sustainability

MDPI

Article

Sustainable Delay Minimization Strategy for Mobile Edge Computing Offloading under Different Network Scenarios

Admoon Andrawes [1,†], Rosdiadee Nordin [1,*,†], Zaid Albataineh [2,†] and Mohammed H. Alsharif [3,†]

[1] Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Build Environment, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia; andrawes2013@gmail.com
[2] Department of Electronics Engineering, Yarmouk University, Irbid 21163, Jordan; zaid.bataineh@yu.edu.jo
[3] Department of Electrical Engineering, College of Electronics and Information Engineering, Sejong University, Seoul 05006, Korea; malsharif@sejong.ac.kr
* Correspondence: adee@ukm.edu.my; Tel.: +60-177-701-312
† These authors contributed equally to this work.

Abstract: The development of mobile edge computing (MEC) is expected to offer better performance in mobile communications than the current cloud computing architecture. MEC involves offering the closest access to the data source or physical mobile network environment. The network services are able to respond faster, thus satisfying the demands of the mobile network industry when deploying various potential business applications in real-time. Since the harvested mobile data are transferred to the edge server to make calculations, data transfers and faults in the mobile network can be swiftly pinpointed and removed accurately. Nevertheless, there are still problems in the practical application of the systems, specifically in reducing delays and lessening energy consumption. Because of non-orthogonal multiple access (NOMA) superior spectrum efficiencies, it is best to combine NOMA with MEC for simultaneous support of multiple access for end users, thus reducing transmission latencies and lowering energy consumption. Combining MEC and NOMA would offer many advantages, including superior energy savings, reductions in latency, massive connectivity, and the potential of combining with additional transmission technologies, such as millimetre-wave (mmWave) and M-MIMO. In this paper, designing wireless resource allocation is crucial for an economically viable low-latency wireless network, which can be realised using the Karush–Kuhn–Tucker (KKT) approach to obtain the optimal solution for partial and full offloading network traffic scenarios to minimize the total latency of the MEC network. The convergence and performance for orthogonal multiple access (OMA), pure-NOMA (P-NOMA), and hybrid-NOMA (H-NOMA) are also compared under different network traffic offloading scenarios. The significant results from this study showed the convergence of the optimal resource allocation in the case of full and partial offloading. The results demonstrated that the P-NOMA reduces the total offloading delay by about 11%.

Keywords: mobile edge computing; NOMA; full offloading; partial offloading

Citation: Andrawes, A.; Nordin, R.; Albataineh, Z.; Alsharif, M.H. Sustainable Delay Minimization Strategy for Mobile Edge Computing Offloading under Different Network Scenarios. *Sustainability* **2021**, *13*, 12112. https://doi.org/10.3390/su132112112

Academic Editor: Manuel Fernandez-Veiga

Received: 10 August 2021
Accepted: 5 October 2021
Published: 2 November 2021

1. Introduction

As 5G technology has begun its rollout, certain transmission techniques—for example, non-orthogonal multiple access (NOMA) and massive multiple-input multiple-output (M-MIMO)—can offer vast improvements in the spectrum efficiency and are well suited to the requirements of low-latency/high-reliability services for smart distribution networks [1–3]. However, in the event of the smart distribution networks experiencing electrical fault/trips, line short-circuits, and switching equipment failure, connectivity to traditional cloud computing will be disrupted due to its dependency on various layers of telecommunication equipment. This can cause a delay in responding to accidents and the length of time needed to locate and restore faults, thus making the 5G distribution network unreliable. NOMA is regarded as a vital enabling multiple access technology for Fifth Generation (5G) generation wireless networks based on its higher levels of spectral efficiency [3]. The

NOMA fundamentally alters the way that multiple access techniques will be designed in the future [4,5]. In comparison with conventional orthogonal multiple access (OMA), in which orthogonal bandwidth resource blocks are allocated to users, NOMA users are encouraged to participate in sharing the same spectrum, where sophisticated transceiver designs, e.g., successive interference cancellation (SIC) and superposition coding, will be employed for handling multiple access interference.

A novel concept has recently been put forward by the European Telecommunications Standards Institute (ETSI) to bring servers closer to end users, named mobile edge computing (MEC). It should be noted that edge servers lack the computational and storage capacities of cloud services (CSs), and their advantage is that they are close to the network users [6–8]. In [9], a computation offloading strategy has been proposed, which employs mobile cloud computing (MCC) to minimize energy usage by the mobile device under delay constraints. The main contributions of this paper can be summarized as follows:

- A MEC-NOMA network is developed and simulated based on two considered network traffic offloading scenarios. The first scenario is related to full offloading, where users offload their entire computation tasks to the MEC server. The second scenario is related to partial offloading, where users are able to execute a part of their processing tasks locally, while the rest is offloaded to the MEC server.
- A two-user NOMA network is assumed to reduce the system's complexity, especially the complexity in the SIC technique. The optimization problem for the two offloading scenarios is developed and validated based on three networks: OMA, P-NOMA, and H-NOMA. The Lagrangian approach is used to solve the optimization problems with Karush–Kuhn–Tucker (KKT) conditions.
- A closed-form solution for the partial task factor has been derived to simplify the optimization problem in the case of partial network offloading to formulate an extensive numerical solution to minimize task delay for different networks.

The rest of this paper is organized as follows. Section 2 discusses recent studies related to MEC offloading with different research directions, including delay minimization and energy consumption. Section 3 describes the system model in this study, including full and partial offloading with OMA, P-NOMA, and H-NOMA. In Section 4, performance analysis and problem formulation for delay minimization with different scenarios are presented. Numerical results are presented in Section 5, and conclusions are made in Section 6.

2. Related Works

Recent research on resource allocation to reduce task delays and energy consumption has been heavily focused on NOMA-MEC systems. MEC is viewed as central to the coming generation of wireless networks, owing to reductions in energy consumption and latency [10]. As Figure 1 illustrates, MEC is capable of supporting a range of users, including self-driving (or autonomous) vehicles, mobile devices, and IoT Devices around the network's edge, with base stations (BSs) fitted with MEC servers offering cloud-like computing capacity for mobile devices dealing with high levels of computation and low latency [11]. MEC networks can offload most computing tasks to the base station, which is equipped with an MEC server to undertake remote computing. Once the BS has computed the task, the results can be sent back to mobile devices [12,13]. With MEC offloading, tasks may be binary offloaded (i.e., the computational task may not be segmented, so either local computation must be applied or the entire task must be completely offloaded to the MEC server) or partially offloaded [14].

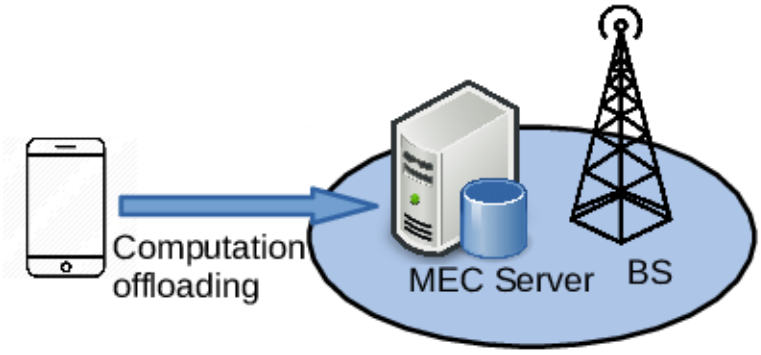

Figure 1. MEC architecture.

One multiple access scheme that has wide popularity is power-domain NOMA, in which SIC techniques are employed at the receiver, allowing users with substantial channel gain to remove interference from users with lower channel gain or poor channel quality [15]. This allows multiple users to employ transmission signals simultaneously, experiencing less interference than the OMA system. As a result of high spectrum efficiency, NOMA resource optimisation performs better than NOMA, especially on energy efficiency and system sum rates [16,17]. Inspired by NOMA's superior performance compared to OMA, it has been suggested that NOMA uplink/downlink transmission should be implemented with MEC networks to allow multiple users simultaneous signal transmission with less interference [18]. Some researchers have recently examined how multi-user MEC networks and NOMA can be combined [19–21]. The authors in [18] investigated weighted-sum energy minimisation problems in multi-user NOMA-MEC systems, while the authors in [22] undertook research on the minimisation of energy usage for NOMA-MEC networks, assuming that all users could access resource blocks of multiple bandwidths. The authors in [23] designed a joint time allocation/power scheme in NOMA-MEC aimed at reducing energy usage for network traffic offloads. The authors in [24] investigated how delays could be minimised by using a novel resource allocation. As delay minimisation is a vital element of NOMA-MEC, investigations were undertaken as to how task delay could be reduced by optimising resources for transmission power, offloading times, and the assignment of offloading tasks [25–27]. In particular, an algorithm was suggested to minimise the overall delay for computational tasks by optimising downloading duration, offloading workload, and offloading duration [25]. The current research focuses on minimising task delays by applying NOMA uplink transmissions to MEC partial offloading. This research is motivated by two central concepts. Firstly, in terms of communications, minimising delay is vital for a MEC network, especially during the offloading phase, which can be responsible for lengthy delays. Secondly, although researchers have investigated delay minimisation in NOMA-MEC [28–30], no real insight has been achieved into delay minimisation in NOMA-MEC networks. The authors in [28] proposed a hybrid NOMA-MEC system, but this research considered a system of greater complexity that aims to minimise task latency using pure NOMA-MEC and partial offloading. Most of the existing studies focus on communication resource allocation, such as sub-channel allocation or offloading power, or computational resources, such as task assignment [31,32]. Within NOMA-MEC, there is a requirement to consider the methods of allocating network resources to optimize energy consumption. The authors in [31] optimized the offloading tasks and offloading power levels of every user based on the order of the SIC decoding to optimize the amount of energy consumed in NOMA-MEC. The authors in [33] successfully reduced the overall energy consumption by optimizing the task offloading partitions, the transmission allocations, and the transmit powers. Time allocation, task assignment, and energy-efficient power allocation were recommended as effective means of reducing the overall energy consumption of MEC networks [34]. A hybrid NOMA-MEC scheme has been proposed and validated to replace the benefits of OMA and NOMA systems [35]. Within this hybrid

scheme, a user is able to offload network-related tasks within a given time slot allocated to an alternative user and subsequently offload the outstanding task within the time slot exclusively occupied by itself [36–38].

Mobile edge computing (MEC) has been suggested as a means of supplying low-latency computing services within efforts to deliver cloud-like computing at the wireless network edge [39,40]. The need for remote transmission to the cloud is avoided because the edge node has sufficient computation capacity to execute the applications. Nonetheless, there is a correlation between an increase in wireless devices and the congestion of the computation task uploading [41,42]. This form of congestion significantly delays the rate of task uploading and results in a general delay in the time it takes to complete computation offloading. As such, it represents a limiting factor for MEC that needs to be addressed through the development of low-latency MEC offloading that can promote the task uploading capacity within the limited bandwidth. Recent studies have suggested using non-orthogonal multiple access (NOMA) as a useful multiple access technique that enables a 5G network to improve the network's capacity [43]. The NOMA enables multiple users to share orthogonal spectrum resources and utilizes signal differences on the power domain to differentiate between users according to the successive interference cancellation (SIC). In comparison to orthogonal multiple access (OMA), NOMA can leverage enhanced spectrum efficiency and network connectivity to boost MEC's task uploading capacity. As such, the NOMA-enabled MEC represents a promising approach as part of the efforts to develop future wireless networks [44,45]. A performance analysis completed by [44] found that NOMA-enabled MEC was superior in energy efficiency and computation offloading.

Unlike conventional cloud computing, MEC reduces the transmission latency by reducing the physical distance between the user and the server [46–48]. However, the majority of IoT devices have a lack of computational tools and limited battery life. As such, it is challenging for them to perform an intensive task within a given time if they rely on local computing alone. An MEC could be employed to offload the raw sensor data of IoT devices to the BS. The MEC server can apportion powerful computing resources to compute the task, while the IoT device can download the outcomes within the delay constraint [49–51]. NOMA represents a viable Multiple Access (MA) technology to support this immense growth in traffic within B5G/6G due to its spectral efficiency gain [52–54]. Every user is serviced within conventional Orthogonal Multiple Access (OMA) via a dedicated frequency or time resource block. Unlike OMA, NOMA can significantly enhance spectral efficiency because it enables more than one user to multiplex on a single frequency band at different power levels during the transmission [55,56]. Successive interference cancellation (SIC) can be employed at the receiver to reduce the interference that users experience.

Several researchers have explored the integration of MEC and NOMA. Multiple studies described in the previous literature aim to overcome the technical limitations of a NOMA-enabled MEC. Within these studies, researchers have commonly concentrated on energy minimization problems [57,58] and delay minimization problems [59–61]. Within these studies, both downlink and uplink NOMA transmissions have been evaluated as a means of reducing the consumption of energy.

Most research studies have examined the uplink transmission that enables multiple users to offload numerous tasks to a BS at once. For instance, [57,62] described a downlink model that enabled a single user to offload various aspects of tasks to more than one BS. The authors in [63] proposed a framework to minimize the energy consumption of multi-antenna NOMA-assisted MEC. Furthermore, [64] presented a half-duplex framework that considered both the time and energy consumption of offloading to the MEC server and the amount of energy consumed to download the results. The authors in [65] recommended a device-to-device (D2D)-assisted MEC that made it possible for users to collaborate, with the underlying intention of decreasing the computational load placed upon the edge servers. An offloading scheme that was studied in [66] was serviced by heterogeneous networks to reduce the task backlog and enhance the offloading utilities.

Although the majority of existing studies have focused on time and power resource allocation, some researchers have examined the impact of various offloading strategies on optimization [57,67]. Two common offloading models, binary and partial offloading, are two types of offloading strategies commonly employed in past research, including [68,69]. When the binary offloading scheme is employed, the entire task can be offloaded to the BS for remote computation. Alternatively, a local calculation can be performed on a mobile device. Although the partial offloading approach makes it possible for each task to be partially offloaded, the remaining tasks can be computed on a local level. As described in [70,71], the offloading strategy can also be employed to break and distribute a given task to multiple MEC servers. The authors in [69] proposed using a novel hybrid NOMA and OMA model, where both transmissions are applied at different time slots for offloading a task. However, this study was limited to a two-user case.

Based on the previous works, most of the past research concentrates on the resource allocation for MEC networks. In this research, applying the MEC under full and partial network traffic offloading scenarios under different network environments (with other multiple access techniques) is expected to minimize the task delay and provide higher energy efficiency for a sustainable 5G system.

3. System Model

When NOMA assists a MEC system, numerous mobile users can offload their tasks simultaneously on a single time/frequency resource. For instance, we apply the assumption that only a one-time slot is unoccupied at a given moment, and more than one user will offload tasks to the BS. If the OMA transmission were applied in a situation of this nature, one user would need to wait, while the other user transmitted. However, if the NOMA transmission were applied, it would be possible for both users to transmit simultaneously, and this would reduce the latency that results from radio resource shortages. As such, 6G and B5G computing services benefit from using a hybrid of MEC and NOMA.

Mobile users choose to offload their computationally demanding, latency-critical, and invisible operations to the server in a practical network deployment due to their limited computing capabilities. Therefore, the users are sorted based on their computation deadlines, i.e., $D_1 \leq \cdots \leq D_k$. It is also presumed that the MEC server only schedules two users to minimize system complexity. Nevertheless, assume that the number of nats for the task of each user $N_k = N$; user m is more delay demanding than user n, so m is served first—in other words, the m with the lower latency.

As OMA is utilized, where the two users use the same channel with multiplexing in time, each user is given a designated offloading time slot. User m is treated first, since it has a tighter schedule deadline than user n, i.e., n has to wait for m to finish the offloading. As a result, the transmit powers of users, represented by P_m and P_n, are required to comply with offloading as $D_m \log_2\left(1 + \frac{P_m^{OMA}|h_m|^2}{\sigma^2}\right) = N$ and $D_n \log_2\left(1 + \frac{P_n^{OMA}|h_n|^2}{\sigma^2}\right) = N$, respectively, where h_i represents the channel gain of user i, and σ^2 is the noise power.

The two users can offload their tasks to the server simultaneously utilizing the NOMA technique during D_m. It is worth noting that if user D_m's message is decoded in the second stage of perfect successive interference cancellation (SIC), the main possible offloading schemes can occur:

- Full offloading: users offload entire computation tasks to the MEC server, such as tasks related to augmented reality (AR)/virtual reality (VR) applications.
- Partial offloading: the user is able to execute a part of their processing tasks locally, while the rest is offloaded to the MEC server, such as drone flight control applications.

In this work, OMA and P-NOMA are applied to both full and partial offloading. In the case of H-NOMA, m and n offload simultaneously, since forcing n to complete its offloading with time slot D_m can be energy inefficient, so we dedicated a time slot to user n called T_n^* with energy E_n^* . The power with time allocation for the three main scenarios is shown in the following figures. As shown in Figure 2a, each user transmits a certain time allocation

based on the delay constraints. P-NOMA is illustrated in Figure 2b. User m is considered the near user with less power allocation, where both users offload simultaneously with D_m. In the case of H-NOMA, m and n also offload simultaneously, but a dedicated time slot is given to the user n called T_n^* with energy E_n^* as shown in Figure 2c.

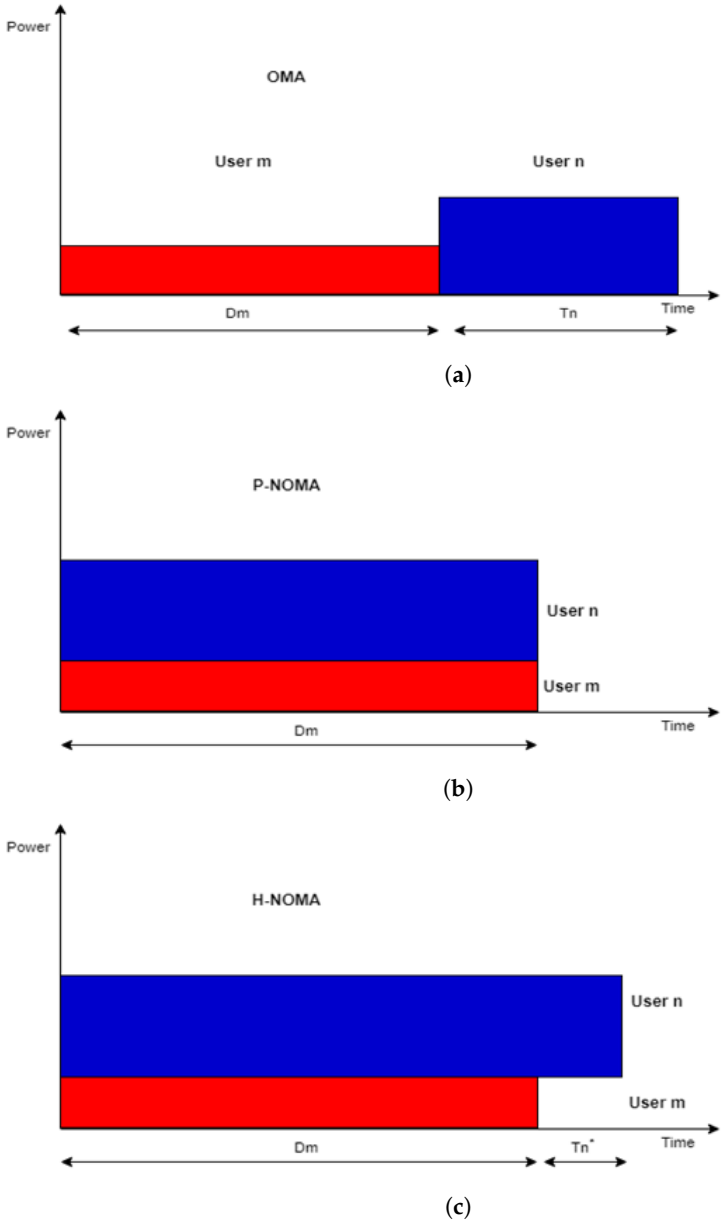

Figure 2. Power-time allocation for (**a**) OMA, (**b**) P-NOMA, (**c**) H-NOMA.

4. Problem Formulation for Delay Minimization

4.1. Full Offloading

In the case of OMA, user m transmits first, where the transmission time for user m is $T_m = D_m$, and where the transmission time for user n is $T_n = D_n - T_m$. The data rate for each user can be written as

$$R_i = B\log_2\left(1 + \frac{P_i|h_i|^2}{\sigma^2}\right), \forall i, i \in \{1, 2\} \quad (1)$$

The problem formulation can be written as

$$\min_{\{P_m, P_n\}} \quad \{T_m^o + T_n^o\} \tag{2a}$$

$$\text{s.t.} \quad N - T_i^o R_i \leq 0, \forall i, i \in \{1, 2\} \tag{2b}$$

$$E_m^o + E_n^o \leq E_{\max} \tag{2c}$$

In the case of P-NOMA, the transmission time for both users is D_m. Since SIC is used to decode the data for a specific user, which is known as the far user (weak channel gain), the data rate for that user is the same as the OMA scenario, while the other user is known as the near user (strong channel gain). In this work, we assume that $h_m \geq h_n$. The data rate for the near user can be written as

$$R_m = B\log_2\left(1 + \frac{P_m|h_m|^2}{P_n|h_n|^2 + \sigma^2}\right) \tag{3}$$

The problem optimization using KKT can be written as

$$\begin{aligned} \min_{P_n, P_m} \quad & T_m + T_n \\ & T_m \geq 0 \\ & T_n \geq 0 \\ & T_m \leq D_m \\ & T_m + T_n \leq D_n \\ & E_m + E_n \leq E_{max} \\ & N - T_i R_i \leq 0, i = 1, 2 \end{aligned} \tag{4}$$

The Lagrangian is given as

$$\mathcal{L}(T_m, T_n, \bar{\mu}) = T_m + T_n + \bar{\mu}^T \bar{g} \tag{5}$$

where $\bar{g}$ is a column vector containing the standardized constraints as the following:

$$\bar{g} = \begin{bmatrix} -T_m \\ -T_n \\ -D_m + T_m \\ -D_m + T_m + T_n \\ -T_m R_m + N \\ -T_n R_n + N \end{bmatrix} \tag{6}$$

and $\bar{\mu} = [\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6]^T$ is the Lagrangian multiplier which is equal to or larger than zero.

The KKT conditions states that the optimal solution of T_m, T_n and its Lagrangian multiplier should satisfy

$$\begin{aligned} & \mu_1 * T_m = 0 \\ & \mu_2 * T_n = 0 \\ & \mu_3 * (T_m - D_m) = 0 \\ & \mu_4 * (T_m + T_n - D_n) = 0 \\ & \mu_7 * (-T_m R_m + N) = 0 \\ & \mu_8 * (-T_n R_n + N) = 0 \end{aligned} \tag{7}$$

After applying the derivative of Lagrangian, the principles of the KKT conditions can be referred to in Chapter 5.5.3 in [69].

Because $T_m + T_n = D_n$, $T_m = D_m$, then, $T_n = D_n - D_m$:

$$\frac{|h_n|^2}{N_o} * \mu_4 \quad = \quad \frac{P_n|h_n|^2 + N_o}{N_o} \ln\left(\frac{P_n|h_n|^2 + N_o}{N_o}\right) - \frac{P_n|h_n|^2}{N_o} \tag{8}$$

Let $x_m \doteq \frac{P_n|h_n|^2+N_o}{N_o} = \frac{P_n|h_n|^2}{N_o} + 1$. Then,

$$\mu_4 = \frac{N_o}{|h_n|^2} f(x_m) \tag{9}$$

where $f(x_m) = x_m \ln(x_m) - (x_m - 1)$.

We can write

$$\mu_3 = \frac{N_o}{|h_m|^2} f(x_m) - \frac{N_o}{|h_n|^2} f(x_n) \tag{10}$$

We can write

$$\begin{aligned} x_m &= 2^{\frac{N}{D_m}} \\ x_n &= 2^{\frac{N}{(D_n - D_m)}} \end{aligned} \tag{11}$$

In summary,

- By N, B, D_n, D_m, we can calculate x_m, x_n.

$$\begin{aligned} x_m &= 2^{\frac{N}{D_m}} \\ x_n &= 2^{\frac{N}{(D_n - D_m)}} \end{aligned} \tag{12}$$

- By x_m, X_n, h_n, h_m, we can calculate P_n, P_m.

$$\begin{aligned} P_m &= \frac{x_m - 1}{(|h_m|^2/N_o)} \\ P_n &= \frac{x_n - 1}{(|h_n|^2/N_o)} \end{aligned} \tag{13}$$

The problem formulation for P-NOMA can be written same as Equations (1) and (2) with different data rates for user m.

Since the transmission time in the case of P-NOMA for both users is D_m and is very noisy, it causes an increase in energy consumption in the system. On the other hand, P-NOMA forces user n to complete its offloading within D_m. For these reasons, H-NOMA is applied to give another time for user n. The average energy efficiency will be improved in this scenario, and more data can be offloaded than the P-NOMA scenario. The problem formulation of this scenario can be written as

$$\min_{\{P_m, P_n, P_n^*\}} \quad \{T_m^o + T_n^o + T_n^*\} \tag{14a}$$

$$\text{s.t.} \quad N - T_m^o R_m \leq 0 \tag{14b}$$

$$N - T_n^o R_n - T_n^* \log_2\left(1 + \frac{P_n^* |h_n|^2}{\sigma^2}\right) \leq 0 \tag{14c}$$

$$E_m^o + E_n^o + E_n^* \leq E_{\max} \tag{14d}$$

4.2. Partial Offloading

In the case of OMA, the problem formulation can be written as

$$\min_{\{P_i, B_i\}} \quad \max\left\{T_i^o + T_i^p\right\}, \forall i \tag{15a}$$

$$\text{s.t.} \quad 0 \leq B_i < 1, \forall i \tag{15b}$$

$$\text{s.t.} \quad N B_i - T_i^o R_i \leq 0, \forall i \tag{15c}$$

$$E_m^o + E_n^o + E_m^p + E_n^p \leq E_{\max} \tag{15d}$$

In the case of partial offloading, the task offloading can be optimized when the offloading time equals the task local computing time. In the case of P-NOMA, the problem formulation can be written the same as Equation (5) with different data rates for user m.

The optimal solution can only be obtained when the offloading times equal each other. Considering a two-user case, the problem can be expressed as follows [67]:

$$\min_{\{\beta_1,\beta_2,p_1,p_2\}} \quad \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1+|h_1|^2 p_1 + |h_2|^2 p_2\right)} \tag{16}$$

$$\text{s.t.} \quad \beta_1 \in [0,1], \beta_2 \in [0,1] \tag{17a}$$

$$\kappa_1(1-\beta_1)L_1C_1\left(f_1^{loc}\right)^2 + \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1+|h_1|^2 p_1 + |h_2|^2 p_2\right)} p_1 \le E_{\max} \tag{17b}$$

$$\kappa_2(1-\beta_2)L_2C_2\left(f_2^{loc}\right)^2 + \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1+|h_1|^2 p_1 + |h_2|^2 p_2\right)} p_2 \le E_{\max} \tag{17c}$$

$$\frac{(1-\beta_1)L_1C_1}{f_1^{loc}} = \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1+|h_1|^2 p_1 + |h_2|^2 p_2\right)} \tag{17d}$$

$$\frac{(1-\beta_2)L_2C_2}{f_2^{loc}} = \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1+|h_1|^2 p_1 + |h_2|^2 p_2\right)} \tag{17e}$$

$$\frac{\beta_1 L_1}{\log_2\left(1+|h_1|^2 p_1\right)} = \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1+|h_1|^2 p_1 + |h_2|^2 p_2\right)} \tag{17f}$$

In the problem above, the objective function is quasiconvex. However, the constraints (17b,c) are not a convex set with respect to $\{\beta_1, \beta_2, p_1, p_2\}$. To simplify this problem, we first deal with equality constraints (17d–f). To solve the above problem and obtain the global optimum, we first equally transform this problem to an equivalent convex form via equality constraints. By using Equation (17f), we can replace the right sides of (17d,e) with the left side of (17f). Then, we have

$$\begin{aligned}
(1-\beta_1)L_1C_1 &= \frac{\beta_1 L_1}{\log_2\left(1+|h_1|^2 p_1\right)} f_1^{loc} \\
\frac{(1-\beta_1)L_1C_1}{f_1^{loc}} &= \frac{(1-\beta_2)L_2C_2}{f_2^{loc}} \\
\frac{(1-\beta_1)L_1C_1}{f_1^{loc}} &= \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1+|h_1|^2 p_1 + |h_2|^2 p_2\right)}
\end{aligned}$$

The above objective function can be rewritten by

$$\begin{aligned}
\frac{(1-\beta_1)L_1C_1}{f_1^{loc}} &= \frac{(1-\beta_2)L_2C_2}{f_2^{loc}} \\
(1-\beta_1)L_1C_1f_2^{loc} &= (1-\beta_2)L_2C_2f_1^{loc} \\
\beta_2 L_2 &= L_2 - \frac{L_1C_1f_2^{loc}}{C_2f_1^{loc}} + \beta_1\frac{L_1C_1f_2^{loc}}{C_2f_1^{loc}}
\end{aligned}$$

and

$$\beta_1 L_1 + \beta_2 L_2 = \beta_1 L_1 A_1 + B_1 \tag{18}$$

where $A_1 = 1 + \frac{C_1 f_2^{loc}}{C_2 f_1^{loc}}$ and $B_1 = L_2 + L_1\frac{C_1 f_2^{loc}}{C_2 f_1^{loc}}$.

Since

$$\frac{\beta_1 L_1 + \beta_2 L_2}{R} = \frac{(1-\beta_1)L_1C_1}{f_1^{loc}}$$

we have

$$\beta_1 L_1 = \frac{L_1C_1R - B_1 f_1^{loc}}{A_1 f_1^{loc} + C_1 R}$$

According to (15), we have

$$\beta_1 L_1 + \beta_2 L_2 = \beta_1 L_1 A_1 + B_1 = \frac{L_1C_1A_1R + B_1C_1R}{A_1 f_1^{loc} + C_1 R}$$

and

$$\frac{\beta_1 L_1 + \beta_2 L_2}{R} = \frac{L_1 C_1 A_1 + B_1 C_1}{A_1 f_1^{loc} + C_1 R} = \frac{L_1 + L_2}{\frac{f_1^{loc}}{C_1} + \frac{f_2^{loc}}{C_2} + R}$$

and it can be rewritten as

$$\frac{L_1 + L_2}{\frac{f_1^{loc}}{C_1} + \frac{f_2^{loc}}{C_2} + B\log_2\left(1 + |h_1|^2 p_1 + |h_2|^2 p_2\right)}$$

Therefore, the problem can be rewritten by

$$\min_{\{p_1, p_2\}} \quad \frac{L_1 + L_2}{\frac{f_1^{loc}}{C_1} + \frac{f_2^{loc}}{C_2} + B\log_2\left(1 + |h_1|^2 p_1 + |h_2|^2 p_2\right)} \tag{19}$$

$$\text{s.t.} \quad 0 \le p_1 \le P_{\max}, 0 \le p_2 \le P_{\max} \tag{20a}$$

$$\kappa_1(1-\beta_1)L_1 C_1 \left(f_1^{loc}\right)^2 + \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1 + |h_1|^2 p_1 + |h_2|^2 p_2\right)} p_1 \le E_{\max} \tag{20b}$$

$$\kappa_2(1-\beta_2)L_2 C_2 \left(f_2^{loc}\right)^2 + \frac{\beta_1 L_1 + \beta_2 L_2}{\log_2\left(1 + |h_1|^2 p_1 + |h_2|^2 p_2\right)} p_2 \le E_{\max} \tag{20c}$$

The optimal p_1^* and p_2^* are obtained, and the optimal β_1^* and β_2^* can be calculated by the following expressions:

$$\beta_1^* = \frac{\log_2\left(1+|h_1|^2 p_1^*\right)}{\frac{f_1^{loc}}{C_1} + \log_2\left(1+|h_1|^2 p_1^*\right)}$$

$$\beta_2^* = 1 - \frac{\left(1-\beta_1^*\right) L_1 C_1 f_2^{loc}}{L_2 C_2 f_1^{loc}}$$

In the case of H-NOMA, the problem formulation can be written as

$$\min_{\{P_i, P_n^*, B_i\}} \quad \max\left\{(T_i^o + T_n^*), T_i^p\right\}, \forall i \tag{21a}$$

$$\text{s.t.} \quad 0 \le B_i < 1, \forall i \tag{21b}$$

$$\text{s.t.} \quad NB_m - T_m^p R_m \le 0, \forall i \tag{21c}$$

$$NB_n - T_n^p R_n - T_n^* \log_2\left(1 + \frac{P_n^* |h_n|^2}{\sigma^2}\right) \le 0 \tag{21d}$$

$$E_m^o + E_n^o + E_m^p + E_n^p + E_n^* \le E_{\max} \tag{21e}$$

5. Discussion and Numerical Results

MEC primarily consists of the following phases. The first phase is the offloading phase, in which a user sends tasks to one or more MEC servers. The second phase is the feedback phase, in which the MEC servers do the offloaded operations and return the results of these calculations to the users. The emphasis of this research is on the influence of NOMA on the first phase of MEC, with the premise that the costs of the second phase of MEC are insignificant for the sake of our analysis. This assumption has frequently been utilized in MEC literature for the following two reasons. Firstly, due to the higher calculation capabilities and the tiny sizes of the computing results, the delay is insignificant in MEC's second stage, i.e., the time for a server to compute an offloaded task and the time for a user to receive the computations from the server. Second, the energy needed to compute the offloaded tasks of a MEC server and the energy consumption of the transmitter during the second phase of MEC can be ignored, as no energy restrictions are placed on MEC servers.

This section illustrates the performance of the three networks (OMA which is based on 4G networks, P-NOMA which is based on 5G networks, H-NOMA which is based on Beyond 5G networks) under two scenarios (full offloading and partial offloading). The simulation parameters can be illustrated in the following table (Table 1).

Table 1. Simulation parameters.

Parameter	Description	Value
σ^2	AWGN spectral density	-174 dBm/Hz
E_{max}	max energy	0.2 J
L_i	number of bits per task	1×10^6
f_i^{loc}	CPU frequency	1×10^6
C_i	number of CPU cycles	1×10^3
B	bandwidth	1 MHz
k_i	capacitance coefficient for each CPU cycle	1×10^{-28}
h_m	fading channel for user m	0.75
h_n	fading channel for user n	0.5

In Figure 3, the performance of the full offloading delay for OMA, P-NOMA, and H-NOMA is shown with respect to the energy at the MEC server-side with certain channel gain for the two users. As shown, using P-NOMA reduces the total offloading delay by about 11% compared to the H-NOMA network. In Figure 4, the performance of the partial offloading delay for the three networks is illustrated. As shown, three different networks in the case of partial offloading provide better performance than three different networks in the case of full offloading, since in the case of partial offloading, the mobile devices are able to execute a part of their processing tasks locally, which reduces the overall latency in the MEC. The gain between partial offloading and full offloading in the case of the H-NOMA network is about 26%.

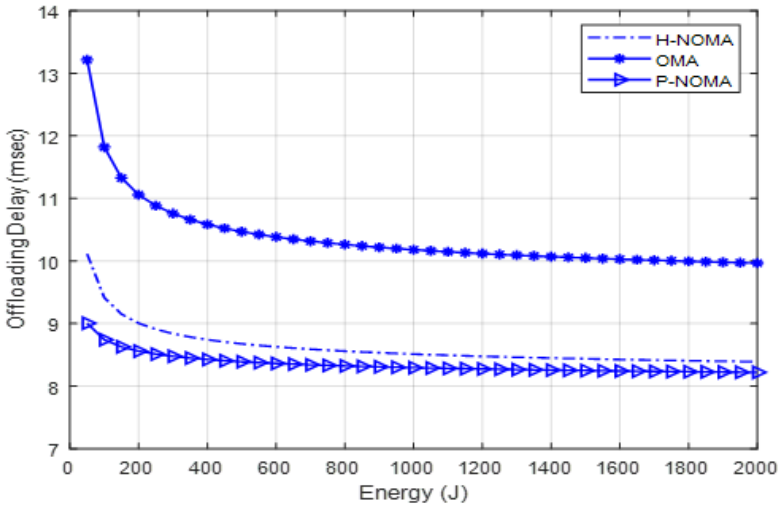

Figure 3. Full offloading delay for three different networks.

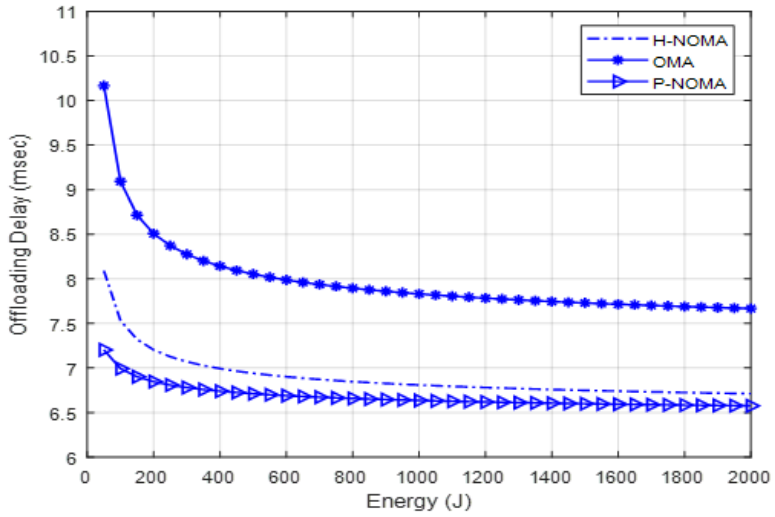

Figure 4. Partial offloading delay for three different networks.

In Figures 5 and 6, the convergence for the three networks is shown for the full and partial offloading, respectively. As shown, both figures need about six iterations to reach a steady state. In the case of full offloading, at N = 15, the offloading delay starts from 9 ms (P-NOMA), 9.6 ms (H-NOMA), and 9.8 ms (OMA), respectively, while in partial offloading, the offloading delay starts from 7.3 ms (P-NOMA), 7.6 ms (H-NOMA), and 7.8 ms (OMA). On the other hand, while the number of nats increases, the offloading delay increases for the three networks. The increasing delay between the networks is expected, since in the case of OMA, user m transmits first, then user n, which leads to more delay, while in the case of P-NOMA, both users offload simultaneously, forcing user n to complete its offloading in the same time allocation. In H-NOMA, it even solves the noisy time slot in Dm and solves the problem in forcing user n to complete its offloading in the same time slot, but it provides more time for user n for offloading, which requires more delay compared to P-NOMA, since the objective of this study is to reduce the total delay for MEC offloading. These observations are shown in Figure 2.

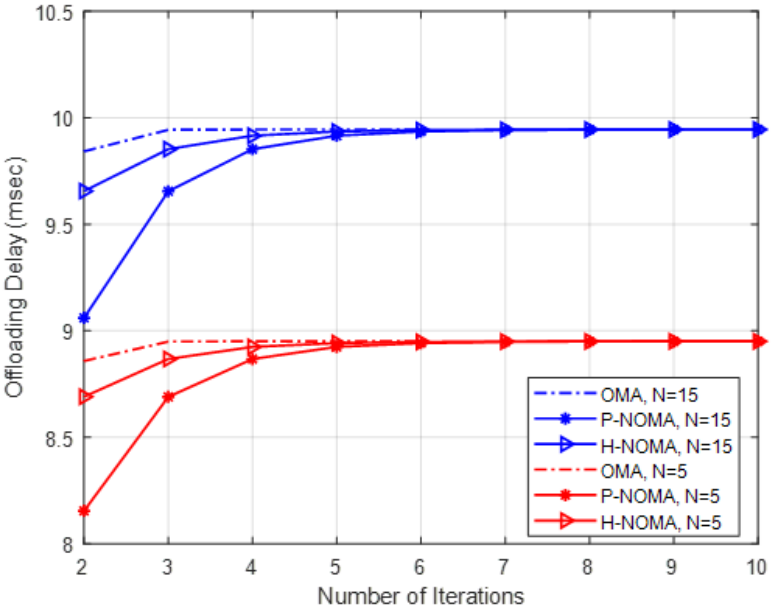

Figure 5. Full offloading delay with iterations for three different networks.

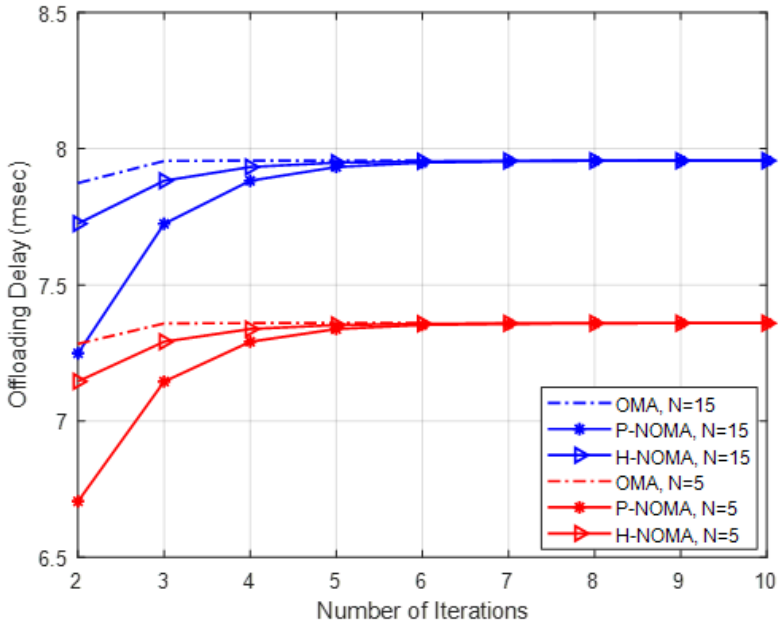

Figure 6. Partial offloading delay with iterations for three different networks.

We initialize the lower limit to 6 ms and the top limit towards the delay to the whole local calculation time. In Figure 5, the convergence fluctuates because the optimum delay is towards the lower limit. The bandwidth is adjusted to B = 1 MHz in Figure 6. Due to the small bandwidth, the download rate is modest, and most tasks are calculated by local users. In this example, the highest limit is near the ideal latency. The delay at the first iteration is therefore considerably below the ideal amount, approaching the top limit, and continues to increase until its convergence via every iteration. From these figures, we can observe that within six rounds, the proposed algorithm is converging, indicating that the method presented is feasible.

6. Conclusions

In this paper, two different network offloading scenarios, namely full offloading and partial offloading, have been considered and simulated under three different mobile network deployments, which are orthogonal multiple access (OMA), pure non-orthogonal multiple access (P-NOMA), and hybrid non-orthogonal multiple access (H-NOMA). Our main aims are to minimize the total offloading delay by leveraging a sustainable mobile edge computing (MEC) architecture. In this work, each user could either compute the task to MEC or achieve partial computation locally. Problem modelling was derived and illustrated for the two offloading scenarios on different multiple access networks to achieve the minimum latency, while achieving sufficient energy for each network. Time allocation for each network has also been illustrated. The problem formulation for each scenario and each network has been deployed and analysed by using simulations. The KKT approach was used to obtain the optimal solution for each case. The convergence of the three networks has been illustrated with a different number of nats. The results achieved show that P-NOMA performs better than H-NOMA in terms of delay. The simulation results have shown that any increase in the number of nats leads to an increase in the average delay. On the other hand, using P-NOMA can reduce total latency by about 10% in comparison to other networks. For future research, this work can be extended to a multi-user scheme with multi-cell optimization by applying a user pairing strategy.

Author Contributions: Conceptualization, A.A., R.N., Z.A., and M.H.A.; methodology, A.A., R.N., and M.H.A.; software, A.A. and Z.A.; validation, A.A., R.N., Z.A., and M.H.A.; formal analysis, A.A., R.N., Z.A., and M.H.A.; investigation, A.A., R.N., Z.A., and M.H.A.; data curation, A.A., R.N., Z.A., and M.H.A.; writing—original draft preparation, A.A. and Z.A.; writing—review and editing, A.A., R.N., Z.A., and M.H.A.; visualization, A.A., R.N., Z.A., and M.H.A.; supervision, R.N. and M.H.A.; project administration, R.N. and M.H.A.; funding acquisition, R.N. and M.H.A. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge the financial support from the Air Force Office of Scientific Research (AFOSR), under the grant ref number FA2386-20-1-4045 (UKM Ref: KK-2020-007), for the open access fee payment.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jameel, F.; Haider, M.; Butt, A. Massive MIMO: A survey of recent advances, research issues and future directions. In Proceedings of the 2017 International Symposium on Recent Advances in Electrical Engineering (RAEE), Islamabad, Pakistan, 24–26 October 2017; pp. 1–6. [CrossRef]
2. Xin, Y.; Wang, D.; Li, J.; Zhu, H.; Wang, J.; You, X. Area spectral efficiency and area energy efficiency of massive MIMO cellular systems. *IEEE Trans. Veh. Technol.* **2016**, *65*, 3243–3254. [CrossRef]
3. Alsharif, M.H.; Nordin, R.; Ismail, M. Intelligent cooperation management of multi-radio access technology towards the green cellular networks for the twenty-twenty information society. *Telecommun. Syst.* **2017**, *65*, 497–510. [CrossRef]
4. Andrawes, A.; Nordin, R.; Abdullah, N.F. Energy-Efficient Downlink for Non-Orthogonal Multiple Access with SWIPT under Constrained Throughput. *Energies* **2020**, *13*, 107. [CrossRef]
5. Andrawes, A.; Nordin, R.; Ismail, M. Wireless Energy Harvesting with Cooperative Relaying under the Best Relay Selection Scheme. *Energies* **2019**, *12*, 892. [CrossRef]
6. Cicirelli, F.; Guerrieri, A.; Spezzano, G.; Vinci, A.; Briante, O.; Iera, A.; Ruggeri, G. Edge computing and social Internet of Things for large-scale smart environments development. *IEEE Internet Things J.* **2018**, *5*, 2557–2571. [CrossRef]
7. Abbas, N.; Zhang, Y.; Taherkordi, A.; Skeie, T. Mobile edge computing: A survey. *IEEE Internet Things J.* **2018**, *5*, 450–465. [CrossRef]
8. Andrawes, A.; Nordin, R.; Ismail, M. Wireless Energy Harvesting with Amplify-andForward Relaying and Link Adaptation under Imperfect Feedback Channel. *J. Telecommun. Electron. Comput. Eng.* **2018**, *10*, 83–90.
9. Andrawes, A.; Nordin, R.; Ismail, M. Energy Harvesting with Link Adaptation under Different Wireless Relaying Schemes. *J. Commun.* **2018**, *13*, 1–6. [CrossRef]
10. Mao, Y.; You, C.; Zhang, J.; Huang, K.; Letaief, K.B. A survey on mobile edge computing: The communication perspective. *IEEE Commun. Surveys Tuts.* **2017**, *19*, 2322–2358. [CrossRef]
11. Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; Xu, L. Edge computing: Vision and challenges. *IEEE Internet Things* **2016**, *3*, 637–646. [CrossRef]
12. Abolfazli, S.; Sanaei, Z.; Ahmed, E.; Gani, A.; Buyya, R. Cloudbased augmentation for mobile devices: Motivation, taxonomies, and open challenges. *IEEE Commun. Surveys Tuts.* **2014**, *16*, 337–368. [CrossRef]
13. Zhang, H.; Qiu, Y.; Chu, X.; Long, K.; Leung, V.C.M. Fog radio access networks: Mobility management, interference mitigation, and resource optimization. *IEEE Trans. Wireless Commun.* **2017**, *24*, 120–127. [CrossRef]
14. You, C.; Huang, K.; Chae, H.; Kim, B. Energy-efficient resource allocation for mobile-edge computation offloading. *IEEE Trans. Wireless Commun.* **2017**, *16*, 1397–1411. [CrossRef]
15. Dai, L.; Wang, B.; Ding, Z.; Wang, Z.; Chen, S.; Hanzo, L. A survey of non-orthogonal multiple access for 5G. *IEEE Commun. Surveys Tuts.* **2018**, *20*, 2294–2323. [CrossRef]
16. Sun, Y.; Ng, D.W.K.; Ding, Z.; Schober, R. Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems. *IEEE Trans. Commun.* **2017**, *65*, 1077–1091. [CrossRef]
17. Fang, F.; Zhang, H.; Cheng, J.; Roy, S.; Leung, V.C.M. Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2874–2885. [CrossRef]
18. Kiani, A.; Ansari, N. Edge computing aware NOMA for 5G networks. *IEEE Internet Things* **2018**, *5*, 1299–1306. [CrossRef]
19. Ding, Z.; Fan, P.; Poor, H.V. Impact of non-orthogonal multiple access on the offloading of mobile edge computing. *IEEE Trans. Commun.* **2019**, *67*, 375–390. [CrossRef]
20. Fadhil, M.; Abdullah, N.F.; Ismail, M.; Nordin, R.; Saif, A.; Al-Obaidi, M. Power Allocation in Cooperative NOMA MU-MIMO Beamforming Based on Maximal SLR Precoding for 5G. *J. Commun.* **2019**, *14*, 676–683. [CrossRef]

21. Shayea, I.; Ergen, M.; Azmi, M.H.; Aldirmaz-Colak, S.; Nordin, R.; Daradkeh, Y.I. Key Challenges, Drivers and Solutions for Mobility Management in 5G Networks: A Survey. *IEEE Access Spec. Sect. Complex Netw. Anal. Eng. 5G 6G* **2020**, *8*, 172534–172552.
22. Fettweis, G.P. The Tactile Internet: Applications and Challenges. *IEEE Veh. Technol. Mag.* **2014**, *9*, 64–70. [CrossRef]
23. Ali, S.S.D.; Zhao, P.H.; Kim, H. Mobile Edge Computing: A Promising Paradigm for Future Communication Systems. In Proceedings of the TENCON 2018—2018 IEEE Region 10 Conference, Jeju, Korea, 28–31 October 2018; pp. 1183–1187. [CrossRef]
24. Ding, Z.; Ng, D.W.K.; Schober, R.; Poor, H.V. Delay minimization for NOMA-MEC offloading. *IEEE Signal Process. Lett.* **2018**, *25*, 1875–1879. [CrossRef]
25. Wu, Y.; Qian, L.P.; Ni, K.; Zhang, C.; Shen, X. Delay-minimization nonorthogonal multiple access enabled multi-user mobile edge computation offloading. *IEEE J. Sel. Signal Process.* **2019**, *13*, 392–407. [CrossRef]
26. Qian, L.P.; Feng, A.; Huang, Y.; Wu, Y.; Ji, B.; Shi, Z. Optimal SIC ordering and computation resource allocation in MEC-aware NOMA NB-IoT networks. *IEEE Internet Things* **2019**, *6*, 2806–2816. [CrossRef]
27. Fang, F.; Xu, Y.; Ding, Z.; Shen, C.; Peng, M.; Karagiannidis, G.K. Optimal task partition and power allocation for mobile edge computing with NOMA. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
28. Chen, X.; Jiao, L.; Li, W.; Fu, X. Efficient multi-user computation offloading for mobile-edge cloud computing. *IEEE/ACM Trans. Netw.* **2016**, *24*, 2795–2808. [CrossRef]
29. Wang, F.; Xu, J.; Wang, X.; Cui, S. Joint offloading and computing optimization in wireless powered mobile-edge computing systems. *IEEE Trans. Wireless Commun.* **2018**, *17*, 1784–1797. [CrossRef]
30. Bi, S.; Zhang, Y.-J.A. Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading. *IEEE Trans. Wireless Commun.* **2018**, *17*, 4177–4190. [CrossRef]
31. Wang, F.; Xu, J.; Ding, Z. Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems. *IEEE Trans. Commun.* **2019**, *67*, 2450–2463. [CrossRef]
32. Wang, S.-C.; Yan, K.-Q.; Liao, W.-P.; Wang, S.-S. Towards a Load Balancing in a Three-Level Cloud Computing Network. In Proceedings of the 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, China, 9–11 July 2010; Volume 1, pp. 108–113. [CrossRef]
33. Pan, Y.; Chen, M.; Yang, Z.; Huang, N.; Shikh-Bahaei, M. Energy efficient NOMA-based mobile edge computing offloading. *IEEE Commun. Lett.* **2019**, *23*, 310–313. [CrossRef]
34. Cao, X.; Wang, F.; Xu, J.; Zhang, R.; Cui, S. Joint computation and communication cooperation for energy-efficient mobile edge computing. *IEEE Internet Things* **2019**, *6*, 4188–4200. [CrossRef]
35. Su, B.; Ni, Q.; Yu, W.; Pervaiz, H. Optimizing Computation Efficiency for NOMA-Assisted Mobile Edge Computing with User Cooperation. *IEEE Trans. Green Commun. Netw.* **2021**, *5*, 858–867. [CrossRef]
36. Ding, Z.; Xu, J.; Dobre, O.A.; Poor, V. Joint power and time allocation for NOMA-MEC offloading. *IEEE Trans. Veh. Technol.* **2019**, *68*, 6207–6211. [CrossRef]
37. He, Y.; Zhao, N.; Yin, H. Integrated networking caching and computing for connected vehicles: A deep reinforcement learning approach. *IEEE Trans. Veh. Technol.* **2018**, *67*, 44–55. [CrossRef]
38. Abataineh, Z. Blind Decoding of Massive MIMO Uplink Systems Based on the Higher Order Cumulants. *Wireless Pers. Commun.* 2018, *103*, 1835–1847. [CrossRef]
39. Cheng, N.; Xu, W.; Shi, W.; Zhou, Y.; Lu, N.; Zhou, H.; Shen, X. Air-ground integrated mobile edge networks: Architecture, challenges, and opportunities. *IEEE Commun. Mag.* **2018**, *56*, 26–32. [CrossRef]
40. Liu, J.; Sheng, M.; Liu, L.; Li, J. Interference management in ultra dense networks: Challenges and approaches. *IEEE Netw.* **2017**, *31*, 70–77. [CrossRef]
41. Javaid, N.; Sher, A.; Nasir, H.; Guizani, N. Intelligence in IoT-based 5G networks: Opportunities and challenges. *IEEE Commun. Mag.* **2018**, *56*, 94–100. [CrossRef]
42. Kucur, O.; Kurt, G.K.; Shakir, M.Z.; Ansari, I.S. Nonorthogonal multiple access for 5G and beyond. *Wireless Commun. Mobile Comput.* **2018**, *2018*, 1–2. [CrossRef]
43. Lin, L.; Zhou, N.; Zhao, Z. Analytical Modeling of NOMA-Based Mobile Edge Computing Systems With Randomly Located Users. *IEEE Commun. Lett.* **2020**, *24*, 2965–2968. [CrossRef]
44. Song, Z.; Liu, Y.; Sun, X. Joint radio and computational resource allocation for NOMA-based mobile edge computing in heterogeneous networks. *IEEE Commun. Lett.* **2018**, *22*, 2559–2562. [CrossRef]
45. Cao, B.; Zhang, L.; Li, Y.; Feng, D.; Cao, W. Intelligent offloading in multi-access edge computing: a state-of-the-art review and framework. *IEEE Commun. Mag.* **2019**, *57*, 5662. [CrossRef]
46. Gu, Q.; Wang, G.; Liu, J.; Fan, R.; Fan, D.; Zhong, Z. Optimal offloading with non-orthogonal multiple access in mobile edge computing. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–5.
47. Chen, M.; Hao, Y. Task offloading for mobile edge computing in software defined ultra-dense network. *IEEE J. Sel. Area. Commun.* **2018**, *36*, 587597. [CrossRef]
48. Wang, P.; Yao, C.; Zheng, Z.; Sun, G.; Song, L. Joint task assignment, transmission, and computing resource allocation in multilayer mobile edge computing systems. *IEEE Internet Things J.* **2019**, *6*, 28722884. [CrossRef]

49. Wu, Y.; Ni, K.; Zhang, C.; Qian, L.P.; Tsang, D.H.K. NOMA-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation. *IEEE Trans. Veh. Technol.* **2018**, *67*, 1224412258. [CrossRef]
50. Vaezi, M.; Amarasuriya, G.; Liu, Y.; Arafa, A.; Fang, F.; Ding, Z. Interplay between NOMA and Other Emerging Technologies: A Survey. *arXiv* **2019**, arXiv:1903.10489.
51. Zhu, L.; Xiao, Z.; Xia, X.; Wu, D.O. Millimeter-wave communications with non-orthogonal multiple access for B5G/6G. *IEEE Access* **2019**, *7*, 116123–116132. [CrossRef]
52. Tariq, F.; Khandaker, M.; Wong, K.-K.; Imran, M.; Bennis, M.; Debbah, M. A Speculative Study on 6G. *arXiv* **2019**, arXiv:1902.06700.
53. Liu, Y.; Qin, Z.; Elkashlan, M.; Ding, Z.; Nallanathan, A.; Hanzo, L. Non-orthogonal multiple access for 5G and beyond. *Proc. IEEE* **2017**, *105*, 2347–2381. [CrossRef]
54. Ding, Z.; Lei, X.; Karagiannidis, G.K.; Schober, R.; Yuan, J.; Bhargava, V.K. A survey non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE J. Sel. Area. Commun.* **2017**, *35*, 21812195. [CrossRef]
55. Fang, F.; Ding, Z.; Liang, W.; Zhang, H. Optimal energy efficient power allocation with user fairness for uplink MC-NOMA systems. *IEEE Wireless Commun. Lett.* **2019**, *8*, 1133–1136. [CrossRef]
56. Qian, L.; Wu, Y.; Ouyang, J.; Shi, Z.; Lin, B.; Jia, W. Latency Optimization for Cellular Assisted Mobile Edge Computing via Non-Orthogonal Multiple Access. *IEEE Trans. Veh. Technol.* **2020**, *69*, 5494–5507. [CrossRef]
57. Mach, P.; Becvar, Z. Mobile edge computing: A survey on architecture and computation offloading. *IEEE Commun. Surv. Tutorials* **2017**, *19*, 1628–1656. [CrossRef]
58. Huynh, L.; Pham, Q.; Nguyen, T.; Hossain, M.; Shin, Y.; Huh, E. Joint Computational Offloading and Data-Content Caching in NOMA-MEC Networks. *IEEE Access* **2021**, *9*, 12943–12954. [CrossRef]
59. Zeng, M.; Nguyen, N.-P.; Dobre, O.A.; Poor, H.V. Delay Minimization for NOMA-Assisted MEC under Power and Energy Constraints. *IEEE Wirel. Commun. Lett.* **2019**, *8*, 1657–1661. [CrossRef]
60. Mollanoori, M.; Ghaderi, M. Uplink scheduling in wireless networks with successive interference cancellation. *IEEE Trans. Mobile Comput.* **2014**, *13*, 1132–1144. [CrossRef]
61. Fang, F.; Wang, K.; Ding, Z. Optimal Task Assignment and Power Allocation for Downlink NOMA MEC Networks. In Proceedings of the 2019 IEEE Globecom Workshops (GC Wkshps), Waikoloa, HI, USA, 9–13 Decembe 2019; pp. 1–6. [CrossRef]
62. Albataineh, Z.; Hayajneh, K.; Bany Salameh, H.; Dang, C.; Dagmseh, A. Robust Massive MIMO Channel Estimation for 5G Networks Using Compressive Sensing Technique. *Int. J. Electron. Commun.* **2020**, *120*, 153197. [CrossRef]
63. Wang, F.; Xu, J.; Ding, Z. Optimized Multiuser Computation Offloading with Multi-Antenna NOMA. In Proceedings of the 2017 IEEE Globecom Workshops (GC Wkshps), Singapore, 4–8 December 2017. [CrossRef]
64. Diao, X.; Zheng, J.; Wu, Y.; Cai, Y. Joint computing resource, power, and channel allocations for d2d-assisted and noma-based mobile edge computing. *IEEE Access* **2019**, *7*, 92439257. [CrossRef]
65. Li, Y.; Xia, S.; Zheng, M.; Cao, B.; Liu, Q. Lyapunov optimization based trade-off policy for mobile cloud offloading in heterogeneous wireless networks. *IEEE Trans. Cloud Comput* **2019**. [CrossRef]
66. Albataineh, Z. Low-Complexity Near-Optimal Iterative Signal Detection Based on MSD-CG Method for Uplink Massive MIMO Systems. *Wireless Pers. Commun.* **2021**, *116*, 2549–2563. [CrossRef]
67. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
68. Mao, Y.; Zhang, J.; Song, S.; Letaief, K. Stochastic joint radio and computational resource management for multi-user mobile edge computing systems. *IEEE Trans. Wireless Commun.* **2017**, *16*, 5994–6009. [CrossRef]
69. Chen, J.; Chen, S.; Wang, Q.; Cao, B.; Feng, G.; Hu, J. iraf, A deep reinforcement learning approach for collaborative mobile edge computing networks. *IEEE Internet Things J.* **2019**, *6*, 70117024. [CrossRef]
70. Fang, F.; Xu, Y.; Ding, Z.; Shen, C.; Peng, M.; Karagiannidis, G.K. Optimal Resource Allocation for Delay Minimization in NOMA-MEC Networks. *IEEE Trans. Commun.* **2020**, *68*, 7867–7881. [CrossRef]
71. Xu, L.; Yu, X.; Gulliver, T.A. Intelligent Outage Probability Prediction for Mobile IoT Networks Based on an IGWO-Elman Neural Network. *IEEE Trans. Veh. Technol.* **2021**, 70, 1365–1375. [CrossRef]