Non-Orthogonal Multiplexing of Ultra-Reliable and Broadband Services in Fog-Radio Architectures

RAHIF KASSAB[1], (Student Member, IEEE), **OSVALDO SIMEONE**[1], (Fellow, IEEE),
PETAR POPOVSKI[2], (Fellow, IEEE), AND **TOUFIQUL ISLAM**[3], (Member, IEEE)

[1]Centre for Telecommunications Research, King's College London, London WC2B 4BG, U.K.
[2]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
[3]Intel Corporation, Santa Clara, CA 95054, USA

Corresponding author: Rahif Kassab (rahif.kassab@kcl.ac.uk)

ABSTRACT The fifth generation (5G) of cellular systems is introducing ultra-reliable low-latency communication (URLLC) services alongside more conventional enhanced Mobile BroadBand (eMBB) traffic. Furthermore, the 5G cellular architecture is evolving from a base station-centric deployment to a fog-like setup that accommodates a flexible functional split between cloud and edge. In this paper, a novel solution is proposed that enables the non-orthogonal coexistence of URLLC and eMBB services by processing URLLC traffic at the edge nodes, while eMBB communications are handled centrally at a cloud processor as in a cloud-radio access network system. This solution guarantees the low-latency requirements of the URLLC service by means of edge processing, e.g., for vehicle-to-cellular use cases, as well as the high-spectral efficiency for eMBB traffic via centralized baseband processing. Both uplink and downlink are analyzed by accounting for the heterogeneous performance requirements of eMBB and URLLC traffic and by considering practical aspects, such as fading, the lack of channel state information for URLLC transmitters, rate adaptation for eMBB transmitters, finite fronthaul capacity, and different coexistence strategies, such as puncturing.

INDEX TERMS eMBB, URLLC, NOMA, 5G, C-RAN, F-RAN, fog networking.

I. INTRODUCTION

With the advent of 5G, wireless cellular systems are undergoing an evolution in terms of both services and network architecture. Conventional enhanced Mobile BroadBand (eMBB) services, mostly aimed at consumers, will share radio and network resources with Ultra-Reliable Low-Latency Communications (URLLC) and machine-type traffic, which cater to vertical industries [1]. Furthermore, the supporting cellular network architecture will evolve from a traditional base station-centric deployment to a fog-like set-up [2], [3] with computation and communication resources at both edge and cloud. Thanks to network softwarization, this architecture will enable network functionalities to be distributed among edge and cloud elements[1] depending on their latency and reliability requirements [2]–[4]. An extreme instance of this

type of network architectures is Cloud-Radio Access Network (C-RAN), in which all processing, apart from radio-frequency components, is carried out in the cloud [5].

The coexistence of the heterogeneous services eMBB and URLLC is typically envisioned to rely on the orthogonal allocation of spectral resources to the two traffic types. Orthogonal multiplexing ensures the isolation of eMBB and URLLC traffic. This, in turn, enables the application of "slicing", a Radio Resource Management approach introduced in 5G that carries out separate designs to meet the heterogeneous performance metrics and guarantees of the two services. As an alternative to orthogonal resource allocation, Non-Orthogonal Multiple Access (NOMA) is grounded by classical results in information theory that prove the capacity-achieving property of non-orthogonal transmissions in multiple access channels and of superposition coding in broadcast channels, modeling single-cell Uplink (UL) and Downlink (DL) scenarios, respectively [6]. Motivated by

---

[1]These are defined as distributed and central units by 3GPP [2], respectively.

these information-theoretic results, NOMA has been proposed as a key component of 5G systems as a mean to share radio resources among transmissions belonging to *homogeneous* devices [7]–[9]. A first question that motivates this work is: *Can NOMA improve the performance of the coexistence between the heterogenuous eMBB and URLLC services?*

An illustration of a fog-based multi-cell system, also known as Fog-RAN (F-RAN) [4], [10], [11] is provided in Fig. 1. In this system, each cell contains an Edge Node (EN), along with its connected computing platform, as well as multiple eMBB and URLLC users. All ENs have a finite-capacity digital fronthaul link to a cloud processor. The mentioned flexibility in the allocation of functions between ENs and cloud afforded by network softwarization motivates the second question that guides this work: *How can the physical layer network functionalities be split between edge and cloud in order to improve the coexistence of eMBB and URLLC traffic?*

A. MAIN CONTRIBUTIONS

In this work, we address the first question by considering the performance of both eMBB and URLLC traffic in the F-RAN multi-cell system of Fig. 1 under both orthogonal and non-orthogonal multiple access. Following [12], we refer to this latter approach as *Heterogeneous NOMA* (H-NOMA), in order to highlight the key distinction with respect to conventional NOMA of accommodating services with heterogeneous, rather than homogeneous, performance requirements.
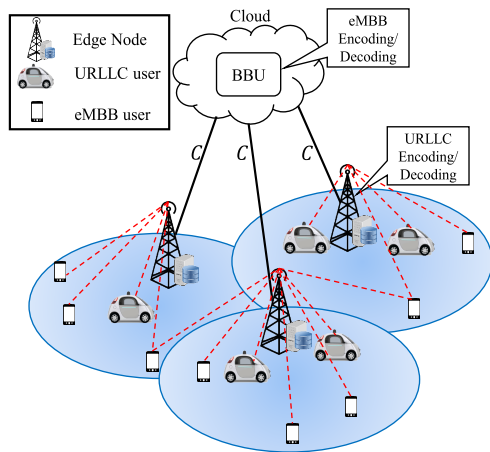

FIGURE 1. An F-RAN multi-cell system with coexisting eMBB and URLLC.

Moreover, we tackle the second question by proposing a novel cloud-edge functional split for the coexistence of URLLC and eMBB traffic. Accordingly, URLLC traffic is handled at the ENs in order to meet the low-latency requirements, URLLC devices may be vehicles in vehicle-to-cellular use cases [13], or they may be devices serving automation chains in Industry 4.0 scenarios characterized by automation and communication-based manufacturing [14], [15]. In contrast, eMBB traffic is processed by the centralized BaseBand

Unit (BBU) in the cloud as in a C-RAN architecture in order to enhance spectral efficiency thanks to the cloud's interference management capabilities [16]. The proposed hybrid edge-cloud solution fully leverages the unique features of F-RAN systems in order to cater to the heterogeneous requirements of URLLC and eMBB systems.
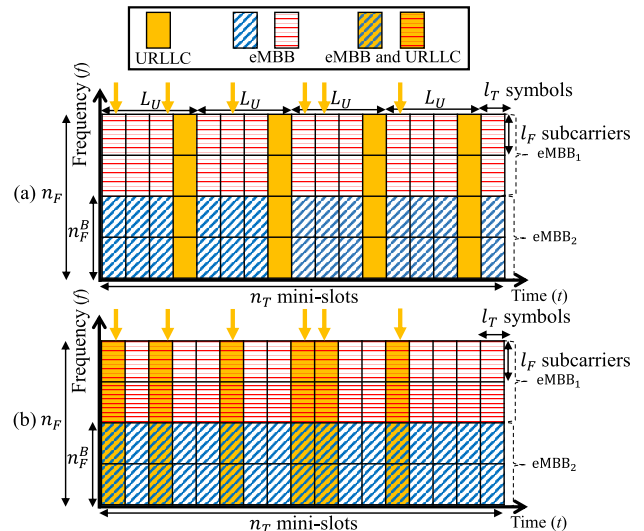

FIGURE 2. Time-frequency resource allocation for:
(a) Heterogeneous-Orthogonal Multiple Access Scheme (H-OMA); and
(b) Heterogeneous-Non Orthogonal Access Scheme (H-NOMA); Hatched areas correspond to eMBB transmissions. Downward arrows denote generation of URLLC packets to or from different URLLC users.

We analyze and compare the performance of conventional Heterogeneous-Orthogonal Multiple Access (H-OMA) techniques with H-NOMA in the *multi-cell F-RAN* system of Fig. 1 for both Uplink (UL) and Downlink (DL). As illustrated in Fig. 2, the communication model assumes random activation of URLLC users, which implies possible collisions in the UL and blockages in the DL, and scheduled access for eMBB users. We aim at understanding the possible tradeoffs between the eMBB and URLLC performance as a function of key parameters, such as the capacity of the digital fronthaul links connecting cloud and ENs, as well as the deployed strategy including puncturing, treating URLLC interference as noise, and superposition coding (see Fig. 5 for the UL and Fig. 6 for the DL). For a preview of the main results, we refer to Fig. 9, which shows the per-cell eMBB and URLLC achievable rates as function of the URLLC traffic load for both H-OMA and H-NOMA.

B. RELATED WORK

In the past few years, NOMA has been widely investigated as a solution to increase the spectral efficiency of cellular networks for both UL and DL. The key idea is to allow simultaneous transmissions on the UL and superposition coding on the DL. As some representative examples, in the UL case, [17] shows that NOMA with Successive Interference Cancellation (SIC) at the base stations can significantly enhance cell-edge users' throughput, while paper [18]

proposes a NOMA scheme based on joint processing at the base stations. As for the DL, NOMA was demonstrated to achieve superior performance in terms of ergodic sum rate of a cellular network with randomly deployed users in [19]. Furthermore, in [20], OMA techniques were compared with NOMA under SIC in terms of system-level performance by taking into account key LTE functionalities such as Hybrid Automatic Repeat reQuest (HARQ) and time/frequency domain scheduling. Other related works include [21]–[23].

While all the work discussed above assumes homogeneous services, coexistence between heterogeneous services such as eMBB and URLLC has been mostly studied under the assumption of orthogonal resource allocation. For example, in [24] a null-space-based spatial preemptive scheduler for URLLC and eMBB traffic is proposed that aims at guaranteeing URLLC quality of service while maximizing the eMBB ergodic capacity. In [25], the joint user-base station association and orthogonal resource allocation problem was considered for the DL of a fog-network in the presence of eMBB and URLLC services.

The non-orthogonal coexistence of heterogeneous services has been much less studied. In [26], the joint scheduling of eMBB and URLLC has been investigated with the goal of maximizing the utility of eMBB traffic while satisfying the quality of service requirements of URLLC traffic. The problem formulation accounts for the different time scales of traffic generation of the two services. From an information-theoretic standpoint, the coexistence of heterogeneous services was studied under the rubric of unequal error protection in simplified abstract settings in [27]. A communication-theoretic model for the coexistence of eMBB-URLLC and eMBB-mMTC was introduced in [28] for a *single-cell* model with decoding at the base station. To the best of our knowledge, the multi-cell case was only studied by some of the authors in [29] by considering *only the UL* and a simplified Wyner-type channel model with *no fading and inter-cell interference limited to neighboring cells*. We also refer to [30] that considers a setup with the same limitations as [29] but with analog fronthaul links. Another related theoretical work for the UL Wyner model is [31], in which higher-latency messages are decoded by means of cooperation between adjacent cells, while lower-latency messages are decoded without cooperation.

C. ORGANIZATION OF THE PAPER

The rest of the paper is organized as follows. In the next section, we detail the system model, while Sec. III describes the signal model and the performance metrics. In Sec. IV and V, the performance of H-OMA is evaluated for UL and DL respectively, while in Sec. VI and VII, the performance of H-NOMA is analyzed for UL and DL respectively. Finally, numerical results are presented in Sec. VIII, and conclusions are drawn in Sec. IX.

Notation: Bold upper-case characters denote matrices and bold lower-case characters denote vectors. $\mathbb{E}_X[\cdot]$ represents

the expectation of the argument with respect to the distribution of the random variable X. $\mathbf{A}^H$ denotes the Hermitian transpose of matrix $\mathbf{A}$. $X \sim \mathcal{B}ern(p)$ indicates a Bernoulli distribution with parameter p. $Y \sim \mathcal{B}in(n, p)$ indicate a Binomial random variable distribution with parameters n and p. $I_z(X; Y)$ denotes the mutual information between random variables X and Y for the given constant value z of random variable Z, i.e., $I_z(X; Y) = I(X; Y|Z = z)$. $|\mathbf{A}|$ is the determinant of matrix $\mathbf{A}$.

II. SYSTEM MODEL

As illustrated in Fig. 1, we study UL and DL communications in a cellular network with an F-RAN architecture that encompasses both eMBB and URLLC users. Each one of the M cells contains an Edge Node (EN) and multiple eMBB and URLLC users. All ENs are connected to a Baseband Unit (BBU) in the cloud by mean of orthogonal fronthaul links of capacity C bit/s/Hz, or equivalently C bits for each symbol of the wireless channel. The RAN uses Frequency Division Duplex (FDD) in order to facilitate grant-free URLLC transmissions, as detailed below.

F-RAN Topology and Operation: As illustrated in Fig. 1, we assume that the URLLC users are located close to the ENs, and hence URLLC communications take place with non-negligible power only with the EN in the same cell. As a result, URLLC users do not cause interference to ENs in other cells while transmitting in the UL, and they only receive transmissions from the same-cell EN in the DL. This condition can be ensured by allowing URLLC transmissions only from users with large average channel gain to the target EN, so that the high reliability requirement of URLLC traffic can be satisfied. As an example, as seen in Fig. 1, the EN may serve a nearby vehicle for transmission of time-sensitive control information in vehicle-to-cellular use cases [13]. Alternatively, in mission-critical or Industry 4.0 scenarios, ENs can be deployed in locations that contain URLLC devices. The eMBB users, instead, need not guarantee this condition, and are assumed to be in arbitrary positions with potentially non-negligible channel gains to all ENs for both UL and DL.

As illustrated in Fig. 1, for the UL, due to latency constraints, the URLLC signals are decoded locally at the EN, while the eMBB traffic is decoded centrally at the BBU as in a C-RAN architecture [32], [33]. In a similar manner, in the DL, the eMBB traffic is assumed to be generated at the cloud, e.g., as a result of web searches or broadband streaming, and C-RAN precoding and quantization are applied [34], [35]. In contrast, URLLC traffic is generated at the edge, with each EN serving same-cell URLLC users, in line with the use cases mentioned above. Note that these assumptions imply that the higher layers of eMBB and URLLC services are implemented separately at cloud and edge, respectively.

Frame structure: As illustrated in Fig. 2, we consider a radio interface that operates in frames of n_T mini-slots and n_F frequency channels. The time frequency plane is organized into Resource Units (RUs), and each RU spans one mini-slot of index $t \in \{1, \ldots, n_T\}$ and one frequency channel of

index $f \in \{1, \ldots, n_F\}$, and it contains l_T symbols in time domain and l_F subcarriers.[2] We index as (f, t) the RU located at frequency channel f and mini-slot t.

eMBB traffic is scheduled and a single eMBB user is assigned to all n_T minislots in a frame and a set $n_F^B \le n_F$ frequency channels, so that there are $\lfloor n_F/n_F^B \rfloor$ scheduled eMBB users per-cell. In contrast, URLLC transmissions are assumed to be grant-free, and packets are randomly generated in each mini-slot for URLLC users. As a result, the number of active URLLC users per-cell in each frame is random. Furthermore, due to latency constraints, each URLLC transmission can span only a single mini-slot, and hence the blocklength of an URLLC transmission is equal to $n_F l_F l_T$ symbols.
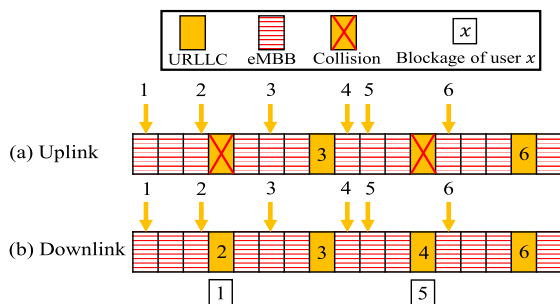

FIGURE 3. Illustration of (a) collisions in the UL and (b) blockages in the DL for URLLC transmissions under H-OMA when $L_U = 4$. Downward arrows denote the generation of a URLLC packet.

As illustrated in Fig. 2 and Fig. 3, we assume that at most one URLLC packet per cell and per mini-slot may be generated at an URLLC user in the UL or at the EN in the DL. Each packet is generated at, or intended for, a different URLLC user. The probability of generation of such packet is a_U and packet generation is independent across mini-slots and cells. We note that, in practice, this condition requires an access control protocol for the spectral resources under study that ensures that no more than one URLLC packet is generated within a cell in a mini-slot.

Channel and Channel State Information (CSI) Model: For both UL and DL, we consider Rayleigh fading channels that are constant over time in the considered frame, but vary independently from one frequency channel to another. Accordingly, the complex channel gain $h_{i,j}^f$ between the i-th EN and an eMBB user in the j-th cell at RU (f, t) is modeled as $h_{i,j}^f \sim \mathcal{CN}(0, \alpha_{i,j}^2)$, where $\alpha_{i,j}^2$ accounts for path loss and large scale fading. The channel gains $h_{i,j}^f$ are i.i.d. over the frequency index f; independent for different pairs (i, j); and constant during the n_T mini-slots of the considered frame. In a similar manner, the channel gains between an URLLC user in cell i and the i-th EN are modeled as $g_i^f(t) \sim \mathcal{CN}(0, \beta_i^2)$, where β_i^2 reflects the path loss and large-scale fading. Note that the dependence on mini-slot index t is kept for URLLC transmissions in order to highlight that, under the given assumptions,

[2]As an example, in 3GPP release 15 [36], the RU consists of l_F =12 subcarriers and l_T =14 symbols.

different URLLC users transmit, or are served in each mini-slot t.

Following a standard path-loss model, we write $\alpha_{i,j}^2 = c_B(d_{B,R}/d_{i,j})^\gamma$ and $\beta_i^2 = c_U(d_{U,R}/d_i)^\gamma$, where $d_{i,j}$ is the distance between i-th EN and the j-th eMBB user and d_i is the distance between the i-th URLLC user and the EN in the i-th cell; γ is the path loss exponent and constants c_B and c_U are used to set the signal-to-noise ratio (SNR) levels at the reference distances $d_{B,R}$ and $d_{U,R}$.

Due to latency constraints, CSI is assumed to be unavailable at the transmitter side in the communications between a URLLC user and an EN, while receiver CSI is available. This assumption reflects the fact that, in an FDD system, transmitter CSI would require feedback from the receiver. This would limit reliability since it would add another potential cause of error, and it would increase latency. In contrast, CSI is conventionally assumed to be available at both the transmitter and the receiver for the eMBB traffic.

A. HETEROGENEOUS ORTHOGONAL AND NON-ORTHOGONAL MULTIPLE ACCESS

In this work, we consider two access schemes, namely H-OMA and H-NOMA. As discussed in Sec. I, these schemes allow the sharing of time and frequency resources between eMBB and URLLC services. As seen in Fig. 2(a), under H-OMA, URLLC packets can only occupy preallocated mini-slots over which eMBB transmissions are not allowed. In particular, a mini-slot is allocated for transmission of URLLC traffic every L_U mini-slots. Parameter L_U hence represents the access latency, i.e., the maximum number of mini-slots a URLLC packet has to wait before transmission. We note that, for the DL, it would also be possible to consider a dynamic schedule of eMBB and URLLC transmissions (see [26]).

In the UL, as illustrated in Fig. 3(a), if multiple URLLC users in a cell generate a packet within the L_U mini-slots between two allocated mini-slots, then a *collision* occurs in the allocated mini-slot. In this case, all packets are discarded due to latency constraints. In the DL, instead, as illustrated in Fig. 3(b), when multiple URLLC packets are generated at an EN, the EN can select one such packet uniformly at random and discard, hence *blocking* from access, all the other packets. Collisions in the UL and blockages in the DL contribute to the overall error rate for URLLC.

In contrast, H-NOMA enables eMBB and URLLC traffic to share the same radio resources. More precisely, as shown in Fig. 2(b), URLLC packets are transmitted in the same mini-slot in which they get generated. It follows that H-NOMA has the minimal access latency of $L_U = 1$ at the price of possible interference between eMBB and URLLC signals. Furthermore, under the assumed model, no collisions or blockages occur with H-NOMA.

III. SIGNAL MODEL AND PERFORMANCE METRICS

In this section, we detail the signal models for UL and DL, as well as the performance metrics of interest. Throughout the

analysis, we focus our attention on the analysis of eMBB and URLLC traffic that occupies the frame shown in Fig. 2. We first concentrate on detailing the signal transmitted and received in any of the symbols of an RU (f, t) and then describe the performance metrics of interest. Throughout, we avoid introducing an explicit notation for the indices pointing to one of the $l_F l_T$ symbols in each RU, and hence refer generically with the index (f, t) to any symbol within RU (f, t).

A. UPLINK SIGNAL MODEL

As discussed in the previous section, with H-OMA, one mini-slot is exclusively allocated to URLLC users in the UL every L_U mini-slots (see Fig. 2(a)). We denote as $y_k^f(t)$ the signal received by each EN_k at any symbol within the RU (f, t). This can be written as

$$y_k^f(t) = h_{k,k}^f x_k^f(t) + \sum_{j \neq k} h_{k,j}^f x_j^f(t) + z_k^f(t), \quad (1)$$

for all t different from $L_U, 2L_U, \ldots$ and where $x_k^f(t)$ denotes any symbol transmitted in RU (f, t) by the eMBB user that is active in cell k over the given frequency channel f; $z_k^f(t) \sim \mathcal{CN}(0, 1)$ is complex Gaussian noise with zero mean and unit variance, which is i.i.d. across the indices k, f, and t, and across the symbols in an RU. Furthermore, for all mini-slots allocated to URLLC users, the received signal for mini-slot $t = L_U, 2L_U, \ldots$ when there is no collision is

$$y_k^f(t) = g_k^f(t) u_k^f(t) + z_k^f(t), \quad (2)$$

where $u_k^f(t)$ denotes the signal sent by an URLLC user in the k-th cell. We recall that, in contrast to the eMBB channel, the URLLC channel coefficients $g_k^f(t)$ depend on the time index t due to the assumption that URLLC packets are generated at different URLLC users.

Unlike H-OMA, under H-NOMA, URLLC users transmit immediately in the mini-slot in which a packet is generated. Accordingly, the signal $y_k^f(t)$ received at each EN_k at RU (f, t) can be written as

$$y_k^f(t) = h_{k,k}^f x_k^f(t) + \sum_{j \neq k} h_{k,j}^f x_j^f(t) + A_k(t) g_k^f(t) u_k^f(t) + z_k^f(t), \quad (3)$$

where $A_k(t) \sim \mathcal{B}ern(a_U)$ is the indicator variable that equals to one if an URLLC packet is generated in cell k at mini-slot $t = 1, 2, \ldots$.

The power constraint for eMBB and URLLC users are defined respectively as

$$\mathbb{E}[|x_k^f(t)|^2] \leq P_B \text{ and } \mathbb{E}[|u_k^f(t)|^2] \leq P_U, \quad (4)$$

where the average in (4) is taken over all uniformly selected information messages.

It will be convenient to write the signal models in matrix form. To this end, the $M \times M$ channel matrix for eMBB users at RU (f, t) is denoted by $\mathbf{H}^f$ with k-th row given by the $1 \times M$ vector $\mathbf{h}_{(k)}^f = [h_{k,1}^f, \ldots, h_{k,M}^f]$. The $M \times M$

channel matrix for URLLC users is diagonal due to the discussed lack of inter-cell interference and is denoted as $\mathbf{G}^f(t) = \text{diag}(g_1^f(t), \ldots, g_M^f(t))$. Consequently, we can write the signals (3) received at RU (f, t) across all ENs under H-NOMA as

$$\mathbf{y}^f(t) = \mathbf{H}^f \mathbf{x}^f(t) + \mathbf{A}(t) \mathbf{G}^f(t) \mathbf{u}^f(t) + \mathbf{z}^f(t), \quad (5)$$

where $\mathbf{A}(t) = \text{diag}(A_1(t), \ldots, A_M(t))$, $\mathbf{x}^f(t) = [x_1^f(t), \ldots, x_M^f(t)]^\mathsf{T}$, $\mathbf{u}^f(t) = [u_1^f(t), \ldots, u_M^f(t)]^\mathsf{T}$ and $\mathbf{z}^f(t) = [z_1^f(t), \ldots, z_M^f(t)]^\mathsf{T}$. Models (1)-(2) can be written in matrix form in an analogous way.

Following the general assumptions introduced in Sec. II, the BBU and the ENs are assumed to have available the channel matrices $\mathbf{H}^f$ and $\mathbf{G}^f(t)$ for both eMBB and URLLC users. Note that providing CSI to the BBU causes a fronthaul overhead that can be considered negligible as the coherence interval size $l_F \times n_T$ increases (see [37]). eMBB users are informed by the BBU about the transmission rate at which to operate, while URLLC users have no CSI. As a result, URLLC transmitters adapt their rates only to the distribution of the channel, while the eMBB transmitters adjust their rates to the current channel realization.

B. DOWNLINK SIGNAL MODEL

In the DL, for both H-OMA and H-NOMA, the signal $y_k^f(t)$ received at an eMBB user in cell k at RU (f, t) can be written as

$$y_k^f(t) = h_{k,k}^f x_k^f(t) + \sum_{j \neq k} h_{k,j}^f x_j^f(t) + z_k^f(t), \quad (6)$$

where $x_k^f(t)$ denotes the symbol transmitted by the k-th EN; and $z_k^f(t)$ is Gaussian noise received at the eMBB users, with $z_k^f(t) \sim \mathcal{CN}(0, 1)$, which is i.i.d. across the indices k, f and t and across all symbols in an RU. As we will detail in Sec. V, under H-OMA, the signal $x_k^f(t)$ is either intended for an URLLC user or an eMBB, while for H-NOMA the signal $x_k^f(t)$ may carry the superposition of URLLC and eMBB signals. We also write (6) in vector form as

$$y_k^f(t) = \mathbf{h}_{(k)}^f \mathbf{x}^f(t) + z_k^f(t), \quad (7)$$

where $\mathbf{x}^f(t) = [x_1^f(t), \ldots, x_M^f(t)]^\mathsf{T}$ and $\mathbf{h}_{(k)}^f = [h_{k,1}^f, \ldots, h_{k,M}^f]$.

In contrast, the signal received by the k-th URLLC user at RU (f, t) is given as

$$u_k^f(t) = g_k^f(t) x_k^f(t) + z_k^f(t), \quad (8)$$

where $z_k^f(t) \sim \mathcal{CN}(0, 1)$ represents Gaussian noise. We recall that, for the same reason as in the UL, the URLLC users' CSI depend on the mini-slot index t. In all cases, the power constraint P for each EN in the DL is defined as

$$\mathbb{E}[|x_k^f(t)|^2] \leq P. \quad (9)$$

Finally, following the general channels assumptions described in Sec. II, all eMBB and URLLC users are
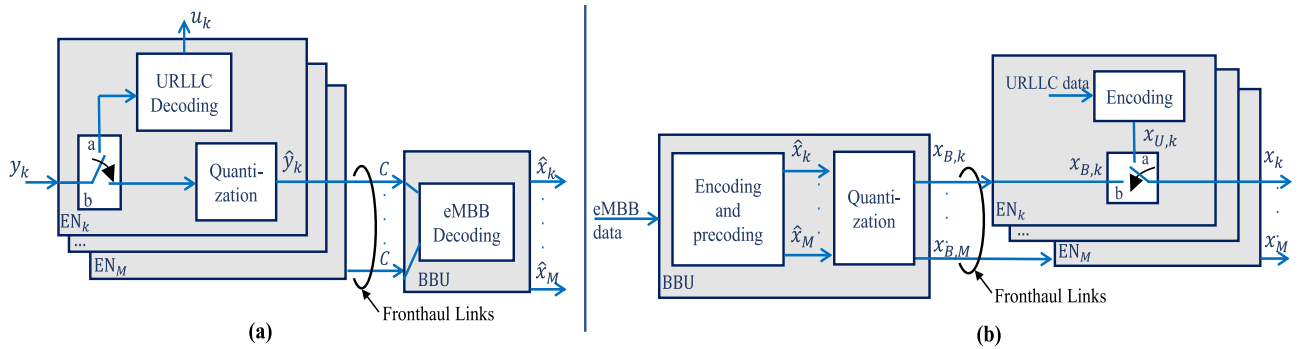
FIGURE 4. Block diagrams for the ENs and the BBU under H-OMA for (a) UL and (b) DL. Switches in both cases move to position *a* every $L_U - 1$ mini-slots.

assumed to have available the local channels and signal-to-interference-plus-noise ratio (SINR). The BBU is informed about the eMBB channel matrices $\mathbf{H}^f$. Finally, as in the UL scenario the URLLC rate is adjusted to the statistics of the channels, while the eMBB rate is adjusted to the current channel realization.

In the remainder of the paper, we will drop the dependence on t when no confusion may arise.

C. PERFORMANCE METRICS

We are interested in the following performance metrics. For eMBB users, we study the average per-cell sum-rate R_B, where the average is taken over the fading distribution. The transmission rates are adapted to the fading realizations thanks to the availability of transmitter CSI. The per-cell sum-rate measures the sum-rate, or, sum-spectral efficiency, in bit/s/Hz across all eMBB users in the system normalized by the number M of cells.

As for URLLC users, we define the access latency L_U as the maximum number of mini-slots an URLLC user has to wait before receiving a generated packet. By construction, for H-NOMA, we have $L_U = 1$ which corresponds to the minimum access latency when a packet is transmitted in the same mini-slot in which it is generated. Furthermore, following 3GPP [38, Sec. 7.9], we define URLLC reliability as the probability to successfully transmit a packet within a given time constraint, here L_U. Accordingly, we explicitly define a constraint on the URLLC error probability $\Pr[\mathcal{E}_U]$ as

$$\Pr[\mathcal{E}_U] \leq \epsilon_U \qquad (10)$$

for some desired error level ϵ_U. This probability can be interpreted as the average fraction of URLLC devices whose quality-of-service requirements are met.

The error event $\mathcal{E}_U$ accounts for two possible types of error, namely collision or blockage and decoding failure. As illustrated in Fig. 3, a collision or a blockage, which only applies to H-OMA, happens in UL or DL respectively, when two or more packets are generated in the L_U mini-slots between two transmission opportunities. In contrast, decoding failure occurs when an URLLC packet is transmitted, and hence it is not subject to collision or blockage,

but decoding fails at the receiver. For a given outage probability, due to the absence of CSI at the transmitter, open-loop transmission with no rate adaptation is assumed, and we adopt the maximum transmission rate under an outage probability constraint, or outage capacity, as the rate metric of interest [39].

IV. UPLINK ORTHOGONAL MULTIPLE ACCESS

In this section, we consider the UL system performance in terms of eMBB rate R_B and URLLC rate R_U for a fixed URLLC access latency L_U and URLLC probability of error requirement ϵ_U when assuming the conventional H-OMA. The operation of the ENs and of the BBU are illustrated in Fig. 4(a).

A. URLLC PERFORMANCE

As discussed in Sec. II, due to latency constraints, URLLC packets are decoded at the local EN upon reception in the transmission mini-slot $t = L_U, 2L_U, \ldots$ (see Fig. 1). For a given decoding error probability ϵ_U^D, the outage capacity of the k-th URLLC user is given as

$$R_{U,k}(\epsilon_U^D) = \sup\left\{R_U : P_{out,k}(R_U) \leq \epsilon_U^D\right\}, \qquad (11)$$

where $P_{out,k}$ denotes the outage probability

$$P_{out,k}(R_U) = \Pr\left[\frac{1}{n_F}\sum_{f=1}^{n_F}\log(1 + S_{U,k}^f) \leq R_U\right], \qquad (12)$$

and $S_{U,k}^f = |g_k^f|^2 P_U$ is the signal-to-noise ratio (SNR) at the EN. The per-cell sum-outage capacity is obtained as $R_U = 1/M \sum_{k=1}^{M} R_{U,k}(\epsilon_U^D)$.

Following Sec. III.C, the probability of error of an URLLC packet can be written as

$$\Pr[\mathcal{E}_U] = \sum_{n=1}^{L_U-1} p(n) + \epsilon_U^D p(0), \qquad (13)$$

where $p(n) = \Pr[N_U(L_U) = n]$ is the distribution of random variable $N_U(L_U) \sim \text{Bin}(L_U - 1, a_U)$. The binomial random variable $N_U(L_U)$ represents the number of additional URLLC packets generated by the URLLC users during the remaining

$L_U - 1$ mini-slots between two transmission opportunities. The first term in (13) is the probability that a packet is lost due to collisions, which occurs if $n \geq 1$ additional packets are generated. The second term in (13) is the probability of a decoding error at the receiver.

B. eMBB RATE

Unlike delay-constrained URLLC traffic, eMBB messages are decoded jointly at the cloud in order to leverage the centralized interference management capabilities of the BBU. To this end, following the standard C-RAN operation, each EN quantizes and compresses the received signal y_k^f for the mini-slots $t \notin \{L_U, 2L_U, \ldots\}$ by using point-to-point compression (see [32], [33], [40]), and forwards the resulting signal to the cloud over the fronthaul links (see Fig. 4(a)). Using (1), for each frequency channel f, the quantized signal received at the BBU from EN_k can be written as

$$\hat{y}_k^f = y_k^f + q_k^f, \qquad (14)$$

where $q_k^f \sim \mathcal{CN}(0, \sigma_{q,k}^2)$ represents the quantization noise with power $\sigma_{q,k}^2$. From classical results in rate-distortion theory, we have the following relationship between the quantization noise power $\sigma_{q,k}^2$ and the fronthaul capacity [40], [41]:

$$C = (1 - L_U^{-1}) \frac{1}{n_F} \sum_{f=1}^{n_F} I_{\mathbf{H}^f}(y_k^f; \hat{y}_k^f) \qquad (15a)$$

$$= (1 - L_U^{-1}) \frac{1}{n_F} \sum_{f=1}^{n_F} \log\left(1 + \frac{1 + \sum_{j=1}^{M} |h_{k,j}^f|^2 P_B}{\sigma_{q,k}^2}\right). \qquad (15b)$$

In (15), the factor $(1 - L_U^{-1})$ captures the fact that a fraction $(1 - L_U^{-1})$ of all mini-slots of the wireless channel are occupied by eMBB transmissions. The value of $\sigma_{q,k}^2$ can be obtained by solving (15b) via numerical methods.

Considering the signals $\hat{\mathbf{y}}^f = [\hat{y}_1^f, \ldots, \hat{y}_M^f]$ received by the BBU from all M ENs, the eMBB per-cell sum-rate for a given channel realizations $\mathbf{H}^f$ can be finally written as

$$R_B = \frac{(1 - L_U^{-1})}{M} \frac{1}{n_F} \sum_{f=1}^{n_F} I_{\mathbf{H}^f}(\mathbf{x}^f; \hat{\mathbf{y}}^f) \qquad (16a)$$

$$= \frac{(1 - L_U^{-1})}{M} \frac{1}{n_F} \sum_{f=1}^{n_F} \log\left|\mathbf{I} + P_B(\mathbf{I} + \mathbf{D}_q)^{-1} \mathbf{H}^f (\mathbf{H}^f)^{\mathsf{H}}\right| \qquad (16b)$$

where $\mathbf{D}_q = \text{diag}(\sigma_{q,1}^2, \ldots, \sigma_{q,M}^2)$. The average per-cell sum-rate is obtained by averaging (16b) over the channel realizations $\mathbf{H}^f$. A closed form expression for (16b) can be obtained for the case of no fading under the Wyner model [42].

V. DOWNLINK ORTHOGONAL MULTIPLE ACCESS

In this section, we consider the performance under H-OMA for the DL. Recalling that, as seen in Fig. 2(a), one every L_U

mini-slots is allocated to URLLC users, the signal sent by the k-th EN for any symbol of the RU (f, t) can be written as

$$x_k^f(t) = \begin{cases} x_{U,k}^f(t) & \text{for } t = L_U, 2L_U, \ldots \\ x_{B,k}^f(t) & \text{otherwise,} \end{cases} \qquad (17)$$

where $x_{U,k}^f(t)$ and $x_{B,k}^f(t)$ are the signals intended for URLLC and eMBB users, respectively, over the given RU. Note that we have $x_{U,k}^f(t) = 0$ if no URLLC packet was generated in mini-slots $t, t-1, \ldots, t - L_U + 1$. As a result of the power constraint (9), we have the conditions $\mathbb{E}[|x_{U,k}^f(t)|^2] \leq P$ and $\mathbb{E}[|x_{B,k}^f(t)|^2] \leq P$.

A. URLLC PERFORMANCE

The rate analysis of the performance of URLLC traffic under H-OMA in the DL yields the same results (11)-(12) as for the UL with the caveat that P_U should be replaced by the EN power constraint P. This is because in both cases, under H-OMA, URLLC links are interference free. Furthermore, the probability of error (13) should be modified as

$$\Pr[\mathcal{E}_U] = \sum_{n=1}^{L_U-1} p(n) \frac{n}{n+1} + \sum_{n=0}^{L_U-1} p(n) \frac{1}{n+1} \epsilon_U^D, \qquad (18)$$

since, in the DL, in case multiple URLLC packets are generated at an EN in the L_U mini-slots per transmission opportunity, one packet can be selected at random and delivered to the corresponding user by the EN. In (18), the first term is the probability that more than one additional packets are generated and the packet of interest is blocked from access (see Fig. 3). The second term accounts instead for the decoding error of the transmitted packet.

B. eMBB RATE

Conventional C-RAN transmission based on linear precoding at the BBU and fronthaul quantization is carried out to communicate with eMBB users (see [34], [35]). To elaborate, we define as $s_k^f \sim \mathcal{CN}(0, 1)$ the independent encoded symbols intended for the eMBB user active in cell k over frequency channel f at a given mini-slot. The assumption reflects the use of standard Gaussian random codebooks. As illustrated in Fig. 4(b) the BBU carries out linear precoding separately for each frequency channel f, producing the $M \times 1$ vector

$$\hat{\mathbf{x}}_B^f = \mathbf{V}^f \mathbf{s}^f, \qquad (19)$$

where we have defined the vectors $\mathbf{s}^f = [s_1^f, s_2^f, \ldots, s_M^f]^{\mathsf{T}}$, $\hat{\mathbf{x}}_B^f = [\hat{x}_{B,1}^f, \hat{x}_{B,2}^f, \ldots, \hat{x}_{B,M}^f]^{\mathsf{T}}$ and $\mathbf{V}^f$ is the $M \times M$ channel precoding matrix for all eMBB users active on frequency channel f. We write $\mathbf{V}^f = [\mathbf{v}_1^f, \ldots, \mathbf{v}_M^f] = [\mathbf{v}_{(1)}^f, \ldots, \mathbf{v}_{(M)}^f]^{\mathsf{T}}$, where $\mathbf{v}_j^f \in \mathbb{C}^{M \times 1}$ and $\mathbf{v}_{(j)}^f \in \mathbb{C}^{1 \times M}$ are the j-th column and j-th row of matrix $\mathbf{V}^f$, respectively.

Assuming the standard C-RAN operation where the BBU compresses and forward the eMBB signal to each EN, the signal received at all ENs from the BBU over each frequency

channel f can be written as

$$\mathbf{x}_B^f = \hat{\mathbf{x}}_B^f + \mathbf{q}^f, \tag{20}$$

where we have defined $\mathbf{q}^f = [q_1^f, \ldots, q_M^f]^\mathsf{T}$ and $q_k^f \sim \mathcal{CN}(0, \sigma_{q,k}^2)$ represents the quantization noise with power $\sigma_{q,k}^2$. The quantization noise is independent across the EN index k and frequency channel f. Consequently, the received signal (7) at the k-th eMBB user can be written as

$$y_k^f = \mathbf{h}_{(k)}^f \sum_{j=1}^M \mathbf{v}_j^f s_j^f + \mathbf{h}_{(k)}^f \mathbf{q}^f + z_k^f. \tag{21}$$

In order to obtain the quantization noise's power $\sigma_{q,k}^2$, in a manner similar to (15), we impose the conventional rate distortion condition [41]

$$C = (1 - L_U^{-1}) \frac{1}{n_F} \sum_{f=1}^{n_F} I_{\mathbf{H}^f}(\hat{x}_{B,k}^f; x_{B,k}^f)$$
$$= (1 - L_U^{-1}) \frac{1}{n_F} \sum_{f=1}^{n_F} \log\left(1 + \frac{\|\mathbf{v}_{(k)}^f\|^2}{\sigma_{q,k}^2}\right), \tag{22}$$

for all ENs $k \in 1, \ldots, M$, which follows from the fronthaul capacity constraint of each EN k.

Based on the derivations above, the eMBB achievable per-cell sum-rate for all eMBB users in cell k for given channel realizations $\mathbf{H}^f$ can be written as

$$R_{B,k} = (1 - L_U^{-1}) \frac{1}{n_F} \sum_{f=1}^{n_F} \log\left(1 + \frac{|\mathbf{h}_{(k)}^f \mathbf{v}_{(k)}^f|^2}{1 + \sigma_{\text{eff},k}^2}\right) \tag{23a}$$

where the effective noise $\sigma_{\text{eff},k}^2 = \sum_{j=1}^M |h_{k,j}^f|^2 \sigma_{q,j}^2 + \sum_{\substack{j=1 \\ j \neq k}}^M |\mathbf{h}_{(k)}^f \mathbf{v}_j^f|^2$ accounts both for the disturbance due to fronthaul quantization and for eMBB inter-cell interference.

Based on the available CSI, precoding can be optimized at the BBU by maximizing the eMBB per-cell sum-rates as

$$\text{maximize } R_B = \frac{1}{M} \sum_{k=1}^M R_{B,k}$$

$$\text{subject to } R_{B,k} \leq (23a) \quad \text{for all } k$$
$$P \geq \|\mathbf{v}_{(k)}^f\|^2 + \sigma_{q,k}^2 \quad \text{for all } k \text{ and } f$$
$$C \geq \frac{1 - L_U^{-1}}{n_F} \sum_{f=1}^{n_F} \log\left(1 + \frac{\|\mathbf{v}_{(k)}^f\|^2}{\sigma_{q,k}^2}\right) \quad \text{for all } k, \tag{24}$$

where the maximization is over the variables $\{\mathbf{V}^f\}_{f=1}^{n_F}$, $\{\sigma_{q,k}^2\}_{k=1}^M$, $\{R_{B,k}\}_{k=1}^M$. The second constraint represents the power constraint at each EN, while the third constraint results from the fronthaul capacity limitations. The problem is non-convex, but it can be tackled using standard methods based on Semidefinite Relaxation (SDR) [43] and Concave-Convex Procedure (CCP) [44]. Accordingly, by performing the change of variables $\Omega_k^f = \mathbf{v}_k^f(\mathbf{v}_k^f)^\mathsf{H}$, adding the constraint $\Omega_k^f \succeq 0$ and relaxing the constraint $\text{rank}(\Omega_k^f) = 1$,

the problem falls in the class of difference of convex problems (DC) [44] and thus CCP can be used to solve it as in, e.g., [45, Sec. IV]. In order to ensure the rank constraint, we adopt the standard approach of considering the dominant eigen vector $\mathbf{v}_k^f$ of the solution matrices Ω_k^f (see [43]).
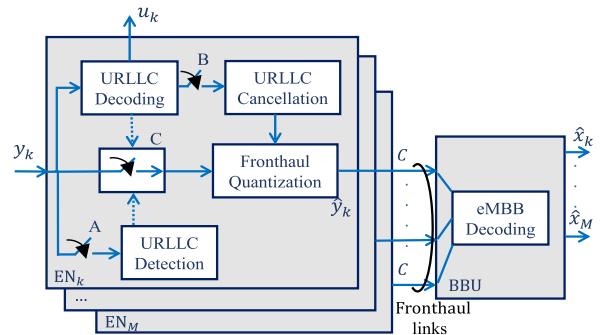

FIGURE 5. Block diagrams for the ENs and the BBU under H-NOMA for UL with (i) TIN, obtained with switches A and B open and switch C closed; (ii) puncturing, with switch A closed and switch B open; (iii) SIC with switch A open and switch B closed, where switch C remains closed when there is an error in URLLC decoding or detection.

VI. UPLINK NON-ORTHOGONAL MULTIPLE ACCESS

In this section we consider the UL performance under H-NOMA. As discussed in Section II, with H-NOMA, eMBB and URLLC users may interfere with each other. For eMBB users, interference is dealt with at the BBU when jointly decoding the eMBB signals. To this end, as illustrated in Fig. 5, three decoding strategies are studied, namely Treating URLLC Interference as Noise (TIN), puncturing, and Successive Interference Cancellation (SIC). With TIN, ENs quantize and forward the received signals to the BBU on the fronthaul links, and the BBU decodes the eMBB transmissions while treating URLLC signal as noise. Under puncturing, whenever an URLLC user is active in a mini-slot, the receiving EN discards the corresponding eMBB signal received in the same mini-slot prior to quantizing the received signal and forwarding it to the BBU over the fronthaul links. Consequently, the BBU decodes only interference-free eMBB mini-slots. Finally, with the more advanced SIC decoder, the ENs decode and cancel the URLLC transmission before fronthaul quantization. In contrast, for URLLC transmissions, due to reliability and latency constraints, the ENs cannot wait for the entire eMBB frame to be received, and hence URLLC decoding cannot benefit from interference cancellation of the eMBB signal. Therefore, the only affordable decoding strategy for URLLC transmissions is treating eMBB transmissions as noise.

A. URLLC RATE

With H-NOMA, as illustrated in Fig. 2(b), URLLC users transmit in any mini-slot in which a packet is generated with no additional access latency. We hence have the minimal access latency of $L_U = 1$. As for the probability of error, an error can only occur when decoding fails, since

$$R_B = \frac{1}{Mn_F} \sum_{f=1}^{n_F} I_{\mathbf{H}^f, \mathbf{G}^f}(\mathbf{x}^f; \hat{\mathbf{y}}^f | \mathbf{A}) = \frac{1}{Mn_F} \sum_{f=1}^{n_F} \mathbb{E}_{\mathbf{A}} \left[\log \frac{\det\left(\mathbf{I} + \mathbf{D}_q + P_B \mathbf{H}^f (\mathbf{H}^f)^{\mathsf{H}} + P_U \mathbf{A} \mathbf{G}^f (\mathbf{G}^f)^{\mathsf{H}}\right)}{\det\left(\mathbf{I} + \mathbf{D}_q + P_U \mathbf{A} \mathbf{G}^f (\mathbf{G}^f)^{\mathsf{H}}\right)} \right]. \tag{28}$$

no collisions may occur under the given assumptions (see Sec. III). Hence, the probability of error coincides with the decoding error probability ϵ_U^D and the reliability constraint (10) imposes the inequality $\epsilon_U^D \leq \epsilon_U$. For a given reliability level ϵ_U, the URLLC outage capacity is thus given as in (11) and (12) with $\epsilon_U^D = \epsilon_U$, and with the caveat that the signal-to-noise-plus-interference ratio $S_{U,k}$ at the k-th EN can be written as

$$S_{U,k}^f = \frac{|g_k^f|^2 P_U}{1 + \sum_{j=1}^{M} |h_{k,j}^f|^2 P_B}, \tag{25}$$

where $\sum_{j=1}^{M} |h_{k,j}^f|^2 P_B$ accounts for interference from eMBB users in all cells active over frequency channel f.

B. eMBB RATE UNDER TREATING URLLC INTERFERENCE AS NOISE

Turning to the eMBB performance, we first study the standard C-RAN solution whereby the EN quantizes and forwards all the received signals to the BBU. The BBU decodes the eMBB messages while treating URLLC signals as noise. Under these assumptions, the signal $\hat{\mathbf{y}}^f = [\hat{y}_1^f, \dots, \hat{y}_M^f]^{\mathsf{T}}$ received at the cloud from all ENs at RU (f, t) can be written in matrix form using (5), as

$$\hat{\mathbf{y}}^f = \mathbf{y}^f + \mathbf{q}^f = \mathbf{H}^f \mathbf{x}^f + \mathbf{A} \mathbf{G}^f \mathbf{u}^f + \mathbf{z}^f + \mathbf{q}^f, \tag{26}$$

where $\mathbf{q}^f = [q_1^f, \dots, q_M^f]$ and $q_k^f \sim \mathcal{CN}(0, \sigma_{q,k}^2)$. In (26), the URLLC activation matrix $\mathbf{A} = \mathrm{diag}(A_1, \dots, A_M)$ contains i.i.d. $\mathcal{B}ern(a_U)$ variables. In order to obtain the quantization noises' powers $\sigma_{q,k}^2$, in a similar manner as in (15), we impose the fronthaul capacity constraint for $k = 1, \dots, M$ as

$$C = \frac{1}{n_F} \sum_{f=1}^{n_F} I_{\mathbf{H}^f, \mathbf{G}^f}(y_k^f; \hat{y}_k^f | A_k)$$
$$= \frac{1}{n_F} \sum_{f=1}^{n_F} \mathbb{E}_{A_k} \left[\log \left(1 + \frac{1 + A_k |g_k^f|^2 P_U + \sum_{j=1}^{M} |h_{k,j}^f|^2 P_B}{\sigma_{q,k}^2} \right) \right]. \tag{27}$$

We note that equation (27) assumes that the BBU is able to detect the presence of URLLC transmissions. This is reflected in the expectation over the URLLC users' activations. Finally, the eMBB per-user rate for given channel realizations $\mathbf{H}^f$ is given by (28) where we recall the notation $\mathbf{D}_q = \mathrm{diag}(\sigma_{q,1}^2, \dots, \sigma_{q,M}^2)$. The expectation in (28), as shown at the top of this page, can in practice be computed exactly by summing over the 2^M possible values for matrix $\mathbf{A}$ as long as M is not too large. Otherwise, stochastic approximation methods can be used.

C. eMBB RATE UNDER PUNCTURING

With puncturing, as seen in Fig. 5, whenever an URLLC user is active in a given cell k, and RU (f, t), the signal $y_k^f(t)$ received at the EN is discarded and not forwarded to the BBU. Note that, with the assumed grant-free URLLC transmissions, this requires the detection of URLLC user's activity prior to fronthaul quantization, e.g., based on the detection of URLLC references sequences. A similar approach is under consideration within 3GPP [46].

To elaborate, we assume that each EN detects correctly that there are transmissions of URLLC devices. The assumption is well justified by the high reliability of URLLC transmissions. The EN compresses and forwards only the signals received during mini-slots free of interference from URLLC transmissions. Under this assumption, the signal $\tilde{y}_k^f$ received at the cloud from EN_k over RU (f, t) can be written as

$$\tilde{y}_k^f = (1 - A_k)\left(h_{k,k}^f x_k^f + \sum_{j \neq k} h_{k,j}^f x_j^f \right) + z_k^f + q_k^f. \tag{29}$$

According to (29), the received signal $\tilde{y}_k^f(t)$ carries no information, i.e., only noise, when an URLLC user is active $(A_k = 1)$ in the corresponding mini-slot. Otherwise, when $A_k = 0$, the signal contains the contributions of the eMBB users and of the quantization noise $q_k^f(t) \sim \mathcal{CN}(0, \sigma_{q,k}^2)$. In matrix form, the signal in (29) received across all ENs over RU (f, t) can be equivalently written as

$$\tilde{\mathbf{y}}^f = (\mathbf{I} - \mathbf{A}) \mathbf{H}^f \mathbf{x}^f + \mathbf{z}^f + \mathbf{q}^f. \tag{30}$$

In order to enable decoding, the BBU at the cloud needs to be informed not only of the signals (29) for all the mini-slots with $A_k = 0$ for all ENs k, but also of the location of such mini-slots. To this end, each EN collects the i.i.d. binary vector containing the n_T Bernoulli variables $A_k \sim \mathcal{B}ern(1 - a_U)$. The number of bits needed to be communicated from EN_k to the BBU in order to ensure the lossless reconstruction of this sequence is given by $n_T H(a_U)$ bits, where $H(a_U) = -a_U \log a_U - (1 - a_U) \log(1 - a_U)$ is the binary entropy function [41]. Based on the discussion above, imposing fronthaul capacity constraint yields the condition

$$n_T n_F l_T l_F C = n_T l_T l_F (1 - a_U)$$
$$\times \sum_{f=1}^{n_F} I_{\mathbf{H}^f}(y_k^f; \tilde{y}_k^f | A_k = 0) + n_T H(a_U), \tag{31}$$

where $I_{\mathbf{H}^f}(y_k^f; \tilde{y}_k^f | A_k = 0) = \log(1 + (1 + \sum_{j=1}^{M} |h_{k,j}^f|^2 P_B)/\sigma_{q,k}^2)$ and we recall that $n_T n_F l_T l_F C$ is the total number of bits per frame available for transmission on each fronthaul link, and the mutual information term accounts for the compression of the received signals over the

$n_T n_F l_T l_F (1 - a_U)$ symbols unaffected by URLLC interference (i.e., with $A_k = 0$). The quantization noise's powers $\sigma_{q,k}^2$ can be obtained by solving (31) using numerical means for all $k = 1, \ldots, M$. Following (28), the eMBB per user rate for given channel realization $\mathbf{H}^f$ is finally given by

$$R_B = \frac{1}{Mn_F} \sum_{f=1}^{n_F} \mathbb{E}_{\mathbf{A}} \left[\log \left| \mathbf{I} + P_B (\mathbf{I} + \mathbf{D}_q)^{-1} (\mathbf{I} - \mathbf{A}) \mathbf{H}^f (\mathbf{H}^f)^{\mathsf{H}} \right| \right]. \tag{32}$$

A closed form expression for (16b) can be obtained for the case of no fading under the Wyner model [29], [42].

D. eMBB RATE UNDER SUCCESSIVE INTERFERENCE CANCELLATION OF URLLC USERS

We finally study a more complex receiver architecture whereby SIC of URLLC packets is carried out at the ENs prior to fronthaul quantization. More specifically, as seen in Fig. 5, if an URLLC user is active and its message is decoded correctly at the receiving EN, the URLLC message is canceled by the EN. If decoding is unsuccessful, the signal received in the corresponding mini-slots is treated as in puncturing. Accordingly, with this scheme, each EN quantizes the received signals only for the minislots that are either free of URLLC transmissions or that contain URLLC messages that were successfully decoded and canceled at the EN. As a result, the received signal at the BBU from k-th EN can be written as (29) but with an erasure probability of $a_U \epsilon_U^D$ instead of a_U. This is because a mini-slot is dropped if an URLLC user in the cell is active, which happens with probability a_U, and if its transmission is decoded incorrectly, which happens with probability ϵ_U^D. As a result, the eMBB rate can be evaluated as (32) with the caveat that the random variables A_k for $k = 1, \ldots, M$ are i.i.d. $\mathcal{Bern}(a_U \epsilon_U^D)$.

VII. DOWNLINK NON-ORTHOGONAL MULTIPLE ACCESS

In this section we consider H-NOMA for the DL scenario. We follow an approach similar to the UL in Sec. V by allowing for different interference management strategies between URLLC and eMBB. A key new aspect in the DL is that interference arises from the URLLC transmissions originating at the ENs, which are a priori unknown to the BBU. Accordingly, as illustrated in Fig. 6, we consider two interference management strategies at the ENs, namely puncturing and superposition coding. Under puncturing, in any mini-slot in which an URLLC packet is generated, the EN transmits only the URLLC packet, dropping any eMBB information. Instead, with superposition coding, the EN transmits a superposition of both eMBB and URLLC signals to both users.

A. PUNCTURING

For both puncturing and superposition coding, the BBU precodes the eMBB signals and forwards them to the ENs over the fronthaul links. Under puncturing, whenever an URLLC packet is generated at an EN in a given mini-slot, the EN discards the eMBB signal received for the same mini-slot
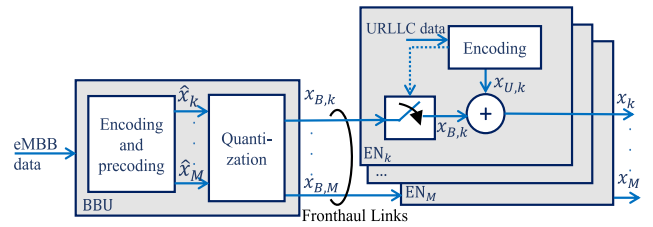

FIGURE 6. Block diagrams for the ENs and the BBU under under H-NOMA for DL with (*i*) puncturing, where the switch is open whenever a URLLC packet is encoded and closed otherwise; and (*ii*) superposition coding, with the switch being always closed.

from the BBU. Note that this does not affect the scheduling decision made at the BBU for eMBB traffic, whose packet still spans the frame, with the exclusion of the punctured mini-slots. Consequently, the transmitted signal x_k^f by EN_k can be written as

$$x_k^f = (1 - A_k)(x_{B,k}^f + q_k^f) + A_k x_{U,k}^f, \tag{33}$$

where we recall that $A_k \sim \mathcal{Bern}(a_U)$ is the binary random variable denoting the generation of a URLLC packet at the EN in mini-slot t.

URLLC Rate: URLLC users' outage capacity is the same as in the H-OMA case discussed in Sec. V due the absence of inter-cell interference at URLLC users. However, with H-NOMA, the probability of error is equal to the decoding error probability due to the absence of collisions between URLLC packets. As a result, the URLLC rate is given by (11) with $\epsilon_U^D = \epsilon_U$.

eMBB Rate: By assuming linear precoding at the BBU with precoding matrix $\mathbf{V}^f$ as in Sec. V, the signal received by the k-th eMBB user in a given mini-slot can be written in a manner similar to (21) as

$$y_k^f = \mathbf{h}_{(k)}^f (\mathbf{I} - \mathbf{A}) \left(\sum_{j=1}^M \mathbf{v}_j^f s_j^f + \mathbf{q}^f \right) + \mathbf{h}_{(k)}^f \mathbf{A} \mathbf{x}_U^f + n_k^f, \tag{34}$$

with definitions given in Sec. III. According to (34), an eMBB user receives useful information only from the ENs in cells j for which no URLLC packet is generated i.e., $A_j = 1$. The per-user eMBB rate at the k-th user and for given channel realizations $\mathbf{H}^f$ can be written as (35), as shown at the bottom of the next page, where the term $\sum_{j=1}^M |h_{k,j}^f|^2 A_j P$ accounts for URLLC interference and $\sum_{\substack{j=1 \\ j \neq k}}^M |\mathbf{h}_{(k)}^f (\mathbf{I} - \mathbf{A}) \mathbf{v}_j^f|^2$ accounts for eMBB interference. We remark that achievability of (35) requires the capability of eMBB users to detect URLLC transmissions, e.g., using a specific preamble in the URLLC minislots. Furthermore, we note that, with puncturing, eMBB and URLLC transmissions are orthogonal within each cell, but inter-service interference still arises due to the asynchronous URLLC packet generation across cells.

In a manner similar to (23), optimal linear precoding can be carried out by maximizing the sum-rates at the BBU. An interesting aspect of this problem is that while the channel matrix is known at the BBU, the effective channels $(\mathbf{I} - \mathbf{A}) \mathbf{H}^f$

are not known to the BBU due to the presence of the random URLLC activation matrix $\mathbf{A}$. The sum-rate maximization problem can be formulated and tackled in a manner similar to (24).

B. SUPERPOSITION CODING

Under this strategy, each EN transmits a superposition of the signal $x^f_{U,k}$ intended for URLLC users and the signal $x^f_{B,k}$ for eMBB users. We fix the power of the signal intended to the URLLC user to $\mathbb{E}[|x^f_{U,k}|^2] = P_U \le P$. When designing the beamforming matrices $\mathbf{V}^f$, the BBU assumes an available power of P since it is not aware of the URLLC activations A_k for $k = 1, \ldots, M$. We hence have the constraint $\|\mathbf{v}^f_{(k)}\|^2 \le P - \sigma^2_{q,k}$ as for puncturing. Accordingly, the signal x^f_k transmitted by the k-th EN over RU (f, t) can be written as

$$x^f_k = (1 + A_k(\sqrt{\delta} - 1))(x^f_{B,k} + q^f_k) + A_k x^f_{U,k}, \quad (36)$$

with the scaling factor $\delta = 1 - P_U/P$. The signal (36) is such that, when $A_k = 0$, only the eMBB signal $x^f_{B,k} + q^f_k$ is transmitted; and, when $A_k = 1$, the transmitted signal is given by the superposition $\sqrt{\delta}(x^f_{B,k} + q^f_k) + x^f_{U,k}$, where the factor δ guarantees the EN power constraint. Note that this strategy is a generalization of puncturing which is obtained by setting $P_U = P$ in (36).

URLLC Performance: The URLLC rate and corresponding outage probability can be obtained using (11) and (12) by setting the following value of the signal-to-noise-plus-interference ratio at the k-th EN

$$S^f_{U,k} = \frac{|g^f_k|^2 P_U}{1 + |g^f_k|^2 (P - P_U)}, \quad (37)$$

where the disturbance term $|g^f_k|^2(P - P_U)$ represents the eMBB interference towards the URLLC user.

eMBB Performance: The signal y^f_k received at the k-th eMBB user on frequency f can be written as

$$y^f_k = \mathbf{h}^f_{(k)}\left(\left(\mathbf{I} + \mathbf{A}(\Delta - \mathbf{I})\right)\left(\sum_{j=1}^{M} \mathbf{v}^f_j s^f_j + \mathbf{q}^f\right) + \mathbf{A}x^f_U\right) + z^f_k, \quad (38)$$

where $\Delta = \delta\mathbf{I}$. Assuming again that the eMBB users can detect URLLC packets, the eMBB per-user rate can be written as (39), as shown at the bottom of this page, where $W_k = 1 + A_k(\delta - 1)$. The sum-rate maximization problem can be formulated and tackled in a manner similar to (24).
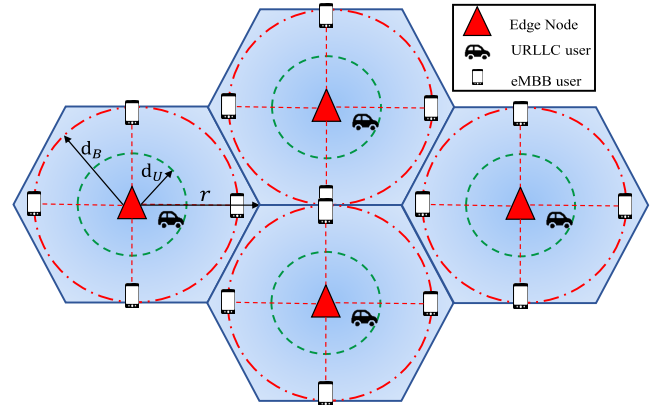

FIGURE 7. Simulations setup with $M = 4$ cells, four eMBB users and one URLLC user per-cell, a cell radius of $r = 2$ km and an URLLC zone (small green circles) with radius of $d_U = 0.1$ km.

VIII. NUMERICAL RESULTS
A. SIMULATIONS SET-UP

As illustrated in Fig. 7, we consider a system of $M = 4$ cells where each cell contains four eMBB users and one URLLC user. The cells' radius is $r = 2$ km. In order to focus on worst-case performance guarantees, eMBB users are located on the circle of radius $d_B = r\sin(\pi/3)$ km, as shown in Fig. 7. In contrast, URLLC users are arbitrary placed on the circle of radius $d_U = 0.1$ km. Furthermore, we set $n_F = 4$ frequency channels in total, with $n^B_F = 1$ frequency channel for each eMBB user. The constraint on the URLLC probability of error is $\epsilon_U = 10^{-3}$. As in 3GPP release 15 [36], we set $l_F = 12$ subcarriers and $l_T = 14$ symbols. For the UL, the eMBB power P_B is fixed to 6.4 dBm, while the URLLC power is $P_U = 23$ dBm. As for the DL, the transmission power of each EN is set to $P = 24.77$ dBm, and, for superposition coding, we fix $P_U = 23$ dBm. The constant c_B in the path loss formula is chosen so as to obtain an average SNR of 3 dB for eMBB at the reference distance $d_{B,R} = d_B$ for both UL and DL with transmission powers P_B in the UL and P in the DL [47]. The constant c_U is instead chosen so as to obtain an average URLLC SNR equal to 10 dB at the reference distance $d_{U,R} = d_U$ for both UL and DL, with transmission powers $P_U = 23$ dBm for both UL and DL [47]. Finally, unless stated otherwise, we assume throughout this section the values $C = 2$ bit/s/Hz, $\gamma = 3$, $a_U = 0.5 \times 10^{-3}$ and $L_U = 2$ for H-OMA.

$$R_{B,k} = \frac{1}{n_F} \sum_{f=1}^{n_F} \mathbb{E}_{\mathbf{A}}\left[\log\left(1 + \frac{|\mathbf{h}^f_{(k)}(\mathbf{I} - \mathbf{A})\mathbf{v}^f_k|^2}{1 + \sum_{j=1}^{M} |h^f_{k,j}|^2((1 - A_j)\sigma^2_{q,j} + A_j P) + \sum_{\substack{j=1 \\ j \ne k}}^{M} |\mathbf{h}^f_{(k)}(\mathbf{I} - \mathbf{A})\mathbf{v}^f_j|^2}\right)\right], \quad (35)$$

$$R_{B,k} = \frac{1}{n_F} \sum_{f=1}^{n_F} \mathbb{E}_{\mathbf{A}}\left[\log\left(1 + \frac{|\mathbf{h}^f_{(k)}(\mathbf{I} + \mathbf{A}(\Delta - \mathbf{I}))\mathbf{v}^f_k|^2}{1 + \sum_{j=1}^{M} |h^f_{k,j}|^2(W_j\sigma^2_{q,j} + A_j P_U) + \sum_{\substack{j=1 \\ j \ne k}}^{M} |\mathbf{h}^f_{(k)}(\mathbf{I} + \mathbf{A}(\Delta - \mathbf{I}))\mathbf{v}^f_j|^2}\right)\right]. \quad (39)$$
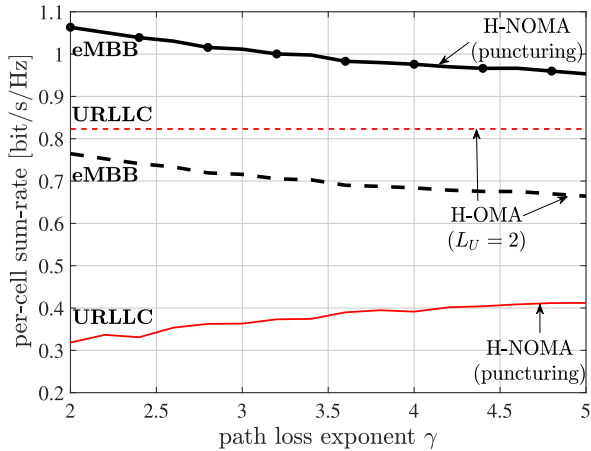
FIGURE 8. UL eMBB average per-cell sum-rate and URLLC per-cell sum-outage capacity as a function of the path loss exponent γ under H-OMA with $L_U = 2$ and H-NOMA with puncturing for the setup in Fig. 7 ($\epsilon_U = 10^{-3}$, $a_U = 0.5 \times 10^{-3}$, $C = 2$ bit/s/Hz).

FIGURE 9. UL eMBB average per-cell sum-rate and URLLC per-cell sum-outage capacity as a function of the packet generation probability a_U for URLLC traffic for H-OMA with $L_U = 2$ and for H-NOMA with different decoding strategies for the set-up in Fig. 7 ($\epsilon_U = 10^{-3}$, $C = 2$ bit/s/Hz, $\gamma = 3$).

B. UPLINK

We first consider the UL. In Fig. 8, we plot the eMBB and URLLC per-cell sum-rate as function of the path loss exponent γ. The average power of the direct channels from each eMBB user to the same cell EN are independent of γ due to the assumption that the reference distance $d_{B,R}$ coincides with the distance d_B. Consequently, decreasing the value of γ effectively increases the inter-cell channel gains for eMBB users (see Sec. II). For H-NOMA, we consider the simplest form of processing, namely puncturing, as studied in Sec. VI.C. We first observe that, in the given scenario with small a_U, H-OMA offers a higher URLLC transmission rate due to the absence of interference of eMBB users, but this comes at the price of the higher URLLC access latency of $L_U = 2$ mini-slots. In contrast, H-NOMA provides the minimal access latency of $L_U = 1$, while supporting a lower URLLC rate that decreases for lower values of γ due to the increasing eMBB interference power. Furthermore, for eMBB traffic, H-NOMA provides a larger rate due to the larger number of available mini-slots. Finally, under both H-NOMA and H-OMA, the eMBB rate increases for decreasing γ thanks to the joint decoding carried out at the BBU, which can benefit from the inter-cell signal paths.

In Fig. 9, we further investigate the per-user eMBB and URLLC rates as a function of URLLC traffic generation probability a_U. The URLLC users' rate under H-OMA is seen to decrease quickly as a function of a_U. This is because, as a_U increases, the error probability in (13) becomes limited by the probability that an URLLC packet is undergoes a collision due to an insufficient number of transmission opportunities. For H-NOMA, the URLLC rate is instead not affected by a_U. As for eMBB, for $a_U \leq 0.86$, treating URLLC signals as noise achieves the worst eMBB rate among the H-NOMA schemes. In fact, in this regime, if the fronthaul capacity is small, it is preferable not to waste fronthaul resources by quantizing samples affected by URLLC interference.
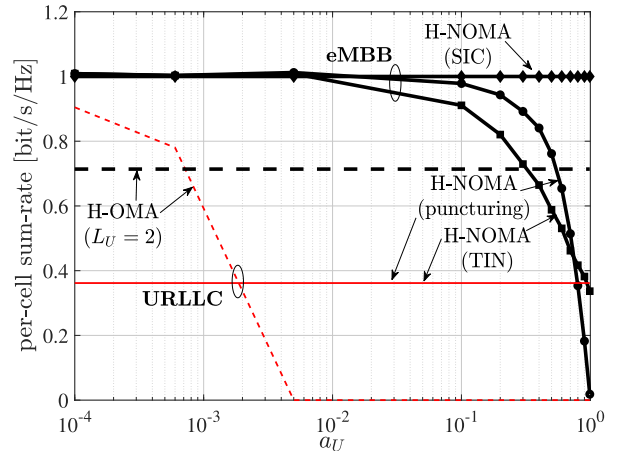
In contrast, for larger values of a_U, puncturing becomes the worst-performing H-NOMA strategy, since the achievable eMBB rate becomes limited by the small number of useful received signal samples forwarded to the BBU. Finally, the more complex SIC scheme always provides the largest per-user eMBB rate thanks to the high probability of cancellation of URLLC signals at the EN.
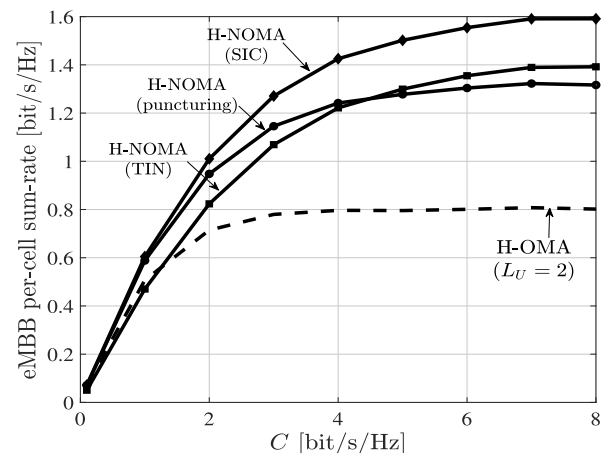

FIGURE 10. UL eMBB average per-cell sum-rate as a function of the fronthaul capacity C for H-OMA with $L_U = 2$ and for H-NOMA with different decoding strategies for the set-up in Fig. 7 ($\epsilon_U = 10^{-3}$, $\gamma = 3$, $a_U = 0.2$).

In Fig. 10, we plot the eMBB per-cell sum-rate rate as a function of the fronthaul capacity C for $a_U = 0.2$. We first note that, for small values of C (in our case, $C \lesssim 4$ bit/s/Hz), puncturing is preferable to treating URLLC as noise, since, as explained above, it avoids wasting the limited fronthaul resources on samples that are corrupted by URLLC interference. In this regime, puncturing provides close performance to SIC, with the added benefit of a lower complexity and power consumption at the ENs. For larger fronthaul
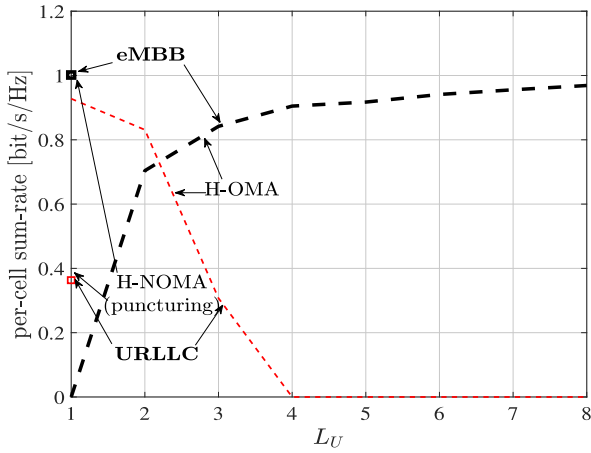
FIGURE 11. UL eMBB average per-cell sum-rate and URLLC per-cell sum-outage capacity as a function of the URLLC access latency L_U for H-OMA and for H-NOMA with different decoding strategies ($\epsilon_U = 10^{-3}$, $C = 2$ bit/s/Hz, $\gamma = 3$, $a_U = 0.5 \times 10^{-3}$).

FIGURE 12. DL eMBB average per-cell sum-rate and URLLC per-cell sum-outage capacity as a function of the packet generation probability a_U for URLLC traffic for H-OMA with $L_U = 2$ and for H-NOMA with different decoding strategies ($\epsilon_U = 10^{-3}$, $C = 2$ bit/s/Hz, $\gamma = 3$).

capacities, the quantization noise tends to zero, and thus treating URLLC as noise outperforms puncturing, given that it allows the BBU to make full use of the received signals. Moreover, H-NOMA with SIC provides the largest rate. Finally, both Fig. 9 and Fig. 10 indicate that, with a sufficiently powerful decoder, such as SIC, the eMBB rate can be improved under H-NOMA as compared to H-OMA.

In Fig. 11, we study the trade-off between the eMBB and URLLC per-user rates as a function of the access latency L_U. Under H-OMA, the URLLC per-user rate decreases when the access latency L_U grows due to the increased probability of URLLC packet collision. In order to compensate for this contribution to the probability of error in (13), one needs to reduce the probability of decoding error ϵ_U^D, causing the rate to decrease (see (11)-(12)). In contrast to H-OMA, H-NOMA provides minimal and constant URLLC latency equal to $L_U = 1$, but at the price of a lower URLLC rate due to interference from eMBB transmission.

C. DOWNLINK

Comparison between H-OMA and H-NOMA is qualitatively similar to the UL and hence we focus here on aspects that are specific to the DL. In Fig. 12, we investigate the average eMBB per-cell sum-rate and URLLC per-cell sum-outage capacity as a function of the URLLC traffic generation probability a_U. As for the UL, the URLLC rate under H-OMA is seen to decrease as function of a_U. However, due to possibility to avoid collisions due to EN scheduling, the DL performance is limited only by blockages and hence the rate degradation is more graceful than for the UL. Another interesting aspect is the comparison between puncturing and superposition coding. Superposition coding is seen to offer a higher eMBB rate due to the larger number of mini-slots available for eMBB transmissions, but a lower URLLC performance owing to interference with eMBB signals which is absent with puncturing. Finally, unlike for the UL results in Fig. 9, H-OMA is observed to provide a larger eMBB
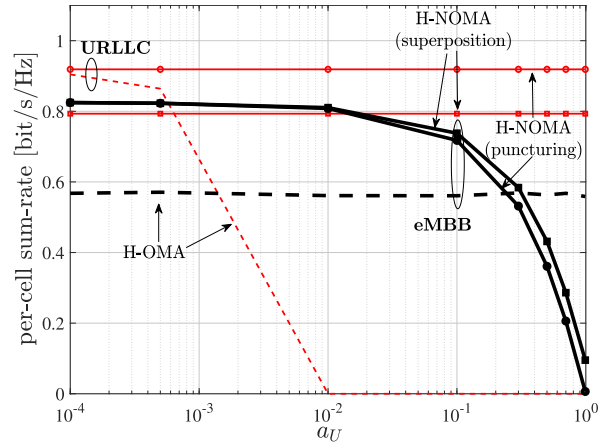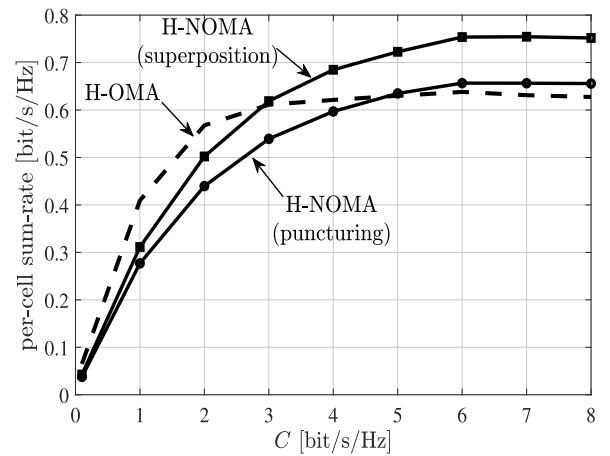

FIGURE 13. DL eMBB average per-cell sum-rate as a function of the fronthaul capacity C for H-OMA with $L_U = 2$ and for H-NOMA with different decoding strategies ($\epsilon_U = 10^{-3}$, $\gamma = 3$, $a_U = 0.4$).

rate than H-NOMA for larger values of a_U. This is because, as discussed, SIC cannot be effectively carried out in the DL.

We further explore the comparison between H-OMA and H-NOMA in Fig. 13, where we plot the eMBB average per-cell sum-rate as function of the fronthaul capacity C for a large URLLC packet generation probability $a_U = 0.4$. The main observation here is that, unlike the UL (cf. Fig. 10), H-NOMA is outperformed by H-OMA for small values of C, here for $C \leq 3$bit/s/Hz. In fact, in the UL, H-NOMA is able to avoid using fronthaul resources for mini-slots that are affected by URLLC interference by leveraging either puncturing or SIC at the ENs. In contrast, in the DL, the BBU is unaware of the URLLC activation and hence it cannot prevent transmitting mini-slots that will eventually either be dropped at the EN, if puncturing is used, or affected by URLLC interference, if superposition coding is adopted. That said, if C is large enough, H-NOMA under both puncturing and superposition coding is able to outperform H-OMA.

IX. CONCLUDING REMARKS

This work has investigated for the first time a multi-cell F-RAN architecture in which eMBB and URLLC services may share the radio resources non-orthogonally with URLLC traffic being processed at the edge and eMBB traffic at the cloud as in a C-RAN architecture. Non-orthogonal transmission was seen to offer potentially significant gains for both eMBB and URLLC services, despite creating inter-service interference. Beside the smaller URLLC access latency, the rate gains of the resulting Heterogeneous-NOMA (H-NOMA) strategy stem from its capability to use more efficiently spectral resources for eMBB traffic, while reducing collisions and blockages for URLLC data. Nevertheless, when the URLLC activation probability is large or the fronthaul capacity is small, the advantages of H-NOMA hinge on an effective management of URLLC interference on eMBB signals. This can be done in the UL by means of puncturing or successive interference cancellation at the ENs prior to fronthaul compression. In contrast, URLLC interference management is more complex in the DL due to the lack of a priori knowledge of the central unit about URLLC activations. Among directions for future work, we mention the inclusion in the study of massive machine-type traffic.

REFERENCES

[1] P. Popovski *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar./Apr. 2018.

[2] *NR and NG-RAN Overall Description-Release 15*, document TS 38.300, 3GPP, Jun. 2018.

[3] "View on 5G architecture," 5G PPP Architecture Working Group, White Paper 1.0, Jul. 2016. [Online]. Available: https://5g-ppp.eu/white-papers/

[4] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: Toward an information theory of fog radio access networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 44–50, Aug. 2016.

[5] T. Q. S. Quek, M. Peng, O. Simeone, and W. Yu, Eds., *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[7] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Downlink NOMA transmission for low-latency short-packet communications," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[8] Z. Ding, R. Schober, and H. V. Poor, "On the design of MIMO-NOMA downlink and uplink transmission," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[9] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor. (2018). "Joint power and time allocation for NOMA-MEC offloading." [Online]. Available: https://arxiv.org/abs/1807.06306

[10] J. Kang, O. Simeone, J. Kang, and S. Shamai (Shitz), "Control-data separation with decentralized edge control in fog-assisted uplink communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3686–3696, Jun. 2018.

[11] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul./Aug. 2016.

[12] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi. (Apr. 2018). "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view." [Online]. Available: https://arxiv.org/abs/1804.05057

[13] *How 5G Connectivity and New Technology Could Pave the Way for Self-Driving Cars*, MIT Technol. Rev., Cambridge, MA, USA, Aug. 2018.

[14] A. E. Kalor, R. Guillaume, J. J. Nielsen, A. Mueller, and P. Popovski, "Network slicing in industry 4.0 applications: Abstraction methods and end-to-end analysis," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 5419–5427, Dec. 2018.

[15] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson, and J. Ansari, "Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G," in *Proc. IEEE ETFA*, Berlin, Germany, Sep. 2016, pp. 1–8.

[16] D. Boviz, C. S. Chen, and S. Yang, "Effective design of multi-user reception and fronthaul rate allocation in 5G cloud RAN," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1825–1836, Aug. 2017.

[17] T. Takeda and K. Higuchi, "Enhanced user fairness using non-orthogonal access with SIC in cellular uplink," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2011, pp. 1–5.

[18] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2014, pp. 781–785.

[19] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[20] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. 24th IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 611–615.

[21] Z. Q. Al-Abbasi and D. K. C. So, "Resource allocation in non-orthogonal and hybrid multiple access system with proportional rate constraint," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6309–6320, Oct. 2017.

[22] Y. Zhang, H.-M. Wang, T.-X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[23] L. Lv, J. Chen, Q. Ni, Z. Ding, and H. Jiang, "Cognitive non-orthogonal multiple access with cooperative relaying: A new wireless frontier for 5G spectrum sharing," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 188–195, Apr. 2018.

[24] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38451–38463, 2018.

[25] S. F. Abedin, G. R. Alam, S. M. A. Kazmi, N. H. Tran, D. Niyato, and C. S. Hong, "Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 489–502, Jan. 2019.

[26] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *CoRR*, vol. abs/1712.05344, Dec. 2017.

[27] T. Keresztfalvi and A. Lapidoth, "Multiplexing zero-error and rare-error communications over a noisy channel with feedback," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1608–1612.

[28] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, to be published.

[29] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study," in *Proc. IEEE GLOBECOM*, Abu Dhabi, United Arab Emirates, Dec. 2018.

[30] A. Matera, R. Kassab, O. Simeone, and U. Spagnolini, "Non-orthogonal eMBB-URLLC radio access for cloud radio access networks with analog fronthauling," *Entropy*, vol. 20, no. 9, p. 661, 2018.

[31] H. Nikbakht, M. Wigger, and S. Shamai (Shtiz). (Jan. 2018). "Mixed delay constraints in Wyner's soft-handoff network." [Online]. Available: https://arxiv.org/abs/1801.05681

[32] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, 2014.

[33] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[34] Y. Zhou and W. Yu, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4138–4151, Aug. 2016.

[35] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[36] *NR—Physical Channels and Modulations—Release 15*, document TS 38.211, 3GPP, Jun. 2018.

[37] J. Kang, O. Simeone, J. Kang, and S. Shamai (Shitz), "Fronthaul compression and precoding design for C-RANs over ergodic fading channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5022–5032, Jul. 2016.

[38] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document TR 38.913, 3GPP, Mar. 2017.

[39] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[40] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[41] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[42] O. Simeone *et al.*, "Cooperative wireless cellular systems: An information-theoretic view," *Found. Trends Commun. Inf. Theory*, vol. 8, nos. 1–2, pp. 1–177, 2012.

[43] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[44] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optim. Eng.*, vol. 17, no. 2, pp. 263–287, Jun. 2016.

[45] S.-H. Park, O. Simeone, and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

[46] "3GPP TSG RAN WG1 meeting 92bis," 3GPP, Sophia Antipolis, France, Tech. Rep. 38.212, Apr. 2018.

[47] P. Joshi, D. Colombi, B. Thors, L.-E. Larsson, and C. Törnevik, "Output power levels of 4G user equipment and implications on realistic RF EMF exposure assessments," *IEEE Access*, vol. 5, pp. 4545–4550, 2017.

OSVALDO SIMEONE received the M.Sc. degree (Hons.) and the Ph.D. degree in information engineering from the Politecnico di Milano, Milan, Italy, in 2001 and 2005, respectively. From 2006 to 2017, he was a Faculty Member with the Electrical and Computer Engineering Department, New Jersey Institute of Technology, where he was affiliated with the Center for Wireless Information Processing. He is currently a Professor of information engineering with the Centre for Telecommunications Research, Department of Informatics, King's College London. He has co-authored two monographs, an edited book published by Cambridge University Press, and more than one hundred research journal papers. His research interests include wireless communications, information theory, optimization, and machine learning. He is a Fellow of the IET and the IEEE. He was a co-recipient of the 2018 IEEE Signal Processing Best Paper Award, the 2017 JCN Best Paper Award, the 2015 IEEE Communication Society Best Tutorial Paper Award, and the Best Paper Awards of the IEEE SPAWC 2007 and the IEEE WRECOM 2007. He was awarded a Consolidator Grant by the European Research Council (ERC), in 2016. His research has been supported by the U.S. NSF, the ERC, the Vienna Science and Technology Fund, as well as by a number of industrial collaborations. He currently serves on the editorial board of the *IEEE Signal Processing Magazine*, and he is a Distinguished Lecturer of the IEEE Information Theory Society.

PETAR POPOVSKI received the Dipl.Ing and Magister Ing. degrees in communication engineering from the University of Saints Cyril and Methodius in Skopje, and the Ph.D. degree from Aalborg University, in 2005. He is currently a Professor of wireless communications with Aalborg University. He holds over 30 patents and patent applications. He has over 300 publications in journals and conference proceedings, and has edited books. His research interests include the area of wireless communication and networking, and communication/information theory. He is a Fellow of the IEEE. He received an ERC Consolidator Grant, in 2015, the Danish Elite Researcher Award, in 2016, the IEEE Fred W. Ellersick Prize, in 2016, and the IEEE Stephen O. Rice Prize, in 2018. He is currently a Steering Committee Member of the IEEE SmartGridComm and previously served as a Steering Committee Member of the IEEE Internet of Things Journal. He is currently an Area Editor of the IEEE Transactions on Wireless Communications.

RAHIF KASSAB received two engineering degrees from Telecom ParisTech and Lebanese University, in 2017, and the master's degree in advanced communication networks jointly from École Polytechnique and Telecom ParisTech. He is currently pursuing the Ph.D. degree with the Department of Informatics, King's College London, U.K. His research interests include communication theory, optimization, and statistics. His industrial experience includes a six-month internship at Nokia Bell Labs, Paris, France. He received the 2018 IEEE Globecom Student Travel Grant, the Ile-de-France Masters Scholarship, from 2015 to 2017, and a Ph.D. Fellowship, from 2018 to 2020, awarded from King's College London and funded by the European research Council. He served as a Reviewer for many journals, including the IEEE Journal on Selected Areas in Communications, the IEEE Journal of Selected Topics in Signal Processing, and the IEEE Transactions on Mobile Computing.

TOUFIQUL ISLAM (S'10–M'15) received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, and the Ph.D. degree in electrical and computer engineering from the University of British Columbia, Vancouver, BC, Canada, in 2006, 2008, and 2014, respectively. He is currently a Wireless Systems Engineer with Intel Corporation working on wireless systems design and research for 5G/NR and beyond. Previously, he was a Research Engineer with the Huawei Canada Research Center, Ottawa, ON, Canada. His research interests include flexible air interface design for 5G and beyond, energy efficiency, URLLC, wireless caching, buffer-aided relaying, and applications of machine learning techniques in communications.

• • •