Stable Throughput Regions of Opportunistic NOMA and Cooperative NOMA With Full-Duplex Relaying

Yong Zhou, *Member, IEEE*, Vincent W.S. Wong, *Fellow, IEEE*, and Robert Schober, *Fellow, IEEE*

***Abstract*— In this paper, we consider downlink non-orthogonal multiple access (NOMA) transmission with dynamic traffic arrival for spatially random users of different priorities. By exploiting limited channel state information, we propose an opportunistic NOMA scheme to enable NOMA for high- and low-priority users when high-priority users experience good channel conditions. Opportunistic NOMA improves the transmission opportunities of low-priority users while reducing the adverse effect of NOMA on high-priority users. Moreover, we propose a cooperative NOMA scheme with full-duplex relaying, where low-priority users act as full-duplex relays to assist the high-priority users. The high-priority user constructively combines the signal and its delayed version transmitted by the base station and a selected relay, respectively. The adopted relay selection scheme takes into account the users' spatial distribution, queue status, and channel conditions. By using tools from queueing theory and stochastic geometry, we derive the stable throughput regions of both proposed schemes. Furthermore, we derive the conditions under which the proposed NOMA schemes achieve larger stable throughput regions than orthogonal multiple access (OMA). At the expense of a higher implementation complexity and with appropriate parameter setting, cooperative NOMA with full-duplex relaying achieves a larger stable throughput region than opportunistic NOMA, which in turn outperforms OMA.**

***Index Terms*— Non-orthogonal multiple access, stable throughput, dynamic traffic arrival, full-duplex relaying, spatially random users.**

I. Introduction

TO MEET the rapidly increasing traffic demand caused by the proliferation of mobile devices and data intensive applications, non-orthogonal multiple access (NOMA) [2] has been proposed as a promising technique to enhance the spectral efficiency of the fifth generation (5G) cellular network. With NOMA, multiple users can simultaneously be served by exploiting the power domain rather than the time and frequency domains as in orthogonal multiple access (OMA). By appropriately allocating the transmit power at the base station to multiple users with diverse channel conditions, NOMA can achieve a balance between network throughput and user fairness.

NOMA has recently received considerable research interest [3]–[9]. Specifically, the system-level performance of downlink NOMA transmission is evaluated in [3], which shows that transmit power allocation and user pairing are two important design aspects of NOMA. An optimal power allocation strategy is proposed in [4] to maximize the sum rate of multiple-input multiple-output (MIMO) NOMA networks. The authors in [5] formulate a joint transmit power and subcarrier allocation problem for maximization of the sum rate of multi-carrier NOMA networks and solve the problem using matching theory. The impact of user pairing on the performance of NOMA is investigated in [6], which shows that NOMA achieves a better performance when the paired NOMA users experience more distinct channel conditions. The authors in [7] derive the outage probability of MIMO-NOMA for both uplink and downlink transmission. In addition, the outage probability of a cooperative NOMA scheme is analyzed in [8], where a relay is selected to forward packets to paired NOMA users having different priorities and the low-priority user is served in an opportunistic manner. However, all of the aforementioned studies focus on resource allocation and performance analysis for NOMA with backlogged traffic.

Full-duplex communication can enhance the spectral efficiency by allowing the radios to simultaneously transmit and receive on the same frequency channel. The main challenge for realizing full-duplex communication is the self-interference due to signal leakage, which significantly degrades the performance gain achieved by full-duplexing [10]. Nevertheless, with the advancement of analog and digital self-interference cancelation techniques, full-duplex radios have been successfully implemented [11]. The rate region of full-duplex links in orthogonal frequency division multiplexing systems is analyzed in [12]. The authors in [13] develop a joint power and subcarrier allocation policy to maximize the weighted sum throughput of multi-carrier NOMA systems, where the full-duplex base station simultaneously serves multiple uplink and downlink users. Furthermore, full-duplex relaying has recently attracted significant interest [14], [15]. The authors in [14] compare the spectral efficiency of half- and full-duplex relaying strategies, and propose a joint opportunistic mode

Manuscript received February 9, 2017; revised October 3, 2017; accepted May 2, 2018. This work was supported by the Natural Sciences and Engineering Research Council of Canada. This paper was presented in part at the IEEE International Conference on Communications, Paris, France, May 2017 [1]. The associate editor coordinating the review of this paper and approving it for publication was R. K. Ganti. *(Corresponding author: Vincent W.S. Wong.)*

Y. Zhou was with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada. He is now with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: zhouyong@shanghaitech.edu.cn).

V. W.S. Wong is with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: vincentw@ece.ubc.ca).

R. Schober is with the Institute for Digital Communications, Friedrich-Alexander University of Erlangen–Nuremberg, 91058 Erlangen, Germany (e-mail: robert.schober@fau.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2018.2837014

selection and transmit power adaptation scheme to optimize spectral efficiency. However, the performance of full-duplex relaying in NOMA systems with dynamic traffic arrival and spatially random relays has not been studied yet.

Different from the aforementioned studies, we consider downlink NOMA transmission with dynamic traffic arrival and spatially random users of different priorities. For dynamic traffic arrival, the *stable throughput region* [16]–[19] is an important performance metric and defined as the set of achievable packet arrival rates given that all queues are stable. However, according to the NOMA principle, a low-priority user is allowed to share the frequency channel and transmit power with a high-priority user, which may reduce the reception reliability of the high-priority user and lead to queue instability. In NOMA, the low-priority user, which is allocated a lower transmit power, needs to decode the signal intended for the high-priority user first before decoding its own signal. Hence, the low-priority user can act as a relay and assist the transmission of the high-priority user. However, when half-duplex relaying is used, an additional time slot is required for packet forwarding, which reduces the spectral efficiency. Full-duplex relaying has the potential to mitigate this disadvantage. The performance gain achieved by full-duplex relaying can be further improved by relay selection, where the selection should take into account the residual self-interference, the queue status, and the spatial distribution of the potential relays. Considering dynamic traffic arrival together with NOMA leads to interacting queues, which complicates the performance analysis. In particular, the service process of a given queue depends on the status of the other queue, as the status of both queues determines whether NOMA can be enabled. Furthermore, channel state information (CSI) plays an important role in designing user pairing and transmit power allocation strategies. As full CSI is difficult to obtain in practice, the impact of limited CSI [20] on the performance of NOMA should be investigated.

To address the aforementioned issues, we first propose an *opportunistic NOMA* scheme exploiting limited CSI, where NOMA for high- and low-priority users is enabled only if the channel gain between the base station and the high-priority user does not fall below a certain threshold. NOMA for the low-priority users is also enabled by exploiting the differences of the low-priority users' distances to the base station. By appropriately setting the threshold to trigger NOMA, the opportunistic NOMA scheme improves the transmission opportunities of the low-priority users without degrading the performance of the high-priority users. Furthermore, we propose a *cooperative NOMA scheme with full-duplex relaying*, where the low-priority users act as full-duplex relays to help forward packets to the high-priority users. By exploiting cooperative diversity to enhance the probability of successful packet reception at the high-priority users, the number of packet retransmissions for the high-priority users is reduced, which in turn further improves the transmission opportunities of the low-priority users. The main contributions of this paper are summarized as follows:

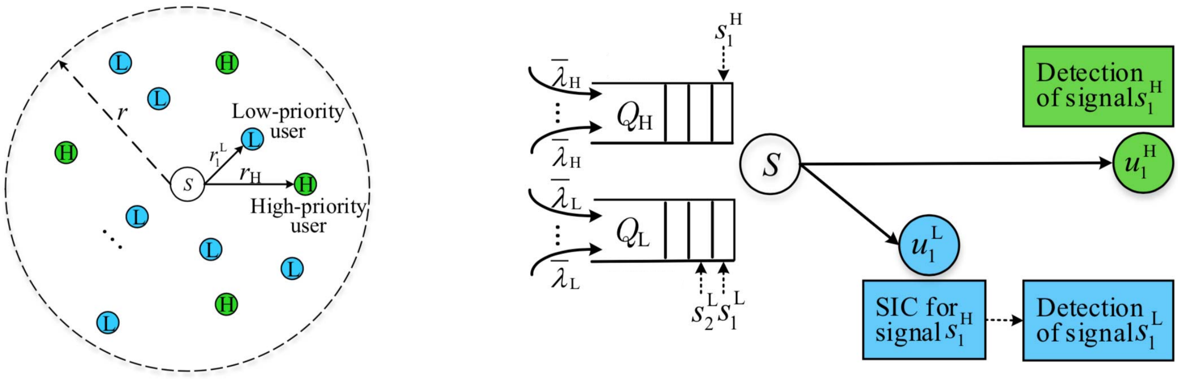
- We develop a theoretical performance analysis framework for downlink NOMA transmission with dynamic traffic arrival and spatially random users of different priorities. This analytical framework provides a better understanding of the benefits and limitations of NOMA.
- We decouple the interacting queues caused by dynamic traffic arrival and NOMA by allowing empty queues to contribute dummy packets. Tools from queueing theory and stochastic geometry are applied to characterize the stable throughput region of opportunistic NOMA.
- We derive the stable throughput region of cooperative NOMA with full-duplex relaying, taking into account the residual self-interference, spatially random low-priority users, and relay selection. Studying both opportunistic NOMA and cooperative NOMA with full-duplex relaying provides insights regarding the tradeoff between network performance and implementation complexity. We also derive the conditions under which the proposed NOMA schemes achieve larger stable throughput regions than OMA.
- Simulation results validate the analysis of the probabilities of successful packet reception. Numerical results show that, with appropriate parameter setting, both proposed NOMA schemes can outperform OMA, and cooperative NOMA with full-duplex relaying can achieve a larger stable throughput region than opportunistic NOMA at the expense of a higher implementation complexity. The impact of the relevant design and system parameters (e.g., the threshold to trigger NOMA and the power allocation coefficients) on the stable throughput regions of the proposed NOMA schemes is also evaluated.

The rest of this paper is organized as follows. We describe the system model in Section II. In Section III, we present the opportunistic NOMA scheme and derive its stable throughput region. We describe the cooperative NOMA scheme with full-duplex relaying and characterize its stable throughput region in Section IV. In Section V, we present the conditions under which the proposed NOMA schemes achieve larger stable throughput regions than OMA. Numerical results are provided in Section VI. Finally, Section VII concludes this paper.

II. System Model

Consider a downlink transmission scenario consisting of one base station and multiple users, as shown in Fig. 1(a). Base station S is located at the center of the circular network coverage area with radius r. Over a single frequency channel, time is divided into slots of constant durations. Users are categorized into two groups with different priorities, i.e., K low-priority users in set $\mathcal{U}^{\mathrm{L}}$ and M high-priority users in set $\mathcal{U}^{\mathrm{H}}$. The locations of low-priority users are assumed to follow a binomial point process (BPP) [21], [22]. Specifically, for each time slot, K low-priority users are independently and uniformly distributed within the network coverage area. On the other hand, the high-priority users are located r_{H} meters away[1]

[1] The proposed framework can be extended to the case where the high-priority users also have random distances to the base station by first conditioning on the distance and then taking the expectation over the high-priority user distance distribution. The resulting analytical expressions involve an additional integral compared to the results obtained for the fixed high-priority user distance considered in this paper. Fixed user distances were also assumed in other works in the literature, e.g., [23]–[25], as this approach simplifies the analytical expressions without compromising the insights that can be obtained, as demonstrated in [26].



(a) (b)

Fig. 1. (a) Illustration of the network topology for downlink transmission with spatially random users of different priorities, where base station S serves M high-priority users and K low-priority users. (b) Illustration of the queueing and signal reception models for downlink transmission with dynamic traffic arrival. Base station S transmits the first packet from queue Q_{H} and the first packet from queue Q_{L} to high-priority user u_1^{H} and low-priority user u_1^{L} using NOMA, respectively.

from base station S in a random direction. Base station S and all users have a single antenna.

Base station S is equipped with two queues of infinite size, denoted as Q_{H} and Q_{L}, which store the packets to be transmitted to the high- and low-priority users, respectively, as shown in Fig. 1(b). The packet arrival at base station S for each user follows an independent and stationary process. For ease of presentation, the average arrival rates of users having the same priority are assumed to be identical, but the analysis can be extended to a general scenario with diverse average arrival rates. The average arrival rates of queues Q_{H} and Q_{L} are given by $\lambda_{\mathrm{H}} = M\overline{\lambda}_{\mathrm{H}}$ and $\lambda_{\mathrm{L}} = K\overline{\lambda}_{\mathrm{L}}$ (packets per time slot), where $\overline{\lambda}_{\mathrm{H}}$ and $\overline{\lambda}_{\mathrm{L}}$ denote the average arrival rates for each high- and low-priority user, respectively. Packets for users having the same priority have the same size in bits and are served in a first-in first-out (FIFO) manner. Each packet is transmitted in one time slot.

The channel between any two transceivers suffers from path loss and Rayleigh fading. A packet can be successfully decoded only if the received signal-to-interference-plus-noise ratio (SINR) is not smaller than a required reception threshold. Upon successfully (or erroneously) receiving a packet from base station S, the corresponding receiver sends feedback that indicates the packet success or failure to base station S via an error- and delay-free control channel. After successful reception, the packet is removed from the queue at base station S. Otherwise, base station S retransmits the packet until it is successfully decoded. We denote $Q_{\mathrm{H}}(t)$ and $Q_{\mathrm{L}}(t)$ as the queue lengths of Q_{H} and Q_{L} in time slot t, respectively. A queue is said to be stable if its queue length has a limiting distribution as time goes to infinity. For high-priority queue, we have $\lim_{t\to\infty} \mathbb{P}\left(Q_{\mathrm{H}}(t) < l\right) = F(l)$ and $\lim_{l\to\infty} F(l) = 1$. If the arrival and service processes of a queue are jointly stationary and ergodic, by Loynes' theorem [27], the sufficient condition for the stability of queue Q_{H} is that $\lambda_{\mathrm{H}} < \mu_{\mathrm{H}}$, where μ_{H} (packets per time slot) is the average service rate of queue Q_{H}. The network is stable when both queues Q_{H} and Q_{L} are stable. In this work, the stable throughput region is defined as the set of arrival rates of queues Q_{H} and Q_{L} that lead to a stable network for fixed power allocation coefficients and threshold to trigger NOMA. The full stable throughput region refers to the union of the stable throughput regions over all possible values of the power allocation coefficients and threshold to trigger NOMA.

In order to reduce the implementation complexity, we consider the case when two users are paired for NOMA transmission. Such a two-user NOMA scheme is included in the 3rd Generation Partnership Project (3GPP) standard [28] and considered in [4] and [6]–[8]. We denote the intended receivers of the first packet from queue Q_{H} and the first packet from queue Q_{L} by u_1^{H} and u_1^{L}, respectively. When NOMA is performed to serve users u_1^{H} and u_1^{L} in time slot t, the superimposed signal transmitted by base station S is $\alpha_{\mathrm{H}}\sqrt{P_S}s_1^{\mathrm{H}}(t) + \alpha_{\mathrm{L}}\sqrt{P_S}s_1^{\mathrm{L}}(t)$, where P_S denotes the transmit power of base station S, α_{H} and α_{L} denote the transmit power allocation coefficients for the high- and low-priority users, respectively, and $s_1^{\mathrm{H}}(t)$ and $s_1^{\mathrm{L}}(t)$ denote the signals intended for users u_1^{H} and u_1^{L} in time slot t, respectively, with $\mathbb{E}\left(\left|s_1^{\mathrm{H}}(t)\right|^2\right) = \mathbb{E}\left(\left|s_1^{\mathrm{L}}(t)\right|^2\right) = 1$. Here, $\mathbb{E}(\cdot)$ denotes statistical expectation. The paired NOMA users are ordered according to their priorities for being served [8]. As user u_1^{H} has a higher priority, we have $\alpha_{\mathrm{H}} > \alpha_{\mathrm{L}}$ and $\alpha_{\mathrm{H}}^2 + \alpha_{\mathrm{L}}^2 = 1$. Before transmission begins, the base station informs user u_1^{L} that it is expected to perform successive interference cancelation (SIC) by sending a corresponding control information, which includes information about the allocated transmit power and is attached to the user's scheduling information, as suggested in [28, pp. 15].

The superimposed signal received at user $u_1^a, a \in \{\mathrm{H}, \mathrm{L}\}$, in time slot t is given by

$$y_1^a(t) = \left(\alpha_{\mathrm{H}} s_1^{\mathrm{H}}(t) + \alpha_{\mathrm{L}} s_1^{\mathrm{L}}(t)\right)\sqrt{P_S}h_1^a(t)\sqrt{\ell\left(x_1^a\right)} + n_1^a(t), \tag{1}$$

where $h_1^a(t)$ denotes the Rayleigh fading channel gain between base station S and user u_1^a in time slot t, $n_1^a(t)$ denotes the additive white Gaussian noise (AWGN) at user u_1^a with zero mean and variance σ^2 in time slot t, x_1^a denotes the location of user u_1^a, $\ell(x_1^a) = \left(1 + (r_1^a)^\beta\right)^{-1}$ and r_1^a denote the non-singular path loss and the distance between base station S and user u_1^a, respectively, and β denotes the path loss exponent. Hence, $\left|h_1^a(t)\right|^2$ is an exponential random variable with unit mean.

After receiving the signal from base station S, high-priority user u_1^{H} treats the signal intended for low-priority user u_1^{L} as interference and decodes its own signal based on SINR

$$\Gamma_{\mathrm{H1|L1}}(t, \alpha_{\mathrm{H}}) = \frac{\alpha_{\mathrm{H}}^2 P_S \left|h_1^{\mathrm{H}}(t)\right|^2 \ell\left(x_1^{\mathrm{H}}\right)}{\alpha_{\mathrm{L}}^2 P_S \left|h_1^{\mathrm{H}}(t)\right|^2 \ell\left(x_1^{\mathrm{H}}\right) + \sigma^2}, \tag{2}$$

where $\Gamma_{\mathrm{H1|L1}}(t, \alpha_{\mathrm{H}})$ denotes the SINR of signal $s_1^{\mathrm{H}}(t)$ observed at high-priority user u_1^{H} when paired with low-priority user u_1^{L} in time slot t.

Low-priority user u_1^{L} first decodes the signal intended for high-priority user u_1^{H} with SINR

$$\Gamma_{\mathrm{H1\to L1}}(t, \alpha_{\mathrm{H}}) = \frac{\alpha_{\mathrm{H}}^2 P_S \left|h_1^{\mathrm{L}}(t)\right|^2 \ell\left(x_1^{\mathrm{L}}\right)}{\alpha_{\mathrm{L}}^2 P_S \left|h_1^{\mathrm{L}}(t)\right|^2 \ell\left(x_1^{\mathrm{L}}\right) + \sigma^2}, \tag{3}$$

where $\Gamma_{\mathrm{H1\to L1}}(t, \alpha_{\mathrm{H}})$ denotes the SINR of signal $s_1^{\mathrm{H}}(t)$ observed at user u_1^{L} in time slot t.

If low-priority user u_1^{L} successfully decodes signal $s_1^{\mathrm{H}}(t)$, i.e., $\Gamma_{\mathrm{H1\to L1}}(t, \alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}$, where $\Gamma_{\mathrm{th}}^{\mathrm{H}}$ denotes the threshold required to successfully decode the packets intended for the high-priority users, then low-priority user u_1^{L} removes signal $s_1^{\mathrm{H}}(t)$ from received signal $y_1^{\mathrm{L}}(t)$ by applying SIC, and decodes its own signal with signal-to-noise ratio (SNR)

$$\Gamma_{\mathrm{L1}}(t, \alpha_{\mathrm{L}}) = \frac{\alpha_{\mathrm{L}}^2 P_S \left|h_1^{\mathrm{L}}(t)\right|^2 \ell\left(x_1^{\mathrm{L}}\right)}{\sigma^2}, \tag{4}$$

where $\Gamma_{\mathrm{L1}}(t, \alpha_{\mathrm{L}})$ denotes the SNR of signal $s_1^{\mathrm{L}}(t)$ observed at user u_1^{L} in time slot t.

When NOMA is enabled, users u_1^{H} and u_1^{L} can successfully decode their own signals if events $\{\Gamma_{\mathrm{H1|L1}}(t, \alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}\}$ and $\{\Gamma_{\mathrm{H1\to L1}}(t, \alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \Gamma_{\mathrm{L1}}(t, \alpha_{\mathrm{L}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{L}}\}$ occur, respectively, where $\Gamma_{\mathrm{th}}^{\mathrm{L}}$ denotes the threshold required to successfully decode the packets intended for the low-priority users. Base station S simultaneously serves users u_1^{H} and u_1^{L}, at the cost of reducing the probability of successful packet reception at high-priority user u_1^{H}. Specifically, by sharing the frequency channel and transmit power, the received SINR at high-priority user u_1^{H} decreases, i.e., $\Gamma_{\mathrm{H1|L1}}(t, \alpha_{\mathrm{H}}) < \Gamma_{\mathrm{H1}}(t, 1) = P_S \left|h_1^{\mathrm{H}}(t)\right|^2 \ell\left(x_1^{\mathrm{H}}\right) / \sigma^2$. Hence, to guarantee the stability of queue Q_{H}, NOMA cannot always be enabled, especially when the average arrival rate λ_{H} is large.

To facilitate our analysis, for the remainder of this paper, we make the following assumptions. The protocol overhead due to feedback from the users to the base station is much smaller than the packet size and is neglected. The fading coefficients are assumed to remain invariant during one time slot and vary independently over different time slots and across different links, as in [23]–[26] and [29]. At the end of each time slot $t \in \mathbb{Z}^+$, the locations of the low-priority users change according to a high mobility random walk model within the network coverage area as in [23]–[25] and [29]. Hence, the displacement theorem [30] can be applied and the user locations are independent across time slots, which enables the derivation of tractable performance results, providing useful insights on the network performance.

III. Opportunistic NOMA

In this section, we propose an opportunistic NOMA scheme to improve the transmission opportunities of low-priority users while reducing the adverse effect of NOMA on high-priority users, and characterize the stable throughput region. We assume that only limited instantaneous CSI is available at base station S. First, when queue Q_{H} is non-empty in time slot t, one bit of information is sent back from high-priority user u_1^{H} to base station S. In particular, high-priority user u_1^{H} sends feedback 1 to base station S if the instantaneous channel gain, $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell\left(x_1^{\mathrm{H}}\right)$, is not less than a threshold, θ, and sends feedback 0 to base station S otherwise. Second, when queue Q_{H} is empty in time slot t, users u_1^{L} and u_2^{L} send back their distances to base station S, where u_2^{L} denotes the intended receiver of the second packet from queue Q_{L} when available. Based on the limited CSI, NOMA is enabled by base station S in an opportunistic manner.

We denote the opportunistic NOMA system as Φ^{ON}, where base station S transmits the first packet from queue Q_{H} whenever it is non-empty due to its high priority to be served. The packet transmissions depend on the status of queues Q_{H} and Q_{L}, and are discussed in the following.

Case 1: If $Q_{\mathrm{H}}(t) > 0$ and $Q_{\mathrm{L}}(t) > 0$, then base station S transmits the first packet from queue Q_{H} and the first packet from queue Q_{L} to users u_1^{H} and u_1^{L}, respectively, using NOMA with fixed power allocation coefficients $\left(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2\right)$ when $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) \geq \theta$, and transmits the first packet from queue Q_{H} to user u_1^{H} using OMA with power P_S when $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) < \theta$.

Case 2: If $Q_{\mathrm{H}}(t) > 0$ and $Q_{\mathrm{L}}(t) = 0$, then base station S transmits the first packet from queue Q_{H} to user u_1^{H} using OMA[2] with power P_S.

Case 3: If $Q_{\mathrm{H}}(t) = 0$ and $Q_{\mathrm{L}}(t) > 0$, then base station S transmits the first and second packets from queue Q_{L} to users u_1^{L} and u_2^{L}, respectively, using NOMA when the first two packets are intended for different users (i.e., $u_1^{\mathrm{L}} \neq u_2^{\mathrm{L}}$), and transmits the first packet from queue Q_{L} to user u_1^{L} using OMA with power P_S when the first two packets are intended for the same user (i.e., $u_1^{\mathrm{L}} = u_2^{\mathrm{L}}$) or $Q_{\mathrm{L}}(t) = 1$.

The average service rate of queue Q_{H} depends on the status of queue Q_{L}. When queue Q_{L} is empty, base station S transmits the first packet from queue Q_{H} to user u_1^{H} using OMA. When queue Q_{L} is non-empty, base station S transmits the first packet from queue Q_{H} to user u_1^{H} using NOMA with probability $\mathbb{P}\left(\left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) \geq \theta\right) = \exp\left(-\theta\left(1 + r_{\mathrm{H}}^\beta\right)\right)$. Similarly, the average service rate of queue Q_{L} also depends

[2]Note that NOMA for different high-priority users is not enabled in this paper, as the probability that different high-priority users experience very different channel conditions is low. However, different user channel conditions are crucial for achieving a gain with NOMA [6]. A similar setting is also considered in [7].

on the status of queue Q_{H}. Hence, queues Q_{H} and Q_{L} interact with each other and their average service rates cannot be directly calculated. In this context, stochastic dominance [31] is a useful tool and can be used to decouple the interacting queues and to characterize the stable throughput region. By using stochastic dominance, we construct two *dominant systems* Φ_1^{ON} and Φ_2^{ON} based on the original opportunistic NOMA system Φ^{ON}. In the following, we derive the stable throughput regions of dominant systems Φ_1^{ON} and Φ_2^{ON}, and then show that the stable throughput region of the original opportunistic NOMA system Φ^{ON} is equal to the union of the stable throughput regions of dominant systems Φ_1^{ON} and Φ_2^{ON}.

A. Stable Throughput Region of Dominant System Φ_1^{ON}

Dominant system Φ_1^{ON}: If queue Q_{L} is empty, then queue Q_{L} contributes a dummy packet when high-priority user u_1^{H} sends feedback 1 to base station S, while queue Q_{H} acts in the same manner as in the original opportunistic NOMA system Φ^{ON}.

In dominant system Φ_1^{ON}, the service process of queue Q_{H} depends on the condition of the channel between base station S and user u_1^{H}. Base station S transmits the first packet from queue Q_{H} to user u_1^{H} using OMA and NOMA when $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) < \theta$ and $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) \geq \theta$, respectively. Note that the average probability of successful packet reception at each high-priority user is the same. Hence, the average service rate of queue Q_{H}, denoted as $\mu_{\mathrm{H}}^{\mathrm{ON1}}$, is given by

$$\begin{aligned}\mu_{\mathrm{H}}^{\mathrm{ON1}} &= \mathbb{P}\left(\Gamma_{\mathrm{H1}}(t,1) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) < \theta\right) \\ &\quad + \mathbb{P}\left(\Gamma_{\mathrm{H1|L1}}(t,\alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) \geq \theta\right), \quad (5)\end{aligned}$$

where the first and second terms of the right-hand side of (5) represent the probabilities of successful packet reception at high-priority user u_1^{H} when OMA and NOMA are enabled, denoted as $q_{\mathrm{H1}}^{\mathrm{OMA}}(\theta)$ and $q_{\mathrm{H1|L1}}^{\mathrm{ON}}(\alpha_{\mathrm{H}},\theta)$, respectively. The following lemma provides the stability condition for queue Q_{H} in dominant system Φ_1^{ON}.

Lemma 1: In dominant system Φ_1^{ON}, queue Q_{H} is stable if

$$\begin{aligned}\lambda_{\mathrm{H}} < \mu_{\mathrm{H}}^{\mathrm{ON1}} &= \exp\left(-\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)\right) - \exp\left(-\theta\left(1+r_{\mathrm{H}}^{\beta}\right)\right) \\ &\quad + \exp\left(-\max\left\{\frac{\rho_{\mathrm{H}}}{\alpha_{\mathrm{H}}^2 - \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2}, \theta\right\}\left(1+r_{\mathrm{H}}^{\beta}\right)\right), \quad (6)\end{aligned}$$

where $\rho_{\mathrm{H}} = \Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2/P_S$, $\theta > \rho_{\mathrm{H}}$, and $\alpha_{\mathrm{H}}^2 > \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2$.

Proof: Please refer to Appendix A. ■

The service process of queue Q_{L} in dominant system Φ_1^{ON} depends on the status of queue Q_{H}. If queue Q_{H} is non-empty, then base station S transmits the first packet from queue Q_{L} to user u_1^{L} using NOMA when $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) \geq \theta$. If queue Q_{H} is empty, then base station S transmits the first and second packets from queue Q_{L} to users u_1^{L} and u_2^{L} using NOMA when those two packets are intended for different users (which occurs with probability $1-\frac{1}{K}$), and transmits the first packet from queue Q_{L} to user u_1^{L} using OMA when the first two packets are intended for the same user (which occurs with probability $\frac{1}{K}$). As all low-priority users follow the same location distribution, the average probability of successful packet reception at each low-priority user is the same. The average service rate of queue Q_{L}, denoted as $\mu_{\mathrm{L}}^{\mathrm{ON1}}$, can be expressed as

$$\begin{aligned}\mu_{\mathrm{L}}^{\mathrm{ON1}} &= \mathbb{P}(Q_{\mathrm{H}}(t)>0)\mathbb{P}\left(\left|h_1^{\mathrm{H}}(t)\right|^2 \ell\left(x_1^{\mathrm{H}}\right) \geq \theta\right) q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}}) \\ &\quad + \mathbb{P}(Q_{\mathrm{H}}(t)=0)\left(\left(1-\frac{1}{K}\right) q_{\mathrm{L1L2}}^{\mathrm{ON}} + \frac{1}{K} q_{\mathrm{L1}}^{\mathrm{OMA}}\right), \quad (7)\end{aligned}$$

where $\mathbb{P}(Q_{\mathrm{H}}(t)>0) = \lambda_{\mathrm{H}}/\mu_{\mathrm{H}}^{\mathrm{ON1}}$, $q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}})$ is the probability of successful packet reception at user u_1^{L} with power allocation coefficient α_{L} when paired with user u_1^{H}, $q_{\mathrm{L1L2}}^{\mathrm{ON}}$ is the summation of the probabilities of successful packet reception at users u_1^{L} and u_2^{L} using NOMA, and $q_{\mathrm{L1}}^{\mathrm{OMA}}$ denotes the probability of successful packet reception at user u_1^{L} using OMA. For two paired low-priority users (i.e., u_1^{L} and u_2^{L}), the transmit power allocation coefficients for the users closer to and farther from the base station are denoted as α_{n} and α_{f}, respectively, with $\alpha_{\mathrm{n}}^2+\alpha_{\mathrm{f}}^2=1$. The following lemma presents the stability condition for queue Q_{L} in dominant system Φ_1^{ON}.

Lemma 2: In dominant system Φ_1^{ON}, queue Q_{L} is stable if

$$\begin{aligned}\lambda_{\mathrm{L}} < \mu_{\mathrm{L}}^{\mathrm{ON1}} &= \frac{\lambda_{\mathrm{H}}}{\mu_{\mathrm{H}}^{\mathrm{ON1}}} \exp\left(-\theta\left(1+r_{\mathrm{H}}^{\beta}\right)\right) q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}}) \\ &\quad + \left(1-\frac{\lambda_{\mathrm{H}}}{\mu_{\mathrm{H}}^{\mathrm{ON1}}}\right)\eta, \quad (8)\end{aligned}$$

where $\mu_{\mathrm{H}}^{\mathrm{ON1}}$ is given in (6),

$$\begin{aligned}q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}}) &= \frac{2}{r^2\beta} N_1^{-2/\beta} \exp(-N_1)\, \gamma\left(\frac{2}{\beta}, N_1 r^{\beta}\right), \quad (9) \\ \eta &= \left(1-\frac{1}{K}\right)\left(q_{\mathrm{Lf|Ln}}^{\mathrm{ON}}(\alpha_{\mathrm{f}}) + q_{\mathrm{Ln|Lf}}^{\mathrm{ON}}(\alpha_{\mathrm{n}})\right) \\ &\quad + \frac{1}{K} q_{\mathrm{L1}}^{\mathrm{OMA}}, \quad (10) \\ q_{\mathrm{Lf|Ln}}^{\mathrm{ON}}(\alpha_{\mathrm{f}}) &= \frac{4}{r^4\beta} N_2^{-4/\beta} \exp(-N_2)\, \gamma\left(\frac{4}{\beta}, N_2 r^{\beta}\right), \quad (11) \\ q_{\mathrm{Ln|Lf}}^{\mathrm{ON}}(\alpha_{\mathrm{n}}) &= \frac{2}{r^2\beta} N_3^{-2/\beta} \exp(-N_3)\, \gamma\left(\frac{2}{\beta}, N_3 r^{\beta}\right) \\ &\quad - \frac{4}{r^4\beta} N_3^{-4/\beta} \gamma\left(\frac{4}{\beta}, N_3 r^{\beta}\right), \quad (12) \\ q_{\mathrm{L1}}^{\mathrm{OMA}} &= \frac{2}{r^2\beta} \rho_{\mathrm{L}}^{-2/\beta} \exp(-\rho_{\mathrm{L}})\, \gamma\left(\frac{2}{\beta}, \rho_{\mathrm{L}} r^{\beta}\right), \quad (13)\end{aligned}$$

$N_1 = \max\left\{\frac{\rho_{\mathrm{H}}}{\alpha_{\mathrm{H}}^2-\Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2}, \frac{\rho_{\mathrm{L}}}{\alpha_{\mathrm{L}}^2}\right\}$, $N_2 = \frac{\rho_{\mathrm{L}}}{\alpha_{\mathrm{f}}^2-\Gamma_{\mathrm{th}}^{\mathrm{L}}\alpha_{\mathrm{n}}^2}$, $N_3 = \max\left\{\frac{\rho_{\mathrm{L}}}{\alpha_{\mathrm{f}}^2-\Gamma_{\mathrm{th}}^{\mathrm{L}}\alpha_{\mathrm{n}}^2}, \frac{\rho_{\mathrm{L}}}{\alpha_{\mathrm{n}}^2}\right\}$, $\rho_{\mathrm{L}} = \Gamma_{\mathrm{th}}^{\mathrm{L}}\sigma^2/P_S$, $\alpha_{\mathrm{H}}^2 > \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2$, $\alpha_{\mathrm{f}}^2 > \Gamma_{\mathrm{th}}^{\mathrm{L}}\alpha_{\mathrm{n}}^2$, and $\gamma(w,v) = \int_0^v e^{-z} z^{w-1} \mathrm{d}z$ is the lower incomplete Gamma function [32].

Proof: Please refer to Appendix B. ■

Dominant system Φ_1^{ON} is stable if both queues Q_{H} and Q_{L} are stable, i.e., both $\lambda_{\mathrm{H}} < \mu_{\mathrm{H}}^{\mathrm{ON1}}$ and $\lambda_{\mathrm{L}} < \mu_{\mathrm{L}}^{\mathrm{ON1}}$ hold. As a result, based on Lemmas 1 and 2, the stable throughput region of dominant system Φ_1^{ON}, denoted as $\mathcal{R}_1^{\mathrm{ON}}$, is given by

$$\mathcal{R}_1^{\mathrm{ON}} = \Bigg\{ (\lambda_{\mathrm{H}}, \lambda_{\mathrm{L}}) : \frac{\lambda_{\mathrm{L}}}{\eta} + \frac{\left(\eta - \exp\left(-\theta\left(1+r_{\mathrm{H}}^{\beta}\right)\right) q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}})\right)\lambda_{\mathrm{H}}}{\eta \mu_{\mathrm{H}}^{\mathrm{ON1}}} < 1, \text{ for } 0 \leq \lambda_{\mathrm{H}} < \mu_{\mathrm{H}}^{\mathrm{ON1}} \Bigg\}. \tag{14}$$

B. Stable Throughput Region of Dominant System Φ_2^{ON}

Dominant system Φ_2^{ON}: If queue Q_{H} is empty, then queue Q_{H} contributes a dummy packet, while queue Q_{L} acts in the same manner as in the original opportunistic NOMA system Φ^{ON}.

In dominant system Φ_2^{ON}, base station S transmits the first packet from queue Q_{L} to user u_1^{L} using NOMA when $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) \geq \theta$. The average service rate of queue Q_{L} can be expressed as $\mu_{\mathrm{L}}^{\mathrm{ON2}} = \exp\left(-\theta\left(1+r_{\mathrm{H}}^{\beta}\right)\right) q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}})$. Queue Q_{L} in dominant system Φ_2^{ON} is stable if $\lambda_{\mathrm{L}} < \mu_{\mathrm{L}}^{\mathrm{ON2}}$. The service process of queue Q_{H} in dominant system Φ_2^{ON} depends on queue Q_{L}. If queue Q_{L} is empty, then base station S transmits the first packet from queue Q_{H} to user u_1^{H} using OMA. If queue Q_{L} is non-empty, then base station S transmits the first packet from queue Q_{H} to user u_1^{H} using NOMA and OMA when $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell\left(x_1^{\mathrm{H}}\right) \geq \theta$ and $\left|h_1^{\mathrm{H}}(t)\right|^2 \ell\left(x_1^{\mathrm{H}}\right) < \theta$, respectively. The average service rate of queue Q_{H} in dominant system Φ_2^{ON} is given by

$$\mu_{\mathrm{H}}^{\mathrm{ON2}} = \mathbb{P}(Q_{\mathrm{L}}(t)=0)\mu_{\mathrm{H}}^{\mathrm{OMA}} + \mathbb{P}\left(Q_{\mathrm{L}}(t)>0\right)\left(q_{\mathrm{H1}}^{\mathrm{OMA}}(\theta) + q_{\mathrm{H1|L1}}^{\mathrm{ON}}(\alpha_{\mathrm{H}}, \theta)\right), \tag{15}$$

where $\mu_{\mathrm{H}}^{\mathrm{OMA}} = \exp\left(-\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)\right)$, $\mathbb{P}(Q_{\mathrm{L}}(t)=0) = 1 - \lambda_{\mathrm{L}}/\mu_{\mathrm{L}}^{\mathrm{ON2}}$, and $q_{\mathrm{H1}}^{\mathrm{OMA}}(\theta)$ and $q_{\mathrm{H1|L1}}^{\mathrm{ON}}(\alpha_{\mathrm{H}}, \theta)$ are given in (38) and (39), respectively. Queue Q_{H} in dominant system Φ_2^{ON} is stable if $\lambda_{\mathrm{H}} < \mu_{\mathrm{H}}^{\mathrm{ON2}}$.

The stable throughput region of dominant system Φ_2^{ON}, denoted as $\mathcal{R}_2^{\mathrm{ON}}$, is given by

$$\mathcal{R}_2^{\mathrm{ON}} = \Bigg\{ (\lambda_{\mathrm{H}}, \lambda_{\mathrm{L}}) : \frac{\lambda_{\mathrm{H}}}{\mu_{\mathrm{H}}^{\mathrm{OMA}}} + \frac{\left(\mu_{\mathrm{H}}^{\mathrm{OMA}} - q_{\mathrm{H1|L1}}^{\mathrm{ON}}(\alpha_{\mathrm{H}}, \theta) - q_{\mathrm{H1}}^{\mathrm{OMA}}(\theta)\right)\lambda_{\mathrm{L}}}{\exp\left(-\theta\left(1+r_{\mathrm{H}}^{\beta}\right)\right)\mu_{\mathrm{H}}^{\mathrm{OMA}} q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}})} < 1, \text{ for } 0 \leq \lambda_{\mathrm{L}} < \exp\left(-\theta\left(1+r_{\mathrm{H}}^{\beta}\right)\right) q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}}) \Bigg\}. \tag{16}$$

The following theorem presents the stable throughput region of the original opportunistic NOMA system Φ^{ON}.

Theorem 1: The stable throughput region of the original dominant NOMA system Φ^{ON} for fixed power allocation coefficients and threshold θ, denoted as $\mathcal{R}^{\mathrm{ON}}$, is equal to the union of the stable throughput regions of dominant systems Φ_1^{ON} and Φ_2^{ON}, i.e., $\mathcal{R}^{\mathrm{ON}} = \mathcal{R}_1^{\mathrm{ON}} \cup \mathcal{R}_2^{\mathrm{ON}}$.

Proof: Please refer to Appendix C. ■

Due to the complexity of analytically deriving the full stable throughput region, we resort to numerical analysis to obtain the full stable throughput region in Section VI, as in [17] and [18].

IV. Cooperative NOMA With Full-Duplex Relaying

The proposed opportunistic NOMA scheme enhances the stable throughput region by providing more transmission opportunities to the low-priority users without improving the performance of the high-priority users. In this section, we propose a cooperative NOMA scheme with full-duplex relaying to improve the reception reliability of the high-priority users with the help of the low-priority users. By exploiting the cooperative diversity gain, the transmission opportunities of the low-priority users can be further increased. The cooperative NOMA system with full-duplex relaying, denoted as Φ^{FCN}, is described as follows.

Case 1: If $Q_{\mathrm{H}}(t) > 0$ and $Q_{\mathrm{L}}(t) > 0$, then base station S transmits the first packet from queue Q_{H} and the first packet from queue Q_{L} to users u_1^{H} and u_1^{L}, respectively, using cooperative NOMA with fixed power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$. Before transmission begins, base station S informs low-priority user u_1^{L} to act as a full-duplex relay. In accordance with the NOMA decoding strategy, low-priority user u_1^{L} decodes signal $s_1^{\mathrm{H}}(t)$ intended for high-priority user u_1^{H} before performing SIC. By utilizing suitable channel coding (e.g., convolutional coding), low-priority user u_1^{L} can decode signal $s_1^{\mathrm{H}}(t)$ after a delay of δ symbol durations. Hence, after δ symbol durations, low-priority user u_1^{L}, which is assumed to be a full-duplex node, simultaneously receives the superimposed signal from the base station and forwards the delayed version of signal $s_1^{\mathrm{H}}(t)$ to high-priority user u_1^{H} [14], [33]. Full-duplex relaying prototypes have been reported in the literature, e.g., [34]. High-priority user u_1^{H} constructively combines and decodes the signal transmitted by base station S and its delayed version forwarded by user u_1^{L}.[3] At the end of time slot t, low-priority user u_1^{L} performs SIC to remove the contribution of signal $s_1^{\mathrm{H}}(t)$ from its received signal, and then decodes its own signal $s_1^{\mathrm{L}}(t)$. The delay δ can be made much smaller than the packet size, and hence, it is neglected for the analysis in this paper. On the other hand, if user u_1^{L} cannot successfully decode signal $s_1^{\mathrm{H}}(t)$, then user u_1^{L} remains silent in time slot t and decodes the signals without suffering from the self-interference caused by full-duplex relaying.

Case 2: If $Q_{\mathrm{H}}(t) > 0$ and $Q_{\mathrm{L}}(t) = 0$, then base station S transmits the first packet from queue Q_{H} to high-priority user u_1^{H} using cooperative OMA. Among all low-priority users, the low-priority user that can decode signal $s_1^{\mathrm{H}}(t)$ from base station S and has the best channel condition with respect to high-priority user u_1^{H} is selected as the best relay. The best relay forwards the delayed version of the signal to user u_1^{H} in the same time slot. Various efficient relay selection schemes have been proposed in the literature. High-priority user u_1^{H}

[3] The constructive combination of the signals from the direct and full-duplex forwarding links has recently been implemented in [34] based on a constructive filter and a Viterbi-style decoder.

constructively combines and decodes the signal received from base station S and its delayed version received from the best relay. If no low-priority user can successfully decode signal $s_1^{\mathrm{H}}(t)$, then user u_1^{H} decodes signal $s_1^{\mathrm{H}}(t)$ only based on the signal transmitted by base station S.

Case 3: If $Q_{\mathrm{H}}(t)=0$ and $Q_{\mathrm{L}}(t)>0$, then base station S transmits the first and second packets from queue Q_{L} to users u_1^{L} and u_2^{L}, respectively, using NOMA when the first two packets are intended for different users, and transmits the first packet from queue Q_{L} to user u_1^{L} using OMA with power P_S when the first two packets are intended for the same user or $Q_{\mathrm{L}}(t)=1$.

Queues Q_{H} and Q_{L} in the cooperative NOMA system with full-duplex relaying Φ^{FCN} interact with each other, as the average service rate of queue Q_{H} (Q_{L}) depends on the status of queue Q_{L} (Q_{H}). When queue Q_{L} is non-empty, base station S transmits the first packet from queue Q_{H} using cooperative NOMA. When queue Q_{L} is empty, base station S transmits the first packet from queue Q_{H} using cooperative OMA. The probabilities of successful packet reception at user u_1^{H} under these two conditions are different. Thus, their average service rates cannot be directly calculated. To decouple the interacting queues and facilitate the derivation of the stable throughput region, we construct two dominant systems, denoted as Φ_1^{FCN} and Φ_2^{FCN}, by using the concept of stochastic dominance, as discussed in the following.

A. Stable Throughput Region of Dominant System Φ_1^{FCN}

Dominant system Φ_1^{FCN}: If queue Q_{L} is empty, then queue Q_{L} contributes a dummy packet, while queue Q_{H} acts in the same manner as in the cooperative NOMA system with full-duplex relaying Φ^{FCN}. In dominant system Φ_1^{FCN}, a randomly selected low-priority user u_1^{L} acts as a full-duplex relay in time slot t when the following condition is satisfied:

$$\begin{aligned}\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) &= \frac{\alpha_{\mathrm{H}}^2 P_S\left|h_1^{\mathrm{L}}(t)\right|^2\ell(x_1^{\mathrm{L}})}{\alpha_{\mathrm{L}}^2 P_S\left|h_1^{\mathrm{L}}(t)\right|^2\ell(x_1^{\mathrm{L}})+\zeta P_{\mathrm{L}}+\sigma^2}\\ &\geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \end{aligned}\tag{17}$$

where $\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})$ denotes the SINR of signal $s_1^{\mathrm{H}}(t)$ observed at user u_1^{L} in time slot t when cooperative NOMA is enabled, ζ denotes the residual self-interference-to-power ratio due to imperfect self-interference cancelation, and P_{L} is the transmit power of the low-priority users.

The service process of queue Q_{H} depends on the value of $\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})$. Base station S transmits the first packet from queues Q_{H} to user u_1^{H} using NOMA and cooperative NOMA when $\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})<\Gamma_{\mathrm{th}}^{\mathrm{H}}$ and $\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})\geq\Gamma_{\mathrm{th}}^{\mathrm{H}}$, respectively. Hence, the average service rate of queue Q_{H} in dominant system Φ_1^{FCN}, denoted as $\mu_{\mathrm{H}}^{\mathrm{FCN1}}$, is given by

$$\begin{aligned}\mu_{\mathrm{H}}^{\mathrm{FCN1}} &= \mathbb{P}\left(\Gamma_{\mathrm{H1|L1}}(t,\alpha_{\mathrm{H}})\geq\Gamma_{\mathrm{th}}^{\mathrm{H}},\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})<\Gamma_{\mathrm{th}}^{\mathrm{H}}\right)\\ &\quad+\mathbb{P}\left(\Gamma_{\mathrm{H1|L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})\geq\Gamma_{\mathrm{th}}^{\mathrm{H}},\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})\geq\Gamma_{\mathrm{th}}^{\mathrm{H}}\right),\end{aligned}\tag{18}$$

where $\Gamma_{\mathrm{H1|L1}}(t,\alpha_{\mathrm{H}})$ is given in (2). The SINR of signal $s_1^{\mathrm{H}}(t)$ observed at user u_1^{H} in time slot t when cooperative NOMA is enabled, denoted as $\Gamma_{\mathrm{H1|L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})$, can be expressed as

$$\begin{aligned}&\Gamma_{\mathrm{H1|L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})\\ &\quad=\frac{\alpha_{\mathrm{H}}^2 P_S\left|h_1^{\mathrm{H}}(t)\right|^2\ell(x_1^{\mathrm{H}})+P_{\mathrm{L}}\left|g_{1,1}^{\mathrm{HL}}(t)\right|^2\ell(x_1^{\mathrm{H}}-x_1^{\mathrm{L}})}{\alpha_{\mathrm{L}}^2 P_S\left|h_1^{\mathrm{H}}(t)\right|^2\ell(x_1^{\mathrm{H}})+\sigma^2},\\ &\qquad\text{if}\quad \Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}})\geq\Gamma_{\mathrm{th}}^{\mathrm{H}},\end{aligned}\tag{19}$$

where $g_{1,1}^{\mathrm{HL}}(t)$ and $\ell(x_1^{\mathrm{H}}-x_1^{\mathrm{L}})$ denote the Rayleigh fading channel gain and non-singular path loss between users u_1^{H} and u_1^{L} in time slot t, respectively. The following lemma provides the stability condition for queue Q_{H} in dominant system Φ_1^{FCN}.

Lemma 3: In dominant system Φ_1^{FCN}, queue Q_{H} is stable if

$$\begin{aligned}\lambda_{\mathrm{H}}<\mu_{\mathrm{H}}^{\mathrm{FCN1}}&=\exp\left(-\frac{\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)}{\alpha_{\mathrm{H}}^2-\Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2}\right)\\ &\quad\times\left(1-\frac{2}{r^2\beta}N_4^{-2/\beta}\exp\left(-N_4\right)\gamma\left(\frac{2}{\beta},N_4 r^{\beta}\right)\right)\\ &\quad+C(\alpha_{\mathrm{H}})+\frac{2N_5}{r^2\beta}N_4^{-2/\beta}\exp(-N_4)\gamma\left(\frac{2}{\beta},N_4 r^{\beta}\right),\end{aligned}\tag{20}$$

where $N_4=\frac{(\zeta P_{\mathrm{L}}+\sigma^2)\Gamma_{\mathrm{th}}^{\mathrm{H}}}{(\alpha_{\mathrm{H}}^2-\Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2)P_S}$, $\ell(x_1^{\mathrm{H}}-x_1^{\mathrm{L}})=\left(1+\left(r_{\mathrm{H}}^2+\left(r_1^{\mathrm{L}}\right)^2-2\,r_{\mathrm{H}}r_1^{\mathrm{L}}\cos\tau_1^{\mathrm{L}}\right)^{\beta/2}\right)^{-1}$, $N_5=\exp\left(-\frac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{Z}\right)$, $Z=\left(\alpha_{\mathrm{H}}^2-\Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2\right)P_S\ell(x_1^{\mathrm{H}})$, $\ell(x_1^{\mathrm{L}})=\left(1+\left(r_1^{\mathrm{L}}\right)^{\beta}\right)^{-1}$, $\alpha_{\mathrm{H}}^2>\Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2$, and $C(\alpha_{\mathrm{H}})$ is given in (21), as shown at the bottom of this page.

Proof: Please refer to Appendix D. ■

The service process of queue Q_{L} can also be divided into two cases: a) if queue Q_{H} is non-empty, then base station S transmits the first packet from queue Q_{L} to user u_1^{L} using cooperative NOMA; b) if queue Q_{H} is empty, then base station S transmits the first two packets from queue Q_{L} to users u_1^{L} and u_2^{L} using NOMA when $u_1^{\mathrm{L}}\neq u_2^{\mathrm{L}}$ and the first packet from queue Q_{L} to user u_1^{L} using OMA when $u_1^{\mathrm{L}}=u_2^{\mathrm{L}}$. The average service rate of queue Q_{L}, denoted as $\mu_{\mathrm{L}}^{\mathrm{FCN1}}$, is

$$\begin{aligned}\mu_{\mathrm{L}}^{\mathrm{FCN1}}&=\mathbb{P}\left(Q_{\mathrm{H}}(t)>0\right)q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}})+\mathbb{P}(Q_{\mathrm{H}}(t)=0)\\ &\quad\times\left(\left(1-\frac{1}{K}\right)q_{\mathrm{L1L2}}^{\mathrm{ON}}+\frac{1}{K}q_{\mathrm{L1}}^{\mathrm{OMA}}\right),\end{aligned}\tag{22}$$

$$C(\alpha_{\mathrm{H}})=\frac{1}{\pi r^2}\int_0^r\int_0^{2\pi}\frac{\exp\left(-N_4/\ell(x_1^{\mathrm{L}})\right)}{1-\frac{Z}{P_{\mathrm{L}}\ell\left(x_1^{\mathrm{H}}-x_1^{\mathrm{L}}\right)}}\left(\exp\left(-\frac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{P_{\mathrm{L}}\ell(x_1^{\mathrm{H}}-x_1^{\mathrm{L}})}\right)-N_5\right)r_1^{\mathrm{L}}\mathrm{d}r_1^{\mathrm{L}}\mathrm{d}\tau_1^{\mathrm{L}}\tag{21}$$

$$q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}}) = \mathbb{P}\left(\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \Gamma_{\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{L}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{L}}\right) + \mathbb{P}\left(\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) < \Gamma_{\mathrm{th}}^{\mathrm{H}}, \Gamma_{\mathrm{H1}\to\mathrm{L1}}(t,\alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \Gamma_{\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{L}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{L}}\right) \quad (24)$$

where $\mathbb{P}(Q_{\mathrm{H}}(t) > 0) = \lambda_{\mathrm{H}}/\mu_{\mathrm{H}}^{\mathrm{FCN1}}$, $q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}})$ is the probability of successful packet reception at user u_1^{L} when cooperative NOMA is enabled, and $q_{\mathrm{L1L2}}^{\mathrm{ON}}$ and $q_{\mathrm{L1}}^{\mathrm{OMA}}$ are given in (47) and (48), respectively.

Depending on whether or not user u_1^{L} forwards signal $s_1^{\mathrm{H}}(t)$ to user u_1^{H}, the received SINR of signal $s_1^{\mathrm{L}}(t)$ observed at user u_1^{L} in time slot t can be expressed as

$$\Gamma_{\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{L}}) = \begin{cases} \dfrac{\alpha_{\mathrm{L}}^2 P_S \left|h_1^{\mathrm{L}}(t)\right|^2 \ell(x_1^{\mathrm{L}})}{\zeta P_{\mathrm{L}} + \sigma^2}, & \text{if } \Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \\ \dfrac{\alpha_{\mathrm{L}}^2 P_S \left|h_1^{\mathrm{L}}(t)\right|^2 \ell(x_1^{\mathrm{L}})}{\sigma^2}, & \text{if } \Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) < \Gamma_{\mathrm{th}}^{\mathrm{H}}. \end{cases} \quad (23)$$

As a result, we obtain (24), as shown at the top of this page.

The following lemma provides the stability condition for queue Q_{L} in dominant system Φ_1^{FCN}.

Lemma 4: In dominant system Φ_1^{FCN}, queue Q_{L} is stable if

$$\begin{aligned} \lambda_{\mathrm{L}} < \mu_{\mathrm{L}}^{\mathrm{FCN1}} = \frac{\lambda_{\mathrm{H}}}{\mu_{\mathrm{H}}^{\mathrm{FCN1}}} \Bigg(& \frac{2}{r^2\beta} N_6^{-2/\beta} \exp(-N_6)\, \gamma\left(\frac{2}{\beta}, N_6 r^{\beta}\right) \\ & + \frac{2}{r^2\beta} N_1^{-2/\beta} \exp(-N_1) \gamma\left(\frac{2}{\beta}, N_1 r^{\beta}\right) \\ & - \frac{2}{r^2\beta} N_4^{-2/\beta} \exp(-N_4) \gamma\left(\frac{2}{\beta}, N_4 r^{\beta}\right) \Bigg) \\ & + \left(1 - \frac{\lambda_{\mathrm{H}}}{\mu_{\mathrm{H}}^{\mathrm{FCN1}}}\right)\eta, \end{aligned} \quad (25)$$

where $N_6 = \max\left\{\frac{(\zeta P_{\mathrm{L}}+\sigma^2)\Gamma_{\mathrm{th}}^{\mathrm{H}}}{(\alpha_{\mathrm{H}}^2 - \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2)P_S}, \frac{(\zeta P_{\mathrm{L}}+\sigma^2)\Gamma_{\mathrm{th}}^{\mathrm{L}}}{\alpha_{\mathrm{L}}^2\ P_S}\right\}$, $\alpha_{\mathrm{H}}^2 > \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2$, and η is given in (10).

Proof: Please refer to Appendix E. ∎

Based on the average service rates of queues Q_{H} and Q_{L}, the stable throughput region of dominant system Φ_1^{FCN}, denoted as $\mathcal{R}_1^{\mathrm{FCN}}$, can be expressed as

$$\mathcal{R}_1^{\mathrm{FCN}} = \left\{ (\lambda_{\mathrm{H}}, \lambda_{\mathrm{L}}) : \frac{\left(\eta - q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}})\right)\lambda_{\mathrm{H}}}{\eta\left(q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})\right)} + \frac{\lambda_{\mathrm{L}}}{\eta} < 1, \text{ for } 0 \leq \lambda_{\mathrm{H}} < q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{L}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}}) \right\}. \quad (26)$$

According to (26), stable throughput region $\mathcal{R}_1^{\mathrm{FCN}}$ depends on the self-interference cancelation coefficient.

B. Stable Throughput Region of Dominant System Φ_2^{FCN}

Dominant system Φ_2^{FCN}: If queue Q_{H} is empty, then queue Q_{H} contributes a dummy packet, while queue Q_{L} acts in the same manner as in the cooperative NOMA system with full-duplex relaying Φ^{FCN}. In dominant system Φ_2^{FCN}, base station S transmits the first packet from queue Q_{L} to user u_1^{L} using cooperative NOMA. The average service rate of queue Q_{L}, denoted as $\mu_{\mathrm{L}}^{\mathrm{FCN2}}$, can be expressed as $\mu_{\mathrm{L}}^{\mathrm{FCN2}} = q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}})$. Queue Q_{L} in dominant system Φ_2^{FCN} is stable if $\lambda_{\mathrm{L}} < \mu_{\mathrm{L}}^{\mathrm{FCN2}}$. The service process of queue Q_{H} depends on queue Q_{L}. If queue Q_{L} is empty, then base station S transmits the first packet from queue Q_{H} to user u_1^{H} using cooperative OMA. If queue Q_{L} is non-empty, then base station S transmits the first packet from queue Q_{H} to user u_1^{H} using cooperative NOMA when $\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}$ and using NOMA when $\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) < \Gamma_{\mathrm{th}}^{\mathrm{H}}$. Thus, the average service rate of queue Q_{H} in dominant system Φ_2^{FCN}, denoted as $\mu_{\mathrm{H}}^{\mathrm{FCN2}}$, is given by

$$\mu_{\mathrm{H}}^{\mathrm{FCN2}} = \mathbb{P}(Q_{\mathrm{L}}(t) = 0)\, q_{\mathrm{H1}}^{\mathrm{FC}} + \mathbb{P}(Q_{\mathrm{L}}(t) > 0)\left(q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})\right), \quad (27)$$

where $\mathbb{P}(Q_{\mathrm{L}}(t) = 0) = 1 - \lambda_{\mathrm{L}}/\mu_{\mathrm{L}}^{\mathrm{FCN2}}$, $q_{\mathrm{H1}}^{\mathrm{FC}}$ denotes the probability of successful packet reception at user u_1^{H} when cooperative OMA is enabled, and $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}})$ and $q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$ are given in (49) and (52), respectively. When cooperative OMA is enabled, the low-priority users that can successfully decode signal $s_1^{\mathrm{H}}(t)$ are referred to as *qualified relays*, which form the decoding set in time slot t, denoted as $\Omega(t)$ and given by

$$\Omega(t) = \left\{u_k^{\mathrm{R}} \in \mathcal{U}^{\mathrm{L}} : \Gamma_{\mathrm{H1}\to\mathrm{R}k}^{\mathrm{FC}}(t) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}\right\}, \quad (28)$$

where $u_k^{\mathrm{R}} \in \mathcal{U}^{\mathrm{L}}$ denotes the k-th full-duplex relay and $\Gamma_{\mathrm{H1}\to\mathrm{R}k}^{\mathrm{FC}}(t) = \frac{P_S\left|h_k^{\mathrm{R}}(t)\right|^2 \ell(x_k^{\mathrm{R}})}{\zeta P_{\mathrm{L}} + \sigma^2}$.

We assume that, via coordination signaling between base station S and user u_1^{H} before the packet transmission, each qualified relay knows the instantaneous channel gain between itself and user u_1^{H}. If decoding set $\Omega(t)$ is empty, no low-priority user can help forward signal $s_1^{\mathrm{H}}(t)$ to user u_1^{H}. On the other hand, if decoding set $\Omega(t)$ is non-empty, the qualified relay that has the best channel condition with respect to high-priority user u_1^{H} is selected as the best relay, i.e.,

$$u_b^{\mathrm{R}} = \arg\max_{u_k^{\mathrm{R}} \in \Omega(t)} \left\{ P_{\mathrm{L}} \left|g_{1,k}^{\mathrm{HR}}(t)\right|^2 \ell(x_1^{\mathrm{H}} - x_k^{\mathrm{R}}) \right\}. \quad (29)$$

User u_1^{H} can successfully decode signal $s_1^{\mathrm{H}}(t)$ in time slot t if the received SNR is not less than the reception threshold, i.e.,

$$\Gamma_{\mathrm{H1R}b}^{\mathrm{FC}}(t) = \frac{P_S\left|h_1^{\mathrm{H}}(t)\right|^2 \ell(x_1^{\mathrm{H}}) + P_{\mathrm{L}}\left|g_{1,b}^{\mathrm{HR}}(t)\right|^2 \ell(x_b^{\mathrm{R}} - x_1^{\mathrm{H}})}{\sigma^2} \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \quad (30)$$

where $\Gamma_{\mathrm{H1R}b}^{\mathrm{FC}}(t)$ denotes the SNR of signal $s_1^{\mathrm{H}}(t)$ observed at user u_1^{H} in time slot t when user u_b^{R} acts as the full-duplex relay.

The probability of successful packet transmission is the complement of the outage probability. In this context, an outage occurs when high-priority user u_1^{H} fails to decode the packet after constructively combining the signals transmitted by the base station and the best relay u_b^{R}. By selecting the best relay, this outage event is equivalent to the event that all qualified relays are in outage, which means that no low-priority user satisfies the following condition:

$$\Gamma_{\mathrm{H1}\to\mathrm{R}k}^{\mathrm{FC}}(t) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}} \quad \text{and} \; \Gamma_{\mathrm{H1R}k}^{\mathrm{FC}}(t) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}, \; \forall \; u_k^{\mathrm{R}} \in \mathcal{U}^{\mathrm{L}}. \quad (31)$$

The following lemma presents the stability condition for queue Q_{H} in dominant system Φ_2^{FCN}.

Lemma 5: In dominant system Φ_2^{FCN}, queue Q_{H} is stable if

$$\lambda_{\mathrm{H}} < \mu_{\mathrm{H}}^{\mathrm{FCN2}} = \left(1 - \frac{\lambda_{\mathrm{L}}}{\mu_{\mathrm{L}}^{\mathrm{FCN2}}}\right) q_{\mathrm{H1}}^{\mathrm{FC}} + \frac{\lambda_{\mathrm{L}}}{\mu_{\mathrm{L}}^{\mathrm{FCN2}}} \left(q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})\right), \quad (32)$$

where $\mu_{\mathrm{L}}^{\mathrm{FCN2}} = q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}})$, $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}})$ and $q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$ are given in (49) and (52), respectively, and

$$q_{\mathrm{H1}}^{\mathrm{FC}} = \exp\left(-\rho_{\mathrm{H}}\left(1 + r_{\mathrm{H}}^{\beta}\right)\right) + \sum_{j=1}^{K} \binom{K}{j} (-1)^{j+1} (C(1))^j \times \left(1 - \exp\left(-\rho_{\mathrm{H}}\left(1 + r_{\mathrm{H}}^{\beta}\right)\right)\right)^{1-j}. \quad (33)$$

Proof: Please refer to Appendix F. ■

After deriving the average service rates of queues Q_{H} and Q_{L}, the stable throughput region of dominant system Φ_2^{FCN}, denoted as $\mathcal{R}_2^{\mathrm{FCN}}$, can be expressed as

$$\mathcal{R}_2^{\mathrm{FCN}} = \left\{ (\lambda_{\mathrm{H}}, \lambda_{\mathrm{L}}) : \frac{\lambda_{\mathrm{H}}}{q_{\mathrm{H1}}^{\mathrm{FC}}} + \frac{\left(q_{\mathrm{H1}}^{\mathrm{FC}} - q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) - q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})\right)\lambda_{\mathrm{L}}}{q_{\mathrm{H1}}^{\mathrm{FC}} q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}})} < 1, \; \text{for} \quad 0 \leq \lambda_{\mathrm{L}} < q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}}) \right\}. \quad (34)$$

According to (34), stable throughout region $\mathcal{R}_2^{\mathrm{FCN}}$ depends on the number of low-priority users and the self-interference cancelation coefficient.

Based on the above derivations, the following theorem presents the stable throughput region of the cooperative NOMA system with full-duplex relaying Φ^{FCN}.

Theorem 2: The stable throughput region of the cooperative NOMA system with full-duplex relaying Φ^{FCN} for fixed power allocation coefficients, denoted as $\mathcal{R}^{\mathrm{FCN}}$, is the union of the stable throughput regions of dominant systems Φ_1^{FCN} and Φ_2^{FCN}, i.e., $\mathcal{R}^{\mathrm{FCN}} = \mathcal{R}_1^{\mathrm{FCN}} \cup \mathcal{R}_2^{\mathrm{FCN}}$.

Proof: The proof is similar to that of Theorem 1, and hence, it is omitted here. ■

Similarly, we resort to numerical analysis to obtain the full stable throughput region in Section VI.

V. Comparison of NOMA and OMA

In this section, we derive the stable throughput region of a baseline OMA scheme and the conditions under which the proposed NOMA schemes achieve larger stable throughput regions than the baseline OMA scheme.

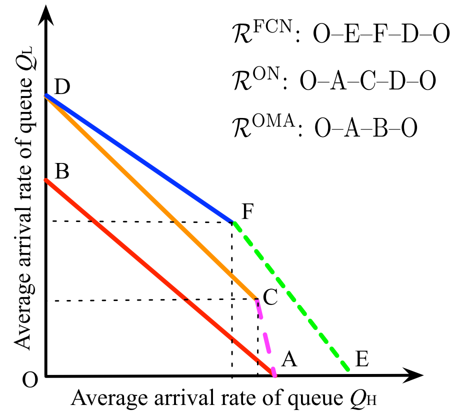


Fig. 2. Stable throughput regions of OMA system Φ^{OMA}, opportunistic NOMA system Φ^{ON}, and cooperative NOMA system with full-duplex relaying Φ^{FCN}.

A. Baseline Orthogonal Multiple Access Scheme

We consider a time division multiple access (TDMA) based OMA system, denoted as Φ^{OMA}, as a baseline, where base station S transmits one packet in one time slot. As queues Q_{H} and Q_{L} do not interact with each other when OMA is utilized, the stability conditions of these two queues can be separately analyzed. Base station S transmits the first packet from queue Q_{H} to high-priority user u_1^{H} whenever queue Q_{H} is not empty, regardless of the status of queue Q_{L}. The average service rate of queue Q_{H} in OMA system Φ^{OMA} is $\mu_{\mathrm{H}}^{\mathrm{OMA}} = \exp\left(-\rho_{\mathrm{H}}\left(1 + r_{\mathrm{H}}^{\beta}\right)\right)$. When queue Q_{H} is empty, base station S transmits the first packet from queue Q_{L} to user u_1^{L}. The average service rate of queue Q_{L} is given by $\mu_{\mathrm{L}}^{\mathrm{OMA}} = \mathbb{P}\left(Q_{\mathrm{H}} = 0\right) \mathbb{P}\left(\Gamma_{\mathrm{L1}}(t, 1) \geq \Gamma_{\mathrm{th}}\right) = \left(1 - \lambda_{\mathrm{H}}/\mu_{\mathrm{H}}^{\mathrm{OMA}}\right) q_{\mathrm{L1}}^{\mathrm{OMA}}$, where $q_{\mathrm{L1}}^{\mathrm{OMA}}$ is given in (48). The stable throughput region of OMA system Φ^{OMA} is given by

$$\mathcal{R}^{\mathrm{OMA}} = \left\{ (\lambda_{\mathrm{H}}, \lambda_{\mathrm{L}}) : \frac{\lambda_{\mathrm{H}}}{\exp\left(-\rho_{\mathrm{H}}\left(1 + r_{\mathrm{H}}^{\beta}\right)\right)} + \frac{\lambda_{\mathrm{L}}}{q_{\mathrm{L1}}^{\mathrm{OMA}}} < 1, \; \text{for} \quad 0 \leq \lambda_{\mathrm{H}} < \exp\left(-\rho_{\mathrm{H}}\left(1 + r_{\mathrm{H}}^{\beta}\right)\right) \right\}. \quad (35)$$

For the queueing model under consideration, the performance comparison between the OMA scheme and the proposed NOMA scheme is fair in the sense that each user is served based on its priority and the order of its packets in the queue, but not based on its CSI.

B. Comparison Between NOMA and OMA

In the following, we present the conditions under which the proposed NOMA schemes achieve larger stable throughput regions than OMA. Based on the stable throughput regions given in (14), (16), (26), (34), and (35), Fig. 2 plots the stable throughput regions of OMA system Φ^{OMA} (i.e., $\mathcal{R}^{\mathrm{OMA}}$: O–A–B–O), opportunistic NOMA system Φ^{ON} (i.e., $\mathcal{R}^{\mathrm{ON}}$: O–A–C–D–O), and the cooperative NOMA system with full-duplex relaying Φ^{FCN} (i.e., $\mathcal{R}^{\mathrm{FCN}}$: O–E–F–D–O). The coordinates of the corner points in Fig. 2 are

$O = (0,0)$, A $= (\mu_{\mathrm{H}}^{\mathrm{OMA}}, 0)$, B $= (0, q_{\mathrm{L1}}^{\mathrm{OMA}})$, C $= \left(\mu_{\mathrm{H}}^{\mathrm{ON1}}, \xi q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}})\right)$, D $= (0, \eta)$, E $= (q_{\mathrm{H1}}^{\mathrm{FC}}, 0)$, and F $= (q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}}), q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}}))$, where $\mu_{\mathrm{H}}^{\mathrm{OMA}} = \exp\left(-\rho_{\mathrm{H}}\left(1 + r_{\mathrm{H}}^{\beta}\right)\right)$ and $\xi = \exp\left(-\theta\left(1 + r_{\mathrm{H}}^{\beta}\right)\right)$.

Proposition 1: The cooperative NOMA scheme with full-duplex relaying achieves a larger stable throughput region than OMA, i.e., $\mathcal{R}^{\mathrm{OMA}} \subset \mathcal{R}^{\mathrm{FCN}}$, when the following conditions hold:

$$q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}}) > q_{\mathrm{L1}}^{\mathrm{OMA}}\left(1 - \frac{q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})}{\mu_{\mathrm{H}}^{\mathrm{OMA}}}\right), \quad (36)$$

$$q_{\mathrm{L1L2}}^{\mathrm{ON}} > q_{\mathrm{L1}}^{\mathrm{OMA}}. \quad (37)$$

Proof: Please refer to Appendix G. ■

As $q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}})$, $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}})$, and $q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$ are functions of α_{L}, it is very difficult to derive a closed-form condition in terms of α_{L} from (36). Nonetheless, we can obtain all possible values of α_{L} that lead to $\mathcal{R}^{\mathrm{OMA}} \subset \mathcal{R}^{\mathrm{FCN}}$ by evaluating (36) numerically, as (37) does not depend on α_{L}. Moreover, the stable throughput region of the cooperative NOMA scheme with full-duplex relaying can be maximized by fixing λ_{H} and then maximizing the corresponding average service rate of queue Q_{L}, i.e., $\mu_{\mathrm{L}}^{\mathrm{FCN1}}$ or $\mu_{\mathrm{L}}^{\mathrm{FCN2}}$, by optimizing the value of α_{L}.

Proposition 2: The opportunistic NOMA scheme achieves a larger stable throughput region than OMA, i.e., $\mathcal{R}^{\mathrm{OMA}} \subset \mathcal{R}^{\mathrm{ON}}$, when $q_{\mathrm{L1L2}}^{\mathrm{ON}} > q_{\mathrm{L1}}^{\mathrm{OMA}}$ and $\xi q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}}) > q_{\mathrm{L1}}^{\mathrm{OMA}}\left(1 - \frac{\mu_{\mathrm{H}}^{\mathrm{ON1}}}{\mu_{\mathrm{H}}^{\mathrm{OMA}}}\right)$ hold.

Proof: The proof is similar to that of Proposition 1, and hence, it is omitted here. ■

VI. Numerical Results

In this section, we evaluate the stable throughput regions of opportunistic NOMA and cooperative NOMA with full-duplex relaying and compare them with the stable throughput region of baseline OMA. The radius of the circular network coverage area is $r = 1.3$ km, where $M = 4$ high-priority users are located $r_{\mathrm{H}} = 1.2$ km away from base station S. The transmit powers (i.e., P_S and P_{L}) and noise power σ^2 are set to be 1 W and -100 dBm, respectively. We consider Rayleigh fading channels and the path loss exponent β is set to be 4. The power allocation coefficients of the far and near users when NOMA is enabled to serve the first two packets from queue Q_{L}, $(\alpha_{\mathrm{f}}^2, \alpha_{\mathrm{n}}^2)$, are set to be $(0.8, 0.2)$.

Fig. 3 shows the impact of the number of low-priority users K and self-interference cancelation coefficient ζ on the probabilities of successful packet reception at the high-priority users when cooperative NOMA and OMA are employed (i.e., $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$ and $q_{\mathrm{H1}}^{\mathrm{FC}}$). The simulation (Sim) results match the analytical (Ana) results well, which validates the performance analysis. We observe that $q_{\mathrm{H1}}^{\mathrm{FC}}$ increases with K, as the probability of selecting a full-duplex relay with good channel condition with respect to user u_1^{H} becomes higher because of the spatial diversity gain. On the other hand, $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$ does not change with K, as the intended receiver of the first packet from queue Q_{L} is selected to act as a full-duplex relay when it can successfully decode signal

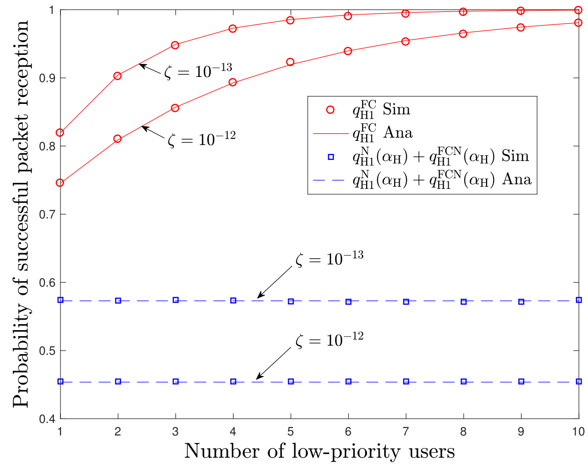


Fig. 3. Probabilities of successful packet reception at user u_1^{H} versus the number of low-priority users K for different self-interference cancelation coefficients, ζ, when $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2) = (0.8, 0.2)$ and $\Gamma_{\mathrm{th}}^{\mathrm{H}} = 2$.

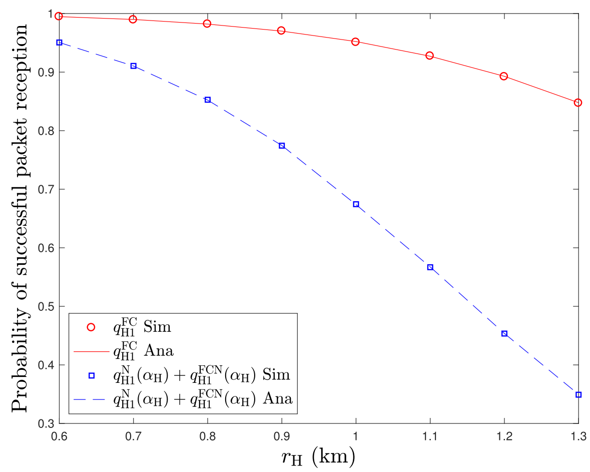


Fig. 4. Probabilities of successful packet reception at user u_1^{H} versus its distance with respect to base station, r_{H}, when $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2) = (0.8, 0.2)$, $\Gamma_{\mathrm{th}}^{\mathrm{H}} = 2$, $\zeta = 10^{-12}$, and $K = 4$.

$s_1^{\mathrm{H}}(t)$ received from base station S, regardless of its channel condition with respect to user u_1^{H}. With better self-interference cancelation (i.e., a smaller value of ζ), the probability of successful packet reception at user u_1^{H} increases for both cooperative NOMA and OMA, as the SINR of signal $s_1^{\mathrm{H}}(t)$ at the low-priority users becomes larger and in turn the probability of selecting a reliable full-duplex relay increases.

Fig. 4 illustrates the impact of the distance between the base station and the high-priority users, r_{H}, on the probabilities of successful packet reception at the high-priority users (i.e., $q_{\mathrm{H1}}^{\mathrm{FC}}$ and $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$). As r_{H} increases, both probabilities decrease because of the larger path loss. As the relay with the best channel condition with respect to user u_1^{H} is selected for cooperative NOMA, the gap between $q_{\mathrm{H1}}^{\mathrm{FC}}$ and $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$ becomes larger as r_{H} increases. In addition, $q_{\mathrm{H1}}^{\mathrm{FC}}$ is always larger than $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$ because of the higher base station transmit power for the high-priority user as well as the higher spatial diversity gain due to relay selection.

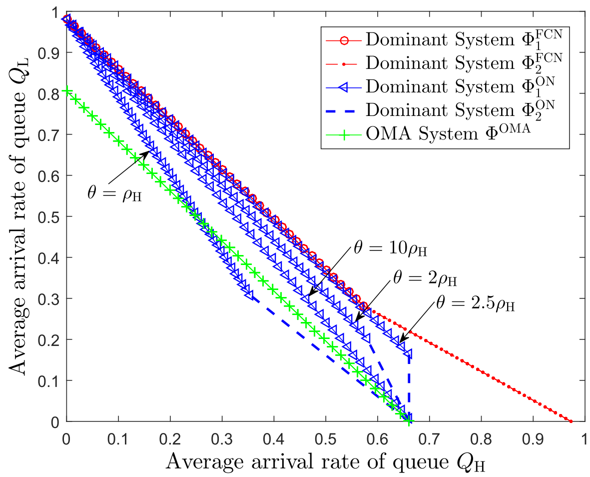


Fig. 5. Stable throughput region for different values of threshold θ when $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2) = (0.8, 0.2)$, $\Gamma_{\mathrm{th}}^{\mathrm{H}} = 2$, $\Gamma_{\mathrm{th}}^{\mathrm{L}} = 2.5$, $K = 4$, and $\zeta = 10^{-13}$.

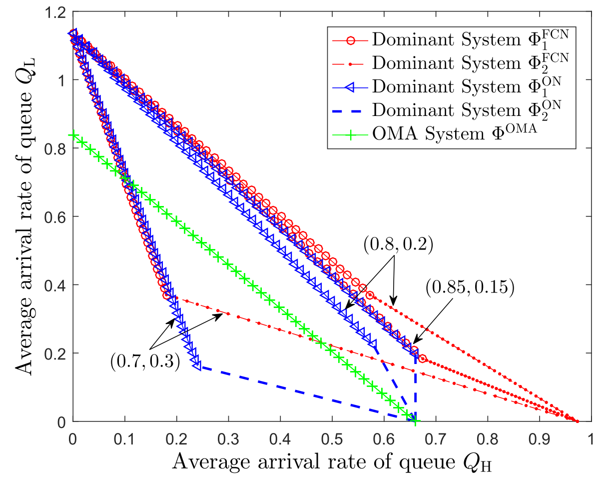


Fig. 6. Stable throughput region for different values of power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$ when $\theta = 2\rho_{\mathrm{H}}$, $\Gamma_{\mathrm{th}}^{\mathrm{H}} = \Gamma_{\mathrm{th}}^{\mathrm{L}} = 2$, $K = 4$, and $\zeta = 10^{-13}$.

Fig. 5 plots the stable throughput region for various values of threshold θ. For opportunistic NOMA system Φ^{ON} and OMA system Φ^{OMA}, the maximum achievable λ_{H} is the same, as the stable throughput region of opportunistic NOMA system Φ^{ON} is equal to the union of the stable throughput regions of dominant systems Φ_1^{ON} and Φ_2^{ON}. The stable throughput region of opportunistic NOMA system Φ^{ON} depends on θ. As θ increases, the probability of enabling NOMA transmission decreases, which reduces the transmission opportunities of the low-priority users. For larger θ, the packet retransmission probability of high-priority users decreases, which in turn provides more transmission opportunities to the low-priority users. With an appropriate choice of θ to balance these two aspects, e.g., $\theta = 2.5\rho_{\mathrm{H}}$ in Fig. 5, opportunistic NOMA can achieve a much larger stable throughput region than OMA. By enabling the low-priority users to act as full-duplex relays and assist the high-priority users, the maximum achievable λ_{H} of the cooperative NOMA system with full-duplex relaying Φ^{FCN} is even larger than that of opportunistic NOMA system Φ^{ON} due to the cooperative diversity gain. The enhanced packet reception reliability for the high-priority users can be exploited to provide more transmission opportunities to the low-priority users. Thereby, the stable throughput region is further enlarged.

Fig. 6 shows the stable throughput region for various values of transmit power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$. When $\left(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2\right) = (0.7, 0.3)$, condition (36) does not hold, and hence, the stable throughput region of the cooperative NOMA system with full-duplex relaying Φ^{FCN} is not larger than that of OMA system Φ^{OMA}. This is because the value of α_{H}^2 is not large enough for the low-priority users to successfully decode the signals intended for the high-priority users, which is the prerequisite for performing SIC. By increasing α_{H}^2 to 0.8, the conditions given in Propositions 1 and 2 hold, and hence the stable throughput regions of both opportunistic NOMA and cooperative NOMA with full-duplex relaying become larger than that of OMA. However, by further increasing α_{H}^2 from 0.8 to 0.85, the stable throughput region of the cooperative NOMA system with full-duplex relaying Φ^{FCN} decreases.

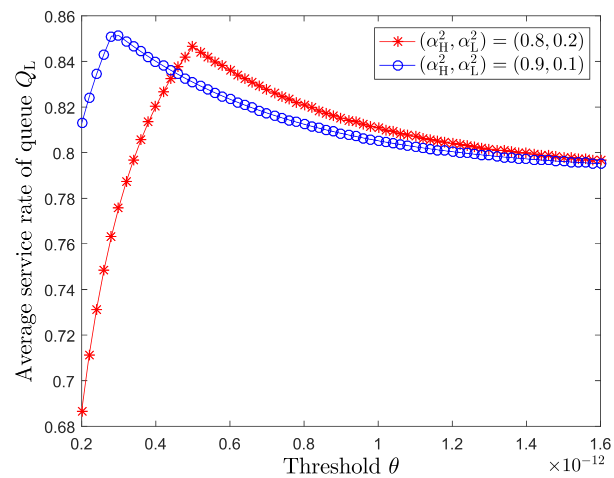


Fig. 7. Average service rate of queue Q_{L} in opportunistic NOMA system versus threshold θ and power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$ when $\lambda_{\mathrm{H}} = 0.2$, $\Gamma_{\mathrm{th}}^{\mathrm{H}} = \Gamma_{\mathrm{th}}^{\mathrm{L}} = 2$, and $K = 4$.

This is because the increased transmission opportunities of the low-priority users cannot compensate for the reduction of successful packet reception at the low-priority users due to the lower transmit power.

Fig. 7 shows the impact of threshold θ and power allocation coefficients on the average service rate of queue Q_{L}. If $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2) = (0.8, 0.2)$, the average service rate of queue Q_{L} increases with θ when $\theta < 0.5 \times 10^{-12}$. By enabling NOMA when the channel gain between base station S and the high-priority users is larger, fewer packet retransmissions are required for high-priority users, which in turn improves the transmission opportunities of low-priority users. The average service rate of queue Q_{L} decreases with θ when $\theta > 0.5 \times 10^{-12}$ and converges to 0.795, as the probability that NOMA is enabled decreases. By increasing α_{H}^2 to 0.9, the optimal threshold θ that maximizes the average service rate of queue Q_{L} becomes smaller, as allocating more transmit power to the high-priority users allows NOMA to be used for smaller channel gain.

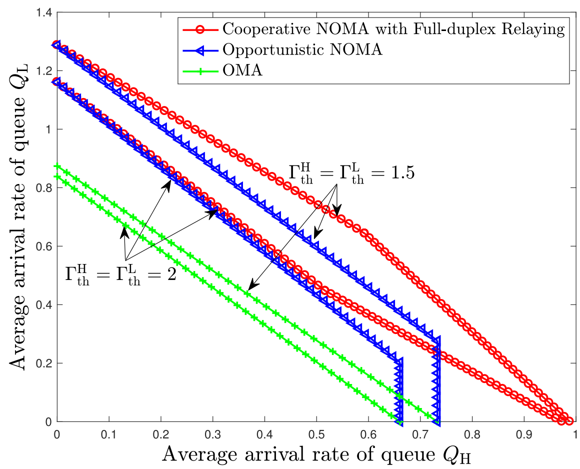


Fig. 8. Full stable throughput region for different values of reception thresholds $\Gamma_{\rm th}^{\rm H}$ and $\Gamma_{\rm th}^{\rm L}$ when $K = 4$ and $\zeta = 10^{-13}$.

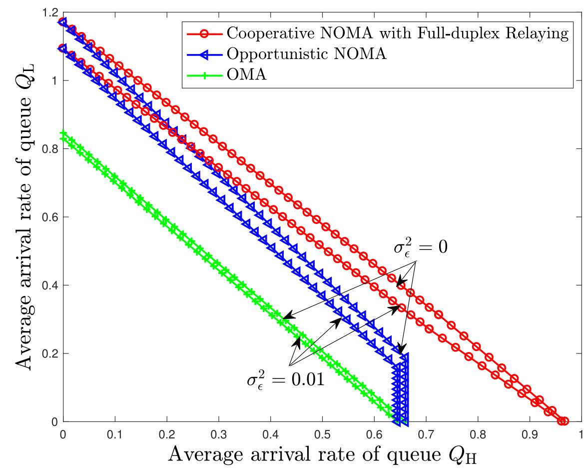


Fig. 10. Impact of imperfect CSI on full stable throughput region when $\Gamma_{\rm th}^{\rm H} = \Gamma_{\rm th}^{\rm L} = 2$, $K = 4$, and $\zeta = 10^{-13}$.

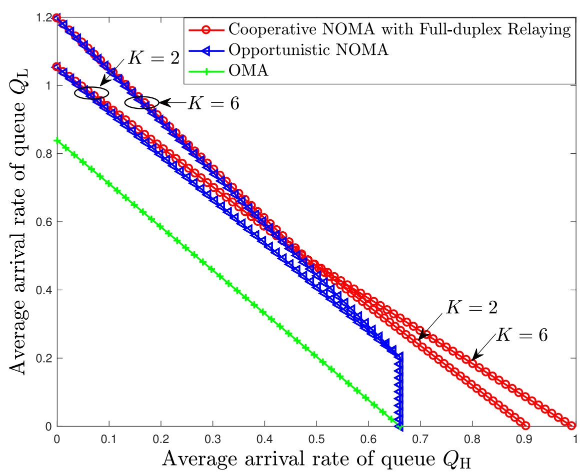


Fig. 9. Full stable throughput region for different numbers of low-priority users K when $\Gamma_{\rm th}^{\rm H} = \Gamma_{\rm th}^{\rm L} = 2$ and $\zeta = 10^{-13}$.

In Fig. 8, we study the impact of reception thresholds $\Gamma_{\rm th}^{\rm H}$ and $\Gamma_{\rm th}^{\rm L}$ on the full stable throughput regions of both proposed NOMA schemes. For a smaller reception threshold, the maximum achievable $\lambda_{\rm H}$ and $\lambda_{\rm L}$ in all schemes increase, as the probability of successful packet reception at each user increases. With a smaller reception threshold, the probability of queue $Q_{\rm H}$ being empty is higher, which leads to more time slots being available for the base station to serve queue $Q_{\rm L}$ using NOMA. Hence, the performance gap between the opportunistic NOMA system (cooperative NOMA system with full-duplex relaying) and the OMA system becomes larger when the reception thresholds are smaller. The average service rate of queue $Q_{\rm L}$ can exceed 1 as NOMA can simultaneously serve two packets from queue $Q_{\rm L}$ when queue $Q_{\rm H}$ is empty.

Fig. 9 shows the impact of the number of low-priority users K on the full stable throughput region. For larger K, the maximum achievable $\lambda_{\rm L}$ of both NOMA schemes increases, as the probability that NOMA can serve the packets from queue $Q_{\rm L}$ increases. The maximum achievable $\lambda_{\rm H}$ of the opportunistic NOMA system does not depend on K. The maximum achievable $\lambda_{\rm H}$ of the cooperative NOMA system with full-duplex relaying increases with K, as more low-priority users are available and the probability of selecting a reliable relay becomes higher. Thus, the full stable throughput regions of both NOMA systems increase with K.

Fig. 10 shows the impact of imperfect CSI estimation on the full stable throughput region. Adopting the model for imperfect CSI in [35], we have $h_i^{\rm a}(t) = \hat{h}_i^{\rm a}(t) + \epsilon_i^{\rm a}(t), i \in \{1,2\}, {\rm a} \in \{{\rm H},{\rm L}\}$, where $\hat{h}_i^{\rm a}(t)$ denotes the estimate of $h_i^{\rm a}(t)$ at user $u_i^{\rm a}$ and $\epsilon_i^{\rm a}(t)$ is the complex Gaussian channel estimation error at user $u_i^{\rm a}$ with zero mean and variance σ_ϵ^2 in time slot t. The value of variance σ_ϵ^2 reflects the accuracy of channel estimation. We first obtain the average service rates of queues $Q_{\rm H}$ and $Q_{\rm L}$ for all considered dominant systems via simulations, which are then used to plot the stable throughput regions in Fig. 10. Results show that, by increasing the value of σ_ϵ^2 from 0 to 0.01, the full stable throughput regions of all considered schemes become smaller. This is because channel estimation errors lead to additional interference as in the SINR expression, which reduces the probability of successful packet reception at each user. In addition, the impact of imperfect CSI on the performance of NOMA is greater compared to OMA, as the SIC at the user being allocated a lower transmit power in NOMA is negatively affected by imperfect CSI.

VII. Conclusion

In this paper, we studied the performance of downlink NOMA transmission with dynamic traffic arrival and spatially random users of different priorities. To reduce the adverse effect of NOMA on high-priority users, we proposed an opportunistic NOMA scheme requiring only limited CSI at the base station. Moreover, we proposed a cooperative NOMA scheme with full-duplex relaying, where the low-priority users assist the high-priority users to enhance the network performance. By utilizing tools from queueing theory and stochastic geometry, we characterized the stable throughput regions of both proposed NOMA schemes by constructing dominant systems to decouple the interacting queues. Simulation results validated the performance analysis. With appropriate parameter setting, the proposed NOMA schemes can significantly improve the transmission opportunities and enhance the stable throughput region compared to OMA.

There are several interesting topics for future work. First, the performance analysis of cooperative NOMA with full-duplex relaying can be extended to multi-cell networks, where the interference from the base stations and the full-duplex relays in other cells has to be taken into account. Second, the proposed performance analysis framework can be extended to the case where each user exploits retransmission diversity [36]. Third, the proposed framework can be extended to the case where more than two users are paired for NOMA transmission.

Appendix A
Proof of Lemma 1

When OMA is enabled, the probability of successful packet reception at user u_1^{H} is given by

$$q_{\mathrm{H1}}^{\mathrm{OMA}}(\theta) = \mathbb{P}\left(\frac{\rho_{\mathrm{H}}}{\ell(x_1^{\mathrm{H}})} \le \left|h_1^{\mathrm{H}}(t)\right|^2 < \frac{\theta}{\ell(x_1^{\mathrm{H}})}\right) \stackrel{(a)}{=} \exp\left(-\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)\right) - \exp\left(-\theta\left(1+r_{\mathrm{H}}^{\beta}\right)\right), \tag{38}$$

if $\theta > \rho_{\mathrm{H}}$, otherwise, $q_{\mathrm{H1}}^{\mathrm{OMA}}(\theta) = 0$, where (a) follows from the exponential distribution of $\left|h_1^{\mathrm{H}}(t)\right|^2$.

On the other hand, when NOMA is enabled, the probability of successful packet reception at high-priority user u_1^{H} can be expressed as

$$\begin{aligned} q_{\mathrm{H1|L1}}^{\mathrm{ON}}(\alpha_{\mathrm{H}}, \theta) &= \mathbb{P}\left(\left|h_1^{\mathrm{H}}(t)\right|^2 \ge \max\left\{\frac{\rho_{\mathrm{H}}}{\alpha_{\mathrm{H}}^2 - \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2}, \theta\right\}\frac{1}{\ell\left(x_1^{\mathrm{H}}\right)}\right) \\ &= \exp\left(-\max\left\{\frac{\rho_{\mathrm{H}}}{\alpha_{\mathrm{H}}^2 - \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2}, \theta\right\}\left(1+r_{\mathrm{H}}^{\beta}\right)\right), \end{aligned} \tag{39}$$

if $\alpha_{\mathrm{H}}^2 > \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2$, otherwise, $q_{\mathrm{H1|L1}}^{\mathrm{ON}}(\alpha_{\mathrm{H}}, \theta) = 0$.

By substituting (38) and (39) into (5), the average service rate of queue Q_{H} is given by

$$\mu_{\mathrm{H}}^{\mathrm{ON1}} = \begin{cases} q_{\mathrm{H1|L1}}^{\mathrm{ON}}(\alpha_{\mathrm{H}}, \theta), & \text{if } \theta \le \rho_{\mathrm{H}}, \\ q_{\mathrm{H1}}^{\mathrm{OMA}}(\theta) + q_{\mathrm{H1|L1}}^{\mathrm{ON}}(\alpha_{\mathrm{H}}, \theta), & \text{if } \theta > \rho_{\mathrm{H}}. \end{cases} \tag{40}$$

According to (40), we set $\theta > \rho_{\mathrm{H}}$ to achieve a higher value of $\mu_{\mathrm{H}}^{\mathrm{ON1}}$. By Loynes' theorem [27], queue Q_{H} is stable if (6) holds.

Appendix B
Proof of Lemma 2

The probability of successful packet reception at low-priority user u_1^{L} when paired with high-priority user u_1^{H} to perform NOMA is given by

$$\begin{aligned} q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}}) &= \mathbb{P}\left(\Gamma_{\mathrm{H1}\to\mathrm{L1}}(t, \alpha_{\mathrm{H}}) \ge \Gamma_{\mathrm{th}}^{\mathrm{H}}, \Gamma_{\mathrm{L1}}(t, \alpha_{\mathrm{L}}) \ge \Gamma_{\mathrm{th}}^{\mathrm{L}}\right) \\ &= \mathbb{P}\left(\left|h_1^{\mathrm{L}}(t)\right|^2 \ge \frac{\rho_{\mathrm{H}}}{\left(\alpha_{\mathrm{H}}^2 - \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2\right)\ell(x_1^{\mathrm{L}})}, \left|h_1^{\mathrm{L}}(t)\right|^2 \ge \frac{\rho_{\mathrm{L}}}{\alpha_{\mathrm{L}}^2\ell(x_1^{\mathrm{L}})}\right) \\ &= \mathbb{E}_{x_1^{\mathrm{L}}}\left[\exp\left(-N_1/\ell(x_1^{\mathrm{L}})\right)\right], \end{aligned} \tag{41}$$

where $\mathbb{E}_{x_1^{\mathrm{L}}}[\cdot]$ denotes the expectation over location coordinate x_1^{L} of low-priority user u_1^{L}. The probability density function (PDF) of the location of low-priority user u_1^{L} is given by $f(x_1^{\mathrm{L}}) = 1/\left(\pi r^2\right)$. Hence, we have

$$\begin{aligned} q_{\mathrm{L1|H1}}^{\mathrm{ON}}(\alpha_{\mathrm{L}}) &= \frac{2}{r^2}\int_0^r \exp\left(-N_1\left(1+\left(r_1^{\mathrm{L}}\right)^{\beta}\right)\right) r_1^{\mathrm{L}}\mathrm{d}r_1^{\mathrm{L}} \\ &= \frac{2}{r^2\beta}N_1^{-2/\beta}\exp\left(-N_1\right)\gamma\left(\frac{2}{\beta}, N_1 r^{\beta}\right). \end{aligned} \tag{42}$$

When queue Q_{H} is empty and the first two packets from queue Q_{L} are intended for different users, base station S transmits the first and second packets from queue Q_{L} using NOMA based on the distances between their intended receivers and the base station. Among these two users, the near and far users are denoted as $u_{\mathrm{n}}^{\mathrm{L}}$ and $u_{\mathrm{f}}^{\mathrm{L}}$ with distances $r_{\mathrm{n}}^{\mathrm{L}}$ and $r_{\mathrm{f}}^{\mathrm{L}}$, respectively, and $r_{\mathrm{n}}^{\mathrm{L}} \le r_{\mathrm{f}}^{\mathrm{L}}$. Hence, we have $\alpha_{\mathrm{f}} \ge \alpha_{\mathrm{n}}$. As users u_1^{L} and u_2^{L} follow the same location distribution, they have the same probability (i.e., 0.5) of being the near or far user. For instance, if $r_1^{\mathrm{L}} \le r_2^{\mathrm{L}}$, then we have $u_{\mathrm{n}}^{\mathrm{L}} = u_1^{\mathrm{L}}$ and $u_{\mathrm{f}}^{\mathrm{L}} = u_2^{\mathrm{L}}$. Otherwise, we have $u_{\mathrm{n}}^{\mathrm{L}} = u_2^{\mathrm{L}}$ and $u_{\mathrm{f}}^{\mathrm{L}} = u_1^{\mathrm{L}}$. The paired NOMA users having the same low priority are ordered based on their distances to the base station. Due to the uniform distribution of users u_1^{L} and u_2^{L}, according to [37], the PDF of the distance between far user $u_{\mathrm{f}}^{\mathrm{L}}$ and base station S is given by

$$f(r_{\mathrm{f}}^{\mathrm{L}}) = 4\left(r_{\mathrm{f}}^{\mathrm{L}}\right)^3 \Big/ r^4, \quad 0 \le r_{\mathrm{f}}^{\mathrm{L}} \le r. \tag{43}$$

When NOMA is enabled, the probability of successful packet reception at user $u_{\mathrm{f}}^{\mathrm{L}}$ is given by

$$\begin{aligned} q_{\mathrm{Lf|Ln}}^{\mathrm{ON}}(\alpha_{\mathrm{f}}) &= \mathbb{P}\left(\Gamma_{\mathrm{Lf|Ln}}(t, \alpha_{\mathrm{f}}) \ge \Gamma_{\mathrm{th}}^{\mathrm{L}}\right) \\ &= \mathbb{E}_{x_{\mathrm{f}}^{\mathrm{L}}}\left[\exp\left(-N_2/\ell(x_{\mathrm{f}}^{\mathrm{L}})\right)\right] \\ &\stackrel{(a)}{=} \frac{4}{r^4}\int_0^r \exp\left(-N_2\left(1+\left(r_{\mathrm{f}}^{\mathrm{L}}\right)^{\beta}\right)\right)\left(r_{\mathrm{f}}^{\mathrm{L}}\right)^3\mathrm{d}r_{\mathrm{f}}^{\mathrm{L}} \\ &= \frac{4}{r^4\beta}N_2^{-4/\beta}\exp\left(-N_2\right)\gamma\left(\frac{4}{\beta}, N_2 r^{\beta}\right), \end{aligned} \tag{44}$$

if $\alpha_{\mathrm{f}}^2 > \Gamma_{\mathrm{th}}^{\mathrm{L}}\alpha_{\mathrm{n}}^2$, otherwise, $q_{\mathrm{Lf|Ln}}^{\mathrm{ON}}(\alpha_{\mathrm{f}}) = 0$, where (a) follows by substituting (43).

Similarly, the PDF of the distance between near user $u_{\mathrm{n}}^{\mathrm{L}}$ and base station S is given by

$$f(r_{\mathrm{n}}^{\mathrm{L}}) = 4\frac{r_{\mathrm{n}}^{\mathrm{L}}}{r^2}\left(1 - \frac{(r_{\mathrm{n}}^{\mathrm{L}})^2}{r^2}\right), \quad 0 \le r_{\mathrm{n}}^{\mathrm{L}} \le r_{\mathrm{L}}. \tag{45}$$

When NOMA is enabled, the probability of successful packet reception at user $u_{\mathrm{n}}^{\mathrm{L}}$ is given by

$$\begin{aligned} q_{\mathrm{Ln|Lf}}^{\mathrm{ON}}(\alpha_{\mathrm{n}}) &= \mathbb{P}\left(\Gamma_{\mathrm{Lf}\to\mathrm{Ln}}(t, \alpha_{\mathrm{f}}) \ge \Gamma_{\mathrm{th}}^{\mathrm{L}}, \Gamma_{\mathrm{Ln}}(t, \alpha_{\mathrm{n}}) \ge \Gamma_{\mathrm{th}}^{\mathrm{L}}\right) \\ &= \mathbb{E}_{x_{\mathrm{n}}^{\mathrm{L}}}\left[\exp\left(-N_3/\ell(x_{\mathrm{n}}^{\mathrm{L}})\right)\right] \\ &\stackrel{(a)}{=} \frac{4}{r^2}\int_0^r \exp\left(-N_3\left(1+(r_{\mathrm{n}}^{\mathrm{L}})^{\beta}\right)\right)\left(r_{\mathrm{n}}^{\mathrm{L}} - \frac{(r_{\mathrm{n}}^{\mathrm{L}})^3}{r^2}\right)\mathrm{d}r_{\mathrm{n}}^{\mathrm{L}} \\ &= \frac{4}{r^2\beta}N_3^{-2/\beta}\exp\left(-N_3\right)\gamma\left(\frac{2}{\beta}, N_3 r^{\beta}\right) \\ &\quad - \frac{4}{r^4\beta}N_3^{-4/\beta}\gamma\left(\frac{4}{\beta}, N_3 r^{\beta}\right), \end{aligned} \tag{46}$$

if $\alpha_{\mathrm{f}}^2 > \Gamma_{\mathrm{th}}^{\mathrm{L}}\alpha_{\mathrm{n}}^2$, otherwise, $q_{\mathrm{Ln|Lf}}^{\mathrm{ON}}(\alpha_{\mathrm{n}}) = 0$, where (a) follows by substituting (45).

Based on (44) and (46), we have

$$q_{\mathrm{L1L2}}^{\mathrm{ON}} = q_{\mathrm{Lf|Ln}}^{\mathrm{ON}}(\alpha_{\mathrm{f}}) + q_{\mathrm{Ln|Lf}}^{\mathrm{ON}}(\alpha_{\mathrm{n}}). \tag{47}$$

When OMA is enabled, the probability of successful packet reception at user u_1^{L} is given by

$$\begin{aligned} q_{\mathrm{L1}}^{\mathrm{OMA}} &= \mathbb{P}\left(\Gamma_{\mathrm{L1}}(t,1) \geq \Gamma_{\mathrm{th}}^{\mathrm{L}}\right) \\ &= \frac{2}{r^2\beta}\rho_{\mathrm{L}}^{-2/\beta}\exp(-\rho_{\mathrm{L}})\gamma\left(\frac{2}{\beta}, \rho_{\mathrm{L}} r^{\beta}\right). \end{aligned} \tag{48}$$

By substituting (9), (47), and (48) into (7), the average service rate of queue Q_{L} in dominant system Φ_1^{ON} can be derived. By Loynes' theorem, queue Q_{L} is stable if (8) holds.

Appendix C
Proof of Theorem 1

Our proof is based on a similar technique as the proofs in [16]–[18]. The dominant systems (i.e., Φ_1^{ON} and Φ_2^{ON}) are modifications of the original opportunistic NOMA system Φ^{ON}. The queue lengths in the dominant systems are never shorter than the queue lengths in the original opportunistic NOMA system Φ^{ON} as an empty queue can contribute dummy packets. The transmission of dummy packets reduces the probability of successful packet reception by generating co-channel interference, but does not contribute to the throughput. Hence, the stability condition obtained for the dominant systems (i.e., Φ_1^{ON} and Φ_2^{ON}) is sufficient for the stability of the original opportunistic NOMA system Φ^{ON}.

As only two queues are considered, the stability condition of the original opportunistic NOMA system Φ^{ON} is determined by the two parallel dominant systems (i.e., Φ_1^{ON} and Φ_2^{ON}). In particular, dominant systems Φ_1^{ON} and Φ_2^{ON} explore all possible choices of the average arrival rates λ_{L} and λ_{H} that can lead to a stable system, respectively. In dominant system Φ_1^{ON}, some λ_{L} would cause queue Q_{L} to be always non-empty. As long as queue Q_{L} always has packets to transmit, queue Q_{L} does not contribute dummy packets and hence the behavior of dominant system Φ_1^{ON} is identical to that of the original opportunistic NOMA system Φ^{ON}. As a result, dominant system Φ_1^{ON} and the original opportunistic NOMA system Φ^{ON} are indistinguishable at the boundary of the stability region (i.e., line CD in Fig. 2). Similarly, dominant system Φ_2^{ON} and the original opportunistic NOMA system Φ^{ON} are also indistinguishable at the boundary of the stability region (i.e., line AC in Fig. 2). Similar indistinguishability arguments are used in [16]–[18]. Thereby, the stability condition obtained for the dominant systems (i.e., Φ_1^{ON} and Φ_2^{ON}) is also necessary for the stability of the original opportunistic NOMA system Φ^{ON}. As a result, we have $\mathcal{R}^{\mathrm{ON}} = \mathcal{R}_1^{\mathrm{ON}} \cup \mathcal{R}_2^{\mathrm{ON}}$.

Appendix D
Proof of Lemma 3

Due to the independence of events $\{\Gamma_{\mathrm{H1|L1}}(t,\alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}\}$ and $\{\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) < \Gamma_{\mathrm{th}}^{\mathrm{H}}\}$, the probability of successful packet reception at user u_1^{H} when NOMA is enabled, denoted as $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}})$, is given by

$$\begin{aligned} &q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) \\ &= \mathbb{P}\left(\Gamma_{\mathrm{H1|L1}}(t,\alpha_{\mathrm{H}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}\right)\mathbb{P}\left(\Gamma_{\mathrm{H1}\to\mathrm{L1}}^{\mathrm{FCN}}(t,\alpha_{\mathrm{H}}) < \Gamma_{\mathrm{th}}^{\mathrm{H}}\right) \\ &= \exp\left(-\frac{\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)}{\alpha_{\mathrm{H}}^2 - \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2}\right)\mathbb{E}_{x_1^{\mathrm{L}}}\left[1 - \exp\left(-\frac{N_4}{\ell(x_1^{\mathrm{L}})}\right)\right] \\ &= \exp\left(-\frac{\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)}{\alpha_{\mathrm{H}}^2 - \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2}\right) \\ &\quad\times\left(1 - \frac{2}{r^2\beta}N_4^{-2/\beta}\exp\left(-N_4\right)\gamma\left(\frac{2}{\beta}, N_4 r^{\beta}\right)\right), \end{aligned} \tag{49}$$

if $\alpha_{\mathrm{H}}^2 > \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2$, otherwise, $q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) = 0$.

The probability of successful packet reception at user u_1^{H} when cooperative NOMA is enabled, denoted as $q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})$, can be expressed as

$$\begin{aligned} q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}}) \overset{(a)}{=} \mathbb{E}_{x_1^{\mathrm{L}}}\Bigg[&\mathbb{P}\left(\left(\alpha_{\mathrm{H}}^2 - \Gamma_{\mathrm{th}}^{\mathrm{H}}\alpha_{\mathrm{L}}^2\right)P_S\left|h_1^{\mathrm{H}}(t)\right|^2\ell(x_1^{\mathrm{H}})\right. \\ &\left.+ P_{\mathrm{L}}\left|g_{1,1}^{\mathrm{HL}}(t)\right|^2\ell(x_1^{\mathrm{H}} - x_1^{\mathrm{L}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2\right) \\ &\times\mathbb{P}\left(\left|h_1^{\mathrm{L}}(t)\right|^2 \geq \frac{N_4}{\ell(x_1^{\mathrm{L}})}\right)\Bigg], \end{aligned} \tag{50}$$

$$\begin{aligned} &\mathbb{P}\left(Z\left|h_1^{\mathrm{H}}(t)\right|^2 + P_{\mathrm{L}}\left|g_{1,1}^{\mathrm{HL}}(t)\right|^2\ell(x_1^{\mathrm{H}} - x_1^{\mathrm{L}}) \geq \Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2\right) \\ &= \int_0^{\frac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{Z}}\int_{\frac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2 - \left|h_1^{\mathrm{H}}(t)\right|^2 Z}{P_{\mathrm{L}}\ell(x_1^{\mathrm{H}} - x_1^{\mathrm{L}})}}^{\infty}\exp\left(-\left|h_1^{\mathrm{H}}(t)\right|^2\right)\exp\left(-\left|g_{1,1}^{\mathrm{HL}}(t)\right|^2\right)\mathrm{d}\left|g_{1,1}^{\mathrm{HL}}(t)\right|^2\mathrm{d}\left|h_1^{\mathrm{H}}(t)\right|^2 \\ &\quad+ \int_{\frac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{Z}}^{\infty}\int_0^{\infty}\exp\left(-\left|h_1^{\mathrm{H}}(t)\right|^2\right)\exp\left(-\left|g_{1,1}^{\mathrm{HL}}(t)\right|^2\right)\mathrm{d}\left|g_{1,1}^{\mathrm{HL}}(t)\right|^2\mathrm{d}\left|h_1^{\mathrm{H}}(t)\right|^2 \\ &= \begin{cases} \dfrac{1}{1 - \frac{Z}{P_{\mathrm{L}}\ell(x_1^{\mathrm{H}} - x_1^{\mathrm{L}})}}\left(\exp\left(-\dfrac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{P_{\mathrm{L}}\ell(x_1^{\mathrm{H}} - x_1^{\mathrm{L}})}\right) - \exp\left(-\dfrac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{Z}\right)\right) + \exp\left(-\dfrac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{Z}\right), & \text{if } Z \neq P_{\mathrm{L}}\ell(x_1^{\mathrm{H}} - x_1^{\mathrm{L}}), \\ \dfrac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{Z}\exp\left(-\dfrac{\Gamma_{\mathrm{th}}^{\mathrm{H}}\sigma^2}{P_{\mathrm{L}}\ell\left(x_1^{\mathrm{H}} - x_1^{\mathrm{L}}\right)}\right), & \text{if } Z = P_{\mathrm{L}}\ell(x_1^{\mathrm{H}} - x_1^{\mathrm{L}}). \end{cases} \end{aligned} \tag{51}$$

$$
\begin{aligned}
&\mathbb{P}\left(\Gamma^{\mathrm{FCN}}_{\mathrm{H1}\rightarrow\mathrm{L1}}(t,\alpha_{\mathrm{H}}) < \Gamma^{\mathrm{H}}_{\mathrm{th}}, \Gamma_{\mathrm{H1}\rightarrow\mathrm{L1}}(t,\alpha_{\mathrm{H}}) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}, \Gamma^{\mathrm{FCN}}_{\mathrm{L1}}(t,\alpha_{\mathrm{L}}) \geq \Gamma^{\mathrm{L}}_{\mathrm{th}}\right)\\
&\quad= \mathbb{P}\left(\max\left\{\frac{\Gamma^{\mathrm{H}}_{\mathrm{th}}}{\alpha^2_{\mathrm{H}} - \Gamma^{\mathrm{H}}_{\mathrm{th}}\alpha^2_{\mathrm{L}}}, \frac{\Gamma^{\mathrm{L}}_{\mathrm{th}}}{\alpha^2_{\mathrm{L}}}\right\}\frac{\sigma^2}{P_S\ell(x^{\mathrm{L}}_1)} \leq \left|h^{\mathrm{L}}_1(t)\right|^2 < \frac{\left(\zeta P_{\mathrm{L}} + \sigma^2\right)\Gamma^{\mathrm{H}}_{\mathrm{th}}}{\left(\alpha^2_{\mathrm{H}} - \Gamma^{\mathrm{H}}_{\mathrm{th}}\alpha^2_{\mathrm{L}}\right)P_S\ell(x^{\mathrm{L}}_1)}\right)\\
&\quad= \mathbb{E}_{x^{\mathrm{L}}_1}\left[\exp\left(-\max\left\{\frac{\rho_{\mathrm{H}}}{\alpha^2_{\mathrm{H}} - \Gamma^{\mathrm{H}}_{\mathrm{th}}\alpha^2_{\mathrm{L}}}, \frac{\rho_{\mathrm{L}}}{\alpha^2_{\mathrm{L}}}\right\}\frac{1}{\ell(x^{\mathrm{L}}_1)}\right) - \exp\left(-\frac{\left(\zeta P_{\mathrm{L}} + \sigma^2\right)\Gamma^{\mathrm{H}}_{\mathrm{th}}}{\left(\alpha^2_{\mathrm{H}} - \Gamma^{\mathrm{H}}_{\mathrm{th}}\alpha^2_{\mathrm{L}}\right)P_S\ell(x^{\mathrm{L}}_1)}\right)\right]\\
&\quad= \frac{2}{r^2\beta}N_1^{-2/\beta}\exp(-N_1)\gamma\left(\frac{2}{\beta}, N_1 r^{\beta}\right) - \frac{2}{r^2\beta}N_4^{-2/\beta}\exp(-N_4)\gamma\left(\frac{2}{\beta}, N_4 r^{\beta}\right), \qquad (53)
\end{aligned}
$$

where (a) follows from the independent channel fading assumption across different links.

By conditioning on location coordinate x^{L}_1, we obtain (51), as shown at the bottom of previous page.

By substituting (51) into (50), we have

$$
\begin{aligned}
&q^{\mathrm{FCN}}_{\mathrm{H1}}(\alpha_{\mathrm{H}})\\
&\quad= \mathbb{E}_{x^{\mathrm{L}}_1}\left[\left(\frac{\left(\exp\left(-\frac{\Gamma^{\mathrm{H}}_{\mathrm{th}}\sigma^2}{P_{\mathrm{L}}\ell\left(x^{\mathrm{H}}_1 - x^{\mathrm{L}}_1\right)}\right) - N_5\right)}{1 - \frac{Z}{P_{\mathrm{L}}\ell\left(x^{\mathrm{H}}_1 - x^{\mathrm{L}}_1\right)}} + N_5\right)\right.\\
&\qquad\left.\times\exp\left(-\frac{N_4}{\ell(x^{\mathrm{L}}_1)}\right)\right]\\
&\quad= C(\alpha_{\mathrm{H}}) + \frac{2N_5}{r^2\beta}N_4^{-2/\beta}\exp(-N_4)\gamma\left(\frac{2}{\beta}, N_4 r^{\beta}\right), \qquad (52)
\end{aligned}
$$

where $C(\alpha_{\mathrm{H}})$ is given in (21), which can be calculated numerically using commercial software (e.g., Mathematica). By substituting (49) and (52) into (18), we obtain the average service rate of queue Q_{H}. Hence, queue Q_{H} is stable if $\lambda_{\mathrm{H}} < \mu^{\mathrm{FCN1}}_{\mathrm{H}} = q^{\mathrm{N}}_{\mathrm{H1}}(\alpha_{\mathrm{H}}) + q^{\mathrm{FCN}}_{\mathrm{H1}}(\alpha_{\mathrm{H}})$.

APPENDIX E
PROOF OF LEMMA 4

The probability of successful packet reception at user u^{L}_1 when cooperative NOMA is enabled, if $\alpha^2_{\mathrm{H}} > \Gamma_{\mathrm{th}}\alpha^2_{\mathrm{L}}$, can be expressed as

$$
\begin{aligned}
&\mathbb{P}\left(\Gamma^{\mathrm{FCN}}_{\mathrm{H1}\rightarrow\mathrm{L1}}(t,\alpha_{\mathrm{H}}) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}, \Gamma^{\mathrm{FCN}}_{\mathrm{L1}}(t,\alpha_{\mathrm{L}}) \geq \Gamma^{\mathrm{L}}_{\mathrm{th}}\right)\\
&\quad= \mathbb{E}_{x^{\mathrm{L}}_1}\left[\exp\left(-\max\left\{\frac{\Gamma^{\mathrm{H}}_{\mathrm{th}}}{\alpha^2_{\mathrm{H}} - \Gamma^{\mathrm{H}}_{\mathrm{th}}\alpha^2_{\mathrm{L}}}, \frac{\Gamma^{\mathrm{L}}_{\mathrm{th}}}{\alpha^2_{\mathrm{L}}}\right\}\frac{\left(\zeta P_{\mathrm{L}} + \sigma^2\right)}{P_S\ell(x^{\mathrm{L}}_1)}\right)\right]\\
&\quad= \frac{2}{r^2\beta}N_6^{-2/\beta}\exp\left(-N_6\right)\gamma\left(\frac{2}{\beta}, N_6 r^{\beta}\right).
\end{aligned}
$$

When NOMA is enabled, the probability of successful packet reception at user u^{L}_1 is given by (53), as shown at the top of this page, where $\alpha^2_{\mathrm{H}} > \Gamma^{\mathrm{H}}_{\mathrm{th}}\alpha^2_{\mathrm{L}}$. By substituting (24), (47), and (48) into (22), we can obtain the average service rate of queue Q_{L}. Hence, queue Q_{L} is stable if $\lambda_{\mathrm{L}} < \mu^{\mathrm{FCN1}}_{\mathrm{L}} = \frac{\lambda_{\mathrm{H}}}{\mu^{\mathrm{FCN1}}_{\mathrm{H}}}q^{\mathrm{FCN}}_{\mathrm{L1|H1}}(\alpha_{\mathrm{L}}) + \left(1 - \frac{\lambda_{\mathrm{H}}}{\mu^{\mathrm{FCN1}}_{\mathrm{H}}}\right)\eta$, where η is given in (10).

APPENDIX F
PROOF OF LEMMA 5

Given that there are K low-priority users, we have

$$
\begin{aligned}
q^{\mathrm{FC}}_{\mathrm{H1}} &= 1 - \mathbb{P}\left(\Gamma_{\mathrm{H1}}(t,1) < \Gamma^{\mathrm{H}}_{\mathrm{th}}\right)\\
&\quad\times\mathbb{E}_{x^{\mathrm{R}}_k}\left[\prod_{u^{\mathrm{R}}_k\in\mathcal{U}^{\mathrm{L}}}\left[1 - \mathbb{P}\left(\Gamma^{\mathrm{FC}}_{\mathrm{H1}\rightarrow\mathrm{R}k}(t) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}\right)\right.\right.\\
&\quad\left.\left.\times\mathbb{P}\left(\Gamma^{\mathrm{FC}}_{\mathrm{H1R}k}(t) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}\left|\Gamma_{\mathrm{H1}}(t,1) < \Gamma^{\mathrm{H}}_{\mathrm{th}}\right.\right)\right]\right], \qquad (54)
\end{aligned}
$$

where $\mathbb{P}\left(\Gamma_{\mathrm{H1}}(t,1) < \Gamma^{\mathrm{H}}_{\mathrm{th}}\right) = 1 - \exp\left(-\rho_{\mathrm{H}}\left(1 + r^{\beta}_{\mathrm{H}}\right)\right)$. As the K low-priority users are uniformly distributed within the network coverage area, we have

$$
\begin{aligned}
&\mathbb{E}_{x^{\mathrm{R}}_k}\left[\prod_{u^{\mathrm{R}}_k\in\mathcal{U}^{\mathrm{L}}}\left[1 - \mathbb{P}\left(\Gamma^{\mathrm{FC}}_{\mathrm{H1}\rightarrow\mathrm{R}k}(t) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}\right)\right.\right.\\
&\quad\left.\left.\times\mathbb{P}\left(\Gamma^{\mathrm{FC}}_{\mathrm{H1R}k}(t) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}\left|\Gamma_{\mathrm{H1}}(t,1) < \Gamma^{\mathrm{H}}_{\mathrm{th}}\right.\right)\right]\right]\\
&\overset{(a)}{=}\left(\frac{1}{\pi r^2}\int_0^r\int_0^{2\pi}\left(1 - \mathbb{P}\left(\Gamma^{\mathrm{FC}}_{\mathrm{H1}\rightarrow\mathrm{R}k}(t) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}\right)\right.\right.\\
&\quad\left.\left.\times\mathbb{P}\left(\Gamma^{\mathrm{FC}}_{\mathrm{H1R}k}(t) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}\left|\Gamma_{\mathrm{H1}}(t,1) < \Gamma^{\mathrm{H}}_{\mathrm{th}}\right.\right)\right)r^{\mathrm{R}}_k\mathrm{d}\tau^{\mathrm{R}}_k\mathrm{d}r^{\mathrm{R}}_k\right)^K\\
&\overset{(b)}{=}\left(\frac{1}{\pi r^2}\int_0^r\int_0^{2\pi}\left(1 - \mathbb{P}\left(\Gamma^{\mathrm{FC}}_{\mathrm{H1}\rightarrow\mathrm{R}k}(t) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}\right)\right.\right.\\
&\quad\left.\left.\times\frac{\mathbb{P}\left(\Gamma^{\mathrm{FC}}_{\mathrm{H1R}k}(t) \geq \Gamma^{\mathrm{H}}_{\mathrm{th}}, \Gamma_{\mathrm{H1}}(t,1) < \Gamma^{\mathrm{H}}_{\mathrm{th}}\right)}{\mathbb{P}\left(\Gamma_{\mathrm{H1}}(t,1) < \Gamma^{\mathrm{H}}_{\mathrm{th}}\right)}\right)r^{\mathrm{R}}_k\mathrm{d}\tau^{\mathrm{R}}_k\mathrm{d}r^{\mathrm{R}}_k\right)^K\\
&\overset{(c)}{=}\left(1 - \frac{C(1)}{1 - \exp\left(-\rho_{\mathrm{H}}\left(1 + r^{\beta}_{\mathrm{H}}\right)\right)}\right)^K, \qquad (55)
\end{aligned}
$$

where (a) follows from the probability generating functional (PGFL) of the BPP [22], (b) follows from the definition of conditional probability, (c) is obtained using similar steps as for deriving (50), and $C(1)$ is given in Lemma 3 by setting $\alpha^2_{\mathrm{H}} = 1$ and $\alpha^2_{\mathrm{L}} = 0$.

By substituting (55) into (54), we have

$$\begin{aligned} q_{\mathrm{H1}}^{\mathrm{FC}} &= 1-\left(1-\exp\left(-\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)\right)\right) \\ &\quad\times\left(1-\frac{C(1)}{1-\exp\left(-\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)\right)}\right)^{K} \\ &\overset{(a)}{=} \exp\left(-\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)\right)+\sum_{j=1}^{K}\binom{K}{j}(-1)^{j+1}(C(1))^{j} \\ &\quad\times\left(1-\exp\left(-\rho_{\mathrm{H}}\left(1+r_{\mathrm{H}}^{\beta}\right)\right)\right)^{1-j}, \end{aligned} \tag{56}$$

where (a) follows from the binomial expansion. By substituting (56) into (27), we can derive the average service rate of queue Q_{H} in dominant system Φ_2^{FCN}. Hence, queue Q_{H} is stable if $\lambda_{\mathrm{H}} < \mu_{\mathrm{H}}^{\mathrm{FCN2}} = \left(1-\lambda_{\mathrm{L}}/\mu_{\mathrm{L}}^{\mathrm{FCN2}}\right) q_{\mathrm{H1}}^{\mathrm{FC}} + \lambda_{\mathrm{L}}/\mu_{\mathrm{L}}^{\mathrm{FCN2}} \left(q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}}) + q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})\right)$.

Appendix G
Proof of Proposition 1

Based on Fig. 2, in order for $\mathcal{R}^{\mathrm{OMA}} \subset \mathcal{R}^{\mathrm{FCN}}$ to hold, points D, E, and F have to be on the right side of line AB. To ensure that point E is on the right side of line AB, condition $q_{\mathrm{H1}}^{\mathrm{FC}} > \mu_{\mathrm{H}}^{\mathrm{OMA}}$ should hold, which is obtained based on the coordinates of points A and E. According to (56), condition $q_{\mathrm{H1}}^{\mathrm{FC}} > \mu_{\mathrm{H}}^{\mathrm{OMA}}$ always holds. In addition, to guarantee that point F is on the right side of line AB, the Y-coordinate of point F should be larger than the Y-coordinate of the point that is on line AB and has the same X-coordinate as point F. Hence, condition $q_{\mathrm{L1|H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{L}}) > q_{\mathrm{L1}}^{\mathrm{OMA}}\left(1-\frac{q_{\mathrm{H1}}^{\mathrm{N}}(\alpha_{\mathrm{H}})+q_{\mathrm{H1}}^{\mathrm{FCN}}(\alpha_{\mathrm{H}})}{\mu_{\mathrm{H}}^{\mathrm{OMA}}}\right)$ should hold, where $\alpha_{\mathrm{L}}^2 = 1-\alpha_{\mathrm{H}}^2$. Similarly, to ensure that point D is on the right side of line AB, condition $\eta > q_{\mathrm{L1}}^{\mathrm{OMA}}$, equivalent to $q_{\mathrm{L1L2}}^{\mathrm{ON}} > q_{\mathrm{L1}}^{\mathrm{OMA}}$, should hold, where $q_{\mathrm{L1L2}}^{\mathrm{ON}}$ is given in (47). As a result, cooperative NOMA with full-duplex relaying achieves a larger stable throughput region than OMA if both (36) and (37) hold.

References

[1] Y. Zhou and V. W.S. Wong, "Stable throughput region of downlink NOMA transmissions with limited CSI," in *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–7.

[2] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE PIMRC*, London, U.K., Sep. 2013, pp. 611–615.

[3] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proc. IEEE Globecom Workshops*, Atlanta, GA, USA, Dec. 2013, pp. 66–70.

[4] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, May 2016.

[5] B. Di, S. Bayat, L. Song, and Y. Li, "Radio resource allocation for downlink non-orthogonal multiple access (NOMA) networks using matching theory," in *Proc. IEEE Globecom*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[6] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[7] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[8] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416–419, Aug. 2016.

[9] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.

[10] D. Kim, H. Lee, and D. Hong, "A survey of in-band full-duplex transmission: From the perspective of PHY and MAC layers," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2017–2046, 4th Quart., 2015.

[11] M. Jain *et al.*, "Practical, real-time, full duplex wireless," in *Proc. ACM MobiCom*, Las Vegas, NV, USA, Sep. 2011, pp. 301–312.

[12] W. Li, J. Lilleberg, and K. Rikkinen, "On rate region analysis of half- and full-duplex OFDM communication links," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1688–1698, Sep. 2014.

[13] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[14] T. Riihonen, S. Werner, and R. Wichman, "Hybrid full-duplex/half-duplex relaying with transmit power adaptation," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 3074–3085, Sep. 2011.

[15] A. Sharma, R. K. Ganti, and J. K. Milleth, "Joint backhaul-access analysis of full duplex self-backhauling heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1727–1740, Mar. 2017.

[16] O. Simeone, Y. Bar-Ness, and U. Spagnolini, "Stable throughput of cognitive radios with and without relaying capability," *IEEE Trans. Commun.*, vol. 55, no. 12, pp. 2351–2360, Dec. 2007.

[17] C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "Cognitive cooperative random access for multicast: Stability and throughput analysis," *IEEE Trans. Control Netw. Syst.*, vol. 1, no. 2, pp. 135–144, Jun. 2014.

[18] S. Kompella, G. Nguyen, C. Kam, J. E. Wieselthier, and A. Ephremides, "Cooperation in cognitive underlay networks: Stable throughput tradeoffs," *IEEE/ACM Trans. Netw.*, vol. 22, no. 6, pp. 1756–1768, Dec. 2014.

[19] I. Krikidis, J. N. Laneman, J. Thompson, and S. McLaughlin, "Protocol design and throughput analysis for multi-user cognitive cooperative systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4740–4751, Sep. 2009.

[20] P. Xu, Y. Yuan, Z. Ding, X. Dai, and R. Schober, "On the outage performance of non-orthogonal multiple access with 1-bit feedback," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6716–6730, Oct. 2016.

[21] M. Haenggi and R. K. Ganti, *Interference in Large Wireless Networks*. Boston, MA, USA: Now, 2009.

[22] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[23] P. H. J. Nardelli, M. Kountouris, P. Cardieri, and M. Latva-Aho, "Throughput optimization in wireless networks under stability and packet loss constraints," *IEEE Trans. Mobile Comput.*, vol. 13, no. 8, pp. 1883–1895, Aug. 2014.

[24] Y. Zhou and W. Zhuang, "Performance analysis of cooperative communication in decentralized wireless networks with unsaturated traffic," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3518–3530, May 2016.

[25] K. Stamatiou and M. Haenggi, "Random-access Poisson networks: Stability and delay," *IEEE Commun. Lett.*, vol. 14, no. 11, pp. 1035–1037, Nov. 2010.

[26] S. Weber and J. G. Andrews, "Transmission capacity of wireless networks," *Found. Trends Netw.*, vol. 5, nos. 2–3, pp. 109–281, 2012, doi: http://dx.doi.org/10.1561/1300000032.

[27] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Cambridge Philos. Soc.*, vol. 58, no. 3, pp. 497–520, 1962.

[28] *Study on Downlink Multiuser Superposition Transmission (MUST) for LTE*, document 3GPP TR 36.859, Jan. 2016.

[29] M. Haenggi, "The local delay in Poisson networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1788–1802, Mar. 2013.

[30] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry Wireless Networks, Part I Theory*. Boston, MA, USA: Now, 2009.

[31] W. Luo and A. Ephremides, "Stability of N interacting queues in random-access systems," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1579–1587, Jul. 1999.

[32] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. New York, NY, USA: Academic, 2014.

[33] D. Bharadia and S. Katti, "FastForward: Fast and constructive full duplex relays," in *Proc. ACM SIGCOMM*, Chicago, IL, USA, Aug. 2014, pp. 1–12.

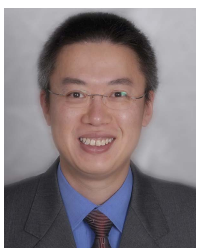
[34] K.-C. Hsu, K. C.-J. Lin, and H.-Y. Wei, “Full-duplex delay-and-forward relaying,” in *Proc. ACM MobiHoc*, Paderborn, Germany, Jul. 2016, pp. 221–230.

[35] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, “On the performance of non-orthogonal multiple access systems with partial channel information,” *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 654–667, Feb. 2016.

[36] M. K. Tsatsanis, R. Zhang, and S. Banerjee, “Network-assisted diversity for random access wireless networks,” *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 702–711, Mar. 2000.

[37] S. Srinivasa and M. Haenggi, “Distance distributions in finite uniformly random networks: Theory and applications,” *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 940–949, Feb. 2010.

Yong Zhou (S’13–M’16) received the B.Sc. and M.Eng. degrees from Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2015. From 2015 to 2017, he was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada. He is currently an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University, China. His research interests include performance analysis and resource allocation of 5G networks. He served as a technical program committee member for several conferences.



Vincent W.S. Wong (S’94–M’00–SM’07–F’16) received the B.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, in 1994, the M.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996, and the Ph.D. degree from The University of British Columbia (UBC), Vancouver, BC, Canada, in 2000. From 2000 to 2001, he was a Systems Engineer with Microsemi. He joined the Department of Electrical and Computer Engineering, UBC, in 2002, where he is currently a Professor. His research areas include protocol design, optimization, and resource management of communication networks, with applications to wireless networks, smart grid, mobile cloud computing, and Internet of Things. He received the 2014 UBC Killam Faculty Research Fellowship. He is an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS. He served as a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and the IEEE WIRELESS COMMUNICATIONS. He also served on the editorial boards for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and the *Journal of Communications and Networks*. He was a Technical Program Co-Chair of the IEEE SmartGridComm 2014 and a Symposium Co-Chair of the IEEE ICC 2018, the IEEE SmartGridComm 2013, the IEEE SmartGridComm 2017, and the IEEE Globecom 2013. He is the Chair of the IEEE Vancouver Joint Communications Chapter. He served as a Chair for the IEEE Communications Society Emerging Technical Sub-Committee on Smart Grid Communications.

Robert Schober (S’98–M’01–SM’08–F’10) received the Diploma and Ph.D. degrees in electrical engineering from the Friedrich-Alexander University of Erlangen–Nuremberg (FAU), Germany, in 1997 and 2000, respectively. From 2002 to 2011, he was a Professor and the Canada Research Chair with The University of British Columbia (UBC), Vancouver, Canada. Since 2012, he has been an Alexander von Humboldt Professor and the Chair for Digital Communication with FAU. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing.

Dr. Schober is a fellow of The Canadian Academy of Engineering and The Engineering Institute of Canada. He received several awards for his research, including the 2002 Heinz Maier-Leibnitz Award of the German Science Foundation (DFG), the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, the 2011 Alexander von Humboldt Professorship, the 2012 NSERC E. W. R. Steacie Fellowship, and the 2017 Wireless Communications Recognition Award by the IEEE Wireless Communications Technical Committee. He is listed as a 2017 Highly Cited Researcher by the Web of Science and a Distinguished Lecturer by the IEEE Communications Society (ComSoc). From 2012 to 2015, he served as the Editor-in-Chief for the IEEE TRANSACTIONS ON COMMUNICATIONS. He is currently the Chair of the Steering Committee of the IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTI-SCALE COMMUNICATION, an Editorial Board Member of the PROCEEDINGS OF THE IEEE, a Member at Large of the Board of Governors of ComSoc, and the ComSoc Director of Journals.