This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



# Delay-Minimization Nonorthogonal Multiple Access enabled Multi-User Mobile Edge Computation Offloading

Yuan Wu *Senior Member IEEE*, Li Ping Qian *Senior Member IEEE*, Kejie Ni, Cheng Zhang, Xuemin (Sherman) Shen *Fellow IEEE*

***Abstract*—The significant advances of cellular systems and mobile Internet services have yielded a variety of computation intensive applications, resulting in great challenge to mobile terminals (MTs) with limited computation resources. Mobile edge computing (MEC), which enables MTs to offload their computation tasks to edge servers located at cellular base stations (BSs), has provided a promising approach to address this challenging issue. Considering the advantage of improving transmission efficiency provided by Nonorthogonal Multiple Access (NOMA), we propose an NOMA-enabled computation offloading scheme, in which a group of MTs offload partial of their computation-workloads to an edge server based on the NOMA-transmission. After finishing all MTs' offloaded computation-workloads, the edge server sends the computation-results back to the MTs based on NOMA. We aim at minimizing the overall delay for completing all MTs' computation requirements, which is achieved by jointly optimizing the MTs' offloaded computation-workloads, and the uploading-duration for the MTs to send their computation-workloads to the BS, and the downloading-duration for the BS to send the computation-results back to the MTs. Despite the nonconvexity of the joint optimization problem, we exploit its layered structure and propose an efficient algorithm to compute the optimal offloading solution. Numerical results are provided to validate the accuracy and efficiency of our proposed algorithm and show the performance advantage of our NOMA-enabled computation-offloading scheme.**

## I. Introduction

The past decades have witnessed a significant development of cellular systems and mobile Internet services, which have raised a variety of mobile applications requiring intensive computation resources (e.g., real-time interactive online gaming and augmented/virtual reality). However, due to the limited computation resources, nowadays mobile terminals suffer from a constrained computational capability, which degrades users' quality of experience when executing the computation-hungry applications. The emerging paradigm of mobile edge computing (MEC), which enables the mobile terminals (MTs) to offload their computation-workloads to edge servers deployed at cellular base stations (BSs), has provided a promising approach to address this challenging issue [1], [2]. The advantage of MEC lies in migrating the intensive computation-workloads from the MTs to nearby edge servers equipped with sufficient computation resources, which thus improves the computation efficiency and reduces the computation-delay. Despite its advantage, MEC invokes the data transmission between the MT and edge server over wireless links. As a result, additional radio resource consumptions are required. Due to the constrained amount of the radio resources available at the MTs and BS, the advantage of MEC might be adversely influenced without a proper management. For instance, when a lot of MTs aggressively offload their computation tasks to the BS[1] (via a common channel), a severe congestion will occur on this channel, which results in a long transmission-delay for the MTs to offload their computation tasks to the BS and increases the overall delay for executing the computation-offloading. Thus, a joint management of the computation-offloading and the resource allocation is required to exploit the benefit of MEC [7]–[24].

Recently, non-orthogonal multiple access (NOMA), which enables a group of MTs to share a same spectrum channel for simultaneous transmissions and exploits the successive interference cancellation (SIC) to mitigate the MTs' co-channel interference, has been considered as a promising approach to improve the spectrum efficiency in cellular radio access networks. Compared with the conventional orthogonal multiple access (OMA), NOMA is able to achieve multi-folded advantages such as improving the throughput and spectrum efficiency and accommodating massive connectivity. As a result, NOMA has attracted lots of research interests [25]–[27]. Thanks to the advantages of NOMA, exploiting NOMA for MEC reduces the delay in data transmissions between the MTs and edge servers and thus reduces the overall delay in computation-offloading.

Nevertheless, to achieve the aforementioned advantage, we need a joint management of the MTs' computation-offloading and the associated radio resource allocation for the NOMA-transmission (including both the MTs' NOMA-transmission and the BS's NOMA-transmission). Specifically, to exploit

Y. Wu, L. Qian, K. Ni, and C. Zhang are with College of Information Engineering, Zhejiang University of Technology, Hangzhou, China (emails:iewuy@zjut.edu.cn, lpqian@zjut.edu.cn, kjni_zjut@163.com, czhang_zjut@163.com). L. Qian is the corresponding author. L. Qian is also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, 210096, China.

X. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail:xshen@bbcr.uwaterloo.ca).

This work was supported in part by the National Natural Science Foundation of China under Grant 61572440, in part by the Zhejiang Provincial Natural Science Foundation of China under Grants LR17F010002 and LR16F010003, in part by the open research fund of National Mobile Communications Research Laboratory, Southeast University (No. 2019D11), and in part by the Natural Sciences and Engineering Research Council, Canada.

[1]In this work, we assume that the edge server is co-located at the BS. Thus, we treat the BS and the edge server interchangeable.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



MEC and reduce the transmission delay, we intend to offload more computation-workloads to the BS as well as set a small uploading-duration (for the MTs to send their computation-workloads to the BS) and a small downloading-duration (for the BS to send back the computation-results). However, such choices require large transmission-rates for the MTs and the BS. Due to the NOMA's nature that allows the MTs' co-channel simultaneous transmissions, the MTs and the BS have to use large transmit-power to support the required large transmission-rates, which might quickly drain out their energy resources. Therefore, how to properly exploit the NOMA for the multi-MT computation-offloading via MEC under the given energy consumption motivates our study.

In this paper, we study the NOMA-enabled multi-MT computation-offloading. We assume that a group of MTs run the application of data-compression as [17] (e.g., compressing a large video-clip) and exploit the MEC to reduce the overall delay for completing all MTs' computation requirements. Similar to [17]–[22], we use the partial-offloading, which allows the MT to partition its total workload into two parts, with one part processed locally and the other offloaded to the edge server. The MTs form an NOMA-cluster and send their offloaded computation-workloads to the BS via the NOMA. After receiving and completing the MTs' offloaded computation-workloads, the BS simultaneously sends the computation-results back to the respective MTs via the NOMA-transmission. To measure the overall delay, we need to jointly consider the MTs' local computation-delay, the computation-delay at the edge server, as well as the round trip transmission-delay between the MTs and the BS (comprised of the transmission-delay for the MTs to send their computation-workloads to the BS and the transmission-delay for the BS to send the computation-results back to the MTs). Our detailed contributions are summarized as follows.

- By exploiting NOMA, we characterize the connection among the MTs' offloaded computation-workloads, the uploading-duration, and the MTs' minimum required transmit-power (in the uplink), as well as the connection among the MTs' computation-results, the downloading-duration, and the BS's minimum transmit-power for sending back the computation-results (in the downlink). By using these two analytical connections, we formulate a joint optimization of the MTs' offloaded workloads, the uploading-duration, and the downloading-duration, with the objective of minimizing the overall delay for completing all MTs' computation requirements, while meeting the MTs' and the BS's energy consumption constraints.
- Despite the non-convexity of the joint optimization problem, we propose an efficient algorithm to compute the optimal offloading solution. We decompose the joint optimization problem into two subproblems that respectively minimize the MTs' uploading-delay and the BS's downloading-delay under the given decision on the MTs' offloaded workloads, and a top-problem that further optimizes the MTs' offloaded workload based on the optimal solutions of the two subproblems. We propose efficient algorithms to solve the two subproblems and the top-problem, respectively, which together find the optimal offloading solution.
- We provide extensive numerical results to validate our proposed algorithms and the proposed NOMA-enabled computation-offloading scheme. We firstly show the accuracy and efficiency of our algorithms, by comparing with the commercial optimization package (i.e., LINGO [48]). We then demonstrate the performance advantage of our NOMA-enabled computation-offloading, by comparing with the time division multiple access (TDMA) based computation-offloading and some fixed computation-offloading method.

The remainder of this paper is organized as follows. We review the related studies in Section II. We present the system model and problem formulation in Section III. We propose efficient algorithms to solve the formulated joint optimization problem in Section IV and present the numerical results in Section V. Finally, we conclude this work and discuss the future directions in Section VI.

## II. Related Studies

We firstly review the related studies about the joint computation and communication resource allocations for multi-MT computation-offloading. In general, these studies can be separated into two main groups. The first group of the related studies focus on the binary computation-offloading in which each MT either offloads its entire computation task to the edge server or executes the computation task locally [7]–[16]. The second group of the related studies focus on the partial computation-offloading in which each MT can offload partial of its computation-workload to the edge server [17]–[24]. Our work belongs to the second stream, and we review these related studies as follows. In [17], Ren *et. al.* investigated the latency-minimization problem for a multi-user TDMA MEC system, by jointly optimizing the computation and communication resource allocations. In [18], You *et. al.* investigated the resource allocation for a multi-user MEC system based on the TDMA and orthogonal frequency-division multiple access (OFDMA), with the objective of minimizing the weighted sum of users' energy consumptions. In [19], Wang *et. al.* proposed a partial computation-offloading that incorporates the dynamic voltage scaling. In [20], Cao *et. al.* proposed a joint computation-and-communication optimization scheme for a cooperative three-nodes model. In [21], a multi-access based computation offloading scheme has been studied. In [22], [23], different incentive-based mechanisms have been exploited for investigating the joint allocation of computation resource and communication resource in MEC. In [24], Chang *et. al.* proposed an energy-efficient computation offloading scheme for a multi-user system, which considers the energy consumption of the computation-offloading and the delay constraint.

We next review the related studies about NOMA and the NOMA-enabled computation offloading.

*(Studies about NOMA):* The advantages of NOMA have attracted lots of research interests in recent years [25]–[27].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



Through analytical study and performance evaluation for different scenarios, the benefits of NOMA against OMA have demonstrated in many studies [28]–[31]. Since our study is closely related to resource (e.g., power and time) management for NOMA, we mainly review those related studies as follows. Power allocation strategies for NOMA have been deeply studies in [32], [33] for achieving different system-wise objectives. As a crucial issue to NOMA, User-pairing strategies have been studied in [34], [35]. In [36] and [37], different joint optimization schemes of the users' transmit-power and sub-channel allocation have been proposed. In [38], Elbamby *et. al.* proposed a joint optimization of the uplink and downlink resource optimization, mode selection, and power allocation for the NOMA systems with in-band full-duplex BSs. In [39], Qian *et. al.* proposed a joint optimization of the cell association and power control for NOMA small-cell networks. In [40], Bai *et. al.* proposed a novel hybrid transmission method by using the NOMA transmission and the multi-user diversity. Resource allocations for exploiting NOMA in relay networks have been investigated in [41], [42]. Several recent studies focused on the exploitation of simultaneous wireless information and power transfer for NOMA [43], [44].

*(Studies about NOMA-enabled computation-offloading):* In [3], Ding *et. al.* developed the analytical results to demonstrate that NOMA can reduce the latency and energy consumption of MEC offloading. In [4], Kiani *et. al.* proposed an edge-computing aware NOMA technique which exploits NOMA to reduce the uplink energy consumptions. The authors formulated an NOMA-based optimization framework which minimizes the users' energy consumptions by optimizing the user-clustering and the computation and communication resource allocations. In [5], Wang *et. al.* studied a multi-user MEC system that exploits the multi-antenna NOMA transmission for offloading. A joint optimization was proposed to minimize the total energy consumption of all users subject to the computation latency constraint. In [6], Ding *et. al.* proposed a novel hybrid NOMA-transmission model for two users' computation offloading via the NOMA transmission.

## III. System Model and Proposed NOMA-enabled Computation Offloading

As shown in Figure 1, we consider a group of MTs $\mathcal{I} = \{1, 2, ..., I\}$, with each MT $i$ having a total file size of $S_i^{\text{tot}}$ (in bits) to be compressed. In the following, we refer $S_i^{\text{tot}}$ as MT $i$'s total computation requirement. Each MT $i$ offloads partial of its computation-workload, denoted by $s_i^{\text{up}} \in [0, S_i^{\text{tot}}]$, to the edge server (which is co-located at the BS) and executes the remainder of its workload $S_i^{\text{tot}} - s_i^{\text{up}}$ locally. Figure 1(a) shows that the MTs form an NOMA-cluster to send their respective offloaded workloads to the BS (i.e., the uplink transmission), and Figure 1(b) shows that the BS uses NOMA to send the computation-results back to the respective MTs (i.e., the downlink transmission). We denote MT $i$'s local computation-rate (i.e., the CPU compression-rate in the unit of bits/second)[2] as $V_i^{\text{loc}}$ and denote the computation-rate of the edge server as $V^{\text{ser}}$.

We use $t^{\text{up}}$ to denote the allocated uplink-duration for the MTs to simultaneously upload their $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ to the BS and $t^{\text{do}}$ to denote the allocated downloading-duration for the BS to send back the computation-results. In this work, we consider that the MTs form an NOMA-cluster to send the offloaded computation-workloads to the BS simultaneously and further receive the computation-results from the BS. Thus, to facilitate our analysis of each MT's overall delay, we assume that the BS uses the same uplink transmission-duration $t^{\text{up}}$ for all MTs to offload their computation-workloads, meaning that different MTs have different uplink NOMA transmission-rates due to their different $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$. Similarly, we assume that the BS uses the same downlink transmission-duration $t^{\text{do}}$ to send the computation-results back to the respective MTs, meaning that different MTs have different downlink NOMA transmission-rates. We notice that, in a very recent work [6], Ding *et. al.* proposed a hybrid NOMA-transmission model in which two users (in an NOMA-cluster) may have different transmission-durations to complete their respective computation offloading. This provides us a very interesting direction to further extend our current model in this paper.

In the next two subsections, we will characterize the MTs' transmit-power and the BS's transmit-power for the NOMA-transmission.

### A. MTs' NOMA-transmissions to the BS

As shown in Figure 1(a), the MTs form an NOMA-cluster to offload their computation-workloads $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ to the BS, i.e., the scenario of uplink NOMA transmission. According to [45], in the uplink NOMA, an arbitrary decoding order can be used for SIC. For the sake of easy presentation, we assume that in the uplink, the BS uses the decoding order according to the indices in $\mathcal{I}$, namely, decoding MT $I$, MT $I-1$, ..., MT 2, and MT 1 sequentially (notice that our following analysis and design are applicable to an arbitrary uplink decoding order by changing the MTs' indices in $\mathcal{I}$). We use $p_i$ to denote MT $i$'s transmit-power to the BS. Based on the above considered decoding order, MT $i$'s uploading-rate to the BS is given by:

$$R_{i\text{B}}^{\text{up}} = W^{\text{up}} \log_2 \left( 1 + \frac{p_i g_{i\text{B}}}{\sum_{j=1}^{i-1} p_j g_{j\text{B}} + W^{\text{up}} n_0} \right), \forall i \in \mathcal{I}, \quad (1)$$

where $g_{i\text{B}}$ denotes the uplink channel power gain from MT $i$ to the BS, and $W^{\text{up}}$ denotes the uplink channel-bandwidth, and $n_0$ denotes the spectral power density of the background noise.

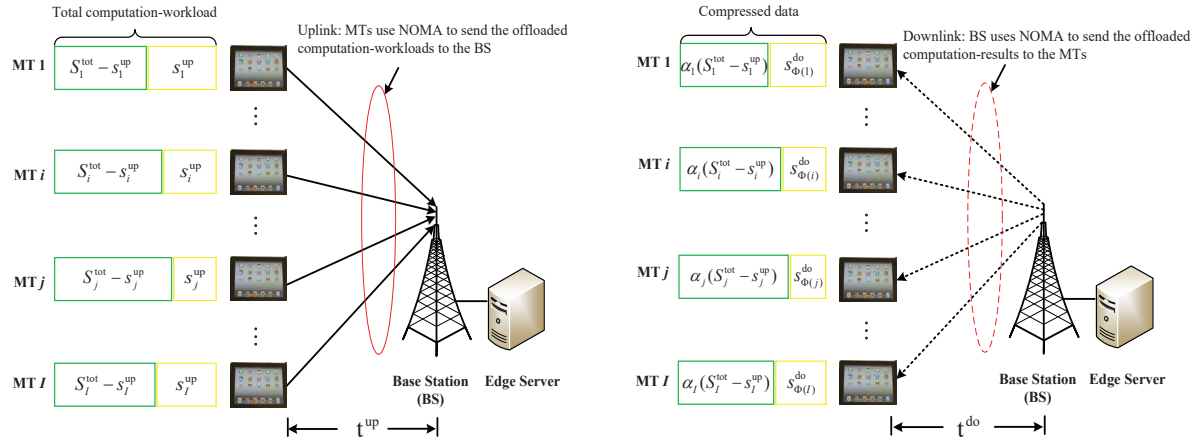
Based on (1), we next quantify the minimum transmit-power required by each MT to transmit to the BS. The result is given in the following proposition.

**Proposition 1:** Given the MTs' offloaded computation-workloads $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ and the uploading-duration $t^{\text{up}}$, each MT

[2]For the sake of clear presentation, we express MT $i$'s computation rate $V_i^{\text{loc}}$ in the unit of bits/second in this work. Notice that $V_i^{\text{loc}}$ can be calculated as $V_i^{\text{loc}} = f_i / C_i^{\text{loc}}$, where $f_i$ denotes MT $i$'s CPU frequency in Hz (i.e., cycles/second), and $C_i^{\text{loc}}$ denotes the consumed CPU cycles for each bit, i.e., in the unit of cycles/bit.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing






(a) MTs' transmission to the BS

(b) BS's transmission to the MTs

Fig. 1: System Model. (a) Uplink: The MTs form an NOMA-cluster to send their respective offloaded computation-workloads to the BS. (b) Downlink: The BS uses NOMA to send the computation-results back to the respective MTs. We explain $\{\Phi(i)\}$ (which is used for indexing the MTs in the downlink) in Subsection III-B.

$i$'s minimum transmit-power can be written as

$$p_i^{\min} = \frac{W^{\text{up}} n_0}{g_{i\text{B}}} \left( 2^{\frac{s_i^{\text{up}}}{t^{\text{up}}} \frac{1}{W^{\text{up}}}} - 1 \right) 2^{\frac{1}{t^{\text{up}}} \frac{1}{W^{\text{up}}} \sum_{j=1}^{i-1} s_j^{\text{up}}}, \forall i \in \mathcal{I}. \quad (2)$$

*Proof:* According to Proposition 2 [42], MT $i$'s minimum transmit-power is written as[3]

$$p_i^{\min} = \frac{W^{\text{up}} n_0}{g_{i\text{B}}} \gamma_{i\text{B}} \prod_{j=1}^{i-1} (1 + \gamma_{j\text{B}}), \forall i \in \mathcal{I}, \quad (3)$$

in which $\gamma_{i\text{B}} = \frac{p_i g_{i\text{B}}}{\sum_{j=1}^{i-1} p_j g_{j\text{B}} + W^{\text{up}} n_0}$ denotes the received signal to interference plus noise ratio (SINR) for MT $i$'s transmission to the BS. Furthermore, given MT $i$'s offloaded computation-workload $s_i^{\text{up}}$ and the uploading-duration $t^{\text{up}}$, we can derive MT $i$'s required SINR $\gamma_{i\text{B}}$ as

$$\gamma_{i\text{B}} = 2^{\frac{s_i^{\text{up}}}{t^{\text{up}}} \frac{1}{W^{\text{up}}}} - 1, \forall i \in \mathcal{I}. \quad (4)$$

By substituting (4) into (3), we then obtain (2). ■

Based on Proposition 1, we denote each MT $i$'s minimum required transmit-power as a function of the upload-duration $t^{\text{up}}$ and the uploaded computation-workloads $\{s_i^{\text{up}}\}_{i \in \mathcal{I}}$ as follows:

$$p_i^{\min}(t^{\text{up}}, \{s_j^{\text{up}}\}_{j \in \mathcal{I}, j \le i}) = \frac{W^{\text{up}} n_0}{g_{i\text{B}}} \left( 2^{\frac{s_i^{\text{up}}}{t^{\text{up}}} \frac{1}{W^{\text{up}}}} - 1 \right) 2^{\frac{1}{t^{\text{up}}} \frac{1}{W^{\text{up}}} \sum_{j=1}^{i-1} s_j^{\text{up}}}, \forall i \in \mathcal{I}, \quad (5)$$

which will be used in our following problem formulation.

### B. BS's NOMA-Transmission to the MTs

As shown in Figure 1(b), after finishing all MTs' offloaded workloads, the BS uses NOMA to simultaneously send the computation-results back to the respective MTs (i.e., the downlink NOMA transmission). Due to the SIC, we need to order the MTs in $\mathcal{I}$ according to the downlink channel power gains from the BS to the MTs. However, since the downlink channel power gain from the BS to MT $i$ is usually different from the uplink channel power gain from MT $i$ to the BS, we cannot use the same decoding order used in (1) for the downlink NOMA. Thus, we introduce another index-set $\mathcal{K}$ to denote the same group of MTs in $\mathcal{I}$. In $\mathcal{K}$, the MTs are now ordered as follows

$$h_{\text{B}1} > h_{\text{B}2} > h_{\text{B}3} > ... > h_{\text{B}k} > ... > h_{\text{B}I}, \quad (6)$$

where $h_{\text{B}k}$ denotes the downlink channel power gain from the BS to the $k$-th MT in $\mathcal{K}$.

We emphasize that both set $\mathcal{I}$ and set $\mathcal{K}$ refer to the same group of the MTs. In $\mathcal{K}$, the MTs are ordered according to (6). Thus, each MT has an index-tuple $(i, k)$, meaning that the MT is the $i$-th MT in $\mathcal{I}$ (from the uplink's perspective), and also is the $k$-th MT in $\mathcal{K}$ (from the downlink's perspective). In particular, to give a connection between $\mathcal{I}$ and $\mathcal{K}$, we introduce a one-to-one mapping $k = \Phi(i), \forall i \in \mathcal{I}$. In other words, if the MT is the $i$-th one in $\mathcal{I}$, then it is the $\Phi(i)$-th one in $\mathcal{K}$. In this work, we assume that all the channel power gains are relative static and known in advance, and thus the mapping $\{\Phi(i)\}_{i \in \mathcal{I}}$ is also known (as shown in Figure 1(b)). We use $\mathcal{I}$ (and the associated indices $i, j \in \mathcal{I}$) when discussing the MTs' uplink NOMA-transmissions, and use $\mathcal{K}$ (and the associated indices $k, m \in \mathcal{K}$) when discussing the BS's NOMA-transmission.

We use $q_k$ to denote the BS's transmit-power to MT $k$. Thus, based on (6), the BS's downlink-rate to MT $k \in \mathcal{K}$ is

$$R_{\text{B}k}^{\text{do}} = W^{\text{do}} \log_2 \left( 1 + \frac{q_k h_{\text{B}k}}{h_{\text{B}k} \sum_{m=1}^{k-1} q_m + W^{\text{do}} n_0} \right), \forall k \in \mathcal{K}, \quad (7)$$

where $W^{\text{do}}$ denotes the downlink channel bandwidth.

Recall that we consider that the MTs are running the application of data-compression. We use $\alpha_i$ to denote MT $i$'s compression-ratio. For each MT $i \in \mathcal{I}$, based on the introduced mapping $\Phi(i)$, there exists the following relationship between MT $i$'s offloaded computation-workload $s_i^{\text{up}}$ and the

[3]The proof of eq. (3) is essentially based on the deduction. We skip the details here due to the limited space. Interested readers can refer to Proposition 2 [42] for the details.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



associated computation-result to be sent from the BS to MT $k = \Phi(i)$:

$$s_k^{\text{do}} = \alpha_i s_i^{\text{up}}, \text{ with } k = \Phi(i) \in \mathcal{K}. \quad (8)$$

Notice that based on (8) and the mapping $\{\Phi(i)\}_{i\in\mathcal{I}}$, the MTs' offloaded computation-workloads $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ and the computation-results $\{s_k^{\text{do}}\}_{k\in\mathcal{K}}$ are actually interchangeable.

Based on (7), we quantify the BS's minimum total transmit-power to send the computation-results $\{s_k^{\text{do}}\}_{k\in\mathcal{K}}$ to the MTs. The result is given in the following proposition.

**Proposition 2:** Given the computation-results $\{s_k^{\text{do}}\}_{k\in\mathcal{K}}$ to be sent back the MTs and the downloading-duration $t^{\text{do}}$, the BS's minimum total transmit-power can be written as:

$$q_{\text{B}}^{\text{min,tot}} = W^{\text{do}} n_0 \sum_{k=1}^{I} \left( \frac{1}{h_{\text{B}k}} - \frac{1}{h_{\text{B}(k-1)}} \right) 2^{\frac{1}{t^{\text{do}}} \frac{1}{W^{\text{do}}} \sum_{m=k}^{I} s_m^{\text{do}}} - \frac{W^{\text{do}} n_0}{h_{\text{B}I}}. \quad (9)$$

*Proof:* According to Proposition 1 [42], the BS's minimum total transmit-power is

$$q_{\text{B}}^{\text{min,tot}} = W^{\text{do}} n_0 \sum_{k=1}^{I} \left( \frac{1}{h_{\text{B}k}} - \frac{1}{h_{\text{B}(k-1)}} \right) \prod_{m=k}^{I} (1 + \beta_{\text{B}m}) - \frac{W^{\text{do}} n_0}{h_{\text{B}I}}. \quad (10)$$

$\beta_{\text{B}k} = \frac{q_k h_{\text{B}k}}{h_{\text{B}k} \sum_{m=1}^{k-1} q_m + W^{\text{do}} n_0}$ denotes the received SINR for the BS's transmission to MT $k$. In particular, given the MTs' computation-results $\{s_k^{\text{do}}\}_{k\in\mathcal{K}}$ and the downloading-duration $t^{\text{do}}$, we derive each MT's required $\beta_{\text{B}k}$ as

$$\beta_{\text{B}k} = 2^{\frac{s_k^{\text{do}}}{t^{\text{do}}} \frac{1}{W^{\text{do}}}} - 1, \forall k \in \mathcal{K}. \quad (11)$$

By putting (11) into (10), we thus obtain (9). ∎

Based on (8) and $\{\Phi(i)\}_{i\in\mathcal{I}}$, $\{s_k^{\text{do}}\}_{k\in\mathcal{K}}$ and $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ can be treated interchangeable, namely, knowing $s_i^{\text{up}}$ enables us to know $s_k^{\text{do}}$ with $k = \Phi(i)$. As a result, for the sake of easy presentation, we denote the BS's minimum total transmit-power as a function of the downloading-duration $t^{\text{do}}$ and the MTs' offloaded computation-workloads $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ as follows

$$q_{\text{B}}^{\text{min,tot}}(t^{\text{do}}, \{s_i^{\text{up}}\}_{i\in\mathcal{I}}) = W^{\text{do}} n_0 \sum_{k=1}^{I} \left( \frac{1}{h_{\text{B}k}} - \frac{1}{h_{\text{B}(k-1)}} \right) 2^{\frac{1}{t^{\text{do}}} \frac{1}{W^{\text{do}}} \sum_{m=k}^{I} s_m^{\text{do}}} - \frac{W^{\text{do}} n_0}{h_{\text{B}I}}, \quad (12)$$

which will be used in our following problem formulation.

### C. Overall Delay for all MTs and Problem Formulation

For MT $i$, its local computation and the computation-offloading can be executed in a parallel manner. As a result, from MT $i$'s perspective, the overall delay for completing its total computation-workload $S_i^{\text{tot}}$ can be written as:

$$d_i^{\text{overall}} = \max \left\{ \frac{S_i^{\text{tot}} - s_i^{\text{up}}}{V_i^{\text{loc}}}, \frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}} + t^{\text{up}} + t^{\text{do}} \right\}, \forall i \in \mathcal{I}. \quad (13)$$

$\frac{S_i^{\text{tot}} - s_i^{\text{up}}}{V_i^{\text{loc}}}$ denotes MT $i$'s local computation delay. $\frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}} + t^{\text{up}} + t^{\text{do}}$ denotes MT $i$'s delay in executing the partial computation-offloading, which includes the round trip delay $t^{\text{up}} + t^{\text{do}}$ and the computation-delay at the edge server $\frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}}$[4]. Thanks to NOMA, all MTs can simultaneously send their offloaded computation-workloads to the BS, and the BS can send the computation-results back to the MTs simultaneously. Therefore, we aim at minimizing the overall delay for completing all MTs' computation requirements, i.e, minimizing $\max_{i\in\mathcal{I}}\{d_i^{\text{overall}}\}$. Notice that in this work, we do not consider the computation delay for the operations of the SIC in the system model.

We next discuss the constraints which will be used in our following problem formulation. First, we consider that each MT $i$'s total energy consumption for transmitting its computation-workload to the BS cannot exceed MT $i$'s maximum transmission energy-budget $E_i^{\max}$, i.e.,

$$t^{\text{up}} p_i^{\min}(t^{\text{up}}, \{s_j^{\text{up}}\}_{j\in\mathcal{I}, j\le i}) \le E_i^{\max}, \forall i \in \mathcal{I}. \quad (14)$$

Second, the BS's total energy consumption for transmitting the computation-results back to the MTs cannot exceed the BS's maximum transmission energy-budget $E_{\text{B}}^{\max}$, i.e.,

$$t^{\text{do}} q_{\text{B}}^{\text{min,tot}}(t^{\text{do}}, \{s_i^{\text{up}}\}_{i\in\mathcal{I}}) \le E_{\text{B}}^{\max}, \quad (15)$$

Third, we consider that each MT $i$'s total energy consumption for its local computation cannot exceed MT $i$'s computation energy-budget denoted by $Q_i^{\max}$, i.e.,

$$\rho_i \frac{S_i^{\text{tot}} - s_i^{\text{up}}}{V_i^{\text{loc}}} \le Q_i^{\max}, \forall i \in \mathcal{I}, \quad (16)$$

where parameter $\rho_i$ denotes the power consumption of MT $i$'s computation-processing unit. Recall that $V_i^{\text{loc}}$ denotes MT $i$'s local computation-rate. Similarly, the BS's total energy computation for executing all MTs' offloaded computation-workloads cannot exceed the BS's computation energy-budget $Q_{\text{B}}^{\max}$, i.e.,

$$\rho_{\text{B}} \frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}} \le Q_{\text{B}}^{\max}, \quad (17)$$

where parameter $\rho_{\text{B}}$ denotes the power consumption of the BS's computation-processing unit, and $V^{\text{ser}}$ denotes the computation-rate of the edge server.

Based on the above modelling, we formulate the following optimization problem that aims at minimizing the overall delay for completing all MTs' computation requirements (here "ODM" refers to "Overall Delay Minimization"):

$$\begin{aligned}
\text{(ODM):} \quad & \min\{\max_{i\in\mathcal{I}}\{d_i^{\text{overall}}\}\} \\
\text{subject to:} \quad & \text{constraints } (14), (15), (16), \text{ and } (17), \\
& 0 \le s_i^{\text{up}} \le S_i^{\text{tot}}, \forall i \in \mathcal{I}, \quad (18) \\
& 0 \le t^{\text{up}} \le T^{\text{up,max}}, \quad (19) \\
& 0 \le t^{\text{do}} \le T^{\text{do,max}}, \quad (20) \\
\text{variables:} \quad & \{s_i^{\text{up}}\}_{i\in\mathcal{I}}, t^{\text{up}}, \text{ and } t^{\text{do}}.
\end{aligned}$$

[4]In this work, to exploit the benefit of simultaneous data transmissions from/to the group of MTs, we assume that the edge server collects the offloaded computation-workloads from all MTs and processes these computation-workloads as a whole task. This assumption is viable, since after collecting all MTs' offloaded computation-workloads via the uplink NOMA-transmission, the edge server can sequentially process different MTs' offloaded computation-workloads. After completing all these computation-workloads, the edge server then sends the computation-results back to the respective MTs simultaneously via the downlink NOMA-transmission.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



Constraint (19) means that the uploading-duration $t^{\text{up}}$ cannot exceed the maximum-delay $T^{\text{up,max}}$, and constraint (20) means that the downloading-duration $t^{\text{do}}$ cannot exceed the maximum-delay $T^{\text{do,max}}$. Problem (ODM) is a typical non-convex optimization problem. Thus, there exists no general algorithm that can efficiently solve Problem (ODM). We focus on proposing an efficient algorithm to solve Problem (ODM) in the following. Notice that, to focus on our key objective of modeling and optimizing the MTs' overall delay in executing the computation offloading via both uplink and downlink NOMA transmissions, we assume the perfect SIC in the NOMA transmission. As an important future direction, we will further take into account the enabling condition for the perfect SIC in the problem formulation and evaluate the consequent impact.

Before leaving this section, we mention that for the purpose of performance comparison, we also use the TDMA enabled computation-offloading scheme in the following Section V. In the TDMA enabled computation-offloading scheme, the MTs sequentially execute the partial computation offloading, and we thus optimize the associated offloaded computation-workload, the uploading-duration, and the downloading-duration for each MT individually. For each MT $i$, our objective is to minimize MT $i$'s overall delay $d_i^{\text{TDMA}} = \max\{\frac{S_i^{\text{tot}}-s_i^{\text{up}}}{V_i^{\text{loc}}}, t_i^{\text{up}} + t_i^{\text{do}} + \frac{s_i^{\text{up}}}{V^{\text{ser}}}\}$, where $t_i^{\text{up}}$ denotes MT $i$'s uploading-duration and $t_i^{\text{do}}$ denotes MT $i$'s downloading-duration. Notice that since the MTs in $\mathcal{I}$ sequentially execute the computation-offloading in the TDMA manner, the overall delay for all MTs is given by $\sum_{i\in\mathcal{I}} d_i^{\text{TDMA}}$.

## IV. Proposed Algorithm to Solve Problem (ODM)

### A. Decomposition of Problem (ODM)

Our idea to solve Problem (ODM) is to exploit its layered structure. Suppose that all MTs' offloaded computation-workloads $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ are given in advance. The coupling constraints (14) and (15) can be separated into the one about the uploading-duration $t^{\text{up}}$ and another one about downloading-duration $t^{\text{do}}$. In other words, given $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$, we can separately minimize $t^{\text{up}}$ (i.e., the following subproblem (Sub-UDM)) and minimize $t^{\text{do}}$ (i.e., subproblem (Sub-DDM)), without loss of any optimality of the original Problem (ODM).

- *(Subproblem (Sub-UDM))*: Given $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$, we aim at minimizing the uploading-duration $t^{\text{up}}$ by solving the following subproblem (here, "Sub-UDM" refers to "Uploading-Duration Minimization Subproblem").

$$\begin{aligned}&\text{(Sub-UDM):} && t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} = \min t^{\text{up}}\\ &\text{subject to:} && p_i^{\min}(t^{\text{up}}, \{s_j^{\text{up}}\}_{j\le i, j\in\mathcal{I}}) \le \frac{E_i^{\max}}{t^{\text{up}}}, \forall i\in\mathcal{I}, \quad (21)\\ &\text{variables:} && 0 \le t^{\text{up}} \le T^{\text{up,max}}.\end{aligned}$$

  We include $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ in the subscript of $t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$ to denote that the minimum uploading-duration can be regarded as a function of $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$.

- *(Subproblem (Sub-DDM))*: Given $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$, we can thus compute the corresponding $\{s_k^{\text{do}}\}_{k\in\mathcal{K}}$ according to (8). Then, we aim at minimizing the downloading-duration $t^{\text{do}}$ by solving (here, "Sub-DDM" refer to "Downloading-Duration Minimization Subproblem"):

$$\begin{aligned}&\text{(Sub-DDM):} && t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} = \min t^{\text{do}}\\ &\text{subject to:} && p_{\text{B}}^{\min}(t^{\text{do}}, \{s_i^{\text{up}}\}_{i\in\mathcal{I}}) \le \frac{E_{\text{B}}^{\max}}{t^{\text{do}}}, \quad (22)\\ &\text{variables:} && 0 \le t^{\text{do}} \le T^{\text{do,max}}.\end{aligned}$$

  We include $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ in the subscript of $t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$ to denote that the minimum downloading-duration can be regarded as a function of $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$.

Based on $t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$ and $t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$, we then minimize $\max_{i\in\mathcal{I}}\{d_i^{\text{overall}}\}$, by optimizing the MTs' offloaded computation-workloads $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$, which corresponds to the following top-problem (Top-OCWO) (here, "OCWO" refers to "Offloaded Computation-Workloads Optimization").

$$\begin{aligned}&\text{(Top-OCWO):} \min\{\max_{i\in\mathcal{I}}\{d_i^{\text{overall}}\}\}\\ &\text{subject to: } d_i^{\text{overall}} =\\ &\max\left\{\frac{S_i^{\text{tot}}-s_i^{\text{up}}}{V_i^{\text{loc}}}, \frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}} + t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} + t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}\right\}, \forall i \quad (23)\\ &\text{constraints (16) and (17),}\\ &\text{variables: } \{d_i^{\text{overall}}\}_{i\in\mathcal{I}}, \text{ and } 0 \le s_i^{\text{up}} \le S_i^{\text{tot}}, \forall i\in\mathcal{I}.\end{aligned}$$

Based on the above vertical decomposition of top-problem (Top-OCWO) and its two subproblems, i.e., Subproblem (Sub-UDM) and Subproblem (Sub-DDM), the minimum overall delay for all MTs provided by top-problem (Top-OCWO) suffices to be the one for the original Problem (ODM).

### B. Proposed Algorithm to Solve Problem (Sub-UDM)

We propose an algorithm to solve Problem (Sub-UDM) and find the minimum uploading-duration under the given MTs' $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$. However, Problem (Sub-UDM) is a typical non-convex optimization problem. To solve this problem, we introduce

$$x = \frac{1}{t^{\text{up}}}. \quad (24)$$

Using (24) and some equivalent manipulations, we can transform Problem (Sub-UDM) into:

$$\begin{aligned}&\text{(Sub-UDM-E):} \; x^*_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} = \arg\max\; x\\ &\text{subject to: } \frac{W^{\text{up}} n_0}{g_{i\text{B}}}\left(2^{x\frac{s_i^{\text{up}}}{W^{\text{up}}}} - 1\right) 2^{x\frac{1}{W^{\text{up}}}\sum_{j=1}^{i-1} s_j^{\text{up}}} -\\ &\qquad E_i^{\max} x \le 0, \forall i\in\mathcal{I}, \quad (25)\\ &\qquad x \ge \frac{1}{T^{\text{up,max}}}, \quad (26)\\ &\text{variable: } x.\end{aligned}$$

Problem (Sub-UDM-E) aims at finding the maximum value of $x$ (i.e., $x^*_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$) that can meet constraints (25) and (26). Since (25) can be separated with respect to different MTs, we express

$$x^*_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} = \min_{i\in\mathcal{I}}\{x_i^{\text{largest}}\}, \quad (27)$$

where the auxiliary variable $x_i^{\text{largest}}$ is given by

$$x_i^{\text{largest}} = \arg\max\{x \ge \frac{1}{T^{\text{up,max}}} | Q_i(x) \ge 0\}, \forall i\in\mathcal{I}. \quad (28)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



Specifically, according to (25), function $Q_i(x)$ is given by:

$$Q_i(x) = E_i^{\max} x \frac{g_{i\mathrm{B}}}{W^{\mathrm{up}} n_0} - \left(2^{x\frac{s_i^{\mathrm{up}}}{W^{\mathrm{up}}}} - 1\right) 2^{x\frac{1}{W^{\mathrm{up}}}\sum_{j=1}^{i-1} s_j^{\mathrm{up}}}, \quad \forall i \in \mathcal{I}. \quad (29)$$

To find $x_i^{\text{largest}}$, we identify the following property.

**Proposition 3:** For MT $i$, its function $Q_i(x)$ is a unimodal function.

*Proof:* We can derive the first order derivative of $Q_i(x)$ as

$$Q_i^{'}(x) = E_i^{\max}\frac{g_{i\mathrm{B}}}{W^{\mathrm{up}} n_0} - \left(2^{x\frac{s_i^{\mathrm{up}}}{W^{\mathrm{up}}}} - 1\right) 2^{x\frac{1}{W^{\mathrm{up}}}\sum_{j=1}^{i-1} s_j^{\mathrm{up}}} (\ln 2)\left(\frac{1}{W^{\mathrm{up}}}\sum_{j=1}^{i-1} s_j^{\mathrm{up}}\right) - 2^{x\frac{1}{W^{\mathrm{up}}}\sum_{j=1}^{i-1} s_j^{\mathrm{up}}} 2^{x\frac{s_i^{\mathrm{up}}}{W^{\mathrm{up}}}} (\ln 2)\frac{s_i^{\mathrm{up}}}{W^{\mathrm{up}}}. \quad (30)$$

Eq. (30) shows $Q_i^{'}(x)$ decreases in $x$. Thus, there are two possible cases for $x \in [\frac{1}{T^{\mathrm{up,max}}}, \infty)$, i.e., **Case-I** $Q_i(x)$ is decreasing if $Q_i^{'}(\frac{1}{T^{\mathrm{up,max}}}) < 0$, or **Case-II** $Q_i(x)$ firstly increases and then decreases if $Q_i^{'}(\frac{1}{T^{\mathrm{up,max}}}) \geq 0$. Thus, $Q_i(x)$ is unimodal. ∎

Using Proposition 3 and the two cases discussed in the above proof of Proposition 3, we can find $x_i^{\text{largest}}$ (defined in (28)) based on the following three cases.

**Proposition 4:** For each MT $i$, we can find $x_i^{\text{largest}}$ in the following three cases:

- **Case-I(a)**: if $Q_i^{'}(\frac{1}{T^{\mathrm{up,max}}}) < 0$ and $Q_i(\frac{1}{T^{\mathrm{up,max}}}) < 0$, then $x_i^{\text{largest}}$ does not exist, which means that Problem (Sub-UDM-E) is infeasible.
- **Case-I(b)**: if $Q_i^{'}(\frac{1}{T^{\mathrm{up,max}}}) < 0$ and $Q_i(\frac{1}{T^{\mathrm{up,max}}}) \geq 0$, then $x_i^{\text{largest}} \in [\frac{1}{T^{\mathrm{up,max}}}, \infty)$ is uniquely determined by $Q_i(x_i^{\text{largest}}) = 0$.
- **Case-II**: if $Q_i^{'}(\frac{1}{T^{\mathrm{up,max}}}) > 0$, then $x_i^{\text{largest}} \in [x_i^{\triangle}, \infty)$ can be uniquely determined by $Q_i(x_i^{\text{largest}}) = 0$, where $x_i^{\triangle} \in [\frac{1}{T^{\mathrm{up,max}}}, \infty)$ is uniquely determined by $Q_i^{'}(x_i^{\triangle}) = 0$.

*Proof:* Based on Proposition 3, if $Q_i^{'}(\frac{1}{T^{\mathrm{up,max}}}) < 0$, then $Q^{'}(x)$ is monotonically decreasing when $x \in [\frac{1}{T^{\mathrm{up,max}}}, \infty)$. In this case, if $Q_i(\frac{1}{T^{\mathrm{up,max}}}) < 0$, then $x_i^{\text{largest}}$ does not exist (i.e., **Case-I(a)**). Otherwise (i.e., $Q_i(\frac{1}{T^{\mathrm{up,max}}}) \geq 0$), then there exists a unique point in the interval of $[\frac{1}{T^{\mathrm{up,max}}}, \infty)$ such that $Q_i(x) = 0$, and $x_i^{\text{largest}}$ corresponds to such a point (i.e., **Case-I(b)**).

On the other hand, if $Q_i^{'}(\frac{1}{T^{\mathrm{up,max}}}) \geq 0$ (i.e., **Case-II**), there exists a unique point in the interval of $[\frac{1}{T^{\mathrm{up,max}}}, \infty)$ such that $Q^{'}(x) = 0$ (since we have $Q^{'}(\infty) < 0$ according to (30)). Let us denote such a point as $x_i^{\triangle}$. Notice that, because of $Q_i(0) = 0$, there always exists $Q_i(x_i^{\triangle}) \geq 0$. As a result, there exists a unique point in the interval of $[x_i^{\triangle}, \infty)$ such that $Q_i(x) = 0$, and $x_i^{\text{largest}}$ corresponds to such a point. We thus finish the whole proof. ∎

Based on Proposition 4, we propose UDM-Algorithm to find $x^*_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})}$.

- Steps 4-5 correspond to Case-II in Proposition 4. Based on the decreasing property of $Q_i^{'}(x)$ when $x \in [\frac{1}{T^{\mathrm{up,max}}}, x^{\mathrm{uppbound}}]$, we firstly use the bisection-search method to find $x_i^{\triangle} \in [\frac{1}{T^{\mathrm{up,max}}}, x^{\mathrm{uppbound}}]$ such that $Q_i^{'}(x_i^{\triangle}) = 0$ (i.e., Step 4). Then, based on the decreasing property of $Q_i(x)$ when $x \in [x_i^{\triangle}, x^{\mathrm{uppbound}}]$, we further adopt the bisection-search method to find $x_i^{\text{largest}} \in [x_i^{\triangle}, x^{\mathrm{uppbound}}]$, such that $Q_i(x_i^{\text{largest}}) = 0$ (i.e., Step 5).
- Step 8 corresponds to Case-I(a) in Proposition 4. Based on the decreasing property of $Q_i(x)$ when $x \in [\frac{1}{T^{\mathrm{up,max}}}, x^{\mathrm{uppbound}}]$, we use the bisection-search method to find $x_i^{\text{largest}} \in [\frac{1}{T^{\mathrm{up,max}}}, x^{\mathrm{uppbound}}]$ such that $Q_i(x_i^{\text{largest}}) = 0$.

We finally output $x^*_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})}$ in Step 15. After obtaining $x^*_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})}$, we can compute the minimum uploading-duration for Problem (Sub-UDM) as follows

$$t^{\mathrm{up},*}_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})} = \frac{1}{x^*_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})}}. \quad (31)$$

*(Complexity of UDM-Algorithm):* Our UDM-Algorithm requires at most $I$ (i.e., the total number of the MTs) iterations. Within each round of the iteration, the bisection-search method requires at most $2\log_2\left(\frac{x^{\mathrm{uppbound}} - \frac{1}{T^{\mathrm{up,max}}}}{\epsilon_{\mathrm{UDM}}}\right)$ iterations to converge, where $x^{\mathrm{uppbound}}$ represents a very large number (e.g., we set $x^{\mathrm{uppbound}} = 10^4$ in UDM-Algorithm), and parameter $\epsilon_{\mathrm{UDM}}$ denotes the tolerable computation-error used in the bisection-search method. As a result, given $\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}}$, our UDM-Algorithm requires at most $2I \times \log_2\left(\frac{x^{\mathrm{uppbound}} - \frac{1}{T^{\mathrm{up,max}}}}{\epsilon_{\mathrm{UDM}}}\right)$ iterations.

Moreover, we identify the following important property regarding $t^{\mathrm{up},*}_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})}$.

**Proposition 5:** The minimum uploading-duration $t^{\mathrm{up},*}_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})}$ is increasing in $\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}}$.

*Proof:* According to eq. (27), $x^*_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})}$ is equal to one of $\{x_i^{\text{largest}}\}_{\in\mathcal{I}}$ exactly. We denote the index of this particular MT as $\hat{i}$, which thus corresponds to $Q_{\hat{i}}(x_{\hat{i}}^{\text{largest}}) = 0$. Recall that Proposition 4 indicates that there must exists $Q_{\hat{i}}^{\prime}(x_{\hat{i}}^{\text{largest}}) < 0$ (i.e., either Case-I(b) or Case-II). As a result, suppose that one element in $\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}}$ slightly increases, which yields that $Q_{\hat{i}}(x_{\hat{i}}^{\text{largest}})$ decreases (according to (29)). The only way to ensure $Q_{\hat{i}}(x_{\hat{i}}^{\text{largest}}) = 0$ to hold again is to reduce $x_{\hat{i}}^{\text{largest}}$. As

**UDM-Algorithm: to solve Problem (Sub-UDM-E) and find $x^*_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})}$**

1: **Initialization:** Set $i = 1$.
2: **while** $i \leq I$ **do**
3: **if** $Q_i^{'}(\frac{1}{T^{\mathrm{up,max}}}) \geq 0$ **then**
4: Use the bisection-search method to find $x_i^{\triangle} \in [\frac{1}{T^{\mathrm{up,max}}}, x^{\mathrm{uppbound}}]$ such that $Q_i^{'}(x_i^{\triangle}) = 0$.
5: Use the bisection-search method to find $x_i^{\text{largest}} \in [x_i^{\triangle}, x^{\mathrm{uppbound}}]$, such that $Q_i(x_i^{\text{largest}}) = 0$.
6: **else**
7: **if** $Q_i(\frac{1}{T^{\mathrm{up,max}}}) \geq 0$ **then**
8: Use the bisection-search method to find $x_i^{\text{largest}} \in [\frac{1}{T^{\mathrm{up,max}}}, x^{\mathrm{uppbound}}]$ such that $Q_i(x_i^{\text{largest}}) = 0$.
9: **else**
10: Problem (Sub-UDM) is infeasible.
11: **end if**
12: **end if**
13: Update $i = i + 1$.
14: **end while**
15: **Output**: $x^*_{(\{s_i^{\mathrm{up}}\}_{i\in\mathcal{I}})} = \min_{i\in\mathcal{I}}\{x_i^{\text{largest}}\}$, if Subproblem (Sub-UDM) is feasible. Otherwise, output Subproblem (Sub-UDM) is infeasible.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



a result, the corresponding $t^{\text{up},*}_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})}$ increases according to (24). ■

Proposition 5 is consistent with the intuition well. Specifically, given the MTs' fixed energy-budgets for transmission, we need to increase the uploading-duration (i.e., to reduce the offloading-rate) when the MTs' uploaded computation-workloads increase.

### C. Proposed Algorithm to Solve Problem (Sub-DDM)

We propose an algorithm to solve Problem (Sub-DDM) and find the minimum downloading-duration under the given MTs' uploaded computation-workloads $\{s^{\text{up}}_i\}_{i\in\mathcal{I}}$ (recall that there exists one-to-one mapping between $\{s^{\text{up}}_i\}_{i\in\mathcal{I}}$ and the computation-results $\{s^{\text{do}}_k\}_{k\in\mathcal{K}}$). We firstly introduce a variable-change as

$$y = \frac{1}{t^{\text{do}}}. \tag{32}$$

By using (32) and some equivalent manipulations, we can transform Problem (Sub-DDM) into:

$$\text{(Sub-DDM-E): } y^*_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})} = \arg\max\ y$$

$$\text{subject to: } W^{\text{do}} n_0 \sum_{k=1}^{I}\left(\frac{1}{h_{\text{B}k}} - \frac{1}{h_{\text{B}(k-1)}}\right) 2^{y\frac{1}{W^{\text{do}}}\sum_{m=k}^{I} s^{\text{do}}_m} - \frac{W^{\text{do}} n_0}{h_{\text{B}I}} - E^{\max}_{\text{B}} y \le 0, \tag{33}$$

$$y \ge \frac{1}{T^{\text{do,max}}}, \tag{34}$$

$$\text{variable: } y.$$

Problem (Sub-DDM-E) aims at finding the maximum value of $y$ that can satisfy (33) and (34). Thus, we express

$$y^*_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})} = y^{\text{largest}} = \arg\max\{y \ge \frac{1}{T^{\text{do,max}}} | R(y) \ge 0\}, \tag{35}$$

where, based on (33), function $R(y)$ is defined as

$$R(y) = E^{\max}_{\text{B}} y + \frac{W^{\text{do}} n_0}{h_{\text{B}I}} - W^{\text{do}} n_0 \sum_{k=1}^{I}\left(\frac{1}{h_{\text{B}k}} - \frac{1}{h_{\text{B}(k-1)}}\right) 2^{y\frac{1}{W^{\text{do}}}\sum_{m=k}^{I} s^{\text{do}}_m}. \tag{36}$$

To find $y^*_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})}$, we identify the following property.

**Proposition 6:** Function $R(y)$ is a unimodal function.

*Proof:* We can derive the first order derivative of $R(y)$ as

$$R^{'}(y) = E^{\max}_{\text{B}} - W^{\text{do}} n_0 \sum_{k=1}^{I}\left(\frac{1}{h_{\text{B}k}} - \frac{1}{h_{\text{B}(k-1)}}\right) 2^{y\frac{1}{W^{\text{do}}}\sum_{m=k}^{I} s^{\text{do}}_m} (\ln 2)\frac{1}{W^{\text{do}}} \sum_{m=k}^{I} s^{\text{do}}_m.$$

The above result shows that $R^{'}(y)$ is decreasing in $y$. Thus, there exists two possible cases for $y \in [\frac{1}{T^{\text{do,max}}}, \infty)$, i.e., **Case-I** function $R(y)$ is monotonically decreasing if $R^{'}(\frac{1}{T^{\text{do,max}}}) < 0$, or **Case-II** function $R(y)$ firstly increases and then decreases if $R^{'}(\frac{1}{T^{\text{do,max}}}) \ge 0$. As a result, $R(y)$ is a unimodal function. ■

Using Proposition 6 and the two cases discussed in the above proof of Proposition 6, we can find $y^{\text{largest}}$ (defined in (35)) as follows.

**Proposition 7:** We can find $y^{\text{largest}}$ in the three cases below:

- **Case-I(a):** if $R^{'}(\frac{1}{T^{\text{do,max}}}) < 0$ and $R(\frac{1}{T^{\text{do,max}}}) < 0$, then $y^{\text{largest}}$ does not exist, which means that Problem (D-DMP) is infeasible.
- **Case-I(b):** if $R^{'}(\frac{1}{T^{\text{do,max}}}) < 0$ and $R(\frac{1}{T^{\text{do,max}}}) \ge 0$, then $y^{\text{largest}} \in [\frac{1}{T^{\text{do,max}}}, \infty)$ is uniquely determined by $R(y^{\text{largest}}) = 0$.
- **Case-II:** if $R^{'}(\frac{1}{T^{\text{do,max}}}) > 0$, then $y^{\text{largest}} \in [y^{\triangle}, \infty)$ can be uniquely determined by $R(y^{\text{largest}}) = 0$, where $y^{\triangle} \in [\frac{1}{T^{\text{do,max}}}, \infty)$ is uniquely determined by $R^{'}(y^{\triangle}) = 0$.

*Proof:* The proof is similar to that for Proposition 4 before. We skip the details here due to the limited space. ■

Based on Proposition 7, we propose the following DDM-Algorithm to solve Problem (Sub-DDM-E) and find $y^*_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})}$. Specifically, Steps 2-3 correspond to Case-II in Proposition 7. Based on the decreasing property of $R^{'}(y)$ when $y \in [\frac{1}{T^{\text{do,max}}}, y^{\text{uppbound}}]$, we firstly use the bisection-search method to find $y^{\triangle} \in [\frac{1}{T^{\text{do,max}}}, y^{\text{uppbound}}]$ such that $R^{'}(y^{\triangle}) = 0$ (i.e., Step 2). Then, based on the decreasing property of $R(y)$ when $y \in [y^{\triangle}, y^{\text{uppbound}}]$, we use the bisection-search method to find $y^{\text{largest}} \in [y^{\triangle}, y^{\text{uppbound}}]$ such that $R(y^{\text{largest}}) = 0$ (i.e., Step 3). Step 6 corresponds to Case-I(b) in Proposition 7. Specifically, based on the decreasing property of $R(y)$ when $y \in [\frac{1}{T^{\text{do,max}}}, y^{\text{uppbound}}]$, we use the bisection-search method to find $y^{\text{largest}} \in [\frac{1}{T^{\text{do,max}}}, y^{\text{uppbound}}]$ such that $R(y^{\text{largest}}) = 0$. DDM-Algorithm finally outputs $y^*_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})}$ at Step 11.

**DDM-Algorithm: to solve Problem (Sub-DDM-E) and find $y^*_{(\{s^{\text{do}}_k\}_{k\in\mathcal{K}})}$**

1: **if** $R^{'}(\frac{1}{T^{\text{do,max}}}) \ge 0$ **then**
2: Use the bisection-search method to find $y^{\triangle} \in [\frac{1}{T^{\text{do,max}}}, y^{\text{uppbound}}]$ such that $R^{'}(y^{\triangle}) = 0$.
3: Use the bisection-search method to find $y^{\text{largest}} \in [y^{\triangle}, y^{\text{uppbound}}]$ such that $R(y^{\text{largest}}) = 0$.
4: **else**
5: **if** $R(\frac{1}{T^{\text{do,max}}}) \ge 0$ **then**
6: Use the bisection-search method to find $y^{\text{largest}} \in [\frac{1}{T^{\text{do,max}}}, y^{\text{uppbound}}]$ such that $R(y^{\text{largest}}) = 0$.
7: **else**
8: Problem (Sub-DDM-E) is infeasible.
9: **end if**
10: **end if**
11: **Output**: $y^*_{(\{s^{\text{do}}_k\}_{k\in\mathcal{K}})} = y^{\text{largest}}$, if Problem (Sub-DDM-E) is feasible. Otherwise, output Problem (Sub-DDM-E) is infeasible.

After obtaining $y^*_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})}$, we compute the downloading-duration for Problem (Sub-DDM) as

$$t^{\text{do},*}_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})} = \frac{1}{y^*_{(\{s^{\text{up}}_i\}_{i\in\mathcal{I}})}}. \tag{37}$$

*(Complexity of DDM-Algorithm):* Our DDM-Algorithm requires at most $2\log_2\left(\frac{y^{\text{uppbound}} - \frac{1}{T^{\text{do,max}}}}{\epsilon_{\text{DDM}}}\right)$ iterations to converge, where $y^{\text{uppbound}}$ represents a very large number (e.g., we set $y^{\text{uppbound}} = 10^4$ in DDM-Algorithm), and parameter $\epsilon_{\text{DDM}}$ denotes the tolerable computation-error used in the bisection-search method. As a result, for each given $\{s^{\text{up}}_i\}_{i\in\mathcal{I}}$, our DDM-Algorithm requires at most $2\log_2\left(\frac{y^{\text{uppbound}} - \frac{1}{T^{\text{do,max}}}}{\epsilon_{\text{DDM}}}\right)$ iterations.

Moreover, we identify the following property.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



**Proposition 8:** The optimum (i.e., the minimum) downloading-duration $t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$ is increasing with respect to $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ (or $\{s_k^{\text{do}}\}_{k\in\mathcal{K}}$, equivalently).

*Proof:* The proof is similar to the proof of Proposition 5 before. We skip the details here due to the limited space. ■

Proposition 8 matches the intuition. Given the BS's energy-budget for transmission, we need to increase the downloading-duration when the MTs' offloaded workloads increase.

### *D. Proposed Algorithm to Solve Problem (Top-OCWO)*

After obtaining $t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$ and $t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$, we next solve Problem (Top-OCWO). We introduce a new variable $\theta$ as

$$\theta = \max_{i\in\mathcal{I}}\{d_i^{\text{overall}}\}. \tag{38}$$

Based on (18), (19), and (20), we can derive the maximum interval for $\theta$ as $\theta \in [0, \theta^{\text{upper}}]$, where

$$\theta^{\text{upper}} = \max_{i\in\mathcal{I}}\left\{\frac{S_i^{\text{tot}}}{V_i^{\text{loc}}}, \frac{\sum_{i\in\mathcal{I}} S_i^{\text{tot}}}{V^{\text{ser}}} + T^{\text{up,max}} + T^{\text{do,max}}\right\}. \tag{39}$$

By using $\theta$, we transform Problem (Top-OCWO) into:

$$\begin{aligned}
&\text{(Top-OCWO-E):}\ \min \theta\\
&\text{subject to: } \theta \geq \frac{S_i^{\text{tot}} - s_i^{\text{up}}}{V_i^{\text{loc}}}, \forall i \in \mathcal{I}, &(40)\\
&\qquad \theta \geq \frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}} + t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} + t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{k\in\mathcal{I}})}, &(41)\\
&\qquad \text{constraints } (16) \text{ and } (17),\\
&\text{variables: } \theta, \text{ and } 0 \leq s_i^{\text{up}} \leq S_i^{\text{tot}}, \forall i \in \mathcal{I}.
\end{aligned}$$

We next propose an algorithm to solve Problem (Top-OCWO-E). The rationale is as follows. Suppose that the value of $\theta$ is given. We then only need to check whether constraints (40)-(41) and (16)-(17) can yield a non-empty feasible region or not. If yes, then we can reduce $\theta$ a little bit. Such a process continues, until we reach a threshold-value of $\theta$ which leads to that the four constraints fail to yield a non-empty feasible region. We thus obtain the optimal objective value (denoted by $\theta^*$) of Problem (Top-OCWO-E). We next illustrate the detailed procedures.

*1) (Procedures to determine the feasibility under a given $\theta$):* We firstly explain how to determine whether the constraints can yield a non-empty feasible region or not, under a given $\theta$. Specifically, under a given $\theta$, constraints (40) and (16) yield:

$$s_i^{\text{up}} \geq \max\{S_i^{\text{tot}} - \theta V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}, \forall i \in \mathcal{I}. \tag{42}$$

Propositions 5 and 8 tell that $t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$ and $t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$ increase in $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$, i.e., the right hand side of (41) is increasing in $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$. By setting $s_i^{\text{up}} = \max\{S_i^{\text{tot}} - \theta V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}, \forall i \in \mathcal{I}$ (i.e., (42) is strictly active) and putting $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$ into (41) and (17), respectively, we obtain the two conditions to check the feasibility of Problem (Top-OCWO-E):

$$\begin{aligned}
\text{(C1): } \theta \geq\ & \frac{\sum_{i\in\mathcal{I}} \max\{S_i^{\text{tot}} - \theta V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}}{V^{\text{ser}}}\\
&+ t^{\text{up},*}_{(\{\max\{S_i^{\text{tot}} - \theta V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}\}_{i\in\mathcal{I}})}\\
&+ t^{\text{do},*}_{(\{\max\{S_i^{\text{tot}} - \theta V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}\}_{i\in\mathcal{I}})} &(43)\\
\text{(C2): } \sum_{i\in\mathcal{I}} & \max\{S_i^{\text{tot}} - \theta V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}\\
&\leq \frac{V^{\text{ser}} Q_{\text{B}}^{\max}}{\rho_{\text{B}}} &(44)
\end{aligned}$$

Notice that given $\theta$, we can use UDM-Algorithm to obtain $t^{\text{up},*}_{(\{\max\{S_i^{\text{tot}} - \theta V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}\}_{i\in\mathcal{I}})}$, and use DDM-Algorithm to obtain $t^{\text{do},*}_{(\{\max\{S_i^{\text{tot}} - \theta V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}\}_{i\in\mathcal{I}})}$. We can thus determine whether constraint (43) is feasible or not. Specifically, if both (43) and (44) are feasible, then (40)-(41) and (16)-(17) can yield a non-empty feasible region under the given $\theta$.

*2) (Procedures to determine $\theta^*$):* We then illustrate how to determine the optimal value of $\theta$, i.e., $\theta^*$. An observation is that the left-hand side of (43) is increasing in $\theta$, while the right hand side of (43) is decreasing in $\theta$ (i.e., based on Proposition 5 and Proposition 8). Based on this property, we can use the bisection-search method to find $\theta^*$ (i.e., the optimal objective value of Problem (Top-OCWO-E)).

*3) (Proposed OCWO-Algorithm):* Based on the above illustrations, we propose the following OCWO-Algorithm to solve Problem (Top-OCWO-E) and find $\theta^*$ (the details are shown on the next page). The key of OCWO-Algorithm is to use the bisection-search method (i.e., the while-loop from Step 3 to Step 17) to find the optimal $\theta^* \in [0, \theta^{\text{upper}}]$ such that (40)-(41) and (16)-(17) are compatible. In each round of iteration, given the currently tested value of $\theta^{\text{cur}}$, we update the value of $s_i^{\text{up}}$ in Step 5, which corresponds to that constraint (42) is strictly active.

- If $\sum_{i\in\mathcal{I}} s_i^{\text{up}} > \frac{V^{\text{ser}} Q_{\text{B}}^{\max}}{\rho_{\text{B}}}$ (i.e., condition (44) fails to be satisfied), we increase $\theta$ in Step 15.
- Based on all MTs' offloaded-workloads $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$, in Step 7, we use UDM-Algorithm to compute the uploading-duration $t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$. In Step 8, we use DDM-Algorithm to compute the downloading-duration $t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$.
- After that, we compare the currently tested $\theta^{\text{cur}}$ and $\frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}} + t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} + t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$. If $\theta^{\text{cur}} \geq \frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}} + t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} + t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$, then the currently tested value of $\theta^{\text{cur}}$ ensures that all constraints lead to a non-empty feasible region. We thus reduce $\theta^{\text{cur}}$ by reducing its upper-bound (i.e., Step 10). Otherwise, the currently tested value of $\theta^{\text{cur}}$ yields that the constraints are incompatible. We thus increase $\theta^{\text{cur}}$ by increasing its lower-bound (i.e., Step 12).

The left-hand side of (43) is increasing in $\theta$, while the right hand side of (43) is decreasing in $\theta$. Thus, the above procedures of the bisection-search are guaranteed to converge.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



**OCWO-Algorithm: to solve Problem (Top-OCWO-E) and find $\theta^*$**

1: **Input:** the tolerable computation-error $\epsilon_{\text{OCWO}}$ for the bisection-search method.
2: **Initialization:** Set $\theta^{\text{cur-upp}} = \theta^{\text{upper}}$ (according to (39)) and $\theta^{\text{cur-low}} = 0$.
3: **while** $|\theta^{\text{cur-upp}} - \theta^{\text{cur-low}}| < \epsilon_{\text{OCWO}}$ **do**
4: Set $\theta^{\text{cur}} = \frac{\theta^{\text{cur-upp}} + \theta^{\text{cur-low}}}{2}$.
5: Set $s_i^{\text{up}} = \max\{S_i^{\text{tot}} - \theta^{\text{cur}} V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}, \forall i \in \mathcal{I}$.
6: **if** $\sum_{i\in\mathcal{I}} s_i^{\text{up}} \leq \frac{V^{\text{ser}} Q_{\text{B}}^{\max}}{\rho_{\text{B}}}$ **then**
7: Given $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$, we use UDM-Algorithm to compute $t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$.
8: Given $\{s_i^{\text{up}}\}_{i\in\mathcal{I}}$, we use DDM-Algorithm to compute $t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$.
9: **if** $\theta^{\text{cur}} \geq \frac{\sum_{i\in\mathcal{I}} s_i^{\text{up}}}{V^{\text{ser}}} + t^{\text{up},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})} + t^{\text{do},*}_{(\{s_i^{\text{up}}\}_{i\in\mathcal{I}})}$ **then**
10: Set $\theta^{\text{cur-upp}} = \theta^{\text{cur}}$.
11: **else**
12: Set $\theta^{\text{cur-low}} = \theta^{\text{cur}}$.
13: **end if**
14: **else**
15: Set $\theta^{\text{cur-low}} = \theta^{\text{cur}}$.
16: **end if**
17: **end while**
18: **Output**: $\theta^* = \theta^{\text{cur}}$ and $s_i^{\text{up}*} = \max\{S_i^{\text{tot}} - \theta^* V_i^{\text{loc}}, S_i^{\text{tot}} - \frac{V_i^{\text{loc}} Q_i^{\max}}{\rho_i}, 0\}, \forall i \in \mathcal{I}$.

Finally, we discuss the complexity of our OCWO-Algorithm. The bisection-search method in OCWO-Algorithm requires at most $\log_2\left(\frac{\theta^{\text{upp}}}{\epsilon_{\text{OCWO}}}\right)$ iterations to converge (and find $\theta^*$), where parameter $\epsilon_{\text{OCWO}}$ denotes the tolerable computation-error used in our OCWO-Algorithm. Moreover, within each round of the iteration, OCWO-Algorithm invokes the uses of UDM-Algorithm and DDM-Algorithm. Our UDM-Algorithm requires the complexity of $2I \times \log_2\left(\frac{x^{\text{uppbound}} - \frac{1}{T^{\text{up,max}}}}{\epsilon_{\text{UDM}}}\right)$ iterations (as explained in Subsection IV-B). Meanwhile, our DDM-Algorithm requires the complexity of $2\log_2\left(\frac{y^{\text{uppbound}} - \frac{1}{T^{\text{do,max}}}}{\epsilon_{\text{DDM}}}\right)$ iterations (as explained at the end of Subsection IV-C). As a result, OCWO-Algorithm requires at most $\log_2\left(\frac{\theta^{\text{upp}}}{\epsilon_{\text{OCWO}}}\right) \times \left(I \times 2\log_2\left(\frac{x^{\text{uppbound}} - \frac{1}{T^{\text{up,max}}}}{\epsilon_{\text{UDM}}}\right) + 2\log_2\left(\frac{y^{\text{uppbound}} - \frac{1}{T^{\text{do,max}}}}{\epsilon_{\text{DDM}}}\right)\right)$ iterations to solve Problem (ODM).

## V. Numerical Results

We show the numerical results to demonstrate the performance of our OCWO-Algorithm and our NOMA-enabled computation-offloading scheme. Specifically, we setup the scenario as follows. The BS (which is co-located with the edge server) is located at (0m,0m). The group of MTs are uniformly distributed within a plane whose central is the BS and the radius is 500m. We use the similar method as [46] to model the channel power gains from the MTs to the BS and that from the BS to the MTs. For instance, the channel power gain from MT $i$ to BS is given by $g_{i\text{B}} = \frac{\varrho_{i\text{B}}}{l_{i\text{B}}^{\kappa}}$, where $l_{i\text{B}}$ denotes the distance between MT $i$ and the BS, and $\kappa$ denotes the power-scaling factor for the path-loss (we set $\kappa = 3$). To capture the fading and shadowing effects, we assume that $\varrho_{i\text{B}}$ follows an exponential distribution with a unit mean. We set each MT's energy-budget for transmission as $E_i^{\max} = 4$J and the BS's energy-budget for transmission as $E_{\text{B}}^{\max} = 10$J. In the following Figure 2 to Figure 9, we set that the BS has a sufficiently large energy-budget $Q_{\text{B}}^{\max}$ such that the BS can accommodate all MTs' entire computation-workloads (specifically, we set $Q_{\text{B}}^{\max} = \rho_{\text{B}} \frac{\sum_{i\in\mathcal{I}} S_i^{\text{tot}}}{V^{\text{ser}}}$), which is the best case of the edge server we can expect. In Figure 10, we will vary $Q_{\text{B}}^{\max}$ to evaluate its impact. Finally, we set the compression-ratio $\alpha_i = 0.3, \forall i \in \mathcal{I}$. Other parameters will be specified in the following.

We firstly evaluate the effectiveness of our OCWO-Algorithm in Table I and Table II. For the purpose of comparison, we use the global-optimum solver provided by LINGO [48] (i.e., a commercial optimization software) to solve Problem (ODM) and obtain the minimum overall delay as the benchmark. Table I shows the comparison for an 8-MTs scenario. The eight MTs' locations are randomly generated as stated before. For each MT $i$, we set its $V_i^{\text{loc}}$ according to a uniform distribution within $[1, 2]$Mbits/second, and we set $V^{\text{ser}} = 10$Mbits/second. We test $W^{\text{do}} = 4$MHz, $W^{\text{do}} = 8$MHz, and $W^{\text{do}} = 12$MHz, and for the three tested cases, we vary $S_i^{\text{tot}}$ from 10Mbits to 100Mbits. In each cell, the first value denotes the minimum overall delay (i.e., $\theta^*$) obtained by our OCWO-Algorithm (or LINGO), and the second value denotes the computational time (in the unit of second) consumed by our OCWO-Algorithm (or LINGO). The results in Table I show that our algorithm can achieve the results almost same as those provided by LINGO's global-optimum solver (with the average error no greater than $0.001\%$), which thus validates the accuracy of our algorithm. Moreover, the comparison on the consumed computational time shows that our OCWO-Algorithm can significantly reduce the computational time compared with LINGO, which thus validates the computational efficiency of our OCWO-Algorithm. Notice that this advantage stems from that we exploit the decomposable structure of Problem (ODM) and decompose it into Problem (Top-OCWO) and two subproblems, based on which we can efficiently compute the optimal solution. Table II further shows the comparison under a 12-MTs scenario, with the other parameter settings same as those in Table I. The results in Table II again show that our algorithm can achieve the results almost same as the benchmark solutions provided by LINGO's global-optimum solver (with the average error no greater than $0.0015\%$), which validates the accuracy of our algorithm. Moreover, the comparison on the consumed computational time in Table II again validates the computational efficiency of our OCWO-Algorithm.

In Figure 2, we plot the MTs' optimal offloaded computation-workloads $\{s_i^{\text{up}*}\}_{i\in\mathcal{I}}$ and different MTs' experienced delays $\{d_i^{\text{overall}*}\}_{i\in\mathcal{I}}$, versus different $S_i^{\text{tot}}$. Specifically, we use a 4-MT scenario with the randomly generated channel power gains as $\{g_{i\text{B}}\}_{i\in\mathcal{I}} = \{7.6681 * 10^{-4}, 6.3200 * 10^{-4}, 1.3010 * 10^{-5}, 3.0625 * 10^{-7}\}$ and $\{h_{\text{B}k}\}_{k\in\mathcal{K}} = \{3.4503 * 10^{-4}, 1.4959 * 10^{-5}, 9.0966 * 10^{-6}, 2.3842 * 10^{-7}\}$, with the index-mapping from $\mathcal{I}$ to $\mathcal{K}$ as $\Phi(1) = 1$, $\Phi(2) = 3$, $\Phi(3) = 2$, and $\Phi(4) = 4$. In addition, we set $W^{\text{up}} = W^{\text{do}} = 8$MHz, and the other parameter-settings are same as those in Table I. Firstly, the results in the top-subplot show that when $S_i^{\text{tot}}$ increases, MT 2's and MT 3's optimal offloaded computation-workloads gradually increase, while

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing

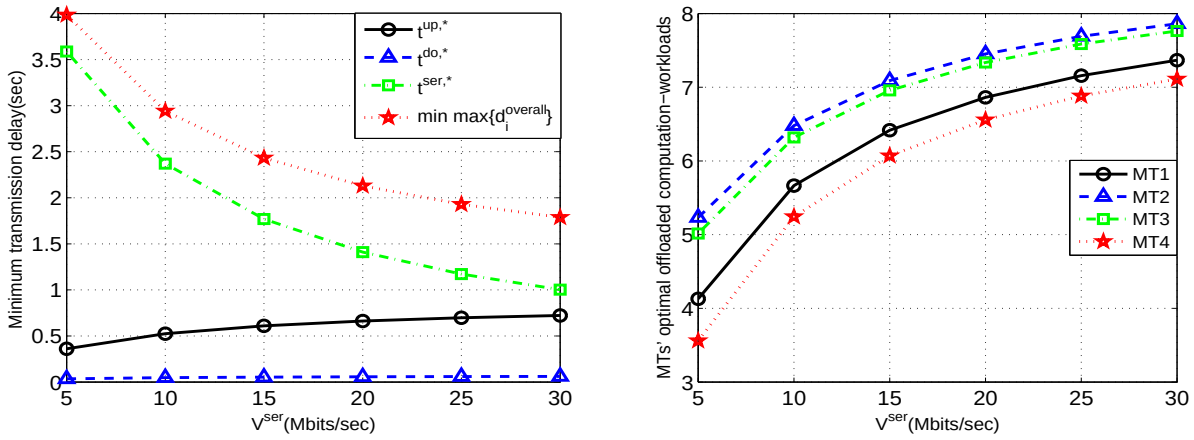


TABLE I: 8-MTs Scenario: We fix $W^{\text{up}} = 8\text{MHz}$, and $T^{\text{up,max}} = T^{\text{do,max}} = 1\text{sec}$

| $W^{\text{do}} = 4\text{MHz}$ | $S_i^{\text{tot}} = 10\text{Mbits}$ | $S_i^{\text{tot}} = 15\text{Mbits}$ | $S_i^{\text{tot}} = 20\text{Mbits}$ | $S_i^{\text{tot}} = 25\text{Mbits}$ | $S_i^{\text{tot}} = 50\text{Mbits}$ | $S_i^{\text{tot}} = 100\text{Mbits}$ | Ave. Error |
|---|---|---|---|---|---|---|---|
| Proposed | 4.0502, 0.0572sec | 6.0985, 0.0588sec | 8.1552, 0.0647sec | 11.7103, 0.0514sec | 29.7220, 0.0541sec | 67.7204, 0.0623sec | 0.0008% |
| LINGO | 4.0503, 3sec | 6.0984, 7sec | 8.1553, 12sec | 11.7103, 3sec | 29.7219, 8sec | 67.7208, 10sec | |
| $W^{\text{do}} = 8\text{MHz}$ | $S_i^{\text{tot}} = 10\text{Mbits}$ | $S_i^{\text{tot}} = 15\text{Mbits}$ | $S_i^{\text{tot}} = 20\text{Mbits}$ | $S_i^{\text{tot}} = 25\text{Mbits}$ | $S_i^{\text{tot}} = 50\text{Mbits}$ | $S_i^{\text{tot}} = 100\text{Mbits}$ | Ave. Error |
| Proposed | 4.0200, 0.0460sec | 6.0518, 0.0612sec | 8.1204, 0.0471sec | 11.7103, 0.0497sec | 29.7220, 0.0530sec | 67.7204, 0.0664sec | 0.0009% |
| LINGO | 4.0199, 2sec | 6.0517, 12sec | 8.1203, 6sec | 11.7104, 4sec | 29.7219, 4sec | 67.7201, 4sec | |
| $W^{\text{do}} = 12\text{MHz}$ | $S_i^{\text{tot}} = 10\text{Mbits}$ | $S_i^{\text{tot}} = 15\text{Mbits}$ | $S_i^{\text{tot}} = 20\text{Mbits}$ | $S_i^{\text{tot}} = 25\text{Mbits}$ | $S_i^{\text{tot}} = 50\text{Mbits}$ | $S_i^{\text{tot}} = 100\text{Mbits}$ | Ave. Error |
| Proposed | 4.0098, 0.0601sec | 6.0360, 0.0542sec | 8.1204, 0.0575sec | 11.7103, 0.0618sec | 29.7220, 0.0620sec | 67.7204, 0.0526sec | 0.0006% |
| LINGO | 4.0097, 4sec | 6.0360, 6sec | 8.1203, 5sec | 11.7103, 3sec | 29.7219, 6sec | 67.7202, 7sec | |

TABLE II: 12-MTs Scenario: We fix $W^{\text{up}} = 8\text{MHz}$, and $T^{\text{up,max}} = T^{\text{do,max}} = 1\text{sec}$

| $W^{\text{do}} = 4\text{MHz}$ | $S_i^{\text{tot}} = 10\text{Mbits}$ | $S_i^{\text{tot}} = 15\text{Mbits}$ | $S_i^{\text{tot}} = 20\text{Mbits}$ | $S_i^{\text{tot}} = 25\text{Mbits}$ | $S_i^{\text{tot}} = 50\text{Mbits}$ | $S_i^{\text{tot}} = 100\text{Mbits}$ | Ave. Error |
|---|---|---|---|---|---|---|---|
| Proposed | 4.6713, 0.0916sec | 7.6502, 0.0833sec | 11.6306, 0.1005sec | 15.6607, 0.0874sec | 35.7582, 0.0949sec | 78.9596, 0.0905sec | 0.0010% |
| LINGO | 4.6714, 11sec | 7.6502, 9sec | 11.6308, 6sec | 15.6609, 9sec | 35.7585, 14sec | 78.9598, 10sec | |
| $W^{\text{do}} = 8\text{MHz}$ | $S_i^{\text{tot}} = 10\text{Mbits}$ | $S_i^{\text{tot}} = 15\text{Mbits}$ | $S_i^{\text{tot}} = 20\text{Mbits}$ | $S_i^{\text{tot}} = 25\text{Mbits}$ | $S_i^{\text{tot}} = 50\text{Mbits}$ | $S_i^{\text{tot}} = 100\text{Mbits}$ | Ave. Error |
| Proposed | 4.6467, 0.0842sec | 7.6482, 0.0858sec | 11.6286, 0.0992sec | 15.6607, 0.0902sec | 35.7582, 0.0895sec | 78.9596, 0.0930sec | 0.0012% |
| LINGO | 4.6468, 9sec | 7.6483, 12sec | 11.6289, 8sec | 15.6609, 16sec | 35.7585, 13sec | 78.9597, 11sec | |
| $W^{\text{do}} = 12\text{MHz}$ | $S_i^{\text{tot}} = 10\text{Mbits}$ | $S_i^{\text{tot}} = 15\text{Mbits}$ | $S_i^{\text{tot}} = 20\text{Mbits}$ | $S_i^{\text{tot}} = 25\text{Mbits}$ | $S_i^{\text{tot}} = 50\text{Mbits}$ | $S_i^{\text{tot}} = 100\text{Mbits}$ | Ave. Error |
| Proposed | 4.6383, 0.0901sec | 7.6422, 0.0813sec | 11.6216, 0.0861sec | 15.6607, 0.0839sec | 35.7582, 0.0919sec | 78.9596, 0.0872sec | 0.0009% |
| LINGO | 4.6383, 15sec | 7.6424, 8sec | 11.6217, 6sec | 15.6609, 10sec | 35.7584, 8sec | 78.9598, 6sec | |




(a) Different components of the delay versus $V^{\text{ser}}$

(b) MTs' offloaded computation-workloads versus $V^{\text{ser}}$

Fig. 3: Optimal offloading solution versus different $V^{\text{ser}}$ (i.e., the computation-rate of the edge server). We set $S_i^{\text{tot}} = 10\text{Mbits}$, and set $W^{\text{up}} = W^{\text{do}} = 8\text{MHz}$. Other parameters are same as those used in Table I.

MT 1's and MT 4's optimal offloaded computation-workloads firstly gradually increase and then start to decrease when $S_i^{\text{tot}}$ becomes very large. This result is consistent with the intuition. Subject to the MTs' and the BS's limited energy-budges for transmission as well as the maximum uploading-duration and downloading duration, it is not beneficial for all MTs to blindly increase their computation-workloads to the BS when $S_i^{\text{tot}}$ increases. That is why we observe that MT 1's and MT 4's optimal offloaded computation-workloads start to decrease when the MTs' computation requirements become very large (in particular, MT 4 even does not offload its computation-workload to the BS when $S_i^{\text{tot}} = 150\text{Mbits}$ and $200\text{Mbits}$). According to the top-subplot, the bottom-subplot shows each MT's experienced overall delay. It is reasonable to observe that the overall delay experienced by each individual MT gradually increases when $S_i^{\text{tot}}$ increases, and meanwhile, all MTs experience the same delay (which thus can minimize the overall relay for all MTs). However, when $S_i^{\text{tot}}$ becomes very large (i.e., $S_i^{\text{tot}} = 150\text{Mbits}$ and $200\text{Mbits}$), MT 4's choice is to execute its total computation-workload locally, since the total amount of the offloaded workloads from MTs 1-3 is very large.

We further evaluate the optimal offloading solution under different $V^{\text{ser}}$ (i.e., the edge server's computation-rate), with the detailed results shown in Figure 3. As shown in Figure 3(a), when $V^{\text{ser}}$ increases, both the optimal uploading-duration and the downloading-duration increase, because the MTs are encouraged to offload more computation-workloads to the BS (to exploit the edge server's large computation-rate). Nevertheless, thanks to the edge server's increasing computation-rate, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



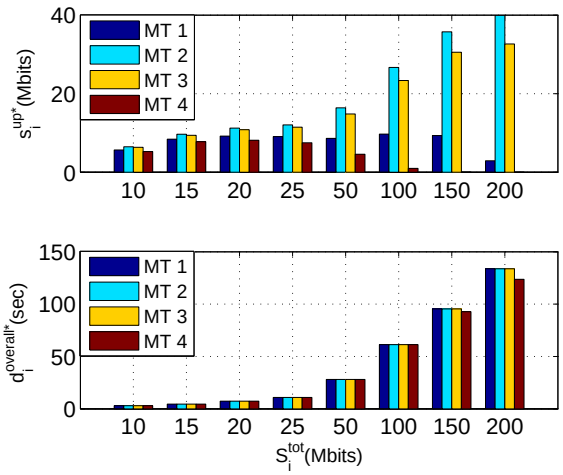



Fig. 2: Illustration of the optimal offloading solution. We use a 4-MTs scenario. Top-subplot: Each MT's optimal offloaded computation-workload. Bottom-subplot: Each MT's experienced delay.

processing-delay at the BS decreases, which thus reduces the overall delay experienced by the MTs. Figure 3(b) verifies that the MTs' optimal offloaded computation-workloads increase when $V^{\text{ser}}$ increases, i.e., the MTs are encouraged to offload more computation-workloads to the BS.

Next, we compare our NOMA-offloading scheme with the fixed computation-offloading scheme, with the detailed results shown in Figures 4 and 5. In the fixed computation-offloading scheme, each MT offloads a fixed portion (e.g., $40\%$, $60\%$, and $80\%$) of its total workload to the BS.

Figure 4 shows the performance of our NOMA-enabled offloading scheme compared with the fixed offloading scheme versus different $S_i^{\text{tot}}$. We use an 8-MTs scenario, and set $W^{\text{up}} = W^{\text{do}} = 8$MHz, and $T^{\text{up,max}} = T^{\text{do,max}} = 1$sec, and $V^{\text{ser}} = 10$Mbits/second and $V_i^{\text{loc}} = 1$Mbits/second. Every point denotes the average result of 200 random realizations of the MTs' locations. The results show that our offloading scheme can reduce the overall delay, compared to the fixed offloading scheme. Moreover, our offloading scheme enables that all the tested cases (i.e., $S_i^{\text{tot}}$ from 4Mbits to 24Mbits) are feasible, while the fixed offloading scheme leads to that the tested cases become infeasible when $S_i^{\text{tot}}$ becomes large.

Figure 5 further shows the comparison with the fixed offloading scheme versus different number of the MTs (we fix $S_i^{\text{tot}} = 10$Mbits for each MT). It is reasonable to observe that the overall delay increases when the total number of the MTs increases[5], and our scheme can reduce the overall delay due to enabling a flexible scheduling of the computation-workload between the edge server and the MTs. In particular, the results show that our NOMA-offloading yields an increasing marginal gain against the fixed scheme when the number of the MTs increases, i.e., our scheme is more beneficial for more MTs.

Next, we compare our computation-offloading scheme with

[5]Notice that for the fixed $40\%$-offloading, the overall delay is bounded by the local computation-delay (i.e., the part of $\frac{S_i^{\text{tot}} - s_i^{\text{up}}}{V_i^{\text{loc}}}$) when the number of the MTs is no greater than 12. Thus, the average overall delay keeps unchanged from $I = 4$ to $I = 12$.

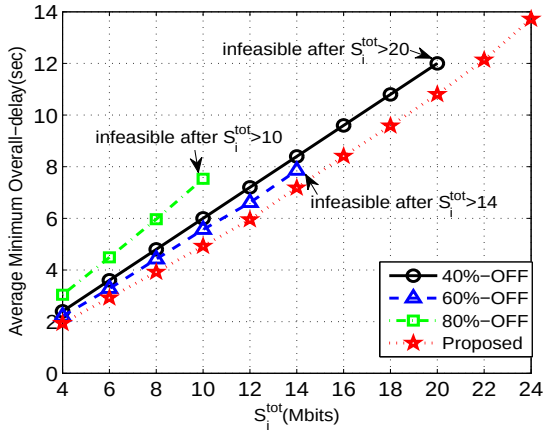



Fig. 4: Comparison with the fixed offloading scheme versus different $S_i^{\text{tot}}$

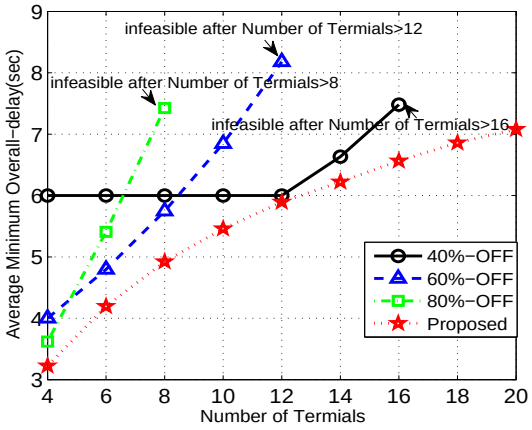



Fig. 5: Comparison with the fixed offloading scheme versus different numbers of the MTs

the TDMA enabled computation-offloading scheme. As we have explained at the end of Section III before, in the TDMA enabled computation-offloading scheme, the MTs sequentially execute the partial computation offloading, and we optimize the associated offloaded computation-workload, the uploading-duration, and the downloading-duration for each MT individually. The objective is to minimize MT $i$'s overall delay $d_i^{\text{TDMA}} = \max\{\frac{S_i^{\text{tot}} - s_i^{\text{up}}}{V_i^{\text{loc}}}, t_i^{\text{up}} + t_i^{\text{do}} + \frac{s_i^{\text{up}}}{V^{\text{ser}}}\}$, where $t_i^{\text{up}}$ and $t_i^{\text{do}}$ denote MT $i$'s uploading-duration and download-duration in the TDMA transmission, respectively. Since the MTs in $\mathcal{I}$ sequentially execute the computation-offloading in the TDMA manner, the overall delay for all MTs is given by $\sum_{i\in\mathcal{I}} d_i^{\text{TDMA}}$. We show the detailed comparison results in the following Figures 6 to 9. Figure 6 shows the performance comparison between our NOMA-offloading scheme and the TDMA-offloading scheme versus different $S_i^{\text{tot}}$. We use an 8-MTs scenario, and set $W^{\text{up}} = W^{\text{do}} = 8$MHz, and set $V^{\text{ser}} = 10$Mbits/second. Every point denotes the average result of 200 random realizations of the MTs' locations. It is reasonable to observe in Figure 6 that when the total computation-workload increases, the overall delays of both our NOMA-offloading scheme and the TDMA-offloading scheme increase. Nevertheless, our NOMA-offloading scheme can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



reduce the overall delay in comparison with the TDMA-offloading scheme. This advantage stems from the benefit of NOMA. Specifically, the NOMA-transmission enables all MTs to simultaneously offload computation-workloads to the BS, and enables the BS to send the computation-results to the respective MTs simultaneously, and further mitigates the co-channel interference by exploiting SIC with proper power allocations. As a result, our NOMA-offloading scheme can effectively reduce the uplink-duration and the downlink-duration, which lowers the overall delay compared with the TDMA-offloading scheme.

Figure 7 shows the performance comparison between our NOMA-offloading scheme and the TDMA-offloading scheme versus different number of the MTs. The results show that the overall delays of both our NOMA-offloading scheme and the TDMA-offloading scheme increase when the number of the MTs increases. Nevertheless, using NOMA can reduce the transmission-delay (for both sending the computation-workloads and computation-results) and thus reduce the overall delay compared with the TDMA-offloading scheme.

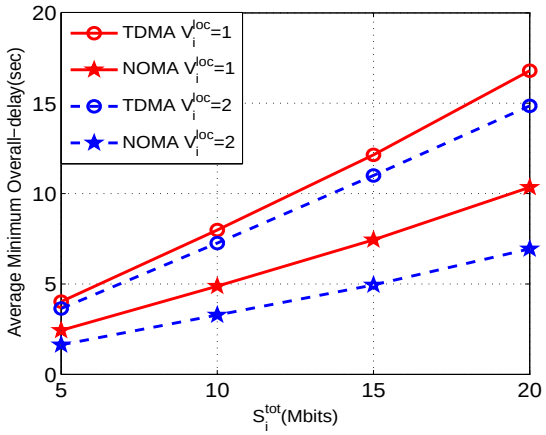



Fig. 6: Comparison with the TDMA scheme versus different $S_i^{\text{tot}}$

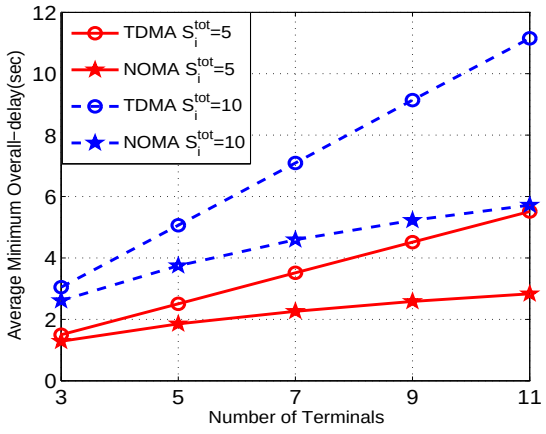



Fig. 7: Comparison with the TDMA scheme versus different number of MTs.

Figure 8(a) shows the comparison between our NOMA-enabled computation-offloading scheme and the TDMA-offloading scheme versus different $V^{\text{ser}}$ (i.e., the edge server's computation-rate). With larger computation-rate provided by the edge server, both our NOMA-offloading scheme and the TDMA-offloading yield smaller overall delays. Meanwhile, the comparison between the two results validate the advantage of using our NOMA-offloading scheme to reduce the overall delay against the TDMA-offloading. Figure 8(b) shows the similar advantage of our NOMA-enabled scheme against the TDMA-offloading versus different $V_i^{\text{loc}}$.

Figure 9(a) shows the comparison between our NOMA-offloading scheme and the TDMA-offloading scheme under different $W^{\text{do}}$ (i.e., the downlink channel bandwidth used by the BS), with the fixed $W^{\text{up}} = 8$MHz. As shown in Figure 9(a), with the increase of $W^{\text{do}}$, both the NOMA-offloading scheme and the TDMA-offloading scheme can provide larger downloading-rates from the BS to the MTs, which lead to that the overall delay decreases. Moreover, thanks to the benefit provided by NOMA, our NOMA-offloading scheme yields a smaller overall delay than the TDMA-offloading scheme for all the tested cases. Figure 9(b) shows the similar advantage of our NOMA-offloading against the TDMA-offloading scheme under different $W^{\text{up}}$.

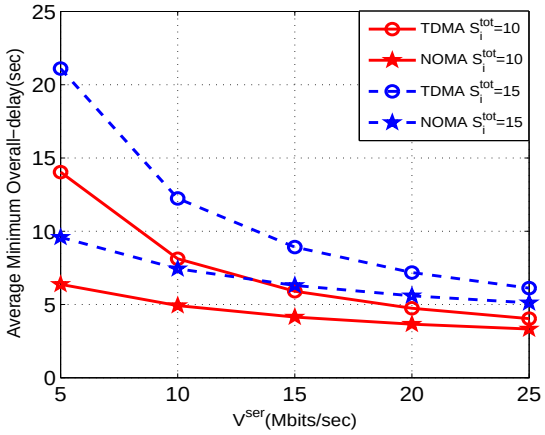
In Figure 10, we demonstrate the advantage of our NOMA-offloading scheme compared with the TDMA-offloading scheme versus different $Q_{\text{B}}^{\max}$ (i.e., the BS's energy-budget for executing the MTs' offloaded computation-workloads). We use an 8-MTs scenario. We set $S_i^{\text{tot}} = 10$Mbits, and set $V^{\text{ser}} = 10$Mbits/second and $V_i^{\text{loc}} = 1$Mbits/second. Figure 10(a) shows the comparison results under different cases of $E_i^{\max}$, and Figure 10(b) shows the comparison results under different cases of $E_{\text{B}}^{\max}$. The results show that both our NOMA-offloading scheme and the TDMA scheme yield a smaller overall delay when $Q_{\text{B}}^{\max}$ increases (i.e., the BS has a larger energy-budget for processing the MTs' offloaded computation-workloads). Meanwhile, our NOMA-offloading scheme can reduce the the overall delay in comparison with the TDMA-offloading scheme. It is because that NOMA enables larger transmission-rates between the MTs and the BS, and thus the computation-capability at the edge server can be better utilized.

## VI. Conclusion

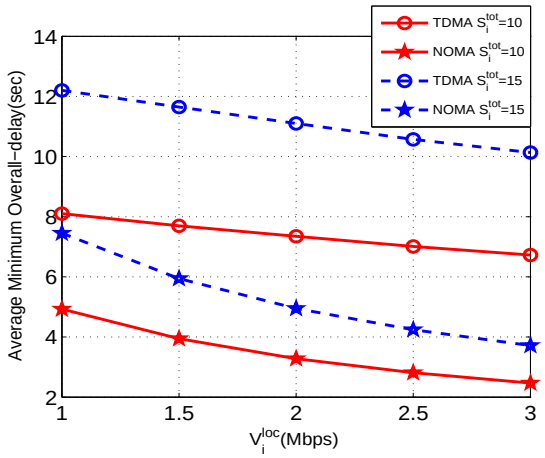
We have studied the optimal NOMA-enabled MEC approach to minimize the overall delay for completing the MTs' computation-requirements. The problem was formulated as a joint optimization of the MTs' offloaded computation-workloads, and the uploading-duration for the MTs to offload their computation-workloads to the BS, and the downloading-duration for the BS to send the computation-results back to the MTs. Exploiting the layered structure of the formulated problem, we proposed the efficient algorithms to compute the optimal computation-offloading solution. Numerical results have been provided to validate our proposed algorithms and demonstrate the advantage of our NOMA-offloading scheme. For our future work, we will investigate the distributed implementation of the NOMA-offloading scheme, in which the MTs determine their respective offloaded computation-workloads a distributed manner, and the BS adjusts the uploading-duration and the downloading-duration accordingly. In addition, it is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2893057, IEEE Journal of Selected Topics in Signal Processing



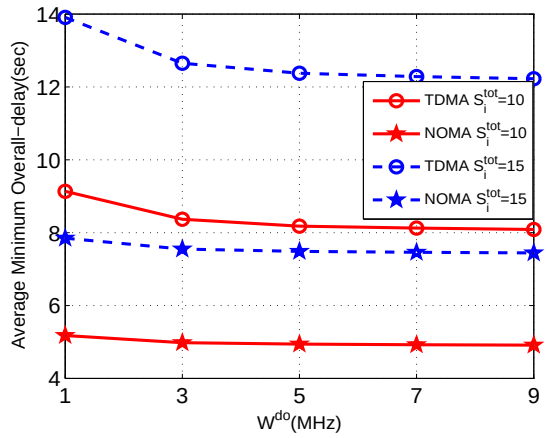



(a) Performance under diff. $V^{\text{ser}}$ with $V_i^{\text{loc}} = 1$Mbits/second

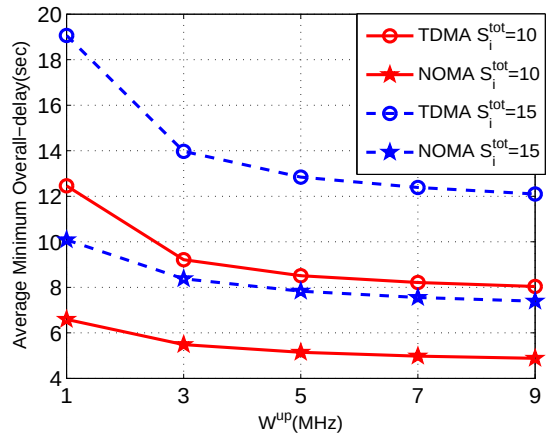



(b) Performance under diff. $V_i^{\text{loc}}$ with $V^{\text{ser}} = 10$Mbits/second

Fig. 8: Performance advantage of the proposed NOMA-enabled offloading scheme compared with the TDMA scheme versus different $V^{\text{ser}}$ (i.e., the BS's computation-rate) and $V_i^{\text{loc}}$ (i.e., the MTs' local computation-rate). We use an 8-MTs scenario, and set $W^{\text{up}} = W^{\text{do}} = 8$MHz.

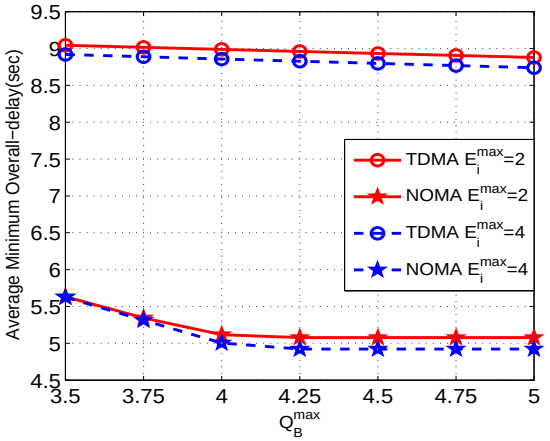



(a) Comparison under different $W^{\text{do}}$ with fixed $W^{\text{up}} = 8$MHz

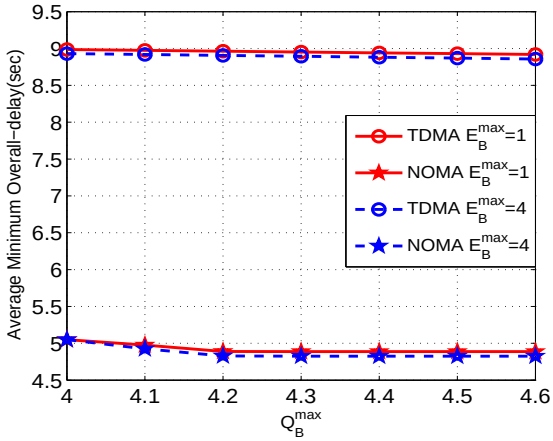



(b) Comparison under different $W^{\text{up}}$ with fixed $W^{\text{do}} = 8$MHz

Fig. 9: Performance advantage of the proposed NOMA-enabled offloading scheme compared with the TDMA scheme versus different $W^{\text{do}}$ (i.e., the downlink channel bandwidth) and $W^{\text{up}}$ (i.e., the uplink channel bandwidth). We use an 8-MTs scenario, and set $V^{\text{ser}} = 10$Mbits/second and $V_i^{\text{loc}} = 1$Mbits/second.




(a) Comparison under different $E_i^{\max}$




(b) Comparison under different $E_{\text{B}}^{\max}$

Fig. 10: Performance of the NOMA-offloading scheme compared with the TDMA scheme versus different $Q_{\text{B}}^{\max}$.

also an interesting future direction for us to study the offloading model in which different MTs may have different transmission-durations in the NOMA transmission.

## REFERENCES


[1] Y. Mao, C. You, J. Zhang, K. Huang, and K.B. Letaief, "A Survey on Mobile Edge Computing: the Communication Perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322-2358, Aug. 2017.

[2] T. Taleb, *et. al.*, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Architecture & Orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657-1681, May 2017.

[3] Z. Ding, P. Fan, and H.V. Poor, "Impact of Non-orthogonal Multiple Access on the offloading of mobile edge computing," April 2018, available at https://arxiv.org/pdf/1804.06712.pdf.

[4] A. Kiani, N. Ansari, "Edge Computing Aware NOMA for 5G Networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299-1306, April, 2018.

[5] F. Wang, J. Xu, and Z. Ding, "Optimized Multiuser Computation Offloading with Multi-antenna NOMA," July 2017, available at https://arxiv.org/pdf/1707.02486.pdf.

[6] Z. Ding, D.W.K. Ng, R. Schober, and H.V. Poor, "Delay Minimization for NOMA-MEC Offloading," *IEEE Signal Processing Letters*, vol. 25, no. 12, Dec. 2018.

[7] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient Mutli-user Computation Offloading for Mobile-edge Cloud Computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795-2808, Oct. 2016.

[8] M.H. Chen, B. Liang, and M. Dong, "Joint Offloading Decision and Resource Allocation for Multi-user Multi-task Mobile Cloud," *in Proc. of IEEE ICC'2016*.

[9] C. Wang, *et. al.*, "Computation Offloading and Resource Allocation in Wireless Cellular Networks With Mobile Edge Computing," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 4924-4938, Aug. 2017.

[10] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative Edge Caching in User-Centric Clustered Mobile Networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1791-1805, Aug. 2018.

[11] T. Dinh, J. Tang, Q. La, and T.Q.S. Quek, "Offloading in Mobile Edge Computing: Task Allocation and Computational Frequency Scaling," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3571-3584, Aug. 2017.

[12] T. Zhao, S. Zhou, X. Guo, and Z. Niu, "Tasks Scheduling and Resource Allocation in Heterogeneous Cloud for Delay-bounded Mobile Edge Computing," in *Proc. of IEEE ICC'2017*.

[13] Y. Zhang, D. Niyato, and P. Wang, "Offloading in Mobile Cloudlet Systems with Intermittent Connectivity," *IEEE Transactions on Mobile Computing*, vol. 14, no. 12, pp. 1536-1233, Dec. 2015.

[14] H. Wu, X. Tao, N. Zhang, and X. Shen, "Cooperative UAV Cluster Assisted Terrestrial Cellular Networks for Ubiquitous Coverage," *IEEE Journal on Sel. Areas in Communications*, vol. 36, no. 9, pp. 2045-2058, Sept. 2018.

[15] J. Liu, Y. Mao, J. Zhang, and K.B. Letaief, "Delay-Optimal Computation Task Scheduling for Mobile-Edge Computing Systems," *in Proc. of ISIT'2016*.

[16] Y. Mao, J. Zhang, S.H. Song, and K.B. Letaief, "Stochastic Joint Radio and Computational Resource Management for Multi-user Mobile-edge Computing Systems," Feb. 2017, available online at https://arxiv.org/abs/1702.00892.

[17] J. Ren, G. Yu, Y. Cai, and Y. He, "Latency for Resource Allocation in Mobile-Edge Computation Offloading," April 2017. available online at: http://arxiv.org/pdf/1704.00163.pdf.

[18] C. You, K. Huang, H. Chae, and B.H. Kim, "Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397-1411, March 2017.

[19] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-Edge Computing: Partial Computation Offloading Using Dynamic Voltage Scaling," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4268-4282, Oct. 2016.

[20] X. Cao, *et. al.*, "Joint Computation and Communication Cooperation for Mobile Edge Computing," in *Proc. of WiOpt'2018*.

[21] Y. Wu, K. Ni, C. Zhang, L. Qian, and D.H.K. Tsang, "NOMA Assisted Multi-Access Mobile Edge Computing: A Joint Optimization of Computation Offloading and Time Allocation," *IEEE Trans. on Vehicular Tech.*, vol. 67, no. 12, pp. 12244-12258, Dec. 2018.

[22] A.-L. Jin, W. Song, and W. Zhuang, "Auction-based Resource Allocation for Sharing Cloudlets in Mobile Cloud Computing," *IEEE Trans. on Emerging Topics in Computing*, available at DOI:10.1109/TETC.2015.2487865.

[23] A. Kiani, N. Ansari, "Toward Hierarchical Mobile Edge Computing: An Auction-Based Profit Maximization Approach," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2082-2091, Dec. 2017.

[24] Z. Chang, Z. Zhou, T. Ristaniemi, and Z. Niu, "Energy Efficient Optimization for Computation Offloading in Fog Computing System," in *Proc. of IEEE GLOBECOM'2017*.

[25] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C.-L. I, and H.V. Poor, "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Communications Magazine*, vol.55, no.2, pp. 185-191, Feb. 2017.

[26] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo "Non-orthogonal Multiple Access for 5G and Beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347-2381, Dec. 2017.

[27] L. Bai, L. Zhu, X. Zhang, W. Zhang, and Q. Yu, "Multi-satellite Relay Transmission in 5G: Concepts, Techniques and Challenges," *IEEE Network*, vol. 32, no. 5, pp. 38-44, Sept. 2018.

[28] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An Optimization Perspective of the Superiority of NOMA Compared to Conventional OMA," *IEEE Trans. on Signal Processing*, vol. 65, no. 19, pp. 5191-5202, Oct. 2017.

[29] Z. Ding, *et. al.* "On the Performance of Nonorthogonal Multiple Access in 5G Systems with Randomly Deployed Users," *IEEE Signal Processing Letter*, vol. 21, no. 12, pp. 1501-1505, Dec. 2014.

[30] Z. Zhang, H. Sun, and R.Q. Hu, "Downlink and Uplink Non-orthogonal Multiple Access in a Dense Wireless Network," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2771-2784, Dec. 2017.

[31] Y. Liu, *et. al.*, "Enhancing the Physical Layer Security of Non-orthogonal Multiple Access in Large-scale Networks", *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1656-1672, Mar. 2017.

[32] M. Hanif, Z. Ding, T. Ratnarajah, and G. Karagiannidis, "A Minorization-Maximization Method for Optimizing Sum Rate in the Downlink of Non-Orthogonal Multiple Access Systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 76-88, Jan. 2016.

[33] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On Optimal Power Allocation for Downlink Non-Orthogonal Multiple Access Systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2744-2757, Dec. 2017.

[34] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of User Clustering in MIMO Non-Orthogonal Multiple Access Systems," *IEEE Communications Letters*, vol. 20, no. 7, pp. 1465-1468, July 2016.

[35] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010-6023, Aug. 2016.

[36] L. Lei, D. Yuan, C.K. Ho, and S. Sun, "Joint Optimization of Power and Channel Allocation with Non-orthogonal Multiple Access for 5G Cellular Systems" in *Proc. of IEEE GLOBECOM'2015*.

[37] B. Di, S. Bayat, L. Song, and Y. Li, "Radio Resource Allocation for Downlink Non-orthogonal Multiple Access (NOMA) Networks using Matching Theory" in *Proc. of IEEE GLOBECOM'2015*.

[38] M.S. Elbamby, *et. al.*, "Resource Optimization and Power Allocation in In-band Full Duplex (IBFD)-enabled Non-orthogonal Multiple Access Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2860-2873, Dec.2017.

[39] L. Qian, *et. al.*, "Joint Uplink Base Station Association and Power Control for Small-Cell Networks with Non-Orthogonal Multiple Access," *IEEE Trans. on Wireless Communs.*, vol. 16, no. 9, pp. 5567-5582, Sept. 2017.

[40] L. Bai, L. Zhu, T. Li, J. Choi, and W. Zhuang, "An Efficient Hybrid Transmission Method: Using Non-orthogonal Multiple Access and Multiuser Diversity," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2276-2288, Oct. 2017.

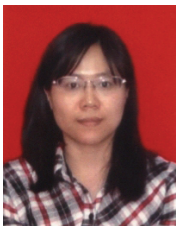
[41] S. Zhang, *et. al.*, "Sub-channel and Power Allocation for Non-orthogonal Multiple Access Relay Networks with Amplify-and-Forward Protocol," *IEEE Trans. on Wireless Communs.*, vol. 16, no. 4, pp. 2249-2261, April 2017.

[42] Y. Wu, L. Qian, H. Mao, X. Yang, and X. Shen, "Optimal Power Allocation and Scheduling for Non-Orthogonal Multiple Access Relay-Assisted Networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 11, pp. 2591-2606, Nov. 2018.

[43] Y. Liu, *et. al.*, "Cooperative Non-Orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer", *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 938-953, April 2018.

[44] Y. Xu, *et. al.*, "Joint Beamforming and Power-Splitting Control in Downlink Cooperative SWIPT NOMA Systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4874-4886, Sept. 2017.
[45] D. Tse and P. Viswanath, "Fundamentals of Wireless Communication" Cambridge University Press, 2005.
[46] R. Zhang, "Optimal Dynamic Resource Allocation for Multi-antenna Broadcasting with Heterogeneous Delay-constrained Traffic," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 2, pp. 243-255, April 2008.
[47] S. Boyd, and L. Vandenberghe, "Convex Optimization," Cambridge University Press, 2004.
[48] L. Schrage, "Optimization Modeling with LINGO," the 5th edition, Lindo System, Jan. 1999.

**Yuan Wu (S'08-M'10-SM'16)** received the Ph.D degree in Electronic and Computer Engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2010. He is currently a full professor in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. During 2016-2017, he was with the Broadband Communications Research (BBCR) group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests focus on resource management for wireless networks, green communications and computing, and smart grid. Dr. Wu was a recipient of the Best Paper Award of the IEEE International Conference on Communications in 2016, and the Best Paper Award of the IEEE Technical Committee on Green Communications & Computing in 2017.



**Li Ping Qian (S'08-M'10-SM'16)** received the PhD degree in Information Engineering from the Chinese University of Hong Hong, Hong Kong, in 2010. She worked as a postdoctoral research associate at the Chinese University of Hong Kong, Hong Kong, during 2010-2011. Since 2011, she has been with College of Information Engineering, Zhejiang University of Technology, China, where she is currently a full Professor. From 2016 to 2017, she was a visiting scholar with the Broadband Communications Research Group, ECE Department, University of Waterloo. Her research interests include wireless communication and networking, resource management in wireless networks, massive IoTs, mobile edge computing, emerging multiple access techniques, and machine learning oriented towards wireless communications. She was a co-recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2011, the Best Paper Award from IEEE ICC 2016, and the Best Paper Award from IEEE Communication Society GCCTC 2017. She is currently on the Editorial Board of IET Communications.

**Kejie Ni** is currently pursuing his M.S. degree in College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interest focuses on resource management for wireless communications and networks and non-orthogonal multiple access.

**Cheng Zhang** is currently pursuing his M.S. degree in College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interest focuses on resource management for wireless communications and networks and non-orthogonal multiple access.

**Xuemin (Sherman) Shen (M'97-SM'02-F'09)** is a University Professor and the Associate Chair for Graduate Studies, Department of Electrical and Computer Engineering, University of Waterloo, Canada. Dr. Shen's research focuses on wireless resource management, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He is the IEEE ComSoc VP Publication, was an elected member of IEEE ComSoc Board of Governor, and the Chair of Distinguished Lecturers Selection Committee. Dr. Shen served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom'16, Infocom'14, IEEE VTC'10 Fall, and Globecom'07, the Symposia Chair for IEEE ICC'10, the Tutorial Chair for IEEE VTC'11 Spring and IEEE ICC'08, the General Co-Chair for ACM Mobihoc'15, and the Chair for IEEE Communications Society Technical Committee on Wireless Communications, and P2P Communications and Networking. He also serves/served as the Editor-in-Chief for IEEE Internet of Things Journal, and IEEE Network, a Founding Area Editor for IEEE Transactions on Wireless Communications; and an Associate Editor for IEEE Transactions on Vehicular Technology and IEEE Wireless Communications, etc. Dr. Shen received the IEEE ComSoc Education Award, the Joseph LoCicero Award for Exemplary Service to Publications, the Excellent Graduate Supervision Award in 2006, and the Premiers Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society.