On the Connectivity Maximization in NOMA-Aided Industrial IoT with Multiple Services

Bo Yin, Jianhua Tang, *Member, IEEE*, Miaowen Wen, *Senior Member, IEEE*, and Weihua Li, *Senior Member, IEEE*

***Abstract*—Industrial Internet-of-Things (IIoT) improving by leaps and bounds has brought new possibilities for industrial manufacturing. Meanwhile, it does bring some serious challenges with the increasing number of devices. In which massive connectivity and multi-service are two major challenges for IIoT. In order to address the two main issues, in this paper, we jointly consider non-orthogonal multiple access (NOMA) and wireless network slicing scenario, where multi-service devices share the same communication resources. To connect devices as many as possible, we formulate the connectivity maximization problem with joint sub-carrier association and power allocation as a mixed-integer nonlinear programming (MINLP) problem, under the constraints of limited communication resources. To solve the problem effectively, we first split the MINLP problem into two subproblems by introducing a power allocation weight. Then, we analyze the theoretical approach for a special case and propose the Layered Access (LA) algorithm for general cases. Furthermore, a Bisection Search (Bisearch) algorithm is devised to find out the optimal power allocation weight. Simulation results show that the proposed LA algorithm has better performance compared to other benchmark schemes.**

***Index Terms*—Connectivity maximization, wireless network slicing, Industrial Internet of Things (IIoT), non-orthogonal multiple access (NOMA), resource allocation.**

I. Introduction

RECENTLY, Industrial Internet of Things (IIoT), based on the Internet of things (IoT) technology, has reshaped the visage of industrial production and accelerated the development of Industry 4.0, while being one of the key factors for improving production efficiency. However, while the development of IIoT creates more possibilities to industrial manufacturing, it also faces more challenges. With IIoT evolving rapidly, dramatically growing number of devices brings an unprecedented ubiquitously connectivity (up to 10^6/km^2 massive devices), which imposes a tremendous pressure on the existing IIoT, such as fieldbus [1]. For example, in [2], it is expected that more than 20 billion IoT devices will be directly connected to IoT communications networks. Besides, the future IIoT technologies are expected to incorporate diverse services, i.e., multi-service[1], to facilitate the coexistence of different devices such as sensors, unmanned monitoring devices and smart meters in IIoT [2]. For instance, industrial robots typically require low-latency and high reliability while environmental sensors such as humidity sensor usually generate small amounts of data and do not require high data rate.

In order to support the rapidly increasing number devices in IIoT, a wide variety of low-power long/short-range wireless connectivity technologies has developed for IIoT applications, such as bluetooth in low energy area networks, Zigbee in automation systems, optical wireless communication, WiFi, low power wide area networks (LPWAN), and widely-deployed cellular technologies[2] [3]–[7]. Nevertheless, these technologies have limits in terms of connection density and coverage when it comes to covering very broad regions as well as a large number of multi-service devices.

The advent fifth generation (5G) mobile communication is expected to support a wide variety of usage scenarios, including enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC) and massive machine type communications (mMTC) which coincides seamlessly with the IIoT [8]. As a result, 5G is perceived as a solution to connect the massive machine type communications devices (MTCDs) with diverse services, driving the deployment of IIoT. From the physical and medium access layer perspective, designing the efficient access protocols in 5G networks, such as 5G multi-service air interface, facilitates the improvement of the ability of network to access massive devices [9]. Consequently, the design and evolution of 5G technologies give the needed breadth and depth to address numerous use cases and bring unprecedented massive connectivity possibilities for IIoT [10].

To further support the massive number and multi-type services devices in IIoT, one needs to more properly schedule physical resources. In this paper, we combine the non-orthogonal multiple access (NOMA) which can accommodate

B. Yin, J. Tang and W. Li are with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 511442, China. M. Wen is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China. J. Tang and W. Li are also with the Pazhou Lab, Guangzhou 510335, China. E-mails: bborg.yin@gmail.com, jtang4@e.ntu.edu.sg, eemwwen@scut.edu.cn, whlee@scut.edu.cn. The corresponding author is Jianhua Tang.

The work of J. Tang was supported in part by the National Nature Science Foundation of China under Grant 62001168 and in part by the Foundation and Application Research Grant of Guangzhou under Grant 202102020515. The work of M. Wen was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021B1515120067. The work of W. Li was supported in part by the National Natural Science Foundation of China under Grant 52275111.

[1]Multi-service refers to the ability of a IIoT network to efficiently support and manage diverse data and services, including real-time control, monitoring, sensing, communication, simultaneously. In this paper, we focus on the means by which devices acquire information.

[2]Such as Long-term Evolution (LTE) and LTE-advanced (LTE-A) - 4G machine-to-machine (M2M) and its standards, e.g., Narrowband IoT (NB-IoT) and LTE machine type communications (MTC) (LTE-M), which are developed for MTC connectivity within LTE systems.

multiple devices on the same radio resource block, and network slicing, which enables multiple logical networks to be created over one physical network.

Compared to the conventional orthogonal multiple access (OMA), such as time division multiple access (TDMA) and orthogonal frequency-division multiple access (OFDMA), NOMA, with high spectral efficiency and massive connectivity, has been expected to be one of the most promising technologies in next generation networks. NOMA can be generally classified into two categories, that is, code-domain NOMA and power-domain NOMA. This paper mainly focus on the power-domain NOMA. By utilizing superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, power-domain NOMA enables the allocation of orthogonal radio resources in the time- or frequency-domain to multiple users or devices at the same time, which satisfies the requirements of mMTC services [11]–[14].

Network slicing, as one of the key technologies of 5G, can overlay multiple virtual/logical networks on top of a shared network domain, that is, a set of shared physical network and computing resources [15]. More specifically, network slicing can improve resource utilization by dynamically adjusting physical resource for different use cases, such as ensuring a specific application or service to obtain priority access in terms of capacity and delivery, or isolating traffic for specific users or device classes. In the case of IIoT networks with limited physical resources and diverse services, slicing networks enables IIoT to maximize the use of network resources and service flexibility and scalability [16], [17].

A. Related Works

With an explosive increase of wireless mobile devices, IIoT faces the challenges associated with the need of energy efficiency, real-time performance, connectivity, multi-service, and with the security and privacy issues [5]. The authors of [18] study the energy-efficient networking problem in cloud services for IIoT networks and propose an intelligent heuristic optimization, which jointly optimizes energy efficiency in data centers and the cloud network. Xia *et al.* [19] constructs a heterogeneous routing model where both source and graph routing coexist, and the proposed scheduling algorithm realize a trade-off between the real-time performance and reliability.

Numerous studies have considered the use of the power-domain NOMA into the 5G enabled IoT/IIoT to achieve massive connectivity [20]–[24]. [20] discusses multiple access approaches for mMTC devices in wireless cellular networks and concludes that NOMA is a potential scheme for supporting a greater number of devices than OMA. The maximum access problem (MAP) for uplink networks with NOMA is explored. In particular, the authors of [21] present a joint admission control, user clustering, channel assignment, and power management optimization MAP which is referred to as the maximal independent set problem in graph theory. In [22], a scheme of uplink power-domain NOMA is proposed to access mMTC devices to support high connection density. In [23], [24], different modes (single-tone and multi-tone) and more general sub-carrier allocation schemes for maximizing the connection density in NB-IoT have been studied, respectively.

Along with the aforementioned uplink power-domain NOMA schemes for connectivity maximization, a variety of research works on downlink power-domain NOMA have also been investigated. Lv *et al.* [25] introduces a millimeter-wave NOMA transmission technique for cellular M2M communication networks for IoT applications and analyzes its downlink performance in terms of device outage probability. In [26], the authors address the problem of connection density maximization for NB-IoT networks utilizing power-domain NOMA in the downlink and propose a stratified device allocation algorithm for optimum connection density under perfect and partial channel state information (CSI) knowledge for all the devices at base station (BS).

On the other hand, Popovski *et al.* [27] propose the concepts of orthogonal and non-orthogonal network slicing on the physical layer and explore the potential advantages of allowing for non-orthogonal sharing of RAN resources in uplink communications from a set of eMBB, mMTC, and URLLC devices to a common base station from the perspective of communication-theoretic. In particular, the paper proposes that heterogeneous nonorthogonal multiple access (H-NOMA) have good performance in terms of performance tradeoffs among the three generic 5G services. The authors of [28] consider the coexistence of eMBB and mMTC in the uplink of the same RAN with a multi-antenna BS scenario, where orthogonal and non-orthogonal network slicing schemes are employed. With the increase in the number of antennas, non-orthogonal slicing outperforms its orthogonal counterpart in terms of data rate and connectivity. Hossain *et al.* [29] propose a NOMA-enabled network slicing technique for a mobile edge computing (MEC) network. The suggested architecture minimizes service delay for MEC users and eliminates needless radio resource allocation via NOMA, which significantly improves the energy efficiency and spectral efficiency of the MEC network. Yin *et al.* [30], [31] use the non-orthogonal network slicing to support the different types of services in IIoT for maximization the connection density. The authors of [27] propose a framework for the orthogonal/non-orthogonal allocation of radio access network (RAN) physical resources to different service devices, such as eMBB, URLLC, and mMTC, from a communication perspective. The investigations in [32] and [33] incorporate two different types of services, i.e., URLLC and eMBB, into a single physical network by wireless network slicing.

B. Our Contributions

In contrast to the previously referenced papers, this work addresses certain important research gaps. First, in the light of the growing and dramatic demand for IIoT, large-scale access with multi-service coexistence capability is indeed crucial for IIoT, which has been previously overlooked in existing studies. Furthermore, even considering both of multi-service and connectivity maximization, the present work differs fundamentally from our previously investigated problems [30], [31]: 1) a high-quality communication scenario (without

interference between different slices), and 2) an enhanced large-scale access algorithm (without limitation of the number of devices to be accessed on a sub-carrier). Specifically, we combine the power-domain NOMA and network slicing where both uplink and downlink NOMA exist in the system simultaneously to maximize the connectivity. We mainly evaluate the connectivity[3] of the downlink while satisfying the quality of service (QoS) of different type devices. The main contributions of this paper are summarized as follows.

1) Based on power-domain NOMA and wireless slicing, we establish a communication system capable of supporting two different services. Within the same service, utilizing power-domain NOMA to enable the massive connectivity while adopting orthogonal wireless slicing for resource isolation among different services. We formulate the connectivity maximization problem as a mixed-integer nonlinear programming (MINLP) problem.
2) By introducing a power allocation weight, we first split the original problem into two subproblems. We analyze the theoretical approach for the case that each sub-carrier can be accessed by at most two devices, and for the more general cases, a low-complexity heuristic Layered Access (LA) algorithm is proposed to efficiently pair downlink devices with different type services to address the decomposed connectivity maximization subproblems.
3) A Bisection Search (Bisearch) algorithm is proposed to find the optimal power allocation weight, namely, power allocation of the BS for different types of devices. The simulation results show that the performance of the proposed algorithms outperforms the benchmark schemes and has lower computational complexity.

The remainder of this paper is organized as follows. Section II describes the basic communication system model and the main constraints of system are given. In Section III, we formulate the connectivity maximization problem as an MINLP problem. Theoretical analysis and the low-complexity heuristic algorithm, LA, are proposed to solve the decomposed MINLP in Section IV. The performance of the proposed algorithms is evaluated in Section V. Finally, Section VI concludes this paper.

Notations: We use calligraphy letters to represent the sets. Vectors and matrices are in the form of lowercase and uppercase bold letters, respectively. $\mathbf{I}_N$ means N-dimensional identity matrix. $\mathbb{E}(\cdot)$ and $\|\cdot\|_2$ stand for the expectation and the Euclidean norm, respectively. $(\cdot)^H$ denotes complex conjugate transpose. The distribution of a circularly symmetric complex Gaussian random variable with zero mean and variance σ^2 is denoted by $\mathcal{CN}(0,\sigma^2)$. The N-dimensional complex is denoted by $\mathbb{C}^N$. $|\mathcal{A}|$ represents the cardinality of set $\mathcal{A}$.

II. System Model And Main Constraints

In this paper, as shown in Fig. 1(a), we consider a power-domain NOMA-assisted multiuser uplink/downlink IIoT scenario, in which a BS is communicating with multiple MTCDs. Within a cell, the BS is equipped M-antennas and each device is equipped with single antenna. We denote the collection of all devices as $\mathcal{C}$. In which, there are two types of services that these devices can provide:

- Type A service: Uploading collected data to the BS and also acquiring data from the BS. For example, the automated guided vehicles (AGVs).
- Type B service: Downloading/receiving data from the BS only. For example, the actuators.

Assume that each device can only provide one type of service and hence we can classify all devices into two disjoint sets, i.e., type A set, $\mathcal{C}^A$, and type B set, $\mathcal{C}^B$. To enable these two distinct types of services in a single IIoT network, we employ two disparate network slices, namely, slice A and slice B, respectively. Thus, in this paper, we can simply regard different slices as different types of services.

Within one transmission time interval (TTI), all devices share the whole spectrum resource which is evenly divided into S_{tol} sub-carriers and each sub-carrier has bandwidth of W Hz. We denote the set of all sub-carriers as $\mathcal{S} \triangleq \{1,\ldots,S_{tol}\}$. Denote $\mathcal{S}^A$ as the set of sub-carriers assigned to slice A and $\mathcal{S}^B$ as that assigned to slice B. In this work, we assume $\mathcal{S}^A$ and $\mathcal{S}^B$ are pre-determined and disjoint, i.e., $\mathcal{S}^A \cup \mathcal{S}^B = \mathcal{S}$ and $\mathcal{S}^A \cap \mathcal{S}^B = \emptyset$. To fully utilize the physical resources, power-domain NOMA is involved for both uplink and downlink:

- NOMA for uplink: Providing non-orthogonal resources to connect type A devices as many as possible.
- NOMA for downlink: Providing non-orthogonal resources to connect different devices (either type A or type B devices) as many as possible.

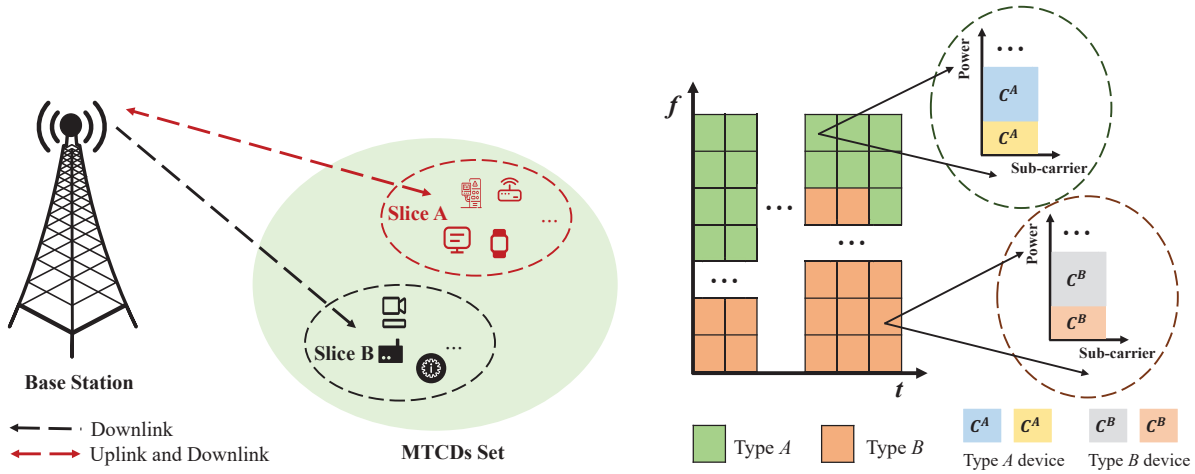
For simplicity, we assume that the channel state information (CSI) is perfectly estimated and known to the BS and devices in this paper[4]. The sub-carrier allocation for different types of services is shown in Fig. 1(b).

A. NOMA Uplink Scheme

This scheme considers an uplink single-input multiple-output (SIMO) scenario to access type A devices. In this situation, type A devices upload data to the BS through power-domain NOMA, namely, the information transmitted by the devices will be multiplexed on the same time-frequency resource. For brevity, define a binary variable $\alpha_{k,s}$ to identify the association between device and sub-carrier, i.e., $\alpha_{k,s}=1$ if the s-th sub-carrier is allocated to the k-th device, and $\alpha_{k,s}=0$ otherwise. The uplink channel coefficients between the k-th device and the BS over the s-th sub-carrier is denoted as $\boldsymbol{h}^u_{k,s}(\in \mathbb{C}^{M\times 1})$, whose small-scale fading is assumed to be frequency-flat. Hence, $\boldsymbol{h}^u_{k,1}=\boldsymbol{h}^u_{k,2}=\cdots \triangleq \boldsymbol{h}^u_k$, and we use $\boldsymbol{h}^u_k$ to replace $\boldsymbol{h}^u_{k,s}$ for brevity in the remaining parts. Assume that x_k is the transmitted symbols of the k-th device satisfying $\mathbb{E}\left(|x_k|^2\right)=1$ and $\mathbb{E}(x_i,x_j)=0 \ \ (\forall i\neq j)$. The

[3]In our paper, connectivity is defined as the capability of a network to connect and interact with multiple devices simultaneously. Specifically, connectivity is characterized by the number of devices that can successfully connect to the network within a transmission interval.

[4]Due to the theoretical upper bound of connectivity is considered as the goal of in this paper, therefore, here we only focus on perfect CSI. In addition, our proposed algorithms (for deterministic problem) can also be extended to the case of imperfect CSI by invoking the Markov inequality.



(a) The system framework.

(b) Resource allocation for different types of services.

Fig. 1. NOMA-aided multi-service IIoT scenario.

M-dimensional uplink received signal vector at the BS is represented as

$$\boldsymbol{y} = \sum_{s\in\mathcal{S}^A}\sum_{i\in\mathcal{C}^A} \alpha_{i,s}\sqrt{p^u_{i,s}}\boldsymbol{h}^u_i x_i + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{C}^{M\times 1}$; $p^u_{k,s}$ is the uplink transmit power for the k-th device over the s-th sub-carrier; $\boldsymbol{n} \sim \mathcal{CN}(0, N_0\mathbf{I}_M)$ represents the additive white Gaussian noise (AWGN) vector at the BS and N_0 is power spectral density of the AWGN. Assuming that the received signal-to-noise ratios (SNR) of the k-th type A device at the BS is in the order of $p^u_{k,s}\left\|\boldsymbol{h}^u_k\right\|_2^2 \geq p^u_{i,s}\left\|\boldsymbol{h}^u_i\right\|_2^2, \forall i \in \{1,\ldots,k-1\}$, according to the optimal minimum mean square error (MMSE)-SIC receiver structure at the BS, then the BS first decodes the message of the k-th device [12], [34]. As a result, for any given $s \in \mathcal{S}^A$, the received instantaneous signal-to-interference-plus-noise ratios (SINR) at the BS corresponding to the k-th device are given by

$$\gamma^u_{k,s} = \frac{\alpha_{k,s}p^u_{k,s}\left\|\boldsymbol{h}^u_k\right\|_2^2}{\sum_{i=1}^{k-1}\alpha_{i,s}p^u_{i,s}\left\|\boldsymbol{h}^u_i\right\|_2^2 + N_0W}, \quad \forall k \in \mathcal{C}^A. \tag{2}$$

Then, the achievable transmission rate for the k-th device can be written as

$$r^u_k = \sum_{s\in\mathcal{S}^A} W\log_2\left(1+\gamma^u_{k,s}\right), \quad \forall k \in \mathcal{C}^A. \tag{3}$$

B. NOMA Downlink Scheme

This scheme considers a downlink multiple-input single-output (MISO) scenario where BS transmits information to type A and B devices simultaneously via power-domain NOMA. Similarly, we define a binary variable $\beta_{k,s}$ to indicate the association, i.e., $\beta_{k,s} = 1$ if the s-th sub-carrier is allocated to the k-th device, and $\beta_{k,s} = 0$ otherwise. Assuming that z_k is the message transmitted by the BS with $\mathbb{E}\left(|z_k|^2\right) = 1$. Denote $\boldsymbol{w}_k \in \mathbb{C}^{M\times 1}$ as the beamforming vector to the k-th device. The received signal y_k of the k-th device can be expressed as

$$y_k = \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{C}} \beta_{k,s}\sqrt{p^d_{k,s}}\boldsymbol{h}^{dH}_k \boldsymbol{w}_i z_i + n_k, \quad k \in \mathcal{C}, \tag{4}$$

where $\boldsymbol{h}^d_k (\in \mathbb{C}^{M\times 1}) \sim \mathcal{CN}(0, \mathbf{I}_M)$ represents the frequency-flat downlink channel coefficients from the BS to k-th device (over every sub-carrier), and $p^d_{k,s}$ represents the downlink transmit power allocated by BS to the k-th device over the s-th sub-carrier; $n_k \sim \mathcal{CN}\left(0,\sigma^2\right)$ is the AWGN noise. For simplicity, the maximum ratio transmission (MRT) linear processing is adopted at BS, and thus the beamforming vectors are designed as [35]

$$\boldsymbol{w}_k = \frac{\boldsymbol{h}^d_k}{\|\boldsymbol{h}^d_k\|_2}, \quad \forall k \in \mathcal{C}. \tag{5}$$

Assuming that the received SNR of k-th type A or B device is in the order of $p^d_{k,s}\left\|\boldsymbol{h}^d_k\right\|_2^2 \geq p^d_{i,s}\left\|\boldsymbol{h}^d_k\right\|_2^2, \forall i \in \{1,\ldots,k-1\}$, the message of the k-th device is decoded first by treating the remaining $k-1$ devices as interference [12]. Then, for any given $s \in \mathcal{S}$, the received instantaneous SINR to the k-th device is calculated by

$$\gamma^d_{k,s} = \frac{\beta_{k,s}p^d_{k,s}\left\|\boldsymbol{h}^d_k\right\|_2^2}{\sum_{i=1}^{k-1}\beta_{i,s}p^d_{i,s}\left\|\boldsymbol{h}^d_k\right\|_2^2 + N_0W}, \quad \forall k \in \mathcal{C}. \tag{6}$$

Thus, the achievable transmission rate for the k-th device can be written as

$$r^d_k = \sum_{s\in\mathcal{S}} W\log_2\left(1+\gamma^d_{k,s}\right), \quad \forall k \in \mathcal{C}. \tag{7}$$

C. Main Constraints

1) Constraints for each sub-carrier: Each sub-carrier can be accessed by at most N devices, i.e.,

$$\begin{aligned} &0 \leq \sum_{k\in\mathcal{C}^A}\alpha_{k,s} \leq N, \quad \forall s \in \mathcal{S}^A, \text{ and,} \\ &0 \leq \sum_{k\in\mathcal{C}}\beta_{k,s} \leq N, \quad \forall s \in \mathcal{S}. \end{aligned} \tag{8}$$

To guarantee that each sub-carrier can be only assigned to either downlink or uplink, the association between $\alpha_{k,s}$ and $\beta_{k,s}$ can be expressed as

$$\Big(\sum_{k\in\mathcal{C}^A}\alpha_{k,s}\Big)\Big(\sum_{k\in\mathcal{C}^A}\beta_{k,s}\Big)=0,\quad s\in\mathcal{S}^A. \tag{9}$$

2) Constraints for each device: For either uplink or downlink, each device can only occupy at most one sub-carrier for transmission, i.e.,

$$\begin{aligned}&0\le\sum_{s\in\mathcal{S}^A}\alpha_{k,s}\le 1,\quad \forall k\in\mathcal{C}^A,\ \text{and,}\\ &0\le\sum_{s\in\mathcal{S}}\beta_{k,s}\le 1,\quad \forall k\in\mathcal{C}.\end{aligned} \tag{10}$$

3) Constraints among devices/sub-carriers: When a type A device is accessed, it takes up both uplink and downlink sub-carriers, i.e.,

$$\sum_{s\in\mathcal{S}^A}\alpha_{k,s}=\sum_{s\in\mathcal{S}^A}\beta_{k,s},\quad \forall k\in\mathcal{C}^A. \tag{11}$$

Thus, $Q\triangleq|\mathcal{S}^A|/2$ sub-carriers are allocated for the uplink of type A service, as well as for the downlink of it.

4) Constraints for power: The uplink transmit power of the k-th device over the s-th sub-carrier should not exceed the maximum transmit power of the device, P_D, i.e.,

$$0\le p^u_{k,s}\le\alpha_{k,s}P_D,\quad \forall k\in\mathcal{C}^A,\ s\in\mathcal{S}^A, \tag{12}$$

and the transmit power allocated by the BS to the k-th device over the s-th sub-carrier should not exceed the maximum power of the BS, P_B, i.e.,

$$0\le p^d_{k,s}\le\beta_{k,s}P_B,\quad \forall k\in\mathcal{C},\ s\in\mathcal{S}. \tag{13}$$

Furthermore, the total downlink transmit power should satisfy

$$0\le\sum_{k\in\mathcal{C}}\sum_{s\in\mathcal{S}}p^d_{k,s}\le P_B. \tag{14}$$

III. Problem Formulation And Decomposition

A. Formulation

Our goal is to maximize the connectivity while satisfying the QoS of different type devices. Hence, the optimization problem can be formulated as follows:

$$\begin{aligned}\mathscr{P}1:\quad &\max_{\substack{\mathcal{P}^u,\mathcal{P}^d,\\ \mathcal{A},\mathcal{B}}}\ \sum_{k\in\mathcal{C}^A}\sum_{s\in\mathcal{S}^A}\alpha_{k,s}+\sum_{k\in\mathcal{C}^B}\sum_{s\in\mathcal{S}^B}\beta_{k,s}\\ &\text{s.t.}\ \ C1: r^u_k\ge\big(\sum_{s\in\mathcal{S}^A}\alpha_{k,s}\big)\bar{\lambda}^u,\quad \forall k\in\mathcal{C}^A,\\ &\qquad C2: r^d_k\ge\big(\sum_{s\in\mathcal{S}}\beta_{k,s}\big)\bar{\lambda}^d,\quad \forall k\in\mathcal{C},\\ &\qquad C3:\alpha_{k,s}\in\{0,1\},\quad \forall k\in\mathcal{C}^A,\ s\in\mathcal{S}^A,\\ &\qquad C4:\beta_{k,s}\in\{0,1\},\quad \forall k\in\mathcal{C},\ s\in\mathcal{S},\\ &\qquad (8)\text{–}(14),\end{aligned}$$

where $\mathcal{P}^u$, $\mathcal{P}^d$, $\mathcal{A}$, and $\mathcal{B}$ are the set of variables $\{p^u_{k,s}\}$, $\{p^d_{k,s}\}$, $\{\alpha_{k,s}\}$, and $\{\beta_{k,s}\}$, respectively; $\bar{\lambda}^u$ and $\bar{\lambda}^d$ is the minimum required data rate of devices in the uplink and downlink, respectively.

Due to the non-convex constraints $C1$ and $C2$, problem $\mathscr{P}1$ is a mixed-integer nonlinear programming (MINLP) problem and is deeply coupled by the variables regarding downlink transmit power. Thus, we next will propose a simple method to decouple this.

B. Decomposition

In this subsection, we decompose problem $\mathscr{P}1$ into two independent subproblems by introducing a power allocation weight θ ($0\le\theta\le 1$), i.e., the power allocated to slice A and B by the BS is θP_B and $(1-\theta)P_B$, respectively. Thus constraints (13) can be rewritten as

$$0\le p^d_{k,s}\le\beta_{k,s}\theta P_B,\quad \forall k\in\mathcal{C}^A,\ s\in\mathcal{S}^A,\ \text{and,} \tag{15}$$

$$0\le p^d_{k,s}\le\beta_{k,s}(1-\theta)P_B,\quad \forall k\in\mathcal{C}^B,\ s\in\mathcal{S}^B, \tag{16}$$

respectively, and constraints (14) can be rewritten as

$$0\le\sum_{k\in\mathcal{C}^A}\sum_{s\in\mathcal{S}^A}p^d_{k,s}\le\theta P_B,\ \text{and,} \tag{17}$$

$$0\le\sum_{k\in\mathcal{C}^B}\sum_{s\in\mathcal{S}^B}p^d_{k,s}\le(1-\theta)P_B, \tag{18}$$

respectively. Now, problem $\mathscr{P}1$ can be decomposed into two independent subproblems as follows:

$$\begin{aligned}\mathscr{P}1\text{-}1:\quad &\max_{\substack{\mathcal{P}^u,\mathcal{P}^d,\\ \mathcal{A},\mathcal{B}}}\ \sum_{k\in\mathcal{C}^A}\sum_{s\in\mathcal{S}^A}\alpha_{k,s}\\ &\text{s.t.}\ \ (8)\text{–}(12),\ (15),\ (17),\ C1,\ C2,\ C3\ \text{and}\ C4,\end{aligned}$$

and

$$\begin{aligned}\mathscr{P}1\text{-}2:\quad &\max_{\mathcal{P}^d,\mathcal{B}}\ \sum_{k\in\mathcal{C}^B}\sum_{s\in\mathcal{S}^B}\beta_{k,s}\\ &\text{s.t.}\ \ (8),\ (10),\ (16),\ (18),\ C2\ \text{and}\ C4.\end{aligned}$$

In which, problem $\mathscr{P}1$-1 is to maximize the connectivity of type A devices, and problem $\mathscr{P}1$-2 is to maximize the connectivity of type B devices. Obviously, both problem $\mathscr{P}1$-1 and $\mathscr{P}1$-2 are still MINLP problems and difficult to obtain the optimal solutions. To balance the computational complexity and system performance, we will propose effective algorithms in the next Section.

IV. Proposed Algorithms

Obviously, the value of N is critical to the performance of the system. As the value of N increases, the number of accessed devices also increases, leading to a corresponding increase in the complexity of the system. Since in most studies, $N=2$ is chosen as a suitable value to balance the performance improvement and decoding complexity in NOMA [21]–[24], therefore, in this section, we first present the approach to find the theoretical bounds for the case of $N=2$ by classifying the set of devices. Since it is hard to obtain the theoretical solutions in the case of $N>2$, we propose an efficient algorithm, namely Layered Access (LA), to solve $\mathscr{P}1$-1 and $\mathscr{P}1$-2, and then devise the Bisection Search

(Bisearch) algorithm to find the suitable weight θ. For brevity, if not specified, all the subscript s in the remaining context are arbitrary, i.e., for any $s \in \mathcal{S}$, due to the frequency-flat channel.

A. A Special Case: $N = 2$

Due to the error propagation and the decoding complexity of SIC (multi-user interference cancellation with complexity $\mathcal{O}(K^3)$, where K is the number of users [12]), we typically consider the case of $N = 2$ in the practical implementation scenarios.

For this case, we first equally divide the type A or B devices into two decoding groups. Taking type A devices as an example, we first sort the devices in an ascending order by the value of $\bar{\lambda}^u/\|\boldsymbol{h}_k^u\|_2$ (or $\bar{\lambda}^d/\|\boldsymbol{h}_k^d\|_2$). Different from [36] and [24], the work in [36] uses channel gain as the only sorting criteria and the work in [24] takes QoS requirement as the only criteria, our proposed sorting criteria considers them both and enables improving the performance of NOMA. That is, pairing two users with significantly distinct channel gains could guarantee a better performance in NOMA, while maximizing the probability of devices being accessed, i.e, the device with smaller QoS requirement can accommodate more interference than the device with larger QoS requirement in terms of the same channel gains and transmit power. We define two decoding sets $\mathcal{C}_1^A$ and $\mathcal{C}_2^A$, i.e.,

$$\begin{aligned} \mathcal{C}_1^A &\triangleq \{k|\bar{\lambda}^u/\|\boldsymbol{h}_k^u\|_2 \leq \Delta,\ k \in \mathcal{C}^A\},\ \text{and,} \\ \mathcal{C}_2^A &\triangleq \{k|\bar{\lambda}^u/\|\boldsymbol{h}_k^u\|_2 > \Delta,\ k \in \mathcal{C}^A\}, \end{aligned} \tag{19}$$

where Δ is the middle value of the set $\{\bar{\lambda}^u/\|\boldsymbol{h}_k^u\|_2 \mid k \in \mathcal{C}^A\}$. For each sub-carrier, it can be ideally accessed by one device from $\mathcal{C}_1^A$ and another device from $\mathcal{C}_2^A$. To access devices from different sets, in the uplink, the signals in $\mathcal{C}_1^A$ are decoded first with interference, and then the signals in $\mathcal{C}_2^A$ are directly decoded without interference while the exactly opposite decoding order is adopted in the downlink. Such a decoding order can improve the performance of NOMA without compromising the fairness of the devices [21]–[24], [36].

Following the decoding order defined above, we next transform constraints $C1$ and $C2$ into linear constraints.

1) Transformation of constraint $C1$: For uplink NOMA, i.e. for $k \in \mathcal{C}^A$, $C1$ can be replaced by [30]

$$\alpha_{k,s}\left(I_{k,s} + N_0 W\right)\left(2^{\frac{\bar{\lambda}^u}{W}} - 1\right) \leq p_{k,s}^u \|\boldsymbol{h}_k^u\|_2^2, \tag{20}$$

where

$$I_{k,s} = \begin{cases} \sum_{i \in \mathcal{C}_2^A} p_{i,s}^u \|\boldsymbol{h}_i^u\|_2^2, & \text{for } k \in \mathcal{C}_1^A, \\ 0, & \text{for } k \in \mathcal{C}_2^A. \end{cases} \tag{21}$$

By introducing a new variable, $\tilde{\alpha}_{k,s} = \alpha_{k,s} I_{k,s}$ ($k \in \mathcal{C}_1^A$), that is, when $\alpha_{k,s} = 1$, $\tilde{\alpha}_{k,s} = I_{k,s}$, otherwise $\alpha_{k,s} = \tilde{\alpha}_{k,s} = 0$, for $k \in \mathcal{C}_1^A$, (20) can be rewritten as

$$\left(\tilde{\alpha}_{k,s} + \alpha_{k,s} N_0 W\right)\left(2^{\frac{\bar{\lambda}^u}{W}} - 1\right) \leq p_{k,s}^u \|\boldsymbol{h}_k^u\|_2^2. \tag{22}$$

According to the Big-M method [24] and (21), $\tilde{\alpha}_{k,s}$ can be equivalently expressed by the following inequality as

$$\begin{aligned} &0 \leq \tilde{\alpha}_{k,s} \leq \sum_{i \in \mathcal{C}_2^A} p_{i,s}^u \|\boldsymbol{h}_i^u\|_2^2, \\ &\tilde{\alpha}_{k,s} \leq \alpha_{k,s} \sum_{i \in \mathcal{C}_2^A} P_D \|\boldsymbol{h}_i^u\|_2^2,\ \text{and,} \\ &(\alpha_{k,s} - 1) \sum_{i \in \mathcal{C}_2^A} P_D \|\boldsymbol{h}_i^u\|_2^2 + \sum_{i \in \mathcal{C}_2^A} p_{i,s}^u \|\boldsymbol{h}_i^u\|_2^2 \leq \tilde{\alpha}_{k,s}. \end{aligned} \tag{23}$$

For $k \in \mathcal{C}_2^A$, since $I_{k,s} = 0$, (20) can be rewritten as

$$\alpha_{k,s} N_0 W \left(2^{\frac{\bar{\lambda}^u}{W}} - 1\right) \leq p_{k,s}^u \|\boldsymbol{h}_k^u\|_2^2. \tag{24}$$

Hence, the constraint $C1$ can be eventually equivalent to linear constraints (22)–(24).

2) Transformation of constraint $C2$: Define

$$\begin{aligned} \mathcal{C}_1 &\triangleq \{k|\bar{\lambda}^d/\|\boldsymbol{h}_k^d\|_2 \leq \hat{\Delta},\ k \in \mathcal{C}\},\ \text{and,} \\ \mathcal{C}_2 &\triangleq \{k|\bar{\lambda}^d/\|\boldsymbol{h}_k^d\|_2 > \hat{\Delta},\ k \in \mathcal{C}\}, \end{aligned} \tag{25}$$

where $\hat{\Delta}$ is the middle value of the set $\{\bar{\lambda}^d/\|\boldsymbol{h}_k^d\|_2 \mid k \in \mathcal{C}\}$. For downlink NOMA, i.e. for $k \in \mathcal{C}$, $C2$ can be similarly replaced by

$$\beta_{k,s}\left(\bar{I}_{k,s} + N_0 W\right)\left(2^{\frac{\bar{\lambda}^d}{W}} - 1\right) \leq p_{k,s}^d \|\boldsymbol{h}_k^d\|_2^2, \tag{26}$$

where

$$\bar{I}_{k,s} = \begin{cases} 0, & \text{for } k \in \mathcal{C}_1, \\ \sum_{i \in \mathcal{C}_1} p_{i,s}^d \|\boldsymbol{h}_k^d\|_2^2, & \text{for } k \in \mathcal{C}_2. \end{cases} \tag{27}$$

In the same way, we also introduce a new variable $\tilde{\beta}_{k,s} = \beta_{k,s}\bar{I}_{k,s}$ ($k \in \mathcal{C}_2$). Then, for $k \in \mathcal{C}_2$, (26) can be rewritten as

$$\left(\tilde{\beta}_{k,s} + \beta_{k,s} N_0 W\right)\left(2^{\frac{\bar{\lambda}^d}{W}} - 1\right) \leq p_{k,s}^d \|\boldsymbol{h}_k^d\|_2^2. \tag{28}$$

Based on the Big-M method and (27), $\tilde{\beta}_{k,s}$ can be equivalently expressed by the following inequality as

$$\begin{aligned} &0 \leq \tilde{\beta}_{k,s} \leq \sum_{i \in \mathcal{C}_1} p_{i,s}^d \|\boldsymbol{h}_k^d\|_2^2, \\ &\tilde{\beta}_{k,s} \leq \beta_{k,s} \sum_{i \in \mathcal{C}_1} P_B \|\boldsymbol{h}_k^d\|_2^2,\ \text{and,} \\ &(\beta_{k,s} - 1) \sum_{i \in \mathcal{C}_1} P_B \|\boldsymbol{h}_k^d\|_2^2 + \sum_{i \in \mathcal{C}_1} p_{i,s}^d \|\boldsymbol{h}_k^d\|_2^2 \leq \tilde{\beta}_{k,s}. \end{aligned} \tag{29}$$

For $k \in \mathcal{C}_1$, due to $\bar{I}_{k,s} = 0$, (26) can be simplified as

$$\beta_{k,s} N_0 W \left(2^{\frac{\bar{\lambda}^d}{W}} - 1\right) \leq p_{k,s}^d \|\boldsymbol{h}_k^d\|_2^2. \tag{30}$$

Therefore, the constraint $C2$ can be equivalently expressed by linear constraints (28)–(30).

3) Transformation of constraint (9): By introducing the binary variables, $y_{i,s} \in \{0, 1\}$ in which $i \in \{1, 2, 3\}$, $s \in \mathcal{S}^A$, (9) can be equivalently linearly characterized by the following

set of inequalities [37]

$$\begin{cases} \sum\limits_{k\in\mathcal{C}^A}\alpha_{k,s}+\sum\limits_{k\in\mathcal{C}^A}\beta_{k,s}\le 2, \\ \sum\limits_{k\in\mathcal{C}^A}\alpha_{k,s}-\sum\limits_{k\in\mathcal{C}^A}\beta_{k,s}\le -1+R(1-y_{1,s}), \\ \sum\limits_{k\in\mathcal{C}^A}\beta_{k,s}-\sum\limits_{k\in\mathcal{C}^A}\alpha_{k,s}\le -1+R(1-y_{2,s}), \\ \sum\limits_{k\in\mathcal{C}^A}\alpha_{k,s}+\sum\limits_{k\in\mathcal{C}^A}\beta_{k,s}\le R(1-y_{3,s}), \\ 1\le\sum\limits_{i}^{3}y_{i,s}, \end{cases} \tag{31}$$

where R is a large positive real number.

Now, problem $\mathscr{P}$1-1 and $\mathscr{P}$1-2 can be reformulated as follows:

$$\mathscr{P}\text{1-3}: \max_{\substack{\mathcal{P}^u,\mathcal{P}^d,\\ \mathcal{A},\mathcal{B},\tilde{\alpha}_{k,s}}} \sum_{k\in\mathcal{C}^A}\sum_{s\in\mathcal{S}^A}\alpha_{k,s}$$

$$\text{s.t. (8), (10)–(12), (15), (17), (22)–(24), (31), } C3 \text{ and } C4,$$

and

$$\mathscr{P}\text{1-4}: \max_{\mathcal{P}^d,\mathcal{B},\tilde{\beta}_{k,s}} \sum_{k\in\mathcal{C}^B}\sum_{s\in\mathcal{S}^B}\beta_{k,s}$$

$$\text{s.t. (8), (10), (16), (18), (28)–(30), and } C4,$$

respectively. Problem $\mathscr{P}$1-3 and $\mathscr{P}$1-4 are essentially mixed-integer linear programming (MILP) problems which can be efficiently solved by toolboxes. For example, one can use Gurobi solver to obtain the optimal solution with the Branch and Bound (B&B) approach.

B. General Cases: $N \ge 2$

Due to the theoretical intractability of the problem with the case of $N > 2$, we next propose a low-complexity heuristic algorithm, namely LA algorithm, to obtain the solution for problems $\mathscr{P}$1-1 and $\mathscr{P}$1-2.

1) Layered Access Algorithm (LA): Since problem $\mathscr{P}$1-1 and $\mathscr{P}$1-2 are independent of each other, and problem $\mathscr{P}$1-1 can be considered as a generalization of problem $\mathscr{P}$1-2, we here only present the LA algorithm for problem $\mathscr{P}$1-1.

From the perspective of QoS, when a device is accessed, the minimum transmit power $p^u_{k,s}$ and $p^d_{k,s}$ are calculated by

$$p^{u(min)}_{k,s}=\left(2^{\frac{\bar{\lambda}^u}{W}}-1\right)\frac{N_0W+T_{k,s}}{\|\boldsymbol{h}^u_k\|_2^2},\quad \forall k\in\mathcal{C}^A, \text{ and,} \tag{32}$$

$$p^{d(min)}_{k,s}=\left(2^{\frac{\bar{\lambda}^d}{W}}-1\right)\frac{N_0W+\bar{T}_{k,s}}{\left\|\boldsymbol{h}^d_k\right\|_2^2},\quad \forall k\in\mathcal{C}, \tag{33}$$

respectively, where

$$T_{k,s}=\sum_{i=1}^{k-1}p^{u(min)}_{i,s}\|\boldsymbol{h}^u_i\|_2^2,\ \bar{T}_{k,s}=\sum_{i=1}^{k-1}p^{d(min)}_{i,s}\left\|\boldsymbol{h}^d_k\right\|_2^2. \tag{34}$$

Algorithm 1 Layered Access Algorithm

Input: θ, P_B, P_D, W, N, Q, $\bar{\lambda}^u$, $\bar{\lambda}^d$, $\boldsymbol{h}^u_k$, $\boldsymbol{h}^d_k$, $\forall k\in\mathcal{C}^A$
Output: $\sum_{k\in\mathcal{C}^A}\sum_{s\in\mathcal{S}^A}\alpha_{k,s}$, p
Initialization: $l\leftarrow 1$, $p\leftarrow 0$, $\alpha_{k,s}\leftarrow 0$, $\boldsymbol{t}^u=\boldsymbol{t}^d\leftarrow\mathbf{0}$
while $l\le N$ **do**
  $s\leftarrow 1$
  while $\mathcal{C}^A\neq\emptyset$ **and** $s\le Q$ **and** $p<\theta P_B$ **do**
    for $k\in\mathcal{C}^A$ **do**
      let $T_{k,s}=\boldsymbol{t}^u(s)$, $\bar{T}_{k,s}=\boldsymbol{t}^d(s)\|\boldsymbol{h}^d_k\|_2^2$
      calculate $p^{u(min)}_{k,s}$ by (32)
      calculate $p^{d(min)}_{k,s}$ by (33)
      if $p^{u(min)}_{k,s}>P_D$ **or** $p^{d(min)}_{k,s}>P_B$ **then**
        $\mathcal{C}^A\leftarrow\mathcal{C}^A\setminus\{k\}$
      end if
    end for
    $k^*\leftarrow\arg\max_{k\in\mathcal{C}^A}\left(p^{u(min)}_{k,s}/p^{d(min)}_{k,s}\right)$
    update $\boldsymbol{t}^u(s)\leftarrow\boldsymbol{t}^u(s)+p^{u(min)}_{k^*,s}\|\boldsymbol{h}^u_{k^*}\|_2^2$
    update $\boldsymbol{t}^d(s)\leftarrow\boldsymbol{t}^d(s)+p^{d(min)}_{k^*,s}$
    $\mathcal{C}^A\leftarrow\mathcal{C}^A\setminus\{k^*\}$
    $\alpha_{k^*,s}\leftarrow 1$, $s\leftarrow s+1$, $p\leftarrow p+p^{d(min)}_{k^*,s}$
  end while
  if $\mathcal{C}^A=\emptyset$ **or** $p\ge\theta P_B$ **then**
    break
  end if
  $l\leftarrow l+1$
end while

The LA algorithm adopts two key principles which ensure that the algorithm is more effective and efficient to determine the most appropriate devices to access:

- For the uplink, a device with larger $p^{u(min)}_{k,s}$ has higher priority to be accessed, since such a device has rare power to accommodate interference[5].
- For the downlink, due to the transmit power limitation at the BS, a device with lower $p^{d(min)}_{k,s}$ has higher priority to be accessed.

The core idea of LA algorithm is to access devices *layer by layer*. Based on the two principles mentioned above, we can now devise the LA algorithm according to the following steps:

Step 1: **Initialize** vectors $\boldsymbol{t}^u$ and $\boldsymbol{t}^d$ ($\in\mathbb{R}^{1\times Q}$) to $\mathbf{0}$, which represents that there is no interference on each sub-carrier, i.e., no devices are connected to the sub-carrier at the beginning.

Step 2: **Calculate** $p^{u(min)}_{k,s}$ and $p^{d(min)}_{k,s}$ for each $s\in\mathcal{S},k\in\mathcal{C}^A$ by substituting $T_{k,s}=\boldsymbol{t}^u(s)$ and $\bar{T}_{k,s}=\boldsymbol{t}^d(s)\|\boldsymbol{h}^d_k\|_2^2$ into (32) and (33) respectively, where $\boldsymbol{t}^u(s)$ and $\boldsymbol{t}^d(s)$ represent the s-th entry of $\boldsymbol{t}^u$ and $\boldsymbol{t}^d$ respectively.

Step 3: **Select** the device with the largest $p^{u(min)}_{k,s}/p^{d(min)}_{k,s}$ as k^* to connect to the s-th sub-carrier.

[5]In this paper, it is assumed that the maximum uplink power, P_D, is consistent across all devices. Given this assumption, a device with higher value of $p^{u(min)}_{k,s}$ implies a lower available power, $P_D-p^{u(min)}_{k,s}$, to handle interference.

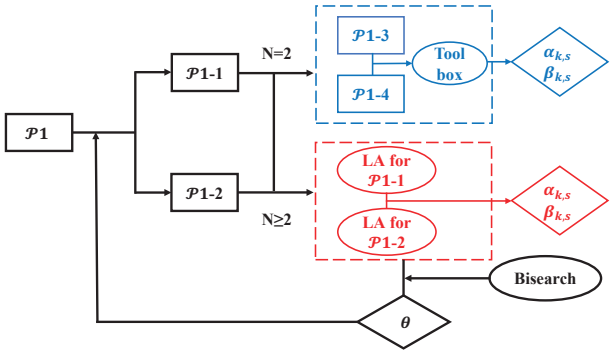


Fig. 2. The logic flow of solving problem $\mathscr{P}1$.

Step 4: **Update** $\boldsymbol{t}^u(s)$ and $\boldsymbol{t}^d(s)$ with $\boldsymbol{t}^u(s) = \boldsymbol{t}^u(s) + p_{k^*,s}^{u(min)} \|\boldsymbol{h}_{k^*}^u\|_2^2$ and $\boldsymbol{t}^d(s) = \boldsymbol{t}^d(s) + p_{k^*,s}^{d(min)}$.

Step 5: **Repeat** the above steps until all Q sub-carriers are occupied, which constitutes a *layer*.

Step 6: **Repeat** the procedure until the number of *layers* reaches N or the total downlink transmit power exceeds θP_B.

We present the LA algorithm in **Algorithm 1**. For the complexity of **Algorithm 1**, we mainly analyze the time complexity of two loops. First, in case of inner loop, each sub-carrier is required to be allocated to the specific device following the above two principles, which builds a *layer* consisting of $|\mathcal{C}^A| \cdot \min\{|\mathcal{C}^A|, Q\}$ operations. And for the outer loop, the number of *layers* is limited to N. As a result, the overall complexity of **Algorithm 1** is $\mathcal{O}(N \cdot |\mathcal{C}^A| \cdot \min\{|\mathcal{C}^A|, Q\})$.

2) Bisection Search Algorithm (Bisearch): In order to achieve an optimal power allocation at the BS between slice A and B, the Bisearch algorithm with low-complexity and high-efficiency is built to find out the value of θ. Bisearch algorithm, consistent with most bisection techniques, is an alternating computation algorithm. For example, initially, let $\theta = 0.5$, i.e. the BS allocates 50% of the power to slice A and B, respectively. If slice B uses only 30% of the total power while the power of slice A is used up, then we update $\theta = 0.6$, and so forth. Define P_a and P_b as the power allocated to slice A and B by the BS, respectively, and define p_a and p_b as the power consumed by slice A and B, respectively, i.e. p in **Algorithm 1**.

We present the Bisearch algorithm in **Algorithm 2**, where $\delta > 0$ is a threshold. The time complexity of **Algorithm 2** can be obtained by two parts analysis. First of all, for each loop, it calls **Algorithm 1** whose complexity is given by the previous $\mathcal{O}(N \cdot |\mathcal{C}^A| \cdot \min\{|\mathcal{C}^A|, Q\})$ once. In addition, the power interval is divided into half for each searching. Thus, the maximum number of iterations X satisfies $2^X = P_B/\delta$. Consequently, the Bisearch algorithm has a time complexity of $\mathcal{O}(\log_2 \frac{P_B}{\delta} \cdot N \cdot |\mathcal{C}^A| \cdot \min\{|\mathcal{C}^A|, Q\})$.

The specific solving flow framework of problem $\mathscr{P}1$ is shown in Fig. 2.

Algorithm 2 Bisection Search Algorithm

Input: δ, θ_0, P_B
Output: θ
Initialization: $\theta \leftarrow \theta_0$, $P_b \leftarrow (1-\theta_0)P_B$, $P_a \leftarrow \theta_0 P_B$
calculate p_b and p_a by **algorithm 1**
while $p_b > P_b$ **or** $p_a > P_a$ **do**
  if $p_b > P_b$ **and** $p_a < P_a$ **then**
    if $(P_a - p_a)/2 < \delta$ **then**
      break
    end if
    $P_b \leftarrow P_b + (P_a - p_a)/2$
    $\theta \leftarrow 1 - P_b/P_B$
    update p_b and p_a by **algorithm 1**
  else if $p_b < P_b$ **and** $p_a > P_a$ **then**
    if $(P_b - p_b)/2 < \delta$ **then**
      break
    end if
    $P_a \leftarrow P_a + (P_b - p_b)/2$
    $\theta \leftarrow P_a/P_B$
    update p_b and p_a by **algorithm 1**
  else
    break
  end if
end while

V. Simulation Results

In this section, we evaluate the performance of the proposed algorithms. We consider a single cell network where all devices are randomly distributed within the radius of 500 m. Unless otherwise specified, the power spectral density of the noise is -174 dBm/Hz and noise figure is 5 dB. P_D and P_B are set to 23 dBm and 30 dBm, respectively. The distance-dependent path loss, $PL(D)$ is generated by [38]

$$PL(D) = 120.9 + 37.6 \log\left(\frac{D}{1000}\right) + L + G, \qquad (35)$$

where D is the distance between devices and the BS in meter, and L is the indoor penetration loss that is assumed to be 20 dB. G is the antenna gain of -4 dB. We assume that, there are 16 sub-carriers, each with bandwidth 3.75 kHz for either uplink or downlink, to serve 48 type A devices and 48 type B devices, respectively.

To estimate the performance of our proposed algorithms, we compare our proposed approach with the following benchmark algorithms[6]:

- N2bound: The result from solving the case of $N = 2$ in problem $\mathscr{P}1$ by the approach in Section IV-A.
- Near-far Pairing (NFP): It pairs the closest (second closest, ...) device with the furthest (second furthest, ...) device, respectively.
- Advanced Near-far Pairing (A-NFP): This scheme extends the NFP to pair the closest, middle, and the furthest devices. If failed, discard the middle device.
- Orthogonal Multiple Access (OMA): Each sub-carrier can only be allocated to a single device, i.e., $N = 1$.

[6]Due to the high complexity of problem $\mathscr{P}1$, especially the combinational nature of sub-carrier association and permutable NOMA decoding order, it is hard to obtain the global optimal solution for the case of $N > 2$, even if by the numerical ways.

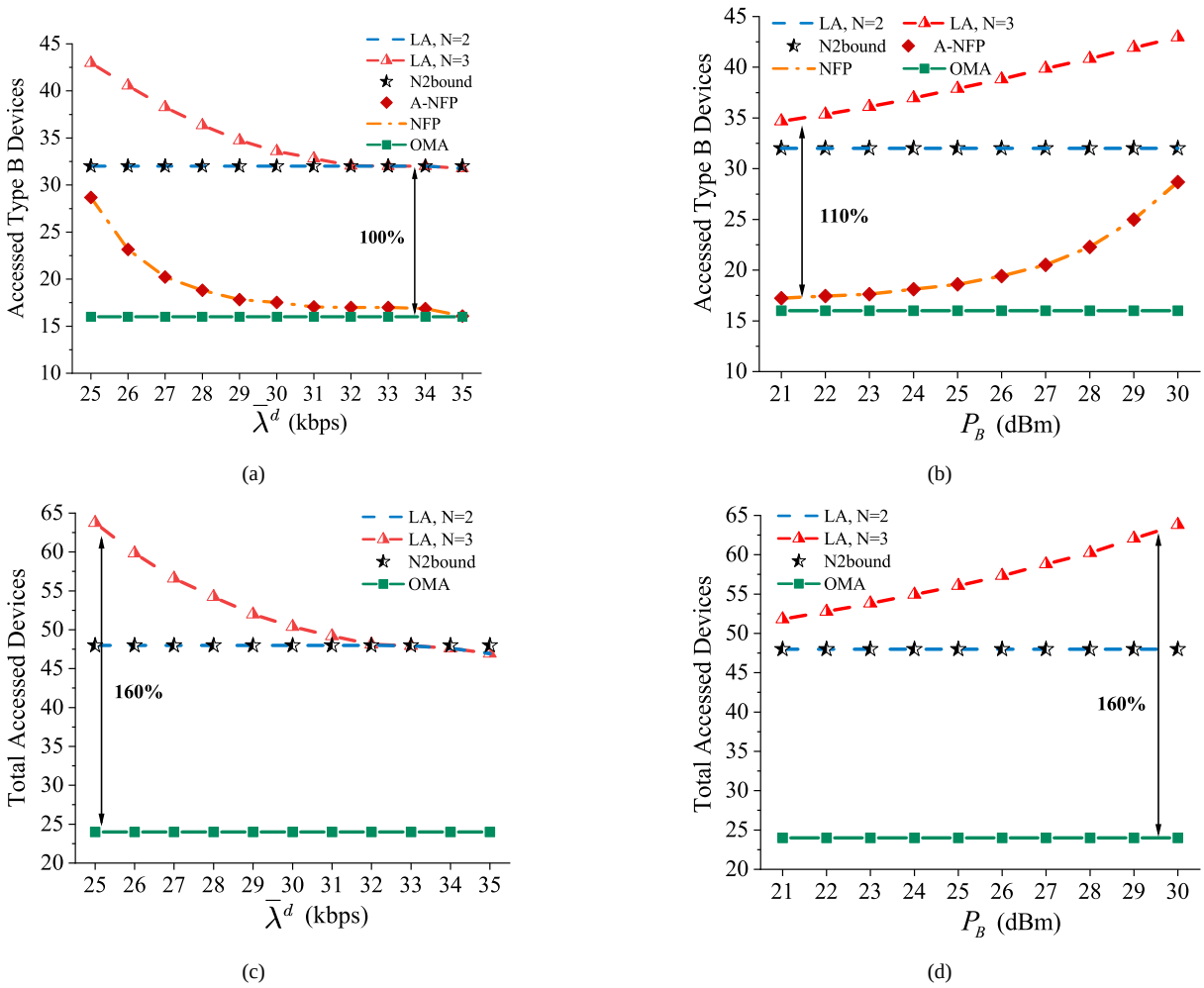


Fig. 3. (a) and (b) are connectivity of type B devices versus $\bar{\lambda}^d$ and P_B, respectively when $\bar{\lambda}^u$ is set to 25 kbps. (c) and (d) are connectivity of type A and B devices versus $\bar{\lambda}^d$ and P_B, respectively when $\bar{\lambda}^u$ and $\bar{\lambda}^d$ are set to 25 kbps. θ is set to 0.6.

Fig. 3 shows the performance of LA algorithm with respect to (w.r.t.) the transmit power and QoS requirements. In particular, Fig. 3(a) and 3(b) depict the connectivity (per TTI) of type B devices versus $\bar{\lambda}^d$ and P_B, respectively, and Fig. 3(c) and 3(d) describe that of both type A and B devices versus $\bar{\lambda}^d$ and P_B, respectively. We can conclude that as $\bar{\lambda}^d$ increases, the number of accessed devices decreases gradually while it is positively correlated with P_B. It can be observed that the performance of the proposed LA ($N = 2$ and $N = 3$) algorithm outperforms the NFP/A-NFP schemes. Specifically, the LA ($N = 3$) algorithm has maximum performance improvements of approximately 100% and 110% w.r.t. connectivity compared to the NFP and A-NFP scheme in Fig. 3(a) and Fig. 3(b), respectively. In addition, the NFP and A-NFP scheme have the same performance, which indicates that the A-NFP scheme can no longer achieve pairing of more than two devices in the case of the downlink transmit power is limited. In Fig. 3(c) and 3(d), since the NFP/A-NFP schemes cannot be applied for scenarios where both uplink and downlink exist simultaneously, their performances are hence not shown. With NOMA, LA ($N = 3$) algorithm accesses up to approximately 160% more devices in both Fig. 3(c) and 3(d), compared to OMA.

Fig. 4 illustrates the performance of LA algorithm w.r.t. $\bar{\lambda}^u$ and P_D on type A devices, and the number of accessed devices versus the total number of devices. Specifically, LA ($N = 3$) algorithm improves the number of accessed devices by up to 31% and 30% compared to LA ($N = 2$) algorithm in the Fig. 4(a) and 4(b), respectively. It can be seen from Fig. 4(c) and 4(d) that with the increase in the total devices, the number of accessed devices increases gradually until it tends to level off. Whereas, it is worth noting that in Fig. 4(c), the NFP/A-NFP schemes decrease rapidly after the total devices reaching at 40. This is because as the total number of devices increases, the NFP/A-NFP schemes do not take transmit power into account, giving higher priority to devices with more transmit power, which leads to a lower number of accessed devices. Furthermore, we can see from both Fig. 3 and Fig. 4 that the LA ($N = 2$) algorithm has a very close performance to the N2bound, which also validates the effectivity of the proposed LA algorithm.

Fig. 5 illustrates the connectivity versus θ, where $\bar{\lambda}^u$ and $\bar{\lambda}^d$ are set to 20 kbps and 30 kbps, respectively. We can observe that the connectivity has a maximum value at approximately

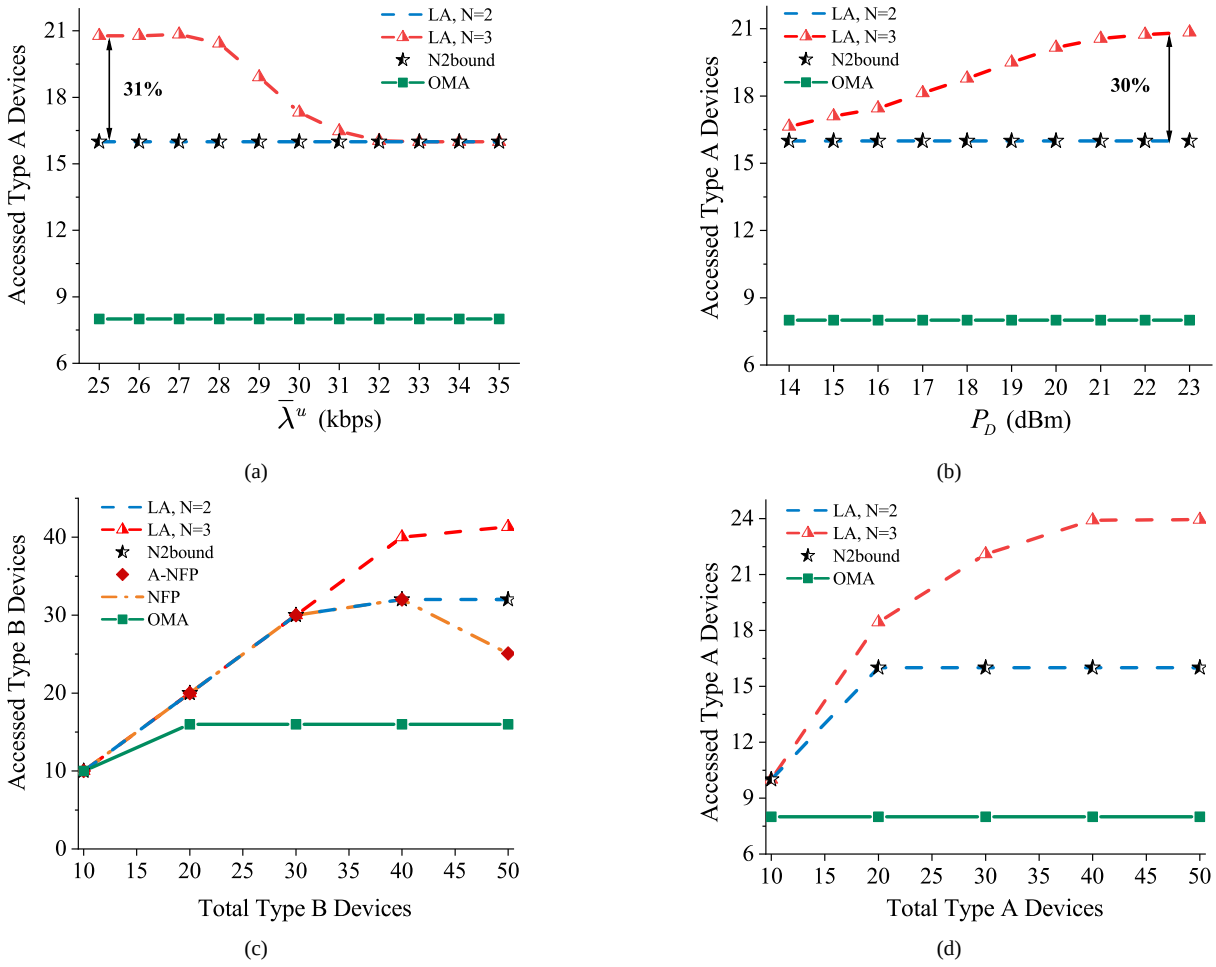


Fig. 4. (a) and (b) are connectivity of type A devices versus $\bar{\lambda}^u$ and P_D, respectively when $\bar{\lambda}^u$ and $\bar{\lambda}^d$ are set to 25 kbps. (c) and (d) are accessed devices of type B and A versus total type B and A devices, respectively when $\bar{\lambda}^u$ and $\bar{\lambda}^d$ are set to 25 kbps.

$\theta = 0.4$ by brute-force search, and the value $\theta = 0.42$ found by Bisearch algorithm is close to it. Fig. 6 shows a more generalized result about θ. In which the value of θ increases as the ratio of sub-carrier resources between type A and B service increases, which means that with the increase in sub-carrier resources, the transmit power should be also allocated more to maximize the connectivity.

VI. Conclusions

In this paper, we leveraged the wireless slicing to support two distinct types of services where power-domain NOMA was employed as the access scheme. We formulated the connectivity maximization problem as an MINLP problem. To address this problem, we first split the MINLP problem into two subproblems, and then proposed a low-complexity LA algorithm for joint sub-carrier association and transmit power allocation which maximizes the connectivity of IIoT with multiple services, and analyzed the theoretical approach for a special case. In addition, we devised a Bisearch algorithm to find the suitable downlink transmit power allocation weights. Finally, the simulation results showed that, compared to the traditional OMA and other benchmark schemes, our proposed algorithms achieved a much better system performance with respect to the connectivity.

In the future work, we will consider more diverse types of services such as high data rate, low-latency and high reliability, all of which can be facilitated by dynamically adjusting the time-frequency resource blocks. Given the complexity of the optimization problem, we believe that machine learning will be a promising approach to address this challenge.

References

[1] M. Shafi, A. F. Molisch, P. J. Smith *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[2] S. Mumtaz, A. Alsohaily, Z. Pang *et al.*, "Massive Internet of Things for industrial applications: Addressing wireless IIoT connectivity challenges and ecosystem fragmentation," *IEEE Ind. Electron. Mag.*, vol. 11, no. 1, pp. 28–33, 2017.

[3] M. R. Palattella, M. Dohler, A. Grieco *et al.*, "Internet of Things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.

[4] M. Aazam, S. Zeadally, and K. A. Harras, "Deploying fog computing in industrial Internet of Things and industry 4.0," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4674–4682, Oct. 2018.

[5] E. Sisinni, A. Saifullah, S. Han *et al.*, "Industrial Internet of Things: Challenges, opportunities, and directions," *IEEE Trans. Ind. Informat.*, vol. 14, no. 11, pp. 4724–4734, Nov. 2018.

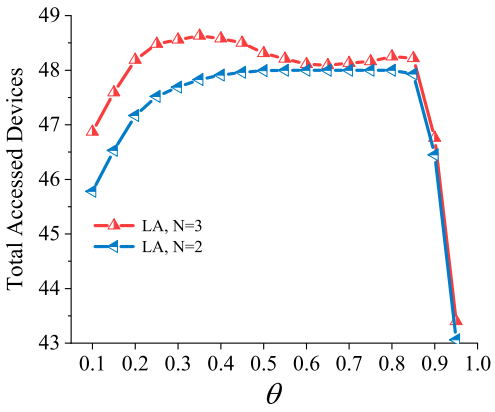


Fig. 5. Connectivity versus θ. $|\mathcal{S}^A|$=$|\mathcal{S}^B|$=16.

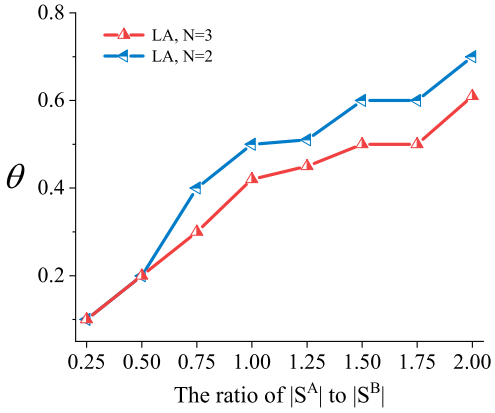


Fig. 6. Optimal θ versus the ratio $|\mathcal{S}^A|/|\mathcal{S}^B|$.

[6] J. Du, F. R. Yu, G. Lu *et al.*, "MEC-assisted immersive VR video streaming over terahertz wireless networks: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9517–9529, Oct. 2020.

[7] J. Ding, M. Nemati, C. Ranaweera *et al.*, "IoT connectivity technologies and applications: A survey," *IEEE Access*, vol. 8, pp. 67 646–67 673, Apr. 2020.

[8] T. Jiang, J. Zhang, P. Tang *et al.*, "3GPP standardized 5G channel model for IIoT scenarios: A survey," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8799–8815, Jun. 2021.

[9] C. Bockelmann, N. K. Pratas, G. Wunder *et al.*, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28 969–28 992, May 2018.

[10] A. Mahmood, L. Beltramelli, S. Fakhrul Abedin *et al.*, "Industrial IoT in 5G-and-beyond networks: Vision, architecture, and design trends," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4122–4137, Sept. 2022.

[11] L. Dai, B. Wang, Z. Ding *et al.*, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quarter. 2018.

[12] L. Dai, B. Wang, Y. Yuan *et al.*, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sept. 2015.

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

[14] R. Abozariba, M. K. Naeem, M. Patwary *et al.*, "NOMA-based resource allocation and mobility enhancement framework for IoT in next generation cellular networks," *IEEE Access*, vol. 7, pp. 29 158–29 172, Jan. 2019.

[15] I. Afolabi, T. Taleb, K. Samdanis *et al.*, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quarter. 2018.

[16] P. V. Kafle, Y. Fukushima, P. Martinez-Julia *et al.*, "Adaptive virtual network slices for diverse IoT services," *IEEE Commun. Stand. Mag.*, vol. 2, no. 4, pp. 33–41, Dec. 2018.

[17] S. Wijethilaka and M. Liyanage, "Survey on network slicing for Internet of Things realization in 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 957–994, 2nd Quarter. 2021.

[18] D. Jiang, Y. Wang, Z. Lv *et al.*, "An energy-efficient networking approach in cloud services for IIoT networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 928–941, May 2020.

[19] C. Xia, X. Jin, C. Xu *et al.*, "Real-time scheduling under heterogeneous routing for industrial Internet of Things," *Computers & electrical engineering*, vol. 86, p. 106740, Jun. 2020.

[20] M. Shirvanimoghaddam and S. Johnson, "Multiple access technologies for cellular M2M communications: An overview," *ZTE commun.*, vol. 14, no. 4, pp. 42–49, Oct. 2016.

[21] D. Zhai and R. Zhang, "Joint admission control and resource allocation for multi-carrier uplink NOMA networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 922–925, Dec. 2018.

[22] Z. Zhang, Y. Hou, Q. Wang *et al.*, "Joint sub-carrier and transmission power allocation for MTC under power-domain NOMA," in *Proc. IEEE ICC*, Kansas City, MO, USA, May 2018, pp. 1–6.

[23] S. Mishra, L. Salan, and C. S. Chen, "Maximizing connection density in NB-IoT networks with NOMA," in *Proc. IEEE VTC*, Antwerp, Belgium, May 2020, pp. 1–6.

[24] A. E. Mostafa, Y. Zhou, and V. W. S. Wong, "Connection density maximization of narrowband IoT systems with NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4708–4722, Oct. 2019.

[25] T. Lv, Y. Ma, J. Zeng *et al.*, "Millimeter-wave NOMA transmission in cellular M2M communications for Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1989–2000, Jun. 2018.

[26] S. Mishra, L. Salan, C. W. Sung *et al.*, "Downlink connection density maximization for NB-IoT networks using NOMA with perfect and partial CSI," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11 305–11 319, Jul. 2021.

[27] P. Popovski, K. F. Trillingsgaard, O. Simeone *et al.*, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, Sept. 2018.

[28] E. N. Tominaga, H. Alves, O. L. A. Lpez *et al.*, "Network slicing for eMBB and mMTC with NOMA and space diversity reception," in *Proc. IEEE VTC*, Helsinki, Finland, Apr. 2021, pp. 1–6.

[29] M. A. Hossain and N. Ansari, "Network slicing for NOMA-enabled edge computing," *IEEE Trans. Cloud Comput*, 2021.

[30] B. Yin, J. Tang, and M. Wen, "Maximizing the connectivity of wireless network slicing enabled industrial Internet-of-Things," in *Proc. IEEE GLOBECOM*, Madrid, Spain, Dec. 2021, pp. 1–6.

[31] B. Yin, J. Tang, and M. Wen, "Connectivity maximization in non-orthogonal network slicing enabled industrial Internet-of-Things with multiple services," *IEEE Trans. Wireless Commun.*, 2023.

[32] J. Tang, B. Shim, T. H. Chang *et al.*, "Incorporating URLLC and multicast eMBB in sliced cloud radio access network," in *Proc. IEEE ICC*, Shanghai, China, May 2019, pp. 1–7.

[33] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019.

[34] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[35] G. Zhu, C. Zhong, H. A. Suraweera *et al.*, "Outage probability of dual-hop multiple antenna AF systems with linear processing in the presence of co-channel interference," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2308–2321, Apr. 2014.

[36] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[37] F. S. Hillier and G. J. Lieberman, *Introduction to operations research*. Tata McGraw-Hill Education, 2012.

[38] 3GPP, "Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)," document 3GPP TR 45.820 V13.1.0, Nov. 2015.

Bo Yin received the B.Eng. degree in information engineering and M.Eng. degree in information and communication engineering from South China University of Technology, Guangzhou, China, in 2019 and 2022, respectively. He is currently pursuing a Ph.D. degree in electronics and ICT engineering with the Department of Information Technology at Ghent University, Belgium. His research interests include network planning, machine learning, and wireless communications.

Jianhua Tang (S'11-M'15) received the B.E. degree in communications engineering from Northeastern University, China, in 2010, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2015. He was a Post-Doctoral Research Fellow with the Singapore University of Technology and Design from 2015 to 2016, and a Research Assistant Professor with the Department of Electrical and Computer Engineering, Seoul National University, South Korea, from 2016 to 2018. He is currently an Associate Professor with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China. His research interests include edge computing/intelligence, network slicing and industrial Internet of Things.

He was honored with the 2020 IEEE Communications Society Stephen O. Rice Prize. He is currently serving as an Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS and the DIGITAL COMMUNICATIONS AND NETWORKS.

Miaowen Wen (SM'18) received the Ph.D. degree from Peking University, Beijing, China, in 2014. From 2019 to 2021, he was with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, as a Post-Doctoral Research Fellow. He is currently an Associate Professor with South China University of Technology, Guangzhou, China. He has published two books and more than 150 journal papers. His research interests include a variety of topics in the areas of wireless and molecular communications.

Dr. Wen was a recipient of the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2020. He served as a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. Currently, he is serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL, AND MULTI-SCALE COMMUNICATIONS, and the IEEE COMMUNICATIONS LETTERS.

Weihua Li (M'12-SM'19) received his Ph.D. degree in mechanical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2003. He is currently the Dean and a Professor with the School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou, China. His research interests include Industrial intelligence, Industrial Big Data, Digital Twins, Intelligent Maintenance & Health Management, and Intelligent Connected Vehicles.

Prof. Li is now serving as the co-chair of Technical Committee (TC-3) on Condition Monitoring & Fault Diagnosis Instrument, IEEE Instrumentation and Measurement Society (IEEE IM Society). He serves as a member of the Editorial Board of IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, IEEE SENSORS JOURNAL, CHINESE JOURNAL OF MECHANICAL ENGINEERING (CJME), JOURNAL OF DYNAMICS, MONITORING AND DIAGNOSTICS (JDMD), and JOURNAL OF VIBRATION ENGINEERING (JVE, IN CHINESE).