Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)

Petar Popovski, *Fellow, IEEE*, Čedomir Stefanović, *Senior Member, IEEE*, Jimmy J. Nielsen, *Member, IEEE*, Elisabeth de Carvalho, *Senior Member, IEEE*, Marko Angjelichinoski, *Student Member, IEEE*, Kasper F. Trillingsgaard, *Student Member, IEEE*, and Alexandru-Sabin Bana, *Student Member, IEEE*

(Invited Paper)

***Abstract*— The future connectivity landscape, and notably, the 5G wireless systems will feature Ultra-Reliable Low Latency Communication (URLLC). The coupling of high reliability and low latency requirements in URLLC use cases makes the wireless access design very challenging, in terms of both the protocol design and of the associated transmission techniques. This paper aims to provide a broad perspective on the fundamental tradeoffs in URLLC, as well as the principles used in building access protocols. Two specific technologies are considered in the context of URLLC: massive MIMO and multi-connectivity, also termed *interface diversity*. This paper also touches on the importance of the proper statistical methodology for designing and assessing extremely high-reliability levels.**

***Index Terms*— Ultra-reliable communication, URLLC, IoT, 5G, access protocols, masssive MIMO, multi-connectivity.**

I. Introduction

DURING the past three decades, wireless connectivity has become a commodity, assumed to be practically always present and visible only when absent. This has naturally increased the confidence in wireless-enabled applications and services, leading to the idea of using wireless at a large scale to support mission-critical communication links. This trend has been termed *ultra-reliable communication (URC)* [1], where the level of connectivity guarantees, e.g. > 99.999 % of the time, matches the cable-based communication systems.

Ultra-reliability has inevitably become a part of the emerging 5G wireless systems. Indeed, 5G aims to cover three generic connectivity types: enhanced Mobile Broadband (eMBB), massive Machine-Type Communication (mMTC) and Ultra-Reliable Low-Latency Communication (URLLC). Obviously, in the context of 5G systems, ultra-reliability is entangled with the requirement for low latency. This makes URLLC very challenging, but also rather restrictive. In the earlier days of ultra-reliable wireless [1], there was a proposal to consider two types of ultra-reliable connectivity: *(i)* URC over a long term, in which the required latency is > 10 ms; *(ii)* URC in a short term, with latency of ≤ 10 ms. URC over a long term is interesting for use cases in which one needs *resilient* wireless connections, such as in disaster scenarios or remote interactions with a larger latency budget, e.g. changing a route of a drone. URC over short term contains URLLC[1] and is meant for applications with very stringent latency requirements, such as communication among machines and robots in Industry 4.0 use cases. However, while URLLC has been established as a concept in the community, URC over long term has been only scarcely present. We will therefore keep the focus in the paper on URLLC, noting that the insights about ultra-reliable connections and the communication-theoretic principles discussed here can be applied to URC defined over both short and long term.

Specifically, we will treat a set of fundamental problems in wireless access for URLLC. The objective is to provide the reader with a framework that can be used to analyze and design ultra-reliable wireless systems. Our previous article [2] can be seen as a predecessor of this work, where we have outlined the principles and the building blocks for wireless access in URLLC. This paper is intended to provide an in-depth treatment of some of the aspects and techniques associated with URLLC. We provide a detailed discussion on the communication-theoretic principles that are underpinning the design of URLLC. Compared to [2], we discuss in details the medium access control (MAC) protocols, use of large number of antennas in massive MIMO for providing high reliability, as well as the concept of interface diversity and multi-connectivity. We also address a fundamental question, largely ignored in the literature so far: what are the statistical requirements to measure and verify ultra-reliability? We note that in this paper we are not treating the details related to transmission of short packets, as these have been discussed to a sufficient level in [2] and [3].

The paper is organized as follows. The next section provides an overview of the URLLC use cases and their requirements,

Manuscript received October 15, 2018; revised March 10, 2019; accepted April 22, 2019. Date of publication May 3, 2019; date of current version August 14, 2019. The work has partly been supported by the European Research Council (ERC Consolidator Grant nr. 648382 WILLOW), partly by the Horizon 2020 project ONE5G (ICT-760809), and by the Danish Council for Independent Research (Det Frie Forskningsråd) DFF-701700271. The associate editor coordinating the review of this paper and approving it for publication was A. Nallanathan. *(Corresponding author: Petar Popovski.)*

The authors are with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: petarp@es.aau.dk; cs@es.aau.dk; jjn@es.aau.dk; edc@es.aau.dk; maa@es.aau.dk; kft@es.aau.dk; asb@es.aau.dk).

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

Digital Object Identifier 10.1109/TCOMM.2019.2914652

[1] URLLC is often associated with latencies of the order of 1 ms, such that 10 ms in this context is too long.

0090-6778 © 2019 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

creating the context for developing wireless access solutions. Section III elaborates on the communication-theoretic principles of URLLC, providing a perspective on the relationship between latency, packet size, bandwidth, as well as provides a finite-blocklength treatment of the problem. This is followed by Section IV on access networking, where a special emphasis is put on the problem of frame synchronization, a procedure that needs to have very high reliability in order to support packet decoding in URLLC scenarios. Section V sheds light on Massive MIMO, a technology that relies on extreme spatial diversity, which makes it a natural candidate for supporting ultra-reliable transmissions. Since the future URLLC devices are likely to have multiple communication interfaces, Section VI is dedicated to ultra-reliability achieved through multi-connectivity, i.e. interface diversity. Section VII treats the fundamental questions related to the statistical aspects of ultra-reliability. An overview of the state-of-the-art in the URLLC literature is provided in Section VIII. The last section concludes the paper and provides a perspective on some open issues.

II. URLLC Use Cases and Requirements

URLLC brings a significant novelty to 5G as a system. Along with mMTC, it makes 5G qualitatively different from the previous mobile wireless generations. Ultra-reliable communication is an potential enabler of a vast set of applications, some yet unknown. To put this in perspective, wireless connectivity and embedded processing have significantly transformed many products by expanding functionality and transcending the traditional product boundaries [4]. For example, a product stays connected to its manufacturer through its lifetime for maintenance and update. Ultra-reliable wireless brings this transformation to the next level, as the availability of a reliable wireless connectivity practically all the time is an important assumption that a system designer should account for when designing a system. For example, ultra-reliable wireless connectivity between two parts of a system removes the need for their physical attachment.

In general, the applications and the use cases of URLLC can be divided into two groups: *(i)* cable replacement and the extensions and *(ii)* native URLLC applications. The ones related to cable replacement are transforming some of the current applications that rely on cabled connections, but also add a new quality due to the flexibility of wireless. An example of this are the digital systems in Industry 4.0, where wireless will replace cabled connections, but also give rise to new types of interactions, e.g. among cooperative robots. On the other hand, a *native* URLLC application is the one that has no precedent in wired communication; an example is vehicle-to-vehicle (V2V) communication.

The general vision of URLLC requirements by 3GPP is presented in [5]:

- A reliability requirement of $1-10^5$ (i.e. 99.999 %) with a user-plane radio latency[2] of 1 ms for a single transmission of 32-byte long packet.
- An average user-plane radio latency of 0.5 ms for both uplink and downlink, without an associated reliability value.

We note that these requirements are solely covering the user-plane, meaning that the end-to-end latency and reliability requirements of an application should also take into account contributions from elements between the 5G network and the end devices.

The above figures, specified by 3GPP, are by far insufficient to describe the variety of use cases and the associated requirements of the verticals that 5G is envisioned to support, as discussed next. Furthermore, as discussed in Section VII, these specifications are also insufficient from a statistical viewpoint.

The automotive 5G URLLC cases represent an important segment of the ongoing 3GPP standardization and can be divided into *assisted*, *co-operative* and *tele-operated driving* [6], [7]. Their user plane reliability requirement is $1-10^5$ with the associated maximum end-to-end (E2E) latency requirement of 5 ms for assisted, 10 ms for co-operative, and 20 ms for tele-operated driving, both in the uplink and downlink. Note that, as a rule of thumb, radio latency can be estimated as 1/10 of E2E latency [6].

Another important set of URLLC use cases is related to monitoring and control of industrial processes, belonging to the emerging paradigm of Industry 4.0. The most important examples are *motion control*, *factory automation* and *process automation* [6], [8]. Motion control pertains to real-time control of machines with moving parts and is characterized by user-plane reliability of $1-10^{-5}$ with E2E latency of 1 ms (i.e. the user-plane radio latency of 0.1 ms). Moreover, this use case is about isochronous transmission of sensory and actuation information in the upink and downlink, respectively, requiring user-plane E2E jitter of 1 μs. Factory automation (also referred to as discrete automation or discrete manufacturing), according to 3GPP [8], requires user-plane reliability of $1-10^{-4}$ with user-plane E2E latency of 10 ms and jitter of 100 μs. However, in some other sources, this use case is characterized with an extreme reliability requirement of $1-10^{-9}$ or more [9]–[12], with a more demanding user-plane latency of 1 ms (for local monitoring and control setups) and 5 ms for (remote setups) and jitter of 1 μs [10]. Process automation, which is related to production of goods in bulk quantities, requires user-plane reliability of $1-10^{-6}$ and E2E latency of 50 ms and jitter of 20 ms, according to 3GPP [8]. Again, industrial sources aim at more stringent values that match the ones for the factory automation [9], [10].

We also mention the category of URLLC use cases that belong to the *tactile Internet*; their common feature is the existence of haptic feedback which puts the most stringent requirements in terms of reliability and latency. As an example, the haptic feedback in tele-surgery may require reliability of $1-10^{-9}$ and round-trip time as low as 1 ms [11].

The novelty of 5G is that reliability and latency are also explicitly involved in mMTC use cases, e.g. in monitoring of non-time critical process and logistics in the contexts of smart cities and factories [6], where user-plane reliability is set to 95 % with a maximum radio latency of 0.5 ms.

[2]Radio latency is measured from the moment of the reception of a packet by layer-2 radio protocol at the transmitting end to the moment of the delivery of the packet to the layer-3 protocol at the receiving end. We stress that the focus of the paper is on the URLLC principles and methods related to user-plane radio latency and reliability, unless otherwise explicitly stated.

Moreover, enhanced Mobile BroadBand (eMBB) service category also features general requirements of user-plane radio latency of 4 ms, both in the uplink and downlink [5]. These latency figures are lower than what 4G is able to provide, where the target user-plane radio latency is 10 ms [13]. We also note that reliability, as defined in 5G standardization, does not exist as a requirement in 4G. In summary, low latency and high reliability seem to be intrinsic to 5G, no matter the actual use case and service category.

Finally, we note that in this section we have focused only on the latency and reliability as the key performance parameters. More information about other performance parameters, such as availability, experienced data rates, payload sizes, as well as about deployment setups, security and other features, can be found in the references mentioned in this section.

III. Communication-Theoretic Principles of URLLC

The objective of this section is to introduce communication-theoretic considerations on the modeling and the fundamental tradeoffs in URLLC.

A. Communication-Theoretic Model

We will build our discussion of design principles and analysis based on the following baseband model of a received signal y

$$y = h\,\alpha\,x + z + w. \tag{1}$$

Here h is the channel coefficient, which in the general MIMO case is a matrix of channel coefficients; α is the activity indicator; x is the transmitted signal; z is the noise; and w is the interference. The value of the activity indicator is $\alpha = 1$ if there is an actual transmission x, otherwise, $\alpha = 0$. All variables h, α, x, z, w are random and contain uncertainty; however, the receiver wishes to learn only α and, if $\alpha = 1$, decode x. The knowledge about the other three variables h, w, z can be partial, statistical, or even non-existing. Let us take an initial look into the nature of these random variables; we will treat h, α and x in details throughout the paper. The most common random disturbance in communication systems is the noise z. The statistics of the noise are usually known and in the most common case is Gaussian, with a known noise power. Some of the most fundamental results in information theory, both in asymptotic case and in the case of packets with finite blocklength, are related to the Gaussian channel with known noise variance.

The situation is substantially different when the interference term w is considered. The knowledge about w depends on the part of the radio spectrum in which the bandwidth B is allocated. In a licensed spectrum, the license-owner pays in order to acquire the right to manage the interference in that spectrum. This does not mean that the interference is non-existent, but is turned into a *known unknown* and the spectrum owner can control or at least influence the interference and its statistics.

On the other hand, if the spectrum is unlicensed, then the statistics of w is largely unknown. Indeed, the open access to the unlicensed spectrum puts constraints on the way a given transmitter may operate, but does not limit the number of independently owned systems that can run in close proximity of each other.[3] The interference in unlicensed, but also sometimes in licensed bands, can be regarded as the most significant "unknown unknown" in the system model and one should use risk-based methods [14] to assess its impact for URLLC communication. This is elaborated in Section VII.

The knowledge of the channel h or at least its statistics is critical in URLLC systems. Even if we consider a non-coherent communication, where the receiver does not need to know or to learn h, the precise knowledge of the statistics of h is crucial to be able to guarantee a certain reliability of communication.

Finally, finding out x is the central task of each receiver and we will treat it throughout the whole paper. The level of knowledge about the activity of the transmitter α depends on the communication scenario. In downlink, the BS is the only transmitting candidate (except in a discovery process) and the receiving device expects to receive the signal, such that for this case we can take $\alpha = 1$. However, in uplink, in general, the BS does not a priori know whether the user is active, which translates into uncertainty about α. Finding out the values α for the users connected to the same BS is the access protocol problem, treated in Section IV, which is very much related to the ultra-reliable performance.

In the rest of the paper, we will treat in details various techniques and aspects of URLLC in the light of the model given by (1).

B. Relating Latency and Reliability

Latency can be defined in different ways and at different layers of the communication protocols. The simplest definition of a latency, treated in this paper, is the delay that a data packet experiences from the ingress of a given protocol layer at the transmitter to the egress of the same layer at the receiver. In applications related to, e.g. remote controls of robots or drones, one is interested in a two-way or round-trip delay.

Under the constraints of a URLLC service, the definition of reliability should be coupled to the latency requirement. In fact, one can say that, when the latency requirement is absent (theoretically infinite), then transmitting at a rate that is lower than channel capacity offers perfect reliability. From the perspective of an application, with a predefined latency constraint, i.e. deadline, we can define the *reliability* of a communication setup as the probability that the latency does not exceed this deadline, and *outage* as the probability that it does. Fig. 1 exemplifies this the generic requirement in terms of latency and reliability, applicable not only to point-to-point link, but also arbitrary communication setup. That is, the blue curve describes the probability of a link or setup to deliver packets within time x. The exact numbers on the deadline and the reliability are dependent on the application. We note that the latency cumulative distribution function (CDF) asymptote

[3] In other words, one can buy and turn on an arbitrary number of WiFi access points in a small space, e.g. room and set them up to transmit at different channels, thereby occupying the whole unlicensed spectrum.

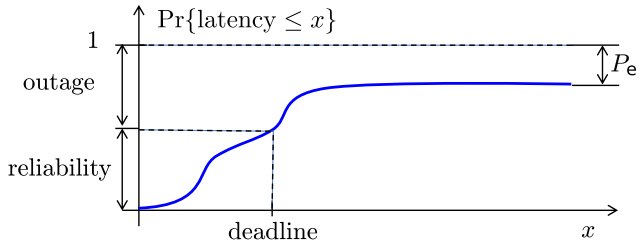


Fig. 1. Relation between outage, reliability, latency, and deadline.

is equal to $1 - P_e$, where P_e is the probability of residual packet loss or packet error. This residual packet loss reflects the fact that some packets will never be delivered due to, for example, limits on the number of retransmissions in link-layer protocols, buffer overflows, synchronization failures, etc.

The strong relation between reliability and latency means that introducing technology improvements with the aim to lower latency will cause the blue curve to have a steeper incline, and thereby result in improvement of reliability, since this is defined for a specific application deadline. Similarly, if the reliability is improved, for example by using multi-connectivity, the blue curve is stretched upwards, meaning that the required reliability level, e.g. 0.99999, can be guaranteed for a lower latency.

C. The Fundamental Tradeoffs and Packet Structure

As URLLC is often associated with transmission of controls and commands over wireless link in a distributed system, one of the basic assumptions about URLLC is that they involve small payloads. This naturally creates the association with the transmission of short packets [3] and the use of finite-blocklength information theory. It is instructive to look at the basic choices and tradeoffs that decide the packet length in an URLLC setting.

We consider a set of five variables: bandwidth, rate, reliability, energy, and latency. Let us at first fix the latency to T. Given the payload size of D bits and the maximal latency T, we can determine the minimal transmission rate R_{bps} in [bps]. Note that in the standard information-theoretic models, the data rate is expressed in terms of bits per channel uses [bpcu], here denoted simply by R. By selecting the bandwidth B, the number of channel uses available for transmission is $2BT$, such that the different types of data rates are related as follows:

$$R = \frac{D}{2BT} = \frac{R_{\text{bps}}}{2B} \quad [\text{bpcu}]. \tag{2}$$

The next variable is the energy, which, in general, refers to the total energy consumed for communicating the packet by the transmitter as well by the receiver, including all forms of diversity used in the system; e.g. this could be the energy invested by the receiver for using multiple antennas in order to harden the received signal. Let us fix the energy used for communicating the packet within the time T. The SINR at the receiver is determined by this energy, along with the channel realization and the interference, which are variables that cannot be chosen. With all these variables fixed, one can determine the achievable reliability of the transmission, denoted by $1-\epsilon$. In an analogous way, if other four variables are fixed, for example rate, reliability, energy and latency, then one can find what is the required bandwidth B.

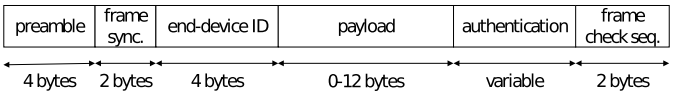


Fig. 2. Example of a packet format used in a low-throughput IoT system. The structure is lagerly inherited from the common packet structure used in broadband systems.

These, rather basic, considerations, are very important to get clarity in describing the models for URLLC. This is illustrated by the following two URLLC aspects:

(1) Given the latency T, the size of the packet blocklength in terms of available channel uses, equal to $N = 2BT$, can be regulated by selecting the bandwidth B. If the bandwidth available for transmission is very large, then the blocklength becomes very large as well. In other words, large bandwidth can move the transmission regime towards asymptotically large packet lengths; however, the data rate becomes very low and so does the spectral efficiency.

(2) Both the sender and the receiver use energy during the communication. Assume there is a single sender, Alice, and two possible receivers, Bob and Carol. If Alice sends to Bob, but not to Carol, then the activity indicator for Bob, denoted by α_B, see (1), is given by $\alpha_B = 1$. For Carol it should be $\alpha_C = 0$. If this is not known in advance, e.g. through pre-scheduling, then Bob and Carol should learn it from Alice's transmission, which requires spending some energy on detection, decoding and carrying out a hypothesis testing about the activity factors α_B and α_C. Alternatively, they can both always set $\alpha_B = \alpha_C = 1$ and receive anything that comes from Alice. In this case, only after decoding the packet, Bob and Carol figure out who is the intended recipient of the packet. This can improve the reliability, since α does not need to be decided separately, but it also increases the receiver energy consumption, as a receiver decodes packets that are not necessarily intended for him/her.

D. URLLC Packet Structure

As already mentioned, a general requirement of 3GPP for URLLC user-plane is a reliability of $1-10^{-5}$ (i.e. probability of error $\epsilon = 10^{-5}$) with latency of $T = 1$ ms for a transmission of a packet of size $D = 32$ bytes. In addition to the data payload, the packet should also contain signaling information/metadata.

An example of a short packet format, taken from low throughput networks [15], is depicted on Fig. 2. This figure illustrates that a significant portion of the packet is taken by metadata (end-device ID, authentication) and resources supporting auxiliary operations (preamble, frame synchronization, frame check sequence). Specifically, the auxiliary operations are synchronization, packet detection and packet integrity verification. As discussed in [2], when the reliability requirements are as high as in URLLC, one can no longer

assume that the transmission of the metadata and the auxiliary procedures are perfectly reliable. Indeed, the probability of success for a given packet π with a structure as the one on Fig. 2 is given by:

$$P_S(\pi) = P_S(A)P_S(M)P_S(D) \tag{3}$$

where $P_S(A), P_S(M)$, and $P_S(D)$ denote the success probability of the auxiliary procedures, metadata and data decoding, respectively. In other words, packet design that is based on separation of the resource for auxiliary procedures, metadata, and data leads to product of the success probability of the different elements, thus deteriorating the overall reliability.

We illustrate the impact of (3) for short packets. For simplicity, we assume that $P_S(A) = 1$, while the packet has $D = 16$ bytes of data and $M = 16$ bytes of metadata. The latency is set to $T = 1$ ms, while the bandwidth B should be determined such as to achieve the desired reliability of $1-10^{-5}$. For the sake of argument, we assume that B is lower than the coherence bandwidth of the system, such that a single coefficient describes the channel and the SNR achieved at the receiver. The channel coefficient is assumed to be known, such that the finite blocklength bounds for the complex AWGN channel hold. Hence, the error probability of receiving b bits of data within $N = 2BT$ channel uses when the SNR is given by γ is well approximated by [3], [16], [17]:

$$\epsilon(N, \gamma, b) = \mathrm{Q}\left(\frac{NC(\gamma) - b + \frac{1}{2}\log_2 N}{\sqrt{NV(\gamma)}}\right) \tag{4}$$

where $C(\gamma) = \frac{1}{2}\log_2(1+\gamma)$ and $V(\gamma) = \frac{\gamma(\gamma+2)}{2(\gamma+1)^2}\log_2^2 e$ denote the channel capacity and dispersion, respectively. Note that the SNR γ also depends on the bandwidth B and thereby on N, the number of available channel uses. Indeed, let us fix a reference SNR to γ_0 for some bandwidth B_0. Then, for equal transmission power (i.e. useful received power), the SNR when the bandwidth is B is given by

$$\gamma = \gamma_0 \frac{B_0}{B} = \gamma_0 \frac{2B_0 T}{N}. \tag{5}$$

We consider two cases:

1) Data and metadata are encoded jointly, the probability of correct packet reception is $(1 - \epsilon(N, \gamma, D + M))$;
2) Data and metadata are encoded separately and the probability of correct packet reception is $[1 - \epsilon(N/2, \gamma, M)][1 - \epsilon(N/2, \gamma, D)]$.

Fig. 3 shows the required bandwidth B as a function of the reference SNR γ_0 in order to achieve the required reliability with the prescribed latency. The reference SNR γ_0 in (5) is fixed to $B_0 = 100$ kHz. Clearly, when the data and the metadata are jointly encoded, the required number N of channel uses is lower compared to the case when they are encoded separately. This results in a lower required bandwidth, as the figure shows.[4]

The trick with increasing the blocklength in order to attain a more efficient transmission, given the reliability and the latency constraints, can also be used in a scenario in which a base station broadcasts to multiple terminals, as in [18]. Namely, given the total number of channel uses for broadcast, the BS can concatenate all packets intended for different users and use the resulting large packet as an input to the encoder. Intuitively, this offers the highest reliability, but the price is that each node needs to spend energy to decode data that it does not need, as it decodes the whole packet before seeing if there is any data intended for that node and, if yes, extract it. This tradeoff is analyzed in [18], and the technique has been termed *concatenate-and-code* in [19], which has extended the work towards cross-layer scheduling.

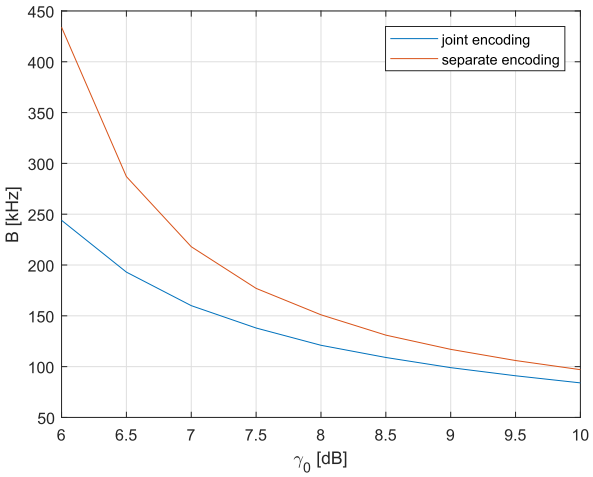


Fig. 3. Minimal required bandwidth to send 16 bytes of data and 16 bytes of metadata for the case of separate and joint encoding, respectively. The latency is set to $T = 1$ ms and the reliability to $1 - 10^{-5}$. Here γ_0 is the reference SNR for a bandwidth of $B_0 = 100$ kHz; the SNR for a bandwidth B is given by $\gamma_B = \gamma_0 \frac{B_0}{B}$.

Nevertheless, the same trick of increasing the blocklength by aggregating data cannot be used when different data chunks are transmitted by different nodes. In other words, if the BS has a data chunk for Alice and Alice has a data chunk from the BS, it is not possible to aggregate both data chunks in order to counter the effect of finite blocklength.

In the most "honest" case for URLLC transmission, the receiver does not have the channel state information. If a coherent transmission is about to take place over bandwidth that is larger than the coherence bandwidth, then the receiver should use pilots to estimate the channel and these pilots consume significant resources. Alternatively, the transmission can be carried out in a non-coherent way. Rigorous studies that take a holistic view on channel estimation and packet transmission are presented in [20], [21].

In summary, increased bandwidth is one of the most straightforward ways to bring diversity to the URLLC transmissions. There are two mechanisms that contribute to it, depending on the relationship between the transmission bandwidth B and the coherence bandwidth of the system B_c:

- $B \leq B_c$: In this case, the SNR of all channel uses/symbols is identical. If the data D to be transmitted is fixed, then reliability increases due to the lower rate per channel use (symbol). On the other hand, the transmission blocklength becomes longer, which improves the

[4]The bandwidth values shown on Fig. 3 are too large for a commonly seen values of coherence bandwidth and here they serve for illustrative purpose only.

transmission efficiency according to the results of the finite blocklength information theory.

- $B > B_c$: In this case the channel uses that are frequency-separated for more than B_c have different SNR statistics, which brings a new degree of diversity, in addition to the one brought by the increased number of channel uses.

Regarding the resource allocation in practical 5G systems, the tradeoffs arising in relation to the definition of time-frequency resources is captured in the flexible numerology used to design the 5G frames [22].

IV. Access Networking

The set of physical, MAC and link-layer protocols are referred to as *access networking*, which has the following tasks: *(i)* resolving the uncertainty in the user activity and inferring the value of α from (1); *(ii)* performing the auxiliary procedures, notably synchronization; *(iii)* decoding metadata and data; and *(iv)* interacting with the higher-layer protocols, where latency is ultimately measured and assessed. Its operation has to be designed according to the target reliability-latency requirements and to the traffic patterns of the supported services. In this section, we consider several generic access networking options for URLLC services.

A. Access Networking for URLLC Services With Deterministic Traffic Arrivals

URLLC services with deterministic traffic arrivals pertain to closed-loop control applications that involve deterministic sensing-actuation cycles with rather short periods and extreme latency-reliability requirements. The examples of such services, which in essence demand isochronous communications, are motion control and factory automation.

For deterministic traffic arrivals, the value of α in (1) is known a priori. A sensible access networking approach in this case is to employ periodic, pre-configured reservation of resources for data transmissions in both uplink and downlink, providing a deterministic timing of the traffic exchanges and, in this way, low-latency guarantees. This static allocation of the resources could be done offline, or using signaling exchanges that take place before the execution of the service starts.

The operation of access networking with static allocation of resources relies critically on precise time synchronization[5] among all involved network elements, which could be achieved using an external synchronization network, such as GPS. If such solution is not viable, due to, e.g. cost or indoor/obscured device location, one could employ synchronization methods that are reliant on dedicated signaling exchanges among the network elements. Nevertheless, achieving and maintaining such high level of synchronism in this way is a challenging task, particularly if the jitter requirements of the service are stringent, which typically is the case, see Section II.

[5]Time synchronization relates to the distribution of absolute time references [23].

The error probability of a packet exchange between a device and the base station in case of static allocation is given by

$$\epsilon_{\text{det}} = 1 - (1 - \epsilon_{\text{sync}})(1 - \epsilon_D)(1 - \epsilon_A) \tag{6}$$

where ϵ_{sync} is the synchronization error, and ϵ_D and ϵ_A are the probabilities that the data and the acknowledgement are not successfully decoded, respectively, which depend only on the choice of transmission parameters. Obviously, the used synchronization and the transmission methods have to ensure that $\epsilon_{\text{det}} < \epsilon_{\text{target}}$, where ϵ_{target} is the target error-probability of an acknowledged packet exchange.

B. Access Networking for URLLC Services With Stochastic Traffic Arrivals

In case of services with stochastic traffic arrivals, the value of α is not known a priori, and it is reasonable to consider options alternative to static allocation of the resources.

1) Four-Step Access: A four-step access procedure consists of the following steps:

1) The device sends a transmission request. The probability of error at the BS for this message is ϵ_R.
2) The device waits to receive access grant denoting the reserved resources; the error probability for this message is ϵ_G.
3) The device sends data in the uplink; the error probability for this transmission is ϵ_D.
4) The devices waits to receive an ACK, whose error probability is ϵ_A.

The overall probability of error for stochastic arrivals in a four-step procedure is

$$\epsilon_{\text{sto4}} = 1-(1-\epsilon_{\text{sync}})(1-\epsilon_R)(1-\epsilon_G)(1-\epsilon_D)(1-\epsilon_A). \tag{7}$$

In contrast to (6), the term ϵ_{sync} here refers to the initial synchronization and not the absolute time synchronization, as the latter is typically not required for URLLC services with stochastic arrivals due to their less stringent requirements. This initial synchronization is established through reception of the synchronization information via downlink broadcast channels. It contains carrier and frequency synchronization that will be exploited by the device for the subsequent uplink transmissions.[6]

For a given target overall error probability ϵ_{target}, meeting the requirement $\epsilon_{\text{sto4}} \leq \epsilon_{\text{target}}$ is more difficult than meeting $\epsilon_{\text{det}} \leq \epsilon_{\text{target}}$, as (7) contains contribution from more steps than (6). Moreover, exchanging four messages that in total contain more metadata than the two messages in static allocation, in principle, implicates higher latency. This is further aggravated by the fact that practical systems feature scheduling limitations, such that messages in the uplink and in the downlink cannot immediately follow each other. However, four-step access makes sense if $\Pr[\alpha = 1] \ll 1$, as it aims to support only the active devices and thus offers an overall better use of resources then the one with the static allocation.

[6]One could thus argue that the four-step procedure actually involves five steps, where the first step is the one related to successful reception of downlink broadcast information. We also note that similar considerations apply to the rest of the considered access procedures.

The error probabilities in four-step access ϵ_i, $i \in \{R, G, D, A\}$ depend on the choice of the transmission parameters, but also on some other aspects of the access procedure operation. Specifically, transmission requests are typically sent using a contention procedure among active devices; an important example is the contention-based random access that is used in LTE. In this case, a request is a randomly chosen access preamble (i.e., metadata), that is subject to the interference w that includes the requests of its contenders, potentially leading to collisions if two or more users choose the same preamble [24]. Thus, special provisions should be made to keep the value of ϵ_R low. A potential approach is to increase the number of resources for contention, and thus statistically decrease the interference generated by the other contenders. However, the standard contention algorithms, like slotted ALOHA, are rather resource-inefficient, especially if target collision probabilities are low. This calls for the application of contention procedures that are better suited to deal with interference, e.g. through use of successive interference cancellation [25], multi-packet reception [26], [27], etc.

Another approach to keep ϵ_R low is to statically allocate the resources for sending the transmission request, no matter whether the user is active ($\alpha = 1$) or not ($\alpha = 0$). An example of such access can be found in LTE contention-free random access, where the user is allocated a specific access preamble. This way, the user cannot experience collisions, but the approach lacks flexibility at the expense of other users that may wish to perform access.

Note that the successful reception of the request enables the BS to estimate the timing offset of the device and consequently instruct the device via grant message to compensate this offset in its subsequent data transmission. In other words, the estimation of the timing offset and its subsequent compensation effectively play the role of *frame synchronization*. In this respect, access requests typically include metadata that fosters estimation of the timing offset.[7]

The probability of not receiving an access grant, ϵ_G, depends on the correct reception of the access request, as well as on the fact whether BS has sufficient data resources to grant to all requests. The later can be influenced by the scheduling and resource allocation policy to other users and services.

For the sake of completeness, we note that the access procedures for the connection establishment in mobile cellular standards actually involve more than 4 steps before the data transmission can take place. For instance, in LTE connection-establishment [29] a device that wants to establish a connection and send data has first to successfully send a series of uplink messages with access request (in contention-based random access fashion) and other metadata[8] used for timing-offset estimation, device's identification and notification of the reason for connection establishment, security context establishment, etc. If a device is only sporadically active, the connection it establishes will become released. This implies that sporadically active devices will have to undergo the connection-establishment procedure each time they have to send data. Obviously, this represents a huge challenge from both latency and reliability perspectives, a topic that has attracted a lot of attention in the recent literature on efficient support of machine-type communications in cellular access [30].

2) Three-Step Access: In the three-step access, sending of a request is skipped and the BS sends directly the access grant to poll the device. In this case, the value of the activity indicator α becomes set to 1, no matter whether the device has experienced a new packet arrival or not. This mode of operation makes sense for services in which the devices are polled when their data becomes needed, or in which the BS can accurately predict when the device wants to send data, i.e. predict when α will change from 0 to 1. Note that an inaccurate prediction results in either resource waste, as resources are allocated when no URLLC message is pending, or an outage, when the resources are not allocated and the message expires until the next transmission opportunity.

The overall probability of error the three-step procedure is:

$$\epsilon_{\text{sto3}} = 1 - (1 - \epsilon_{\text{sync}})(1 - \epsilon_G)(1 - \epsilon_D)(1 - \epsilon_A). \quad (8)$$

In contrast to the four-step access and due to the lack of the timing offset estimation in the uplink, a correct reception of the data transmission in the three-step access is more challenging. Specifically, the BS has to detect where is the start of the data transmission in the dedicated resources, i.e. to acquire frame synchronization directly on the data transmission itself. This implies that the value of ϵ_D in three-step access is potentially higher than in four step access, but the overall latency of the access procedure is lower, given that the grants are timely sent to the devices. We turn to this problem in more details in Section IV-C.

3) Grant-Free Access: The 3-step procedure can be further decreased to a 2-step procedure, termed *grant-free access*, where the transmission of the grant by the BS is skipped. The first transmission is carried out by the device and it contains the actual data that should be sent during the access procedure. The probability of error in this case is:

$$\epsilon_{\text{sto2}} = 1 - (1 - \epsilon_{\text{sync}})(1 - \epsilon_D)(1 - \epsilon_A). \quad (9)$$

Although similar in form, the semantic difference between (9) and (6) is substantial. In case of grant-free access, the data transmission is by default contention-based and subject to potential interference from data transmissions of other devices. The lack of grant leads to a contention, such that ϵ_D in grant-free access is, in general, higher than in schemes employing grants or static allocation of resources, assuming that the same amount of resources and the same scheme is used for data transmission on both cases. The modest performance of the standard contention algorithms necessitates consideration of more advanced solutions that are able to deal with interference, see [25]–[27]. Moreover, similarly to the three-step access, the BS has to acquire frame synchronization. Nevertheless, there are at least two reasons to use grant-free access: (1) it decreases latency and (2) if the URLLC packets are very short, then the overhead brought by the request/grant is very significant and impacts the system efficiency.

[7]In LTE, the access preamble is an Zadoff-Chu sequence; Zadoff-Chu sequences feature favorable auto- and cross-correlation properties [28].

[8]Specifically, ten messages with metadata [30].

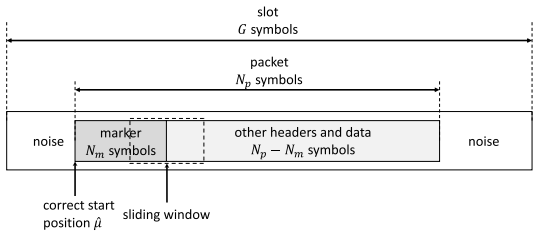


Fig. 4. Frame synchronization: the receiver should detect the correct value of $\hat{\mu}$.

We also remark that the assumptions for grant-free access can be further relaxed by not assuming prior synchronization between the base station and the devices in its domain. A prominent example of such approach is the pure ALOHA. However, when the reliability requirements are very stringent, this type of solutions is infeasible, in particular when the load/interference of the access is high.

C. Frame Synchronization

The task of the frame synchronization is to establish where is the start of the received packet. In cases with static resource allocation and network-wide time synchronization, this task is inherently addressed. In the following text, we briefly discuss achieving frame synchronization on the basis of the initial transmission from a user. We focus on the three-step and grant-free access, where this initial transmission contains data, and note that the task of frame synchronization in four-step access is similar, but can be much simplified, as the transmission request may be specifically designed just for frame synchronization purposes, like in LTE.

A common approach in frame synchronization is to place a *marker*, also known as *synchronization sequence*, at the beginning of the packet, and the task of the receiver is to detect the marker in the observation made in the slot, see Fig. 4. An alternative approach is to employ *blind frame synchronization*, which exploits other forms of redundancy contained in the packet, notably the knowledge of the channel coding algorithm, in order to establish frame synchronization. However, this approach suffers from higher complexity.

For the marker-based approach, the optimal detection algorithm depends on the channel model [31], [32]. In high signal-to-noise-ratio regimes, the optimal detection algorithm can be well approximated with the correlation between the locally generated marker and the subset of the received symbols that are placed in the sliding window. The length of this window is equal to length of the marker and it slides symbol-by-symbol through the slot, see Fig. 4. Upon performing correlations for all window positions, the receiver selects the one with the highest correlation value.

If the impact of noise can be neglected, then the receiver may miss the correct start of the frame if the marker happens to be generated in the rest of the packet by chance. To illustrate this, let us consider a high SNR scenario, in which BPSK modulation is used. Further, assume that the marker length is N_m bits, the total packet length is N_p bits, and that the bit values of 0 and 1 in the rest of the packet are equiprobable,

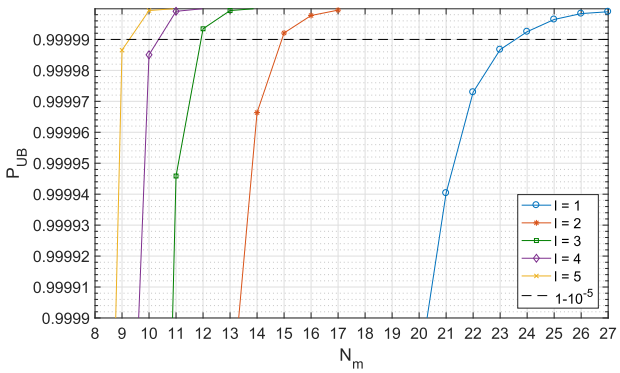


Fig. 5. Upper bound on the probability of correct frame synchronization as function of marker length N_m bits and of list length l, where packet length is $N_\mathrm{p} = 256 + N_\mathrm{m}$ bits.

which is a reasonable assumption. Note that this is different from the AWGN channel discussed in Section III-D, as here the input is set to be binary. The upper bound on the probability of correct frame synchronization can be computed as

$$P_\mathrm{UB} = \sum_i \frac{1}{i+1} \Pr\{C = i\} \tag{10}$$

where C is a random variable denoting the number of times that the marker is randomly reproduced by the packet symbols. In the expression for P_UB we have neglected the impact of the noise samples surrounding the packet, see Fig. 4. P_UB can be computed using the method presented in [33].

Fig. 5 shows P_UB (the line marked with circles) as function of the marker length N_m, assuming that $N_\mathrm{p} = N_\mathrm{m} + 256$ bits, i.e. the packet length is 32 bytes, not taking into account the marker. The presented results are obtained using the marker patterns that follow standard design guidelines [34], [35]. If one wants to achieve the correct frame-synchronization performance of $1 - 10^{-5}$, thus matching the standard URLLC reliability requirement [5], the marker length should be larger than 24 bits, even in the high SNR regime.

A compromise between complexity and performance can be achieved using a two-stage list-based synchronizer. The output of the first stage are l positions that are most probable to contain the marker, obtained using the marker detection algorithm. In the second stage, one of these l positions is selected using a metric that exploits the knowledge of the channel coding algorithm. Analogously to (10), we can calculate an upper bound of the correct frame synchronization related to the probability that the output of the first stage contains the correct frame start position:

$$P_\mathrm{UB}(l) = \Pr\{C < l\} + \sum_{i \geq l} \frac{l}{i+1} \Pr\{C = i\}. \tag{11}$$

Fig. 5 presents the results for increasing values of l, also obtained via the method from [33]. It can be seen that, for a fixed value in P_UB, the increase in l allows to decrease N_m.

Finally, we also mention that frame synchronization can be achieved by using markers whose symbols belong to different alphabets in comparison to the symbols carrying data in the rest of the packet. For example, an option could be to

use Zadoff-Chu sequences as markers, cf. [36]. Nevertheless, the results concerning the marker lengths and/or receiver complexity, presented above, also hold qualitatively in those cases.

V. URLLC in Massive Multi-Antenna Systems

A. The Benefits of Massive Multi-Antenna Systems for URLLC

Multiple antennas at the base station or terminals of a wireless network provide efficient mechanisms at the physical layer to ensure reliable and low latency communications. They offer a powerful complement to the higher layer methods described in this paper. This section focuses on massive antenna systems, characterized by a very large number of antennas at the BS and, possibly at the terminals, at high frequency bands, that have clearly emerged as a major enabler towards the creation of 5G wireless networks [37]. They are largely viewed as essential in magnifying the data rates and/or increasing the number of broadband users that can be simultaneously multiplexed within the same bandwidth. However, they are also fundamental tools in building the two other 5G services, i.e. massive machine type communications [38] and URLLC [39], [40].

The benefits of massive antenna systems lie in their ability to create a very large number of spatial Degrees-of-Freedom (DoF), which determine the following remarkable properties that are beneficial for URLLC:

1) *High SNR links.* This property is due to the array gain.
2) *Quasi-deterministic links, practically immune to fast fading.* This property is rather specific to systems operating below 6 GHz in a rich scattering environment [41]. It is a result of the *channel hardening* phenomenon. Along with the first property, it relaxes the need for strong coding schemes, hence maintaining high reliability for short packets. This can dramatically reduce the need for retransmissions.
3) *High capability for spatial division multiplexing.* In a multi-user system, this property can be exploited to improve the latency incurred due to multiple access, as multiple users can exchange data simultaneously. However, it should be noted that the multi-antenna processing employed to separate the users might induce additional computational delay [42].

This section is dedicated to the exploitation of multiple antennas at the transmitter and receiver to support URLLC. At first, we need to establish the fact that the acquisition of the instantaneous CSI is one of the most severe limitations with respect to URLLC when exploiting multiple antennas; see Section VIII. This is because the CSI acquisition is a major protocol step in massive MIMO, impacting both the reliability and the latency. Taking this into account, we devise beamforming methods that rely mostly on the structure of the channel, that is, the direction of the propagation path. The information about small scale fading is exploited as little as possible. As the structure of the channel varies on a large scale basis, its acquisition is more robust to device mobility.

It should be noted that the basic idea of using the singular vectors of the channel [43] or the structure of the channel [44], [45] (singular vector of the covariance matrix or steering vectors) to build multi-user transceivers is not new. Here we show that this basic idea creates a good basis to build URLLC transmission schemes.

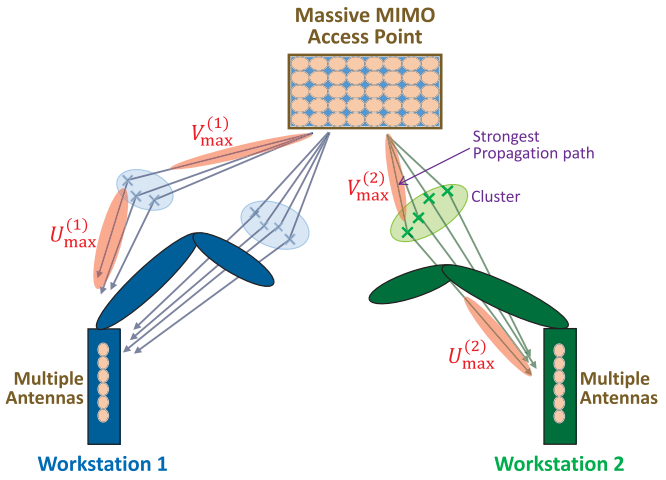


Fig. 6. Factory scenario where a massive MIMO access points serves multiple terminals (workstations).

B. Channel Structure

To illustrate the main concepts of this section, we assume a factory-type environment as pictured in Fig. 6. An access point is equipped with an array that consists of a very large number of antennas while the terminals (workstations) are equipped with one or possible small number of antennas.

Furthermore, we consider the simplified case of two terminals, each receiving a single stream of data from the access point. The results can be easily extended to the general case. We adopt a cluster-based channel model where each cluster is characterized by a group of localized propagation paths defined by their direction of departure and their direction of arrival. Each propagation path is affected independently by an attenuation factor that follows a certain distribution. The channel from the access point to terminal k is described as the sum of the propagation paths over all the clusters (in the sum, we make no distinction between clusters):

$$\mathbf{H}^{(k)} = \sum_{i=1}^{N_P^{(k)}} \alpha_i^{(k)} \mathbf{s}_{i,\mathrm{rx}}^{(k)} \mathbf{s}_{i,\mathrm{tx}}^{(k)H}. \tag{12}$$

$N_P^{(k)}$ is the total number of paths. $\mathbf{s}_{i,\mathrm{tx}}^{(k)}$ and $\mathbf{s}_{i,\mathrm{rx}}^{(k)}$ are the normalized steering vectors which characterize respectively the direction of departure from the BS and direction of arrival to the terminal. When a terminal is equipped with a single antenna, we have $\mathbf{s}_{i,\mathrm{tx}}^{(k)} = 1$. The direction of the propagation paths correspond to long-term statistics, meaning that for a localized movement of the terminal the directions remain unchanged, whereas the coefficients $\{\alpha_i^{(k)}\}$ correspond to small scale fading and vary for small movements.

C. Covariance-Based Design

In order to promote reliability and low latency, the general purpose of the beamforming design is to rely as much as

possible on the structure of the channel (i.e. the propagation path) and as less as possible on the small scale fading properties, while still benefiting from the properties brought by the massive number of antennas. This, in general, is a non-trivial task as those properties are brought by coherent combining of the signals from each antenna, while this combining depends on the small-scale fading.

We adopt a design based on the covariance matrix of the signal of each terminal at the transmitter and receiver. Those covariance matrices reflect the structural properties of the channel.

The singular value decomposition of the covariance matrix at the transmitter for terminal k is:

$$\mathbf{R}_{\text{tx}}^{(k)} = \mathbf{V}^{(k)}\Lambda^{(k)}\mathbf{V}^{(k)H}. \tag{13}$$

The columns of $\mathbf{V}^{(k)}$ comprise the singular vectors denoted as $\mathbf{v}_i^{(k)}$ and $\Lambda^{(k)}$ is a diagonal matrix grouping the non zero singular values. Likewise, we write the covariance matrix at terminal k as:

$$\mathbf{R}_{\text{rx}}^{(k)} = \mathbf{U}^{(k)}\Lambda^{(k)}\mathbf{U}^{(k)H}. \tag{14}$$

The singular vectors of $\mathbf{R}_{\text{rx}}^{(k)}$ are denoted as $\mathbf{u}_i^{(k)}$. The singular vectors associated to the maximal singular value $\mathbf{R}_{\text{tx}}^{(k)}$ and $\mathbf{R}_{\text{rx}}^{(k)}$ are $\mathbf{v}_{\max}^{(k)}$ and $\mathbf{u}_{\max}^{(k)}$.

D. Transceiver Structures

We now examine zero-forcing beamforming designs. They are based on the following principles:

1) **Inter-terminal properties**: The inter-terminal interference can be removed based solely on the singular vectors of the covariance matrix of the interfering terminals defining their signal subspace. Interference is eliminated by projecting the transmitted signal into the space orthogonal to the signal subspace of the interferers. This is advantageous in URLLC as this operation does not depend on instantaneous CSI.

2) **Intra-terminal properties**: Once the inter-terminal interference is removed, transmission to a single terminal might exploit several levels of CSI knowledge at the transmitter.

A general form of the zero-forcing precoder for terminal 1 is as:

$$\mathbf{F}_{\text{ZF}}^{(1)} = \underbrace{P_{\mathcal{V}^{(2)}}^{\perp}}_{\text{Term 1}} \underbrace{\bar{\mathbf{V}}^{(1)}}_{\text{Term 2}} \underbrace{\mathbf{w}}_{\text{Term 3}} \tag{15}$$

Term 1. The first term forces the precoded signal to lie in the signal subspace orthogonal to terminal 2. $\mathcal{V}^{(2)}$ contains the singular vectors of the covariance matrix at the transmitter for terminal 2 associated to non-zero singular values.

Term 2. The columns of the matrix in the second term defines the subspace of the transmit covariance matrix of terminal 1 where the signal of interest lies. $\bar{\mathbf{V}}^{(1)}$ contains the singular vectors of the covariance matrix associated to non-zero singular values.

Term 3. Vector $\mathbf{w}$ defines a linear combination of the columns of $\bar{\mathbf{V}}^{(1)}$. In general, this is a coherent operation requiring the knowledge of the channel projection onto the columns of $\bar{\mathbf{V}}^{(1)}$.

Note that term 1 and term 2 depend only on the long term statistics of the channel.

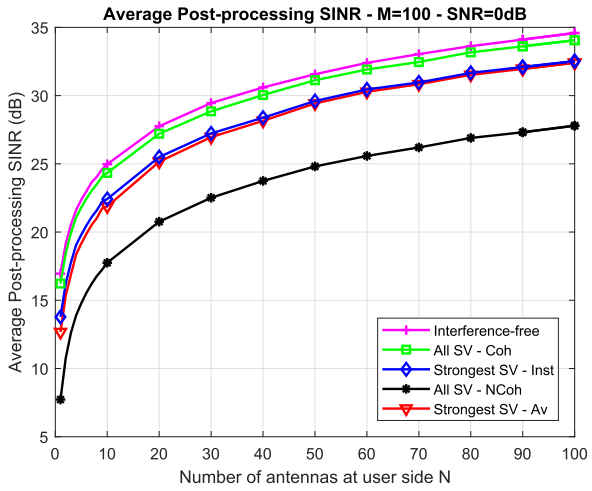


Fig. 7. SINR in a 2-user scenario vs number of antennas at the terminals, $\rho = 0$dB.

E. Beamforming Methods

We test the following transceiver structures that are classified by decreasing level of instantaneous CSI they exploit:

- **Interference free**: as a performance upper bound, we plot the case where the inter-terminal interference is ignored.
- **All SV - Coh**: transceiver according to equation (15) where all effective SVs are considered and coherent combining is performed. Information about the instantaneous CSI is needed.
- **Strongest SV - Inst**: transceiver according to equation (15) with $\mathcal{V}^{(2)} = \mathbf{U}^{(2)}$ and $\bar{V}^{(1)} = \mathbf{v}_{I,\max}^{(1)}$. This strategy necessitates partial instantaneous CSI at the transmitter. Assuming that the receiver applies $\mathbf{u}_{\max}^{(1)}$, the transmitter estimates the projection of $\mathbf{H}^{(1)}$ into the singular vectors $\bar{V}^{(1)}$ and selects the strongest one, denoted as $\mathbf{v}_{I,\max}^{(1)}$.
- **All SV - NCoh**: transceiver according to equation (15) where all SV are considered. Transmission across the singular vectors is performed non-coherently. The transmit power along singular vector $\mathbf{v}_i^{(1)}$ is $\lambda_i^{(1)}$.
- **Strongest SV - Av**: transceiver according to equation (15) with $\mathcal{V}^{(2)} = \mathbf{U}^{(2)}$ and $\bar{\mathbf{V}}^{(1)} = \mathbf{v}_{\max}^{(1)}$.

For the methods relying on the whole set of singular vectors ("All SV"), the receiver estimates the aggregate channel matrix $\mathbf{H}^{(1)}\mathbf{F}_{\text{ZF}}^{(1)}$ and matched filtering is applied [46]. For the other methods, the receiver applies the filter matched to $\mathbf{v}_{\max}^{(1)}$ and only requires the estimation of the projection of the aggregate channel on $\mathbf{v}_{\max}^{(1)}$.

F. Numerical Evaluations

Fig. 7 and Fig. 8 display the post-processing SINR and the Packet Error Rate (PER) associated to the different transceiver structures. The total number of antennas at the access point is $M = 100$ and the SNR is defined as $\rho = P/\sigma_n^2$ where P is the total transmit power and σ_n^2 is the variance of the noise at each receiving antenna. We normalize $\mathbf{F}_{\text{ZF}}$ in (15) so that the transmit power is divided equally among the users.

The multipaths in a single cluster are assigned different delays with an exponential decay that is up to 20dB.

Fig. 7 shows the post-processing SINR as a function of the number of antennas at the terminal side. As expected, we observe a gap between the methods exploiting full CSI and the methods based on second-order statistics or partial CSI. There is little differentiation for the latter methods.

Fig. 8 displays the Packet Error Rate (PER) as a function of the transmission slot for the case of a single antenna per user. The bandwidth is normalized, such that the number of channel uses directly reflects the delay. The payload is composed of 100 bits drawn from a BPSK modulation, hence we transmit 1 bit per channel use. We assume that the duration of the training for the coherent transmission techniques is twice as much as for the methods based on partial and no instantaneous CSI. A transmission slot is defined as the duration to send a packet (payload and overhead) using coherent transmission. Within a transmission slot, two packets can be sent using non-coherent transmission. The case where the users are multiplexed in space (solid lines) or in time (dotted lines) is shown.

For the selected simulation parameters, the following observations can be highlighted:

- The general tendency is that performance gets better with an increased exploitation level about the channel at the transmitter.
- There is a notable exception when the terminals are equipped with multiple antennas and receive diversity is exploited. In Fig. 8, for $N = 4$, the non-coherent strategy ("All SV - NCoh") performs the best. Hence, from a BER perspective, it is preferable to transmit the signal in a non-coherent fashion along each singular vector. The non-coherent transmission is compensated for by a receive coherent processing by multiple antennas that allows the extraction of diversity.
- Depending on the level of CSI exploited at the transmitter, space multiplexing is not always favorable.

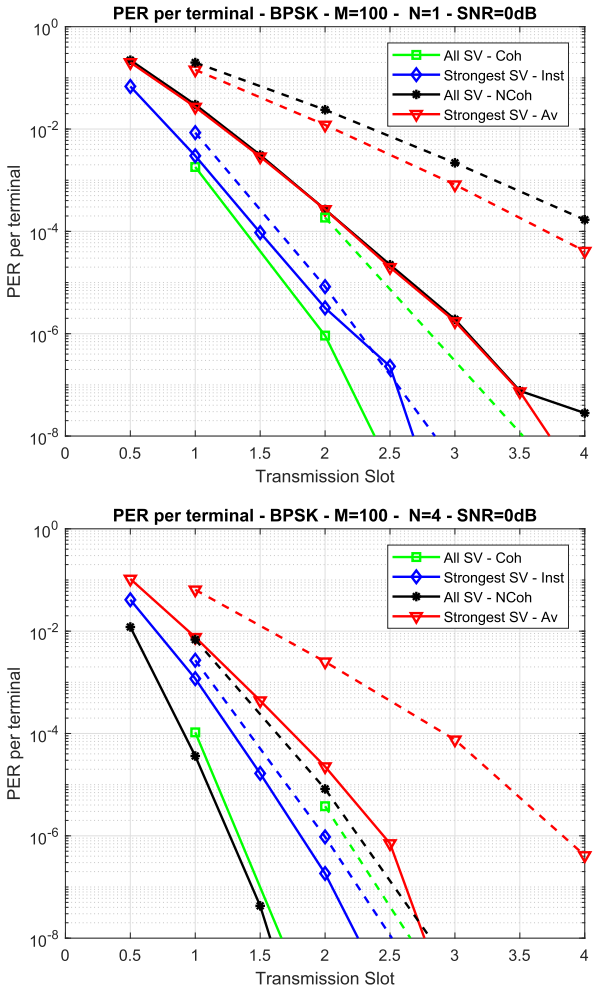


Fig. 8. Packet error rate in a 2-user scenario vs transmission slot, $\rho = 0$dB. Users are spatially multiplexed (solid lines) and time multiplexed (dashed lines).

VI. Multi-Connectivity and Interface Diversity

The mobile devices today have multiple radio interfaces and it is likely that many of the future devices will have that as well. This is also an indicator that the 5G radio interfaces will be deployed along with other radio interfaces. From the perspective of URLLC, the existence of multiple interfaces offers an additional degree of diversity that can be used to fulfil the stringent latency-reliability requirements. This is commonly known as *multi-connectivity* [47], while here we use the terms *link diversity* or *interface diversity* in order to emphasize the diversity role played by the availability of multiple different communication interfaces.

The idea of using multiple links or interfaces simultaneously is fairly natural and it has already emerged in some settings. In the context of 3GPP systems, LTE has supported Multi-Connectivity through Carrier Aggregation (CA) and Dual Connectivity (DC) since rel. 10 and 12, respectively. However, in this case the objective is throughput enhancement. Recently, in Rel. 15, Packet Duplication was introduced by 3GPP to boost reliability [47]. The data packet is duplicated on PDCP and transmitted on independent channels, either from the same eNB on different carriers via CA or from different eNBs using DC.

Packet Duplication in Multi-Connectivity architectures are excellent for mitigating losses due to fading and interference on individual links or temporary scarcity of air interface resources. Nevertheless, the reliability of the end-to-end connectivity relies on the correct functioning of an infrastructure and core network, often belonging to a single operator. While infrastructure and core networks are based on redundant solutions, they are still subject to single Point-of-Failure (PoF), e.g. through equipment misconfiguration. This reliance on a single network infrastructure can be mitigated by providing diversity not only at a link level, but also at a level of a communication interface or a path, as illustrated in Fig. 9.

The concept of Interface Diversity (IFD) was studied in [48]. Interface Diversity provides an independent path from the UE to the internet (cloud), by the use of a different wireless technology and/or a different mobile network operator. That is, IFD can be obtained by equipping a device with, for example, LTE/5G and Wi-Fi interfaces, or LTE/5G interfaces with SIM

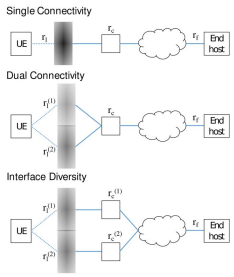


Fig. 9. Dual Connectivity and Interface Diversity architectures.

TABLE I
ASSUMED DEFAULT RELIABILITY PARAMETERS

	LTE/5G	Wi-Fi
r_{l}	0.99	0.9
r_{c}	0.999	0.99
r_{f}	0.9999	0.9999

cards from two physically independent mobile network operators. The key benefit of IFD is that there is no dependency on a single point of failure in the access network part. However, IFD requires that source and destination devices are configured to duplicate packets and handle multiple received copies, respectively. In comparison, dual connectivity is transparent to the source and destination devices, above the MAC layer. It should also be noted that Packet Duplication is the simplest instance of IFD, in which packet replicas are sent over different interfaces; more advanced IFD solutions involve various types of data segmentation and packet-level coding.

A. Reliability Model and Numerical Illustration

Assuming link/component reliabilities as indicated in the figure, we can use a series/parallel systems analogy from reliability engineering to express the end-to-end reliability of the architectures as:

$$R_{\text{single}} = r_{\mathrm{l}} r_{\mathrm{c}} r_{\mathrm{f}} \tag{16}$$

$$R_{\text{DC}} = \left(1 - \prod_{i=1}^{N} (1 - r_{\mathrm{l}}^{(i)})\right) r_{\mathrm{c}} r_{\mathrm{f}} \tag{17}$$

$$R_{\text{IFD}} = \left(1 - \prod_{i=1}^{N} (1 - r_{\mathrm{l}}^{(i)} r_{\mathrm{c}}^{(i)})\right) r_{\mathrm{f}}, \tag{18}$$

where R_{single}, R_{DC} and R_{IFD} are for a single link, N-link Dual Connectivity (DC) and N-interfaces Interface Diversity (IFD). $r_{\mathrm{l}}^{(i)}$ and $r_{\mathrm{c}}^{(i)}$ refer to the reliability of the ith link/interface or core network, respectively. A key assumption for eq. (18) is that the considered interfaces are uncorrelated in the sense that failures are occurring independently. This can be ensured in practice by using different mobile networks that do not share physical infrastructure.

Let us initially consider the two-link/interface instances sketched in Fig. 9. The assumed default parameters are given in Table I.

Fig. 10 shows the resulting end-to-end outage probability when subject to different cellular link outages. The results

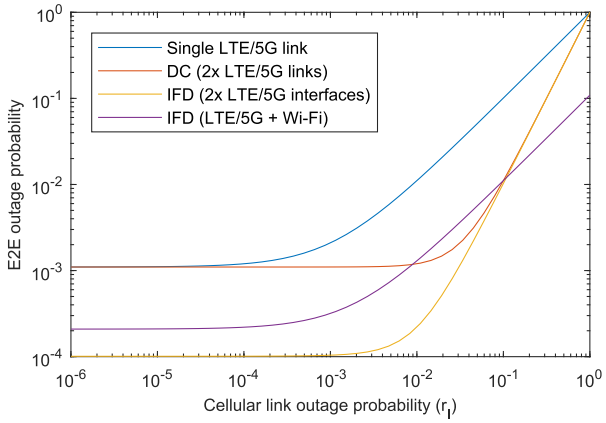


Fig. 10. End-to-end outage probability for varying cellular link outage. Note that the assumed Wi-Fi link reliability is $r_{\mathrm{l}_2} = 0.9$.

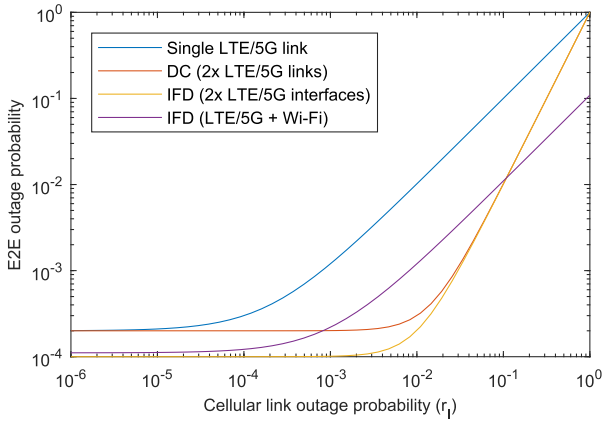


Fig. 11. End-to-end outage probability for varying cellular link outage with $r_{\mathrm{c}} = 0.9999$. Note that the assumed Wi-Fi link reliability is $r_{\mathrm{l}_2} = 0.9$.

show that IFD using two independent networks is always superior or equal in outage compared to DC. Specifically, when the cellular links are good, i.e. outage below 10^{-2}, the outage of IFD is for an order of magnitude better compared to DC. Even the alternative configuration where a LTE/5G is complemented by an inferior, but independent Wi-Fi connection, is outperforming DC for link outages below 10^{-2}. Further, the plot reveals that DC is better than using a single link, unless when the link outage is very low, and the end-to-end outage is instead dominated by the core outage probability. In comparison, consider the plot in Fig. 11, where the mobile network core is assumed to be more reliable. In that case, the difference between DC and IFD is almost negligible. The advantage compared to using just a single link is significant, especially for link outages between 10^{-3} and 10^{-2}.

VII. STATISTICAL ASPECTS OF ULTRA-RELIABLE GUARANTEES

In this section we investigate fundamental statistical questions related to ultra-reliability; namely, how can high reliability be assessed and measured. The ambitious reliability figures in URLLC only make sense when they are related to a statistical model of the context/environment in which the

URLLC system is deployed. However, the statistical model of the wireless context for URLLC is not known *a priori* and sets the stage of the methods of *statistical machine learning*, through which one can estimate the statistical properties of the environment and offer reliability guarantees.

Here we consider these questions in a simple setting: we investigate the impact of limited transmitter-side channel knowledge on the reliability performance of the communication links. Specifically, we consider a one-way communication link where the transmitter sends a packet to a receiver at rate R over a narrowband wireless channel; the model of the received baseband signal at the transmitter is given by (1). To isolate and study only the impact of channel uncertainty, in this section we consider packet errors due to *outage*, defined by the following event:

$$R > \log_2(1+P), \tag{19}$$

where P denotes the received power; from model (1) we have that $P = |h|^2$ where we have normalized the transmit power $|x|^2$ to unity.

Throughout the section, we will assume that the true distribution of the channel is circularly-symmetric $h \sim \mathcal{CN}(0, \sigma^2)$, implying that the received signal is dominated by scattered diffuse components; hence, the received envelope $\sqrt{P} = |h|$ follows Rayleigh distribution and the received power P follows exponential distribution with scale parameter $\theta = 2\sigma^2$ denoting the average channel power. Under the above assumption, the *outage probability* at transmission rate R is given by the cdf of the received power:

$$F(R) = 1 - e^{-\frac{2^R-1}{\theta}}. \tag{20}$$

Thus, transmitting at a specific rate over a Rayleigh channel with average power θ, yields a specific outage probability. The goal of ultra-reliable communication is to choose the *maximum* rate that meets a predetermined reliability criteria. However, as discussed in the following paragraphs, designing the reliability criteria and finding the most favorable rate is strongly linked to how much the transmitter knows about the channel.

A. Naïve Rate-Selection Under Channel Uncertainty

First, consider the benchmark case where the transmitter knows the channel perfectly; this implies that the transmitter knows the average channel power θ also perfectly. Under such circumstances, the transmitter can easily determine the maximum rate as a function of θ at which an outage probability level of ϵ can be guaranteed with *certainty*:

$$R_\epsilon(\theta) = \sup\{R > 0 : \Pr(R > \log_2(1+P)) \le \epsilon\} \tag{21}$$

$$= \log_2(1 - \theta \ln(1-\epsilon)). \tag{22}$$

$R_\epsilon(\theta)$ is also known as ϵ-outage capacity.

Now, consider the situation in which the transmitter has no knowledge of the average power θ. Instead, it collects n independent and noiseless power measurements $x_1, \ldots, x_n$ via *training*. Having acquired $x_1, \ldots, x_n$, a simple but naïve solution would be for the transmitter to compute the Maximum Likelihood (ML) estimate of θ by averaging $x_1, \ldots, x_n$, i.e.

$$\hat{\boldsymbol{\theta}}_{\text{ml}}(x^n) = \frac{1}{n}\sum_{i=1}^{n} x^i, \tag{23}$$

and plug the obtained estimate in (22) to determine the transmission rate $R(x_1, \ldots, x_n)$, which now becomes a random variable. According to (20), every R yields specific outage probability; hence, the sequence of random variables X^n induces a distribution over F and the transmitter can no longer guarantee that the outage probability under transmission rate $R(X^n)$ will be equal of less than ϵ.

B. Probabilistic Rate-Selection Framework: Parametric Channel Models

The above discussion shows that when the transmitter has limited knowledge of the channel, it can only guarantee the reliability *probabilistically*. Formally, this can be done by choosing the most favorable, i.e. the largest *rate-selection function* $R(X^n)$ such that predetermined *statistical reliability constraint* is satisfied. Depending on the specific formulation of the constraint, the channel knowledge status at the transmitter and the actual statistics of the channel, finding the most favorable rate-selection function might be involved problem. In the rest of the section, we will limit our discussion to the following somewhat heuristic but intuitive choice. Specifically, given specific realization x^n of X^n, the transmitter uses the following transmission rate:

$$R(x^n) = R_{\epsilon_n}\left(\hat{\boldsymbol{\theta}}_{\text{ml}}(x^n)\right), \tag{24}$$

for some $\epsilon_n > 0$; our aim is to find ϵ_n for each n such that $R(x^n)$ is maximized under predefined reliability constraint. The outage probability under transmission rate selected according to (24) is still a random variable; however, selecting ε_n according to specific reliability constraint effectively controls the amount of uncertainty in the outage probability. Note that for $\epsilon_n = \epsilon$ we have the naïve solution. Finally, we expect that the transmission rate is consistent, i.e. as $n \to \infty$, $R(x^n)$ converges to the ϵ-outage capacity.

Choosing the specific formulation of the statistical reliability constraint can be done in many ways. Here, we will consider two approaches, described next.

1) Average Reliability (AR): Consider the following constraint:

$$\sup_\theta \Pr[R(X^n) > \log_2(1+Y)] \le \epsilon, \tag{25}$$

where the averaging is performed w.r.t. the joint distribution of Y, X^n. By rewriting the above according to the rule for total probability, by first averaging over Y and then averaging over X^n, we observe that (25) guarantees that the worst-case *mean* of the outage probability $F(X^n)$ will remain below ϵ. Note that this does not guarantee that for some specific realization of X^n the outage probability will not be larger than ϵ. Regarding potential use-cases, constraint (25) is suitable for dynamic environments and can be used when one wishes to optimize the rate of the system and provide reliability guarantees *jointly*

over the training and the transmission, prior to the actual training.

Using the Rayleigh-channel assumption and the rate-selection function (24) and after few elementary computations, we obtain the following:

$$\sup_{\theta} \Pr[R(X^n) > \log_2(1+P)] \tag{26}$$

$$= 1 - \left(1 - \frac{\ln(1-\epsilon_n)}{n}\right)^{-n}. \tag{27}$$

The maximum value of ϵ_n, satisfying (25) is given by:

$$\varepsilon_n = 1 - e^{-n\left((1-\epsilon)^{-\frac{1}{n}}-1\right)}. \tag{28}$$

Clearly, as $n \to \infty$, $\varepsilon_n \to \epsilon$.

2) Probably Correct Reliability (PCR): We consider another, more restrictive reliability constraint that effectively controls the higher order moments of the outage probability via the concept of *meta-probability* [49]; namely, we require:

$$\sup_{\theta} \Pr[\Pr[R(X^n) > \log_2(1+Y)|X^n] > \epsilon] \leq \xi. \tag{29}$$

In other words, we limit the probability that the outage probability given X^n is larger than ϵ. Intuitively, ξ is an upper limit on the willingness of the system to tolerate outages larger than ϵ. So, the probability of outage is guaranteed to be equal or less than ϵ with probability larger than ξ; again note that this does not guarantee that for specific realization of X^n the outage probability will not be larger than ϵ. This approach is more suitable for static environments where channel training is done infrequently. In such circumstances, the average channel power estimate is used to set the rate for multiple, possibly many future transmission cycles; obviously, the transmitter here needs to be more conservative.

Using the Rayleigh-channel assumption and the rate-selection function (24), we obtain the following simple result:

$$\sup_{\theta} \Pr[\Pr[R(X^n) > \log_2(1+Y)|X^n] > \epsilon] \tag{30}$$

$$= 1 - \frac{\gamma\left(n, n\frac{\log(1-\epsilon)}{\log(1-\varepsilon_n)}\right)}{(n-1)!}, \tag{31}$$

with $\gamma(\cdot,\cdot)$ denoting the lower incomplete gamma function. By choosing ε_n satisfying

$$1 - \frac{\gamma\left(n, n\frac{\log(1-\epsilon)}{\log(1-\varepsilon_n)}\right)}{(n-1)!} \leq \xi, \tag{32}$$

we obtain a rate-selection function that satisfies (30) for any θ. Note that in the specific study of Rayleigh channel, the choice of rate using PCR approach does not depend on θ. This is a convenient result, stating that in case of Rayleigh channel, the transmitter only needs to ensure that the rate does not violate the maximum allowed tolerance ξ; such rate will be valid for any ϵ.

3) Evaluation: We evaluate both approaches w.r.t. the ratio between the average achievable throughput using rate $R(X^n)$ and the optimal throughput given that the distribution is known

$$\lambda_\epsilon(\theta) = \frac{\mathbb{E}\left[R(X^n)1_{R(X^n)\leq\log_2(1+Y)}\right]}{R_\epsilon(\theta)(1-\epsilon)}. \tag{33}$$

We evaluate the average throughput via Monte-Carlo simulation with K trials by simply averaging the rates $R(x^n)$ provided that $R(x^n) \leq \log_2(1+y)$ for any pair of realizations y, x^n. We depict $\lambda_\epsilon(\theta)$ as a function of n for different ϵ and for $\theta = 10$ in Fig. 12. We see that $\lambda < 1$ in both approaches, i.e. limited channel knowledge reduces the transmission rate. We also observe that the rate-selection function is consistent, that is as n grows large $\lambda \to 1$. This reduction is dramatically visible for the PCR constraint; this is also expected as (30) is significantly more restrictive than (25). Hence, using meta-probability, although providing stricter and more firm reliability guarantees, reduces the average rate significantly. Besides, the rate converges significantly slower to the respective ϵ-outage capacity.

C. Alternative Channel Models

We note that the rate-selection function cannot be always guaranteed to be consistent. Specifically, when relaying on parametric channel models, the rate is consistent only in the case when the true channel distribution adheres to the adopted model as above, where we assumed that the true channel envelope is Rayleigh distributed. However, if the true channel differs even slightly from the assumed model (e.g. a small specular component is also present), the rate is no longer consistent and both the AR and PCR constraints will be violated, see [50] for in-depth discussions. This can be viewed as a general pitfall of parametric channel models; while they provide fast convergence, they are prone to significant bias which lead to inconsistent rate and severe reliability violations. As an alternative, one can consider non-parametric channel modeling approaches for the scenarios in which there is a model deficit, as for example for wireless channels characterized with statistics for rare events. They indeed are guaranteed to give consistent rates under very mild channel restrictions but they require extensive training; in fact, the number of channel samples necessary to obtain non-negative transmission rate grows as [50]

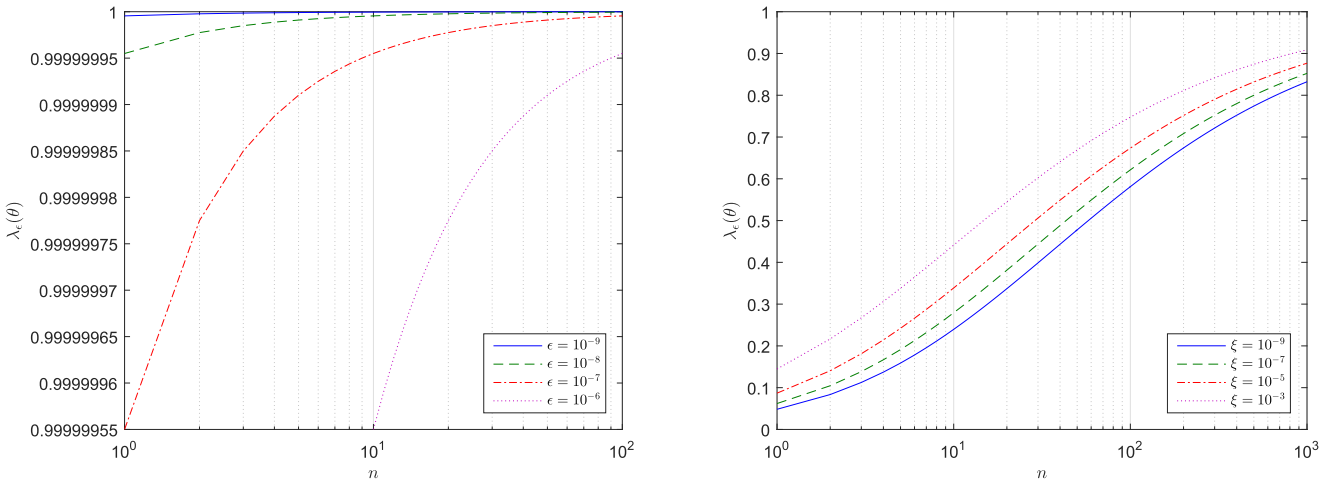
$$n \sim \frac{1}{\epsilon}. \tag{34}$$

Finally, [50] suggests to use power law approximations of the channel tail as third alternative as they provide "the best of the two worlds"; even though they do not guarantee rate consistency due to approximation error, they are significantly less biased than poor parametric models and they require reasonable training samples lengths.

VIII. Related Work

The body of literature on URLLC has been rapidly growing during the recent years and for good reason. While it is possible to achieve URLLC (1 ms latency and 0.99999 reliability) in idealized real-world trials [51] with only minor modifications to the frame structure compared to LTE-Advanced, it is much more difficult to statistically guarantee URLLC performance in both uplink and downlink for multi-user systems with co-existence of different service types.

Here we make a brief account of the literature related to the topics assessed in the paper, noting that this account is by no means exhaustive.



(a) Average reliability constraint

(b) Probably correct reliability constraint

Fig. 12. Parametric rate-selection under Rayleigh channel fading with average power $\theta = 10$.

A. 3GPP Standardization

The basic requirements of the URLLC "usage scenario" stem from 3GPP document TR 38.913 [5], whose initial version was made under Release 14 that belongs to LTE specification series. The first 5G release is Release 15, finalized in June 2018. Release 16 is currently in the focus of 3GPP standardization, and its finalization is expected at the end of 2019.

From the development standpoint, the work in Release 15 focused on the techniques for achieving low (user-plane radio) latency, documented in TS 38.300 [52]. Most of the novelties are concentrated in new and flexible numerologies (i.e. configurations of physical waveform parameters). For the reasons of backward compatibility in terms of scheduling, 3GPP has decided to keep the frame and subframe structure from LTE, with durations of 10 ms and 1 ms respectively. Thus, in order to decrease the duration of the transmission, the new numerologies feature (i) larger subcarrier spacings (SCS), which shortens the symbol and slot durations in respect to LTE; this is an option for frequency bands above 6 GHz, where the delay spread is small, and (ii) introduction of minislots of the length 2,3 and 7 symbols, in contrast to the standard length of 14 symbols; this option is of particular interest for frequency bands below 6 GHz, given that a packet fits into a minislot (expected to be the case for URLLC applications). With respect to the latter, 5G foresees that a URLLC packet can be allocated resource blocks along the frequency axis within the subframe, thus fitting the duration of the minislot. It should be also noted that 5G radio frame can support a mix of numerologies, i.e. a mix of subframes with different SCSs, some of which are compatible with LTE and NB-IoT.

The use of coding schemes with faster processing times with respect to the schemes employed LTE is another novelty in 5G: LDPC codes for the data channels, and polar codes for the control channels.

Further, 5G specifies use of an instant (pre-emptive) scheduling in the downlink. That is, upon arrival of a URLLC packet, any ongoing data transmission is interrupted and this packet gets transmitted. Although effective in reducing latency, the approach degrades performance of other, non-URLLC services. Also, TS 38.300 prescribes that semi-persistent scheduling can be used in the downlink for the periodic pre-assignment of resources for the (initial) transmissions to UEs, which is accompanied by signaling in the downlink control channel whether the pre-assignment is activated or not. SPS can also be used in the uplink, at the risk of the periodically pre-assigned resources being unused if the UE does not have any new packet to transmit. This risk can be reduced if overlapping pre-assignments to multiple UEs are made, which, in turn, calls for carefully assessing the impact of the potential collisions among the UEs on reliability [47]. We also note that 5G, in essence, inherits the connection-establishment procedure from LTE, see Section IV-B.1.

The work on methods for supporting URLLC in 5G has continued in Release 16. Currently, the most important investigations target enhancements related to grant free (i.e. configured grant) transmissions [53]. It was concluded that multiple configurations of configured grants should be supported to at least differentiate among services or traffic types within a cell, in order to enhance reliability and to reduce latency. Also, the configured grant access can be coupled with multiple repetitions (termed K repetitions) in order to improve reliability, where the repetitions can be proactive or reactive, driven by the downlink feedback. Specifically, if a configured grant transmission is detected by the base station, but is not properly decoded, the base station can initiate a grant based HARQ re-transmission by sending a HARQ-ACK. Furthermore, system-level investigations were performed assessing HARQ performance with ideas and realistic assumptions on channel estimation, different number of repetitions, as well as combining of repetitions [54].

Finally, we note that the interested reader can find further details on design and optimization of 5G for URLLC services in [47], [55].

B. URLLC Use Cases and General Works

A detailed description of URLLC use cases that are expected to play important roles in future 5G networks can be found in [6]; these include automotive, Industry 4.0, ITS and

ad-hoc disaster and emergency relief. From 3GPP perspective, automotive URLLC cases, i.e. their service architectures and requirements, are elaborated in [7], while specification [8] elaborates service architectures and requirements for URLLC use cases belonging to Industry 4.0 and intelligent transportation systems (ITS). Another take on emerging mission-critical services in cellular networks, such as tele-surgery, ITS and industrial automation, is given in [11]. Live audio production, also a potential URLLC use case, is described in [56]; notably, this use case emphasizes isochronous communication that is not supported by 4G and the article elaborates the ways to achieve it in 5G networks. A common feature of all mentioned URLLC use cases is that they are described by a single combination of target reliability and latency parameters. In contrast, [1] proposes a more general URLLC service model with reliable-service composition, by which target reliability performance progressively increases with latency.

From the industrial perspective, most of the interest has been shown for the use cases pertaining to Industry 4.0. In this respect, the potential of 5G to revolutionize the manufacturing and process automation has been recognized by industrial players. For instance, 5G Alliance for Connected Industries and Automation has been recently formed [57], with members from operational technology and ICT industries, telcos, academia and authorities, with the goals of ensuring that the 5G standardization adequately addresses the interest of the industries and that the 5G developments are adequately transferred to the industrial domain. We note that there are two real-world trials driven by Ericsson, related to real-time interaction in health-care industry [58] and autonomous management and control of fleet of drones [59]. In both cases, the trials revolve around the emerging 5G technology.

A number of works deals with general treatment of wireless URLLC communications. A survey of challenges and methods related to support of low-latency wireless communications is presented in [60]. An assessment of the sources of diversity and their exploitation to enable wireless URLLC networking is given in [2]. The challenges of high-performance wireless communications for industrial control are in focus of [61]. The study the fundamental energy-latency tradeoff in URLLC systems employing incremental-redundancy hybrid automatic repeat request in a specific context of point-to-point wireless connectivity is presented in [62].

C. Statistical Treatment of URLLC

Traditionally, channel models have been used to study average or cell edge channel conditions. In the case of URLLC, where reliability requirements are in the order of $10^{-5} - 10^{-9}$, it is the extreme tail of the channel model distribution that is important. In [63], the URLLC level behavior of common wireless channel models is investigated. The authors find that in many cases, a simple power law model with fitted parameters can sufficiently characterize the tail. In practical systems, where only limited channel knowledge is available and thus the specific channel model is unknown, the model uncertainty will inevitably impact the overlying communication protocol. In [64], the most critical dimensions of uncertainty are identified and their impact analyzed for two examples of cooperative communication protocols. A different perspective is taken in [14], where, first, relevant metrics and key enables of URLLC are discussed, where after mathematical tools from different scientific disciplines are proposed for evaluating the URLLC properties of different wireless communication system applications. While traditionally dependability metrics such as availability and reliability are expressed as functions of time, [65] proposes an evaluation framework for extending such analyses in the space domain, specifically considering ultra-reliable heterogeneous and homogeneous cellular communication systems.

D. Link and Interface Diversity

It is well-known from reliability engineering that ultra-high reliability can be effectively achieved through the parallel use of independent system components. In communication systems this translates to using multiple channels in parallel to achieve redundancy. 3GPP NR rel. 15 specifies use of Packet Duplication for URLLC services, where two independent transmission paths, from two different BSs, are used simultaneously to increase reliability and lower latency. Due to the involvement of multiple BSs, packet duplication requires modifications in the network architecture [66], [67]. Packet Duplication can be achieved both through Dual Connectivity and Carrier Aggregation, which are compared in [66]. In [68], an information theoretic study of the achievable gain of using joint decoding in a multi-connectivity setting over traditional MSC and MRC schemes is presented. For these schemes, the diversity multiplexing tradeoff that allows to trade-off outage probability and system throughput, is investigated. While the latency is not explicitly quantified in this work, the use of short packets and demonstrated reduction of outage probability enables URLLC. Finally, field trial measurements are used to demonstrate the practical performance. An extension of multi-connectivity is to complement BS links with D2D links, whereby UEs function as relays. A mathematical evaluation framework based on short block length regime for such D2D extended systems is proposed in [69] and where DF and AF relay strategies are compared to traditional BS-oriented MC. While multi-connectivity is typically assuming LTE or 5G links in a 3GPP system, the authors in [70] propose network architecture for enabling multi-RAT multi-connectivity, thereby allowing for example Wi-Fi or LTE-LAA to be exploited for multi-connectivity. As shown in [48], a combination of several different types of communication links can be flexibly used so as to optimize for a service-specific latency/reliability/bandwidth consumption tradeoff.

E. Wireless Access Networking

A vast number of papers discuss cellular access networking for URLLC services, we mention only a handful of them that explicitly deal with uplink communications. An overview of high-level optimization of radio-resource management for both uplink and downlink using the tools of network calculus and effective bandwidth is made in [71]. Finding a scheduling policy when that allows multiple users to meet a

target reliable-latency objective is the topic of [72], where the dynamic programming and knapsack-inspired approaches are used to find optimal and computationally-efficient suboptimal policies, respectively.

A number of works address advanced grant-free and grant-based access schemes for cellular access networks. The benefits of multi-user detection in radio-access uplink are presented in [73]. A scheme that exploit multi-user detection in combination with grant-free access with proactive sending of packet replicas is proposed in [74]. A grant-free scheme for batch arrivals that exploits replicas and SIC-based receiver and offers latency-reliability guarantees is introduced in [75]. Coordinated sharing of uplink resources in a grant-free manner, where the arrived users transmit packet replicas according to predefined activity patterns is proposed in [76]; the performance of the scheme is analyzed for both MMSE- and SIC-based receivers. Comparison of grant-free schemes with proactive and reactive sending of replicas is made in [77]; under the assumptions of an ideal control channel, perfect channel estimation, MMSE-interference rejection combining receiver, and target reliability of $1 - 10^{-5}$, it is shown that latency of grant-free schemes is generally lower than of the grant-based ones, while the choice between proactive and reactive grant-free schemes should be based on the load of the access network. In regards to the grant-based, four-step access, the work [78] proposes a control channel design that increases the reliability of signaling exchanges.

F. Massive MIMO

Multiple antenna systems appear as a natural enabler for URLLC as multiple antenna communications provide high SNR and diversity links as well as spatial multiplexing capability. Those properties contribute to increasing the reliability or the latency or both. In spite of their obvious advantages, however, those mechanisms are still relatively unexplored. A statistical characterization of multiple-input, single-output systems under statistical delay constraints is investigated in [79], using the tools of stochastic network calculus. In [39], for the uplink of a single user massive MIMO system with 64 antennas at the base station, diverse multi-antenna schemes are tested: coherent and non-coherent transceivers, transceivers assuming that the channel is unknown at the transmitter and using space-time codes, where the preference is given to non-coherent transceivers. In [42], the accent is put on the processing delay caused by multiple antenna processing at a base station in a multi-user massive MIMO system. In [40], a multi-user massive MIMO system network is optimized under a probabilistic constraint on the queue size to satisfy URLLC requirements.

One central question in multi-antenna system is the acquisition of instantaneous CSI. It is one of the most severe limitations to achieve URLLC when exploiting multiple antennas in a mobile environment constrained by channel coherence time as well as extreme latency requirements. The most critical acquisition occurs at the antenna array when the CSI is used for transmission mode (CSIT). In frequency division duplex (FDD) systems, CSIT acquisition requires a feedback loop from the terminals inducing a significant latency as the number of links to report is large. In a massive MIMO system below 6 GHz where the terminals have a small number of antennas, this concerns the BS in downlink transmission. In time division duplex (TDD) systems, latency can still be reduced by exploiting channel reciprocity [80], but remains critical. In a mmWave system, with potentially many antennas at the terminal, the issue concerns both side of the communication links.

Acquisition of the CSI at the receiver is usually perceived as less critical than at the transmitter as the delay between channel estimation and data detection is short. However, extreme cases of mobility at the user side might require an alternative to coherent detection, especially if URLLC is the target. Hence, non-coherent detection methods can be an asset in that case. A particularly simple method [81] that greatly benefits from the presence of a massive number of antennas is based on energy detection at the uplink of a massive MIMO system. The principle is to send a single stream of data, collect and aggregate the energy from all antennas. Detection is performed based on the average channel energy across the antenna array, which tends to a deterministic quantity for localized movements of the user and is therefore much more robust to user mobility than coherent detection. In addition, an efficient constellation design has been proposed in [82] that is able to benefit from the advantages of coherent communications at low mobility while switching to energy detection to ensure reliable communications at high mobility.

IX. Discussion and Outlook

In this paper we have discussed the principles of wireless access for Ultra-Reliable Low Latency Communication (URLLC). We have used a communication-theoretic framework to provide discussion on the fundamental tradeoffs. This was followed by elaboration on the important elements in access protocols. Two specific technologies were considered in the context of ultra-reliable communication, massive MIMO and multi-connectivity (interface diversity). We have also touched upon the important question about the proper statistical methodology for designing and assessing ultra-reliable communication. However, there are also important challenges that were not treated in the paper and out of them we would like to single out the problem of design of short-length codes for URLLC, a topic treated in [83], [84].

Here we reiterate the role of Machine Learning (ML) techniques in measuring and ensuring ultra-reliability. The knowledge of the environment in which ultra-reliable connectivity takes place is crucial for guaranteeing high levels of reliability, but also for applying methods to achieve this reliability. As the simplest example, knowing the coherence bandwidth in a given Industry 4.0 setting will lead to a proper selection of the allocated frequency resources. Getting knowledge about the environment requires data-driven technique and ML methods, while it introduces training cost. We believe that one of the main next frontiers in URLLC is proper formulation of the reliability challenge as well as methods to solve it based on ML techniques.

Finally, as an important next step, the community needs to address the issue of coupling high reliability with low latency, as it is done in the context of 5G. Relaxing the latency requirements towards a long term, e.g., beyond 10 or 50 ms, opens the design space for solutions that have good system level characteristics, such as coexistence with the other 5G services, or exhibit a higher energy efficiency. It should also be noted that the latency requirements on the wireless link can be reduced by adopting a holistic system design. For example, requiring 1 ms from the wireless link, while allowing source compression procedures that introduce much larger delay, is certainly not the optimal approach. Hence, one is tempted to define new research problems related to joint source-channel coding and protocol design that are suited to meet end-to-end latency and reliability requirements.

REFERENCES

[1] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. 1st Int. Conf. 5G Ubiquitous Connectivity (5GU)*, Nov. 2014, pp. 146–151.

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

[3] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[4] M. E. Porter and J. E. Heppelmann, "How smart, connected products are transforming competition," *Harvard Bus. Rev.*, vol. 92, no. 11, pp. 64–88, 2014.

[5] *Study on Scenarios and Requirements for Next Generation Access Technologies; (Release 15)*, document 3GPP TR 38.913 V15.0.0, Jun. 2018.

[6] J. Lorca *et al.*, "Deliverable D2.1: Scenarios, KPIs, use cases and baseline system evaluation," E2E-aware Optim. Advancements Netw. Edge 5G New Radio (ONE5G), Tech. Rep. D2.1, Nov. 2017. Accessed: Aug. 25, 2018. [Online]. Available: https://one5g.eu/wp-content/uploads/2017/12/ONE5G D2.1 finalversion.pdf

[7] *Study on Enhancement of 3GPP Support for 5G V2X Services (Release 16)*, document 3GPP TR 22.886 v16.0.0, Jun. 2018.

[8] *Service Requirements for the 5G System; Stage 1 (Release 16)*, document 3GPP TS 22.261 v16.4.0, Jun. 2018.

[9] Ericsson. (2015). Manufacturing Reengineered: Robots, 5G and the Industrial IoT. Ericsson Business Review. Accessed: Oct. 2, 2018. [Online]. Available: https://www.ericsson.com/assets/local/publications/ericsson-business-review/issue-4–2015/ebr-issue4-2015-industrial-iot.pdf

[10] "5G communication networks: Vertical industry requirements," Siemens AG, Munich, Germany, 2016. Accessed: Oct. 2, 2018. [Online]. Available: http://www.virtuwind.eu/_docs/Siemens_PositionPaper_5G_2016.pdf

[11] H. Chen *et al.*, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 119–125, Dec. 2018.

[12] B. Holfeld *et al.*, "Wireless communication for factory automation: An opportunity for LTE and 5G systems," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 36–43, Jun. 2016.

[13] *Requirements Related to Technical Performance for IMT-Advanced Radio Interface(s)*, document ITU-R M.2134, 2008. Accessed: Oct. 6, 2018. [Online]. Available: https://www.itu.int/pub/R-REP-M.2134

[14] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[15] *Low Throughput Networks (LTN); Functional Architecture*, document ETSI GS LTN 002 V1.1.1, 2014. [Online]. Available: http://www.etsi.org/deliver/etsigs/LTN/001099/002/01.01.0160/gsLTN002v010101p.pdf

[16] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, Apr. 2010.

[17] A.-S. Bana, K. F. Trillingsgaard, P. Popovski, and E. de Carvalho, "Short packet structure for ultra-reliable machine-type communication: Tradeoff between detection and decoding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6608–6612.

[18] K. F. Trillingsgaard and P. Popovski, "Downlink transmission of short packets: Framing and control information revisited," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2048–2061, May 2017.

[19] D. Tuninetti, B. Smida, N. Devroye, and H. Seferoglu, "Scheduling on the Gaussian broadcast channel with hard deadlines," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[20] J. Östman, G. Durisi, E. G. Ström, M. C. Coşkun, and G. Liva, "Short packets over block-memoryless fading channels: Pilot-assisted or noncoherent transmission?" *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1521–1536, Feb. 2019.

[21] G. C. Ferrante, J. Ostman, G. Durisi, and K. Kittichokechai, "Pilot-assisted short-packet transmission over multiantenna fading channels: A 5G case study," in *Proc. 52nd Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2018, pp. 1–6.

[22] G. Berardinelli, K. Pedersen, F. Frederiksen, and P. Mogensen, "On the design of a radio numerology for 5G wide area," in *Proc. 11th Int. Conf. Wireless Mobile Commun. (ICWMC)*, 2015, pp. 13–18.

[23] *ITU-T Recommendation G.8260*. Accessed: Aug. 17, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-G.8260-201508-I/

[24] N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. S. Quek, "Analyzing random access collisions in massive IoT networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6853–6870, Oct. 2018.

[25] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, Jun. 2015.

[26] J. Goseling, Č. Stefanović, and P. Popovski, "Sign-compute-resolve for tree splitting random access," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5261–5276, Jul. 2018.

[27] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[28] Y. Wen, W. Huang, and Z. Zhang, "CAZAC sequence and its application in LTE random access," in *Proc. IEEE Inf. Theory Workshop ITW*, Chengdu, China, Oct. 2006, pp. 544–547.

[29] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification; (Release 15)*, document 3GPP TS 36.321 v12.5.0, Apr. 2015.

[30] G. C. Madueño, J. J. Nielsen, D. M. Kim, N. K. Pratas, Č. Stefanović, and P. Popovski, "Assessment of LTE wireless access for monitoring of energy distribution in the smart grid," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 675–688, Mar. 2016.

[31] J. Massey, "Optimum frame synchronization," *IEEE Trans. Commun.*, vol. 20, no. 2, pp. 115–119, Apr. 1972.

[32] P. Robertson, "Optimum frame synchronization of preamble-less packets surrounded by noise with coherent and differentially coherent demodulation," in *Proc. Int. Conf. Commun. (ICC/SUPERCOMM)*, vol. 2, May 1994, pp. 874–879.

[33] C. Stefanovic and D. Bajic, "On the search for a sequence from a predefined set of sequences in random and framed data streams," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 189–197, Jan. 2012.

[34] M. N. Al-Subbagh and E. V. Jones, "Optimum patterns for frame alignment," *IEE Proc. F-Commun., Radar Signal Process.*, vol. 135, no. 6, pp. 594–604, Dec. 1988.

[35] S. W. Golomb and G. Gong, *Signal Design for Good Correlation: For Wireless Communication, Cryptography, and Radar*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[36] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G Networks," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1–7.

[37] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[38] L. Liu and W. Yu, "Massive device connectivity with massive MIMO," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1072–1076.

[39] S. R. Panigrahi, N. Bjorsell, and M. Bengtsson, "Feasibility of large antenna arrays towards low latency ultra reliable communication," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2017, pp. 1289–1294.

[40] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, M. Latva-aho, and C. S. Hong, "Ultra-reliable and low latency communication in mmWave-enabled massive MIMO networks," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2041–2044, Sep. 2017.

[41] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Aspects of favorable propagation in massive MIMO," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 76–80.

[42] W. Tärneberg, M. Karaca, A. Robertsson, F. Tufvesson, and M. Kihl, "Utilizing massive MIMO for the tactile Internet: Advantages and trade-offs," in *Proc. IEEE Int. Conf. Sens., Commun. Netw. (SECON Workshops)*, Jun. 2017, pp. 1–6.

[43] F. Boccardi, H. Huang, and M. Trivellato, "A near-optimum precoding technique for downlink multi-user MIMO transmissions," *Bell Labs Tech. J.*, vol. 13, no. 4, pp. 79–95, 2009.

[44] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[45] M. Chowdhury, A. Manolakos, and A. Goldsmith, "Multiplexing and diversity gains in noncoherent massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 265–277, Jan. 2017.

[46] A.-S. Bana, G. Xu, E. De Carvalho, and P. Popovski, "Ultra reliable low latency communications in massive multi-antenna systems," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 188–192.

[47] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Netw.*, vol. 32, no. 2, pp. 24–31, Mar./Apr. 2018.

[48] J. J. Nielsen, R. Liu, and P. Popovski, "Ultra-reliable low latency communication using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322–1334, Mar. 2018.

[49] M. Haenggi, "The meta distribution of the SIR in poisson bipolar and cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2577–2589, Apr. 2016.

[50] M. Angjelichinoski, K. F. Trillingsgaard, and P. Popovski. (2018). "A statistical learning approach to ultra-reliable low latency communication." [Online]. Available: https://arxiv.org/abs/1809.05515

[51] M. Iwabuchi *et al.*, "5G field experimental trials on URLLC using new frame structure," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[52] *NR; Overall Description; Stage-2*, document 3GPP ETSI TS 138 300 V15.3.1, Jun. 2018.

[53] *Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC)*, document 3GPP 38.824, Feb. 2019.

[54] *Study on New Radio Access Technology–Physical Layer Aspects (Release 14)*, document 3GPP TR 38.802 v14.2.0, Sep. 2017.

[55] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

[56] J. Pilz, B. Holfeld, A. Schmidt, and K. Septinus, "Professional live audio production: A highly synchronized use case for 5G URLLC systems," *IEEE Netw.*, vol. 32, no. 2, pp. 85–91, Mar./Apr. 2018.

[57] *5G Alliance for Connected Industries and Automation (5G-ACIA)*. Accessed: Mar. 08, 2019. [Online]. Available: www.5g-acia.org

[58] *Real-Time Interaction in 5G A Use Case Example from the Healthcare Industry*. Accessed: Mar. 08, 2019. [Online]. Available: www.ericsson.com/assets/local/digital-services/doc/health-care-case-real-time-interaction-in-5g-with-ims-data-channel.pdf,

[59] *Case Study: Operators Collaborate to Drive Mission-Critical Services Globally Using 5G*. Accessed: Mar. 8, 2019. [Online]. Available: https://www.ericsson.com/en/news/2018/2/case-study-bt-verizon-and-erics%son

[60] X. Jiang *et al.*, "Low-latency networking: Where latency lurks and how to tame it," *Proc. IEEE*, vol. 107, no. 2, pp. 208–306, Feb. 2019.

[61] M. Luvisotto, Z. Pang, and D. Dzung, "Ultra high performance wireless control for critical applications: Challenges and directions," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1448–1459, Jun. 2017.

[62] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2475–2485, Nov. 2018. [Online]. Available: http://arxiv.org/abs/1805.01332

[63] P. C. F. Eggers, M. Angjelichinoski, and P. Popovski, "Wireless channel modeling perspectives for ultra-reliable communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2229–2243, Apr. 2019.

[64] V. N. Swamy, P. Rigge, G. Ranade, B. Nikolic, and A. Sahai. (2018). "Wireless channel dynamics and robustness for ultra-reliable low-latency communications." [Online]. Available: https://arxiv.org/abs/1806.08777

[65] H. V. K. Mendis and F. Y. Li, "Achieving ultra reliable communication in 5G networks: A dependability perspective availability analysis in the space domain," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2057–2060, Sep. 2017.

[66] J. Rao and S. Vrzic, "Packet duplication for URLLC in 5G: Architectural enhancements and performance analysis," *IEEE Netw.*, vol. 32, no. 2, pp. 32–40, Mar./Apr. 2018.

[67] A. Aijaz. (2018). "Packet duplication in dual connectivity enabled 5G wireless networks: Overview and challenges." [Online]. Available: https://arxiv.org/abs/1804.01058

[68] A. Wolf, P. Schulz, M. Dörpinghaus, J. C. S. S. Filho, and G. Fettweis. (2017). "How reliable and capable is multi-connectivity?" [Online]. Available: https://arxiv.org/abs/1703.09992

[69] C. She, Z. Chen, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Improving network availability of ultra-reliable and low-latency communications with multi-connectivity," *IEEE Trans. Commun*, vol. 66, no. 11, pp. 5482–5496, Nov. 2018.

[70] S. Chandrashekar, A. Maeder, C. Sartori, T. Höhne, B. Vejlgaard, and D. Chandramouli, "5G multi-RAT multi-connectivity architecture," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, May 2016, pp. 180–186.

[71] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

[72] A. Destounis, G. S. Paschos, J. Arnau, and M. Kountouris, "Scheduling URLLC users with reliable latency guarantees," in *Proc. 16th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2018, pp. 1–8.

[73] S. Saur and M. Centenaro, "Radio access protocols with multi-user detection for URLLC in 5G," in *Proc. 23th Eur. Wireless Conf.*, May 2017, pp. 1–6.

[74] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 182–185, Apr. 2018.

[75] C. Stefanovic, F. Lazaro, and P. Popovski, "Frameless ALOHA with reliability-latency guarantees," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[76] R. Kotaba, C. N. Manchón, T. Balercia, and P. Popovski, "Uplink transmissions in URLLC systems with shared diversity resources," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 590–593, Aug. 2018.

[77] T. Jacobsen *et al.*, "System level analysis of uplink grant-free transmission for URLLC," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[78] H. Shariatmadari, S. Iraji, R. Jantti, P. Popovski, Z. Li, and M. A. Uusitalo, "Fifth-generation control channel design: Achieving ultrareliable low-latency communications," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 84–93, Jun. 2018.

[79] J. Arnau and M. Kountouris, "Delay performance of MISO wireless communications," in *Proc. 16th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2018, pp. 1–8.

[80] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, "Massive MIMO performance with imperfect channel reciprocity and channel estimation error," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3734–3749, Sep. 2017.

[81] L. Jing, E. D. Carvalho, P. Popovski, and À. O. Martínez, "Design and performance analysis of noncoherent detection systems with massive receiver arrays," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5000–5010, Oct. 2016.

[82] A.-S. Bana, M. Angjelichinoski, E. De Carvalho, and P. Popovski, "Massive MIMO for ultra-reliable communications with constellations for dual coherent-noncoherent detection," in *Proc. 22nd Int. ITG Workshop Smart Antennas (WSA)*, Mar. 2018, pp. 1–4.

[83] L. Gaudio, T. Ninacs, T. Jerkovits, and G. Liva, "On the performance of short tail-biting convolutional codes for ultra-reliable communications," in *Proc. 11th Int. ITG Conf. Syst., Commun. Coding (SCC)*, Feb. 2017, pp. 1–6.

[84] M. Shirvanimoghaddam *et al.*, "Short block-length codes for ultra-reliable low latency communications," *IEEE Commun. Mag.*, vol. 57, no. 2, pp. 130–137, Feb. 2019.