson-distributed cellular networks,” in *IEEE Int. Conf. Commun.*, Jun. 2015, pp. 2816–2821.
[20] C.-H. Liu and K. L. Fong, “Fundamentals of the downlink green coverage and energy efficiency in heterogeneous networks,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 12, pp. 3271–3287, Dec. 2016.
[21] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 3rd ed. New York: John Wiley and Sons, Inc., 2013.
[22] C.-H. Liu and L.-C. Wang, “Optimal cell load and throughput in green small cell networks with generalized cell association,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1058–1072, May 2016.
[23] S. B. Lowen and M. C. Teich, “Power-law shot noise,” *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1302–1318, Sep. 1990.
[24] C.-H. Liu and J. G. Andrews, “Ergodic transmission capacity of wireless ad hoc networks with interference management,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2136–2147, Jun. 2012.
[25] J.-S. Ferenc and Z. Néda, “On the size distribution of Poisson voronoi cells,” *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 518–526, 2007.
[26] C.-H. Liu and D.-C. Liang, “Heterogeneous networks with power-domain NOMA: Coverage, throughput and power allocation analysis,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3524–3539, May 2018.
[27] M. Haenggi, *Stochastic Geometry for Wireless Networks*, 1st ed. Cambridge University Press, 2012.
[28] F. Baccelli and B. Błaszczyszyn, “Stochastic geometry and wireless networks: Volume ITTheory,” *Foundations and Trends in Networking*, vol. 3, no. 3-4, pp. 249–449, 2010.