Efficient Transmission Schemes for Low-Latency Networks: NOMA vs. Relaying

Yulin Hu[1], M. Cenk Gursoy[2] and Anke Schmeink[1]
[1]Information Theory and Systematic Design of Communication Systems, RWTH Aachen University, 52062 Aachen, Germany. Email: $hu|schmeink$@umic.rwth-aachen.de
[2]Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244, USA. E-mail: $mcgursoy$@syr.edu

***Abstract*—In this work, we focus on a low-latency multi-user broadcast network operating in the finite blocklength regime and employing a non-orthogonal multiple-access (NOMA) scheme. By letting the user with the stronger channel from the source act as a relay, we propose two relay-assisted transmission schemes, namely relaying and NOMA-relay. We study the finite blocklength performance of the proposed schemes in comparison with the NOMA scheme. Both the average performance of and fairness between users are considered. Our results show that the NOMA scheme is not preferred in the low-latency scenario in comparison to the proposed schemes. In particular, the relaying scheme generally provides the best fairness between users, while the NOMA-relay scheme is able to achieve a higher average throughput by setting the packet size relatively aggressively.**

***Index Terms*—Finite blocklength, NOMA, DF, relaying.**

I. Introduction

Low-latency communication is one of the major concerns in the design of future wireless networks. In particular, there has recently been significant interest in having wireless links to support latency-critical traffic as relevant in several applications involving, e.g., haptic feedback in virtual and augmented reality, E-health, autonomous driving, industrial control applications and cyber physical systems. In the design of fifth generation (5G) cellular networking architectures, this concept is called tactile Internet [1], [2]. Similarly, low latency applications are also widely discussed in the Internet of Things (IoT) [3], [4] and industrial wireless networks [5], [6] for the future industry design, i.e., Industry 4.0. The common characteristic of these discussed scenarios is that the coding blocklengths for wireless transmission are quite short due to the low latency constraint. In this finite blocklength (FBL) regime, especially when the blocklength is short, the error probability (due to noise) becomes considerable [7]. In particular, an accurate approximation of the achievable coding rate was identified in [7] for a single-user transmission via an additive white Gaussian noise (AWGN) channel while taking the error probability into account. Subsequently, the initial work regarding the AWGN channel was extended to Gilbert-Elliott Channels [8], quasi-static fading channels [9]–[11], quasi-static fading channels with retransmissions [12], [13] as well as spectrum sharing networks [14].

On the other hand, non-orthogonal multiple access (NOMA) has recently been considered as a key promising radio access technique and multi-user broadcast for future wireless communications systems. Compared to orthogonal frequency division multiple access (OMA), by applying the successive interference cancellation (SIC), NOMA exploits the channel gain differences between users and thus, schedules multiple users non-orthogonally on the same spectrum resource. As a result, NOMA effectively improves the system performance, especially the spectrum efficiency [15]–[17]. However, most of the existing studies demonstrating the advantage of NOMA are conducted under the ideal assumption of communicating arbitrarily reliably at Shannon's channel capacity (assuming infinite coding blocklength), i.e., an error in transmission is determined by comparing the channel's instantaneous Shannon capacity with the coding rate. In other words, the impact of finite blocklength coding on a multi-user broadcast system has only been recently addressed within a limited scope. For instance, in [18] the FBL performance of a two-user NOMA network is discussed and the advantage of the NOMA scheme is shown by comparing it with the scheme that separately transmits packets to the users (each user having half the blocklength). However, it is still not immediately clear if NOMA is the best transmission scheme in the low-latency multi-user broadcast scenario, since separate transmission of packets is a relatively simple approach. In particular, in our previous work [19]–[21] we investigated the relaying performance with FBL codes and showed the advantages of applying relaying in low-latency scenarios. This motivates us to consider relaying in a NOMA network by letting the user with the stronger channel act as a relay.

Specifically, in this work, we focus on transmission schemes in low-latency broadcast networks in the FBL regime, where the NOMA scheme is applied. Two relay-assisted schemes that consume the same total energy and require the same total number of symbol resources (i.e., the same blocklength) as the NOMA scheme are proposed in Section III. We derive the FBL throughputs of the proposed schemes while also discussing the performance of NOMA as a reference. Moreover, in Section IV we compare both the average performance and the fairness of these schemes via numerical analysis.

II. Preliminaries

A. System model

We consider a simple network with a Source and two users (User 1 and User 2), as shown in Fig. 1. The

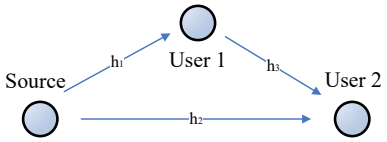

Fig. 1. Considered network where User 1 can perform as a relay.

source has two packets which need to be transmitted to the two users, respectively. Each packet has size D bits. In addition, the transmission of these two packets is performed under a latency constraint, i.e., transmission should be completed in a frame with length of M symbols. Moreover, the total energy budget/constraint for the transmission is $E_{\text{tot}} = Mp_{\text{ave}}$, where p_{ave} is the average power per channel use (symbol).

The channel fading coefficients of the Source-User 1 link, the Source-User 2 link and the User 1-User 2 link are denoted by h_1, h_2 and h_3, respectively. The channels are assumed to experience quasi-static fading, i.e., channels are assumed to be static during each frame of length M symbols and vary independently from one frame to the next. Due to the considered topology in Fig. 1, User 1 is more likely to have a stronger channel from the source than User 2, while this channel gain difference can be exploited by applying NOMA.

B. FBL performance of a single-user transmission

For AWGN channels, [7] derives a tight bound for the coding rate of a single-user transmission system. With blocklength m, block error probability ε and SNR γ, the coding rate (in bits per channel use) is given by $r = \frac{1}{2}\log_2(1+\gamma) - \sqrt{(1-\frac{1}{(1+\gamma)/2m^2})}Q^{-1}(\varepsilon)\log_2 e + \frac{O(\log_2 m)}{m}$, where $Q^{-1}(\cdot)$ is the inverse of the Q-function given by $Q(w) = \int_w^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$. In [10], the above result has been extended to a complex channel model with received SNR γ, where the coding rate (in bits per channel use) is

$$r = \mathcal{R}(\gamma, \varepsilon, m) \approx \mathcal{C}(\gamma) - \sqrt{\frac{V}{m}} Q^{-1}(\varepsilon), \tag{1}$$

where $\mathcal{C}(\gamma)$ is the Shannon capacity. For a known SNR γ, it is given by $\mathcal{C}(\gamma) = \log_2(1+\gamma)$. Moreover, V is the channel dispersion [7, Def.1]. Under a complex AWGN channel, $V = 1 - \frac{1}{(1+\gamma)^2}$. Hence, for a single-user transmission with blocklength m and coding rate r, the decoding (block) error probability at the receiver is given by

$$\varepsilon = \mathcal{P}(\gamma, r, m) = Q\left(\sqrt{\tfrac{m}{V}}\left(\mathcal{C}(\gamma) - r\right)\right). \tag{2}$$

So far, we have introduced the system model and the performance model of a single-user transmission with FBL codes. In the following, we further study multi-user transmission schemes maximizing the FBL throughput under latency and energy constraints.

III. Transmission Schemes in the FBL Regime

In this section, we first discuss the FBL throughput of the well-known NOMA scheme. Subsequently, we propose and study two broadcast transmission schemes by letting the user (User 1) with the better channel from the source perform as a decode-and-forward (DF) relay. The frame structures of these three multiple access schemes are presented in Fig. 2.

A. NOMA

When NOMA is applied, the source transmits two signals x_1 and x_2 for the two users via the same block with length M, as shown in Fig. 2-A. The received signal of User 1 in each frame is given by

$$y_1 = \sqrt{p_1}h_1 x_1 + \sqrt{p_2}h_1 x_2 + n_1, \tag{3}$$

where p_1 and p_2 are the allocated transmit power to the users, i.e., we have $p_1 + p_2 = p_{\text{ave}}$ and the energy consumption is $M(p_1+p_2) = Mp_{\text{ave}} = E_{\text{tot}}$. n_1 is the complex AWGN with power σ^2. In addition, x_1 and x_2 carry information for different packets, while each packet has a size of D. Hence, the coding rate for the transmission to each user is D/M bits per channel use (symbol).

Note that under the NOMA scheme, SIC is assumed to be employed at User 1 to cancel the interference of x_2. Hence, User 1 first tries to decode x_2 while treating x_1 as interference. According to the structure of the received signal shown in (3), the signal to interference plus noise ratio (SINR) at User 1 for decoding x_2 is given by $\gamma_{1,x_2} = \frac{p_2|h_1|^2}{p_1|h_1|^2+\sigma^2}$.

Then, based on (2), the error probability of User 1 decoding x_2 is given by $\mathcal{P}(\gamma_{1,x_2}, \frac{D}{M}, M)$. In other words, with probability $\mathcal{P}(\gamma_{1,x_2}, \frac{D}{M}, M)$, SIC fails. User 1 should decode x_1 while treating x_2 as interference, thus, resulting in an SINR for decoding x_1 given by $\frac{p_1|h_1|^2}{p_2|h_1|^2+\sigma^2}$. On the other hand, if SIC is successful, User 1 decodes x_1 based on an SNR level given by $\frac{p_1|h_2|^2}{\sigma^2}$. Hence, the SNR/SINR of User 1 decoding its own signal x_1 is Bernoulli-distributed, given by

$$\gamma_{1,x_1} = \begin{cases} \frac{p_1|h_2|^2}{\sigma^2} & \text{with Prob. } 1-\mathcal{P}(\gamma_{1,x_2}, D/M, M), \\ \frac{p_1|h_1|^2}{p_2|h_1|^2+\sigma^2} & \text{with Prob. } \mathcal{P}(\gamma_{1,x_2}, D/M, M). \end{cases} \tag{4}$$

Therefore, we obtain the expected error probability of the transmission to User 1 as

$$\begin{aligned} \varepsilon_1 = &\left(1 - \mathcal{P}\left(\gamma_{1,x_2}, \tfrac{D}{M}, M\right)\right)\mathcal{P}\left(\tfrac{p_1|h_1|^2}{\sigma^2}, \tfrac{D}{M}, M\right) \\ &+ \mathcal{P}\left(\gamma_{1,x_2}, \tfrac{D}{M}, M\right)\mathcal{P}\left(\tfrac{p_1|h_1|^2}{p_2|h_1|^2+\sigma^2}, \tfrac{D}{M}, M\right). \end{aligned} \tag{5}$$

Regarding the transmission to User 2, the received signal at User 2 is given by

$$y_2 = \sqrt{p_2}h_2 x_2 + \sqrt{p_1}h_2 x_1 + n_2 \tag{6}$$

where n_2 is also the Gaussian noise with power σ^2. Similarly as User 1, User 2 first tries to decode x_1 (to apply SIC) based on the SINR $\gamma_{2,x_1} = \frac{p_1|h_2|^2}{p_2|h_2|^2+\sigma^2}$. With probability $\mathcal{P}(\gamma_{2,x_1}, D/M, M)$, SIC fails. Then, User 2 decodes x_2 based on the SINR $\gamma_{2,x_2} = \frac{p_2|h_2|^2}{p_1|h_2|^2+\sigma^2}$. If SIC is successful, User 2 decodes x_2 based on an SNR level given by $\frac{p_2|h_2|^2}{\sigma^2}$. As a result, the error probability for the transmission to User 2 is

$$\begin{aligned} \varepsilon_2 = &\left(1 - \mathcal{P}\left(\gamma_{2,x_1}, \tfrac{D}{M}, M\right)\right)\mathcal{P}\left(\tfrac{p_2|h_2|^2}{\sigma^2}, \tfrac{D}{M}, M\right) \\ &+ \mathcal{P}\left(\gamma_{2,x_1}, \tfrac{D}{M}, M\right)\mathcal{P}\left(\tfrac{p_2|h_2|^2}{p_1|h_2|^2+\sigma^2}, \tfrac{D}{M}, M\right). \end{aligned} \tag{7}$$

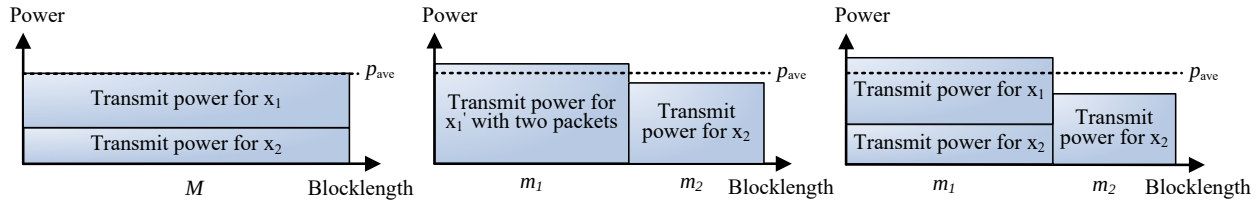

Fig. 2. Frame structures of multiple access schemes considered in this work.

B. Relaying

In this subsection, we propose to apply relaying to this multiple user scenario, as shown in Fig. 2-B. Since User 1 has a stronger link from the source, User 1 is required to perform as a DF relay. In particular, the whole frame is divided into two phases, the backhaul phase with length m_1 and relaying phase with length m_2. Under the latency constraint, we have $m_1+m_2 = M$. In addition, let us denote the transmit powers of the first and second phases as p_s and p_r. Then, we have the energy consumption bound $p_s m_1 + p_r m_2 = M p_{\text{ave}} = E_{\text{tot}}$.

As User 1 acts both as a user and a relay for User 2, it receives a large packet from the source in the first phase. This large packet is a combination of the two packets intended for the two users and the size of the large packet is $2D$. In the first phase, the received signal at the relay (User 1) and User 2 are given by

$$y_{1,1} = \sqrt{p_s} h_1 x_1' + n_1, \tag{8}$$

$$y_{1,2} = \sqrt{p_s} h_2 x_1' + n_2, \tag{9}$$

where x_1' carries the large packet with coding rate $2D/m_1$. Hence, the SNR for the relay to decode the large packet is $\gamma_1 = \frac{p_s|h_1|^2}{\sigma^2}$. Then, the error probability at the relay is $\varepsilon_1 = \mathcal{P}\left(\frac{p_s|h_1|^2}{\sigma^2}, \frac{2D}{m_1}, m_1\right)$. Based on the DF relaying principle, if User 1 decodes the large packet successfully, it will forward User 2's packet (with size D) at coding rate D/m_2 in the next phase. Therefore, in the second phase the received signal at User 2 (conditioned on the correct decoding at User 1) is

$$y_{2,2} = \sqrt{p_r} h_3 x_2 + n_3, \tag{10}$$

with the received SNR $\gamma_2 = \frac{p_r|h_3|^2}{\sigma^2}$. Under this case, the error probability at User 2 is $\mathcal{P}(\frac{p_r|h_3|^2}{\sigma^2}, \frac{D}{m_2}, m_2)$.

On the other hand, with probability ε_1, the relay (User 1) fails in decoding the large packet. Then, User 2 receives nothing in the second phase. In addition, even if User 2 receives $y_{2,2}$, with probability $\mathcal{P}(\frac{p_r|h_3|^2}{\sigma^2}, \frac{D}{m_2}, m_2)$ it is possible that x_2 is decoded incorrectly by User 2. Under these two cases of erroneous decoding, User 2 will thus try to decode the large packet based on the received signal in the first phase $y_{1,2}$. In this case, the SNR at User 2 is $\gamma_2 = \frac{p_s|h_2|^2}{\sigma^2}$, while the corresponding error probability is $\mathcal{P}(\frac{p_s|h_2|^2}{\sigma^2}, \frac{2D}{m_1}, m_1)$.

Finally, the expected error probability of the transmission to User 2 is given by

$$\varepsilon_2 = \left[(1-\varepsilon_1)\mathcal{P}\left(\frac{p_r|h_3|^2}{\sigma^2}, \frac{D}{m_2}, m_2\right) + \varepsilon_1\right] \cdot \mathcal{P}\left(\frac{p_s|h_2|^2}{\sigma^2}, \frac{2D}{m_1}, m_1\right). \tag{11}$$

C. NOMA-Relay: Relaying with NOMA backhaul

A key idea of applying relaying in the considered multi-user scenario is to transmit two packets together to User 1 and let it forward one of the packets to User 2. In Section III-B, we considered stitching the two packets to create a large one and transmitting this large packet. In this subsection, we propose to apply NOMA in the first hop of relaying to transmit the two packets. As shown in Fig. 2-C, the blocklength for the two phases are still denoted by m_1 and m_2, while $m_1+m_2 = M$. In addition, the total energy is allocated for three purposes: p_{s_1} for the source transmitting x_1, p_{s_2} for the source transmitting x_2 and p_r for relay (User 1) forwarding. We have the energy consumption bound $m_1(p_{s_1}+p_{s_2}) + m_2 p_r = M p_{\text{ave}} = E_{\text{tot}}$.

According to the NOMA principle, in the first phase the received signals at User 1 and User 2 are given by

$$y_{1,1} = \sqrt{p_1} h_1 x_1 + \sqrt{p_2} h_1 x_2 + n_1, \tag{12}$$

$$y_{1,2} = \sqrt{p_1} h_2 x_1 + \sqrt{p_2} h_2 x_2 + n_2. \tag{13}$$

Similarly as in Section III-A, the SINR for decoding x_2 at User 1 in the presence of interference from x_1 is $\gamma_{1,x_2} = \frac{p_2|h_1|^2}{p_1|h_1|^2+\sigma^2}$. The distribution of the SINR/SNR of User 1 decoding x_1 is

$$\gamma_{1,x_1} = \begin{cases} \frac{p_{s_1}|h_2|^2}{\sigma^2} & 1-\mathcal{P}(\gamma_{1,x_2}, D/m_1, m_1), \\ \frac{p_{s_1}|h_1|^2}{p_{s_2}|h_1|^2+\sigma^2} & \mathcal{P}(\gamma_{1,x_2}, D/m_1, m_1). \end{cases} \tag{14}$$

Therefore, the expected error probability of the transmission to User 1 is given by

$$\varepsilon_1 = \left(1-\mathcal{P}\left(\gamma_{1,x_2}, \frac{D}{m_1}, m_1\right)\right)\mathcal{P}\left(\frac{p_{s_1}|h_1|^2}{\sigma^2}, \frac{D}{m_1}, m_1\right) + \mathcal{P}\left(\gamma_{1,x_2}, \frac{D}{m_1}, m_1\right)\mathcal{P}\left(\frac{p_{s_1}|h_1|^2}{p_{s_2}|h_1|^2+\sigma^2}, \frac{D}{m_1}, m_1\right). \tag{15}$$

If User 1 (relay) correctly decodes user 2's packet, it forwards the packet in the second phase [1]. Then, the received signal at User 2 in the second phase has the same expression as (10). The received SNR and

[1] In the NOMA-relay scheme, User 1 is able to forward the packet of User 2 if it successfully decodes x_2, i.e., no matter if it fails to decode its own packet via x_1. This is different from the relaying scheme where User 1 needs to decode both packets before forwarding.

the error probability at User 2 are $\gamma_2 = \frac{p_r|h_3|^2}{\sigma^2}$ and $\mathcal{P}\left(\frac{p_r|h_3|^2}{\sigma^2}, \frac{D}{m_2}, m_2\right)$, respectively.

It is also possible that the two-hop relaying does not lead to a correct decoding at User 2. On one hand, with probability $\varepsilon_{1,x_2} = \mathcal{P}(\gamma_{1,x_2}, D/m_1, m_1)$ the relay incorrectly decodes user 2's packet and forwards nothing to User 2. On the other hand, with probability $(1-\varepsilon_{1,x_2})\mathcal{P}(\frac{p_r|h_3|^2}{\sigma^2}, \frac{D}{m_2}, m_2)$ the relay decodes and forwards to User 2 but User 2 fails to decode the forwarded packet. User 2 under these situations will attempt to decode using the received signal in the first phase $y_{1,2}$, i.e., by applying SIC. The SINR obtained from $y_{1,2}$ for User 2 to decode x_1 is $\gamma_{2,x_1} = \frac{p_1|h_2|^2}{p_2|h_2|^2+\sigma^2}$. Similarly as in the derivation of ε_1, the error probability at User 2 conditioned on decoding $y_{1,2}$ by exploring the SIC is given by

$$\begin{aligned}\varepsilon_{2|y_{1,2}} =& \left(1-\mathcal{P}\left(\gamma_{2,x_1}, \tfrac{D}{m_1}, m_1\right)\right)\mathcal{P}\left(\tfrac{p_{s_2}|h_2|^2}{\sigma^2}, \tfrac{D}{m_1}, m_1\right) \\ &+\mathcal{P}\left(\gamma_{2,x_1}, \tfrac{D}{m_1}, m_1\right)\mathcal{P}\left(\tfrac{p_{s_2}|h_2|^2}{p_{s_1}|h_2|^2+\sigma^2}, \tfrac{D}{m_1}, m_1\right).\end{aligned} \tag{16}$$

As a result, the expected error probability of the transmission to User 2 under the NOMR-relay scheme is given by

$$\varepsilon_2 = \left[(1-\varepsilon_{1,x_2})\,\mathcal{P}\left(\tfrac{p_r|h_3|^2}{\sigma^2}, \tfrac{D}{m_2}, m_2\right) + \varepsilon_{1,x_2}\right]\varepsilon_{2|y_{1,2}}. \tag{17}$$

D. Average throughput and max-min throughput

For all the above three schemes, in each frame, the FBL throughputs (in bits/ch.use) of the two users are given by $\mu_1 = D(1-\varepsilon_1)/M$ and $\mu_2 = D(1-\varepsilon_2)/M$. Then, we can define the average throughput as $(\mu_1 + \mu_2)/2$ and the minimal throughput as $\min\{\mu_1, \mu_2\}$ for the two users in one frame. Obviously, these throughputs are influenced by the resource allocation decisions, i.e., powers for the two hops/users and coding blocklengths for the two hops. Hence, the optimal (achievable) average throughput

$$\mu_{\text{ave}} = \max_{\text{blocklength,power}} \tfrac{\mu_1+\mu_2}{2}, \tag{18}$$

and the optimal max-min throughput

$$\mu_{\text{max-min}} = \max_{\text{blocklength,power}} \min\{\mu_1, \mu_2\}. \tag{19}$$

can be determined by solving the corresponding optimization problems above over coding blocklength and power allocations under an average power constraint. Note that μ_{ave} considers the average performance while $\mu_{\text{max-min}}$ focuses on the fairness. Both of them are important for the design of low-latency networks. In this work, we consider both the average throughput and the max-min throughput as performance metrics for the evaluation of these three transmission schemes.

Proposition 1. *Both μ_{ave} and $\mu_{max\text{-}min}$ performances of the NOMA scheme are upper bounded by the NOMA-relay scheme.*

Proof. As shown in Fig. 2, the NOMA scheme can be seen as a special case of NOMA-relay which allocates all the blocklength for the first phase, i.e, $m_1 = M$ and $m_2 = 0$. Hence, if $m_1 = M$ is the optimal solution, then the two schemes have the same resource allocation decision and thus, have the same performance. Otherwise, the NOMA-relay scheme achieves a better performance than the NOMA scheme. □

Note that both (18) and (19) are throughputs within a single frame. The ergodic throughputs can be obtained by averaging the throughputs over the channel fading.

IV. Numerical Analysis

In this section, we provide our numerical results. First the throughput performance within a single frame is considered. Subsequently, we evaluate the ergodic throughput performance over Rayleigh fading.

A. The impact of resource allocation in a transmission frame

We start with the numerical analysis on how the allocation of power and blocklength influences the throughput within a frame. The three schemes are compared with respect to different choices of blocklength and power allocations over the two hops. In particular, the NOMA scheme is not influenced by the blocklength allocation as it just has one hop. In addition, for the NOMA-relay case which needs to further allocate the power of the first hop to x_1 and x_2, we collect the throughput based on the optimal power allocation in the first hop by exhaustive search. In the numerical analysis of this subsection, we set $|h_1|^2 = 10^{-10}, |h_2|^2 = 10^{-12}, |h_3|^2 = 10^{-8}$, $p_{\text{ave}} = 30$ dBm, $\sigma^2 = -80$ dBm, $D = 400$ bits and $M = 800$ symbols.

The results are shown in Fig. 3 and Fig. 4 with respect to the average throughput and max-min throughput, respectively. The figures demonstrate that throughputs (especially the max-min throughput) are roughly quasi-concave in the choice of power/blocklength allocation. In addition, both relaying and NOMA-relay schemes are able to achieve the optimal throughput based on the given packet size, while the NOMA scheme is not preferred, especially from a fairness perspective (a low max-min throughput). Moreover, relaying and NOMA-relay schemes are relatively stable with respect to the resource allocation decision. This indicates that if there is not enough accurate information for optimal resource allocation, the relaying and NOMA-relay schemes are options that likely result in a good throughput.

B. Ergodic throughput over channel fading

In the following, we study the ergodic throughput performance of the three schemes, while the instantaneous throughput performance for each channel realization is maximized employing the optimal resource allocation, which is achieved by performing an exhaustive search. In the analysis, we set $p_{\text{ave}} = 30$ dBm, $\sigma^2 = -80$ dBm. In addition, the average channel gains for the three links are set to $10^{-10}, 10^{-12}, 10^{-8}$.

We vary the packet size and plot the resulting ergodic throughput in Fig. 5. It can be observed from this figure

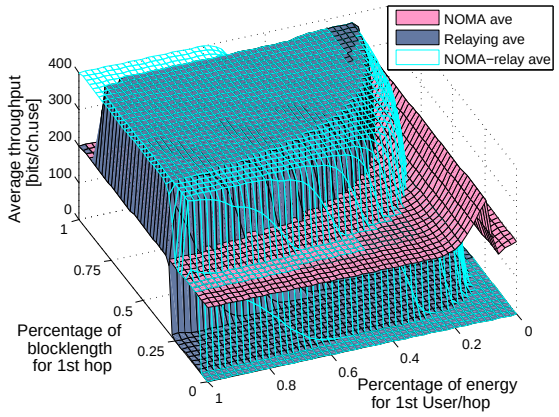

Fig. 3. The impact of blocklength and power allocation on the average throughput within a frame.

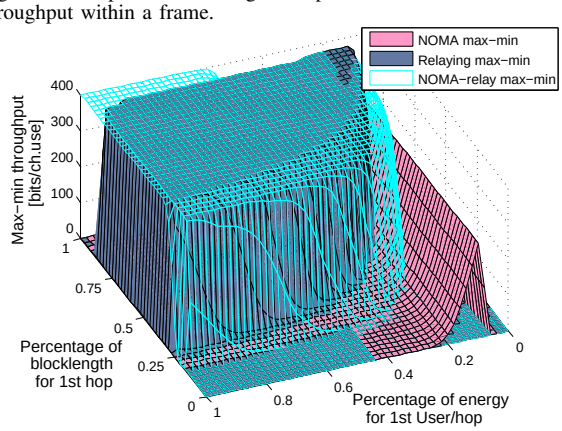

Fig. 4. The impact of blocklength and power allocation on the max-min throughput within a frame.

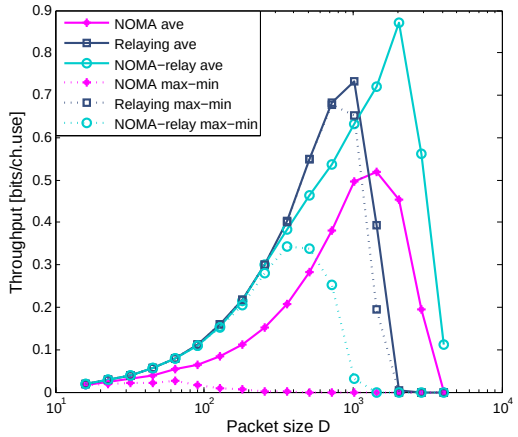

Fig. 5. Ergodic throughput vs. packet size. We set $M = 800$ symbols.

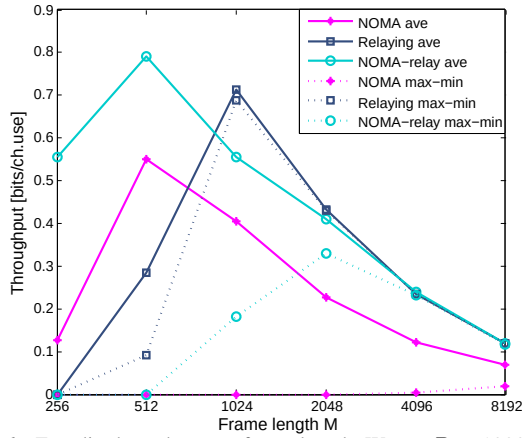

Fig. 6. Ergodic throughput vs. frame length. We set $D = 1000$ bits.

that all the throughput curves are quasi-concave in the packet size. This is because a small packet limits the throughput while a large packet introduces errors. In addition, the optimal choices of the packet size are different for these schemes. In particular, for all the average throughput curves, the optimal packet size of NOMA-relay is higher than those of the NOMA and relaying schemes. As the NOMA-relay scheme allows to set the packet size more aggressively, it has a higher maximal average throughput than the rest of the schemes. On the other hand, the relaying scheme provides the best fairness, i.e., the max-min throughput of relaying is significantly higher than those of NOMA and NOMA-relay schemes. In addition, the NOMA scheme has a lower performance than the NOMA-relay scheme in terms of both the average throughput and the max-min throughput, which confirms our analytical model and Proposition 1. Finally, it is interesting that the optimal choices of packet size maximizing the average throughput and the max-min throughput under the relaying case are quite similar, while this is not true for NOMA and NOMA-relay schemes.

Next, we study the ergodic throughput while varying the frame length M. We set the packet size to 1000 bits, the choice of which is according to Fig. 5 where the average throughputs of the schemes are similar (relaying performance is slightly higher than NOMA-relay) at point $D = 10^3$. The relationship between the ergodic throughput and frame length is illustrated in Fig. 6. We observe that the average throughputs of all schemes are quasi-concave in the frame length. In addition, the NOMA-relay scheme outperforms the NOMA scheme with respect to average throughput. Moreover, the optimal frame lengths maximizing the average throughput are different for the relaying and NOMA-relay schemes, i.e., relaying prefers a longer frame length than the NOMA-relay scheme. Furthermore, the relaying scheme provides a higher fairness performance.

Finally, we investigate the ergodic throughput while varying the system topology (in a meter scale). We fix the location of source at point $(0, 0)$, fix the location of User 1 (relay) at point $(100, 0)$ and change the location (x, y) of User 2 in the area where $x \in [-100, 500], y \in [-300, 300]$. For each location of User 2, we calculate the path loss and obtain the ergodic throughput performance of different schemes. Then, we compare the ergodic throughputs. According to the result of the comparison for each location, we paint a color at the location. In particular, we paint the location by the slate-gray color if the relaying performs better than the NOMA-relay, otherwise we paint it by the light blue color when the NOMA-relay scheme leads to a higher throughput. Note that since the NOMA scheme is definitely not better than the NOMA-relay scheme, we actually only have two colors for painting. The results

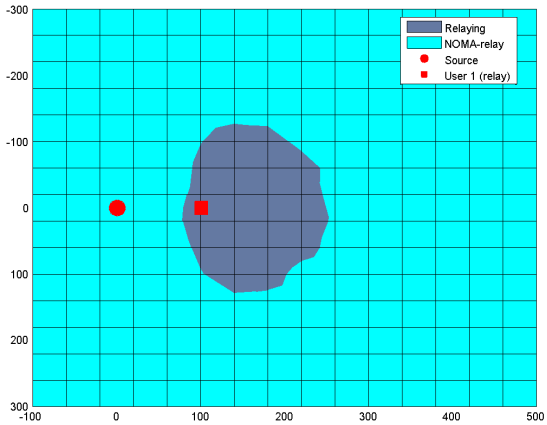

Fig. 7. Average throughput comparison from a topological perspective. In the numerical analysis, we set $D = 400$ bits and $m = 400$ symbols.

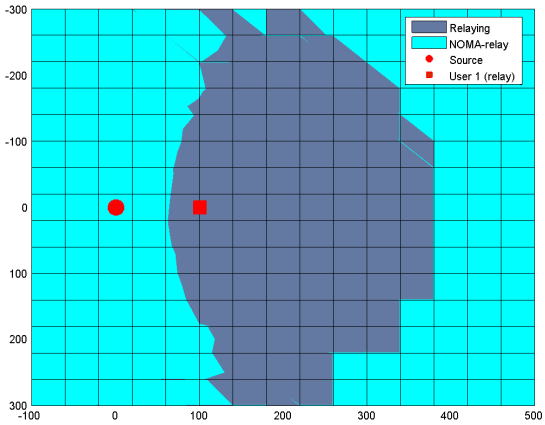

Fig. 8. Max-min throughput comparison from a topological perspective. In the numerical analysis, we set $D = 400$ bits and $m = 400$ symbols.

are provided in Fig. 7 and Fig. 8, where corresponding metrics for the comparison are the average throughput and the max-min throughput, respectively.

The two figures illustrate that relaying is generally preferred when User 2 is not far from the center of the area. For instance, the NOMA-relay is a better choice if User 2 is at the cell edge in a cellular network. In addition, the relaying scheme is more broadly preferred if fairness is more important for the system, which matches well with the results in the previous figures.

V. Conclusion

In this work, we have studied broadcast schemes for source transmissions in a low-latency scenario with FBL codes. We have proposed two relay-assisted transmission schemes and derived their FBL performance in comparison with the NOMA scheme. Via numerical analysis, we have shown that the NOMA scheme is not preferred in low-latency scenarios in comparison to the proposed relaying and NOMA-relay schemes. In particular, the NOMA-relay scheme is able to achieve a higher 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 the fairness is the major concern.

References

[1] M. Simsek, A. Aijaz, M. Dohler, et al., "5G-Enabled tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. .3, pp. 460-473, Mar. 2016

[2] M. Maier, M. Chowdhury, B. P. Rimal and D. P. Van, "The tactile internet: Vision, recent progress, and open challenges," *IEEE Commn. Magazine*, vol. 54, no. 5, pp. 138-145, May 2016.

[3] S. C. Lin and K. C. Chen, "Statistical QoS control of network coded multipath routing in large cognitive machine-to-machine networks," *IEEE Internet Things J.*, vol. 3, no. 4, pp. 619-627, Aug. 2016.

[4] W. Guo, S. Zhou, Y. Chen, S. Wang, X. Chu and Z. Niu, "Simultaneous information and energy flow for IoT relay systems with crowd harvesting," *IEEE Commn. Magazine*, vol. 54, no. 11, pp. 143-149, Nov.2016.

[5] C. Chen, J. Yan, et al., "Ubiquitous monitoring for industrial cyber-physical systems over relay-assisted wireless sensor networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 3, pp. 352-362, Sep. 2015.

[6] S. Girs, A. Willig, E. Uhlemann and M. Björkman, "Scheduling for source relaying with packet aggregation in industrial wireless networks," *IEEE Trans. Ind. Informat.*, vol. 12, no. 5, pp. 1855-1864, Oct. 2016.

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

[8] ——, "Dispersion of the Gilbert-Elliott channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1829–1848, April 2011.

[9] W. Yang, G. Durisi, T. Koch and Y. Polyanskiy "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, July 2014.

[10] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP J. Wireless Commun.*, vol. 2013:290, Dec. 2013.

[11] G. Ozcan and M. C. Gursoy, "Throughput of cognitive radio systems with finite blocklength codes," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2541-2554, Nov. 2013.

[12] P. Wu and N. Jindal, "Coding versus ARQ in fading channels: How reliable should the phy be?" *IEEE Trans. Commun.* , vol. 59, no. 12, pp. 3363–3374, Dec. 2011.

[13] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commn. Lett.*, vol. 3, no. 5, pp. 529–532, Oct. 2014.

[14] ——, "Finite block-length analysis of spectrum sharing networks using rate adaptation," *IEEE Trans. Commun.* , vol. 63, no. 8, pp. 2823–2835, Aug. 2015.

[15] L. Dai, B. Wang, Y. Yuan, et al., "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun Magazine*, vol. 53, no. 9, pp. 74-81, Sep. 2015.

[16] M. Ali, H. Tabassum, E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, no. , pp. 6325-6343, 2016.

[17] X. Yue, Y. Liu, S. Kang and A. Nallanathan, "Performance analysis of NOMA with fixed gain relaying over Nakagami-m fading channels," *IEEE Access*, vol. 5, no. , pp. 5445-5454, 2017.

[18] Xiaofang Sun, et al., "Short-packet communications in non-orthogonal multiple access systems", 21 Apr 2017. arXiv:1704.06420v1

[19] Y. Hu, J. Gross, and A. Schmeink, "On the capacity of relaying with finite blocklength," *IEEE Trans. Veh. Technol.,* vol. 65, no. 3, pp. 1790-1794, Mar. 2016.

[20] ——, "On the performance advantage of relaying under the finite blocklength regime," *IEEE Commn. Letters, IEEE*, vol. 19, no. 5, pp. 779 – 782, May 2016.

[21] 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, pp. 4548-4558, July. 2016.