An Information-Theoretic View of Mixed-Delay Traffic in 5G and 6G

Homa Nikbakht, Michèle Wigger, Malcolm Egan, Shlomo Shamai Shitz, Jean-Marie Gorce, H Vincent Poor

entropy

MDPI

Tutorial

An Information-Theoretic View of Mixed-Delay Traffic in 5G and 6G

Homa Nikbakht [1,*], Michèle Wigger [2,*], Malcolm Egan [1,*], Shlomo Shamai (Shitz) [3,*], Jean-Marie Gorce [1,*] and H. Vincent Poor [4,*]

1 INRIA, INSA, CITI, Université de Lyon, EA3720, 69621 Villeurbanne, France
2 LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France
3 Department of Electrical and Computer Engineering, Technion–IIT, Haifa 3200003, Israel
4 School of Engineering and Applied Science, Princeton University, Princeton, NJ 08544, USA
* Correspondence: homa.nikbakht@inria.fr (N.H.); michele.wigger@telecom-paris.fr (M.W.); malcom.egan@inria.fr (M.E.); sshlomo@ee.technion.ac.il (S.S.); jean-marie.gorce@inria.fr (J.-M.G.); poor@princeton.edu (H.V.P.)

Abstract: Fifth generation mobile communication systems (5G) have to accommodate both Ultra-Reliable Low-Latency Communication (URLLC) and enhanced Mobile Broadband (eMBB) services. While eMBB applications support high data rates, URLLC services aim at guaranteeing low-latencies and high-reliabilities. eMBB and URLLC services are scheduled on the same frequency band, where the different latency requirements of the communications render their coexistence challenging. In this survey, we review, from an information theoretic perspective, coding schemes that simultaneously accommodate URLLC and eMBB transmissions and show that they outperform traditional scheduling approaches. Various communication scenarios are considered, including point-to-point channels, broadcast channels, interference networks, cellular models, and cloud radio access networks (C-RANs). The main focus is on the set of rate pairs that can simultaneously be achieved for URLLC and eMBB messages, which captures well the tension between the two types of communications. We also discuss finite-blocklength results where the measure of interest is the set of error probability pairs that can simultaneously be achieved in the two communication regimes.

Keywords: mixed-delay constraints; URLLC; eMBB

Citation: Nikbakht, H.; Wigger, M.; Egan, M.; Shamai, S.; Gorce, J.-M.; Poor, H.V. An Information-Theoretic View of Mixed-Delay Traffic in 5G and 6G. *Entropy* **2022**, *24*, 637. https://doi.org/10.3390/e24050637

Academic Editor: Song-Nam Hong

Received: 21 March 2022
Accepted: 22 April 2022
Published: 30 April 2022

1. Introduction

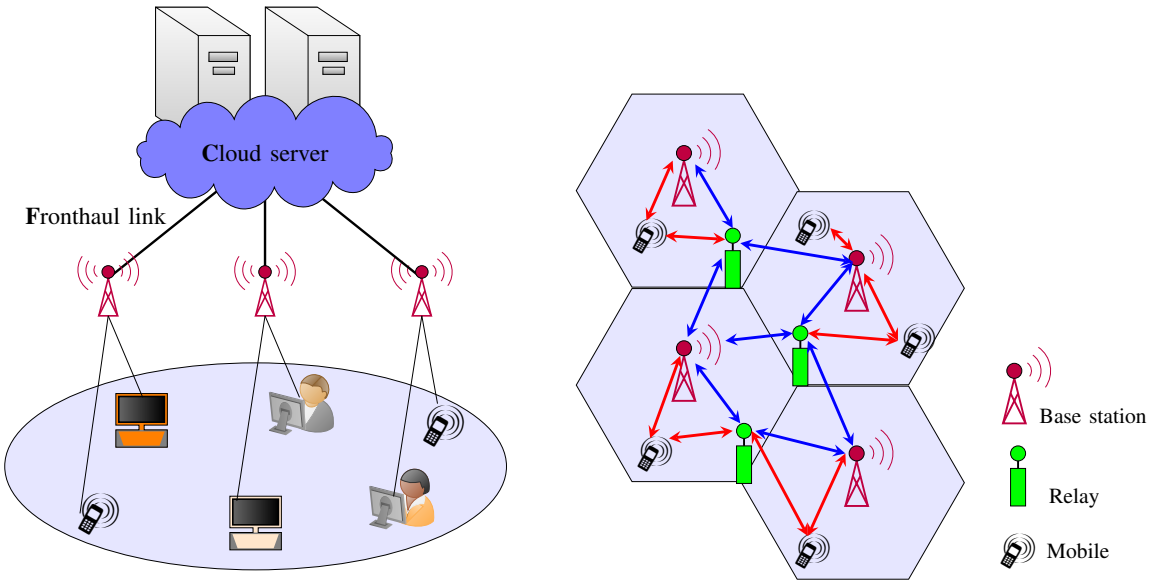
Modern communication networks serve a range of applications with heterogenous characteristics. Indeed, 5G and proposed 6G wireless mobile cellular networks are expected to serve a diverse set of applications including telephony, video-streaming, online gaming, time-critical control for transportation or remote surgery, or massive machine-type applications for sensor networks in the Internet of Things (IoT) [1]. These applications differ in terms of both reliability and latency requirements. A key example is when *Ultra-Reliable Low-Latency Communication (URLLC)* and *enhanced Mobile Broadband (eMBB)* [2] applications utilize the same time-frequency resource blocks.

URLLC is designed to ensure 99.99% reliability at a maximum end-to-end delay of no more than one millisecond [3–10]. It is thus suited for delay- and mission-critical applications such as remote surgery, control of manufacturing sites, or communication to and from autonomous vehicles. On the other hand, eMBB is most prominently used for video streaming and other applications with less stringent delay tolerances [2].

In 5G and proposed 6G systems, URLLC and eMBB users are allocated *network slices*, which correspond to resources within the radio access network. A key challenge is how to design resource allocation and coding schemes given that URLLC and eMBB slices have very different delay requirements. As such, the network must support *mixed delay traffic*. This challenge is further complicated when the radio access network exploits advanced

architectures, such as *cloud radio access networks (C-RANs)* [11,12] (illustrated in Figure 1a) or cooperative networks (illustrated in Figure 1b).

One standard approach is to use smart scheduling and resource allocation algorithms, which interrupt eMBB transmissions to send URLLC data. Various scheduling algorithms have been proposed, which exploit machine learning techniques [13–16] (including deep learning [17,18]) and intelligent reflective surfaces [19] to improve performance.

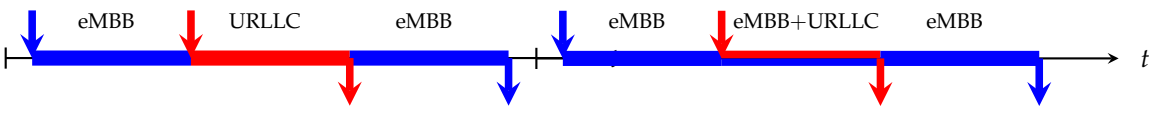


(**a**) C-RAN (**b**) Cellular network with cooperation links

Figure 1. Infrastructures that allow the mitigation of interference.

Nevertheless, an *information theoretic perspective* suggests that performance can be further improved via advanced *joint* coding schemes, which account for the mixed delay requirements of URLLC and eMBB slices. Indeed, by introducing joint coding schemes, data from both URLLC and eMBB slices can be simultaneously transmitted in the same resource block, as illustrated in Figure 2. A key issue is that interference is introduced not only by multiple users within the same time-frequency resource, but also from data from each slice transmitted by the *same* user. Nevertheless, as we show in this survey, by using appropriate joint coding techniques, the presence of interference does not necessarily lead to reductions in performance.

In this survey, we overview recent work on joint coding with mixed delay traffic arising from URLLC and eMBB slices in network architectures ranging from point-to-point to C-RANs.



(**a**) Scheduling approach (**b**) Joint coding approach

Figure 2. URLLC and eMBB transmissions: (**a**) Scheduling approach, (**b**) Joint coding approach.

1.1. Related Works

While in this survey we focus on communication scenarios with different network slices that have different delay constraints, information-theorists have also studied related scenarios with other types of heterogeneous communication requirements. The works most closely related to mixed-delay are [20–22]. Specifically, [20] studies a scenario where two messages are transmitted over a broadcast channel, but only one of them can profit from the cooperation link between the two receivers. The other message has to be

decoded directly based on the legitimate receiver's channel outputs, without cooperation from the other receiver. The motivation in [20] to study such a system was to design a robust communication scheme where the receivers can reliably decode two messages when the cooperation link between the receivers is present, while still being able to reliably decode a single message in case the cooperation link fails. A different interpretation, but with the same mathematical model, is to say that one of the messages needs to be decoded immediately without waiting for the cooperation message from the other receiver, while the other message can tolerate more delay and therefore be decoded also based on the cooperation message. In this sense, the model [20] well suits also a mixed-delay communication scenario with URLLC and eMBB slices as considered in this survey. The works in [21,22] study a scenario with different reliability criteria of two slices, as also characteristic for URLLC and eMBB slices. In particular, [21] imposes the constraint that the data from one slice has to be decoded even under an adversarial attack model (the arbitrarily varying channel [23–25]) whereas the data from the other slice only has to be decoded in the likely event that the channel exhibits an expected behavior.

1.2. Related Surveys and Contributions

A number of surveys, summarized in Table 1, have recently appeared covering varying aspects of coding, resource allocation, and architecture design in 5G and beyond. The surveys in [11,26,27] have focused on system level approaches in order to support network slicing, but do not consider aspects related to coding.

Table 1. Related Surveys.

Survey	Year	Comments
[26]	2020	System level perspective on C-RANs.
[11]	2014	System level perspective on C-RANs.
[27]	2019	Overview of network slicing.
[28]	2015	Overview of 5G cellular interference management.
[29]	2019	Machine learning for interference management.
[30]	2022	Survey on rate-splitting in multiple access networks..
[2]	2018	Overview of eMBB and URLLC from a communications theory perspective.
[31]	2018	Survey on control channel design.
[32]	2021	Survey on communication theoretic aspects of 5G.
[33]	2020	Survey on NOMA.
[34]	2019	Survey on NOMA.
[35]	2018	Survey on NOMA.
[36]	2016	Survey on NOMA.

On the other hand, the surveys in [2,28–30,32–36] focus on communication theoretic aspects of 5G and future 6G systems. In particular, [2,28,29,32] consider various resource allocation techniques and specifically [2] treates resource allocation for eMBB and URLLC slices. The surveys in [33–36] focus on non-orthogonal multiple access (NOMA) schemes based on successive interference cancellation for multiple access networks, while the survey [30] overviews the benefits of rate-splitting techniques on the multiple-access channel.

Despite the importance of joint coding schemes with mixed delay traffic, there has not been a comprehensive survey on this topic. This survey aims to fill this gap by highlighting how joint coding can improve performance of standard scheduling schemes, drawing on fundamental insights from an information theoretic analysis of the network.

The main contributions in this survey are summarized as follows:

(i) An overview of interference mitigation techniques drawn from information theory, with a focus on superposition coding, dirty paper coding, and coordinated multi point transmission and reception.
(ii) A summary of joint coding schemes and recent results on their performance for mixed delay traffic in
  (a) point-to-point networks;
  (b) broadcast networks;
  (c) cooperative networks;
  (d) C-RANs.
(iii) A discussion of open problems in the design of joint coding schemes for mixed delay traffic.

1.3. Outline of the Survey

This survey article is organized as follows. In Section 2 we review known interference mitigation techniques such as superposition coding, dirty-paper coding, and Coordinated Multi-Point (CoMP) transmission and reception. For a more thorough discussion of these tools, we refer to the original articles or standard textbooks [37,38]. We then continue in Section 3 to discuss integrated transmission of URLLC and eMBB messages on P2P channels with a single transmitter and a single receiver, followed by Section 4 which discusses extensions to multi-receiver broadcast channels (BC). Sections 5 and 6 consider mixed-delay transmissions over cooperative cellular interference networks and C-RANs. The survey is concluded with a summary and outlook section in Section 7.

Notation: Throughout the survey, we abbreviate *transmitter* and *receiver* by *Tx* and *Rx*. For *independent and identically distributed* we use *i.i.d.* Random variables are denoted using upper case letters, and realizations thereof by lower case letter, e.g., X and x. Random vectors are denoted with uppercase bold symbols. Fixed constants are often written with sans-serif font, for example $\mathsf{K}, \mathsf{P}, \mathsf{Q}$ or using Greek letters, for example ρ and α. To follow standard notation we however use n to denote the blocklength of transmission. For any positive integer K we use the short-hand notation $[\mathsf{K}] = \{1, \dots, \mathsf{K}\}$. Channels are assumed to be real-valued, extensions to complex channels with independent real and complex components are straightforward. In this sense, $\mathcal{N}(0, \sigma^2)$ denotes the real centered Gaussian distribution of variance σ^2. We also use the usual shorthand notation $Y^n = (Y_1, \dots, Y_n)$.

2. Mixed Delay Traffic and Interference Mitigation

2.1. Coding and Delay

The primary goal of a communications network is to reliably send one or more messages $M_i \in \{1, \dots, \mathsf{M}_i\}$ from one or more Txs to one or more Rxs. To do so, each Tx *encodes* the different messages it wishes to send into a waveform x_k^n, which corresponds to the physical signal sent over the network. In a scenario with *homogeneous delay constraints* i.e., where all messages are sent over the same blocklength n, a Tx k encodes its messages $\{M_i \colon i \in \mathcal{T}_k\}$, where $\mathcal{T}_k$ collects the indices of all messages sent by Tx k, into the codeword $x_k^n(\{M_i \colon i \in \mathcal{T}_k\})$ using a joint encoding function $f_k \colon \{1, \dots, \mathsf{M}_{i_k}\} \times \cdots \times \{1, \dots, \mathsf{M}_{i_{k+1}-1}\} \to \mathbb{R}^n$. After receiving the corresponding output symbols y_k^n, Rx k produces a guess $\{\hat{M}_i \colon i \in \mathcal{R}_k\}$ for each of its desired messages $\{M_i \colon i \in \mathcal{R}_k\}$, where $\mathcal{R}_k$ collects the indices of the messages intended for Rx k, by applying a decoding function $g_k \colon \mathbb{R}^n \to \{1, \dots, \mathsf{M}_{i_k}\} \times \cdots \times \{1, \dots, \mathsf{M}_{i_{k+1}-1}\}$ to its observed outputs y_k^n.

A more complicated scenario typically arises in the mixed-delay scenarios we consider in this survey, because the various messages are created at different times and have different decoding delays. In this case, each message is assigned a creation time a_i and a latest possible decoding time d_i. As a consequence, a Tx k produces its inputs x_k^n using per-symbol encoding functions $\{f_{k,t}\}_t$, where at time-t the function $f_{k,t}$ maps all its previously created messages to an input symbol:

$$x_{k,t} = f_{k,t}(\{M_i \colon i \in \mathcal{T}_k, a_i \leq t\}). \tag{1}$$

Rx k decodes any of its intended messages $\{M_i \colon i \in \mathcal{R}_k\}$ by applying the decoding function g_i to the first d_i channel outputs $y_k^{d_i}$. The decoding function that produces the message guess $\hat{M}_i$ is thus of the form $g_i \colon \mathbb{R}^{d_i} \to \{1, \ldots, \mathsf{M}_i\}$.

Associated with the described mixed-delay encodings are two key parameters:

(i) the length of each codeword, $n_i = d_i - a_i$;
(ii) and the *rate*, defined by

$$R_i = \frac{\log \mathsf{M}_i}{n_i}. \tag{2}$$

In multi-hop scenarios such as experienced in C-RANs or cooperative networks, transmission delay not only depends on the blocklength of communication, but also on the delay introduced from the communication over the addition hops. For example, in the uplink of C-RANs, the total delay experienced for the transmission of a messages is formed by:

- the communication delay over the network from the mobile users to the BSs;
- the processing time of the compression at the BSs as well as the communication delay over the fronthaul links to the cloud processor;
- the decoding processing time at the cloud processor.

In mixed-delay networks, the additional delay introduced by the compression at the BSs and the fronthaul communication might exceed the latest allowed decoding time d_i for certain messages M_i, which thus have to be directly decoded at the BSs. A similar situation is also encountered in the downlink of C-RANs, where URLLC messages should directly be encoded at the BSs and not at the cloud processor so as to avoid the delay introduced by the additional communication hop over the fronthaul link. In the same way, URLLC messages transmitted in cooperative interference networks cannot support the additional communication hops required to establish cooperation between Txs or Rxs. In these networks, the cooperative communication at the Tx side thus can only depend on eMBB messages and the cooperative communication at the Rx side can only serve decoding of eMBB messages. We will provide a more detailed model for the encoding and decoding of mixed-delay messages in Section 5 ahead. A main assumption in our model will be that the communication over the interference network is sufficiently short so that also URLLC messages can tolerate the introduced delay.

As we will overview in this survey, mixed delay constraints require careful design of the joint coding schemes, in particular to mitigate the interference caused by the different slices. In the remainder of this section, we summarize key information-theoretic interference mitigation techniques.

2.2. Superposition Coding

Superposition coding [37,39] was first proposed in the context of broadcast communication. It can be used to send multiple messages to one or more receivers. The main technical feature is that the different messages are encoded into the different layers of a so called superposition code, illustrated in Figure 3 for a code example with three layers. The entries of the layer-1 code are drawn i.i.d. according to a chosen distribution P_{U_0}. *For each layer-1 codeword* $u_0^{(n)}(\ell)$ *a new layer-2 codebook is chosen.* The entries of the layer-2 codewords are drawn independent of each other and the i-th entry follows a conditional distribution $P_{U_1|U_0}$ given the i-th entry of codeword $u_0^{(n)}(\ell)$. For each layer-2 codeword, a new layer-3 codebook is chosen. Entries of this codebook are again independently of each other and drawn according to a conditional distribution $P_{U_2|U_1}$ given the entries in the corresponding layer-2 codeword.

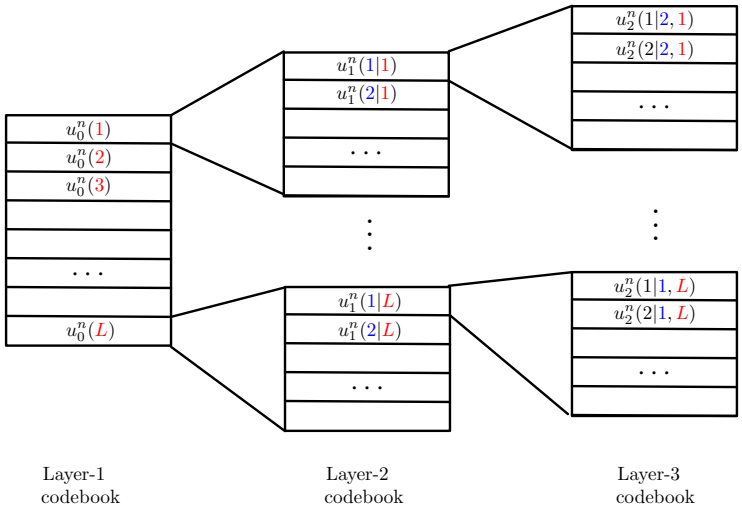


Figure 3. Superposition code with 3 layers.

In a superposition code, each message is not only protected by its corresponding layer, but also by all underlying layers. Given the structure of the code, any receiver that is interested in decoding a given layer also has to decode *all previous layers*, typically in a joint manner. Upper layers are not decoded and cause additional disturbance (noise) on the decoding of lower layers.

Given the described decoding order, in the context of mixed-delay traffic it is possible to send URLLC data on lower layers and eMBB on upper layers but not the other way around, because URLLC messages have to be decoded first, prior to decoding eMBB messages.

A simpler alternative to superposition coding is to encode each message using an independent codebook and to combine the chosen codewords, by means of a predefined mapping, to form the sequence of channel inputs. In particular, for Gaussian channels, messages are encoded into Gaussian codewords and the sum of these codewords is transmitted over the channel. The advantage of this method is that no layering-order of the codewords has to be established a priori. This is particularly convenient in fading channels where the exact channel statistics are not known at time of encoding, and the receiver can decide on the layers to decode after having estimated the realization of the channel. In this sense, the simpler alternative can allow for increased *expected rates* over slowly fading channels where the channel variations are limited over the duration of a single codeword. As we shall see, this approach is also highly beneficial for joint transmission of URLLC and eMBB messages where the fading is almost constant over the duration of an URLLC communication but varies significantly during the transmission of eMBB messages.

2.3. Dirty-Paper Coding (DPC)

If interference is known at a Tx before communication starts, the Tx can mitigate this interference through *Dirty Paper Coding* [40–42] and achieve the full capacity of the channel without interference. To illustrate, consider the Gaussian interference channel

$$Y^n = X^n + W^n + Z^n, \tag{3}$$

where Z^n is an i.i.d. standard Gaussian noise sequence and W^n is memoryless interference sequence, with each component zero-mean Gaussian of power Q. If W^n is unknown to both the Tx and the Rx, the interference simply acts as additional noise, and the capacity of the channel equals $\frac{1}{2}\log(1+\frac{\mathsf{P}}{1+\mathsf{Q}})$. If the Rx knows W^n, then it can subtract this interference from the outputs and as a consequence the capacity of the channel is the same as without interference, i.e., $\frac{1}{2}\log(1+\mathsf{P})$. Costa [40] showed that when W^n is unknown to the Rx but known to the Tx even before the communication starts, then the capacity of the channel is also equal to the interference-free capacity $\frac{1}{2}\log(1+\mathsf{P})$. The coding scheme achieving this

performance was termed dirty-paper coding (DPC) and is described in the following. (For an analysis see [37,40]).

Define the parameter $\alpha := \frac{\mathsf{P}}{1+\mathsf{Q}}$ and the random variable $U = X + \alpha W$ where $W \sim \mathcal{N}(0, \mathsf{Q})$ and $X \sim \mathcal{N}(0, \mathsf{P})$ independent of each other. It can be verified that

$$I(U;Y) = \frac{1}{2}\log\left(\frac{(\mathsf{P}+\mathsf{Q}+1)(\mathsf{P}+\alpha^2\mathsf{Q})}{\mathsf{PQ}(1-\alpha)^2+(\mathsf{P}+\alpha^2\mathsf{Q})}\right) \tag{4}$$

$$I(U;W) = \frac{1}{2}\log\left(\frac{\mathsf{P}+\alpha^2\mathsf{Q}}{\mathsf{P}}\right) \tag{5}$$

$$I(U;Y) - I(U;W) = \frac{1}{2}\log\left(\frac{\mathsf{P}(\mathsf{P}+\mathsf{Q}+1)}{\mathsf{PQ}(1-\alpha)^2+(\mathsf{P}+\alpha^2\mathsf{Q})}\right) = \frac{1}{2}\log(1+\mathsf{P})\mathsf{Q} \tag{6}$$

Fix $\epsilon > 0$ arbitrary small. For each $m \in \left[2^{n(I(U;Y)-I(U;W)-\epsilon)}\right]$ generate a bin with $2^{n(I(U;W)+\epsilon/2)}$ codewords $\left\{U^n(j,m)\colon j \in \left[2^{n(I(U;W)+\epsilon/2)}\right]\right\}$ by picking each component of each codeword i.i.d. according to $\mathcal{N}\left(0, \mathsf{P}+\alpha^2\mathsf{Q}\right)$. Reveal the codebook consisting of all $2^{n(I(U;Y)-I(U;W)-\epsilon)}$ bins to the Rx and the Tx.

Encoding: To encode a message $M = m$, the encoder looks for a codeword $U^n(j,m)$ in bin m that is jointly typical [37] (i.e., has approximately the correct joint empirical distribution) with the interference sequence W^n. The Tx then forms $X^n = U^n(j^*,m) - \alpha W^n$, where j^* indicates the chosen index, and send this sequence X^n over the channel. Note that the Tx declares an error if no codeword $U^n(j,m)$ in bin m is jointly typical with the interference W^n.

Decoding: After observing the sequence Y^n, the Rx looks for a codeword $U^n(j,m)$ that is jointly typical with Y^n. If a single such codeword exists, the Rx declares $\hat{M} = m$, otherwise it declares an error.

It can be shown that with probability tending to 1 as the blocklength $n \to \infty$, the only codeword that is jointly typical with Y^n is codeword $U^n(j^*,m)$ which was selected at the transmitter. The Rx thus not only recovers the correct message $\hat{M} = M$ with high probability, but can also reconstruct the transmitted codeword $U^n(j^*,m)$ with high probability.

2.4. Coordinated Multi Point (CoMP)

Coordinated multi-point (CoMP) refers to a wide set of techniques that enable either a set of distributed Txs to jointly encode messages or a set of distributed Rxs to jointly decode messages. We will be particularly interested in scenarios where CoMP is facilitated through cooperative communication over dedicated links between transmitters or between receivers.

In the case of *CoMP transmission* [43–46], we consider a set of distributed Txs, each having one message to send, and with cooperation links between neighbouring Txs. Before communicating to the Rxs, all Txs convey their messages to a dedicated Tx, called the *master Tx*, which then jointly designs input signals for all Txs (also exploiting its available state-information) and conveys rate-distortion compressed (lossy) versions of these signals to each of the Txs. The Txs reconstruct the compressed signals and send these signals over the channel to the Rxs. If cooperation links are of sufficiently high rates, then the loss of the compression can be maintained at noise-level and does not decrease the *degrees of freedom (DoF)*, i.e., the factor in front of the logarithmic expansion of the asymptotic high-SNR sum-capacity. In this case, the interference channel is intuitively transformed into a multi-antenna single-Tx BC and the DoF is given by the minimum number of Tx and Rx antennas.

In case of *CoMP reception* [47–50], consider a set of distributed Rxs, each wishing to decode one message, and with cooperation links between neighbouring Rxs. Each Rx applies a lossy compression algorithm to its observed output signal and describes the compressed signal over the cooperation links to a dedicated Rx, called *master Rx*. This master Rx reconstructs all the compressed signals and jointly decodes all the messages, which it then sends to their intended Rxs over the cooperation links. If cooperation links are

of sufficiently large rates, then lossy compression of the receive signals can be performed so that the loss is maintained at noise-level, which again has no influence on the DoF of the channel, and corresponds to the DoF of a single-Rx multi-access channel which equals the minimum number of Tx and Rx antennas.

CoMP transmission and reception can only be used to encode and decode eMBB messages, because the communication over the cooperation links induces significant delay. The delay is in fact given by twice the number of hops required on the cooperation links to reach the master Tx/Rx from any other Tx/Rx in the network times the communication duration on a single cooperation hop. Since in practical networks also eMBB communication is delay-limited, CoMP transmission and reception can be performed only on small subsets of Txs and Rxs, where the size depends on the maximum allowed number of communication hops.

3. Point-to-Point Communications

3.1. Introduction

This section focuses on P2P channels with a single Tx and a single Rx. Section 3.2 reviews the superposition coding approach over fading channels in [51,52]. This approach manages to send URLLC messages over single coherence blocks of the fading channel without suffering from a degradation due to the lack of state knowledge at the Tx, and simultaneously also sends eMBB messages over multiple coherence blocks, thus exploiting the ergodic behaviour of the channel. This approach is also known as *broadcast approach* and has been studied for a wide field of applications, see the recent survey paper [53].

Section 3.3 summarizes the results in [54,55], which analyze a similar superposition coding approach but for Gaussian channels and in the finite-blocklength regime. The analysis is based on the concept of "parallel channels" introduced in [56].

3.2. The Broadcast Approach over Fading Channels without Transmitter Channel State Information

This section is based on the results in [51–53]. Consider a P2P channel where a single Tx wishes to send both URLLC and eMBB messages over a fading channel to a single Rx. Latency requirements impose that transmission of URLLC messages spans only a single coherence time of the fading channel, but transmission of eMBB can span multiple coherence blocks and thus profit from channel diversity. In a single-antenna setup, a simple channel model capturing these constraints is as follows:

$$Y_{b,t} = \sqrt{S_b} \cdot X_{b,t} + Z_{b,t}, \quad b = 1, \ldots, \mathsf{B}, \quad t = 1, \ldots, \mathsf{T}, \tag{7}$$

where B denotes the number of blocks, T the channel coherence time, $\{S_b\}$ describe the fading power in the various coherence blocks and are assumed i.i.d. with probability distribution function (pdf) f_S and variance 1, and $\{Z_{b,t}\}$ is a sequence of i.i.d. standard Gaussian noises. The fading power is assumed to be perfectly known at the Rx (e.g., by transmitting pilot signals at the beginning of each block based on which the Rx can estimate the fading power), but not at the Tx. Since each URLLC message can be transmitted only over a single coherence block, the channel inputs are formed as

$$X_{b,t} = f_{b,t}\left(M_b^{(\mathsf{U})}, M^{(\mathsf{e})}\right), \quad b = 1, \ldots, \mathsf{B}, \quad t = 1, \ldots, \mathsf{T}, \tag{8}$$

for some appropriate encoding functions $\{f_{b,t}\}$ satisfying the power constraint

$$\frac{1}{\mathsf{BT}} \sum_{b=1}^{\mathsf{B}} \sum_{t=1}^{\mathsf{T}} |X_{b,t}|^2 \leq \mathsf{P}, \tag{9}$$

and where $M_b^{(\mathsf{U})}$ indicates the URLLC message sent in block b and $M^{(\mathsf{e})}$ the single eMBB message sent over the entire B blocks.

In the following, messages are assumed independent of each other and uniform over message sets $\mathcal{M}_{\mathsf{U}}$ and $\mathcal{M}_{\mathsf{e}}$. In this subsection, $\mathcal{M}_{\mathsf{U}} = \left[2^{\mathsf{T}R_{\mathsf{U}}}\right]$ and $\mathcal{M}_{\mathsf{e}} = \left[2^{\mathsf{TB}R_{\mathsf{e}}}\right]$, where R_{U} and R_{e} denote the URLLC and eMBB rates of transmission.

After each block b, the Rx decodes the URLLC message $M_b^{(\mathsf{U})}$ sent in this block:

$$\hat{M}_b^{(\mathsf{U})} = g_b^{(\mathsf{U})}(Y_{b,1}, \ldots, Y_{b,\mathsf{T}}), \quad b = 1, \ldots, \mathsf{B}, \tag{10}$$

and at the end of the entire communication it also decodes the eMBB message:

$$\hat{M}^{(\mathsf{e})} = g^{(\mathsf{e})}(Y_{1,1}, \ldots, Y_{\mathsf{B},\mathsf{T}}), \tag{11}$$

for decoding functions $\{g_b^{(\mathsf{U})}\}$ and $g^{(\mathsf{e})}$ on appropriate domains.

In [51,52], the authors propose to encode the two message streams using simplified superposition coding where both streams are encoded into independent Gaussian codewords, which are then added up for transmission. More precisely, the URLLC message *in each block* is encoded into multiple layers so that the Rx can decode as many layers as the actual instantaneous fading power S_b permits. (This implies also that depending on the fading realization, certain URLLC messages are not decoded and in practice have to be retransmitted in the next block.) To allow for closed-form expressions, an infinite layering approach is employed with layers that are of infinitesimally small power.

The Tx allocates total power $\beta\mathsf{P}$ to the transmission of the URLLC messages and power $(1-\beta)\mathsf{P}$ to transmit the eMBB messages. The power distribution to the different URLLC layers is described by a power density $\rho(\cdot)$ satisfying $\int_u \rho(u)du = \beta\mathsf{P}$, where $\rho(s')$ indicates the (infinitesimally small) power that is assigned to a given layer that is decoded whenever the fading $S_b \geq s'$. The interference power stemming from non-decoded URLLC messages under state $S_b = s$ is then given by $s \cdot I(s)$ where

$$I(s) := \int_{u=s}^{\infty} \rho(u)\mathrm{d}u. \tag{12}$$

Since URLLC messages are decoded after each block, and eMBB messages only at the end of the last block B, decoding of URLLC messages not only suffers from the interference of non-decoded URLLC messages, but also from the interference of eMBB messages. The power of this latter interference is equal to $(1-\beta)\mathsf{P}$ independent of the block (since the Tx has no knowledge about $\{S_b\}$ it cannot adapt the power).

To decode the eMBB message at the end of the last block B, the Rx first subtracts the contributions of the codewords corresponding to the decoded URLLC messages and then decodes the eMBB message based on this difference while accounting for the interference power created by all non-decoded URLLC messages, which in block b is given by $I(S_b)$.

A careful analysis of the infinite-layering approach, see [52], reveals that the expected rate of the reliably decoded messages (i.e., messages decoded with error probability tending to 0 as the blocklength $\mathsf{T} \to \infty$) can be as high as

$$R^{(\mathsf{U})} = \int_{u=0}^{\infty} (1 - F_S(u)) \frac{u\rho(u)}{1 + u(I(u) + (1-\beta)\mathsf{P})} \mathrm{d}u, \tag{13}$$

where $F_S(\cdot)$ denotes the cumulative distribution function (cdf) associated with the pdf $f_S(\cdot)$. In the denominator of (13), the term $u(I(u) + (1-\beta)\mathsf{P})$ indicates the interference power experienced during the decoding of URLLC messages stemming from the eMBB transmission and the non-decoded URLLC messages.

For a sufficiently large number of blocks B, the following rate is achievable for the eMBB messages:

$$R^{(\mathsf{e})} = \int_{u=0}^{\infty} f_S(u) \log\left(1 + \frac{(1-\beta)\mathsf{P}u}{1 + uI(u)}\right) \mathrm{d}u. \tag{14}$$

Here we find $uI(u)$ in the denominator which describes the interference power of the non-decoded URLLC messages.

Equations (13) and (14) thus determine the maximum achievable (expected) sum-rate $R^{(\mathsf{U})} + R^{(\mathsf{e})}$ in function of the URLLC interference power $I(u)$ (notice that $\rho(u) = -\frac{\mathrm{d}}{\mathrm{d}u} I(u)$), which is a design parameter of the scheme and can be optimized. It is shown in [52] that the optimal interference power function $I(s)$ among all continuously differentiable functions satisfying the boundary conditions $I(0) = \beta \mathsf{P}$ and $I(\infty) = 0$ is given by:

$$I^*(s) = \frac{1}{2}\left(\frac{-b(s) + \sqrt{b^2(s) - 4a(s)c(s)}}{2a(s)}\right), \tag{15}$$

for $a(s) = sf_S(s)$, $b(s) = 2(1-\beta)\mathsf{P}f_S(s)s^2 - (1 - F_s(s))$, and $c(s) = (1-\beta)^2\mathsf{P}^2 f_S(s)s^3$.

Figure 4 compares the sum-rate achieved for this optimal interference power $I^*(s)$ for different power allocation parameters β to a simple outage-based approach where $\rho(s)$ is chosen as a dirac-function at threshold s_{th}, i.e., when the interference power is given by the step function

$$I_{\mathrm{o}}(s) = \mathbb{1}\{s < s_{\mathrm{th}}\}. \tag{16}$$

Here, the optimal value for the threshold s_{th} can be derived analytically and is given by the solution to the following equation

$$f_S(s_{\mathrm{th}}) \log(1 + \beta \mathsf{P} s_{\mathrm{th}}) = (1 - F_S(s_{\mathrm{th}})) \frac{\beta \mathsf{P}}{(1 + \mathsf{P}s_{\mathrm{th}})(1 + (1-\beta)\mathsf{P}s_{\mathrm{th}})}. \tag{17}$$

From Figure 4, we observe that for small URLLC rates $R^{(\mathsf{U})}$, the penalty in eMBB rates $R^{(\mathsf{e})}$ is small when using the suboptimal power allocation corresponding to $I_{\mathrm{o}}(s)$ instead of the optimal allocation corresponding to $I^*(s)$. For larger URLLC rates, this penalty increases.

We further observe that the maximum sum-rate achieved by both power allocations decreases with increasing URLLC rates. The sum-rate for both approaches is more than 5 when $R^{(\mathsf{U})} = 0$. For $R^{(\mathsf{U})} \geq 3.5$ it is around 4 under the optimal power allocation and even vanishes completely under outage power allocation leading to (16). In this high-URLLC-rate regime, the gap to the outer bound is also significant, leaving open the possibility of finding better coding schemes.

The described broadcast approach can further be improved by applying a multi-layering approach also for the transmitted eMBB message. In this approach, different eMBB layers are decoded successively, and after each eMBB decoding step, the Tx decodes further URLLC layers so as to remove their interference for the decoding of subsequent eMBB layers. This additional decoding of URLLC messages at the end of block B cannot be used to improve performance of the URLLC communication, because the admissible delay is exceeded. However, it enables an improvement in the decoding performance of eMBB messages.

Another way to improve this broadcast approach is to combine it with adaptive causal network coding. For example, the work in [57] proposes a novel layering scheme consisting of a base layer and an enhancement layer for data streaming under mixed-delay constraints. The base layer contains URLLC data and the enhancement layer contains eMBB data. In the proposed scheme, the base layer is encoded using a broadcast approach, which allows the Rx to decode the base layer (i.e., URLLC data) with minimum delay required. The enhancement layer is encoded using a priori and posteriori forward error correction so as to be able to control the throughput-delay trade-off of this communication.

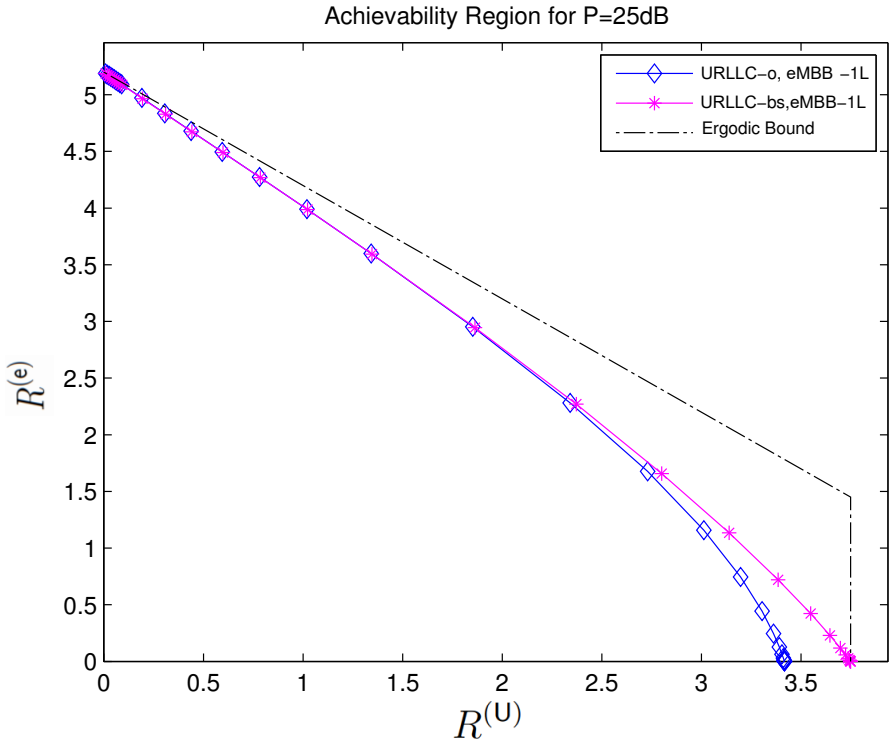


Figure 4. The figure illustrates the set of rate pairs $(R^{(\mathsf{U})}, R^{(\mathsf{e})})$ in function of the power allocation parameter β and of the interference powers in (15) and (16), respectively.

3.3. Finite Block-Length Analysis over Gaussian Channels

This section is based on the results in [54,55]. We again consider a P2P scenario, but where communication is over a non-fading Gaussian channel

$$Y_t = X_t + Z_t, \qquad t = 1, 2, \ldots,$$

for $\{Z_t\}$ an i.i.d. standard Gaussian noise sequence.

The Tx has a single URLLC message and a single eMBB message to send to the Rx, where both messages are assumed to have strict creation times and fixed decoding deadlines. Specifically, transmission of eMBB message $M^{(\mathsf{e})}$ commences at time $t = a_{\mathsf{e}}$ and decoding has to be performed at time $t = d_{\mathsf{e}}$, while the URLLC message can be transmitted starting at time $t = a_{\mathsf{U}}$ and has to be decoded at time $t = d_{\mathsf{U}}$. We thus parametrize the message sets as $\mathcal{M}_{\mathsf{U}} = \left[2^{(d_{\mathsf{U}} - a_{\mathsf{U}})R_{\mathsf{U}}}\right]$ and $\mathcal{M}_{\mathsf{e}} = \left[2^{(d_{\mathsf{e}} - a_{\mathsf{e}})R_{\mathsf{e}}}\right]$. We also denote the transmission window of the eMBB message by $\mathcal{W}_{\mathsf{e}}$ and the transmission window of the URLLC messages by $\mathcal{W}_{\mathsf{U}}$:

$$\mathcal{W}_{\mathsf{e}} \stackrel{\triangle}{=} \{a_{\mathsf{e}}, \ldots, d_{\mathsf{e}}\}, \quad \mathcal{W}_{\mathsf{U}} \stackrel{\triangle}{=} \{a_{\mathsf{U}}, \ldots, d_{\mathsf{U}}\}. \tag{18}$$

Since the URLLC delay $d_{\mathsf{U}} - a_{\mathsf{U}}$ is assumed shorter than the eMBB delay $d_{\mathsf{e}} - a_{\mathsf{e}}$, the following three situations can occur:

- Case 1: URLLC and eMBB transmissions do not overlap. i.e., $\mathcal{W}_{\mathsf{U}} \cap \mathcal{W}_{\mathsf{e}} = \emptyset$.
- Case 2: The eMBB transmission interval includes the URLLC transmission interval, i.e., $\mathcal{W}_{\mathsf{U}} \subset \mathcal{W}_{\mathsf{e}}$.
- Case 3: URLLC and eMBB transmissions overlap, but URLLC transmission is not included in eMBB transmission, i.e., $\mathcal{W}_{\mathsf{e}} \cap \mathcal{W}_{\mathsf{U}} \neq \emptyset$ and $\mathcal{W}_{\mathsf{U}} \not\subset \mathcal{W}_{\mathsf{e}}$.

In Case 1 where the two messages are transmitted during independent time intervals, URLLC and eMBB transmissions can be analyzed independently based on the achievability and converse bounds in [58]. Cases 2 and 3 can be treated similarly. Here, we focus on a subcase of Case 3 where encoding starts with the eMBB message at time $t = a_{\mathsf{e}}$ and

terminates at time $t = d_{\mathsf{U}}$ with the decoding of the URLLC message, see Figure 5. The Tx thus produces channel inputs at times $t \in \{a_{\mathsf{e}}, \ldots, d_{\mathsf{U}}\}$ as follows:

$$X_t = \begin{cases} f_t(M^{(\mathsf{e})}), & t \in \{a_{\mathsf{e}}, \ldots, a_{\mathsf{U}} - 1\} \\ \psi_t(M^{(\mathsf{e})}, M^{(\mathsf{U})}), & t \in \{a_{\mathsf{U}}, \ldots, d_{\mathsf{e}}\} \\ \phi_t(M^{(\mathsf{U})}), & t \in \{d_{\mathsf{e}} + 1, \ldots, d_{\mathsf{U}}\}, \end{cases} \tag{19}$$

where $\{f_t\}, \{\psi_t\}, \{\phi_t\}$ are appropriate encoding functions. Note that the Tx does not know the URLLC message before time $t = a_{\mathsf{U}}$ and therefore channel inputs prior to time $t = a_{\mathsf{U}}$ cannot depend on $M^{(\mathsf{U})}$. It can also be assumed that channel inputs after the eMBB decoding time d_{e} de not depend on the eMBB message $M^{(\mathsf{e})}$. One can therefore think of the transmission taking place over three parallel channels, with respective blocklengths

$$n_1 \triangleq a_{\mathsf{U}} - a_{\mathsf{e}}, \quad n_2 \triangleq d_{\mathsf{e}} - a_{\mathsf{U}} + 1, \quad \text{and} \quad n_3 \triangleq d_{\mathsf{U}} - d_{\mathsf{e}}, \tag{20}$$

where the first channel consists of channel uses $\{a_{\mathsf{e}}, \ldots, a_{\mathsf{U}} - 1\}$ and incorporates only eMBB transmission; the second channel consists of channel uses $\{a_{\mathsf{U}}, \ldots, d_{\mathsf{e}}\}$ and incorporates *joint transmission* of URLLC and eMBB messages; and the third channel consists of channel uses $\{d_{\mathsf{e}} + 1, \ldots, d_{\mathsf{U}}\}$ and incorporates only URLLC transmission. We denote the inputs and outputs of the three parallel channels by

$$\boldsymbol{X}_1 \triangleq \{X_{a_{\mathsf{e}}}, \ldots, X_{a_{\mathsf{U}}-1}\}, \quad \boldsymbol{Y}_1 \triangleq \{Y_{a_{\mathsf{e}}}, \ldots, Y_{a_{\mathsf{U}}-1}\}, \tag{21a}$$

$$\boldsymbol{X}_2 \triangleq \{X_{a_{\mathsf{U}}}, \ldots, X_{d_{\mathsf{e}}}\}, \quad \boldsymbol{Y}_2 \triangleq \{Y_{a_{\mathsf{U}}}, \ldots, Y_{d_{\mathsf{e}}}\}, \tag{21b}$$

$$\boldsymbol{X}_3 \triangleq \{X_{d_{\mathsf{e}}+1}, \ldots, X_{d_{\mathsf{U}}}\}, \quad \boldsymbol{Y}_3 \triangleq \{Y_{d_{\mathsf{e}}+1}, \ldots, Y_{d_{\mathsf{U}}}\}. \tag{21c}$$

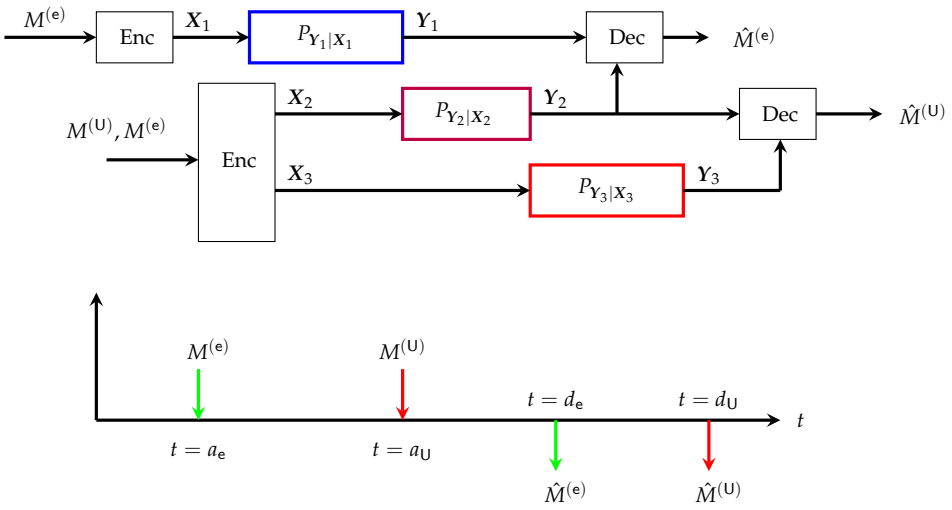


Figure 5. System model for transmission of URLLC and eMBB messgaes in the finite blocklength regime and under heterogeneous decoding deadline.

For the i-th channel with $i \in \{1, 2, 3\}$, the encoding functions satisfy the average block power constraint

$$\frac{1}{n_i} ||\boldsymbol{X}_i||^2 \leq \mathsf{P}_i \tag{22}$$

almost surely. The resulting system model is illustrated in Figure 5, where notice that the three channels $P_{Y_1|X_1}$, $P_{Y_2|X_2}$ and $P_{Y_3|X_3}$ are additive, memoryless, stationary, and Gaussian of variances 1.

The scheme further proposes to combine the eMBB and URLLC transmission over the second channel $P_{Y_2|X_2}$ by means of the simple superposition coding approach described at

the end of Section 3.2, for which the block-2 power P_2 is split into power $\beta\mathsf{P}_2$ for the eMBB transmission and power $(1-\beta_2)\mathsf{P}_2$ for the URLLC transmission, for some $\beta \in [0,1]$. In particular, the channel inputs $\boldsymbol{X}_2$ are formed as

$$\boldsymbol{X}_2 = \boldsymbol{X}_{2,\mathsf{e}} + \boldsymbol{X}_{2,\mathsf{U}}, \tag{23}$$

where $\boldsymbol{X}_{2,\mathsf{e}}$ is a codeword encoding $M^{(\mathsf{e})}$ of average power $\|\boldsymbol{X}_{2,\mathsf{e}}\|^2 = n_2\beta\mathsf{P}_2$, and $\boldsymbol{X}_{2,\mathsf{U}}$ is a codeword encoding $M^{(\mathsf{U})}$ of average power $\|\boldsymbol{X}_{2,\mathsf{U}}\|^2 = n_2(1-\beta)\mathsf{P}_2$.

The Rx first decodes the eMBB message based on the outputs of the first and second channels where it treats the transmission of the URLLC message over the second channel as interference. Subsequently, it decodes the URLLC message based on the outputs of the second and third channel, conditioning on the already decoded eMBB message.

The error probabilities of the described scheme can be analyzed and compared to fundamental lower bounds on the error probabilities, obtained via *meta-converse* arguments [58] with an extension to parallel channels [56]. As shown in [54], the converse and achievability bounds match in specific cases.

Figure 6, from [54], compares the performance of the described superposition coding scheme with a standard scheduling scheme that allocates the first half of the channel uses to eMBB transmission and the second half to URLLC transmission. (Under this scheduling approach $\epsilon_{\mathsf{e}} = \epsilon_{\mathsf{U}}$.) One observes that for the chosen set of parameters, $n_1 = n_2 = n_3 = 20$, $R_{\mathsf{U}} = 1/4$, and $\beta = 0.65$, the superposition coding approach results in almost identical URLLC and eMBB error probabilities ϵ_{e} and ϵ_{U}. Moreover, at medium and high powers P_2, the superposition coding approach significantly outperforms the scheduling approach.

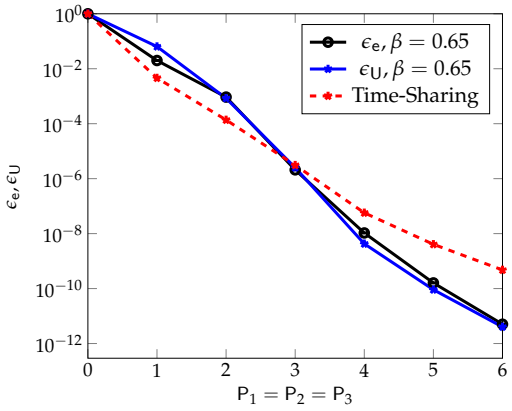


Figure 6. The figure illustrates the average error probabilities of the eMBB and URLLC messages denoted by ϵ_{e} and ϵ_{U} in function of the block transmit powers $\mathsf{P}_1 = \mathsf{P}_2 = \mathsf{P}_3$ and for blocklengths $n_1 = 20, n_2 = 20, n_3 = 20$, URLLC rate $R_{\mathsf{U}} = 1/4$, and power split $\beta = 0.65$.

In [55], the authors extend above coding scheme by using the finite-blocklength dirty paper coding (DPC) scheme in [59] to precancel the interference of the eMBB message on the URLLC transmission. Notice that in finite-blocklength DPC, the joint-typicality check is replaced by a norm condition on the input signal, and the Tx has to sacrifice few channel uses to approximately describes the norm of the interference sequence to the Rx. The error probabilities of this DPC based scheme are analyzed in [55] based on the DPC analysis technique in [59] and the parallel channel extension analysis in [56].

Figure 7, from [55], compares the performances of the proposed DPC based scheme with standard scheduling, and shows that for large transmit powers $\mathsf{P}_1 = \mathsf{P}_2 = \mathsf{P}_3$, the DPC based scheme outperforms scheduling over a wide range of blocklengths n_1.

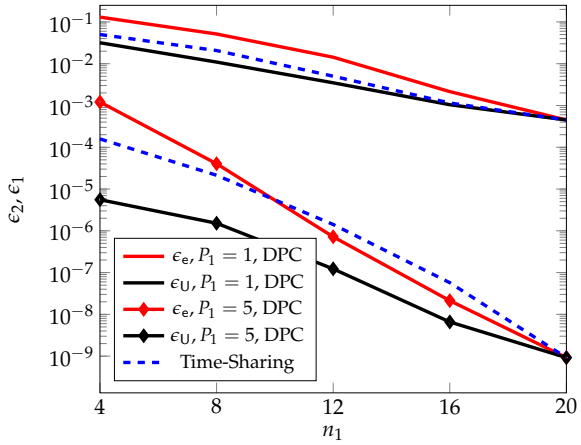


Figure 7. DPC based upper bounds on ϵ_{U} and ϵ_{e} in function of the blocklength n_1 and for average block-powers $\mathsf{P}_1 = \mathsf{P}_2 = \mathsf{P}_3$.

3.4. Summary

This section considered a P2P channel with a single Tx that sends an URLLC message and an eMBB message, where the two types of messages have different decoding delays. In Section 3.2, transmission is over fading channels and URLLC messages have to be transmitted within a single coherence block, whereas eMBB messages can be sent over multiple blocks and thus profit from channel diversity. To compensate for the missing channel state-information at the Tx, an infinite-layer broadcast approach is employed. A closed-form solution for the sum-rate achieved by this broadcast approach was presented, and based on numerical simulations it was observed its maximum sum-rate decreases with increasing URLLC rates. Furthermore, a simplified single-layer power allocation was shown to perform close to the optimal power allocation in the broadcast approach at low URLLC rates.

Section 3.3 studies a related simplified superposition coding or dirty-paper coding schemes but over static Gaussian channels and with fixed creation and decoding times. For this setup, upper and lower bounds on the set of achievable error probability pairs that can simultaneously be achieved for URLLC and eMBB messages was derived in [54]. The obtained results show a performance improvement under these schemes compared to the standard scheduling scheme.

4. Broadcast Channels with Mixed-Delay Traffic

4.1. Introduction

This section focuses on multi-receiver broadcast channels (BC). The results of this section are based on [60] where similarly to Section 3.2, URLLC messages have to be decoded within a single coherence block but eMBB messages can be transmitted over multiple blocks. In contrast to Section 3.2, the fading powers $\{S_{k,b}\}$ of the various blocks are known to the various Rxs and the Tx in advance.

4.2. Broadcast Approach over Fading Channels

In the setup proposed in [60] each Rx might demand a URLLC message, an eMBB message or both. We thus define the two sets $\mathcal{K}^{(\mathsf{U})}$ and $\mathcal{K}^{(\mathsf{e})}$ indicating the sets of users requesting URLLC and eMBB messages, respectively. Notice that the two sets can overlap. The mixed-delay constraint is captured by imposing a fixed rate on all transmitted URLLC messages, whereas eMBB messages can be either of larger or smaller rates. This rate-adaption on eMBB messages depending on the encountered fading powers enables an increase in the system's sum-rate. The corresponding optimization problem can be expressed as

$$\max \sum_{k \in \mathcal{K}^{(\mathsf{U})}} R^{(\mathsf{U})} + \sum_{k \in \mathcal{K}^{(\mathsf{e})}} R_k^{(\mathsf{e})}, \tag{24}$$

where the maximization is only over rate-tuples such that URLLC rates $\{R_{k,b}^{(\mathsf{U})} = R^{(\mathsf{U})}\}_{k\in\mathcal{K}^{(\mathsf{U})}}$ are achievable on each block $b \in \{1,\ldots,\mathsf{B}\}$ and eMBB rates $\{R_k^{(\mathsf{e})}\}_{k\in\mathcal{K}^{(\mathsf{e})}}$ are simultaneously achievable over the entire transmission, all using dirty-paper coding with an optimal precoding order and under an average block power constraint P.

The optimization problem in (24) is cumbersome to solve, and instead [60] proposes the following suboptimal algorithm. Fix the dirty-paper precoding order to first precode the eMBB messages followed by the URLLC messages. This implies that eMBB transmissions act as noise on the URLLC communication but not vice versa. Then choose a target URLLC rate $R^{(\mathsf{U})}$ and find the minimum required average block-power $\beta\mathsf{P}$, for $\beta \in [0,1]$, that ensures achievability of the per-block and per-user URLLC rate $R^{(\mathsf{U})}$. Identify finally the maximum sum-rate $\sum_{k\in\mathcal{K}^{(\mathsf{e})}} R_k^{(\mathsf{e})}$ achievable on the eMBB transmission with average power $(1-\beta)\mathsf{P}$.

Though optimal, dirty-paper coding is difficult to implement in practical systems and is often replaced by the simpler zero-force beamforming. In the context of our multi-user and mixed-delay communication scenario, under zero-force beamforming, it remains to determine the assignment of beams to users and the two communication types. The work in [60] proposes a sophisticated beam assignement algorithm, which assigns stronger sub-channels (beams) to URLLC messages, and weaker channels to eMBB messages. The idea being that eMBB communication can profit from channel diversity over multiple coherence blocks.

Numerical simulations in [60] compare the sum-rate in (24) achieved with dirty-paper coding and with a precoding order and power allocation established according to the suboptimal algorithm described above, with the sum-rate achieved with a beamforming alternative. For both schemes the maximum sum-rate increases with small values of $R^{(\mathsf{U})}$ and reaches a peak when the URLLC rate contributes approximately a third of the sum-rate. Beyond, the sum-rate decays rapidly because the delay constraint on the URLLC message becomes too stringent and limits the overall performance.

4.3. Summary

This section extended the superposition coding approach to fading BCs with multiple Rxs where certain Rxs demand URLLC messages and other eMBB messages. Assuming perfect channel state information, [60] proposes precoding orders or beam assignments for URLLC and eMBB messages, that take into account that URLLC messages have to achieve their desired rates in a single coherence block and therefore cannot exploit the channel diversity offered over multiple blocks. The results in [60] show that for small requested URLLC rates, the sum-rate of the system is not limited by the stringent delay constraint of URLLC messages. For larger URLLC rates this is however the case and URLLC delay constraints limit the overall performance.

5. Cooperative Interference Networks

5.1. Introduction

In this section we consider interference networks where Txs and/or Rxs can cooperate over dedicated cooperation links. This models for example cellular networks where BSs can cooperate over high-rate fiber-optic links, and neighbouring mobiles can cooperate using bluetooth or millimeter wave communication, which take place on different frequency bands than the standard radio communication between mobiles and BSs, and cause no interference.

Cooperation links between Txs are beneficial for eMBB transmissions, because they allow Txs to exchange parts of their messages or their signals so as to enable cooperative signaling over the channel. Cooperation links between Rxs can be used to exchange information about receive signals or decoded messages, allowing the Rxs to better mitigate interference. URLLC transmissions however have to start immediately after the creation of the messages and the additional delays caused by exchanging (parts of) URLLC messages between Txs cannot be tolerated. In the same sense, URLLC messages have to be decoded

before Rxs can learn information about other Rxs' decoded messages or receive signals. Figure 8 illustrates a typical timeline in our model. The actual communication time over the interference network is from time t_0 to time $t_0 + n$ and corresponds to the blocklength of communication. Here t_0 denotes an arbitrary starting time of a block and n refers to the block length. It is dedicated to the transmission of URLLC messages generated just prior to t_0 and of eMBB messages generated prior to $t_0 - \mathsf{D}_{\mathrm{Tx}} \cdot n$, so as to allow the eMBB messages to profit from D_{Tx} rounds of Tx-cooperation. (For simplicity it is assumed that n represents also the length of a cooperation round. The results also extend to scenarios where this is not the case.) Rxs decode the URLLC messages transmitted in this block $[t_0, t_0 + n]$ as soon as the block is terminated, each Rx simply based on its receive signal. Decoding of eMBB messages can be delayed to time $t_0 + (\mathsf{D}_{\mathrm{Rx}} + 1) \cdot n$, until the termination of D_{Rx} Rx-cooperation rounds.

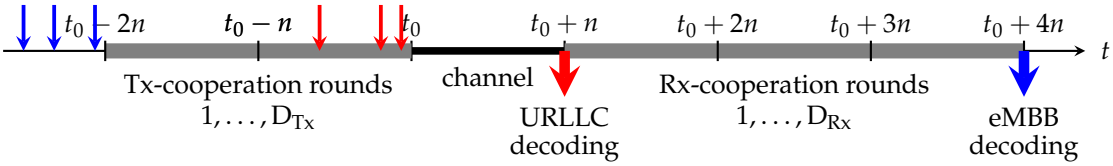


Figure 8. Timeline of cooperation and transmission over the interference network for URLLC and eMBB messages associated to the block from time t_0 to time $t_0 + n$.

Consider a scenario where each Tx sends an URLLC message and an eMBB message to its corresponding Rx. The focus is on the set of *degrees of freedom (DoF)* pairs that are simultaneously achievable for URLLC and eMBB messages, and in particular on how the sum-DoF decreases with increasing URLLC DoFs. This decrease describes the degradation of the overall system performance caused by the stringent delay constraints on URLLC messages, as a function of the URLLC rates. Somehow surprisingly, it can be shown that such a degradation does not exist in a variety of networks even with moderate or large URLLC DoFs.

In the following Section 5.2 we describe the problem setup. In Section 5.3, we present the integrated scheduling and coding scheme for URLLC and eMBB messages in [61], and in Section 5.4 we show that this scheme achieves maximum sum-rate even for moderate or large URLLC rates on a variety of network topologies, thus limiting the degradation of the overall system performance. In Section 5.5 we discuss a random-arrival model for URLLC and eMBB messages, where URLLC and eMBB messages are assigned to users according to some random arrival process. Again based on the coding scheme in [61], it can be shown that even under random arrival messages, the overall system performance is hardly degraded by the strict URLLC delay constraints [62,63].

5.2. Problem Description

Throughout this section, we consider a cellular network, but assume that users of the same cell are scheduled in different frequency bands. Interference thus occurs only from the mobile users in neighbouring cells that are scheduled on the same frequency band. The system therefore decomposes into subsystems with only a single mobile in each cell.

Consider thus an interference network with K cells, each consisting of a single Tx/Rx pair (i.e., a single mobile/BS pair). Networks have a regular interference pattern except at the network borders, with a focus on three different network topologies with short-range interference:

- Wyner's linear symmetric model in Figure 9a, where Txs and Rxs are aligned on two parallel lines and interference is only from the two Txs on the left and the right of any given Tx/Rx pair. This topology models for example situations in a corridor or along a railway line or highway where BSs are aligned. Cooperation links are present between neighbouring Txs and between neighbouring Rxs.

- Wyner's hexagonal model in Figure 9b, where cells are assumed of hexagonal shape. Interference is from the six neighbouring cells. Cooperation links are present between BSs and between mobiles of neighbouring cells.
- Sectorized hexagonal model in Figure 9c, where cells are again of hexagonal shape. In this model, Txs and Rxs use directed antennas, allowing us to divide each cell into three sectors with non-interfering communications, and interference is only from the neighbouring sectors in neighbouring cells, but not from sectors within the same cell. Here, a single mobile user is assumed in each sector, and thus three mobiles in each cell. Cooperation links are present between BSs of neighbouring cells and between mobiles in neighbouring sectors that are not in the same cell.

Each Tx $k \in [\mathsf{K}]$ wishes to convey a pair of independent URLLC and eMBB messages $M_k^{(\mathsf{U})}$ and $M_k^{(\mathsf{e})}$ to its corresponding Rx $k \in [\mathsf{K}]$. URLLC Message $M_k^{(\mathsf{U})}$ is of rate $R_k^{(\mathsf{U})}$ and eMBB message $M_k^{(\mathsf{e})}$ of rate $R_k^{(\mathsf{e})}$. The focus is on the average URLLC and eMBB rates

$$R^{(\mathsf{U})} := \frac{1}{\mathsf{K}} \sum_{k=1}^{\mathsf{K}} R_k^{(\mathsf{U})} \tag{25}$$

$$R^{(\mathsf{e})} := \frac{1}{\mathsf{K}} \sum_{k=1}^{\mathsf{K}} R_k^{(\mathsf{e})}. \tag{26}$$

Consider a cooperation scenario where neighbouring Txs cooperate during $\mathsf{D}_{\mathrm{Tx}} > 0$ rounds and neighbouring Rxs during $\mathsf{D}_{\mathrm{Rx}} > 0$ rounds. The total cooperation delay is constrained as

$$\mathsf{D}_{\mathrm{Tx}} + \mathsf{D}_{\mathrm{Rx}} \leq \mathsf{D}, \tag{27}$$

where $\mathsf{D} \geq 0$ is a given parameter of the system and the values of D_{Tx} and D_{Rx} are design parameters and can be chosen arbitrary such that (27) is satisfied. During the D_{Tx} Tx-cooperation rounds, each Tx can send arbitrary messages to its neighbours depending on the cooperation messages it received in previous rounds and on its eMBB Message $M_k^{(\mathsf{e})}$. In contrast, Tx-cooperation messages cannot depend on URLLC messages, as they are created only shortly before their transmission over the channel. The cooperative communication is assumed noise-free and the total cooperation load over all D_{Tx} Tx-cooperation rounds on each link is limited to $n \cdot \mu_{\mathrm{Tx}} / 2 \log(1 + \mathsf{P})$ bits. Each Tx forms then its channel inputs X_k^n as a function of all its received cooperation messages $\boldsymbol{T}_k$ and both its URLLC message $M_k^{(\mathsf{U})}$ and eMBB message $M_k^{(\mathsf{e})}$:

$$X_k^n = f_k^{(n)}\left(M_k^{(\mathsf{U})}, M_k^{(\mathsf{e})}, \boldsymbol{T}_k\right). \tag{28}$$

Channel inputs at each Tx are subject to an average block-power constraint P.
After receiving its channel outputs

$$Y_k^n = H_{k,k}\, X_k^n + \sum_{\hat{k} \in \mathcal{I}_k} H_{\hat{k},k}\, X_{\hat{k}}^n + Z_k^n, \tag{29}$$

where Z_k^n is i.i.d. standard Gaussian noise, matrix $\mathsf{H}_{\hat{k},k}$ models the channel from Tx $\hat{k}$ to Rx k, which is assumed to be constant during the duration of communication and known by all terminals, each Rx k immediately decodes its intended URLLC message:

$$\hat{M}_k^{(\mathsf{U})} = g_k^{(n)}\left(Y_k^n\right), \tag{30}$$

using some decoding function $g_k^{(n)}$ on appropriate domains. Following this first decoding step, neighbouring Rxs communicate with each other during D_{Rx} Rx-cooperation rounds. In each round, each Rx can send arbitrary messages to its neighbours that depend both on the previously received cooperation messages as well as on its output signals. The

cooperative communication is assumed noise-free, but its total communication load over all D_{Rx} Rx-cooperation rounds on each link is restricted to $n \cdot \mu_{\mathrm{Rx}}/2\log(1+\mathsf{P})$ bits. At the end of these D_{Rx} Rx-cooperation rounds, each Rx k decodes its desired eMBB message as

$$\hat{M}_k^{(\mathsf{e})} = b_k^{(n)}(Y_k^n, \boldsymbol{Q}_k), \tag{31}$$

where $\boldsymbol{Q}_k$ denotes all the Rx-cooperation messages received at Rx k and $b_k^{(n)}$ is an appropriate decoding function.

The focus of this section is on the *Degrees of Freedom (DoF) region* of the described model, i.e., on the set of possible pre-log factors $(\mathsf{S}^{(\mathsf{U})}, \mathsf{S}^{(\mathsf{e})})$ of URLLC and eMBB rates that are simultaneously achievable in the limit of infinite powers $\mathsf{P} \to \infty$:

$$\mathsf{S}^{(\mathsf{U})} := \lim_{\mathsf{P}\to\infty} \frac{R^{(\mathsf{U})}(\mathsf{P})}{\frac{1}{2}\log \mathsf{P}} \tag{32}$$

$$\mathsf{S}^{(\mathsf{e})} := \lim_{\mathsf{P}\to\infty} \frac{R^{(\mathsf{e})}(\mathsf{P})}{\frac{1}{2}\log \mathsf{P}}, \tag{33}$$

where the pairs $(R^{(\mathsf{U})}(\mathsf{P}), R^{(\mathsf{e})}(\mathsf{P}))$ need to be simultaneously achievable for given power P.

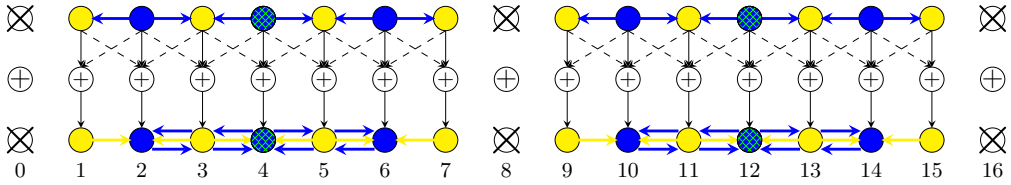
5.3. Coding Schemes

The following coding scheme was presented in [61]. All Txs in the network are scheduled to either send their URLLC message, their eMBB message, or no message at all. The scheduled eMBB and URLLC messages are then jointly transmitted using an integrated URLLC/eMBB coding scheme. Different schedulings can be envisioned to achieve fairness and send all the required messages. Scheduling is described by three sets $\mathcal{T}_{\text{silent}}$, $\mathcal{T}_{\mathsf{U}}$, and $\mathcal{T}_{\mathsf{e}}$, where

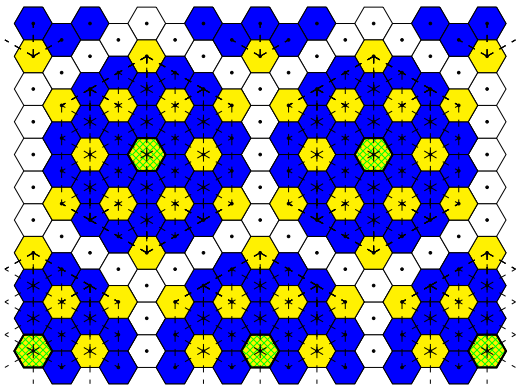
- Txs in $\mathcal{T}_{\text{silent}}$ are silenced and Rxs in $\mathcal{T}_{\text{silent}}$ do not take any action.
- Txs in $\mathcal{T}_{\mathsf{U}}$ send only URLLC messages. Tx/Rx pairs in $\mathcal{T}_{\mathsf{U}}$ are called *URLLC Txs/Rxs*.
- Txs in $\mathcal{T}_{\mathsf{e}}$ send only eMBB messages. Tx/Rx pairs in $\mathcal{T}_{\mathsf{e}}$ are called *eMBB Txs/Rxs*.

Figure 9 illustrates the choices of the $\mathcal{T}_{\text{silent}}$, $\mathcal{T}_{\mathsf{U}}$, and $\mathcal{T}_{\mathsf{e}}$ proposed in [61] for Wyner's linear symmetric network, the hexagonal model, and the sectorized hexagonal model when the maximum number of allowed cooperation rounds is either $\mathsf{D} = 6$ or $\mathsf{D} = 8$. White colour is used for Tx/Rx pairs in $\mathcal{T}_{\text{silent}}$, yellow colour for pairs in $\mathcal{T}_{\mathsf{U}}$, and blue colour for pairs in $\mathcal{T}_{\mathsf{e}}$. The set $\mathcal{T}_{\mathsf{U}}$ is chosen as large as possible so that URLLC transmissions are interfered only by eMBB transmissions and not by other URLLC transmissions.

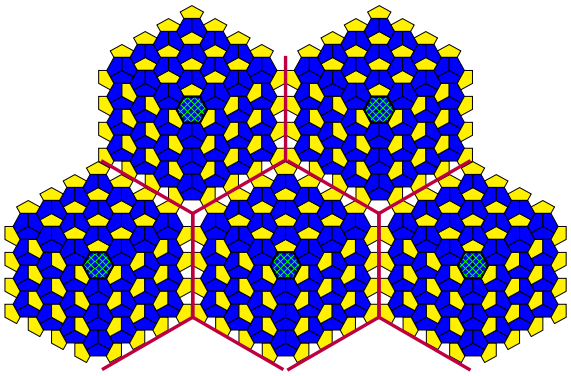
Consider the following joint coding scheme, which integrates both eMBB and URLLC messages. eMBB Txs describe quantized versions of their channel input signals during the Tx-conferencing phase to their neighbouring URLLC Txs, which then precancel the interference on their transmissions. URLLC Rxs can thus decode based on interference-free channels. After decoding, URLLC Rxs describe their decoded messages during the Rx-conferencing phase to the adjacent eMBB Rxs, so as to allow them to pre-subtract the interference from URLLC messages before decoding their intended eMBB messages. As a result, with the proposed scheduling and coding, URLLC messages can be decoded based on interference-free outputs and do not disturb the transmission of eMBB messages. For the transmission of eMBB messages, either CoMP transmission or CoMP reception is used, see Section 2.4, but only on subnets. In fact, with the choice of $\mathcal{T}_{\text{silent}}$ in Figure 9, the networks decompose into small subnets so that each subnet contains a master Tx/Rx that can be reached by any other Tx/Rx in the subnet with no more than $(\mathsf{D}-2)/2$ hops over the cooperation links. This ensures that CoMP transmission or reception in each subnet is possible with only $\mathsf{D}-2$ cooperation rounds. Since a single cooperation round is used to describe eMBB transmit signals to URLLC Txs and a single round is used to describe the decoded URLLC messages to eMBB Rxs, the scheme respects the maximum number D of total cooperation rounds.



(a) Wyner's symmetric network.



(b) Hexagonal network.



(c) Sectorized hexagonal model.

Figure 9. Message assignment for the various network models. White is used for $\mathcal{T}_{\text{silent}}$, yellow for $\mathcal{T}_{\mathsf{U}}$, and blue for $\mathcal{T}_{\mathsf{e}}$. Master Txs (Rxs) are in green pattern. Maximum number of allowed cooperation rounds $\mathrm{D} = 8$.

The coding scheme described above transmits both URLLC and eMBB messages. Variants thereof can be used to transmit only eMBB messages or only URLLC messages. More precisely, since any eMBB message can also be treated as a URLLC message (this would mean imposing stringent delay constraints also on some eMBB messages), the same scheme can also be used to send only eMBB messages. An alternative for sending only eMBB messages, is to silence again a set of Tx/Rx pairs and then directly employ CoMP transmission or CoMP reception on the set of non-silenced Txs/Rxs. Both schemes achieve the same DoF, but depending on the specific network they require larger or smaller cooperation rates μ_{Tx} and μ_{Rx}.

A simple way to send only URLLC messages is to choose a largest possible set of non-interfering Tx/Rx pairs and to silence all other Tx/Rx pairs. For Wyner's linear symmetric network this is optimal. For the two hexagonal models, and for certain channel coefficients, better performance is possible using the interference alignment techniques in [64].

5.4. Results on the Joint eMBB/URLLC DoF region

This subsection presents the achievable eMBB/URLLC DoF region achieved by the schemes in the previous subsection on the three network topologies in Figure 9, and compares them to the outer bounds derived in [61].

First consider Wyner's linear symmetric network. For this network and for sufficiently large cooperation prelog factors μ_{Tx} and μ_{Rx}, all DoF pairs $(\mathsf{S}^{(\mathsf{U})}, \mathsf{S}^{(\mathsf{e})})$ in the DoF region are achieved by the schemes described in the previous subsection or by time-sharing different versions thereof. The DoF region is given by the set of all DoF pairs $(\mathsf{S}^{(\mathsf{U})}, \mathsf{S}^{(\mathsf{e})})$ that satisfy

$$0 \leq \mathsf{S}^{(\mathsf{U})} \leq \frac{1}{2}, \tag{34a}$$

$$0 \leq \mathsf{S}^{(\mathsf{U})} + \mathsf{S}^{(\mathsf{e})} \leq \frac{\mathrm{D}+1}{\mathrm{D}+2}. \tag{34b}$$

One notices that the sum-DoF of the system is limited by the maximum number of allowed cooperation rounds D. Moreover, the stringent delay constraint on URLLC messages does not penalize the maximum achievable sum-DoF, which is equal to $\frac{D+1}{D+2}$, irrespective of $S^{(U)}$.

For smaller cooperation prelog factors μ_{Tx}, μ_{Rx} this is not the case, as can be seen in Figure 10, which shows inner and outer bounds on the DoF region derived in [61]. The inner bound is achieved by time-sharing the coding schemes in Section 5.3, and significantly improves over a pure scheduling approach that time-shares between a system sending only URLLC messages or only eMBB messages. We notice that for $\mu_{Rx} \geq 2.625$ and $\mu_{Tx} \geq 1.125$ the inner and outer bounds match. The inner bound is achieved by the schemes in Section 5.3 employing CoMP reception for eMBB messages. For $\mu_{Rx} \geq 1.125$ and $\mu_{Tx} \geq 2.25$ the inner and outer bounds also match, and are achieved by the same schemes, but employing CoMP transmission. When only one of the two cooperation prelogs μ_{Tx} or μ_{Rx} is large and the other small (e.g., $\mu_{Rx} \geq 4.5$ and $\mu_{Tx} = 0.5$; or $\mu_{Tx} \geq 4.5$ and $\mu_{Rx} = 0.5$) the inner bound matches the outer bound only for $S^{(U)}$ below a given threshold. For URLLC DoFs $S^{(U)}$ exceeding this threshold, the maximum eMBB DoF $S^{(e)}$ achieved by the schemes in Section 5.3 decreases linearly with $S^{(U)}$. For example, for $D = 6$ and $(\mu_{Rx} \geq 4.5, \mu_{Tx} = 0.5)$ or $(\mu_{Tx} \geq 4.5, \mu_{Rx} = 0.5)$, beyond this threshold, when one increases the URLLC DoF $S^{(U)}$ by Δ, then the eMBB DoF $S^{(e)}$ decreases by approximately 1.75Δ and the sum DoF by 0.75Δ. Yet another behavior is observed when both μ_{Rx} and μ_{Tx} are moderate or small, e.g., $\mu_{Rx} = 0.5$ and $\mu_{Tx} = 1$ or $\mu_{Rx} = 1$ and $\mu_{Tx} = 0.5$. In this case, the sum DoF achieved by the inner bound is not at its maximum value, but constant over all regimes of $S^{(U)}$. The overall performance of the system is thus again not limited by the stringent delay constraints on URLLC messages, but simply by the available cooperation rates.

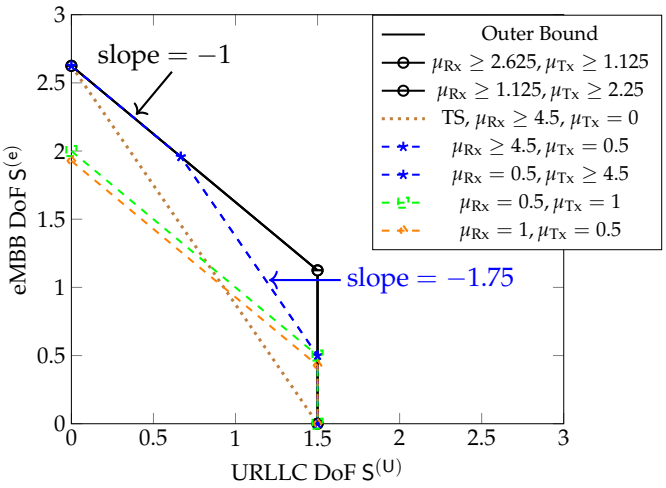


Figure 10. Bounds on DoF region for Wyner's symmetric model for different values of μ_{Rx} and μ_{Tx}, and $D = 6$. The brown dotted line represents the pure scheduling performance.

Figure 11 shows the inner and outer bounds on the DoF region proposed in [61] for the hexagonal model when $D = 8$ and for different values of μ_{Rx} and μ_{Tx}. Unlike in Wyner's symmetric model, the sum DoF achieved by the schemes in Section 5.3 always decreases as $S^{(U)}$ increases, irrespective of the cooperation prelogs μ_{Tx}, μ_{Rx}. Moreover, the maximum $S^{(U)} = \frac{1}{3}$ is only achieved for $S^{(e)} = 0$.

Figure 12 shows inner and outer bounds on the DoF region for the sectorized hexagonal model when $D = 4$. We notice that when both μ_{Rx} and μ_{Tx} are above given thresholds, $(\mu_{Tx} \geq 0.75, \mu_{Rx} \geq 2.25)$, then the combined scheme integrating both URLLC and eMBB messages in Section 5.3 simultaneously achieves maximum URLLC DoF and maximum sum-DoF. If only one of the two cooperation prelogs is very high but the other one small, the scheme achieves maximum sum-DoF only for small URLLC DoFs. The reason is that the

integrated scheme in Section 5.3 that jointly sends URLLC *and* eMBB messages inherently requires both Tx- *and* Rx-cooperation of sufficiently high cooperation prelogs, whereas Tx- *or* Rx-cooperation are sufficient for the scheme that sends only eMBB messages.

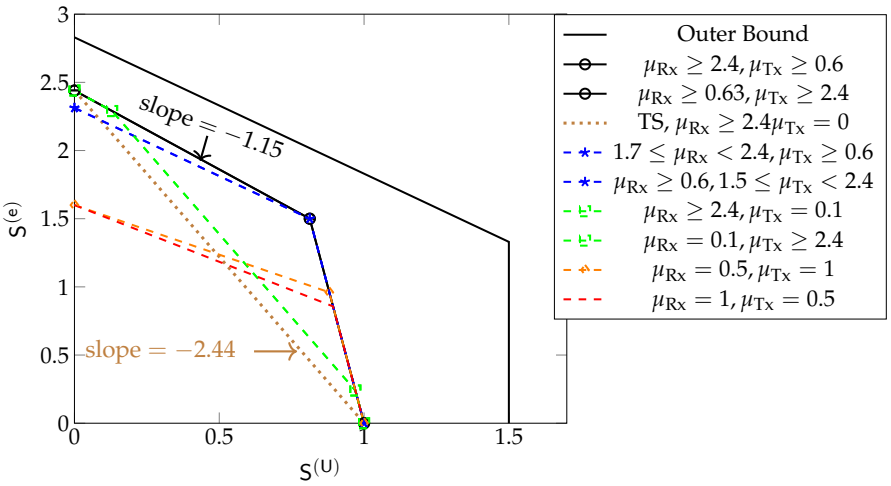


Figure 11. Inner and outer bounds on the DoF region for the hexagonal model for D = 8 and different values of μ_{Rx} and μ_{Tx}. The brown dotted line shows the pure time-sharing region.

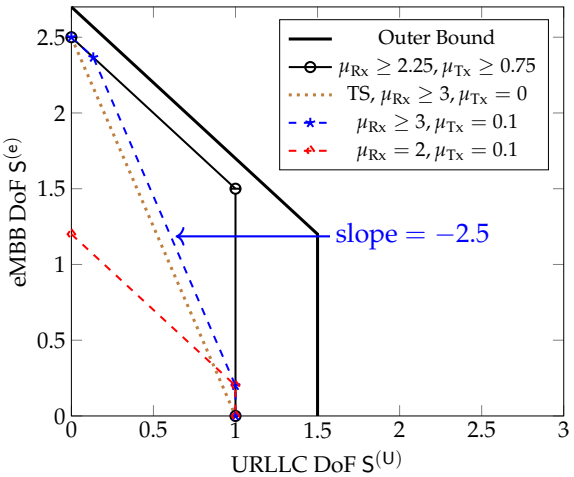


Figure 12. Inner and outer bounds on the DoF region for the sectorized hexagonal model for D = 4 and different values of μ_{Rx} and μ_{Tx}. The brown dotted line indicates the pure scheduling performance.

To summarize, for all three considered network models the joint scheme in Section 5.3 that integrates both URLLC and eMBB messages achieves maximum sum-DoF at high (or maximum) URLLC DoFs whenever the cooperation rates are sufficiently large. In this case, the stringent delay constraints on the URLLC messages do not harm the overall system performance. For smaller cooperation rates either the maximum sum-DoF is decreased or it is the same as with high cooperation prelogs but can only be achieved for small URLLC DoFs.

The described integrated coding scheme inherently requires at least a single cooperation round both at the Tx-side as well as at the Rx-side. The work in [65] also considered a scenario with only Rx- or only Tx-cooperation. It was shown that when only Rxs or only Txs can cooperate, then the ideal performance in (34) is not possible. Instead for sufficiently large cooperation rates the DoF region is given by the set of all rate-pairs $(S^{(U)}, S^{(e)})$ satisfying [65]

$$0 \leq 2S^{(U)} + S^{(e)} \leq 1, \tag{35a}$$

$$0 \leq \mathsf{S}^{(\mathsf{U})} + \mathsf{S}^{(\mathsf{e})} \quad \leq \quad \frac{\mathsf{D}+1}{\mathsf{D}+2}. \tag{35b}$$

The maximum sum-DoF is thus not decreased compared to a scenario with Tx- *and* Rx-cooperation. However, this maximum sum-DoF is only achievable for URLLC DoF $\mathsf{S}^{(\mathsf{U})} \leq \frac{1}{\mathsf{D}+2}$. We conclude that the stringent delay constraint inherently limits the overall system performance for moderate or large URLLC DoFs when only Rxs can cooperate. In fact, in this regime, increasing the URLLC DoF by Δ requires decreasing the eMBB DoF by 2Δ and the sum-DoF by Δ. Similar conclusions also hold for smaller cooperation prelogs and even in the non-asymptotic regime of finite powers [65].

5.5. Random User Activities

In practical systems, URLLC messages (and sometimes even eMBB messages) arrive in a random and bursty fashion and consequently in any given block, some Txs do not have an URLLC message to transmit. We consider the *random user-activity and random arrival model* proposed in [62], where each Tx is active with probability ρ, independent of all other Txs. If a Tx is active, it sends an eMBB message to its corresponding Rx, and moreover, with probability ρ_f, it also sends an additional URLLC message. Both the activity and arrival realizations are assumed to be known to all terminals in the network.

The DoF of *all* URLLC messages in the system is fixed and given by $\mathsf{S}^{(\mathsf{U})}$, whereas the *eMBB* DoF can vary over the variuos eMBB messages, and the quantity of interest is the *expected average* DoF $\mathsf{S}^{(\mathsf{e})}$ over all eMBB messages. (Similarly to the BC scenario in Section 4, the expected average eMBB DoF accounts for the possibility that eMBB messages are sent over multiple URLLC arrival blocks.) The same random-user activity model (but without mixed delays and random message arrivals) was already considered in [66–68], where it was observed that under this model the networks considered in Figure 9 decomposes into non-interfering subnets.

The same decomposition happens in the mixed-delay and random message arrivals model. As a consequence, an independent instance of the schemes in Section 5.3 should be applied to each subnet, where the schemes however have to be further adapted to the random URLLC message arrival situation. In particular, the scheduling (choices of sets $\mathcal{T}_{\text{silent}}, \mathcal{T}_{\mathsf{U}}, \mathcal{T}_{\mathsf{e}}$) needs to be adapted to the actual URLLC messages present in a subnet. The work in [62] proposes such a new scheduling approach, which on Wyner's linear symmetric network time-shares between a scheduling that sends URLLC messages at odd Txs and a second scheduling that sends URLLC messages at even Txs. (This allows us to achieve a symmetric URLLC DoF over all users having a URLLC message to send.) The scheduling for odd URLLC Txs is illustrated in Figure 13 for Wyner's linear symmetric model and specific realizations of the user activities and URLLC message arrivals. In the presented example, Txs 9, 12, 13 are inactive and Txs 1, 3, 7, 11, and 15 have an URLLC message to send. The network thus decomposes into three subnets: the first includes Tx/Rx pairs 1–8, the second includes Tx/Rx pairs 10–11, and the third includes Tx/Rx pairs 14–20. In each subnet, an independent instance of the integrated scheme of Section 5.3 is applied, but where the eMBB message at Tx 19 is treated as URLLC messages to comply with the scheme. In particular, Txs 8 and 20 are silenced because the maximum allowed cooperation delay equals $\mathsf{D} = 8$, and Txs/Rxs 4 and 16 act as master Txs/Rxs in the CoMP scheme.

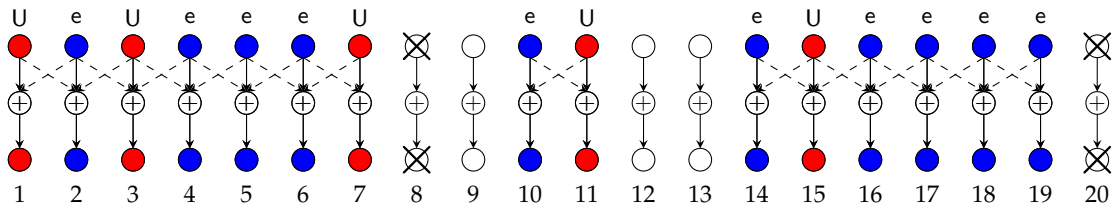


Figure 13. Wyner's symmetric linear network with random user activity and random arrival. $\mathsf{D} = 6$.

For sufficiently large cooperation rates, the approach in [62] achieves all DoF pairs $(\mathsf{S}^{(\mathsf{U})}, \mathsf{S}^{(\mathsf{e})})$ satisfying

$$\mathsf{S}^{(\mathsf{U})} \leq \frac{\rho\rho_f}{2}, \tag{36}$$

$$\mathsf{S}^{(\mathsf{e})} + M \cdot \mathsf{S}^{(\mathsf{U})} \leq \rho - \frac{(1-\rho)\rho^{\mathrm{D}+2}}{1-\rho^{\mathrm{D}+2}}, \tag{37}$$

where

$$M \triangleq 1 + \frac{(1-\rho)^2\rho^{\mathrm{D}+2}}{\rho\rho_f(1-\rho^{\mathrm{D}+2})} + \frac{(1-\rho)^2\rho^{\mathrm{D}+1}(1-\rho_f)^{\frac{\mathrm{D}}{2}}}{\rho\rho_f(1-\rho^{\mathrm{D}+2}(1-\rho_f)^{\frac{\mathrm{D}}{2}+1})}. \tag{38}$$

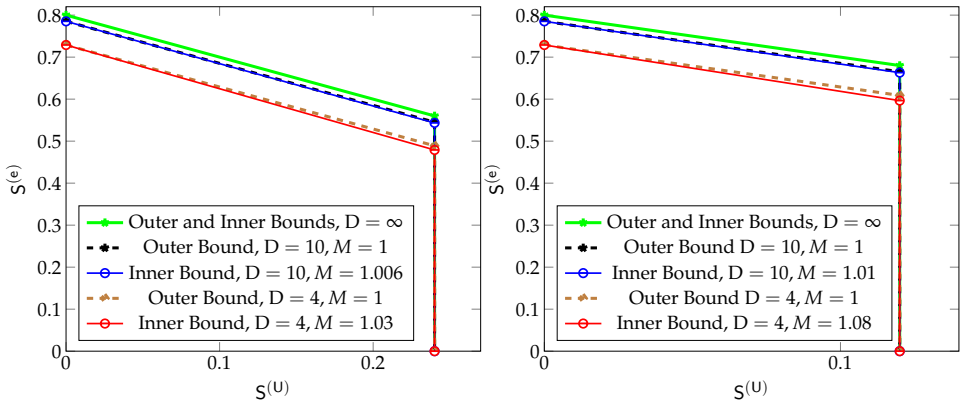
By means of an information-theoretic converse, it can be shown that all DoF pairs $(\mathsf{S}^{(\mathsf{U})}, \mathsf{S}^{(\mathsf{e})})$ not satisfying (36) and

$$\mathsf{S}^{(\mathsf{e})} + \mathsf{S}^{(\mathsf{U})} \leq \rho - \frac{(1-\rho)\rho^{\mathrm{D}+2}}{1-\rho^{\mathrm{D}+2}} \tag{39}$$

cannot lie in the DoF region. Constraints (36) and (39) thus provide an outer bound on the DoF region. Notice that this outer bound and the inner bound given by (36) and (37) only differ in the factors, $M > 1$ or 1, preceding the URLLC DoF $\mathsf{S}^{(\mathsf{U})}$ in the bounds (37) and (39), respectively. These factors are close whenever $\mathrm{D} \geq 10$, and as a consequence also the presented inner and outer bounds are close. For small values of ρ the factors are already close for $\mathrm{D} \geq 4$. Moreover, for small values of ρ and $\mathrm{D} \geq 4$, the right-hand sides of (37) and (39), are approximately equal to ρ, irrespective of D. This indicates that for small values of ρ, increasing the number of cooperation rounds D beyond 4 (and thus further increasing the delay of eMBB messages) does not improve the DoF region of the system. The reason behind this phenomenon is that a large number of cooperation rounds D is only useful in subnets with a large number of consecutive active Txs, and such subnets are very rare when the random user-activity probability ρ is small.

Notice further that by (36), the maximum URLLC DoF both in the inner and outer bounds is $\mathsf{S}^{(\mathsf{U})} = \frac{\rho\rho_f}{2}$, because each Tx sends an URLLC message with probability $\rho\rho_f$ and in the deterministic setup of Section 5.4, the maximum URLLC DoF is 1/2. Notice also that all bounds (36)–(39) increase with the activity parameter ρ. The maximum eMBB DoF in the inner and outer bounds is $\mathsf{S}^{(\mathsf{e})} = \rho - \frac{(1-\rho)\rho^{\mathrm{D}+2}}{1-\rho^{\mathrm{D}+2}}$. In the limit as $\mathrm{D} \to \infty$, it is thus given by ρ, and simply represents the expected fraction of active users. For finite D, the eMBB DoF decreases because to avoid interference to propagate, some of the Txs have to be silenced as in the scheme of Section 5.3. The term $\frac{(1-\rho)\rho^{\mathrm{D}+2}}{1-\rho^{\mathrm{D}+2}}$ thus describes the expected fraction of active but silenced Txs.

Figures 14, illustrate the inner and outer bounds in (36)–(39) for different values of ρ, ρ_f, and D. The most interesting part of the plots is the upper side of the trapezoids (the side lying opposite the two right angles). The slope of this line, which is -1 for the outer bounds and $-M$ for the inner bounds, describes the penalty in maximum eMBB DoF $\mathsf{S}^{(\mathsf{e})}$ incurred when one increases the URLLC Dof $\mathsf{S}^{(\mathsf{U})}$. Thus, on the outer bounds, increasing $\mathsf{S}^{(\mathsf{U})}$ by Δ decreases the maximum $\mathsf{S}^{(\mathsf{e})}$ by Δ and thus the sum DoF stays constant for all values of $\mathsf{S}^{(\mathsf{U})}$. On the inner bounds, the maximum eMBB DoF $\mathsf{S}^{(\mathsf{e})}$ is decreased by $M\Delta > \Delta$ when $\mathsf{S}^{(\mathsf{U})}$ is increased by Δ and the sum DoF thus decreases by $(M-1)\Delta > 0$.



(**a**) $\rho_f = 0.6$. (**b**) $\rho_f = 0.3$.

Figure 14. Inner and outer bounds on the DoF region for $\rho = 0.8$ and different values of D.

A similar model, but with only Rx-cooperation was studied in [63] for Wyner's soft-handoff model, Wyner's linear symmetric model, and the hexagonal model. Similarly to the setup with deterministic user activities and arrivals, with only Rx-cooperation but no Tx-cooperation the stringent delay requirement on URLLC messages harm the overall performance of the system and maximum sum DoF is only achieved when transmitting only eMBB messages. For networks with regular interference structures as in Figure 9 and when $\rho\rho_f \gg 1$, e.g., because URLLC messages are rare, it was shown in [63] that DoF pairs $(\mathsf{S}^{(\mathsf{U})}, \mathsf{S}^{(\mathsf{e})})$ satisfying the following equality are achievable

$$\mathsf{S}^{(\mathsf{e})} = \rho - (1 + \ell\rho)\mathsf{S}^{(\mathsf{U})}, \tag{40}$$

for $\mathsf{S}^{(\mathsf{U})} \leq \frac{\rho\rho_f}{\ell}$ and ℓ denoting the number of interference signals experienced at each Rx. (For example, for Wyner's linear symmetric network $\ell = 2$ and for the hexagonal model $\ell = 6$).

5.6. Summary

This section considered cooperative interference networks where only eMBB messages can profit from these cooperation links, but not URLLC messages because they have to be transmitted and decoded without further due. A general coding scheme was presented that manages to exploit the cooperation links for the transmission of eMBB messages in a way that allows us to attain the optimal overall performance (sum DoF) of the system despite the transmission of URLLC messages. Achieving this performance requires cooperation both at the Tx and Rx side, one of the two is not sufficient. Moreover, a careful scheduling of URLLC and eMBB messages to users had to be performed. In practice this scheduling is performed at a system level in the sense that applications randomly generate URLLC messages. The proposed scheme in [62] was adapted to such random and bursty arrivals of the URLLC messages, with only a small penalty in overall system performance.

6. Cloud Radio Access Networks (C-RANs)

6.1. Introduction

In this section we consider cloud radio access networks (C-RANs) where BSs are connected to a cloud processor via high-rate fronthaul links and in the uplink communication all transmitted messages are jointly decoded at the central processor so as to be able to alleviate the effect of interference [12]. URLLC messages are however not compatible with this new architecture as they have to be decoded directly at the BSs because communication over an additional hop to the cloud processor would violate their stringent delay constraints. As in the previous sections, we wish to investigate how this restriction affects the overall performance of the systems, and more specifically the sum-rates and rate pairs can simultaneously be achieved for URLLC and eMBB transmissions.

This section reviews two pieces of work on C-RAN with mixed-delay traffic. The work in [69], discussed in Section 6.2 considers a fading network and focuses on an information-theoretic discussion of the problem comparing inner and outer bounds on the fundamental performance limits of such systems. The work in [70,71], discussed in Section 6.3, takes a communication-theoretic approach. It decomposes communication in minislots and then compares performance of different communication strategies. In this latter work, the network is assumed static.

6.2. Fading C-RAN Model

The work in [69] considers the uplink of a C-RAN, and models the network from the mobile users to the BSs by an i.i.d. fading Wyner soft-handoff model, see Figure 15. That means, BSs are aligned on a line and each cell contains a single mobile user. This latter assumption stems again from the orthogonal frequency access applied by various mobile users in a cell. In Wyner's soft-handoff model, mobile users are assumed to be located on cell borders and thus interfere only on the communication in this neighbouring and closeby cell. At a given time $t \in [n]$, the signal received at any BS $k \in [\mathsf{K}]$ is thus described as

$$Y_{k,t} = G_{k,t}X_{k,t} + F_{k,t}X_{k-1,t} + Z_{k,t}, \tag{41}$$

where $X_{k,t}$ and $X_{k-1,t}$ are the signals sent by mobile users k and $k-1$ at time t; $\{Z_{k,t}\}$ are i.i.d standard Gaussian noise; and the sequence of channel coefficients

$$\left\{(G_{1,t}, G_{2,t}, \ldots, G_{\mathsf{K},t}, F_{1,t}, F_{2,t}, \ldots, F_{\mathsf{K},t})\right\}_{t=1}^{n} \tag{42}$$

is i.i.d. over time and distributed according to a given K-tuple distribution $P_{G_1 \cdots G_{\mathsf{K}} F_1 \cdots F_{\mathsf{K}}}$. This K-tuple distribution is the marginal distribution of a given stationary and ergodic process $\{(G_k, F_k)\}_{k=-\infty}^{\infty}$ satisfying $\mathbb{E}\left[|G_0|^2\right] < \infty$ and $\mathbb{E}\left[|F_0|^2\right] < \infty$. The fading coefficients $\{(G_{k,t}, F_{k,t})\}$ are known perfectly at BS k but not at the mobile users.

Each mobile user k sends both an URLLC message $M_k^{(\mathsf{U})}$ and an eMBB message $M_k^{(\mathsf{e})}$. It thus produces its channel inputs as $\mathbf{X}_k = f_k(M_k^{(\mathsf{U})}, M_k^{(\mathsf{e})})$, and so that they satisfy an average block-power constraint P. URLLC messages are directly decoded at the BSs based on the observed signals in (41). eMBB messages are decoded at the cloud processor, which perfectly observes the symbols sent over the fronthaul links by the BSs, where each BS k generates its symbols $\boldsymbol{L}_k$ by employing a compression function f_k to its observed outputs Y_k^n. The compression has to account for the capacity of the fronthaul links, which is assumed to be $C = \mu \frac{1}{2}\log(1+\mathsf{P})$ for each link, where μ is termed the fronthaul prelog.

For the described model, ref. [69] characterizes the set of all achievable average expected URLLC and eMBB DoFs (across users and fadings) as

$$2\mathsf{S}^{(\mathsf{U})} + \mathsf{S}^{(\mathsf{e})} \leq 1 \tag{43a}$$

$$\mathsf{S}^{(\mathsf{e})} \leq \mu. \tag{43b}$$

The entire DoF region can be achieved by a resource scheduling approach that time-shares URLLC and eMBB transmissions. Notice that the eMBB DoF is limited by the fronthaul prelog μ because all eMBB messages have to be decoded at the cloud center. For small fronthaul prelogs μ this restriction limits the DoF of the system, which can be improved by allowing BSs to also decode part of the eMBB messages.

As can be inferred from (43), the DoF region of the described C-RAN does not depend on the fading processes $\{F_{k,t}\}$ and $\{G_{k,t}\}$. This contrasts the behaviour at finite powers P where the set of achievable average URLLC and eMBB rates depends on the law of this fading process, as can be seen in Figure 16. The performance described in this Figure 16 is attained by a superposition coding scheme, and significantly improves over a pure scheduling scheme. We further notice from the figure that for small values of URLLC rates $R^{(\mathsf{U})}$, when $R^{(\mathsf{U})}$ increases by Δ, then the maximum eMBB rate $R^{(\mathsf{e})}$ decreases approximately

by the same amount Δ, and thus the sum-rate remains constant. For larger values of $R^{(\mathsf{U})}$, the maximum $R^{(\mathsf{e})}$ decreases approximately by 3Δ if $R^{(\mathsf{U})}$ is increased by Δ. Thereby the loss is larger for random than for static fading coefficients.

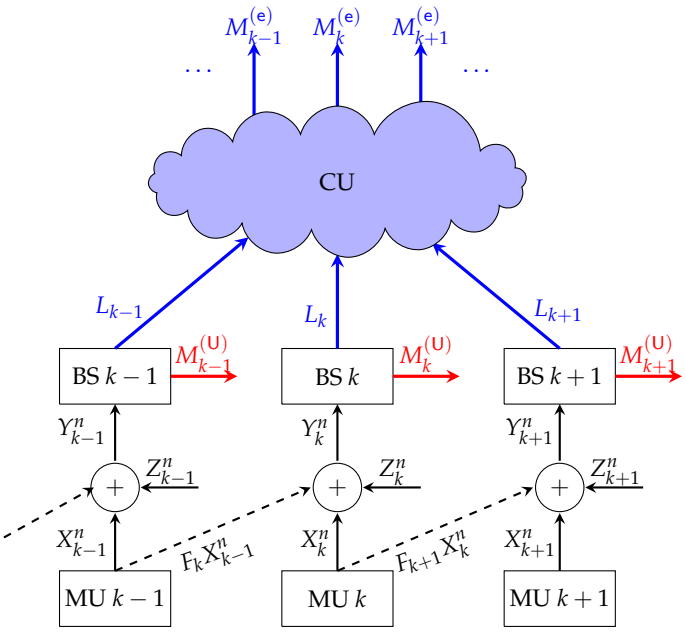


Figure 15. C-RAN with URLLC and eMBB transmissions and the mobile-to-BS network modeled by Wyner's soft-handoff model.

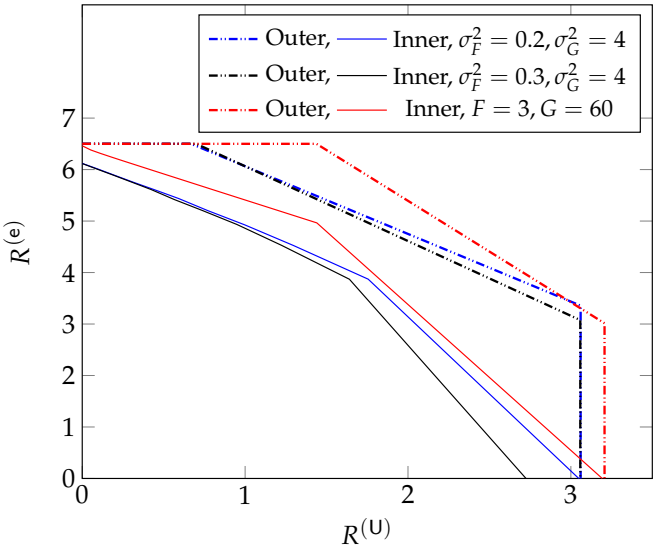


Figure 16. Inner and outer bounds on the achievable $R^{(\mathsf{U})}$ and $R^{(\mathsf{e})}$ rate region presented in [69] for $\mathsf{P} = 100$, $\mathsf{C} = 6.5$, Gaussian i.i.d. fadings of variances σ_F^2 and σ_G^2 or for constant non-time varying fadings $F = 3$ and $G = 60$.

6.3. Static CRAN Model with Slotted Communication

Mixed delay constraints in C-RANs were also studied in [70], where on a system-wide level communication is divided into minislots. In each minislot, each mobile user generates an URLLC message with probability q and attempts to send it over the network during the next minislot that is dedicated to URLLC communication. If a user generates multiple URLLC messages before the next URLLC minislot, it drops all but one URLLC message, which is then sent in this minislot. eMBB messages are sent over multiple minislots and

share the available resources in eMBB slots. Moreover, in [70], URLLC communication is assumed to be from mobile users close to the BSs, and their communication does not suffer from intercell-interference. eMBB users are assumed on the network border as in [69] and communication suffers from intercell interference as described by Wyner's symmetric model.

The performances of different coding schemes are compared in [70]. In all the schemes, eMBB messages are transmitted using standard multi-access codes, since they are jointly decoded at the cloud processor. URLLC messages are transmitted using standard Gaussian codebooks. If eMBB and URLLC messages are sent in the same slots, then eMBB communication (which lasts several minislots) is considered as noise in the decoding of URLLC messages.

- *OMA:* URLLC and eMBB messages are sent using *orthogonal multiaccess (OMA)*, i.e., a pure scheduling approach where the minislots dedicated for URLLC and eMBB communications are disjoint. Specifically, here every L_{U}-th minislot is dedicated for URLLC transmission and the other minislots for eMBB transmissions.
- *NOMA–puncturing:* eMBB and URLLC messages are sent using *non-orthogonal multiaccess (NOMA)*, i.e., eMBB and URLLC are sent over the same minislots. In particular, URLLC messages are transmitted in the minislot following their generation. To avoid this URLLC communication interfering with eMBB communication, BSs compress only the signals they receive in minislots where no URLLC communication is taking place from their corresponding mobile user.
- *NOMA–treating URLLC as noise:* eMBB and URLLC messages are sent using NOMA. URLLC communication is treated as noise for eMBB decoding. Therefore, BSs compress all their output signals, and send all this compression information to the cloud processor.
- *NOMA–SIC:* As in the previous item, except that BSs perform *successive interference cancellation (SIC)* on their decoded URLLC messages. That means, if URLLC decoding is successful, they subtract the URLLC signal from their outputs before compressing it for transmission to the cloud processor.

The performance of eMBB transmissions is measured by the information-theoretic rate that is achievable in the asymptotic regime of large blocklengths. URLLC transmissions are performed over single minislots and thus of much smaller blocklength. Their performances are measured using a finite blocklength rate-expression [58]

$$R_{\mathsf{U}} = \log(1+S_{\mathsf{U}}) - \sqrt{\frac{S_{\mathsf{U}}}{n(1+S_{\mathsf{U}})}}\mathcal{Q}^{-1}(\epsilon_{\mathsf{U}}), \tag{44}$$

where S_{U} denotes the interference power at a BS and ϵ_{U} has to be chosen sufficiently small so that the overall error probability (including the packet drops in case of OMA) does not exceed a desired threshold. Figure 17 compares the performances of these schemes in function of the URLLC message generation probability q. We observe that for the OMA approach performance degenerates quickly even for $L_{\mathsf{U}} = 2$ because the probability of URLLC message drop is too large. The URLLC performance is identical for all three NOMA approaches. The NOMA SIC approach achieves the best performance for the eMBB message.

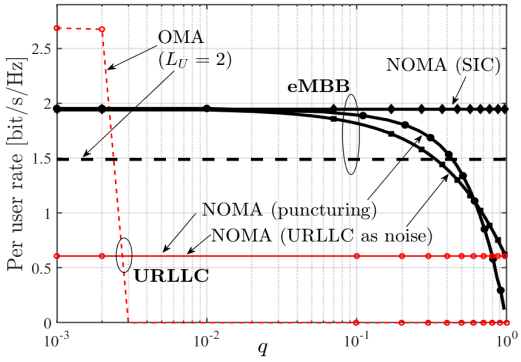


Figure 17. eMBB and URLLC per-user rates under OMA with L_U and NOMA for different decoding strategies as function of q in C-RAN [70].

In [71], these results are also extended to the downlink scenario. In this case, eMBB messages are created at the cloud processor and can profit from joint encoding to mitigate interference. URLLC messages are created directly at the BSs and their communications thus suffer from interference.

6.4. Summary

The last setup considered in this paper are C-RAN architectures where eMBB messages are jointly decoded at the cloud processor, whereas URLLC messages have to be decoded immediately at the BSs. Similarly to the P2P, BC, and cooperative network scenarios, for moderate powers, the overall system performance of the C-RAN with mixed-delay traffic in Section 6.2 decreases for large URLLC rates. This degradation seems to be more pronounced in fading environments than in static Gaussian environments. In the asymptotic high-power regime, however such a degradation is not observed, and at small URLLC DoFs $\mathsf{S}^{(\mathsf{U})}$, the sum-DoF is even increasing in $\mathsf{S}^{(\mathsf{U})}$. In certain scenarios it is thus possible to improve overall system performance by decoding part of the eMBB messages directly at the BSs and not at the cloud processor. Section 6.3 considers a non-fading environment and random generation of URLLC messages, for which it applies finite-blocklength performance measures. It is shown that for moderate or high URLLC generation rates, a NOMA scheme that first subtracts the contribution of the URLLC communication from the receive signals at the BSs, and then compresses and sends these differences over the fronhaul links, outperforms similar OMA and NOMA schemes.

7. Conclusions and Outlook

In this survey, we have reviewed joint coding schemes that integrate transmissions of URLLC and eMBB traffic and compared them to pure scheduling schemes in terms of rate, error probability and degree of freedom pairs that the schemes simultaneously achieve for URLLC and eMBB messages. A wide range of communication scenarios including P2P channels, BCs, cooperative interference networks, and C-RANs have been considered. The results have shown that joint coding schemes can significantly outperform the standard scheduling approach. As we have seen, in certain scenarios optimal system performance can be achieved under any URLLC rate. For other scenarios however, a large URLLC rate penalizes the overall system performance, showing that in these situations the stringent URLLC decoding constraint degrades system performance.

We conclude this survey with some lines of potential future research.

- The presented works have considered perfect channel state information (CSI) at the Rxs, and sometimes even at the Txs, where naturally CSI is more difficult to obtain. An interesting model for mixed-delay traffic is where CSI can be used for encoding and decoding of eMBB messages but not of URLLC messages [72,73]. The motivation behind such a model is that the processing of pilot and feedback signals required

to gather CSI at the Rxs and Txs introduces inadmissibly large delays for URLLC communication.

- So far, firm finite-blocklength results for mixed-delay traffic have been mostly limited to the P2P case; see [74] for an exception. Extensions to multi-user network scenarios is an important future research direction.
- In practical scenarios, both URLLC and eMBB messages are randomly generated by higher layer applications. This naturally leads to potential bottlenecks where not-yet-transmitted messages have to be buffered, similar to [75]. In this context, a thorough analysis of the behavior of the buffered contents and the required size of these buffers, is of significant practical interest.
- Other heterogeneous requirements on URLLC and eMBB traffic could be introduced in the study of mixed-delay traffic. For example, different security requirements as in [76] or different reliability constraints as in [21].
- Langberg and Effros [77] introduced the notion of *time-rate region* which describes the fraction of the blocklengths required for the transmissions of the various messages to the different Rxs in a network scenario under given message communication rates. A natural question is whether the interference mitigation techniques discussed in this survey can improve the inner bound on the time-rate region for general networks obtained in [77], which is obtained through a reduction to standard network information theory problems.
- Finally, mixed-delay traffic where different messages are transmitted over different blocklengths is inherently also connected to variable-rate and variable-length coding [20,77–79]. For example, the variable-rate channel coding framework of [78] includes the variable-to-variable scenario where depending on the specific system configuration and channel state realization, a receiver can decode a message of variable-size (similarly to the broadcast approaches in Sections III-B and IV) and decoding is performed after a variable number of channel uses. An interesting line of future research is to extend this scenario to multiple messages and mixed-delay traffic where URLLC and eMBB messages are decoded with different delays, and to study the four-dimensional tradeoff between URLLC and eMBB variable-rates and variable-delays.

Author Contributions: The authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by INRIA Nokia Bell Labs ADR "Network Information Theory", the French National Agency for Research (ANR) under grant ANR-16-CE25-0001-ARBURST, the European Union's Horizon 2020 Research And Innovation Program, grant agreements no. 715111, the US-Israel Binational Science Foundation (BSF) under grant BSF-2018710, and the U.S. National Science Foundation (NSF) within the Israel-US Binational program under grant CCF-1908308.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The works of H.N., M.E. and J.-M.G. have been supported by INRIA Nokia Bell Labs ADR "Network Information Theory" and by the French National Agency for Research (ANR) under grant ANR-16-CE25-0001-ARBURST. The work of M.W. has been supported by the European Union's Horizon 2020 Research And Innovation Program, grant agreements no. 715111. The work of S.S. (Shitz) has been supported by the US-Israel Binational Science Foundation (BSF) under grant BSF-2018710. The work of H.V.P. has been supported by the U.S. National Science Foundation (NSF) within the Israel-US Binational program under grant CCF-1908308.

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

References

1. Tataria, H.; Shafi, M.; Molisch, A.F.; Dohler, M.; Sjöland, H.; Tufvesson, F. 6G wireless systems: Vision, requirements, challenges, insights, and opportunities. *Proc. IEEE* **2021**, *109*, 1166–1199. [CrossRef]
2. Popovski, P.; Trillingsgaard, K.F.; Simeone, O.; Durisi, G. 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view. *IEEE Access* **2018**, *6*, 55765–55779. [CrossRef]
3. Popovski, P.; Stefanović, Č.; Nielsen, J.J.; de Carvalho, E.; Angjelichinoski, M.; Trillingsgaard, K.F.; Bana, A. Wireless access in ultra-reliable low-latency communication (URLLC). *IEEE Trans. Commun.* **2019**, *67*, 5783–5801. [CrossRef]
4. Ge, X. Ultra-reliable low-latency communications in autonomous vehicular networks. *IEEE Trans. Veh. Technol.* **2019**, *68*, 5005–5016. [CrossRef]
5. Shirvanimoghaddam, M.; Mohamadi, M.S.; Abbas, R.; Minja, A.; Yue, C.; Matuz, B.; Han, G.; Lin, Z.; Li, Y.; Johnson, S.; et al. Short block-length codes for ultra-reliable low latency communications. *IEEE Commun. Mag.* **2019**, *57*, 130–137. [CrossRef]
6. Zhang, X.; Wang, J.; Poor, H.V. Statistical QoS-driven energy-efficiency optimization for URLLC over 5G mobile wireless networks in the finite blocklength regime. In Proceedings of the Conference on Information Sciences and Systems, Baltimore, MD, USA, 20–22 March 2019; pp. 1–6.
7. Bairagi, A.K.; Munir, M.S.; Alsenwi, M.; Tran, N.H.; Alshamrani, S.S.; Masud, M.; Han, Z.; Hong, C.S. Coexistence mechanism between eMBB and URLLC in 5G wireless networks. *IEEE Trans. Commun.* **2021**, *69*, 1736–1749. [CrossRef]
8. Xiao, K.; Liu, X.; Han, X.H.; Hao, P.; Zhang, J.; Zhou, D.; Wei, X. Flexible multiplexing mechanism for coexistence of URLLC and eMBB services in 5G networks. *ZTE Commun.* **2021**, *19*, 82–90.
9. Bennis, M.; Debbah, M.; Poor, H.V. Ultrareliable and low-latency wireless communication: Tail, risk, and scale. *Proc. IEEE* **2018**, *106*, 1834–1853. [CrossRef]
10. Aleksandra, C.; Lehrmann, C.H.; Ying, Y.; Lara, S.; Georgios, K.; Stübert, B.M.; Lars, D. 6G wireless networks: Vision, requirements, architecture, and key technologies. *IEEE Veh. Technol. Mag.* **2019**, *14*, 28–41.
11. Aleksandra, C.; Lehrmann, C.H.; Ying, Y.; Lara, S.; Georgios, K.; Stübert, B.M.; Lars, D. Cloud RAN for mobile networks—A technology overview. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 405–426.
12. Peng, M.; Li, Y.; Zhao, Z.; Wang, C. System architecture and key technologies for 5G heterogeneous cloud radio access networks. *IEEE Netw.* **2015**, *29*, 6–14. [CrossRef]
13. Zhang, J.; Xu, X.; Zhang, K.; Zhang, B.; Tao, X.; Zhang, P. Machine learning based flexible transmission time interval scheduling for eMBB and uRLLC coexistence scenario. *IEEE Access* **2019**, *7*, 65811–65820. [CrossRef]
14. Bairagi, A.K.; Munir, M.S.; Alsenwi, M.; Tran, N.H.; Hong, C.S. A matching based coexistence mechanism between eMBB and URLLC in 5G wireless networks. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, New York, NY, USA, 8–12 April 2019.
15. Elsayed, M.; Erol-Kantarci, M. AI-enabled radio resource allocation in 5G for URLLC and eMBB users. In Proceedings of the IEEE 2nd 5G World Forum (5GWF), Dresden, Germany, 30 September–2 October 2019; pp. 590–595.
16. Alsenwi, M.; Tran, N.H.; Bennis, M.; Pandey, S.R.; Bairagi, A.K.; Hong, C.S. Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 4585–4600. [CrossRef]
17. Khan, H.; Butt, M.M.; Samarakoon, S.; Sehier, P.; Bennis, M. Deep learning assisted CSI estimation for joint URLLC and eMBB resource allocation. In Proceedings of the IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 7–11 June 2020; pp. 1–6.
18. Li, J.; Zhang, X. Deep reinforcement learning-based joint scheduling of eMBB and URLLC in 5G networks. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 1543–1546. [CrossRef]
19. Almekhlafi, M.; Arfaoui, M.A.; Elhattab, M.; Assi, C.; Ghrayeb, A. Joint scheduling of eMBB and URLLC services in RIS-aided downlink cellular networks. In Proceedings of the International Conference on Computer Communications and Networks (ICCCN), Athens, Greece, 19–22 July 2021; pp. 1–9.
20. Huleihel, W.; Steinberg, Y. Channels with cooperation links that may be absent. *IEEE Trans. Inf. Theory* **2017**, *63*, 5886–5906. [CrossRef]
21. Keresztfalvi, T.; Lapidoth, A. Semi-robust communications over a broadcast channel. *IEEE Trans. Inf. Theory* **2019**, *65*, 5043–5049. [CrossRef]
22. Keresztfalvi, T.; Lapidoth, A. Multiplexing zero-error and rare-error communications over a noisy channel. *IEEE Trans. Inf. Theory* **2019**, *65*, 2824–2837. [CrossRef]
23. Blackwell, D.; Breiman, L.; Thomasian, A.J. The capacities of certain channel classes under random coding. *Ann. Math. Stat.* **1960**, *31*, 558–567. [CrossRef]
24. Csiszár, I.; Narayan, P. Capacity and decoding rules for classes of arbitrarily varying channels. *IEEE Trans. Inf. Theory* **1989**, *35*, 752–769. [CrossRef]
25. Lapidoth, A.; Narayan, P. Reliable communication under channel uncertainty. *IEEE Trans. Inf. Theory* **1998**, *44*, 2148–2177. [CrossRef]
26. Morais, F.Z.; da Costa, C.A.; Alberti, A.M.; Both, C.B.; Righi, R.R. When SDN meets C-RAN: A survey exploring multi-point coordination, interference, and performance. *J. Netw. Comput. Appl.* **2020**, *162*, 102655. [CrossRef]
27. Zhang, S. An overview of network slicing for 5G. *IEEE Wirel. Commun.* **2019**, *26*, 111–117. [CrossRef]

28. Hossain, E.; Hasan, M. 5G cellular: Key enabling technologies and research challenges. *IEEE Instrum. Meas. Mag.* **2015**, *18*, 11–21. [CrossRef]
29. Sun, Y.; Peng, M.; Zhou, Y.; Huang, Y.; Mao, S. Application of machine learning in wireless networks: Key techniques and open issues. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3072–3108. [CrossRef]
30. Mao, Y.; Dizdar, O.; Clerckx, B.; Schober, R.; Popovski, P.; Poor, H.V. Rate-splitting multiple access: Fundamentals, survey, and future research trends. *arXiv* **2022**, arXiv:2201.03192.
31. Shariatmadari, H.; Iraji, S.; Jantti, R.; Popovski, P.; Li, Z.; Uusitalo, M.A. Fifth-generation control channel design: Achieving ultra reliable low-latency communications. *IEEE Veh. Technol. Mag.* **2018**, *13*, 84–93. [CrossRef]
32. Vaezi, M.; Azari, A.; Khosravirad, S.R.; Shirvanimoghaddam, M.; Azari, M.M.; Chasaki, D.; Popovski, P. Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road towards 6G. *arXiv* **2021**, arXiv:2107.03059.
33. Makki, B.; Chitti, K.; Behravan, A.; Alouini, M.S. A survey of NOMA: Current status and open research challenges. *IEEE Open J. Commun. Soc.* **2020**, *1*, 179–189. [CrossRef]
34. Vaezi, M.; Amarasuriya, G.; Liu, Y.; Arafa, A.; Fang, F.; Ding, Z. Interplay between NOMA and other emerging technologies: A survey. *IEEE Trans. Cogn. Commun. Netw.* **2019**, *5*, 900–919. [CrossRef]
35. Dai, L.; Wang, B.; Ding, Z.; Wang, Z.; Chen, S.; Hanzo, L. A survey of non-orthogonal multiple access for 5G. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2294–2323. [CrossRef]
36. Islam, S.M.R.; Avazov, N.; Dobre, O.A.; Kwak, K. Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges. *IEEE Commun. Surv. Tutor* **2017**, *19*, 721–742. [CrossRef]
37. Veeravalli, V.V.; Gamal, A.E. *Interference Management in Wireless Networks: Fundamental Bounds and the Role of Cooperation*; Cambridge University Press: Cambridge, UK, 2018.
38. Cover, T.M.; Tomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
39. Cover, T. Broadcast channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 2–14. [CrossRef]
40. Costa, M. Writing on dirty paper (Corresp.). *IEEE Trans. Inf. Theory* 1983, 29, 439–441. [CrossRef]
41. Erez, U.; Brink, S.t. A close-to-capacity dirty paper coding scheme. *IEEE Trans. Inf. Theory* **2005**, *51*, 3417–3432. [CrossRef]
42. Cohen, A.S.; Lapidoth, A. Generalized writing on dirty paper. In Proceedings of the IEEE International Symposium on Information Theory, Lausanne, Switzerland, 30 June–5 July 2002; p. 227.
43. Yu, W.; Kwon, T.; Shin, C. Multicell coordination via joint scheduling, beamforming, and power spectrum adaptation. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 1–14. [CrossRef]
44. Chen, J.; Mitra, U.; Gesbert, D. Optimal UAV relay placement for single user capacity maximization over terrain with obstacles. In Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications, Cannes, France, 2–5 July 2019; pp. 1–5.
45. Nikbakht, H.; Wigger, M.; Shamai, S. Multiplexing gain region of sectorized cellular networks with mixed delay constraints. In Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications, Cannes, France, 2–5 July 2019.
46. Levy, N.; Shamai, S. Clustered local decoding for Wyner-type cellular models. *IEEE Trans. Inf. Theory* **2009**, *55*, 4976–4985. [CrossRef]
47. Zhou, L.; Yu, W. Uplink multicell processing with limited backhaul via per-base-station successive interference cancellation. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 1981–1993. [CrossRef]
48. Simeone, O.; Somekh, O.; Poor, H.V.; Shamai, S. Local base station cooperation via finite-capacity links for the uplink of linear cellular networks. *IEEE Trans. Inf. Theory* **2009**, *55*, 190–204. [CrossRef]
49. Simeone, O.; Levy, N.; Sanderovich, A.; Somekh, O.; Zaidel, B.M.; Poor, H.V.; Shamai, S. Cooperative wireless cellular systems: An information-theoretic view. In *Foundations and Trends in Communications and Information Theory*; Now Publishers Inc.: Hanover, MA, USA, 2012; Volume 8, pp. 1–177.
50. Egan, M.; Collings, I.B. Low complexity quantization codebooks for CoMP. In Proceedings of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications, London, UK, 8–11 September 2013; pp. 1024–1028.
51. Cohen, K.M.; Steiner, A.; Shamai, S. The broadcast approach under mixed delay constraints. In Proceedings of the IEEE International Symposium on Information Theory, Cambridge, MA, USA, 1–6 July 2012; pp. 209–213.
52. Cohen, K.M.; Steiner, A.; Shamai, S. Broadcasting with mixed delay demands. In Proceedings of the IEEE 27th Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel, 14–17 November 2012; pp. 1–5.
53. Tajer, A.; Steiner, A.; Shamai, S. The broadcast approach in communication networks. *Entropy* **2021**, *23*, 120. [CrossRef]
54. Nikbakht, H.; Egan, M.; Gorce, J.-M. Joint channel coding of consecutive messages with heterogeneous decoding deadlines in the finite blocklength regime. In Proceedings of the IEEE Wireless Communications and Networking Conference, Austin, TX, USA, 10–13 April 2022.
55. Nikbakht, H.; Egan, M.; Gorce, J.-M. Dirty Paper Coding for Consecutive Messages with Heterogeneous Decoding Deadlines in the Finite Blocklength Regime. [Research Report] Inria–Research Centre Grenoble–Rhône-Alpes. Available online: https://hal.inria.fr/hal-03556888 (accessed on 1 February 2022).
56. Erseghe, T. Coding in the finite-blocklength regime: Bounds based on Laplace integrals and their asymptotic approximations. *IEEE Trans. Inf. Theory* **2016**, *62*, 6854–6883. [CrossRef]
57. Cohen, A.; Médard, M.; Shamai, S. Broadcast approach meets network coding for data streaming. *arXiv* **2022**, arXiv:2202.03018v1.

58. Polyanskiy, Y.; Poor, H.V.; Verdu, S. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [CrossRef]
59. Scarlett, J. On the dispersions of the Gel'fand—Pinsker channel and dirty paper coding. *IEEE Trans. Inf. Theory* **2015**, *61*, 4569–4586. [CrossRef]
60. Lin, P.H.; Lin, S.C.; Jorswieck, E.A. Early decoding for Gaussian broadcast channels with heterogeneous blocklength constraints. In Proceedings of the IEEE International Symposium on Information Theory, Melbourne, Australia, 12–20 July 2021; pp. 3243–3248.
61. Nikbakht, H.; Wigger, M.; Shamai, S. Coordinated multi point transmission and reception for mixed delay traffic. *IEEE Trans. Commun.* **2021**, *69*, 8116–8131. [CrossRef]
62. Nikbakht, H.; Wigger, M.; Shamai, S.; Gorce, J.-M. Cooperative encoding and decoding of mixed delay traffic under random-user activity. In Proceedings of the IEEE Information Theory Workshop, Kanazawa, Japan, 17–21 October 2021; pp. 1–6.
63. Nikbakht, H.; Wigger, M.; Shamai, S. Random user activity with mixed delay traffic. In Proceedings of the IEEE Information Theory Workshop, Riva del Garda, Italy, 11–14 April 2021.
64. Jafar, S.A. *Interference Alignment: A New Look at Signal Dimensions in a Communication Network*; Now Publishers Inc.: Hanover, MA, USA, 2011; Volume 7, pp. 1–134.
65. Nikbakht, H.; Wigger, M.; Shamai, S. Multiplexing gains under mixed-delay constraints on Wyner's soft-handoff model. *Entropy* **2020**, *22*, 182. [CrossRef]
66. Somekh, O.; Simeone, O.; Poor, H.V.; Shamai, S. The two-tap input-erasure Gaussian channel and its application to cellular communications. In Proceedings of the Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 23–26 September 2008.
67. Levy, N.; Shamai, S. 'Information theoretic aspects of users' activity in a Wyner-like cellular model. *IEEE Trans. Inf. Theory* **2010**, *56*, 2241–2248. [CrossRef]
68. Somekh, O.; Simeone, O.; Poor, H.V.; Shamai, S. Throughput of cellular uplink with dynamic user activity and cooperative base-stations. In Proceedings of the IEEE Information Theory Workshop, Taormina, Italy, 11–16 October 2009.
69. Nikbakht, H.; Wigger, M.; Hachem, W.; Shamai, S. Mixed delay constraints on a fading C-RAN uplink. In Proceedings of the IEEE Information Theory Workshop, Visby, Sweden, 25–28 August 2019.
70. 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 IEEE Global Communications Conference, Abu Dhabi, United Arab Emirates, 9–13 December 2018.
71. Kassab, R.; Simeone, O.; Popovski, P.; Islam, T. Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures. *IEEE Access* **2019**, *7*, 13035–13049. [CrossRef]
72. Wang, J.; Yuan, B.; Huang, L.; Jafar, S.A. GDoF of interference channel with limited cooperation under finite precision CSIT. *arXiv* **2019**, arXiv:1908.00703.
73. Chan, Y.; Wang, J.; Jafar, S.A. Towards an extremal network theory–robust GDoF gain of transmitter cooperation over TIN. *IEEE Trans. Inf. Theory* **2020**, *66*, 3827–3845. [CrossRef]
74. Mary, P.; Gorce, J.; Unsal, A.; Poor, H.V. Finite blocklength information theory: What is the practical impact on wireless communications? In Proceedings of the IEEE Global Communications Conference, Washington, DC, USA, 4–8 December 2016.
75. Steiner, A.; Shamai, S. On queuing and multilayer coding. *IEEE Trans. Inf. Theory* **2010**, *56*, 2392–2415. [CrossRef]
76. Zou, S.; Liang, Y.; Lai, L.; Poor, H.V.; Shamai, S. Degraded broadcast channel with secrecy outside a bounded range. *IEEE Trans. Inf. Theory* **2018**, *64*, 2104–2120. [CrossRef]
77. Langberg, M.; Effros, M. Beyond capacity: The joint time-rate region. *arXiv* **2021**, arXiv:2101.12236v1.
78. Verdu, S.; Shamai, S. Variable-rate channel capacity. *IEEE Trans. Inf. Theory* **2010**, *56*, 2651–2667. [CrossRef]
79. Shulman, N.; Feder, M. Static broadcasting. In Proceedings of the IEEE International Symposium on Information Theory, Sorrento, Italy, 25–30 June 2000.