entropy

MDPI

Article

Non-Orthogonal eMBB-URLLC Radio Access for Cloud Radio Access Networks with Analog Fronthauling

Andrea Matera [1,*], Rahif Kassab [2], Osvaldo Simeone [2] and Umberto Spagnolini [1]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, 20133 Milano, Italy; umberto.spagnolini@polimi.it

[2] Centre for Telecommunications Research (CTR), Department of Informatics, King's College London, London WC2B 4BG, UK; rahif.kassab@kcl.ac.uk (R.K.); osvaldo.simeone@kcl.ac.uk (O.S.)

* Correspondence: andrea.matera@polimi.it; Tel.: +39-22-399-3462

Received: 17 July 2018; Accepted: 31 August 2018; Published: 2 September 2018

Abstract: This paper considers the coexistence of Ultra Reliable Low Latency Communications (URLLC) and enhanced Mobile BroadBand (eMBB) services in the uplink of Cloud Radio Access Network (C-RAN) architecture based on the relaying of radio signals over analog fronthaul links. While Orthogonal Multiple Access (OMA) to the radio resources enables the isolation and the separate design of different 5G services, Non-Orthogonal Multiple Access (NOMA) can enhance the system performance by sharing wireless and fronthaul resources. This paper provides an information-theoretic perspective in the performance of URLLC and eMBB traffic under both OMA and NOMA. The analysis focuses on standard cellular models with additive Gaussian noise links and a finite inter-cell interference span, and it accounts for different decoding strategies such as puncturing, Treating Interference as Noise (TIN) and Successive Interference Cancellation (SIC). Numerical results demonstrate that, for the considered analog fronthauling C-RAN architecture, NOMA achieves higher eMBB rates with respect to OMA, while guaranteeing reliable low-rate URLLC communication with minimal access latency. Moreover, NOMA under SIC is seen to achieve the best performance, while, unlike the case with digital capacity-constrained fronthaul links, TIN always outperforms puncturing.

Keywords: network slicing; RoC; URLLC; eMBB; C-RAN

1. Introduction

Accommodating the heterogeneity of users' requirements is one of the main challenges that both industry and academia are facing in order to make 5G a reality [1]. In fact, next-generation wireless communication systems must be designed to provision different services, each of which with distinct constraints in terms of latency, reliability, and information rate. In particular, 5G is expected to support three different macro-categories of services, namely enhanced Mobile BroadBand (eMBB), massive Machine-Type Communications (mMTC), and Ultra-Reliable and Low-Latency Communications (URLLC) [2–4].

eMBB service is meant to provide very-high data-rate communications as compared with current (4G) networks. This can be generally achieved by using codewords that spread over a large number of time-frequency resources, given that latency is not an issue. mMTC supports low-rate bursty communication between a massive number of uncoordinated devices and the network. Finally, URLLC is designed to ensure low-rate ultra-reliable radio access for a few nodes, while guaranteeing very low-latency. As a result, URLLC transmissions need to be localized in time, and hence URLLC packets should be short [5].

The coexistence among eMBB, mMTC and URLLC traffic types can be ensured by slicing the Radio Access Network (RAN) resources into non-overlapping, or orthogonal, blocks, and by assigning

distinct resources to different services. With the resulting Orthogonal Multiple Access (OMA), the target quality-of-service guarantees can be achieved by designing each service separately [6,7]. However, when URLLC or mMTC traffic types are characterized by short and bursty transmissions at random time instants, resources allocated statically to these services are likely to be unused for most of time, and thus wasted. A more efficient use of radio resources can be accomplished by Non-Orthogonal Multiple Access (or NOMA), which allows multiple services to share the same physical resources.

By enabling an opportunistic shared use of the radio resources, NOMA can provide significant benefits in terms of spectrum efficiency, but it also poses the challenge of designing the system so that the heterogeneous requirements of the services are satisfied despite the mutual interference. The objective of this paper is to address this issue by considering a Cloud-RAN (C-RAN) architecture characterized by analog fronthaul links, referred to as Analog Radio-over-X, which is introduced in the next section.

1.1. C-RAN Based on Analog Radio-over-X Fronthauling

The advent of 5G is introducing advanced physical layer technologies and network deployment strategies such as massive MIMO, mmWave, small-cell densification, mobile edge computing, etc. (see [8,9] for an overview). C-RAN is an enabling technology that is based on the colocation of the Base Band Unit (BBU) of Edge Nodes (ENs) that are densely distributed in a given indoor or outdoor area. This solution has the advantages of allowing for centralized BBU signal processing, providing network scalability, increasing spectral efficiency, and reducing costs.

The most typical C-RAN architecture relies on digital optical fronthaul links to connect ENs to BBUs. This solution, known as Digital Radio-over-Fiber (D-RoF), is adopted in current 4G mobile networks, and is based on the transmission of in-phase and quadrature baseband signals, upon digitization and packetization according to the CPRI protocol [10].

Over the last years, several alternative C-RAN architectures have been proposed that redistribute the RAN functionalities between BBU and ENs, obtaining different trade-offs in terms of bandwidth and latency requirements, advanced Cooperative Multi-Point processing capabilities, and EN cost and complexity [11]. For scenarios with stringent cost and latency constraints, a promising solution is to use analog fronthauling.

With analog fronthauling, focusing on the uplink, the ENs directly relay the radio signals to the BBUs after frequency translation and, possibly, signal amplification. This has the advantages of avoiding any bandwidth expansion due to digitization; guaranteeing ENs synchronization; minimizing latency; reducing hardware cost; and improving energy efficiency [12–14]. A C-RAN architecture based on analog fronthauling is also known in the literature as Analog Radio-over-X (A-RoX), where X depends on the technology employed for the fronthaul, which can be either Fiber (A-RoF [12]), Radio (A-RoR [15]), or Copper (A-RoC [13]), as depicted in Figure 1.

In particular, A-RoF provides an effective example of analog fronthauling, due to its capability to support the transport of large bandwidths [12,16]. However, A-RoF requires the deployment of a fiber optic infrastructure whose installation is not always feasible, e.g., in dense urban areas. In such scenarios, a possible solution is to rely on the A-RoR concept, thus employing point-to-point wireless links, mainly based on mmWave or THz bands, with several advantages in terms of flexibility, resiliency, hardware complexity and cost [15,17]. Another application scenario where the installation of fiber links may be too expensive to provide satisfactory business cases is indoor coverage. For indoor deployments, A-RoC [18–20] has been recently proved to be an attractive solution, especially from the deployment costs perspective [21], since it leverages the pre-existing Local Area Network (LAN) cabling infrastructure of building and enterprises. Moreover, LAN cables are equipped with four twisted-pairs with a transport capability up to 500 MHz each, or 2 GHz overall, for radio signals, thus providing enough bandwidth for analog fronthaul applications [22]. Over the last years, A-RoC based on LAN cables has become a standard solution for in-building commercial C-RAN deployments, allowing to extend the indoor coverage over distances longer than 100 m [23].

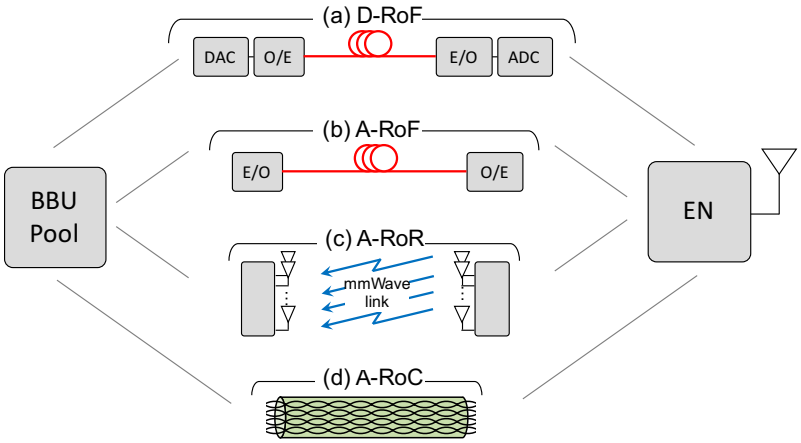

Figure 1. C-RAN architecture overview for uplink direction: (**a**) Digital Radio-over-Fiber, (**b**) Analog Radio-over-Fiber, (**c**) Analog Radio-over-Radio, (**d**) Analog Radio-over-Copper.

In this paper we study the coexistence of URLLC and eMBB services under both OMA and NOMA assuming a C-RAN multi-cell architecture based on analog fronthauling (see Figure 1) by using information theoretical tools.

1.2. Related Works

The concept of NOMA is well-known from the information theoretic literature [24], and its application to 5G dates back to [25], where authors demonstrated for a single-cell scenario that superimposing multiple users in the same resources achieves superior performance with respect to conventional LTE networks, provided that the resulting interference is properly taken care of. The extension of NOMA to multi-cell networks is presented in [26], which addresses several multi-cell NOMA challenges, including coordinated scheduling, beamforming, and practical implementation issues related to successive interference cancellation.

In contrast with most of the works on NOMA, which deal with homogeneous traffic conditions (see [27] for a recent review), here the focus is on NOMA techniques in the context of heterogeneous networks, such as the forthcoming 5G wireless systems, as discussed in [28–30]. In fact, NOMA represents an attractive solution to meet the distinct requirements of 5G services, as it improves spectral efficiency (eMBB), enables massive device connectivity (mMTC), and allows for low-transmission latency (URLLC) [31].

In [29] a communication-theoretic model was introduced to investigate the performance trade-offs for eMBB, mMTC and URLLC services in a single-cell scenario under both OMA and NOMA. This single-cell model has been later extended in [32] to the uplink of a multi-cell C-RAN architecture, in which the BBU communicates with multiple URLLC and eMBB users belonging to different cells through geographically distributed ENs. In the C-RAN system studied in [32], while the URLLC signals are locally decoded at the ENs due to latency constraints, the eMBB signals are quantized and forwarded over limited-capacity digital fronthaul links to the BBU, where centralized joint decoding is performed.

None of the aforementioned works considers the coexistence of different 5G services in a C-RAN architecture based on analog fronthaul links which is the focus of this paper.

1.3. Contributions

In this paper we study for the first time the coexistence between URLLC and eMBB services in the uplink of a C-RAN system with analog fronthauling in which the URLLC signals are still decoded locally at the EN, while the eMBB signals are forwarded to the BBU over analog fronthaul links.

In particular, the main contributions of this paper are three-fold:

- We extend the uplink C-RAN theoretic model proposed in [32] to the case of analog fronthaul, assuming that the fronthaul links are characterized by multiple, generally interfering, channels that carry the received radio signals;
- By leveraging information theoretical tools, we investigate the performance trade-offs between URLLC and eMBB services under both OMA and NOMA, by considering different interference management strategies such as puncturing, considered for the standardization of 5G New Radio [33,34], Treating Interference as Noise (TIN), and Successive Interference Cancellation (SIC);
- The analysis demonstrates that NOMA allows for higher eMBB information rates with respect to OMA, while guaranteeing a reliable low-rate URLLC communication with minimal access latency. Moreover, differently from the case of conventional digital C-RAN architecture based on limited-capacity fronthaul links [32], in analog C-RAN, TIN always outperforms puncturing, while the best performance is still achieved by NOMA with SIC.

1.4. Organization

The remainder of the paper is organized as follows. The considered system model is introduced in Section 2. Section 3 details the fronthaul signal processing techniques employed to cope with the impairments of the A-RoC fronthaul links. The eMBB rand URLLC information rates are discussed in Sections 4 and 5 for OMA and NOMA, respectively. Numerical results are presented in Section 6. And Section 7 concludes the paper.

1.5. Notation

Bold upper- and lower-case letters describe matrices and column vectors, respectively. Letters $\mathbb{R}$, and $\mathbb{C}$ refer to real and complex numbers, respectively. We denote matrix inversion, transposition and conjugate transposition as $(\cdot)^{-1}$, $(\cdot)^{T}$, $(\cdot)^{H}$. Matrix $\mathbf{I}_n$ is an identity matrix of size n and $\mathbf{1}_n$ is a column vector made by n "1s". Symbol $\otimes$ denotes the Kronecker operator, $\mathrm{vec}(\cdot)$ is the vectorization operator, and $\mathbb{E}[\cdot]$ is the statistical expectation. Notation $\mathrm{diag}(A_1, A_2, ..., A_n)$ denotes a diagonal matrix with elements $A_1, A_2, ..., A_n$ on the main diagonal. The Q-function $\mathcal{Q}(\cdot)$ is the complementary cumulative distribution function of the standardized normal random variable, and $\mathcal{Q}^{-1}(\cdot)$ is its inverse.

2. System Model

The C-RAN architecture under study is illustrated in Figure 2. In this system, the BBUs communicate with multiple user equipments (UEs) belonging to M cells through M single-antenna Edge Nodes (ENs). The BBUs are co-located in the so-called BBU pool so that joint processing can be performed, while the ENs are geographically distributed. In particular, we assume here that cells are arranged in a line following the conventional circulant Wyner model ([35], Chapter 2), and each cell contains two single-antenna UEs with different service constraints: one eMBB user and one URLLC user.

Due to the strict latency constraints of URLLC traffic, the signal for the URLLC UEs is decoded on-site at the EN, while the eMBB signals are forwarded to the BBU through a multi-channel analog fronthaul. In this hybrid cloud-edge architecture, the mobile operator equips the EN with edge computing capabilities in order to provision the services required by the URLLC user directly from the EN. Following the A-RoX concept, the end-to-end channel from the eMBB UEs and the BBU pool is assumed to be fully analog: the EN performs only signal amplification and frequency translation to comply with fronthaul capabilities and forwards the signals to the BBU, where centralized decoding is performed. In practice, as detailed later in the paper (see Sections 4 and 5), we assume that each EN hosts a digital module, responsible for URLLC signal decoding, and an analog module, responsible for the mapping of received radio signal over the analog fronthauling, which is identified as Analog-to-Analog (A/A) mapping.

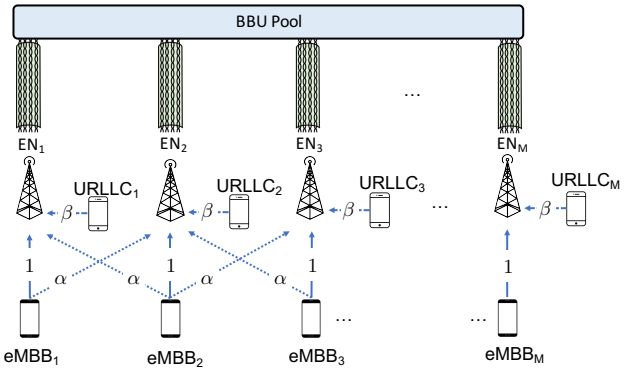

Figure 2. Model of the uplink of C-RAN system based on Analog Radio-over-Copper (A-RoC) fronthauling.

While the technology used for the analog fronthauling can be either fiber-optics (A-RoF), wireless (A-RoR) or cable (A-RoC), the system model proposed in this paper reflects mainly the last two solutions. Furthermore, we will focus on A-RoC, and we will use the corresponding terminology to fix the ideas (see Figure 2).

2.1. RAN Model

We consider the same Wyner-type radio access model of [32], which is described in this subsection. The Wyner model is an abstraction of cellular systems that captures one of the main aspects of such settings, namely the locality of inter-cell interference. The advantage of employing such a simple model is the possibility to obtain analytical insights, which is a first mandatory step for the performance assessment under more realistic operating conditions [35]. As illustrated in Figure 3, the direct channel gain from the eMBB UE and the EN belonging to the same cell is set to one, while the inter-cell eMBB channel gain is equal to $\alpha \in [0, 1]$. Furthermore, the URLLC UEs have a channel gain equal to $\beta > 0$. The URLLC user is assumed to be in the proximity of the EN, and thus it does not interfere with the neighboring cells. The eMBB user, instead, is assumed to be located at the edges of the cell in order to consider worst-case performance guarantees. As a result, each eMBB user interferes with both left and right neighboring cells, following the standard Wyner model [35]. All channel gains are assumed to be constant over the considered radio resources shown Figure 3, and known to all UEs and ENs.

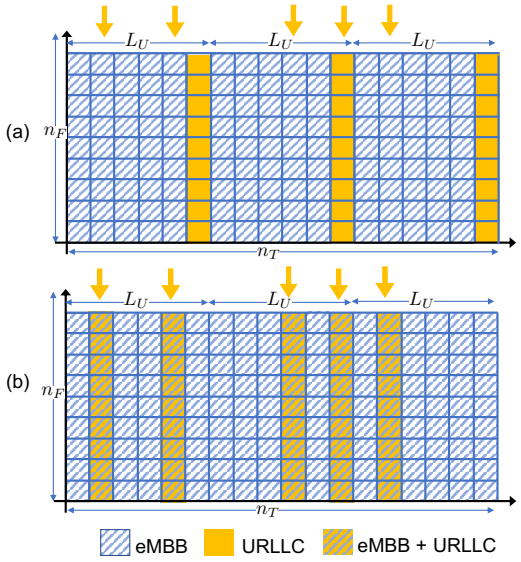

Figure 3. Time-frequency resource allocation: (**a**) Orthogonal Multiple Access (OMA) and (**b**) Non-Orthogonal Multiple Access (NOMA). Downwards arrows denote arrival of URLLC packets.

As illustrated in Figure 3, we assume that the time-frequency plane is divided in n_T minislots, indexed as $t \in [1, n_T]$, where each minislot is composed of n_F frequency channels, indexed as $f \in [1, n_F]$, for a total of $n_F n_T$ time-frequency radio resources. Each radio resource accommodates the transmission of a single symbol, although generalizations are straightforward. The eMBB UEs transmit over the entire $n_F \times n_T$ time-frequency frame. In contrast, due to the latency constraints of the URLLC traffic, each URLLC transmission is limited to the n_F frequency channels of a single minislot, and URLLC packets are generally small compared to the eMBB frame, which requires the condition $n_T \gg 1$. As illustrated in Figure 3, each URLLC UE generates an independent packet in each minislot with probability q. This packet is transmitted at the next available transmission opportunity in a grant-free manner.

In the case of OMA, one minislot is exclusively allocated for URLLC transmission every L_U minislots. Parameter L_U is considered here as the worst-case access latency. Accordingly, if more than one packet is generated within the L_U minislots between two transmission opportunities, only one of those packets is randomly selected for transmission and all the remaining are discarded. The signal $Y_k^f(t)$ received at the k-th EN at the f-th frequency under OMA is

$$Y_k^f(t) = \begin{cases} \beta A_k(t) U_k^f(t) + Z_k^f(t), & \text{if } t = L_U,\ 2L_U,\ \ldots \\ X_k^f(t) + \alpha X_{k-1}^f(t) + \alpha X_{k+1}^f(t) + Z_k^f(t), & \text{otherwise} \end{cases} \tag{1}$$

where $X_k^f(t)$ and $U_k^f(t)$ are the signals transmitted at time t and subcarrier f by the k-th eMBB UE and URLLC UE, respectively; $Z_k^f(t) \sim \mathcal{CN}(0,1)$ is the unit-power zero-mean additive white Gaussian noise; and $A_k(t) \in \{0,1\}$ is a binary variable indicating whether or not the URLLC UE is transmitting at time t.

In case of NOMA, the URLLC UE transmits its packet in the same slot where it is generated by the application layer, so that the access latency is always minimal, i.e., $L_U = 1$ minislot. Under NOMA, the signal $Y_k^f(t)$ received at the k-th EN at the f-th frequency is

$$Y_k^f(t) = X_k^f(t) + \alpha X_{k-1}^f(t) + \alpha X_{k+1}^f(t) + \beta A_k(t) U_k^f(t) + Z_k^f(t). \tag{2}$$

According to the circulant Wyner model, in (1) and (2), we assume that $[k-1] = M$ for $k = 1$ and $[k+1] = 1$ for $k = M$, in order to guarantee symmetry.

For both OMA and NOMA, the power constraints for the k-th eMBB and URLLC users are defined within each radio resource frame respectively as

$$\frac{1}{n_F n_T} \sum_{t=1}^{n_T} \sum_{f=1}^{n_F} \mathbb{E}\left[\left| X_k^f(t) \right|^2 \right] \leq P_B, \tag{3}$$

and

$$\frac{1}{n_F} \sum_{f=1}^{n_F} \mathbb{E}\left[\left| U_k^f(t) \right|^2 \right] \leq P_U, \tag{4}$$

where the temporal average are taken over all symbols within a codeword.

Models (1) and (2) can be written in matrix form as

$$\mathbf{Y}(t) = \mathbf{X}(t)\mathbf{H} + \beta \mathbf{U}(t)\mathbf{A}(t) + \mathbf{Z}(t), \tag{5}$$

where matrix $\mathbf{Y}(t) = [\mathbf{y}_1(t),\ \mathbf{y}_2(t),\ \ldots,\ \mathbf{y}_M(t)] \in \mathbb{C}^{n_F \times M}$ gathers all the signals received at all the M ENs over all the n_F frequencies, and the k-th column $\mathbf{y}_k(t) \in \mathbb{C}^{n_F \times 1}$ denotes the signal received across all the radio frequencies at the k-th EN. The channel matrix $\mathbf{H} \in \mathbb{R}^{M \times M}$ is circulant with the first column given by vector $[1,\ \alpha,\ 0,\ \ldots,\ 0,\ \alpha]^T$; matrices $\mathbf{U}(t) \in \mathbb{C}^{n_F \times M}$ and $\mathbf{X}(t) \in \mathbb{C}^{n_F \times M}$ collect the signals transmitted by URLLC and eMBB UEs, respectively; and $\mathbf{Z}(t) \in \mathbb{C}^{n_F \times M}$ is the overall noise

matrix. Finally, $\mathbf{A}(t)$ is a diagonal matrix whose k-th diagonal element is a Bernoulli random variable distributed as $A_k(t) \sim \mathcal{B}(q)$, $\forall k = 1, 2, ..., M$.

2.2. Space-Frequency Analog Fronthaul Channel

In the considered analog fronthaul architecture, the k-th EN forwards the signal $\mathbf{y}_k(t)$ received by the UEs to the BBU over a wired-access link in a fully analog fashion. As depicted in Figure 2, we focus our attention on a multichannel link that is possibly affected by inter-channel interference. While the model considered here can apply also to wireless multichannel links, as in Figure 1, we adopt here the terminology of Analog Radio-over-Copper (A-RoC) as an important example in which A-RoX is affected by fronthauling inter-link interference. Accordingly, each of the cables employed for the fronthaul contains l_S twisted-pairs, i.e., l_S space-separated channels, indexed as $c \in [1, l_S]$. Each pair carries a bandwidth equal to $l_F \leq n_F$ frequency channels of the RAN, indexed as $f' \in [1, l_F]$, so that a total of $l_S l_F$ space-frequency resource blocks are available over each cable, as shown in Figure 4. Furthermore, we assume that each analog fronthaul link has enough resources to accommodate the transmission of the whole radio signal at each EN, i.e., $l_S l_F \geq n_F$.

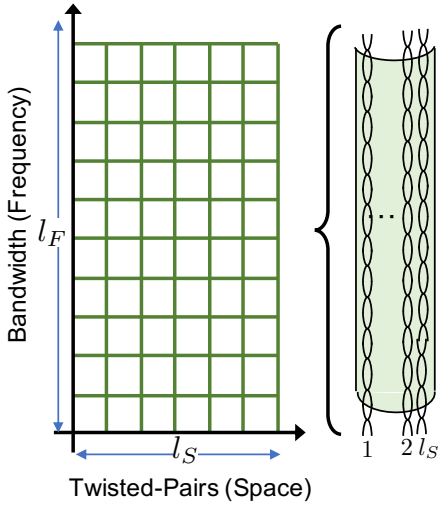

Figure 4. Space-frequency cable resource allocation.

The fronthaul channel between each EN and the BBU is described by the matrix $\mathbf{H}_c \in \mathbb{R}^{l_S \times l_S}$, which accounts for direct channel gains on each cable, given by the diagonal elements $[\mathbf{H}_c]_{ii}$, and for the intra-cable crosstalk, described by the off-diagonal elements $[\mathbf{H}_c]_{ij}$, with $i \neq j$. We assume here that the channel coefficients in $\mathbf{H}_c$ do not depend on frequency f'. Furthermore, in keeping the modeling assumptions of the Wyner model, we posit that the direct channel gains for all the pairs are normalized to 1, while the crosstalk coefficients between any pair of twisted-pairs are given by a coupling parameter $\gamma \geq 0$. It follows that the fronthaul channel matrix can be written as

$$\mathbf{H}_c = \gamma \mathbf{1}_{l_S} \mathbf{1}_{l_S}^T + (1 - \gamma) \mathbf{I}_{l_S}, \tag{6}$$

where $\mathbf{1}_n$ denotes a column vector of size n of all ones and $\mathbf{I}_n$ is the identity matrix of size n. We note that, in case of wireless fronthaul links such as in A-RoR, the coefficient γ accounts for the mutual interference between spatially separated radio links. As a result, one typically has $\gamma > 0$ when considering sub-6 GHz frequency bands, while the condition $\gamma = 0$ may be reasonable in the mmWave or THz bands, in which communication is mainly noise-limited due to the highly directive beams [36].

For a given time t, the symbols $\mathbf{y}_k(t)$ received at EN k-th over all the n_F radio frequency channels are transported to the BBU over the $l_S l_F$ cable resource blocks, where the mapping between radio and cable resources is to be designed (see Section 3.1) and depends on the bandwidth l_F available at each twisted-pair.

In this regards, we define the fraction $\mu \in [1/l_S, 1]$ of the radio bandwidth n_F that can be carried by each pair, referred to as *normalized cable bandwidth*, as

$$\mu = \frac{l_F}{n_F}. \tag{7}$$

As a result, the quantity

$$\eta = \mu \cdot l_S \tag{8}$$

expresses the bandwidth amplification factor (or redundancy) over the cable fronthaul, as $\eta \geq 1$. To simplify, we assume here that both $1/\mu$ and η are integer numbers. The two extreme situations with $\mu = 1$, or $\eta = l_S$, and $\mu = 1/l_S$, or $\eta = 1$, are shown in Figure 5 for $l_S = 4$ twisted-pairs and $n_F = 8$ subcarriers. For the first case, one replica of the whole radio signal $\mathbf{y}_k(t)$ can be transmitted over all of the l_S pairs, and the bandwidth amplification over cable is $\eta = l_S$. In the second case, disjoint fractions of the received bandwidth can be forwarded on each pair and the bandwidth amplification factor is $\eta = 1$.

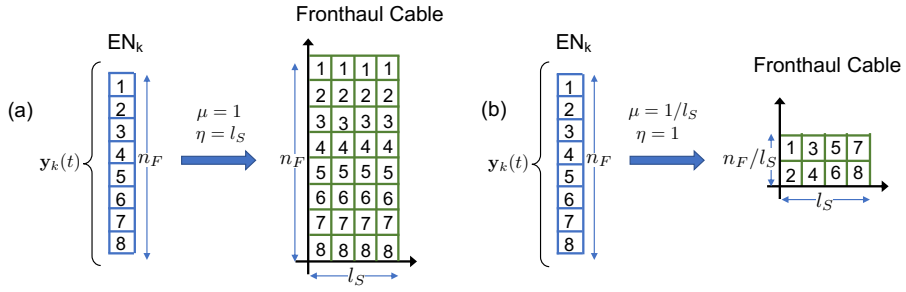

Figure 5. Mapping of radio resources over cable resources: (**a**) maximum normalized cable bandwidth (full redundancy), $\mu = 1$, or $\eta = l_S$; (**b**) minimal normalized cable bandwidth (no redundancy), $\mu = 1/l_S$, or $\eta = 1$.

We now detail the signal model for fronthaul transmission. To this end, let us define the $l_F \times l_S$ matrix $\tilde{\mathbf{Y}}_k$ containing the signal to be transmitted by the EN k-th to the BBU over the copper cable as

$$\tilde{\mathbf{Y}}_k = [\tilde{\mathbf{y}}_k^1(t), \tilde{\mathbf{y}}_k^2(t), ..., \tilde{\mathbf{y}}_k^{l_S}(t)], \tag{9}$$

where the k-th column $\tilde{\mathbf{y}}_k^c(t) \in \mathbb{C}^{l_F \times 1}$ denotes the signal transmitted on twisted-pair c across all the l_F cable frequency resources. The signal $\tilde{\mathbf{R}}_k \in \mathbb{C}^{l_F \times l_S}$ received at the BBU from the k-th EN across all the cable space-frequency resources is then computed as

$$\tilde{\mathbf{R}}_k(t) = \tilde{\mathbf{Y}}_k(t)\mathbf{H}_c + \tilde{\mathbf{W}}_k(t), \tag{10}$$

where $\tilde{\mathbf{W}}_k(t) = [\tilde{\mathbf{w}}_k^1(t), \tilde{\mathbf{w}}_k^2(t), ..., \tilde{\mathbf{w}}_k^{l_S}(t)] \in \mathbb{C}^{l_F \times l_S}$ is the additive white Gaussian cable noise uncorrelated over cable pairs and frequencies, i.e., $\tilde{\mathbf{w}}_k^c(t) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{l_F})$ for all pairs $c = 1, 2, ..., l_S$.

As commonly assumed in wireline communications to control cable radiations [37], the power of the cable symbol $\left[\tilde{\mathbf{y}}_k^c(t)\right]_{f'}$ transmitted from EN$_k$ over twisted-pair c at frequency f' is constrained to P_c by the (short-term: One can also consider the "long-term" power constraint

$n_T^{-1} \sum_t \mathbb{E}[|[\tilde{\mathbf{y}}_k^c(t)]_{f'}|^2] \leq P_c$, $\forall c \in [1, l_S]$, $f' \in [1, l_F]$ with minor modifications to the analysis and final results) power constraint

$$\mathbb{E}\left[\left|[\tilde{\mathbf{y}}_k^c(t)]_{f'}\right|^2\right] \leq P_c \quad \forall c \in [1, l_S], f' \in [1, l_F], t \in [1, n_T]. \tag{11}$$

In the following, we will omit the time index t, when no confusion arises.

2.3. Performance Metrics

The performance metrics used to evaluate the interaction between eMBB and URLLC services in the considered A-RoC-based C-RAN architecture are detailed in the following.

2.3.1. eMBB

Capacity enhancement is the main goal of the eMBB service, which is envisioned to provide very high-rate communication to all the UEs. Therefore, for eMBB UEs, we are interested in the per-UE rate defined as

$$R_B = \frac{\log_2(M_B)}{n_T n_F}, \tag{12}$$

where M_B is the number of codewords in the codebook of each eMBB UE.

2.3.2. URLLC

Differently from eMBB, URLLC service is mainly focused on low-latency and reliability aspects. Due to the short length of URLLC packets, in order to guarantee ultra reliable communications, we need to ensure that the error probability for each URLLC UE, denoted as $\Pr[E_U]$, is bounded by a predefined value ϵ_U (typically smaller than 10^{-3}) as

$$\Pr[E_U] \leq \epsilon_U. \tag{13}$$

Concerning latency, we define the maximum access latency L_U as the maximum number of minislots that an URLLC UE has to wait before transmitting a packet. Finally, although rate enhancement is not one of the goals of URLLC service, it is still important to evaluate the per-UE URLLC rate that can be guaranteed while satisfying the aforementioned latency and reliability constraints. Similarly to (12), the per-UE URLLC rate is defined as

$$R_U = \frac{\log_2(M_U)}{n_F}, \tag{14}$$

where M_U is the number of URLLC codewords in the codebook used by the URLLC UE for each information packet.

3. Analog Fronthaul Signal Processing

The analog fronthaul links employed in the C-RAN system under study pose several challenges in the system design, which are addressed in this section. Firstly, the radio signal received at each EN needs to be mapped over the corresponding fronthaul resources in both frequency and space dimension. Secondly, the signal at the output of each fronthaul link needs to be processed in order to maximize the Signal-to-Noise Ratio (SNR) for all UE signals. Finally, the power constraints in (11) must be properly enforced. All these requirements are to be addressed by all-analog processing in order to meet the low-complexity and latency constraints of the analog fronthaul. In the rest of this section, we discuss each of these problems in turn.

3.1. Radio Resource Mapping over Fronthaul Channels

To maximize the SNRs for all the signals forwarded over the fronthaul by symmetry, we need to ensure that: (i) all the received signals are replicated η times across the cable twisted-pairs, where we recall that η is the bandwidth amplification factor defined in (8); and (ii) cable cross-talk interference among different radio frequency bands is minimized. In fact, as the transmitted power at the cable input is limited by the constraints in (11), a simple and effective way to cope with the impairments of the analog fronthaul links using the only analog-processing capability is by introducing redundancy in the fronthaul transmission. To this end, without loss of generality, we assume the following mapping rule between the n_F radio signals at each EN and the $l_S l_F$ cable resources.

Let us consider the radio signal $\mathbf{y}_k = \left[Y_k^1, Y_k^2, ..., Y_k^{n_F}\right]^T$ received at the k-th EN. For a given normalized cable bandwidth μ, the n_F frequency channels of the radio signal $\mathbf{y}_k$ can be split into $1/\mu$ sub-vectors of size $l_F = \mu n_F$ as

$$\mathbf{y}_k = \begin{bmatrix} \mathbf{y}_k^1 \\ \mathbf{y}_k^2 \\ \vdots \\ \mathbf{y}_k^{\frac{1}{\mu}} \end{bmatrix}, \tag{15}$$

where each vector $\mathbf{y}_k^j$ contains a disjoint fraction of the radio signal bandwidth n_F. Each vector $\mathbf{y}_k^j$ can be transmitted over η twisted-pairs, where we recall that η is the fronthaul redundancy factor. To this end, each signal $\mathbf{y}_k^j$ in (16) is transmitted over η consecutive cable twisted-pairs.

To formalize the described mapping, the first step consists in reorganizing the signal $\mathbf{y}_k$ into a $l_F \times \frac{1}{\mu}$ matrix as

$$\mathbf{Y}_k = \text{vec}_{\frac{1}{\mu}}^{-1}(\mathbf{y}_k) = \begin{bmatrix} \mathbf{y}_k^1 & \mathbf{y}_k^2 & ... & \mathbf{y}_k^{\frac{1}{\mu}} \end{bmatrix}, \tag{16}$$

where the operator $\text{vec}_{\frac{1}{\mu}}^{-1}(\cdot) : \mathbb{C}^{n_F} \to \mathbb{C}^{n_F \cdot \mu \times \frac{1}{\mu}}$ acts as the inverse of the vectorization operator $\text{vec}(\cdot)$, with the subindex $1/\mu$ denoting the number of columns of the resulting matrix. Then, the overall cable signal $\tilde{\mathbf{Y}}_k$ transmitted by EN k-th in (9) can be equivalently written as

$$\tilde{\mathbf{Y}}_k = \left[\underbrace{\mathbf{y}_k^1 \quad \mathbf{y}_k^1 \quad \cdots \quad \mathbf{y}_k^1}_{\eta}, \quad \underbrace{\mathbf{y}_k^2 \quad \mathbf{y}_k^2 \quad \cdots \quad \mathbf{y}_k^2}_{\eta} \quad \cdots \quad \underbrace{\mathbf{y}_k^{\frac{1}{\mu}} \quad \mathbf{y}_k^{\frac{1}{\mu}} \quad \cdots \quad \mathbf{y}_k^{\frac{1}{\mu}}}_{\eta} \right], \tag{17}$$

or, in compact form, as

$$\tilde{\mathbf{Y}}_k = \mathbf{Y}_k \otimes \mathbf{1}_\eta^T, \tag{18}$$

where $\otimes$ denotes the Kronecker product (for a review of Kronecker product properties in signal processing we refer the reader to [38]). Notice that in case of full normalized cable bandwidth, i.e., $\mu = 1$ (corresponding to $\eta = l_S$), the signal $\tilde{\mathbf{Y}}_k$ transmitted over the fronthaul cable simplifies to $\tilde{\mathbf{Y}}_k = \mathbf{y}_k \otimes \mathbf{1}_{l_S}^T$, which implies that the radio signal $\mathbf{y}_k$ is replicated over all the l_S twisted-pairs. On the contrary, when the normalized cable bandwidth is minimal, i.e., $\mu = 1/l_S$ (corresponding to $\eta = 1$), the signal $\tilde{\mathbf{Y}}_k$ does not contain any redundancy, and disjoints signals are transmitted over all pairs, so that the cable signal $\tilde{\mathbf{Y}}_k$ equals the matrix radio signal in (16) as $\tilde{\mathbf{Y}}_k = \mathbf{Y}_k$.

Remark—Practical Implementation Issues. *The easiest practical implementation of the proposed analog radio resource mapping at the EN is by grouping the subcarriers onto a specific frequency portion of the cable, as described in [18–20,22]. As an example, let us assume that the EN is equipped with 5 antennas, that each antenna receives a 20-MHz radio signal, and that the analog fronthauling disposes of 4 links with 100 MHz bandwidth each. In this case, the above references have shown that it is possible to freely map, or to replicate, in an all-analog fashion the 5 × 20 MHz bands onto the overall 4 × 100 MHz = 400 MHz fronthaul bandwidth. This example corresponds to a special case of the model studied in this paper, obtained by setting* $\mu = 1$,

i.e., the whole radio signal bandwidth received at the ENs is mapped/replicated over the analog fronthauling. More generally, this paper posits the possibility to carry out the fronthaul mapping at a finer granularity, i.e., at a subcarrier level. In this case, filtering operations would in practice be mandatory in order to extract groups of subcarriers. This operation can be implemented in principle still by analog filters, whose design is left as future works.

3.2. Signal Combining at the Fronthaul Output

As discussed, depending on the fronthaul bandwidth, a number η of noisy replicas of the radio signals received at each EN are relayed to the BBU over η different twisted-pairs. Hence, in order to maximize the SNRs for all signals, Maximum Ratio Combining (MRC) [39] is applied at the cable output as

$$\mathbf{R}_k = \tilde{\mathbf{R}}_k \mathbf{G}, \tag{19}$$

where $\mathbf{R}_k \in \mathbb{C}^{l_F \times \frac{1}{\mu}}$ is the signal received at BBU from the k-th EN after the combiner and $\mathbf{G} \in \mathbb{R}^{l_S \times \frac{1}{\mu}}$ is the MRC matrix. Under the assumptions here, MRC coincides with equal ratio combining and hence matrix $\mathbf{G}$ can be written as

$$\mathbf{G} = \frac{1}{\eta} \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{1}_\eta \right). \tag{20}$$

As an example, in the case of maximum redundancy, i.e., $\eta = l_S$, the MRC matrix $\mathbf{G} = l_S^{-1} \mathbf{1}_{l_S}$ combines the analog signals received over all pairs, since they carry the same information signal. On the contrary, in the case of minimal normalized bandwidth $\mu = 1/l_S$ or $\eta = 1$, the matrix $\mathbf{G}$ equals the identity matrix as $\mathbf{G} = \mathbf{I}_{l_S}$, since no combining is possible.

The signal $\mathbf{r}_k \in \mathbb{C}^{n_F}$ received at the BBU from EN_k across the n_F subcarriers is thus obtained by vectorizing matrix $\mathbf{R}_k$ in (19) as

$$\mathbf{r}_k = \text{vec}(\mathbf{R}_k). \tag{21}$$

The relationship between the signal $\mathbf{r}_k$ (21) obtained at the output of the combiner and the radio received signal $\mathbf{y}_k$ in (15) is summarized by the block-scheme in Figure 6, and it is

$$\mathbf{r}_k = \text{vec} \left[\left(\left(\text{vec}_{\frac{1}{\mu}}^{-1} (\mathbf{y}_k) \otimes \mathbf{1}_\eta^T \right) \mathbf{H}_c + \tilde{\mathbf{W}}_k \right) \mathbf{G}_k \right]. \tag{22}$$

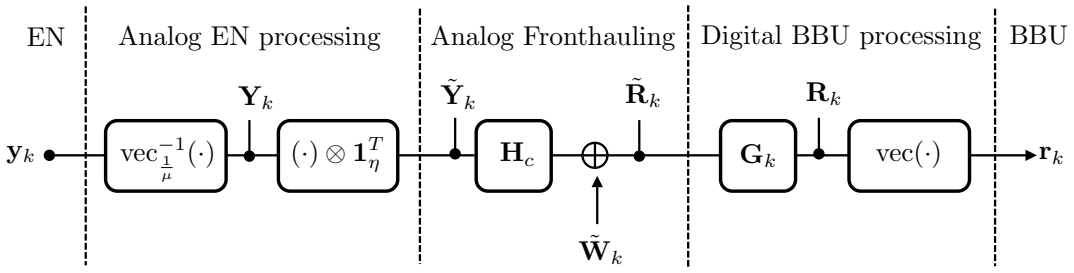

Figure 6. Relationship between the signal $\mathbf{r}_k$ (21) obtained at the output of the combiner and the radio received signal $\mathbf{y}_k$ in (15).

Finally, we collect the overall signal $\mathbf{R} \in \mathbb{C}^{n_F \times M}$ received at the BBU from all ENs across all frequencies in the matrix

$$\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \ ..., \ \mathbf{r}_M]. \tag{23}$$

After some algebraic manipulations, it is possible to express Equation (23) in a more compact form, which is reported in the following Lemma 1.

Lemma 1. *In the given C-RAN architecture with analog fronthaul links, for a given bandwidth amplification factor $\eta \geq 1$, the signal $\mathbf{R} \in \mathbb{C}^{n_F \times M}$ received at the BBU from all ENs across all radio frequencies after MRC can be written as*

$$\mathbf{R} = \left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right) \mathbf{Y} + \mathbf{W}, \tag{24}$$

where

$$\mathbf{H}_c^{\eta} = \gamma\eta \mathbf{1}_{\frac{1}{\mu}} \mathbf{1}_{\frac{1}{\mu}}^T + (1-\gamma)\mathbf{I}_{\frac{1}{\mu}} \tag{25}$$

is the equivalent fronthaul channel matrix; $\mathbf{Y}$ is the signal received at all ENs across all frequency radio channels in (5); and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M]$ is the equivalent cable noise at the BBU after MRC, with the k-th column distributed as $\mathbf{w}_k \sim \mathcal{CN}\left(\mathbf{0}, \frac{1}{\eta}\mathbf{I}_{n_F}\right)$ for all $k = 1, 2, ..., M$.

Proof. see Appendix A. □

To gain some insights, it is useful again to consider the two extreme cases of maximum redundancy, i.e., $\eta = l_S$, and no redundancy, i.e., $\eta = 1$. In the former case, the equivalent channel (25) equals the scalar $\mathbf{H}_c^{\eta} = 1 + \gamma(l_S - 1)$. This demonstrates the effect of transmitting a replica of the whole radio signal over all pairs. In fact, the useful signal is received at the BBU not only through the direct path, which has unit gain, but also from the remaining $l_S - 1$ interfering paths, each with gain γ, which constructively contribute to the overall SNR after the combiner. More precisely, it can be observed that, in case of full redundancy, the SNR of the radio signal $\mathbf{Y}$ at the BBU is increased by the analog fronthaul links by a factor of $(1 + \gamma(l_S - 1))^2/(1/\eta) = l_S(1 + \gamma(l_S - 1))^2$. As a result, in this case, for a coupling factor $\gamma > 0$, the SNR at the BBU increases with the cube of the number of fronthaul links l_S. In contrast, for $\mu = 1/l_S$, the equivalent fronthaul channel reflects the fact that signals forwarded over the different pairs interfere with each other, and is equal to $\mathbf{H}_c^{\eta} = \mathbf{H}_c$. The beneficial effect of redundantly transmitting radio signals over different pairs is reflected also in the power of the noise after the combiner, which is reduced proportionally to the bandwidth amplification factor η.

3.3. Fronthaul Power Constraints

To enforce the cable power constraints in (11), it is necessary to scale the radio signal $\mathbf{Y}$ in (24) by a factor of λ prior to the transmission over the fronthaul. This is given as

$$\lambda = \sqrt{\frac{P_c}{\delta P_B(1 + 2\alpha^2) + 1}}, \tag{26}$$

where δ is equal to $\delta = \left(1 - L_U^{-1}\right)^{-1}$ for OMA, accounting for the fact that only $L_U - 1$ minislots are devoted to the eMBB UE, while it equals $\delta = 1$ for NOMA, since the eMBB transmission spreads over all L_U minislots. To simplify the notation, in the following we will account for the gain λ by scaling the noise over the cable after MRC in Equation (25) accordingly as

$$\mathbf{w}_k(t) \sim \mathcal{CN}\left(\mathbf{0}, \frac{1}{\eta\lambda^2}\mathbf{I}\right). \tag{27}$$

4. Orthogonal Multiple Access (OMA)

As described in Section 2.1, under OMA over the radio channel, one minislot every L_U is exclusively allocated to URLLC UEs, while eMBB UEs transmit over the remaining minislots. In this way, URLLC UEs never interfere with eMBB transmissions. If more than one URLLC packet is generated at a user between two URLLC transmission opportunities, only one of such packets (randomly selected) is transmitted, while the others are discarded, causing a blockage error. Due to the latency constraints, URLLC signals are digitized and decoded locally at the ENs, while the

eMB signals are first mapped over the fronthaul lines, and then analogically forwarded to the BBU, as mathematically summarized in Figure 7. In this section, we derive the expressions for the eMBB and URLLC rates under OMA for a given URLLC access latency L_U and, in the case of URLLC, for a fixed URLLC target error probability ϵ_U.

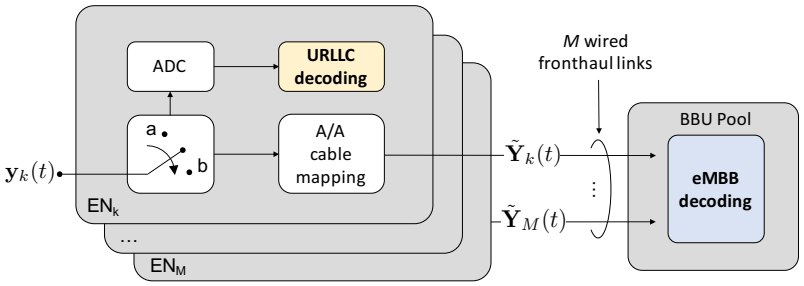

Figure 7. Block diagram of the operation of the ENs and BBU for Orthogonal Multiple Access (OMA). A/A stands for Analog-to-Analog.

4.1. URLLC Rate

To evaluate the per-UE URLLC rate under OMA and for a given URLLC target error probability ϵ_U, we follow the approach in [32], which is reviewed here. URLLC packets are generally short due to the strict latency constraints, and the maximum achievable rate can be computed by leveraging results from finite blocklength information theory. To this end, fix a given blocklength n_F and URLLC error decoding probability ϵ_U^D. Notice that this probability is different from the general URLLC target error probability ϵ_U, as detailed later in this section. According to [40], the URLLC rate can be well approximated by

$$R_U = \log_2(1+\beta^2 P_U) - \sqrt{\frac{V}{n_F}} Q^{-1}(\epsilon_U^D), \tag{28}$$

where

$$V = \frac{\beta^2 P_U}{1+\beta^2 P_U} \tag{29}$$

is the channel dispersion and $\mathcal{Q}^{-1}(\cdot)$ is the inverse Q-function (see Section 1.5).

The error probability for URLLC packets is the sum of two contributions. The first represents the probability that an URLLC packet is discarded due to blockage, given that only one URLLC packet can be transmitted within the required L_U worst-case latency; while the second is the probability that the packet is transmitted but not successfully decoded. Accordingly, the overall error probability can be computed as

$$\Pr[E_U] = \sum_{n}^{L_U-1} p(n)\frac{n}{n+1} + \sum_{n}^{L_U-1} p(n)\frac{1}{n+1}\epsilon_U^D, \tag{30}$$

where $p(n) = \Pr[N_U(L_U) = n]$ is the distribution of the binomial random variable $N_U(L_U) \sim \text{Bin}(L_U - 1, q)$ representing the number of additional packets generated by the URLLC UE during the remaining minislots between two transmission opportunities. The decoding error probability ϵ_U^D in (28) can be obtained from the URLLC reliability constraint in (13), i.e., $\Pr[E_U] = \epsilon_U$.

4.2. eMBB Rate

The eMBB signals received at the ENs are forwarded over the analog fronthaul to the BBU, where centralized digital signal processing and decoding are performed. In the case of OMA, the eMBB signal is free from URLLC interference, hence signal $\mathbf{Y}$ received by all ENs over all radio channels in (5) can be written as

$$\mathbf{Y} = \mathbf{XH} + \mathbf{Z}. \tag{31}$$

By substituting (31) in (24), it is possible to compute the expression for the eMBB per-UE rate under OMA, as shown in Lemma 2. Notice that, unlike the case of URLLC packets, the $n_F n_T$ blocklength of eMBB packets allows for the use of standard asymptotic Shannon theory in the computation of eMBB information rate.

Lemma 2. *In the given C-RAN architecture with analog fronthaul links, for a given bandwidth amplification factor $\eta \geq 1$, the eMBB user rate under OMA is given as*

$$R_B = \mu \frac{1 - L_U^{-1}}{M} log \left(det \left(\mathbf{I} + \bar{P}_B \mathbf{R}_{z_{eq}}^{-1} \mathbf{H}_{eq} \mathbf{H}_{eq}^T \right) \right), \tag{32}$$

where $\bar{P}_B = P_B \left(1 - L_U^{-1}\right)^{-1}$ is the transmission power of eMBB users under OMA, $\mathbf{H}_{eq} = \mathbf{H} \otimes \mathbf{H}_c^{\eta}$ is the overall channel matrix comprising both the radio channel $\mathbf{H}$ and the equivalent cable channel $\mathbf{H}_c^{\eta}$ defined in Lemma 1, and $\mathbf{R}_{z_{eq}} = \mathbf{I}_M \otimes \mathbf{H}_c^{\eta} \mathbf{H}_c^{\eta} + \frac{1}{\lambda^2 \eta} \mathbf{I}_{\frac{M}{\mu}}$ is the overall wireless plus cable noise at the BBU.

Proof. see Appendix B. □

As a first observation, the eMBB rate (32) linearly scales with the normalized bandwidth μ. This shows that a potential performance degradation in terms of spectral efficiency can be incurred in the presence of fronthaul channels with bandwidth limitations, i.e., with $\mu < 1$. This loss is pronounced in the presence of significant inter-channel interference, i.e., for large γ. In fact, a large γ increases the effective noise power as per expression of matrix $\mathbf{R}_{z_{\text{eq}}}$. It is also important to point out that in the considered C-RAN system based on the analog relaying of radio signals, the overall noise at the BBU is no longer white as it accounts both for the white cable noise and the wireless noise, where the latter is correlated when there is some bandwidth redundancy, i.e., when $\mu \geq 1/l_S$ or $\eta > 1$.

5. Non-Orthogonal Multiple Access (NOMA)

In NOMA, URLLC UEs transmit in the same minislot where the packet is generated, and hence the access latency is minimal and limited to $L_U = 1$ minislot. However, the URLLC signals mutually interfere with the eMBB transmission, which spans the whole time-frequency resource plane. Due to URLLC latency constraints, the eMBB signals necessarily need to be treated as noise while decoding URLLC packets at the ENs. On the contrary, several strategies can be adapted in order to deal with the interfering URLLC signal. Beside puncturing, considered for 5G NR standardization [33,34], this work considers two other techniques, namely Treating Interference as Noise (TIN) and Successive Interference Cancellation (SIC), as detailed in the rest of this section.

5.1. URLLC Rate under NOMA

The URLLC per-UE rate for NOMA can be computed by leveraging results from finite blocklength information theory similarly to the OMA case, but accounting for the additional eMBB interference [41]. The URLLC per-UE rate under NOMA is thus well approximated by [32]

$$R_U = \log_2(1 + S_U) - \sqrt{\frac{V}{n_F}} Q^{-1}(\epsilon_U^D),$$

where

$$S_U = \frac{\beta^2 P_U}{1 + (1 + 2\alpha^2) P_B} \tag{33}$$

is the Signal-to-Interference-plus-Noise Ratio (SINR) for the URLLC UE, and the channel dispersion V is given as

$$V = \frac{S_U}{1 + S_U}. \tag{34}$$

Notice that in NOMA the incoming URLLC packet is always transmitted, and hence an URLLC error occurs only if the decoding of such packet fails, which happens with probability ϵ_U^D. This implies that under NOMA, the probability of URLLC error is given by

$$\Pr[E_U] = \epsilon_U^D, \tag{35}$$

hence imposing the condition $\epsilon_U^D \leq \epsilon_U$ by the requirement (13).

5.2. eMBB Rate by Puncturing

To carry out joint decoding at the BBU of the eMBB signals under NOMA, the standard approach is to simply discard at the eMBB decoder those signals that are interfered by URLLC. As shown in Figure 8, this technique, referred to as puncturing, is based on the detection of URLLC transmissions at the BBU: if a URLLC transmission is detected in the signal received from EN_k, such signal is discarded. Otherwise, the interference-free eMBB signals are jointly decoded at the BBU.

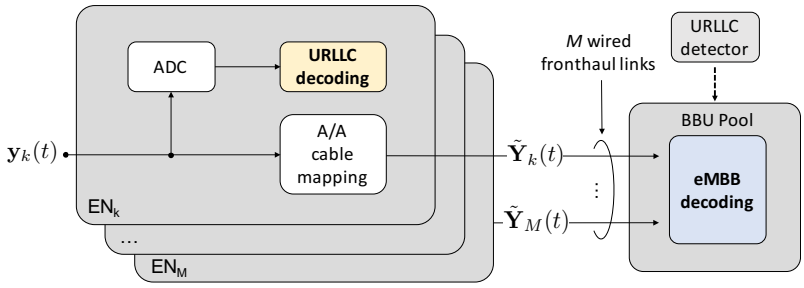

Figure 8. Block diagram of the operation of the ENs and BBU for Non-Orthogonal Multiple Access (NOMA) by puncturing and Treating Interference as Noise (TIN). A/A stands for Analog-to-Analog.

Considering the aforementioned assumptions, the signal model for puncturing can be equivalently described by assuming that if the signal Y_k^f received at the EN_k on frequency f is interfered by an URLLC transmission, then such signal is discarded, and the BBU receives only noise. This is mathematical described by

$$Y_k^f = B_k(X_k^f + \alpha X_{k+1}^f + \alpha X_{k-1}^f) + Z_k^f, \tag{36}$$

where the Bernoulli variable $B_k = 1 - A_k \sim \mathcal{B}(1-q)$ indicates the absence ($B_k = 1$) or presence ($B_k = 0$) of URLLC transmissions in the given minislot. The signal in (36) received across all ENs and frequencies can be written in matrix form as

$$\mathbf{Y} = \mathbf{XHB} + \mathbf{Z}, \tag{37}$$

with definitions given in Section 2 and with $\mathbf{B} = \text{diag}(B_1, B_2, ..., B_M)$. The rate for the eMBB UE under NOMA by puncturing is reported in Lemma 3 and can be derived by substituting signal (37) in Equation (24).

Lemma 3. *In the given C-RAN architecture with analog fronthaul links, for a given bandwidth amplification factor $\eta \geq 1$, the eMBB user rate under NOMA by puncturing yields*

$$R_B = \frac{\mu}{M}\mathbb{E}_{\mathbf{B}}\left[log\left(det\left(\mathbf{I} + P_B\mathbf{R}_{z_{eq}}^{-1}\mathbf{H}_{B,eq}\mathbf{H}_{B,eq}^T\right)\right)\right], \tag{38}$$

where $\mathbf{R}_{z_{eq}}$ is defined as in Lemma 2; $\mathbf{H}_{B,eq} = \mathbf{BH} \otimes \mathbf{H}_c^{\eta}$ is the equivalent wireless plus cable channel in case of puncturing; and we have $\mathbf{B} = diag(B_1, B_2, ..., B_M)$, with B_k being i.i.d. $\mathcal{B}(1-q)$ variables .

Proof. Lemma 3 can be proved by following similar steps as for the proof of Lemma 2 with two minor differences: *(i)* the radio channel matrix $\mathbf{H}$ is right multiplied by the random matrix $\mathbf{B}$ and *(ii)* the capacity is computed by averaging over the distribution of $\mathbf{B}$. □

Differently from the rate (32) achieved by OMA, under NOMA, the eMBB transmission spreads over all the minislots, so that there is no scaling factor $1 - L_U^{-1}$ in front of the rate expression (38) to account for the resulting loss in spectral efficiency. In case of NOMA by puncturing, the noise covariance is exactly as the one in the eMBB OMA rate in Equation (32), since the eMBB signal, if not discarded at the BBU, is guaranteed to be URLLC interference-free. The overall rate is computed by averaging over all the possible realizations of the random matrix $\mathbf{B}$, which left-multiplies the radio channel matrix $\mathbf{H}$ and accounts for the probability that the entire signal is discarded due to an incoming URLLC packet.

In the case of C-RAN with digital limited-capacity fronthaul links, as discussed in [32], it is advantageous to carry out the operation of detecting and, eventually, discarding the eMBB signal at the ENs. In fact, with a digital fronthaul, only the undiscarded minislots can be quantized, hence devoting the limited fronthaul resources to increase the resolution of interference-free eMBB samples [32]. The same does not apply to the analog fronthaul considered here, as signals are directly relayed to the BBU without any digitization.

5.3. eMBB Rate by Treating Interference as Noise

In the case of analog fronthaul, an enhanced strategy to jointly decode the eMBB signals under NOMA at the BBU is to treat the URLLC interfering transmissions as noise at the eMBB decoder, instead of discarding the corresponding minislot as in puncturing. The block diagram is the same as for puncturing and shown in Figure 8. Accordingly, based on the signals received over the fronthaul links, the BBU first detects the presence of URLLC transmission so as to properly select the decoding metric. Then, based on this knowledge, joint decoding is performed by TIN.

Lemma 4. *In the given C-RAN architecture with analog fronthaul links, for a given bandwidth amplification factor $\eta \geq 1$, the eMBB user rate under NOMA by treating URLLC interference as noise yields*

$$R_B = \frac{\mu}{M}\mathbb{E}_{\mathbf{A}}\left[log\left(det\left(\mathbf{I} + P_B\mathbf{R}_{A,z_{eq}}^{-1}\mathbf{H}_{eq}\mathbf{H}_{eq}^{T}\right)\right)\right], \tag{39}$$

where $\mathbf{R}_{A,z_{eq}} = \mathbf{R}_{z_{eq}} + \beta^2 P_U\left(\mathbf{A} \otimes \mathbf{H}_c^{\eta}\mathbf{H}_c^{\eta}\right)$ is the overall noise plus URLLC interference at the BBU; matrix $\mathbf{A}$ is as in (5); and matrices $\mathbf{R}_{z_{eq}}$ and $\mathbf{H}_{eq}$ are the same as in Lemma 2.

Proof. see Appendix C. □

Differently from the two previous cases, in the case of NOMA under TIN, the noise covariance matrix $\mathbf{R}_{A,z_{\text{eq}}}$ needs to account also for the interfering URLLC transmissions, whose packet arrival probability is described by matrix $\mathbf{A}$. The achievable rate is then computed by taking the average over the random matrix $\mathbf{A}$. This average reflects the long-blocklength transmissions of the eMBB users.

5.4. eMBB Rate by Successive Interference Cancellation

Finally, a more complex receiver architecture can be considered at the BBU, whereby interference is cancelled out from the useful signal. This technique, referred to as Successive Interference Cancellation (SIC), is based on the idea that, if an URLLC signal is successfully decoded at the EN_k, it can be cancelled from the overall received signal $\mathbf{y}_k$ prior to the relaying over the cable, so that an ideally interference-free eMBB signal is forwarded to the BBU. We also assume that, if the URLLC signal is not successfully decoded, signal $\mathbf{y}_k$ is discarded.

As a practical note, SIC must be performed in the analog domain, thus complicating the system design. Practical complications are not considered in the analysis here. As shown in Figure 9, if the

URLLC signal is successfully decoded at EN_k, this needs first to be Digital-to-Analog Converted (DAC) and then cancelled from the analog signal $\mathbf{y}_k$. Therefore, signal $\mathbf{y}_k$ needs to be suitably delayed in order to wait for the cascade of ADC, decoding, and DAC operations to be completed at the URLLC decoder. Being latency not an issue for eMBB traffic, in this work we assume to employ ideal ADC/DAC, so that the delay in Figure 9 is assumed as ideally zero.

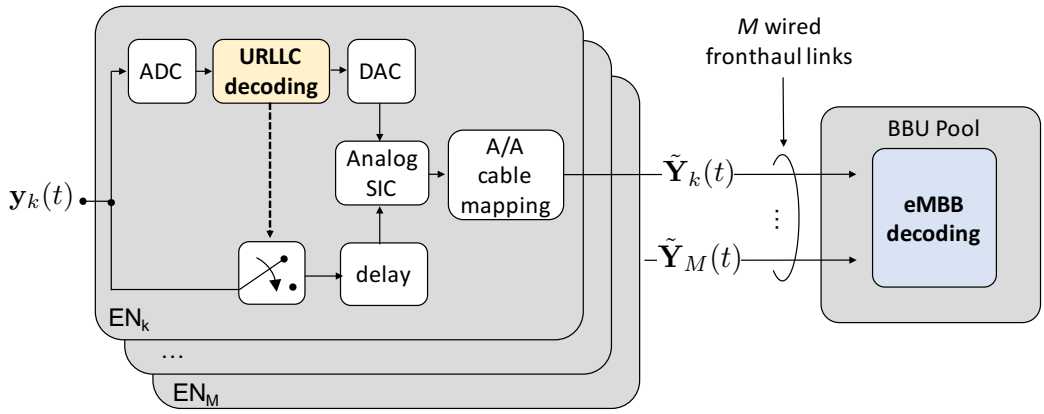

Figure 9. Block diagram of the operation of the ENs and BBU for Non-Orthogonal Multiple Access (NOMA) by Successive Interference Cancellation (SIC). A/A stands for Analog-to-Analog.

To account for imperfect SIC, the amplitude of the residual URLLC interference on eMBB signal is assumed to be proportional to a factor $\rho \in [0, 1]$. Accordingly, perfect SIC corresponds to $\rho = 0$, and no SIC to $\rho = 1$.

The signal received at the BBU from EN_k can be thus written as

$$Y_k^f = (1 - A_k E_k)(X_k^f + \alpha X_{k+1}^f + \alpha X_{k-1}^f) + \rho\beta A_k (1 - E_k) U_k^f + Z_k^f, \tag{40}$$

where the Bernoulli variable $E_k \sim \mathcal{B}(q\epsilon_U^D)$ indicates whether there has been an error in decoding the URLLC packet (i.e., $E_k = 1$), or it has been successfully decoded (i.e., $E_k = 0$), and A_k is the same as above. It is easy to show that the factor $(1 - A_k E_k)$ multiplying the eMBB signal indicates that the eMBB signal (40) is discarded only when the two following events simultaneously happen: (i) there is a URLLC transmission (i.e., $A_k = 1$, whose probability is q), and (ii) such URLLC transmission is not successfully decoded ($E_k = 1$, whose probability is ϵ_U^D). In turn, the factor $A_k(1 - E_k)$ multiplying the URLLC signal implies that, if there is a URLLC transmission (i.e., $A_k = 1$) and such transmission is successfully decoded at the EN_k (i.e., $E_k = 0$), then the URLLC signal is mitigated by analog SIC so that only a ρ-fraction of it is forwarded to the BBU and impairs the eMBB transmission.

The signal in (40) received across all ENs and frequencies in the case of NOMA by SIC can be equivalently written in matrix form as

$$\mathbf{Y} = \mathbf{XH}\,(\mathbf{I} - \mathbf{AE}) + \rho\beta\mathbf{UA}\,(\mathbf{I} - \mathbf{E}) + \mathbf{Z}, \tag{41}$$

where $\mathbf{E} = \text{diag}(E_1, E_2, ..., E_M)$. The eMBB UE rate for NOMA by SIC can thus be computed by substituting signal (41) in (24) and the final result is in Lemma 5.

Lemma 5. *In the given C-RAN architecture with analog fronthaul links, for a given bandwidth amplification factor $\eta \geq 1$, the eMBB user rate under NOMA by SIC yields*

$$R_B = \frac{\mu}{M}\mathbb{E}_{\mathbf{A},\mathbf{E}}\left[log\left(det\left(\mathbf{I} + P_B \mathbf{R}_{AE,\tilde{z}_{eq}}^{-1}\mathbf{H}_{AE}\mathbf{H}_{AE}^T\right)\right)\right], \tag{42}$$

where $\mathbf{H}_{AE} = ((\mathbf{I} - \mathbf{AE})\mathbf{H}) \otimes \mathbf{H}_c^{\eta}$ *is the equivalent wireless plus cable channel in case of SIC;* $\mathbf{E} = diag(E_1, E_2, ..., E_M)$ *is a diagonal matrix whose k-th entry* $E_k \sim \mathcal{B}(q\epsilon_U^D)$ *accounts for the probability that the URLLC signal is not successfully decoded at* EN_k*; and* $\mathbf{R}_{AE,z_{eq}} = \mathbf{R}_{z_{eq}} + \rho^2\beta^2 P_U \left((\mathbf{A}(\mathbf{I} - \mathbf{E})) \otimes \mathbf{H}_c^{\eta}\mathbf{H}_c^{\eta}\right)$ *is the overall noise plus residual URLLC interference.*

Proof. Lemma 5 can be proved by following similar steps as for the proofs of the previous Lemmas. □

SIC describes a more complex ENs architecture in which the URLLC signals, if successfully decoded, are successively canceled from the eMBB signals at the EN. However, in the case of imperfect interference cancellation, i.e., $\rho > 0$, the eMBB signal is still impaired by some residual URLLC interference, which is accounted for by the overall noise covariance $\mathbf{R}_{AE,z_{\text{eq}}}$ in (42), similarly to TIN. The URLLC arrival probability and the probability of successful decoding of the URLLC packets are reflected by random matrices $\mathbf{A}$ and $\mathbf{E}$, respectively.

6. Numerical Results

Numerical results based on the previous theoretical discussion are shown in this section with the aim of providing some useful intuitions about the performance of C-RAN systems based on analog RoC in the presence of both URLLC and eMBB services. Unless otherwise stated, we consider the following settings: $M = 6$ ENs, $n_F = 60$ subcarriers (This choice is motivated by the fact that, while still resembling the properties of URLLC short-packet transmissions, $n_F = 60$ is a sufficiently long packet size to ensure tight lower and upper bounds for the limited-blocklength channel capacity [5]), $P_B = 7$ dB, $P_U = 10$ dB, URLLC channel gain $\beta^2 = 1$, $P_C = 7$ dB, $l_S = 4$, and, conventionally, $\epsilon_U = 10^{-3}$. In the case of OMA, the worst-case access latency for URLLC users is set to $L_U = 2$ minislots.

Figure 10 shows URLLC and eMBB per-UE rates for both OMA and NOMA by varying the fronthaul crosstalk interference power γ^2. We consider two values for the normalized bandwidth μ of each copper cable, namely $\mu = 1/l_S = 1/4$ and $\mu = 1$. Please note that the first value corresponds to the minimal bandwidth, while the latter enables each twisted-pair to carry the whole signal bandwidth. For reference, the eMBB rates obtained in the case of ideal fronthaul are shown for both OMA and NOMA. The inter-cell interference power is set to $\alpha^2 = 0.2$, and the URLLC arrival probability to $q = 10^{-3}$. For NOMA, we consider here only puncturing.

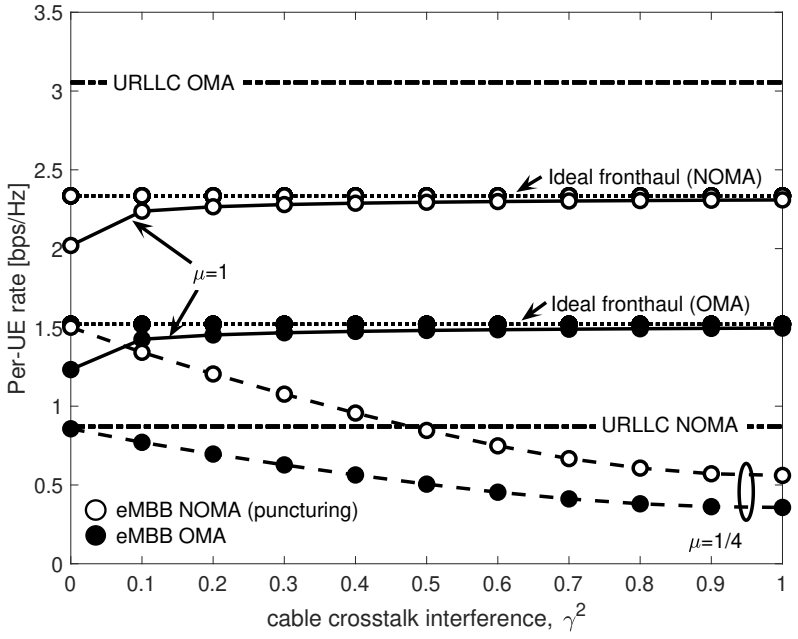

Figure 10. URLLC and eMBB per-UE rates as a function of fronthaul crosstalk interference power γ^2 for OMA and NOMA with puncturing.

The URLLC rates do not depend on cable crosstalk interference γ^2, since URLLC packets are decoded at the EN and thus never forwarded to the BBU over the fronthaul. Furthermore, with NOMA, the access latency of URLLC is minimum, i.e., $L_U = 1$, while for OMA it equals $L_U = 2$. However, Figure 10 shows that the price to pay for this reduced latency is in terms of transmission rate, which is lower than in the OMA case. On the contrary, in the case of eMBB, NOMA allows for a communication at higher rates than those achieved by OMA, thanks to the larger available bandwidth when q is small enough.

On the subject of eMBB rates, it is interesting to discuss the interplay between the normalized cable bandwidth μ and crosstalk power γ^2. For $\mu = 1$, the same signal is transmitted over all the $l_S = 4$ fronthaul twisted-pairs, and hence the four spatial paths sum coherently over the cable, thus turning crosstalk into a benefit. Hence, the eMBB rates under both OMA and NOMA increase with γ^2 and ultimately converge to those achieved over the ideal fronthaul (In this work, we consider the same interference gain γ for all fronthaul links. In practice, the performance boost shown in Figure 10 for increasing γ and for $\mu = 1$ would still be present, albeit to a different extent dependent on the channel realization, even if considering complex channel gain. However, this would require the use of more complex precoding techniques, such as Tomlinson-Harashima [37], which require an estimate of the fronthaul channel. Since for wired fronthauling the channel is nearly static and time-invariant, channel state information can be easily obtained [42]). For $\mu = 1/4$ instead, disjoint portions of the radio signal are transmitted over different interfering twisted-pairs, and the performance progressively decrease with γ^2. The leftmost portion of Figure 10 suggests that for mild cable interference, even when the cable bandwidth is small (i.e., $\mu = 1/4$), it is still possible to provide communication with acceptable performance degradation, i.e., with a ≈ 1 bps/Hz loss from the ideal fronthaul case for NOMA, and an even smaller loss for OMA. However, when the cable crosstalk increases, the rate degradation is severe, and fair performance are achieved only if the cable bandwidth is large enough to accommodate the redundant transmission of radio signals over all pairs, i.e., $\mu = 1$.

Figure 11 shows URLLC and eMBB rates as a function of the URLLC packet arrival probability q. eMBB rates under OMA are compared with those achieved by NOMA under puncturing, TIN and SIC. We consider here full cable bandwidth availability $\mu = 1$ and $\gamma^2 = 1$, so that, as in Figure 10, the rates achieved for both OMA and NOMA for low q (say, $q < 10^{-2}$) coincide with those achieved over the ideal fronthaul. Inter-cell interference is set to $\alpha^2 = 0.2$. As noted in [32], under OMA, when q increases, the probability of an URLLC packet to be dropped due to blockage becomes very high, preventing URLLC transmission from meeting the strict reliability constraints, and results in a vanishing URLLC rate. This is unlike in NOMA, whereby the URLLC rate is not affected by q, and the access latency is minimal, i.e., $L_U = 1$. For eMBB under NOMA, TIN always outperforms puncturing. This is because TIN does not discard any received minislot, thus contributing to the overall eMBB rate. The result is in contrast with the conventional digital capacity-constrained fronthaul considered in [32]. In fact, in the latter, for sufficient low q, it is preferable not to waste fronthaul capacity resources by quantizing samples received in minislots affected by URLLC interference in order to increase the resolution of the interference-free samples. Additional gains are achieved by SIC, which takes advantages of the high reliability, and thus high probability to be cancelled, of the URLLC signal at the EN.

Implementing SIC in a fully analog fashion is practically not trivial, and there is generally some residual URLLC interference. The effect of residual interference on the achievable eMBB rate is investigated in Figure 12 for $q = 0.3$, $\alpha^2 = 0.4$, $\gamma^2 = 0.5$, $\mu = 1$ and for different power of URLLC UE P_U. Once again, in case of perfect interference cancellation, i.e., $\rho = 0$, SIC approaches the ideal fronthaul performance, while for more severe values of the residual interference power ρ, the eMBB performance progressively decreases. Nevertheless, even in the worst-case SIC scenario, i.e., $\rho = 1$, the achievable rates are never worse than those achieved by TIN irrespective of the value of P_U. This is once again due to the high reliability of URLLC transmission. It is in fact easy to prove that for $\rho = 1$ and low values of ϵ_U^D, the SIC eMBB rate in (42) converges to the one of TIN in (39).

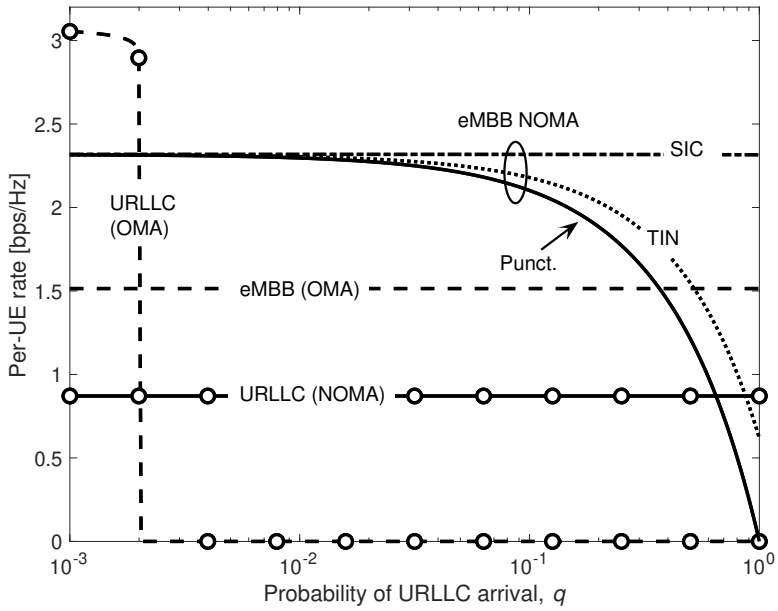

Figure 11. URLLC and eMBB rates vs. probability of URLLC arrival q for OMA and NOMA by puncturing, treating interference as noise (TIN), and successive interference cancellation (SIC).

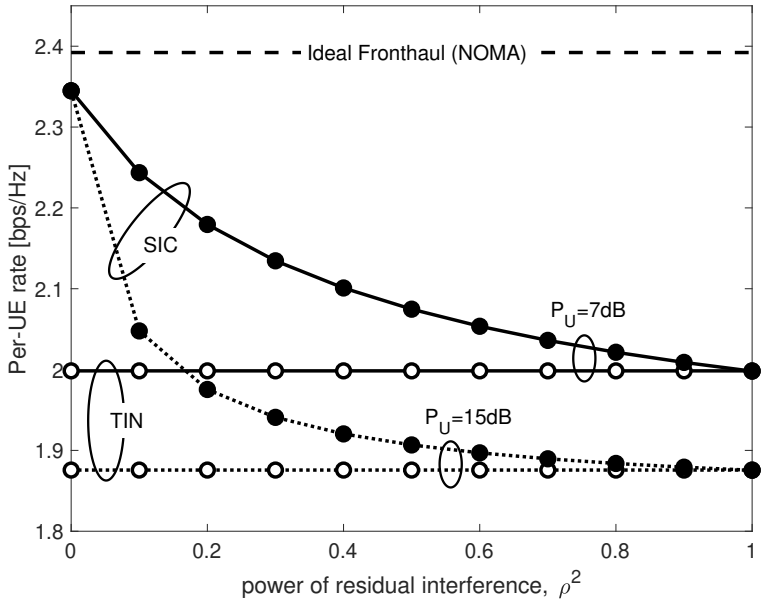

Figure 12. eMBB rates for NOMA by SIC vs. power of residual URLLC interference ρ^2.

For completeness, Figure 13 shows the trade-off between eMBB and URLLC per-UE rates as a function of the access latency L_U for OMA and NOMA with puncturing, and considering $q = 10^{-3}$, $\alpha^2 = 0.2$ and $\gamma^2 = 0.5$. The behavior of the RoC-based C-RAN system versus the access latency L_U is similar to the one observed for digital capacity-constrained fronthaul [32] for both $\mu = 1$ and $\mu = 1/4$. While under OMA it is not possible to achieve a non-zero URLLC rate at even relatively low access latency such as $L_U > 3$, NOMA provides a reliable communication with constant minimal $L_U = 1$ access latency, but with lower rate. For eMBB, NOMA achieves an higher per-UE rate regardless of the value of the normalized bandwidth μ.

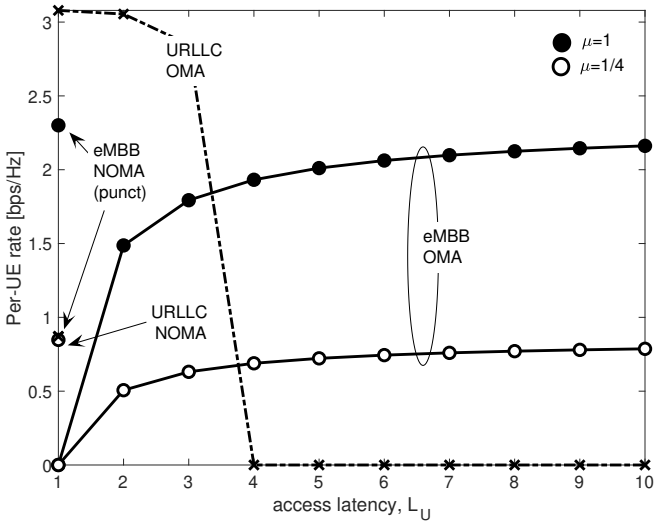

Figure 13. URLLC and eMBB per-UE rates as a function of access latency L_U for OMA and NOMA with puncturing.

7. Conclusions

This paper considers the coexistence of eMBB and URLLC services in the uplink of an analog C-RAN architecture from an information theoretic perspective. The rate expressions for URLLC and eMBB users under Orthogonal and Non-Orthogonal Multiple Access (OMA and NOMA, respectively) have been derived considering Analog Radio-over-Copper (A-RoC) as a sample scenario, although the proposed model can be easily adapted to other analog fronthaul technologies. For eMBB signals, performance have been evaluated in terms of information rate, while for URLLC we also took into account worst-case access latency and reliability. In case of NOMA, different decoding strategies have been considered in order to mitigate the impact of URLLC transmission on eMBB information rate. In particular, the performance achieved by puncturing, considered for 5G standardization, have been compared with those achieved by Treating URLLC Interference as Noise (TIN), and by Successive URLLC Interference Cancellation (SIC).

The analysis showed that NOMA allows for higher eMBB information rates with respect to OMA, while guaranteeing a reliable low-rate URLLC communication with minimal access latency. Furthermore, numerical results demonstrated that, differently from the digital C-RAN architecture based on limited-capacity fronthaul links, for analog C-RAN, TIN always outperforms puncturing, and SIC achieves the best performance at the price of an higher decoder complexity.

As work in progress, the theoretical model can be extended to account for fading channels or geometric mmWave-link channel models.

Similarly, a frequency-dependent cable channel can be considered by making the cable crosstalk coefficient γ increase with cable frequency [22]. Another interesting research direction is to consider the case in which the BBU has no knowledge about the incoming signal, i.e., it is not able to detect the URLLC transmissions, so that it is impossible for the BBU to choose the proper metric for joint signal decoding [43,44]. Finally, the overall system can be extended to the case of multiple users per-cell, where both ENs and users are equipped with multiple antennas.

Author Contributions: Conceptualization, A.M., O.S. and U.S.; Formal analysis, A.M., O.S. and U.S.; Methodology, A.M.; Software, A.M.; Supervision, O.S. and U.S.; Validation, R.K.; Writing—original draft, A.M.; Writing—review & editing, A.M., R.K., O.S. and U.S.

Funding: R.K. and O.S. have received funding from the European Research Council (ERC) under the European Union Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Lemma 1

The proof of Lemma 1 is structured in two main steps detailed in the following: (i) MRC at the cable output, (ii) cable signal vectorization.

Appendix A.1. Maximum Ratio Combining at the Cable Output

The signal $\mathbf{R}_k \in \mathbb{C}^{l_F \times \frac{1}{\mu}}$ after the MRC at the cable output is

$$\begin{aligned} \mathbf{R}_k &= \tilde{\mathbf{R}}_k \mathbf{G} \\ &= \left(\tilde{\mathbf{Y}}_k \mathbf{H}_c + \tilde{\mathbf{W}}_k\right) \mathbf{G}, \end{aligned} \tag{A1}$$

where the definitions of $\tilde{\mathbf{R}}_k$, $\mathbf{G}$, $\tilde{\mathbf{Y}}_k$ and $\mathbf{H}_c$ follow from and (10), (20), (18) and (6), respectively. Thus, Equation (A1) can be rewritten as

$$\begin{aligned} \mathbf{R}_k &= \left(\mathbf{Y}_k \otimes \mathbf{1}_\eta^T\right) \left(\gamma \mathbf{1}_{l_S} \mathbf{1}_{l_S}^T + (1-\gamma)\mathbf{I}_{l_S}\right) \mathbf{G} + \tilde{\mathbf{W}}_k \mathbf{G} \\ &= \underbrace{\gamma \left(\mathbf{Y}_k \otimes \mathbf{1}_\eta^T\right) \mathbf{1}_{l_S} \mathbf{1}_{l_S}^T \mathbf{G}}_{\mathbf{R}_k'} + \underbrace{(1-\gamma) \left(\mathbf{Y}_k \otimes \mathbf{1}_\eta^T\right) \mathbf{G}}_{\mathbf{R}_k''} + \underbrace{\tilde{\mathbf{W}}_k \mathbf{G}}_{\mathbf{W}_k}, \end{aligned} \tag{A2}$$

where $\mathbf{W}_k = \tilde{\mathbf{W}}_k \mathbf{G}$ is the noise post MRC. The first term in (A2) can be rewritten as

$$\begin{aligned} \mathbf{R}_k' &= \frac{\gamma}{\eta} \left(\mathbf{Y}_k \otimes \mathbf{1}_\eta^T\right) \mathbf{1}_{l_S} \mathbf{1}_{l_S}^T \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{1}_\eta\right) \\ &\overset{(a)}{=} \gamma \left(\mathbf{Y}_k \otimes \mathbf{1}_\eta^T\right) \mathbf{1}_{l_S} \mathbf{1}_{\frac{1}{\mu}}^T \\ &\overset{(b)}{=} \gamma\eta \mathbf{Y}_k \mathbf{1}_{\frac{1}{\mu}} \mathbf{1}_{\frac{1}{\mu}}^T, \end{aligned} \tag{A3}$$

where $\overset{(a)}{=}$ comes from the fact that $\mathbf{1}_{l_S}^T \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{1}_\eta\right) = \left(\mathbf{1}_{\frac{1}{\mu}}^T \otimes \mathbf{1}_\eta^T\right) \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{1}_\eta\right) = \eta \mathbf{1}_{\frac{1}{\mu}}^T$ due to the mixed-product property of Kronecker product operator $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ (see [38], Chapter 2), and, similarly, $\overset{(b)}{=}$ is obtained by $\left(\mathbf{Y}_k \otimes \mathbf{1}_\eta^T\right) \mathbf{1}_{l_S} \mathbf{1}_{\frac{1}{\mu}}^T = \left(\mathbf{Y}_k \otimes \mathbf{1}_\eta^T\right) \left(\mathbf{1}_{\frac{1}{\mu}} \mathbf{1}_{\frac{1}{\mu}}^T \otimes \mathbf{1}_\eta\right) = \eta \mathbf{Y}_k \mathbf{1}_{\frac{1}{\mu}} \mathbf{1}_{\frac{1}{\mu}}^T$. Using similar arguments, the second term in (A2) simplifies to

$$\begin{aligned} \mathbf{R}_k'' &= \frac{1-\gamma}{\eta} \left(\mathbf{Y}_k \otimes \mathbf{1}_\eta^T\right) \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{1}_\eta\right) \\ &= (1-\gamma)\mathbf{Y}_k. \end{aligned} \tag{A4}$$

Finally, substituting (A3) and (A4) in (A2) we obtain

$$\mathbf{R}_k = \mathbf{Y}_k \mathbf{H}_c^\eta + \mathbf{W}_k, \tag{A5}$$

where

$$\mathbf{H}_c^\eta = \gamma\eta \mathbf{1}_{\frac{1}{\mu}} \mathbf{1}_{\frac{1}{\mu}}^T + (1-\gamma)\mathbf{I}_{\frac{1}{\mu}} \tag{A6}$$

is the equivalent cable channel matrix accounting for the bandwidth amplification factor over cable η, and $\mathbf{Y}_k$ is the radio signal reorganized in matrix form as in (16).

Appendix A.2. Cable Signal Vectorization

The vector signal $\mathbf{r}_k \in \mathbb{C}^{n_F \times 1}$ received at the BBU from the k-th EN over all the radio frequency channels can be obtained by vectorizing matrix $\mathbf{R}_k$ in (A5) as

$$\begin{aligned} \mathbf{r}_k &= \mathrm{vec}(\mathbf{Y}_k \mathbf{H}_c^{\eta} + \mathbf{W}_k) \\ &= \left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right) \mathrm{vec}(\mathbf{Y}_k) + \mathrm{vec}(\mathbf{W}_k) \\ &= \left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right) \mathbf{y}_k + \mathbf{w}_k, \end{aligned} \tag{A7}$$

where $\mathbf{y}_k$ is the radio signal received at EN k-th. The overall cable noise vector $\mathbf{w}_k$ can be rewritten as

$$\begin{aligned} \mathbf{w}_k &= \mathrm{vec}(\mathbf{W}_k) \\ &= \mathrm{vec}(\tilde{\mathbf{W}}_k \mathbf{G}) \\ &= \left(\mathbf{G}^T \otimes \mathbf{I}_{l_F}\right) \mathrm{vec}(\tilde{\mathbf{W}}_k) \\ &= \frac{1}{\eta} \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{1}_{\eta}^T \otimes \mathbf{I}_{l_F}\right) \tilde{\mathbf{w}}_k. \end{aligned} \tag{A8}$$

It is important to notice that since the cable noise $\tilde{\mathbf{W}}_k$ is white Gaussian and uncorrelated over cable pairs (see (10)), $\tilde{\mathbf{w}}_k = \mathrm{vec}(\tilde{\mathbf{W}}_k)$ is also white Gaussian distributed as $\tilde{\mathbf{w}}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{l_S l_F})$. Hence, the covariance $\mathbf{R}_w$ of the overall cable noise vector $\mathbf{w}_k$ yields

$$\begin{aligned} \mathbf{R}_w &= \mathbb{E}\left[\mathbf{w}_k \mathbf{w}_k^H\right] \\ &= \frac{1}{\eta^2} \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{1}_{\eta}^T \otimes \mathbf{I}_{l_F}\right) \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{1}_{\eta}^T \otimes \mathbf{I}_{l_F}\right)^T \\ &\overset{(a)}{=} \frac{1}{\eta} \left(\mathbf{I}_{\frac{1}{\mu}} \otimes \mathbf{I}_{l_F}\right) \\ &= \frac{1}{\eta} \mathbf{I}_{n_F}, \end{aligned} \tag{A9}$$

where the equality $^{(a)}$ comes again from the mixed-product property of Kronecker product. Equation (A9) shows that the MRC allows to take advantages from the signal redundancy over the cable, which results in a reduction of the cable noise power by a factor of η.

The proof is completed by gathering the signals $\mathbf{r}_k$ (A7) received at the BBU from all ENs as

$$\begin{aligned} \mathbf{R} &= [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_M] \\ &= \left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right) [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M] + [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M] \\ &= \left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right) \mathbf{Y} + \mathbf{W}, \end{aligned} \tag{A10}$$

where $\mathbf{Y}$ is the signal received by all ENs over all radio channels in (5).

Appendix B. Proof of Lemma 2

By substituting signal $\mathbf{Y}$ in (31) received by all ENs over all radio channels in case of OMA in Equation (24), we obtain

$$\mathbf{R} = \left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right) \mathbf{X}\mathbf{H} + \left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right) \mathbf{Z} + \mathbf{W}. \tag{A11}$$

To compute the per-UE eMBB rate, a further vectorization is needed, leading to

$$\begin{aligned}\mathbf{r} &= \mathrm{vec}(\mathbf{R}) \\ &= \mathrm{vec}\left(\left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right)\mathbf{X}\mathbf{H}\right) + \mathrm{vec}\left(\left(\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right)\mathbf{Z}\right) + \mathrm{vec}(\mathbf{W}) \\ &= \left(\mathbf{H}^T \otimes \mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right)\mathbf{x} + \left(\mathbf{I}_M \otimes \mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right)\mathbf{z} + \mathbf{w} \\ &= \bar{\mathbf{H}}_{\mathrm{eq}}\mathbf{x} + \bar{\mathbf{z}}_{\mathrm{eq}},\end{aligned} \tag{A12}$$

where $\bar{\mathbf{H}}_{\mathrm{eq}} = \mathbf{H}^T \otimes \mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}$ is the overall equivalent channel comprising both cable and radio channels over all ENs, and $\bar{\mathbf{z}}_{\mathrm{eq}} = \left(\mathbf{I}_M \otimes \mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right)\mathbf{z} + \mathbf{w}$ is the overall noise vector comprising both the vectorized radio noise $\mathbf{z} = \mathrm{vec}(\mathbf{Z}) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{n_F M})$ and the vectorized cable noise $\mathbf{w} = \mathrm{vec}(\mathbf{W}) \sim \mathcal{CN}(\mathbf{0}, \frac{1}{\lambda^2\eta}\mathbf{I}_{n_F M})$, where we recall that the scaling λ is due to the cable power constraints. Hence, the covariance of the equivalent noise $\bar{\mathbf{z}}_{\mathrm{eq}}$ yields

$$\begin{aligned}\bar{\mathbf{R}}_{z_{\mathrm{eq}}} &= \mathbb{E}\left[\bar{\mathbf{z}}_{\mathrm{eq}}\bar{\mathbf{z}}_{\mathrm{eq}}^H\right] \\ &= \left(\mathbf{I}_M \otimes \mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right)\left(\mathbf{I}_M \otimes \mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}\right)^H + \frac{1}{\lambda^2\eta}\mathbf{I}_{n_F M} \\ &= \mathbf{I}_M \otimes \mathbf{H}_c^{\eta}\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F} + \frac{1}{\lambda^2\eta}\mathbf{I}_{n_F M}.\end{aligned} \tag{A13}$$

The eMBB per-UE rate under OMA is computed by

$$\begin{aligned}R_B &= \frac{(1 - L_U^{-1})}{n_F M} I(\mathbf{r}, \mathbf{x}) \\ &= \frac{(1 - L_U^{-1})}{n_F M} \log\left(\det\left(\mathbf{I} + \bar{P}_B \bar{\mathbf{R}}_{\bar{z}_{\mathrm{eq}}}^{-1} \bar{\mathbf{H}}_{\mathrm{eq}} \bar{\mathbf{H}}_{\mathrm{eq}}^T\right)\right),\end{aligned} \tag{A14}$$

where $\bar{\mathbf{H}}_{\mathrm{eq}}\bar{\mathbf{H}}_{\mathrm{eq}}^T = \mathbf{H}\mathbf{H} \otimes \mathbf{H}_c^{\eta}\mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}$. Finally, using the determinant property of the Kronecker product $|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^m |\mathbf{B}|^n$, Equation (A14) simplifies to

$$\begin{aligned}R_B &= \frac{l_F(1 - L_U^{-1})}{n_F M} \log\left(\det\left(\mathbf{I} + \bar{P}_B \mathbf{R}_{\bar{z}_{\mathrm{eq}}}^{-1} \mathbf{H}_{\mathrm{eq}} \mathbf{H}_{\mathrm{eq}}^T\right)\right) \\ &= \mu \frac{1 - L_U^{-1}}{M} \log\left(\det\left(\mathbf{I} + \bar{P}_B \mathbf{R}_{z_{\mathrm{eq}}}^{-1} \mathbf{H}_{\mathrm{eq}} \mathbf{H}_{\mathrm{eq}}^T\right)\right),\end{aligned} \tag{A15}$$

where $\mathbf{H}_{\mathrm{eq}} = \mathbf{H} \otimes \mathbf{H}_c^{\eta}$ and $\mathbf{R}_{z_{\mathrm{eq}}} = \mathbf{I}_M \otimes \mathbf{H}_c^{\eta}\mathbf{H}_c^{\eta} + \frac{1}{\lambda^2\eta}\mathbf{I}_{\frac{M}{\mu}}$, thus concluding the proof.

Appendix C. Proof of Lemma 3

By substituting the signal (5) received at the k-th EN under NOMA into (A10) and by following similar steps as for the proof of Lemma 2, the overall vector signal received at the BBU comprising both cable and radio channels over all ENs under NOMA by treating interference as noise yields

$$\mathbf{r} = \bar{\mathbf{H}}_{\mathrm{eq}}\mathbf{x} + \beta\bar{\mathbf{A}}_{\mathrm{eq}}\mathbf{u} + \bar{\mathbf{z}}_{\mathrm{eq}}, \tag{A16}$$

where the definitions of $\bar{\mathbf{H}}_{\mathrm{eq}}$ and $\bar{\mathbf{z}}_{\mathrm{eq}}$ are the same as in (A12), $\bar{\mathbf{A}}_{\mathrm{eq}} = \mathbf{A} \otimes \mathbf{H}_c^{\eta} \otimes \mathbf{I}_{l_F}$ accounts for the relay of URLLC signal over the cable, and $\mathbf{u}$ is the vectorization of the URLLC signal matrix $\mathbf{U}$ in (5). Similarly to the proof of Lemma 2, the eMBB per-UE rate under NOMA by treating URLLC interference as noise can be computed by

$$\begin{aligned} R_B &= \frac{1}{n_F M} I\left(\mathbf{r}, \mathbf{x} \,|\mathbf{A}\right) \\ &= \frac{\mu}{M} \mathbb{E}_{\mathbf{A}} \left[\log\left(\det\left(\mathbf{I} + P_B \mathbf{R}_{A,z_{\text{eq}}}^{-1} \mathbf{H}_{\text{eq}} \mathbf{H}_{\text{eq}}^T\right)\right)\right] \end{aligned} \tag{A17}$$

where the average is taken over all the possible values of matrix $\mathbf{A}$, P_B is the power of the eMBB user under NOMA and $\mathbf{H}_{\text{eq}}$ is defined as in (A15). Finally, the covariance $\mathbf{R}_{A,z_{\text{eq}}}$ of the noise plus URLLC interference yields

$$\mathbf{R}_{A,z_{\text{eq}}} = \beta^2 P_U \quad \mathbf{A}_{\text{eq}} \mathbf{A}_{\text{eq}}^T + \mathbf{R}_{z_{\text{eq}}},$$

where $\mathbf{A}_{\text{eq}} = \mathbf{A} \otimes \mathbf{H}_c^{\eta}$ and $\mathbf{R}_{z_{\text{eq}}}$ is the same as in (A15). The proof is completed by noticing that $\mathbf{A}_{\text{eq}} \mathbf{A}_{\text{eq}}^T = (\mathbf{A} \otimes \mathbf{H}_c^{\eta})(\mathbf{A} \otimes \mathbf{H}_c^{\eta})^H = \mathbf{A} \otimes \mathbf{H}_c^{\eta} \mathbf{H}_c^{\eta}$, since matrix $\mathbf{A}$ is idempotent and $\mathbf{H}_c^{\eta}$ symmetric.

References

1. Shafi, M.; Molisch, A.F.; Smith, P.J.; Haustein, T.; Zhu, P.; Silva, P.D.; Tufvesson, F.; Benjebbour, A.; Wunder, G. 5G: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1201–1221. [CrossRef]
2. 5G PPP Architecture Working Group. View on 5G Architecture. 2016. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-July-2016.pdf (accessed on 31 August 2018).
3. Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s). 2017. Available online: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf (accessed on 31 August 2018).
4. Takeda, F. Study on New Radio (NR) Access Technology—Physical Layer Aspects. 2017. Available online: http://www.tech-invite.com/3m38/tinv-3gpp-38-802.html#e-3-1 (accessed on 31 August 2018).
5. Durisi, G.; Koch, T.; Popovski, P. Toward massive, ultrareliable, and low-latency wireless communication with short packets. *Proc. IEEE* **2016**, *104*, 1711–1726. [CrossRef]
6. Zhang, H.; Liu, N.; Chu, X.; Long, K.; Aghvami, A.H.; Leung, V.C.M. Network slicing based 5G and future mobile networks: mobility, resource management, and challenges. *IEEE Commun. Mag.* **2017**, *55*, 138–145, [CrossRef]
7. Popovski, P.; Nielsen, J.J.; Stefanovic, C.; de Carvalho, E.; Strom, E.; Trillingsgaard, K.F.; Bana, A.S.; Kim, D.M.; Kotaba, R.; Park, J.; et al. Wireless access for ultra-reliable low-latency communication: Principles and building blocks. *IEEE Netw.* **2018**, *32*, 16–23. [CrossRef]
8. Boccardi, F.; Heath, R.W.; Lozano, A.; Marzetta, T.L.; Popovski, P. Five disruptive technology directions for 5G. *IEEE Commun. Mag.* **2014**, *52*, 74–80. [CrossRef]
9. Wong, V.; Schober, R.; Ng, D.; Wang, L.E. *Key Technologies for 5G Wireless Systems*; Cambridge University Press: Cambridge, UK, 2017.
10. CPRI Specification V.6.1 (2014-07-01). 2014. Available online: http://www.cpri.info/downloads/CPRI_v_6_1_2014-07-01.pdf (accessed on 31 August 2018).
11. Bartelt, J.; Rost, P.; Wubben, D.; Lessmann, J.; Melis, B.; Fettweis, G. Fronthaul and backhaul requirements of flexibly centralized radio access networks. *IEEE Wirel. Commun.* **2015**, *22*, 105–111. [CrossRef]
12. Wake, D.; Nkansah, A.; Gomes, N.J. Radio over fiber link design for next generation wireless systems. *J. Lightw. Technol.* **2010**, *28*, 2456–2464. [CrossRef]
13. Gambini, J.; Spagnolini, U. Radio over telephone lines in femtocell systems. In Proceedings of the 21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Istanbul, Turkey, 26–30 September 2010; pp. 1544–1549.
14. Gambini, J.; Spagnolini, U. Wireless over cable for femtocell systems. *IEEE Commun. Mag.* **2013**, *51*, 178–185. [CrossRef]
15. Bartelt, J.; Fettweis, G. Radio-over-radio: I/Q-stream backhauling for cloud-based networks via millimeter wave links. In Proceedings of the 2013 IEEE Globecom Workshops (GC Wkshps), Atlanta, GA, USA, 9–13 December 2013; pp. 772–777.

16. Combi, L.; Gatto, A.; Martinelli, M.; Parolari, P.; Spagnolini, U. Pulse-Width optical modulation for CRAN front-hauling. In Proceedings of the 2015 IEEE Globecom Workshops (GC Wkshps), San Diego, CA, USA, 6–10 December 2015; pp. 1–5.
17. Dat, P.T.; Kanno, A.; Yamamoto, N.; Kawanishi, T. 5G transport networks: The need for new technologies and standards. *IEEE Commun. Mag.* **2016**, *54*, 18–26. [CrossRef]
18. Matera, A.; Spagnolini, U. On the optimal Space-Frequency to Frequency mapping in indoor single-pair RoC fronthaul. In Proceedings of the 2017 European Conference on Networks and Communications (EuCNC), Oulu, Finland, 12–15 June 2017; pp. 1–5.
19. Matera, A.; Combi, L.; Naqvi, S.H.R.; Spagnolini, U. Space-frequency to space-frequency for MIMO radio over copper. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; pp. 1–6.
20. Matera, A.; Spagnolini, U. Analog MIMO-RoC Downlink with SF2SF. *IEEE Wirel. Commun. Lett.* **2018**. [CrossRef]
21. Tonini, F.; Fiorani, M.; Furdek, M.; Raffaelli, C.; Wosinska, L.; Monti, P. Radio and transport planning of centralized radio architectures in 5G indoor scenarios. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1837–1848. [CrossRef]
22. Naqvi, S.H.R.; Matera, A.; Combi, L.; Spagnolini, U. On the transport capability of LAN cables in all-analog MIMO-RoC fronthaul. In Proceedings of the 2017 IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, USA, 19–22 March 2017; pp. 1–6.
23. Lu, C.; Berg, M.; Trojer, E.; Eriksson, P.E.; Laraqui, K.; Tridblad, O.V.; Almeida, H. Connecting the dots: Small cells shape up for high-performance indoor radio. *Ericsson Rev.* **2014**, *91*, 38–45. [CrossRef] [PubMed]
24. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 2012.
25. Saito, Y.; Kishiyama, Y.; Benjebbour, A.; Nakamura, T.; Li, A.; Higuchi, K. Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access. In Proceedings of the 2013 IEEE 77th Vehicular Technology Conference (VTC Spring), Dresden, Germany, 2–5 June 2013; pp. 1–5.
26. Shin, W.; Vaezi, M.; Lee, B.; Love, D.J.; Lee, J.; Poor, H.V. Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges. *IEEE Commun. Mag.* **2017**, *55*, 176–183. [CrossRef]
27. Ding, Z.; Lei, X.; Karagiannidis, G.K.; Schober, R.; Yuan, J.; Bhargava, V.K. A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2181–2195. [CrossRef]
28. Anand, A.; de Veciana, G.; Shakkottai, S. Joint scheduling of URLLC and eMBB traffic in 5G wireless networks. *arXiv* **2017**, arXiv:1712.05344.
29. Popovski, P.; Trillingsgaard, K.F.; Simeone, O.; Durisi, G. 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View. *arXiv* **2018**, arXiv:1804.05057.
30. Vaezi, M.; Ding, Z.; Poor, H. *Multiple Access Techniques for 5G Wireless Networks and Beyond*; Springer: New York, NY, USA, 2018.
31. Dai, L.; Wang, B.; Yuan, Y.; Han, S.; I, C.-L.; Wang, Z. Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun. Mag.* **2015**, *53*, 74–81. [CrossRef]
32. Kassab, R.; Simeone, O.; Popovski, P. Coexistence of URLLC and eMBB services in the C-RAN Uplink: An Information-Theoretic Study. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, UAE, 9–13 December 2018, pp. 1–6.
33. Making 5G NR a Reality. 2016. Available online: https://www.qualcomm.com/media/documents/files/whitepaper-making-5g-nr-a-reality.pdf (accessed on 31 August 2018).
34. eMBB and URLLC Multiplexing for NR. R1-1705051. 2017. Available online: https://portal.3gpp.org/ngppapp/CreateTdoc.aspx?mode=view&contributionId=775991 (accessed on 31 August 2018).
35. Simeone, O.; Levy, N.; Sanderovich, A.; Somekh, O.; Zaidel, B.M.; Poor, H.V.; Shamai, S. Cooperative wireless cellular systems: An information-theoretic view. *Found. Trends Commun. Inf. Theory* **2012**, *8*, 1–177. [CrossRef]
36. Rappaport, T.S.; Sun, S.; Mayzus, R.; Zhao, H.; Azar, Y.; Wang, K.; Wong, G.N.; Schulz, J.K.; Samimi, M.; Gutierrez, F., Jr. Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access* **2013**, *1*, 335–349. [CrossRef]

37. Hekrdla, M.; Matera, A.; Wang, W.; Wei, D.; Spagnolini, U. Ordered Tomlinson-Harashima Precoding in G.fast Downstream. In Proceedings of the 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, USA, 6–10 December 2015; pp. 1–6.
38. Spagnolini, U. *Statistical Signal Processing in Engineering*; John Wiley & Sons: New York, NY, USA, 2018.
39. Jakes, W.C.; Cox, D.C. *Microwave Mobile Communications*; John Wiley & Sons: New York, NY, USA, 1994.
40. Polyanskiy, Y.; Poor, H.; Verdu, S. Channel Coding Rate in the Finite Blocklength Regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [CrossRef]
41. Scarlett, J.; Tan, V.Y.; Durisi, G. The dispersion of nearest-neighbor decoding for additive non-Gaussian channels. *IEEE Trans. Inf. Theory* **2017**, *63*, 81–92. [CrossRef]
42. Zafaruddin, S.; Bergel, I.; Leshem, A. Signal processing for gigabit-rate wireline communications: An overview of the state of the art and research challenges. *IEEE Signal Process. Mag.* **2017**, *34*, 141–164. [CrossRef]
43. Zhang, W. A general framework for transmission with transceiver distortion and some applications. *IEEE Trans. Commun.* **2012**, *60*, 384–399. [CrossRef]
44. Merhav, N.; Kaplan, G.; Lapidoth, A.; Shitz, S.S. On information rates for mismatched decoders. *IEEE Trans. Inf. Theory* **1994**, *40*, 1953–1967. [CrossRef]