Double QoS Guarantee for NOMA-Enabled Massive MTC Networks

Ting Qi, *Member, IEEE,* Wei Feng, *Senior Member, IEEE,*
Yunfei Chen, *Senior Member, IEEE,* and Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*—Massive connections and diverse quality of service (QoS) requirements pose a major challenge for machine-type communication (MTC) networks. In this paper, to satisfy various QoS requirements of a massive number of MTC devices (MTCDs), the devices are divided into multiple clusters based on the QoS characteristics. The cluster access control and intra-cluster resource allocation problems are studied to satisfy the double delay requirements in the access and data transmission phases in a cross-layer approach. Specifically, we formulate an access control problem to maximize the access efficiency with constraints on access and transmission delays. An efficient algorithm is proposed to adaptively adjust the access time intervals and backoff factors of the clusters for different numbers of active MTCDs and transmission rates. Given the access parameters, non-orthogonal multiple access is adopted in resource allocation to maximize the system utility function while guaranteeing the delay requirements for each accessed MTCD. An efficient sequential convex programming iterative algorithm is proposed to solve the NP-hard nonconvex problem with two typical utility objectives: total throughput and consumed power. Simulation results show that the proposed scheme can achieve better performance in terms of access efficiency, delay, throughput, and consumed power than other schemes. The impacts of various parameters, including delay and traffic rate, on the performance are disclosed.**

***Index Terms*—Access management, clustering, delay guarantee, machine-type communication (MTC), resource allocation.**

I. Introduction

With the development of wireless communication and the internet of things, machine-type communication (MTC) is becoming an important part of future communication systems [1]. MTC connects all types of machines and equipment in our daily life and industrial processes in a communication network. Application scenarios represented by smart cities, smart transportation, and smart homes supported by MTC networks can promote the digital transformation and upgrading of traditional industries, nurture new information products and services, and generate huge economic benefits [2].

Massive connections are the basic characteristic of MTC. The connection density in the 6G era is expected to reach 100 million per square kilometer. At that time, the access capability of the cellular network will face significant challenges [3], [4]. Meanwhile, since the MTC network has a wide range of application areas, these MTC devices (MTCDs) may have a large difference in quality of service (QoS), including delay and rate. For example, MTC applications such as industrial production automation control and tactile networks have stringent latency requirements, and their end-to-end latency requirement can reach the millisecond level [5]. Therefore, existing cellular systems designed mainly for human-to-human communications are not suitable for MTC. To meet these challenges, it is urgent to explore new access methods to accommodate the massive number of MTCDs with diverse QoS requirements.

A. Related Works

To meet the double requirements of massive connections and QoS, radio resource management, including access control, radio resource allocation and power management, are effective tools for MTC [6]. To alleviate the access congestion problem caused by a massive number of MTCDs and improve access efficiency, the throughput of MTC based on a double-queue model was maximized by tuning the backoff parameters in [7]. To save the radio resource, a simultaneous preamble and data transmission approach was investigated in [8]. A recursive access class barring (ACB) technique was proposed to effectively accommodate bursty and non-periodic MTC traffic in [9].

Inspired by the idea of non-orthogonal multiple access (NOMA), the non-orthogonal random access (NORA) scheme taking advantage of the difference in arrival time was used to reduce the preamble collision probability [10]. The authors in [11] proposed a novel random access procedure based on sparse code multiple access to further improve the access performance. A throughput optimization problem for NORA was investigated subject to constraints on the ACB factor, the number of MTCDs, and successful transmission probability in [12]. In [13], a frame structure adaptive to the number of users with successive decoding was proposed. The average delay performance was analyzed and optimized.

This work is partly supported by Natural Science Foundation of Jiangsu Province of China under Grant No. BK20200759; Jiangsu Postdoctoral Science Foundation(No.SBH20003); the open research fund of Key Lab of Broadband Wireless Communication and Sensor Network Technology (Nanjing University of Posts and Telecommunications), Ministry of Education under Grant No.JZNY201903; NUPTSF under Grant No.NY219013; Natural Science Research of Higher Education Institutions of Jiangsu Province under Grant No. 19KJB510009; and the National Natural Science Foundation of China under Grant 61901229.

T. Qi is with the Key Laboratory of Ministry of Education in Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: qit18@njupt.edu.cn). W. Feng is with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (email: fengwei@tsinghua.edu.cn). Y. Chen is with the School of Engineering, University of Warwick, Coventry, U. K (e-mail: Yunfei.Chen@warwick.ac.uk). A. Nallanathan is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, 4617 London, London, U. K (e-mail: a.nallanathan@qmul.ac.uk).

To meet the diverse QoS requirements of a massive number of MTCDs, a group-based or cluster-based access method is widely adopted for nonhomogenous MTC networks [14]–[17]. A group-based massive access management was presented in [14], where MTCDs with identical QoS characteristics are grouped into the same cluster. It was shown that group-based access where the data packets from the same group are aggregated and forwarded by the MTC gateways can improve the network performance [15], [16]. The authors in [17] proposed an adaptive access management method for adjusting the allowable access time of each cluster according to the delay requirements.

For the access layer, the random access and clustering problem were considered in [18]–[22]. Dividing MTCDs into delay-sensitive and delay-tolerant types, the throughput of delay-tolerant MTCDs was maximized with the delay constraints for delay-sensitive MTCDs in [18]. The authors in [19] investigated the random access opportunity (RAO) allocation problem to maximize the access efficiency with a random access delay constraint and proposed a dynamic access control mechanism. A joint ACB, cell selection and power allocation algorithm was proposed in [21]. When concentrating on energy consumption, a clustering and resource allocation algorithm with maximizing energy efficiency was presented in [20]. Multiuser sequencing and scheduling schemes can significantly reduce energy consumption, especially when the delay bound is stringent [22].

In addition to the QoS in the access phase, some research works focused on guaranteeing the QoS in the transmitting phase [23]–[26]. Uplink resource scheduling and power allocation were optimized to guarantee the transmission QoS for each individual MTCD in the LTE networks in [23]. When using NOMA in the data transmission phase, resource allocation can be more flexible, and the utilization of resources can be promoted [24], [27]. In [25], the performances in terms of coverage and throughput were studied in NOMA-based human-type communication and MTC hybrid networks. For short-packet communication in MTC, energy efficiency can be improved with NOMA [26].

B. Motivation and Contributions

Despite the many works mentioned above on random access and data transmission in MTC networks, it can be seen that QoS is mainly guaranteed in the access or transmission phase separately. Motivated by this, we propose an efficient access management and resource allocation method that guarantees the access and transmission QoS jointly. Specifically, in the access phase, a massive number of MTCDs with diverse QoS requirements are grouped into different clusters and the access time intervals and back-off factors of these clusters are jointly adjusted to maximize the random access efficiency while satisfying the access and transmission delay requirements. In the transmission phase, an efficient resource allocation method is proposed to maximize the system utility function with constraints on delay and number of MTCDs sharing one resource block with NOMA. Simulation results demonstrate the superiority of the proposed scheme in terms of access efficiency, delay, throughput and consumed power. The main contributions of the paper are as follows.

1) Different from the existing schemes that focus mainly on access or transmission separately, we consider access and transmission delay requirements simultaneously in the random access phase. We propose a random access management scheme for clustering-based MTC networks that dynamically adjusts the access time intervals and back-off factors to optimize the access efficiency while guaranteeing the double delay requirements.
2) To further provide the transmission delay guarantee for each MTCD, we formulate a NOMA-based resource allocation problem to optimize the defined utility function with constraints on the delay and number of MTCDs sharing one resource block. An efficient sequential convex programming iterative algorithm is proposed for utility with respect to throughput and consumed power.
3) The formulated access management problem is decomposed, and the optimal solution and feasible condition are derived. For the intractable NP-hard resource allocation problem, the proposed algorithm is efficient for the two common objectives. Simulation results demonstrate the superiority of the proposed scheme and disclose the significant impact of system parameters, such as delay, traffic rate, transmission rate and number of active MTCDs, on the access and transmission performance in massive MTC networks.

The rest of the paper is organized as follows. In Section II, the system model of the massive MTC networks is presented. The time-controlled massive access management framework is introduced. Then, the access management problem for the clusters is formulated, analyzed and solved in Section III. In Section IV, the intra-cluster resource allocation problem is investigated. The proposed access scheme and resource allocation method are combined to provide a practical joint access management scheme in Section V. Simulation results are presented in Section VI. Finally, Section VII concludes the paper.

II. System Model

Consider a single-cell MTC system with a massive number of devices of various QoS requirements. As suggested by 3GPP, these massive numbers of devices can be grouped into clusters for efficient access management. In this paper, we group devices into M clusters based on their QoS characteristics and requirements. MTCDs in the same cluster have identical QoS characteristics. This paper considers latency as the measure of QoS since it is the main QoS concern in many MTC applications.

As shown in Fig. 1, radio resources are divided into time-frequency domain resource blocks (RBs) shared by MTCDs. Orthogonal frequency division multiplexing is adopted, and several subcarriers constitute a subchannel. An RB contains a subchannel of bandwidth B along with a time slot (TS). Due to the large number of MTCDs, it may cause signaling congestion even with small packets of data. To reduce access collision, devices in clusters are allowed to access in a time-controlled manner. The access of devices in each cluster is

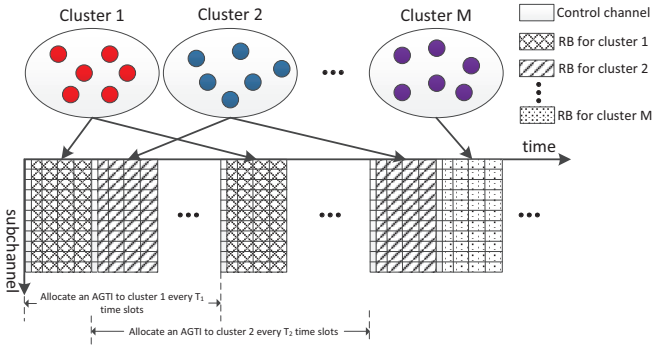

Fig. 1. Diagram of time-controlled massive access management.

permitted within its access grant time interval (AGTI), which includes L RAOs for random access. The RAO consists of a preamble and a physical random access channel (PRACH), which is a type of control channel. Thus, the number of RAOs equals the product of the number of preambles and PRACHs available for random access. In addition, each AGTI contains N RBs for data transmission with a fixed transmission time denoted by T_0 time slots. Since different clusters may have different QoS requirements, the AGTI may be allocated with different time intervals among clusters. Let T_m denote the TS interval cluster m to be allocated an AGTI. Note that $T_m > T_0$, and T_m is an integer multiple of T_0.

Let C_m be the number of active MTCDs that prepare to access in cluster m. To control access and avoid congestion, MTCDs need to pass through the ACB procedure during their AGTI before transmitting the preamble. Denote $\theta_m \in (0, 1]$ as the ACB factor. Each active MTCD generates a random number q ($0 \leq q \leq 1$) and compares q with θ_m. If $q \leq \theta_m$, then the MTCD passes the ACB check and proceeds to the random access procedure. Otherwise, it is barred during the current AGTI and waits for the next one. Therefore, the number of MTCDs that pass barring and access at the same time is $C_m\theta_m$. Denote P_m^{suc} as the probability that an MTCD can successfully gain access when competing with other $C_m\theta_m - 1$ contenders. Thus, P_m^{suc} is the probability that an MTCD randomly selects an RAO without collision with other MTCDs. Then, we have

$$P_m^{\text{suc}} = L \cdot \left(\frac{1}{L}\right)\left(1-\frac{1}{L}\right)^{C_m\theta_m-1} = \left(1-\frac{1}{L}\right)^{C_m\theta_m-1}. \tag{1}$$

Denote P_m^{acc} as the probability that an MTCD passes the ACB check and gains access successfully. Since the probability of passing the ACB check is θ_m, we find that P_m^{acc} is the product of θ_m and P_m^{suc}, i.e.,

$$P_m^{\text{acc}} = P_m^{\text{suc}}\theta_m. \tag{2}$$

We define the random access efficiency of cluster m, denoted by E_m, as the efficiency regarding the RAO utilization. It is measured by the number of MTCDs that can be accessed successfully per RAO with the resource provided within an AGTI. Thus, E_m is given as

$$E_m = \frac{C_m P_m^{\text{acc}}}{L}. \tag{3}$$

The random access efficiency of the system E is the sum of efficiencies of all clusters, given by

$$\begin{aligned} E &= \sum_{m=1}^{M} E_m \\ &= \sum_{m=1}^{M} C_m\theta_m \left(\frac{1}{L}\right)\left(1-\frac{1}{L}\right)^{C_m\theta_m-1}. \end{aligned} \tag{4}$$

Since MTCDs in cluster m perform access requests every T_m times, the average random access delay denoted by T_m^0 can be derived as

$$\begin{aligned} T_m^0 &= \sum_{r=0}^{\infty} T_m(r+1)P_m^{\text{acc}}(1-P_m^{\text{acc}})^r \\ &= \frac{T_m}{P_m^{\text{acc}}}. \end{aligned} \tag{5}$$

Considering the essential difference in delay requirements, we assume that different clusters have different delay requirements. Denote T_m^{req} as the access delay requirement of cluster m, and we can write its access delay constraint as $T_m^0 \leq T_m^{\text{req}}$.

In addition to the delay in the access phase, in this paper we also focus on the latency problem in the transmission phase. The transmission QoS requirement for cluster m is defined by two parameters (d_m, ϵ_m), where d_m is the predefined maximum tolerable end-to-end latency threshold and ϵ_m represents the acceptable probability that the delay violates d_m, i.e., $\Pr\{D_m > d_m\} \leq \epsilon_m$, where D_m is the packet delay for the MTCDs in cluster m.

The transmission delay is mainly determined by the resource allocation and traffic queue. In each cluster, assume that NOMA is used by devices for efficient uplink data transmission. All devices are equipped with one antenna. Granted access devices are scheduled to utilize RBs in terms of their traffic arrival rate, channel condition and delay requirement. A queuing model is usually used to determine the packet transmission delay with a known traffic arrival process. Assume that the traffic arrival process is a Poisson process with arrival rate λ_m for devices in cluster m. For a stationary service process, we assume that through proper resource allocation, the transmission rate of a cluster is R_m in an AGTI. Therefore, for a long time across numerous AGTI periods, the average transmission rate of cluster m is $\frac{T_0}{T_m}R_m$. According to the queue model [17], the delay-bound violation probability can be transformed to the queue overflow probability as

$$\Pr\{D_m > d_m\} = \Pr\{Q_m > U_m\}, \tag{6}$$

where Q_m represents the transmission queue length of cluster m in the current time, and U_m is the corresponding queue upper bound beyond which the latency threshold will be violated with

$$U_m = \frac{T_0}{T_m}R_m d_m. \tag{7}$$

For stationary arrival and service processes, large deviation theory can be adopted to derive that the queue overflow probability decreases exponentially with the threshold U_m, i.e., (6) can be approximated as [17, Eq.(6)]

$$Pr\{D_m > d_m\} \approx \exp(-\frac{T_0}{T_m}\alpha_m R_m d_m), \tag{8}$$

where α_m is a positive constant determining the QoS requirement level called the QoS exponent. Substituting (8) into $\Pr\{D_m > d_m\} \leq \epsilon_m$, the transmission delay requirement is derived as

$$-\frac{\ln \epsilon_m}{\alpha_m d_m} \leq \frac{T_0}{T_m} R_m. \tag{9}$$

The left side of (9) can be seen as the minimum transmission rate to satisfy the transmission delay requirement, while the right side of (9) represents the average practical data rate. For a Poisson arrival process, α_m is related to the arrival rate of the Poisson process. Let λ_m be the total traffic arrival rate of cluster m, and we have

$$\alpha_m = \ln\left(1 - \frac{\ln \epsilon_m}{\lambda_m d_m}\right). \tag{10}$$

Note that the traffic arrival process of each MTCD also follows a Poisson distribution. Let $\overline{\lambda}_m$ represent the traffic arrival rate of each MTCD in cluster m. Thus, λ_m is the sum of arrival rates of all the MTCDs in cluster m, which is proportional to the number of successfully accessed MTCDs. Based on this, we have

$$\lambda_m = \overline{\lambda}_m C_m P_m^{\text{acc}}. \tag{11}$$

In the following section, we will investigate the problem of optimizing the access and transmission efficiency while guaranteeing the double requirements of access and transmission delay.

III. Access Management Optimization

Using the system model discussed, we formulate the access management optimization problem as

$$\max_{\{T_m\},\{\theta_m\}} \quad E \tag{12a}$$

$$\text{s.t.} \quad T_m^0 \leq T_m^{\text{req}} \quad \forall m, \tag{12b}$$

$$-\frac{\ln \epsilon_m}{\alpha_m d_m} \leq \frac{T_0}{T_m} R_m \quad \forall m, \tag{12c}$$

$$\sum_{m=1}^{M} \frac{T_0}{T_m} \leq 1, \frac{T_m}{T_0} \in \mathbb{N}^+, \tag{12d}$$

where (12b) and (12c) represent access and transmission delay constraints respectively, $\frac{T_0}{T_m}$ is the percentage of access time of cluster m, (12d) means that the total percentage of access time is no more than 1, and T_m is an integer times of T_0.

By observing the objective function in (12a), it is evident that E is affected by T_m but is not a function of T_m. On this basis, by fixing T_m and replacing the optimization variables $\{\theta_m\}$ with $\{P_m^{\text{acc}}\}$, we can simplify the problem to a subproblem given by

$$\max_{\{P_m^{\text{acc}}\}} \sum_{m=1}^{M} \frac{C_m}{L} P_m^{\text{acc}} \tag{13a}$$

$$\text{s.t.} \quad \frac{T_m}{P_m^{\text{acc}}} \leq T_m^{\text{req}} \quad \forall m, \tag{13b}$$

$$-\frac{\ln \epsilon_m}{\alpha_m d_m} \leq \frac{T_0}{T_m} R_m \quad \forall m. \tag{13c}$$

From (1) and (2), by deriving the derivative of P_m^{acc} with respect to θ_m and setting it to zero, we have

$$\frac{\partial P_m^{\text{acc}}}{\partial \theta_m} = \left(1 - \frac{1}{L}\right)^{C_m \theta_m - 1} \left[1 + C_m \theta_m \log(1 - \frac{1}{L})\right] = 0. \tag{14}$$

Solving the above equation, we obtain the stationary point

$$\widetilde{\theta}_m = \frac{1}{C_m \log(1 - \frac{1}{L})^{-1}}. \tag{15}$$

Because $\frac{\partial^2 P_m^{\text{acc}}}{\partial \theta_m^2} < 0$, by substituting (15) into (2), we can obtain the maximum extremum value of P_m^{acc} without any constraints as

$$\widetilde{P}_m^{\text{acc}} = \frac{1}{eC_m(1 - \frac{1}{L}) \log(1 - \frac{1}{L})^{-1}}. \tag{16}$$

Thus, the optimal solution to (13) is as follows.

***Theorem** 1:* Problem (13) is feasible when T_m satisfies

$$T_m(\exp(\phi_m) - 1) \leq -\frac{T_m^{\text{req}} \ln \epsilon_m}{\overline{\lambda}_m C_m d_m}, \tag{17}$$

$$T_m \leq T_m^{\text{req}} \widetilde{P}_m^{\text{acc}}, \tag{18}$$

where $\phi_m = -\frac{T_m \ln \epsilon_m}{T_0 d_m R_m}$. The optimal solution is obtained as

$$P_m^{\text{acc}*} = \begin{cases} \widetilde{P}_m^{\text{acc}} & \widetilde{P}_m^{\text{acc}} \in [P_m^L, P_m^U] \ \& \ \widetilde{\theta}_m \leq 1 \\ P_m^U & \widetilde{P}_m^{\text{acc}} > P_m^U \ \& \ \theta'_m \leq 1 \\ (1 - \frac{1}{L})^{C_m - 1} & \widetilde{P}_m^{\text{acc}} \in [P_m^L, P_m^U] \ \& \ \widetilde{\theta}_m > 1; \\ & \text{or } \widetilde{P}_m^{\text{acc}} > P_m^U \ \& \ \theta'_m > 1 \end{cases} \tag{19}$$

where P_m^U is defined as

$$P_m^U = -\frac{\ln \epsilon_m}{\overline{\lambda}_m C_m d_m (\exp(\phi_m) - 1)}, \tag{20}$$

and θ'_m is the solution to the following equation:

$$\theta'_m (1 - \frac{1}{L})^{C_m \theta'_m - 1} = P_m^U. \tag{21}$$

Proof: Constraint (13b) gives the lower bound of P_m^{acc}, i.e.,

$$P_m^{\text{acc}} \geq \frac{T_m}{T_m^{\text{req}}} \triangleq P_m^L. \tag{22}$$

By transforming (13c), we can obtain $\alpha_m \geq \phi_m$ and then combining with (10), we obtain the upper bound $P_m^{\text{acc}} \leq P_m^U$. The objective function increases with P_m^{acc} and is a concave function of θ_m. Thus, the optimal solution is achieved in the following cases:

- For the case of $\widetilde{P}_m^{\text{acc}}$ falling in the interval $[P_m^L, P_m^U]$, if $\widetilde{\theta}_m \leq 1$, the optimal ACB factor is $\theta_m^* = \widetilde{\theta}_m$ and $P_m^{\text{acc}*} = \widetilde{P}_m^{\text{acc}}$; otherwise, $\theta_m^* = 1$ and $P_m^{\text{acc}*} = (1 - \frac{1}{L})^{C_m - 1}$.
- For the case of $\widetilde{P}_m^{\text{acc}} > P_m^U$, the optimum is achieved at P_m^U if $\theta'_m \leq 1$ holds, i.e., $\theta_m^* = \theta'_m$ and $P_m^{\text{acc}*} = P_m^U$; otherwise, $\theta_m^* = 1$ and $P_m^{\text{acc}*} = (1 - \frac{1}{L})^{C_m - 1}$.
- For the case $P_m^L > P_m^U$ or $\widetilde{P}_m^{\text{acc}} < P_m^L$, the problem is insoluble due to the unreasonable parameter settings, from which we obtain the feasible conditions (17) and (18).

Summarizing the above cases, we obtain the optimal solution in (19). ■

Given P_m^{acc}, we observe that decreasing T_m benefits both delay constraints (12b) and (12c). However, the service time is shared among clusters as constrained by (12d). Therefore, fixing P_m^{acc}, the optimal T_m can be obtained by maximizing the sum percentage of access time of all clusters as

$$\max_{\{T_m\}} \sum_{m=1}^{M} \frac{T_0}{T_m} \tag{23a}$$

$$\text{s.t.} \quad T_m \le T_m^{\text{req}} P_m^{\text{acc}} \quad \forall m, \tag{23b}$$

$$T_m \le -T_0 R_m \frac{\alpha_m d_m}{\ln \epsilon_m} \quad \forall m, \tag{23c}$$

$$\sum_{m=1}^{M} \frac{T_0}{T_m} \le 1, \frac{T_m}{T_0} \in \mathbb{N}^+. \tag{23d}$$

To simplify problem (23), define $n_m = T_m/T_0$ as the access period of the cluster and $R_m^{\min} = -\frac{\ln \epsilon_m}{\alpha_m d_m}$, we transform the problem as

$$\max_{\{n_m\}} \sum_{m=1}^{M} \frac{1}{n_m} \tag{24a}$$

$$\text{s.t.} \quad 1 < n_m \le \min\left\{ \frac{T_m^{\text{req}} P_m^{\text{acc}}}{T_0}, \frac{R_m}{R_m^{\min}} \right\} \forall m, \tag{24b}$$

$$\sum_{m=1}^{M} \frac{1}{n_m} \le 1, n_m \in \mathbb{N}^+. \tag{24c}$$

The transformed problem (24) is a nonlinear integer optimization problem. We use the genetic algorithm described in [28] to solve it. Then, by alternatively optimizing problems (13) and (24), we can obtain the locally optimal solution to (12). The access management algorithm is illustrated in Algorithm 1.

The time complexity of the genetic algorithm with tournament selection is $\mathcal{O}(M)$ [28]. As shown from Line 7 to Line 21 in Algorithm 1, each iteration requires computing $P_m^{\text{acc(i)}}$ for the M clusters, the complexity of which is $\mathcal{O}(M)$. Therefore, the total time complexity of each iteration for Algorithm 1 is $\mathcal{O}(M)$.

In this section, we solve the access control problem among the clusters, and the transmission delay obtains the initial guarantee from the perspective of the whole cluster. Next, we will further guarantee the transmission delay requirement for each MTCD through resource allocation among MTCDs in a cluster.

Algorithm 1 Access management optimization algorithm

1: Initialize $P_m^{\text{acc}(1)}, \forall m$ for each cluster and set the iteration number $i = 1$.
2: Calculate $\widetilde{\theta}_m$ and $\widetilde{P}_m^{\text{acc}}$ by (15) and (16) respectively.
3: **repeat**
4: **for** $m = 1$ to M **do**
5: With fixed $\{P_m^{\text{acc(i)}}\}$, solve problem (24) using the genetic algorithm to obtain $n_m^{(i+1)}$;
6: Calculate the optimal $T_m^{(i+1)} = n_m^{(i+1)} T_0$;
7: With fixed $\{T_m^{(i+1)}\}$, calculate P_m^U and P_m^L;
8: **if** $\widetilde{P}_m^{\text{acc}} \le P_m^U$ **then**
9: **if** $\widetilde{\theta}_m < 1$ **then**
10: Set $P_m^{\text{acc(i+1)}} = \widetilde{P}_m^{\text{acc}}$, $\theta_m^{(i+1)} = \widetilde{\theta}_m$;
11: **else**
12: Set $\theta_m^{(i+1)} = 1$ and $P_m^{\text{acc(i+1)}} = (1 - \frac{1}{L})^{C_m - 1}$;
13: **end if**
14: **else**
15: Set $P_m^{\text{acc(i+1)}} = P_m^U$ and solve equation (21) to obtain θ'_m;
16: **if** $\theta'_m < 1$ **then**
17: $\theta_m^{(i+1)} = \theta'_m$;
18: **else**
19: Set $\theta_m^{(i+1)} = 1$ and $P_m^{\text{acc(i+1)}} = (1 - \frac{1}{L})^{C_m - 1}$;
20: **end if**
21: **end if**
22: **end for**
23: Update $i \leftarrow i + 1$;
24: **until** Convergence is achieved, i.e., $|E(\{\theta_m^{(i+1)}\}) - E(\{\theta_m^{(i)}\})| < \delta$

IV. Intra-cluster Resource allocation

After optimizing the access control problem among clusters, we investigate the resource allocation problem within a cluster.

Assume that $K_m(t)$ devices are granted access in cluster m at time t. For convenience, we omit the indexs m and t as K. Let $x_{k,n}$ be the RB assignment status that takes the value of 1 or 0 when the nth RB is or is not occupied by the kth device. The channel gain of the kth device upon the nth RB is denoted by $h_{k,n}$. We consider a block fading channel, where the channel gain remains constant within one RB but varies independently from one to another. According to the uplink NOMA principle, devices are allowed to transmit through the same RB. At the receiver, successive interference cancellation (SIC) is implemented to separate the superimposed device signals. The set of devices that share the nth RB for the current scheduling time is represented by Ω_n. The maximum number of devices on the same RB is set by U, i.e., $|\Omega_n| \le U$. The detecting order on the nth RB is denoted by a permutation π_n. Signals of devices are detected in the order of $\pi_n(1), \pi_n(2), \cdots, \pi_n(|\Omega_n|)$, which means device $\pi_n(i)$ will see interference from device $\pi_n(i+1)$ to $\pi_n(|\Omega_n|)$. Usually, the detecting order is determined by the descending order of the received signal-to-noise ratio of devices on RB n. Let $I_{k,n}$ be the interference power to the kth device upon the nth RB as

$$I_{k,n} = \sum_{j=\pi_n^{-1}(k)+1}^{|\Omega_n|} x_{j,n} p_{j,n} |h_{j,n}|^2, \tag{25}$$

where $\pi_n^{-1}(k)$ is the detecting order of device k in π_n, i.e., $\pi_n^{-1}(k) = i$ if $\pi_n(i) = k$. Therefore, the achievable data rate of the kth device on the nth RB can be obtained as

$$r_{k,n} = B \log_2 \left(1 + \frac{p_{k,n} |h_{k,n}|^2}{I_{k,n} + \sigma^2} \right), \tag{26}$$

where $p_{k,n}$ is the transmission power of the kth device on the nth RB, and σ^2 is the Gaussian white noise power.

We consider two utilities of the radio resources. To improve the total throughput of the cluster, we define the utility function

as

$$U(\boldsymbol{X},\boldsymbol{P})=\sum_{k=1}^{K}\sum_{n=1}^{N}x_{k,n}r_{k,n}, \tag{27}$$

where $\boldsymbol{X}=[x_{k,n}]$, $\boldsymbol{P}=[p_{k,n}]$. To consider the power consumption, the utility function is defined as

$$U(\boldsymbol{X},\boldsymbol{P})=-\sum_{k=1}^{K}\sum_{n=1}^{N}x_{k,n}p_{k,n}. \tag{28}$$

To guarantee QoS for devices in each cluster, we formulate an optimization problem to optimize resource allocation within each cluster. The optimization problem is

$$\max_{\boldsymbol{X},\boldsymbol{P}}\ U(\boldsymbol{X},\boldsymbol{P}) \tag{29a}$$

$$\text{s.t.}\quad \sum_{n=1}^{N}x_{k,n}r_{k,n}\geq r_{\min}, \tag{29b}$$

$$\sum_{k=1}^{K_m}x_{k,n}\leq U, x_{k,n}\in\{0,1\}\quad \forall n, \tag{29c}$$

$$\sum_{n=1}^{N}x_{k,n}p_{k,n}\leq P_0,\ p_{k,n}\geq 0\quad \forall k, \tag{29d}$$

where P_0 is the total available power for each device, and $r_{\min}$ is the minimum rate to guarantee that each MTCD satisfies the transmission delay of its cluster, i.e., the maximum delay is d_m with violating probability not exceeding ϵ_m. From the analysis in (6)-(9), we have

$$r_{\min}=-\frac{\ln\epsilon_m}{\alpha_0 d_m}\frac{T_m}{T_0}, \tag{30}$$

where α_0 is the QoS exponent related to the traffic arrival rate $\overline{\lambda}_m$ of the MTCD, given by

$$\alpha_0=\ln\left(1-\frac{\ln\epsilon_m}{\overline{\lambda}_m d_m}\right). \tag{31}$$

Constraint (29c) indicates that at most U devices can share one RB. Note that this condition includes the case when some RBs are occupied by only one device, leading to orthogonal transmission on these RBs. Note that this is a coupled mixed integer nonconvex optimization problem due to the binary constraint in (29c) and the interference term in the data rate. It is difficult to solve since even the degraded problem without QoS guarantee is proven to be NP-hard [29]. We propose a framework to optimize the problem iteratively.

For nonnegative variables γ and $\overline{\gamma}$, we have the following inequality:

$$B\log_2(1+\gamma)\geq b\log_2\gamma+c, \tag{32}$$

where b and c are parameters of $\overline{\gamma}$ given by

$$b=\frac{B\overline{\gamma}}{1+\overline{\gamma}},\quad c=B\log_2(1+\overline{\gamma})-\frac{B\overline{\gamma}}{1+\overline{\gamma}}\log_2\overline{\gamma}, \tag{33}$$

respectively. The bound (32) is tight at $\gamma=\overline{\gamma}$. As a result, for each iteration, we approximate the rate (26) by the lower bound

$$\overline{r}_{k,n}=b_{k,n}\log_2\gamma_{k,n}+c_{k,n}, \tag{34}$$

Algorithm 2 Intra-cluster Resource Allocation Algorithm

1: Initialize the power allocation vector $\boldsymbol{P}^{(1)}$ and set the iteration number $i=1$.
2: **repeat**
3: Calculate the SINR $\gamma_{k,n}$ with $\boldsymbol{P}^{(i)}$;
4: Set $\overline{\gamma}_{k,n}=\gamma_{k,n}$ and then compute $b_{k,n}$ and $c_{k,n}$ according to (33);
5: Construct the convex problem (36) and solve it to obtain the optimal solution $\boldsymbol{Q}^{(i+1)}$;
6: Update the power allocation matrix $\boldsymbol{P}^{(i+1)}=2^{\boldsymbol{Q}^{(i+1)}}$;
7: Perform the RB assignment according to Strategy 1 and obtain the RB allocation matrix $\boldsymbol{X}^{(i+1)}$;
8: Update $i\leftarrow i+1$;
9: **until** Convergence is achieved, i.e., $|U(\boldsymbol{P}^{(i+1)},\boldsymbol{X}^{(i+1)})-U(\boldsymbol{P}^{(i)},\boldsymbol{X}^{(i)})|<\delta$

where $\gamma_{k,n}=\frac{p_{k,n}|h_{k,n}|^2}{I_{k,n}+\sigma^2}$ denotes the signal-to-interference plus noise ratio (SINR). By letting $p_{k,n}=2^{q_{k,n}}$ and combining with (26), (34) can be transformed into

$$\overline{r}_{k,n}=b_{k,n}\left[q_{k,n}+\log_2|h_{k,n}|^2-\log_2\left(\sum_{j=\pi_n^{-1}(k)+1}^{|\Omega_n|}2^{q_{j,n}}|h_{j,n}|^2+\sigma^2\right)\right]+c_{k,n}. \tag{35}$$

Given the RB allocation, we obtain the power allocation subproblem

$$\max_{\boldsymbol{Q}}\ \overline{U}(\boldsymbol{Q}) \tag{36a}$$

$$\text{s.t.}\quad \sum_{n=1}^{N}\overline{r}_{k,n}\geq r_{\min}\quad \forall k, \tag{36b}$$

$$\sum_{n=1}^{N}2^{q_{k,n}}\leq P_0\quad \forall k, \tag{36c}$$

where $\boldsymbol{Q}=[q_{k,n}]$, $\overline{U}(\boldsymbol{Q})$ is the transformed utility function. To improve the throughput, $\overline{U}(\boldsymbol{Q})$ is written as

$$\overline{U}(\boldsymbol{Q})=\sum_{k=1}^{K}\sum_{n=1}^{N}\overline{r}_{k,n}, \tag{37}$$

while for the aim of the power, $\overline{U}(\boldsymbol{Q})$ is transformed as

$$\overline{U}(\boldsymbol{Q})=-\sum_{k=1}^{K}\sum_{n=1}^{N}2^{q_{k,n}}. \tag{38}$$

The following theorem states the characteristic of problem (36).

***Theorem** 2:* Problem (36) is a convex optimization problem.

Proof: It can be observed that $\overline{r}_{k,n}$ is a concave function of $\boldsymbol{Q}$ since the log-sum-exp function is convex. Thus, constraint (36b) forms a convex set, and objective (37) is concave. Due to the convexity of the exponential function, constraint (36c) is also a convex set, and objective (38) is concave. Therefore, for both utility functions, problem (36) is a convex optimization problem. ■

For the convex problem (36), the optimal solution can be found by a standard convex problem solver. Then, we obtain $\boldsymbol{P}$ for the current iteration and calculate the individual utility function, which is defined as $u_{k,n} = r_{k,n}$ or $u_{k,n} = -p_{k,n}$ for RB n. Then, the RB is assigned by the following strategy (Strategy 1):

- If no more than U devices have positive power on RB n, then allocate the RB to these devices;
- If more than U devices have positive power on RB n, then allocate the RB to the top U devices with the highest individual utility function, i.e., set $x_{k,n} = 1$ if $u_{k,n}$ is the top U maximum for $n \in \{1, \ldots, N\}$; otherwise, set $x_{k,n} = 0$ and $p_{k,n} = 0$.

Then, we use the updated $\boldsymbol{P}$ and $\boldsymbol{X}$ to reconstruct the subproblem (36) and solve it. The proposed sequential convex programming iterative algorithm (SCPIA) for intra-cluster resource allocation is shown in Algorithm 2. It can be observed that $U(\boldsymbol{P}^{(i)}, \boldsymbol{X}^{(i)}) = \overline{U}(\boldsymbol{Q}^{(i)})$, and after each iteration, we have $\overline{U}(\boldsymbol{Q}^{(i)}) \geq \overline{U}(\boldsymbol{Q}^{(i-1)})$. Therefore, we obtain $U(\boldsymbol{P}^{(i)}, \boldsymbol{X}^{(i)}) \geq U(\boldsymbol{P}^{(i-1)}, \boldsymbol{X}^{(i-1)})$, i.e., Algorithm 2 monotonically increases the value of $U(\boldsymbol{P}, \boldsymbol{X})$ in (29) at each iteration and finally converges.

According to [30], the time complexity of solving the convex problem (36) using interior point methods with a ν-self-concordant barrier is $\mathcal{O}(KN\sqrt{\nu}\log\frac{\nu}{\varepsilon})$, where ε is the error tolerance for algorithm termination. The RB assignment strategy involves sorting K devices for each RB. Accordingly, the complexity is $\mathcal{O}(KN)$. Therefore, the time complexity of each iteration for Algorithm 2 is $\mathcal{O}(KN)$.

Note that in the resource allocation, the parameter T_m is passed from the access control layer and influences the constraint related to the transmission delay. The throughput of the cluster, represented by R'_m, is obtained and used to update the access control results. R'_m is calculated by $R'_m = \sum_{k=1}^{K}\sum_{n=1}^{N} x^*_{k,n} r_{k,n}(\boldsymbol{P}^*)$, where $\boldsymbol{X}^*, \boldsymbol{P}^*$ is the optimal solution to problem (29). In the next section, we will try to combine access control with intra-cluster resource allocation and propose a joint access management scheme.

V. Joint Access Management Scheme

We combine the proposed access management in Section III and the intra-cluster resource allocation in Section IV. A QoS guaranteed joint access management (Q-JAM) scheme is proposed in Algorithm 3. Considering a long access time, for the access management layer, transmission rate R_m is determined by the average total throughput after resource allocation for cluster m since the channel changes from one AGTI to another. Specifically, let $R'_m(t)$ be the throughput at time t for cluster m, and R_m is calculated by

$$R_m = \frac{1}{t_0}\sum_{\tau=t-t_0}^{t-1} R'_m(\tau), \tag{39}$$

where t_0 is the length of the sliding window for evaluating the average throughput. At the resource allocation stage, the number of granted MTCDs is determined by the product of the number of active MTCDs and the successful access probability, i.e., we have

$$K_m = \lfloor C_m P_m^{\text{acc}} \rfloor = \lfloor C_m \left(1 - \frac{1}{L}\right)^{C_m\theta_m - 1} \theta_m \rfloor, \tag{40}$$

where $\lfloor x \rfloor$ means obtaining the maximum integer not greater than x. The procedure of Q-JAM is summarized as follows.

Assume the channel follows independent and identical Rayleigh fading across RBs. For the joint Q-JAM scheme, we define the average delay violation probability as

$$P_v = \frac{1}{M}\sum_{m=1}^{M} \frac{K_m^v}{K_m}, \tag{41}$$

where K_m and K_m^v are the number of granted access MTCDs and the number of MTCDs whose rate is lower than the minimum rate that guarantees the transmission delay. We will then give the numerical simulation results on the proposed scheme.

Algorithm 3 QoS guaranteed joint access management (Q-JAM) scheme

```
Initialize the transmission rate {R'_m(1)}.
for t = 1 to T do
    Calculate {R_m} according to (39) for each cluster;
    Perform the access management optimizing algorithm according to Algorithm 1 to obtain optimal {T_m} and {θ_m};
    Calculate {K_m} according to (40) for each cluster;
    for m = 1 to M do
        Perform the intra-cluster resource allocation according to Algorithm 2;
        Update the total throughput R'_m(t) = Σ_{k=1}^{K} Σ_{n=1}^{N} x_{k,n} r_{k,n};
    end for
end for
```

TABLE I
Simulation Parameters for Access Management

Parameters	Cluster 1	Cluster 2	Cluster 3
Access delay requirements	10	20	30
Maximum transmission delay (ms)	1	10	100
Acceptable delay violating probability	0.01	0.05	0.1
Transmission rate of cluster (kbps)	10	40	90
Active MTCDs	5	10	20
Traffic arrival rate (kbps)	0.5	1	1.5

VI. Simulation Results

In this section, we evaluate the performance of the proposed access management method and the intra-cluster resource allocation algorithm. Moreover, we investigate the relationship between intra-cluster resource allocation and inter-cluster access management.

A. Performance of Access Management Method

Assume that there are 3 clusters with different traffic characteristics, including delay requirements and traffic arrival. Set $T_0 = 1$, and other simulation parameters are listed in Table I.

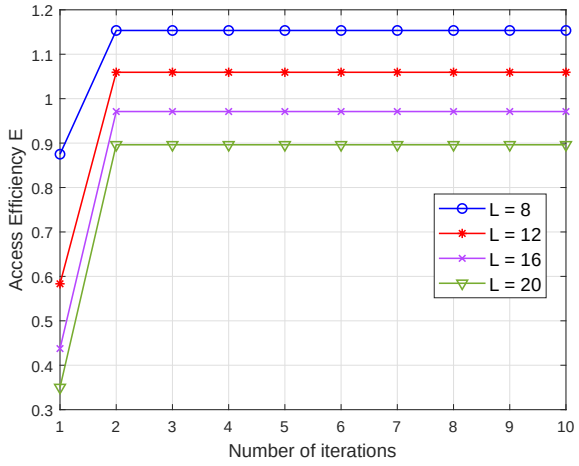

Fig. 2. Access efficiency E versus number of iterations.

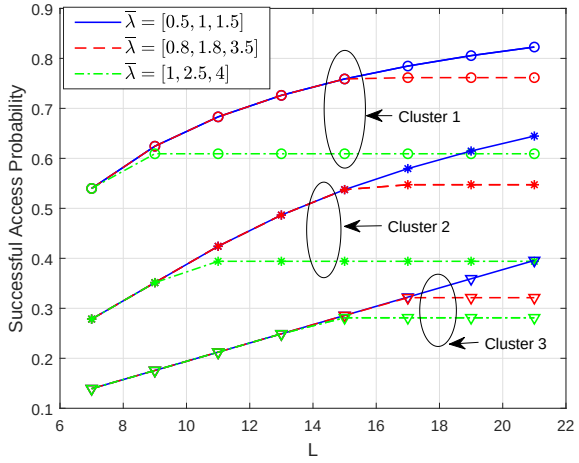

Fig. 3. Successful access probability P_m^{acc} versus number of RAOs L.

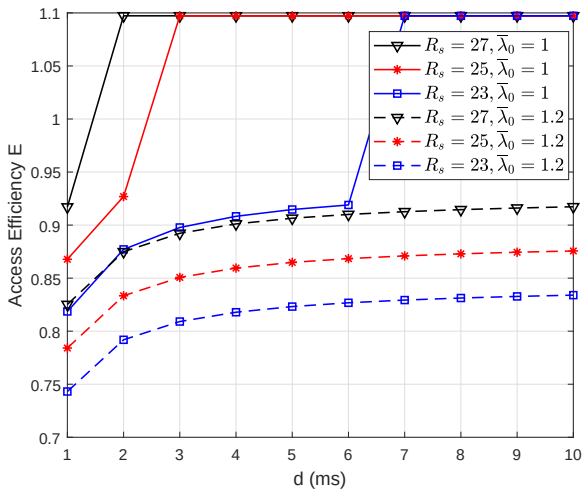

Fig. 4. Access efficiency E versus transmission delay d.

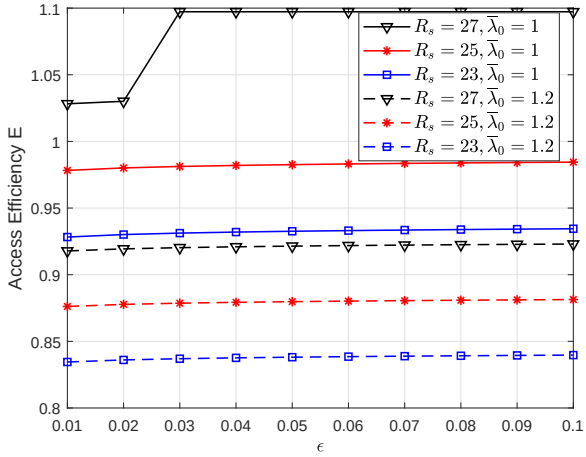

Fig. 5. Access efficiency E versus acceptable delay-violating probability ϵ.

Fig. 2 shows that the access efficiency E converges to the maximum as the iteration proceeds for different L. The average time cost of one iteration in Algorithm 1 is evaluated by MATLAB to be 0.067 seconds. It can be seen that E declines with L, which means the average access efficiency per RAO decreases as the number of RAOs increases.

Denoting the average traffic arrival rate of all clusters as a vector $\overline{\lambda} = [\overline{\lambda_1}, \ldots, \overline{\lambda_M}]$ and setting the value of $\overline{\lambda}$ (kbps) as $[0.5, 1, 1.5], [0.8, 1.8, 3.5], [1, 2.5, 4]$ and keeping the other parameters unchanged, we plot the curve of successful access probability P_m^{acc} for each m versus L in Fig. 3. It can be seen that P_m^{acc} is a nondecreasing function of L. Specifically, P_m^{acc} increases with L when L is in a certain interval while remaining constant when L is higher than a threshold. In addition, both the threshold value and achievable successful access probability decrease as $\overline{\lambda}_m$ increases. This is because more RAOs can facilitate successful access but the benefit is limited by the delay constraint (13c), which leads to the upper bound of P_m^{acc}, as analyzed in (20). It can also be observed that the cluster with more stringent delay requirements has a higher successful access probability. Thus, through our access management algorithm, various delay requirements of different clusters are satisfied by allocating different access time intervals and ACB factors.

To investigate the influence of transmission delay, we set the maximum transmission delay of the three clusters to be identical, i.e., $d_1 = d_2 = d_3 = d$. Meanwhile, set $L = 20$, the active MTCDs $C_1 = C_2 = C_3 = 15 \triangleq C$, the transmission rate and traffic arrival rate of each cluster to be the same as R_s and $\overline{\lambda}_0$, respectively, and the other parameters are from Table I. Fig. 4 shows the optimized results for access efficiency E for varying maximum transmission delay d with different transmission rates ($R_s = 27$, $R_s = 25$, $R_s = 23$ kbps) and different traffic arrival rates ($\overline{\lambda}_0 = 1$ and $\overline{\lambda}_0 = 1.2$ kbps). As a whole, E improves as d increases, which means looser transmission delay enhances access efficiency. However, the degree of access efficiency improvement is related to the settings of transmission and traffic arrival rate, which determines when the optimal solution is achieved as analyzed in Theorem 1. Specifically, the optimal $P_m^{\text{acc}*}$ is obtained at P_m^U for $m = 2$ and $m = 3$ with $\overline{\lambda}_0 = 1.2$ kbps. As a result, E

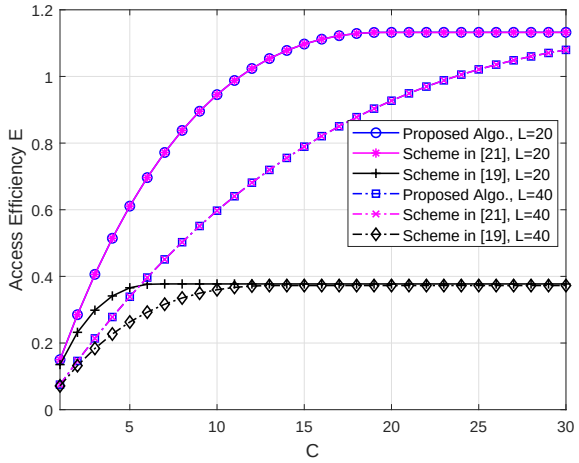

Fig. 6. Access efficiency E versus number of active MTCDs C for different access control schemes.

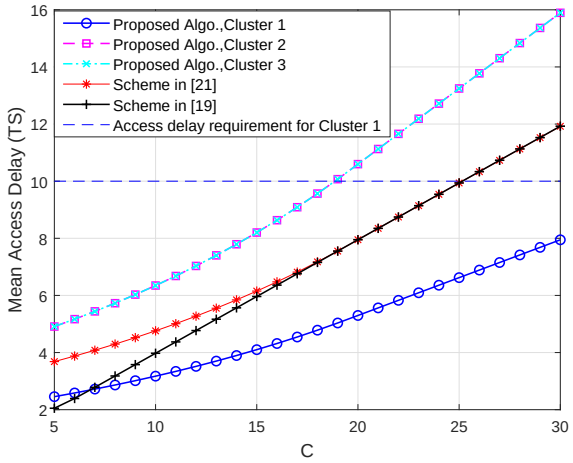

Fig. 7. Mean access delay versus number of active MTCDs C for different access control schemes.

rises slowly as d increases for such a high traffic arrival rate. In contrast, E rises quickly with a relatively low traffic arrival rate of 1 kbps. This is because the optimal $P_m^{\text{acc}*}$ is obtained at $(1-\frac{1}{L})^{C_m-1}$ as the delay constraint relaxes. Since this optimal value is irrelevant to d, E remains fixed when d is beyond a certain threshold. Moreover, E peaks more quickly with smaller d for larger transmission rate settings.

Now, set the other parameters as above and the acceptable delay-violating probability of each cluster to be identical, i.e., $\epsilon_m = \epsilon, \forall m$ and the maximum transmission delay as $1, 10, 100$ ms. The curve of E versus ϵ is presented in Fig. 5. For a large transmission rate ($R_s = 27$ kbps) and low traffic arrival rate ($\overline{\lambda}_0 = 1$ kbps), E jumps to the top at $\epsilon = 0.03$ since from that point the optimal $P_m^{\text{acc}*}$ is obtained at $(1-\frac{1}{L})^{C_m-1}$ for all m. For the other R_s and $\overline{\lambda}_0$, $P_m^{\text{acc}*}$ is obtained at P_m^U, which is a logarithmic form of ϵ. This leads to a slight increase in E.

We compare the proposed access control algorithm with the schemes in [21] and [19]. The MTCDs are partitioned into virtual clusters and the AGTIs are equally allocated among clusters. The ACB factor for each cluster is optimized in [21], while all the MTCDs are considered as a whole, and the ACB factor is optimized in [19]. We set the transmission rate and traffic arrival rate of all clusters to 40 and 0.5 kbps, respectively, and Fig. 6 illustrates the results of access efficiency versus the number of active MTCDs C for $L = 20$ and $L = 40$. It can be seen that the optimized access efficiencies of the proposed scheme and the one in [21] are nearly equal, and are much better than the scheme in [19]. This indicates that clustering helps alleviate access collisions and improve access efficiency. Regarding the results of the mean access delay shown in Fig. 7, which is calculated by (5), we find that the proposed algorithm can well guarantee the stringent access delay requirement for Cluster 1, whose maximum access delay is 10 TS. This is because the proposed scheme can flexibly adjust the AGTI allocation time for clusters with different delay requirements.

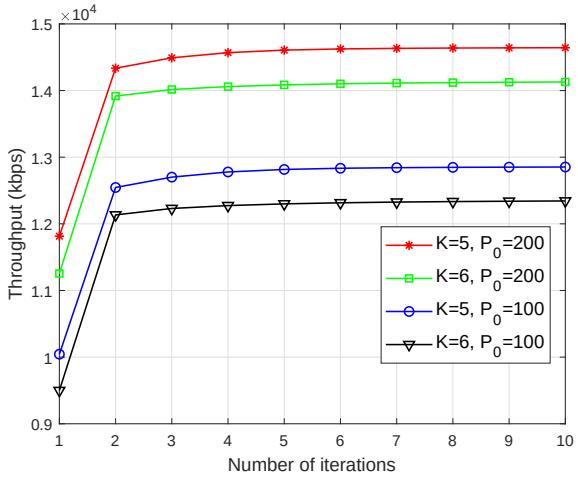

Fig. 8. Throughput versus number of iterations.

B. Performance of Intra-cluster Resource Allocation

Take one cluster as an example. Assume the access period of the cluster is $n_m = T_m/T_0 = 2$ and the channel suffers from Rayleigh fading. The simulation parameters are listed in Table II. It is shown in Fig. 8 that the proposed resource allocation algorithm converges when the utility function is set as the total throughput for the maximum power $P_0 = 100$ and $P_0 = 200$ mW and the number of granted devices $K = 5$ and $K = 6$. Similarly, the algorithm converges when minimizing the total consumed power. The relevant figure is omitted to save space. By adopting the CVX tool to solve the convex problem (36) in MATLAB, the average time cost of one iteration in Algorithm 2 is 6.2 seconds.

We compare the utility performance of SCPIA with the iterative water-filling algorithm for orthogonal multiple access (IWA-OMA) [31] and greedy-based RB allocation plus Lagrange-dual based power allocation (GA-LDPA) [23]. Setting $K = 5$ and the throughput as the utility function, as shown in Fig. 9, the proposed SCPIA achieves the highest throughput performance. The performance gain of SCPIA is due to the

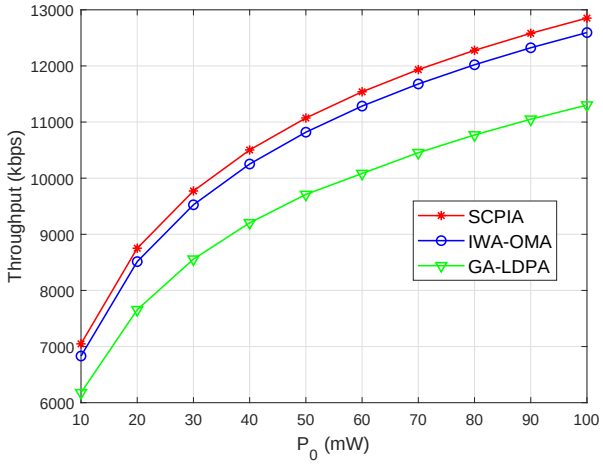

Fig. 9. Throughput versus total available power P_0.

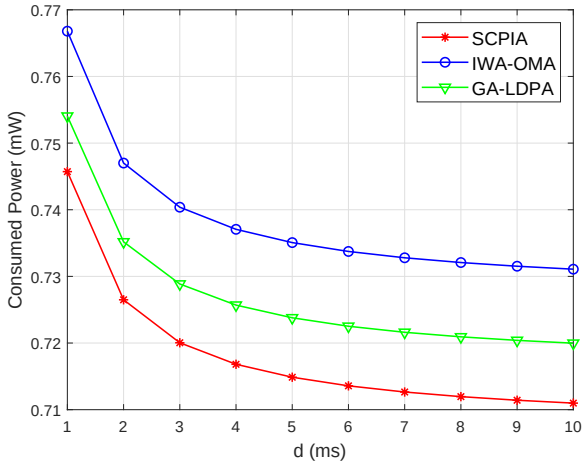

Fig. 10. Consumed power versus transmission delay d.

TABLE II
SIMULATION PARAMETERS FOR INTRA-CLUSTER RESOURCE ALLOCATION

Parameters	Values
Bandwidth of each RB	$B = 180$ kHz
Number of RBs in an AGTI	$N = 10$
Power of Gaussian noise	1 mW
Traffic arrival rate	50 kbps
Maximum tolerable latency	1 ms
The acceptable delay-bound violation probability	$\epsilon = 1\%$
Maximum number of devices on the same RB	$U = 2$

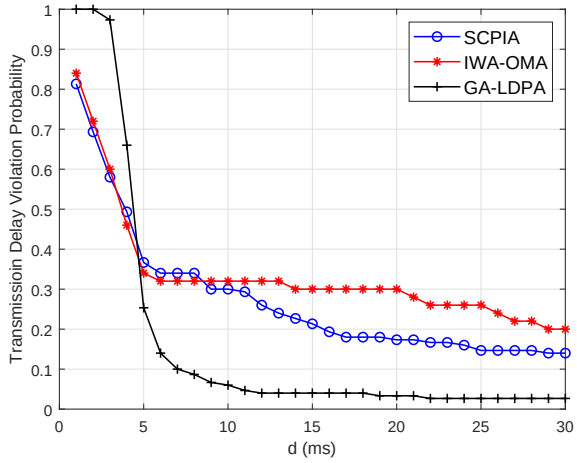

Fig. 11. Transmission delay violation probability versus transmission delay d.

gain of more flexible usage of radio resources of NOMA compared with IWA-OMA. On the other hand, SCPIA and IWA-OMA outperform GA-LDPA since RB is allocated to the device with the best channel gain in GA-LDPC, which leads to a locally optimal solution compared with the joint optimization of RB and power allocation. It can be seen that the throughput of the system is improved as P_0 increases because the optimal solution is achieved at the equality of equation (29d). However, the growth rate of throughput decreases with increasing P_0 due to the logarithmic characteristic of the data rate. In Fig. 10, we investigate the power consumption versus the maximum transmission delay d by setting the power consumption as the optimization objective. It can be seen that SCPIA outperforms the other two schemes in that it consumes less power to satisfy the minimum rate requirement derived from the transmission delay. The consumed power decreases with d, and the declining rate gradually reduces, which means that as the delay requirement becomes more stringent, the required power increases dramatically.

Set $K = 10$, $N = 20$, $P_0 = 50$ mW, $B = 180$ Hz and the traffic arrival rate as 600 bps. Perform resource allocations with a random channel one hundred times. Then, count the number of MTCDs whose rate is less than the minimum rate that guarantees the transmission delay, the ratio of which to the total number of accessed MTCDs is defined as the transmission delay violation probability, as illustrated in Fig. 11. This indicates that the transmission delay violation probability decreases exponentially with the maximum transmission delay d, which is consistent with (8). It can also be seen that the proposed algorithm, SCPIA, outperforms the other two algorithms in the low latency regime and stays at the moderate level as d increases. GA-LDPA has better performance with respect to fairness, so it achieves a lower transmission delay violation probability, but it is at the expense of throughput.

C. Performance of Joint Access Management Scheme

In this section, we investigate the performance of the QoS guaranteed joint access management (Q-JAM) scheme.

We consider three clusters with transmission delay requirements set the same as shown in Table I. To evaluate the transmission delay violation probability, the access delay requirements are set to 30 ms. The traffic arrival rates are all set to $\overline{\lambda}_0 = 10$ kbps, and the number of RAOs is $L = 30$. For each cluster, the bandwidth of RB is $B = 5$ kHz, maximum power $P_0 = 20$ mW and other intra-cluster related parameters are set in Table II. The simulator runs for 100 AGTIs and $t_0 = 5$, i.e., R_m is obtained by the average of the previous five cluster throughputs. Fig. 12 shows the transmission delay violation probability versus the number of active MTCDs.

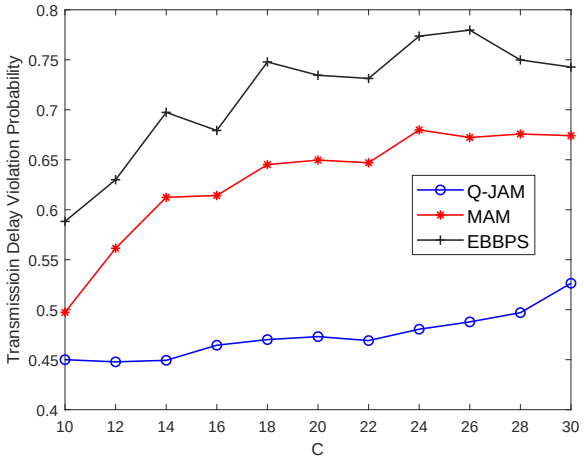

Fig. 12. Transmission delay violation probability versus number of active MTCDs C.

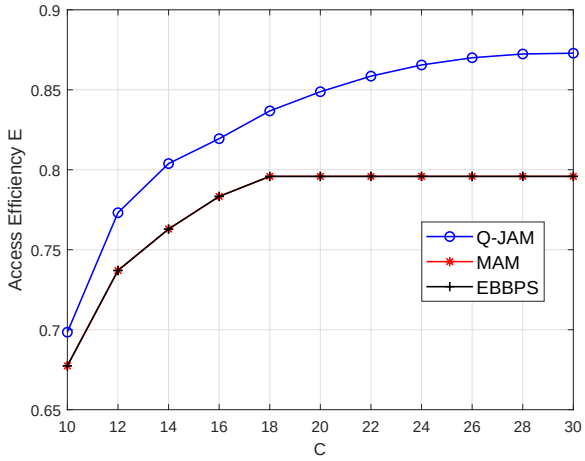

Fig. 13. Access efficiency E versus number of active MTCDs C.

For comparison, the massive access management (MAM) and the EB-based period scheduling (EBBPS) [17] are used as benchmark algorithms, where $T_m, \forall m$ are inversely proportional to $\overline{\lambda}_m$ and $r_{\min}$, respectively, for MAM and EBBPS. We can see from the figure that the proposed Q-JAM scheme can well guarantee the delay requirements by controlling the number of granted MTCDs by optimizing T_m and θ_m, which improves the transmission rate of each accessed MTCD to address the stringent delay requirements due to the high-speed traffic arrival rates. However, for MAM and EBBPS, more MTCDs are granted as C increases since T_m and θ_m cannot be jointly adjusted, which leads to a higher delay violation probability.

Fig. 13 illustrates the access efficiency E versus C, where the traffic arrival rates are set as $\overline{\lambda} = [2, 4, 6]$ kbps, the transmission rates $\{R_m\}$ are $[80, 100, 120]$ kbps and the transmission delay requirements are in Table I. It can be seen that the proposed Q-JAM scheme can significantly improve the access efficiency E while guaranteeing the delay requirements. This is because by jointly controlling the access time T_m and ACB factors θ_m, Q-JAM can optimize the total access efficiency instead of optimizing the individual efficiency of each cluster. Moreover, the access efficiency of Q-JAM increases with the number of active MTCDs C, while MAM and EBBPS remain constant from $C = 18$ onward.

VII. Conclusion

In this paper, we considered the random access and resource allocation problem in clustering-based massive MTC networks. An access control problem was formulated to maximize the random access efficiency with the double delay constraints for the access and transmission stage. The problem was analyzed and solved by an iterative algorithm, which can adaptively adjust the access time interval and back-off factors of these clusters. The resource allocation was formulated to maximize the system utility function with the constraints of delay and number of MTCDs sharing one resource block with NOMA and solved by the proposed sequential convex programming iterative algorithm. Simulation results demonstrated the superiority of the proposed scheme in terms of access efficiency, delay, throughput and consumed power. The impact of system parameters was also discussed. This showed that higher delay requirements can reduce the access efficiency, while the damage extent is relevant to the transmission and traffic rate. In a practical changing channel, it is helpful to consider the transmission delay in the cluster access control stage.

References

[1] M. T. Islam, A.-e. M. Taha, and S. Akl, "A survey of access management techniques in machine type communications," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 74–81, 2014.

[2] S. Chen, R. Ma, H.-H. Chen, H. Zhang, W. Meng, and J. Liu, "Machine-to-machine communications in ultra-dense networks – a survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1478–1503, 2017.

[3] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[4] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 426–471, 2020.

[5] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar, and J. J. P. C. Rodrigues, "Tactile internet for smart communities in 5G: An insight for NOMA-based solutions," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 3104–3112, 2019.

[6] N. Xia, H.-H. Chen, and C.-S. Yang, "Radio resource management in machine-to-machine communications a survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 1, pp. 791–828, 2018.

[7] W. Zhan and L. Dai, "Massive random access of machine-to-machine communications in LTE networks: Modeling and throughput optimization," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2771–2785, 2018.

[8] J. Choi and J. Ding, "Co-existing preamble and data transmissions in random access for MTC with massive MIMO," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7576–7586, 2021.

[9] H. S. Jang, H. Jin, B. C. Jung, and T. Q. S. Quek, "Resource-optimized recursive access class barring for bursty traffic in cellular IoT networks," *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11 640–11 654, 2021.

[10] Y. Liang, X. Li, J. Zhang, and Z. Ding, "Non-orthogonal random access for 5G networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4817–4831, 2017.

[11] S. Moon, H.-S. Lee, and J.-W. Lee, "SARA: Sparse code multiple access-applied random access for IoT devices," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 3160–3174, 2018.

[12] Y. Wang, T. Wang, Z. Yang, D. Wang, and J. Cheng, "Throughput-oriented non-orthogonal random access scheme for massive MTC networks," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1777–1793, 2020.

[13] Z. Chen, Q. Yao, H. H. Yang, and T. Q. S. Quek, "Massive wireless random access with successive decoding: Delay analysis and optimization," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 457–471, 2019.

[14] S.-Y. Lien and K.-C. Chen, "Massive access management for QoS guarantees in 3GPP machine-to-machine communications," *IEEE Communications Letters*, vol. 15, no. 3, pp. 311–313, 2011.

[15] L. Liang, L. Xu, B. Cao, and Y. Jia, "A cluster-based congestion-mitigating access scheme for massive M2M communications in internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2200–2211, 2018.

[16] T. Wang, Y. Wang, C. Wang, Z. Yang, and J. Cheng, "Group-based random access and data transmission scheme for massive MTC networks," *IEEE Transactions on Communications*, pp. 1–1, 2021.

[17] P. Si, J. Yang, S. Chen, and H. Xi, "Adaptive massive access management for QoS guarantees in M2M communications," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3152–3166, July 2015.

[18] C. Zhang, X. Sun, J. Zhang, X. Wang, S. Jin, and H. Zhu, "Throughput optimization with delay guarantee for massive random access of M2M communications in industrial IoT," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 077–10 092, 2019.

[19] C.-Y. Oh, D. Hwang, and T.-J. Lee, "Joint access control and resource allocation for concurrent and massive access of M2M devices," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4182–4192, 2015.

[20] R. Chai, C. Liu, and Q. Chen, "Energy efficiency optimization-based joint resource allocation and clustering algorithm for M2M communication systems," *IEEE Access*, vol. 7, pp. 168 507–168 519, 2019.

[21] R. Chai, Z. Ma, C. Liu, and Q. Chen, "Service characteristics-oriented joint acb, cell selection, and resource allocation scheme for heterogeneous M2M communication networks," *IEEE Systems Journal*, vol. 13, no. 3, pp. 2641–2652, 2019.

[22] S. A. Alvi, X. Zhou, S. Durrani, and D. T. Ngo, "Sequencing and scheduling for multi-user machine-type communication," *IEEE Transactions on Communications*, vol. 68, no. 4, pp. 2459–2473, 2020.

[23] F. Ghavimi, Y. Lu, and H. Chen, "Uplink scheduling and power allocation for M2M communications in SC-FDMA-based LTE-A networks with QoS guarantees," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6160–6170, July 2017.

[24] L. Miuccio, D. Panno, and S. Riolo, "Joint control of random access and dynamic uplink resource dimensioning for massive MTC in 5G NR based on SCMA," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5042–5063, 2020.

[25] M. Kamel, W. Hamouda, and A. Youssef, "Uplink performance of NOMA-based combined HTC and MTC in ultradense networks," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7319–7333, 2020.

[26] S. Han, X. Xu, Z. Liu, P. Xiao, K. Moessner, X. Tao, and P. Zhang, "Energy-efficient short packet communications for uplink NOMA-based massive MTC networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 066–12 078, 2019.

[27] T. Lv, Y. Ma, J. Zeng, and P. T. Mathiopoulos, "Millimeter-wave NOMA transmission in cellular M2M communications for internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1989–2000, 2018.

[28] K. Deep, K. P. Singh, M. L. Kansal, and C. Mohan, "A real coded genetic algorithm for solving integer and mixed integer optimization problems," *Applied Mathematics and Computation*, vol. 212, no. 2, pp. 505–518, 2009.

[29] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[30] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[31] Y. Fu, L. Salaün, C. W. Sung, and C. S. Chen, "Subcarrier and power allocation for the downlink of multicarrier NOMA systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11 833–11 847, 2018.