Backhauling Many Devices: Relay Schemes for Massive Random Access Networks

Dennis Ogbe
Virginia Tech
dogbe@vt.edu

David J. Love
Purdue University
djlove@purdue.edu

Chih-Chun Wang
Purdue University
chihw@purdue.edu

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9322431

***Abstract*—The number of devices connected to Internet of Things (IoT) and massive machine-type communication (mMTC) networks is expected to increase exponentially in the next generation of wireless communication systems, resulting in a new type of "massive" random access network. However, most of the work in this emerging field considers the *single-hop* setting with direct communication between the users and a fully-equipped base station. In contrast, this work explores the massive random access problem in a *two-hop relay* setting where users access the network through a femto- or pico-cell relay which itself only has a limited amount of bandwidth/power. We present two low-complexity relaying schemes designed to minimize power consumption and discuss their tradeoffs using numerical simulations.**

I. Introduction

One of the next frontiers of beyond-5G communication standards is the emergence of so-called *massive* networks of low-power and low-complexity devices, evolving the current paradigms of the Internet of Things (IoT) and massive machine-type communications (mMTC) into a hyper-connected *Internet of Everything* (IoE) [1]. Characteristics of these networks include user densities larger than that of current networks by multiple orders of magnitude, small message sizes, bursty transmission patters, and grant-free medium access. For a comprehensive review of random access and multiple-access techniques, see e.g., [2].

It has become clear that traditional communication standards are unsuited for these massive random access networks and new communications schemes need to be carefully designed to fit these new requirements. Fortunately, recent work in this area has begun to establish novel system models, communication schemes, and fundamental limits. For example, the work in [3] adapted the traditional information-theoretic multi-access channel (MAC) model to a massive number of users which scale as a function of the length of a frame of data. In addition, the works in [4], [5] departed from the traditional MAC model by formulating the massive random access communication task such that the total number of users in the network K_{tot} may grow infinitely, while the active number of users per frame K_{a} is held constant, with the ultimate goal of their analysis being the minimization of the users' transmit power.

One commonality of the previous work in this area is the focus on what we call the *single-hop* setting, in which the users, which may be thought of as a set of IoT or sensor nodes, wish to communicate their bursty stream of short packets directly to a base station. However, with the trend towards heterogeneous network architectures expected to continue to gain traction in 5G and beyond-5G networks, we can expect that in real-world settings, many users will communicate with a fully-equipped base station exclusively through femto- or pico-cells acting as relays. This trend has inspired recent work in the area of transmission schemes for multi-hop relay networks [6], [7]. As part of this line of research, this work represents a first step on the path of combining the massive random access set-up with the practical model of utilizing a relay between the users and the final destination.

More specifically, the contributions of this work can be summarized as follows. First, we extend the recently developed model of the single-hop Gaussian massive random access channel (G-mRAC) [4] to the two-hop relay setting. We then present two relaying schemes which extend a recently developed practical low-complexity transmission scheme [5] for the single-hop G-mRAC to the two-hop relay setting. Finally, we perform numerical studies of the power consumption of our proposed schemes and discuss the insights gained from these studies.

II. System Model and Problem Set-up

A. System Model

We consider the Gaussian massive random access channel from [4], [5] extended by an additional hop between the users and the destination, which we will refer to as the *relay channel* or *relay-destination channel* in this text. This scenario is sketched in Fig. 1. We assume a total of K_{tot} users in the network who all wish to convey messages to the destination. However, the users cannot reach the destination directly and are thus required to utilize the help of the relay. Communication is assumed to be frame-synchronized with frames of blocklength n symbols, and all nodes in the network have knowledge of the frame boundaries. During each frame, $K_{\mathrm{a}} \leq K_{\mathrm{tot}}$ users are considered active with each having $k \ll n$ bits of information to convey to the destination. We note that since K_{tot} may grow to infinity and is not needed for the remainder of the discussion, its value can be ignored. At the beginning of each frame, the K_{a} users each choose a message to be conveyed to the destination. We assume that all users share the same message set $\mathcal{M}$, that all messages are

This work was supported in parts by NSF under Grant CCF-1422997, Grant CCF-1618475, and Grant CCF-1816013.

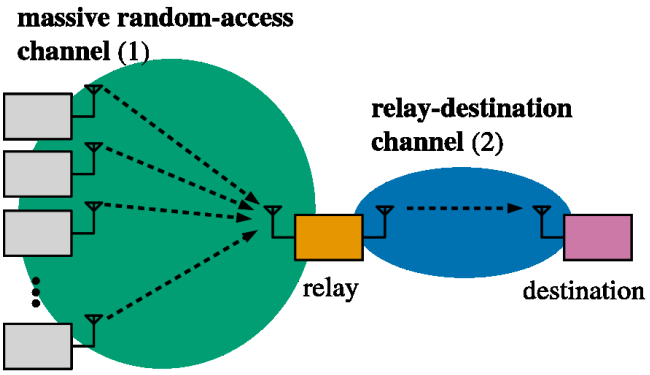

Fig. 1. System model. During one frame, K_{a} nodes wish to convey messages to the destination node via the relay.

equiprobable, and that all users employ the same codebook $\mathcal{C}$. We denote the message selected by the i-th user as $M_i \in \mathcal{M}$, the codeword selected by this user as $\mathbf{x}_i \in \mathcal{C} \subset \mathbb{R}^n$, and write the frame of n symbols observed by the relay as

$$\mathbf{y}^{(r)} = \sum_{i=1}^{K_{\mathrm{a}}} \mathbf{x}_i + \mathbf{z}, \tag{1}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is additive white Gaussian noise (AWGN). Furthermore, we assume that all nodes share an equal power constraint $\mathsf{E}\left[\|\mathbf{x}_i\|^2\right] \leq nP_x$. At this stage we note that the model in (1) implies a channel gain of unity for every user. In a real-world system, this can be achieved by computing a channel estimate at each user by utilizing the same signal that is used to convey the frame boundaries and exploiting channel reciprocity and using power normalization.

The communication task of the relay node is to form a transmit signal $\mathbf{x}^{(r)} = f_r(\mathbf{y}^{(r)}) \in \mathbb{R}^n$, which, after being observed by the destination node through the relay-destination channel, is input to the decoder at the destination. We require the power constraint $\mathsf{E}\left[\left\|\mathbf{x}^{(r)}\right\|^2\right] \leq nP_r$ at the relay and write the observation of one frame at the destination as

$$\mathbf{y}^{(d)} = \mathbf{x}^{(r)} + \mathbf{w}, \tag{2}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is AWGN.

In a similar manner as in the single-hop case outlined in [4], [5], the decoding task of the destination is to construct an unordered list of messages of size J, denoted as $\mathcal{L}(\mathbf{y}^{(d)}) = (m_1, \dots m_J)$, which, in the ideal case, contains all of the messages selected by the users. However, since some messages could be duplicated, we allow for $J \leq K_{\mathrm{a}}$. An error for user i is declared if $M_i \notin \mathcal{L}(\mathbf{y}^{(d)})$. The error probability of this set-up is then defined as

$$\mathsf{P_e} = \frac{1}{K_{\mathrm{a}}} \sum_{i=1}^{K_{\mathrm{a}}} \Pr\left[M_i \notin \mathcal{L}(\mathbf{y}^{(d)})\right]. \tag{3}$$

At this stage, we note that the motivating examples behind this work are the massive, city-scale IoT or sensor networks envisioned as part of future wireless communication standards[1]. In this light, the relay-destination channel is modeling the backhaul link between a low-power wireless pico- or femto-cell unit and a fully-equipped base station with a fiber connection. In this setting, the power consumption of the users and the relay nodes becomes a major aspect of the overall system design. Thus, the goal of this line of work is to design relaying schemes which minimize either the user's transmit power P_x or the relay's transmit power P_r.

We furthermore note that the model for the random-access channel, first introduced in [4], [5], differs substantially from the conventional information-theoretic model of the multiple-access channel (MAC). The conventional MAC model allows the users to utilize individual codebooks and power constraints and defines the *joint* error probability as the error metric as opposed to the *per-user* error probability from (3). In addition, there are also issues like power-control (to compensate near-far effects), synchronization (all nodes send simultaneously), and admission control (how to estimate K_{a} accurately). However, as an analytical first step and following the existing set-up from [5], we assume the simplified model as described in (1)—(3).

Finally, since this paper presents a preliminary exploration of the relay-specific aspects of this model, we assume that communication over the random-access channel (1) is handled via the single-hop scheme introduced in [5], and we concern ourselves mainly with the task of designing the transmission scheme over the relay channel (2). We thus now briefly describe the random-access scheme from [5]—which we from now on refer to as *Ordentlich-Polyanskiy scheme* or short *OP scheme*—, before presenting two possible relaying techniques which can be used to adapt this scheme to the relay setting described in this section.

B. The OP Scheme

From a high-level perspective, the OP scheme can be seen as a variation of slotted ALOHA which allows collisions of up to T users to be successfully decoded at the receiver. The description in this section is condensed and highlights the overall principle of [5]. The n symbols of one frame are divided into V sub-blocks, where $T \geq 1$ and $\alpha \in [0, 1]$ are design parameters such that $V = K_{\mathrm{a}}/(\alpha T)$ and are optimized to minimize the users' transmit power P_x. The common codebook $\mathcal{C}$ is constructed as a concatenated binary code mapped to a BPSK constellation and is described as follows. Given a message M_i, the encoding function of the i-th user produces a binary codeword $\mathbf{c}_i \in \{0, 1\}^{\bar{n}}$ of length $\bar{n}$, where $\bar{n} = \frac{n}{V}$ is the block length of one sub-block after dividing the total block length n into the V sub-blocks. The user then chooses one of the V sub-blocks uniformly at random and transmits its codeword during this sub-block using a BPSK constellation, i.e.,

$$\mathbf{x}_i = 2\sqrt{VP_x}\left(\mathbf{c}_i - \frac{1}{2}\right). \tag{4}$$

[1]The work in [4], [5] assumes a setting in which $k = 100$ bits, $n = 30000$ symbols, $\mathsf{P_e} \leq 0.05$ and $20 \leq K_{\mathrm{a}} \leq 300$. For simplicity, this work assumes the same parameters.

The codeword $\mathbf{c}_i$ is obtained from the following concatenated code. The inner code $\mathcal{C}_{\text{lin}}$ is a systematic binary linear code of length $\bar{n}$ and rate R_{lin}. The outer code $\mathcal{C}_{\text{BAC}}$ is a BCH code of rate $R_{\text{BAC}} = 1/T$. Let $R = R_{\text{lin}} \cdot R_{\text{BAC}}$ (units: bits/symbol) denote the rate of the concatenated code. The two-stage encoding then proceeds as follows. First, the BCH outer code maps the $k = \log_2(M)$ input bits to a a binary codeword $\mathbf{c}_{\text{BAC},i} \in \{0,1\}^{Tk}$ of length Tk. Then, the inner linear code maps the Tk bits of $\mathbf{c}_{\text{BAC},i}$ to an output codeword $\mathbf{c}_i \in \{0,1\}^{\bar{n}}$ of length $\bar{n}$.

Decoding is performed in two stages on a sub-block basis, i.e., the decoder produces a list of messages for each sub-block and the union of these lists over all sub-blocks gives the result for the entire frame. In the following, only the first sub-block is considered. Suppose that $\{i_1, \ldots, i_L\}$ are the L active users which transmitted during the first sub-block. Denote the $\bar{n}$ received symbols of the first sub-block at the receiver (in the set-up of [5], this is the final destination; however, in our set-up, this would be the relay node) as $\mathbf{y}_1 \in \mathbb{R}^{\bar{n}}$. In the first stage of decoding (dubbed the Compute-and-Forward (CoF) stage in [5] due to its similarities to the concepts of [8]), the decoder computes

$$\mathbf{y}_{\text{CoF},1} = \left[\frac{1}{2\sqrt{VP_x}}\mathbf{y}_1 + \frac{L}{2}\right] \bmod 2 \tag{5}$$

$$= \left[\sum_{j=1}^{L} \mathbf{c}_{i_j} + \tilde{\mathbf{z}}_1\right] \bmod 2 \tag{6}$$

$$= \left[\mathbf{c}_1^{\oplus} + \tilde{\mathbf{z}}_1\right] \bmod 2, \tag{7}$$

where $\mathbf{c}_1^{\oplus} = \left[\sum_{j=1}^{L} \mathbf{c}_{i_j}\right] \bmod 2 \in \mathcal{C}_{\text{lin}}$ is the modulo-2 sum[2] of the transmitted codewords and

$$\tilde{\mathbf{z}}_1 \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{4VP_x}\mathbf{I}\right). \tag{8}$$

Since $\mathbf{c}_1^{\oplus} \in \mathcal{C}_{\text{lin}}$, this step effectively transforms the multiple-access channel (1) into the modulo-2 AWGN channel (7), where the "input" $\mathbf{c}_1^{\oplus}$ is drawn from the binary systematic linear code $\mathcal{C}_{\text{lin}}$. Thus, the decoder can produce an estimate $\hat{\mathbf{c}}_1^{\oplus}$ using the decoding function of $\mathcal{C}_{\text{lin}}$. Denote the error event associated with this stage as $E^{(\text{lin})} = \{\hat{\mathbf{c}}_1^{\oplus} \neq \mathbf{c}_1^{\oplus}\}$.

In the case where $E^{(\text{lin})}$ did not occur, due to the linearity of $\mathcal{C}_{\text{lin}}$, the first $k_{\text{lin}} \triangleq \bar{n}R_{\text{lin}}$ symbols of $\hat{\mathbf{c}}_1^{\oplus}$ correspond to

$$\left(\hat{\mathbf{c}}_1^{\oplus}\right)_{[1:k_{\text{lin}}]} \triangleq \mathbf{y}_{\text{BAC},1} = \left[\sum_{j=1}^{L} \mathbf{c}_{\text{BAC},i_j}\right] \bmod 2, \tag{9}$$

which is equivalent to the input-output relationship of a binary adder channel (BAC). One of the key insights of the OP scheme is that a certain class of BCH codes of rate $1/T$ can be used to decode the individual terms of this sum[3] and thus reproduce a list of the original messages with a negligible error probability, provided that $L \leq T$. This gives rise to the following two error events. First, denote the *over-the-air collision* error event for user i as $E_i^{(\text{coll.})} = \{L > T\}$. Second, denote the list of L messages which satisfy (9) and is output by the decoder as $\tilde{\mathcal{L}}(\mathbf{y}_{\text{BAC},1})$ and then define the error event of the BCH decoding stage as

$$E^{(\text{BAC})} = \left\{\tilde{\mathcal{L}}(\mathbf{y}_{\text{BAC},1}) \neq \{\mathbf{c}_{\text{BAC},i_1}, \ldots, \mathbf{c}_{\text{BAC},i_L}\}\right\}. \tag{10}$$

The overall error probability for the first sub-block is then bounded using the union bound as

$$\mathsf{P}_{\mathsf{e},i,1} \leq \Pr\left[E_i^{(\text{coll.})}\right] + \Pr\left[E^{(\text{lin})}\right] + \Pr\left[E^{(\text{BAC})} \middle| \overline{E_i^{(\text{coll.})}} \cap \overline{E^{(\text{lin})}}\right]. \tag{11}$$

Note that due to the symmetry of the code construction across users and sub-blocks, the overall error probability satisfies $\mathsf{P}_{\mathsf{e}} \leq \mathsf{P}_{\mathsf{e},i,1}$.

To give an example scenario of this code construction, suppose a network is designed to support $K_{\text{a}} = 100$ active users per frame and we have $n = 30000$ symbols and $k = 100$ bits. Furthermore, suppose the number of sub-blocks was chosen as $V = 50$ and the maximum number of users per sub-block was chosen as $T = 5$. This means that the length of one sub-block is given as $\bar{n} = 600$ symbols and the rate of the linear code must satisfy $R_{\text{lin}} = 5/6$. Since $R_{\text{BAC}} = 1/5$, an active user i first selects the 500-bit codeword $\mathbf{c}_{\text{BAC},i}$ corresponding to the message $M_i \in [1, \ldots, 2^{100}]$ before computing $\mathbf{c}_i$ (length: 600 bits) using $\mathcal{C}_{\text{lin}}$ and then $\mathbf{x}_i$ according to (4). At the beginning of the frame, user i then selects one out of the V sub-blocks at random. The probability that user i chooses sub-block v is $1/V$. The average number of users per sub-blocks is $K_{\text{a}}/V = 2$, but the probability of more than 4 other users choosing sub-block v and thus colliding with user i without a chance of successful decoding is ≈ 0.049 (see (23)).

At this stage, we note that [5] describes a straightforward multi-level extension of this scheme which allows for increased spectral efficiency and introduces the number of levels τ as additional design parameter to be optimized. While the description of this extension is out of the scope of this paper, we did include this optimization in the numerical studies of Section IV.

C. Introducing a Relay

The extension of the OP scheme to a two-hop setting is best described with the help of Fig. 2, which outlines the sequence of events during the transmission of one frame of data from the users to the destination. We assume full-duplex relaying and, for simplicity, assume that the processing time at the relay is negligible in comparison with the transmission time of one sub-block. The modified input-output model can then be described as follows.

As described in Section II-B, each active user chooses one of the V sub-blocks uniformly at random and transmits according to (4) (i.e., (1) in Fig. 2). As above, let $\{i_1, \ldots, i_L\}$ be the L active users which choose the first sub-block. The

[2]We use the "floored-division" definition of the modulo operation, i.e., $r = [s] \mod p$ implies $r = s - p\lfloor \frac{s}{p} \rfloor$ for a scalar $s \in \mathbb{R}$ and $[\mathbf{s}] \mod p \triangleq ([s_1] \mod p, \ldots, [s_n] \mod p)^{\mathsf{T}}$ for a vector $\mathbf{s} \in R^n$.

[3]Note that [5] assumes that L is known at the receiver and gives a practical justification for this assumption.

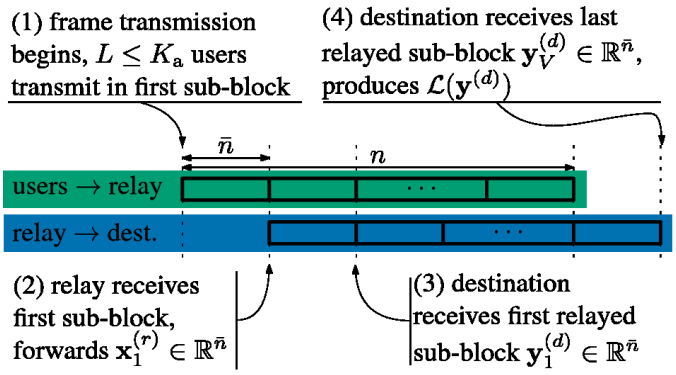

Fig. 2. Sequence of events during one frame transmission.

observation of the first sub-block at the relay is denoted as $\mathbf{y}_1^{(r)}$ and given as

$$\mathbf{y}_1^{(r)} = \sum_{j=1}^{L} \mathbf{x}_{i_j} + \mathbf{z}_1, \tag{12}$$

where $\mathbf{z}_1 \in \mathbb{R}^{\bar{n}} \sim \mathcal{N}(0, \mathbf{I})$ is AWGN and the transmit signals $\mathbf{x}_{i_j}$ are constructed using the concatenated code described in Section II-B. Once $\mathbf{y}_1^{(r)}$ is received at the relay, it constructs a transmit signal $\mathbf{x}_1^{(r)} = f_r(\mathbf{y}_1^{(r)}) \in \mathbb{R}^{\bar{n}}$ and forwards it to the destination node (i.e., (2) in Fig. 2). We denote the observation corresponding to the first sub-block at the destination as

$$\mathbf{y}_1^{(d)} = \mathbf{x}_1^{(r)} + \mathbf{w}_1, \tag{13}$$

where $\mathbf{w}_1 \in \mathbb{R}^{\bar{n}} \sim \mathcal{N}(0, \mathbf{I})$ is AWGN (i.e., (3) in Fig. 2). This relaying sequence continues until the last sub-block is received at the destination (i.e., (4) in Fig. 2). Denote the full frame as transmitted by the relay node and as received by the destination as $\mathbf{x}^{(r)} = ((\mathbf{x}_1^{(r)})^{\mathsf{T}}, \ldots, (\mathbf{x}_V^{(r)})^{\mathsf{T}})^{\mathsf{T}}$ and $\mathbf{y}^{(d)} = ((\mathbf{y}_1^{(d)})^{\mathsf{T}}, \ldots, (\mathbf{y}_V^{(d)})^{\mathsf{T}})^{\mathsf{T}}$, respectively. Then, note that we require only the destination node to produce the list of messages $\mathcal{L}(\mathbf{y}^{(d)})$, thus allowing the relay node to forward individual sub-blocks during the frame transmission. Furthermore, recall that the power constraint of the relay is given as $\mathsf{E}\left[\left\|\mathbf{x}^{(r)}\right\|^2\right] \leq nP_r$. Given this description of the relaying model, we want to investigate the following question: How to design the relaying function $f_r(\cdot)$ to minimize either P_x, P_r, or both? The following section introduces two candidates and discusses their tradeoffs, and we perform numerical studies using these techniques in Section IV.

III. Description of the Relaying Schemes

In this section, we introduce two possible candidates for the relaying function $f_r(\cdot)$. The first scheme draws inspiration from the well-known *Amplify-&-Forward* technique [9] and enables low-complexity relay implementations. The second scheme, a novel technique dubbed *Compute-Encode-&-Forward*, exploits the coding structure of the OP scheme and requires some additional computing and processing capability at the relay.

A. Amplify-&-Forward (AF)

In this simple scheme, the relay scales its received signal by a constant factor and forwards it to the destination without any additional processing, effectively creating an equivalent single-hop channel with decreased signal-to-noise ratio. To illustrate this, we focus without loss in generality on the first sub-block and write the corresponding relayed signal as

$$\mathbf{x}_1^{(r,\mathsf{AF})} = \sqrt{\rho}\, \mathbf{y}_1^{(r)}, \tag{14}$$

where the scale factor ρ is chosen to satisfy the power constraint $\mathsf{E}\left[\left\|\mathbf{x}^{(r,\mathsf{AF})}\right\|^2\right] \leq nP_r$. The received signal at the destination then becomes

$$\mathbf{y}_1^{(d,\mathsf{AF})} = \sqrt{\rho} \sum_{j=1}^{L} \mathbf{x}_{i_j} + \tilde{\mathbf{w}}_1, \tag{15}$$

where $\tilde{\mathbf{w}}_1 \triangleq \sqrt{\rho}\, \mathbf{z}_1 + \mathbf{w}_1$ is the effective AWGN at the destination with $\tilde{\mathbf{w}}_1 \sim \mathcal{N}(\mathbf{0}, (1+\rho)\mathbf{I})$. The decoding process proceeds exactly as described in Section II-B, with the exception that a different scaling factor must be used in the CoF stage (5) resulting in

$$\mathbf{y}_{\mathrm{CoF},1}^{(d,\mathsf{AF})} = \left[\frac{1}{2\sqrt{\rho V P_x}} \mathbf{y}_1^{(d,\mathsf{AF})} + \frac{L}{2}\right] \bmod 2 \tag{16}$$

$$= \left[\mathbf{c}_1^{\oplus} + \tilde{\mathbf{z}}_1^{(\mathsf{AF})}\right] \bmod 2, \tag{17}$$

where the noise of the effective modulo-2 AWGN channel now satisfies

$$\tilde{\mathbf{z}}_1^{(\mathsf{AF})} \sim \mathcal{N}\left(\mathbf{0}, \frac{1+\rho}{4\rho V P_x}\mathbf{I}\right). \tag{18}$$

We observe that a simple scheme like the one described in this section is in accordance with the low-power requirement of IoT and sensor networks. In this spirit, we assume that the scale factor ρ is constant (rather than computed on-the-fly for every sub-block) and chosen to enforce the power constraint. With this restriction, it can be shown (the proof is omitted due to lack of space) that in order to satisfy $\mathsf{E}\left[\left\|\mathbf{x}^{(r,\mathsf{AF})}\right\|^2\right] \leq nP_r$, we need

$$\rho \leq \frac{P_r}{P_x K_{\mathrm{a}} + 2\sqrt{P_x K_{\mathrm{a}}} + 1}. \tag{19}$$

B. Compute-Encode-&-Forward (CEF)

This scheme assumes increased computing and processing capabilities at the relay node. Here, we assume that the relay performs the CoF phase for each sub-block, i.e., focusing as before only on the first sub-block, we assume that the relay computes $\hat{\mathbf{c}}_1^{\oplus}$ using the decoder of $\mathcal{C}_{\mathrm{lin}}$ as described in Section II-B. Note that the information necessary for the BCH decoding stage at the destination are the $k_{\mathrm{lin}} \triangleq \bar{n} R_{\mathrm{lin}}$ initial bits of $\hat{\mathbf{c}}_1^{\oplus}$. Denote these bits as $\bar{\mathbf{y}}_{\mathrm{BAC},1} \in \{0,1\}^{k_{\mathrm{lin}}}$.

To convey these bits to the destination node, the relay then employs a code designed for the relay-destination channel, denoted as $\mathcal{C}_r$, which is designed to uphold the power constraint $\mathsf{E}\left[\left\|\mathbf{x}_1^{(r,\mathsf{CEF})}\right\|^2\right] \leq \bar{n} P_r$ with block length $\bar{n}$. Denote the encoding and decoding operations of $\mathcal{C}_r$ as $\mathcal{C}_r(\cdot)$ and

$\mathcal{C}_r^{-1}(\cdot)$, respectively. We then have $\mathbf{x}_1^{(r,\mathsf{CEF})} = \mathcal{C}_r(\bar{\mathbf{y}}_{\mathrm{BAC},1})$ and define $\hat{\mathbf{y}}_{\mathrm{BAC},1} \triangleq \mathcal{C}_r^{-1}\left(\mathbf{y}_1^{(d,\mathsf{CEF})}\right)$ at the destination node. We denote the corresponding error event as $E^{(\mathrm{relay})} = \{\hat{\mathbf{y}}_{\mathrm{BAC},1} \neq \bar{\mathbf{y}}_{\mathrm{BAC},1}\}$. Note that in the event that $E^{(\mathrm{relay})}$ and $E^{(\mathrm{lin})}$ did not occur, we have

$$\hat{\mathbf{y}}_{\mathrm{BAC},1} = \left[\sum_{j=1}^{L} \mathbf{c}_{\mathrm{BAC},i_j}\right] \bmod 2, \tag{20}$$

which, as described in Section II-B, is input to the BCH decoding stage.

In contrast to the AF scheme, the error analysis for this scheme changes in comparison to the baseline single-hop OP scheme. Due to the additional error event $E^{(\mathrm{relay})}$, we now have

$$\begin{aligned}\mathsf{P}_{\mathsf{e},i,1} \leq{}& \Pr\left[E_i^{(\mathrm{coll.})}\right] + \Pr\left[E^{(\mathrm{lin})}\right] \\ &+ \Pr\left[E^{(\mathrm{relay})} \,\middle|\, \overline{E^{(\mathrm{lin})}} \cap \overline{E_i^{(\mathrm{coll.})}}\right] \\ &+ \Pr\left[E^{(\mathrm{BAC})} \,\middle|\, \overline{E^{(\mathrm{lin})}} \cap \overline{E_i^{(\mathrm{coll.})}} \cap \overline{E^{(\mathrm{relay})}}\right]. \end{aligned} \tag{21}$$

Finally, instead of attempting to characterize this new relay error probability $\epsilon^{(\mathrm{relay})} \triangleq \Pr\left[E^{(\mathrm{relay})} \,\middle|\, \overline{E^{(\mathrm{lin})}} \cap \overline{E_i^{(\mathrm{coll.})}}\right]$ for some fixed code, we note that the finite-length behavior of $\mathcal{C}_r$ can be conveniently bounded using the normal approximation [10] in a similar manner as it was done to evaluate $\Pr\left[E^{(\mathrm{lin})}\right]$ in [5]. Used in this context, the approximation states that the rate of $\mathcal{C}_r$ over the AWGN channel (13) at block length $\bar{n}$ with power constraint P_r for a fixed average error probability $\epsilon^{(\mathrm{relay})}$ can be approximated as

$$R_r \approx C(P_r) - \sqrt{\frac{V(P_r)}{\bar{n}}} Q^{-1}(\epsilon^{(\mathrm{relay})}), \tag{22}$$

where $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function $Q(x) = \int_x^\infty \frac{1}{\sqrt{2x}} e^{-t^2/2}\,\mathrm{dt}$ and $C(P) = \frac{1}{2}\log_2(1+P)$ and $V(P) = \frac{P}{2}\frac{P+2}{(P+1)^2}\log_2^2 e$ are the capacity and channel dispersion of a real-valued AWGN channel with SNR P_r, respectively. Since this approximation is known to be quite accurate, we use it in this work to determine the minimum transmit power P_r for the required code rate R_{lin}.

IV. Numerical Studies

In this section, we perform numerical studies of the power consumption of our proposed schemes. Both schemes require the computation of the minimum required user transmit power P_x in the OP scheme over the random access channel (12). To compute this, we follow the formula from [5], which is summarized in the following subsection.

A. Numerical Evaluation of the Random-Access Channel

The fixed input parameters are the block length n, the number of information bits at the users k, the number of active users per frame K_{a}, and the target error probability $\epsilon^{(\mathrm{ra})}$. Since it was shown in [5] that the error of the BCH decoding (i.e., the second term in (11)) is negligible, we choose $\epsilon^{(\mathrm{coll.})} \triangleq \Pr\left[E_i^{(\mathrm{coll.})}\right]$ and $\epsilon^{(\mathrm{lin})} \triangleq \Pr\left[E^{(\mathrm{lin})}\right]$ such that $\epsilon^{(\mathrm{ra})} = \epsilon^{(\mathrm{coll.})} + \epsilon^{(\mathrm{lin})}$. We then proceed to find P_x using the following brute-force search. For every $T \leq T_{\mathsf{max}}$ for some reasonably-chosen[4] value of T_{max} we choose α to be the solution of the equation

$$\epsilon^{(\mathrm{coll.})} = \Pr\left[\mathrm{Binomial}\left[K_{\mathrm{a}} - 1, \frac{\alpha T}{K_{\mathrm{a}}}\right] \geq T\right], \tag{23}$$

since we have K_{a} active users with each user choosing a particular sub-block with probability $1/V = \alpha T/K_{\mathrm{a}}$. A fixed α gives the required rate of the linear code $\mathcal{C}_{\mathrm{lin}}$, since

$$R_{\mathrm{lin}} = R/R_{\mathrm{BAC}} = \frac{k}{\bar{n}} \cdot T = \frac{k \cdot K_{\mathrm{a}}}{\alpha n} \tag{24}$$

for a single-level code. Note that in the simulations R_{lin} is further optimized as a function of the number of levels τ for the multi-level construction from [5] whose description is omitted due to lack of space. As a final step for a specific (T, α, τ) tuple, it remains to find the value of P_x which achieves the specified R_{lin} for block length $\bar{n}$ and error probability $\epsilon^{(\mathrm{lin})}$. This is done using the normal approximation of the effective modulo-2 AWGN channel (7). Finally, we output the minimum of all computed P_x values over all (T, α, τ) tuples.

B. Amplify-&-Forward (AF)

The AF scheme is evaluated in a similar manner as described in Section IV-A, with the only difference being the statistics of the additive noise in the effective modulo-2 AWGN channel (7), which now depend on the relay's power constraint P_r as described in Section III-A, given by (18). Fig. 3, shows the minimum P_x as a function of P_r for an increasing number of active users.

The first observation from Fig. 3 is that, in general, increasing the number of active users to be supported by the network requires increasing the transmit power at the users. This is in line with the single-hop baseline scheme from [5] and is due to the fact that increasing the number of users increases the total number of bits to be recovered during one frame, thus increasing the required code rate.

Another interesting insight from Fig. 3 is the behavior of the required user transmit power for small and large relay transmit powers. For decreasing P_r, we observe a seemingly exponentially increasing P_x. However, since we limited our evaluations to a maximum number of $\tau_{\mathsf{max}} = 5$ coding levels, effectively limiting the maximum rate and thus the maximum P_x, the curves are cut off at some minimum P_r. On the other hand, as we increase the relay transmit power P_r, we observe diminishing returns for sufficiently large values. This limiting behavior is not surprising however, since for fixed (T, α, τ), the required transmit power P_x is obtained by solving the normal approximation of the modulo-2 AWGN (17) and it is straightforward to show that in the limit as $P_r \to \infty$, the distribution of the noise in this channel converges to (8).

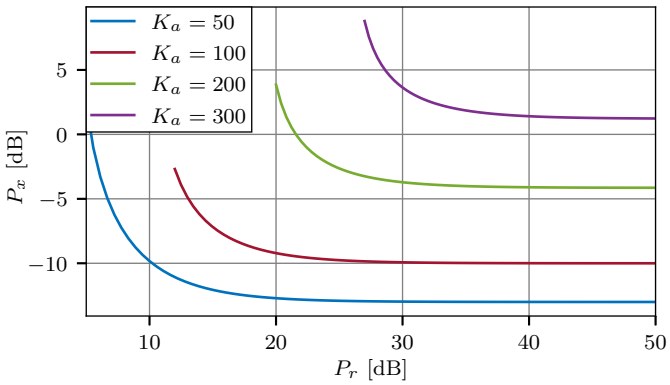

Fig. 3. User transmit power vs. relay transmit power with the AF scheme for a varying number of active users. Parameters: $k = 100$, $n = 30000$, $\mathsf{P_e} = 0.05$

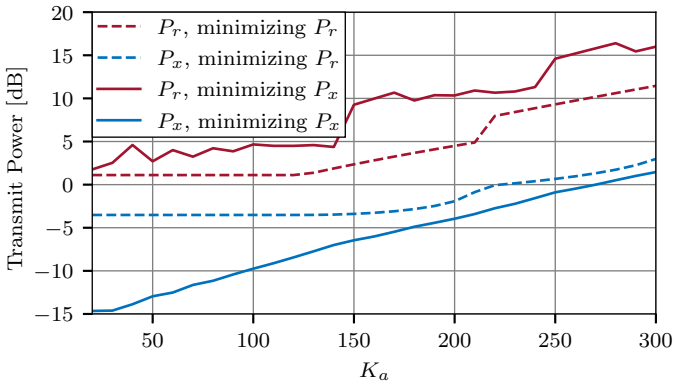

Fig. 4. Minimum user or relay transmit powers for increasing number of active users. Parameters: $k = 100$, $n = 30000$, $\mathsf{P_e} = 0.05$, $\eta = 0.1$

C. Compute-Encode-&-Forward (CEF)

To evaluate the CEF scheme from Section III-B, we introduce the additional parameter $\eta \in [0, 1]$ such that $\epsilon^{(\text{relay})} = \eta\mathsf{P_e}$ and $\epsilon^{(\text{ra})} = (1-\eta)\mathsf{P_e}$. In contrast to the AF study from the previous subsection, the relay transmit power in this subsection is not fixed and has to be computed using the normal approximation. To do this, we proceed with the same brute-force computation described in Section IV-A (see (23), (24), and the normal approximation of (7)) over the (T, α, τ)-space using $\epsilon^{(\text{ra})} = (1-\eta)\mathsf{P_e}$, which results in the required P_x for a specific (T, α, τ)-tuple. Then, as an additional step, we solve the normal approximation of the relay-destination channel (13) by solving (22) for $R_r = R_{\text{lin}}$ and $\epsilon^{(\text{relay})} = \eta\mathsf{P_e}$, giving the required P_r for this (T, α, τ)-tuple.

Figure 4 then presents two sets of curves. For an increasing number of active users K_a, the solid lines show the value of P_x and the corresponding value of P_r when selecting the (T, α, τ)-tuple which produced the minimum user transmit power P_x for a given value of K_a. Similarly, the dashed lines show the values of P_x and P_r when selecting for the minimum relay transmit power P_r.

[4]In our simulations, we chose $T_{\max} = 30$. We did not observe any optimal values at this edge.

We first notice that the solid line for P_x when minimizing P_x essentially reproduces the result from [5] with the linear increase in transmit power as a function of the number of active users. This is an expected outcome of this study. The corresponding P_r curve seems to follow this overall trend with increased variability between the increments in K_a. In addition, we observe an interesting behavior for the dashed lines representing the minimization of P_r. In the large-K_a regime, the required value of P_x seems to approach its corresponding solid line, indicating that system designers could significantly reduce the power consumption of their relays for only a moderate cost in the power consumption of the users by designing their transmission scheme to minimize P_r instead of P_x. Here, for $K_\text{a} = 300$, choosing the "minimizing P_r" strategy results in a ≈ 5 dB reduction in P_r at a cost of ≈ 1 dB increase in P_x when compared to the alternative.

V. Concluding Remarks

This work represents the first step toward combining two important aspects of beyond-5G networks: massive random access systems and multi-hop relaying. In this paper, we introduced the system model and gained some initial insights based on previous work. However, we have merely opened the door towards many interesting research problems in the future. First, a clear direction for future work is to design relaying schemes "from the ground up" with relaying in mind. Second, there is a need to design schemes for non-AWGN and fading channels. Finally, the extension of this simple two-hop model to more hops is another interesting direction for future work as the size and heterogeneity of networks beyond 5G continues to increase.

References

[1] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," 2019. [Online]. Available: http://arxiv.org/abs/1910.12678

[2] R. De Gaudenzi, O. del Río Herrero, S. Cioni, and A. Mengali, "Random access versus multiple access," in *Multiple Access Techniques for 5G Wireless Networks and Beyond*. Springer, 2019, pp. 535–584.

[3] X. Chen, T. Chen, and D. Guo, "Capacity of gaussian many-access channels," *IEEE Trans. Inform. Theory*, vol. 63, no. 6, pp. 3516–3539, June 2017.

[4] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2523–2527.

[5] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access gaussian channel," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2528–2532.

[6] C. Wang, D. J. Love, and D. Ogbe, "Transcoding: A new strategy for relay channels," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2017, pp. 450–454.

[7] D. Ogbe, C. Wang, and D. J. Love, "On the optimal delay amplification factor of multi-hop relay channels," in *2019 IEEE International Symposium on Information Theory (ISIT)*, July 2019, pp. 2913–2917.

[8] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct 2011.

[9] J. N. Laneman, D. N. C. 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.

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