Relaying-Enabled Ultra-Reliable Low Latency Communications in 5G

Yulin Hu, M. Cenk Gursoy and Anke Schmeink

***Abstract*—Supporting ultra-reliable and low-latency communications (URLLC) has become one of the major considerations in the design of 5G systems. In the literature, it has been shown that cooperative relaying is an efficient strategy to improve the reliability of transmissions, support higher rates, and lower the latency. However, prior studies have demonstrated the performance advantages of relaying generally under the ideal assumption of communicating arbitrarily reliably at Shannon's channel capacity, which is not an accurate performance indicator for relaying in URLLC networks in which transmission is required to be completed within a strict time span and coding schemes with relatively short blocklengths need to be employed. In this article, we address the performance modeling and optimization of relaying-enabled URLLC networks. We first discuss the accurate performance modeling of relay-enabled 5G networks. In particular, we provide a comprehensive summary on the performance advantage of applying relaying in 5G URLLC transmissions in comparison to the case of direct transmission (without relaying). Both a noise-limited scenario and an interference-limited scenario are discussed. Then, we present tools for performance optimization utilizing the knowledge of either perfect or average channel side information. Finally, we summarize the proposed optimization schemes and discuss potential future research directions.**

I. Introduction

Low-latency and high reliability have become two major concerns in the design of future wireless networks. In particular, researchers and designers of next-generation wireless networks are increasingly interested in having wireless links carry latency-critical traffic with ultra-high reliability as relevant in certain applications, including, for instance, haptic feedback in virtual and augmented reality, E-health, autonomous driving, industrial control applications and cyber physical systems. In the design of 5G New Radio, this concept is called ultra-reliable low latency communication (URLLC) [1]. The common characteristic of 5G URLLC networks is that the coding blocklengths for wireless transmission are quite short due to the low latency constraint.

On the other hand, relaying is well known in the literature [2] as an efficient way to mitigate wireless fading by reducing path-loss and exploiting spatial diversity. Specifically, two-hop relaying protocols significantly improve the capacity and quality of service. However, prior studies on relaying are conducted under the ideal assumption of communicating arbitrarily reliably at Shannon's channel capacity achieved when the coding blocklength grows with no bound. In such cases, error/outage in a transmission is determined by comparing the channel's instantaneous Shannon capacity and the given coding rate. In other words, the results in these works only hold in the so-called infinite blocklength (IBL) regime, i.e., code blocks have unbounded lengths. These results are likely also accurate for scenarios with finite but significantly long blocklengths. However, they do not reflect the performance in URLLC applications where the blocklengths are quite short due to latency requirements.

In the finite blocklength (FBL) regime, the data transmission is no longer arbitrarily reliable. Especially when the blocklength is short, the error probability (due to noise) becomes significant even if the rate is selected from below the Shannon limit. Taking this into account, an accurate approximation of the achievable coding rate under the finite blocklength assumption for an additive white Gaussian noise (AWGN) channel was derived in [3], [4] for a single-hop transmission system. Subsequently, the initial work for AWGN channels was extended to quasi-static fading channels [5].

According to the results in [4], the relationship between the coding rate r and error probability ε is given by

$$r=\mathcal{R}\left(\gamma,\varepsilon,m\right)=\log(1+\gamma)-\log e\sqrt{\frac{\gamma(\gamma+2)}{\left(\gamma+1\right)^2 m}}Q^{-1}\left(\varepsilon\right)+\frac{\log m}{m}+\frac{o(1)}{m}, \tag{1}$$

where m is the blocklength, γ is the signal-to-noise ratio (SNR) and $Q\left(x\right)=\int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$ is the Gaussian Q-function. From (1), the (block) error probability can be expressed as:

$$\varepsilon=\mathcal{P}\left(\gamma,r,m\right)=Q\left(\frac{\log(1+\gamma)+\frac{\log m}{m}+\frac{o(1)}{m}-r}{\log e\sqrt{\frac{\gamma(\gamma+2)}{(\gamma+1)^2 m}}}\right). \tag{2}$$

Based on the above characterizations, in the FBL regime the achievable coding rate can be seen to increase with increased blocklength. However, there exists a performance gap between the Shannon capacity and FBL achievable rate. This performance gap in a single-hop system is numerically shown in Fig. 1. The figure illustrates that the gap is more significant for short blocklengths, and the FBL performance degrades significantly as the blocklength decreases. Note that (in comparison to direct transmission) two-hop relaying actually halves the blocklength of the transmission (if equal time allocation is considered). Then, the following interesting questions arise: What is the accurate performance of relaying in the FBL regime? Does relaying pay off less in the design of future 5G URLLC networks with FBL codes? In particular, note that both orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) schemes are key enabling technologies in 5G, corresponding to a noise-limited

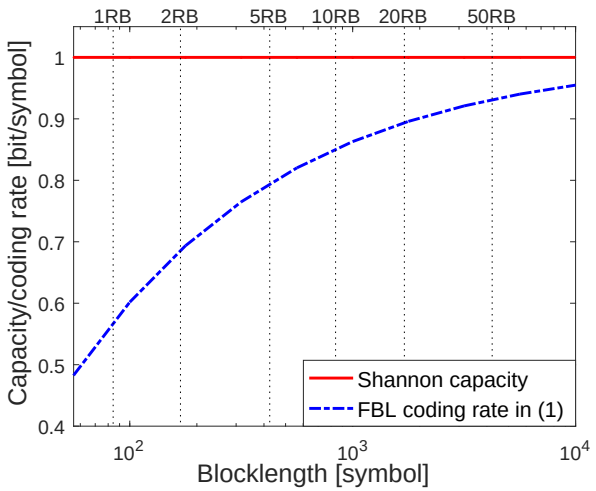

Fig. 1. Performance gap between the IBL and FBL regimes in a single-hop system under the a static channel (the channel is assumed to be constant in time). The FBL (achievable) rate is calculated according to (1) while the target error probability is set as 10^{-4}. The LTE baseline blocklengths are also provided in the figure as references, where 1 RB $=$ 84 symbols (following LTE release 13 where an RB comprises 12 adjacent subcarriers over 7 consecutive OFDM symbols.

scenario and an interference-limited scenario, respectively. Is relaying still a promising approach for URLLC transmissions in these two 5G scenarios? Moreover, how to optimize the FBL performance in relaying-enabled URLLC networks, e.g., by scheduling, resource allocation and re-transmission? These questions motivate us to address the FBL performance of relaying with the goal of contributing to the design of URLLC networks. In particular, we are interested in the trade-off introduced by applying relaying in the FBL regime between shortening the blocklength and providing a stronger signal. Therefore, the aim of this article is three-fold:

- To provide a comprehensive summary on the current state of the art in relaying-enabled URLLC networks. The performance advantages of applying relaying are discussed under both a noise-limited scenario and an interference-limited scenario.
- To address optimal scheduling, optimal resource allocation and optimal hybrid automatic repeat request (HARQ) for the relaying-enabled URLLC networks considering scenarios with perfect CSI and average CSI, respectively.
- To discuss future research directions within a high-level scope.

In the remainder of the article, we first discuss the current state-of-the-art regarding FBL performance of relay-enabled URLLC networks. Following this, we present a case study to illustrate in further detail the importance of optimal scheduling and resource allocation for improving the FBL performance. Finally, we provide concluding remarks and outline important future research directions.

II. Performance analysis of applying relaying in URLLC networks with FBL codes

In this section, we review existing works addressing the performance of relaying in the FBL regime. First, a noise-limited scenario is considered, i.e., noise along with fading is the key impairment of the wireless links while assuming no interference to be present. In addition, the FBL performance of applying relaying in NOMA networks is discussed, where the received signal at a user is subject to interference, resulting in an interference-limited scenario.

A. Performance advantages of relaying in a noise-limited scenario

Under the noise-limited scenario, the FBL performance of relaying is for the first time addressed in [6]. This paper considers a single half-duplex decode and forward (DF) relay network employing FBL codes, where the coding rate of each hop of relaying is assumed to be fixed at r. The overall error probability of this two-hop relaying is derived following the FBL coding rate characterization in [3]. In particular, applying (2) to each hop of relaying, the error probabilities of the first and second hop of relaying is given by $\varepsilon_i = \mathcal{P}(\gamma_i, r, m)$ where γ_i is the corresponding SNR. Hence, the overall error probability of this two hop relaying is given by $\varepsilon_{\rm O} = \varepsilon_1 + \varepsilon_2 - \varepsilon_1\varepsilon_2$. Then, the FBL throughput of the network is determined as $\mu = r(1-\varepsilon_{\rm O})/2$, where the division by 2 is due to the half-duplex operation of the relay. Note that relaying has two hops and the hop which has the worse channel quality (resulting in a lower reliability) is the so-called bottleneck hop. It is proven in [6] that the overall error probability is strictly increasing in the error probability of the bottleneck hop. In addition, the FBL throughput is shown to be concave in the coding rate and quasi-concave in the overall error probability. Therefore, the FBL throughput can be optimized by determining an appropriate overall target error probability level. More importantly, the performance advantages of relaying in the FBL regime is demonstrated in [6], i.e., the performance loss due to halving the blocklengh is much lower than expected, while the performance degradation in direct transmission is more significant.

These performance improvements achieved in the FBL regime is further studied in [7], in which the performance differences between relaying and direct transmission in the FBL regime are investigated specifically in a scenario where these two transmission schemes have the same IBL performance. In other words, in this specific scenario we analyze the conditions under which relaying is more efficient in the FBL regime than in the IBL regime, i.e., relaying has a higher FBL throughput than direct transmission. In particular, these two transmission schemes are compared under the same total blocklength constraint with value $2m$, i.e., the blocklength of direct transmission is $2m$ while the blocklength at each hop of relaying is m. It is proved that relaying is definitely superior to direct transmission in the FBL regime if the error probability of the bottleneck link of relaying is higher than the overall error probability of relaying. In particular, if the error probability of each link of relaying is within a range of practical interest, i.e., is lower than 0.5, relaying outperforms direct transmission in the FBL regime. More interestingly, the performance advantage of relaying is more significant for shorter coding blocklengths. From a topological perspective, this performance advantage of relaying leads to a broader

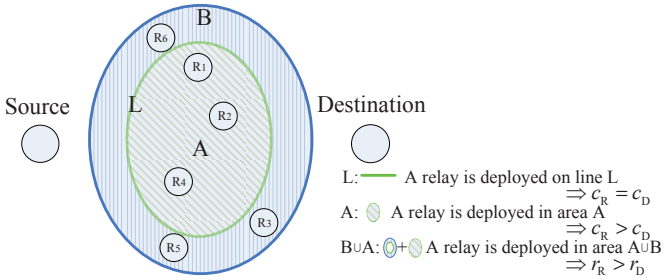

Fig. 2. The performance advantage of relaying in the FBL regime (from a perspective of topology). C_{R} and r_{R} are the Shannon capacity and FBL coding rate of relaying, while C_{D} and r_{D} are the Shannon capacity and FBL coding rate of direct transmission.

area for deploying/selecting a relay, as shown in Fig. 2. In the figure, if a relay is deployed in region A, relaying outperforms direct transmission in both the FBL regime and the IBL regime. On the other hand, if a relay is deployed in the (ring shaped) region B, relaying has a higher FBL-throughput but a lower Shannon capacity in comparison to direct transmission. Therefore, the key contribution of this work is that we have analyzed the conditions under which region B exists.

The FBL performance of a multiple relay scenario with best single relay (BSR) selection is studied in [8]. The evaluation in this work confirms the substantial benefits of BSR compared to direct transmission in the FBL regime. In comparison to the single relay network, having more relay candidates is more beneficial in the FBL regime than in the IBL regime. Moreover, the performance improvement by applying BSR is more significant in short blocklength scenarios.

B. Performance advantages of relaying in NOMA: An interference-limited scenario.

In addition to the noise-limited scenario, the FBL performance of relaying is discussed under a scenario with interference. In 5G, NOMA has been recently considered as a key promising radio access technique and a multi-user broadcast scheme [9]. Due to scheduling multiple users non-orthogonally on the same spectrum resource, inter-user interference is introduced to the network. In this section, we consider this interference-limited scenario in 5G low-latency networks and review the performance advantages of applying relaying in such a network.

In reference [10], the FBL performance of NOMA is addressed. In particular, this work focuses on a multi-user broadcast URLLC network operating in the finite blocklength regime and employing a NOMA scheme. By letting the user with the stronger channel from the source act as a relay, two relay-enabled transmission schemes, namely relaying and NOMA-relay, are proposed. The frame structures of NOMA and the proposed schemes are presented in Fig. 3, while the detailed description of the proposed schemes are as follows:

- **Relaying:** Without loss of generality, it is assumed that one user (called User 1) has a stronger link from the source than the other user (called User 2). Then, User 1 is required to perform as a DF relay. The whole frame with length M is divided into two phases, i.e., the backhaul phase with length m_1 and relaying phase with length m_2. Hence, $m_1 + m_2 = M$. As User 1 acts both as a user and a relay for User 2, it receives a large packet from the source, which is a combination of the two packets intended for the two users. Based on the DF relaying principle, if User 1 decodes the large packet successfully, it will forward User 2's packet in the second phase.
- **NOMA-relay (relaying with NOMA backhaul)**: This scheme is proposed to further apply NOMA in the first hop of relaying to transmit the two packets. If User 1 decodes the signal, x_2, intended for User 2 successfully (regardless of whether it decodes its own signal x_1 correctly or not), it forwards the packet to User 2. As shown in Fig. 3-C, the NOMA scheme in Fig. 3-A can be seen as a special case of NOMA-relay which allocates all the blocklength to the first phase, i.e, $m_1 = M$ and $m_2 = 0$. Hence, if $m_1 = M$ is the optimal solution maximizing the FBL performance, then the two schemes have the same resource allocation decision and thus, have the same performance. Otherwise, the NOMA-relay scheme definitely achieves a better performance than the NOMA scheme.

The FBL performance of the above proposed two protocols together with a typical NOMA protocol are studied in [10]. Although the performance advantages of NOMA are widely discussed under the assumption of infinite blocklengths (following the Shannon capacity bound), it is shown in [10] that the NOMA scheme is not preferred in the low-latency short blocklength scenario in comparison to the proposed relaying and NOMA-relay schemes. Moreover, by comparing these three schemes, it is observed that the NOMA-relay scheme is able to achieve the highest average throughput by setting the packet size relatively aggressively, while the relaying scheme generally provides the best fairness performance and leads to (although not the best) a relatively competitive average performance. From a topological perspective, the NOMA-relay scheme is the best choice if User 2 is at the cell edge, while the relaying scheme is broadly preferred when fairness is the major concern.

Heretofore, we have reviewed the performance advantages of applying relaying in 5G URLLC networks with different noise and interference assumptions. Next, we address optimal design considerations.

III. Optimal Design on Relaying-Enabled URLLC Networks

In this section, we discuss the optimal design of efficient relaying-enabled URLLC networks operating with FBL codes. We first review existing studies on optimal scheduling on the coding rate or the selection of the error probability. Then, we propose resource allocation protocols for the system design. In particular, both the perfect CSI and average CSI scenarios are considered in the design of relaying-enabled URLLC networks.

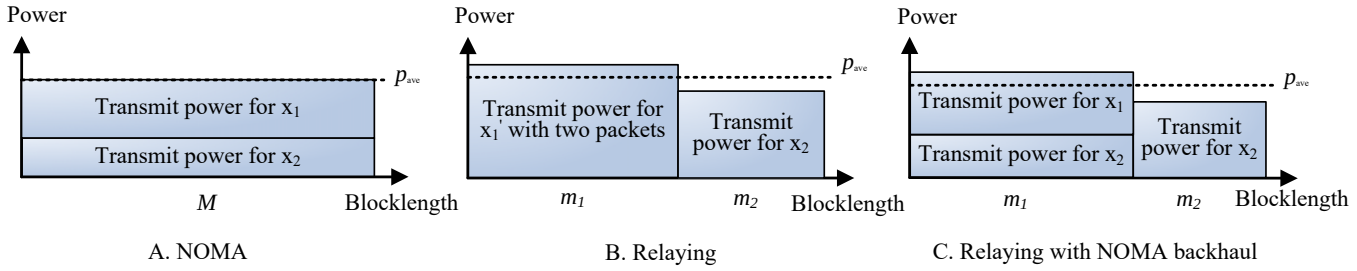

Fig. 3. Frame structures of multiple access schemes considered. Signals x_1, x_2 carry the data packet for user 1 and user 2, respectively, while the signal x_1' carries the two packets at the same time.

A. Optimal scheduling

1) Perfect-CSI driven scheduling: Under the perfect CSI assumption, the QoS-constrained performance of a DF relaying network has been investigated in [11]. The QoS-constrained performance is studied in the work by determining the effective capacity expression, which is a well-known QoS-constrained performance metric that characterizes the (maximum) arrival rate of a data flow to a buffer and relates the stochastic characterization of the service rate of the queuing system to queue-length or delay constraints of the flow. According to (1), for a given channel SNR, the coding rate is decreasing in the target error probability and increasing in the blocklength. Moreover, the queuing behavior at the source and relay buffers are also affected by the coding blocklengths and error probabilities. As a result, the decisions on target error probabilities and the blocklengths of the two hops strongly influence the QoS-constrained performance. In the above-mentioned work, the performance is maximized by a joint target error probability selection and blocklength allocation. In particular, authors in [11] characterize the convexity of the problem and propose an optimal numerical search algorithm in order to determine the optimal parameter setting more efficiently compared to a direct search in the three-dimensional bounded space.

2) Average-CSI-driven scheduling: If only the average CSI is available, the source is not able to instantaneously adjust the transmission rate along with the channel fading. In this scenario, a simple system operation is proposed in [12] by introducing a factor based on which the source weights the average CSI and determines the coding rate accordingly. In the mentioned work, both the physical-layer throughput and the QoS-constrained performance have been addressed, while the derivation of the QoS-constrained performance is also facilitated by the effective capacity formulation. In particular, it is proved that both the physical-layer throughput and the effective capacity are quasi-concave in the proposed weight factor in the considered relaying-enabled URLLC network operating with FBL codes.

In addition, this study shows that in the average CSI scenario, the FBL throughput of relaying is slightly increasing in the blocklength while the effective capacity is significantly decreasing in the blocklength. Hence, determining the optimal coding blocklength is critical for the design of QoS-supporting relaying systems. Moreover, the proposed weight factor provides a good tradeoff between the error probability and the coding rate in the considered relaying-enabled URLLC network. It allows us to optimize both the physical-layer performance (throughput) and the QoS-constrained performance (effective capacity) of the network. In particular, the optimal values of the weight factor for maximizing the throughput and for maximizing the effective capacity are different. Moreover, under the condition of having a similar Shannon capacity performance, relaying outperforms direct transmission in the FBL regime. This actually confirms the performance advantage of applying relaying in URLLC networks operating with FBL codes. More importantly, this performance advantage of relaying under the average CSI scenario is more significant than under the perfect CSI scenario. Finally, the performance loss due to finite blocklength (i.e., the performance gap between the performance in the IBL regime and in the FBL regime) is negligible under the average CSI scenario in comparison to the one under the perfect CSI scenario.

B. Optimal resource allocation

1) Perfect-CSI driven power allocation: In this subsection, we provide a brief case study on the perfect-CSI driven resource allocation algorithm for a relaying-enabled URLLC network. In particular, we propose an optimal power allocation policy to maximize the FBL throughput by optimally allocating power between the source and the relay and over frames (time). In addition, the optimization is required to satisfy an average (over time) power constraint $p_{\rm ave}$, i.e., $\mathbb{E}_{\mathbf{z}}\{p_1+p_2\} \leq p_{\rm ave}$, where p_1 and p_2 are the transmitted powers determined for the source and the relay while the vector $\mathbf{z}=(z_1,z_2)$ includes the instantaneous channel gains of the two hops.

Assume that the transmissions of the two hops are subject to the reliability constraints, i.e., $\varepsilon_1 \leq \varepsilon_{\rm th}, \varepsilon_2 \leq \varepsilon_{\rm th}$. In addition, transmissions over both hops have the same blocklength m. If the coding rate for the two-hop transmission is r, the error probabilities of the two hops are $\varepsilon_1 = \mathcal{P}(r,\gamma_1,m)$ and $\varepsilon_2 = \mathcal{P}(r,\gamma_2,m)$, while the SNR γ_i is determined by p_i, $i=1,2$. According to the model provided in Section II-A, the overall error probability $\varepsilon_{\rm O}$ is actually a function of p_1 and p_2. For a DF relay network with half-duplex operation, the problem of reliability maximization by performing optimal power allocation across frames and between the source and the relay can be formulated as $\min_{p_1,p_2} \varepsilon_{\rm O}(p_1,p_2)$ with constraints $\mathbb{E}_{\mathbf{z}}\{p_{\rm S}+p_{\rm R}\} \leq p_{\rm ave}$ and $\varepsilon_i \leq \varepsilon_{\rm th}, i=1,2$.

Here, we briefly discuss the optimal solution of the above problem in a reliable transmission scenario in which $\varepsilon_i \leq$

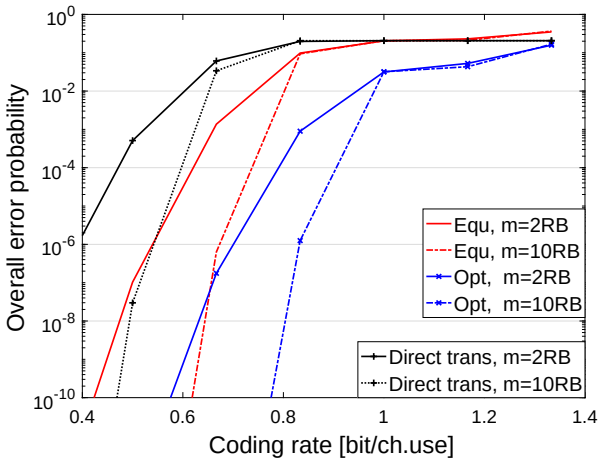

Fig. 4. The reliability comparison between the proposed optimal policy (Opt) and the equal power allocation policy (Equ). Both a short-blocklength ($m =$ 2RB) scenario and a relatively long-blocklength ($m =$ 10RB) scenario are considered, while 1 RB $= 84$ symbols (following LTE release 13 where an RB comprises 12 adjacent subcarriers over 7 consecutive OFDM symbols). In the numerical study, the average received SNR of two hops of relaying is set as 10 dB, while the SNR of the direct link transmission is set to 5 dB.

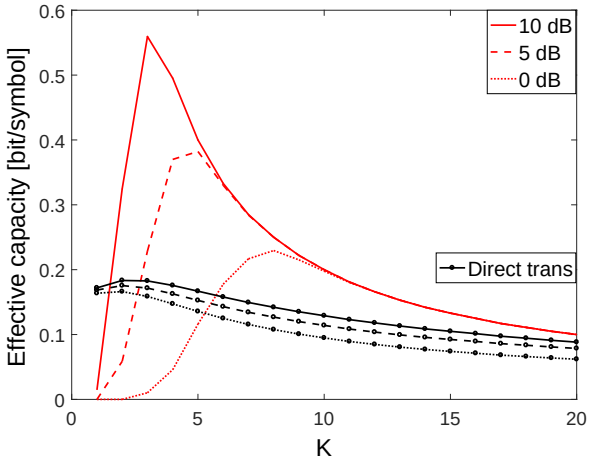

Fig. 5. The QoS-constrained performance (effective capacity) vs. the allowed maximum transmission and retransmission times K. In the analysis, we set $M = 4000$ symbols and $r = 1$ bit/symbol while the QoS exponent for determining the effective capacity is set to 10^{-2}. In addition, the SNR of the direct link is set to be 2 dB less than each hop of relaying.

$\varepsilon_{\rm th} \leq 0.1$. Then, we have $\varepsilon_1 + \varepsilon_2 \gg \varepsilon_1\varepsilon_2$. Hence, $\varepsilon_{\rm O} \approx \varepsilon_1 + \varepsilon_2$. It is clear that ε_i is convex in p_i in the region $\varepsilon_i \leq \varepsilon_{\rm th} \leq 0.1$. Hence, we can show that $\varepsilon_1 + \varepsilon_2$ is convex in (p_1, p_2). As a result, the above problem can be solved efficiently, using convex optimization tools. In addition, based on the above analysis regarding the error probability, it is easy to further prove that the throughput, given by $\mu = r\left(1 - \varepsilon_{\rm O}\left(p_1, p_2\right)\right)/2$, is concave in (p_1, p_2) under the considered reliable transmission scenario, i.e., can also be optimized efficiently. We provide numerical results for the reliability performance in Fig. 4 to demonstrate the improvements in the performance with the proposed optimal power allocation policy in comparison to the equal power allocation policy. In addition, we observe in Fig. 4 that ultra-reliable communication with error rates, for instance, ranging from 10^{-8} to 10^{-10} can be sustained by the appropriate choice of coding rates and coding blocklengths. Moreover, the performance of direct transmission (with double the blocklength of each hop of relaying) is also provided in the figure, which again confirms the performance advantage of applying relaying in such URLLC networks.

2) Average-CSI driven retransmission: When no instantaneous CSI is available to guide the resource allocation, an efficient approach to improve the reliability performance is to let the system employ HARQ with incremental redundancy (IR), which requires the source to allocate the total blocklength resource to several subblocks for the initial transmission and retransmissions. Let us assume that transmission rate is fixed at mr (bits/block) at the transmitter, where m is the blocklength of each hop, and r is the fixed coding rate in bits/symbol. In addition, the deadline constraint for each packet is M (in the unit of symbols). Note that each two-hop relaying transmission/retransmission costs $2m$ symbols. Hence, each packet is encoded into (at most) K codeword blocks, where $K \leq \left\lfloor \frac{M}{2m} \right\rfloor$. During each transmission/retransmission, the transmitter sends one codeword block to the destination via a two-hop relay channel, consuming $2m$ symbols. If the receiver decodes the received packet correctly, it sends an acknowledgment (ACK) feedback to the transmitter through an error-free feedback link, and a new packet is sent in the next time block. If the receiver cannot decode the packet, a retransmission request is sent through the feedback link, and another codeword block of the same packet is sent in the next time block. For simplicity, we assume an ideal ARQ protocol in our analysis, in which the transmitter gets the feedback immediately at the end of each time block without any delay.

In the HARQ-IR scheme, additional information is sent in each retransmission and the receiver combines all received code blocks in the same transmission period to decode the transmitted packets. At the end of the $k^{\rm th}$ trial (where $k \leq K$) in a transmission/retransmission period, the receiver combines the k received codeword blocks to decode the packet, which is equivalent to decoding a codeword with k subblocks and each subblock has a length of m symbols from the perspective of achievable rate. The error probability at the end of the $k^{\rm th}$ trial is given by

$$\varepsilon_{i,k} = Q\left(\frac{\sum_{l=1}^{k}\log_2\left(1+\gamma_{i,l}\right) + \frac{\log m}{m} + \frac{o(1)}{m} - r}{\log_2 e\sqrt{\sum_{l=1}^{k}\frac{(2+\gamma_{i,l})\gamma_{i,l}}{m(\gamma_{i,l}+1)^2}}}\right) \quad (3)$$

for given received SNRs $\{\gamma_{i,l}, i = 1, 2, l = 1, ..., k\}$.

Hence, the overall error probability for the $k^{\rm th}$ trial is $\varepsilon_{{\rm O},k} = \varepsilon_{1,k} + \varepsilon_{2,k} - \varepsilon_{1,k}\varepsilon_{2,k}$. Note that the whole transmission process is required to be finished within M symbols, i.e., allowing (at most) K transmission and retransmissions. An outage (or deadline violation) event occurs if the receiver experiences K decoding errors successively in a transmission period. The outage probability can be expressed as $P_{\rm out} = \mathbb{E}_{\mathbf{z}}\left\{\varepsilon_{{\rm O},K}\right\}$. In addition, as the coding rate is fixed, the FBL throughput and the effective capacity can be further determined based on $P_{\rm out}$.

It is clear that the choice of the blocklength m for each transmission/retransmission introduces a tradeoff between FBL performance loss (due to short blocklengths)

and allowed number of retransmissions. More importantly, if the QoS-constrained performance (effective capacity) is the objective, a flexible frame structure (with short blocklength) is more preferred, which results in significant FBL performance loss. Therefore, it is interesting to determine the optimal K for the HARQ-IR to minimize $P_{\rm out}$, or to maximize the FBL throughput or effective capacity. Note that K is an integer with a relatively small value and it can be efficiently determined by numerical search algorithms.

Here, we provide the numerical results of QoS-constrained performance as an example. The results are shown in Fig. 5, where we consider fading scenarios with different average SNRs. The results confirm that the system performance can be optimized by choosing an appropriate value of K. More interestingly, the optimal decisions on K are not the same for scenarios with different average channel qualities. In addition, the sharpness of these curves are different. In particular, determining the optimal K is more beneficial for the network with relatively better average channel qualities. Finally, the performance of direct transmission (with double the blocklength of each hop of relaying) is also provided in the figure. The optimal direct transmission performance is achieved at a different choice of K in comparison to the relaying case while the relaying provides a significantly higher achievable performance than the direct transmission.

IV. Conclusions and Open Research Directions

In this article, we have addressed major technical issues on the performance analysis and the system design of relay-enabled URLLC networks. We first elaborated the performance model. Despite halving the blocklength, the performance improvement of applying relaying is significant in URLLC networks operating with short blocklengths codes. In particular, relaying shows performance advantages in comparison to direct transmission and provides significant performance improvements in both a noise-limited scenario and an interference-limited scenario applying NOMA. Then, optimal system design tools have been presented for relay-enabled URLLC networks. We first discussed the optimal scheduling of the coding rate or the error probability for systems with perfect CSI and average CSI, respectively. Subsequently, we have proposed to optimize the system by performing perfect-CSI-driven power allocation. Finally, we have introduced an average-CSI-driven retransmission scheme to improve the FBL performance.

We finally conclude the article by discussing a set of promising research directions in the area.

- First, for the relay-driven URLLC networks, one of the most urgent issues is the frame structure design. In particular, two tradeoffs need to be addressed in the design, specifically the tradeoff between long blocklength vs. short blocklength with HARQ and the tradeoff between overhead of frequent and accurate CSI feedback vs. inaccurate CSI with longer transmission blocklength.
- Second, common scenarios in URLLC networks feature multiple nodes being densely deployed, potentially allowing multiple nodes to perform as relays, i.e., relaying transmission via two or more hops. It is essential to study the FBL performance of such networks and propose optimal system designs, regarding e.g., relay selection, scheduling and resource allocation policies.
- The third direction is to consider optimal system designs on scheduling and resource (i.e., power/blocklength) allocation for relay-assisted NOMA scenarios supporting URLLC transmissions.
- 5G URLLC networks with energy harvesting (EH) is another promising research direction [13]. In particular, simultaneous wireless information and power transfer (SWIPT) has been shown to be an efficient technique in the design of URLLC networks with EH nodes [14]. An interesting topic is perhaps to consider resource allocation, i.e., to optimally allocate blocklength/symbols and determine the optimal EH ratio.
- Finally, the performance advantage of relaying discussed in this work are mainly based on a relatively simple system model. It is challenging and interesting to address the URLLC network design in cellular networks in the presence of more practical constraints, e.g., backhaul limits [15], CSI inaccuracy, multiple users and/or types of data traffic with different priorities and security concerns. Moreover, it is also interesting to investigate which level of URLLC key performance indicators (KPIs) can be achieved by applying relaying in a network with the above practical assumptions.

References

[1] M. Simsek *et al.*, "5G-Enabled tactile internet," *IEEE JSAC*, vol. 34, no.3, Mar. 2016, pp. 460-73.

[2] S. Karmakar and M. Varanasi, "The diversity-multiplexing tradeoff of the dynamic decode-and-forward protocol on a MIMO half-duplex relay channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, Oct. 2011, pp. 6569-90.

[3] Y. Polyanskiy, H. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, May 2010, pp. 2307-59.

[4] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the AWGN channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, May 2015, pp. 2430–2438.

[5] G. Ozcan and M. C. Gursoy, "Throughput of cognitive radio systems with finite blocklength codes," *IEEE JSAC*, vol. 31, no. 11, Nov. 2013, pp. 2541-54.

[6] Y. Hu, J. Gross and A. Schmeink, "On the capacity of relaying with finite blocklength," *IEEE Trans. Vehic. Tech.,* vol. 65, no. 3, Mar. 2016, pp. 1790-94.

[7] Y. Hu, J. Gross and A. Schmeink, "On the performance advantage of relaying under the finite blocklength regime," *IEEE Commn. Letters*, vol. 19, no. 5, May 2016, pp. 779-82.

[8] Y. Hu *et al.*, "Finite Blocklength Performance of a Multi-Relay Network with Best Single Relay Selection", *Proc. IEEE ISWCS*, 2017. (Best Paper Award)

[9] B. Xu *et al.*, "Resource allocation in energy cooperation enabled two-tier NOMA hetnets towards green 5G," *IEEE JSAC* vol. 35, no. 12, Dec. 2017, pp. 2758–70.

[10] Y. Hu, M. C. Gursoy and A. Schmeink, "Efficient Transmission Schemes for Low-Latency Networks: NOMA vs. Relaying", *Proc. IEEE PIMRC*, 2017. (Best Paper Award)

[11] Y. Li, M. C. Gursoy and S. Velipasalar, "Throughput of two-hop wireless channels with queueing constraints and finite blocklength codes," *Proc. IEEE ISIT*, 2016.

[12] Y. Hu, A. Schmeink and J. Gross, "Blocklength-limited performance of relaying under quasi-static Rayleigh channels", *IEEE Trans. Wireless Commn.* vol.15, no.7, July 2016, pp. 4548-58.

[13] H. Zhang *et al.*, "Energy efficient user association and power allocation in millimeter wave based ultra dense networks with energy harvesting base stations," *IEEE JSAC* vol. 35, no. 9, June 2017, pp. 1936-47.
[14] O. L. A. Lopez *et al.*, "Ultra reliable short message relaying with wireless power transfer," *IEEE ICC*, 2017.
[15] H. Zhang *et al.*, "Fronthauling for 5G LTE-U ultra dense cloud small cell networks," *IEEE Wireless Commn.* , vol. 23, no. 6, Dec. 2016, pp. 48-53.

BIOGRAPHIES

Yulin Hu received his M.Sc.E.E degree from USTC, China, in 2011. He successfully defended his dissertation in 2015 and received his Ph.D.E.E. degree with great honor from RWTH Aachen University where he is a research fellow since 2016. From May to July 2017, he was a visiting scholar with Prof. M. Cenk Gursoy in Syracuse University, USA. His research interests are in information theory, optimal design of wireless communication systems. He is currently serving as an editor for Physical Communication (Elsevier).

M. Cenk Gursoy received his Ph.D. degree in electrical engineering from Princeton University in 2004. He is currently a professor in the EECS Department at Syracuse University. His research interests are in the general areas of wireless networks and information theory. He has been serving as an editor of IEEE Transactions on Green Communications and Networking, IEEE Transactions on Communications, and IEEE Transactions on Vehicular Technology. He received an NSF CAREER Award in 2006.

Anke Schmeink received the Diploma degree in mathematics with a minor in medicine and the Ph.D.E.E degree from RWTH Aachen University, Germany, in 2002 and 2006, respectively. She worked as a research scientist for Philips Research before joining RWTH Aachen University in 2008 where she is an associate professor since 2012. She is a member of the Young Academy at the NRW Academy of Science. Her research interests are in information theory, systematic design of communication systems and bioinspired signal processing.