Encoding of control information and data for downlink broadcast of short packets

Trillingsgaard, Kasper Fløe; Popovski, Petar

Published in:
2016 Information Theory and Applications Workshop (ITA)

DOI (link to publication from Publisher):
10.1109/ITA.2016.7888159

Publication date:
2017

Document Version
Accepted author manuscript, peer reviewed version

Encoding of Control Information and Data for Downlink Broadcast of Short Packets

Kasper Fløe Trillingsgaard and Petar Popovski
Department of Electronic Systems, Aalborg University
9220 Aalborg, Denmark

***Abstract*—It is almost an axiom that every cellular wireless system, including the upcoming 5G systems, should be based on data transmissions organized in frames. The frame design is based on heuristics, consisting of a frame header and data part. The frame header contains control information that specifies the sizes of the data packets and provides pointers to their location within the data part. In this paper we show that this design heuristics is suboptimal when the messages in the data part are short. We consider a downlink scenario represented by an AWGN broadcast channel with K users, while the sizes of the messages to the users are random variables. Each data packet encodes a message to one user. However, if the message sizes are small, there is a significant overhead caused by the header and the data packets can not be encoded efficiently. This calls for revision of the established heuristics for framing control information and data. We show that grouping messages of multiple users allows more efficient encoding from a transmitter perspective. On the other hand, it has the undesirable implication that it requires each user to decode the messages of a whole group of users. We assume that the power spend by each user is proportional to the number of channel uses it needs to decode. Using recent results in finite blocklength analysis, we investigate the trade-offs between total transmission time from the transmitter perspective and the average power spend at each user. Our approach shows that the space of feasible protocols is significantly enlarged and thereby allows the designer to trade-off between average total transmission time and the average power spend by each user.**

I. Introduction

Modern high-speed wireless networks heavily depend on reliable and efficient transmission of large data packets through the use of coding and information theory. The advent of machine-to-machine (M2M), vehicular-to-vehicular (V2V), and various streaming systems has spawned a renewed interest in developing information theoretical bounds and codes for communication of short packets [1][2]. Additionally, these applications often have tight reliability and latency constraints compared to a typical wireless systems today. Communication at shorter blocklengths introduces several new challenges which are not present when considering communication of larger data packets. For example, the overhead caused by control signals and header data is insignificant if large data packets are sent, and hence this overhead is often neglected in the analysis of protocols. However, more stringent latency requirements lead to shortened blocklengths for transmission, such that the size of the control information and header data may approach, or even exceed, the size of the actual data in the packet. This is especially true for multiuser systems such as broadcast channels, two-way channels, or multiple access channels, where the header data must include information about the packet structure, security, and user address information for identification purposes.

The fundamentals of communication of short packets have recently been addressed by Polyanskiy, Poor, and Verdú (2010) [3]. Here, it was shown that the maximal coding rate of a fixed-length block code in a traditional point-to-point setting is tightly approximated by

$$R^*(n,\epsilon) = C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log n}{n}\right) \qquad (1)$$

where C is the Shannon capacity, V is the channel dispersion, n is the blocklength, ϵ is the desired probability of error, and $Q^{-1}(\cdot)$ denotes the inverse Q-function. Approximations such as (1) are useful in the design of modern communication problems because the specifics of code selection can be neglected in the optimization of protocol parameters.

In this paper, we consider downlink transmission with an AWGN broadcast channel that consists of a transmitter and K users. There is a message from the transmitter to the k-th user with a certain probability $1-q$ (in this case user k is *active*). The size of the message is itself a random variable which implies that the transmitter needs to convey information about which users are active, the structure of the transmission, and sizes of the messages. An interesting observation from (1) is that larger data packets are encoded more efficiently. This introduces an interesting trade-off with two extremes: (1) in a broadcast setting one can either encode all messages in one large packet which is efficiently encoded or (2) one can encode each message separately as is the norm in modern wireless protocols. In (1) , the average total transmission time seen from the transmitter is minimized. However, all users need to receive for the whole period to be able to decode their message, which is undesirable for devices that are power-constrained. The latter approach (2), depicted on Fig. 1, uses codes which are less efficient, and thus the average total transmission time is larger. On the other hand, each user only needs to decode the information intended for that user. The key point, however, is that these design considerations enlarge the design space and enable the designer to trade-off between transmitter resources and user resources. Despite this, practically all wireless systems solely use the approach (2). The purpose of this paper is to explore this design trade-off. Specifically, by grouping multiple users together, we encode larger amount of information bits jointly, which implies that the rate at which the information bits of the groups can be encoded is larger. The disadvantage of grouping users is that each user

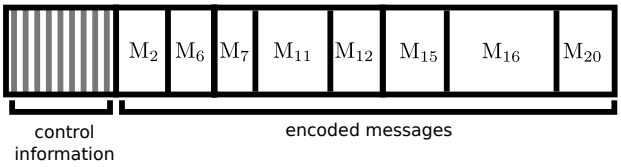

Fig. 1. Conventional approach to downlink broadcasting. An initial packet contains control information that defines the structure of the remaining part of the transmission. Each message are encoded separately.

needs receive for larger proportion of the total transmission time.

This paper is organized as follows. The next section describes the system model. Section III briefly addresses the approximations of the finite blocklength bounds, Section IV discusses design considerations for protocol design for downlink broadcast for short packets and describes our proposed protocol. Finally, we evaluate the proposed protocol in Section V and conclude the paper in Section VI.

II. System model

We consider an AWGN broadcast channel with one transmitter and K users. In the t-th time slot, the k-th user receive

$$Y_{k,t} \triangleq \sqrt{\gamma_k} X_t + Z_{k,t}. \tag{2}$$

where $Z_{k,t} \sim \mathcal{N}(0,1)$ and $X_t \in \mathbb{R}$ is the channel input. Throughout the paper, we assume $\gamma_k = \gamma$. The message M_k destined to the k-th user is nonempty with probability $1-q \in (0,1)$, and we say that the k-th user is *active* if there is a message destined to that user. The size of the message M_k in bits is denoted by $D_k \in \mathbb{Z}_+$, which is a discrete random variable distributed i.i.d. according to the probability mass function $P_D(\cdot)$ given by

$$P_D(d) = \begin{cases} q & \text{if } d = 0 \\ \frac{1-q}{S} & \text{if } d \in \{\alpha, \cdots, \alpha S\} \end{cases}. \tag{3}$$

for some $\alpha \in \mathbb{N}$ and $S \in \mathbb{N}$. The average message size of an active user is therefore $\mathbb{E}[D_k | D_k > 0] = \alpha(S+1)/2$.

Based on the message sizes D_k, the transmitter computes the total transmission time T which is also a random variable. The transmitter encodes the message $\{\mathrm{M}_k\}$ into a sequence of channel inputs using the encoder function $f_t(\mathrm{M}_1, \cdots, \mathrm{M}_K)$ such that

$$X_t \triangleq f_t(\mathrm{M}_1, \cdots, \mathrm{M}_K) \tag{4}$$

for $t \in \{1, \cdots, T\}$ and $X_t = 0$ for $t \in \{T+1, \cdots\}$.

At user k, we define the ON-OFF function $g_{k,t} : (\mathbb{R} \cup \{\mathrm{e}\})^{t-1} \to \{0,1\}$ that in turn defines the sequence

$$\bar{Y}_{k,t} \triangleq \begin{cases} Y_{k,t}, & g_{k,t}(\bar{Y}_k^{t-1}) = 1 \\ \mathrm{e}, & \text{otherwise} \end{cases}. \tag{5}$$

The ON-OFF function defines stopping times $T_k \triangleq \min\{n \geq 1 : \forall t > n, g_{k,t}(\bar{Y}_k^{t-1}) = 0\}$ for which we require $T_k < \infty$. Additionally, we define the decoding function $h_{k,t}(\bar{Y}_k^t)$ which estimates the message M_k based on $\bar{Y}_k^t$. The intuition is that a certain user can only use the channel outputs if the corresponding user is ON. This is modeled by the ON-OFF function which replaces the t-th channel output with an erasure if the user is OFF at that time. The ON-OFF functions are causal in the sense that the decision of whether the users are ON at time t depends on previous channel outputs, $\bar{Y}_k^{t-1}$. The stopping times T_k represent the time index of the last nonerasure channel output in the sequence $\bar{Y}_{k,t}$. For our applications, the stopping times T_k are less than or equal T with high probability. We merely define T_k to emphasize that T is a random variable which is not known by the users, and hence the users need to obtain this information through the sequence $\bar{Y}_{k,t}$. In a conventional system, control information in the initial packet defines the structure of the remaining transmission. Hence, after decoding the control information in the initial packet successfully, the user knows T_k and when to be ON and OFF to receive the message intended for that user.

The ON-OFF function also defines the average power consumption of the k-th user which we define by

$$P_k \triangleq \mathbb{E}\left[\sum_{i=1}^{T_k} \mathbb{1}\left\{g_{k,i}(\bar{Y}_k^{i-1}) = 1\right\}\right] \tag{6}$$

where $\mathbb{1}\{\text{condition}\}$ denotes the indicator function. Note that $\mathbb{E}[P_1] = \mathbb{E}[P_k]$, for $k \in \{1, \cdots, K\}$, since the message sizes D_k are distributed identically and $\gamma_k = \gamma$. Finally, the active users need to decode the their messages with reliability larger than or equal $1-\epsilon$ such that

$$\mathrm{P}\left[h_{k,T_k}(\bar{Y}_k^{T_k}) \neq \mathrm{M}_k | D_k > 0\right] \leq \epsilon \tag{7}$$

for $k \in \{1, \cdots, K\}$ and $\epsilon \in (0,1)$.

Our objective is to explore trade-offs between the competing goals of minimizing $\mathbb{E}[T]$ and $\mathbb{E}[P_1]$. We do this by investigating a class of feasible protocols.

III. Finite blocklength approximation

In the analysis of the proposed protocol, we apply recent results in finite blocklength information theory. Polyanskiy, Poor, and Verdú [3] showed that the maximal achievable coding rate of a code with fixed blocklength n and reliability $1-\epsilon' \in (0,1)$ over an AWGN channel is tightly approximated by

$$R^*(n, \epsilon') \approx C - \sqrt{\frac{V}{n}} Q^{-1}(\epsilon') + \frac{1}{2}\log_2 n \tag{8}$$

where the channel capacity C and the channel dispersion V are given by

$$C \triangleq \frac{1}{2}\log_2(1+P) \tag{9}$$

$$V \triangleq \frac{P(P+2)}{2(P+1)^2}\log_2(\exp(1))^2 \tag{10}$$

respectively. One can obtain tight upper and lower bounds for $R^*(n, \epsilon')$ using the achievability and converse bounds in [3]. In

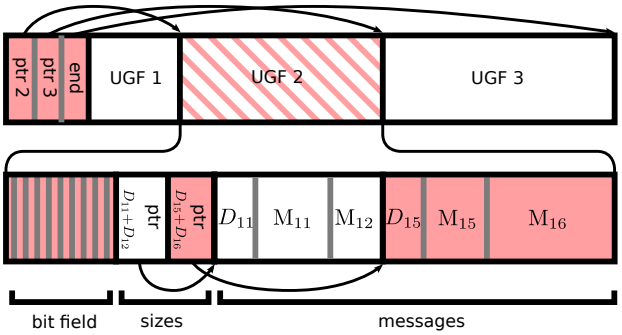

Fig. 2. Proposed protocol for $K = 30$, $L_B = 10$, and $L = 2$. In this case $\{11, 12, 15, 16\} \subseteq \mathcal{U}$ are among the active users. Grey separators means that data on both sides are encoded jointly. The red shaded regions correspond to the packets that the users 15 and 16 need to decode.

present paper, however, we resort to the approximation (8). An important property is that (8) is concave in n which implies that long packets are encoded more efficient than short packets.

We assume that a user is able to detect whether an error occurs during decoding of a packet. This assumption is suggested by the results in [4], and it is crucial to ensure the integrity of our protocols. In practical systems, one can use CRC checks to ensure integrity of packets with very high probability. We also assume that a user needs to receive all channel uses of a packet to allow for decoding. In other words, if k bits are encoded into n channel uses, the user needs to receive all n channel uses to decode any of the k bits.

In the design of our protocol, we rely on the approximation of $R^*(n, \epsilon')$. Specifically, the transmitter divides the total transmission time T into several smaller packets, each encoded separately at the maximal coding rate approximated by (8). We also define $N(k, \epsilon') \triangleq \min\{n \geq 0 : nR^*(n, \epsilon') \geq k\}$ for $k \geq 1$ and $N(0, \epsilon') \triangleq 0$, which is smallest number of channel uses that allows the encoding of k bits with reliability $1 - \epsilon'$.

One can easily obtain lower bounds on $\mathbb{E}[T]$ and $\mathbb{E}[P_k]$ by assuming the messages sizes are large:

$$\mathbb{E}[T] \geq \frac{1}{C}\mathbb{E}\left[\sum_{k=1}^{K} D_k\right] = \frac{K\alpha(1-q)(S+1)}{2C} \tag{11}$$

and

$$\mathbb{E}[P_k] \geq \frac{1}{C}\mathbb{E}[D_k] = \frac{\alpha(1-q)(S+1)}{2C}. \tag{12}$$

For sufficiently large α, the control information becomes negligible, and hence for the conventional approach both $\mathbb{E}[T]$ and $\mathbb{E}[P_k]$ simultanously approach the lower bounds in (11) and (12).

IV. Protocol design

There are various ways in which the messages $\{M_k\}$ can be conveyed to the respective users. Our approach is to design a protocol in which the transmitter forms multiple packets which are encoded separately. For each of these packets, we apply the finite blocklength approximation in (8) to find the optimal rate at which they can be encoded. We assume that the users are not provided with any control information such as the active users and $\{D_k\}$. Thus, the transmitter needs to encode packets about which users are active, the packet sizes of M_k, D_k, and the structure of the transmission. Clearly, this leaves us with a large space of feasible protocols. Here we introduce one class of protocols.

We first discuss what information, the transmitter needs to convey:

1) *Messages* $\{\mathrm{M}_k\}$: The message M_k only needs to be received by the k-th user, but as discussed previously, messages can be grouped and encoded jointly.
2) *Message sizes* $\{D_k\}$: The k-th user needs to know the message size D_k before attempting to decode the actual message M_k (otherwise, the user does not know how many channel uses the message M_k takes).
3) *Reciever activity* $\mathcal{U}$: It is necessary to convey whether the k-th user is active. In total, it requires K information bits to convey this information to all users.[1] As K information bits may represent a significant overhead, it may be beneficial to encode user activity bits in multiple packets such that each user only needs to decode one such packet.

In the proposed protocol, depicted in Fig. 2, users are grouped into $\lceil K/L_B \rceil$ *user groups* with at most L_B users in each user group. User acitivity, messages, and message sizes associated to each of these user groups are conveyed sequentially in *user group frames* (UGF). A transmission is initiated by a packet that jointly encodes the total transmission time (equivalent to the an end of transmission pointer) along with the $\lceil K/L_B \rceil - 1$ time indices that points to the time indices where the 2-th, 3-th, ..., and $\lceil K/L_B \rceil$-th UGF begin. This packet is transmitted with a reliability $1 - \epsilon_4$. The first UGF trivially begins after the initiating packet.

Let $\mathcal{K} \triangleq \{1, \cdots, K\}$ and let the users in the u-th user group be $\mathcal{K}_u \subseteq \mathcal{K}$. Then, the UGF for the u-th user group is constructed as follows. Initially, the transmitter divides the active users $\mathcal{U}_u \subseteq \mathcal{K}_u$ of the u-th user group into subgroups $\mathcal{U}_{u,i} \subseteq \mathcal{U}_u$, $i \in \{1, \cdots, \lceil |\mathcal{U}_u|/L \rceil\}$ of at most L users such that all groups except the last one always contains L users. The transmitter and users can agree on how to partition the users into subgroups for every set $\mathcal{U}_u$. The set of users $\mathcal{U}_{u,i} \subseteq \mathcal{U}_u$ is referred to as the i-th subgroup of the u-th user group. The main idea of our protocol is to jointly encode each of the subgroups.

A UGF consists of the following types of packets

1) *Bit field packet*: A bit field, encoding the the information $\{\mathbb{1}\{D_k = 0\}\}_{k \in \mathcal{K}_u}$. Hence, the packet consists of $|\mathcal{K}_u|$ information bits which are encoded with reliability ϵ_1.
2) *Size packets*: After grouping the active users of the u-th user group, $\mathcal{U}_u$, into $\lceil |\mathcal{U}_u|/L \rceil$ subgroups, the transmitter constructs a packet for each subgroup. For the i-th subgroup, the transmitter conveys a packet consisting of $\sum_{k \in \mathcal{U}_{u,i}} D_k$ along with a pointer to the packet that jointly encodes $\{\mathrm{M}_k\}_{k \in \mathcal{U}_{u,i}}$. Since $\sum_{k \in \mathcal{U}_{u,i}} D_k$ can take at most

[1] For the case $q \neq 1/2$, one can apply compression to reduce the number of information bits. This is, however, left for future work.

$L(S-1)+1$ distinct values, the size packet for the i-th subgroup needs to convey $\lceil \log_2(L(S-1)+1) \rceil + \text{ptr}$ information bits which are encoded with reliablity ϵ_2. Here, ptr denotes the number of bits needed to convey a pointer to a time index. The size packets are transmitted sequentially.

3) *Message packets*: Next, the transmitter encodes the messages of each subgroup, $\{\mathrm{M}_k\}_{k\in\mathcal{U}_{u,i}}$, along with the messages sizes of $|\mathcal{U}_{u,i}|-1$ of the messages. We only need $|\mathcal{U}_{u,i}|-1$, since the sum of the sizes, $\sum_{k\in\mathcal{U}_{u,i}} D_k$, is already successfully received in the size packet described above. This requires $(|\mathcal{U}_{u,i}|-1)\lceil \log_2 S \rceil + \sum_{k\in\mathcal{U}_{u,i}} D_k$ information bits. These information bits are encoded with reliability ϵ_3.

In order to decode the packet destined to user k, it needs to decode four packets successfully. If one or more of these packets are not successfully decoded, the user can not decode the packet containing the message destined to that user. Thus, the reliabilities need to be chosen such that $(1-\epsilon_1)(1-\epsilon_2)(1-\epsilon_3)(1-\epsilon_4)$ is kept above or equal to $1-\epsilon$ to fulfill the reliability constraint in (7). If the k-th user is inactive, it only needs to decode the initial packet containing pointers to the UGFs and the bit field packet. It thereby achieves a reliability of $(1-\epsilon_4)(1-\epsilon_1)$. We also point out that the described protocol reduces to a variant of the conventional protocol when $L=1$.

We remark that the protocol specified above is one class among a large space of feasible protocols. For large q is may be beneficial to use a different approach for conveying user activity. For example, one could encode the number of active users in an initial packet and encode an additional packet with the user identification numbers. Regarding the size packets, one can also encode all size packets jointly in each UGF jointly to enhance encoding efficiency at the expense of higher average power consumption at the users.

Assuming that L_B divides K, we may sum up the block-lengths of all the packets to obtain the average total transmission time of the protocol

$$\begin{aligned}\mathbb{E}[T_{L,L_B}] &= \frac{K}{L_B} N(L_B,\epsilon_1) + N\left(\frac{K}{L_B}\text{ptr},\epsilon_4\right) \\ &\quad + \frac{K}{L_B}\mathbb{E}\left[\sum_{i=1}^{\lceil |\mathcal{U}_1|/L\rceil} \left(N(\lceil \log_2(L(S-1)+1)\rceil + \text{ptr},\epsilon_2)\right.\right. \\ &\quad \left.\left. + N\left((|\mathcal{U}_{1,i}|-1)\lceil \log_2 S\rceil + \sum_{k\in\mathcal{U}_{1,i}} D_k,\epsilon_3\right)\right)\right]. \quad (13)\end{aligned}$$

Here, we have used that $|\mathcal{U}_i|$ are identically distributed for each i. For the average power, we have

$$\begin{aligned}\mathbb{E}[P_{L,L_B}] &= N(L_B,\epsilon_1) + N\left(\frac{K}{L_B}\text{ptr},\epsilon_4\right) \\ &+ \frac{1}{L_B}\mathbb{E}\left[\sum_{i=1}^{\lceil |\mathcal{U}_1|/L\rceil} |\mathcal{U}_{1,i}|\left(N(\lceil \log_2(L(S-1)+1)\rceil + \text{ptr},\epsilon_2)\right.\right. \\ &\quad \left.\left. + N\left((|\mathcal{U}_{1,i}|-1)\lceil \log_2 S\rceil + \sum_{k\in\mathcal{U}_{1,i}} D_k,\epsilon_3\right)\right)\right]. \quad (14)\end{aligned}$$

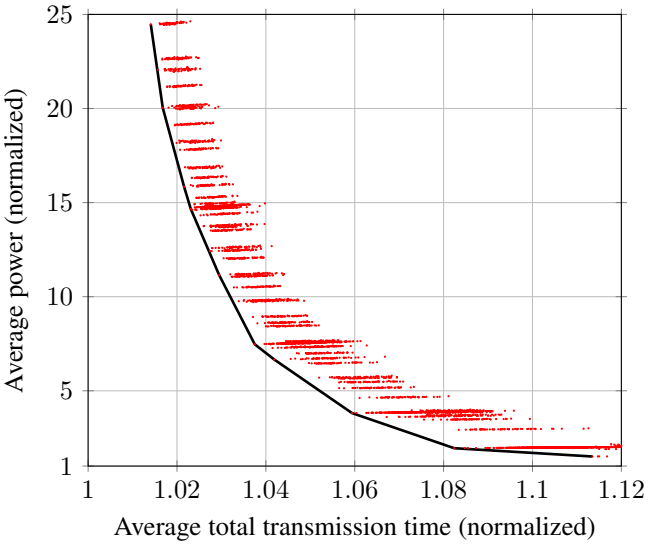

Fig. 3. Trade-off between average transmission time and average power for the case $K=64$, $P=1$, $q=0.5$, $\alpha=1000$, $\epsilon=10^{-3}$, and $S=2$. Red dots are simulation points.

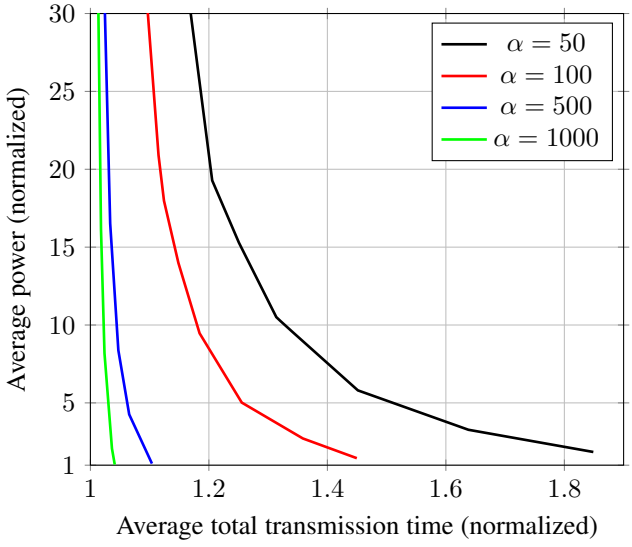

Fig. 4. Trade-off between average total transmission time and average power for the parameters are $K=128$, $P=1$, $q=0.5$, $\epsilon=10^{-3}$, and $S=4$.

The specified protocol leaves the parameters ϵ_1, ϵ_2, ϵ_3, ϵ_4, L, and L_B to be specified. Now, we can trace the optimal trade-off between T_{L,L_B} and P_{L,L_B} by solving the optimization problem

$$\min_{\substack{L,L_B,\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4:\\ \prod_{j=1}^4(1-\epsilon_j)\geq 1-\epsilon}} \mathbb{E}[T_{L,L_B}] + \beta\mathbb{E}[P_{L,L_B}]\,. \quad (15)$$

for a range of values of $\beta \geq 0$. The optimization problem is clearly not convex, and hence we find an approximate solution in the next section using a grid search for practical values of K, ϵ, q, and P_D.

V. Numerical results

In order to solve the optimization problem in (15), we compute $\mathbb{E}[T_{L,L_B}]$ and $\mathbb{E}[P_{L,L_B}]$ using 5000 Monte Carlo simulations of the protocol for $L \in \{1, 2, \cdots, K\}$ and L_B equal to all powers of two between 1 and K. We evaluate $\epsilon_1, \cdots, \epsilon_4$ over the four-dimensional grid $10 \times 10 \times 10 \times 10$ grid. The average total transmission time $\mathbb{E}[T_{L,L_B}]$ and average power $\mathbb{E}[P_{L,L_B}]$ are normalized according to the lower bounds in (11) and (12), respectively. The normalization implies that any simulation point must be in the square $[1, \infty) \times [1, \infty)$. The trade-off between average total transmission and average power is computed as the lower convex envelope of the simulation points. This is depicted in Fig. 3, where the simulations points are shown as red dots and the lower convex envelope is the black curve. For the computation, we use $\mathrm{ptr} = 16$ bits. Although the lower convex envelope is not directly achievable using our protocol, it can be achieved by time sharing between two sets of protocol parameters. Note that the lower-most point of the trade-off curve corresponds to the conventional extreme case where the messages of each user are encoded separately. The gap to 1 is thus due overhead from control information.

Our results are depicted in Fig. 4 for the parameters $K = 128$, $P = 1$, $q = 0.5$, $S = 4$, $\epsilon = 10^{-3}$, and $\alpha \in \{50, 100, 500, 1000\}$. We observe that one can reduce the average total transmission time by grouping users as proposed. Smaller values of α implies that messages are encoded less efficient, and hence grouping becomes an interesting option.

VI. Conclusions

In this paper, we have addressed the problem of downlink transmission of short packets to K users. Our main objective has been to highlight some of the challenges faced when the messages are small. Specifically, we used recent finite blocklength approximations to visualize the trade-offs between the average power of the each user and the average total transmission time seen from the transmitter. To show this trade-off, we have designed a practical protocol that groups messages and thereby achieves more efficient coding rates. The key element in the protocol design is the encoding of control information.

References

[1] P. Popovski, "Ultra-reliable communication in 5g wireless systems," in *International Conference on 5G for Ubiquitous Connectivity*, Nov. 2014, pp. 146–151.

[2] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultra-reliable, and low-latency wireless: The art of sending short packets," pp. 1–12, 2015.

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

[4] V. Y. F. Tan and P. Moulin, "Fixed error asymptotics for erasure and list decoding," *arXiv*, pp. 1–18, Feb. 2013. [Online]. Available: http://arxiv.org/abs/1301.7464v2