ENERGY MINIMIZATION OF MULTI-USER LATENCY-CONSTRAINED BINARY COMPUTATION OFFLOADING

Mahsa Salmani and Timothy N. Davidson

Department of Electrical and Computer Engineering
McMaster University, Hamilton, ON, Canada

ABSTRACT

Computation offloading expands the range of computationally-intensive and latency-constrained tasks that mobile users can execute. In a multi-user setting, the system selects the offloading users and allocates resources to them so that the overall energy consumption is minimized. We show herein that when the users have different latencies, the appropriate signalling architecture has a time-slotted structure with different subsets of the offloading users transmitting in each time slot. Furthermore, for multiple access schemes that exploit the full capabilities of the channel we analytically determine the optimal signalling architecture and we develop a highly-efficient algorithm for the power and rate allocations in each time slot. Our numerical results illustrate the advantages of the proposed system over those that employ a single-slot architecture and over time division multiple access.

1. INTRODUCTION

The opportunity of offloading computational tasks provided by Mobile Edge Computing resources not only expands the range of applications that mobile users can execute, it also enables them to reduce the energy and time they need to complete their tasks [1–3]. An efficient use of the offloading opportunity involves an effective allocation of the available communication resources to the offloading users [4–6]. Such an allocation depends on the multiple access scheme that is employed [7], and the nature of the users' computational tasks, e.g., [8]. Typically, the allocation problem seeks to minimize the energy that the users expend to complete their computational tasks while meeting the latency constraint on each task and the constraints on the available resources. Accordingly, this paper will focus on such a latency-constrained energy optimization problem in the case of binary offloading decisions; i.e., each user has an indivisible task that is either fully offloaded or completed locally.

In the energy minimization problems for multi-user offloading systems that have been addressed in existing work, e.g., [9–11], the users are constrained to offload their tasks over a single time slot. However, in general, the users' tasks have different description lengths and latency constraints. Accordingly, the users may finish offloading at different times. Removing the users that have completed their offloading from the set of offloading users reduces the interference imposed on the remaining users, and the newly available resources can be reallocated to the users that are still offloading.

In this paper, we exploit the differences between the latency constraints of the users by introducing a 'time-slotted" signalling structure in which different sets of users are offloading in different time slots. In this setting, the energy-minimizing offloading problem involves determining the users that should offload, the subset of the offloading users that should transmit in each time slot (and, implicitly, the number of time slots), the duration of each slot, and the power and rate allocations for each slot. Since we consider indivisible tasks, the offloading decisions are binary and the resource allocation problem has a combinatorial structure. However, those decisions admit a tree structure, and we propose a tailored pruned greedy search for the decisions; see Section 5. Each node in the tree corresponds to a "complete" offloading problem for a given set of offloading users. Given that that set of offloading users is further decomposed into subsets for each time slot, even the complete offloading problem initially appears to be formidable. However, we show in Sections 3 and 4 that for a multiple access scheme that exploits the full capabilities of the channel (FullMA) [12], the complete offloading problem can be decoupled, leading to simple closed-form expressions for the sets of offloading users in each slot and the optimal slot durations. In particular, when K_o users are offloading only K_o time slots are required. Furthermore, for each time slot we obtain a closed-form expression for the optimal power allocation in terms of the rates, and an efficient block coordinate descent algorithm that is guaranteed to provide a stationary solution for the rates. Our numerical results illustrate the advantages of the proposed system over systems that are constrained to operate in one time slot or are constrained to use time division multiple access (TDMA).

2. SYSTEM MODEL

Let us consider a K-user offloading system in which each user seeks to complete an indivisible computational task within its own specified deadline. The single-antenna users may offload their tasks to a single-antenna access point that is equipped with sufficiently large computational resources. The channel from each user to the access point is assumed to be constant over the duration of the transmission and known by the access point. We adopt a discrete-time signal model with symbol interval T_s, and if s_k denotes the transmitted signal from user k and h_k denotes the corresponding channel, the received signal at the access point at a given symbol instant can be written as $y = \sum_{k=1}^{K} h_k s_k + v$, where v is a Gaussian random variable with zero mean and variance σ^2. For later convenience we define $\alpha_k = \frac{|h_k|^2}{\sigma^2}$.

We consider a computational model in which the full description of a task must be received before execution begins, and the results are sent back to the k^{th} user when its task has been fully executed. Under this model, if t_{UP_k} denotes the time it takes for the k^{th} user to offload its task, t_{exe_k} denotes the time it takes for the access point to execute that task, and t_{DL_k} denotes the time it takes to send the result back to the user k, the latency constraint of the k^{th} user is

$$t_{\text{off}_k} = t_{\text{UP}_k} + t_{\text{exe}_k} + t_{\text{DL}_k} \leq L_k, \tag{1}$$

where L_k denotes the maximum allowable latency for user k. Un-

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant RGPIN-2015-06631.

der the assumption of sufficiently large computational resources in the access point, and sufficiently large communication resources for sending the results back to the users, we assume that $T_k = t_{\text{exe}_k} + t_{\text{DL}_k}$ is a constant for each user.

2.1. Single-time-slot Signalling Structure

To provide some context for the time-slotted signalling structure that will be developed in Section 3, we observe that minimal energy offloading problems have previously been addressed in a variety of different ways for the case in which the users are constrained to transmit over a single time slot, with a single rate and power being assigned to each user [9–11, 13]. To illustrate the approach in [13], let us consider the case of "complete" offloading, in which a subset $\mathcal{S}_o \subseteq \{1, 2, \ldots, K\}$ of the users has been selected for offloading. If we let B_k denote the description length of user k's task, ℓ_k denote the number of symbol instants (channel uses) over which user k transmits, R_k and P_k denote the rate and the power of that user (in units per channel use), and $\tilde{L}_k = L_k - T_k$, then the minimum energy complete offloading problem in [13] can be written as

$$\min_{\{P_k\},\{R_k\},\{\ell_k\}} \quad \textstyle\sum_{k\in\mathcal{S}_o} \ell_k P_k \tag{2a}$$

$$\text{s.t.} \quad \ell_k R_k = B_k,\ T_s\ell_k \le \tilde{L}_k,\ 0 \le P_k,\ \forall k \tag{2b}$$

$$\{R_k\}_{\mathcal{S}_o} \in \mathcal{R}_{\mathcal{S}_o}(\{P_k\}_{\mathcal{S}_o}), \tag{2c}$$

where $\mathcal{R}_{\mathcal{S}_o}(\cdot)$ is the achievable rate region for the chosen multiple access scheme. In [13] we obtained closed-form and quasi-closed-form expressions for the optimal solutions to (2) when a FullMA scheme and when the TDMA scheme are employed, respectively. In addition to its computational advantages, an advantage of the approach in [13] over those in [9–11] is that each user is able to exploit the full extent of its latency. However, the fact that a single rate and power are assigned to each user means that the system cannot take advantage of the reduction in interference that arises when a user completes its transmission. That observation motivates the following development of a time-slotted signalling structure.

3. TIME-SLOTTED SIGNALLING STRUCTURE

In a general multi-user offloading system, the users will have tasks with different latency constraints, and hence some may finish offloading before the others. In that case, the remaining (offloading) users would be able to reduce their energy consumption if we could design the signalling architecture so that they can take advantage of the reduction in interference when a user completes its offloading. To do so, we introduce a "time-slotted" signalling structure, in each time slot of which a different subset of the offloading users will transmit. If $K_o = |\mathcal{S}_o|$ users have been selected for offloading, then the generic complete offloading problem has $2^{K_o} - 1$ time slots.

In order to formulate the variant of the energy minimization problem in (2) for the time-slotted system, let $\mathcal{S}_i \subseteq \mathcal{S}_o$ denote the subset of users offloading in the i^{th} time slot, and let τ_i denote the length (in channel uses) of that time slot. If P_{ki} and R_{ki} denote the power and data rate (in units per channel use) for an offloading user, k, in time slot i, then the offloading energy consumption of user k is $\sum_i \tau_i P_{ki}$. We will let $\gamma_{ki} \in [0, 1]$ denote the fraction of the description of the k^{th} user's task that is offloaded in time slot i, and hence the number of bits that user k offloads in time slot i is $\gamma_{ki}B_k = \tau_i R_{ki}$, and for any offloading user $\sum_i \gamma_{ki}B_k = \sum_i \tau_i R_{ki} = B_k$. Finally, if we define $\bar{i}_k$ as the index of the last time slot in which user k is offloading, then the transmission time for each user, which is the summation of the lengths of the time slots during which the user still has fractions of bits to offload, can be written as $t_{\text{UP}_k} = T_s \sum_{i=1}^{\bar{i}_k} \tau_i$.

In this section and in Section 4, we will focus on the "complete" offloading problem for a given set of offloading users, $\mathcal{S}_o$. An algorithm for selecting $\mathcal{S}_o$ will be briefly described in Section 5. In this setting, the energy minimization problem is

$$\min_{\substack{\{R_{ki}\},\{P_{ki}\},\\ \{\tau_i\},\{\mathcal{S}_i\}}} \ \textstyle\sum_i \sum_{k\in\mathcal{S}_o} \tau_i P_{ki} \tag{3a}$$

$$\text{s.t.} \quad 0 \le \tau_i R_{ki} \le B_k, \quad \textstyle\sum_i \tau_i R_{ki} = B_k, \quad \forall k, i, \tag{3b}$$

$$T_s \textstyle\sum_{i=1}^{\bar{i}_k} \tau_i \le \tilde{L}_k, \quad \forall k, \tag{3c}$$

$$0 \le P_{ki}, \quad \forall k, i, \tag{3d}$$

$$\{R_{ki}\}_{\mathcal{S}_i} \in \mathcal{R}_{\mathcal{S}_i}\left(\{P_{ki}\}_{\mathcal{S}_i}\right), \quad \forall i, \tag{3e}$$

where, without loss of generality, we order the users in $\mathcal{S}_o$ so that $\tilde{L}_1 \le \tilde{L}_2 \le \cdots \le \tilde{L}_{K_o}$. In the problem in (3), there are $(2^{K_o} - 1)!$ choices for the set of transmitting subsets $\{\mathcal{S}_i\}$, and for each choice there are $\sum_{m=1}^{K_o} \binom{K_o}{m}(2m + 1)$ remaining design variables. That is a lot more than the $3K_o$ variables in (2). In the following section, we will show that for the full multiple access scheme we can achieve an optimal solution to (3) using only K_o time slots. We will also determine optimal offloading sets $\mathcal{S}_i$ and the optimal lengths for each time slot. That will reduce the number of design variables to $K_o(K_o + 1)$; see (8) below. In Section 4, we will provide closed-form solutions for the powers and consequently halve that number. We will subsequently construct a block coordinate descent algorithm for the rates that cyclically solves K_o problems with $(K_o - i + 1)$ variables, $i = 1, 2, \ldots, K_o$.

3.1. Optimized Time-slotted Signalling Structure for FullMA

To begin our reduction of the generic $2^{K_o} - 1$ time slots, we first observe that all the time slots in which user k transmits must occur before that user's latency, $\tilde{L}_k$. That is necessary for the problem to be feasible. Furthermore, from the perspective of user k, the time slots in which it transmits can be arranged in an arbitrary order. Therefore, one optimal ordering of the time slots groups all slots involving user 1 at the beginning (since we have ordered the users according to their latency constraints), and within that group orders the slots in non-increasing order of the number of transmitting users. We then group together the remaining time slots that involve user 2, and order them analogously. This procedure is continued until there is a single remaining time slot for user K_o. An example of this arrangement is illustrated for $K_o = 3$ in Fig. 1.

Although we have resolved the ordering of the time slots, there remain $2^{K_o} - 1$ of them in general, and hence a large number of rates, powers and lengths to design, as quantified above. We will now show that if the full multiple access scheme is used in each time slot we can restrict attention to K_o time slots, without loss of generality. To do so, we first observe that if we assume that the time slot lengths are long enough, the achievable rate region for a FullMA scheme approaches the capacity region. (Adaptations to the finite block length regime can be based on work in [14, 15].) If $\mathcal{N}_i$ denotes an arbitrary subset of $\mathcal{S}_i$, the capacity region corresponding to the i^{th} time slot is the region bounded by constraints of the form [12]

$$0 \le \textstyle\sum_{k\in\mathcal{N}_i} R_{ki} \le \log\big(1 + \sum_{k\in\mathcal{N}_i} \alpha_k P_{ki}\big), \quad \forall \mathcal{N}_i \subseteq \mathcal{S}_i. \tag{4}$$

In the $K_o = 2$ case, we were able to exploit the structure of this region to reduce the number of time slots for the FullMA case from the generic 3 to 2 [7]. In the following we will show that for $K_o = 3$, the time slots in Fig. 1 can be reduced to 3. The principles that we will use can be extended to construct a formal induction proof that for K_o offloading users, without loss of generality, we can restrict

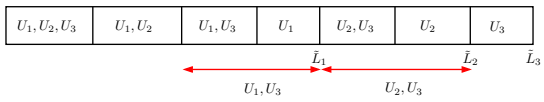

Fig. 1: Latency-sorted time-slotted structure for a 3-user offloading system.

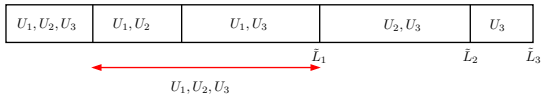

Fig. 2: Reduced time-slotted structure for a 3-user offloading system.

Fig. 3: Equivalent time slots in a 3-user FullMA offloading system.

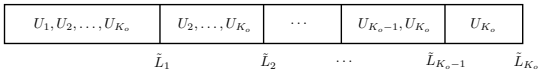

Fig. 4: Optimized time-slotted structure for a K_o-user FullMA system.

attention to the K_o time slots in Fig. 4. In addition to determining the sets $\mathcal{S}_i$, which we will index by $i = 1, 2, \ldots, K_o$ in the order in Fig. 4, that analysis also shows that the optimal time slot lengths are

$$\tau_i^\star = \tfrac{\tilde{L}_i - \tilde{L}_{i-1}}{T_s}, \quad \text{where } \tilde{L}_0 = 0. \tag{5}$$

Hence when all the users have the same latency, the optimized system has a single active time slot.

3.1.1. Optimized Time Slots for K_o=3

Our construction begins with the following observation from [7]: for $K_o = 2$, a three-time-slot FullMA system with the 5th, 6th and 7th time slots in Fig. 1 is equivalent to the two-time-slot system that consists of the 4th and 5th time slots in Fig. 2. This equivalence is in the sense that any energy consumption achieved by the former system can also be achieved by the latter. To extend that result to the three-user case we need to show that the two-time-slot system in Fig. 3a is equivalent to the single-time-slot system in Fig. 3b. To do so, we let $\tilde{R}_{ki}$, $\tilde{P}_{ki}$ denote the rate and power of user k in time slot $i \in \{a, b\}$ in Fig. 3a, respectively. These rates and powers satisfy the FullMA achievable rate region constraints. We also let R'_k and P'_k denote the rate and power of the k^{th} user in the single interval in Fig. 3b, respectively, where because there is only one time slot in Fig. 3b the index corresponding to the time slot is removed. In order for the two structures in Fig. 3 to be equivalent, the energy and the number of transmitted bits of each user should be equal, i.e.,

$$\begin{aligned}
\bar{\tau}_a\tilde{P}_{1a} + \bar{\tau}_b\tilde{P}_{1b} &= (\bar{\tau}_a + \bar{\tau}_b)P'_1 \\
\bar{\tau}_a\tilde{P}_{2a} = (\bar{\tau}_a + \bar{\tau}_b)P'_2, \quad \bar{\tau}_b\tilde{P}_{3b} &= (\bar{\tau}_a + \bar{\tau}_b)P'_3 \\
\bar{\tau}_a\tilde{R}_{1a} + \bar{\tau}_b\tilde{R}_{1b} &= (\bar{\tau}_a + \bar{\tau}_b)R'_1 \\
\bar{\tau}_a\tilde{R}_{2a} = (\bar{\tau}_a + \bar{\tau}_b)R'_2, \quad \bar{\tau}_b\tilde{R}_{3b} &= (\bar{\tau}_a + \bar{\tau}_b)R'_3
\end{aligned}$$

The solution of that set of linear equations is

$$P'_1 = \tfrac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\tilde{P}_{1a} + \tfrac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\tilde{P}_{1b} \tag{6a}$$

$$P'_2 = \tfrac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\tilde{P}_{2a}, \quad P'_3 = \tfrac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\tilde{P}_{3b} \tag{6b}$$

$$R'_1 = \tfrac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\tilde{R}_{1a} + \tfrac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\tilde{R}_{1b} \tag{6c}$$

$$R'_2 = \tfrac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\tilde{R}_{2a}, \quad R'_3 = \tfrac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\tilde{R}_{3b} \tag{6d}$$

What remains is to show that these powers and rates satisfy the rate region constraints for Fig. 3b. To do so, we rewrite those constraints in terms of the rates and powers of the two time slots in Fig. 3a. The constraint on the rate of the first user is then

$$\tfrac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\tilde{R}_{1a} + \tfrac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\tilde{R}_{1b} \le \log_2\big(1 + \tfrac{\alpha_1\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\tilde{P}_{1a} + \tfrac{\alpha_1\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\tilde{P}_{1b}\big). \tag{7}$$

Since the rates of the first user in the two time slots in Fig. 3a satisfy the rate region constraints, using the concavity of the logarithm we can show that the inequality in (7) holds. All other rate region constraints in Fig. 3b can be established in an analogous way.

4. COMPLETE COMPUTATION OFFLOADING

We have shown that for a K_o-user FullMA complete offloading system the time-slotted structure illustrated in Fig. 4 is an optimal structure. Accordingly, the energy minimization problem in (3) for a system that employs a FullMA scheme can be written as

$$\min_{\{R_{ki}\},\{P_{ki}\}} \ \textstyle\sum_{k=1}^{K_o}\sum_{i=1}^{k} \tau_i^\star P_{ki} \tag{8a}$$

$$\text{s.t.} \quad 0 \le \tau_i^\star R_{ki} \le B_k, \quad \forall k, i, \tag{8b}$$

$$\textstyle\sum_{i=1}^{k} \tau_i^\star R_{ki} = B_k, \quad \forall k, \tag{8c}$$

$$\textstyle\sum_{k\in\mathcal{N}_i} R_{ki} \le \log_2\big(1 + \sum_{k\in\mathcal{N}_i} \alpha_k P_{ki}\big), \forall \mathcal{N}_i \subseteq \mathcal{S}_i, \tag{8d}$$

where the constraints $0 \le P_k$ have been omitted because they are implicit in the combination of (8b) and (8d).

4.1. Closed-form Solutions for Transmission Powers

To solve (8) we decompose the problem into an inner problem over the powers and an outer problem over the rates:

$$\begin{array}{ll} \min\limits_{\{R_{ki}\}} & \min\limits_{\{P_{ki}\}} \ (8a) \\ \text{s.t. } (8b), (8c) & \text{s.t. } (8d). \end{array} \tag{9}$$

The inner optimization problem can be written as

$$\min_{\{P_{ki}\}} \ \textstyle\sum_{k=1}^{K_o}\sum_{i=1}^{k} \tau_i^\star P_{ki} \tag{10a}$$

$$\text{s.t.} \ \textstyle\sum_{k\in\mathcal{N}_i} R_{ki} \le \log_2\big(1 + \sum_{k\in\mathcal{N}_i} \alpha_k P_{ki}\big), \forall \mathcal{N}_i \subseteq \mathcal{S}_i. \tag{10b}$$

It can be seen from the constraints in (10b) that, for a given set of rates, the power of each user in each time slot depends only on the rates of the users that are offloading in that specific time slot. Hence, for a given set of rates, the problem in (10) can be partitioned into K_o decoupled problems over the K_o separate time slots in Fig. 4. The problem of finding the optimal transmission powers in each time slot has similar structure to the problem studied in our previous work [13, 16] for a single-time-slot K-user offloading system. Accordingly, we can obtain closed-form optimal solutions for the powers in each time slot for a given set of transmission rates in that time slot. To do so, the offloading users in each time slot are sorted according to the channel gains, $\rho_k = \frac{1}{\alpha_k}$, using the permutation π that ensures that $\rho_{\pi(K)} \le \rho_{\pi(K-1)} \le \cdots \le \rho_{\pi(1)}$. We have shown in [16] that the closed-form optimal solutions for the power of each user in terms of the rates of the other users in that time slot can be obtained as

$$P_{\pi(k)i} = \big(\tfrac{2^{R_{\pi(k)i}}-1}{\alpha_{\pi(k)}}\big)2^{\sum_{j=1}^{k-1} R_{\pi(j)i}}. \tag{11}$$

4.2. Optimal Solutions for Transmission Rates

Having derived the closed-form solutions for the optimal powers in (11), the outer problem in (9) can be written as

$$\min_{\{R_{\pi(k)i}\}} \quad \sum_{k=1}^{K_o}\sum_{i=1}^{k}\tau_i^\star\left(\frac{2^{R_{\pi(k)i}}-1}{\alpha_{\pi(k)}}\right)2^{\sum_{j=1}^{\pi(k-1)}R_{\pi(j)i}} \tag{12a}$$

$$\text{s.t.} \quad 0 \le \tau_i^\star R_{\pi(k)i} \le B_k, \quad \forall k,i \tag{12b}$$

$$\sum_{i=1}^{k}\tau_i^\star R_{\pi(k)i} = B_k, \quad \forall k. \tag{12c}$$

It can be seen from (12) that by finding the closed-form solutions for the powers in terms of the rates by exploiting the structure of the achievable rate region, the rates of each user have been decoupled from the rates of the other users. Indeed, if we fix the set of rates of users $\ell \neq k$ and define

$$A_{\pi(k)i} = \frac{2^{\sum_{j=1}^{\pi(k-1)}R_{\pi(j)i}}}{\alpha_{\pi(k)}} + \sum_{j=k+1}^{K} 2^{\sum_{\substack{m=1,\\ m\neq k}}^{j} R_{\pi(m)i}}\left(\frac{2^{R_{\pi(j)i}}-1}{\alpha_{\pi(j)}}\right), \tag{13}$$

the transmission rates of user k in its offloading time slots can be obtained by solving the following problem

$$\min_{\{R_{\pi(k)i}\}} \quad \sum_{i=1}^{k}\tau_i^\star A_{\pi(k)i}\,2^{R_{\pi(k)i}} \tag{14a}$$

$$\text{s.t.} \quad 0 \le \tau_i^\star R_{\pi(k)i} \le B_k, \quad \forall i, \tag{14b}$$

$$\sum_{i=1}^{k}\tau_i^\star R_{\pi(k)i} = B_k. \tag{14c}$$

This problem is convex and it can be efficiently solved by applying a generalized water-filling approach; cf. [17]. Accordingly, by employing an iterative block coordinate descent algorithm, a stationary solution (cf. [18]) for the sets of rates for all the users can be achieved. Finally, by substituting the obtained solutions for the rates we can obtain the optimal values for the powers using (11).

5. BINARY COMPUTATION OFFLOADING

Now that we have obtained a stationary solution for the energy minimization problem for a given set of offloading users in the case of a FullMA scheme, we can apply a tree-search strategy to find an appropriate set of offloading users. While a wide variety of tree-search algorithms are available, in [13] we have developed a customized pruned greedy search algorithm to solve the binary offloading problem for a single-time-slot binary offloading system. That customized algorithm can also be applied in the time-slotted case.

6. NUMERICAL RESULTS

We now illustrate the performance of the proposed time-slotted FullMA system with the customized greedy search algorithm. We will compare the total energy consumption to that of the single-time-slot systems in [9, 13], and to a TDMA-based system that is optimized in an analogous way [13]. We will consider a cell of radius 1,000m over which the users are uniformly distributed. We consider a slow-fading channel model with a path-loss exponent of 3 and independent Rayleigh distributed small-scale fading. The receiver noise variance is set to $\sigma^2 = 10^{-13}$. The energy consumption in each experiment is averaged over 100 channel realizations. The symbol interval is $T_s = 10^{-6}$s, and $T_k = 0.2$s for all users.

In our first experiment we consider a four-user system with different latencies, $[L_1, L_2, L_3, L_4] = [1.2, 1.4, 2.4, 2.7]$s, and we examine the energy consumption as the (different) description lengths of the tasks grow (in proportion); $[B_1, B_2, B_3, B_4] = \zeta \times [2, 1, 3, 4] \times 10^6$ bits. We apply the customized greedy algorithm (see [13]) to find a good set of offloading users, and the algorithms in Section 4 and [13] to find corresponding power and rate allocations. The first observation from the results in Fig. 5 is that the proposed time-slotted structure for the FullMA scheme can significantly reduce the energy consumption over the single-time-slot case. It can also be seen that the greedy search algorithm in Section 5 finds a good set of offloading users, as it did for the single-time-slot case in [13]. Furthermore, Fig. 5 shows that using the full capabilities of the channel enables a significant reduction in the energy consumption over the TDMA scheme.

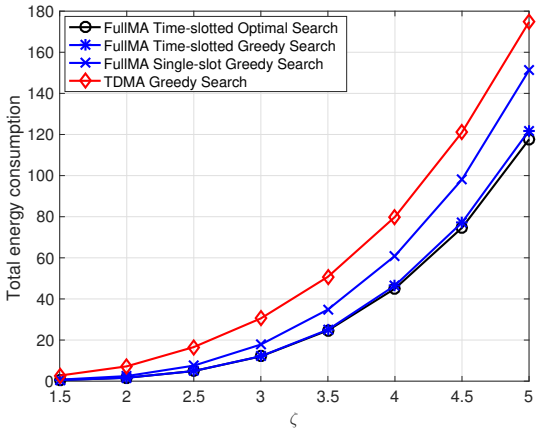

Fig. 5: Average energy consumption of a four-user binary offloading system with different latency constraints as the description lengths increase.

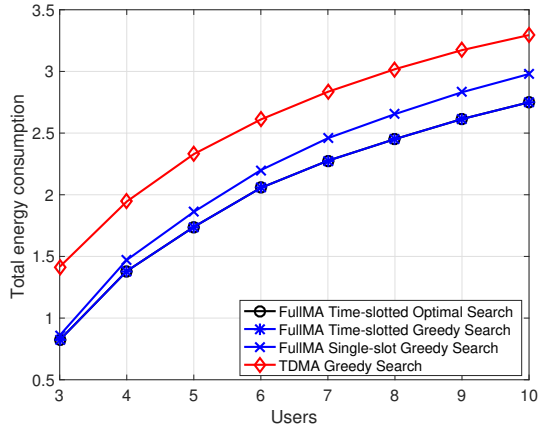

Fig. 6: Average energy consumption of a binary offloading system for different number of users with different latencies.

In our second experiment, we examine the total energy consumption as the number of users increases. The description lengths of the tasks are $B_k = 5 \times 10^6$ bits, and the latency constraint of user k is $L_k = 1 + 0.7k$ (s). The significant reduction in the energy consumption of the proposed time-slotted FullMA scheme compared to the single-time-slot FullMA scheme and the TDMA scheme is once again apparent from Fig. 6. Furthermore, the customized greedy search algorithm obtains a close-to-optimal user selection even as the number of users grows.

7. REFERENCES

[1] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *IEEE Computer*, vol. 43, no. 4, pp. 51–56, 2010.

[2] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Networks and Applications*, vol. 18, no. 1, pp. 129–140, 2013.

[3] M. Barbera, S. Kosta, A. Mei, and J. Stefa, "To offload or not to offload? The bandwidth and energy costs of mobile cloud computing," in *Proc. IEEE Intl. Conf. Comput. Commun.*, Turin, Apr. 2013, pp. 1285–1293.

[4] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Info. Process. over Network*, vol. 1, no. 2, pp. 89–103, 2015.

[5] O. Munoz, A. Pascual-Iserte, and J. Vidal, "Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4738–4755, 2015.

[6] M.-H. Chen, B. Liang, and M. Dong, "Resource sharing of a computing access point for multi-user mobile cloud offloading with delay constraints," to appear in *IEEE Trans. Mobile Comput.*, also available: https://arxiv.org/abs/1712.00030.

[7] M. Salmani and T. N. Davidson, "Multiple access computational offloading: Communication resource allocation in the two-user case (extended version)," 2018. [Online]. Available: https://arxiv.org/abs/1805.04981v2

[8] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, 2016.

[9] F. Wang, J. Xu, and Z. Ding, "Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems," 2018. [Online]. Available: https://arxiv.org/abs/1707.02486v3

[10] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, 2018.

[11] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, 2017.

[12] A. El-Gamal and T. M. Cover, "Multiple user information theory," *Proc. IEEE*, vol. 68, no. 12, pp. 1466–1483, 1980.

[13] M. Salmani and T. N. Davidson, "Uplink resource allocation for multiple access computational offloading," 2018. [Online]. Available: https://arxiv.org/abs/1809.07453

[14] E. MolavianJazi and J. N. Laneman, "Simpler achievable rate regions for multiaccess with finite blocklength," in *Proc. IEEE Int. Symp. Information Theory*, 2012, pp. 36–40.

[15] ——, "A finite-blocklength perspective on Gaussian multi-access channels," 2013. [Online]. Available: https://arxiv.org/abs/1309.2343v1

[16] M. Salmani and T. N. Davidson, "Multiple access binary computational offloading in the K-user case," May 2018, to appear in the *Conf. Rec. 52nd Asilomar Conf. Signals, Syst. Comput.*

[17] C. Xing, Y. Jing, S. Wang, S. Ma, and H. V. Poor, "New viewpoint and algorithms for water-filling solutions in wireless communications," 2018. [Online]. Available: https://arxiv.org/abs/1808.01707

[18] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, 2016.