# Energy and Delay Aware Two-hop NOMA-Enabled Massive Cellular IoT Communications

Hesham G. Moussa, *Student Member, IEEE,* Weihua Zhuang, *Fellow, IEEE,*

***Abstract*—Providing energy efficient and delay-aware channel access in cellular networks is essential to many anticipated massive Internet of Things (IoT) applications. However, as the number of devices increases, the contention over the limited network radio resources increases, leading to network congestion. The congestion increases channel access delay and energy consumption of IoT devices, and reduces the number of supported devices. Node clustering and data aggregation are potential approaches to support massive number of devices, while meeting the various service quality requirements of diverse applications. As the number of devices increases, optimizing the node clustering and data aggregation process becomes critical as many tradeoffs arise among different network performance metrics. In this paper, we present a novel NOMA-enabled two-stage transmission architecture to enable massive cellular IoT communications. Concepts from queuing theory and stochastic geometry are jointly exploited to derive tractable models for different network performance parameters such as coverage probability, two-hop access delay, and the number of served devices per transmission frame. The established models characterize relations among various network parameters, and hence facilitate the design of two-stage transmission architecture. Numerical results demonstrate that the proposed solution improves the overall access delay and energy efficiency as compared to traditional OMA-based clustered networks. They also highlight that, for the scenario considered, there is an optimal number of aggregators at which the tradeoffs among the different network performance measures are optimized.**

***Index Terms*—Machine type communications, data aggregation, node clustering, delay analysis, Internet of Things (IoT).**

## I. Introduction

Massive machine type communication (mMTC) is envisioned to play a major role in efficiently supporting massive internet of things (IoT) applications in future cellular networks. Under mMTC, a massive number of heterogeneous devices communicate with each other with minimal human intervention, while satisfying a wide range of quality of service (QoS) requirements of various applications. Yet, a major challenge brought on by the massive number of connected devices lies in the possible congestion of the random access channel in the cellular network. This challenge is often referred to as overload problem and is shown to greatly impact the delay and energy efficiency performance of the access network [1].

Manuscript submitted for review on September 17, 2019. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (NSERC Canadian FloodNet).

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: h3moussa@uwaterloo.ca; wzhuang@uwaterloo.ca)

Many solutions have been proposed for the overload problem such as access class barring (ACB) [2], extended access baring (EAB) [3], randomized back-off schemes [4], prioritized random access [5], and many others [6], [7]. The most promising solutions to enable mMTC and avoid overloading the access channel are node clustering and data aggregation [8], [9]. With clustering and data aggregation, devices do not contend directly on the available resources at the base station (BS); instead, they connect to a middle layer of data aggregators (DAs) which relay the data on their behalf in a two-hop fashion. By doing so, the number of devices contending for BS resources decreases, reducing the chances for congestion. Furthermore, devices consume less energy as the transmission distance is significantly shortened [10]–[12].

Given these potentials, recently, node clustering and data aggregation have been investigated as possible solutions for enabling energy efficient mMTC while maintaining acceptable network performance in terms of delay and throughput [13], [14]. For instance, in [9], to study the impact of clustering on the outage probability under different channel models, the clustering problem is formulated to answer three main questions: how many clusters to form, what should be the transmit power of the devices, and how the clustering decision depends on the networking environment. The authors conclude that, with the proper choice of the number of clusters to form, optimized network design can be achieved. In [10], Guo et al. consider a two-phase transmission system where the DAs take the responsibility of not only relaying data, but also handling resource scheduling among active devices. It is shown that by properly choosing the number of aggregators, the number of successfully served devices can be enhanced. In [11], the clustering problem is formulated as an optimization problem in order to determine the number of clusters which results in the lowest energy consumption of the devices, with a hierarchical transmission system of two or more hops. In [12], energy efficiency of cellular networks for massive IoT applications is studied under a drone assisted scenario. In [15], different deployment strategies for the DAs are studied in terms of their impact on energy consumption and coverage probability. In [8], an $n$-CSMA medium access for in-cluster transmission is proposed, and its impact on the performance of clustered networks is studied, including the correlation between cluster size and overall network performance in terms of energy and throughput. In [16], a two-stage access for cellular IoT is considered, and queuing theory is used to analyze the delay performance as a function of the device density. All the existing works emphasize on the effectiveness

of node clustering and data aggregation for enabling mMTC, while maintaining acceptable network performance. They also show that, when dealing with a massive number of devices, data aggregation becomes challenging and optimization of different design parameters is essential.

On the other hand, non-orthogonal multiple access (NOMA) has recently gained attention as a potential medium sharing scheme to improve the spectral efficiency and to increase the number of supported devices in future 5G networks [17]. In NOMA, thanks to multi-user detection and interference mitigation techniques such as successive interference cancellation (SIC) [18], multiple devices are allowed to share the same radio channel simultaneously, leading to a higher user capacity and better performance than traditional orthogonal multiple access (OMA) [19]. Thus, a NOMA-based data aggregation framework presents a promising solution to enable mMTC. Nevertheless, most of the existing works on data aggregation for mMTC communications have focused on optimizing the network performance with OMA based channel access [10], [11], and only a few have considered NOMA [20]. While these studies prove the potential of NOMA as a method for supporting massive connectivity in future cellular networks, further research on the potentials of NOMA-enabled clustered operation is needed.

Thus, in an effort to further explore the potentials of node clustering and data aggregation in the context of mMTC, in this paper, we aim at shedding some light on the impact of clustering on different network performance metrics such as number of supported devices, energy consumption, and delay. We propose a new two-hop non-orthogonal multiple access (NOMA) enabled clustered transmission framework to support massive cellular IoT applications. As we are considering NOMA, all of the above network parameters depend on the the severity of the interference, which is a function of the number of devices sharing the same resource channel as well as the transmission power of the devices. Accordingly, to be able to quantify the network performance metrics, we first start by characterizing the in-network interference components and define the coverage probability in the system. We then utilize techniques from both queuing theory and stochastic geometry to achieve tractable analytical results for the various network parameters as functions of the defined coverage probability. Our proposed framework captures the unique characteristics of future IoT applications, including the small sized data packets generated by the devices, massive number of devices in the network, and limited radio resources. The main contributions of this work are summarized as follows:

- We present a novel NOMA-enabled two-hop network model for massive cellular IoT communications. We develop a general analytical framework to obtain approximating yet accurate mathematical models for the coverage probability, average number of served devices, overall average access delay, and average energy consumption of a device and a DA;
- Compared to the literature, our derived expressions are more comprehensive and thus can be used to study the impact of various design parameters (such as device density, DA density, device transmission power, and available radio resources) on delay, coverage probability, and energy efficiency performance of massive cellular IoT applications;
- The accuracy of the analytic models are corroborated via computer simulations. Compared with traditional OMA-based clustered networks, with the proper choice of the number of clusters, our proposed solution improves the overall network performance;

The rest of this paper is organized as follows: In Section II, we describe the system model for the NOMA-enabled two-stage transmission. Section III details the coverage probability analysis for the two stages. In Section IV, the network delay performance is investigated. Section V provides the energy consumption analysis of both devices and DAs. Section VI presents the numerical and simulation results, and summarizes the main findings. Section VII concludes the study. The main mathematical symbols are listed in Table I for easy reference.

TABLE I. Summary of important mathematical symbols.

| Parameter | Description |
|---|---|
| $\mathcal{C}_d$ | Device-DA coverage probability |
| $\mathcal{C}_{tot}$ | Device-BS total coverage probability |
| $\mathcal{D}$ | Packet size in bits |
| $\mathcal{E}_d^f$ ($\mathcal{E}_a^f$) | Fixed amount of energy consumed by a device (DA) in the case of failed access attempt |
| $\mathcal{E}_d$ ($\mathcal{E}_a$) | Average energy consumption of a device (DA) to transmit a single data packet |
| $\mathcal{K}$ | Average number of DAs in the Voronoi cell of a BS |
| $\mathcal{M}_a$ | RV denoting number of devices in a cluster |
| $M_{max}$ | Maximum number of devices a DA can serve in a transmission frame |
| $\mathcal{N}$ | Total number of available NOMA sub-channels |
| $\mathcal{Q}_d^r$ ($\mathcal{Q}_a^r$) | Power received at a DA (BS) after power control |
| $\mathcal{Q}_d^t$ ($\mathcal{Q}_a^t$) | Power transmitted by a device (DA) |
| $\mathcal{R}_d^{sjn}$ ($\mathcal{R}_a^k$) | Average achievable packet rate of a device (DA) |
| $S^n$ | Number of scheduled devices on the $n^{th}$ subchannel |
| $T_a$, $T_r$, $T_f$ | Length of aggregation, relaying and full transmission frame respectively |
| $\gamma_d$, ($\gamma_a$) | Packet arrival rate at a device (DA) |
| $\Delta t_{e2e}$ | Average end-to-end system delay |
| $\Theta(\Xi_{ij}^{ns})$, $\Theta(\Xi_a^k)$ | Device (DA) transmission success probability given the device ranking $j$ |
| $\bar{\Theta}$ | Average transmission success probability of a device |
| $\Lambda_s$ | NOMA sub-channel scheduling probability |
| $\lambda$ | entity density (BS($_b$), DA($_a$), device($_d$)) |
| $\mu_d^{sjn}$ ($\mu_a$) | Packet departure rate from a device (DA) |
| $\Xi_{ij}^{ns}$ ($\Xi_a^k$) | SIR of the received transmission from a device (DA) |
| $\tau_a$ ($\tau_b$) | SIR threshold for successful transmission |
| $\Phi$ | Set of locations of network components (BS ($_b$), DA ($_a$), device ($_d$)) |
| $\Psi_n^h$ ($\Psi_n^l$)/$\Psi_a^k$ | Location set of the primary (secondary) interfering devices/DA on the $n^{th}$ NOMA subchannel/$k^{th}$ uplink sub-channel |

## II. System Model

### A. Spatial system model

Consider a layer of cellular BSs, whose locations can be modelled by a homogeneous Poisson point process (HPPP) $\Phi_b = b_0, b_1, ...$ with density $\lambda_b$, where $b_i$ denotes the location of the $i^{th}$ BS in the network. The cellular network has a massive number of identical and battery powered IoT devices, whose locations can be modelled by an independent HPPP $\Phi_d = d_0, d_1, ...$ with density $\lambda_d$, where $d_i$ denotes the location of the $i^{th}$ IoT device in the network. Data packet generation at each IoT device follows a Poisson process with parameter $\gamma$ and each generated data packet is of size $\mathcal{D}$ bits.

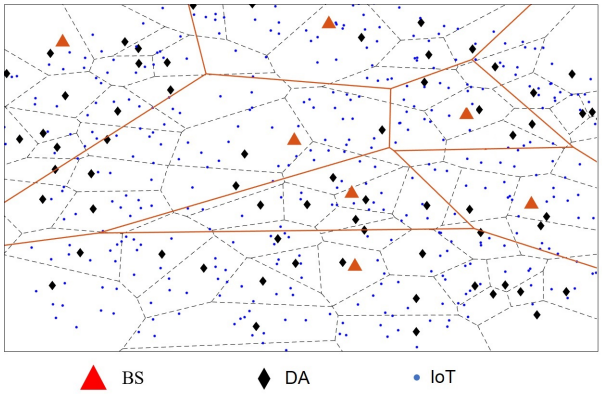



Fig. 1: An illustration of BS, DA, and IoT device locations as well as the Voronoi tessellation formed by the coverage of the BSs (solid lines) and that formed by the coverage of the DAs (dashed lines).

The cellular network is overlaid with a layer of stationary DAs that aggregate and relay data from IoT devices to BSs. The DAs are randomly scattered across the network coverage area such that their locations can be modelled as an independent HPPP $\Phi_a = a_0, a_1, ...$ with density $\lambda_a$, where $\lambda_b < \lambda_a << \lambda_d$ and $a_i$ denotes the location of the $i^{th}$ DA in the network. Active IoT devices transmit their data packets simultaneously towards the core network only via two-hop communication, i.e., IoT devices connect to DAs, and DAs relay aggregated data packets to BSs. Maximum received signal strength association policy is used by the devices and DAs to connect to their preferred access point. Figure 1 shows an example of node locations and coverage areas.

### B. Transmission frame structure and transmission process

Time is partitioned into transmission frames of equal length $T_f$. Each transmission frame is further divided into two phases: an aggregation phase of length $T_a$ and a relay phase of length $T_r (= T_f - T_a)$. Frequency division multiple access (FDMA) is used to split the channel between the devices and DAs into $\mathcal{N}$ sub-channels of equal bandwidth $\omega_n$. These sub-channels are used by all the DAs in the network. During the aggregation phase, devices transmit data to their serving DAs via power domain NOMA (PD-NOMA), where at most two devices from the same cluster are allowed to non-orthogonally share the same in-cluster sub-channel. The maximum number of devices that a DA can support in a single transmission frame is $M_{max} = 2\mathcal{N}$. Full channel state information (CSI) is assumed at the DAs, such that appropriate device pairing on the sub-channels is achieved in order to obtain benchmark results [20]. At the end of the aggregation phase, scheduled devices return to the pool of active devices awaiting service in the new transmission frame.

DAs employ successive interference cancellation (SIC) to decode the superimposed messages and cancel their mutual interference in a sequential manner. Multiplexed transmissions from devices on a sub-channel are ranked in a descending order based on their received signal strength at the DA. The DA decodes the signal from the higher ranked (stronger signal) device first, hereafter indexed by $j = h$. For the message from the higher ranked device to be successfully decoded, it should be received with an SIR above threshold $\tau_{ah}$. For the transmission from the lower ranked device, hereafter indexed by $j = l$, to be successfully decoded, two conditions should be met: i) the message from the higher ranked multiplexed device was successfully decoded, and ii) the message from the lower ranked device is received with an SIR above threshold $\tau_{al}$.

By the end of the data aggregation phase, all IoT devices switch off their communication modules and go to sleep to reduce energy consumption. The relay phase then begins where DAs transmit the aggregated data to the BS in a single hop fashion. We make a simplifying assumption that the number of DAs in any cell is equal to $\mathcal{K}$, which is the average number of DAs per cell given certain DA and BS node densities. A shared uplink channel (SUCH) of bandwidth $\Omega_b$ is dedicated for uplink transmissions from the DAs, which is divided in frequency into $\mathcal{K}$ orthogonal sub-channels of equal bandwidth. The $\mathcal{K}$ uplink sub-channels are shared by all BSs in the network. Only a single DA from the same Voronoi cell is allowed to occupy a sub-channel and thus DAs only experience inter-cell interference. A DA is considered connected to its serving BS if its transmitted data packets are received with SIR above threshold $\tau_b$.

### C. Wireless transmission model

Transmitted signals in both phases experience propagation attenuation according to a general power-law path-loss model. The signal power decays at rate $D^{-\alpha}$, where $D$ is the propagation distance and $\alpha$ is the path-loss exponent. Consider a Rayleigh fading channel that introduces a random instantaneous power gain, $g$, which follows an exponential distribution with unity mean (i.e. $g \sim \exp\{1\}$). Channel gains are distance independent, independent of each other and identically distributed (i.i.d.). The network is interference limited due to the massive number of devices. As in our previous work [21], consider that DAs employ fractional power control (FPC) when transmitting their payload over their scheduled uplink sub-channel. Accordingly, the power received at the BS located at the origin from an associated DA located at $a$ after FPC is $Q_a^r = q_a ||a||^{(\epsilon-1)\alpha} g$, where $q_a$ is the nominal transmission power of any DA in the network, $||a||$ is the Euclidean distance between the DA and the BS, and $\epsilon \in \{0, 1\}$ is the power control factor. On the other hand, IoT devices employ full power inversion such that the received power at a DA located at the origin from an associated device located at $d$ is $Q_d^r = q_d g$, where $q_d$ is the nominal transmission power of any IoT device in the network. As in previous studies, we do not consider the maximum transmission power limitation for both DAs and devices [21].

In the following few sections, we attempt to derive mathematical models for coverage probability, end-to-end delay, and energy consumption for the preceding described system model. First, in Section-III, by borrowing tools from stochastic geometry and using the spatially modeling of the locations of the BSs, DAs and active devices as three independent

homogeneous PPPs, we characterize interference components experienced by both the devices and the DAs. The Laplace transforms for the interference components are then derived and used to find expressions for the average device-DA and DA-BS transmission success probabilities. These probabilities are then used to derive both the device-DA coverage probability and the total coverage probability in this two-hop network. In this work, coverage probability is defined as the probability that a device is able to successfully transmit its packet to the core network via a two-hop network architecture. A device that is able to transmit its data packets successfully is referred to as a covered device. Second, using these coverage probabilities and by modeling this two-hop network as a two-stage tandem queue, in Section-IV, we derive the arrival and departure processes from which the end-to-end average packet delay can be derived. Last, in Section-V, the coverage probabilities and the end-to-end delay results for the two-stage tandem queue are used to derive expressions for the average energy consumption of a typical device and a typical delay to transmit a single data packet.

## III. Coverage probability

Coverage probability refers to the probability that a generated data packet from an arbitrary IoT device is successfully received by its serving BS [10]. Different from existing studies [20], the main challenge in our work is the PPP locations of IoT devices and DAs. The coverage probability consists of three parts: i) NOMA sub-channel scheduling probability - the probability of an arbitrary active device being scheduled in the current transmission frame, such that the device is able to transmit its data packet to the serving DA; ii) Device-DA coverage probability - the probability that a transmitted packet from a scheduled device is received at its serving DA with SIR above threshold $\tau_a \in \{\tau_{ah}, \tau_{al}\}$, depending on the ranking of the device; and iii) DA-BS coverage probability - the probability that a transmitted aggregated data packet from a DA is received at its serving BS with SIR above threshold $\tau_b$. The aggregation and relaying phases are correlated as the data transmitted by a DA depends on the number of scheduled devices and their SIR performance. Nonetheless, for the sake of tractability, we assume that the two phases are independent [10]. In what follows, we focus on a cluster that has its DA located at the origin (with index $i = 0$); however, the analysis is applicable to any cluster in the network. We consider the case when the network is full, i.e., at the beginning of the considered transmission frame, all devices in the network have at least one data packet to transmit.

### A. NOMA sub-channel scheduling probability

As mentioned in the system model, the number of devices in a cluster is random and dependent on both the density of devices and DAs. On the other hand, the number of available NOMA sub-channels is fixed and so is the maximum number of devices, $M_{max}$, that can be supported by a DA per transmission frame. Consequently, in the case that the number of devices in a cluster is greater than $M_{max}$, some devices may not be scheduled for transmission in the current frame due to insufficient resources. Thus, we define NOMA sub-channel scheduling probability as the probability that a device with data to transmit is scheduled for transmission on a sub-channel in the current frame. To determine this scheduling probability, we first characterize the distribution of the device number in a cluster. Let random variable (RV) $\mathcal{M}_a$ denote the number of devices associated with the DA over area $\mathcal{V}_a$ of the Voronoi cell. As in our previous work [21], by approximating the PDF of $\mathcal{V}_a$ by a generalized Gamma distribution, we obtain the probability mass function (PMF) of $\mathcal{M}_a$, given by

$$\begin{aligned} P(\mathcal{M}_a = m) &= \int_0^\infty \frac{(\lambda_d v)^m e^{-\lambda_d v}}{m!} f_{\mathcal{V}_a}(v, \lambda_a, c)\, \mathrm{d}v \\ &\approx \frac{(\lambda_a)^c (\lambda_d)^m \Gamma(m+c)}{\Gamma(c)(\lambda_a+\lambda_d)^{m+c} m!} \end{aligned} \tag{1}$$

where $f_{\mathcal{V}_a}(v, \lambda_a, c) = \frac{(\lambda_a^c v^{c-1} e^{-\lambda_a v})}{\Gamma(c)}$, $c = 3.575$ is a constant defined for the Voronoi tessellation in $\mathbb{R}^2$ [22] and $\Gamma(c) = \int_0^\infty t^{c-1} e^{-t}\, \mathrm{d}t$.

Let $\Lambda_s = \frac{M_{max}}{\max\{\mathcal{M}_a, M_{max}\}}\big|_{\mathcal{M}_a=m}$ be the conditional scheduling probability given the number of active devices in the current transmission frame, $\mathcal{M}_a = m$. Averaging over the PDF given in (1), the NOMA sub-channel scheduling probability is given by

$$\bar{\Lambda}_s = \sum_0^{M_{max}} \frac{(\lambda_a)^c (\lambda_d)^m \Gamma(m+c)}{\Gamma(c)(\lambda_a+\lambda_d)^{m+c} m!} + \sum_{M_{max}+1}^{\infty} \frac{M_{max}}{m} \frac{(\lambda_a)^c (\lambda_d)^m \Gamma(m+c)}{\Gamma(c)(\lambda_a+\lambda_d)^{m+c} m!}. \tag{2}$$

### B. Device-DA coverage probability

To determine device-DA coverage probability, the SIR should be carefully characterized. We focus on a device from a cluster centered at the origin. As per the system model, each DA schedules its associated devices across the sub-channels in a sequential manner until a maximum of two devices are scheduled per sub-channel. The number of devices in the cluster is random, leading to randomness in the number of scheduled devices on a sub-channel. Let RV $S^n$ denote the number of devices scheduled on the $n^{th}$ sub-channel, for $n = 1, 2, ..., \mathcal{N}$. The PMF of $S^n$, $P(S^n = s)$, is given by (see appendix A)

$$\begin{cases} \sum_{m=0}^{\mathcal{N}-1} (1 - \frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m), & s = 0 \\ \sum_{m=\mathcal{N}}^{M_{max}-1} (2 - \frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m) + \sum_{m=0}^{\mathcal{N}-1} (\frac{m}{\mathcal{N}}) P(\mathcal{M}_a = m), & s = 1 \\ \sum_{m=M_{max}}^{\infty} P(\mathcal{M}_a = m) + \sum_{m=\mathcal{N}}^{M_{max}-1} (\frac{m}{\mathcal{N}} - 1) P(\mathcal{M}_a = m), & s = 2 \\ 0, & s > 2. \end{cases} \tag{3}$$

A device can experience two types of interference components: intra-cluster interference ($I_1$), and inter-cluster interference ($I_2$). $I_1$ can only be experience by a higher ranked device when sharing a sub-channel with a lower ranked device. On the other hand, $I_2$ can be experienced by any device regardless of its ranking. $I_2$ is composed of primary component ($I_{2h}$) due to higher ranked devices from adjacent clusters, and secondary component ($I_{2l}$) due to lower ranked devices from adjacent clusters. Locations of all higher and lower ranked inter-cluster

interfering devices can be modeled by two independent and thinned PPPs $\Psi_n^h = x_{0h}^n, x_{1h}^n, ...$ and $\Psi_n^l = x_{0l}^n, x_{1l}^n, ...$, with device densities $\lambda_d^h = (P(S^n = 1) + P(S^n = 2))\lambda_a$ and $\lambda_d^l = P(S^n = 2)\lambda_a$ respectively, where $x_{ij}^n$ denotes the location of the $j$ ranked interfering device from the $i^{th}$ cluster on the $n^{th}$ sub-channel, for $j \in \{h, l\}$, $i = 1, 2, ...$ and $n = 1, 2, ..., \mathcal{N}$. Note that, $\lambda_d^h$ follows from the fact that a device is classified as higher ranked if it is scheduled alone on a sub-channel or co-occupies the sub-channel with a lower ranked device, whereas $\lambda_d^l$ happens only when a sub-channel has more than one scheduled device. Thus, for an arbitrary device scheduled on the $n^{th}$ sub-channel and associated with the DA located at the origin, $I_{2h}$ and $I_{2l}$ are given by

$$I_{2h} = \sum_{x_{ih}^n \in \Psi_n^h} \mathbb{1}(||x_{ih}^n - a_k|| < ||x_{ih}^n||)||x_{ih}^n - a_k||^{\alpha}||x_{ih}^n||^{-\alpha} g_{ih}^{n0} \quad (4)$$

$$I_{2l} = \sum_{x_{il}^n \in \Psi_n^l} \mathbb{1}(||x_{il}^n - a_k|| < ||x_{il}^n||)||x_{il}^n - a_k||^{\alpha}||x_{il}^n||^{-\alpha} g_{il}^{n0} \quad (5)$$

where $\mathbb{1}(\cdot)$ is the indicator function, the condition $||x_{ij}^n - a_k|| < ||x_{ij}^n||$ is to ensure that an interfering device located at $x_{ij}^n$ is closer to its serving DA located at $a_k$ than to the origin, and $g_{ij}^{n0}$ denotes the i.i.d exponentially distributed channel gain for the link between the $j$ ranked interfering device on the $n^{th}$ sub-channel from the $i^{th}$ cluster and the origin.

For the following, we focus on a device associated with the DA located at the origin (DA at the origin is indexed by $i = 0$); however, the analysis can be generalized to any device-DA pair. Let $\Xi_{0j}^{ns}$ denote the received SIR at the origin from a $j$ ranked device located at $x_{0j}^n$ and scheduled on the $n^{th}$ sub-channel that has $s$ multiplexed devices, given by

$$\Xi_{0j}^{ns} = \begin{cases} \frac{g_{0h}^{n0}}{I_2}, & s = 1 \ \& \ j = h \\ \frac{g_{0h}^{n0}}{I_1 + I_2}, & s = 2 \ \& \ j = h \\ \frac{g_{0l}^{n0}}{I_2}, & s = 2 \ \& \ j = l \end{cases} \quad (6)$$

where $I_1 = g_{0l}^{n0}$ and $I_2 = I_{2h} + I_{2l}$. $g_{0j}^{n0}$ denotes the i.i.d exponentially distributed channel gain for the link between the $j$ ranked device from the cluster indexed $i = 0$ and the origin. As we are considering full power inversion, the received power from the device of interest is $g_{0j}^{n0}$.

As discussed, for a device to successfully transmit to the serving DA, the received signal should have SIR above $\tau_a \in \{\tau_{ah}, \tau_{al}\}$, depending on the device's rank. Accordingly, the conditional transmission success probability of a device associated with the DA located at the origin, $\Theta(\Xi_{0j}^{ns})$, given the NOMA ranking $j$ of the device and the number of scheduled devices $s$ on the $n^{th}$ sub-channel, is given by

$$\Theta(\Xi_{0j}^{ns}) = \begin{cases} P\left(\Xi_{0h}^{n1} > \tau_{ah}\right), & s = 1 \ \& \ j = h \\ P\left(\Xi_{0h}^{n2} > \tau_{ah}\right), & s = 2 \ \& \ j = h \\ P\left(\Xi_{0l}^{n2} > \tau_{al} \cap \Xi_{0h}^{n2} > \tau_{ah}\right), & s = 2 \ \& \ j = l. \end{cases} \quad (7)$$

Due to complexity of the interference components, there is no closed form expression of $\Theta(\Xi_{0j}^{ns})$. As shown in Appendix B, the probability terms in (7) have Laplace transform for the interference components experienced by the devices. For $\alpha \neq 4$, the integrals within the Laplace transform cannot be solved and hence a closed form expression is unattainable. On the other hand, in the case of $\alpha = 4$, a closed form expression for the conditional transmission success probability of a device can be found, given by (see Appendix B)

$$\Theta(\Xi_{0h}^{n1}) = \prod_{j \in \{h,l\}} \exp\left\{-\frac{\sqrt{\tau_{ah}}\lambda_d^j}{\lambda_d}\left(\tfrac{\pi}{2} - \arctan\left(\left(\sqrt{\tau_{ah}}\right)\right)\right)\right\} \quad (8)$$

$$\Theta(\Xi_{0h}^{n2}) = \mathcal{J}\left(\prod_{j \in \{h,l\}} \exp\left\{-\frac{\sqrt{\tau_{ah}}\lambda_d^j}{\lambda_d}\left(\tfrac{\pi}{2} - \arctan\left(\left(\sqrt{\tau_{ah}}\right)\right)\right)\right\}\right) \quad (9)$$

$$\Theta(\Xi_{0l}^{n2}) = \mathcal{J}\left(\prod_{j \in \{h,l\}} \exp\left\{-\frac{\sqrt{A}\lambda_d^j}{\lambda_d}\left(\tfrac{\pi}{2} - \arctan\left(\left(\sqrt{A}\right)\right)\right)\right\}\right) \quad (10)$$

where $A = \tau_{ah} + \tau_{al}(1 + \tau_{ah})$, and $\mathcal{J} = \frac{1}{1+\tau_{ah}}$ is a result of the fact that in the case when two devices co-occupying a channel, they experience both intra- and inter-cluster interference components. It should be noted that for other $\alpha$ values, numerical evaluations can be used to compute the probabilities.

To find the device-DA transmission success probability, we need to average $\Theta(\Xi_{0j}^{ns})$ with respect to the number of scheduled devices on a sub-channel. That is, the long term device-DA transmission success probability, $\bar{\Theta}$, is given by

$$\bar{\Theta} = P(S^n = 1)\Theta(\Xi_{0h}^{n1}) + \frac{P(S^n=2)}{2}(\Theta(\Xi_{0h}^{n2}) + \Theta(\Xi_{0l}^{n2})). \quad (11)$$

Accordingly, the device-DA coverage probability, $C_d$, defined as the probability that an active device is able to successfully transmit its data to the serving DA in the current transmission frame, is given by

$$C_d = \bar{\Lambda}_s \cdot \bar{\Theta}. \quad (12)$$

### C. DA-BS coverage probability

A dedicated uplink channel, SUCH, is utilized for DA-BS transmissions, consisting of $\mathcal{K}$ sub-channels of equal bandwidth. Let RV $K$ denote the number of DAs in the coverage area $\mathcal{V}_b$ of a BS, which has the PMF given by [21]

$$P(K = k) \approx \frac{(\lambda_b)^c(\lambda_a)^m\Gamma(m + c)}{\Gamma(c)(\lambda_b + \lambda_a)^{m+c}m!}. \quad (13)$$

Thus, we have

$$\mathcal{K} = E_k[K] \approx \left\lceil \frac{\kappa\zeta\Gamma(c + 1)}{(1 - \zeta)^{c+1}} \right\rceil \quad (14)$$

where $\kappa = [(\lambda_b)^c/(\Gamma(c)(\lambda_b + \lambda_a)^c)]$ and $\zeta = [\lambda_a/(\lambda_b + \lambda_a)]$.

DAs experience only inter-cell interference. Accordingly, locations of interfering DAs on the $k^{th}$ sub-channel can be modelled by an independent PPP $\Psi_a^k = a_1^k, a_2^k, ...$ with density $\lambda_a^k = \lambda_b$. Similar to device-DA transmission success probability, for a DA to have successful transmission, the received signal at its serving BS should have SIR above $\tau_b$. Let $\Xi_a^k$ denote the received SIR at the BS located at the origin from its associated aggregator located at $a_0^k$ and transmitting on the $k^{th}$ sub-channel, given by

$$\Xi_a^k = \frac{||a_0^k||^{(\epsilon-1)\alpha} g_{00}^a}{I_a} \quad (15)$$

where $I_a = \sum_{a_i^k \in \Psi_a^k} ||a_i^k - b_i||^{\epsilon\alpha} ||a_i^k||^{-\alpha} g_{i0}^a$ is the inter-cell interference experienced by the DA of interest, and $g_{i0}^a$ is the channel gain for the link between the DA associated with the $i^{th}$ BS and the origin, for $i = 0, 1, 2, ...$ where $i = 0$ refers to the BS at the origin.

Let $\Theta(\Xi_a^k)$ denote the DA-BS coverage probability for the DA located at $a_0^k$ and associated with the BS located at the origin and transmitting on the $k^{th}$ sub-channel, which is given by (for details, see Appendix D in [21])

$$\Theta(\Xi_a^k) = P\left\{\Xi_a^k > \tau_b\right\} = 2\pi\lambda_b \int_0^\infty r e^{-(\pi\lambda_b r^2)} \mathscr{L}_{I_a}\left\{\tau_b r^{\alpha(1-\epsilon)}\right\} dr \quad (16)$$

where $\mathscr{L}_{I_a}\{x\}$ is the Laplace transform of $x$ with respect to $I_a$, given by

$$\mathscr{L}_{I_a}\left\{\tau_b r^{\alpha(1-\epsilon)}\right\} = E_{I_a}\left[e^{\tau_b r^{\alpha(1-\epsilon)} I_a}\right] = \exp\left\{-2\pi\lambda_a^k \int_y^\infty \left(1 - \int_0^\infty \left(\frac{2\pi\lambda_a y e^{-\pi\lambda_a y^2}}{1 + \tau_b r^{\alpha(1-\epsilon)} y^{\alpha\epsilon} x^{-\alpha}}\right) dy\right) x\, dx\right\}$$

where $r = ||a_0^k||$ is used to denote the distance between the DA of interest and the origin, $y = ||a_i^k - b_i||$ is used to denote the distance between an arbitrary DA located at $a_i^k$ and its serving BS located at $b_i$, and $x = ||a_i^k||$ is used to denote the distance between an arbitrary DA located at $a_i^k$ and the origin. Note that, $r$, $y$ and $x$ are RVs following Rayleigh distributions with parameters $\lambda_b$, $\lambda_a$ and $\lambda_a^k$ respectively [21], [23].

As a result, the long term device-BS total coverage probability for a device in a NOMA-enabled two-hop network with maximum of two paired devices on a sub-channel is given by

$$C_{tot} = C_d \cdot \Theta(\Xi_a^k). \quad (17)$$

## IV. Delay performance

Average transmission delay is the expected time duration from the instant that a packet is generated at an IoT device to the instant that it is successfully received at the serving BS. In our system, the end-to-end delay is composed of two parts: the first hop delay which is the expected time that a packet spends in the queue of an IoT device until it is successfully received at the serving DA, and the second hop delay which is the expected time that the packet spends in the queue of a DA until it is successfully transmitted towards the serving BS. We model the two-hop transmission paradigm as a Tandem queue, as shown in Figure 2 where $\gamma$ is the packet arrival rate at a device, $\gamma_a$ is the packet arrival rate at a DA, and $\mu_a$ is the departure rate from a DA.

Packets are independently generated at each IoT device according to a Poisson process with parameter $\gamma$. Assuming infinite queue space, the arrival and departure processes at an IoT device can be modelled as M/G/1 queue, where service rate is dependent on the device's ranking on the sub-channel as well as the SIR value. On the other hand, for a DA, the packet arrival flow consists of the aggregation of all the service processes from its scheduled devices in the current transmission frame, while its service rate depends on the SIR value of its transmissions at the serving BS. Consequently, the arrival and departure processes at a DA can be modeled as G/G/1 queue under infinite queue space. In this section, we establish mathematical models for the arrival and departure processes at the devices and DAs based on queuing theory and stochastic geometry. The models are then used to find the overall average transmission delay based on the two-hop tandem queuing model.

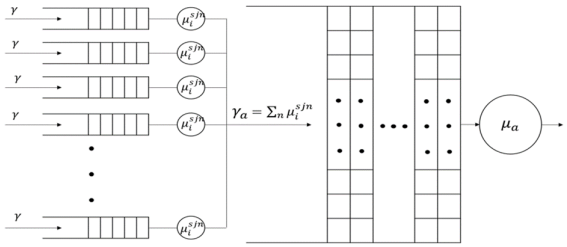



Fig. 2: Tandem queue model of the proposed two-hop NOMA-enabled transmission network

### A. Device-DA queue analysis

For a packet to be transmitted from a device to the DA, the device first needs to access the DA. Accessing the DA in a transmission frame means that the device is scheduled on a NOMA sub-channel and is in the DA coverage. Let $\Gamma_d$ denote the number of transmission frames needed for a device to successfully access the serving DA and transmit its data, which follows a geometric distribution with the probability of success given by device-DA coverage probability in (12). Thus, the average number of transmission frames needed is given by $\bar{\Gamma}_d = \frac{1}{C_d}$. As an approximation, we consider the rounded up representation of $\bar{\Gamma}_d$ and thus, packet generation and transmission at a device, over the period of $\Gamma_d^* = T_f \cdot \lceil \bar{\Gamma}_d \rceil$, can be modeled as an M/G/1 queue with packet arrival rate $\gamma$. As for the departure rate, a device, scheduled on a NOMA sub-channel, transmits its data during the aggregation period of the transmission frame during which it accessed its DA. Let $\mu_d^{sjn}$ denote the packet service rate of the $j$ ranked device scheduled on the $n^{th}$ sub-channel that has $s$ multiplexed devices during the period of $\Gamma_d^*$, given by

$$\mu_d^{sjn} = \left\lfloor \frac{T_a \mathcal{R}_d^{sjn}}{\mathcal{D}} \right\rfloor, s \in \{1,2\}, j \in \{h,l\}, n = 1, 2, ..., \mathcal{N} \quad (18)$$

where we use the flooring function to avoid fractions and to emphasize that a device transmits a packet if and only if the whole packet can be received by the serving DA in the current transmission frame. $\mathcal{R}_d^{sjn}$ denotes the average achievable bit rate by a $j$ ranked device located at $d$ and scheduled on the $n^{th}$ sub-channel with $s$ scheduled devices. Using Shannon's channel capacity formula, coverage probabilities given in (7), and SIRs given in (6), $\mathcal{R}_d^{sjn}$, conditioned on the device is in coverage (i.e., $\Xi_{0j}^{ns} > \tau \in \{\tau_{ah}, \tau_{al}\}$), can be derived as

$$\mathcal{R}_d^{1hn} = E\left[\omega_n \log_2\left(1 + \Xi_{0h}^{n1}\right) \middle| \Xi_{0h}^{n1} > \tau_{ah}\right]. \quad (19)$$

$$\mathcal{R}_d^{2hn} = E\left[\omega_n \log_2\left(1 + \Xi_{0h}^{n2}\right) \middle| \Xi_{0h}^{n2} > \tau_{ah}\right]. \quad (20)$$

$$\mathcal{R}_d^{2ln} = E\left[\omega_n \log_2\left(1 + \Xi_{0l}^{n2}\right) \middle| \Xi_{0l}^{n2} > \tau_{al} \cap \Xi_{0h}^{n2} > \tau_{ah}\right] \quad (21)$$

Notice that, as the closed form expression for the PDF of the SIR is not available, the expectation terms in of the data rate cannot be solved. However, what the expectation signifies is the fact that, when a device is in the coverage of its serving DA, it achieves a certain average bit rate. This average bit rate is lower bounded by a fixed bit rate allocation given by

$$\mathcal{L}_d^{sjn} = \begin{cases} \omega_n \log_2(1+\tau_{ah}), & s=1 \;\&\; j=h \\ \omega_n \log_2(1+\tau_{ah}), & s=2 \;\&\; j=h \\ \omega_n \log_2(1+\tau_{al}), & s=2 \;\&\; j=l \end{cases} \quad (22)$$

Another possible approximation of $\mathcal{R}_d^{sjn}$ is the unconditional average achievable bit rate which can be derived by removing the condition from the expectation terms, and making use of the definition $E[X] = \int_{t>0} P(X > t)\, dt$, as the logarithm function is non-negative. Accordingly, the unconditional achievable bit rate, $\mathcal{F}_d^{sjn}$, is given by (see appendix C)

$$\mathcal{F}_d^{1hn} = \int_{t>0} \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt \quad (23)$$

$$\mathcal{F}_d^{2hn} = \int_{t>0} E\left[\exp\{-\mathcal{B} g_{0l}^{n0}\}\right] \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt \quad (24)$$

$$\mathcal{F}_d^{2ln} = \int_{t>0} \mathscr{L}_{I_{2h}}\{\mathcal{B}\}\, \mathscr{L}_{I_{2l}}\{\mathcal{B}\}\, dt \quad (25)$$

where $\mathcal{B} = 2^{\frac{t}{\omega_n}} - 1$.

Therefore, for the purpose of this analysis, we assume that the average achievable bit rate by a device is given by

$$\mathcal{R}_d^{sjn} = \max\{\mathcal{L}_d^{sjn}, \mathcal{F}_d^{sjn}\}, \quad \text{for } s \in \{1,2\} \;\&\; j \in \{h,l\}. \quad (26)$$

### *B. DA-BS queue analysis*

Similar to the case of a device, for a DA to be able to transmit data to the serving BS, it first needs to access the BS. Let $\Gamma_a$ denote the number of transmission frames needed for a DA to be able to transmit its data successfully to the serving BS, which is modelled as a geometric distribution with success probability $\Theta(\Xi_a^k)$ given in (16). Accordingly, the average number of transmission frames needed for a DA to successfully access its serving BS is given by $\bar{\Gamma}_a = 1/\Theta(\Xi_a^k)$. To be able to model the behaviour of packet arrivals and departures at a DA, we focus on the period $\Gamma_a^* = T_f \cdot \lceil \bar{\Gamma}_a \rceil$, which is an approximation for the average access delay in terms of an integer number of transmission frames.

Once a DA accesses the BS, it transmits its data during the relay phase of the current transmission frame. Therefore, during the period of $\Gamma_a^*$, a DA can transmit only for a period of $T_r < \Gamma_a^*$. Since the transmission rate of a DA depends on the amount of allocated resources and its SIR value at the serving BS, using Shannon's channel capacity formula, SIR given in (15), and interference model described in Section-III, the average achievable bit rate, $\mathcal{R}_a^k$, conditioned on the DA being covered by the serving BS, by a DA located at $a$ and transmitting its data on the $k^{th}$ sub-channel to the BS located at the origin is given by

$$\mathcal{R}_a^k = E\left[\frac{\Omega_b}{\mathcal{K}} \log_2(1+\Xi_a^k) \,\middle|\, \Xi_a^k > \tau_b\right]. \quad (27)$$

Similar to the case of a device, the conditional expectation given in (27) cannot be solved as the PDF of $\Xi_a^k$ is not available. Thus, the average achievable bit rate by a DA in coverage is approximated by its lower bound, $\mathcal{L}_a^k = \frac{\Omega_b}{\mathcal{K}} \log_2(1+\tau_b)$, or the unconditional average achievable bit rate, given by

$$\mathcal{F}_a^k = E\left[\frac{\Omega_b}{\mathcal{K}} \log_2(1+\Xi_a^k)\right] \stackrel{a}{=} \int_0^\infty \left(2\pi\lambda_a \int_0^\infty y e^{-(\pi\lambda_a y^2)} \mathscr{L}_{I_a}\left\{\mathcal{B}_a y^{\alpha(1-\epsilon)}\right\} dy\right) dt \quad (28)$$

where we use the definition $E[X] = \int_{t>0} P(X > t)\, dt$ to compute the expectation of the logarithmic term with respect to $\Xi_a^k$, and the Laplace term in (a) follows from the DA-BS coverage probability given in (16) with $\mathcal{B}_a = 2^{\frac{t\mathcal{K}}{\Omega_b}} - 1$ substituted for $\tau_b$. Accordingly, the average achievable bit rate by a covered DA is given by

$$\mathcal{R}_a^k = \max\{\mathcal{L}_a^k, \mathcal{F}_a^k\}. \quad (29)$$

The average number of packets that a DA can transmit during the period $\Gamma_a^*$ is given by

$$\mu_a = \left\lfloor \frac{T_r \mathcal{R}_a^k}{\mathcal{D}} \right\rfloor \quad (30)$$

where, just like in (8), the floor function $\lfloor \cdot \rfloor$ is to ensure that a packet is transmitted from a DA if and only if it can be fully received at the BS during the current transmission frame.

Packets arrive at a DA from its scheduled devices in each transmission frame. The average number of packets arriving at a DA depends on the number of associated devices, the probability of successful scheduling on a NOMA sub-channel, and the SIR values of the received signals, and the length of the aggregation phase of each transmission frame. Accordingly, let $\gamma_a^i$ denote the packet arrival rate at a DA during the $i^{th}$ transmission frame, given by

$$\gamma_a^i = \mathcal{N} \cdot \left(P(S^n=1)\mu_d^{1hn} + P(S^n=2)(\mu_d^{2hn} + \mu_d^{2ln})\right). \quad (31)$$

Following from the preceding analysis, for the period of $\Gamma_a^*$, packet arrivals and transmissions at a DA can be modelled as a G/G/1 queue with packet departure rate given in (30), and packet arrival rate given by

$$\gamma_a \approx \gamma_a^i * \lceil \bar{\Gamma}_a \rceil. \quad (32)$$

In summary, the proposed NOMA-enabled two-hop architecture can be modelled as a tandem queue with packet arrival rate, $\gamma$, at the first queue (i.e., at a device), arrival process at the second queue (i.e., at a DA) having $\gamma_a$ packet arrival rate, and departure process from the second queue with departure rate of $\mu_a$. Based on the performance measures of the two-stage tandem queues, the expected end-to-end system delay experienced by a packet transmitted by a device located at $d$ to a BS at the origin, going through a DA located at $a$, under the first in first out service rule, is given by [24]

$$\Delta t_{e2e} = \frac{\gamma_a + \mu_a - 2\gamma_d}{\gamma_a\mu_a - (\gamma_a+\mu_a)\gamma_d + (\gamma_d)^2}. \quad (33)$$

## V. Energy consumption

We define energy consumption of a network component as the energy consumed by a device or a DA in the process of successfully transmitting/relaying a single data packet of size $\mathcal{D}$ bits towards the serving BS. We model the devices and DAs as wireless transceivers [11]. A transceiver consists of four major energy consuming blocks: transmission block (TX), receiver block (RX), local oscillator block (LO), and power amplification (PA) block. The power consumption of a transceiver when acting as a transmitter, $Q_T$, and as a receiver, $Q_R$, are given respectively by

$$Q_T = Q_{PA} + Q^i_{LO} + Q_{TX} \tag{34}$$

$$Q_R = \mathcal{M}_a Q^i_{LO} + Q_{RX} + \mathbb{1}(S^n = 2)Q_{SIC} \tag{35}$$

where the power consumption of the RX and TX blocks are denoted by $Q_{RX}$ and $Q_{TX}$ respectively; $Q^i_{LO} \in \{Q^d_{LO}, Q^a_{LO}\}$ where $Q^d_{LO}$ and $Q^a_{LO}$ are non-negative fixed power consumption constants of the LOs of a device and a DA respectively; PA power consumption is denoted by $Q_{PA} \in \{Q^t_d, Q^t_a\}$ for a device and a DA respectively, where $Q^t_d = q_d||d-a||^\alpha$ and $Q^t_a = q_a||a-b||^{\epsilon\alpha}$; $Q_{SIC}$ is a fixed amount of power consumption by a DA to perform SIC on a sub-channel with more than one scheduled device.

### A. Energy consumption of a device

Energy consumption of a device to transmit a data packet is divided into two parts: energy consumed to successfully access the serving DA, and energy consumed to successfully transmit the data packet. We focus on a device with a single data packet in its queue.

Every time the device is scheduled on a NOMA sub-channel, it attempts to access its serving DA. If the device fails to access, it gives up its scheduled sub-channel and waits until the next time it is scheduled. The device attempts access $\lceil \bar{\Gamma}_d \rceil$ times until it gains access. Out of those attempts, the device is scheduled $W$ times, which is a binomial random variable with parameters $\lceil \bar{\Gamma}_d \rceil$ and $\bar{\Lambda}_s$. Considering that every time a scheduled device attempts to transmit its data packet and fails, it consumes a fixed amount of energy denoted by $\mathcal{E}^f_d$, the access energy consumption of a device is $\mathcal{E}_{d,1} = \lceil \bar{\Gamma}_d \rceil \cdot \bar{\Lambda}_s \cdot \mathcal{E}^f_d$.

Once the device accesses the DA, it successfully transmits its data packet over the allocated sub-channel. Let $\Delta t_x$ denote the average time a device takes to transmit a single data packet of size $\mathcal{D}$ to its serving DA, given by

$$\Delta t_x = \frac{\mathcal{D}}{P(S^n = 1)\mathcal{R}^{1h}_d + P(S^n = 2)(\mathcal{R}^{2h}_d + \mathcal{R}^{2l}_d)}. \tag{36}$$

The associated average transmission energy consumption is

$$\mathcal{E}_{d,2} = \Delta t_x \left( E[Q^t_d] + Q^d_{LO} + Q_{TX} \right) \tag{37}$$

where $E[Q^t_d] = \int_0^\infty 2q_d \pi \lambda_a r^{(\alpha+1)} e^{-\pi\lambda_a r^2} dr$, as $Q^t_d$ is a function of the distance between the device and the DA having a Rayleigh distribution with parameter $\lambda_d$.

Thus, the total average energy consumption of a device to transmit a single data packet of size $\mathcal{D}$ bits in the proposed NOMA-enabled two hop network is $\mathcal{E}_d = \mathcal{E}^1_d + \mathcal{E}^2_d$.

### B. Energy consumption of a DA

Different from an IoT device, a DA consumes energy in two forms: as a receiver while receiving data from its cluster members, and as a transmitter while relaying the aggregate data packets to the BS. We focus on a DA located at $a$ and associated with the BS at the origin.

The total time for a device to transmit a single data packet of size $\mathcal{D}$ bits to the DA is $\Delta t_x$. The DA needs to stay awake for this time period to successfully receive the data packet. Thus, let $\mathcal{E}_{a,1}$ denote the amount of energy that the DA consumes to receive a single data packet, given by

$$\mathcal{E}_{a,1} = \Delta t_x \left( Q^a_{LO} + Q_{RX} + P(S^n = 2)Q_{SIC} \right). \tag{38}$$

The DA then attempts to relay the data packet and in the process consumes energy in two forms: energy consumed to successfully access the BS, and energy consumed to successfully transmit the packet. The average number of transmission frames needed for the DA to successfully access the BS is $\lceil \bar{\Gamma}_a \rceil$. Consider that a DA consumes a fixed amount of power, $\mathcal{E}^f_a$, in every access attempt. Then, the average access energy consumption is $\mathcal{E}_{a,2} = \mathcal{E}^f_a \lceil \bar{\Gamma}_a \rceil$.

Once the DA successfully accesses the BS, it transmits the data packet. The transmission energy consumption for the data packet is given by

$$\mathcal{E}_{a,3} = \Delta t_a \left( E_Y[Q^t_a] + Q^a_{LO} + Q_{TX} \right) \tag{39}$$

where $E_Y[Q^t_a] = \int_0^\infty 2q_a \pi \lambda_b y^{(\epsilon\alpha+1)} e^{-\pi\lambda_b y^2} dy$, where we used $y$ to denote the random transmission distance between the DA and the BS which follows a Rayleigh distribution with parameter $\lambda_b$, and $\Delta t_a = \mathcal{D}/\mathcal{R}^a_k$.

Therefore, the total energy consumption of a DA to relay a single data packet of size $\mathcal{D}$ bits in the proposed NOMA-enabled two hop network is $\mathcal{E}_a = \mathcal{E}_{a,1} + \mathcal{E}_{a,2} + \mathcal{E}_{a,3}$.

## VI. Numerical Results and Discussion

We evaluate the developed models via computer simulations for the scheduling probability, device-DA coverage probability, total coverage probability, system delay, device energy consumption, and DA energy consumption, as functions of DA density $\lambda_a$ and thresholds $\tau_{ah}$, $\tau_{al}$ and $\tau_b$. The numerical results are validated by means of independent system level simulations. Table II lists parameter values used.

TABLE II. Simulation parameter values

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $\mathcal{D}$ | 10 bits/packet | $\mathcal{E}^f_d$ | 0.05 mJ |
| $\mathcal{E}^f_a$ | 0.05 mJ | $\mathcal{N}$ | 10 |
| $Q_{SIC}$ | 0.1 mW | $Q^d_{LO}$ ($Q^a_{LO}$) | 0.1 mW |
| $Q_{RX}$ ($Q_{TX}$) | 0.1 mW | $q_d$ ($q_a$) | 1 |
| $T_f$,($T_a$,$T_r$) | 10 ms (5 ms, 5 ms) | $\alpha$ | 4 |
| $\gamma$ | 10 packets/s | $\epsilon$ | 0.8 |
| $\lambda_b$ | 1k dev/km$^2$ | $\lambda_d$ | 2 BS/km$^2$ |
| $\Omega_b$ | 50 kHz | $\omega_n$ | 5 kHz |

Simulations are performed based on the system model in Section II using MATLAB for a 200 $km$ x 200 $km$ 2D plane. Each result is an average of 1000 independent Monte Carlo simulation runs. Locations of BSs, DAs, and devices are randomly generated based on their deployment densities. The

wireless channel is simulated with the 3GPP non-line-of-sight (NLOS) path loss model given in [25]. To simulate Rayleigh fading, random channel gains are generated according to an exponential distribution with unity mean and incorporated into the received signal power of all transmissions. Maximum received signal strength association rule is employed. Through extensive simulations, the derived models are shown to be functional and accurate as evident by the close match between the numerical and simulation results.

Figure 3 shows the scheduling probability of the devices for NOMA and OMA configurations versus DA density $\lambda_a$ between 95 and 435 DAs/$km^2$. For both NOMA and OMA, the scheduling probability increases as DA density increases. As expected, the scheduling probability is higher for the NOMA as compared to OMA, thanks to its ability to allow multiplexing of more than one device on the same sub-channel.

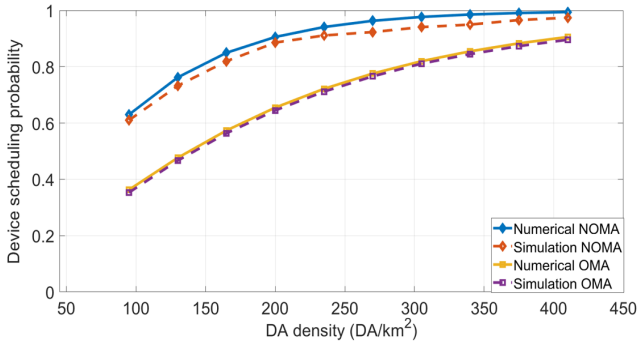



Fig. 3: Scheduling probability of the devices as a function of DA density ($\lambda_a$)

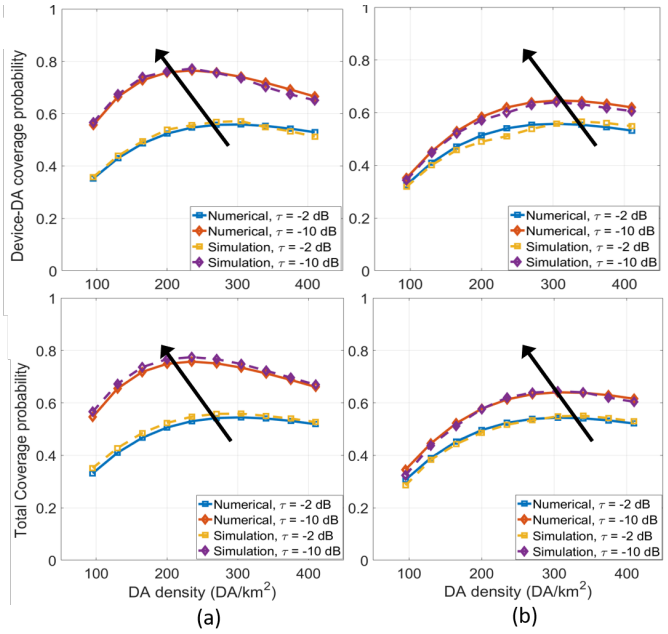



Fig. 4: Device-DA coverage probability (top) and total coverage probability (bottom) as a function of DA density ($\lambda_a$) and network thresholds ($\tau$); (a) NOMA; (b) OMA

Figure 4 shows device-DA coverage probability ($C_d$) and total coverage probability ($C_{tot}$) versus DA density $\lambda_a$, and network thresholds $\tau_{ah} = \tau_{al} = \tau_b = \tau$. The $C_{tot}$ values are slightly lower than those of $C_d$, as $C_{tot}$ takes into account the performance of both hops. However, it is clear that the device-DA coverage is more dominant in our system, which is expected due to the extensive interference among the devices and the high contention over the limited number of in-cluster sub-channels. As both coverage probabilities follow the same trends, we focus our analysis only on $C_d$. For both NOMA and OMA, as $\lambda_a$ increases from 95 DAs/$km^2$ to 435 DAs/$km^2$, $C_d$ increases to a maximum and then drops in a concave manner. This concave pattern can be attributed to the fact that, as $\lambda_a$ increases, the number of interfering devices on a sub-channel increases, leading to a decrease in the probability of successful transmission from a device to the serving DA. For NOMA, the maximums occur at densities of 305 DA/$km^2$, for $\tau = -2$ $dB$, and 235 DA/$km^2$, for $\tau = -10$ $dB$, whereas in the case of OMA, the maximum for both $\tau$ settings happens at 305 DA/$km^2$. Accordingly, we can conclude that there is a $\lambda_a$ that maximizes $C_d$, given a fixed density of devices for both NOMA and OMA. Also, a lesser number of clusters is needed in the case of NOMA to maximize performance. On the other hand, for both configurations, the coverage probability improves as $\tau$ drops, and the improvement for NOMA is more significant.

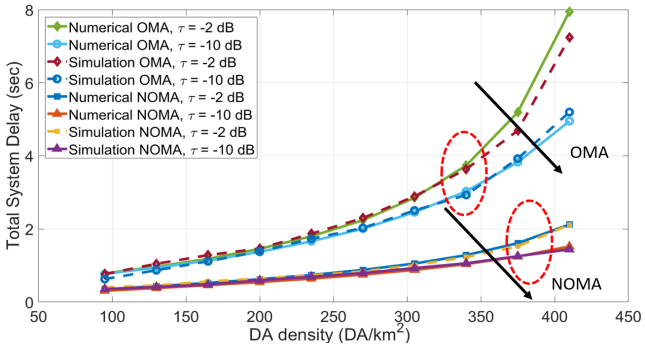



Fig. 5: Total system delay as a function of DA density ($\lambda_a$) and network thresholds ($\tau$)

Figure 5 shows the total system delay versus DA density for multiple network thresholds for both NOMA and OMA. NOMA has a lower average system delay as compared to OMA. The difference is further aggravated at higher DA densities. Thanks to its ability to multiplex multiple devices on the sub-channels, NOMA not only allows scheduling a larger number of devices per transmission frame, but also provides a better coverage probability, especially for lower network threshold settings. Accordingly, packet departure rate from a device is higher for NOMA, leading to shorter queuing delays and overall improvement in system delay. On the other hand, the delay increases as $\lambda_a$ increases as more DAs share a limited bandwidth, leading to a lower achievable bit rate in packet departure from a DA.

Figure 6 shows the average energy consumption of device to transmit a packet of size $\mathcal{D}$ bits to the serving DA for both NOMA and OMA configurations versus $\lambda_a$ and $\tau$. For both

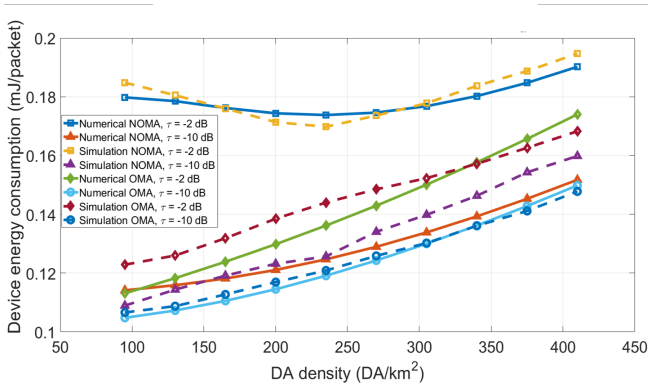



Fig. 6: Average device energy consumption as a function of DA density ($\lambda_a$) and network thresholds ($\tau$)

NOMA and OMA, energy consumption increases as $\lambda_a$ and $\tau$ values increase except for NOMA at $\tau = -2dB$. The higher the $\tau$, the higher the number of transmission frames required before a device can access the serving DA. At higher $\lambda_a$ values, interference among devices on a sub-channel increases, leading to a higher number of failed attempts and in turn higher levels of energy wastage. For NOMA at $\tau = -2dB$, the impact of $\tau$ is more dominant than the effect of $\lambda_a$, because devices experience higher interference than that under OMA. Thus, as $\tau$ increases, the SIR performance decreases at a faster rate for NOMA. Further, for the same $\tau$ setting, devices under NOMA configuration consume more energy to transmit a data packet compared to under OMA for all DA densities, because the average achievable bit rate by a device is lower in the case of NOMA due to intra-cluster interference. A lower bit rate results in a longer time that a device takes to transmit a single data packet, leading to higher device energy consumption.

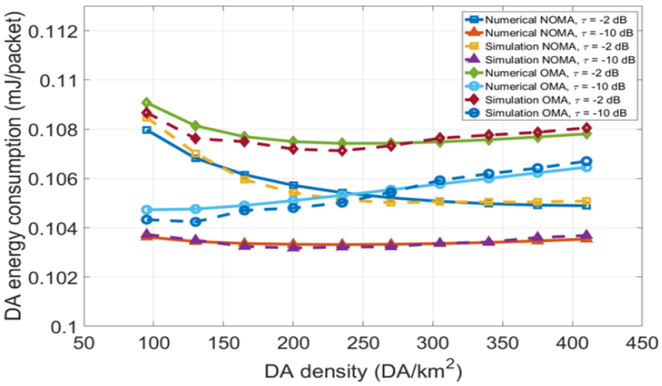



Fig. 7: Average DA energy consumption as a function of DA density ($\lambda_a$) and network thresholds ($\tau$)

The average energy consumed by a DA to relay a data packet of size $\mathcal{D}$ bits, can be of a concern, especially for DAs selected from the IoT devices. As shown in Figure 7, for the same $\tau$ setting, DAs consume less amount of energy under NOMA. As $\lambda_a$ increases, DA energy consumption either decreases (in the case of $\tau = -2$ $dB$), or is almost stable (in the case of $\tau = -10$ $dB$). On the other hand, for OMA, as DA density increases, DA energy consumption increases for all $\tau$. Overall, NOMA is more energy efficient for the DAs.

Based on the preceding results, for a two-hop NOMA enabled data aggregation architecture, the following general conclusions can be made: i) NOMA for in-cluster transmissions can improve the scheduling probability of the devices in the case of limited transmission resources; ii) NOMA can improve the total coverage probability of the two-hop network and in turn increase the number of supported devices per transmission frame; iii) with NOMA, end-to-end delay can be improved with the proper choice of network parameters such as DA density and amount of available resources; and iv) with a proper interference mitigation mechanism for the in-cluster NOMA transmissions and the appropriate choice of the DA density, NOMA can improve the overall energy efficiency of both DAs and devices. Based on the results presented in this study, our proposed NOMA enabled architecture has many potentials and further investigations are needed to investigate the energy and delay optimality design of two-hop NOMA-enabled transmission architecture for future massive cellular IoT applications.

## VII. Conclusion

In this paper, we propose a novel two-hop NOMA enabled data aggregation architecture to enable massive cellular IoT applications. Concepts and techniques from stochastic geometry and queuing theory are jointly exploited to derive tractable models for various network performance measures including scheduling probability, coverage probability, system delay, and energy consumption. We abstract the network topology by spatially modeling the locations of the BSs, DAs and active devices using three independent homogeneous PPPs, and characterize intra- and inter-cluster interference components experienced by the devices and characterize the inter-cellular interference experienced by the DAs. All derived models are corroborated via Monte Carlo simulations. Numerical results demonstrate that, the DA density and network thresholds highly impact network performance. In comparison with the traditional two-hop OMA based architecture, the proposed NOMA architecture provides better scheduling and coverage probability, supports a larger number of devices per transmission frame, has a lower average end-to-end system delay, and improves the energy consumption of the devices and DAs.

## References

[1] M. Gerasimenko, V. Petrov, O. Galinina, S. Andreev, and Y. Koucheryavy, "Impact of machine-type communications on energy and delay performance of random access channel in LTE-advanced," *Trans. Emerging Telecom. Tech.*, vol. 24, no. 4, pp. 366–377, Jun. 2013.

[2] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks," in *Proc. IEEE ICC*, May 2016, pp. 1–6.

[3] M.-H. Fong, P. K. Jain, and H.-N. Choi, "Extended access barring," feb 2016, US Patent 9,264,979.

[4] X. Yang, A. Fapojuwo, and E. Egbogah, "Performance analysis and parameter optimization of random access backoff algorithm in LTE," in *Proc. IEEE VTC*, Sep. 2012, pp. 1–5.

[5] T. Lin, C. Lee, J. Cheng, and W. Chen, "PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467–2472, Jun. 2014.

[6] 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, Dec. 2013.

[7] A. Larmo and R. Susitaival, "RAN overload control for machine type communications in LTE," in *Proc. IEEE Globecom Workshops*, Dec. 2012, pp. 1626–1631.

[8] G. Miao, A. Azari, and T. Hwang, "$E^2$ -MAC: Energy efficient medium access for massive M2M communications," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4720–4735, Nov. 2016.

[9] T. Kwon and J. M. Cioffi, "Random deployment of data collectors for serving randomly-located sensors," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2556–2565, Jun. 2013.

[10] J. Guo, S. Durrani, X. Zhou, and H. Yanikomeroglu, "Massive machine type communication with data aggregation and resource scheduling," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 4012–4026, Sep. 2017.

[11] D. Malak, H. S. Dhillon, and J. G. Andrews, "Optimizing data aggregation for uplink machine-to-machine communication networks," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 1274–1290, Mar. 2016.

[12] G. Hattab and D. Cabric, "Energy-efficient massive IoT shared spectrum access over UAV-enabled cellular networks," *arXiv preprint arXiv:1808.08006*, 2018.

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

[14] M. K. Elhattab, M. M. Elmesalawy, and I. I. Ibrahim, "Opportunistic device association for heterogeneous cellular networks with H2H/IoT co-existence under QoS guarantee," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1360–1369, Oct. 2017.

[15] G. Hattab and D. Cabric, "Performance analysis of uplink cellular IoT using different deployments of data aggregators," in *Proc. IEEE Globcom*, Dec. 2018, pp. 1–6.

[16] J. Kim, H. Lee, D. M. Kim, and S. Kim, "Delay performance of two-stage access in cellular internet-of-things networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3521–3533, Apr. 2018.

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

[18] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (noma) for cellular future radio access," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, June 2013, pp. 1–5.

[19] P. Wang, J. Xiao, and L. P, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Veh. Technol. Mag.*, vol. 1, no. 3, pp. 4–11, 2006.

[20] O. L. Alcaraz López, H. Alves, P. H. Juliano Nardelli, and M. Latva-aho, "Aggregation and resource scheduling in machine-type communication networks: A stochastic geometry approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4750–4765, Jul. 2018.

[21] H. G. Moussa and W. Zhuang, "Rach performance analysis for large-scale cellular IoT applications," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3364–3372, Apr. 2019.

[22] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons, 2009.

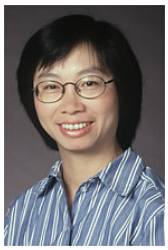
[23] T. D. Novlan, H. S. Dhillon, and J. G. Andrews, "Analytical modeling of uplink cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2669–2679, Jun. 2013.

[24] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer Science & Business Media, 2013, vol. 46.

[25] Q. H. Chu, J. M. Conrat, and J. C. Cousin, "Propagation path loss models for LTE-advanced urban relaying systems," in *Proc. IEEE APSURSI*, Jul. 2011, pp. 2797–2800.


**Hesham Moussa** (S'10) received the B.S. and M.S. degrees from the American University of Sharjah, Sharjah, UAE, in 2013 and 2015, respectively. He is currently pursuing the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada.

His current research interests include artificial intelligence in wireless communication, resource allocation, medium access control, QoS provisioning in wireless networks, performance optimization for wireless machine-to-machine, and Internet of Things communication paradigms.



**Weihua Zhuang** (M'93-SM'01-F'08) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, since 1993, where she is a Professor and a Tier I Canada Research Chair in Wireless Communication Networks.

Dr. Zhuang was a recipient of the 2017 Technical Recognition Award from the IEEE Communications Society Ad Hoc and Sensor Networks Technical Committee, and a co-recipient of several Best Paper Awards from IEEE conferences. She was the Editor-in-Chief of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2007 to 2013, the Technical Program Chair/Co-Chair of IEEE VTC Fall 2017 and Fall 2016, and the Technical Program Symposia Chair of IEEE Globecom 2011. She is an elected member of the Board of Governors and Vice President - Publications of the IEEE Vehicular Technology Society. She was an IEEE Communications Society Distinguished Lecturer from 2008 to 2011. Dr. Zhuang is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada.