Optimizing Resources Allocation for Fog Computing-based Internet of Things Networks

Xi Li*†, Yiming Liu*†, Hong Ji*, Heli Zhang*, and Victor C.M. Leung†
*Key Lab. of Universal Wireless Comm., Beijing Univ. of Posts and Telecom., P.R. China
†Depart. of Electrical and Computer Eng., Univ. of British Columbia, Vancouver, BC, Canada

***Abstract*—In wireless Internet of Things (IoT) networks with resource-constrained devices, fog computing has been introduced to deal with the computation-intensive applications at the edges of the networks. While fog computing decreases the computation delay and fronthaul traffic data, it also brings the severe challenge on complex resource allocation of the available computation and communication resources under stringent quality of service (QoS) requirements. In this paper, we investigate the problem of tasks scheduling and heterogeneous resource allocation for multiple devices in wireless IoT networks. The IoT devices that collect a massive amount of data need to make proper offloading decision to transfer the data to the fog computing nodes (FNs). Moreover, to support a massive number of device connections and transfer a huge amount of data with low latency and limited resource, we consider the deployment of non-orthogonal multiple access (NOMA) in IoT networks, which enables multiple IoT devices to simultaneously transmit data to the same FN in the same time, frequency and code domain. We jointly optimize the allocation of resource blocks and transmit power of multiple IoT devices, subject to respective QoS requirements. Furthermore, the optimization problem is formulated as a mixed-integer nonlinear programming problem to minimize the system energy consumption. Since it is a NP-hard problem, we introduce an improved genetic algorithm (IGA) to solve it. Simulation results show that the proposed scheme achieves good performance in throughput, delay, outage probability and energy consumption.**

***Index Terms*—Fog computing, Internet of things, non-orthogonal multiple access, resource allocation, energy consumption.**

I. Introduction

Encouraged by the fast development of the fifth generation (5G) and beyond mobile communication systems, Internet of Things (IoT) [1] networks have been researched to enable future smart systems such as smart grids, smart homes and smart cities with ubiquitously connected devices. As reported by Cisco, it is expected that more than 50 Billion devices will be connected to the Internet by 2020 [2]. In order to support the popular computation-intensive applications, such as image/video processing, virtual reality, inter-active games, etc. [3] [4], the IoT networks need to collect and process with lots of data in the limited time. On one hand, many of the IoT devices have severe constraints in battery life, processing ability and storage resources, so that they could not complete these computation tasks by themselves. On the other hand, the amount of data generated by various sensors and mobile devices is rapidly increasing and becoming huge, and there are great pressure on communication and computation in the network under different quality of service (QoS) requirements. Then, fog computing is proposed as an attractive solution to extend the cloud computing paradigm to the local networks [5]. By taking full advantage of the available resources in the edge nodes, the combination of fog computing and IoT networks has attracted extensive attention in both academia and industry.

In the fog computing, various edge nodes can cooperate to share computing, caching and communication resources so as to complete some computation tasks locally without interacting with the cloud computing center via the fronthaul links. These fog nodes (FNs) include but are not limited to the base stations (BSs), Wi-Fi access points (APs), relays and routers. The computation tasks proposed in the user terminals can be offloaded to the FNs when necessary. In this paper, we investigate the optimal heterogenous resource allocation problem in fog computing-based IoT networks to fulfill the computation and latency requirements of the IoT applications, with the target to reduce energy consumption of the local computation in the battery-powered IoT devices. While fog computing addresses the computation problem for the IoT devices, the huge amount of data offloaded from multiple IoT devices to FNs require substantial communication resources to satisfy both the bandwidth and latency requirements. Then, both computation and communication resources need to be jointly scheduled to improve the system performance.

In the IoT networks, it is common that there are multiple IoT devices requiring to access to the same FN. In this paper, non-orthogonal multiple access (NOMA), which is considered as the key technique in 5G, is adopted to support low-latency offloading, and improve the spectrum energy (SE) and energy efficiency (EE) at the same time [6].The key idea of NOMA is to exploit the power domain for multiple access, i.e., multiple transmissions at different power levels could simultaneously share the same channel in the same time, frequency and/or code domains [7]. Then, successive interference cancellation (SIC) is applied at the receiver to separate and decode the superimposed signals [8]. Therefore, there are different requirements and processing procedures for the transmitter and the receiver. In this paper, we consider the offloading problem from the massive IoT devices to the FNs over the uplink in the typical IoT networks. In this case, multiple IoT devices transmit data using different power levels on the same spectrum resource to the FN via the collocated BS. Although some neighbor IoT devices may experience similar fading channel conditions, their transmit power levels may be controlled by the FN to enable differentiation of these signals in the power domain at the receiver as required by the NOMA technique [9]. In this manner, the receiver can apply SIC successfully and decode these superimposed signals. Therefore, by exploiting the channel gain differences or allocating different transmit power levels to enable multiple access in the power domain from multiple IoT devices, IoT networks employing NOMA on the uplink could substantially increase their capacities by offloading data to FNs.

While the introduction of fog computing and NOMA into wireless IoT networks has the potential to bring substantial benefits, the performance and optimization of such networks have not been widely studied, and there are still many open issues that need to be addressed [10]. Consider multiple IoT devices to offload computation tasks to the same FN with different service/application- dependent QoS requirements. How to properly assign the partitioned tasks from a set of IoT devices to the given FN, and then allocate the required computation and communication resources, are important and challenging problems. Furthermore, with increasing concern of sustainability leading to significant research activities on green communications and networking, minimizing the energy consumption of IoT networks with a massive number of connected devices and supporting diverse applications is another important research problem. In this context, the interference levels among the IoT devices sharing the same NOMA uplink need to be properly controlled to maximize EE while satisfying QoS requirements. Therefore, it is important to address the complex and challenging problem of joint allocation of heterogenous resources, considering several constraints and many impacting factors, as discussed above.

In this paper, we tackle the problem of optimization for heterogenous resources with multiple IoT devices offloading in fog computing-based IoT networks, with the goal of minimizing the system energy consumption. Various computation and communication resources are considered jointly within both FNs and IoT devices under NOMA uplinks. Comparing within existing research work, this problem still needs further investigation.

A. Related Works

Recently, IoTs have gained wide popularity in many domains such as smart home, health monitoring, environmental and agricultural applications [11]–[13]. For processing a huge amount of data generated by IoT devices, various ideas have been proposed to address many problems about fog computing in wireless IoT networks. In [14], it proposed a service placement policy in IoT networks based on graph partitions to increase the service availability and QoS satisfaction. In [15], the authors discussed the application partitioning rationale of wearable devices in mobile cloud and fog computing for computation offloading. In order to solve the user association and resource allocation problem for broadband IoT applications in fog computing, a two-side matching game was formulated based on the determination of QoS requirements priorities in [16]. The authors in [17] introduced drone base stations to mitigate the heavy traffic loads of macro base stations and designed algorithms for optimal drone placement and user association in fog IoT networks. The fair offloading among several FNs was investigated in [18] with consideration on energy consumption. The authors proposed an algorithm about power control and virtual machines rental costs in the fog-aided IoT networks in [19]. While the scenarios with multiple fog nodes have attracted wide attention, another typical scenario with only one FN and several IoT devices also has many interesting problem under investigation. The authors in [20] proposed a joint resource allocation and coordinated computation offloading algorithm for the fog radio access network where one fog node is deployed for serving multiple users. The authors in [21] developed a game theoretical model and proposed a decentralized algorithm for allocating the computational tasks among several local devices and one edge cloud. In [22], the authors investigated the computation offloading problem in a mixed fog and cloud computing system, which is composed of a small-cell based fog node, a powerful cloud center, and a group of users. However, these works adopt orthogonal multiple access technique which may suffer severe spectrum competition and result in heavier interference and longer delay.

On the other hand, there are increasing research activities focused on NOMA to improve the radio resource utilization in wireless IoT networks. System-level and link-level simulations in [23] indicated obvious benefits of NOMA over orthogonal multiple access (OMA) in terms of SE as well as EE. In [24], an energy-efficient transmission method has been proposed for NOMA systems. In [9], the authors proposed a random NOMA strategy for massive IoT, where multiple devices were allowed to transmit over the same sub-band. In [25], the authors adopted NOMA to support the communications of massive IoT devices and analyzed the practical challenges as well as future research directions. In [26] and [27], the authors analyzed the EE of downlink NOMA for practical heterogeneous cloud radio access networks and illustrated that the power available in the cloud, the propagation environment and cell types could have significant impacts on the EE performance. Considering NOMA and edge computing jointly, the authors in [28] proposed an edge computing-aware NOMA framework to reduce users' uplink energy consumption. However, they only focused on the optimization of transmission energy consumption, but neglected the energy consumption of task execution. Moreover, although they investigated the problem in multi-user single-cell scenario, they only considered the single task model for each user offloading to the eNB equipped with a cloudlet, which is not applicable for IoT networks.

Although several recent research works have been carried out on fog computing and NOMA, respectively, there still needs further study that jointly considering fog computing and NOMA for heterogenous resources allocation in wireless IoT networks. The authors in [29] discussed the IoT device clustering and energy management problem in NOMA system. They focused on the effective clustering method based on wireless channel condition of the NOMA links. As to the resource allocation, only transmit power was considered for spectrum efficiency and fairness. In [30], another condition that a IoT device offloads its computation tasks to multiple FNs by NOMA downlink was investigated, and the corresponding problem was formulated under delay and energy constraints. However, as to the heterogeneous resources allocation in fog computing-based IoT networks with NOMA to minimizing system energy consumption, the existing work has not tackled this problem with thorough consideration. That is the motivation of this paper.

B. Main Contributions

In this paper, we investigate the heterogenous resources allocation problem to minimize the system energy consumption in the fog computing-based wireless IoT networks with NOMA. The major contributions of this paper are summarized as follows.

- We consider a general scenario of fog computing-based wireless IoT networks with NOMA, which supports multiple IoT devices to offload computation tasks simultaneously to the same FN. Heterogenous resources of computation and communication in various devices and the FN are jointly considered under respective QoS requirements. The offloading decision as well as the allocation of resource blocks (RBs) and transmit power are optimized as a complex problem to minimize system energy consumption.
- The formulated problem is a mixed integer nonlinear programming (MINLP) problem. In order to achieve a good balance between performance and complexity, we introduce an improved genetic algorithm (IGA) to solve this problem and obtain the suboptimal solution.
- Simulation results are presented to evaluate the performance of the proposed scheme. It is shown that compared with existing algorithms, the proposed scheme has better performance in throughput, delay, outage probability and energy consumption.

The remainder of our work is organized as follows. Section II presents the network model, communication model and computation model. Then we formulate the optimization problem and provide detailed explanations in Section III, and solve this MINLP problem by IGA. Simulation results are presented in Section IV with the discussion about these results. Finally, in Section V we conclude the work of this paper and give the future research directions.

II. System Model

In this section, we introduce the network model of the considered scenario, as well as the communication model and computation model for the fog computing-based wireless IoT networks with NOMA.

A. Network Model

We consider a general system model for fog computing-based wireless IoT networks as shown in Fig. 1. Dedicated FNs are deployed to provide offloading services for M IoT devices. In practice, these FNs are usually played by the network edge nodes such as routers, switches, access points, or BSs. There are a massive number of devices in IoT networks, including wearable devices, smart phones, cameras, sensors, etc. They may generate a huge amount of data and support many computation-based applications with respective deployment constraints. The authors in [31] discussed the analysis methods of huge amount of data for intelligent energy networks and provided a comprehensive and solid reference in the data processing methods. Considering the limited battery life, computing and storage capabilities of these devices, the generated data might be processed locally or by FNs. Here we assume that each FN has its own authority areas. They would manage the computation offloading of the IoT devices in the area according to the respective QoS requirements. In this paper, we consider that there are two computation modes, i.e., to compute locally in the IoT device, or to offload computation tasks to an FN. Either modes would be chosen to finish the task.

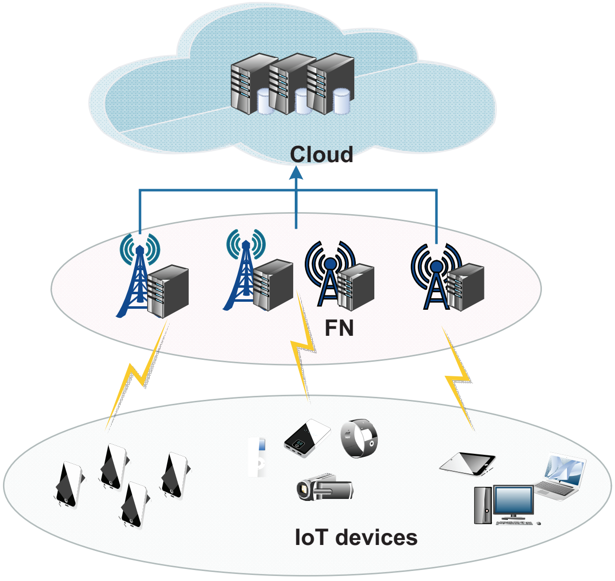

Fig. 1: System model of fog computing-based wireless IoT networks.

It is assumed that the total frequency band is divided into K RBs, and multiple IoT devices can occupy the same spectrum resource to transfer data to the same FN by NOMA. In this paper, we assume that the fading channel follows a Rayleigh distribution with mean one, which implies that the channel gain is exponentially distributed with mean one [32]. As pointed in [33], we assume μ captures the effects of path loss and fading and model μ as zero-mean, independent, circularly symmetric complex Gaussian random variables with variance one, so that μ is Rayleigh distributed and μ^2 is exponentially distributed with parameter one. In addition, we assume that perfect channel state information (CSI) is available, enabling the receiver to perform the interference cancellation perfectly. Therefore, SIC could be applied successfully and multiple signals could be decoded and received as expected.

Note that in practice, interference cancellation may not be realized perfectly as it is impacted by channel estimation errors. This may result in incorrect user ordering for decoding, which in turn affects the SIC decoding accuracy. Then the average data rate of NOMA system degrades due to the fact that imperfect CSI introduces not only extra interference on the desired signal but also an incorrect decoding order [34]. To deal with this problem caused by imperfect CSI, the authors in [35] investigate the dynamic-ordered decoding scheme with SIC for uplink NOMA system to improve the performance. Moreover, channel estimation techniques have been well researched for many years. Some effective designs have been proposed for striking a good tradeoff between complexity and performance. From [36], [37], interference cancellation can perform well with the instantaneous CSI and different channel gains between the transmitters and receivers. In this paper, we focus on the joint optimization of computation and communication resources, and assume perfect interference cancellation in the receiver. The condition with imperfect interference cancellation would be investigated in our future work.

The complexity of SIC at the receiver is directly proportional to the number of accessing devices. Therefore, most existing research consider two devices as example. However, there are also some researchers discussing the decoding issues of SIC in multi-user NOMA systems [9], [37], and then pointed out the feasibility of adopting NOMA in IoT networks with massive devices. We assume that there are up to L IoT devices ($L < M$) on the same RB for tasks offloading. We also consider that the wireless channel does not vary during the transmission of a packet and perfect instantaneous CSI is available.

We denote m as the index for the m-th IoT device where $m \in \{1, 2, \cdots, M\}$, n as the index for the n-th task where $n \in \{1, 2, \cdots, N\}$ and k as the index for the k-th RB where $k \in \{1, 2, \cdots, K\}$.

B. Communication Model

In this subsection, we present the communication model in the wireless IoT networks with NOMA. In an uplink NOMA system, a set of IoT devices would send their data to the FN on the same RB at the same time by different power levels. Suppose that the m-th IoT device offloads computation tasks to the FN for processing by transmitting the signal x_m^k with transmission power p_m^k on the k-th RB to the FN. The received signals y_m^k from the m-th IoT device on the k-th RB can be written as

$$y_m^k = \sqrt{p_m^k} h_m^k x_m^k + \sum_{i \neq m, i \in M} \sqrt{p_i^k} h_i^k x_i^k + z_m^k, \tag{1}$$

where the first term is the desired signal, in which h_m^k represents the channel gain for the m-th IoT device connecting to the FN on the k-th RB; the second term represents the interference from other IoT devices on the same RB; the last term z_m^k is additive white gaussian noise (AWGN) with zero mean and variance δ^2.

In the fog computing-based IoT networks with NOMA, multiple IoT devices transmit their offloading data to the same FN on the same RB. These signals may cause interference to each other, which could be resolved by performing SIC in the receiver for separating and decoding the superimposed signals. This requires the receiver to control the transmit power of these IoT devices to enable differentiation of these signals in the power domain. Therefore, although some neighbor IoT devices may experience similar channel fading, their received signals could also be distinguished by the receiver successfully [9], [38]. In order to efficiently apply SIC at the FN, we introduce a specific power constraint in the following problem formulation.

By applying SIC technique, the receiver would first decode the strongest received signal by treating others as interference, and then remove it before decoding the second strongest signal. Consequently, the IoT device whose signal strength is the highest experiences interference from all the other IoT devices sharing the same uplink. Then the similar procedure is followed by the second strongest signal, which has in fact become the strongest signal in the second stage. When all but one of the signals is detected, the weakest signal from the IoT device is decoded without suffering from any interference anymore. All the IoT devices connected to the same FB on the k-th RB are sorted into a descending order according to the channel gains, and can be expressed as

$$|h_1^k|^2 \geq |h_2^k|^2 \geq ... \geq |h_M^k|^2 \quad \forall k \in K. \tag{2}$$

According to these orders of RBs, the FN can successfully decode the superposed signals. Therefore, in the FN, the received signal-to-interference-plus-noise ratio (SINR) of the m-th IoT device on the k-th RB is given by

$$\mathit{SINR}_m^k(\mathbf{p}) = \frac{p_m^k |h_m^k|^2}{\delta^2 + \sum\limits_{i=m+1}^{M} p_i^k |h_i^k|^2}. \tag{3}$$

The corresponding data rate of the m-th IoT device that transmits to the FN on the k-th RB can be denoted as

$$R_m^k(\mathbf{p}) = \log_2(1 + \ \mathit{SINR}_m^k). \tag{4}$$

As a result, the achievable data rate of the m-th IoT device is

$$R_m = \sum_{k \in K} b_m^k \mathcal{R}_m^k, \tag{5}$$

where b_m^k denotes the result of RBs allocation for IoT devices, with $b_m^k = 1$ indicating the k-th RB is allocated to the m-th IoT device for offloading; otherwise, $b_m^k = 0$.

C. Computation Model

We consider that the m-th IoT device has a computation workload that is partitioned into a series of tasks. These tasks could be completed either locally on the device or remotely on the FN by offloading via wireless links. In general, the computing resource of the FN can be modeled as a multi-dimensional vector, representing the capabilities of the central processing units (CPUs), memory and network interfaces. For the ease of analysis, we only consider scalar computing capability in this paper. Then, we denote C_m as the total computing capability in terms of the number of CPU cycles per second of the m-th IoT device, and denote C_e as the total computing capability of the FN.

The computation workload needs to be partitioned into a sequence of tasks for scheduling. These tasks are denoted as $F_{mn} = (A_{mn}, D_{mn}), m \in \{1, 2, ...M\}, n \in \{1, 2, ...N\}$, in which A_{mn} presents the size of computation input data and D_{mn} presents the total number of CPU cycles required to accomplish the n-th task of the m-th device. We assume that the IoT devices have two choices to perform computation tasks, namely local computing and the FN offloaded computing. For a given partitioned task F_{mn}, a_{mn} represents that the FN chooses the m-th IoT device to provide the offloading service. Note that the computational tasks offloaded by different devices are dynamic since each device randomly sends its data through the selected RB. Then the number of coded symbols that need to be transmitted over each RB is random correspondingly. To deal with this issue, Raptor codes [39], [40], which generates as many coded symbols as required by the FN, could be used in NOMA-based IoT networks [9]. Moreover, the FNs are deployed practically in a certain region, where includes massive sensors or other devices. Usually, these sensors and devices are the same or similar types with close distance, so the amount and types of collected data for computing offloading in the same period do not vary too much. It is convenient to design the transmission of NOMA technique with fog computing in IoT networks. Next, we evaluate the computation cost in terms of both energy consumption and processing time for both local and FN computing modes.

1) Local Computing: For the local computing approach, the IoT devices process their computation tasks locally by individual computing resources. The computation execution time of the n-th task of the m-th IoT device by local computing can be defined as the ratio of the total required number of CPU cycles to the assigned local computing resources, which is given by

$$T^l_{mn} = \frac{D_{mn}}{C_{mn}}, \quad \forall m \in M, n \in N. \tag{6}$$

Then, according to the measurements of energy consumption [41], the energy for processing task n of the m-th device locally can be given by

$$\mathcal{P}^l_{mn} = \xi D_{mn} C^2_{mn}, \quad \forall m \in M, n \in N, \tag{7}$$

where ξ is the coefficient presenting the consumed energy per CPU cycle. It depends on the average switched capacitance and the average activity factor [41].

2) FN Computing: For the FN computing approach, the IoT device offloads its tasks to the authorized FN through wireless links. Then, the FN with sufficient computation and storage capabilities would process these tasks for the IoT devices, and send the computing results as required. During the procedure, it would incur an extra overhead in terms of time and energy for transmitting the relative data between the IoT devices and the FN through wireless links. Note that the amount of data in uplink transmissions from IoT devices to the FNs is usually very huge, and in this paper we focus on the uplink offloading communications here. According to the communication model, we can compute the transmission time and energy consumption of IoT devices for sending the computation input data, respectively. The computation capability for executing task n in terms of the number of CPU cycles per second of the FN is denoted as C_{0n}. Suppose that the m-th IoT device offloads its computation task F_{mn} to the FN, the transmission time of offloading the n-th task of the m-th IoT device is defined as the ratio of the sizes of the offloaded tasks to the transmission data rate, which is denoted by

$$T^c_{mn,t} = \frac{A_{mn}}{R_m}, \quad \forall m \in M, n \in N. \tag{8}$$

Similarly, the computation execution time of the n-th task of the m-th IoT device in the FN is given by

$$T^c_{mn,e} = \frac{D_{mn}}{C_{0n}}, \quad \forall m \in M, n \in N. \tag{9}$$

On the other hand, the energy consumption for offloading task n of IoT device m is given by

$$\mathcal{P}^c_{mn} = \sum_{k=1}^{K} T_{mn,t} p^k_m + \eta D_{mn} C^2_{0n}, \quad \forall m \in M, n \in N, \tag{10}$$

where η is the coefficient for the consumed energy per CPU cycle of the FN. It depends on the average switched capacitance and the average activity factor. The first term represents the transmission energy consumption and the second term represents the computing energy consumption of the FN.

III. Optimizing Heterogeneous Resource Allocation Problem

Based on the above system models, we investigate the problem of optimization of computation and communication resources. System energy consumption is adopted as the optimization goal, and the QoS requirements of the IoT devices are also considered in the constraints. Since the proposed problem is an MINLP problem, IGA algorithm is introduced to solve it with low complexity.

A. Problem Formulation

From the above analysis, we could see that to support computation-intensive applications in wireless IoT networks, we need to consider several problems and constraints.

1) Where to process the computation task: Here we give two methods as **local computing** and **FN computing**. The first one is based on the local resources of the IoT devices and has no transmission delay or extra costs. However, due to the limited capabilities of the devices, this method might put a heavy burden on a device, which might even be unable to finish the task. The latter one could utilize the available computing and storage capabilities of the FN, but with extra costs on communication resources to transmit offloading data. Also, when there are massive devices, the FNs have to control the access and manage the available resources.

2) How to get the computation result efficiently: For every computing task, it should be finished before the time deadline. In some conditions, the computation result needs to be returned to the IoT devices, such as inter-active games in smart phones. In other conditions, the computation result would be sent forward to the dedicated servers for further processing, such as environment sensing for weather forecast. Here we focus on the offloading problem, and it is obvious that our work could be applied to the both conditions aforementioned without loss of generality. For the offloading cases, it requires that firstly, the device could access to the FN by available wireless links for offloading; secondly, the FN has sufficient computation capability at present to process the tasks in time.

3) The necessary resources during the procedure: It is obvious that both computation resources in the IoT devices and the FN, and communication resources of uplink should be considered. The limited wireless transmission resource could support a larger number of devices simultaneously by adopting NOMA. Note that because the FN usually has a maximum access users limit, there is still an accessing problem for the networks.

4) The energy consumption of the system: The computation procedure and wireless offloading transmission would require sufficient power to support. Since the computation energy consumption has direct relation with the task itself, we pay more attention on the wireless offloading. As discussed before, the uplink NOMA-based access may improve the system energy efficiency. The transmission power allocation is very important for the entire network energy consumption.

To enable efficient tasks offloading in fog computing-based wireless IoT networks with NOMA, we formulate the optimization problem as a joint computation and communication resources allocation problem to reduce the system energy consumption under the QoS requirements of the IoT devices. The computation resources are distributed among the FN and multiple IoT devices, which have respective computation capabilities. The communication resources include available RBs and transmit power for the multiple IoT devices. Therefore, it is a complicated problem about heterogeneous resources allocation.

$$
\begin{aligned}
\min_{a_{mn},b_m^k,p_m^k} \quad & \sum_{m=1}^{M}\sum_{n=1}^{N}\{(1-a_{mn})\mathcal{P}_{mn}^{l}+\sum_{k=1}^{K}a_{mn}b_m^k\mathcal{P}_{mn}^{c}\},\\
\texttt{s.t.}\quad & C1: a_{mn}\in\{0,1\}, b_m^k\in\{0,1\} \quad \forall m,n,k\\
& C2: \sum_{m=0}^{M}a_{mn}=1 \quad \forall n\\
& C3: \sum_{m=0}^{M}b_m^k\leq L \quad \forall k\\
& C4: p_m^k-\sum_{i=m+1}^{M}p_i^k\geqslant p_{thr} \quad \forall m,k\\
& C5: \sum_{m=1}^{M}p_m\leq P_{max}\\
& C6: a_{mn}(T_{mn,t}^{c}+T_{mn,e}^{c})\leq T_{mn,req} \quad \forall m,n\\
& C7: (1-a_{mn})T_{mn}^{l}\leq T_{mn,req} \quad \forall m,n.
\end{aligned}
\tag{11}
$$

Hence, based on the discussion above, we have the problem formulated as follows in (11). Here constraint C1 includes two Boolean variables. a_{mn} indicates the result of joint computing offloading decision and task assignments and b_m^k indicates the result of RBs allocation for IoT devices. We consider each device has a set of independent tasks, which denoted as $n\in 1,\cdots,N$. The tasks of different users are independent and processed individually. For an arbitrary device m, it has a set of tasks N and each of them $n\in 1,\cdots,N$ is processed by two ways, i.e., local computing at its device m or edge computing at the FN. Then, we derive the constraint C2 that means that each task in the system can be either assigned to the FN for offloading or executed on IoT device locally. Note that each device can execute more than one tasks. Constraint C3 indicates each RB in the system can be assigned up to L IoT devices. Furthermore, in order to support different IoT devices transmission in the same RBs by NOMA, there is additional restriction. Different from the OMA techniques, in the discussed NOMA-based system, the fog node applies SIC technique to decode and subtract useful signals. SIC is achieved in the receiver by decoding the strongest signal first, subtracting it from the combined signal, and then decoding the second strongest signal and repeating the procedure until all the signals successfully decoded. Thus, it is very important to guarantee those received signals have different strength for successfully decoding by proper transmit power control. This is reflected in Constraint C4. p_{thr} is the minimum power difference required to distinguish between the signal to be decoded and the remaining non-decoded message signals. Constraint C5 requires that the overall power consumption should be less than the maximum of available power P_{max}. Constraints C6 and C7 limit the maximum of the tolerated delay $T_{mn,req}$ for offloading the tasks and executing locally, respectively.

B. Low-complexity Sub-optimal Solution

The problem (11) is an MINLP problem, for which it is extremely difficult to obtain a globally optimal solution with low complexity. For ease of practical implementation, we propose a close-to-optimal solution method with lower complexity by leveraging IGA.

Genetic algorithms (GAs) are heuristic algorithms that inspired by the natural selection and evolutionary genetics [42]. With a good balance between the complexity and the effectiveness, GAs provide close-to-optimal solution to NP-hard problem. The chromosome (also known as the individual) represents a possible solution of the objective formula (11), and it can be designed with binary, real or integer representation. Specifically, in order to meet several constraints in formula (11), we introduce the penalty function and update the process of IGA, such as crossover and mutation. Although we cannot get the optimal solution, we still achieve the sub-optimal solution which is very close to the optimal one with the limited iteration, as shown in simulation results of Fig. 6.

In this paper, we adopt the real representation of the chromosomes to reduce the complexity of encoding and decoding chromosomes. The length of the chromosome is related to the number of tasks, IoT devices and RBs as $M\times N+M\times K$. The

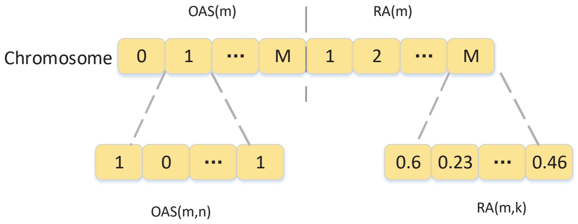

Fig. 2: The structure of the chromosome.

structure of the chromosome includes two parts: offloading decision and resource allocation, as shown in Fig. 2. For the first part, each gene of the chromosome represents the decision of the IoT devices for accessing to the FN and offloading task, which is denoted as $OAS_m(n)$. When the FN chooses the m-th IoT device for processing the n-th task, $OAS_m(n) = 1$; otherwise, $OAS_m(n) = 0$. For the second part, each gene of the chromosome represents the resource allocation results, including the RBs assignment and power allocation, which are denoted as $RA_m(k)$. To deal with the combinatorial constraint of resource allocation, we relax b_m^k to be real-valued variables. Then we introduce a variable $c_m^k = b_m^k \times p_m^k$ to interpret the RBs assignment and power allocation. Obviously, the IoT device will not allocate any power over an RB if the RB is not occupied by the IoT device. Then, $p_m^k = \frac{c_m^k}{b_m^k}$, except for $b_m^k = 0$.

In general, IGA has four operations as selection, crossover, mutation and fitness evaluation, as shown in Fig. 3. Next, we give the process of solving the proposed problem by using IGA.

1) Initialization

A union of chromosomes is called a population, and the initial population should provide possible solutions that are diverse enough to enable the optimal solution to be found eventually. Here, we choose chromosomes randomly as the first generation.

The initial population size is S. And each chromosome contains two parts: OAS part, and RA part. Each of them is comprised of $M * N$ genes sequence and $M * K$ genes sequence, respectively.

2) Evaluation

After initializations, the fitness values are derived for each chromosome of the current population based on the formula (11). Its purpose is to select better chromosomes to reproduce the next generation as parents. Judging whether a chromosome is fine or not is based on its own fitness. The lower the fitness value is, the better the chromosome is. However, in our case, not all the possible chromosomes are feasible solutions, due to the requirements of the constraints of formula (11). To satisfy these constraints, we adopt the idea of penalty function solving constrained optimization problem from [43]. Then a penalty function $penalty(m, \mathbf{g})$ is introduced to measure the constraints violation and refer to inequality constraints (the equality constraints are transformed into inequality constraints). The basic idea is that the feasible solutions have superiority over unfeasible ones, and the infeasible solutions are penalized to provide a search direction towards the feasible region. For example, to fulfill constraint C4, $penalty(m, \mathbf{g})$ should include the item $\{p_m^k - \sum_{i=m+1}^{M} p_i^k - p_{thr}\}$. The fitness function of chromosome $\mathbf{g}$ follows the rules defined in [44] and is shown as

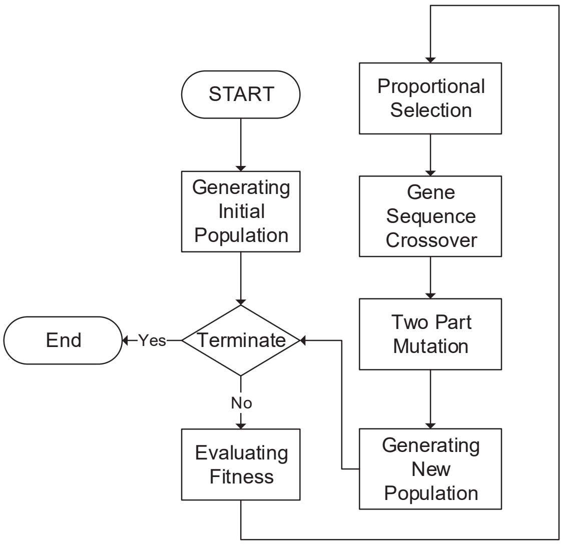

Fig. 3: Flow diagram of IGA.

$$Fit(\mathbf{g}) = \begin{cases} \mathcal{E}_{tot} & \mathbf{g} \text{ is feasible} \\ \mathcal{E}_{tot} + \lambda \sum_{m=1}^{M} penalty(m, \mathbf{g}) & \mathbf{g} \text{ is not feasible} \end{cases} \tag{12}$$

where $\mathcal{E}_{tot} = \sum_{m=1}^{M} \sum_{n=1}^{N} \{(1 - a_{mn})\mathcal{P}_{mn}^l + \sum_{k=1}^{K} a_{mn} b_m^k \mathcal{P}_{mn}^c\}$ is the objective function in problem (11), λ is the penalty factor representing the degree of penalty.

By making individuals with better fitness reproduce offsprings, GA could search in both feasible and infeasible regions, and then obtain global sub-optimal solution more easily.

In order to preserve good chromosomes to yield better offsprings, we employ the roulette wheel method which selects two individuals from the survival chromosomes to produce two new offsprings. The selection probability p_s is defined as

$$p_s = F_s \Big/ \sum_{s=1}^{S} F_s \,, \tag{13}$$

where F_s denotes the fitness value of the s-th population.

3) Crossover

To ensure the IGA process to converge more efficiently, we adopt the gene sequence level crossover method, which means the smallest unit of crossover in each step is a gene sequence like OAS_m or RA_m [45]. That is to say, all of the genes that belong to a certain gene sequence, like all the $OAS_m(n), n \in \{1, ..., N\}$ that belong to OAS_m and all the $RA_m(k), k \in \{1, ..., K\}$ that belong to RA_m, will crossover with the other gene sequence as a whole package,

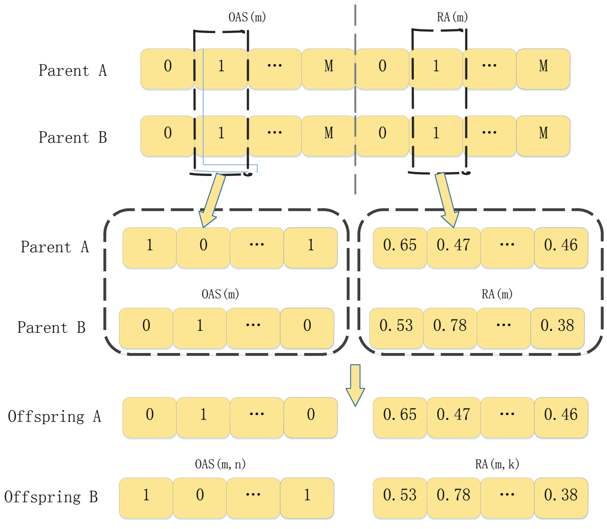

Fig. 4: The process of crossover operation.

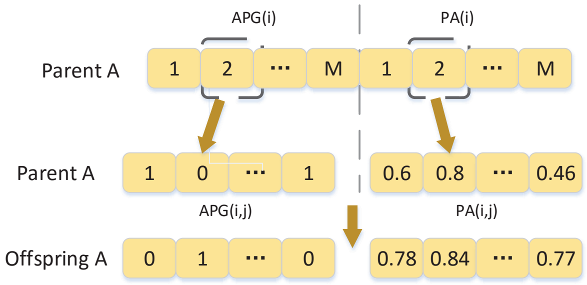

Fig. 5: The process of mutation.

respectively. Here, we use a sequence $\alpha_m, m \in \{1, ..., M\}$ which follows Bernoulli distribution to determine whether the crossover happens on OAS_m and RA_m. And this process can be expressed as

$$\begin{aligned} OAS_a^{g+1} &= (1-\alpha_m) * OAS_a^g + \alpha_m * OAS_b^g \\ RA_a^{g+1} &= \alpha_m * RA_a^g + (1-\alpha_m) * RA_b^g. \end{aligned} \tag{14}$$

For example, when $\alpha_2 = 1$, the crossover of parents would happen to OAS_2 and RA_2. All the $OAS_2(n), n \in \{1, ..., N\}$ and $RA_2(k), k \in \{1, ..., K\}$ would exchange between parent A and parent B, as shown in Fig. 4. While if $\alpha_2 = 0$, the crossover of parents will not happen. Please note that all the OAS_m and RA_n are exchanged according to formula (14) in the crossover stage.

4) Mutation

As each chromosome contains two parts as OAS_m and RA_m, the mutation is divided into two parts: the integer mutation for the offloading decision and the uniform mutation for the resource allocation. We randomly generate a 2-element mutation mask sequence $\beta_m, m \in \{1, ..., M\}$ comprising of 1 and 0. If the element of the mutation mask is 1, then the gene will have mutation. While if the element of the mutation mask is 0, the gene will not have mutation.

For the OAS part, if the mutation happens, the gene will be exchanged by the opposed elements.

For the RA part, if the mutation happens, the gene will be replaced in the offspring and can be expressed as:

$$RA^{g+1}(m,k) = RA_a^g(m,k) + \beta(m) * (p_{max}(m) - p_{tot}(m)), \tag{15}$$

where $p_{max}(m)$,$p_{tot}(m)$ are the maximum power of the m-th IoT device and the total transmit power of the m-th IoT device, respectively.

For example, if $\beta_2 = 1$, the mutation of parents would happen to OAS_2 and RA_2. All the $OAS_2(j)$ are exchanged by the opposed elements and $RA_2(j)$ is exchanged based on (15), as shown in Fig. 5. Please note that all the OAS_m and RA_n will exchange according to formula (15) in the mutation stage.

5) Iteration

If the fitness values from (11) of all chromosomes are smaller than the maximum one in the previous iteration, then we will replace the chromosome with the smallest fitness value by the best one in the previous iteration. The goal of this step is to ensure that the fitness values form a non-decreasing sequence as required for convergence of IGA. The iteration is repeated until the maximum number of generation is reached. And the chromosome with the maximum fitness value is then chosen as our solution.

IGA repeats the aforementioned steps 1 to 5 until the number of the generation exceeds the limitation, to get the desired solution. The detailed steps are enumerated in **Algorithm 1**.

C. Algorithm Complexity

Here we analyze the complexity of the proposed scheme. The complexity of IGA is mainly decided by its encoding, selection, fitness calculation, crossover and mutation operations with IGA's generation number and population size. In the proposed scheme, the computation complexity is linear to the number of generation, population, tasks, IoT devices and RBs and is denoted as $\mathcal{O}(GSM(N+K))$. The FN is responsible for the user scheduling and resource allocation. For a given FN, the number of served IoT devices is limited with consideration on the complexity of SIC process. Then, for a limited number of served IoT devices, the FN can implement the proposed method efficiently to control the user accessing and resource allocation. Reference [46] points out that a good convergence performance could be achieved by setting proper parameters. Finally, we set a proper population size in the initialization, which affects the convergence speed: the greater population size is, the faster population converges.

For comparison, we analyze the complexity of the optimal exhaustive search algorithm. Its complexity could be considered from two parts, i.e., the task scheduling and the resource allocation. For the task scheduling part, there are M IoT devices and each of them has N tasks to process, then the FN searches all tasks of candidate IoT devices. Since the number of offloaded tasks sets is $2^N - 1$ and there are M IoT devices, the complexity can be expressed by $\mathcal{O}(N \cdot 2^M)$. For the resource allocation part, the optimal scheme searches all required IoT devices over each RB, the complexity can be expressed by $\mathcal{O}(K^M)$. $\Lambda = \lfloor \frac{P_{max}}{\varpi} \rfloor$ is the quantity of the scale of transmission power after discretization, and the value

Algorithm 1 *Solving the proposed optimization problem based on IGA*

```
Initialization:
a) The FN collects channel state information and the task requirements of all the IoT devices in its authority;
b) Initialize the maximum number of iterations $\Gamma_{max}$ and set iteration number $\tau = 0$;
c) Initialize the chromosomes[1:N+K] by selecting random variable $N + K$ $RA$ solutions and the fitness value $bF = 0$;
d) Initialize the crossover probability $p_c$ and the mutate probability $p_m$.
while $\tau \le \Gamma_{max}$ do
    $\tau = \tau + 1$;
    $fitness \leftarrow$ calculate fitness values of each chromosome in $chromosomes$;
    $selectionP \leftarrow$ calculate the selection probability of each chromosome according to $fitness$;
    $(child1, child2) \leftarrow$ randomly choose two chromosomes according to $selectionP$;
    if $rand() < p_c$ then
        $(s_0Child1, s_0Child2) \leftarrow$ the index of all the $s_0$ in $child1$ and $child2$
        $(newChild1, newChild2) \leftarrow$ remove all the $s_0$ in $child1$ and $child2$;
        $(child1, child2) \leftarrow$ insert the $s_0$ into $child1$ and $child2$ according to $(s_0Child1, s_0Child2)$;
    end if
    if $rand() < p_m$ then
        if $rand() < 0.5$ then
            $(child1, child2) \leftarrow$ randomly insert a $s_0$ into $child1$ and $child2$;
        else
            $(child1, child2) \leftarrow$ randomly remove a $s_0$ inside $child1$ and $child2$;
        end if
    end if
    $fitness \leftarrow$ calculate fitness values of each chromosome in $chromosomes$;
    $chromosomes \leftarrow$ replace two chromosomes with the smallest fitness values by $child1$ and $child2$;
    if $max(fitness) \ge bF$ then
        $RA \leftarrow$ the chromosome with the maximum fitness value;
    end if
end while
Output the resource allocation results $RA$.
```

TABLE I: The simulation parameters

Simulation parameters	Value
Carrier center frequency	2.5GHz
The bandwidth	10MHz
The radius of the FN	100m
Fading	Rayleigh flat fading
Path loss exponent	4
Power spectral density of noise	-174dBm$/Hz$
Data size for task F_{mn}	[0.1-1]Mbits
The number of required CPU cycles for task F_{mn}	[0.1-1]GHz
Computation capacity of the FN	20GHz
Computation capacity of the IoT device	[0.7 - 1] GHz
The local computing energy consumption	$[2-4]*10^{-11}$J/cycle
The FN computing energy consumption	$1*10^{-11}$J/cycle
The maximum transmission power	$300mW$
Crossover probability p_c	0.8
Mutate probability p_m	0.1

of Λ is related to the discretization interval ϖ. $\Upsilon = \lfloor \frac{1}{\sigma} \rfloor$ is the quantity of the scale of time slot after discretization, and the value of Λ is related to the discretization interval σ. Then the complexity can be expressed by $\mathcal{O}(K^{(M+\Lambda+\Upsilon)})$. Thus, the total complexity is $\mathcal{O}(N \cdot 2^M + K^{(M+\Lambda+\Upsilon)})$. From the discussion we could see that the proposed IGA-based algorithm has much lower complexity and is better suited to solving the formulated problem in practical deployments.

IV. Simulation Results and Discussions

In this section, we evaluate the performance of the proposed optimization scheme of heterogeneous resource allocation in fog computing-based wireless IoT networks with NOMA via Monte Carlo simulations. There is a FN with covered radius of 100 meters. Multiple IoT devices are deployed randomly in this area. The important simulations parameters are shown in Table I.

We choose three computation schemes for comparisons: the local computation scheme that all the tasks are processed locally at the IoT device, labeled as ”all at the local”; the fully computation offloading scheme that all the tasks are offloaded to the FN in wireless IoT networks, labeled as ”all offloading”. Then, to evaluate the performance of NOMA, the traditional OMA scheme is considered as the baseline where the multiple IoT devices cannot transmit signals to the same FN on the same RB, labeled as ”OMA scheme”. For fairness in the comparisons, every scheme has the similar system configurations as Table I.

A. Convergence Performance

In this subsection, we firstly investigate the convergence of the proposed IGA solution. Fig. 6 illustrates the energy consumption of the proposed IGA versus the iterations for different populations. For formula (11) which is a NP-hard problem, the exhaustive search (ES) algorithm could get the optimal solution but with high complexity. Meanwhile, the proposed IGA-based solution has sub-optimal solution with low complexity. Therefore, we select ES scheme as a benchmark for comparison. It can be seen that the proposed algorithm converges fast. With the increasing number of iterations, the proposed algorithm approaches the ES scheme

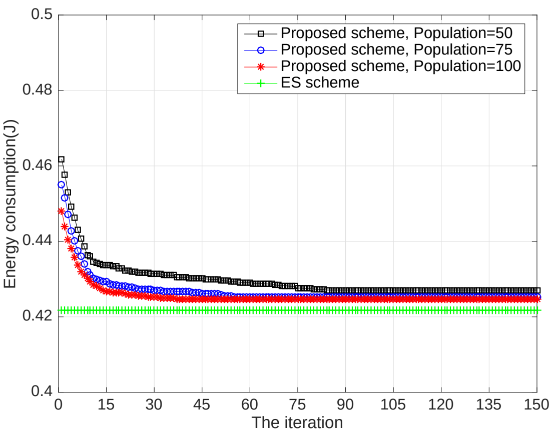

Fig. 6: Energy consumption versus the iterations.

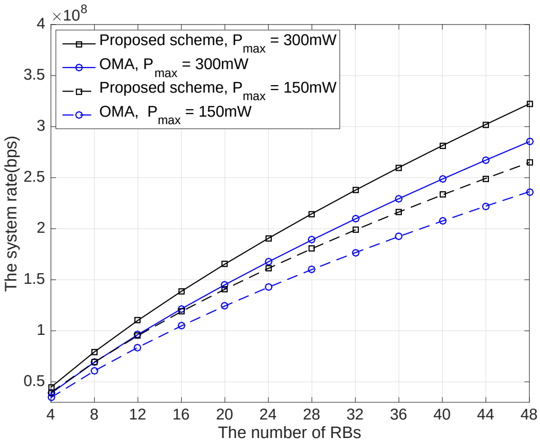

Fig. 8: System rate versus the number of RBs.

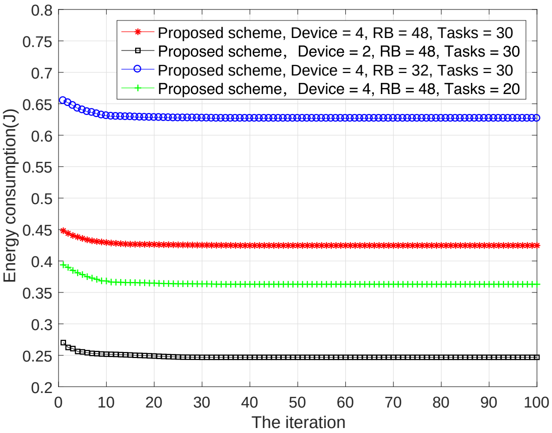

Fig. 7: Convergence performance with devices/tasks/RBs.

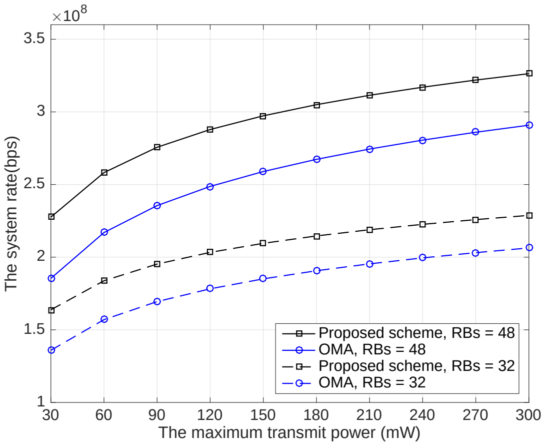

Fig. 9: System rate versus the maximum of transmit power.

quickly. The gap between the proposed algorithm and ES scheme gets more narrow. In addition, the result shows that the larger the population is, the fewer iterations are needed to converge. And we can find that when the population is larger than 100, only 40 iterations are required. Therefore, in the rest of the simulations, we set the generation as 40 and the population as 100 to ensure the convergence of the proposed scheme. On the other hand, we discuss the impact of number of devices, that of tasks and that of RBs on the convergence performance in Fig.7. The population is set to 100. We could see from the curves that, when RBs and tasks are the same, more devices indicates higher energy consumption for the convergence. Similarly, when devices and RBs are the same, the necessary energy for convergence is increased along with the increasing of the tasks. While the devices and tasks are the same, more RBs means more transmission resources for offloading, which results in better convergence performance.

B. The Performance of NOMA

In this subsection, we evaluate the effectiveness of the proposed scheme with NOMA. For comparison, we consider a typical OMA scheme as the benchmark. Figs. 8 and 9 compare the system throughput with different number of RBs and different maximum transmit power, respectively. In Fig. 8, it shows that in all the cases the system rate increases with the number of RBs. The proposed scheme has a higher throughput than the OMA scheme, and this superiority is more obvious when the number of RBs is further increased. In general, the increasing trends are due to the multi-connectivity gain, and the proposed scheme provides more significant diversity gain by supporting multiple IoT devices on the same RB. With the increase of maximum transmit power, the system throughput increases as well, as shown in Fig. 9. The system rate in all the cases increases slower when the transmit power is getting higher because of the increased interference among multiple IoT devices.

Figs. 10 and 11 compare the outage probability with different SNR and numbers of RBs, respectively. Outage probability is an important metric to characterize the systems. We consider the outage probability in uplink systems based on the sum rate [47]. In Fig. 10, it could be observed that the outage probability of all the cases decreases significantly with the increment of SNR. With the increasing number of RBs, it is obvious that the outage probability for all the schemes decrease gradually in Fig. 11. Increasing the number of RBs can increase the received data rate through scheduling

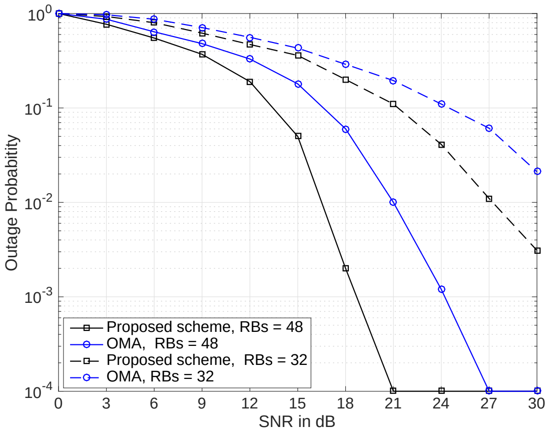

Fig. 10: Outage probability versus SNR.

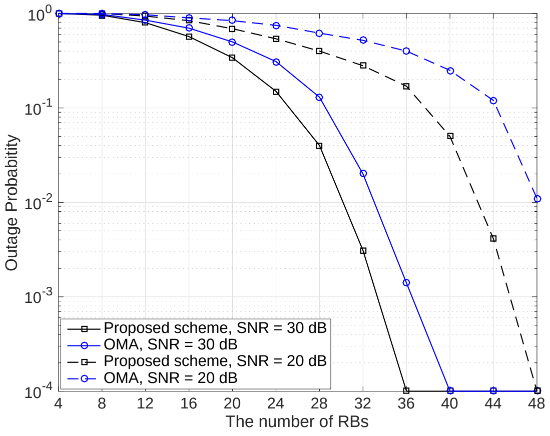

Fig. 11: Outage probability versus the number of RBs.

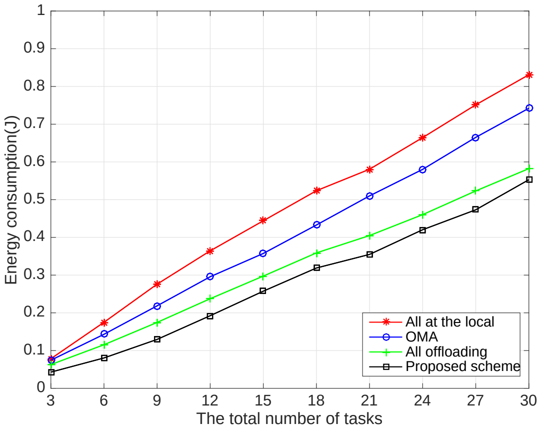

Fig. 12: Energy consumption versus the number of tasks.

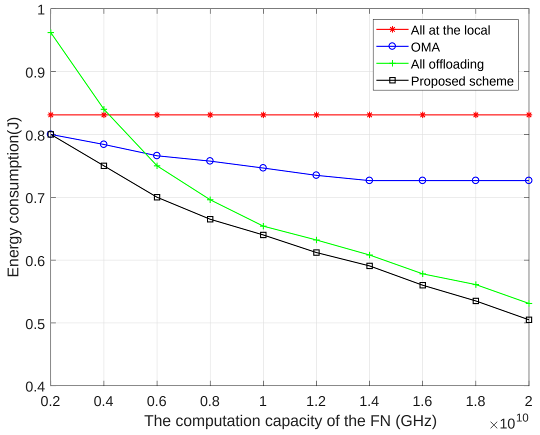

Fig. 13: Energy consumption versus the computation capacity of the FN.

the devices and allocating resources properly. Furthermore, NOMA scheme exploits the channel resources more efficiently than OMA, and hence is more beneficial at higher data rate. In addition, the outage probability based on the sum rate for uplink NOMA scheme has a higher diversity gain compared with OMA scheme. Then, for the outage probability based on the sum rate, NOMA can achieve better outage performance than OMA.

C. Energy consumption Performance

In this subsection, we focus on the energy consumption performance of the proposed scheme. Figs. 12, 13 and 14 compare the energy consumption with different number of tasks, different computation capacity of the FN, and different number of RBs, respectively. These figures demonstrate that the proposed scheme in fog computing-based wireless IoT networks with NOMA outperforms other compared schemes.

Fig. 12 illustrates the energy consumption versus the total number of tasks. We can observe that the energy consumption increases with the increment of tasks. The energy consumption of the scheme that executes all the tasks locally is the highest compared with the other schemes. The reason is that the FN has a much higher computation energy efficiency than that of the IoT devices. In addition, when the radio resources are sufficient for offloading, the energy consumption for transmission can be reduced greatly. Thus, other three compared schemes, such as all offloading, the proposed scheme and the similar scheme with OMA, can reduce energy consumption by utilizing the powerful computation and storage resource of the FN. Furthermore, compared with OMA scheme, the proposed scheme with NOMA achieves better performance. That is because the transmission data rate with NOMA is much higher than the data rate of the OMA scheme. Then, the transmission energy cost is reduced considerably.

Fig. 13 illustrates the energy consumption versus the computation capacity of the FN. With the increasing computation capacity of the FN, the energy consumptions of offloading schemes are decreased greatly. The local computation scheme keeps a constant value since the IoT devices process all the computation tasks locally. When the computation capacity of the FN is limited, the energy consumption of the all offloading computation scheme is higher. That is because the lower computation capacity results in more serious contention and more waiting time for processing. Then, the energy consumption is much higher than other schemes. Furthermore, considering

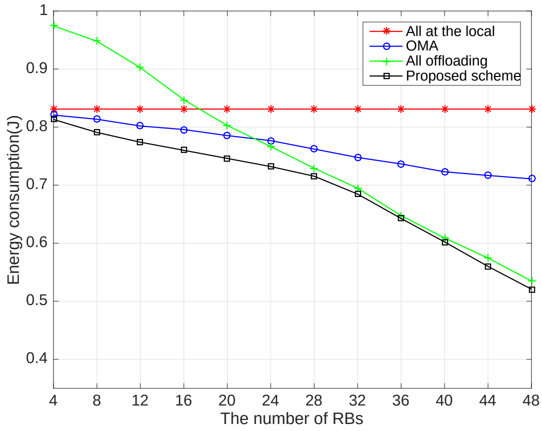

Fig. 14: Energy consumption versus the number of RBs.

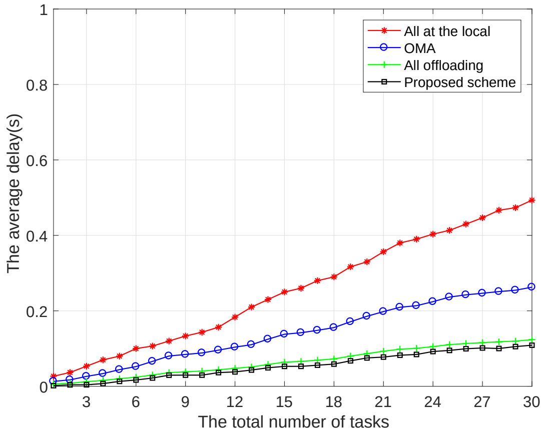

Fig. 15: Average delay versus the number of tasks.

the computation on the IoT devices and on the FN jointly, the proposed scheme has the lowest energy consumption by assigning the tasks to the IoT devices and the FN dynamically according to the available resources.

Fig. 14 shows the energy consumption versus the total number of RBs. Except for the local computation scheme, the energy consumption is gradually decreased with the increased number of RBs. When the number of RBs is small, the proposed scheme with NOMA and the offloading with OMA scheme prefer to compute the tasks on the IoT devices rather than offload to the FN, and then process the tasks under lower energy cost. Furthermore, as the number of RBs is increased gradually, the proposed scheme assigns the tasks to the FN by applying NOMA and achieves the lowest energy consumption.

D. Average Delay Performance

In this subsection, we focus on the average delay of the proposed scheme. The average delay for the computation tasks plays an important role in meeting the requirement of IoT applications as well as improving resource utilization. Fig. 15, 16 and 17 compare the average delay performance with the different number of tasks, the different computation capacity of the FN and the different number of RBs, respectively. These figures demonstrate that the proposed scheme has better performance than the compared schemes.

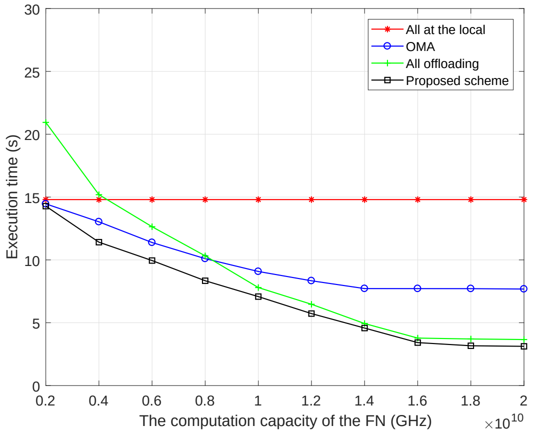

Fig. 16: Average delay versus the computation capacity of the FN.

Fig. 15 illustrates the average delay versus the total number of tasks. It can be observed that the average delay increases along with the number of tasks. It is noted that the scheme that executes all the tasks locally has the longest average delay. The reason is that the computation capability of the IoT devices is often almost one order of magnitude smaller than that of the FN. When a large number of tasks are involved, the IoT devices are not able to process these tasks in a short time. It results in serious delay and a poor performance. Furthermore, by utilizing the NOMA technique, the proposed scheme achieves a lower delay compared with that of the OMA scheme. The NOMA technique enables multiple IoT devices to transmit the data to the same FN on the same RB, therefore, it can obtain significant capacity gains of the access links compared with OMA. Thus, the proposed scheme has the lowest average delay.

Fig. 16 shows the average delay versus the computation capacity of the FN. Since all the tasks are computed at the IoT devices locally, the average delay of the local computation scheme is not affected by the computation capacity of the FN. It is noted that with the increasing computation capacity of the FN, the average delay of the OMA scheme is first decreased and then becomes almost constant. Since in the OMA scheme each RB can be used by only one IoT device, not all the IoT devices could access to the FN for offloading due to the limited spectrum resources. Obviously, the average delay of the fully offloading scheme is higher when the computation capacity of the FN is limited. The reason is that the transmission delay for offloading takes much longer without sufficient radio resources.

Fig. 17 illustrates the total average delay versus the total number of RBs. All the curves of the offloading schemes that include the proposed scheme, the offloading with OMA scheme and the fully offloading scheme, are decreasing significantly along with the increasing number of RBs. That is because the transmission time for offloading is reduced greatly when the available spectrum resources are increased. Fur-

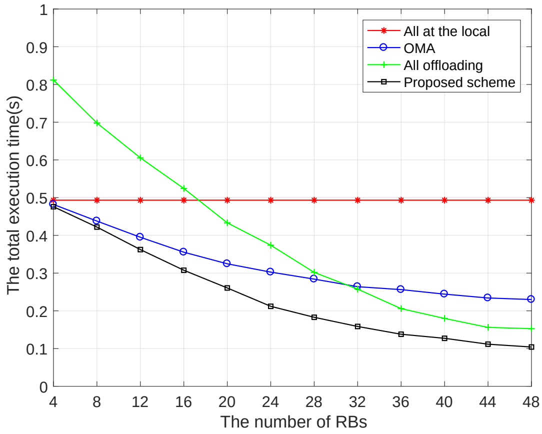

Fig. 17: Average delay versus the number of RBs.

thermore, considering the tasks scheduling jointly at the IoT devices and the FN, the proposed scheme prefers to compute the tasks locally when the available spectrum resource is inadequate. Then, the proposed scheme has a lower average delay compared to the fully offloading scheme. Finally, compared with OMA scheme, the proposed scheme could achieve better capacity gain by transmitting data to the FN for the multiple IoT devices and reducing the transmission time.

V. Conclusion and Future Work

In this paper, we have focused on the optimization of computation and communication resource allocation in fog computing-based wireless IoT networks with NOMA. We have considered a general scenario with massive IoT devices, and modeled the cost and energy consumption for both local computing and offloading computing tasks to FN. We have found that the system energy consumption and the average delay could be impacted by the different computing modes, and the proposed scheme that could make an optimal decision for choosing the proper computing mode could achieve a good performance. Moreover, in order to support multiple IoT devices in the uplink for offloading, we have analyzed the accessing and resource allocation problem in wireless IoT networks with NOMA uplinks. Constraints for interference, delay and other practical deployments have been discussed and applied. The formulated optimization problem is an MINLP problem, and then IGA is introduced to solve it with low-complexity. We have also observed from the simulation results that under similar offloading strategy, NOMA technique could bring better performance than OMA technique in system throughput and outage probability. The work in this paper is the first to give a joint modeling and optimization of fog computing-based wireless IoT networks with NOMA, and leads the way for further developments in the area. For future work, we will consider the efficient computation offloading strategy with cooperation among multiple FNs. Moreover, we will investigate the computation offloading problem with NOMA under imperfect interference cancellation, which is important for the implementation in the practical condition. The multiple optimization objects is another important issue for improving the overall network performance, which is also considered as another research direction of the future work.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No.61771070 and Grant No.61671088).

References

[1] I. U. Din, M. Guizani, S. Hassan, B.-S. Kim, M. K. Khan, M. Atiquzzaman, and S. H. Ahmed, "The internet of things: A review of enabled technologies and future challenges," *IEEE Access*, vol. 7, pp. 7606–7640, 2019.

[2] CISCO, "Fog computing and the internet of things: Extend the cloud to where the things are," *https://www.cisco.com/c/dam/en us/solutions/trends/iot/docs/computing-overview.pdf*, 2015.

[3] Z. Yin, F. R. Yu, S. Bu, and Z. Han, "Joint cloud and wireless networks operations in mobile cloud computing environments with telecom operator cloud," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 4020–4033, July 2015.

[4] X. Peng, J. Ren, L. She, D. Zhang, J. Li, and Y. Zhang, "BOAT: A block-streaming app execution scheme for lightweight IoT devices," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1816–1829, June 2018.

[5] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: issues and challenges," *Ieee Network*, vol. 30, no. 4, pp. 46–53, 2016.

[6] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Network*, vol. 31, no. 4, pp. 8–14, July 2017.

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

[8] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, February 2017.

[9] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

[10] A. Kiani and N. Ansari, "Edge computing aware noma for 5g networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299–1306, Apr. 2018.

[11] T. Qiu, Y. Zhang, D. Qiao, X. Zhang, M. L. Wymore, and A. K. Sangaiah, "A robust time synchronization scheme for industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. PP, no. 99, pp. 1–1, 2017.

[12] Y. Zhang, C. Jiang, J. Wang, Z. Han, J. Yuan, and J. Cao, "Green Wi-Fi implementation and management izhu andn dense autonomous environments for smart cities," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1552–1563, Dec. 2018.

[13] M. A. Rahman, M. M. Rashid, M. S. Hossain, E. Hassanain, M. F. Alhamid, and M. Guizani, "Blockchain and iot-based cognitive edge framework for sharing economy services in a smart city," *IEEE Access*, 2019.

[14] I. Lera, C. Guerrero, and C. Juiz, "Availability-aware service placement policy in fog computing based on graph partitions," *IEEE Internet of Things Journal*, 2018.

[15] C. Fiandrino, N. Allio, D. Kliazovich, P. Giaccone, and P. Bouvry, "Profiling performance of application partitioning for wearable devices in mobile cloud and fog computing," *IEEE Access*, 2019.

[16] S. F. Abedin, M. G. R. Alam, S. A. Kazmi, N. H. Tran, D. Niyato, and C. S. Hong, "Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 489–502, 2019.

[17] Q. Fan and N. Ansari, "Towards traffic load balancing in drone-assisted communications for IoT," *IEEE Internet of Things Journal*, 2018.

[18] G. Zhang, F. Shen, Z. Liu, Y. Yang, K. Wang, and M.-T. Zhou, "FEMTO: Fair and energy-minimized task offloading for fog-enabled IoT networks," *IEEE Internet of Things Journal*, 2018.

[19] J. Yao and N. Ansari, "QoS-aware fog resource provisioning and mobile device power control in IoT networks," *IEEE Transactions on Network and Service Management*, 2018.

[20] K. Liang, L. Zhao, X. Zhao, Y. Wang, and S. Ou, "Joint resource allocation and coordinated computation offloading for fog radio access networks," *China Communications*, vol. 13, no. Supplement2, pp. 131–139, N 2016.

[21] S. Joilo and G. Dn, "Decentralized algorithm for randomized task allocation in fog computing systems," *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 85–97, Feb 2019.

[22] J. Du, L. Zhao, X. Chu, F. R. Yu, J. Feng, and C. I, "Enabling low-latency applications in lte-a based mixed fog/cloud computing systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1757–1771, Feb 2019.

[23] Y. Liu, G. Pan, H. Zhang, and M. Song, "On the capacity comparison between MIMO-NOMA and MIMO-OMA," *IEEE Access*, vol. 4, pp. 2123–2129, 2016.

[24] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2852–2857, Mar 2017.

[25] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.

[26] H. Q. Tran, P. Q. Truong, C. V. Phan, and Q. T. Vien, "On the energy efficiency of NOMA for wireless backhaul in multi-tier heterogeneous cran," in *2017 International Conference on Recent Advances in Signal Processing, Telecommunications Computing (SigTelCom)*, Jan 2017, pp. 229–234.

[27] Q. T. Vien, T. A. Le, B. Barn, and C. V. Phan, "Optimising energy efficiency of non-orthogonal multiple access for wireless backhaul in heterogeneous cloud radio access network," *IET Communications*, vol. 10, no. 18, pp. 2516–2524, 2016.

[28] A. Kiani and N. Ansari, "Edge computing aware noma for 5g networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299–1306, Apr. 2018.

[29] X. Shao, C. Yang, D. Chen, N. Zhao, and F. R. Yu, "Dynamic IoT device clustering and energy management with hybrid NOMA systems," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4622–4630, 2018.

[30] Z. Wei and H. Jiang, "Optimal offloading in fog computing systems with non-orthogonal multiple access," *IEEE Access*, vol. 6, pp. 49 767–49 778, 2018.

[31] Z. Ma, J. Xie, H. Li, Q. Sun, Z. Si, J. Zhang, and J. Guo, "The role of data analysis in the development of intelligent energy networks," *IEEE Network*, vol. 31, no. 5, pp. 88–95, 2017.

[32] M. Kamel, W. Hamouda, and A. Youssef, "Performance analysis of multiple association in ultra-dense networks," *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3818–3831, Sep. 2017.

[33] J. N. Laneman and G. W. Wornell, "Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks," in *Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE*, vol. 1, Nov 2002, pp. 77–81 vol.1.

[34] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 654–667, Feb. 2016.

[35] Y. Gao, B. Xia, Y. Liu, Y. Yao, K. Xiao, and G. Lu, "Analysis of the dynamic ordered decoding for uplink NOMA systems with imperfect CSI," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2018.

[36] B. Ling, C. Dong, J. Dai, and J. Lin, "Multiple decision aided successive interference cancellation receiver for NOMA systems," *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 498–501, Aug. 2017.

[37] Y. Gao, B. Xia, K. Xiao, Z. Chen, X. Li, and S. Zhang, "Theoretical analysis of the dynamic decode ordering SIC receiver for uplink NOMA systems," *IEEE Communications Letters*, vol. 21, no. 10, pp. 2246–2249, Oct. 2017.

[38] Z. Yang, W. Xu, Y. Pan, C. Pan, and M. Chen, "Energy efficient resource allocation in machine-to-machine communications with multiple access and energy harvesting for IoT," *IEEE Internet of Things Journal*, vol. 5P, no. 1, pp. 229–245, Feb. 2018.

[39] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive multiple access based on superposition raptor codes for cellular m2m communications," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 307–319, Jan. 2017.

[40] A. Shokrollahi, "Raptor codes," *IEEE transactions on information theory*, vol. 52, no. 6, pp. 2551–2567, 2006.

[41] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," in *Usenix Conference on Hot Topics in Cloud Computing*, 2010, pp. 4–4.

[42] H. Ahmadi and Y. H. Chew, "Adaptive subcarrier-and-bit allocation in multiclass multiuser ofdm systems using genetic algorithm," in *2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept 2009, pp. 1883–1887.

[43] A. Auslender, "Penalty and barrier methods: a unified framework," *SIAM Journal on Optimization*, vol. 10, no. 1, pp. 211–230, 1999.

[44] K. Deb, "An efficient constraint handling method for genetic algorithms," *Computer methods in applied mechanics and engineering*, vol. 186, no. 2-4, pp. 311–338, 2000.

[45] C. You, K. Huang, and H. Chae, "Energy efficient mobile cloud computing powered by wireless energy transfer," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1757–1771, May 2016.

[46] H. S. Lang, S. C. Lin, and W. H. Fang, "Subcarrier pairing and power allocation with interference management in cognitive relay networks based on genetic algorithms," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7051–7063, Sept 2016.

[47] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink noma systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.