Blocklength-Limited Performance of Relaying under Quasi-Static Rayleigh Channels

Yulin Hu, *Student Member, IEEE*, Anke Schmeink, *Member, IEEE* and James Gross, *Member, IEEE*

***Abstract*—In this paper, the blocklength-limited performance of a relaying system is studied, where channels are assumed to experience quasi-static Rayleigh fading while at the same time only the average channel state information (CSI) is available at the source. Both the physical-layer performance (blocklength-limited throughput) and the link-layer performance (effective capacity) of the relaying system are investigated. We propose a simple system operation by introducing a factor based on which we weight the average CSI and let the source determine the coding rate accordingly. In particular, we show that both the blocklength-limited throughput and the effective capacity are quasi-concave in the weight factor. Through numerical analysis, we investigate the relaying performance with average CSI while considering perfect CSI scenario and direct transmission as contrasts. We observe that relaying is more efficient than direct transmission in the finite blocklength regime. Moreover, this performance advantage of relaying under the average CSI scenario is more significant than under the perfect CSI scenario. Finally, the speed of convergence (between the blocklength-limited performance and the performance with infinite blocklengths) in relaying system is faster in comparison to the direct transmission under both the average CSI scenario and the perfect CSI scenario.**

***Index Terms*—Finite blocklength, decode-and-forward, relaying, throughput, effective capacity, average CSI.**

I. Introduction

In wireless communications, relaying [1]–[3] is well known as an efficient way to mitigate wireless fading by exploiting spatial diversity. Specifically, two-hop decode-and-forward (DF) relaying protocols significantly improve the capacity and quality of service [4]–[8]. However, typically these studies were based on the ideal assumption of communicating arbitrarily reliably at Shannon's channel capacity, i.e., coding is assumed to be performed over an infinite blocklength.

In the finite blocklength regime, especially when the blocklength is short, the error probability (due to noise) becomes significant. To address this issue, an accurate approximation of achievable coding rate was identified in [9] for a single-hop transmission system while taking the error probability into account. Subsequently, these initial works regarding AWGN channels were extended to Gilbert-Elliott Channels [10], quasi-static fading channels [11], [12], quasi-static fading channels with retransmissions [13], [14] as well as spectrum sharing networks [15]. However, all these works focused on single-hop non-relaying systems, leaving the analysis of relaying in the finite blocklength regime an open problem.

In a two-hop relaying network, relaying exploits spatial diversity but at the same time halves the blocklength of the transmission (if equal time division is considered). Considering the results of [9] (ie. the loss due to finite blocklength increases as the blocklength decreases), studying the relaying performance with finite blocklengths is interesting, as it points to a trade-off between diversity gain vs. finite blocklength loss. In our recent work [16], we addressed in general analytical performance models for relaying with finite blocklengths. We investigate the blocklength-limited throughput (BL-throughput) of relaying, where the BL-throughput is defined by the average number of correctly decoded bits at the destination per channel use. We observe by simulations in [16] that the performance loss (due to finite blocklength) of relaying is much smaller than expected, while the performance loss of direct transmission is larger. In other words, the diversity gain introduced by relaying is more significant than the loss due to halving the blocklength. This surprising observation shows the performance advantage of relaying in the finite blocklength regime (in comparison to direct transmission). We further show the reason of this performance advantage in [17] that relaying has a higher SNR at each hop (in comparison to direct transmission) which makes it set the coding rate more aggressively.

Our previous work [16] [17] assumed static channels and perfect channel state information (CSI) at the source. In this paper, we generalize our previous work [16] [17] to a scenario with quasi-static Rayleigh channels while only the average CSI is available at the source. Different from the perfect CSI scenario, based on the average CSI the source is not able to determine an appropriate coding rate to fit the instantaneous channel. Thus, the analysis and improvement of the blocklength-limited performance of such a relaying system (under quasi-static fading channels but only with average CSI) becomes interesting and also challenging. To the best of our knowledge, these issues have not been studied in detail so far.

Under the described relaying system (with quasi-static fading channels and average CSI), we investigate both the physical-layer performance, i.e., BL-throughput, and the link-layer performance, i.e., effective capacity[1]. We introduce a factor based on which we weight the average CSI and let the source determine the coding rate accordingly. In particular, we prove that both the BL-throughput and the effective capacity are quasi-concave in the weight factor. By numerical investigations, we show that the BL-throughput of relaying is slightly increasing in the blocklength while the effective capacity is significantly decreasing in the blocklength. Moreover, we study the performance advantage of relaying in the finite blocklength regime: Under the condition of having similar Shannon capacity performance, relaying outperforms

[1]The effective capacity is an accepted performance model that accounts for transmission and queuing effects [18]. It characterizes the (maximum) constant arrival rate of a flow to a queuing system and relates the stochastic characterization of the service of the queuing system to a queue-length or delay constraints of the flow. It has been widely applied to the analysis of wireless systems [12], [19], [20].

direct transmission in the finite blocklength regime. More importantly, this performance advantage under the average CSI scenario is larger than under the perfect CSI scenario. Finally, we find that the performance loss due to finite blocklength (the gap between BL-throughput and Shannon/outage capacity) is negligible under the average CSI scenario in comparison to the perfect CSI scenario.

The paper is organized as follows. Section II describes the system model and briefly recaps basics of the finite blocklength regime. In Section III, we consider the physical-layer performance and derive the BL-throughput of the relaying system. Subsequently, in Section IV the link-layer performance of the relaying system is studied, where the distribution of the service process increment and the maximum sustainable date rate are investigated. Section V presents a numerical study. Finally, we conclude our work in Section VI.

II. Preliminaries

A. System Model

We consider a simple relaying scenario with a source, a destination and a decode-and-forward (DF) relay as schematically shown in Figure 1. The links between the above transceivers

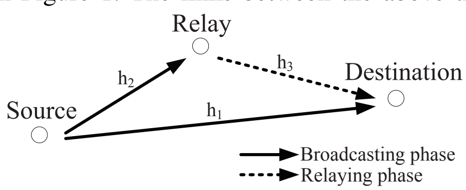

Fig. 1. Illustration of the system scenario.

are referred to as the direct link (from the source to the destination), the backhaul link (from the source to the relay) and the relaying link (from the relay to the destination). In general, we assume the direct link to be much weaker than the backhaul link as well as the relaying link. The entire system operates in a slotted fashion where time is divided into transmission periods of length $2m$ (symbols). Each transmission period contains two frames (each frame with length m), which corresponds to the two hops of relaying. The blocklength of the coding over the channel in each frame is as long as the frame length m.

During a transmission period i, first a broadcasting frame is employed, followed by a relaying frame. During the broadcasting frame, the source transmits data to the relay and the destination. The received signals at the destination and the relay in the broadcasting frame of transmission period i are given by $\mathbf{y}_{1,i} = p_{\text{tx}} h_{1,i} \mathbf{x}_i + \mathbf{n}_{1,i}$ and $\mathbf{y}_{2,i} = p_{\text{tx}} h_{2,i} \mathbf{x}_i + \mathbf{n}_{2,i}$. Next, if the data is decoded correctly and forwarded by the relay, the received signal at the destination in the relaying frame of transmission period i is given by $\mathbf{y}_{3,i} = p_{\text{tx}} h_{3,i} \mathbf{x}_i + \mathbf{n}_{3,i}$. The transmitted signal $\mathbf{x}$ and received signals $\mathbf{y}_{1,i}$, $\mathbf{y}_{2,i}$ and $\mathbf{y}_{3,i}$ are complex m-dimensional vectors. Besides, the transmit power at either the relay or the source is denoted by p_{tx}. In addition, the noise vectors of these links in transmission period i are denoted by $\mathbf{n}_{1,i}$, $\mathbf{n}_{2,i}$ and $\mathbf{n}_{3,i}$, which are independent and identically distributed (i.i.d.) complex Gaussian vectors: $\mathbf{n} \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_m\right)$, $\mathbf{n} \in \{\mathbf{n}_{1,i}, \mathbf{n}_{2,i}, \mathbf{n}_{3,i}\}$, where $\mathbf{I}_m$ denotes an $m \times m$ identity matrix. Moreover, $h_{1,i}$, $h_{2,i}$ and $h_{3,i}$ are the channels (scalars) of the direct link, backhaul link and relaying link during transmission period i, respectively.

In this work, we consider quasi-static Rayleigh fading channels where the channels remain constant within each transmission period (i.e., during the the two consecutive frames/hops) and vary independently from one period to the next. Hence, the instantaneous channel gain of each link has two components, i.e., the average channel gain and the random fading. We denote average channel gains of these three links over the transmission periods by $|\bar{h}_1|^2$, $|\bar{h}_2|^2$ and $|\bar{h}_3|^2$. We assume these channels to experience Rayleigh fading where the envelope $\sqrt{z}$ of the channel fading (response) is Rayleigh distributed with the probability density function (PDF) $f\left(\sqrt{z}\right) = 2\frac{\sqrt{z}}{\sigma^2} e^{-z/\lambda^2}$. Typically, the channel fading is modeled as having unit power gain. This requires that $\lambda^2 = 1$. Hence, the PDF of z, the gain due to Rayleigh fading, is given by the exponential distribution: $f(z) = e^{-z}$. We denote by $z_{k,i}$ ($k = 1, 2, 3$) the gains due to Rayleigh fading (during transmission period i) of the direct link, the backhaul link and the relaying link. Hence, we have $|h_{k,i}|^2 = z_{k,i} |\bar{h}_k|^2, k = 1, 2, 3$. Moreover, these channel fading gains at different links during the same transmission period are assumed to be independent and identically distributed (i.i.d).

Finally, the destination is assumed to apply maximal ratio combining (MRC) where the combined channel gain is given by $|h_{1,i}|^2 + |h_{3,i}|^2$. Thus, the received signal-to-noise ratio (SNR) at the relay and the received SNR at the destination in transmission period i under maximum ratio combining are given by $\gamma_{2,i} = \frac{|h_{2,i}|^2 p_{\text{tx}}}{\sigma^2}$ and $\gamma_{\text{MRC},i} = \frac{\left(|h_{1,i}|^2 + |h_{3,i}|^2\right) p_{\text{tx}}}{\sigma^2}$.

B. Blocklength-Limited Performance of Single-Hop Transmission Scenario with Perfect CSI (at the source)

For the real additive white Gaussian noise (AWGN) channel, [8, Theorem 54] derives an accurate approximation of the coding rate of a single-hop transmission system. With blocklength m, block error probability ε and SNR γ, the coding rate (in bits per channel use) is given by: $r \approx \frac{1}{2}\log_2\left(1+\gamma\right) - \sqrt{\frac{V_{\text{real}}}{m}} Q^{-1}\left(\varepsilon\right)$, where $Q^{-1}(\cdot)$ is the inverse Q-function while the Q-function is given by $Q\left(w\right) = \int_w^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$. In addition, V_{real} is the channel dispersion of a real Gaussian channel given by $V_{\text{real}} = \frac{\gamma}{2} \frac{\gamma+2}{(1+\gamma)^2} (\log_2 e)^2$.

In each transmission period a quasi-static fading channel with fading coefficient h can be viewed as an AWGN channel with channel gain $|h|^2$. Therefore, the above result with a real AWGN channel can be reasonably extended to a complex quasi-static fading channel model [11]–[15]: With a channel gain $|h|^2$ the coding rate (in bits per channel use) is given by:

$$r = \mathcal{R}(|h|^2, \varepsilon, m) \approx \mathcal{C}(|h|^2) - \sqrt{V_{\text{comp}}/m} \cdot Q^{-1}\left(\varepsilon\right), \quad (1)$$

where $\mathcal{C}\left(|h|^2\right)$ is the Shannon capacity function of a complex channel with gain $|h|^2$: $\mathcal{C}(|h|^2) = \log_2(1 + \frac{|h|^2 p_{\text{tx}}}{\sigma^2})$. In addition, the channel dispersion of a complex Gaussian channel is twice the one of a real Gaussian channel: $V_{\text{comp}} = 2V_{\text{real}} = (1 - \frac{1}{(1+\gamma)^2})(\log_2 e)^2 = (1 - 2^{-2\mathcal{C}(|h|^2)})(\log_2 e)^2$.

Then, for a single hop transmission of a transmission period of quasi-static fading channel, with blocklength m and coding rate r, the (decoding) error probability at the receiver is:

$$\varepsilon = \mathbb{P}(|h|^2, r, m) \approx Q\left(\frac{\mathcal{C}(|h|^2) - r}{\sqrt{V_{\text{comp}}/m}}\right). \quad (2)$$

Considering the channel fading, the expected error probability over channel fading is given by [11]:

$$\mathop{\mathbb{E}}_{z}[\varepsilon] = \mathop{\mathbb{E}}_{z}\left[\mathbb{P}(|h|^2, r, m)\right] \approx \mathop{\mathbb{E}}_{z}\left[Q\left(\frac{\mathcal{C}(|h|^2) - r}{\sqrt{V_{\text{comp}}/m}}\right)\right], \quad (3)$$

where $\mathop{\mathbb{E}}_{z}[*]$ is the expectation over the distribution of z.

In the remainder of the paper, we investigate the performance of relaying under quasi-static fading channels by applying the above approximations. As these approximations have been shown to be tight for sufficiently large values of m [9]–[11], for simplicity we will assume them to be equal in our analysis and numerical evaluation where we consider sufficiently large value of m at each hop of relaying.

III. Physical-layer Performance of Relaying with Average CSI

With only average CSI, if the source determines the coding rate directly based on it, this likely results in that $|h_{k,i}|^2 < |\bar{h}_k|^2$, i.e., the instantaneous channel gain at transmission period i is lower than the average. As a result, the error probability increases significantly. Therefore, we propose the source to choose a relatively lower coding rate which is obtained by the weighted average channel gains $\eta|\bar{h}_k|^2$ ($k = 1, 2, 3$). In addition, we assume that $0 < \eta \leq \hat{z}$, where $\hat{z}$ is the median of z. It follows that for any link j with probability 0.5 value η is lower than the channel fading gain z_k, i.e., $\Pr\{\eta < z_k\} \leq 0.5$. Hence, although the instantaneous channel gain $z_k|\bar{h}_k|^2$ is still possible to be lower than the weighted one $\eta|\bar{h}_k|^2$, the probability of this becomes much lower when $\eta < \hat{z}$ and is bounded by 0.5, i.e., $\Pr\{z_k|\bar{h}_k|^2 < \eta|\bar{h}_k|^2\} = \Pr\{\eta < z_k\} \leq 0.5$.[2]

Recall that we assume MRC to be applied at the destination. Hence, the coding rates at different hops of relaying are required to be the same. This coding rate r is determined by the source based on the weighted average CSI according to (1). As the overall performance of the considered two-hop relaying system is actually mainly subject to the bottleneck link which is either the backhaul link or the combined link, the coding rate is determined by $r = \mathcal{R}(\eta \cdot \min\{|\bar{h}_2|^2, |\bar{h}_1|^2 + |\bar{h}_3|^2\}, \varepsilon^\circ, m)$, where ε° is a constant error probability and has a value of practical interest. According to (1), r is strictly increasing in the weight factor η. In other words, a large η means a high expectation on the channel quality and results in a high coding rate.

Once the coding rate r is determined, it will not be changed during transmission periods. In other words, from one transmission period to the next the coding rate is fixed and therefore error probabilities of different links of relaying vary along with the channel fading. Regarding the overall error of relaying, in this work we treat the decoding error at the destination based on the combined channel gain as the overall error. Although it is theoretically possible that an error occurs in the two-hop relaying though the direct transmission (during the broadcasting phase) is correct, the probability of this is negligible as on the one hand we mainly consider error probabilities of practical interest (i.e., the overall error probability of relaying is not significant) and on the other hand we assume the direct link to be much weaker than the backhaul link as well as the relaying link. Therefore, the overall error probability of relaying during transmission period i is:

$$\varepsilon_{\mathcal{R},i} = \varepsilon_{2,i} + (1 - \varepsilon_{2,i}) \cdot \varepsilon_{\text{MRC},i}, \quad (4)$$

where $\varepsilon_{\text{MRC},i} = \mathbb{P}(|h_{1,i}|^2 + |h_{3,i}|^2, r, m)$, $\varepsilon_{2,i} = \mathbb{P}(|h_{2,i}|^2, r, m)$.

Under the studied two-hop relaying scenario where the coding rate at each hop is r, the (source-to-destination) equivalent coding rate is actually $r/2$. Therefore, the expected BL-throughput of relaying during transmission period i (the number of correctly received bits at the destination per channel use) is $C_{\text{BL},i} = r(1 - \varepsilon_{\mathcal{R},i})/2$. Then, we have the (average) BL-throughput of relaying over time, which is actually the expected valueof $C_{\text{BL},i}$ over the transmission periods:

$$C_{\text{BL}} = \mathop{\mathbb{E}}_{i}[C_{\text{BL},i}] = r(1 - \mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}])/2. \quad (5)$$

Hence, the major challenge for determining C_{BL} is to obtain the expectation of the overall error probability over the transmission periods, which is actually equal to the expectation over channel fading. Recall that all the channels are independent from each other, hence the expected value of the overall error probability of relaying is further given by:

$$\mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}] = \mathop{\mathbb{E}}_{z_1,z_2,z_3}[\varepsilon_{\mathcal{R}}] = \mathop{\mathbb{E}}_{z_2}[\varepsilon_2] + (1 - \mathop{\mathbb{E}}_{z_2}[\varepsilon_2]) \mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\text{MRC}}], \quad (6)$$

where $\mathop{\mathbb{E}}_{z_2}[\varepsilon_2]$ and $\mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\text{MRC}}]$ are the expectation values (over fading) of error probabilities of the backhaul link and the combined link. Based on (3), they are given by:

$$\begin{aligned}\mathop{\mathbb{E}}_{z_2}[\varepsilon_2] &= \int_0^\infty e^{-z_2} \varepsilon_2 dz_2 = \int_0^\infty \mathbb{P}(z_2|\bar{h}_2|^2, r, m) e^{-z_2} dz_2 \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \int_{\sqrt{m}\alpha(z_2)}^\infty e^{-\frac{t^2 + 2z_2}{2}} dt dz_2,\end{aligned} \quad (7)$$

$$\begin{aligned}\mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\text{MRC}}] &= \int_0^\infty \int_0^\infty \varepsilon_{\text{MRC}}\, e^{-z_1 - z_3} dz_1 dz_3 \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \int_0^\infty \mathbb{P}(z_1|\bar{h}_1|^2 + z_3|\bar{h}_3|^2, r, m) e^{-z_1 - z_3} dz_1 dz_3 \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \int_0^\infty \int_{\sqrt{m}\alpha(z_1,z_3)}^\infty e^{-\frac{t^2 + 2z_1 + 2z_3}{2}} dt dz_1 dz_3,\end{aligned} \quad (8)$$

where $\alpha(z_2) = \frac{\mathcal{C}(z_2|\bar{h}_2|^2) - r}{\sqrt{\frac{1}{m}\left(1 - 2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)}\log_2 e}$ and $\alpha(z_1, z_3) = \frac{\mathcal{C}(z_1|\bar{h}_1|^2 + z_3|\bar{h}_3|^2) - r}{\sqrt{\frac{1}{m}\left(1 - 2^{-2\mathcal{C}(z_1|\bar{h}_1|^2 + z_3|\bar{h}_3|^2)}\right)}\log_2 e}$.

So far, we derived the BL-throughput of relaying under the studied system. We then have the following theorem regarding the BL-throughput.

Proposition 1. *Under a relaying scenario with quasi-static*

[2]The setup (letting $\hat{z}$ be the upper limit of η) facilitates the proof of Theorem 1 and Theorem 2. In addition, in practice it is unreasonable to choose a coding rate which likely exceeds the Shannon limit with a probability higher than 0.5. We will also show in the simulation that this setup is not impacting the system operation/optimization as the optimal value of η is about 0.2 which is much lower than $\hat{z} \approx 0.7$ (Rayleigh channels).

Rayleigh channels where only the average CSI is available at the source, the BL-throughput is concave in the coding rate.

Proof: See Appendix A. ■

Recall that the coding rate chosen by the source is strictly increasing in the weight factor η. Combined with Theorem 1, we have an important corollary of Theorem 1:

Corollary 1. *Consider a relaying scenario with quasi-static Rayleigh channels and only the average CSI being available at the source. If the source determines the coding rate according to the weighted average CSI, the BL-throughput is quasi-concave in the weight factor η.*

Proof: See Appendix B. ■

Therefore, only with the average CSI there is a unique optimal value of η, which maximizes the BL-throughput of relaying in the finite blocklength regime.

IV. Link-Layer Performance of Relaying with Average CSI

In this section, we study the link-layer performance of relaying based on the *effective capacity*, which is a well-known performance model that accounts for transmission and queuing effects in (wireless) networks [18]. We first briefly review the effective capacity model and extend the expression of the maximum sustainable data rate into a blocklength-limited relaying scenario. Subsequently, we derive the maximum sustainable data rate (MSDR) of the studied relaying system in the finite blocklength regime based on the model.

A. Maximum Sustainable Data Rate

The effective capacity characterizes the (maximum) arrival rate of a flow to a queuing system and relates the stochastic characterization of the service of the queuing system to the queue-length or delay constraints of the constant rate flow. In the finite blocklength regime, decoding errors may occur. If a decoding error occurs during transmission period i, the service process increment (effectively transmitted information) s_i of the two-frame relaying equals zero. On the contrary, if no error occurs during frame i, the service process increment equals $s_i = m \cdot r$, where r is the coding rate (in bits per channel use) employed over a block of m symbols in each hop/frame of relaying. During period i, the cumulative service process is $S_i = \sum_{n=0}^{i} s_n$. Assume that the queue is stable as the average service rate is larger than the average arrival rate. Hence, the random queue length Q_i at period i converges to the steady-state random queue length Q. To characterize the long-term statistics $\Pr\{Q\}$ of the queue length, the framework of the effective capacity gives us the following upper bound:

$$\Pr\{Q > x\} \leq K \cdot e^{-\theta \cdot x}, \tag{9}$$

where K is the probability that the queue is non-empty and θ is the so called QoS exponent. Based on [21], for a constant rate source with r bits per two-hop transmission period, the exponent θ has to fulfill the following constraint:

$$r < -\Lambda(-\theta)/\theta. \tag{10}$$

$\Lambda(\theta)$ is the log-moment generating function of the cumulative service process S_i defined as:

$$\Lambda(\theta) = \lim_{i\to\infty} \frac{1}{i} \log \mathbb{E}\left[e^{\theta\cdot(S_i - S_0)}\right]. \tag{11}$$

The ratio $-\Lambda(-\theta)/\theta$ is called the effective capacity. Denote by D_i the random queuing delay of the head-of-line bit during period i. If the constant arrival rate at the source is r, with a queue length of $Q = q$, a current delay of the head-of-line bit is given by $D = q/r$. This yields the following bound for the steady-state delay distribution based on Equation (9):

$$\Pr\{D > d\} \leq K \cdot e^{-\theta \cdot r \cdot d}. \tag{12}$$

If the service process s_i can be assumed to be independent and identically distributed (i.i.d.), a convenient simplification is to obtain the log-moment generating function via the central limit theorem. Then, the effective capacity can be obtained by [22].

$$-\frac{\Lambda(-\theta)}{\theta} = -\lim_{i\to\infty} \frac{1}{i\cdot\theta} \log \mathbb{E}\left[e^{-\theta\cdot\Sigma_1^i s_i}\right] = \mathbb{E}_i[s_i] - \frac{\theta}{2} \operatorname*{Var}_i[s_i]. \tag{13}$$

Therefore, the queuing performance of the system is determined by the mean and the variance of the random increment of the service process s_i. Combining (10) and (12), the maximum arrival rate at the source R_{MS} in (bits per transmission period) that can be supported by the random service process is obtained, which has been first proposed by [23]: $\frac{\mathbb{E}_i[s_i]}{2} + \frac{1}{2}\sqrt{(\mathbb{E}_i[s_i])^2 + \frac{2\ln(P_{\mathrm{d}})}{d} \cdot \operatorname*{Var}_i[s_i]}$. In the formula, $\{d, P_{\mathrm{d}}\}$ is QoS requirement pair of the service, where in [23] d is the delay constraint with transmission period as unit and P_{d} is the constraint of delay violation probability. Now, we extend the above MSDR into the studied relaying system where the length of each transmission period is $2m$:

$$R_{\mathrm{MS}} \approx \frac{\mathbb{E}_i[s_i]}{4m} + \frac{1}{4m}\sqrt{\left(\mathbb{E}_i[s_i]\right)^2 + \frac{4m\ln(P_{\mathrm{d}})}{d} \cdot \operatorname*{Var}_i[s_i]}. \tag{14}$$

Based on (14), the major challenge for determining the MSDR R_{MS} is to obtain the mean and variance of the service process increment of the studied relaying system.

B. Mean and Variance of the Service Process Increment

Recall that the service process increment of a transmission period is either zero or rm. In other words, it is Bernoulli-distributed. Moreover, the probability of the error event of this Bernoulli distribution is actually the expected overall error probability of relaying which is given by (6). Based on the characteristic of the Bernoulli distribution, we immediately have the mean and variance of service process increments:

$$\mathbb{E}_i[s_i] = rm \cdot (1 - \mathbb{E}_{z_1,z_2,z_3}[\varepsilon_{\mathcal{R}}]), \tag{15}$$

$$\operatorname*{Var}_i[s_i] = r^2 m^2 \mathbb{E}_{z_1,z_2,z_3}[\varepsilon_{\mathcal{R}}] \cdot (1 - \mathbb{E}_{z_1,z_2,z_3}[\varepsilon_{\mathcal{R}}]). \tag{16}$$

Substituting (15) and (16) into (14), we have

$$R_{\mathrm{MS}}=\frac{r(1-\underset{z_1,z_2,z_3}{\mathbb{E}}[\varepsilon_{\mathcal{R}}])}{4}+\frac{r}{4}\sqrt{(1-\underset{z_1,z_2,z_3}{\mathbb{E}}[\varepsilon_{\mathcal{R}}])^2+\frac{4m\ln(P_{\mathrm{d}})}{d}\underset{z_1,z_2,z_3}{\mathbb{E}}[\varepsilon_{\mathcal{R}}](1-\underset{z_1,z_2,z_3}{\mathbb{E}}[\varepsilon_{\mathcal{R}}])}. \tag{17}$$

Obviously, the performance of MSDR R_{MS} is subject to the coding rate r, as $\underset{i}{\mathbb{E}}[\varepsilon_{\mathcal{R},i}]$ is also a function of r. In fact, the following theorem holds regarding the relationship between MSDR and the coding rate.

Proposition 2. *Under a relaying scenario with quasi-static Rayleigh channels where only the average CSI is available at the source, the MSDR is concave in the coding rate r.*

Proof: See Appendix D. ■

Similar to Theorem 1, Theorem 2 has a corollary:

Corollary 2. *Consider a relaying scenario with quasi-static Rayleigh channels and only the average CSI being available at the source. If the source determines the coding rate according to the weighted average CSI, the MSDR is quasi-concave in the weight factor η.*

Proof: The proof of Corollary 2 (based on Theorem 2) is similar to the proof of Corollary 1 (based on Theorem 1). ■

Hence, the link-layer performance (MSDR) also can be optimized by choosing an appropriate η.

V. Numerical Result and Discussion

In this section, we first show the appropriateness of our theoretical model. Subsequently, we evaluate the performance of the studied relaying system (with average CSI) in comparison to direct transmission and relaying with perfect CSI.

For the numerical results, we consider the following parameterization of the system model: First of all, we consider the cases with blocklength $m \geq 100$ (at each hop of relaying), for which the approximation is tight enough[3]. In addition, the codeword length of each link is set to be the same as the blocklength. Secondly, we consider an outdoor urban scenario and the distances of the backhaul, relaying and direct links are set to 200 m, 200 m and 360 m. Thirdly, we set the transmit power p_{tx} equal to 30 dBm and noise power to -90 dBm, respectively. In addition, we utilize the well-known COST [24] model for calculating the path-loss while the center frequency is set to equal 2 GHz. Regarding the channel, we only consider quasi-static channel model (i.e. a block-fading model with independent realizations from slot to slot) in the simulation. Hence, the numerical results for all validations and evaluations are based on the average/ergodic performance over the random channel fading. Moreover, as we consider the link-layer performance, in the simulation the QoS constraints $\{d, P_{\mathrm{d}}\}$ are set to $\{10^4$ symbols, $10^{-2}\}$. Finally, to observe the relaying performance we mainly vary the following parameters in the simulation: Blocklength and weight factor η.

[3]The choice of $m \geq 100$ as the minimum possible length is motivated by [10, Fig. 2] where the relative difference of the approximate and the exact achievable rates is less than 2% for the cases with $m \geq 100$.

A. Appropriateness of Our Theoretical Model

In Figure 2 we show the relationship between the relaying performance and the coding rate. We plot the related (ergodic) Shannon capacity of relaying as a reference. As shown in the figure, the Shannon capacity is not influenced by the coding rate. More importantly, it is shown that the BL-throughput and the MSDR are concave in the coding rate, which matches Theorem 1 and 2. For the low coding rate region, both the BL-throughput and the MSDR increase approximately linearly as the coding rate increases. This is due to the fact that errors hardly occur with a low coding rate. Hence, the bottleneck of the performance is the coding rate. However, as the coding rate further increases, the error probability becomes more and more significant. As a result, the error probability then becomes the major limit of the system performance. Therefore, both the BL-throughput and the MSDR decrease for the higher coding rate region.

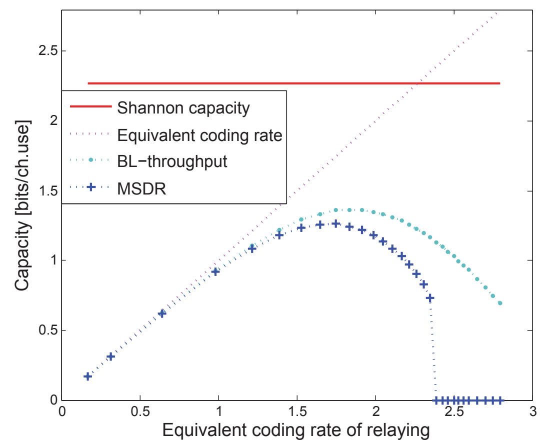

Fig. 2. Relaying performance with average CSI (at the source).

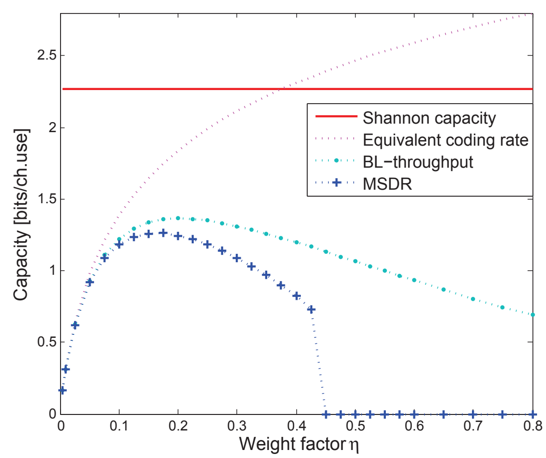

Fig. 3. The relaying performance versus the weight factor.

Figure 3 validates Corollary 1 and 2 that both the BL-throughput and the MSDR are quasi-concave in the weight factor η. Hence, the weight factor represents a trade-off between the coding rate and the error probability of the studied relaying system (only with average CSI at the source). Based on this tradeoff, both the physical-layer performance and the link-layer performance of the system can be optimized. Moreover, it can also be observed from Figure 3 that the optimal values of the weight factor for maximizing the BL-throughput and for maximizing the MSDR are not the same. In addition, we also find in the simulation (not shown here) that the optimal η for maximizing the MSDR is subject to the QoS constraints. In particular, the stricter the constraints are, the smaller the optimal η is. Hence, the parameterization of the weight factor η for a QoS-sensitive flow and a non-QoS-support system should be treated differently.

In Figure 4, we show how the weight factor introduces the tradeoff between the coding rate and the error probability and further influences the BL-throughput and the MSDR. As

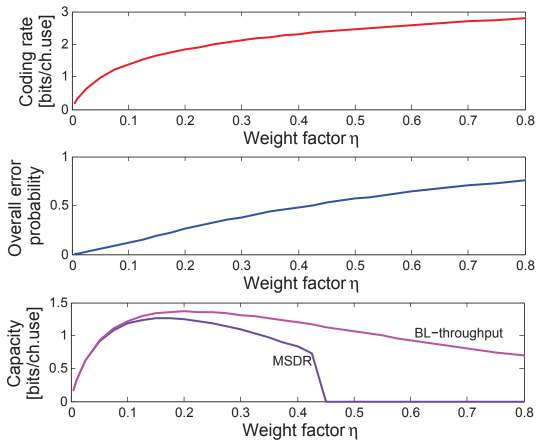

Fig. 4. The tradeoff between error probability and coding rate introduced by the weight factor, which further influences the BL-throughput and the MSDR. In the simulation, the blocklength at each hop of relaying is 500 symbols.

shown in the figure, by increasing the weight factor the coding rate increases as well as the error probability. In particular, the coding rate increases rapidly at the beginning but slowly later on. At the same time, the error probability increases approximately linearly. As a result, the BL-throughput/MSDR first increases and then deceases. In other words, they are quasi-concave in the weight factor η.

B. Numerical Investigation

In the following, we further evaluate the studied relaying system. In the evaluation, the performance under the perfect CSI scenario[4] is considered as a comparison scheme.

1) Relaying performance vs. blocklength under the average CSI scenario: We first show the relationship between the performance of the studied relaying system and the blocklength in Figure 5. The figure shows that the BL-throughput of the studied relaying system is slightly increasing in the blocklength while the MSDR is significantly decreasing in the blocklength. The explanation is as follows. On one hand, based on (1) and (2) both error probability and coding rate are influenced by the blocklength m. In particular, with a fixed coding rate at each link a long blocklength leads to a low error probability of each link. Obviously, this results in a

[4]In [16] we showed that the BL-throughput of relaying in increasing in the coding rate r (based on the perfect CSI) for AWGN channels. This also holds for maximizing the instantaneous BL-throughput at each transmission period under quasi-static channels. In this section, the BL-throughput with the perfect CSI is obtained by maximizing the instantaneous BL-throughput for each transmission period based on the perfect CSI.

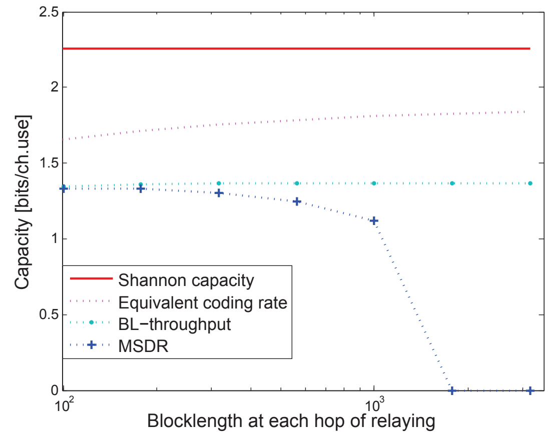

Fig. 5. The relaying performance versus blocklength.

low overall error probability of relaying and therefore a high BL-throughput. This is the reason why the BL-throughput is slightly increasing in the blocklength. On the other hand, a long blocklength means that a single transmission (i.e., two-hop relaying) costs a long time. This reduces the number of allowed retransmission attempts under a given absolute delay constraint. For example, consider a relaying system where the blocklength of each hop of relaying is 500 symbols (two-hop relaying transmission period is 10^3 symbols). To support a service with a delay constraint $d = 10^4$ symbols, the maximal number of transmission attempts (including an initial transmission and retransmissions), which not violates the delay constraint, is 10 times. This number would be 5 if the blocklength is doubled. In other words, a long blocklength reduces the flexibility of a QoS-support system due to limiting the number of transmission attempts. At the same time, the gain from increasing blocklength becomes tiny as the blocklength increases (e.g., even ragarding the physical-layer performance, the BL-throughput is approximately a constant during the long blocklength region). As a result, the MSDR reduces significantly for larger blocklengths.

2) Relaying vs. direct transmission under average CSI scenario: Observing the performance advantage of relaying: Under the average CSI scenario, we compare the performance of the studied relaying system with direct transmission. To make a fair comparison, we set the coding rate of direct transmission to be equal to the equivalent coding rate of relaying. In addition, we set the blocklength of direct transmission to be twice as large as the blocklength at each hop of relaying. More importantly, for the fair comparison in the finite blocklength regime we consider the scenario where relaying and direct transmission have a similar Shannon capacity. Under the above setup, we compare the two schemes in Figure 6 and Figure 7 where we vary the coding rate and blocklength, respectively. Based on these two figures, we observe that both the BL-throughput and the MSDR of relaying significantly outperform direct transmission, although these two schemes have a similar Shannon capacity. In other words, under the average CSI scenario relaying shows a significant advantage (in comparison to the direct transmission) under the finite blocklength regime.

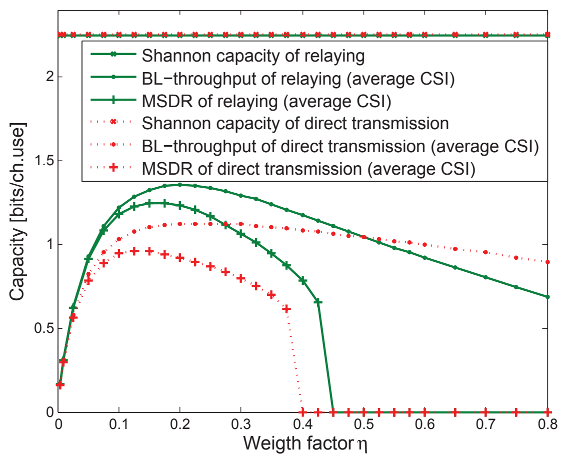

Fig. 6. The performance comparison between relaying (with average CSI) and direct transmission (with average CSI) while varying the weight factor η. In the simulation, the blocklength at each hop of relaying is 500 symbols.

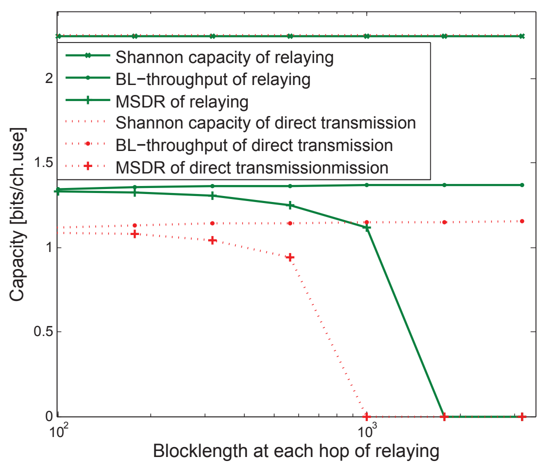

Fig. 7. The performance comparison between relaying (with average CSI) and direct transmission (with average CSI) under dynamic blocklength ($\eta = 0.2$).

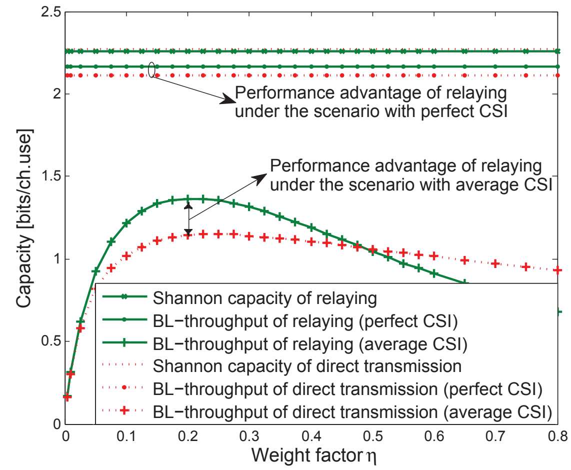

Fig. 8. The performance comparison between relaying with average CSI and relaying with perfect CSI while varying the weight factor ($\eta = 0.2$).

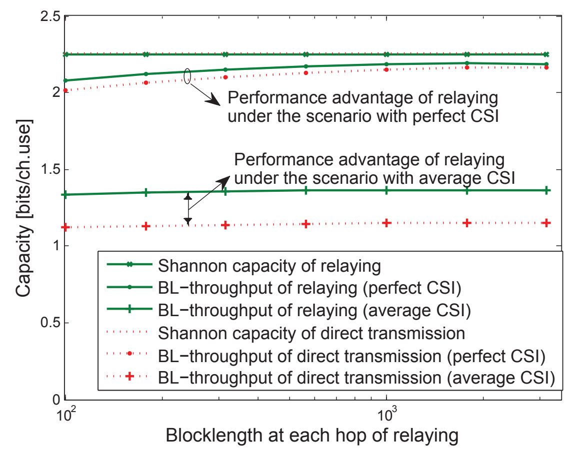

Fig. 9. The performance comparison between relaying with average CSI and relaying with perfect CSI under dynamic blocklength ($\eta = 0.2$).

3) Perfect CSI scenario vs. average CSI scenario with respect to the performance advantage of relaying: The previous subsection shows the performance advantage of relaying under the average CSI scenario. In this subsection, we compare this performance advantage for the case of average CSI with the case of perfect CSI. We show the comparison in Figure 8 and Figure 9 where the weight factor and the blocklength are varied respectively. In general, these figures show that relaying has better blocklength-limited performance than direct transmission under both the perfect CSI scenario and the average CSI scenario. However, we further observe that the performance advantage of relaying under the average CSI scenario is bigger than under the perfect CSI scenario, i.e., the performance improvement by relaying (comparing relaying with direct transmission) under the average CSI scenario is significantly higher than the one under the perfect CSI scenario. Recall that in [16] we have shown that with perfect CSI relaying is more beneficial in the finite blocklength regime in comparison to in the infinite blocklength regime (i.e., the scenario has an assumption of infinite blocklengths, under which the Shannon capacity is accureate). The observation from Figure 8 and Figure 9 further indicates that in the finite blocklength regime relaying is more beneficial for the average CSI scenario in comparison to the perfect CSI scenario.

4) Perfect CSI scenario vs. average CSI scenario with respect to the performance loss due to finite blocklength: Finally, we compare the performance loss due to finite blocklength under perfect CSI scenario and the one under average CSI scenario. Under the perfect CSI scenario, the performance loss due to finite blocklength is actually the performance gap between the Shannon capacity and the BL-throughput (with perfect CSI). On the other hand, this performance loss under the average CSI scenario is observed by comparing the BL-throughput (with average CSI) with the outage capacity. In particular, the outage capacity [25] is a performance metric in the infinite blocklength regime. It is given by $r\,(1 - \mathrm{Pr}_{\mathrm{out}}(r))$, where $\mathrm{Pr}_{\mathrm{out}}(r)$ is the outage probability. To calculate the outage capacity, the weighting CSI operation is also considered in the simulation, i.e., the packet size is chosen based on the Shannon capacity of the weighted CSI.

Under the above setup, we first show the numerical results of the comparison in Figure 10 where we fix the blocklength and vary the weight factor. The figure shows that the performance loss due to finite blocklength is considerable under the perfect CSI scenario. However, we find that under the average CSI scenario the performance loss due to finite blocklength is negligible. In the simulation, we also observe that (not

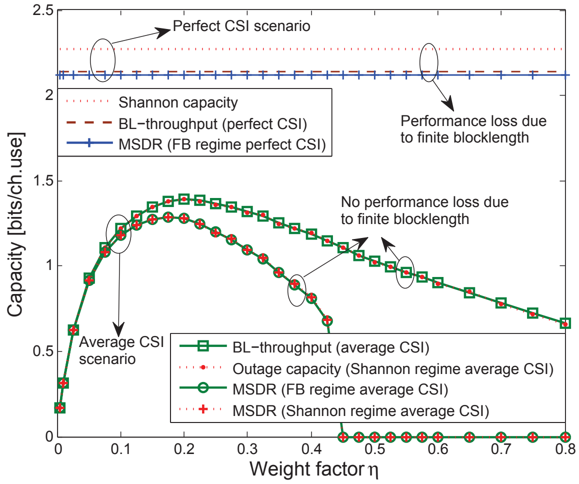

Fig. 10. Relaying with infinite blocklengths vs. relaying with finite blocklengths (blocklength at each hop of relaying is 500 symbols).

shown here) only with average CSI at the source the outage probability (in the infinite blocklength regime) and the average error probability (in the finite blocklength regime) are equal.

The observations from Figure 10 are based on the setup that the blocklength equals $m = 500$ symbols. Recall that the BL-throughput is limited by the blocklength and the performance in the infinite blocklength regime is not influenced by the blocklength. Hence, the performance loss due to finite blocklength should also be subject to the blocklength. We further investigate this performance loss in Figure 11 where we vary the blocklength. Again, we observe that the performance

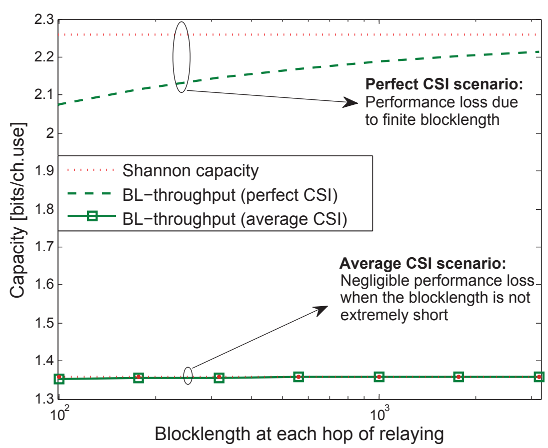

Fig. 11. The performance comparison between relaying in the infinite blocklength regime (i.e., with infinite blocklengths) and relaying in the finite blocklength regime while varying the blocklength. In the simulation, $\eta = 0.2$.

loss under the average CSI scenario is much smaller than the one under the perfect CSI scenario. In particular, with a short blocklength (e.g., blocklength $m > 100$) the performance loss due to finite blocklength is negligible under the average CSI scenario.

From Figure 11 we observe that the BL-throughput of relaying with average CSI and the outage capacity converge quickly. It should be mentioned that this is not a unique characteristic of relaying. It is studied and observed in [11] that the speed of convergence between the BL-throughput and the outage capacity for a non-relaying direct transmission is also fast. This motivates us to further compare the speed of convergence of these two transmissions schemes. Based on the comparison (not shown here), we observe that relaying has a slightly faster speed of convergence in comparison to direct transmission while in general both these two schemes converge quickly between the BL-throughput with average CSI and the outage capacity. This observation is different from the case with perfect CSI shown in Figure 9 where speeds of the convergence (between the BL-throughput and the Shannon capacity) for both relaying and direct transmission are relatively slow. More interestingly, Figure 9 shows that the speed of convergence of relaying is also faster than direct transmission under the perfect CSI scenario. Hence, we conclude that (although halving the blocklength) relaying surprisingly has a faster speed of convergence (between the blocklength-limited performance and the performance in the infinite blocklength regime) in comparison to direct transmission under both the average CSI scenario and the perfect CSI scenario. This is actually the performance advantage of relaying in the finite blocklength regime from a perspective of the speed of convergence (to the Shanon/outage capacity).

VI. Conclusion

Under the finite blocklength regime, we investigated the physical-layer performance (BL-throughput) as well as the link-layer performance (MSDR) of a relaying system with quasi-static fading channels and average CSI at the source. We proposed a simple system operation by introducing a weight factor of the average CSI to determine the coding rate. The BL-throughput and the MSDR of the studied relaying system are investigated. We first showed that both the BL-throughput and the MSDR are concave in the coding rate. More importantly, it was showed that the BL-throughput and the MSDR are quasi-concave in the weight factor. In other words, the performance of the studied system can be easily optimized based on the proposed operation.

We conclude a set of guidelines for the design of efficient relaying systems (in the finite blocklength regime) from numerical analysis. Firstly, under the average CSI scenario the BL-throughput of relaying is slightly increasing in the blocklength while the effective capacity is significantly decreasing in the blocklength. Secondly, the weight factor we proposed introduces a tradeoff between the coding rate and the error probability. Based on this tradeoff, both the physical-layer performance and the link-layer performance of the system can be optimized. Moreover, the optimal values of the weight factor for maximizing the BL-throughput and for maximizing the MSDR are different. Thirdly, under the condition of having similar Shannon capacity performance, relaying outperforms direct transmission in the finite blocklength regime. More importantly, this performance advantage of relaying under the average CSI scenario is more significant than under the perfect CSI scenario. In addition, the performance loss due to finite blocklength (i.e., the performance gap between the performance in the infinite blocklength regime and in the finite blocklength regime) is negligible under the average CSI scenario in comparison to the one under the perfect

CSI scenario. Finally, the speed of convergence (between the blocklength-limited performance and the performance in the infinite blocklength regime) in relaying system is faster in comparison to the direct transmission under both the average CSI scenario and the perfect CSI scenario.

Appendix A
Proof of the Theorem 1

Proof: Based on Equation (5) we immediately have

$$\frac{\partial C_{\mathrm{BL}}}{\partial r} = \frac{1}{2}(1-\mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}]) - \frac{r}{2}\frac{\partial \mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial r}, \tag{18}$$

$$\frac{\partial^2 C_{\mathrm{BL}}}{\partial^2 r} = -\frac{\partial \mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial r} - \frac{r}{2}\frac{\partial^2 \mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r}. \tag{19}$$

In the following, we prove Theorem 1 by showing $\frac{\partial^2 C_{\mathrm{BL}}}{\partial^2 r}<0$. According to Equation (6), we have:

$$\begin{aligned}\frac{\partial \mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial r} &= \frac{\partial \mathop{\mathbb{E}}_{z_2}[\varepsilon_2]}{\partial r}\left(1-\mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]\right) \\ &+ \frac{\partial \mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]}{\partial r}\left(1-\mathop{\mathbb{E}}_{z_2}[\varepsilon_2]\right) \geqslant 0,\end{aligned} \tag{20}$$

$$\begin{aligned}\frac{\partial^2 \mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r} &= \frac{\partial^2 \mathop{\mathbb{E}}_{z_2}[\varepsilon_2]}{\partial^2 r}\left(1-\mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]\right) \\ &- 2\frac{\partial \mathop{\mathbb{E}}_{z_2}[\varepsilon_2]}{\partial r}\frac{\partial \mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]}{\partial r} \\ &+ \frac{\partial^2 \mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]}{\partial^2 r}\left(1-\mathop{\mathbb{E}}_{z_2}[\varepsilon_2]\right).\end{aligned} \tag{21}$$

Hence, $\frac{\partial^2 C_{\mathrm{BL}}}{\partial^2 r}<0$ if $\frac{\partial^2 \mathbb{E}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r}>0$.

Recall that the coding rate is determined based on $\eta \bar{h}_k^2$ ($j=1,2,3$), where $0<\eta\leq\hat{z}$. In other words, the coding rate is definitely lower than the Shannon capacity of a link (either the backhaul link or the combined link) if the fading gain of this link is higher than η. Consider for example the backhaul link with fading z_2. We have $r<\mathcal{C}(z_2|\bar{h}_2|^2), z_2\in(\eta,+\infty)$ and $r>\mathcal{C}(z_2|\bar{h}_2|^2), z_2\in(0,\eta)$. Denote the integral part in (22) as $\Delta(z_2)$ for short. Hence we have $\frac{\partial^2 \mathbb{E}_{z_2}[\varepsilon_2]}{\partial^2 r} = \int_0^{\eta}\Delta(z_2)dz_2 + \int_{\eta}^{\hat{z}}\Delta(z_2)dz_2 + \int_{\hat{z}}^{+\infty}\Delta(z_2)dz_2$ where $\int_0^{\eta}\Delta(z_2)dz_2 < 0$ $\int_{\eta}^{\hat{z}}\Delta(z_2)dz_2>0$ and $\int_{\hat{z}}^{+\infty}\Delta(z_2)dz_2>0$.

As median, $\hat{z}$ satisfies $\int_0^{\hat{z}}e^{-z_2}dz_2=\int_{\hat{z}}^{+\infty}e^{-z_2}dz_2=\frac{1}{2}$.

The error probability ε_2 satisfies $0\leq\varepsilon_2(z_2)\leq 1$, and in particular $0\leq\varepsilon_2(z_2)\leq\frac{1}{2}$ if the instantaneous channel gain z_2 is higher than the weighted average channel gain (based on which the source determines the coding rate), i.e., $z_2>\hat{z}_2\geq\eta$.

$\Rightarrow \mathop{\mathbb{E}}_{z_2}[\varepsilon_2] = \int_0^{\hat{z}}e^{-z_2}\varepsilon_2(z_2)\,dz_2 + \int_{\hat{z}}^{\infty}e^{-z_2}\varepsilon_2(z_2)\,dz_2 < \int_0^{\hat{z}}e^{-z_2}dz_2 + \int_{\hat{z}}^{\infty}e^{-z_2}\frac{1}{2}dz_2 = \frac{1}{2}+\frac{1}{4}=\frac{3}{4}$.

In addition, $\frac{\partial^2 \mathbb{E}_{z_2}[\varepsilon_2]}{\partial^2 r}$ can be given as:

$$\begin{aligned}\frac{\partial^2 \mathop{\mathbb{E}}_{z_2}[\varepsilon_2]}{\partial^2 r} &= \mathop{\mathbb{E}}_{z_2}\left[\frac{\partial^2\varepsilon_2}{\partial^2 r}\right] \\ &= \int_0^{+\infty}\frac{m^{\frac{3}{2}}\left(\mathcal{C}(z_2|\bar{h}_2|^2)-r\right)e^{-\frac{m\left(\mathcal{C}(z_2|\bar{h}_2|^2)-r\right)^2}{2\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)(\log_2 e)^2}}}{\sqrt{2\pi}\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)^{\frac{3}{2}}(\log_2 e)^3}e^{-z_2}dz_2.\end{aligned} \tag{22}$$

Moreover, the following relationship holds based on $\hat{z}$:

$$\begin{aligned}&\int_{\hat{z}}^{+\infty}\frac{\left(\mathcal{C}(z_2|\bar{h}_2|^2)-r\right)e^{-\frac{\left(\mathcal{C}(z_2|\bar{h}_2|^2)-r\right)^2}{\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)}}}{\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)^{\frac{3}{2}}}e^{-z_2}dz_2 \\ &> \int_0^{\hat{z}}\frac{\left|\mathcal{C}(z_2|\bar{h}_2|^2)-r\right|e^{-\frac{\left(\mathcal{C}(z_2|\bar{h}_2|^2)-r\right)^2}{\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)}}}{\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)^{\frac{3}{2}}}e^{-z_2}dz_2 \\ &\geq \int_0^{\eta}\frac{\left|\mathcal{C}(z_2|\bar{h}_2|^2)-r\right|e^{-\frac{\left(\mathcal{C}(z_2|\bar{h}_2|^2)-r\right)^2}{\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)}}}{\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)^{\frac{3}{2}}}e^{-z_2}dz_2 > 0\end{aligned}$$

$\Rightarrow \int_{\hat{z}}^{+\infty}\Delta(z_2)dz_2 - \left|\int_0^{\eta}\Delta(z_2)dz_2\right| > 0.$

$\Rightarrow$ We have $\frac{\partial^2 \mathbb{E}_{z_2}[\varepsilon_2]}{\partial^2 r}>0$ for the backhaul link.

Regarding the combined link, the combined channel gain is $|\bar{h}_{\mathrm{MRC}}|^2=z_1|\bar{h}_1|^2+z_3|\bar{h}_3|^2$. Therefore, we have

$$\begin{aligned}&\int_{\hat{z}}^{+\infty}\int_{\hat{z}}^{+\infty}\frac{\sqrt{2\pi}\left(\mathcal{C}\left(|\bar{h}_{\mathrm{MRC}}|^2\right)-r\right)e^{-\frac{m\left(\mathcal{C}\left(|\bar{h}_{\mathrm{MRC}}|^2\right)-r\right)^2}{\left(1-2^{-2\mathcal{C}\left(|\bar{h}_{\mathrm{MRC}}|^2\right)}\right)(\log_2 e)^2}-z_1-z_3}}{\left(\frac{1-2^{-2\mathcal{C}\left(|\bar{h}_{\mathrm{MRC}}|^2\right)}}{m}\right)^{\frac{3}{2}}(\log_2 e)^3}dz_1dz_3 \\ &\geq\int_0^{\eta}\int_0^{\eta}\frac{\sqrt{2\pi}\left|\mathcal{C}\left(|\bar{h}_{\mathrm{MRC}}|^2\right)-r\right|e^{-\frac{m\left(\mathcal{C}\left(|\bar{h}_{\mathrm{MRC}}|^2\right)-r\right)^2}{2\left(1-2^{-2\mathcal{C}\left(|\bar{h}_{\mathrm{MRC}}|^2\right)}\right)(\log_2 e)^2}-z_1-z_3}}{4\left(\frac{1-2^{-2\mathcal{C}\left(|\bar{h}_{\mathrm{MRC}}|^2\right)}}{m}\right)^{\frac{3}{2}}(\log_2 e)^3}dz_1dz_3 \\ &> 0\end{aligned}$$

Similarly, we have $\frac{\partial^2 \mathbb{E}_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]}{\partial^2 r}>0$.

Then, to prove $\frac{\partial^2 \mathbb{E}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r}>0$, the following two cases are considered, which differ in the relationship between the average channel gains of the backhaul link and the combined link are considered. The first case is $|\bar{h}_2|^2\leq|\bar{h}_1|^2+|\bar{h}_3|^2$, where the average channel gain of the backhaul link is lower than the combined link. As the fading of different links are i.i.d., under this case the instantaneous channel gain of the backhaul link is more likely to be lower than the combined link.

$|\bar{h}_2|^2\leq|\bar{h}_1|^2+|\bar{h}_3|^2$

$\Rightarrow 3/4 > \mathop{\mathbb{E}}_{z_2}[\varepsilon_2] \geq \mathop{\mathbb{E}}_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]$ and it is easy to prove $\frac{\partial^i \mathbb{E}_{z_2}[\varepsilon_2]}{\partial^i r} \geq \frac{\partial^i \mathbb{E}_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]}{\partial^i r}, i=1,2,...+\infty$ based on (2), (7) and (8).

$\Rightarrow$ Based on (21), $\frac{\partial^2 \mathbb{E}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r}$ is bounded by:

$$\frac{\partial^2 \mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r}>\frac{1}{4}\frac{\partial^2 \mathop{\mathbb{E}}_{z_2}[\varepsilon_2]}{\partial^2 r}-2\left(\frac{\partial \mathop{\mathbb{E}}_{z_2}[\varepsilon_2]}{\partial r}\right)^2.$$

Consider m is the blocklength, where $m\gg 1$, we have:

$$\frac{\partial^2 \mathop{\mathbb{E}}_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r}>\frac{m^{\frac{3}{2}}}{8\sqrt{2\pi}(\log_2 e)^3}\int_0^{+\infty}\frac{\left(\mathcal{C}(z_2|\bar{h}_2|^2)-r\right)e^{-\frac{m\left(\mathcal{C}(z_2|\bar{h}_2|^2)-r\right)^2}{2\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)(\log_2 e)^2}}}{\left(1-2^{-2\mathcal{C}(z_2|\bar{h}_2|^2)}\right)^{\frac{3}{2}}}e^{-z_2}dz_2$$

$$= \frac{m^{\frac{3}{2}}}{8\sqrt{2\pi}(\log_2 e)^3} \int_0^{+\infty} \Phi(z_2)\, e^{-z_2} dz_2.$$

Based on the same idea during the above proof of $\frac{\partial^2 \mathop{\mathbb{E}}\limits_{z_2}[\varepsilon_2]}{\partial^2 r} > 0$, it is easy to have $\int_{\hat{z}}^{+\infty} \Phi(z_2) dz_2 - \left|\int_0^{\eta} \Phi(z_2) dz_2\right| > 0$ and $\int_{\eta}^{\hat{z}} \Phi(z_2)\, dz_2 > 0$

$\Rightarrow \int_0^{\infty} \Phi(z_2)\, dz_2 > 0$

$\Rightarrow \frac{\partial^2 \mathop{\mathbb{E}}\limits_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r} > 0$ under the case $|\bar{h}_2|^2 \le |\bar{h}_1|^2 + |\bar{h}_3|^2$.

Under the other case $|\bar{h}_2|^2 > |\bar{h}_1|^2 + |\bar{h}_3|^2$, it can be proved $\frac{\partial^2 \mathop{\mathbb{E}}\limits_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r} > 0$ similarly (based on $\frac{\partial^2 \mathop{\mathbb{E}}\limits_{z_1,z_3}[\varepsilon_{\mathrm{MRC}}]}{\partial^2 r} > 0$).

$\Rightarrow \frac{\partial^2 C_{\mathrm{BL}}}{\partial^2 r} < 0$

$\Rightarrow C_{\mathrm{BL}}$ is a strictly concave in the coding rate.

■

Appendix B
Proof of the Corollary 1

Proof: r is strictly increasing in η, $0 < \eta \le \hat{z}$.

$\Rightarrow \forall\ x < y,\ x, y \in (0, \hat{z})$ and $\lambda \in [0, 1]$, we have $r|_{\eta=x} < r|_{\eta=\lambda x+(1-\lambda)y} < r|_{\eta=x}$.

Based on Theorem 1, C_{BL} is concave in r,

$\Rightarrow \min\left\{C_{\mathrm{BL}}\left(r|_{\eta=x}\right), C_{\mathrm{BL}}\left(r|_{\eta=y}\right)\right\} \leqslant C_{\mathrm{BL}}\left(r|_{\eta=\lambda x+(1-\lambda)y}\right)$.

$\Rightarrow C_{\mathrm{BL}}$ is quasi-concave in η, $0 < \eta \le \hat{z}$.

■

Appendix C
Proof of the Theorem 2

Proof: Comparing (17) and (5), we have $R_{\mathrm{MS}} = \frac{C_{\mathrm{BL}}}{2} + R^*$, where $R^* = \frac{r}{4}\sqrt{(1 - \mathop{\mathbb{E}}\limits_{z_1,z_2,z_3}[\varepsilon_{\mathcal{R}}])^2 + \frac{4m\ln(P_{\mathrm{d}})}{d} \mathop{\mathbb{E}}\limits_{z_1,z_2,z_3}[\varepsilon_{\mathcal{R}}](1 - \mathop{\mathbb{E}}\limits_{z_1,z_2,z_3}[\varepsilon_{\mathcal{R}}])}$.

As we have proved that C_{BL} is concave in r in Theorem 1, Theorem 2 holds if R^* is concave, too.

Note that $\frac{4m\ln(P_{\mathrm{d}})}{d}$ is not influenced by r. It is actually a negative constant as $\ln(P_{\mathrm{d}}) < 0$. We denote this constant by φ ($\varphi < 0$) for short. To facilitate the proof we also denote $\mathop{\mathbb{E}}\limits_{z_1,z_2,z_3}[\varepsilon_{\mathcal{R}}]$ by $\bar{\varepsilon}_{\mathcal{R}}$ for short.

Then, we have:

$R^* = \frac{r}{4}\sqrt{(1-\bar{\varepsilon}_{\mathcal{R}})^2 + \varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})} = \frac{r}{4}\sqrt{1 + (\varphi-2)\bar{\varepsilon}_{\mathcal{R}} + (1-\varphi)\,\bar{\varepsilon}_{\mathcal{R}}^{\,2}}$

As R^* is a square root function, it should be satisfied that:

$$(1 - \bar{\varepsilon}_{\mathcal{R}})^2 + \varphi\bar{\varepsilon}_{\mathcal{R}} \cdot (1 - \bar{\varepsilon}_{\mathcal{R}}) \ge 0. \tag{23}$$

$\bar{\varepsilon}_{\mathcal{R}}$ is the expectation of the error probability over fading

$\Rightarrow 0 < \bar{\varepsilon}_{\mathcal{R}} < 1$

$\Rightarrow 1 - \bar{\varepsilon}_{\mathcal{R}} > 0$. Combining this with (23)

$\Rightarrow 1 - \bar{\varepsilon}_{\mathcal{R}} + \varphi\bar{\varepsilon}_{\mathcal{R}} \ge 0$,

$\Rightarrow 1 \ge (1 - \varphi)\,\bar{\varepsilon}_{\mathcal{R}}$.

$\frac{\partial R^*}{\partial r} = \frac{1}{4}\sqrt{1 + (\varphi - 2)\,\bar{\varepsilon}_{\mathcal{R}} + (1 - \varphi)\,\bar{\varepsilon}_{\mathcal{R}}^{\,2}} + \frac{r}{8}\frac{(\varphi-2)+2(1-\varphi)\bar{\varepsilon}_{\mathcal{R}}}{\sqrt{1+(\varphi-2)\bar{\varepsilon}_{\mathcal{R}}+(1-\varphi)\bar{\varepsilon}_{\mathcal{R}}^{\,2}}}\frac{\partial\bar{\varepsilon}_{\mathcal{R}}}{\partial r}$

$\frac{\partial^2 R^*}{\partial^2 r} = \frac{1}{8}\frac{(\varphi-2)+2(1-\varphi)\bar{\varepsilon}_{\mathcal{R}}+r(1-\varphi)}{\left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{1}{2}}}$
$+\frac{r}{16}\frac{-((\varphi-2)+2(1-\varphi)\bar{\varepsilon}_{\mathcal{R}})^2}{\left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{3}{2}}}\left(\frac{\partial\bar{\varepsilon}_{\mathcal{R}}}{\partial r}\right)^2 + \frac{r}{8}\frac{(\varphi-2)+2(1-\varphi)\bar{\varepsilon}_{\mathcal{R}}}{\left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{1}{2}}}\frac{\partial^2\bar{\varepsilon}_{\mathcal{R}}}{\partial^2 r}$

As we have $\varphi < 0$ and $0 < \bar{\varepsilon}_{\mathcal{R}} < 1$

$\Rightarrow 0 < (1 - \bar{\varepsilon}_{\mathcal{R}})^2 + \varphi\bar{\varepsilon}_{\mathcal{R}}(1 - \bar{\varepsilon}_{\mathcal{R}}) < (1 - \bar{\varepsilon}_{\mathcal{R}})^2 < 1$

$\Rightarrow 0 < \left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{3}{2}} < \left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{1}{2}} < 1$

$\Rightarrow$ we have:

$\frac{\partial^2 R^*}{\partial^2 r} < \frac{1}{8}\frac{(\varphi-2)+2(1-\varphi)\bar{\varepsilon}_{\mathcal{R}}+r(1-\varphi)}{\left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{1}{2}}}$
$+\frac{r}{16}\frac{-((\varphi-2)+2(1-\varphi)\bar{\varepsilon}_{\mathcal{R}})^2}{\left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{1}{2}}}\left(\frac{\partial\bar{\varepsilon}_{\mathcal{R}}}{\partial r}\right)^2 + \frac{r}{8}\frac{(\varphi-2)+2(1-\varphi)\bar{\varepsilon}_{\mathcal{R}}}{\left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{1}{2}}}\frac{\partial^2\bar{\varepsilon}_{\mathcal{R}}}{\partial^2 r}$

As $1 \ge (1 - \varphi)\,\bar{\varepsilon}_{\mathcal{R}}$, we have:

$\frac{\partial^2 R^*}{\partial^2 r} < \frac{2\varphi+2r\left(1-\varphi-\frac{\varphi^2}{2}\left(\frac{\partial\bar{\varepsilon}_{\mathcal{R}}}{\partial r}\right)^2+\varphi\frac{\partial^2\bar{\varepsilon}_{\mathcal{R}}}{\partial^2 r}\right)}{16\left((1-\bar{\varepsilon}_{\mathcal{R}})^2+\varphi\bar{\varepsilon}_{\mathcal{R}}(1-\bar{\varepsilon}_{\mathcal{R}})\right)^{\frac{1}{2}}}$

As shown in Appendix A, $\frac{\partial^2 \mathop{\mathbb{E}}\limits_{i}[\varepsilon_{\mathcal{R},i}]}{\partial^2 r} > 0$. In addition, $\left(\frac{\partial\bar{\varepsilon}_{\mathcal{R}}}{\partial r}\right)^2 \sim \mathrm{O}(m)$ and $\frac{\partial^2\bar{\varepsilon}_{\mathcal{R}}}{\partial^2 r} \sim \mathrm{O}\left(m^{3/2}\right)$. Moreover, φ is a constant with reasonable value[5] $\varphi \in (-30m, -0.05m)$. Therefore, $\frac{\partial^2 R^*}{\partial^2 r} < 0$ holds.

Hence, R_{MS} is concave in r.

■

References

[1] J. Laneman, D. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[2] S. Karmakar and M. Varanasi, "The diversity-multiplexing tradeoff of the dynamic decode-and-forward protocol on a MIMO half-duplex relay channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6569–6590, Oct. 2011.

[3] Y. Huang, F. Al-Qahtani, C. Zhong et al., "Performance analysis of multiuser multiple antenna relaying networks with co-channel interference and feedback delay," *IEEE Trans. Commun.*, vol. 62, no. 1, pp. 59–73, Jan. 2014.

[4] F. Parzysz, M. Vu and F. Gagnon, "Impact of propagation environment on energy-efficient relay placement: Model and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2214–2228, Feb. 2011.

[5] Y. Hu and L. Qiu, "A novel multiple relay selection strategy for LTE-advanced relay systems," in *Proc. IEEE Vehicular Technology Conference (Spring)*, Budapest, Hungary, May 2011.

[6] Y. Hu, J. Gross and A. Schmeink, "QoS-Constrained energy efficiency of cooperative ARQ in multiple DF relay systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 848-859, Feb. 2016

[7] Y. Hu and J. Gross, "On the outage probability and effective capacity of multiple decode-and-forward relay system," in *Proc. IFIP Wireless Days*, Dublin, Ireland, Nov. 2012.

[8] M. Bhatnagar, "On the capacity of decode-and-forward relaying over Rician fading channels," *IEEE Commun. Lett.*, vol. 17, no. 6, pp. 1100–1103, June 2013.

[9] 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.

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

[11] 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.

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

[13] 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.

[5]Recall that φ is subject to the QoS constraints {delay (in symbols), delay violation probability}. Considering extremely loose constraints $\{100m, 10^{-0.5}\}$ and extremely strict constraints $\{2m, 10^{-7}\}$ (recall that the length of a transmission period of relaying is $2m$), we have $\varphi \in (-30m, -0.05m)$.

[14] 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.
[15] ——, "Finite block-length analysis of spectrum sharing networks using rate adaptation," *IEEE Trans. Commun.* , vol. 63, no. 8, pp. 2823–2835, Aug 2015.
[16] 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.
[17] ——,"On the performance advantage of relaying under the finite block-length regime," *IEEE Commn. Lett.*, vol. 19, no. 5, pp. 779 - 782, May 2015.
[18] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.
[19] S. Ren and K. Letaief, "Maximizing the effective capacity for wireless cooperative relay networks with qos guarantees," *IEEE Trans. Commun.* , vol. 57, no. 7, pp. 2148–2159, July 2009.
[20] J. Harsini and M. Zorzi, "Effective capacity for multi-rate relay channels with delay constraint exploiting adaptive cooperative diversity," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3136–3147, Sep. 2012.
[21] A. Kumar, D. Manjunath, and J. Kuri, "Communication Networking: An Analytical Approach", *Morgan Kaufmann Publishers*, 2004.
[22] B. Soret, M. Aguayo-torres, and J. Entrambasaguas, "Multiuser capacity for heterogeneous qos constraints in uncorrelated rayleigh channels," in *IEEE Information Theory Workshop*, pp. 278–282, Oct. 2009
[23] B. Soret, M. Aguayo-Torres, and J. Entrambasaguas, "Capacity with explicit delay guarantees for generic sources over correlated Rayleigh channel," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 1901 –1911, Jun. 2010.
[24] A. F. Molisch, *Wireless Communications*, 2011.
[25] S. Verdu and T. S. Han, "A general formula for channel capacity, "*IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

Yulin Hu received the B.Eng. degree in communications engineering from Haerbin Institute of Technology, Weihai, China, in 2007 and the M.Sc. degree in engineering of communication and information system from University of Science and Technology of China, China, in 2011. Recently, he received the Ph.D. degree in electrical engineering and information technology from RWTH Aachen University, Germany, Jan. 2016. Now, he is a Post-doctoral researcher with the Institute for Theoretical Information Technology, RWTH Aachen University, Germany. His research interests are mainly with cooperative communication and relaying, including effective capacity, energy efficiency and finite-blocklength-limited performances analysis as well as optimal system design.

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

James Gross is an associate professor with the Electrical Engineering School of KTH Royal Institute of Technology, Stockholm since November 2012. He also serves as vice director for the ACCESS Linneaus Centre and is a member of the board of KTH's Innovative Centre for Embedded Systems. His research interests are broadly in the area of mobile systems & networks, with a focus on critical machine-to-machine communications, cellular networks, resource allocation as well as performance evaluation methods (in particular stochastic network calculus as well as network simulation). Prior to joining KTH, he was assistant professor and head of the Mobile Network Performance Group at RWTH Aachen University from 2008 – 2012 as well as a member of the DFG-funded UMIC research centre of RWTH. James studied at TU Berlin and UC San Diego, and received his PhD from TU Berlin in 2006. James has published about 100 (peer-reviewed) papers in international journals and conferences. His work has been awarded multiple times, among them the best paper awards at ACM MSWiM 2015, the best demo paper award at IEEE WoWMoM 2015, the best paper award at IEEE WoWMoM 2009 and the best paper award at European Wireless 2009. In 2007, James was the recipient of the ITG/KuVS dissertation award for his PhD thesis.