A Multi-Rate Queue Management for Delay-Constrained Non-Orthogonal Multiple Access (NOMA) based Secure Cognitive Radio Network

K. Adli Mehr[1]; J. Musevi Niya[2,*]; N. Akar[3]

1- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, e-mail: k.adlimehr@tabrizu.ac.ir,
2- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, e-mail: niya@tabrizu.ac.ir,
3-Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey, e-mail: akar@ee.bilkent.edu.tr
[*]Corresponding author

Received day month year, Revised day month year, Accepted day month year.

Abstract
In this article, we consider a secure cognitive radio network (CRN) deploying non-orthogonal multiple access (NOMA). A mixed delay constrained multicast and unicast traffic is received by the intended CRN receivers, while keeping the traffic secret from the eavesdroppers. Physical layer security (PLS) is deployed to secure the confidential messages of both primary user (PU) and secondary user (SU). A queue management policy (QMP) is proposed to enhance the quality of service (QoS) of CRN. The proposed QMP adaptively uses all or some of the SU's resources towards the transmission of the PU's packet, this decision being based on the packet's delay experienced in the PU queue. Exact delay distribution of PU traffic is derived via a novel multi-regime Markov fluid queue (MRMFQ) model. Thanks to the closed form expressions, the proposed QMP is optimized to provide the highest attainable throughput for SU, while satisfying PU's QoS constraints. It is shown via numerical examples that NOMA based CRN consistently outperforms orthogonal multiple access (OMA) based counterpart. We also show that the performance improvements gained by the proposed QMP depends on the intensity of PU traffic as well as the channel conditions. Moreover, a heuristic suboptimal parameter selection procedure with significantly lesser computational complexity is proposed for less capable devices.

Keywords
NOMA, Physical Layer Security, Queue Management, Multi-Regime Markov Fluid Queue, Cognitive Radio.

1. Introduction

Wireless next generation networks (WNGNs) such as 5G cellular networks, promises ubiquitous connectivity for everyone and everything. As expected, there will be over 25 billion devices connected to cellular networks by 2020, a great portion of them being IoT terminals [1]. It is a notable growth compared to the number of devices operating in the current wireless networks. Considering limited available spectrum, it is required for the wireless next generation networks to adopt a novel multiple access technology to realize spectral efficient massive access. Non-orthogonal multiple access (NOMA) has been introduced as the enabling technology to facilitate the major requirements of WNGNs such as massive connectivity, high data rate, and low latency [2,3,4] . In contrast to traditional orthogonal multiple access (OMA), where the communication resources is allocated to each user in the different time, frequency, or code domains, NOMA superposes multiple users' signals at the transmitter side by exploiting the power or code domain [5]. This concept has been integrated into the cognitive radio [6, 7] as well, which has emerged as a promising solution to cope with spectrum scarcity in WNGN systems. In cognitive radio networks (CRNs), second tier network users with cognition capabilities, dubbed as Secondary Users (SU), have been introduced into the existing communication infrastructure. SUs intelligently monitor the channel and transmit their messages in such a way that the communications of the existing network nodes, named Primary Users (PU), are not affected. There are three distinct deployment paradigms for cognitive radio, namely underlay, interweave, and overlay paradigms [8], where the first and the third paradigms, i.e. underlay and overlay paradigms, are NOMA-based; then PUs and SUs can transmit simultaneously at the same time, frequency, or code domains [8]. In the underlay paradigm, the transmission by SUs is allowed, if the interference generated by these cognitive users in the primary receiver is below a predefined threshold. While, in the overlay paradigm, the secondary transmitter completely cooperates with the primary transmitter to deliver the PU's messages to the intended primary receivers in exchange for its own opportunity for transmission [8]. In this paradigm, the primary message is known at the secondary transmitter causally or non-causally. For instance, a cognitive radio system where the cognitive transmitter has acquired the message of the primary user in the previous phases, or, a cognitive radio system, where both primary and secondary transmitters belong to the same authority (such as cloud empowered

cellular network [10]) are examples of overlay cognitive radio systems.

Since several heterogeneous users share the same spectrum, the secrecy of the messages becomes a critical issue in NOMA, due to the broadcast nature of wireless channels [11]. In order to address the security issue in NOMA, deployment of physical layer security (PLS) is proposed in this paper. Due to inherent interference, PLS is an ideal solution to provide message secrecy for NOMA based systems. In contrast to traditional cryptographic methods, security of physical layer methods is proven via information-theoretic methods. The information-theoretic methods improve the secrecy of the conveyed messages by enhancing the received signal quality at the intended users, while imposing excessive interference to the eavesdroppers in order to deteriorate their decoding capabilities [12]. However, the improved security of PLS is provided in exchange of reduced achievable rate. The NOMA-based cognitive radio system achieves lower rates compared to its insecure counterpart. On the other hand, the demand for higher quality of service (QoS) for the traffics delivered via wireless media is increasing rapidly, which further challenges the proper deployment of secure NOMA-based cognitive radio system. Due to nature of NOMA, inter-user interference inevitably increased compared to conventional OMA, which may result in service interruptions and degraded QoS. This issue is more emphasized in NOMA based CRNs, since SUs should operate totally transparent to PUs and no performance degradation is accepted for PUs. In order to address the QoS challenges, NOMA-based system should operate in harmony with higher layer of the network. In this paper, we introduce two QMPs to efficiently deal with the packet queues held in PU and SU. The main goal of the QMP is to provide as many secure transmission opportunities as possible for the SU (provide maximum secure channel utilization) while satisfying the PU's delay constraint given in terms of the probability that a packet's queuing delay exceeding a given threshold being less than a certain value. In this case, not only the QoS criteria of the PU is satisfied, which was challenged by the reduced achievable rate of PLS, but also, the SU can deliver its messages within the spectrum that originally belongs to PU. The first proposed QMP constitutes an OMA based system, where the PU leverages all of SU's resources unless the PU's queue is empty. The SU is only allowed to transmit its own packets whenever the PU's queue is empty. That is, no simultaneous transmission occurs in this scenario. The second proposed QMP, which is a NOMA based system, adaptively uses all or some of the SU's resources (through a quantity called the power allocation factor), towards the transmission of the PU's packet depending on packet's delay experienced in the PU queue, i.e., delay-aware QMP. In particular, when the delay is less than a delay threshold, i.e., first regime, the SU uses a certain ratio (to be determined so as to maximize the SU's throughput) of its power to transmit its own packets along with the PU's packets, simultaneously. However, when the queue delay experienced by its packets is above the threshold, i.e., second regime, all of SU's resources are devoted to the transmission of PU packets. When the PU queue is empty, the SU is allowed to transmit at the maximum rate. The performance of both proposed QMPs are thoroughly investigated and compared. Moreover, based on the analytical expressions derived using MRMFQ for NOMA-based QMP, the optimal power allocation factor and delay threshold values are numerically obtained to maximize the throughput of the SU while meeting the PU's delay constraint.

In summary, the main contribution of this paper is three-fold: 1) securing a NOMA-based CRN using PLS and deriving its achievable rate region in AWGN channel, 2) proposing a QMP for NOMA-based secure CRN in order to satisfy the QoS constraints of PU and deriving exact performance metrics analytically, which shows the superior performance compared to its OMA-based counterpart, and 3) optimizing the performance of the proposed QMP in terms of power allocation and delay criteria.

The remainder of the paper is organized as follows: In Section 2, related works is addressed. Section 3 presents the system model and the secure achievable rate region. Stochastic model of the MRMFQ in NOMA system is presented in Section 4. In Section 5, parameter optimization of the proposed delay-aware QMP is addressed. Section 6 is devoted to numerical (both simulation and analysis) results. Finally, we conclude.

2. Related Works

Physical Layer Security (PLS) of the unicast NOMA systems is considered in different scenarios in the literature. Cognitive interference channel with two secure messages is considered in [17], where the primary and secondary transmitters send two confidential messages simultaneously while trying to conceal it from the eavesdropper, being the other user's receiver. In [18], security of the primary message is provided against the secondary receiver using PLS in a NOMA based CRN. Inter-user interference is manipulated in [19] to secure the confidential messages against active eavesdroppers in a massive NOMA system.

Secrecy of transmitted messages is enhanced in a NOMA cellular IoT network by deploying stochastic geometry in [20]. Moreover, multicast NOMA systems, where one information bearing message is intended for multiple receivers, are considered in the literature. Beamforming with superposition coding is deployed in [21] to deliver two types of messages, one with a high priority to be decoded by both of two users, and a low priority type intended for one receiver only. Ding et al. investigated the spectral efficiency and security of a NOMA system with simultaneous multi-cast and unicast transmissions [22]. A secure NOMA based cognitive interference channel is considered in [23], where in an overlay scenario SU helps PU to deliver its messages to both primary and secondary users, while transmitting its message, which should be kept secret from the primary receiver, simultaneously. Nevertheless, in all these studies, the authors focused on obtaining the achievable rate regions and did not consider the delay performance of the network. On the other hand, performance of QoS constrained wireless communication systems with PLS is investigated in a few studies. In [24], a cross-layer resource allocation problem is investigated in a wireless single-hop uplink scenario by taking the information-

theoretic secrecy as a QoS metric. A similar problem is studied in [25] to control the uplink of a cellular network by deploying hybrid ARQ (Automatic Repeat reQuest). They maximize a network utility function while keeping the queues stable and meeting secrecy constraint in a block-fading channel. In [26], a secret key queue is deployed to secure the private information in a single-hop wireless communication system over a block fading channel while meeting the delay constraints. Similarly, fair rate allocation is studied in [27] for a secure broadcast channel where unintended receivers are considered as internal eavesdroppers. Instead of devoting all available resources to the user with the best channel condition, secure transmission of less favorable users are facilitated simultaneously by using secret keys, for the sake of fairness. Furthermore, a dynamic network control mechanism is proposed in [28] for a multi-hop wireless network while taking privacy and delay into account. They investigate the optimal rate allocation to preserve the confidentiality of the transmitted data while meeting delay constraints. Moreover, there are a few studies considering cognitive radio in conjunction with higher layer functionalities such as scheduling, routing, application-level throughput, etc. QoS-constrained SU in an underlay cognitive radio relay channel is studied in [29] and the so called effective capacity [30] is derived. This scenario is extended to an energy-constrained SU in [31] where a proper band from a pool of available spectrum bands is to be chosen. The reference [32] considers optimal power allocation for a delay-constrained SU in the underlay secure cognitive radio. In [33], queuing analysis of a cooperative cognitive radio with QoS-constrained PU is considered. Optimum admission control parameters of the primary packets at the SU is derived to maximize the SU's throughput while meeting PU's QoS conditions. Despite the rich literature in regards with the underlay cognitive radio paradigm with QoS-constrained SU, the overlay cognitive radio and QoS constraints of the PU did not get enough attention. Recently, Adli Mehr et al. have considered the QoS of the PU in a two-user CRN [16]. However, the security of the messages conveyed in CRN is neglected. In contrast, PLS is deployed in this paper to secure the confidential messages of both PU and SU, in the presence of the eavesdropper.

3. System Model and Secure Achievable Rate Region

The secure cognitive radio system is depicted in Fig. 1 where g_{ij} refers to channel gains. This model is inspired by the secure overlay cognitive interference channel model presented in [23] with an additional eavesdropper to take into account the security of the PU's messages as well. Since, it can easily be converted to other overlay cognitive radio models by removing the security criterion, it can serve as a general model for overlay cognitive radio.

In this model, the delay-constrained PU intends to multicast its messages (or packets) securely to both primary and secondary receivers in a timely manner in cooperation with the secondary transmitter, while keeping it secret from the external eavesdropper. On the other hand, the SU intends to unicast its messages to the secondary receiver, while keeping the primary receiver and the eavesdropper ignorant of them. It is assumed that the secondary transmitter has non-causal knowledge about the PU's packet. In this scenario, PU and SU, using NOMA, transmit their messages at the same time, frequency, or code. To be precise, SU creates its codeword in cooperation with PU, such that the simultaneous transmission does not affect the PU communication adversely, while delivering its own message. Let P_1 (P_2) be transmission power of the PU (SU) in milliwatts (mW). The SU devotes a portion of its power, i.e., ζP_2, to help the PU deliver its messages whereas the remaining power, i.e., $(1-\zeta)P_2$, is used for transmitting SU's packets.

The achievable rate of the PU and SU, denoted by r_1 and r_2, respectively, as a function of ζ in units of bits per second per Hertz (bps/Hz) can be calculated based on the following. For this purpose, let

$$r_{1j} = C\left(\frac{|g_{1j}|^2 P_1 + |g_{2j}|^2 \zeta P_2}{|g_{2j}|^2 (1-\zeta)P_2 + \sigma_j^2}\right),$$
$$r_{2j} = C\left(\frac{|g_{2j}|^2 (1-\zeta)P_2}{\sigma_j^2}\right), for\ j = 1,2,3. \quad (1)$$

where $C(x) = \frac{1}{2}log(1+x)$, g_{ij}s are the corresponding channel coefficients, and σ_j^2 is the variance of the received noise at receiver j.

The achievable secrecy rate of the PU (r_1) is derived as the difference of the achievable rate of the worst legitimate receiver and that of the eavesdropper, and that of the SU (r_2) is calculated as the difference of the achievable rate of the secondary receiver and that of the strongest illegitimate receiver, see for example [34]:

$$r_1 \le min(r_{11}, r_{12}) - r_{13},\ \ r_2 \le r_{22} - min(r_{21}, r_{22}.) \quad (2)$$

Note that in this scenario, the eavesdropper does not acquire neither PU's nor SU's messages. However, in order to consider the worst case scenario, it is assumed that the eavesdropper knows the PU's message and can cancel its interference out, while it tries to decode the SU's message.

As an illustrative example, the achievable secrecy rate region which is derived by varying the value of ζ is depicted in Fig. 2 for a sample scenario with $g_{11} = 1$, $g_{12} = 0.5$, $g_{13} = 0.25$, $g_{12} = 0.6$, $g_{22} = 1$, and $g_{23} = 0.25$, unit noise variances, $P_1 = 1\ mW$, and $P_2 = 2\ mW$. We denote the region of achievable rates with A, which is essentially the region where points located below the curve in Fig. 2. Also, the set of all points on the curve is denoted by $\mathfrak{C}$. The PU permits the SU to use its licensed band as long as the delay constraints are not violated.

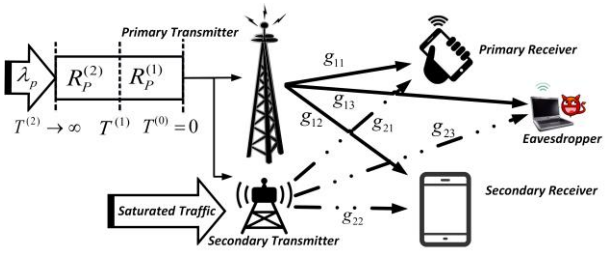

Fig. 1. Secure Overlay Cognitive Radio System with an Eavesdropper.

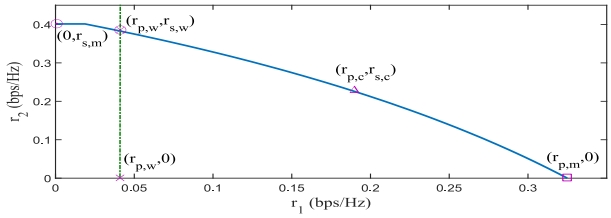

Fig. 2. Achievable secrecy rate region of secure overlay CRN.

A QMP controls the interaction between the PU and the SU transmitters by deciding which and when secure achievable rate pair to use for the transmission of their packets. Obviously, a QMP will select the rate pairs in $\mathfrak{C}$, in order to achieve the best performance. Any rate pair in the curve $\mathfrak{C}$ is denoted by $(r_{p,c}, r_{s,c})$. The maximum achievable rate of the PU (SU) when all of the resources of the SU are allocated to PU's (SU's) packets is denoted by $r_{p,m}$ ($r_{s,m}$), which is located in the intersection of curve $\mathfrak{C}$ and r_1 (r_2) axis. In this article, we suggest two QMPs for the two user CRN with the delay-constrained PU and an eavesdropper, which are described in the following. First, OMA based system, whenever the PU has a packet to transmit, all SU's resources are devoted to the transmission of the packet, i.e., PU's message is transmitted with rate $r_{p,m}$. When the PU's queue is empty, the SU transmits its own packets at the maximum rate, i.e., $r_{s,m}$. Thus, the system with OMA-based system alternates between $(0, r_{s,m})$ and $(r_{p,m}, 0)$, depending on whether the PU's queue is empty or not, respectively. No simultaneous transmission occurs. Second, NOMA-based system, the queue delay experienced by each packet in the primary transmitter is monitored. When a packet gets to be transmitted, its queuing delay is checked. If this delay is below a certain threshold (denoted by $T^{(1)}$), i.e., regime 1, the SU simultaneously transmits its own messages and the PU's packet at rates $r_{s,c}$ and $r_{p,c}$, respectively. If the queue delay is above the threshold $T^{(1)}$, i.e., regime 2, all of SU's resources are devoted to transmit only PU's packets at rate $r_{p,m}$ while the SU's rate is zero. Whenever the primary transmitter's queue becomes empty, the SU achieves the rate $r_{s,m}$ by deploying all its resources. With the appropriate choice of the parameter ζ (or equivalently the rate pair $(r_{p,c}, r_{s,c})$ along with the choice of the threshold $T^{(1)}$, the performance of NOMA based system can be controlled. Note that when $T^{(1)} = 0$, all PU packets are served at a rate $r_{p,m}$, i.e., the system reduces to the OMA based system.

The PU messages are assumed to arrive at the PU transmitter according to a Poisson point process with rate λ_p packets per milliseconds (ms), a random length of each packet with exponential distribution and the mean length L, and the band-limited channel with a bandwidth of $B\ kHz$. Therefore, the service time of a message to be served at rate $r_{i,j}$ for $i \in \{p, s\}$ and $j \in \{c, m, w\}$ during its transmission has an exponential distribution with mean $1/R_{i,j} = L/Br_{i,j}\ ms$. Also, we assume that the SU's traffic is saturated, i.e., it always has packets to transmit. In the next section, we will derive the distribution of the steady-state delay of the PU packets along with the throughput of the SU for both QMPs.

4. Stochastic Model of the MRMFQ in NOMA

In OMA-based system, the packets at the PU transmitter are served with a fixed rate, which equals to $r_{p,m}$. Assuming a Poisson point process for arriving of the PU packets with rate λ_p, we can model the queue of the PU as the well-known $M/M/1$ queue model [13]. Since the secondary user transmits only when the PU queue is empty (with rate $r_{s,m}$), the throughput of the SU, denoted by S_2, can be calculateded as follows in bps:

$$S_2^O = \left(1 - \lambda_p / R_{p,m}\right) B r_{s,m} \tag{3}$$

NOMA-based system uses a delay-aware service mechanism to manage the PU queue, i.e., the transmission rate of both primary and secondary users depends on the delay experienced by the Head of the Line (HoL) message in the PU queue. These mechanisms cannot be modeled with conventional queue models like $M/M/1$. Hence, we use multi-regime Markov fluid queues for modeling NOMA-based system and derive the steady-state distribution of the queue delay. In NOMA-based system, there are two possible service rates to choose for the primary user, when a packet gets to be sent. We define the following thresholds: $0 = T^{(0)} < T^{(1)} < T^{(2)} \to \infty$. Consider a packet at the HoL of the PU queue which is just about to be transmitted at time t. In NOMA based system, a service rate decision is to be made based on the delay already experienced by the packet (queue delay denoted by $D^Q(t)$). Whenever $T^{(0)} < D^Q(t) < T^{(1)}$, the system is in regime 1 and the packet is to be served with rate $R_p^{(1)} = R_{p,c}$, and whenever $T^{(1)} < D^Q(t) < T^{(2)}$, the system is in regime 2 and the packet is to be served with rate $R_p^{(2)} = R_{p,m}$ by devoting all of the resources of the SU to PU's message. Intuitively, it should hold that $R_p^{(2)} > R_p^{(1)}$ for meeting the delay constraints. In order to model the queuing system of the primary user as a Markov fluid queue, three auxiliary random processes are defined. First, the sojourn time process $S(t)$ which is the overall time spent in the system including service time for the packet to be served by the server. If there are no packets in service at time t, then $S(t) = 0$. The unfinished work process $A(t)$ denotes the amount of time needed to serve the waiting packets including the one in the service at time t. It is obvious, a packet that had arrived at time t with $T^{(0)} < A(t) < T^{(1)}$ ($T^{(1)} < A(t) < T^{(2)}$) will eventually be served at rate $R_p^{(1)}$ ($R_p^{(2)}$). The typical paths for the two processes $S(t)$ and $A(t)$ are given in Fig. 3 (a) and Fig. 3 (b), respectively, for an example scenario. The job arrival instants are indicated by the small arrows. Because of abrupt jumps, neither the sojourn time process, nor the unfinished work process are suitable to be modeled as a fluid queue [35]. Then, the random process $X(t)$ is introduced, by replacing the abrupt jumps of the sojourn time process with linear decrements corresponding to

negative unity drifts (see Fig. 3 (c)). Also, it is clear that the steady-state distribution of the process $S(t)$ $(A(t))$ can be derived from that of $X(t)$ by eliminating the states corresponding to negative (positive) drifts.

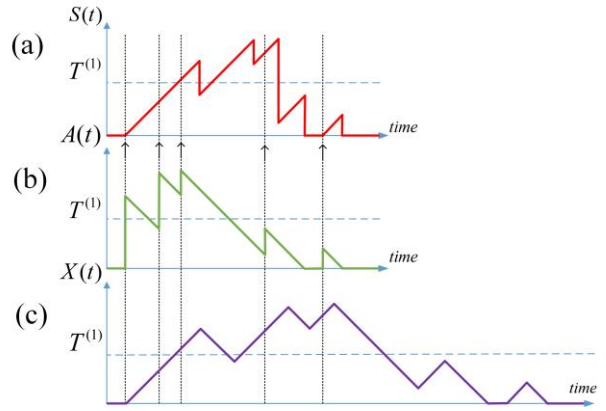

Fig. 3. Sample paths of: (a) the sojourn time process $S(t)$, (b) the unfinished work process $A(t)$, and (c) $X(t)$.

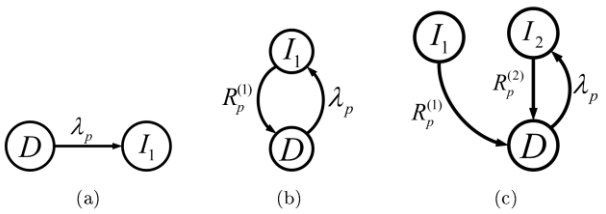

Fig. 4. Sample State transition diagrams of $Z(t)$ for (a) boundary $X(t) = 0$, (b) first regime, and (c) second regime.

We first focus on the MRMFQ model for $X(t)$ for which we define two service states 1 and 2 during where the packets are served with rate $R_p^{(1)}$ and $R_p^{(2)}$, respectively, and $X(t)$ increases with drift equal to 1. Upon the completion of the service of the current job in states 1 and 2, the system transits into the third state called *D*. During this newer state *D*, $X(t)$ decreases with slope one for an exponentially distributed amount of time with mean$1/\lambda_p$, so the delay of the new HoL message is reduced by an amount corresponding to its inter-arrival time. If $T^{(0)} < X(t) < T^{(1)}$ $(T^{(1)} < X(t) < T^{(2)})$, the system transition from state *D* to state 1 (state 2) occurs. Also, hitting zero by $X(t)$ in state *D* is possible.

In this case, upon the arrival of a new packet to the system, this packet is going to be served by $R_p^{(1)}$.Therefore, only transition from the boundary $X(t) = 0$, occurs out of state *D* to state 1 with rate λ_p. In summary, the background modulating Markov system denoted by $Z(t)$ has three different states, namely, 1, 2, and *D*. Therefore, the sample path followed by $X(t)$ can well be modeled as the modulated process of an MRMFQ, namely the process $(X(t), Z(t))$, with two regimes and three states. This MRMFQ is characterized by two infinitesimal generator matrices for regimes 1 and 2, denoted by $Q^{(i)}\ for\ i = 0,1$, two infinitesimal generator matrices for the finite boundaries, denoted by $\bar{Q}^{(i)}\ for\ i = 1,2$, and the corresponding drift matrices which are denoted by $R^{(i)}\ for\ i = 1,2$, and $\bar{R}^{(i)}\ for\ i = 1,2$ for the regimes and boundaries, respectively. See [15] for an elaborate description of MRMFQs. The state transition diagrams of $Z(t)$ are depicted in Fig. 4 for both regimes and boundary $X(t) = 0$ [16].

Based on the above description, the infinitesimal generator matrices for the two regimes of MRMFQ model can be obtained as follows by considering the state ordering $(2,1,D)$ [16]:

$$Q^{(1)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -R_p^{(1)} & R_p^{(1)} \\ 0 & \lambda_p & -\lambda_p \end{bmatrix},$$
$$Q^{(2)} = \begin{bmatrix} -R_p^{(2)} & 0 & R_p^{(2)} \\ 0 & -R_p^{(1)} & R_p^{(1)} \\ \lambda_p & 0 & -\lambda_p \end{bmatrix} \quad (4)$$

where $X(t)$ increases in the service state 1, system may transits from state 1 to state *D* in regime 2. Furthermore, the transition matrix at threshold $T^{(1)}$ is equal to $Q^{(2)}$. $\bar{Q}^{(0)}$ is similar to $Q^{(1)}$ except that there is no transition from state 1 to state *D* at the boundarypoint $X(t) = 0$. Then, it holds that

$$\bar{Q}^{(0)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \lambda_p & -\lambda_p \end{bmatrix} \quad (5)$$

On the basis of the above MRMFQ description, the drift matrices for each regime and the boundary are written as $R^{(1)} = \bar{R}^{(1)} = R^{(2)} = diag[1,1,-1]$ and $\bar{R}^{(0)} = diag[1,1,0]$ where diag denotes a diagonal matrix [16]. The joint probability density function (pdf) vector of $X(t)$ for regime k for $k = 1,2$, when $T^{(k-1)} \le D_Q(t) < T^{(k)}$ is defined as follows:

$$f_i^{(k)}(x) = \lim_{t\to\infty} \frac{d}{dx} Pr\{X(t) < x, Z(t) = i\},$$
$$f^{(k)}(x) = \left(f_2^{(k)}(x), f_1^{(k)}(x), f_D^{(k)}(x)\right). \quad (6)$$

Similarly, the steady state mass accumulation vector for boundary points $T^{(0)}$ and $T^{(1)}$ is defined as follows:

$$c_i^{(k)} = \lim_{t\to\infty} Pr\{X(t) = T^k, Z(t) = i\},$$
$$c^{(k)} = \left(c_2^{(k)}, c_1^{(k)}, c_D^{(k)}\right). \quad (7)$$

Following the same procedure as [14], the following set of differential equations holds for the joint pdf vector of $X(t)$:

$$\frac{d}{dx} f^{(i)}(x) R^{(i)} = f^{(i)}(x) Q^{(i)} \text{ for } i = 1,2. \quad (8)$$

In order to derive the boundary conditions, we first identify the set of states with positive, negative and zero drift for each regime and boundary. $S_+^{(k)}$, $S_-^{(k)}$, and $S_0^{(k)}$represent the states with positive, negative, and zero drift at regime *k*, respectively, and $\bar{S}_+^{(k)}$, $\bar{S}_-^{(k)}$, and $\bar{S}_0^{(k)}$ represent the states with positive, negative, and zero drift at boundary *k*, respectively. According to aforementioned infinitesimal generator matrices and drift matrices of the system model, these sets are populated as follows:

$S_+^{(1)} = \{I_1, I_2\}, S_-^{(1)} = \{D\}, S_0^{(1)} = 0,$
$S_+^{(2)} = \{I_1, I_2\}, S_-^{(2)} = \{D\}, S_0^{(2)} = 0,$
$\bar{S}_+^{(0)} = \{I_1, I_2\}, \bar{S}_-^{(0)} = 0, \bar{S}_0^{(0)} = 0,$
$\bar{S}_+^{(1)} = \{I_1, I_2\}, \bar{S}_-^{(1)} = \{D\}, \bar{S}_0^{(1)} = 0.$

Hence, the intermediate boundary point $T^{(1)}$ is an emitting state [14, 15] and the boundary conditions for the differential equations presented in (8) are as follows:

$$c_1^{(0)} = c_1^{(0)} = 0 \quad (9a)$$
$$c_m^{(0)} = 0, \text{ for } m \in \{1,2,D\} \quad (9b)$$
$$f^{(1)}(0+)R^{(1)} = c^{(0)}\bar{Q}^{(0)}, \quad (9c)$$
$$f^{(2)}\left(T^{(1)}+\right)R^{(2)} - f^{(1)}\left(T^{(1)}-\right)R^{(21} = c^{(1)}\bar{Q}^{(1)} \quad (9d)$$
$$\left(\sum_{k=1}^{2}\int_{T^{(k-1)}+}^{T^{(k)}-} f^{(k)}(x) + \sum_{k=0}^{1} c^{(k)}\right)(1,1,1)^T = 1. \quad (9e)$$

As the boundary conditions for the differential equations are not determined, this problem should be considered as a boundary value problem. The spectral solution to MRMFQs are presented in [15]. For our model, where there is no state with zero drift neither in the regimes, nor in the boundaries, the general solution to (8) is given as follows:

$$f^{(k)}(x) = \sum_i a_i^{(k)} e^{\lambda_i^{(k)} x}\phi_i^{(k)}, \text{ for } T^{(k-1)} < x < T^{(k)}, 1 \le k \le 2. \quad (10)$$

where $\left(\lambda_i^{(k)}, \phi_i^{(k)}\right)$ is the ith eigenvalue-left eigenvector pair of the matrix $Q^{(k)}\left(R^{(k)}\right)^{-1}$ [36] which are derived for the first and second regime as follows:

$$\begin{cases} \lambda_1^{(1)} = \lambda_p - R_p^{(1)} \rightarrow \phi_1^{(1)} = (0,1,1) \\ \lambda_2^{(1)} = 0 \rightarrow \phi_2^{(1)} = \left(0,1,\frac{R_p^{(1)}}{\lambda_p}\right) \\ \lambda_3^{(1)} = 0 \rightarrow \phi_3^{(1)} = (1,0,0) \end{cases}$$

$$\begin{cases} \lambda_1^{(2)} = -R_p^{(1)} \rightarrow \phi_1^{(2)} = \left(\frac{\lambda_p}{R_p^{(2)} - R_p^{(1)}}, \frac{R_p^{(2)} - R_p^{(1)} - \lambda_p}{R_p^{(2)} - R_p^{(1)}}, 1\right) \\ \lambda_2^{(2)} = \lambda_p - R_p^{(2)} \rightarrow \phi_2^{(1)} = (1,0,1) \\ \lambda_3^{(3)} = 0 \rightarrow \phi_3^{(1)} = \left(\frac{\lambda_p}{R_p^{(2)}}, 0, 1\right) \end{cases}$$

So, the pdf of the delay for the first and second regime is derived as follows:

$$f^{(1)}(x) = a_1^{(1)} e^{\lambda_i^{(1)} x}\phi_1^{(1)} + a_2^{(1)}\phi_2^{(1)} + a_3^{(1)}\phi_3^{(1)},$$
$$f^{(2)}(x) = a_1^{(2)} e^{\lambda_i^{(2)} x}\phi_1^{(2)} + a_2^{(2)} e^{\lambda_2^{(2)} x}\phi_2^{(2)} + a_3^{(2)}\phi_3^{(2)}.$$

In order to solve the differential equation, there are 6 unknown *a* coefficients, and 6 unknown *c* coefficients that should be determined using the boundary conditions. Since the queue model is of infinite size, the stability condition should also be satisfied for the second regime as $\pi^{(2)}R^{(2)}(1,1,1)^T < 0$ where $\pi^{(2)}$ is the steady state vector of $Q^{(2)}$. Furthermore, in order to acquire a bounded distribution for the second regime, the coefficients corresponding to the zero eigenvalues and the eigenvalues in the right-hand side of the imaginary axis must equal to zero. To be precise, imposing (9a) and (9b) results in:

$c_1^{(0)} = c_2^{(0)} = c_1^{(1)} = c_2^{(1)} = c_D^{(1)} = a_2^{(1)} = a_3^{(1)} = a_3^{(2)} = 0.$

Then, the equation (9c) yields the following:

$a_1^{(1)} = c_D^{(0)}\lambda_p, a_2^{(1)} = 0, a_3^{(1)} = 0.$

In order to derive a bounded distribution, $a_3^{(2)}$ should be equal to zero, since $a_3^{(2)}$ is the coefficient associated with the eigenvalue at the origin in the last regime. Hence, employing (9d) results in:

$$\begin{cases} a_1^{(2)} = \left(\frac{R_p^{(2)} - R_p^{(1)} - \lambda_p}{R_p^{(2)} - R_p^{(1)}}\right) c_D^{(0)}\lambda_p exp\left(\left(\lambda_1^{(1)} - \lambda_1^{(2)}\right)T^{(1)}\right) \\ a_2^{(2)} = \left(\frac{-\lambda_p}{R_p^{(2)} - R_p^{(1)} - \lambda_p}\right) c_D^{(0)}\lambda_p exp\left(\left(\lambda_1^{(1)} - \lambda_1^{(2)}\right)T^{(1)}\right) \\ a_3^{(2)} = 0 \end{cases}$$

Furthermore, we impose the stability condition as $\lambda_p < R_p^{(2)}$. Considering the stability condition in (10) and $R_p^{(1)} > 0$ (since rate can not be negative), we replace all unknown variables in $c_D^{(0)}$ in equation (9e) and derive $c_D^{(0)}$ as follows:

$$c_D^{(0)} = \left[\frac{R_p^{(1)} + \lambda_p}{R_p^{(1)} - \lambda_p} + \frac{2\lambda_p\left(R_p^{(2)} - R_p^{(1)}\right)}{R_p^{(1)}\left(R_p^{(2)} - R_p^{(1)} - \lambda_p\right)} e^{\lambda_p T^{(1)}} - \frac{2\lambda_p\left(R_p^{(2)} - R_p^{(1)}\right)\left(R_p^{(2)} - \lambda_p\right)}{\left(R_p^{(1)} - \lambda_p\right)\left(R_p^{(2)} - \lambda_p\right)\left(R_p^{(2)} - R_p^{(1)} - \lambda_p\right)} e^{-\left(R_p^{(1)} - \lambda_p\right)T^{(1)}}\right]^{-1}$$

On the other hand, we have the following closed-form expression for the *a* coefficients:

$$a_1^{(1)} = c_D^{(0)}\lambda_p,$$
$$a_1^{(2)} = \frac{\left(R_p^{(2)} - R_p^{(1)} - \lambda_p\right)}{\left(R_p^{(2)} - R_p^{(1)}\right)} c_D^{(0)}\lambda_p e^{\lambda_p T^{(1)}},$$
$$a_1^{(3)} = \frac{(-\lambda_p)}{\left(R_p^{(2)} - R_p^{(1)} - \lambda_p\right)} c_D^{(0)}\lambda_p e^{\left(R_p^{(2)} - R_p^{(1)}\right)T^{(1)}}. \quad (11)$$

and

$a_2^{(1)} = a_3^{(1)} = a_3^{(2)} = 0$

which give the probability distribution functions as follows:

$$f^{(1)}(x) = a_1^{(1)} e^{-\left(R_p^{(1)} - \lambda_p\right)x}(0,1,1),$$
$$f^{(2)}(x) = a_1^{(2)} e^{-R_p^{(1)} x}\left(\frac{\lambda_p}{R_p^{(2)} - R_p^{(1)}}, \frac{R_p^{(2)} - R_p^{(1)} - \lambda_p}{R_p^{(2)} - R_p^{(1)}}, 1\right) + a_2^{(2)} e^{-\left(R_p^{(2)} - \lambda_p\right)x}(1,0,1).$$

Since the unfinished work process $A(t)$ determines the amount of delay that newly arriving jobs (which arrive to the system according to a Poisson process) will experience, the steady-state probability distribution of state *D* can be used to obtain the quantities of interest, by a direct consequence of the PASTA (Poisson Arrivals See Time Averages) property. Therefore, in order to obtain the steady-state distribution of $A(t)$ from that of the fluid process $X(t)$, we censor out the service states and subsequently normalize the steady-state distributions.

In mathematical terms, we calculate the steady-state cumulative distribution function (CDF) of $A(t)$ denoted by $F(x)$, which is equal to the distribution of the queue

delay of a newly arrived packet, from that of $(X(t), Z(t))$ as follows:

$$F(x) = \lim_{t\to\infty} Pr\{A(t) \le x\} = \lim_{t\to\infty} \frac{Pr\{Z(t)=D, X(t)\le x\}}{Pr\{Z(t)=D\}} = \frac{c_D^{(0)} + \int_{0+}^{x} f^{(1)}(x)dx + \int_{T^{(1)}}^{\infty} f^{(2)}(x)dx}{c_D^{(0)} + \int_{0+}^{T^{(1)}} f^{(1)}(x)dx + \int_{T^{(1)}}^{\infty} f^{(2)}(x)dx} =$$

$$\begin{cases} \frac{c_D^{(0)}}{\Delta_2}, & x = 0, \\ \frac{\frac{a_1^{(1)}}{R_p^{(1)}-\lambda_p}\left(1-e^{-\left(R_p^{(1)}-\lambda_p\right)x}\right)}{\Delta_2}, & 0 < x \le T^{(1)}, \\ \frac{\Delta_1}{\Delta_2}, & x > T^{(1)}. \end{cases} \quad (12)$$

where

$$\Delta_1 = \frac{a_1^{(2)}}{R_p^{(1)}}\left(e^{-R_p^{(1)}x} - e^{-R_p^{(1)}T^{(1)}}\right) - \frac{a_1^{(1)}}{R_p^{(1)}-\lambda_p}\left(1-e^{-\left(R_p^{(1)}-\lambda_p\right)T^{(1)}}\right) + \frac{a_2^{(2)}}{R_p^{(2)}-\lambda_p}\left(e^{-\left(R_p^{(2)}-\lambda_p\right)x} - e^{-\left(R_p^{(2)}-\lambda_p\right)T^{(1)}}\right)$$

$$\Delta_2 = c_D^{(0)} + \frac{a_1^{(1)}}{R_p^{(1)}-\lambda_p}\left(1-e^{-\left(R_p^{(1)}-\lambda_p\right)T^{(1)}}\right) + \frac{a_1^{(2)}}{R_p^{(1)}}e^{-R_p^{(1)}T^{(1)}} + \frac{a_2^{(2)}}{R_p^{(2)}-\lambda_p}e^{-\left(R_p^{(2)}-\lambda_p\right)T^{(1)}}$$

Equation (12) provides a closed form expression for the cumulative distribution of queue delay for the packets of the PU in NOMA based QMP and this expression constitutes the main contribution of this paper.

Based on the analytical expression derived for the queue delay CDF, the SU's throughput is calculated as follows. If the PU's queue is empty, the secondary transmits at rate $Br_{s,m}$.

On the other hand, if the PU's queue is not empty, the secondary transmits at rate $Br_{s,w}$ or zero, depending on the delay experienced by the HoL packet of PU is below or above $T^{(1)}$, respectively. Therefore,

$$S_2^N = Br_{s,m}F(0) + Br_{s,w}\left(F\left(T^{(1)}\right) - F(0)\right) + B_0\left(1 - F\left(T^{(1)}\right)\right) = Br_{s,m}F(0) + Br_{s,w}\left(F\left(T^{(1)}\right) - F(0)\right). \quad (13)$$

5. Optimization of NOMA based system

To derive the closed-form theoretical expressions for the performance metrics of NOMA-based system, we select the optimum system parameters, namely the parameter ζ which controls the primary and secondary rate pairs as illustrated in Fig. 2 and the delay threshold $T^{(1)}$. In the CRN, the SU desires to achieve the highest throughput possible while meeting the PU delay constraint. At the first stage, simultaneous optimization of the power allocation factor ζ and the threshold $T^{(1)}$ is considered. Thus, the optimization problem is the maximization of the SU's throughput while satisfying the PU's delay constraint with appropriate choices of ζ and $T^{(1)}$.

The closed-form expressions of the SU's throughput and PU packets' delay distribution facilitates numerical calculation of the optimum parameters by deploying standard NLP (Non-Linear Programming) methods. First, the intersection point of r_{11} with r_{12}, and r_{21} with r_{23} is determined for a given channel condition. Hence, the min and max functions in the rate region can be eliminated by dividing the region into at most 3 sub-regions. Then, the optimum point is calculated for each sub-region by using NLP. Finally, the global optimum point is obtained by comparing the optimum points in each sub-region; see Alg. 1 for the pseudo-code for *NOMA*†.

Alg. 1: Pseudo-code of *NOMA*†

Input: Channel gains and transmission powers
Output: $\bar{T}^{(1)\dagger}$, $\zeta^{\dagger}$

1. Find $\zeta^{(1)}$ such that $r_{11} = r_{12}$ and find $\zeta^{(2)}$ such that $r_{21} = r_{23}$,
2. Divide the support set of ζ into 3 sub-regions:
   a) $S_1 = \left[0, min\left\{\zeta^{(1)}, \zeta^{(2)}\right\}\right]$
   b) $S_2 = \left[min\left\{\zeta^{(1)}, \zeta^{(2)}\right\}, max\left\{\zeta^{(1)}, \zeta^{(2)}\right\}\right]$
   c) $S_3 = \left[max\left\{\zeta^{(1)}, \zeta^{(2)}\right\}, 1\right]$
3. Perform NLP in each sub-region in order to solve the following problem for $i \in \{1,2,3\}$:
   $\left[T_i^{(1)}, \zeta^{(1)}\right] = \arg \max_{T^{(1)}, \zeta} S_2^N$
   $s.t.\, Pr\{D_q > D_T\} < \varepsilon, \zeta \in S_i$
4. Choose the $\left[T_i^{(1)}, \zeta^{(1)}\right]$ pair among the three candidates that provide the maximum SU throughput as $\left[T^{(1)\dagger}, \zeta^{\dagger}\right]$

Two-dimensional NLP used for *NOMA*† is computationally costly. Hence, a heuristic sub-optimal version is proposed, which is named as *NOMA*‡, to reduce the complexity of the optimization problem. This heuristic method stems from the observation that the optimal value of ζ does not vary much beyond a certain value of $T^{(1)}$.

In *NOMA*‡, the PU's delay constraint is neglected in the first step, and the optimal ζ is calculated by deploying a one-dimensional NLP, using a sufficiently large value for $T^{(1)}$. Then, given the optimal ζ and corresponding $r_p^{(1)}$ which is denoted by $r_p^{(1)\ddagger}$, the optimal value of $T^{(1)}$ is calculated. For this purpose, we first derive the maximum value of $T^{(1)}$, denoted by $\bar{T}^{(1)\ddagger}$, that satisfies the delay constraint $Pr\{D_q > D_T\} = \varepsilon$. Since the SU's throughput is an increasing function of $T^{(1)}$ except at the origin, the optimum choice for $T^{(1)}$ denoted by $T^{(1)\ddagger}$ will either be $\bar{T}^{(1)\ddagger}$ or zero, depending on whichever yields higher SU throughput; the pseudo-code of this second version is given in Alg. 2.

Alg. 2: Pseudo-code of *NOMA*‡

Input: Channel gains and transmission powers
Output: $\bar{T}^{(1)\ddagger}$, $\zeta^{\ddagger}$

1. Same as the steps 1-3 of Alg. 1,
2. Set $T^{(1)}$ to a sufficiently large value,
3. Perform one-dimensional NLP just for ζ for each sub-region in order to solve the following optimization problem for $i \in \{1,2,3\}$:
   $\zeta_i = \arg \max_{\zeta} S_2^N$
   $s.t. \zeta \in S_i$
4. Choose the ζ_i pair among the three candidates that provide the maximum SU throughput as $\zeta^{\ddagger}$,
5. Obtain $r_p^{(1)\ddagger}$ corresponding $\zeta_i^{\ddagger}$,
6. Find the value of $T^{(1)}$ given $r_p^{(1)\ddagger}$ that satisfies the delay constraint with equality, i.e., $\bar{T}^{(1)\ddagger}: Pr\{D_q > D_T\} = \varepsilon$
7. Choose $T^{(1)\ddagger}$ as either 0 or $\bar{T}^{(1)\ddagger}$ depending on whichever yields higher SU throughput: $T^{(1)\ddagger} = \arg \max_{T^{(1)} \in \{0, \bar{T}^{(1)\ddagger}\}} S_2^N$

6. Simulations and Numerical Results

We assume that the channel bandwidth is $B = 1\,MHz$, and $L = 1\,kbits$. The time unit is set to ms. Therefore, the packet service times are exponentially distributed with mean $1/R_{i,j} = 1/r_{i,j}$ for $i \in \{p,s\}$ and $j \in \{c,m,w\}$. The numerical results are obtained by *MATLAB 2014a* running on a machine with Intel core-i5 4300U *CPU* and 8 GB of *RAM*. In order to validate the analytical method, the analytical results are compared against simulations for the particular setting of $R_p^{(1)} = r_{p,w} = 0.1327$ and $R_p^{(2)} = r_{p,m} = 1.2266$ in the first numerical example. In Fig. 5, the CDF of the queuing delay in the PU queue obtained by both the analytical model as well as simulations are plotted for four different values of λ_p. Analytical results are in total agreement with the simulation results, which validates the analytic model. Next, we consider the performance of the OMA based and NOMA based QMPs in the following four separate scenarios, associated with different channel conditions:

1. In the poor PU scenario, $g_{12} = 0.5, g_{12} = 0.45;\ g_{21} = 0.7$.
2. In the poor PU scenario, $g_{12} = 0.8, g_{12} = 0.25;\ g_{21} = 0.7$.
3. In the poor PU scenario, $g_{12} = 0.5, g_{12} = 0.25;\ g_{21} = 0.3$.
4. In the poor PU scenario, $g_{12} = 0.5, g_{12} = 0.25;\ g_{21} = 0.7$.

while the remaining channel coefficients are set to $g_{11} = 1$, $g_{22} = 1$, and $g_{23} = 0.25$ and $P_i = 1\,mW$ for $i = 1,2$. These channel coefficients are obtainable through well-known channel estimation processes implemented in most communication protocols. The delay constraint parameters are set to $D_T = 80;\ 320\ ms$ and $\varepsilon = 10^{-2}$. The NLP steps in *NOMA*† and *NOMA*‡ are conducted using the fmincon function in MATLAB. Furthermore, in the second step of Alg. 2, the sufficiently large number value is chosen as 600 ms in this example. The throughput of the SU, while the primary's delay constraint is satisfied, is considered as the performance metric in all comparisons. This metric not only provides a tool to compare the performance of the SU, but also, implicitly determines the maximum admissible rate for the PU traffic, while satisfying the de lay constraint. The performance of OMA based and NOMA based methods are demonstrated in Fig. 7. The obtained result shows the superiority of NOMA based method in the sense of SU's throughput.

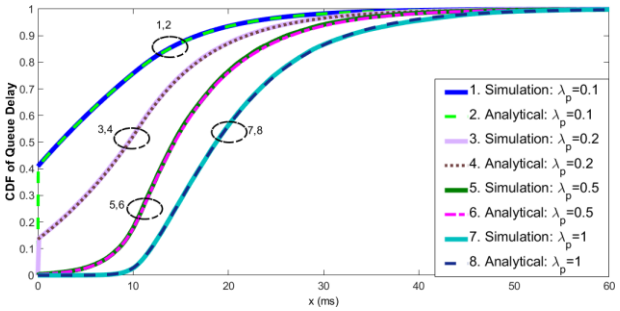

Fig. 5: CDF of the delay obtained by both simulations and the proposed analytical method.

As the PU channel condition improves, the performance of NOMA based method improves with respect to OMA based one, since the system operates more frequently in regime 1 and PU can handle its packets on its own. Moreover, as depicted in Fig. 6, the performance of NOMA based methods improves as the delay constraint of the PU is relax. Note that the performance of NOMA based methods reduces to OMA based counterpart as the conditions, i.e., channel gains or arrival rate, turn out to be unfavorable. In these situations, the system benefits more by transmitting each message one at a time (OMA) instead of simultaneous transmission of primary and secondary messages (NOMA). This situation happens when $T^{(1)}$ is chosen as zero. This scenario is highlighted in the poor PU channel condition (Fig. 7 (a)) where NOMA based methods perform identical to OMA based one independent of the delay tolerance of primary packets. Moreover, it is illustrated that the proposed heuristic method *NOMA*‡ performs very close to *NOMA*† while requiring less computational power, because the latter is a $O(n^2)$ algorithm, while the former is a $O(n)$ algorithm. In particular, the av erage run time of *NOMA*† and *NOMA*‡, which are the average values of over 50 runs, are 79.5743 and 6.5552 seconds, respectively, which translates into more than 90% reduction in computation time with *NOMA*‡. The secondary throughput and their relative gains with respect to OMA based method (denoted by G^k for $k \in \{NOMA^{\dagger}, NOMA^{\ddagger}\}$) for both versions of *NOMA* are given in Tables 1, 2, and 3, for good primary, poor secondary and good secondary conditions, respectively. The *NOMA*† appears to be superior to all the other methods.

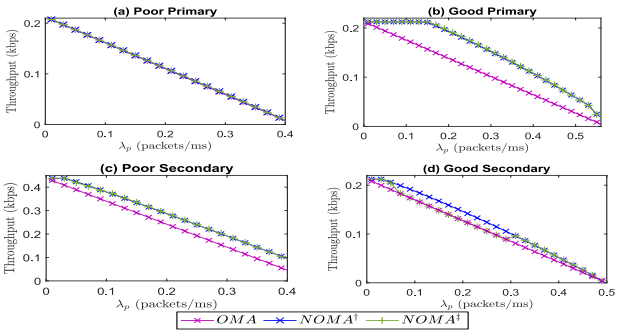

Fig. 6: SU throughput with OMA based and NOMA based methods for the four channel scenarios

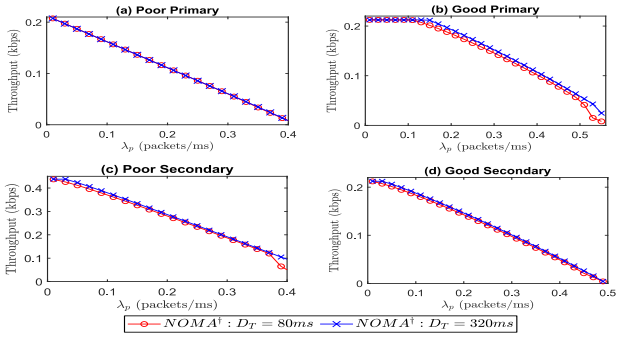

Fig. 7: Comparing the performance of $NOMA^{\dagger}$ for $D_T = 80\ ms$ and $D_T = 320\ ms$.

Table I. The SU throughput in the good primary scenario.

λ_p	S_2^O	$S_2^{N\dagger}$	$S_2^{N\ddagger}$	$G^{N\dagger}$(%)	$G^{N\ddagger}$(%)
0.09	0.174	0.2123	0.2123	22.01149	22.01149
0.13	0.1642	0.2123	0.2123	29.29354	29.29354
0.19	0.1419	0.1968	0.1968	38.68922	38.68922
0.29	0.1049	0.1564	0.1564	49.09438	49.09438
0.31	0.09749	0.1478	0.1478	51.60529	51.60529
0.39	0.06784	0.112	0.112	65.09434	65.09434
0.49	0.03079	0.06365	0.06365	106.9178	106.723

Table II. The SU throughput in the poor secondary scenario.

λ_p	S_2^O	$S_2^{N\dagger}$	$S_2^{N\ddagger}$	$G^{N\dagger}$(%)	$G^{N\ddagger}$(%)
0.01	0.428	0.4378	0.4378	2.28972	2.28972
0.05	0.3888	0.4231	0.4191	8.822016	7.79321
0.09	0.3496	0.3885	0.3867	11.127	10.61213
0.11	0.33	0.3704	0.3704	12.24242	12.24242
0.13	0.3104	0.3521	0.3521	13.43428	13.43428
0.17	0.2712	0.3148	0.3148	16.0767	16.0767
0.21	0.232	0.2771	0.2771	19.43966	19.43966
0.29	0.1536	0.2011	0.2011	30.92448	30.92448
0.39	0.05565	0.105	0.1044	88.67925	87.60108
0.43	0.01645	0.06635	0.05821	303.3435	253.8602

Table III. The SU throughput in the good primary scenario.

λ_p	S_2^O	$S_2^{N\dagger}$	$S_2^{N\ddagger}$	$G^{N\dagger}$(%)	$G^{N\ddagger}$(%)
0.01	0.2081	0.2123	0.2123	2.01826	2.01826
0.05	0.1911	0.2062	0.2043	7.9016	6.90734
0.09	0.1741	0.1916	0.1741	10.05169	0
0.11	0.1656	0.1838	0.1656	10.9903	0
0.13	0.1571	0.1758	0.1571	11.9032	0
0.17	0.1401	0.1592	0.1401	13.6331	0
0.27	0.09768	0.1148	0.09768	17.5266	0
0.37	0.05521	0.06691	0.06691	21.1918	21.1918
0.47	0.01275	0.01566	0.01564	23.1132	22.95597

The gain of $NOMA^{\dagger}$ depends on the channel conditions and the arrival rate of the PU traffic. We also observe from Tables 1, 2, and 3 that the gain of $NOMA^{\dagger}$ grows consistently as the arrival rate increases and this gain reaches over 300% for relatively high p values in the poor secondary scenario. We also observe that $NOMA^{\ddagger}$ performance is very close to $NOMA^{\dagger}$ for most of the cases, which makes it an attractive choice considering its significantly lower computational complexity. Note that for poor PU channel conditions, all three algorithms perform identically since the optimization is conducted for the parameters of the first regime which is practically not used due to very low performing PU. That is, in poor PU channel conditions, all three algorithms prefer to devote all available resources to empty the PU queue first and then assign the remaining resources to the SU. The results associated with the poor channel scenario are deliberately not tabulated.

In summary, the proposed $NOMA^{\dagger}$ method provides a robust mechanism for the PU to handle its delay sensitive traffic in a resource limited environment by leveraging SU's resources. The SU, which does not possess licensed bands, helps the PU to deliver its messages in a timely manner, in exchange for transmission privileges. It is shown that the SU achieves higher throughputs compared to conventional OMA method. Furthermore, as the PU delay constraint is relaxed, the SU throughput improves.

7. Conclusion

The performance of NOMA is investigated in conjunction with upper layer of network stack. A novel QMP is proposed to handle the delay sensitive traffic of PU in a secure NOMA based CRN, which conveys mixed multicast and unicast traffic. The security of the primary's multicast traffic as well as secondary's unicast traffic is guaranteed by deploying PLS. The proposed QMP, which deploys a delay-aware adaptive mechanism, attempt to maximize the SU throughput while meeting the PU's delay criterion. Through a novel MRFMQ based model, the closed form expressions for the exact delay distribution of the PU traffic are derived and validated for this QMP. Moreover, analytical expressions are employed to optimally tune its parameters. The proposed methods are simulated and compared to the conventional OMA based method in a secure CRN. It is shown that by intelligent manipulation of the PU's delay constraint, NOMA consistently outperforms OMA based methods. More general traffic models for PU and SU are left for future research.

8. References

[1] Daniel Minoli and Benedict Occhiogrosso, "Practical Aspects for the Integration of 5G Networks and IoT Applications in Smart Cities Environments", Wireless Communications and Mobile Computing, Vol. 2019, pp. 1-30, August 2019.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," IEEE Commun. Mag., vol. 53, no. 9, pp. 74-81, Sep. 2015.

[3] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in Proc. IEEE 77th Veh. Technol. Conf., Dresden, Germany, Jun. 2013, pp. 1-5.

[4] F. Dehghani, J. Pourrostam, "Improving Jain Fairness Index and Optimizing Transmitter Power Consumption in NOMA Systems", Tabriz Journal of Electrical Engineering, vol. 49, no. 2, pp. 577-586, 2019 (in persian).

[5] Z. Ding, Z. Zhao, M. Peng, and H. V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming," IEEE Trans. Commun., vol. 65, no. 7, pp. 3151-3163, Jul. 2017.

[6] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal," IEEE Personal Communications, Vol. 6, No. 4, pp. 13-18, August 1999.
[7] M. Torabi, N. Mohammadi, "Performance Evaluation of Cooperative Amplify-and-Forward Relaying Systems in a Spectrum Sharing Cognitive Radio with Relay Selection", Tabriz Journal of Electrical Engineering, online preprint, 2020 (in persian).
[8] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," Proceedings of the IEEE, Vol. 97, No. 5, pp. 894-914, May 2009.
[9] L. Xu, A. Nallanathan, X. Pan, J. Yang, and W. Liao, "Security-aware resource allocation with delay constraint for NOMA-based cognitive radio network," IEEE Trans. Inf. Forensics Security, vol. 13, no. 2, pp. 366-376, Feb. 2018.
[10] S. A. R. Zaidi, C. D. McLernon, M. Ghogho, and M. A. Imran, "Cloud empowered cognitive inter-cell interference coordination for small cellular networks," IEEE International Conference on Communication Workshop (ICCW 2015), pp. 2218-2224, 2015.
[11] R. Giuliano, F. Mazzenga, A. Neri, and A. M. Vegni, "Security access protocols in IoT capillary networks," IEEE Internet Things J., vol. 4, no. 3, pp. 645-657, June 2017.
[12] A. D. Wyner, "The wire-tap channel," The Bell System Technical Journal, vol. 54, no. 8, pp.1355–1387, Oct 1975.
[13] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of Queueing Theory*, 5th Edition, Wiley, New York, 2018.
[14] H. E. Kankaya and N. Akar, "Solving multi-regime feedback fluid queues," Stochastic Models, Vol. 24, No. 3, pp. 425-450, 2008.
[15] N. Akar and K. Sohraby, "Infinite- and finite-buffer Markov fluid queues: A unified analysis," Journal of Applied Probability, Vol. 41, No. 2, pp. 557-569, June 2004.
[16] K. Adli Mehr, J. Musevi Niya, and N. Akar, "Queue management for two-user cognitive radio with delay-constrained primary user," Computer Networks, Vol 142, pp. 1-12, 2018.
[17] H. G. Bafghi, S. Salimi, B. Seyfe, and M. R. Aref, "Cognitive interference channel with two confidential messages," International Symposium on Information Theory Its Applications, pp. 952-956, 2010.
[18] F. Gabry, N. Li, N. Schrammar, M. Girnyk, L. K. Rasmussen, and M. Skoglund, "On the optimization of the secondary transmitter's strategy in cognitive radio channels with secrecy," IEEE Journal on Selected Areas in Communications, Vol. 32, No. 3, pp. 451-463, March 2014.
[19] X. Chen and et al, "Exploiting Inter-User Interference for Secure Massive Non-Orthogonal Multiple Access," IEEE Journal on Selected Areas in Communications, Vol. 36, No. 4, pp. 788 - 801, April 2018.
[20] S. Zhang et al, "Enhancing the Physical Layer Security of Uplink Non-Orthogonal Multiple Access in Cellular Internet of Things," IEEE Access, Vol. 6, pp. 58405 – 58417, October 2018.
[21] J. Choi, et al, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," IEEE Trans. Commun., vol. 63, no. 3, pp. 791-800, March 2015.
[22] Ding et al, "On the Spectral Efficiency and Security Enhancements of NOMA Assisted Multicast-Unicast Streaming," IEEE Trans. Commun., vol. 65, no. 7, pp. 3151 - 3163, July 2017.
[23] Y. Liang, A. Somekh-Baruch, H. V. Poor, S. Shamai, and S. Verdu, "Capacity of cognitive interference channels with and without secrecy," IEEE Transactions on Information Theory, Vol. 55, No. 2, pp. 604-619, February 2009.
[24] C. E. Koksal, O. Ercetin, and Y. Sarikaya, "Control of wireless networks with secrecy," IEEE/ACM Transactions on Networking, Vol. 21, No. 1, pp. 324-337, February 2013.
[25] Y. Sarikaya, O. Ercetin, and C. E. Koksal, "Confidentiality-preserving control of uplink cellular wireless networks using hybrid ARQ," IEEE/ACM Transactions on Networking, Vol. 23, No. 5, pp. 1457-1470, October 2015.
[26] Z. Mao, C. E. Koksal, and N. B. Shroff, "Achieving full secrecy rate with low packet delays: An Optimal Control Approach," IEEE Journal on Selected Areas in Communications, Vol. 31, No. 9, pp. 1944-1956, September 2013.
[27] Z. Mao, C. E. Koksal, and N. B. Shroff, "Fair rate allocation for broadcast channel with confidential messages," IEEE Global Conference on Signal and Information Processing, pp. 799-802, 2013.
[28] Y. Sarikaya, C. E. Koksal, and O. Ercetin, "Dynamic network control for confidential multi-hop communications," IEEE/ACM Transactions on Networking, Vol. 24, No. 2, pp. 1181-1195, April 2016.
[29] L. Musavian, S. Aïssa, and S. Lambotharan, "Effective capacity for interference and delay constrained cognitive radio relay channels," IEEE Transactions on Wireless Communications, Vol. 9, No. 5, pp. 1698-1707, May 2010.
[30] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," IEEE Transactions on Wireless Communications, Vol. 24, No. 4, pp. 630-643, July 2003.
[31] Y. Yang, S. Aïssa, and K. N. Salama, "Spectrum band selection in delay-QoS constrained cognitive radio networks," IEEE Transactions on Vehicular Technology, Vol. 64, No. 7, pp. 2925-2937, July 2015.
[32] L. Ma, Y. Ma, and P. Ma, "Delay-QoS-driven secrecy power allocation in underlay secure cognitive radio system," IEEE 83rd Vehicular Technology Conference (VTC 2016), pp. 1_5, 2016.
[33] A. M. Elmahdy, A. El-Keyi, T. ElBatt, and K. G. Seddik, "Optimizing cooperative cognitive radio networks performance with primary QoS provisioning," IEEE Transactions on Communications, Vol. 65, N0. 4, pp. 1451-1463, April 2017.
[34] M. Bloch, J. Barros, M. R. D. Rodrigues, and S. W. McLaughlin, "Wireless information-theoretic security," IEEE Transactions on Information Theory, Vol. 54, No. 6, pp. 2515-2534, June 2008.

[35] C. Tunc and N. Akar, "Performance modeling of delay-based dynamic speed scaling," 9th International Conference on Matrix-Analytic Methods in Stochastic Models, Budapest, Hungry, 2016.

[36] W. Scheinhardt, N. Van Foreest, and M. Mandjes, "Continuous feedback fluid queues," Operations Research Letters, Vol. 33, No. 6, pp. 551-559, November 2005.