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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

Optimal Design and Orchestration of Mobile Edge Computing with Energy Awareness

Panagiotis D. Diamantoulakis *Senior Member, IEEE,* Pavlos S. Bouzinis, *Student Member, IEEE,* Panagiotis Sarigannidis, *Member, IEEE,* Zhiguo Ding, *Fellow, IEEE,* and George K. Karagiannidis *Fellow, IEEE*

Abstract—The wireless networks beyond the fifth generation (5G) are envisioned to be the platform that will support a vast amount of diversified data-driven applications with stringent requirements in terms of computational accuracy, delay, and energy efficiency. The fulfillment of this objective can be achieved by the convergence of communication and computing networks, enabling the exploitation of edge computing resources and the joint orchestration of the corresponding resources. Mobile edge computing (MEC), which refers to the use of edge serves for offloading tasks from mobile devices, is a particularly promising approach to provide the required computational performance for emerging internet-of-things applications, such as the smart grids, smart industry, healthcare, and smart farming. In this work, we propose the use of an advanced multiple access technique and its joint design with adaptive task offloading, in order to reduce delay and energy consumption. More specifically, the use of generalized hybrid orthogonal/non-orthogonal multiple access (OMA/NOMA) for MEC is introduced, which is theoretically superior to other alternatives from the existing literature. In more detail, the proposed scheme is based on the joint utilization of dynamic user scheduling among OMA/NOMA phases and variable decoding order during the successive interference cancellation in NOMA phase. Also, the system's orchestration is optimized for both full and partial task offloading. Specifically, in full offloading scenario, the user scheduling, time allocation, and power control are jointly optimized. Regarding partial offloading, the computational resources, i.e., the clock speed of the local processors and the number of offloaded bits, are jointly optimized with the communication resources, taking into account the constraint of the energy that is consumed for both local processing and task offloading, which is particularly challenging due to the non-convex nature of the corresponding optimization problem. All optimization problems are efficiently solved by either using closed-form solutions that provide useful insights or low-complexity algorithms. Finally, simulation results demonstrate the effectiveness of the proposed techniques and provide useful insights on the system's performance, in terms of average delay and energy consumption.

Index Terms—Mobile edge computing, task offloading, energy consumption, delay, multiple access

1 Introduction

The main objective for both industry and academia regarding wireless networks beyond the fifth generation (5G) is to increase the networks' capabilities to serve a massive amount of diversified mobile applications. Relative examples could be several emerging applications supported by artificial intelligence, such as smart grids, manufacturing, transportation, healthcare, and smart farming [1], [2]. To this direction, the convergence of communication and computing networks is envisioned, taking into account the trade-off among computational accuracy, delay, and efficient use of available energy [3]. This framework is based on programmable and flexible architectures, which are mainly based on edge-centric computing, network function virtualization, and software defined networking [4], [5]. It deserves to be mentioned that edge-centric computing is one of the key technologies for the next generation internet-of-things (IoT) [6]. Making data computation distributed and data governance decentralized can offer important advantages, mainly in terms of energy efficiency, delay, reliability, and privacy.

A promising approach for facilitating the convergence between wireless communication and computing is mobile edge computing (MEC). The concept of MEC was firstly proposed by the European Telecommunications Standard Institute (ETSI) in 2014, and was defined as a new platform that "provides information technology and cloud computing capabilities within the Radio Access Network (RAN) in close proximity to mobile subscribers" [7]. The original definition of MEC refers to the use of base stations (BSs) for offloading computation tasks from mobile devices [8]. Emerging applications supported by artificial intelligence demand low-latency and energy-efficient computing, while MEC can substantially contribute to fulfilling this requirement [9]. Since mobile devices have limited computational and energy resources, they may not be capable of completing computationally intensive tasks in the required deadline. To this end, MEC aims to employ computing facilities at the edge of mobile networks, enabling mobile devices to offload their computation tasks in this edge server [10], [11], [12], [13]. The efficient use of MEC depends on two interrelated factors, namely the utilized multiple access protocol and the efficient use of the computational and communication

P. Diamantoulakis, P. S. Bouzinis, and G. Karagiannidis are with Wireless Communication and Information Processing Group (WCIP), Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece, E-mails: {padiaman, mpouzinis, geokarag}@auth.gr

P. Sarigiannidis is with Department of Informatics and Telecommunications Engineering, University of Western Macedonia, 501 00 Kozani, Greece (e-mail: psarigiannidis@uowm.gr).

Z. Ding is with the School of Electrical and Electronic Engineering, the University of Manchester, Manchester, E-mail: zhiguo.ding@manchester.ac.uk

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957406.

resources [14], [15], [16]. Hence, their joint optimization can lead to substantial improvement of the overall performance, e.g., in terms of delay or energy consumption.

1.1 State-of-the-Art & Motivation

Energy sustainability of MEC can be improved by using non-orthogonal multiple access (NOMA), which has been recognized as an important technique for future wireless networks [17]. Several different forms of NOMA have been proposed, such as power-domain NOMA, multicarrier NOMA, distributed uplink NOMA, sparse code multiple access, and pattern division multiple access, etc. However, all of them are based on the same key concept, according to which more than one users are served in each orthogonal resource block, e.g., a time slot, a frequency channel, a spreading code, or an orthogonal spatial degree of freedom [18], [19], [20], [21]. Thus, in contrast to orthogonal multiple access (OMA), e.g., time/frequency/code division multiple access, where users that belong to the same cell are served by using different resource blocks in order to avoid interference, NOMA does not try to eliminate intra-cell interference. This approach, if carefully designed and optimized, has the potential to increase spectral efficiency and connectivity, with the latter being crucial in IoT applications [18], [19], [20]. These advantages are mainly due to the fact that NOMA avoids the exclusive use of resource blocks by users with poor channel conditions [19].

Among the different forms of NOMA, power-domain NOMA has received considerable attention due to its potential to achieve the capacity of the broadcast and multiple access channel, which correspond to downlink and uplink, respectively [19], [22]. This technique is based on the transmission of a superimposed mixture containing multiple messages. Since the power domain is utilized to achieve multiple access, multiuser detection techniques are required to retrieve the sources' signals at the receiver, such as successive interference cancellation (SIC) [23], [24]. In SIC, the signal of one user is decoded by treating the signals of other users as interference, and subtracted from the received signals if successfully decoded. Therefore, SIC can mitigate the interference due to simultaneous utilization of the system's resources. Interestingly, in uplink power-domain NOMA, it has been shown that fairness in terms of data rate allocation and energy efficiency can be improved when SIC with time-sharing is used, which is able to achieve any point of the capacity region of the multiple access channel [25], [26]. Thus, NOMA with time-sharing can be seen as a generalization of uplink NOMA with fixed decoding order, so that a user, whose message suffers from strong interference for a specific decoding order, can experience a better reception reliability for another decoding order, during the implementation of SIC. Moreover, it deserves to be noted that due to the increasing complexity of SIC with the number of multiplexed messages, a particularly interesting form of power-domain NOMA is with hybrid user-pairing, where up to two users can use the same resource block, while different pairs of users are multiplexed by using orthogonal resources [27]. Note that scheduling two users to perform NOMA is also aligned with how NOMA is implemented in Long Term Evolution Advanced (LTE-A) [28].

As it has been shown in [29], [30], [31], [32] and, the application of NOMA to MEC has the potential to assist in reducing the energy consumption and avoiding severe delay. However, it has been proved that the stand-alone use of NOMA does not necessarily improve performance in MEC computing systems, when the users do not have the same delay requirement. Thus, the use of hybrid NOMA-OMA schemes has been proposed [33], [34]. More specifically, in [33] and in [34], assuming a two users scenario and focusing on the special case that the delay and the transmit power for one of the two users is fixed, the delay and the energy consumption of the delay-tolerant user was minimized, respectively. To this end, in both works it was assumed that each frame is divided in two consecutive phases. In the first phase, both users transmit information by using NOMA, while the second phase is allocated to the delay-tolerant user in order to complete its offloading. Also, solely fixed decoding order has been considered, assuming that the message of the user with the fixed delay is decoded last, during the SIC process, i.e., in the first phase, regardless of the channel conditions and its impact on both users' performance.

Moreover, [29], [30], [33], [34] focused on a full offloading scenario, where the whole task is executed at the edge server. However, this is not always the case, since the task can be partitioned into two parts, with one executed at the device locally while the residual can be offloaded for edge execution, i.e., partial-offloading [8]. Therefore, the full offloading strategy may not be the optimal for minimizing the completion time of a task, since it may occur to increased overhead due to the communication between the devices and server. As a matter of fact, partial offloading has also been considered as an alternative strategy for MEC systems. Furthermore, partial offloading leverages the parallelism between devices and edge server, which renders it suitable for latency constrained applications [35]. For instance, in [36], [37], [38], [39], the use of NOMA for minimizing the energy consumption in MEC systems with offloading decision was examined. Furthermore, [40], [41] focused on minimizing the completion time of tasks, in a NOMA-enabled MEC network, by adopting the partial offloading technique.

Meanwhile, the device's central processing unit (CPU) clock speed for the local computation of a task's part, can be adaptively adjusted through the dynamic voltage scaling technique [35]. Thus, by controlling the local computational speed, it is possible to reduce the energy consumption or shorten the computation completion time, while the trade-off among these considered metrics can be balanced. The authors in [35], [42] jointly optimized the offloading decisions and resources, while they explored the trade-off between task execution time and energy consumption, by adaptively adjusting the CPU speed. Furthermore, in [43], the weighted sum of energy consumption and delay was minimized, in a device-to-device-assisted MEC with controllable computational speed. However, all the aforementioned works that focused on a NOMA-aided MEC system with partial offloading considered a fixed computational speed of the devices except [36], where it was shown that partial offloading with controllable CPU clock speed can reduce delay [36]. More specifically, in [36] a multi-user, multi-antenna MEC system with NOMA was investigated, in

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

order to minimize the weighted sum energy consumption of users, subject to their computation latency constraints. However, authors did not consider a hybrid NOMA scheme.

To the best of our knowledge, a hybrid NOMA configuration for MEC systems, which exploits partial offloading with controllable CPU clock speed, has not been yet investigated. Also, it is noted that all the aforementioned works on MEC with hybrid NOMA focus on fixed decoding order during the SIC process. However, the system's performance could be improved by using NOMA with time-sharing. It is highlighted that when NOMA with time-sharing is used, the resource allocation becomes more challenging and substantially different to NOMA with fixed decoding order. Moreover, exploiting a more efficient multiple access scheme and its joint optimization with the use of local computation resources has the potential to offer a meaningful performance improvement. The considered improvement, could be characterized in terms of average sum delay or energy consumption for completing computationally intensive tasks. However, both criteria are of paramount importance in the general case that the delay and the consumed energy are not predetermined for any of the users, as in [33], [34], which is a scenario that has not been taken into account in the optimal design of MEC with hybrid NOMA.

1.2 Contribution

The main scope of this work is to improve the performance of MEC in terms of task completion delay and energy consumption. To this end, we introduce a novel generalized hybrid NOMA protocol for MEC systems, while we jointly optimize the orchestration of the available communication and computation resources considering all the available degrees of freedom. All optimization problems are efficiently solved and useful insights are provided. More specifically, the contribution of this paper can be summarized as follows:

- MEC with a novel generalized hybrid NOMA-OMA protocol is proposed, according to which two phases and time-sharing during the SIC process are used. During the first phase both considered users can transmit their messages by using NOMA with time-sharing, while in the second phase solely one of the users can transmit its information by using OMA. The user that is scheduled to offload part of its task at each phase is not predetermined, but is dynamically assigned. In addition, the time that is allocated to each phase and the corresponding transmit power of each user are optimized. It is noted that hybrid NOMA-OMA with fixed decoding order can be viewed as a special case of the proposed protocol. Also, since, compared to NOMA with fixed decoding order, NOMA with time-sharing can achieve every point of the rate region of the multiple access channel, the proposed technique can serve in future research as a performance upper bound in order to evaluate MEC with other MA techniques.
- The weighted sum of users' delay is minimized when the proposed hybrid NOMA-OMA protocol is used, taking into account the energy constraints of the mobile users. Also, the extension of the provided analysis to the case of users' energy consumption minimization is considered. Moreover, both full offloading and partial offloading are taken into account. In the case of partial offloading, the CPU clock speed of the local processors and the number of offloaded bits are dynamically adjusted, which contributes in achieving shorter task completion time, without consuming more energy. All optimization problems are efficiently solved by either using closed-form solutions that provide useful insights or low-complexity algorithms. In more detail, as regards the full offloading scenario, first, we provide analytical solutions for the delay minimization problem, as well as interesting insights related with the values of the Lagrange multipliers. Also, for the energy minimization problem we show that the optimal energy consumption of each user is determined by a two-branch closed-form solution, with the decision function depending solely on the channel gains and the weights that are associated with the users' priority. On the other hand, in order to solve the challenging non-convex delay minimization problem for the case of partial offloading, the initial problem is transformed to an equivalent difference-of-convex (DC) structured one, which is efficiently solved by a successive convex approximation procedure. It deserves to be noted that the proposed solving method has the potential to facilitate the solution of similar problems in MEC computing with partial offloading.
- Finally, simulation results are provided which illustrate the effectiveness of the proposed techniques and solutions and provide useful insights on the system's performance. For example, it is shown that the use of generalized hybrid NOMA and partial offloading with control of local processors can lead to substantial reduction of delay and energy consumption compared to the use of NOMA with fixed decoding order and full offloading, respectively. Also, it is demonstrated that the average number of iterations upon convergence of the successive convex approximation procedure is particularly low, making the proposed solving method suitable for practical implementation.

1.3 Structure

The rest of the paper is organized as follows. Section II describes the communication and task offloading model, while it defines the number of bits that can be offloaded, the delay, and the energy consumption. In section III, the case of full offloading is optimized, considering both the weighted sum delay and energy consumption minimization. In section IV, the weighted sum delay is minimized for the case of partial offloading. Following that, in section V, the simulation results are demonstrated and discussed. Finally, conclusions are summarized in Section VII.

2 System Model and Proposed Protocol

An MEC offloading scenario is considered, in which two users offload their computations to an MEC server in two phases, as it is shown in Fig. 1. Perfect channel state information is assumed at the MEC server, which also performs the optimization. Let T_1 and T_2 denote the duration of the 1-st and 2-nd phase, respectively. Also, $\tilde{g}_j = |h_j|^2\omega_j$, is the channel gain of j-th user, $\forall j \in \{1,2\}$, including both small scale fading h_j, and path-loss factor ω_j. Moreover, $p_{i,j}$ and N_j, denote the transmit power of the i-th user

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

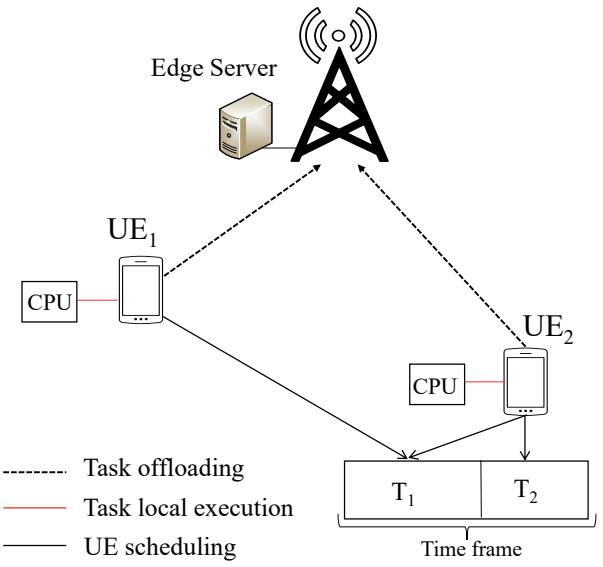

Fig. 1. An example for the users scheduling

during the i-th phase and the number of bits that need to be processed, respectively. Furthermore, in accordance to Fig. 1, where the existence of a CPU at each user is also depicted, two different cases for offloading are taken into account, namely full offloading and partial offloading. It is noted that the time and the energy costs for the server to send the outcomes of the tasks to the users are omitted, since the size of the outcomes is typically very small. Moreover, the energy consumption for the computation at the server is also omitted, as the server is not energy constrained [33].

2.1 Proposed protocol

According to the proposed protocol, termed as generalized hybrid NOMA-OMA, during the 1-st phase both users can transmit their messages by using NOMA with time-sharing, according to which the order of decoding for the users can change for specific fractions of the 1-st phase duration. On the other hand, in the 2-nd phase solely one of the users can transmit its information by using OMA. The subset of users that is scheduled to offload part of its task at each phase is not predetermined, but is dynamically assigned. In addition, the time that is allocated to each phase and the corresponding transmit power of each user are optimized. Taking into account the provided degrees of freedom, the proposed protocol can be seen as a generalization of NOMA and OMA. For the sake of clarity, it deserves to be mentioned that in NOMA, both users offload their task solely during the 1-st phase,while in OMA the 1-st phase is solely accessed by one user. Also, the proposed protocol can be seen as a generalization of hybrid NOMA-OMA with fixed decoding both users transmit information during the 1-st phase, in which one of the user's messages is always decoded first regardless of the channel conditions or the users' requirements and energy constraints.

When NOMA with time-sharing is used during the 1-st phase, the capacity of the multiple access channel is achieved and the number of bits that can be offloaded by the j-th user, denoted by $\tilde{N}_{1,j}$, is constrained by the capacity region of multiple access channel [25], [44] and is given by

$$\begin{aligned} &T_1 B \log_2\left(1+g_j p_{1,j}\right) \geq \tilde{N}_{1,j},\ \forall j \in \{1,2\}, \\ &T_1 B \log_2\left(1+\sum_{j=1}^{2} g_j p_{1,j}\right) \geq \sum_{j=1}^{2} \tilde{N}_{1,j}, \end{aligned} \tag{1}$$

where $g_j = \frac{\tilde{g}_j}{BN_0}$ is the normalized channel gain, with N_0 being the power spectral density of the additive white gaussian noise (AWGN), while B denotes the available bandwidth. Moreover, the number of bits that can be offloaded by the j-th user during the 2-nd phase, denoted by $\tilde{N}_{2,j}$ are constrained by

$$s_j T_2 B \log_2(1+g_j p_{2,j}) \geq \tilde{N}_{2,j}, \tag{2}$$

where $s_j \in \{0,1\}, \forall j \in \{1,2\}$ is a binary variable and

$$s_1 + s_2 \leq 1, \tag{3}$$

since solely one of the two users can offload data during the 2-nd phase. A specific example for the users scheduling is provided in Fig. 1, according to which the 2-nd phase is used by the 2-nd user, i.e., $s_1 = 0, s_2 = 1$. It is highlighted though that in the considered analysis the value of $s_j, \forall j \in \{1,2\}$ is subject to optimization. To what follows, the expressions for the delay and the energy consumption that correspond to the proposed protocol for both the cases of full and partial offloading are discussed.

2.1.1 Full offloading

In this scenario, it is assumed that the tasks of both users are processed solely at the edge server. It is assumed that the individual delay for each user is solely imposed by the offloading delay, i.e., $T_{\mathrm{FO},j} = T_{\mathrm{o},j}$, which is given by

$$T_{\mathrm{o},j} = T_1 + s_j T_2, \tag{4}$$

from which it becomes apparent that user j with $s_j \neq 0$ experiences higher delay. Also, the energy that is consumed by each user, i.e., $E_{\mathrm{FO},j}$, corresponds to the energy consumption for data offloading, which can be written as

$$E_{\mathrm{o},j} = T_1 p_{1,j} + T_2 p_{2,j}. \tag{5}$$

Due to the fact that when full offloading is used, the whole task needs to be offloaded, thus, it must hold that

$$\tilde{N}_{1,j} + \tilde{N}_{2,j} \geq N_j, \tag{6}$$

which can be achieved when

$$\tilde{N}_{1,j} + s_j T_2 B \log_2(1+g_j p_{2,j}) \geq N_j. \tag{7}$$

2.1.2 Partial offloading

In this scenario, it is assumed that part of each user's task can be processed locally, with the rest being offloaded to the server. The execution latency for the j-th user's sub-task $A_j(L_j, X_j)$ that is processed locally can be calculated as [8]

$$T_{\mathrm{loc},j} = \frac{L_j X_j}{f_j}, \tag{8}$$

where L_j is the sub-task input-data size in bits, f_j is CPU clock speed of the local processor, X_j (in CPU cycles per

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

bit) is the computation workload. Next, we assume that each user is able to adaptively adjust its CPU clock speed $f_j \in (0, f_{\max}], \forall j \in \{1,2\}$, by using the dynamic voltage scaling technique. We further assume that users can simultaneously offload and process tasks, considering that the data transmission between the users and the edge server can be done in parallel with the local CPU computation. Therefore, the delay for completing the j-th user's task, is given by

$$T_{\mathrm{PO},j} = \max\left(T_{\mathrm{o},j}, T_{\mathrm{loc},j}\right), \quad \forall j \in \{1,2\}, \tag{9}$$

since the total delay of a user is determined by the maximum duration of either the task's offloading or local processing. Note that by setting the locally processed sub-task L_j, as $L_j = 0$, the full offloading scenario is being enforced, since no computations are performed locally.

The energy consumption of a CPU cycle is given by $k_j f_j^2$, where k_j is a constant parameter related to the hardware architecture. For the computation sub-task $A_j(L_j, X_j)$, the energy consumption can be derived by [8]

$$E_{\mathrm{loc},j} = k_j L_j X_j f_j^2, \quad \forall j \in \{1,2\}. \tag{10}$$

Thus, the total amount of energy that is consumed by each user is given by

$$E_{\mathrm{PO},j} = E_{\mathrm{o},j} + E_{\mathrm{loc},j}. \tag{11}$$

Also, it is noted that when partial offloading is used, taking into account the number of bits that are processed locally, the constraint in (6) is replaced by

$$\tilde{N}_{1,j} + \tilde{N}_{2,j} + L_j \geq N_j. \tag{12}$$

2.2 Optimization objectives

We adopt two performance metrics to optimize the resources, namely the *weighted sum of users' delay* and the *weighted sum of users' energy consumption*. The weighted sum of users' delay is defined as

$$\mathcal{T}_{\mathrm{m}} = \sum_{j\in\{1,2\}} w_j T_{\mathrm{m},j}, \tag{13}$$

where $T_{\mathrm{PO},j}$ and $T_{\mathrm{FO},j}$ denote the individual delay of user j for the case of partial and full offloading, respectively, i.e., $\mathrm{m} \in \{\mathrm{PO}, \mathrm{FO}\}$. Also, $0 \leq w_j \leq 1$ is a positive constant provided by the upper layers that facilitates the assignment of different priorities to different users and and provides certain notions of fairness. Furthermore, the weighted sum of users' energy consumption is defined as

$$\mathcal{E}_{\mathrm{m}} = \sum_{j\in\{1,2\}} w_j E_{\mathrm{m},j}, \tag{14}$$

where $E_{\mathrm{m},j}$ is the amount of energy that is consumed by each user.

3 Full Offloading

In these section, the case of full offloading is considered, according to which each user's delay and energy consumption are given by (4) and (5), respectively. Also, two different optimization objectives are taken into account, namely energy-constrained delay minimization and delay-constrained energy minimization.

3.1 Energy-constrained delay minimization

The weighted sum of users' delay minimization can be formulated as

$$\begin{aligned}
\min_{\mathbf{T},\mathbf{p},\tilde{\mathbf{N}},\mathbf{s}} \quad & \mathcal{T}_{\mathrm{FO}} \\
\text{s.t.} \quad & \mathrm{C}_1: T_1 B \log_2\left(1 + g_j p_{1,j}\right) \geq \tilde{N}_{1,j}, \\
& \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_2: T_1 B \log_2\left(1 + \sum_{j=1}^{2} g_j p_{1,j}\right) \geq \sum_{j=1}^{2} \tilde{N}_{1,j}, \\
& \mathrm{C}_3: \tilde{N}_{1,j} + s_j T_2 B \log_2(1 + g_j p_{2,j}) \geq N_j, \\
& \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_4: E_{\mathrm{o},j} \leq E_j, \forall j \in \{1,2\}, \\
& \mathrm{C}_5: s_j \in \{0,1\}, \forall j \in \{1,2\}, \\
& \mathrm{C}_6: \textstyle\sum_{j=1}^{2} s_j \leq 1, \\
& \mathrm{C}_7: p_{i,j}, s_j, T_i \geq 0, \quad \forall i,j \in \{1,2\},
\end{aligned} \tag{15}$$

where $\mathbf{T}, \mathbf{p}, \tilde{\mathbf{N}}$, and $\mathbf{s}$ denote the vectors that include the variables T_i, $p_{i,j}$, $\tilde{N}_{i,j}$, and s_j, respectively. Also, $\mathrm{C}_1, \mathrm{C}_2$, represent the constrained number of offloaded bits in the 1-st phase, as defined in (1), while C_3 guarantees the successful offloading of the whole users' tasks until the end of the 2-nd phase. Furthermore, C_4 is associated with the maximum available energy of each user, E_j. Finally, C_6 implies that at most one user is able to transmit information during T_2.

After properly manipulating the constraints $\mathrm{C}_1, \mathrm{C}_2$ with C_3, the optimization problem in (15) can be rewritten as

$$\begin{aligned}
\min_{\mathbf{T},\mathbf{p},\mathbf{s}} \quad & \mathcal{T}_{\mathrm{FO}} \\
\text{s.t.} \quad & \mathrm{C}_1: T_1 B \log_2\left(1 + g_j p_{1,j}\right) \\
& \quad + T_2 s_j \log_2(1 + g_j p_{2,j}) \geq N_j, \\
& \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_2: T_1 B \log_2\left(1 + \textstyle\sum_{j=1}^{2} g_j p_{1,j}\right) \\
& \quad + \textstyle\sum_{j=1}^{2} s_j T_2 \log_2(1 + g_j p_{2,j}) \\
& \quad \geq \textstyle\sum_{j=1}^{2} N_j, \\
& \mathrm{C}_3: T_1 p_{1,j} + T_2 p_{2,j} \leq E_j, \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_4: s_j \in \{0,1\}, \forall j \in \{1,2\}, \\
& \mathrm{C}_5: \textstyle\sum_{j=1}^{2} s_j \leq 1, \\
& \mathrm{C}_6: p_{i,j}, s_j, T_i \geq 0, \quad \forall i,j \in \{1,2\}.
\end{aligned} \tag{16}$$

Next, to efficiently solve (16), we consider three different cases, i.e., one for each possible deployment of the users' scheduling in the two phases, which can be solved in parallel.

3.1.1 *Case 1:* $s_1 = s_2 = 0$

In this case, both users solely transmit their messages in the 1-st phase, hence the pure NOMA scheme is only utilized. Following that, the optimization problem can be written as

$$\begin{aligned}
\min_{T_1,\mathbf{p}} \quad & T_1 \\
\text{s.t.} \quad & \mathrm{C}_1: T_1 B \log_2\left(1 + g_j p_{1,j}\right) \geq N_j, \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_2: T_1 B \log_2\left(1 + \sum_{j=1}^{2} g_j p_{1,j}\right) \geq \sum_{j=1}^{2} N_j, \\
& \mathrm{C}_3: T_1 p_{1,j} \leq E_j, \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_4: p_{1,j}, T_1 \geq 0, \quad \forall j \in \{1,2\}.
\end{aligned} \tag{17}$$

To solve (17), it is useful to notice that using the constraint C_3 with equality preserves optimality. To give further insight on this, let's assume that there exists a solution

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

for $p_{1,j}$ which satisfies $p'_{1,j} < p^*_{1,j}$ and $T_1 p'_{1,j} < E_j$ and achieves lower delay. By observing C_1, C_2 and considering that $\log_2(\cdot)$ is an ascending function with respect to $p_{1,j}$, we conclude that in order to fulfill the constraints, T'_1 should satisfy $T'_1 \geq T^*_1$. This contradicts to the assumption since the goal is to minimize T_1. Following that, the optimal T_1 is given by the most stringent constraint among C_1 and C_2, when this holds with equality. Next, with direct calculations we conclude that the optimal value of T_1 is given by

$$T^*_1 = \max_{k \leq l, k, l \in \{1,2\}} \left(\frac{-\ln(2) \sum_{j=k}^{l} N_j}{B(\mathcal{W}_{-1}(-a_{k,l}\exp(-a_{k,l}))+a_{k,l})} \right), \quad (18)$$

where $a_{k,l}$ is given by

$$a_{k,l} = \frac{\sum_{j=k}^{l} N_j \ln(2)}{B \sum_{j=k}^{l} g_j E_j}, \quad (19)$$

while $\mathcal{W}_{-1}(\cdot)$ is the secondary real branch of the Lambert W function [45]. Moreover, it is noted that the optimal value for $p_{1,j}$ is not necessarily unique. Specifically, the minimum delay is achieved for

$$p^*_{1,j} = \frac{E_j}{T^*_1}, \quad \forall j \in \{1,2\}, \quad (20)$$

as well as for any other value of $p_{1,j}$ that satisfies C_1, C_2, and C_3, by replacing T_1 with the closed-form of T^*_1, for which case they become linear.

3.1.2 Case 2: $s_1 = 1$ and $s_2 = 0$

In this case, the optimization problem can be written as

$$\begin{aligned} \underset{\mathbf{T},\mathbf{p}}{\mathbf{min}} \quad & T_1(w_1+w_2) + T_2 w_1 \\ \mathbf{s.t.} \quad & \mathrm{C}_{1\mathrm{a}} : T_1 B \log_2(1+g_1 p_{1,1}) \\ & \quad +T_2 B \log_2(1+g_1 p_{2,1}) \geq N_1, \\ & \mathrm{C}_{1\mathrm{b}} : T_1 B \log_2(1+g_2 p_{1,2}) \geq N_2, \\ & \mathrm{C}_2 : T_1 B \log_2\left(1+\textstyle\sum_{j=1}^{2} g_j p_{1,j}\right) \\ & \quad +T_2 B \log_2(1+g_1 p_{2,1}) \geq \textstyle\sum_{j=1}^{2} N_j, \\ & \mathrm{C}_{3\mathrm{a}} : T_1 p_{1,1} + T_2 p_{2,1} \leq E_1 \\ & \mathrm{C}_{3\mathrm{b}} : T_1 p_{1,2} \leq E_2, \\ & \mathrm{C}_4 : p_{1,1}, p_{2,1}, p_{1,2}, T_1, T_2 \geq 0. \end{aligned} \quad (21)$$

The problem is non-convex, due to the coupling of $\mathbf{T}$ and $\mathbf{p}$. However, by setting, $E_{1,1} = T_1 p_{1,1}$, $E_{2,1} = T_2 p_{2,1}$, and $E_{1,2} = T_1 p_{1,2}$ the optimization problem in (21) can be rewritten as

$$\begin{aligned} \underset{\mathbf{T},\mathbf{E},z}{\mathbf{min}} \quad & T_1(w_1+w_2) + T_2 w_1 \\ \mathbf{s.t.} \quad & \mathrm{C}_{1\mathrm{a}} : T_1 B \log_2\left(1+g_1\tfrac{E_{1,1}}{T_1}\right) \\ & \quad +T_2 B \log_2\left(1+g_1\tfrac{E_{2,1}}{T_2}\right) \geq N_1, \\ & \mathrm{C}_{1\mathrm{b}} : T_1 B \log_2\left(1+g_2\tfrac{E_{1,2}}{T_1}\right) \geq N_2, \\ & \mathrm{C}_2 : T_1 B \log_2\left(1+\tfrac{z}{T_1}\right) \\ & \quad +T_2 B \log_2\left(1+g_1\tfrac{E_{2,1}}{T_2}\right) \geq \textstyle\sum_{j=1}^{2} N_j, \\ & \mathrm{C}_{3\mathrm{a}} : E_{1,1} + E_{2,1} \leq E_1 \\ & \mathrm{C}_{3\mathrm{b}} : E_{1,2} \leq E_2, \\ & \mathrm{C}_4 : z = g_1 E_{1,1} + g_2 E_{1,2}, \\ & \mathrm{C}_5 : E_{1,1}, E_{2,1}, E_{1,2}, T_1, T_2, z \geq 0. \end{aligned} \quad (22)$$

where $\mathbf{E}$ denotes the vector that include the variables $E_{i,j}$ Also, in (22), the auxiliary variable z was introduced, which corresponds to the constraint C_4. It's easy to verify that the problem in (22) is convex, since the objective function is linear while the constraints are convex and affine. Furthermore, optimality requires that $\mathrm{C}_{3\mathrm{a}}$ and $\mathrm{C}_{3\mathrm{b}}$ hold with equality. Following that, the Lagrange dual decomposition will be applied [46], where the Lagrangian is written as follows

$$\begin{aligned} &\mathcal{L}(T_1, T_2, E_{2,1}, z, \boldsymbol{\lambda}) = T_1(w_1+w_2) + w_1 T_2 + \lambda_1 \times \\ &\Big(N_1 - T_1 B \log_2\Big(1+g_1\tfrac{E_{1,1}}{T_1}\Big) - T_2 B \log_2\Big(1+g_1\tfrac{E_{2,1}}{T_2}\Big)\Big) \\ &+\lambda_2\Big(N_2 - T_1 B \log_2\Big(1+g_2\tfrac{E_2}{T_1}\Big)\Big) \\ &+\lambda_3\Big(N_1+N_2 - T_1 B \log_2\Big(1+\tfrac{z}{T_1}\Big) \\ &-T_2 B \log_2\Big(1+g_1\tfrac{E_{2,1}}{T_2}\Big)\Big) + \lambda_4(z - g_1 E_{1,1} - g_2 E_{1,2}), \end{aligned} \quad (23)$$

where $E_{1,1} = E_1 - E_{2,1}$, $E_{1,2} = E_2$ and $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3, \lambda_4]$ is the Lagrange multiplier vector, with $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$, corresponding to the constraints $\mathrm{C}_{1\mathrm{a}}, \mathrm{C}_{1\mathrm{b}}$, C_2 and C_4 respectively.

Given the Lagrangian function in (23), the complementary slackness conditions require [46]

$$\begin{aligned} &\lambda^*_1\Big(N_1 - T^*_1 B \log_2\Big(1+g_1\tfrac{E^*_{1,1}}{T^*_1}\Big) \\ &-T^*_2 B \log_2\Big(1+g_1\tfrac{E^*_{2,1}}{T^*_2}\Big)\Big) = 0, \end{aligned} \quad (24)$$

$$\lambda^*_2\left(N_2 - T^*_1 B \log_2\left(1+g_2\tfrac{E_2}{T^*_1}\right)\right) = 0, \quad (25)$$

and

$$\begin{aligned} &\lambda^*_3\Big(N_1+N_2 - T^*_1 B \log_2\Big(1+\tfrac{z^*}{T^*_1}\Big) \\ &-T^*_2 B \log_2\Big(1+g_1\tfrac{E^*_{2,1}}{T^*_2}\Big)\Big) = 0. \end{aligned} \quad (26)$$

In addition, according to the Karush-Kuhn-Tucker conditions, the optimal variables satisfy

$$\nabla\mathcal{L}(T^*_1, T^*_2, E^*_{2,1}, z^*, \boldsymbol{\lambda}^*) = 0, \quad (27)$$

$$z^* = g_1(E_1 - E^*_{2,1}) + g_2 E_2. \quad (28)$$

Firstly, taking $\frac{\partial\mathcal{L}}{\partial T_2} = 0$ yields to

$$w_1 - B(\lambda_1+\lambda_3)\frac{-\frac{g_1 E_{2,1}}{T_2} + \left(1+\frac{g_1 E_{2,1}}{T_2}\right)\ln\left(1+\frac{g_1 E_{2,1}}{T_2}\right)}{\ln(2)\left(1+\frac{g_1 E_{2,1}}{T_2}\right)} = 0, \quad (29)$$

while with direct calculations, the optimal value of T_2 is given by

$$T^*_2 = \frac{g_1 E^*_{2,1}}{a}, \quad (30)$$

with a being given by

$$a = -1 - \left[\mathcal{W}_0\left(-\exp\left(-1 - \frac{w_1\ln(2)}{B(\lambda^*_1+\lambda^*_3)}\right)\right)\right]^{-1} \quad (31)$$

and $\mathcal{W}_0$ being the principal real branch of the Lambert W function. Following that, $\frac{\partial\mathcal{L}}{\partial z} = 0$, gives

$$z^* = \frac{B\lambda^*_3 - \ln(2)\lambda^*_4}{\ln(2)\lambda^*_4} T^*_1, \quad \lambda^*_4 \neq 0, \quad (32)$$

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

while $\frac{\partial \mathcal{L}}{\partial E_{2,1}} = 0$, gives

$$E_{2,1}^* = \frac{cE_1 g_1 + T_1^*(c - Bg_1\lambda_1^*)}{cg_1}, \tag{33}$$

where

$$c = \left((\lambda_1^* + \lambda_3^*)\frac{Bg_1}{1+a} + g_1\lambda_4^*\right)\ln(2). \tag{34}$$

Next, we consider the case $\lambda_2^* \neq 0$, which indicates that the constraint $\mathrm{C_{1b}}$ will be active. Therefore, from (25), we conclude to

$$T_1^* B \log_2\left(1 + g_2\frac{E_2}{T_1^*}\right) = N_2. \tag{35}$$

Note that this case, implies that user 2 experiences no interference from user 1, as can been seen in (35). As a matter of fact, user 2 will be decoded second for the whole T_1 duration, which brings on the standard fixed decoding order scenario. Hence, in this case, the time-sharing strategy is reduced to the fixed decoding order. Next, we consider the case $\lambda_2^* = 0$. By combining (28) and (32), T_1^* can be calculated. Therefore, taking into account the two aforementioned cases, i.e., $\lambda_2 \neq 0$ and $\lambda_2 = 0$, the optimal T_1^* can be written as

$$T_1^* = \begin{cases} -\frac{N_2 E_2 g_2 \ln(2)}{N_2\ln(2) + BE_2 g_2 \mathcal{W}_{-1}(b)}, & \lambda_2^* \neq 0, \\ \frac{cg_1 g_2 E_2}{cg_1\frac{B\lambda_3^* - \ln(2)\lambda_4^*}{\ln(2)\lambda_4^*} + g_1(c - Bg_1\lambda_1^*)}, & \lambda_2^* = 0, \end{cases} \tag{36}$$

where

$$b = -\frac{N_2\ln(2)2^{\frac{-N_2}{BE_2 g_2}}}{BE_2 g_2}. \tag{37}$$

It is noted that for the case of $\lambda_2^* \neq 0$, closed-form solutions in terms of the Lagrange multipliers have been derived, while the Lagrange multipliers can be calculated iteratively with the aid of the sub-gradient method [46], in order to find $E_{2,1}^*, T_2^*$ and z^*. Also, for the case of $\lambda_2^* = 0$, we will prove by contradiction that, λ_1^* and λ_3^* satisfy $\lambda_1^* \neq 0$ and $\lambda_3^* \neq 0$. Firstly, let $\lambda_3^* = 0$. From (32) we conclude to $z^* = -T_1^*$, which holds only when $z^* = T_1^* = 0$, while this is an infeasible solution. Thus, λ_3^* satisfies $\lambda_3^* \neq 0$. Next, we assume that $\lambda_1^* = 0$. From (32) and (33), we derive

$$\frac{z^*}{T_1^*} = a = g_1\frac{E_{2,1}^*}{T_2^*}, \tag{38}$$

while $\frac{\partial \mathcal{L}}{\partial T_1} = 0$, for $\lambda_2^* = \lambda_1^* = 0$, gives

$$\sum_{j=1}^{2} w_j - B\lambda_3^* \frac{-\frac{z^*}{T_1^*} + \left(1 + \frac{z^*}{T_1^*}\right)\ln\left(1 + \frac{z^*}{T_1^*}\right)}{\ln(2)\left(1 + \frac{z^*}{T_1^*}\right)} = 0. \tag{39}$$

From (29), (39), (38) we conclude to $w_2 = 0$, which is not valid, since w_2 is an arbitrary constant. Thus, $\lambda_1^* \neq 0$. Since $\lambda_1^* \neq 0$ and $\lambda_3^* \neq 0$, by manipulating the complementary slackness conditions (24), (26), we conclude to

$$T_1^* B \log_2\left(1 + \frac{g_2 E_2}{T_1^* + g_1(E_1 - E_{2,1}^*)}\right) = N_2. \tag{40}$$

Given T_1^* from the second branch of (36) and $E_{2,1}^*$ from (33), (40) can be used in order to update the Lagrange multipliers with the aid of the sub-gradient method and finally calculate the optimal variables.

3.1.3 *Case 3:* $s_1 = 0$ *and* $s_2 = 1$

The solution is similar to the case 2, since the considered problems exhibit symmetric structure.

3.2 Delay-constrained energy minimization

It is noted that the analysis and all the optimization problems presented in this paper can easily be extended to the case of energy minimization. As an example, the problem for minimizing the weighted sum of users' energy consumption when full offloading is used, i.e., $\mathcal{E}_{\mathrm{FO}}$, subject to their latency constraints, by setting $E_{i,j} = T_i p_{i,j}$ as in (22), can be formulated as follows

$$\begin{aligned} \min_{\mathbf{T},\mathbf{E},\mathbf{s}} \quad & w_1(E_{1,1} + s_1 E_{2,1}) + w_2(E_{1,2} + s_2 E_{2,2}) \\ \text{s.t.} \quad & \mathrm{C}_1: T_1 B\log_2\left(1 + g_j\tfrac{E_{1,j}}{T_1}\right) \\ & \quad + T_2 s_j \log_2\left(1 + g_j\tfrac{E_{2,j}}{T_2}\right) \geq N_j, \\ & \quad \forall j \in \{1,2\}, \\ & \mathrm{C}_2: T_1 B\log_2\left(1 + \textstyle\sum_{j=1}^{2} g_j\tfrac{E_{1,j}}{T_1}\right) \\ & \quad + \textstyle\sum_{j=1}^{2} s_j T_2 \log_2(1 + g_j\tfrac{E_{2,j}}{T_2}) \\ & \quad \geq \textstyle\sum_{j=1}^{2} N_j, \\ & \mathrm{C}_3: T_1 + s_j T_2 \leq T^{(j)}, \quad \forall j \in \{1,2\}, \\ & \mathrm{C}_4: s_j \in \{0,1\}, \forall j \in \{1,2\}, \\ & \mathrm{C}_5: \textstyle\sum_{j=1}^{2} s_j \leq 1, \\ & \mathrm{C}_6: E_{i,j}, s_j, T_i \geq 0, \quad \forall i,j \in \{1,2\}, \end{aligned} \tag{41}$$

where $T^{(j)}$ denotes the delay deadline of the j-th user.

This optimization problem can be solved similarly as the one in (15). Thus, in order to avoid repetition solely the closed-form solutions for the case of $s_1 = s_2 = 0$ will be presented, which is particularly interesting due to existence of closed-form solutions. In this case, it can be rewritten as

$$\begin{aligned} \min_{T_1,\mathbf{E}} \quad & w_1 E_{1,1} + w_2 E_{1,2} \\ \text{s.t.} \quad & \mathrm{C}_1: T_1 B\log_2\left(1 + g_j\tfrac{E_{1,j}}{T_1}\right) \geq N_j, \ \forall j \in \{1,2\}, \\ & \mathrm{C}_2: T_1 B\log_2\left(1 + \textstyle\sum_{j=1}^{2} g_j\tfrac{E_{1,j}}{T_1}\right) \geq \textstyle\sum_{j=1}^{2} N_j, \\ & \mathrm{C}_3: T_1 \leq T^{(j)}, \quad \forall j \in \{1,2\}, \\ & \mathrm{C}_4: E_{1,j}, T_1 \geq 0, \quad \forall j \in \{1,2\}, \end{aligned} \tag{42}$$

To solve (42), first let's assume that there exists an optimal solution of T_1, denoted by T_1', which satisfies $T_1' < T^{(j)}, \forall j \in \{1,2\}$. By observing C_1, C_2 and considering that the functions appearing in these constraints are ascending with respect to T_1, we conclude that in order to fulfill the constraints $E_{1,1}'$ and $E_{1,2}'$ would increase. This contradicts to the assumption since the goal is to minimize the weighted sum of consumed energy. Thus, T_1^* is given by the most stringent inequality among the ones in C_3.

Next, we assume that C_2 holds with equality, from which it holds that

$$E_{1,1} = T_1^*\frac{2^{\frac{N_1+N_2}{T_1^* B}} - 1}{g_1} - \frac{g_2 E_{1,2}}{g_1}. \tag{43}$$

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

Thus, for given optimal T_1^* and by using (43), the optimization problem in (42) can be reformulated as

$$\begin{aligned} \min_{E_{1,2}} \quad & (w_2 - \tfrac{w_1 g_2}{g_1}) E_{1,2} \\ \text{s.t.} \quad & \mathrm{C}_1 : E_{1,2} \geq \tfrac{(2^{\frac{N_2}{T_1^* B}} - 1) T_1^*}{g_2}, \\ & \mathrm{C}_2 : E_{1,2} \leq 2^{\frac{N_1}{T_1^* B}} \tfrac{(2^{\frac{N_2}{T_1^* B}} - 1) T_1^*}{g_2}. \end{aligned} \tag{44}$$

It's easy to verify that the upper bound of $E_{1,2}$ is always greater than the lower one. Thus, when $w_2 - \frac{w_1 g_2}{g_1} \geq 0$, $E_{1,2}^*$ will be equal to the lower bound, while when $w_2 - \frac{w_1 g_2}{g_1} < 0$, $E_{1,2}^*$ will be equal to the upper bound, in order to minimize the objective.

Next, it is assumed that C_2 holds with strict inequality. We will show that this case leads to a non-optimal solution. Firstly, let C_1 hold with equality, $\forall j \in \{1,2\}$, i.e.,

$$E_{1,j} = \frac{(2^{\frac{N_j}{T_1^* B}} - 1) T_1^*}{g_j}, \quad \forall j \in \{1,2\}. \tag{45}$$

By using (45) and C_2, we conclude to

$$(2^{\frac{N_1}{T_1^* B}} - 1)(1 - 2^{\frac{N_2}{T_1^* B}}) \geq 0, \tag{46}$$

which is not valid, since this product is always negative for $N_1, N_2 \neq 0$. Hence, C_1 cannot hold with equality, $\forall j \in \{1,2\}$. Following that, let C_1 hold with equality only for $j = 1$. Therefore, the solution for $E_{1,1}$ is

$$E'_{1,1} = \frac{(2^{\frac{N_1}{T_1^* B}} - 1) T_1^*}{g_1} \tag{47}$$

and (42) can be reformulated as

$$\begin{aligned} \min_{E_{1,2}} \quad & E_{1,2} \\ \text{s.t.} \quad & \mathrm{C}_1 : E_{1,2} > \tfrac{(2^{\frac{N_2}{T_1^* B}} - 1) T_1^*}{g_2}, \\ & \mathrm{C}_2 : E_{1,2} > 2^{\frac{N_1}{T_1^* B}} \tfrac{(2^{\frac{N_2}{T_1^* B}} - 1) T_1^*}{g_2}. \end{aligned} \tag{48}$$

Following that, we conclude that the optimal $E'_{1,2}$ should satisfy

$$E'_{1,2} > 2^{\frac{N_1}{T_1^* B}} \frac{(2^{\frac{N_2}{T_1^* B}} - 1) T_1^*}{g_2}. \tag{49}$$

By comparing this solution with the one of (44) for $w_2 < \frac{w_1 g_2}{g_1}$, it is obvious that is not optimal, since $E'_{1,1} = E^*_{1,1}$ and $E'_{1,2} > E^*_{1,2}$. Based on this, it becomes evident that the solution of $E'_{1,j}$ in (47) and (49) is suboptimal also for $w_2 \geq \frac{w_1 g_2}{g_1}$, since in this case the solution of (44) for $w_2 < \frac{w_1 g_2}{g_1}$ is dominated by the solution of (44) for $w_2 \geq \frac{w_1 g_2}{g_1}$. Similar conclusions can be drawn if C_1 holds with equality only for $j = 2$. As a consequence, C_2 will always hold with equality.

In conclusion, the optimal values of T_1 and $\mathbf{E}$ are

$$T_1^* = \min_{j \in \{1,2\}} \left(T^{(j)} \right), \tag{50}$$

$$E^*_{1,1} = \begin{cases} 2^{\frac{N_2}{T_1^* B}} \frac{(2^{\frac{N_1}{T_1^* B}} - 1) T_1^*}{g_1}, & w_2 \geq \frac{w_1 g_2}{g_1}, \\ \frac{(2^{\frac{N_1}{T_1^* B}} - 1) T_1^*}{g_1}, & w_2 < \frac{w_1 g_2}{g_1}, \end{cases} \tag{51}$$

and

$$E^*_{1,2} = \begin{cases} \frac{(2^{\frac{N_2}{T_1^* B}} - 1) T_1^*}{g_2}, & w_2 \geq \frac{w_1 g_2}{g_1}, \\ 2^{\frac{N_1}{T_1^* B}} \frac{(2^{\frac{N_2}{T_1^* B}} - 1) T_1^*}{g_2}, & w_2 < \frac{w_1 g_2}{g_1}. \end{cases} \tag{52}$$

The closed-form solutions in (51) and (52) are particularly interesting, since it is proved that the decision function for optimizing the energy consumption for each user depends solely on the channel gains ($g_j, j \in \{1,2\}$) and the weights ($w_j, j \in \{1,2\}$). It is also noted that the expression for the optimal value for $E^*_{1,1}$ given in (51) has the same form with $E^*_{1,2}$ in (52), in the case of the complementary event regarding the ordering of the weighted channel gains.

4 Partial Offloading with Controllable CPU Clock Speed

In these section, the case of partial offloading is considered, according to which each user's delay and energy consumption are given by (9) and (11), respectively. The main focus of the provided analysis is on energy-constrained delay minimization, however, the problem formulation and solution can easily be extended to the case of delay-constrained energy minimization. It is noted that partial offloading is particularly interesting because it calls for the joint optimization of different types of resources, i.e., communications and computational resources, and introduces some non-trivial trade-offs. For example, by increasing the CPU speed of the local processor, f_j, reduces latency without increasing the energy that is used for information transmission and potentially increasing interference, but it also increases the energy that is consumed locally. The optimization problem for minimizing the weighted sum delay of completing the users' tasks, when the partial offloading scenario is considered, can be formally written as

$$\begin{aligned} \min_{\mathbf{T},\mathbf{p},\tilde{\mathbf{N}},\mathbf{s},\mathbf{L},\mathbf{f}} \quad & \mathcal{T}_{\mathrm{PO}} \\ \text{s.t.} \quad & \mathrm{C}_1 : T_1 B \log_2 (1 + g_j p_{1,j}) \geq \tilde{N}_{1,j}, \forall j \in \{1,2\}, \\ & \mathrm{C}_2 : T_1 B \log_2 \left(1 + \sum_{j=1}^{2} g_j p_{1,j} \right) \geq \sum_{j=1}^{2} \tilde{N}_{1,j}, \\ & \mathrm{C}_3 : \tilde{N}_{1,j} + L_j + s_j T_2 B \log_2 (1 + g_j p_{2,j}) \geq N_j, \\ & \quad \forall j \in \{1,2\}, \\ & \mathrm{C}_4 : E_{\mathrm{PO},j} \leq E_j, \forall j \in \{1,2\}, \\ & \mathrm{C}_5 : s_j \in \{0,1\}, \forall j \in \{1,2\}, \\ & \mathrm{C}_6 : \textstyle\sum_{j=1}^{2} s_j \leq 1, \\ & \mathrm{C}_7 : p_{i,j}, s_j, T_i, f_j \geq 0, \forall i,j \in \{1,2\}, \\ & \mathrm{C}_8 : f_j \leq f_{\max}, \forall j \in \{1,2\}, \end{aligned} \tag{53}$$

where $\mathbf{L}, \mathbf{f}$ denote the vectors which correspond to the variables, $L_j, f_j, \forall j \in \{1,2\}$, respectively. C_3 guarantees the successful processing of the whole users' tasks until the end of the 2-nd phase, taking into account that part of which is performed locally. Also, C_4 ensures that the amount of the consumed energy for offloading computation tasks to the edge server, plus the consumed energy for local processing of the residual tasks, cannot exceed the available energy of the user, E_j, $\forall j \in \{1,2\}$. It should be noted, that the length of the locally processed sub-task A_j is also subject to the optimization, specifying the number of bits, i.e., L_j, that will

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

be executed locally. Finally, C_8 indicates the maximum CPU clock speed that each user can utilize.

Following that, the problem can be written in its equivalent epigraph form as

$$\begin{aligned}
\min_{\mathbf{T},\mathbf{p},\mathbf{N},\mathbf{s},\mathbf{L},\mathbf{f},\mathbf{y}} \quad & w_1 y_1 + w_2 y_2 \\
\text{s.t.} \quad & \mathrm{C}_1 : T_1 B \log_2 \left(1+\rho g_j p_{1,j}\right) \\
& \quad + s_j T_2 B \log_2(1+\rho g_j p_{2,j}) + L_j \geq N_j, \\
& \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_2 : T_1 B \log_2 \left(1+\rho \sum_{j=1}^{2} g_j p_{1,j}\right) \\
& \quad + \textstyle\sum_{j=1}^{2} s_j T_2 B \log_2(1+\rho g_j p_{2,j}) \\
& \quad + \textstyle\sum_{j=1}^{2} L_j \geq \sum_{j=1}^{2} N_j, \\
& \mathrm{C}_3 : \textstyle\sum_{i=1}^{2} T_i p_{i,j} + k_j L_j X_j f_j^2 \leq E_j, \forall j \in \{1,2\}, \\
& \mathrm{C}_4 : s_j \in \{0,1\}, \forall j \in \{1,2\}, \\
& \mathrm{C}_5 : \textstyle\sum_{j=1}^{2} s_j \leq 1, \\
& \mathrm{C}_6 : p_{i,j}, s_j, T_i, f_j \geq 0, \quad \forall i,j \in \{1,2\} \\
& \mathrm{C}_7 : y_j \geq T_1 + s_j T_2, \forall j \in \{1,2\} \\
& \mathrm{C}_8 : y_j \geq \tfrac{L_j X_j}{f_j}, \quad \forall j \in \{1,2\} \\
& \mathrm{C}_9 : f_j \leq f_{\max}, \quad \forall j \in \{1,2\}
\end{aligned} \tag{54}$$

where $\mathbf{y} = \{y_1, y_2\}$. The problem in (54) is non-convex, due to the presence of the f_j^2 term in C_3 and the coupling of $\mathbf{p}$ and $\mathbf{T}$ in C_1, C_2. Therefore, by setting $E_{1,j} = T_1 p_{1,j}$, $E_{2,j} = T_2 p_{2,j}$, $L_j = \exp(\tilde{L}_j)$, and $f_j = \exp(\tilde{f}_j)$, the problem can be re-formulated as follows

$$\begin{aligned}
\min_{\mathbf{T},\mathbf{E},\mathbf{s},\tilde{\mathbf{L}},\tilde{\mathbf{f}},\mathbf{y}} \quad & w_1 y_1 + w_2 y_2 \\
\text{s.t.} \quad & \mathrm{C}_1 : T_1 B \log_2 \left(1+g_j \tfrac{E_{1,j}}{T_1}\right) \\
& \quad + s_j T_2 B \log_2 \left(1+g_j \tfrac{E_{2,j}}{T_2}\right) + \exp(\tilde{L}_j) \\
& \quad \geq N_j, \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_2 : T_1 B \log_2 \left(1+\sum_{j=1}^{2} g_j \tfrac{E_{1,j}}{T_1}\right) \\
& \quad + \textstyle\sum_{j=1}^{2} s_j T_2 B \log_2 \left(1+g_j \tfrac{E_{2,j}}{T_2}\right) \\
& \quad + \textstyle\sum_{j=1}^{2} \exp(\tilde{L}_j) \geq \sum_{j=1}^{2} N_j, \\
& \mathrm{C}_3 : E_{1,j} + E_{2,j} + k_j X_j \exp(\tilde{L}_j + 2\tilde{f}_j) \leq E_j, \\
& \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_4 : s_j \in \{0,1\}, \forall j \in \{1,2\}, \\
& \mathrm{C}_5 : \textstyle\sum_{j=1}^{2} s_j \leq 1, \\
& \mathrm{C}_6 : E_{i,j}, s_j, T_i \geq 0, \quad \forall i,j \in \{1,2\}, \\
& \mathrm{C}_7 : y_j \geq T_1 + s_j T_2, \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_8 : y_j \geq \exp(\tilde{L}_j - \tilde{f}_j) X_j, \quad \forall j \in \{1,2\}, \\
& \mathrm{C}_9 : \exp(\tilde{f}_j) \leq f_{\max}, \quad \forall j \in \{1,2\},
\end{aligned} \tag{55}$$

and C_3 is now convex. However, the problem is still non-convex, even for fixed values of s_j, since C_1 and C_2 are non-convex. However, the left-hand-side term of C_1 and C_2 is clearly a difference of concave (DC) functions, since $\exp(\tilde{L}_j)$ can be written as $-(-\exp(\tilde{L}_j))$. To this end, by exploiting the DC structure of the problem, we use successive convex approximation procedure, that approximates the exponential term of $\tilde{L}_j$, by using its first order Taylor series approximation. Thus,

$$\exp(\tilde{L}_j) \simeq g(\tilde{L}_j, \bar{L}_j) = \exp(\bar{L}_j)(1+\tilde{L}_j - \bar{L}_j), \quad \forall j \in \{1,2\}, \tag{56}$$

where $g(\tilde{L}_j, \bar{L}_j)$ is the first order Taylor series approximation of the function $\exp(\tilde{L}_j)$, around $\bar{L}_j$. Following that, $\exp(\tilde{L}_j)$ can be replaced in C_1 and C_2 by $g(\tilde{L}_j, \bar{L}_j)$.

Algorithm 1 Solution of optimization problem in (53)

1: **Initialize** A, ϵ, $\bar{\mathbf{L}}$
2: **while** $A > \epsilon$ **do**
3: **Solve** problem in (55), for fixed $\mathbf{s}$
4: **Update** $\tilde{\mathbf{L}}^*$
5: $A = \|\tilde{\mathbf{L}}^* - \bar{\mathbf{L}}\|_2^2$
6: $\bar{\mathbf{L}} \leftarrow \tilde{\mathbf{L}}^*$
7: **end while**
8: **Output** $\mathbf{T}^*$, $\mathbf{E}^*$, $\tilde{\mathbf{L}}^*$, $\tilde{\mathbf{f}}^*$, $\mathbf{y}^*$.

In the continue, we develop the Algorithm 1, which in a practical scenario runs in the MEC server, which has the required computing capabilities. To this end, the primary non-convex problem in (53) is approximated iteratively by convex optimization problems [47], given the DC transformation of the primary problem, to the one in (55). It should be noted that this algorithm will be executed for all possible values of the user-scheduling vector $\mathbf{s}$, while the optimal case will be finally saved. Moreover, in each iteration stage, standard convex-optimization methods, such as interior point, may be employed in order to solve (55). It is known that the interior point method present a polynomial-time complexity [46]. Furthermore, the successive convex approximation, which is executed throughout the "while" loop, has a linear convergence rate [48]. Therefore, the algorithm presents a polynomial-time complexity. In addition, the optimal primal variables $\mathbf{L}^*$ and $\mathbf{f}^*$ are given by $\mathbf{L}^* = \exp(\tilde{\mathbf{L}}^*)$ and $\mathbf{f}^* = \exp(\tilde{\mathbf{f}}^*)$, respectively.

5 Simulation Results and Discussion

For the simulations results, we assume that the available bandwidth B is 1MHz. We define the metric $E_{0,j} = \frac{E_j \omega_j}{B N_0}$, as the average received energy in each second, which incorporates the noise power spectral density and the path-loss factor ω_j. Next, we set $E_0 = E_{0,1} = E_{0,2}$, unless specified otherwise, while in the Monte Carlo simulations the small scale fading is given by the complex random variable $h_j \sim \mathcal{CN}(0,1)$. Finally, the weighting factors have been set as $w_1 = w_2 = 1$.

Regarding the considered benchmark, the user scheduling in each phase is predetermined while the decoding order among users in the 1-st phase, is considered to be fixed. In addition, without loss of generality, we assume for the benchmark that user 2 is scheduled to access the channel for T_1 time duration, while user 1 is able to transmit in both phases, i.e., for a duration $T_1 + T_2$. Although the considered benchmark is based on [33], [34], its performance is superior. To give further insight on the selection of the considered benchmark, it is noted that since a fixed delay and transmit power has been assumed for one of the users in [33], [34], a direct comparison of the proposed protocol to [33], [34] would be unfair against [33], [34] and would heavily depend on the selection of the fixed value of delay. Thus, compared to the considered benchmark as well as [33], [34], in the proposed protocol the user which is scheduled to offload

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

TABLE 1
Characteristics of the proposed protocol and benchmarks.

Characteristics	Proposed Protocol	Fixed decoding order (in Fig. 2)	Benchmark
User Scheduling between NOMA and OMA phases	✓	✓	×
Variable decoding order during NOMA phase	✓	×	×

its tasks in each phase, is not predetermined. Also, when NOMA scheme is utilized in T_1 phase, the decoding order among users is not fixed, while the time-sharing technique is considered, which enlarges the users' capacity region [25]. Moreover, it should be noted that the fixed decoding order scheme, as illustrated in Fig. 2, is also a special case of the proposed protocol, since it enables user scheduling but not time-sharing. Table 1 summarizes the characteristics of the proposed protocol and the baseline schemes.

The illustrative example in Fig. 2 exhibits the two-fold gain of the proposed protocol for a specific channels realization. For this example, we set $N_1 = N_2 = 0.5\,\text{Mbit}$ and $|h_1|^2 = 0.1$, $|h_2|^2 = 3$. It can be observed, that the proposed protocol overlaps with the fixed decoding order one (red line), when $E_0 > 24\text{dB}$. This, lies to the fact that the maximum performance can be achieved by using the decoding order that corresponds to one of the corner points of the capacity region, i.e., by using a fixed decoding order. Hence, the performance gain of the proposed protocol over the benchmark, is achieved only due to the user scheduling. On the other hand, for values lower than 24dB the delay is minimized for an intermediate point of the capacity region, which can achieved by altering the decoding order within a frame, by using the time-sharing technique. Therefore, the proposed protocol's gain may occur due to both the dynamic user scheduling and the decoding order strategy.

Moreover, to generalize the conclusions that have been derived form from Fig. 2 regarding the comparison of the proposed protocol to the benchmark, Fig. 3 is provided. In more detail, Fig. 3 demonstrates the performance of the proposed protocol in comparison with the benchmark protocol, which have been extracted via Monte Carlo simulations. From Fig. 3, it is observed, the proposed protocol outperforms the benchmark, in terms of minimum average delay deadline. We further observe that, as the number of user's 1 offloaded bits increases, the performance gain of the proposed protocol is being reduced. This is reasonable, since user 1 is more likely to be scheduled for transmission in both phases, similarly to the benchmark protocol, owing this to the increased number of bits to be offloaded. Finally, it is a general observation that as the energy consumption requirement becomes more stringent, the average sum delay increases. This is due to the fact that the number of offloaded-bits is an increasing function of both the time delay and the consumed energy. Thus, in order to offload the same number of bits with reduced energy consumption, the time delay increases.

Next, the impact of users' distance from the BS is investigated via Monte Carlo simulations. Specifically, we assume that the user which is closer to the BS, has an average received energy E_0, while for the far user it has

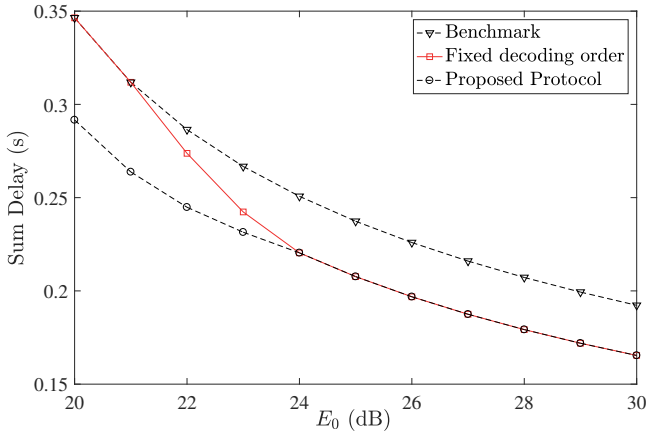

Fig. 2. The two-fold gain of the proposed protocol in terms of delay reduction for the full offloading scenario.

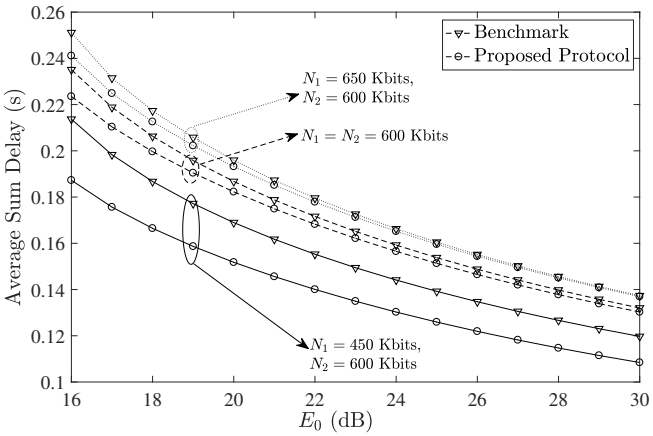

Fig. 3. Sum delay of the full offloading scenario.

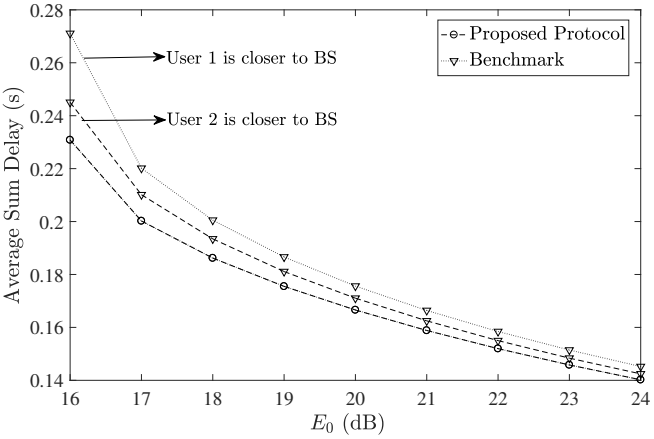

Fig. 4. Impact of users' position on delay for the full offloading scenario.

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

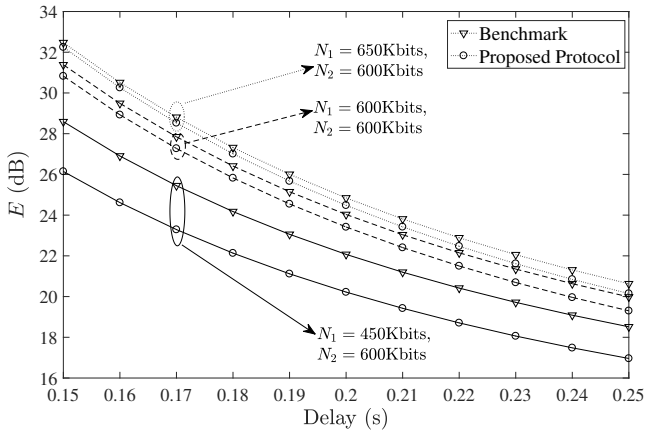

Fig. 5. Average sum energy consumption versus delay for the full offloading scenario.

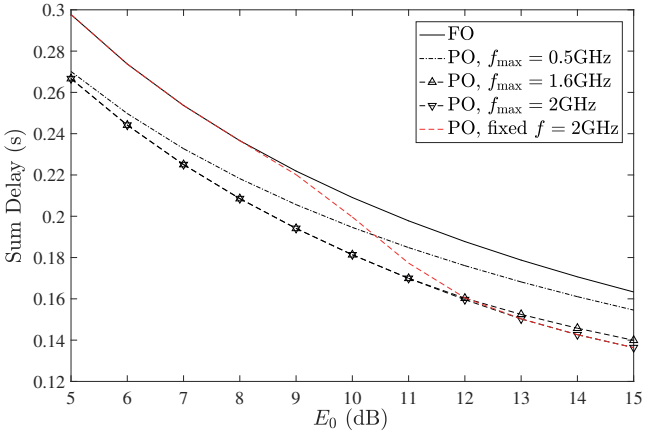

Fig. 6. Sum delay of full offloading and partial offloading.

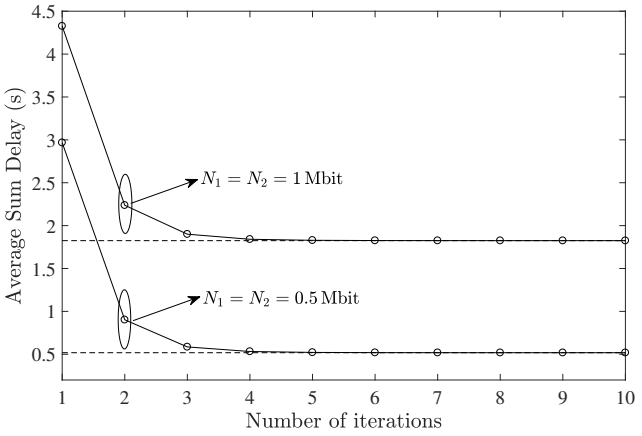

Fig. 7. Convergence evaluation of Algorithm 1.

been set as $E_0/4$. In this case, the number of offloaded bits is $N_1 = N_2 = 0.5\,\text{Mbit}$. From Fig. 4, we observe that the proposed protocol's offloading delay is identical, for both cases of users location, and is less than the benchmark's one. This happens because in the proposed protocol the scheduling of the users in each phase is enabled, in contrast to the benchmark where the pre-assignment of a specific user to the 2-nd phase in combination with the use of fixed decoding order during the 1-st phase increases the sum delay, especially if the channel gain of the user that can solely perform task offloading during the 1-*st* phase is relatively low. Hence, the benchmark's delay is more prone to the case where user 1 is located closer to the BS.

In Fig. 5, the sum energy consumption versus the delay deadline is demonstrated, via Monte Carlo simulations. It is considered that both users present equal delay deadline, i.e, $T^{(1)} = T^{(2)}$. Once again, the prevalence of the proposed protocol for various cases of offloading workload, is clearly seen.

In the continue, the performance of partial offloading scenario of the proposed protocol is examined. For the CPU clock speed of users, it is assumed that $f_j \in (0, f_{\max}]$. Fig. 6, illustrates the performance of partial offloading for various values of $f_{\max}$, which correspond to the use of mobile devices with relatively limited computational capabilities. The speed of local processor has been set $X = 1500$ cycles per bit [10]. In addition, the value of constant parameter k_j, normalized by the product of the noise power spectral density and the path-loss factor, has been set as $\frac{k_j}{BN_0} = 10^{-26}$. The number of the computation tasks' bits is $N_1 = N_2 = 0.5\text{Mbit}$, while the fading channel coefficients are $|h_1|^2 = |h_2|^2 = 1$. Following that, we observe that partial offloading can achieve shorter delay deadline, compared to the full offloading scheme. This outcome was expected, considering that full offloading is a sub-case of partial offloading. It is also worth noticing, that the sweeping of the maximum clock speed $f_{\max}$, from 1.6GHz to 2GHz, has almost no impact in the performance. This implies, that the optimal operational CPU clock speed is not necessarily the highest available one. As can been seen, by fixing the clock speed at $f = 2\text{GHz}$, there is no contribution to the performance gain in the lower regime of E_0. This observation lies to the fact that computation tasks, which are executed with higher CPU clock speed, are more energy-intensive, while devices are energy constrained. Hence, the significance of dynamically regulating the CPU clock speed is corroborated.

Finally, Fig. 7, illustrates the evolution of the average minimum delay in the partial offloading scenario that is derived by using Monte Carlo, when algorithm 1 is executed, with tolerance $\epsilon = 10^{-6}$. The number of iterations refers to the execution of the "while" loop in algorithm 1. We observe that the algorithm presents a relatively fast convergence to the optimal solution. This fact validates the effectiveness of the proposed algorithm, which is based on successive convex approximation, as an efficient method for solving the non-convex problem in (53).

6 Conclusions

In this paper, we have investigated and optimized the performance of mobile edge computing with generalized hybrid NOMA. More specifically, the weighted sum of users' delay was minimized, with respect to their energy constraints. The system's orchestration has been optimized

for both full and partial offloading, in which case, apart from the user scheduling, power control, and time allocation, each user's CPU clock speed was also optimized. All optimization problems were efficiently solved by either using closed-form solutions or efficient algorithms. Moreover, the extension of the analysis to the case of energy minimization has also been considered. Finally, the effectiveness of the proposed techniques has been verified by simulation results, which also provided useful insights on the system's performance. For the case of full offloading, it was shown that the joint use of dynamic user scheduling and time-sharing during the successive interference cancellation process can lead to substantial reduction of delay compared to the considered benchmarks, especially when energy consumption is retained relatively low. Moreover, the application of the proposed protocol becomes very efficient when combined with partial offloading and offers important gains compared to the case of full offloading, especially when each user's CPU clock speed and the resources that are used for communication purposes are jointly optimized.

In general, this work has showcased that the joint orchestration of advanced communication protocols and computing resources can offer substantial improvement in terms of delay and energy consumption reduction. It is highlighted that since the proposed generalized hybrid multiple access scheme is theoretically superior to other alternatives, it can serve in future research as a performance upper bound in order to evaluate MEC with other multiple access techniques. Also, the proposed solving approach has the potential to facilitate the solution of similar problems in MEC systems with partial offloading. Moreover, the introduced multiple access scheme could serve as a baseline for more complex network configurations, such as multi-user multi-carrier systems. Finally, the investigation of the proposed protocol can be extended to the case that different users are interested to different and potentially conflicting performance metrics.

REFERENCES

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: Ai empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[2] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: Deep learning for the internet of things with edge computing," *IEEE Netw.*, vol. 32, no. 1, pp. 96–101, 2018.

[3] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, 2019.

[4] H. Li, K. Ota, and M. Dong, "Eccn: Orchestration of edge-centric computing and content-centric networking in the 5g radio access network," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 88–93, 2018.

[5] C. Zhang, M. Dong, and K. Ota, "Fine-grained management in 5G: DQL based intelligent resource allocation for network function virtualization in c-ran," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 428–435, 2020.

[6] A. Brékine *et al.*, "Building a Roadmap for the Next Generation Internet of Things. Research, Innovation and Implementation 2021 – 2027 (Scoping Paper)," M. Brynskov, F. M. Facca, and G. Hrasko, Eds., Sep. 2019.

[7] M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, A. Neal *et al.*, "Mobile-edge computing introductory technical white paper," *White paper, mobile-edge computing (MEC) industry initiative*, pp. 1089–7801, 2014.

[8] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, Fourthquarter 2017.

[9] Q. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020.

[10] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, 2016.

[11] K. Li, "Computation offloading strategy optimization with multiple heterogeneous servers in mobile edge computing," *IEEE Trans. Sustain. Comput.*, pp. 1–1, 2019.

[12] ——, "A game theoretic approach to computation offloading strategy optimization for non-cooperative users in mobile edge computing," *IEEE Trans. Sustain Comput.*, pp. 1–1, 2018.

[13] C. Yang, X. Chen, Y. Liu, W. Zhong, and S. Xie, "Efficient task offloading and resource allocation for edge computing-based smart grid networks," in *Proc. IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[14] A. A. Al-Habob, O. A. Dobre, A. G. Armada, and S. Muhaidat, "Task scheduling for mobile edge computing using genetic algorithm and conflict graphs," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8805–8819, 2020.

[15] A. A. Al-Habob, A. Ibrahim, O. A. Dobre, and A. G. Armada, "Collision-free sequential task offloading for mobile edge computing," *IEEE Commun. Lett.*, vol. 24, no. 1, pp. 71–75, 2020.

[16] P. X. Nguyen, D. H. Tran, O. Onireti, P. T. Tin, S. Q. Nguyen, S. Chatzinotas, and H. V. Poor, "Backscatter-assisted data offloading in OFDMA-based wireless powered mobile edge computing for iot networks," *IEEE Internet Things J.*, 2021.

[17] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, 2017.

[18] ——, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, 2017.

[19] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, 2017.

[20] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2017.

[21] P. D. Diamantoulakis and G. K. Karagiannidis, "Performance analysis of distributed uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 788–792, 2021.

[22] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 536–568, 2020.

[23] P. Wang, J. Xiao, and L. P, "Comparison of Orthogonal and Non-Orthogonal Approaches to Future Wireless Cellular Systems," *IEEE Veh. Technol. Mag.*, vol. 1, no. 3, pp. 4–11, Sep. 2006.

[24] G. Mazzini, "Power division multiple access," in *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, Florence, Italy, 1998, pp. 543–546.

[25] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-powered communications with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8422–8436, 2016.

[26] K. N. Pappi, P. D. Diamantoulakis, and G. K. Karagiannidis, "Distributed uplink-NOMA for cloud radio access networks," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2274–2277, 2017.

[27] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, 2016.

[28] J. M. Meredith, "Study on downlink multiuser superposition transmission for lte," in *TSG RAN Meeting*, vol. 67, 2015.

[29] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 375–390, 2018.

[30] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1299–1306, 2018.

[31] K. Wang, Z. Ding, D. K. C. So, and G. K. Karagiannidis, "Stackelberg game of energy consumption and latency in MEC systems with NOMA," *IEEE Trans. Commun.*, 2021.

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/TSUSC.2021.3103476, IEEE Transactions on Sustainable Computing

[32] J. Du, W. Liu, G. Lu, J. Jiang, D. Zhai, F. R. Yu, and Z. Ding, "When mobile edge computing (MEC) meets non-orthogonal multiple access (NOMA) for the internet of things (iot): System design and optimization," *IEEE Internet Things J.*, 2020.
[33] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1875–1879, Dec 2018.
[34] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint power and time allocation for NOMA–MEC offloading," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6207–6211, Jun. 2019.
[35] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. on Commun.*, vol. 64, no. 10, pp. 4268–4282, 2016.
[36] F. Wang, J. Xu, and Z. Ding, "Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2450–2463, 2018.
[37] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient NOMA-based mobile edge computing offloading," *IEEE Commun. Lett.*, vol. 23, no. 2, pp. 310–313, 2018.
[38] M. Eliodorou, C. Psomas, I. Krikidis, and S. Socratous, "Energy efficiency for MEC offloading with NOMA through coalitional games," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, 2019, pp. 1–6.
[39] H. Li, F. Fang, and Z. Ding, "Joint resource allocation for hybrid NOMA-assisted MEC in 6G networks," *Digital Communications and Networks*, 2020.
[40] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. K. Karagiannidis, "Optimal task partition and power allocation for mobile edge computing with NOMA," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, 2019, pp. 1–6.
[41] Y. Wu, L. P. Qian, K. Ni, C. Zhang, and X. Shen, "Delay-minimization nonorthogonal multiple access enabled multi-user mobile edge computation offloading," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 392–407, 2019.
[42] X. Lyu, H. Tian, C. Sengul, and P. Zhang, "Multiuser joint task offloading and resource optimization in proximate clouds," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3435–3447, 2016.
[43] X. Diao, J. Zheng, Y. Wu, and Y. Cai, "Joint computing resource, power, and channel allocations for D2D-assisted and NOMA-based mobile edge computing," *IEEE Access*, vol. 7, pp. 9243–9257, 2019.
[44] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.
[45] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the lambertw function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.
[46] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[47] A. M. Abdelhady, O. Amin, A. Chaaban, and M.-S. Alouini, "Downlink resource allocation for multichannel TDMA visible light communications," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Washington, DC, USA, 2016, pp. 1–5.
[48] H. A. Le Thi and T. P. Dinh, "DC programming in communication systems: challenging problems and methods," *Vietnam journal of computer science*, vol. 1, no. 1, pp. 15–28, 2014.

Panagiotis D. Diamantoulakis (SM IEEE) received the Diploma (five years) and PhD from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki (AUTH), Greece, in 2012 and 2017, respectively. Since 2017, he works as a Post-doctoral Fellow in Wireless Communications & Information Processing (WCIP) Group at AUTH and, since 2021, he is also a visiting Assistant Professor in the Key Lab of Information Coding and Transmission at Southwest Jiaotong University (SWJTU), China. From 2018 to 2020, he also worked as visiting Post-doctoral Researcher in the Key Lab of Information Coding and Transmission at SWJTU and in the Institute for Digital Communications (IDC) of the Telecommunications Laboratory (LNT) at Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany. His current research interests include resource allocation in wireless communications, optimization theory and applications in wireless networks and smart grids, game theory, wireless power transfer and optical wireless communications, He is a Working Group Member in the Newfocus COST Action "European Network on Future Generation Optical Wireless Communication Technologies". He serves as an Associate Editor for IEEE Wireless Communications Letters, IEEE Open Journal of the Communications Society, Physical Communications (Elsevier), and Frontiers in Communications and Networks. He was also an Exemplary Reviewer of IEEE Communications Letters in 2014 and IEEE Transactions on Communications in 2017 and 2019 (top 3% of reviewers).

Pavlos S. Bouzinis received the Diploma Degree (5 years) in Electrical and Computer Engineering from the Aristotle University of Thessaloniki (AUTH), Greece, in 2019, where he is currently pursuing his PhD with the Department of Electrical and Computer Engineering. Also, he is a member of the Wireless Communications & Information Processing (WCIP) Group. His current research interests include resource allocation in wireless networks, optimization theory, wireless power transfer, and non-orthogonal multiple access.

Dr. Panagiotis G. Sarigiannidis is an Assistant Professor in the Department of Electrical and Computer Engineering in the University of Western Macedonia, Kozani, Greece since 2016. He received the B.Sc. and Ph.D. degrees in computer science from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and 2007, respectively. He has published over 170 papers in international journals, conferences and book chapters, including IEEE Communications Surveys and Tutorials, IEEE Internet of Things, IEEE Transactions on Broadcasting, IEEE Systems Journal, IEEE Wireless Communications Magazine, IEEE/OSA Journal of Lightwave Technology, IEEE Access, and Computer Networks. He has been involved in several national, European and international projects. He is currently the project coordinator of three H2020 projects, namely a) H2020-DS-SC7-2017 (DS-07-2017), SPEAR: Secure and PrivatE smArt gRid, b) H2020-LC-SC3-EE-2020-1 (LC-SC3-EC-4-2020), EVIDENT: bEhaVioral Insgihts anD Effective eNergy policy acTions, and H2020-ICT-2020-1 (ICT-56-2020), c) TERMINET: nexT gEneRation sMart INterconnectEd ioT, while he coordinates the Operational Program MARS: sMart fArming with dRoneS (Competitiveness, Entrepreneurship, and Innovation). He also serves as a principal investigator in the H2020-SU-DS-2018 (SU-DS04-2018-2020), SDN-microSENSE: SDN-microgrid reSilient Electrical eNergy SystEm and in the Erasmus+ KA2 ARRANGE-ICT: pArtneRship foR AddressiNG mEgatrends in ICT (Cooperation for Innovation and the Exchange of Good Practices). His research interests include telecommunication networks, internet of things and network security. He is an IEEE member and participates in the Editorial Boards of various journals, including International Journal of Communication Systems and EURASIP Journal on Wireless Communications and Networking.

Zhiguo Ding (Fellow, IEEE) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2000, and the Ph.D. degree in electrical engineering from Imperial College London, London, U.K., in 2005. From 2005 to 2018, he was with Queen's University Belfast, Belfast, U.K., Imperial College, Newcastle University, Newcastle upon Tyne, U.K., and Lancaster University, Lancashire, U.K. Since 2018, he has been a Professor of communications with the University of Manchester, Manchester, U.K. From 2012 to 2020, he was an Academic Visitor with Princeton University, Princeton, NJ, USA. His research interests include 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He is the Area Editor of the IEEE Open Journal of the Communications Society, the Editor of IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology, and Journal of Wireless Communications and Mobile Computing, and from 2013 to 2016, he was the Editor of the IEEE Wireless Communication Letters, IEEE Communication Letters. He was the recipient of the Best Paper Award in IET ICWMC-2009 and IEEE WCSP-2014, EU Marie Curie Fellowship 2012–2014, Top IEEE TVT Editor 2017, IEEE Heinrich Hertz Award 2018, IEEE Jack Neubauer Memorial Award 2018, IEEE Best Signal Processing Letter Award 2018, and the Web of Science Highly Cited Researcher 2019.

George K. Karagiannidis (M'96-SM'03-F'14) was born in Pithagorion, Samos Island, Greece. He received the University Diploma (5 years) and PhD degree, both in electrical and computer engineering from the University of Patras, in 1987 and 1999, respectively. From 2000 to 2004, he was a Senior Researcher at the Institute for Space Applications and Remote Sensing, National Observatory of Athens, Greece. In June 2004, he joined the faculty of Aristotle University of Thessaloniki, Greece where he is currently Professor in the Electrical & Computer Engineering Dept. and Head of Wireless Communications & Information Processing (WCIP) Group. He is also Honorary Professor at South West Jiaotong University, Chengdu, China. His research interests are in the broad area of Digital Communications Systems and Signal processing, with emphasis on Wireless Communications, Optical Wireless Communications, Wireless Power Transfer and Applications and Communications & Signal Processing for Biomedical Engineering. Dr. Karagiannidis has been involved as General Chair, Technical Program Chair and member of Technical Program Committees in several IEEE and non-IEEE conferences. In the past, he was Editor in several IEEE journals and from 2012 to 2015 he was the Editor-in Chief of IEEE Communications Letters. Currently, he serves as Associate Editor-in Chief of IEEE Open Journal of Communications Society. Dr. Karagiannidis is one of the highly-cited authors across all areas of Electrical Engineering, recognized from Clarivate Analytics as Web-of-Science Highly-Cited Researcher in the six consecutive years 2015-2020.