Understanding Inter- and Intra-Cluster Concurrent Transmissions for IoT Uplink Traffic in MIMO-NOMA Networks: A DTMC Analysis

Abhishek Kumar, *Graduate Student Member, IEEE*, Jorge Martinez-Bauset, Frank Y. Li, Carmen Florea, *Member, IEEE*, and Octavia A. Dobre, *Fellow, IEEE*

***Abstract*—To enable concurrent transmissions for Internet of things (IoT) traffic in multi-antenna beyond fifth generation networks, non-orthogonal multiple access (NOMA) mechanisms appear as a promising approach. For NOMA-enabled transmissions, IoT devices are grouped into clusters in order to exploit the benefit of concurrent transmissions. However, how to facilitate transmissions from both intra- and inter-cluster is not an easy task and the performance of such concurrent transmissions is so far not well understood from a mathematical point of view, especially when error-prone channel conditions are considered. In this paper, we propose two random access schemes which enable intra- and inter-cluster concurrent transmissions for uplink IoT traffic with and without access control. To assess the performance of such systems, we develop two analytical models based on discrete-time Markov chains (DTMCs) that mimic the behavior of such transmissions. Our models deal with cluster-level performance considering dynamic packet arrivals and the transmissions from devices belonging to the same or different clusters. Through extensive simulations, we validate the accuracy of the analytical models and evaluate the system- and cluster-level performance in terms of throughput and delay under various traffic load conditions and network configurations.**

***Index Terms*—Massive Internet of things, uplink traffic, intra- and inter-cluster concurrent transmission, two-dimensional discrete-time Markov chain, performance evaluation.**

I. Introduction

WHILE the fifth generation communication systems are being deployed worldwide intensively, there is a surge of research efforts towards the six generation communication systems. It is expected that these systems will support ultra-high connectivity, and that massive Internet of things (mIoT) or massive machine-type communications (mMTC) will play a greater role than in the fifth generation era. According to [2] [3], the estimated device density for mMTC/mIoT applications will increase from 10^6 devices/km^2 to 10^7 devices/km^2 in years to come. As such, it is imperative to design effective medium access control mechanisms for effective and fair radio resource allocation.

For mobile systems from the first to the fifth generation, orthogonal multiple access (OMA) based medium access mechanisms – where a dedicated amount of radio (time and frequency) resources is allocated to each device – have been a dominant solution. For future systems, non-orthogonal multiple access (NOMA) mechanisms which allow concurrent transmissions of multiple users based on the same radio resource represent a popular trend, deserving further exploration. NOMA mechanisms deploy successive interference cancellation (SIC) so that it might be able to retrieve one or more of the received signals successfully when multiple concurrent transmissions arrive to a receiver. In contrast, when an OMA mechanism is employed, all concurrent transmissions using the same radio resource will be lost at the receiver.

Empowered with a more advanced resource sharing capability, NOMA outperforms OMA in many cases, achieving higher accumulative system capacity [2]. However, to obtain any benefit for a NOMA scheme, the received signal strength difference has to be sufficiently large [4]. As the number of users in a system increases, the complexity of user pairing algorithms becomes prohibitively high and the advantage of NOMA mechanisms diminishes. As pointed out in [3], it is unlikely that downlink NOMA will be adopted in mIoT/mMTC applications. On the other hand, uplink NOMA mechanisms exhibit great potential to be considered in the near future mIoT/mMTC applications. For example, the 3rd generation partnership project (3GPP) intends to develop a grant-free uplink transmission scheme which may support resource sharing for mIoT/mMTC traffic [5].

To accommodate massive transmissions for mIoT traffic, beamforming through multiple antennas deployed at base station (BS) appears as a powerful technique. By beamforming, a beam of signals is directed to a group of devices with the same or a similar angle towards the BS [6]. In this way, the number of devices covered by each beam will be greatly reduced since distinct beams are assigned to different groups of devices. However, to enable NOMA transmissions for devices covered

Manuscript received June 19, 2023; revised August 23, 2023; accepted December 5, 2023. The research leading to these results has received funding from the European Economic Area (EEA) Norway (NO) Grants 2014-2021, under project contract no. 42/2021, RO-NO-2019-0499 - "A Massive MIMO Enabled IoT Platform with Networking Slicing for Beyond 5G IoV/V2X and Maritime Services (SOLID-B5G)". The work of Jorge Martinez-Bauset was also supported by Grants PGC2018-094151B-I00 and PID2021-123168NB-I00, funded by MCIN/AEI, Spain/10.13039/501100011033 and the European Union A way of making Europe/ERDF, and Grant TED2021-131387B-I00, funded by MCIN/AEI, Spain /10.13039/501100011033 and the European Union NextGenerationEU/RTRP. A part of this work has been presented at the IEEE International Conference on Communications (ICC), May 2023 [1]. (Corresponding author: Frank Y. Li.)

Abhishek Kumar and Frank Y. Li are with the Department of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway (email: {abhishek.kumar; frank.li}@uia.no).

Jorge Martinez-Bauset is with the Departamento de Comunicaciones, Universitat Politècnica de València (UPV), 46022 València, Spain (email: jmartinez@upv.es).

Carmen Florea is with the Department of Telecommunications, National University of Science and Technology Politehnica Bucharest (UNSTPB), 061071 Bucharest, Romania (email: carmen.voicu@upb.ro).

Octavia A. Dobre is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1B 3X9, Canada (e-mail: odobre@mun.ca).

by the same beam is not an easy task as both intra-cluster and inter-cluster interference[1] exists. Furthermore, how to mathematically model the behavior of such transmissions is largely an un-chartered topic.

A. Related Work

1) NOMA transmission schemes for uplink traffic: Uplink IoT traffic is characterized by small data in terms of packet size and traffic intensity as well as by uncoordinated transmissions [7]. As such, access mechanisms following the *grant-free* principle appear as a better fit [8]. In the literature, grant-free NOMA schemes have been proposed considering various use cases and network conditions for both downlink and uplink traffic (see [9] and the references therein). In addition, more schemes have been emerging recently. For instance, packet combining including chase and incremental redundancy combining was proposed in [10] for uplink grant-free NOMA transmissions. In [11], a NOMA-based priority access scheme which considers emergency over regular devices was proposed. Moreover, NOMA coded slotted ALOHA-based transmission for sixth generation enabled IoT that does not rely on power control was proposed in [12]. In another study which focused on the analysis of data error rate, device grouping for data transmissions was determined based on the transmission status of uplink traffic from active devices [13]. In our earlier work [14], a grant-free scheme with dynamic slot allocation was explored, by considering the co-existence of two types of traffic with different service requirements. Another recent study considered IoT devices with and without energy harvesting and proposed two slotted ALOHA protocols for uplink traffic supported by an optimal decoding order and joint decoding [15]. Furthermore, an uncoordinated multiple access scheme which performs inter- and intra-slot SIC-based packet decoding through iterative collision resolution was proposed in [16] and an optimization problem was formulated therein for efficiency maximization.

2) Clustering and NOMA uplink transmissions: Clustering deals with how to associate devices to suitable clusters, in order to exploit the benefits of NOMA. In [17], an access scheme which optimizes user clustering, power and resource allocation in a hybrid NOMA-OMA system was proposed, however, for downlink traffic. To share available radio resources, signature-based access schemes performed in a non-orthogonal manner may apply [19]. Targeting at massive connectivity for narrowband IoT traffic, [18] investigated power domain NOMA-based user clustering and formulated an optimization problem for throughput maximization. For uplink transmissions in a multiple-input multiple-output (MIMO) network, a strong-weak user pairing scheme which depends on channel path loss was proposed in [20]. In [21], expressions of rate coverage were derived for uplink transmission in NOMA-enabled large-scale cellular networks considering both intra- and inter-cluster interference. Furthermore, optimal user pairing for uplink traffic considering both single- and multi-antennas with fixed transmit power for devices has been investigated in [22]. More recently, user clustering and power control for uplink traffic were jointly studied in [23] as an optimization problem when intra- and inter-cluster interference was taken into account.

3) Modeling of NOMA uplink transmissions: To assess the performance of NOMA access schemes, various tools, mostly theoretical from an information theory or optimization perspective, have been applied. However, Markov modeling based mathematical approaches are comparatively less investigated. In [24], the authors proposed a NOMA and hybrid automatic repeat request (HARQ) transmission scheme for both coordinated and uncoordinated transmissions, and then developed a Markov model to evaluate the performance of the NOMA-HARQ scheme. The scheme was further extended in [25] to allow a low-power user to postpone its transmission when a negative acknowledgment is received and a Markov model was developed therein. In [26], a random NOMA scheme with cross-slot SIC was proposed where the packet recovery process was modeled as a Markov process. However, *so far no Markov model that reveals the behavior of both intra- and inter-cluster concurrent transmissions for uplink traffic in MIMO-NOMA networks exists*.

In Table I, we present a comprehensive comparison of our work versus a representative subset of state-of-the-art studies closely related to this topic. The comparison covers various aspects of uplink concurrent transmissions for IoT traffic in MIMO-NOMA networks, spreading from network scenario, traffic pattern, access control, SIC decoding threshold, to tools for mathematical analysis.

B. Contributions

Despite a huge amount of efforts on NOMA transmissions, user clustering under a single beam for uplink transmissions in MIMO-NOMA systems has not been studied in depth [1]. From the traffic perspective, little attention has been paid for concurrent transmissions when both inter- and intra-cluster interference exists, especially when error-prone channel conditions are considered.

In this paper, we investigate multi-antenna based power-domain NOMA systems for uplink IoT traffic, focusing on concurrent transmissions from devices distributed in different clusters which are served by a single beam. In the envisaged network scenario, dynamic packet arrivals and transmissions from devices belonging to the same or different clusters occur. We propose two random access schemes, referred to as Scheme I and Scheme II respectively, for uplink data transmissions that consider both inter- and intra-cluster interference as well as error-prone channel conditions. Following Scheme I, devices transmit their packets as long as they have a packet to transmit. Following Scheme II, device transmission is constrained by access control that the BS imposes to all the devices.

To evaluate the performance of such a system, we develop analytical models that track the evolution of the number of active devices at each cluster over time. Through extensive simulations, we validate the accuracy of the developed analytical models and demonstrate the performance advantage of

[1]While intra-cluster interference means the interference generated by concurrent transmissions from other devices in the same cluster, inter-cluster interference is caused by the transmissions from any other cluster(s).

TABLE I. Concurrent transmission schemes for IoT traffic in MIMO-NOMA networks: A qualitative comparison of our work versus a representative subset of the state-of-the-art studies

Solution versus Features	Our schemes	[15]	[23]	[24]	[13]	[18]
Network scenario	Multiple clusters under single beam	Multiple devices under single beam	Multiple clusters under single beam	Multiple devices under single beam	Multiple devices under single beam	Multiple clusters under multiple beams
Transmission direction	Uplink only	Uplink only	Uplink only	Uplink only	Uplink and downlink	Uplink only
Traffic pattern	Dynamic	Static	Static	Dynamic	Static	Static
Access control mechanism	No (Scheme I); Yes (Scheme II)	No	No	No	No	No
Power control	No	No	Yes	Yes	No	Yes
Error-prone channel combined with DTMC	Yes	No	No	Yes	No	No
SINR difference threshold for SIC decoding	10 dB (default); 2 and 6 dB for comparison	Gradient	Gradient	Gradient	Gradient	Gradient
Sources of interference	Inter- and intra-cluster	Inter-device	Inter- and intra-cluster	Inter-device	Inter-device	Inter-cluster
Mathematical analysis	DTMC	Probability and algebra	Probability and algebra	Probability, algebra, and Markov modeling	Probability and algebra	Probability and algebra

‡Further explanations of this table: *Power control* indicates whether devices transmit at a fixed power level or not; *Gradient* means that SIC decoding is regarded as successful even though the SINR difference is just slightly higher than 1 (0 dB) and less than 2 (3 dB). We argue, however, that these marginal thresholds are impractical for service provisioning in real-world systems; *Sources of interference* identify the transmissions of other devices that are considered as interference to the ongoing transmission of a given device. These concurrent transmissions are originated by devices located inside the same cluster and also from devices associated with other clusters.

the proposed access schemes in terms of cluster and system throughput, and access delay.

In brief, the novelty and main contributions of this paper are summarized as follows:

- Two random access schemes for uplink IoT traffic in error-prone MIMO-NOMA networks have been developed where the success of a transmission depends on both inter- and intra-cluster interference as well as channel conditions. In the studied network, data transmissions are time slotted and time slots are grouped into frames. The proposed schemes enable concurrent packet transmissions from different clusters in the same frame.Moreover, the designed access schemes can be configured to achieve either maximal system throughput or inter-cluster throughput fairness.
- Two novel DTMCs that model the operation of two random access schemes have been developed. As it will be shown for Scheme I, at each system state, the transition probabilities for one cluster are independent from the access behavior of the devices in other clusters. Then, the system transition probabilities adopt a product-form. For Scheme II, the state transition probabilities for one cluster depend on the access behavior of the devices of other clusters.
- The preciseness of the developed models has been validated through extensive discrete-event based simulations. The performance of the schemes at both cluster and system levels has been assessed under various network configurations and traffic conditions in terms of throughput and access delay. Our results also shed light on the effects of multiple factors on NOMA performance, including inter- and intra-interference, number of antennas, and SIC threshold.

In a nutshell, the uniqueness of this paper is represented by a combination of two random access scheme design that considers channel condition, inter-/intra-cluster interference, dynamic traffic conditions, and Markov modeling. With this approach, we achieve insightful understanding of cluster level behavior for uplink NOMA-enabled transmissions. To the best of our knowledge, the models developed in this paper are the first DTMC-based analytical models that mirror the behavior of NOMA uplink transmissions considering both inter- and intra-cluster interference in error-prone environments as well as transmission dependence from different clusters.

The remainder of the paper is organized as follows. The envisaged network scenario and the assumptions for this study are presented in Section II. Then the proposed random access schemes are explained in Section III. To evaluate the performance of the access schemes, two DTMC-based analytical models are developed in Section IV based on which the performance metrics are defined in Section V. In Section VI, extensive simulations are performed to validate the accuracy of the models and to assess the performance of the schemes with various network configurations and traffic load conditions. Finally, the paper is concluded in Section VII.

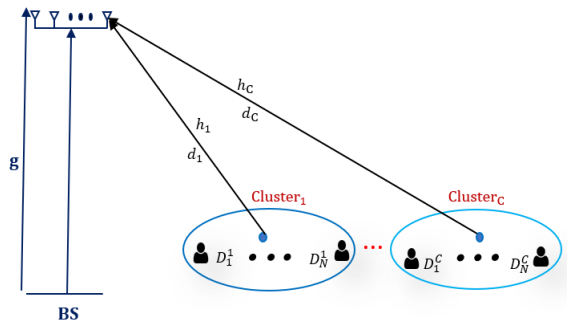

Fig. 1. Illustration of cluster-based concurrent uplink transmissions in a NOMA-MIMO network where D_i^j stands for the i^{th}-device in cluster j with C clusters and N devices per cluster in total.

II. Network Scenario and Assumptions

In this section, we present the network scenario and the assumptions made for scheme design and model development.

A. Network Topology

Consider that a BS equipped with multiple antennas provides services to a number of IoT devices. All the devices in this network are battery-powered, each equipped with a single antenna. The network topology remains static and devices are deployed across the region of interest, grouped into clusters, and positioned at specific geographic locations. Although multiple beams may be formed by the BS, we concentrate on uplink transmissions for devices covered by the same antenna beam. Furthermore, the number of antennas in this network is comparatively small and it does not satisfy the condition that characterizes a massive MIMO (mMIMO) setup where each device has a separate mMIMO beamforming vector [27]. However, developing algorithms for device clustering or beamforming is beyond the scope of this paper.

Although multiple devices may be deployed in a network, not all of them are always active. *A device is regarded as active when it has one packet to transmit.* For each transmission, the same amount of transmit power is allocated to all the devices in our network. Given the simplicity of the IoT devices, no automatic power control is introduced in our network [28].

Whether a transmission is successful or not depends on the co-existence of other concurrent transmission(s) as well as channel condition. Following the power-domain NOMA principle, when two or more concurrent transmissions occur, one or more of them may survive when their received signal-to-interference-plus-noise ratio (SINR) differences are greater than a given threshold.

B. Signal Transmission

As illustrated in Fig. 1, there are M antennas mounted at the BS, covering multiple devices grouped into clusters. The antennas are installed at a height of g meters above the ground. Consider that the number of clusters in the network is C. Cluster j contains $N_j = N$ devices, the center of cluster j is located d_j meters away from the BS, and a cluster with a lower index number is closer to the BS. Devices belonging to the same cluster are distributed randomly within a certain radius from the cluster center. The cluster location and device identities are known to the BS.

By allowing concurrent transmissions from multiple devices, the total received signal ψ at the BS including the transmissions from the N devices in each of the C clusters is obtained by

$$\psi = \sum_{j=1}^{C} \sum_{i=1}^{N} \mathbf{H}_i^j x_i^j + n_a, \tag{1}$$

where x_i^j, $\mathbf{H}_i^j$, and n_a stand for the transmitted signal by the i^{th}-device in cluster j, the complex channel gain vector between the BS and that device, and the additive noise, respectively. Furthermore, x_i^j is expressed as follows:

$$x_i^j = \sqrt{P_t} s_i^j, \tag{2}$$

where P_t refers to the transmit power for each device which is identical for all the devices and s_i^j is the signal to be transmitted from the i^{th}-device in cluster j. Assuming a Gaussian distribution with zero mean and variance σ^2 for noise, we have $n_a \sim \mathcal{CN}\left(0, \sigma^2\right)$. Furthermore, we assume that the complex channel gain vector, denoted by $\mathbf{H}_i^j \sim \mathcal{CN}\left(\mathbf{0}_M, \mathbf{I}_M\right)$, follows Rayleigh fading with zero-mean complex Gaussian distribution.

C. Packet Reception

Whether the transmission of a device is successful or not depends on the channel condition as well as the interference level generated by the other concurrent transmission(s).

1) Channel condition: The wireless channel considered in this study is error-prone. By error-prone, it is meant that there is no guarantee that a packet will be successfully received even if it is the only ongoing transmission over the channel.

More specifically, the channel gain between a device and the BS is decided by the sum of path loss and channel fading. In this study, the path loss is calculated as $PL = 128 + 37.6 \log_{10} d$, where d is the distance between a device and the BS (in kilometers). To quantify fading, we deploy the Rayleigh model with a standard deviation of 8 dB. Furthermore, the noise power spectral density is -174 dBm/Hz. Whether a single transmission is successful or not depends on the received signal strength at the BS. If it is stronger than -104 dBm, the transmission is regarded as successful, otherwise failed.

2) Concurrent transmissions: When two or more concurrent transmissions occur, the BS may be able to recover one or more of the original signals by deploying SIC.

Here it is worth mentioning that, different from many reference studies which calculate the sum rate for NOMA transmissions based on a marginal gradient threshold between signal and interference, the threshold for a successful transmission considered in this study is significantly large, based on experiments obtained from real-life [29].

More specifically, for each radio resource, the BS determines the SINR of each received signal, considering the signals from all other concurrent transmissions as interference. The signal with the strongest SINR will be decoded first,

provided that its SINR is above a pre-configured threshold[2]. Then, the decoded signal is subtracted from the received signals, and the decoding process repeats with one less signal. The decoding process terminates when the highest SINR of the remaining signals is below the threshold [1].

III. Enabling Concurrent Transmissions via Two Random Access Schemes

Data transmission in the envisaged network scenario is time slotted and time slots are grouped into frames with identical frame length. At the beginning of a frame, active devices from the same or different clusters independently select a time slot with equal probability to transmit their packets.

A. Introduction to Access Control

When the amount of radio resources dedicated to the access of IoT services is small and the number of devices is large, access congestion may occur, resulting in network failure and service degradation. To resolve this issue, various control methods based on the principle of access class barring have been proposed [32] [33]. These methods aim at increasing the successful probability of an access attempt by randomly delaying access requests of devices based on a barring rate and a barring time. These access control parameters are periodically broadcasted by a BS.

In this study, we design and implement two grant-free based random access schemes that follow the same principle of access class barring. The schemes aim at orchestrating the operation of devices for uplink radio resource access such that collisions are minimized.

The proposed access schemes follow the *immediate first transmission* principle [14], and an active device decides whether to transmit or not at the begining of a frame based on the broadcasted barring rate, that we refer to as the *access probability*.

B. Random Access Schemes

Let r be the most recent access probability broadcasted by the BS. Two random access schemes are defined as follows:

- Scheme I: An active device will always transmit a packet at the beginning of the current frame, regardless of channel status or potential collision with other possible concurrent transmission(s). In Scheme I the access control is disabled, or $r = 1$.
- Scheme II: At the beginning of each frame, an active device will access with probability r, or it will defer its access attempt with probability $1-r$ until the next frame. The procedure is repeated at every frame *in a memoryless fashion*, but using the most recent value of r. In Scheme II the access control is enabled, or $0 \leq r < 1$.

It is assumed that at each frame active devices are capable of generating a random variate that is uniformly distributed within the range of $[0, 1]$. Within each frame, only one packet will be transmitted per active device. Before a packet transmission, devices do not perform carrier sensing, nor do they check or care about the status of other devices in the same network.

For each device, a packet that arrives during the current frame will be stored in its buffer, if sufficient free memory is available. After a successful packet transmission, if the device remains active, it will contend in the next frame to transmit its next packet in the queue. Otherwise, it will contend to retransmit the same packet in the next frame. For both schemes, the newly activated and backlogged devices have the same access behavior at the beginning of each new frame [30].

For Scheme II, we further develop two implementations for random access control:

– Scheme II-A: The BS broadcasts one unique value of r that applies to all clusters;
– Scheme II-B: The BS broadcasts a specific value of r for each cluster.

Scheme II-B is designed considering that devices located closer to the BS experience less path loss. Accordingly, the BS may assign a higher access probability to devices that are farther away, with the purpose of achieving fairer access among devices belonging to different clusters. On the other hand, the BS may also give higher access probability (priority) to a given cluster due to quality-of-service (QoS) requirements, for instance.

IV. Analytical Models for Random Access

In this section, we develop two DTMCs that model the evolution of the system state over time for the two random access schemes, respectively.

The system state is defined by a vector that describes the number of active devices in each cluster. For the sake of simplicity, we initially consider two clusters, with two devices per cluster, both clusters covered by the same beam in a network where devices share a single radio resource for medium access. In what follows, we first elaborate the state transition probabilities of the DTMCs for this simple case, and later we generalize the models to any number of devices per cluster, any number of clusters, and any number of radio resources.

A. Model Assumptions

In addition to the assumptions mentioned in Section II, the following assumptions are made for our model development:

- The queue size is finite. Similar to the model we developed in [31], the proposed model deals with the packet at the head of the queue.
- During each frame, packets independently arrive to devices, following a Bernoulli distribution with parameter a_c where $0 \leq a_c \leq 1$ for cluster c.
- If one arrival and one departure events occur at the same frame, we consider that the packet at the head of the queue is transmitted first, and then the newly arrived packet is buffered in the queue.
- A failed packet is retransmitted until it is successfully received. The number of retransmissions allowed is unlimited.

[2]According to real-life experiments performed in [29], this SINR difference threshold, denoted by β should be $\beta \geq 10$ dB.

B. Scheme I: State and Transition Probabilities

For the simple scenario with two clusters, two devices per cluster and one resource unit to contend per frame, let $\boldsymbol{n} = (n_1, n_2)$ represent the system state, where n_c is the number of active devices in cluster c. Denote by $S_s^c(\boldsymbol{n})$ the probability that $s = 0, 1, \ldots, n_c$ packets are successfully transmitted from cluster c in state $\boldsymbol{n}$. For simplicity and when it does not lead to ambiguity, we use the compact notation S_s to represent $S_s^c(\boldsymbol{n})$, and a to represent a_c. Note that the successful transmission probabilities $S_s^c(\boldsymbol{n})$ only depend on the physical layer conditions (to be further elaborated in Subsection VI-C).

1) State transitions in a single cluster: The state transition probabilities are defined in the transition matrix $\boldsymbol{P}$,

$$\begin{aligned} P(0,:) &= \left[(1-a)^2,\, 2(1-a)a,\, a^2\right], \\ P(1,:) &= \Big[S_1(1-a)^2,\, S_0(1-a) + 2S_1 a(1-a), \\ &\qquad S_0 a + S_1 a^2\Big], \\ P(2,:) &= \Big[S_2(1-a)^2,\, S_1(1-a) + 2S_2 a(1-a), \\ &\qquad S_0 + S_1 a + S_2 a^2\Big], \end{aligned} \tag{3}$$

where $P(n,:) = [P_{n0}, P_{n1}, P_{n2},]$ is row n of matrix $\boldsymbol{P}$, and P_{nm} is the probability that the cluster transits from n active devices to m, $m = 0, 1, 2$, active ones. It can be easily shown that this transition matrix is stochastic, i.e., the sum of row elements is equal to 1.

As an example, Table II shows how the transition probability P_{11} from frame t to frame $t+1$ is determined. Let $k\ (i,j)$ denote the cluster state, where k is the total number of active devices in the cluster, $i, j = \{0, 1\}$ is the state of each of the 2 devices, where '0' represents an inactive device and '1' an active one, and $k = i + j$.

TABLE II. Computation of P_{11}

t	$t+1$	Transition probability
1 $(0,1)$	1 $(0,1)$	$S_0\,(1-a) + S_1\, 2a\,(1-a)$
	1 $(1,0)$	
1 $(1,0)$	1 $(0,1)$	
	1 $(1,0)$	

A transition is possible when one the following two events occur: i) the packet transmitted by the active device fails (with probability S_0), and the inactive device does not become active (with probability $1-a$); ii) the packet transmitted by the active device is successful (with probability S_1), and one of the two devices receives a packet (with probability a) while the other device does not receive a packet (with probability $1-a$).

2) State transitions in the entire system: For a given state $\boldsymbol{n}$, the operation of each cluster is independent from the operation of the other. Then, the system transition probability from state $\boldsymbol{n} = (n_1, n_2)$ to $\boldsymbol{m} = (m_1, m_2)$ can be expressed by a product-form as,

$$P_{\boldsymbol{nm}} = P^1_{n_1 m_1}(\boldsymbol{n}) \cdot P^2_{n_2 m_2}(\boldsymbol{n}), \tag{4}$$

where $P^c_{n_c m_c}(\boldsymbol{n})$ is the transition probability of cluster c from state n_c to state m_c when the system is in state $\boldsymbol{n}$.

As an example, the system transition probability from state $\boldsymbol{n} = (1,0)$ (i.e., 1 active device in cluster 1 and 0 in cluster 2 in frame t, respectively) to $\boldsymbol{m} = (2,1)$ (2 active devices in cluster 1 and 1 in cluster 2 in frame $t+1$, respectively) is obtained as,

$$\begin{aligned} P^1_{12} &= S^1_0(1,0)\, a_1 + S^1_1(1,0)\, a_1^2, \\ P^2_{01} &= 2(1-a_2)\, a_2, \\ P_{10,21} &= P^1_{12} \cdot P^2_{01}. \end{aligned} \tag{5}$$

3) General expressions: We further extend the model for Scheme I to any number of devices per cluster, any number of clusters, and any number of radio resources. Recall that we refer to $\boldsymbol{n} = (n_1, \ldots, n_C)$ and $\boldsymbol{m} = (m_1, \ldots, m_C)$ as the system state at frames t and $t+1$, respectively, where C is the number of clusters. Let $\boldsymbol{v}_c(n_c) = (v_{c1}, \ldots, v_{cR})$ be an allocation of transmitting devices from cluster c to each of the R radio resources, such that $\sum_{k=1}^{R} v_{ck} = n_c$. Denote by $\boldsymbol{V}^c(n_c) = \{\boldsymbol{v}_c(n_c)\}$ the set of all allocations of transmitting devices from cluster c to the R radio resources and by $\boldsymbol{V}(\boldsymbol{n}) = \left\{\boldsymbol{V}^1(n_1), \ldots, \boldsymbol{V}^C(n_C)\right\}$ the set of all possible allocations of the $\boldsymbol{n}$ transmitters to the R radio resources.

We define $\boldsymbol{w}_k = (v_{1k}, \ldots, v_{Ck})$ as the number of transmitting devices from each cluster using radio resource k, $k = 1, \ldots, R$, where v_{ck} are the devices transmitting in radio resource k from cluster c. Let $\widehat{\boldsymbol{S}}^c_k(\boldsymbol{w}_k)$ represent the probability distribution of successful transmissions from cluster c in radio resource k when $\boldsymbol{w}_k$ devices transmit.

Denote by $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_R]$ one of the possible allocations of the $\boldsymbol{n}$ transmitters to each of the R radio resources and by $\widetilde{\boldsymbol{S}}^c(\boldsymbol{W})$ the probability distribution of the total number of successful transmissions from cluster c, when the allocation of transmitters to radio resources is $\boldsymbol{W}$. Element $\widetilde{S}^c_s(\boldsymbol{W})$ of $\widetilde{\boldsymbol{S}}^c(\boldsymbol{W})$ is the probability that cluster c has s successful transmissions, $s = 0, \ldots, n_c$. Note that $\widetilde{\boldsymbol{S}}^c(\boldsymbol{W})$ can be obtained by the convolution of all $\widehat{\boldsymbol{S}}^c_k(\boldsymbol{w}_k)$ distributions[3]. Then, $\widetilde{\boldsymbol{S}}^c(\boldsymbol{W}) = \widehat{\boldsymbol{S}}^c_1(\boldsymbol{w}_1) \otimes \ldots \otimes \widehat{\boldsymbol{S}}^c_R(\boldsymbol{w}_R)$. Also, $\widetilde{\boldsymbol{S}}^c(\boldsymbol{W})$ might need to be truncated to the first $n_c + 1$ elements and re-normalized.

Let us define,

$$\alpha^c_{s_c}(n_c, m_c) = \binom{N_c - n_c + s_c}{m_c - n_c + s_c} a_c^{m_c - n_c + s_c} (1-a_c)^{N_c - m_c},$$

where $\alpha^c_{s_c}(n_c, m_c)$ is the probability that cluster c transits from state n_c to m_c conditioned on having s_c successful packet transmissions, $s_c \geq \max(0, n_c - m_c)$.

Then, the transition probability for cluster c from state $\boldsymbol{n}$ to

[3]Consider two independent discrete random variables, Y_1 and Y_2, with distributions $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ respectively, where $y_n(k) = P(Y_n = k)$, $n = 1, 2$, $k \geq 0$. Define $Z = Y_1 + Y_2$. The distribution $\boldsymbol{z}$ of Z can be obtained by the convolution of the distributions of Y_1 and Y_2, $\boldsymbol{z} = \boldsymbol{y}_1 \otimes \boldsymbol{y}_2$, where $\otimes$ is the convolution operator. Element $z(i)$ can be obtained as, $z(i) = \sum_{k=0}^{i} y_1(k) \cdot y_2(i-k)$.

state $\boldsymbol{m}$ can be expressed as,

$$P^c_{n_c m_c} = \sum_{\boldsymbol{V}(\boldsymbol{n})} \widetilde{P}\left(\boldsymbol{v}\left(\boldsymbol{n}\right)\right) \sum_{s_c=\delta_c}^{n_c} \widetilde{S}^c_{s_c}\left(\boldsymbol{W}\right)\ \alpha^c_{s_c}\left(n_c, m_c\right)\ ,$$

$$\widetilde{P}\left(\boldsymbol{v}\left(\boldsymbol{n}\right)\right) = \prod_{c=1}^{C} P\left(\boldsymbol{v}_c\left(n_c\right)\right)\ ,$$

$$P\left(\boldsymbol{v}_c\left(n_c\right)\right) = \frac{n_c!}{v_{c1}! \cdot v_{c2}! \cdot \ldots \cdot v_{cR}!} \left(1/R\right)^{n_c}\ ,$$

where $\delta_c = \max\left(0, n_c - m_c\right)$ is the minimum number of active devices that must transmit from cluster c in the current frame to make the transition possible, and $P\left(\boldsymbol{v}_c\left(u_c\right)\right)$ is the probability that the allocation of transmitters from cluster c to the R radio resources is $\boldsymbol{v}_c\left(n_c\right)$. As observed from $P\left(\boldsymbol{v}_c\left(u_c\right)\right)$, each device selects a radio resource with equal probability. Note that for each element of the set $\boldsymbol{V}\left(\boldsymbol{n}\right)$ the allocation $\boldsymbol{W}$ can be determined. Then,

$$P_{\boldsymbol{nm}} = \prod_{c=1}^{C} P^c_{n_c m_c}\,.$$

As an example, for the case $C = 2$ and $R = 2$, we obtain

$$P^c_{n_c m_c} = \sum_{v_{11}=0}^{n_1} \frac{n_1!\ \left(1/2\right)^{n_1}}{v_{11}!\left(n_1 - v_{11}\right)!} \sum_{v_{21}=0}^{n_2} \frac{n_2!\ \left(1/2\right)^{n_2}}{v_{21}!\left(n_2 - v_{21}\right)!} \sum_{s_c=\delta_c}^{n_c} \widetilde{S}^c_{s_c}\left(\boldsymbol{W}\right)\ \alpha^c_{s_c}\left(n_c, m_c\right)\ ,$$

where $\boldsymbol{w}_1 = (v_{11}, v_{21})$, $\boldsymbol{w}_2 = (v_{12}, v_{22})$, $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2]$, and $\widetilde{\boldsymbol{S}}^c\left(\boldsymbol{W}\right) = \widehat{\boldsymbol{S}}^c_1\left(\boldsymbol{w}_1\right) \otimes \widehat{\boldsymbol{S}}^c_2\left(\boldsymbol{w}_2\right)$.

C. Scheme II: State and Transition Probabilities

When Scheme II is employed, an active device in cluster c might postpone its transmission to the next frame based the outcome of the access probability check. This check is computed using the latest access probability r_c broadcasted by the BS. When an active device postpones its transmission, this event increases the success probability of the packets being concurrently transmitted by other active devices in the same and in other clusters. Clearly, the transition probabilities for one cluster depend on the access behavior of devices in other clusters.

1) State transitions in a single cluster: As an example, consider the system state $\boldsymbol{n} = (1, 2)$, with 1 active device in cluster 1 and 2 active ones in cluster 2, for the simple scenario with two clusters, two devices per cluster and one radio resource. For clarity, let us focus only on cluster 2. Table III illustrates the transition probability P_{21}.

TABLE III. Cluster 2. Computation of P_{21}

t	$t+1$	Transition probability
2 (1, 1)	1 (0, 1)	$r_2^2 S_2^2\left(1,2\right) 2a_2\left(1-a_2\right)$
	1 (1, 0)	$+r_2^2 S_1^2\left(1,2\right)\left(1-a_2\right)$
		$+2r_2\left(1-r_2\right) S_1^2\left(1,1\right)\left(1-a_2\right)$

As shown in this table, a transition is possible when one of the following three events occur: i) both devices access, both transmissions are successful and a packet arrives to one of the two devices; ii) both devices access, only one transmission is successful and no packet arrival occurs; iii) only one of two the devices accesses, its transmission is successful and no packet arrival occurs in the transmitting device.

Note that now $S^c_s\left(i,j\right)$ *does not* refer to the probability that cluster c has s successful transmissions in state $\boldsymbol{n} = (i, j)$, but to the probability that cluster c has s successful transmissions when the numbers of transmitting devices from each cluster are (i, j), respectively.

2) General expressions: For Scheme II, we also extend the model to any number of devices per cluster, any number of clusters, and any number of radio resources. The same as in Scheme I, $\boldsymbol{n} = (n_1, \ldots, n_C)$ and $\boldsymbol{m} = (m_1, \ldots, m_C)$ denote the system state at frames t and $t+1$, respectively, and C the number of clusters. Let $\boldsymbol{u}\left(\boldsymbol{n}\right) = (u_1, \ldots, u_C)$ be a possible selection of the number devices from each cluster that transmit at frame t, when the system is in state $\boldsymbol{n}$. Note that the transmitters are selected after passing the access probability check. Define $\boldsymbol{v}_c\left(u_c\right) = (v_{c1}, \ldots, v_{cR})$ as an allocation of transmitting devices from cluster c to each of the R radio resources, such that $\sum_{k=1}^{R} v_{ck} = u_c$. Let $\boldsymbol{V}^c\left(u_c\right) = \{\boldsymbol{v}_c\left(u_c\right)\}$ be the set of all allocations of transmitting devices from cluster c to the R radio resources and $\boldsymbol{V}\left(\boldsymbol{u}\right) = \left\{\boldsymbol{V}^1\left(u_1\right), \ldots, \boldsymbol{V}^C\left(u_C\right)\right\}$ be the set of all possible allocations of the $\boldsymbol{u}\left(\boldsymbol{n}\right)$ transmitters to the R radio resources.

We define,

$$\theta_c\left(n_c, u_c\right) = \binom{n_c}{u_c} r_c^{u_c}\left(1 - r_c\right)^{n_c - u_c}\ ,\ \ \theta_c\left(0,0\right) = 1\,, \quad (6)$$

$$\widetilde{\theta}\left(\boldsymbol{n}, \boldsymbol{u}\right) = \prod_{c=1}^{C} \theta_c\left(n_c, u_c\right)\ , \quad (7)$$

where $\theta_c\left(n_c, u_c\right)$ is the probability that u_c active devices, out of n_c ones, transmit from cluster c in the current frame.

Consider the set of all possible transmitter selections in state $\boldsymbol{n}$ that meet $u_c \geq \delta_c$, and $u_c \leq n_c$, $\forall c$, where $\delta_c = \max\left(0, n_c - m_c\right)$ is the minimum number of active devices that must transmit from cluster c in the current frame to make the transition possible, and denote it by $\boldsymbol{U}\left(\boldsymbol{\delta}, \boldsymbol{n}\right) = \{\boldsymbol{u}\left(\boldsymbol{n}\right)\}$. $\boldsymbol{S}\left(\boldsymbol{U}\left(\boldsymbol{\delta}, \boldsymbol{n}\right)\right)$ is the set of all possible successful transmissions $\boldsymbol{s} = (s_1, \ldots, s_C)$ for each element of $\boldsymbol{U}\left(\boldsymbol{\delta}, \boldsymbol{n}\right)$.

For each element of $\boldsymbol{U}\left(\boldsymbol{\delta}, \boldsymbol{n}\right)$, the set $\boldsymbol{V}\left(\boldsymbol{u}\right)$ can be determined. For each element of $\boldsymbol{V}\left(\boldsymbol{u}\right)$, each of the possible allocations of transmitters to each of the R radio resources $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_R]$ can be determined. Furthermore, for each allocation $\boldsymbol{W}$, the distribution of successful packet transmissions from cluster c, $\widetilde{\boldsymbol{S}}^c\left(\boldsymbol{W}\right) = \widehat{\boldsymbol{S}}^c_1\left(\boldsymbol{w}_1\right) \otimes \ldots \otimes \widehat{\boldsymbol{S}}^c_R\left(\boldsymbol{w}_R\right)$, can be computed. Given an allocation $\boldsymbol{W}$, we can compute the probability of each element of $\boldsymbol{S}\left(\boldsymbol{U}\left(\boldsymbol{\delta}, \boldsymbol{n}\right)\right)$ as, $S^{\boldsymbol{W}}_{\boldsymbol{s}} = \widetilde{S}^1_{s_1}\left(\boldsymbol{W}\right) \cdot \ldots \cdot \widetilde{S}^C_{s_C}\left(\boldsymbol{W}\right)$, where $\widetilde{S}^c_{s_c}\left(\boldsymbol{W}\right)$ is element s_c of the distribution $\widetilde{\boldsymbol{S}}^c\left(\boldsymbol{W}\right)$.

Then, the transition probability from state $\boldsymbol{n}$ to state $\boldsymbol{m}$ is

given by,

$$P_{\boldsymbol{n},\boldsymbol{m}} = \sum_{\boldsymbol{U}(\boldsymbol{\delta},\boldsymbol{n})} \widetilde{\theta}\left(\boldsymbol{n},\boldsymbol{u}\right) \sum_{\boldsymbol{V}(\boldsymbol{n})} P\left(\boldsymbol{v}\left(\boldsymbol{u}\right)\right) \sum_{\boldsymbol{S}(\boldsymbol{U}(\boldsymbol{\delta},\boldsymbol{n}))} S_{\boldsymbol{s}}^{\boldsymbol{W}}\, \widetilde{\alpha}_{\boldsymbol{s}}\left(\boldsymbol{n},\boldsymbol{m}\right) , \tag{8}$$

$$\widetilde{\alpha}_{\boldsymbol{s}}\left(\boldsymbol{n},\boldsymbol{m}\right) = \prod_{c=1}^{C} \alpha_{s_c}^{c}\left(n_c, m_c\right) . \tag{9}$$

As an example, for the case $C = 2$ and $R = 2$, we obtain

$$P_{\boldsymbol{n},\boldsymbol{m}} = \sum_{u_1=\delta_1}^{n_1} \sum_{u_2=\delta_2}^{n_2} \theta_1\left(n_1,u_1\right)\theta_2\left(n_2,u_2\right) \sum_{v_{11}=0}^{u_1} \frac{u_1!\left(1/2\right)^{u_1}}{v_{11}!\cdot\left(u_1-v_{11}\right)!} \sum_{v_{21}=0}^{u_2} \frac{u_2!\left(1/2\right)^{u_2}}{v_{21}!\cdot\left(u_2-v_{21}\right)!} \sum_{s_1=\delta_1}^{u_1} \sum_{s_2=\delta_2}^{u_2} \widetilde{S}_{s_1}^{1}\left(\boldsymbol{W}\right)\widetilde{S}_{s_2}^{2}\left(\boldsymbol{W}\right)\widetilde{\alpha}_{\boldsymbol{s}}\left(\boldsymbol{n},\boldsymbol{m}\right) , \tag{10}$$

where $v_{12} = u_1 - v_{11}$, $v_{22} = u_2 - v_{21}$, $\boldsymbol{w}_1 = (v_{11}, v_{21})$, $\boldsymbol{w}_2 = (v_{12}, v_{22})$, $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2]$, $\widetilde{\boldsymbol{S}}^{1}\left(\boldsymbol{W}\right) = \widehat{\boldsymbol{S}}_1^{1}\left(\boldsymbol{w}_1\right) \otimes \widehat{\boldsymbol{S}}_2^{1}\left(\boldsymbol{w}_2\right)$, $\widetilde{\boldsymbol{S}}^{2}\left(\boldsymbol{W}\right) = \widehat{\boldsymbol{S}}_1^{2}\left(\boldsymbol{w}_1\right) \otimes \widehat{\boldsymbol{S}}_2^{2}\left(\boldsymbol{w}_2\right)$, and $\widetilde{\alpha}_{\boldsymbol{s}}\left(\boldsymbol{n},\boldsymbol{m}\right) = \alpha_{s_1}^{1}\left(n_1,m_1\right)\alpha_{s_2}^{2}\left(n_2,m_2\right)$.

V. Performance Metrics

In this section, we first define two performance metrics and then derive their expressions based on the developed analytical models presented in the previous section.

A. Performance Metric Definition

For performance evaluation, the following two performance metrics are defined in this study:

- *Cluster throughput* and *system throughput*: While cluster throughput is defined as the average number of packets successfully transmitted by a cluster per frame, system throughput specifies the total number of packets successfully transmitted per frame in the entire network, i.e., including all clusters.
- *Access delay* is defined as the average number of frames it takes for a device to transmit a packet successfully. Note that access delay includes the frames in which a device defers the packet transmission (for Scheme II only) as well as the frames used for retransmissions caused by collision or poor channel condition.

B. Performance Analysis

For the sake of expression simplicity, we derive performance metric expressions for a system with two clusters ($C = 2$) and two radio resources ($R = 2$). Let $\boldsymbol{\pi} = \{\pi_{n_1 n_2}\}$ be the stationary distribution of the system, and $\pi_{n_1 n_2}$ the stationary probability of finding n_1 active devices in cluster 1 and n_2 devices in cluster 2, i.e., the fraction of frames the system is found in state $\boldsymbol{n} = (n_1, n_2)$. The stationary distribution can be obtained by solving the system of linear equations $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, and $\boldsymbol{\pi}\mathbf{1} = 1$, where $\mathbf{P} = [\mathbf{P}_{\boldsymbol{nm}}]$ is the transition probability matrix and $P_{\boldsymbol{nm}}$ is the transition probability from system state $\boldsymbol{n} = (n_1, n_2)$ to state $\boldsymbol{m} = (m_1, m_2)$.

For generality, we derive expressions for a system operating with Scheme II. Note that the same expressions apply to a system with access control disabled (i.e., Scheme I) by configuring $r_c = 1$.

1) Cluster and system throughput: Let us define,

$$\overline{S}^{1}\left(u_1,u_2\right) = \sum_{v_{11}=0}^{u_1}\sum_{v_{21}=0}^{u_2} \widetilde{\varphi}\left(u_1,v_{11},u_2,v_{21}\right) \sum_{s_1=0}^{u_1} s_1\, \widetilde{S}_{s_1}^{1}\left(\boldsymbol{W}\right) , \tag{11}$$

$$\overline{S}^{2}\left(u_1,u_2\right) = \sum_{v_{11}=0}^{u_1}\sum_{v_{21}=0}^{u_2} \widetilde{\varphi}\left(u_1,v_{11},u_2,v_{21}\right) \sum_{s_2=0}^{u_2} s_2\, \widetilde{S}_{s_2}^{2}\left(\boldsymbol{W}\right) , \tag{12}$$

$$\varphi\left(u_c,v_{c1}\right) = \frac{u_c!}{v_{c1}!\cdot\left(u_c-v_{c1}\right)!}\left(1/2\right)^{u_c} , \tag{13}$$

where $\widetilde{\varphi}\left(u_1,v_{11},u_2,v_{21}\right) = \varphi\left(u_1,v_{11}\right)\cdot\varphi\left(u_2,v_{21}\right)$, and $\overline{S}^{c}\left(u_1,u_2\right)$ can be interpreted as the average number of successes for cluster c, for all possible allocations $\boldsymbol{W}$, conditioned on the number of transmitting devices being (u_1, u_2).

Let γ_c be the throughput of cluster c expressed as the average number of packets successfully transmitted per frame, and $\gamma = \gamma_1 + \gamma_2$ the system throughput.

$$\gamma_1 = \sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2} \pi_{n_1 n_2} \sum_{u_1=0}^{n_1}\sum_{u_2=0}^{n_2} \widetilde{\theta}\left(n_1,u_1,n_2,u_2\right)\overline{S}^{1}\left(u_1,u_2\right) , \tag{14}$$

$$\gamma_2 = \sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2} \pi_{n_1 n_2} \sum_{u_1=0}^{n_1}\sum_{u_2=0}^{n_2} \widetilde{\theta}\left(n_1,u_1,n_2,u_2\right)\overline{S}^{2}\left(u_1,u_2\right) , \tag{15}$$

where N_c is the number of devices in cluster c, and $\widetilde{\theta}\left(n_1,u_1,n_2,u_2\right) = \theta_1\left(n_1,u_1\right)\cdot\theta_2\left(n_2,u_2\right)$.

The system throughput, γ_s, for a system with two cluster is determined as,

$$\gamma_s = \gamma_1 + \gamma_2 . \tag{16}$$

2) Access delay: Let D_c be the average delay for packets in cluster c expressed in frames, i.e., the average number of frames since a packet arrives to a device buffer until it is successfully transmitted. Let Q_c be the average number of active devices in cluster c,

$$Q_1 = \sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2} n_1\,\pi_{n_1 n_2} , \quad Q_2 = \sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2} n_2\,\pi_{n_1 n_2} . \tag{17}$$

Then, by Little's Law,

$$D_c = Q_c/\gamma_c . \tag{18}$$

VI. Numerical Results and Discussions

In this section, we first summarize the experimental setup and then validate the accuracy of the developed DTMC-based analytical models by comparing the values of the performance parameters obtained from the analytical models and the ones obtained by simulations. Afterwards, the performance of the proposed access schemes is assessed through various network configurations with different number of devices per cluster and

TABLE IV. Network parameter configurations [1] [29]

Parameter	Value
System type	Single cell
Number of clusters (C)	2
Number of radio resources (R)	1∼3
Number of antennas at the BS (M)	1 or 2
Distance between cluster 1 and the BS (d_1)	150∼450 m
Distance between cluster 2 and the BS (d_2)	900 m
Cluster radius	25 m
Antenna height (g)	30 m
Number of devices per cluster (N)	2 or 3
Path loss model	$128 + 37.6 \log_{10} d$
Standard deviation for shadow fading	8 dB
Transmit power (P_t) (for each device)	23 dBm
Receiver sensitivity	-104 dBm
SINR difference threshold for SIC (β)	10 dB
Noise power spectral density (σ^2)	-174 dBm/Hz

different number of radio resources. Also, a performance comparison between NOMA and OMA is also presented. Finally, we study other aspects that may affect network performance.

A. Experimental Setup

To mimic the behavior of devices according to the principles of the proposed access schemes, we have developed a custom-built discrete-event simulator in MATLAB. Based on this simulator, we perform experiments considering a single-cell multi-antenna network where one BS serves two clusters, both covered by a single beam. As shown in Fig. 1, each cluster is composed of multiple devices. In the basic experiment, all devices of both clusters share a single radio resource ($R = 1$). Unless otherwise stated, Table IV specifies the physical layer parameters used in our simulations. The transmission success probabilities considering channel state as well as SIC are shown in Table V and these values are obtained by simulations.

All the devices in the studied network dynamically generate packets following a Bernoulli distribution with parameter a. In this network, the access of devices to the radio resources is governed by the proposed access schemes, i.e., Scheme I and Scheme II that are presented in Subsection III-B. For Scheme II, we investigate two implementation alternatives of the access probability: i) a unique one, that is stored by all devices in the network; and ii) a cluster specific one, that is stored by all the devices of the same cluster.

The BS dynamically configures the to-be-broadcasted access probabilities for the purpose of achieving a given QoS objective, e.g., throughput maximization. In systems where devices have multiple radio resources for possible access, a transmitting device selects one of the resources with equal probability at each new frame. The packets received by the BS on each radio resource are processed according to the SIC principle explained above. In our experiments, we consider the signals received by the BS over each radio resource as independent. As explained earlier, a packet success probability depends on the number of transmitters from each cluster using the same radio resource, i.e., the inter- and intra-cluster interference, as well as the channel state.

TABLE V. Successful probability distributions (error-prone channel)

State (n_1, n_2)	Cluster 1				Cluster 2			
	S_0	S_1	S_2	S_3	S_0	S_1	S_2	S_3
(0,0)	1	0	0	0	1	0	0	0
(0,1)	1	0	0	0	0.567	0.433	0	0
(0,2)	1	0	0	0	0.477	0.490	0.033	0
(0,3)	1	0	0	0	0.521	0.424	0.055	0
(1,0)	0.163	0.837	0	0	1	0	0	0
(1,1)	0.401	0.599	0	0	0.804	0.196	0	0
(1,2)	0.564	0.436	0	0	0.766	0.227	0.007	0
(1,3)	0.675	0.325	0	0	0.788	0.200	0.012	0
(2,0)	0.606	0.274	0.120	0	1	0	0	0
(2,1)	0.677	0.249	0.074	0	0.941	0.059	0	0
(2,2)	0.734	0.220	0.046	0	0.929	0.070	0.001	0
(2,3)	0.780	0.191	0.029	0	0.935	0.062	0.003	0
(3,0)	0.741	0.167	0.079	0.013	1	0	0	0
(3,1)	0.779	0.164	0.049	0.008	0.984	0.016	0	0
(3,2)	0.810	0.155	0.031	0.004	0.980	0.020	0	0
(3,3)	0.836	0.143	0.019	0.002	0.981	0.019	0	0

To obtain the numerical results presented later on in this section, we have considered different network configurations and various factors that may affect network performance. For each experiment, we configure network parameters, including i) the physical layer parameters defined in Table V; ii) the number devices per cluster; iii) the load; and iv) the access scheme parameters (access probabilities), and assess the performance when one of them varies.

B. Network Configuration and Model Validation

In the studied network, cluster 1 is located closer to the BS than cluster 2. For both clusters, the radius of each cluster is 25 meters, and devices belonging to the same cluster are uniformly distributed across the cluster. For the sake of clarity, we configure the same arrival probability and an identical number of devices per cluster for all clusters. Therefore, $a_1 = a_2 = a$ and $N_1 = N_2 = N$ when $C = 2$. Except the results shown in Subsection VI-H, each cluster in our network contains two or three devices sharing one radio resource. In Subsection VI-H, we further demonstrate the performance for a larger network with $N = 8$ and $R = 3$.

The obtained analytical and simulation results are compared in Figs. 2–7 (for $R = 1$) and Figs. 12–13 (for $R = 2$). From the curves shown in these figures, it is evident that the analytical and simulation results match precisely with each other. As such, the accuracy of the developed analytical models is verified. For the sake of illustration clarity, not all figures shown in the rest of this section include both analytical and simulation results.

C. Successful Probabilities over an Error-Prone Channel

As explained in Subsection II-C, when multiple packet transmissions share the same radio resource in the same frame, the number of successful packet receptions depends on the received signal strength versus interference and noise. A signal is successfully decoded when it is stronger than the sensitivity level and the SIC decoding criterion ($SINR \geq \beta$ dB) is met.

For a system with two clusters, three devices per cluster, and one radio resource, there are 16 system states in total. Table V specifies the successful probabilities for each system

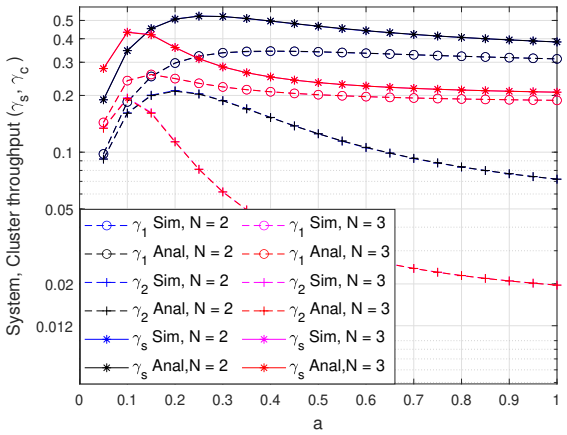

Fig. 2. System, cluster throughput as traffic load a varies: Scheme I.

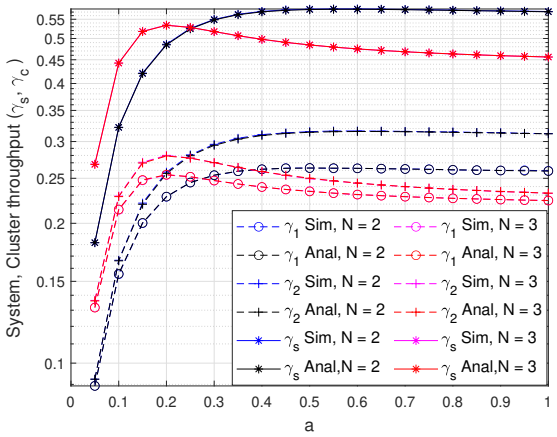

Fig. 4. System, cluster throughput for Scheme II-B: Fair access.

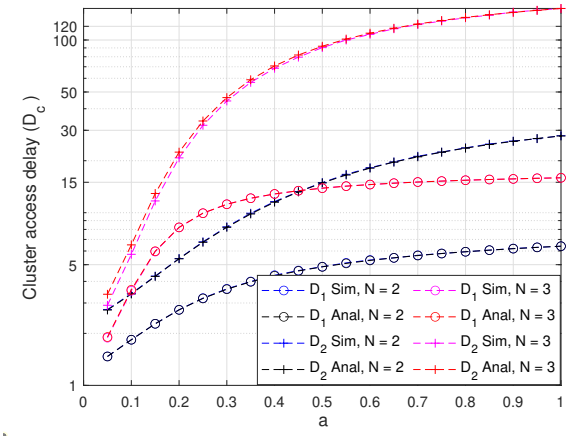

Fig. 3. Access delay for Scheme I as traffic load a varies.

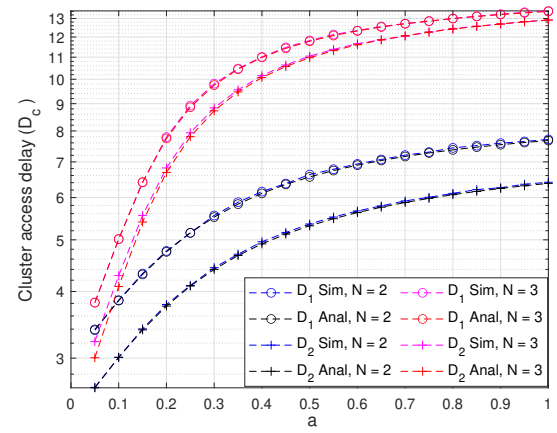

Fig. 5. Access delay for Scheme II-B: Fair access.

state. These values are obtained through extensive simulations when the network is configured as $d_1 = 450$ m, $d_2 = 900$ m, $M = 1$, $r_c = 1$, and $c = 1, 2$, respectively.

To clarify the distribution of the number of successful transmissions, $S_s^c(\boldsymbol{n})$, we take a look at two examples: i) in state (0,1), although there is only one packet transmission from cluster 2, only 43.3% of the transmissions succeeded whereas 56.7% failed. This is because the received signal at the BS is generally weak due to the long distance attenuation from cluster 2; ii) in state (2,1), the probabilities that cluster 1 achieves 0, 1, or 2 successful transmissions are 0.677, 0.249, and 0.074, respectively. On the other hand, the probabilities that cluster 2 achieves 0 or 1 successful transmission are 0.941 and 0.059, respectively.

D. Performance of Scheme I: Throughput and Delay

In this subsection, we present the performance of Scheme I with $N = 2$ or 3 devices per cluster. The other parameters are configured as $d_1 = 450$ m, $d_2 = 900$ m, $M = 1$, $r_1 = r_2 = 1$, $a \in [0.05, 1]$, respectively.

1) System and cluster throughput: Let us first examine the throughput performance of Scheme I as the packet arrival probability, a, increases. As shown in Fig. 2, both cluster and system throughput increases quickly with a when traffic load is light. At $a = 0.25$ and $a = 0.10$, the system throughput reaches its peak value of $\gamma_s = 0.55$ (for $N = 2$ in black) and $\gamma_s = 0.42$ (for $N = 3$ in red) respectively. Then, γ_s decreases monotonically as traffic load further increases. This is because more irresolvable collisions[4] happen when the traffic load becomes heavier.

With respect to the cluster throughput, it is expected that cluster 1 achieves higher throughput than cluster 2. This is because the signals sent from cluster 1 are much stronger than that from cluster 2. When two or more inter-cluster transmissions occur concurrently, the transmission(s) from cluster 1 has/have much better chance to survive than the ones from cluster 2. When comparing the throughput of $N = 2$ versus that of $N = 3$ (shown in Fig. 2), it is generally true that the set of $N = 2$ achieves higher throughput. This is because less interference is generated when there are fewer devices in a cluster.

2) Access delay: In Fig. 3, we illustrate the access delay performance for both clusters as a increases. As expected, a longer delay is experienced with a higher traffic load. In accordance with the cluster throughput performance, a shorter access delay is achieved by cluster 1. This is because packet success probabilities for transmissions from cluster 1 are higher than those from cluster 2, leading to shorter delay. When comparing the access delay performance of $N = 2$

[4]An irresolvable collision means that two or more devices transmit concurrently, but the received signal strength difference is not large enough for SIC decoding. Therefore, both/all transmissions are regarded as failed.

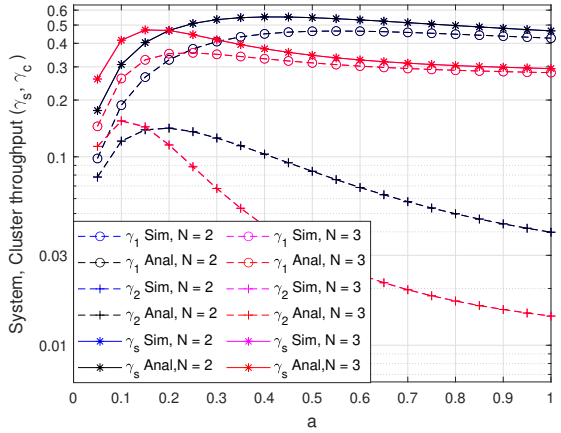

Fig. 6. System, cluster throughput for Scheme II-B: QoS.

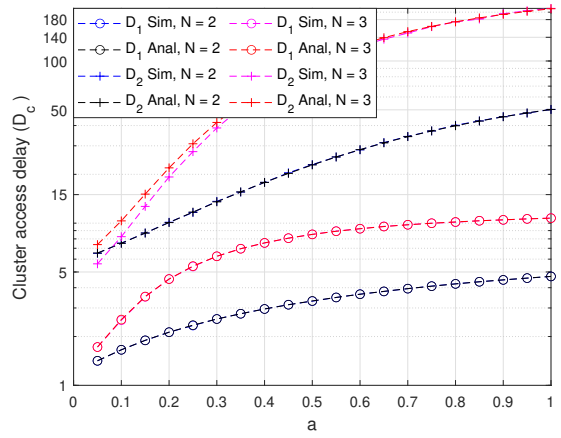

Fig. 7. Access delay for Scheme II-B: QoS.

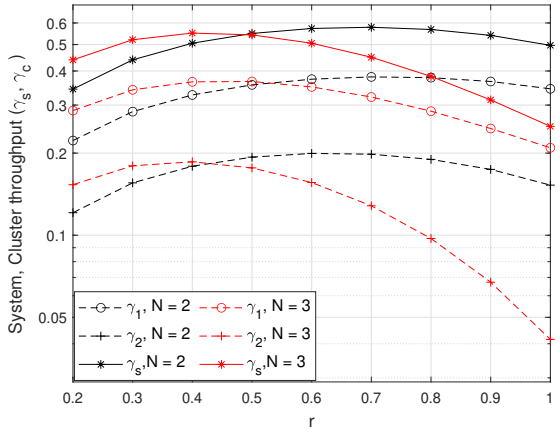

Fig. 8. Scheme II-B: The effect of access control on throughput.

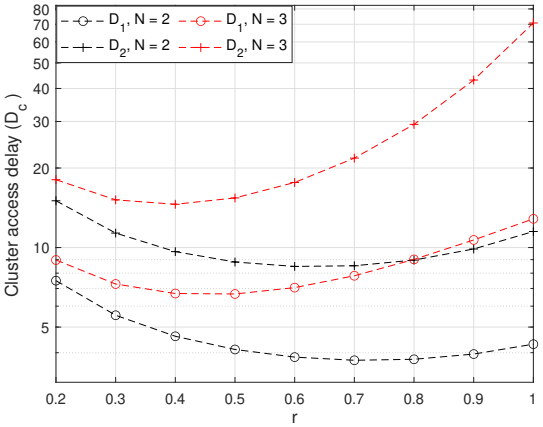

Fig. 9. Scheme II-B: The effect of access control on delay.

versus of $N = 3$, it is evident that much shorter delay is achieved with fewer number of devices in each cluster. This is due to the fact that more retransmissions are needed for a successful transmission, as many collisions are irresolvable in the $N = 3$ case.

E. Performance of Scheme II: Throughput and Delay

As mentioned earlier, there are two implementations of Scheme II, identical access probability for all clusters (II-A) and individual probability for each cluster (II-B). Unless otherwise stated, the network configuration is the same as presented in Subsection VI-D.

1) System and cluster throughput (with higher priority to cluster 2): By configuring $r_1 = 0.4$ and $r_2 = 1$, more access restrictions are imposed to cluster 1. As shown in Fig. 4, better fairness has been achieved as the difference between the throughput achieved by these two clusters is much smaller in comparison with what is achieved in Scheme I. The reason for this behavior is that cluster 2 devices obtain more opportunities for transmission with this configuration, coupled with the fact that cluster 1 devices get less opportunities. When comparing the performance of $N = 2$ versus of $N = 3$, higher throughput is achieved with $N = 2$ as less interference is experienced with fewer number of devices per cluster. In the meantime, the gap between cluster throughput is even narrower with $N = 3$. Observe that with $N = 3$ the throughput reduction is larger for cluster 2 than for cluster 1. Clearly, with $N = 3$, cluster 2 transmissions suffer more acutely from irresolvable collisions than cluster 1 transmissions.

2) System and cluster throughput (with higher priority to cluster 1): As mentioned in Subsection III-B, higher priority may also be given to cluster 1 for example due to QoS considerations. Fig. 6 reveals the system and cluster throughput when access control is configured as $r_1 = 1$ and $r_2 = 0.4$. As shown in this figure, it is clear that much higher cluster throughput has been achieved for cluster 1. With a higher traffic load, the difference between them becomes larger. When the traffic load is very heavy (approaching $a = 1$), cluster 2 throughput is paltry, transmitting successfully merely 4 and 1.5 packets per 100 frames for $N = 2$ and $N = 3$, respectively.

3) Access delay: The obtained access delays for Scheme II-B with preference to cluster 2 (with $r_1 = 0.4$ and $r_2 = 1$) is illustrated in Fig. 5. With a constrain on cluster 1 devices, the access delays from both clusters are quite close to each other, revealing that fair access is also achieved with respect to access delay. With a more precise configuration of r_1, identical access delay could be achieved.

When higher priority is given to cluster 1 devices (with $r_1 = 1$ and $r_2 = 0.4$), much shorter access delay is obtained for cluster 1 devices (as shown in Fig. 7). The reason for this behavior is that with $r_2 = 0.4$, cluster 2 devices will access the radio resource less frequently than cluster 1 devices, leading to

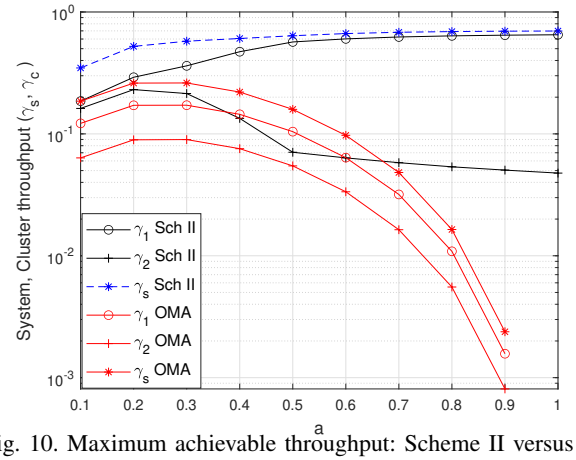

Fig. 10. Maximum achievable throughput: Scheme II versus OMA.

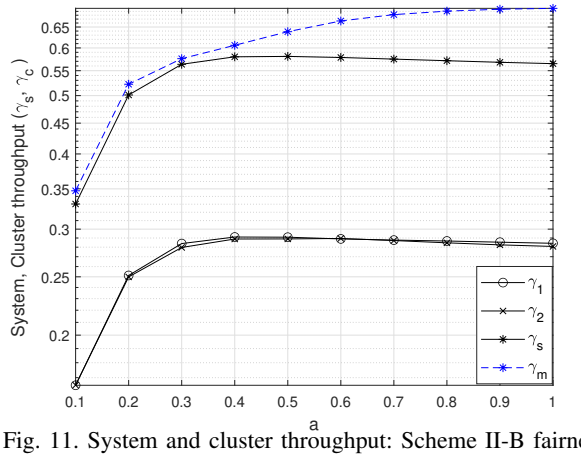

Fig. 11. System and cluster throughput: Scheme II-B fairness.

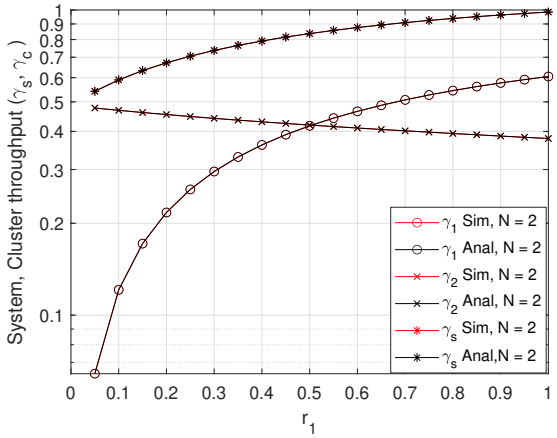

Fig. 12. System, cluster throughput for Scheme II-B: $R = 2$.

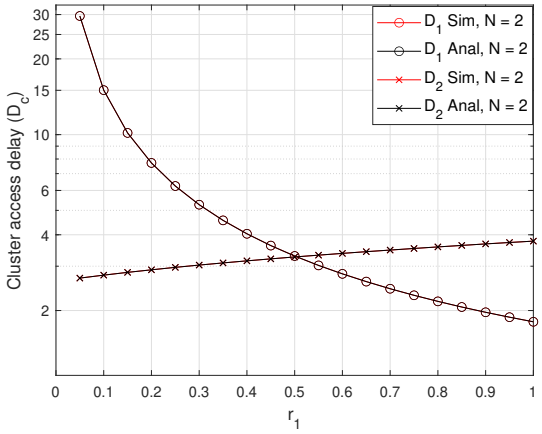

Fig. 13. Cluster access delay for Scheme II-B: $R = 2$.

longer delay. In the meantime, less interference is generated to cluster 1 meaning that cluster 1 transmission will more likely be successful, resulting in short access delay.

4) The effect of access control probability: To investigate the impact of access control, we configure $r_1 = r_2 = r$ and let r vary under a light traffic load ($a = 0.4$) while keeping the other parameters the same as the other results in this subsection. The obtained throughput and delay performance is illustrated in Figs. 8 and 9, respectively.

Let us observe the set of curves in Fig. 8 with $N = 2$ (in black) first. With a low r value, devices intend to postpone their transmission attempts, leading to a light traffic load in the network. Although collisions occur seldom, the obtained cluster and system throughput is low as the injected traffic is low. When more transmission attempts pursue with a larger r, higher throughput is obtained until a certain point ($r = 0.7$ for system throughput) at which the highest throughput is achieved. After this saturation point, throughput goes down as devices are more likely to transmit immediately (i.e., r becomes larger). Without access control ($r = 1$), heavy collisions happen, resulting in lowest throughput. When comparing the results of $N = 3$ (in red) versus of $N = 2$ (in black), we observe that the saturation point is shifted towards a lower value of r. This is because collisions may occur more often when the device population becomes larger.

With respect to the delay performance which is shown in Fig. 9, the same trend can be observed. That is, inferring access control too early with a light traffic load (devices intend to postpone their transmission attempts) or too late with a heavy traffic load (many retransmissions due to irresolvable collisions) could lead to longer delay. In general, we prefer not to introduce access control when the traffic load is light and recommend more stringent access control when traffic load becomes heavier.

F. Comparison of NOMA and OMA

In this subsection, we first compare the throughput achieved by NOMA (Scheme II) versus by OMA. Then, we show that by properly configuring the access probability for a given traffic load and network configuration, optimal performance can be achieved. Configure the system as $R = 1$, $N = 2$, $d_1 = 450$ m, and $d_2 = 900$ m. We fine-tune $(r1, r2)$ for each traffic load to maximize the system throughput. The maximum throughput shown in Figs. 10 and 11 is obtained by exhaustive search of the appropriate access probabilities.

Fig. 10 illustrates the obtained cluster and system throughput as a varies, for both Scheme II and OMA. In both cases, an error-prone channel which is elaborated in Subsection VI-C has been adopted. For packet transmissions based on OMA, concurrent transmissions from two or more devices are regarded as collision(s) and the packets being transmitted are lost. In this case, a packet transmission may be successful if

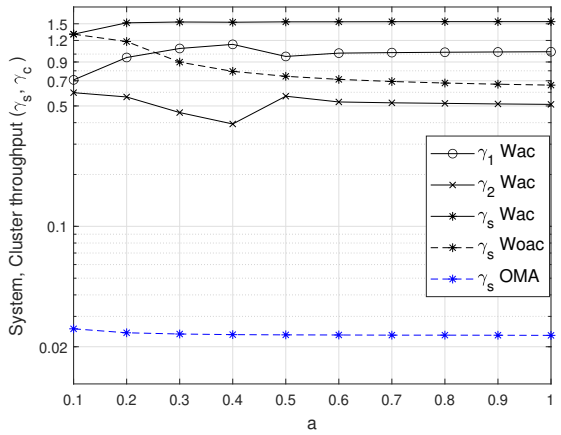

Fig. 14. System and cluster throughput: $N = 8, R = 3$, where wac and woac stand for with and without access control respectively.

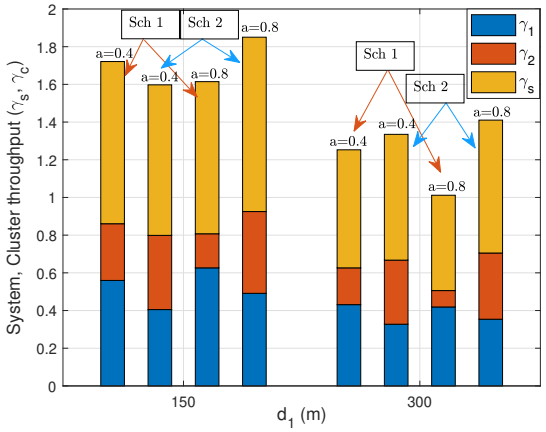

Fig. 15. The effect of inter-cluster interference on throughput.

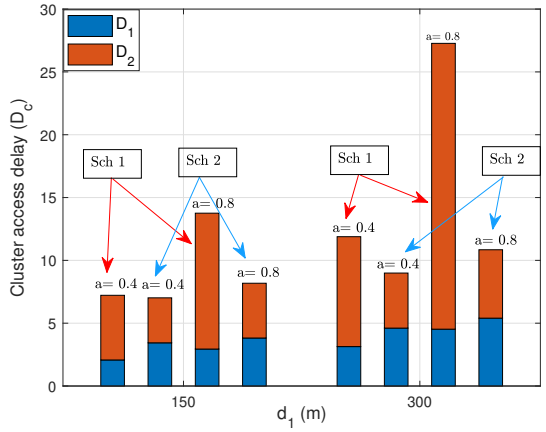

Fig. 16. The effect of inter-cluster interference on access delay.

and only if it is the sole ongoing transmission in a time slot and the received signal strength satisfies the threshold for receiver sensitivity.

As evident in Fig. 10, the system throughput of our proposed scheme is always higher than that of the conventional OMA, regardless of the traffic condition. With light traffic, the distinction between NOMA and OMA throughput is not significant as collisions rarely occur. As traffic load increases, NOMA with optimal access control exhibits a significant performance advantage over its OMA counterpart. With heavy traffic load, Scheme II still obtains very high throughput thanks to its capability of access control and interference tolerance by SIC.

G. Fairness-based Access Control

As mentioned previously, access probabilities in Scheme II-B can be configured to achieve throughput fairness among clusters. Fig. 11 depicts the throughput achieved by each cluster for different loads, when the access probabilities of each cluster have been chosen to maximize fairness. Fig. 11 also shows the system throughput γ_s (black), i.e., the sum of the cluster throughput, and the system throughput γ_m (blue) that would be achieved if the access probabilities of each cluster have been tuned to maximize throughput. Observe that for the studied evaluation scenario, throughput fairness among clusters has been achieved at the expense of a tolerable sacrifice on the system throughput.

H. Performance with Multiple Resources

In this subsection, we further explore the performance of Scheme II-B when multiple radio resources are available for packet transmissions to the BS.

1) Model validation with multiple resources: To demonstrate the versatility of our models, we define another evaluation scenario with the following network configuration: $d_1 = 450$ m, $d_2 = 900$ m, $N = 2$, $a = 0.4$, $r_2 = 1$, $M = 1$, $C = 2$, $R = 2$, and variable r_1.

Figs. 12 and 13 reveal that the throughput and access delay obtained by the analytical model match the ones obtained by simulations, validating in this way the precision of the proposed analytical model for a large number of radio resources. Observe in Fig. 12 that in the evaluated scenario, the maximum throughput is achieved when access control is disabled ($r_1 = r_2 = 1$). Recall that for $R = 1$ the maximum throughput was achieved with the access control enabled as shown in Fig. 8. Clearly, adding more radio resource reduces the intra- and inter-cluster interference.

2) Performance with a larger device population: Furthermore, Fig. 14 illustrates the evolution of the cluster and system throughput in a larger network with $R = 3$ and $N_1 = N_2 = 8$ with identical traffic load to both clusters. As evident from this figure, at a maximum load ($a = 1$), when the access probabilities have been configured to maximize throughput, the system throughput (γ_s wac) is higher than the throughput obtained with access control disabled (γ_s woac) by a factor of 2.3 , and it is higher than the throughput obtained by OMA by two orders of magnitude, approximately.

Observe also that to maximize the system throughput it is more effective to increase the access probability of cluster 1 devices and decrease the access probability of cluster 2 devices up to a certain load ($a = 0.4$). However, as load keeps on increasing, it is more beneficial to allocate similar access probabilities to both clusters. That is, while inter-cluster interference is low or moderate, it is more reasonable to maximize throughput to allocate a larger access share to the closest cluster. On the other hand, as inter-cluster interference becomes more significant, it is more preferable to allocate similar access probabilities to both clusters.

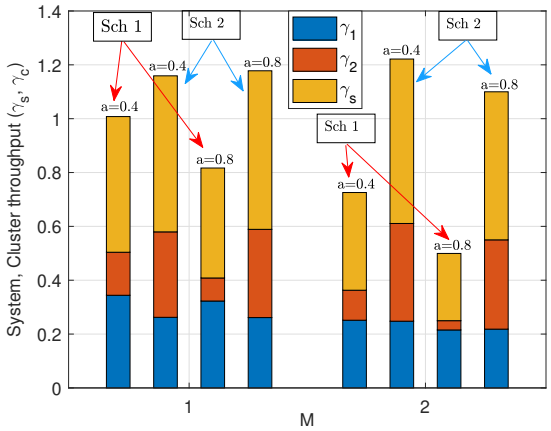

Fig. 17. The effect of BS antenna number on throughput.

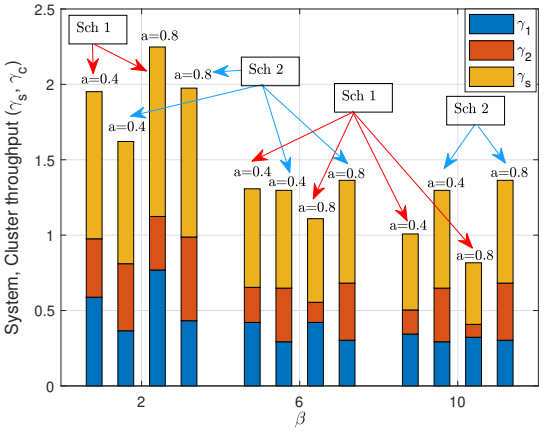

Fig. 19. The effect of SIC threshold on throughput.

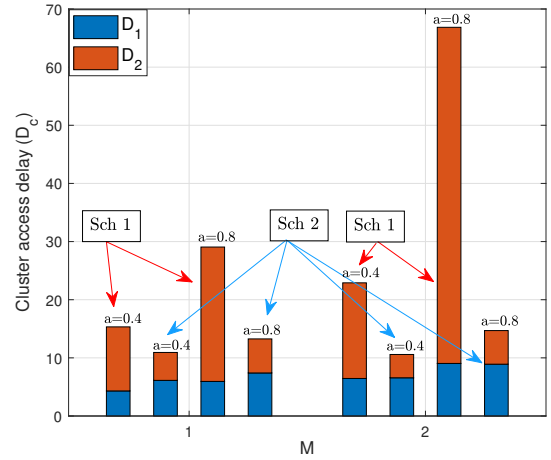

Fig. 18. The effect of BS antenna number on delay.

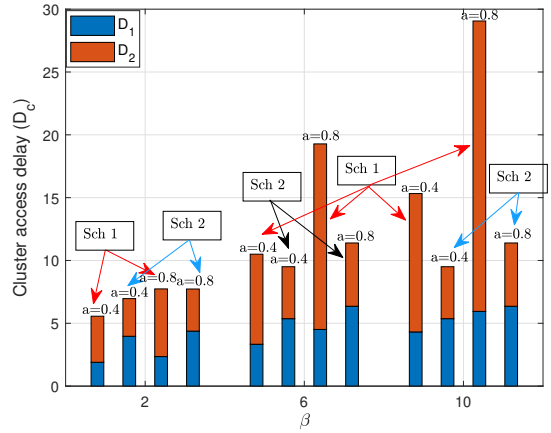

Fig. 20. The effect of SIC threshold on access delay.

I. Other Factors that Have an Impact on the Performance

Finally, we investigate the effects of a few other factors that may play a role in the performance of the schemes.

1) The effect of variable inter-cluster interference: To investigate the effect of inter-cluster interference, we reconfigure d_1 with different values as $d_1 = 150$ and $d_1 = 300$ m respectively, while keeping the BS-cluster 2 distance as $d_2 = 900$ m. The throughput and delay performance under light ($a = 0.4$) and heavy ($a = 0.8$) traffic load is illustrated in Figs. 15 and 16, respectively, as a histogram. For Scheme II shown in these figures, it is meant for the fair access implementation of Scheme II-B, as $r_1 = 0.4$ and $r_2 = 1$.

With a shorter distance d_1, the inter-cluster interference becomes weaker as the distance between these two clusters becomes greater. Consequently, higher system throughput is obtained for both schemes. With light traffic, Scheme I achieves higher system throughput since very few collisions happen in this case, and imposing restrictions on transmission attempts may not be necessary. On the other hand when traffic load is heavy, postponing a certain amount of transmission attempts would help reduce collisions, leading to higher system throughput. The same trend is observed when $d_1 = 300$ m.

For cluster throughput, we notice that γ_1 is higher than γ_2 under both traffic load conditions. However, higher cluster 2 throughput is obtained when a restriction is introduced to cluster 1. In this way, cluster 2 contributes more to system throughput in Scheme II as compared in Scheme I. Furthermore, the delay performance (shown in Fig. 16) exhibits similar trends as we observe for system and cluster throughput.

2) The effect of MIMO: Let us investigate the effect of multiple antennas by configuring $M = 2$ at the BS and compare its performance versus that of $M = 1$. The throughput and delay results are illustrated in Figs. 17 and 18, respectively.

With light traffic ($a = 0.4$), lower system throughput is achieved when Scheme I is employed. This is due to the fact that deploying more antennas leads to deteriorated SINR compared with the single antenna case. When Scheme II is introduced, however, less transmissions are initiated from cluster 1 and more antennas would lead to stronger signals received from cluster 2, yielding higher cluster throughput from cluster 2. Consequently, the system throughput is somehow increased. With heavy traffic ($a = 0.8$), the performance trend of Scheme I is similar to light traffic as no access control is imposed. For Scheme II with heavy traffic, lower throughput is observed with $M = 2$. This is due to stronger interference caused by more antennas, showing that *MIMO may not bring throughput benefit when traffic load is heavy*.

The same observation becomes more evident when we consider access delay, which is shown in Fig. 18. When traffic load is heavy, collisions occur more often with $M = 2$. This is

again due to the fact that stronger interference from cluster 2 is introduced when more antennas are employed. With access control (i.e., for Scheme II), much shorter delay is achieved thanks to less interference. With this network configuration, the benefit of deploying multiple antennas is very limited.

3) The effect of SIC threshold values: When demonstrating the benefits of NOMA, most existing studies illustrate typically the sum rate accumulated from multiple transmissions, where a certain amount of data rate can still be obtained through SIC even though the difference between strong and weak signals is very small [15]. We argue, however, that *null data rate could be achieved if the required threshold β is lower than a certain value*. The threshold value in the other figures in this paper is adopted from real-life experiments performed in [29], as $\beta = 10$ dB. In Figs. 19 and 20, we re-configure this threshold to $\beta = 2$ and $\beta = 6$ dB respectively, and compare the performance with that of $\beta = 10$ dB.

As illustrated in Fig. 19, a general trend of the throughput performance coincides with our intuition. That is, the smaller the β value, the higher the system throughput. The cluster throughput performance appears also more or less the same as what is observed with $\beta = 10$ dB. Similar performance applies to access delay as well (shown in Fig. 20). This is because small β values facilitate successful decoding of concurrent transmissions by SIC when the interference is still strong (SINR$= 1.58$ or 3.98 respectively, when $\beta = 2$ and $\beta = 6$ dB). In reality, a much larger SINR value, with a threshold as $\beta = 10$ dB for instance, has to be imposed before conducting successful SIC decoding.

VII. Conclusions

To understand the behavior of uplink transmissions in an error-prone MIMO-NOMA network where a single beam covers IoT devices grouped into clusters, we have proposed two random access schemes with and without access control and evaluated their performance using dedicated DTMC-based analytical models. The developed models represent precisely the behavior of intra- and inter-cluster transmissions as both dependent and independent inter-cluster transmissions are integrated in our models. The accuracy of the models is validated through extensive simulations and the performance of the schemes is evaluated with various network configurations and under different traffic load conditions. The main takeaways of this work are: 1) To improve network performance for NOMA-enabled uplink concurrent transmissions in massive IoT networks, device clustering and parameter configuration with respect to inter- and intra-interference are scenario-oriented and must be carefully tuned; 2) Access control is a powerful mechanism that can be fine-tuned to achieve maximal aggregated throughput or throughput fairness among clusters. 3) Deploying more antennas does not necessarily bring benefit for throughput improvement when multiple devices and clusters exist; and 4) Power-domain NOMA with marginal SINR threshold does not achieve system throughput as many early studies claim. To demonstrate the benefit of NOMA, realistic threshold values for SIC decoding, for instance the one adopted in this study, should apply.

References

[1] A. Kumar, F. Y. Li, and J. Martinez-Bauset, "Performance evaluation of cluster-based concurrent uplink transmissions in MIMO-NOMA networks," in *Proc. IEEE ICC*, pp. 1–6, May 2023.

[2] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[3] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road toward 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 1117–1174, 2nd Quart., 2022.

[4] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 55, pp. 179–189, Jan. 2020.

[5] *Study on New Radio (NR) Access Technology*, document TS38.912 R17, v17.0.0, 3GPP, Mar. 2022.

[6] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR: The Next Generation Wireless Access Technology*, 2nd ed., London, UK: Academic Press, 2021.

[7] M. Mohammadkarimi, O. A. Dobre, and M. Z. Win, "Massive uncoordinated multiple access for beyond 5G," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 2969–2986, May 2022.

[8] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, Feb. 2014.

[9] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 3rd Quart., 2020.

[10] F. Jabbarvaziri, N. M. Balasubramanya, L. Lampe, "HARQ-based grant-free NOMA for mMTC uplink," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 8372–8386, May 2021.

[11] I. N. A. Ramatryana and S. Y. Shin, "Priority access in NOMA-based slotted ALOHA for overload 6G massive IoT," *IEEE Commun. Lett.*, vol. 26, no. 12, pp.3064-3068, Dec. 2022.

[12] Y. Ma, Z. Yuan, W. Li, and Z. Li, "Novel solutions to NOMA-based modern random access for 6G-enabled IoT," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15382–15395, Oct. 2021.

[13] D. Cai, P. Fan, Q. Zou, Y. Xu, Z. Ding, and Z. Liu, "Active device detection and performance analysis of massive non-orthogonal transmissions in cellular Internet of things," *Sci. China Inf. Sci.*, vol. 65, art. no. 182301, pp. 1–18, Aug. 2022.

[14] T. N. Weerasinghe, V. Casares-Giner, I. A. M. Balapuwaduge, and F. Y. Li, "Priority enabled grant-free access with dynamic slot allocation for heterogeneous mMTC traffic in 5G NR networks," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3192–3206, May 2021.

[15] S. A. Tegos, P. D. Diamantoulakis, A. S. Lioumpas, P. G. Sarigiannidis, and G. K. Karagiannidis, "Slotted ALOHA with NOMA for the next generation IoT," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6289–6301, Oct. 2020.

[16] Y.-C. Huang, S.-L. Shieh, Y.-P. Hsu, and H.-P. Cheng, "Iterative collision resolution for slotted ALOHA with NOMA for heterogeneous devices," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 2948–2961, May 2021.

[17] A. Mahmoudi, B. Abolhassani, S. M. Razavizadeh, and H. H. Nguyen, "User clustering and resource allocation in hybrid NOMA-OMA systems under Nakagami-m fading," *IEEE Access*, vol. 10, pp. 38709–38728, Apr. 2022.

[18] A. Shahini and N. Ansari, "NOMA aided narrowband IoT for machine type communications with user clustering," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7183–7191, Aug. 2019.

[19] M. Mohammadkarimi, M. A. Raza, and O. A. Dobre, "Signature-based nonorthogonal massive multiple access for future wireless networks: Uplink massive connectivity for machine-type communications," *IEEE Veh. Tech. Mag.*, vol. 13, no. 4, pp. 40–50, Dec. 2018.

[20] W. A. Al-Hussaibi and F. H. Ali, "Efficient user clustering, receive antenna selection, and power allocation algorithms for massive MIMO NOMA systems," *IEEE Access*, vol. 7, pp. 31865–31882, Feb. 2019.

[21] H. Tabassum, E. Hossain, and Md. J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using Poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.

[22] M. A. Sedaghat and R. R. Müller, "On user pairing in uplink NOMA," *IEEE Wireless Trans. Commun.*, vol. 17, no. 5, pp. 3474–3486, May 2018.

[23] M. Liu, J. Zhang, K. Xiong, M. Zhang, P. Fan, and K. B. Letaief, "Effective user clustering and power control for multiantenna uplink NOMA transmission," *IEEE Wireless Trans. Commun.*, vol. 21, no. 7, pp. 8995–9009, Nov. 2022.
[24] F. Ghanami, G. A. Hodtani, B. Vucetic, and M. Shirvanimoghaddam, "Performance analysis and optimization of NOMA with HARQ for short packet communications in massive IoT," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4726–4748, Mar. 2021.
[25] F. Ghanami, S. Poursheikhali, and M. Shirvanimoghaddam, "Dynamic NOMA-HARQ in the finite blocklength regime: Performance analysis and optimization," *IEEE Commun. Lett.*, Early access article, 2023.
[26] Z. Zhang, Y. Li, G. Song, C. Yuen, and Y.-L. Guan, "Random NOMA with cross-slot successive interference cancelation packet recovery," *IEEE Wireless Commun. Lett.*, vol. 9, no. 7, pp.1065-1069, Jul. 2020.
[27] K. Senel, H. V. Cheng, E. Björnson, and E. G. Larsson, "What role can NOMA play in massive MIMO?" *IEEE J. Sel. Topics Signal Process*, vol. 13, no. 3, pp. 597–611, Jun. 2019.
[28] Z. Ding, R. Schober, and H. V. Poor, "Unveiling the importance of SIC in NOMA systems – Part 1: State of the art and recent findings," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2373–2377, Nov. 2020.
[29] Y. Qi, X. Zhang, and M. Vaezi, "Over-the-air implementation of NOMA: New experiments and future directions," *IEEE Access*, vol. 9, pp.135828–135844, Sep. 2021.
[30] V. Casares-Giner, J. Martinez-Bauset, and C. Portillo, "Performance evaluation of framed slotted ALOHA with reservation packets and succesive interference cancelation for M2M networks," *Comput. Netw.*, vol. 155, pp. 15–30, May 2019.
[31] D. Ghose, F. Y. Li, and V. Pla, "MAC protocols for wake-up radio: Principles, modeling and performance analysis," *IEEE Trans. Ind. Informat.*, vol. 14, no. 5, pp. 2294–2306, May 2018.
[32] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the accurate performance evaluation of the LTE-A random access procedure and the access class barring scheme," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7785–7799, Dec. 2017.
[33] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J.-R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Trans. Veh. Tech.*, vol. 6, no. 4, pp. 3505–3520, Apr. 2018.

Abhishek Kumar (Student Member, IEEE) obtained his B.E. degree in electronics and communication engineering from Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV), Bhopal, India, in 2017. He further pursued his M.Tech. degree in electronics and communication engineering from the Indian Institute of Technology (IIT) Bhubaneswar, India, in 2021. He is currently working towards a Ph.D. degree at the Department of Information and Communication Technology, University of Agder (UiA), Norway. His research focuses on MAC processes, routing protocols, and resource management in beyond 5G mobile systems, wireless networks, and IoT networks.

Jorge Martinez-Bauset received the Ph.D. degree from the Universitat Politècnica de València (UPV), Valencia, Spain, in 1997. He is currently a Professor with the UPV. From 1987 to 1991, he was with QPSX Communications, Perth, Australia, working with the team that designed the first IEEE 802.6 MAN. Since 1991, he has been with the Department of Communications, UPV. His research interests include performance evaluation and traffic control for multiservice networks. He was the recipient of the 1997 Alcatel Spain Best Ph.D. Thesis Award in access networks.

Frank Y. Li received the Ph.D. degree from the Department of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2003. He was a Senior Researcher at the Department of Technology Systems, University of Oslo, before joining the Department of Information and Communication Technology, University of Agder (UiA), in 2007, as an Associate Professor and then a Full Professor. From Aug. 2017 to Jul. 2018, he was a Visiting Professor with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. During the past years, he has been an active participant in multiple Norwegian and EU research projects. His research interests include MAC mechanisms and routing protocols in 5G and beyond mobile systems and wireless networks, Internet of things, mesh and ad hoc networks, wireless sensor networks, D2D communications, cooperative communications, cognitive radio networks, green wireless communications, dependability and reliability in wireless networks, QoS, resource management, and traffic engineering in wired and wireless IP-based networks, and the analysis, simulation, and performance evaluation of communication protocols and networks. He was listed as a Lead Scientist by the European Commission DG RTD Unit A.03- Evaluation and Monitoring of Programmes in Nov. 2007.

Carmen Florea (Member, IEEE) received her M.Sc degree in mobile communications in 2011 and the Ph.D. degree, in the area of multiuser MIMO and its applications in wireless communications the National University of Science and Technology Politehnica Bucharest (UNSTPB), Romania, in 2014. She has been an Assistant Professor (2011–2014), Lecturer (2014–2022), and from 2022 she is an Associate Professor at the Department of Telecommunications, Faculty of Electronics, Telecommunications and Information Technology, UNSTPB. She is a participant in several research projects focused on wireless networks with an activity related to signal processing. Her domains of interest are multiple access systems & techniques, analog and digital transmission systems, channel coding for wireless systems.

Octavia A. Dobre (Fellow, IEEE) received the Dipl.Ing. and Ph.D. degrees from the Polytechnic Institute of Bucharest, Bucharest, Romania, in 1991 and 2000, respectively. She was with New Jersey Institute of Technology, Newark, NJ, USA, from 2002 to 2005. She joined Memorial University of Newfoundland, St. John's, NL, Canada, in 2005, where she is currently a Professor and a Research Chair. She was a Visiting Professor with Massachusetts Institute of Technology, Cambridge, MA, USA and Université de Bretagne Occidentale, Brest, France. Her research interests encompass wireless communication and networking technologies, as well as optical and underwater communications. She has (co)authored over 400 refereed papers in these areas. Dr. Dobre obtained the Best Paper Awards at various conferences, including IEEE ICC, IEEE GLOBECOM, IEEE WCNC, and IEEE PIMRC. She also served as a general chair, technical program co-chair, tutorial co-chair, and technical co-chair for symposia at numerous conferences. She was a Fulbright Scholar, a Royal Society Scholar, and a Distinguished Lecturer of the IEEE Communications Society. She serves as the Director of Journals of the Communications Society. She was the inaugural Editor-in-Chief (EiC) of the IEEE Open Journal of the Communications Society, and the EiC of the IEEE Communications Letters. She is an Elected Member of the European Academy of Sciences and Arts, and a Fellow of the Engineering Institute of Canada and Canadian Academy of Engineering.