Multi-hop Relaying: An End-to-End Delay Analysis

Anas Chaaban and Aydin Sezgin

***Abstract*—The impact of multi-hopping schemes on the communication latency in a relay channel is studied. The main aim is to characterize conditions under which such schemes decrease the communication latency given a reliability requirement. Both decode-forward (DF) and amplify-forward (AF) with block coding are considered, and are compared with the point-to-point (P2P) scheme which ignores the relay. Latency expressions for the three schemes are derived, and conditions under which DF and AF reduce latency are obtained for high signal-to-noise ratio (SNR). Interestingly, these conditions are more strict when compared to the conditions under which the same multi-hopping schemes achieve higher long-term (information-theoretic) rates than P2P. It turns out that the relation between the source-destination SNR and the harmonic mean of the SNR's of the channels to and from the relay dictates whether multi-hopping reduces latency or not.**

I. Introduction

Low latency is an important requirement in many communication scenarios (security, emergency, multimedia, entertainment, etc.). In such scenarios, a number of bits has to be transmitted from a source to a destination with a given reliability (error probability) within a deadline, leading to the so-called delay-limited communication [2]–[5]. Methods for reducing communication latency are thus of practical interest. Several methods have been examined for this purpose, such as finite block-length channel coding [6] and feedback [7]–[10].

Future cellular system architectures will benefit from small-cell deployment [11] (encompassing micro-, pico-, and femto-cells) to facilitate meeting the growing demand for higher data rates. Such small-cell deployment is based on serving a user using an intermediate base-station acting as a relay that covers a small area. This situation can be modeled by the so-called relay channel (RC). Analyzing the delay in this case is of practical interest. In particular, it is interesting to know whether and when the relay node has a positive impact on the delay in the RC. By answering this question, we can identify conditions which can be used by a macro base-station to (i) decide whether to incorporate a relay node in the transmission based on the delay requirement, or (ii) select a relay from a given set of relays that leads to the lowest delay. The question we examine in this paper is thus: When can a relay node reduce end-to-end delay?

One way to look at this problem involves analyzing the communication rate that can be achieved in a network with relays. Indeed a relay can increase the communication rate [12]. In this context, the achievable rates are derived under the requirement that the error probability approaches zero as the length of a transmission block goes to infinity. Moreover, the relaying gain (in terms of rate) is achieved by transmitting over an infinite number of blocks [12]–[16]. In this case, decoding at the receivers ends after the end of the last transmission block. Clearly, this increased communication rate is attained at the expense of a large latency. This perspective which does not take communication delay into account is thus not suitable for low-latency communications.

Thus, this problem has to be approached from a different perspective. Namely, the latency has to be calculated for a given number of bits to be transmitted under a given error probability requirement. This is captured by the error exponent framework [17]. This framework has been applied earlier to RCs [18]–[23]. For instance, [18] studied the optimal number of hops in a separated amplify-forward (AF) half-duplex relay network from an error exponent point of view. Separated here means that a physical channel exists only between neighboring nodes. Similarly, [19] considered a separated multi-hop network with a comparison of the error exponents of concatenated coding and pass-or-decode schemes. The paper [20] studied the error exponent of a separated AF two-way relay channel, and optimized the exponent with respect to rate and power allocations, while [21] studied a separated parallel relay channel, and compared decode-forward (DF), compress-forward, and quantize-forward from an error exponent perspective.

In contrast to [18]–[21], [23] studied the transmission delay in a non-separated multi-hop network as a function of the number of hops and transmission blocks. The resulting interference due to the non-separated nature of the network has been treated as additional noise in [23], which reduces the achievable rate. Alternatively, this interference can be exploited by coherent combining [12] or canceled by backward decoding. In [22] a non-separated relay channel was studied, and error-exponents for half-duplex AF and DF, and full-duplex DF with block coding and coherent combining were derived. Tan [24] provides an in-depth study of the error exponents of partial DF and compress-forward with block-Markov encoding over a discrete memoryless RC from an information-theoretic point of view. In particular, [24] studies the error exponents of these coding schemes using the method of types, and also derives an upper bound on the error exponent of the RC. Note that while the schemes considered in [24] are better than aforementioned schemes (AF, DF, etc.), they are rather more sophisticated from a practical point of view. In this paper, we restrict ourselves to simple DF and AF schemes as described next.

We define the latency in this paper, or the end-to-end delay, as the time from the beginning to the end of transmission of a message (file) consisting of B bits, where the probability that the wrong message is received does not exceed a reliability

A. Chaaban is with the Department of Computer, Electrical, and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi-Arabia (email: anas.chaaban@kaust.edu.sa).

A. Sezgin is with the Institute of Digital Communication Systems, Ruhr-Universität Bochum (RUB), Germany (email: aydin.sezgin@rub.de).

Parts of the paper have been presented in the ICCSPA 2015 [1].

requirement ϵ. This message can be sent over one block or over multiple blocks. Note that in transmission where a message is split into multiple blocks, computing the per-block error-exponent is not sufficient for characterizing the end-to-end delay. In this case, decoding ends after the end of all transmission blocks, and thus this has to be taken into account. Now for a given error probability requirement, the latency of communication given by the block length is bounded [25], [26]. With the aforementioned block coding structure in mind, it might seem at first that multi-hop relaying increases latency. However, we show in this paper that under some conditions on the SNR's, the end-to-end latency is reduced in comparison to the P2P channel even when block coding is taken into account.

Our approach towards computing the end-to-end delay in the RC is based on the following steps:

1) First, the bits to be delivered are divided into several blocks.
2) Then, a per-block error probability is defined, so that the message-error probability does not exceed the error probability requirement.
3) Next, the per-block error probability is divided between the uplink to the relay and the downlink to the destination.
4) Having the number of bits and error probability requirement per block, we can find the required block-length using the error exponent framework. With this, we obtain the length of uplink and downlink blocks, which are equal in AF but not necessarily so in DF.
5) Finally, we take the number of blocks into account to find the delay from the beginning of the first block till the end of the last block. This delay is a function of the number of bits, error probability requirement, the SNR's, and has to be minimized with respect to the number of blocks.

At this point, a comparison with the work in [22] is due. The authors of [22] studied the per-block error exponent of full-duplex DF (step 4 above). However, the impact of the number of blocks was not considered in [22], and hence the end-to-end delay was not discussed. Indeed, the result of [22] can be embedded in our framework (steps 1 to 5 above) to obtain the corresponding end-to-end delay. Another difference between our work and [22] is that we allow the lengths and error probabilities of the uplink and downlink blocks to be different when using DF, which makes the design more flexible.

We build on the above framework to analyze the latency of two multi-hopping type relaying schemes based on DF and AF, and derive conditions under which these schemes are beneficial in terms of delay. Here, multi-hopping refers to schemes which do not exploit the existence of the source-destination channel for coherent combining of signals at the destination. The reason we restrict ourselves to these schemes is three-fold. First, these are simple schemes, and more suited to practical applications. Second, these two schemes, combined with the P2P scheme (which ignores the relay) achieve the capacity of the RC within a constant gap. This can be verified using methods similar to [27, Appendix A]. Finally, studying the latency of these simple schemes provides an upper bound on the latency that can be achieved by more sophisticated schemes [24]. Thus, by studying the simpler multi-hopping DF and AF schemes, we can find *sufficient conditions* under which multi-hop relaying with backward decoding reduces latency.

Our contributions can thus be summarized as follows:

- We derive expressions for the latency of multi-hopping DF and AF.
- We compare the latency of these schemes to that of the P2P scheme (benchmark).
- We approximate the latency expressions at high SNR, and identify *sufficient* conditions guaranteeing that relaying reduces latency.

It is important here to highlight an interesting interplay between different parameters in the analysis. The DF and AF schemes transmit the information to the destination by distributing the information over multiple blocks. The number of block is a design parameter that should be chosen to minimize latency. Increasing the number of blocks reduces the number of bits per block on one hand, and results in a stricter reliability requirement per block on the other hand. This follows since the transmission will be erroneous if at least one block is erroneous. Thus, if the overall reliability requirement is ϵ, the per-block reliability requirement is ϵ/L where L is the number of blocks. From this point of view, the best choice of L is not clear. We show that one transmission block is optimal for a small number of information bits to be transmitted, while multiple blocks yield better latency for a large number of bits. In both cases, the latency of DF and AF can be lower than that of P2P despite the multi-block transmission structure. Still in some cases (conditions on the SNR's), P2P yields lower latency. Thus, it is interesting find conditions under which DF and AF reduce latency.

By studying the schemes at high SNR, we obtain the following conclusion. If the relay increases the capacity of the P2P channel, it does not necessarily reduce the latency of transmission. On the other hand, if either DF or AF reduces latency, then its long-term achievable rate has to be higher than that of P2P. Roughly speaking, while DF increases the information-theoretic capacity of in comparison to P2P if both the source-relay and relay-destination SNR's are larger than the source-destination SNR, DF reduces latency if the *harmonic mean* of the source-relay and relay-destination SNR's (in dB) is larger than twice the source-destination SNR (in dB). Thus, DF reduces latency under a stricter condition. AF reduces latency if the SNR of AF is larger than twice the source-destination SNR. These conditions are practically interesting for relay activation/deactivation and for relay selection as mentioned earlier.

These aspects will be discussed in detail throughout the paper. In the next section, we introduce the system model of the RC and provide the problem formulation.

II. System Model and Problem Formulation

We consider a relay channel (RC) as shown in Fig. 1 where the source node wants to send a message m of B bits to the destination node with the aid of a full-duplex relay. At time instant $i \in \{1, \cdots, N\}$, the source sends the real-valued

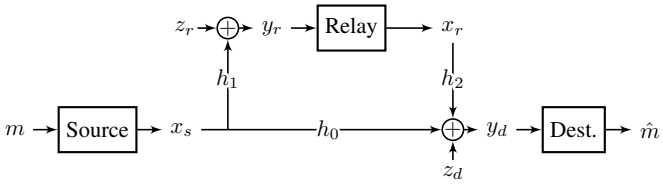


Fig. 1. A mathematical model of the Gaussian relay channel.

signal $x_s(i)$ to the relay and the destination. These nodes in turn observe the following received signals, respectively,

$$y_r(i) = h_1 x_s(i) + z_r(i), \tag{1}$$
$$y_d(i) = h_0 x_s(i) + h_2 x_r(i) + z_d(i). \tag{2}$$

Here, $x_r(i)$ is the relay transmit signal constructed from the relay's received signal up to time instant $i-1$, i.e., $y_r(1), \cdots, y_r(i-1)$. The variables $z_r(i)$ and $z_d(i)$ are independent Gaussian noises with zero mean and unit variance. The scalars $h_0, h_1, h_2 \in \mathbb{R}$ are the source-destination, source-relay, and relay-destination channel coefficients, respectively. It is assumed that h_0, h_1, and h_2 maintain the same value throughout the transmission. Each of the source and the relay have a power constraints given by $\mathbb{E}[x_s^2] \leq P$ and $\mathbb{E}[x_r^2] \leq P_r$.

After N transmissions, the destination decodes $\hat{m}$ from $y_d(1), \cdots, y_d(N)$. The induced error probability of this procedure is $P_e = \text{Prob}\{m \neq \hat{m}\}$. The transmission has to satisfy

$$P_e < \epsilon, \tag{3}$$

where $\epsilon > 0$ is a pre-defined reliability requirement.

The main questions we would like to answer in this paper are: What is the latency of this communication? And when does the relay have a positive impact on this latency? More precisely, we aim for finding the value of N that has to be chosen so that the message m with B bits can be delivered to the destination with an error probability less that ϵ, while taking the number of transmission blocks in to account.

An instrumental quantity for this study is the error exponent of a coding scheme [28]. Gaussian coding with rate R bits per transmission over a Gaussian P2P channel with a signal-to-noise ratio SNR achieves an error exponent given by [25]

$$E_r(R, \mathsf{SNR}) = \max_{\rho \in [0,1]} \left[\frac{\rho}{2} \log_2 \left(1 + \frac{\mathsf{SNR}}{1+\rho} \right) - \rho R \right]. \tag{4}$$

Given this error exponent, the latency of communicating B bits over this channel with reliability ϵ is upper bounded by[1]

$$n(B, \mathsf{SNR}, \epsilon) = \min_{\rho \in [0,1]} \frac{\rho B - \ln(\epsilon)}{\frac{\rho}{2} \log_2 \left(1 + \frac{\mathsf{SNR}}{1+\rho} \right)}. \tag{5}$$

The function $n(B, \mathsf{SNR}, \epsilon)$ will be used frequently in the paper for bounding the latency of a transmission scheme over a RC. Next, we summarize the main contribution of the paper.

[1]This can be obtained by using (4), the asymptotic equivalence $P_e \approx e^{-NE_r}$ [25], and $B = nR$.

III. Summary of Main Results

In the following sections, we will derive the achievable latency of using Gaussian coding in the RC with multi-hopping DF and AF at the relay (henceforth simply DF and AF). We are going to prove that DF and AF achieve a latency of

$$N_{DF} = \min_{\substack{L \in \mathbb{N} \setminus \{0\} \\ \delta \in (0,1)}} \max \left\{ \begin{array}{l} n_1(L,\delta) + L n_2(L,\delta), \\ L n_1(L,\delta) + n_2(L,\delta) \end{array} \right\}, \tag{6}$$

$$N_{AF} = \min_{L \in \mathbb{N} \setminus \{0\}} (L+1) \cdot n_3(L), \tag{7}$$

respectively, where

$$n_1(L,\delta) = n(B/L, \mathsf{SNR}_1, (1-\delta)\epsilon/L) \tag{8}$$
$$n_2(L,\delta) = n(B/L, \mathsf{SNR}_2, \delta\epsilon/L) \tag{9}$$
$$n_3(L) = n(B/L, \mathsf{SNR}_{AF}, \epsilon/L), \tag{10}$$

and $\mathsf{SNR}_1 = h_1^2 P$, $\mathsf{SNR}_2 = h_2^2 P_r$, and $\mathsf{SNR}_{AF} = \frac{\mathsf{SNR}_1 \mathsf{SNR}_2}{1+\mathsf{SNR}_1+\mathsf{SNR}_2}$. Here, the parameter L is the number of transmission blocks and δ is a trade-off factor between the error probabilities of the uplink and downlink[2] in DF (which allows different length of uplink and downlink blocks). Notice the strict error probability requirement represented by ϵ/L arising due to the block structure.

Despite their simplicity, the latency of DF and AF can be lower than that of P2P given by $N_{P2P} = n(B, \mathsf{SNR}_0, \epsilon)$ where $\mathsf{SNR}_0 = h_0^2 P$, if the relay is properly placed (see Figure 2). In this figure, we show the best scheme in terms of latency as a function of the relay position. The source (square) is located at the origin $(0,0)$ and the destination (diamond) at $(1,0)$ (normalized units of distance). We assume consider a wireless channel with a path-loss exponent of 3. The channels h_1 and h_2 depend on the relay position. If the relay is located in the region marked by x's, then P2P achieves lower latency than both DF and AF. However, if the relay is located in the region marked by o's, then DF achieves lower latency than both AF and P2P. These positions marked by o are potential positions where a relay might be placed in a cellular communications scenario for instance, since the relay is normally located between the transmitter and the receiver.

Note that the region marked with o's is a sub-set of the region bounded between the two black curves, where DF achieves higher information-theoretic rate[3] given by $R \approx \log(\min\{\mathsf{SNR}_1, \mathsf{SNR}_2\})$ than P2P which achieves $R \approx \log(\mathsf{SNR}_0)$. This interestingly means that if DF increases the achievable rate, it does not necessarily reduce latency. However, in the inner region marked by o's, DF indeed reduces latency in comparison to P2P. While DF provides lower latency than AF, the latter has the advantage of reduced computational requirements at the relay node. Thus, in cases where the relay has computational limitation, AF can be a favored scheme in the region marked by +'s.

To obtain a closer look on the conditions under which a relay reduces latency in a RC, we consider the high SNR regime. We have the following statement regarding DF.

[2]Throughout the paper, we refer to the source-relay channel as the uplink channel, and the relay-destination channel as the downlink channel.

[3]An information-theoretic rate is computed under the condition that $P_e \to 0$ as $N \to \infty$.

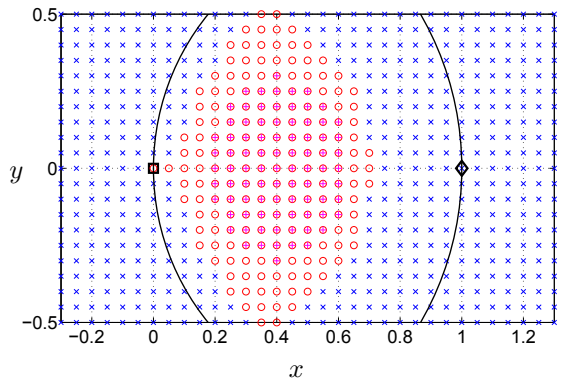


Fig. 2. The best scheme as a function of the relay position for a relay channel with a path-loss exponent of 3, $P = 20$dB (relative to noise variance), $P_r = 2P$, $\epsilon = 10^{-3}$, and $B = 10$kbit. The square and the diamond denote the source and the destination, respectively. The ∘ and × denote relay positions at which DF provides lower latency than P2P, and vice versa, respectively. The + denotes relay positions where AF performs better than P2P. The region between the two black curves intersecting with the positions of the source and destination is where DF has a higher achievable rate than P2P.

Proposition 1. *At high* SNR*, a sufficient condition under which DF has a lower latency than P2P is given by*

$$M_h(\mathsf{SNR}_1|_{dB}, \mathsf{SNR}_2|_{dB}) > 2(\mathsf{SNR}_0|_{dB}), \tag{11}$$

where $M_h(x, y)$ is the harmonic mean *of x and y defined as $M_h(x, y) = \frac{2xy}{x+y}$, and $x|_{dB}$ is defined as $10\log_{10}(x)$, i.e., the value of x in dB.*

This statement leads to the following interesting conclusion. If the high SNR achievable rate of DF given by $\frac{1}{2}\log_2(\min\{\mathsf{SNR}_1, \mathsf{SNR}_2\})$ is higher than that of the P2P scheme given by $\frac{1}{2}\log_2(\mathsf{SNR}_0)$, then DF does not necessarily provide lower latency than P2P! Note that while the superiority of DF in terms of the information-theoretic achievable rate is dictated by the bottle-neck SNR between SNR_1 and SNR_2, i.e., the condition $\min\{\mathsf{SNR}_1, \mathsf{SNR}_2\} > \mathsf{SNR}_0$, its superiority in terms of latency is dictated by both SNR's as seen in (11).

A similar statement holds for AF, for which we have the following statement.

Proposition 2. *At high* SNR*, a sufficient condition under which AF has a lower latency than P2P is given by*

$$\left.\frac{M_h(\mathsf{SNR}_1, \mathsf{SNR}_2)}{2}\right|_{dB} > 2(\mathsf{SNR}_0|_{dB}). \tag{12}$$

Note that here, the condition is on the dB value of the harmonic mean, contrary to (11) in which the condition is on the harmonic mean of the dB values of the SNRs.

Both the DF and AF schemes can reduce the latency of transmission, but under a stricter condition than merely having a larger achievable rate.

It turns out that in general, transmission using DF or AF should be carried out over only one transmission block (for each of the uplink and the downlink) for a small payload, but over several blocks for large payload. Interestingly, although the use of multiple transmission blocks imposes a stricter error probability requirement per block, the overall transmission can still have lower latency than P2P. Next, we describe the three main transmission schemes of this paper.

IV. Transmission Schemes

The number of transmissions required to satisfy (3) depends on the scheme being used over the RC. The benchmark for our work is the scheme without a relay. The reason to choose this scheme as a benchmark is to check when the relay can in fact decrease the latency of this communication. If the relay is inactive, then the RC becomes a point-to-point channel (P2P) with $\mathsf{SNR}_0 = h_0^2 P$. The optimal code for this P2P channel is a random Gaussian code [29]. In this case, the source encodes the message m into a sequence of length N_{P2P} whose components are i.i.d. Gaussian with zero mean and variance P. The destination decodes after observing N_{P2P} received symbols. The latency of this scheme is given by N_{P2P}, which has to be chosen such that the error probability is below ϵ. Next, we describe schemes that incorporate the relay.

A. Decode-forward

In decode-forward (DF), the relay decodes the signal sent by the source, and forwards it to the destination in the next transmission block. Here, we use a simple variant of DF which does not incorporate superposition block-Markov encoding [12], [30]. This simplification is made since the channel capacity can be achieved within a constant gap by DF and P2P without superposition block-Markov encoding. This can be verified using methods similar to [27]. Furthermore, this simplifies the analysis of the problem at hand, and provides an upper bound on the latency of more sophisticated DF schemes.

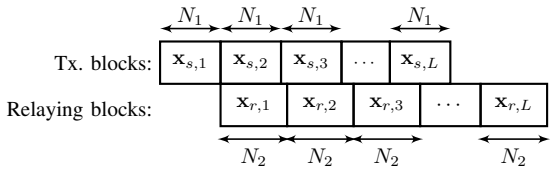
The source splits m into L equal parts, denoted $m_1, \cdots, m_L$, each with B' bits, i.e., $B' = B/L$. If B is not divisible by L, we zero-pad B to the length $\tilde{B} = L\left\lceil \frac{B}{L} \right\rceil$. We assume that B is large enough (in comparison to L) so that the number of padded zeros $\tilde{B} - B$ is negligible with respect to B, and hence $\tilde{B}/L \approx B/L = B'$. Then, the source encodes each message m_ℓ, $\ell = 1, \cdots, L$, into a codeword $\mathbf{x}_{s,\ell}$ of length N_1 using a Gaussian code with power P. Afterwards, the source sends $\mathbf{x}_{s,\ell}$ in the ℓ-th transmission block.

The relay waits until it has received N_1 symbols, after which it decodes $\hat{m}_{r,1}$ (which is equal to m_1 unless an error occurs). Thus, the channel from the source to the relay is treated as a P2P channel with signal-to-noise ratio

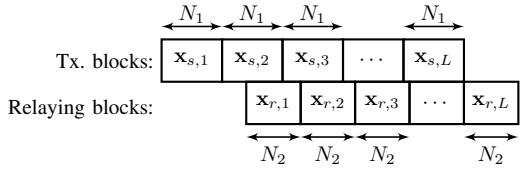
$$\mathsf{SNR}_1 = h_1^2 P.$$

The relay then encodes $\hat{m}_{r,1}$ into $\mathbf{x}_{r,1}$ using a Gaussian code with power P_r and length N_2, and sends $\mathbf{x}_{r,1}$ in the first relaying block. The first relaying block begins at time instant $\tau + 1$ where $\tau \geq N_1$ is to be determined later. The relay proceeds similarly by decoding $\hat{m}_{r,\ell}$ after the end of the ℓ-the transmission block, and forwarding it in the ℓ-th relaying block, until all message parts have been sent. The whole process takes $N_{DF} = \tau + LN_2$.

The destination listens and stores the received signals during the whole transmission and relaying time, and starts decoding



(a) $N_1 < N_2$.



(b) $N_1 \geq N_2$.

Fig. 3. The block structure of transmission using decode-forward at the relay.

backwards at the end of the transmission.[4] The destination starts by decoding the last message part $\hat{m}_L$ (which is equal to the L-th relay message $\hat{m}_{r,L}$ unless an error occurs). We require that $\mathbf{x}_{r,L}$ is received free of interference (from $\mathbf{x}_{s,L}$) at the relay. This is achieved by ensuring that the transmission of $\mathbf{x}_{s,L}$ from the source is completed before the transmission of $\mathbf{x}_{r,L}$ from the relay starts, as shown in Fig. 3. Hence, $LN_1 \leq \tau + (L-1)N_2$ which yields $\tau \geq LN_1 - (L-1)N_2$. Thus, by choosing $\tau = \max\{N_1, LN_1 - (L-1)N_2\}$ (see Fig. 3), the overall duration of communication in DF becomes

$$N_{DF} = \max\{N_1 + LN_2, LN_1 + N_2\}.$$

By using this procedure, the channel from the relay to the destination becomes a P2P channel with signal-to-noise ratio

$$\mathsf{SNR}_2 = h_2^2 P_r.$$

After decoding $\hat{m}_L$, the destination constructs $\hat{\mathbf{x}}_{s,L}$ and uses it to cancel the contribution of $\mathbf{x}_{s,L}$ from its received signal. Then the destination decodes the last but one message part $\hat{m}_{L-1}$. This process continues until all message parts have been recovered at the destination. Notice that if $\hat{m}_L = m_L$, then perfect cancellation of $\mathbf{x}_{r,L-1}$ can be carried out. Otherwise, interference cancellation can not be done perfectly. In this case, an error might occur while decoding $\hat{m}_{L-1}$. This error propagates till block $\ell = 1$. The impact of this on the error probability will be discussed in Section V.

The value of L has to be chosen so that the latency of the whole transmission is minimized. Choosing $L = 1$ leads to a large message with B bits, and hence large block sizes N_1 and N_2 and a latency of $N_1 + N_2$. On the other hand, increasing L leads to smaller messages m_ℓ (less bits), and hence smaller N_1 and N_2, but more blocks. The overall latency of the communication depends on N_1, N_2, and L, which in turn depend on the reliability constraint $P_e < \epsilon$ (3). More on that will follow in Section V.

Note that DF requires decoding at the relay. Consequently, a reliability requirement has to be guaranteed not only at the destination, but also at the relay. The reliability requirement at the relay can be avoided by refraining from decoding at the relay, by using AF instead.

[4]We assume that the processing delay at the destination is negligible compared to the transmission delay.

B. Amplify-forward

In AF, the source encodes similar to DF, by splitting m into L messages $(m_1, \cdots, m_L)$ and sending the messages in L blocks. We denote the length of the codeword used by the source by N_3. Similar to DF, the relay waits until time instant $\tau \geq N_3$, and then starts transmission at time instant $\tau + 1$.

The relay scales the received signal $y_r(i)$ at time i by

$$\alpha = \sqrt{P_r/(1 + h_1^2 P)}$$

and sends it in time instant $i + \tau$. This scaling guarantees satisfying the power constraint at the relay. This leads to an equal length of transmission and relaying blocks, i.e., the length of the relaying block is also N_3.

The destination receives a superposition of the transmit signal and the relay signal. It starts decoding from the last block. To guarantee that the last relaying block is free of interference, we need to choose $\tau = N_3$. During the L-th relaying block, the destination receives

$$\mathbf{y}_{d,L} = h_2 \alpha (h_1 \mathbf{x}_{s,L} + \mathbf{z}_{r,L}) + \mathbf{z}_{d,L}, \tag{13}$$

where $\mathbf{x}_{s,L}$ is the source signal corresponding to m_L, $\mathbf{z}_{r,L} = (z_r([L-1]N_3 + 1), \cdots, z_r(LN_3))$ is the relay noise during the L-th transmission block, and $\mathbf{z}_{d,L} = (z_d(LN_3 + 1), \cdots, z_d([L+1]N_3))$ is the noise at the destination during the L-th relaying block. The destination decodes $\hat{m}_L$ (which is equal to m_L unless an error occurs) from (13) which resembles a P2P channel with a signal-to-noise ratio of

$$\mathsf{SNR}_{AF} = \frac{\mathsf{SNR}_1 \mathsf{SNR}_2}{1 + \mathsf{SNR}_1 + \mathsf{SNR}_2}. \tag{14}$$

After decoding, $\hat{m}_L$ is used to cancel the contribution of $\mathbf{x}_{s,L}$ from the $(L-1)$-th relaying block, which works if $\hat{m}_L = m_L$. Next, the destination decodes $\hat{m}_{L-1}$. This proceeds until all messages are decoded. The overall latency of this scheme is

$$N_{AF} = (L+1)N_3,$$

where L and N_3 have to be chosen to satisfy the reliability constraint $P_e < \epsilon$. Next, we compute the latency of these schemes.

V. Performance Analysis

A. Latency of the P2P scheme

Recall that the P2P scheme has an SNR of $\mathsf{SNR}_0 = h_0^2 P$. The B bits can be delivered in this case to the destination with reliability ϵ if N_{P2P} is chosen such that (5)

$$N_{P2P} \geq n(B, \mathsf{SNR}_0, \epsilon).$$

B. Latency of the DF scheme

The DF scheme delivers the B bits in L blocks, each containing $B' = B/L$ bits. Here, an error occurs if the event $E_\ell = \{\hat{m}_\ell \neq m_\ell\}$ occurs for any $\ell = 1, \cdots, L$, since an error propagates[5] due to backward decoding and interference cancellation (cancellation of the transmit signal $\mathbf{x}_{s,L}$ after decoding $\hat{m}_L$). Denote the block-error probability by $P_{e,b}$. Thus, an error occurs in DF if $E_1 \cup E_2 \cup \cdots \cup E_L$ occurs, and we can upper bound the probability of this event by

$$P_{e,DF} = \text{Prob}(E_1 \cup E_2 \cup \cdots \cup E_L) \leq L P_{e,b}, \tag{15}$$

by the union bound. Therefore, if we can guarantee that $P_{e,b} < \frac{\epsilon}{L} = \epsilon'$, then we guarantee that $P_{e,DF} < \epsilon$. From this calculation, we conclude that by increasing L, we decrease the number of bits B' that should be delivered per block, but we get a stricter reliability requirement per block given by $P_{e,b} < \epsilon'$, which in turn increases the block length. Whether this increase in L has an advantage strongly depends on the parameters of the system B, ϵ, SNR_1, and SNR_2.

We conclude that we can only tolerate a block error probability $P_{e,b} < \epsilon'$. However, when does a block error occur in DF? A block error occurs if $m_\ell \neq \hat{m}_\ell$. This in turn occurs if either the event E_{b1} or E_{b2} occur, where E_{b1} corresponds to the case that the relay decodes correctly while the destination does not: $E_{b1} = \{m_\ell = \hat{m}_{r,\ell} \wedge m_\ell \neq \hat{m}_\ell\}$, and E_{b2} corresponds to the case where both the relay and the destination decode incorrectly: $E_{b2} = \{m_\ell \neq \hat{m}_{r,\ell} \wedge m_\ell \neq \hat{m}_\ell\}$. The probability of E_{b1} can be upper bounded by the probability of error at the destination $P_{e,d} = \text{Prob}\{m_\ell \neq \hat{m}_\ell\}$. On the other hand, the probability of E_{b2} can be upper bounded by the probability of error at the relay $P_{e,r} = \text{Prob}\{m_\ell \neq \hat{m}_{r,\ell}\}$. Therefore, we can write $P_{e,b} < P_{e,r} + P_{e,d}$.

If we set $P_{e,d} < \delta\epsilon'$ and $P_{e,r} < (1-\delta)\epsilon'$, $\delta \in (0,1)$, then we guarantee that $P_{e,b} < \epsilon'$. The advantage of this trade-off parameter δ is that it allows exploiting the better channel among h_1 and h_2 to allow higher error probability tolerance at the weaker channel.

Having bounded $P_{e,r}$ and $P_{e,d}$, now we can write the length of blocks N_1 and N_2 by using (5) as follows

$$N_1 \geq n_1(L,\delta) = n(B', \mathsf{SNR}_1, (1-\delta)\epsilon') \tag{16}$$

$$N_2 \geq n_2(L,\delta) = n(B', \mathsf{SNR}_2, \delta\epsilon'). \tag{17}$$

The strict error probability requirement of DF becomes obvious in these expressions of N_1 and N_2. The error probability tolerance is reduced from ϵ to ϵ' due to the block structure, and further reduced to $\delta\epsilon'$ and $(1-\delta)\epsilon'$ due to decoding each message twice.

The parameter δ and L have to be chosen so that the latency of DF N_{DF} is minimized. The minimum latency of DF can thus be written as

$$N_{DF} = \min_{\substack{L \in \mathbb{N}\setminus\{0\} \\ \delta \in (0,1)}} \max \left\{ \begin{array}{l} n_1(L,\delta) + L n_2(L,\delta), \\ L n_1(L,\delta) + n_2(L,\delta) \end{array} \right\}. \tag{18}$$

[5]This is a worst case consideration since an error in block L might, but does not necessarily, lead to an error in block $L-1$.

C. Latency of the AF scheme

Recall that AF requires decoding only at the destination. This relaxes the error probability requirement since we do not need the parameter δ anymore. However, this comes at the expense of a reduced SNR. Following similar arguments as before, we can write the block size of AF as

$$N_3 \geq n_3(L) = n(B', \mathsf{SNR}_{AF}, \epsilon'). \tag{19}$$

The minimum latency of AF can thus be written as

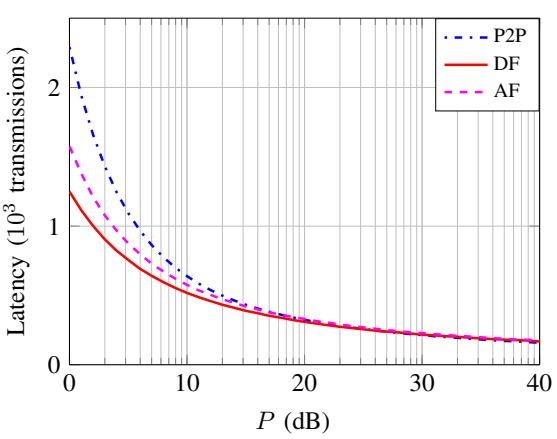
$$N_{AF} = \min_{L \in \mathbb{N}\setminus\{0\}} (L+1) \cdot n_3(L). \tag{20}$$

D. Comparison

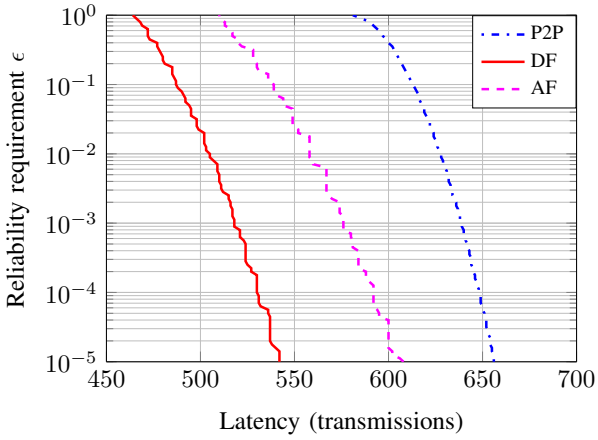
Intuitively, if $\mathsf{SNR}_0 \gg \mathsf{SNR}_1, \mathsf{SNR}_2$, then the latency of the P2P is smaller than that of the DF scheme. Similar statement holds for AF. That is, if the uplink and downlink channels are weak, then DF and AF have a negative impact on the latency. However, if the relay channels are strong enough, then a reduction of the latency can be achieved by DF and AF. This can be seen in Figure 4(a) which shows the latency of each of the P2P, DF, and AF schemes as a function of the power P for a RC with $h_0 = h_2 = 1$, $h_1 = 2$. This setting models a scenario where the relay is close to the source, and the source and relay are equidistant from the destination. Furthermore, it is assumed that $P_r = 16P$ which models scenarios where the relay is e.g. a fixed device which is mounted on a building having access to abundant power (12dB more than the source). In this figure, a message of size $B = 1$kbit and a reliability requirement of $\epsilon = 10^{-3}$ are considered. At $P = 10$dB (relative to noise variance), $N_{P2P} = 639$ transmissions, while $N_{DF} = 518$ transmissions thus reducing the latency by $\approx 20\%$. The performance of the three schemes becomes close at high P since the difference in their SNR's (on a logarithmic scale) becomes negligible in comparison to $\log_2(P)$. More on that will follow in Section VI.

Figure 4(b) gives a closer look at the error probability for the same scenario at $P = 10$dB. This figure shows the latency reduction achieved by relaying. Among the 3 schemes, the best in this case is DF. However, if less computational complexity is required at the relay, then AF can also be used to reduce the latency of the communication.

In Figure 5, the latency is plotted as a function of the message size B. In this figure, we can see that the performance of DF is close to that of P2P at small B. Recall that the block structure of DF has an advantage and a disadvantage. The advantage is the decreased number of bits to be delivered per block. The disadvantage is that these bits have to be delivered will a lower error probability. At low B, the advantage is lost since B becomes negligible in comparison to $\ln(\epsilon)$. In other words, the function $n(B, \mathsf{SNR}, \epsilon)$ approaches $n(0, \mathsf{SNR}, \epsilon)$ at low B, and thus, dividing the number of bits to delivered per block by L is irrelevant. This explains the behaviour of DF at low B in Figure 5. However, at high B, this advantage becomes prominent, and DF becomes better than P2P. In this example, at $B = 10$kbit we have a decrease in latency from ≈ 6000 transmissions for P2P to ≈ 4400 transmissions for DF, a drop of $> 25\%$.



(a) $B = 1$kbit, $\epsilon = 10^{-3}$.



(b) $B = 1$kbit, $P = 10$dB.

Fig. 4. The performance of the P2P, DF, and AF schemes for a relay channel with $h_0 = h_2 = 1$, $h_1 = 2$, $P_r = 16P$.

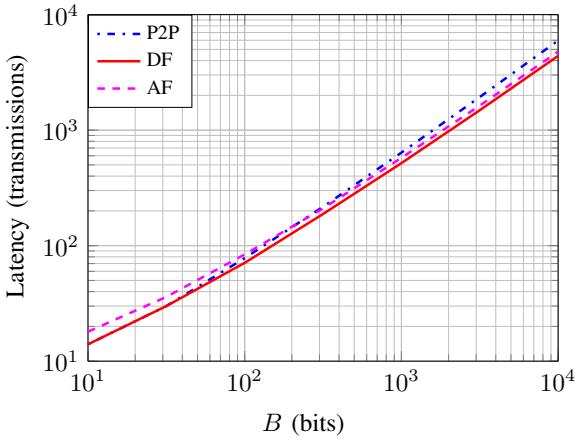


Fig. 5. The latency of the P2P, DF, and AF schemes for a relay channel with $h_0 = h_2 = \frac{h_1}{2} = 1$, $P = 10$dB, $P_r = 16P$, under reliability $\epsilon = 10^{-3}$

Recall that our DF scheme is a multi-hop scheme which does not exploit the existence of the direct link h_0. The error exponent of a DF scheme which exploits this direct channel has been studied in [22]. This channel is exploited by coherently combining the received signals from the source and the relay at the destinations, where the lengths of uplink and downlink blocks are equal. On the other hand, the simpler multi-hop DF scheme we consider in this paper has more flexibility in choosing the block lengths. From this point of view, it is important to compare the two schemes. Figure 6 shows a comparison for a relay channel where the three nodes are co-linear, and the source-destination, source-relay, and relay-destination distance is 1, d, and $1-d$ units of distance, respectively. The path-loss exponent is 4, the file size is 1kbit, the error probability requirement is 10^{-3}, and $P = P_r = 10$dB. The figure shows the end-to-end delay as a function of the number of blocks L for multi-hop DF (DF-M) studied in this paper. It also shows the delay obtained by embedding the per-block error exponent of coherent DF (DF-C) from [22] in our framework. We note that the advantage of coherent combining appears when the relay is closer to the source, while the advantage of unequal transmission/relaying block lengths appears when the relay is closer to the destination. Recall that we have to choose L which minimizes the end-to-end delay. By choosing the optimal L, this figure shows that multi-hop DF (DF-M) achieves good performance, while having lower complexity than coherent DF (DF-C).

Next, we analyse the performance of the three schemes described in Section IV at high SNR, in order to obtain the statements of Propositions 1 and 2.

VI. High SNR Analysis

We start by approximating the P2P latency at high SNR.

A. Latency of the P2P scheme

Let us first approximate the function $n(B, \mathsf{SNR}, \epsilon)$ at high SNR. This function is defined as follows (5)

$$n(B, \mathsf{SNR}, \epsilon) = \min_{\rho \in [0,1]} \frac{\rho B - \ln(\epsilon)}{\frac{\rho}{2} \log_2\left(1 + \frac{\mathsf{SNR}}{1+\rho}\right)}. \tag{21}$$

Note that at high SNR, the denominator of (21) can be approximated for any ρ as

$$\frac{\rho}{2} \log_2\left(1 + \frac{\mathsf{SNR}}{1+\rho}\right) \approx \frac{\rho}{2} \log_2(\mathsf{SNR}). \tag{22}$$

Substituting in (21) yields

$$n(B, \mathsf{SNR}, \epsilon) \approx \min_{\rho \in [0,1]} \frac{\rho B - \ln(\epsilon)}{\frac{\rho}{2} \log_2(\mathsf{SNR})} \tag{23}$$

$$= \frac{2B - 2\ln(\epsilon)}{\log_2(\mathsf{SNR})}. \tag{24}$$

Now we use this approximation to write the latency of the P2P scheme at high SNR. This can be approximated as

$$N_{P2P} \approx \frac{2B - 2\ln(\epsilon)}{\log_2(\mathsf{SNR}_0)}. \tag{25}$$

Next, we use the approximation in (24) to express the latency of DF and AF at high SNR. To this end, we consider large P and P_r leading to large SNR_1, SNR_2, and SNR_{AF}.

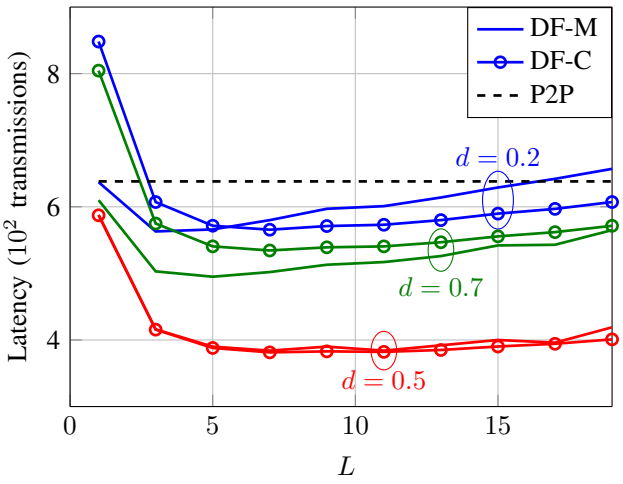


Fig. 6. The end-to-end delay of multi-hop DF (DF-M) and coherent DF (DF-C) [22] as a function of the number of blocks L for sending 1kbit with an error probability requirement of 10^{-3}. The powers at the source and the relay are 10dB, and the relay is placed on the line between the source and the destination at a distance d from the source and $1-d$ from the destination. The path-loss exponent is assumed to be 4.

B. Latency of the DF scheme

At high SNR, the block length parameters of the DF scheme can be approximated as

$$n_1(L,\delta) \approx \frac{2\frac{B}{L} - 2\ln((1-\delta)\frac{\epsilon}{L})}{\log_2(\mathsf{SNR}_1)}, \tag{26}$$

$$n_2(L,\delta) \approx \frac{2\frac{B}{L} - 2\ln(\delta\frac{\epsilon}{L})}{\log_2(\mathsf{SNR}_2)}. \tag{27}$$

To approximate the latency of DF, we need to minimize $\max\{n_1(L,\delta) + Ln_2(L,\delta), Ln_1(L,\delta) + n_2(L,\delta)\}$ over $L \in \mathbb{N}\setminus\{0\}$ and over $\delta \in (0,1)$. But before we proceed, this is a good point to discuss the impact of L on latency. Let us examine the behaviour of $f(L) \triangleq n_1(L,\delta) + Ln_2(L,\delta)$ as a function of L. We have

$$\begin{aligned} f(L) &= \frac{2\frac{B}{L} - 2\ln((1-\delta)\frac{\epsilon}{2L})}{\log_2(\mathsf{SNR}_1)} + L\frac{2\frac{B}{L} - 2\ln(\delta\frac{\epsilon}{2L})}{\log_2(\mathsf{SNR}_2)} \\ &= \frac{2B}{\mu L} - \frac{2}{\mu}\ln\left((1-\delta)\frac{\epsilon}{2L}\right) + \frac{2B}{\nu} - \frac{2L}{\nu}\ln\left(\delta\frac{\epsilon}{2L}\right), \end{aligned} \tag{28}$$

where we used μ and ν to denote $\log_2(\mathsf{SNR}_1)$ and $\log_2(\mathsf{SNR}_2)$ for clarity. The derivative of $f(L)$ by dL is thus

$$\frac{df}{dL} = -\frac{2B}{\mu L^2} + \frac{2}{L}\left(\frac{1}{\mu} + \frac{L}{\nu}\right) - \frac{2}{\nu}\ln\left(\delta\frac{\epsilon}{2L}\right). \tag{29}$$

We can notice that all terms of this derivative are positive except the first one. A similar behaviour holds for $Ln_1(L,\delta) + n_2(L,\delta)$. This leads to the following conclusion. If B is large enough, then $\frac{df}{dL}$ is negative for small L and positive for large L, which implies that the optimum L is larger than 1. Thus, at high SNR and high B, it is best to divide B into several blocks to minimize latency. On the other hand, for small B, $\frac{df}{dL}$ is always positive, and thus choosing $L=1$ is optimal.

Next, we bound N_{DF} by choosing $\delta = \frac{1}{2}$ and $L=1$ by

$$N_{DF} \le 2\left(B - \ln\left(\frac{\epsilon}{2}\right)\right)\left(\frac{1}{\log_2(\mathsf{SNR}_1)} + \frac{1}{\log_2(\mathsf{SNR}_2)}\right).$$

C. Latency of the AF scheme

At high P and P_r, SNR_{AF} is also high. The block length of the AF scheme is given by $N_3 \ge n_3(L)$ where

$$n_3(L) = n(B', \mathsf{SNR}_{AF}, \epsilon') \approx \frac{2\frac{B}{L} - 2\ln\left(\frac{\epsilon}{L}\right)}{\log_2(\mathsf{SNR}_{AF})}. \tag{30}$$

The total latency of the AF scheme is thus given by

$$N_{AF} = (L+1)\cdot n_3(L) \approx (L+1)\frac{2\frac{B}{L} - 2\ln\left(\frac{\epsilon}{L}\right)}{\log_2(\mathsf{SNR}_{AF})}. \tag{31}$$

The behaviour of N_{AF} as a function of L is similar to $f(L)$ in (28), i.e., it is decreasing and then increasing for large B, and only increasing for small B. Thus, the optimal L is 1 for small B and larger than 1 for larger B. We can bound N_{AF} by setting $L=1$ as follows

$$N_{AF} \le \frac{4B - 4\ln(\epsilon)}{\log_2(\mathsf{SNR}_{AF})}. \tag{32}$$

D. Comparison

Although we have set $L=1$ to upper bound the latency of DF and AF, the resulting latency upper bound of both scheme can be lower than the latency of P2P at high SNR. To show this, let us start by collecting the latency of the three schemes:

$$\begin{aligned} N_{P2P} &\approx \frac{2B - 2\ln(\epsilon)}{\log_2(\mathsf{SNR}_0)} \\ N_{DF} &\le 2\left(B - \ln\left(\frac{\epsilon}{2}\right)\right)\left(\frac{1}{\log_2(\mathsf{SNR}_1)} + \frac{1}{\log_2(\mathsf{SNR}_2)}\right) \\ N_{AF} &\le \frac{4B - 4\ln(\epsilon)}{\log_2(\mathsf{SNR}_{AF})}. \end{aligned}$$

By comparing N_{P2P} and the upper bound for N_{DF}, we obtain the statement of Proposition 1. Namely, at high SNR, if

$$\frac{1}{\log_2(\mathsf{SNR}_1)} + \frac{1}{\log_2(\mathsf{SNR}_2)} < \frac{1}{\log_2(\mathsf{SNR}_0)}, \tag{33}$$

then DF has a lower latency than P2P. This follows by neglecting $\ln(2)$ from the upper bound on N_{DF} at high SNR. Note that (33) implies

$$\frac{\log_2(\mathsf{SNR}_1)\log_2(\mathsf{SNR}_2)}{\log_2(\mathsf{SNR}_1) + \log_2(\mathsf{SNR}_2)} > \log_2(\mathsf{SNR}_0), \tag{34}$$

which implies that the harmonic mean of $\log_2(\mathsf{SNR}_1)$ and $\log_2(\mathsf{SNR}_2)$ is larger than $2\log_2(\mathsf{SNR}_0)$, i.e., $M_h(\log_2(\mathsf{SNR}_1), \log_2(\mathsf{SNR}_2)) > 2\log_2(\mathsf{SNR}_0)$. Equivalently,

$$M_h(\mathsf{SNR}_1|_{\mathrm{dB}}, \mathsf{SNR}_2|_{\mathrm{dB}}) > 2(\mathsf{SNR}_0|_{\mathrm{dB}})$$

where $x|_{\mathrm{dB}}$ is the dB value of x. We emphasize here that this is a sufficient condition under which relaying using DF reduces latency. More sophisticated DF schemes might be able to reduce latency under a more relaxed condition.

Condition (34) is interesting especially in light of the long-term achievable rate of DF. Namely, for $B \to \infty$, DF can achieve a rate which can be approximated at high SNR by

$$R_{DF} = \min\left\{\frac{1}{2}\log_2(\mathsf{SNR}_1), \frac{1}{2}\log_2(\mathsf{SNR}_2)\right\}, \tag{35}$$

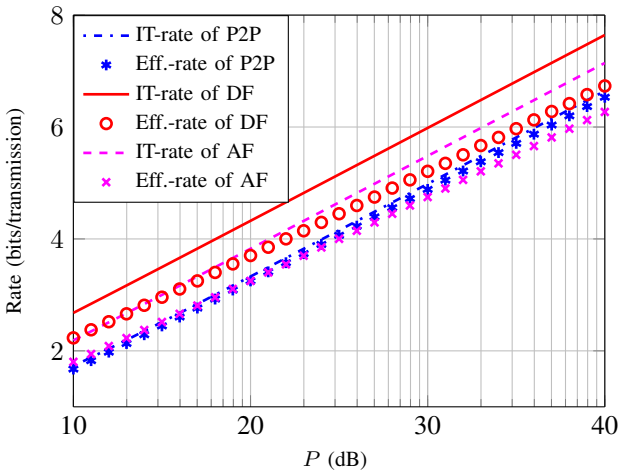


Fig. 7. Information-theoretic rate (with $P_e \to 0$) and effective rate (with $P_e < \epsilon$) for a RC with $h_0 = 1$, $h_1 = h_2 = 2$, and $P = P_r$. A payload of $B = 10$kbit is considered with an error probability requirement of $\epsilon = 10^{-3}$.

bits per transmission. On the other hand, P2P achieves a rate $\frac{1}{2}\log_2(\mathsf{SNR}_0)$ bits per transmission at high SNR. Thus, DF achieves higher rates than P2P if $\min\{\log_2(\mathsf{SNR}_1), \log_2(\mathsf{SNR}_2)\} > \log_2(\mathsf{SNR}_0)$, i.e.,

$$\min\{\mathsf{SNR}_1|_{\text{dB}}, \mathsf{SNR}_2|_{\text{dB}}\} > \mathsf{SNR}_0|_{\text{dB}}. \tag{36}$$

This condition indicates that the relative performance of DF with respect to P2P is determined by the bottleneck among SNR_1 and SNR_2 for infinite transmission ($B \to \infty$), while it is determined by both SNR's for finite transmission $B < \infty$ as can be seen from (34).

By comparing conditions (33) and (36), we notice that DF provides lower latency only if it achieves higher rates[6] than P2P, but the converse is not true. If DF achieves higher rates than P2P then it does not necessarily provide lower latency. For instance, if $\mathsf{SNR}_1 = 36$dB, $\mathsf{SNR}_2 = 33$dB, and $\mathsf{SNR}_0 = 30$dB, then DF achieves higher rate than P2P (on the long-term), but yields higher latency (for finite B). This is due to the discrepancy between the information-theoretic rate of a scheme, and its effective rate (B/N) for finite B under a reliability requirement. Figure 7 shows that information-theoretic rates (IT-rate) of the P2P, DF, and AF schemes as a function of P for a RC with $h_0 = 1$, $h_1 = h_2 = 2$, and $P = P_r$. In the same plot, we show the effective rate (Eff.-rate) for transmitting $B = 10$kbit of information using the three schemes with a reliability requirement $\epsilon = 10^{-3}$. In this example, it can be seen that while DF and AF perform better than P2P in the IT-sense, they loose their advantage from an effective rate point of view as P increases.

A similar comparison between AF and P2P leads to the statement of Proposition 2. Namely, at high SNR, if

$$\frac{2}{\log_2(\mathsf{SNR}_{AF})} < \frac{1}{\log_2(\mathsf{SNR}_0)}, \tag{37}$$

[6]in the information-theoretic sense, i.e., with $P_e \to 0$ as $N \to \infty$

then AF has a lower latency than P2P. Note that at high SNR, we can write (cf. (14))

$$\mathsf{SNR}_{AF} \approx \frac{\mathsf{SNR}_1\mathsf{SNR}_2}{\mathsf{SNR}_1 + \mathsf{SNR}_2} = \frac{M_h(\mathsf{SNR}_1, \mathsf{SNR}_2)}{2}.$$

Therefore, condition (37) can be written as

$$\left.\frac{M_h(\mathsf{SNR}_1, \mathsf{SNR}_2)}{2}\right|_{\text{dB}} > 2(\mathsf{SNR}_0|_{\text{dB}}), \tag{38}$$

as given in Proposition 2. Similar to the discussion on DF above, at high SNR, AF achieves higher long-term rate than P2P if $\frac{1}{\log_2(\mathsf{SNR}_{AF})} < \frac{1}{\log_2(\mathsf{SNR}_0)}$ but achieves lower latency only if condition (37) holds, which is stricter.

It can be easily shown that the upper bound on N_{AF} is always larger than that of N_{DF} at high SNR. Nevertheless, in cases where both AF and DF achieve lower latency than P2P, AF remains a good candidate if low processing complexity at the relay is required.

VII. Conclusion

We have derived the transmission latency of multi-hopping DF and AF in a relay channel and compared it with the latency of the P2P channel scheme which ignores the relay. We have also derived sufficient conditions on the SNR's under which these simple schemes reduce latency. The obtained conditions are more stringent than the conditions under which these schemes outperform the P2P scheme in terms of information-theoretic achievable rates. The obtained conditions can be used in practice by a network controller, e.g., to decide whether to involve a relay node in a given transmission or not. They can also be used as a relay selection criterion, where among a set of relays, the relay leading to the lowest latency is incorporated in the transmission.

As an extension for this work, it would be interesting to examine the impact of relays on the latency in a fading channel. In a block fading channel, a transmission has be optimized for a given block before the channel state changes. For such a scenario, the number of bits to be delivered per block has to be chosen such that the reliability requirement is met in the given block under the given channel state. Thus, it is interesting to study a dynamic payload allocation which minimizes the delay over such a channel. This work can also be extended towards studying the latency of multi-way communications [31], [32], or the latency under different reliability metrics such as guaranteed MSE [33].

References

[1] A. Chaaban and A. Sezgin, "When Can a Relay Reduce End-to-End Communication Delay?" in *Proc. of the ICCSPA 2015*, Sharjah, UAE, Feb. 2015.

[2] G. Cocco, D. Gündüz, and C. Ibars, "Streaming transmission over block fading channels with delay constraint," *IEEE Trans. on Wireless Communications*, vol. 12, no. 9, pp. 4315 – 4327, Sept. 2013.

[3] E. A. Jorswieck and H. Boche, "Delay-limited capacity: multiple antennas, moment constraints, and fading statistics," *IEEE Trans. on Wireless Communications*, vol. 6, no. 12, pp. 4204 – 4208, Dec. 2007.

[4] E. A. Jorswieck, H. Boche, and A. Sezgin, "Delay-limited capacity and maximum throughput of spatially correlated multiple antenna systems under average and peak-power constraints," in *IEEE Info. Theory Workshop (ITW)*, Oct. 2004, pp. 440–445.

[5] G. Caire, R. Muller, and R. Knopp, "Multiuser diversity in delay-limited cellular wideband systems," in *Proc. of the International Zurich Seminar on Communications*, Zurich, 2006, pp. 178 – 181.
[6] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. on Info. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[7] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback–I: No bandwidth constraint," *IEEE Trans. on Info. Theory*, vol. IT-12, no. 2, pp. 172–182, Apr. 1966.
[8] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback–II: Band-limited signals," *IEEE Trans. on Info. Theory*, vol. IT-12, no. 2, pp. 183–189, Apr. 1966.
[9] Y.-H. Kim, A. Lapidoth, and T. Weissman, "On the reliability of Gaussian channels with noisy feedback," in *Proc. of the 41st Annual Allerton Conference on Communication, Control, and Computing*, Monticello , Illinois, Sep. 2006, pp. 364–371.
[10] M. V. Burnashev and H. Yamamoto, "Noisy feedback improves the Gaussian channel reliability function," in *Proc. of IEEE International Symposium on Info. Theory (ISIT)*, Honolulu, HI, USA, Jul. 2014.
[11] A. Laya, K. Wang, A. A. Widaa, J. Alonso-Zarate, J. Markendahl, and L. Alonso, "Device-to-device communications and small cells: enabling spectrum reuse for dense networks," *IEEE Wireless Communications*, vol. 21, no. 4, pp. 98 – 105, Aug. 2014.
[12] T. M. Cover and A. El-Gamal, "Capacity theorems for the relay channel," *IEEE Trans. on Info. Theory*, vol. IT-25, no. 5, pp. 572–584, Sep. 1979.
[13] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. on Info. Theory*, vol. 51, no. 9, pp. 3037–3063, Sep. 2005.
[14] A. Chaaban and A. Sezgin, "On the generalized degrees of freedom of the Gaussian interference relay channel ," *IEEE Trans. on Info. Theory*, vol. 58, no. 7, pp. 4432–4461, July 2012.
[15] Y. Tian and A. Yener, "The Gaussian interference relay channel: improved achievable rates and sum rate upper bounds using a potent relay," *IEEE Trans. on Info. Theory*, vol. 57, no. 5, pp. 2865–2879, May 2011.
[16] I. Marić, R. Dabora, and A. J. Goldsmith, "Relaying in the presence of interference: Achievable rates, interference forwarding, and outer bounds," *IEEE Trans. on Info. Theory*, vol. 58, no. 7, pp. 4342–4354, July 2012.
[17] R. G. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.
[18] H. Q. Ngo and E. G. Larsson, "Linear multihop amplify-and-forward relay channels: Error exponent and optimal number of hops," *IEEE Trans. on Wireless Communications*, vol. 10, no. 11, pp. 3834–3842, Nov. 2011.
[19] W. Zhang and U. Mitra, "Multi-hopping strategies: An error-exponent comparison," in *Proc. of IEEE International Symposium on Info. Theory (ISIT)*, Nice, France, Jun 2007.
[20] H. Q. Ngo, T. Q. S. Quek, and H. Shin, "Amplify-and-forward two-way relay networks: Error exponents and resource allocation," *IEEE Trans. on Communications*, vol. 58, no. 9, pp. 2653–2666, Sept. 2010.
[21] E. Yilmaz, R. Knopp, and D. Gesbert, "Error exponents for backhaul-constrained parallel relay channels," in *Proc. of the IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2010.
[22] Q. Li and C. N. Georghiades, "On the error exponent of the wideband relay channel," in *Proc. of the European Signal Processing Conference (EUSIPCO)*, Florence, Italy, Sept. 2006.
[23] N. Wen and R. Berry, "Reliability constrained packet-sizing for linear multi-hop wireless networks," in *Proc. of IEEE International Symposium on Info. Theory (ISIT)*, Toronto, Canada, July 2008.
[24] V. Y. F. Tan, "On the reliability function of the discrete memoryless relay channel," *IEEE Trans. on Info. Theory*, vol. 60, no. 4, pp. 1550–1573, Feb. 2015.
[25] M. Agarwal, D. Guo, and M. L. Honig, "Error exponent for Gaussian channels with partial sequential feedback," *IEEE Trans. on Info. Theory*, vol. 59, no. 8, pp. 4757–4766, Aug. 2013.
[26] I. Marić, "Low latency communications," in *Presented at the Information Theory and Applications Workshop (ITA 2013), arXiv:1302.5662*, San Diego, CA, USA, Feb. 2013.
[27] A. S. Avestimehr, S. N. Diggavi, and D. N. C. Tse, "Wireless network information flow: A deterministic approach," *IEEE Trans. on Info. Theory*, vol. 57, no. 4, pp. 1872–1905, Apr. 2011.
[28] Y. Xiang and Y.-H. Kim, "On the AWGN channel with noisy feedback and peak energy constraint ," in *Proc. of IEEE International Symposium on Info. Theory (ISIT)*, Austin, TX, June 2010, pp. 256–259.
[29] T. Cover and J. Thomas, *Elements of information theory (Second Edition)*. John Wiley and Sons, Inc., 2006.
[30] O. Sahin and E. Erkip, "Achievable rates for the Gaussian interference relay channel," in *Proc. of 2007 GLOBECOM Communication Theory Symposium*, Washington D.C., Nov. 2007.
[31] A. Chaaban and A. Sezgin, "Multi-way communications: An information theoretic perspective," *Foundations and Trends in Communications and Information Theory*, vol. 12, no. 3-4, pp. 185–371, 2015.
[32] ——, "The approximate capacity region of the Gaussian Y-Channel via the deterministic approach," *IEEE Trans. on Info. Theory*, vol. 61, no. 2, pp. 939–962, Feb. 2015.
[33] E. Jorswieck, B. Ottersten, A. Sezgin, and A. Paulraj, "Guaranteed performance region in fading orthogonal space-time coded broadcast channels," *EURASIP Journal on Wireless Communications and Networking*, no. 1, 2008.

Anas Chaaban (S'09 - M'14) received his Maîtrise ès Sciences degree in electronics from the Lebanese University, Lebanon, in 2006. He received his M.Sc. degree in communications technology and his Dr.-Ing. (Ph.D.) degree in Electrical Engineering and Information Technology from the University of Ulm and the Ruhr-University of Bochum, Germany, in 2009 and 2013, respectively.

During 2008-2009, he was with the Daimler AG research group on machine vision, Ulm, Germany. He was a Research Assistant with the Emmy-Noether Research Group on Wireless Networks at the University of Ulm, Germany, during 2009-2011, which relocated to Ruhr-Universität Bochum, Germany, in 2011. He was a postdoctoral researcher at the Ruhr-Universität Bochum, Germany, in 2013-2014, and joined King Abdullah University of Science and Technology as a postdoctoral researcher in 2015. He received the best poster award at the IEEE Comm. Theory Workshop in 2011, and the best paper award at ICCSPA in 2015. His research interests are in the areas of information theory and wireless communications.

Aydin Sezgin (S'01 - M'05 - SM'13) received the Dipl.-Ing. (M.S.) degree in communications engineering and the Dr.-Ing. (Ph.D.) degree in electrical engineering from the TFH Berlin in 2000 and the TU Berlin, in 2005, respectively.

From 2001 to 2006, he was with the Heinrich-Hertz-Institut (HHI), Berlin. From 2006 to 2008, he was a Post-doc and Lecturer at the Information Systems Laboratory, Department of Electrical Engineering, Stanford University. From 2008 to 2009, he was a Post-doc at the Department of Electrical Engineering and Computer Science at the University of California Irvine. From 2009 to 2011, he was the Head of the Emmy-Noether-Research Group on Wireless Networks at the Ulm University. In 2011, he was full professor at TU Darmstadt, Germany. He is currently a full professor at the Department of Electrical Engineering and Information Technology at Ruhr-University Bochum, Germany.

He served as Associate Editor for IEEE Transactions on Wireless Communications 2009-2014. Sezgin is the winner of of the ITG-sponsorship award in 2006. He is the first recipient of the prestigious Emmy-Noether grant by the German Research Foundation (DFG) in communication engineering in 2009. He has co-authored a paper that received the best poster award at the IEEE Comm. Theory Workshop in 2011. He has also co-authored a paper that received the best paper award at ICCSPA in 2015.