Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Satisfying Strict Deadlines for Cellular Internet of Things through Hybrid Multiple Access

Onur Berkay Gamgam* | Ezhan Karasan

[1]Electrical and Electronics Engineering, Bilkent University, Turkey

Correspondence
*Onur Berkay Gamgam, Universiteler, Ankara, 06800, Turkey. Email: onur.gamgam@bilkent.edu.tr

Present Address
Universiteler, Ankara, 06800, Turkey

Summary

Latency-constrained aspects of cellular Internet of Things (IoT) applications rely on Ultra-Reliable and Low Latency Communications (URLLC) which highlight research on satisfying strict deadlines. In this study, we address the problem of latency constrained communications with strict deadlines under average power constraint using Hybrid Multiple Access (MA) which consists of both Orthogonal MA (OMA) and power domain Non-Orthogonal MA (NOMA) as transmission scheme options. We aim to maximize the timely throughput, which represents the average number of successfully transmitted packets before deadline expiration, where expired packets still waiting in the buffer are dropped. We use Lyapunov stochastic optimization methods to develop a dynamic power assignment algorithm for minimizing the packet drop rate while satisfying time average power constraints. Numerical results show that Hybrid MA improves the timely throughput compared to conventional OMA by up to 46% and on the average by more than 21% while satisfying average power constraints.

KEYWORDS:
Non-Orthogonal Multiple Access (NOMA), Deadline-constrained communications, Dynamic algorithms, Lyapunov optimization, Power efficient algorithms

1 | INTRODUCTION

The evolution from human-to-human (H2H) oriented communication ecosystem towards enabling human-to-machine (H2M) and machine-to-machine (M2M) communication types highlights the need for massive connectivity, high throughput and low latency communications. This leads the emergence of Internet of Things (IoT) applications requiring ultra-reliable and low-latency communications (URLLC)[1]. High connectivity demand for the considerably increasing number of interconnected devices under limited resources highlights the cellular IoT as a promissing new frontier for managing these challenges[2,3]. For instance,

Machine Type Communication (MTC) based cellular IoT applications focus on delivery of small sized packets between significantly large number of devices with latency, reliability and connectivity oriented constraints[4]. Reliable communication for IoT with small sized packets requires short block-length transmission which further increases problem complexity[5]. Moreover, recent standardization studies on 5G communications highlight the extremely stringent latency and reliability requirements for these emerging applications. For instance, latency requirements for motion control, mobile automation and electric power grid applications are denoted as up to 0.5 ms, 1 ms and 10 ms, respectively[6]. These challenging URLLC constraints considered for IoT applications focus attention on the research for the latency constrained communications with strict deadlines[7,8].

The time critical aspects of cellular IoT applications bring out the concept of deadline as the maximum allowable time duration for the successful delivery of a data packet. If the data packet is not fully transmitted within the deadline duration, then it is considered as useless and dropped out of the system[9]. For the performance evaluation of deadline constrained systems, the notion of timely throughput is proposed, which represents the long term average rate of data packets that are successfully delivered within their deadlines[10].

Another critical aspect of cellular IoT applications is the increasing number of connected devices[11]. The number of Machine to Machine (M2M) devices connected to the global network in 2023 is expected to increase to 29.3 billion[12]. The limitations of widely used OMA schemes introduce new challenges in terms of increasing efficiency of available resources in order to satisfy the emerging massive connectivity demand. NOMA is able to adapt resources according to the traffic load and user channel state information, therefore, spectrum and energy efficiency can be increased under various conditions[13]. Moreover, NOMA increases connectivity in the system by increasing the number of concurrent transmissions on the same spectral resource[14]. Yet, the advantage of NOMA in terms of system capacity depends on diversity of user channel conditions and number of connected users. In order to address various needs of emerging applications, a Hybrid MA scheme consisting of adaptively switching between OMA and NOMA in the time domain is considered by 3rd Generation Partnership Project (3GPP)[15,16,17], which is a collaborative project to develop globally applicable specifications for mobile systems.

The focus of this study is to maximize timely throughput using Hybrid MA for cellular IoT applications. Although, the potential of Hybrid MA is widely studied in the scope of information freshness[18,19,20,21], to the best of our knowledge, there is a lack of study on Hybrid MA in the scope of timely throughput for cellular IoT applications.

1.1 | Related work

In latency constrained cellular IoT applications, information freshness and timely delivery of critical information is crucial for efficiency and system operability. Previous research on latency constrained communication focus on several different metrics: (i) average latency, (ii) satisfaction of strict deadlines, and (iii) information freshness. In the first type of studies, the aim is to minimize average latency by minimizing average queue length, which is based on the Little's theorem[7,8,22]. In[22], a joint dynamic

power control and user pairing algorithm is proposed under a Hybrid MA scheme for power efficient and delay constrained communications. The dynamic algorithm is based on the $drift - plus - penalty$[23] technique, which is a Lyapunov based stochastic network optimization method.

The second type of studies aim to meet strict deadlines for the problem of latency constrained communications. In[24], the authors introduce $BT-Problem$ for sending B bits of data within T duration of deadline while minimizing power utilization, and they solve it using a continuous time model. In[25], timely data transmission under deadline constraints using NOMA is considered and high computational complexity of scheduling and resource allocation tasks is addressed with deep learning techniques. In[26,27], Fountoulakis et al. considered the packet drop rate minimization with limited power budget. Fixed sized packet arrivals are served in a packet per slot manner with OMA transmission scheme through a wireless medium modeled as binary channel. A penalty metric based on the remaining packet deadline until expiration is proposed. The induced penalty reaches the maximum value when the packet expires. A dynamic power assignment algorithm is developed with $drift - plus - penalty$[23] technique for the maximization of timely throughput under average power constraints.

The third type of studies is concerned with satisfying information freshness, for which, Age of Information (AoI) is used as a general performance metric[21]. In[28], AoI is considered for the increasing connectivity in massive MTC applications which demand diverse latency requirements. In[18,29], the potential of NOMA is investigated for information freshness along with increased system connectivity. In[19], NOMA is considered for the task of timely information updates. In[20], Hybrid MA is considered for AoI. These recent studies demonstrate the research community's interest on NOMA for timely information freshness.

1.2 | Contributions

In this study, we address the problem of latency constrained communications with strict deadlines under time average power constraints in OMA and NOMA based Hybrid MA to be used by cellular IoT applications. We propose a dynamic algorithm which allocates user power in real-time to satisfy time average power constraints while maximizing timely throughput by minimizing the packet drop rate. Main contributions of this study are as follows:

- We use a realistic model which is appropriate for the cellular IoT scenario. We extend the stochastic network optimization framework for packets with deadlines under average power constraints, proposed by Fountoulakis et al.[26]. The scope of the extensions covers the time-varying arrival content, OMA and NOMA based Hybrid MA, fragmentation of packets and modelling the wireless medium as a fading channel. Moreover, we consider short packets with Finite Blocklength (FBL) codes, which is appropriate for latency critical cellular IoT applications.

- We introduce a novel degree of freedom to the objective function to adjust its increment pattern as remaining deadline diminishes. Our aim is to investigate the relation between the remaining packet deadline and the packet drop rate. The prioritization of packets with the proposed technique is called Remaining Deadline based Parametric Prioritization Approach (RDPPA).

- We consider constraints on time average power utilization. We propose a dynamic algorithm using Lyapunov stochastic optimization to satisfy time average constraints while minimizing the objective of packet drop rate.

- The proposed dynamic algorithm leverages optimal power allocations for OMA and NOMA transmission schemes. Using convex optimization techniques, we derive optimum transmission schemes according to the observed channel and queue state information.

Our key numerical results show that the Hybrid MA outperforms OMA-only based systems by increasing the timely throughput on the average by more than 21% while satisfying time average power constraints. In delay constrained wired systems, Earliest Deadline First (EDF) is the optimal scheduling algorithm[30]. We show that for fading channels, the drop rate is minimized using RDPPA when packets are prioritized considering the remaining deadline as well as the channel state. In this way, a non-earliest deadline packet of a user with a strong channel condition can be eligible for transmission in order to minimize overall packet drop rate, instead of an earliest deadline packet of another user with a weak channel condition. RDPPA controls the relation between power allocations and Channel-Queue State Information (CQSI) in the system to minimize overall packet drop rate.

In the rest of the paper, the system model is explained first. Then, the optimization problem for power allocation using Hybrid MA is presented with the proposed solution. This is followed by a demonstration of optimal power assignment for Hybrid MA. Finally, the numerical results are presented with an elaborate analysis of system parameters' effects.

2 | SYSTEM MODEL

We consider a downlink broadcast scenario for cellular IoT applications in which a single-antenna access point (AP) transmitting time critical data to N stationary single-antenna users within its coverage area. The set of users is denoted as $\mathcal{N} \triangleq \{1, \cdots, N\}$. Since AP is equipped with a single antenna, there is a single available output link. Therefore, the output link is allocated either for user $\{i\}$'s OMA transmission, or users $\{i, j\}$'s two-user NOMA transmission on each time slot. We consider a discrete-time system where time duration of each slot is denoted as τ and $\mathcal{B}$ represents the transmission bandwidth. Therefore, transmission is performed within FBL of $\tau\mathcal{B}$. Conventional Shannon capacity is based on infinite blocklength, thus it is not applicable for this scenario. Polyanski *et al.*[31] proposed a framework for tightly approximating tranmsission rate R^* in FBL regime for blocklength

$\tau\mathcal{B}$ and block error rate (BLER) ϵ, as follows:

$$R^* \approx \log_2(1+\eta) - \sqrt{\frac{\mathcal{V}(\eta)}{\tau\mathcal{B}}} \cdot \frac{Q^{-1}(\epsilon)}{\ln 2} \tag{1}$$

where η is Signal-to-Noise Ratio (SNR), $\mathcal{V}(\eta) = 1 - (1+\eta)^{-2}$ is the channel dispersion and $Q^{-1}(.)$ is the inverse Q-function. The power budget for the transmitter is denoted as P_0. $\mathbf{P}(t) = \{P_i(t)\}_{i\in\mathcal{N}}$ corresponds to the transmitter powers allocated for users in slot t. We consider continuous power levels such that $0 \leq P_i(t) \leq P_0$ for $\forall\, i,t$. $\mathcal{P}^O(t)$ and $\mathcal{P}^H(t)$ denote the set of all available power assignments for OMA-only and Hybrid MA at time t, respectively. Let $\overline{P_i}$ be the average power utilization over all time slots for user i and $\overline{P} \triangleq (1/N)\sum_{i=1}^{N} \overline{P_i}$ be the overall average power consumption.

$$\overline{P_i} \triangleq \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} P_i(\tau), \qquad \forall\, i \in \mathcal{N} \tag{2}$$

We consider an average power consumption constraint $\gamma_i \in (0, P_0]$ for each user i, such that $\overline{P_i} \leq \gamma_i$. R denotes the radius of the circular coverage area. Let r_i denote the distance between the transmitter and user i, which is uniformly selected: $r_i \sim U[0, R]$. Let h_i denote the channel gain of user i. Wireless communication link between the transmitter and users is modeled as a Rayleigh fading channel[22]. The Random Variables (RVs) representing the fast fading component of each user's channel in a slot are independent and identically distributed (i.i.d.). Finally, we assume that the channels between the transmitter and users are static during a time slot, but they alter from slot to slot. This assumption is justified for cellular IoT applications where mobility is typically low. We assume that channel state information is available at the beginning of each slot, based on the available channel estimation methods for low mobility applications[32,33].

User i's arriving data packets are stored in queue i. Let $m_i \in \mathbb{Z}_+$ be the deadline for the arriving packets of user i in terms of slot count. Let $a_i(t) \sim Ber(\pi_i)$ be a Bernoulli RV with arrival rate π_i representing the arrival probability of a packet for user i at slot t. Let $u_i(t) \sim U[\Lambda]$ be the bit count of the arriving packet for user i in the t^{th} slot and it is uniformly selected from a finite set of positive integers Λ, which represents available packet sizes. The RVs considered for the arrival processes are i.i.d. The arrival process of user i in the t^{th} slot is denoted as $\lambda_i(t) \triangleq a_i(t) \cdot u_i(t)$, in bits per slot. The queue backlog for user i in the t^{th} slot is denoted as $Q_i(t)$, where $Q_i(0) = 0$. The packets in a queue are processed in the first-in first-out manner, so that, only the packet at the head of the queue is considered for transmission. Let $d_i(t)$ be the number of slots left before expiration and $q_i(t)$ be the number of data bits left in the t^{th} slot for the packet at the head of queue i.

The departure process represents transmitted number of bits in a slot. Let $\Psi_i(t) = j$ be the user paired with user i in the slot t. If $i = j$, OMA is employed for user i, else, users $\{i, j\}$ are paired for a NOMA transmission. Let $R_i(t, P_i(t), \Psi_i(t))$ be the data rate of user i in bits per second in the slot t. The departure rate for user i in the t^{th} slot is denoted as $\mu_i(t) \triangleq \tau \cdot R_i(t, P_i(t), \Psi_i(t))$, in bits per slot. The drop event occurs for the packet at the head of queue i when $d_i(t) = 1$ and $\mu_i(t) < q_i(t)$. The number of dropped bits for user i is denoted as $D_i(t) = \mathbb{1}\{d_i(t) = 1\} \cdot \max[q_i(t) - \mu_i(t), 0]$. The queue dynamics is presented in (3), based

on the arrival and departure processes, queue backlog and number of dropped bits. $\overline{D_i}$ represents the packet drop rate for user i and $\overline{D} \triangleq (1/N)\sum_{i=1}^{N}\overline{D_i}$ is the average drop rate.

$$Q_i(t+1) \triangleq \max[Q_i(t) - \mu_i(t), 0] + \lambda_i(t) - D_i(t) \tag{3}$$

$$\overline{D_i} \triangleq \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{1}\{D_i(\tau) > 0\}, \qquad \forall\, i \in \mathcal{N} \tag{4}$$

3 | OPTIMIZATION PROBLEM FORMULATION FOR POWER ALLOCATION USING HYBRID MA

The problem of minimizing packet drop rate with deadlines under users' average power constraints[26] using Hybrid MA is presented as follows:

$$\min_{\mathbf{P}(t)} \quad \sum_{i=1}^{N}\overline{D_i} \tag{5a}$$

$$\text{s.t.} \quad \overline{P_i} \le \gamma_i, \forall\, i \in \mathcal{N} \tag{5b}$$

$$\mathbf{P}(t) \in \mathcal{P}^H(t) \tag{5c}$$

We consider two different Transmission Modes (TM) for a packet. The first one is for completely transmitting a packet per slot, which is called Complete TM (CTM). A binary decision is made to either fully transmit content of a packet or not transmit it at all. The second one is called Fragmented TM (FTM), where a data packet is fragmented at the source for being transmitted in different slots and reassembled at the destination. A packet is considered to be successfully transmitted only when all its fragments are successfully transmitted before deadline expiration. Let $\phi(t) \in \{\Phi_C, \Phi_F\}$ be the occupied TM at time t, where Φ_C and Φ_F represent CTM and FTM, respectively. The function $\varphi(t, \phi(t))$ quantifies the queue reduction ratio as follows:

$$\varphi(t,\phi(t)) \triangleq \begin{cases} \mathbb{1}\left\{q_i(t) - \mu_i(t) > 0\right\} & \text{, if } \phi(t) = \Phi_C \\ (q_i(t) - \mu_i(t))/q_i(t) & \text{, if } \phi(t) = \Phi_F \end{cases} \tag{6}$$

The cost of a packet drop contributes to the minimization of the objective function over the infinite horizon, however, the decision variable $\mathbf{P}(t)$ is optimized slot-by-slot. The future values of CQSI are unknown due to their random nature. Therefore, it is not possible to predict future values of (5a). In[26], Fountoulakis et al. introduced the function $f_i(t)$ whose future values are affected by the current decision $P_i(t)$ and the relative difference between the packet deadline, m_i, and the number of remaining future slots, $d_i(t) - 1$, before expiration of packet at the head of queue. In this paper, we propose a novel function $\mathcal{F}_i(t, \alpha_i(t), \phi(t))$ as:

$$\mathcal{F}_i(t,\alpha_i(t),\phi(t)) \triangleq \left(\frac{m_i - (d_i(t)-1)}{m_i}\right)^{\alpha_i(t)} \varphi(t,\phi(t)) \tag{7}$$

where $\alpha_i(t)$ is a non-negative exponent parameter proposed to adjust the importance of remaining deadline in the objective function for user i at slot t. The $f_i(t)$ in[26] is equal to $\mathcal{F}_i(t, 1, \Phi_C)$, showing that $\mathcal{F}_i$[†] has two additional degrees of freedom, $\alpha_i(t)$ and $\phi(t)$. Note that $0 \leq \mathcal{F}_i \leq 1$. While $\mathcal{F}_i = 1$ represents the packet drop event, $\mathcal{F}_i = 0$ indicates that the packet of user i is served completely before expiration. Between these two extreme cases, the remaining deadline of user i's packet, $d_i(t)$, is mapped to a penalty value which elevates as the remaining deadlines reduces. We propose RDPPA to investigate the relative importance of packets' remaining deadlines in terms of average dropping rate by rapidly and slowly elevating the penalty towards 1 for $\alpha_i(t) < 1$ and $\alpha_i(t) > 1$ cases, respectively. Therefore, $\mathcal{F}_i$ provides information that helps us to predict future consequences of our actions. Let $\overline{\mathcal{F}_i}$ be the time average of $\mathcal{F}_i$. We define the following new problem:

$$\min_{\mathbf{P}(t)} \quad \sum_{i=1}^{N} \overline{\mathcal{F}_i} \triangleq \sum_{i=1}^{N} \left(\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathcal{F}_i \right) \tag{8a}$$

$$\text{s.t.} \quad \overline{P_i} \leq \gamma_i, \forall\, i \in \mathcal{N} \tag{8b}$$

$$\mathbf{P}(t) \in \mathcal{P}^H(t) \tag{8c}$$

The problem definition presented in (8) is a minimization problem with constraints in the form of time averages, which can be solved using $drift - plus - penalty$ technique[23]. The time average constraints in (8b) are transformed into virtual-queues $(X_i(t), \forall i \in \mathcal{N})$ in (9), where arrivals are $P_i(t)$ and respective service rates are γ_i. The problem becomes a queue stability problem with a penalty metric, which is minimized. The strong stability of these virtual queues guarantees the respective time average constraints.

$$X_i(t+1) \triangleq \max[X_i(t) - \gamma_i\ ,\ 0] + P_i(t) \tag{9}$$

Let $\mathbf{X}(t)$ be the vector of $X_i(t), \forall i \in \mathcal{N}$. Let $L(\mathbf{X}(t)) \triangleq 1/2 \sum_{i=1}^{N} X_i^2(t)$ be defined as the quadratic Lyapunov function and $\Delta(\mathbf{X}(t))$ be the conditional Lyapunov drift with respect to the random channel states and arrivals:

$$\Delta(\mathbf{X}(t)) \triangleq \mathrm{E}[L(\mathbf{X}(t+1)) - L(\mathbf{X}(t)) \mid \mathbf{X}(t)] \tag{10}$$

At every time slot t, the problem in (8) is solved by determining $\mathbf{P}(t)$ in order to minimize the $drift - plus - penalty$ expression[23], $\Delta(\mathbf{X}(t)) + V \sum_{i=1}^{N} \mathrm{E}[\mathcal{F}_i \mid \mathbf{X}(t)]$, where $V > 0$ is a weight parameter to scale the tradeoff between the average power constraint and the penalty related to the deadline. The dynamic policy is obtained by applying the principle of opportunistically minimizing an expectation[23] on the upper bound analysis[26] of the $drift - plus - penalty$ expression. We observe QCSI and

[†]For simplicity of the notation, $\mathcal{F}_i(t, \alpha_i(t), \phi(t))$ is referred as $\mathcal{F}_i$.

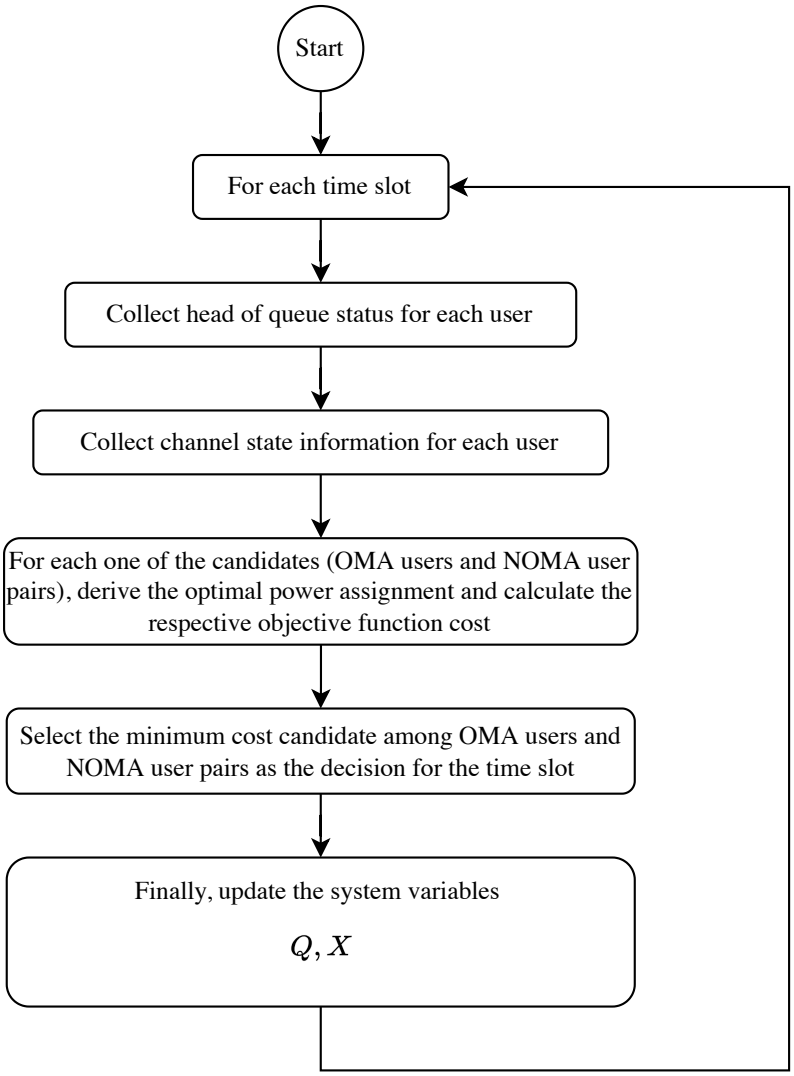

Figure 1 Flow chart of the proposed algorithm.

determine $\mathbf{P}(t)$ by solving the following $drift - plus - penalty$ problem at each time slot t:

$$\min_{\mathbf{P}(t)} \mathcal{M}(\mathbf{P}(t)) = \sum_{i=1}^{N} \left(V\mathcal{F}_i + X_i(t)(P_i(t) - \gamma_i) \right) \tag{11a}$$

$$\text{s.t.} \quad \mathbf{P}(t) \in \mathcal{P}^H(t) \tag{11b}$$

Due to our single transmitter model, there are N possible OMA transmissions and $\binom{N}{2}$ possible NOMA transmissions. Thus, the scheduling complexity is $O(N^2)$. At first, optimal power assignment for each transmission scheme is calculated individually and then the optimization metric in (11) is exhaustively minimized. One of the minimizing candidates is selected randomly for the sake of fairness among users. Flow chart of the proposed algorithm is presented in Fig 1.

4 | OPTIMAL POWER ASSIGNMENT FOR HYBRID MA

4.1 | Optimal Power Assignment for OMA

In this part, optimal power assignment is explained given that OMA transmission scheme is occupied for user i, such that $\Psi_i = i$. Based on [31], departure rate $\mu_i(P_i)$** for user i using OMA with FBL $\tau\mathcal{B}$ and BLER ϵ_O is given by:

$$\mu_i(P_i) = \tau\mathcal{B}\log_2\left(1 + \frac{|\,h_i\,|^2\,P_i}{\mathcal{B}N_0}\right) - \sqrt{\tau\mathcal{B}\mathcal{V}_i}\cdot\frac{Q^{-1}(\epsilon_O)}{\ln 2} \tag{12}$$

where N_0 is the noise power spectral density. As a reliable communication demands SNR to be sufficiently high, we consider $\mathcal{V}_i \approx 1$ in this high SNR regime. Let P_i^{req} and P_i^{one} be the required power value for transmission of q_i bits and 1 bit, respectively, so that $\mu_i(P_i^{req}) = q_i$ and $\mu_i(P_i^{one}) = 1$. Moreover, let P_i^{min} be the minimum required power to successfully perform an OMA transmission under deadline constraint, so that, at least 1 and q_i bits must be transmitted for $d_i > 1$ and $d_i = 1$ cases, respectively. Then, P_i^{min} can be defined as follows:

$$P_i^{min} = P_i^{req}\mathbb{1}\left\{d_i = 1\right\} + P_i^{one}\mathbb{1}\left\{d_i > 1\right\} \tag{13}$$

Let $P_i^{\max} \triangleq \min(P_i^{req}, P_0)$ and $\mathcal{P}_i$ be the available power region for an OMA transmission of user i. $\mathcal{P}_i = \{0\} \cup [P_i^{min}, P_i^{\max}]$ when $P_i^{min} \leq P_i^{\max}$. Otherwise, $\mathcal{P}_i = \{0\}$, since the respective packet will definitely be dropped when $d_i = 1$. Let the objective function for user i under OMA transmission scheme be $\mathcal{M}_i^O(P_i)$ and the resulting OMA power optimization problem is stated as:

$$P_i^O = \arg\min_{P_i} \mathcal{M}_i^O(P_i) \triangleq \mathcal{M}(\mathbf{P}) \tag{14a}$$

$$\text{s.t.} \quad P_i \in \mathcal{P}_i,\ {}^{P_j=0}_{\forall j\neq i} \tag{14b}$$

where P_i^O is the optimal power value for user i under OMA transmission scheme. The solution of the optimization problem above is presented in Theorem 1.

Theorem 1. Optimal power allocation for OMA can be achieved with FTM ¶. P_i^O is given by:

$$P_i^O = \begin{cases} \overline{P_i^O} & \text{, if } \ \mathcal{M}_i^O(0)\Big|_{\Phi_F} > \mathcal{M}_i^O(\overline{P_i^O})\Big|_{\Phi_F} \\ 0 & \text{, otherwise} \end{cases} \tag{15}$$

**Since $\mathbf{P}(t)$ is optimized slot-by-slot, in the rest of the paper the time parameter t is removed for simplification, and $\mathbf{P}$ indicates $\mathbf{P}(t)$. Similarly, the time index is removed from all other parameters that depend on time.

¶The occupied TM is represented as $\Big|_{\phi}$ within equations for simplification. Thus, FTM and CTM are indicated as $\Big|_{\Phi_F}$ and $\Big|_{\Phi_C}$, respectively.

where

$$\overline{P_i^O} = \begin{cases} P_i^{\max} & \text{, if } \quad P_i^{\max} \leq P_i^* \\ P_i^* & \text{, if } \quad P_i^{min} \leq P_i^* < P_i^{\max} \\ P_i^{\min} & \text{, if } \quad P_i^* < P_i^{min} \end{cases} \tag{16a}$$

$$P_i^* = \Gamma_i/(X_i \ln 2) - (\mathcal{B}N_0)/ \mid h_i \mid^2 \tag{16b}$$

$$\Gamma_i = V(1-(d_i-1)/m_i)^{\alpha_i}(\tau\mathcal{B})/q_i \tag{16c}$$

Proof. The proof can be found in Appendix A. □

4.2 | Optimal Power Assignment for NOMA

In this part, power assignment is explained given that NOMA transmission scheme is occupied for the paired users $\{i, j\}$, such that $\Psi_i = j$ and $\Psi_j = i$. Assume that $|h_j|^2 > |h_i|^2$, thus the channel of user i is weaker. In theory, allocation of higher power to the user with weaker channel is not necessary in NOMA[34]. Therefore, allocated power values are only constrained with total power budget: $P_i + P_j \leq P_0$. Based on[31], departure rates $\mu_{i,(i,j)}(P_i, P_j)$ and $\mu_{j,(i,j)}(P_j)$ with FBL $\tau\mathcal{B}$ and BLER ϵ_N are given by:

$$\mu_{i,(i,j)} = \tau B \log_2\left(1 + \frac{|h_i|^2 P_i}{|h_i|^2 P_j + BN_0}\right) - \sqrt{\tau\mathcal{B}\mathcal{V}_i}\frac{Q^{-1}(\epsilon_N)}{\ln 2} \tag{17}$$

$$\mu_{j,(i,j)} = \tau B \log_2\left(1 + \frac{|h_j|^2 P_j}{BN_0}\right) - \sqrt{\tau\mathcal{B}\mathcal{V}_j}\frac{Q^{-1}(\epsilon_N)}{\ln 2} \tag{18}$$

where $\mathcal{V}_i \approx 1$ and $\mathcal{V}_j \approx 1$ in the high SNR regime. Let $\{P_{j,(i,j)}^{req}, P_{j,(i,j)}^{one}\}$ and $\{P_{i,(i,j)}^{req}, P_{i,(i,j)}^{one}\}$ be the required power values for transmitting $\{q_j, 1\}$ and $\{q_i, 1\}$ bits, respectively, of user pair $\{i, j\}$ within a slot under the NOMA transmission scheme. Moreover, let $P_{k,(i,j)}^{min}$, $k \in \{i, j\}$ be the minimum required power to successfully perform a NOMA transmission under deadline constraint, as follows:

$$P_{k,(i,j)}^{min} = P_{k,(i,j)}^{req}\mathbb{1}\left\{d_k = 1\right\} + P_{k,(i,j)}^{one}\mathbb{1}\left\{d_k > 1\right\} \tag{19}$$

Thus, we have $\mu_{j,(i,j)}(P_{j,(i,j)}^{req}) = q_j$, $\mu_{j,(i,j)}(P_{j,(i,j)}^{one}) = 1$, $\mu_{i,(i,j)}(P_{i,(i,j)}^{req}, P_{j,(i,j)}^{min}) = q_i$, and $\mu_{i,(i,j)}(P_{i,(i,j)}^{one}, P_{j,(i,j)}^{min}) = 1$. Let $P_{(i,j)}^{\max} \triangleq \min(P_0\ , (P_{j,(i,j)}^{req} + P_{i,(i,j)}^{req}))$, $P_{(i,j)}^{\min} \triangleq (P_{j,(i,j)}^{min} + P_{i,(i,j)}^{min})$, and $\mathcal{P}_{(i,j)}$ be the available total power region for a NOMA transmission of user pair $\{i, j\}$. $\mathcal{P}_{(i,j)} = \{0\} \cup [P_{(i,j)}^{\min}, P_{(i,j)}^{\max}]$ when $P_{(i,j)}^{\min} \leq P_{(i,j)}^{\max}$. Otherwise, $\mathcal{P}_{(i,j)} = \{0\}$, since the respective packet will definitely be dropped when $d_i = 1$ or $d_j = 1$. Let $\mathcal{M}_{(i,j)}^N(P_i, P_j) = \mathcal{M}(\mathbf{P})$ where $P_k = 0,\ \forall k \neq \{i, j\}$ be the objective function for $\{i, j\}$ under NOMA scheme. Assume that FTM is selected. Then, optimization problem of power assignment for FTM based NOMA

transmission of user pair $\{i, j\}$ can be expressed as:

$$\{P^N_{i,(i,j)}\Big|_{\Phi_F}, P^N_{j,(i,j)}\Big|_{\Phi_F}\} = \arg\min_{P_i,P_j} \mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F} \tag{20a}$$

$$s.t.\ \ P_i + P_j \in \mathcal{P}_{(i,j)} \tag{20b}$$

where $\{P^N_{i,(i,j)}\Big|_{\Phi_F}, P^N_{j,(i,j)}\Big|_{\Phi_F}\}$ are optimal power values under FTM based NOMA transmission schemes for users i and j, respectively. $\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F}$ can be expressed as:

$$\begin{aligned}\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F} &= -\Gamma_i \log_2\left(1 + \frac{|h_i|^2 P_i}{|h_i|^2 P_j + \mathcal{B}N_0}\right) \\ &- \Gamma_j \log_2\left(1 + \frac{|h_j|^2 P_j}{\mathcal{B}N_0}\right) + X_i P_i + X_j P_j + C^N_{(i,j)}\end{aligned} \tag{21}$$

where $C^N_{(i,j)}$ is a constant. Note that $\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F} \geq 0$ for $P_k \geq P^{min}_{k,(i,j)}$, $k \in \{i, j\}$. Moreover, the reduction ratio in (7) is non-negative for $P_k \in [P^{min}_{k,(i,j)}, P^{req}_{k,(i,j)}]$, $k \in \{i, j\}$. Since $\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F}$ in (21) is not convex, the optimization problem in (20) is not convex, too. In order to solve this problem, an auxilary variable $\theta = P_i + P_j$ is introduced and solution process of the problem is divided in two consecutive sub-problems. At first, optimal value of θ is calculated. Then, optimal value of P_j is calculated for given θ. The first sub-problem is to find the optimal value of θ as follows:

$$\theta^N = \arg\min_{\theta} g(\theta, P_j) \tag{22a}$$

$$s.t.\ \ \theta \in [P^{min}_{(i,j)}, P^{max}_{(i,j)}] \tag{22b}$$

where θ^N is the optimal value and $g(\theta, P_j)$ is given as follows:

$$\begin{aligned}g(\theta, P_j) &= -\Gamma_i \log_2\left(\frac{|h_i|^2\theta + \mathcal{B}N_0}{|h_i|^2 P_j + \mathcal{B}N_0}\right) + X_i\theta \\ &- \Gamma_j \log_2\left(1 + \frac{|h_j|^2 P_j}{\mathcal{B}N_0}\right) + \left(X_j - X_i\right) P_j + C^N_{(i,j)}\end{aligned} \tag{23}$$

The solution for the optimal θ^N is presented below.

Theorem 2. Optimal θ^N for FTM based NOMA transmission scheme is given by:

$$\theta^N = \begin{cases} P^{\max}_{(i,j)} & \text{, if } \ P^{\min}_{(i,j)} \leq P^{\max}_{(i,j)} \leq \theta^* \\ \theta^* & \text{, if } \ P^{\min}_{(i,j)} \leq \theta^* \leq P^{\max}_{(i,j)} \\ P^{\min}_{(i,j)} & \text{, if } \ \theta^* \leq P^{\min}_{(i,j)} \leq P^{\max}_{(i,j)} \end{cases} \tag{24}$$

where

$$\theta^* = \frac{\Gamma_i}{X_i \cdot \ln 2} - \frac{\mathcal{B} \cdot N_0}{|h_i|^2} \tag{25}$$

Proof. Proof can be found in Appendix B. □

The second sub-problem is to find the optimal value of P_j, given the auxiliary variable θ^N, as follows:

$$P^{'}_{j,(i,j)}\Big|_{\Phi_F} = \arg\min_{P_j} g(\theta^N, P_j) \tag{26a}$$

$$s.t.\ \ P_i + P_j = \theta^N \tag{26b}$$

where $P^{'}_{j,(i,j)}\Big|_{\Phi_F}$ is the optimal value. The solution of the second sub-problem to find $P^{'}_{j,(i,j)}\Big|_{\Phi_F}$ is presented below.

Theorem 3. Suppose that $\Gamma_j \geq \Gamma_i$. Then, optimal $P^{'}_{j,(i,j)}$ for FTM based NOMA transmission scheme for $\theta \in [P^{min}_{(i,j)}, P^{max}_{(i,j)}]$ is given by:

$$P^{'}_{j,(i,j)}\Big|_{\Phi_F} = \begin{cases} P^{min}_j & \text{, if } \dfrac{dg(P_j)}{dP_j}\Big|_{P_j=P^{min}_j} \geq 0 \\ P^{\dagger}_j & \text{, if } \dfrac{dg(P_j)}{dP_j}\Big|_{P_j=P^{\dagger}_j} \leq 0 \\ P^{\ddagger}_j & \text{, otherwise} \end{cases} \tag{27}$$

where $P^{\dagger}_j = \min((\theta - P^{min}_{i,(i,j)}), P^{req}_{j,(i,j)})$, $\dfrac{d\ g(P_j)}{d\ P_j}\Big|_{P_j=P^{*}_j} = 0$ and $P^{\ddagger}_j = \min(P^{*}_j, P^{\dagger}_j)$.

Proof. Proof can be found in Appendix C. □

The CTM based objective function $\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_C}$ can be written as:

$$\begin{aligned} \mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_C} &= X_i P_i + X_j P_j + C^N_{(i,j)} \\ &+ \sum_{k\in\{i,j\}} \left(\frac{\Gamma_k q_k}{\tau \mathcal{B}}\right)\left(\mathbb{1}\left\{P_k < P^{req}_{k,(i,j)}\right\} - 1\right) \end{aligned} \tag{28}$$

Note that $\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F} \leq \mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_C}$ when the pair $\{P_i, P_j\}$ is $\{0, 0\}$ or $\{P^{req}_{i,(i,j)}, P^{req}_{j,(i,j)}\}$. Moreover, $\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_C}$ increases linearly with P_i and P_j. Since FTM based objective functions in (21) is convex for $\Gamma_j \geq \Gamma_i$, we can conclude that $\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F} \leq \mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_C}$ for $\Gamma_j \geq \Gamma_i$. Therefore, Theorem 3 provides optimal power $P^{'}_{j,(i,j)}$ under the assumption that $\Gamma_j \geq \Gamma_i$. The solution for remaining cases $\Gamma_j < \Gamma_i$ is derived using the approach proposed in [26]. The optimization problem of power assignment for CTM based NOMA transmission can be expressed as:

$$\arg\min_{P_i, P_j}\ \mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_C} \tag{29a}$$

$$s.t.\ \ P_i + P_j \in \mathcal{P}_{(i,j)} \tag{29b}$$

where $\{P^{'}_{i,(i,j)}\Big|_{\Phi_C}, P^{'}_{j,(i,j)}\Big|_{\Phi_C}\}$ are optimal power values under CTM based NOMA transmission schemes for users i and j, respectively. The solution of the respective optimization problem in (29) is given below.

Theorem 4. Let users $\{i, j\}$ be such that $|h_j| > |h_i|$. The optimal $P'_{j,(i,j)}\Big|_{\Phi_C}$ and $P'_{i,(i,j)}\Big|_{\Phi_C}$ for NOMA transmission scheme under CTM is given by: When $(P^{req}_{i,(i,j)} + P^{req}_{j,(i,j)} \leq P^{\max}_{(i,j)})$,

$$\{P'_{i,(i,j)}\Big|_{\Phi_C}, P'_{j,(i,j)}\Big|_{\Phi_C}\} = \{P^{req}_{i,(i,j)}, P^{req}_{j,(i,j)}\}, \tag{30}$$

otherwise,

$$\{P'_{i,(i,j)}\Big|_{\Phi_C}, P'_{j,(i,j)}\Big|_{\Phi_C}\} = \{0, 0\}. \tag{31}$$

Proof. Proof can be found in Appendix D. □

In this study, power allocation decision under NOMA transmission scheme in a Hybrid MA scenario for a user pair $\{i, j\}$, such that $|h_j| > |h_i|$, is as follows:

$$\{P_i, P_j\} = \begin{cases} \{P'_i, P'_j\} & \text{, if } \mathcal{M}^N(0,0) > \mathcal{M}^N(P'_i, P'_j) \\ \{0, 0\} & \text{, otherwise} \end{cases} \tag{32}$$

where $\{P'_i, P'_j\}$ is as follows:

$$\{P'_i, P'_j\} = \begin{cases} \{(\theta^N - P'_{j,(i,j)}\Big|_{\Phi_F}), P'_{j,(i,j)}\Big|_{\Phi_F}\} & \text{, if } (\Gamma_j \geq \Gamma_i) \\ \{P'_{i,(i,j)}\Big|_{\Phi_C}, P'_{j,(i,j)}\Big|_{\Phi_C}\} & \text{, otherwise.} \end{cases} \tag{33}$$

5 | NUMERICAL RESULTS AND DISCUSSIONS

In this section, we comparatively evaluate the MA performances in terms of timely throughput. For the simulations, let α, γ, m, π be the system parameters such that $\alpha = \alpha_i(t)$, $\gamma = \gamma_i$, $m = m_i$, $\pi = \pi_i$ for all t and $i \in \mathcal{N}$. In this case, timely throughput can be represented as $\pi - \overline{D}$. The target BLER is considered as 10^{-5}. Due to successive interference cancellation (SIC), NOMA related target BLER is set as $\epsilon_N = 5 \cdot 10^{-6}$, so that overall system target BLER is ensured. The default values of all parameters used in obtaining the numerical results are given in Table 1, unless otherwise stated. Simulations are performed for 1000 random seeds and their averages are reported. In order to assess fairness among $\overline{D_i}$, we considered Jain's Fairness Index (FI)[35] as $FI = (\sum_i \overline{D_i})^2/(N \sum_i \overline{D_i}^2)$. Let $\mathbb{D}^O(\pi, N)$ represent the set of all achievable per user drop rates using OMA for a given $\{\pi, N\}$ pair. Then, lower bound $\inf \mathbb{D}^O(\pi, N) = \max(0, \pi N - 1)/N$ is the minimum average unserved arrival rate per user. As N increases, $\inf \mathbb{D}^O(\pi, N)$ increases towards π. Therefore, in order to clearly analyse the impact of other parameters on the system, we select $N = 5$ in Table 1. The problem becomes intractable for low values of m due to resultant unavoidable unfairness among users. For high values of m, the problem becomes relaxed. In accordance with the selection for N, we select $m = 5$ in Table 1. The algorithms proposed in this paper are as follows. FTM-based OMA is indicated as soft, others as hard.

Table 1 Simulation parameters and default values

V (Weight parameter)	100
Λ (Packet size)	$(160 \cdot w)_{w=1}^{20}$ bits
α	0.1
γ (Average power constraint for all users)	0.6 W
π (Arrival rate for all users)	0.3
R (Cell radius)	50 m
$\#Slots$ (Simulation slot count)	10^4
τ (Slot duration)	0.1 ms
$\mathcal{B}$ (Bandwidth)	1 MHz
P_0 (Power budget)	3 W
N (User count)	5
m (Deadline slot count)	5
Path loss	$35.3 + 37.6log_{10}(r_i)$ dB
Fast fading component	$CN(0, 1)$
N_0 (Noise power spectral density)	$-174dBm/Hz$

- soft-OMA (s-OMA): Optimal power assignment for OMA with FTM is considered and Theorem 1 is used. This algorithm performs RDPPA.
- Hybrid-MA (H-MA): Optimal power assignment for Hybrid MA is considered. The content of "Hybrid" consists of the NOMA and OMA using (32) and Theorem 1, respectively. This algorithm performs RDPPA.
- NOMA: Optimal power assignment for only NOMA is considered, using (32).

The following algorithms are compared with the above algorithms as baseline references:

- hard-OMA (h-OMA): Optimal power assignment for OMA with CTM is considered. The CTM based objective function (A4) is minimized. The dynamic power allocation approach presented in[26] is used.
- EDF[30]: It uniformly selects one of the users with shortest remaining expiration time and performs OMA using available power budget without any power constraint.
- P_{inf}-OMA (p-OMA): This algorithm performs OMA using infinite power budget P_{inf} without any power constraint for a uniformly selected user. p-OMA drop rate achieves inf $\mathbb{D}^O(\pi, N)$ for any given $\{\pi, N\}$.

In Figs. 2,4,6, $\overline{D}$ with different values of α under $\pi \in \{0.1, 0.3, 0.5\}$ are presented, respectively. Moreover, the results with $V \in \{10, 100, 1000\}$ are presented. Since h-OMA, EDF and p-OMA algorithms are independent of α, results for each of these algorithms are the same for all α values.

Observations on Figs. 2,4,6 show the consistent relation between α and $\overline{D}$ under different traffic levels. Moreover, $\overline{D}$ converges at around $V = 100$ and increasing V towards 1000 does not make much difference on $\overline{D}$. Therefore, $V = 100$ is considered as

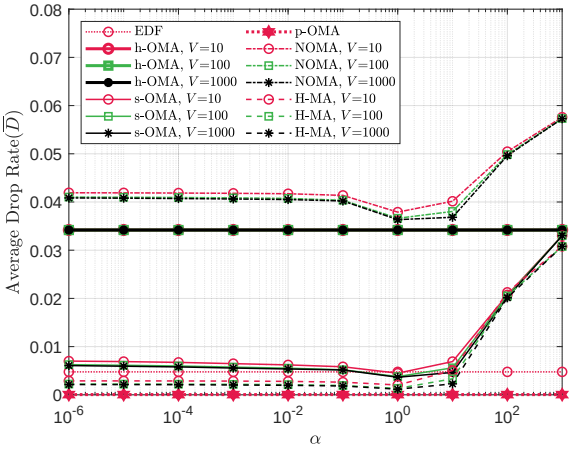

Figure 2 Average drop rate as a function of α with $\pi = 0.1$.

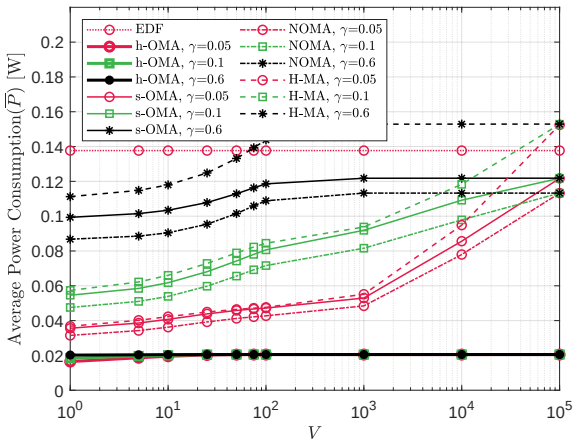

Figure 3 Average power consumption as a function of γ with $\pi = 0.1$.

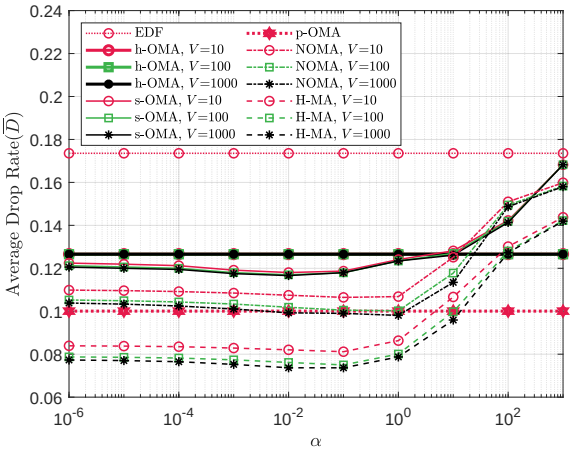

Figure 4 Average drop rate as a function of α with $\pi = 0.3$.

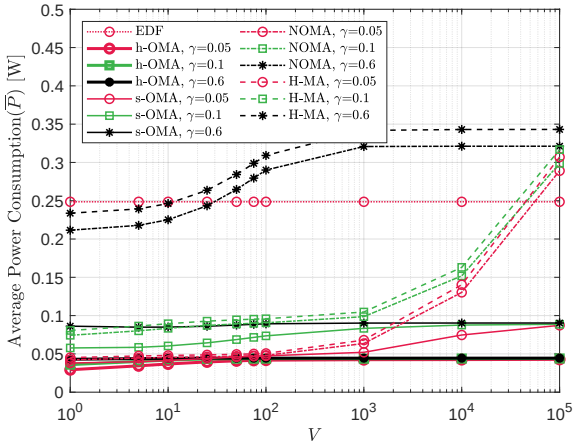

Figure 5 Average power consumption as a function of γ with $\pi = 0.3$.

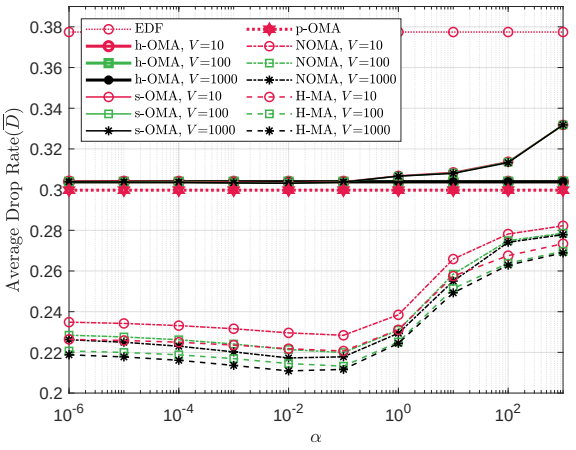

Figure 6 Average drop rate as a function of α with $\pi = 0.5$.

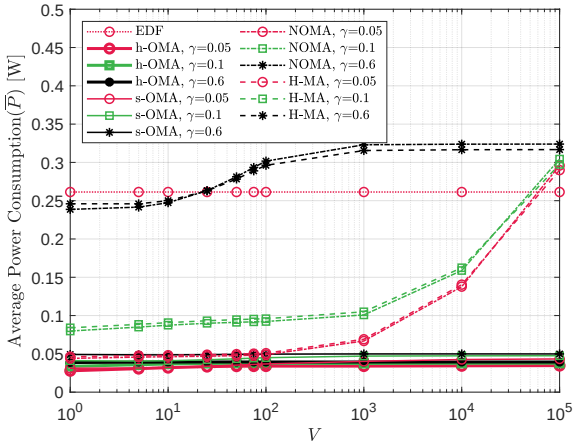

Figure 7 Average power consumption as a function of γ with $\pi = 0.5$.

the most suitable choice in terms of the tradeoff between $\overline{D}$ and γ. In Fig. 2, p-OMA can fully serve the arrival rate of $\pi = 0.1$. H-MA and s-OMA perform close to it with $\alpha = 1$. In this low traffic level, EDF performs well, too. Interestingly, h-OMA and NOMA perform considerably worse compared to other algorithms. The reason is that, low arrival rate $\pi = 0.1$ results in smaller number of packets with small sizes. In such a case, binary decision of CTM is not a good choice and the existence probability of a packet pair suitable for an effective NOMA transmission is reduced. In Fig. 4 with $\pi = 0.3$, p-OMA cannot fully serve the arriving traffic. However, H-MA with $\alpha = 0.1$ outperforms the minimum achievable $\overline{D}$ with OMA. Since the packet diversity increases with the increasing arrival rate, h-OMA starts to perform close to the best performance of s-OMA with $\alpha = 0.01$. Moreover, NOMA starts to perform closer to H-MA. Finally, in Fig. 6 with $\pi = 0.5$, H-MA with $\alpha = 0.1$ performs about 30% better than p-OMA. These results show that with increasing traffic load and packet size diversity, NOMA capable H-MA and NOMA significantly outperform other algorithms.

In all traffic levels, H-MA performs the best, showing that Hybrid MA is a robust approach and RDPPA with proper selection of α increases timely throughput. The best performance for H-MA and s-OMA under different traffic levels are obtained when $\alpha \in \{0.01, 0.1, 1\}$. Therefore, $\alpha = 0.1$ is considered for the rest of the simulations. Finally, EDF performs significantly worse as the traffic level increases, showing that simple algorithms like EDF are not viable.

In Figs. 3,5,7, $\overline{P}$ with different values of γ and $\pi \in \{0.1, 0.5\}$ are presented, respectively. Since EDF is independent of V, the results of EDF are the same for all V values. H-MA is the most sensitive algorithm to the given average power utilization constraint and NOMA consumes slightly less power. Significantly high values of V forces the algorithms to ignore the provided average power constraint in Figs. 3,5,7. For all of the arrival rates $\pi \in \{0.1, 0.5\}$, H-MA, NOMA and s-OMA start to violate provided average power constraint as V increases from 100, which is observed also for $\pi = 0.3$ in Fig. 5. Based on these observations, it can be concluded that $V = 100$ is suitable for balancing the tradeoff between $\overline{D}$ and $\overline{P}$.

The percentage of increase in timely throughput achieved by H-MA compared to s-OMA and NOMA for different π and γ values under $\alpha = 0.1$ and $V = 100$ is presented in Tables 2 and 3, respectively. In 2, increase on timely throughput using H-MA with respect to s-OMA is small for $\pi = 0.1$. As traffic level increases with $\pi = 0.5$, H-MA increases timely throughput up to 46%. Based on these results, we can conclude that H-MA increases timely throughput compared to s-OMA on the average by nearly 21.27% while satisfying average power constraints for all arrival rates. On the other hand, increase on timely throughput using H-MA with respect to NOMA is high for $\pi = 0.1$ and it decreases as traffic level increases with $\pi = 0.5$. The reason for this inversely proportional behavior is that, low traffic levels can be handled by s-OMA, whereas the existence probability of a suitable NOMA user pair is proportional to the increasing traffic level. Finally, these results show that H-MA is also robust under varying traffic levels.

In Table 5, mean and variance of the $\overline{D_i}$'s FI over 1000 simulations under the design parameters $\alpha = 0.1$ and $V = 100$ are presented for traffic level $\pi = 0.3$. The OMA based algorithms, h-OMA and s-OMA, are slightly more fair in terms of average

Table 2 In H-MA and NOMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

		π		
		0.1	0.3	0.5
γ	0.05	3.9085 %	18.0011 %	39.1193 %
	0.1	3.6803 %	20.7381 %	43.9200 %
	0.6	3.5240 %	26.6297 %	50.0514 %

Table 3 In H-MA and s-OMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

		π		
		0.1	0.3	0.5
γ	0.05	66.7241 %	13.0315 %	4.2492 %
	0.1	65.4238 %	13.0389 %	3.7622 %
	0.6	64.4219 %	12.7149 %	2.3997 %

Table 4 Mean and variance of FI for different algorithms and γ values under $\pi = 0.1$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.86292	0.86292	0.86293	0.013101	0.01311	0.013111
s-OMA	0.71412	0.722	0.86293	0.015277	0.015563	0.015428
NOMA	0.92265	0.92813	0.93506	0.001781	0.0016488	0.0015408
H-MA	0.77739	0.79863	0.82947	0.01901	0.019065	0.018031

Table 5 Mean and variance of FI for different algorithms and γ values under $\pi = 0.3$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.91076	0.92004	0.92665	0.0048109	0.0037833	0.0031237
s-OMA	0.89435	0.90751	0.92444	0.0049643	0.0037439	0.0023825
NOMA	0.87862	0.88264	0.9001	0.0061223	0.006124	0.00579
H-MA	0.85498	0.86147	0.88792	0.011707	0.011834	0.010868

packet dropping rate compared to NOMA based H-MA and NOMA algorithms, since their average FI is higher and variance of FI is lower. Same relation is observed also for $\pi = 0.1$ and $\pi = 0.5$ traffic levels in Tables 4 and 6, respectively. Thus, we can conclude that, NOMA based H-MA significantly increases overall performance at the cost of slight decrease in FI.

In Fig. 8, average incomplete bit-rate ($\overline{I}$) with different values of N under the arrival rate $\pi = 0.3$ is shown. $\overline{I}$ refers to the average data rate, in bits per slot, of dropped packets for which all fragments originated from FTM could not be successfully transmitted by the packet deadline. $\overline{I}$ value of H-MA is lower than $\overline{I}$ value of NOMA for small values of N, which shows the benefit of Hybrid MA in terms of avoiding packet dropping by suitable decisions between OMA and NOMA transmission. $\overline{I}$

Table 6 Mean and variance of FI for different algorithms and γ values under $\pi = 0.5$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.97091	0.97978	0.98588	0.00053733	0.00026329	0.00012353
s-OMA	0.97026	0.97949	0.98592	0.00054142	0.00027355	0.00011804
NOMA	0.91289	0.91341	0.92093	0.0049273	0.0049922	0.0047763
H-MA	0.90627	0.9089	0.92179	0.0063586	0.0062572	0.0056853

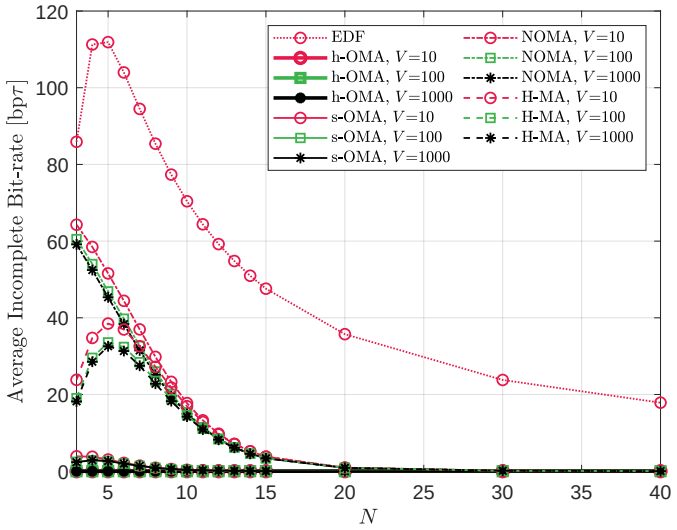

Figure 8 Average incomplete bit-rate as a function of N with $\pi = 0.3$.

for s-OMA, NOMA and H-MA converges to 0 as N increases, due to the increase in the existence probability of a small packet from Λ on a strong channel. For small N, such as $N = 5$, we observe high $\overline{I}$ for NOMA and H-MA. Although s-OMA, NOMA and H-MA algorithms reduce $\overline{D}$ compared to baseline algorithms, they increase $\overline{I}$ due to dynamic nature of the opportunistic scheduling.

6 | CONCLUSION

In this study, we address the problem of latency constrained communications with strict deadlines under average power constraint using OMA-NOMA based Hybrid MA. We develop a dynamic algorithm that assigns user power in real-time to minimize the packet drop rate under average power constraints. Numerical results show that Hybrid MA increases timely throughput compared to OMA-only case by up to 46% and on the average by 21.27% while satisfying average power constraints. RDPPA introduces a novel degree of freedom for the packet drop rate minimization by prioritizing packets in the system considering both their remaining deadlines as well as the channel states. In order to further increase the performance of the proposed s-OMA and H-MA, techniques to reduce the incomplete transmission rate can be studied as a future work.

□

APPENDIX

A PROOF OF THEOREM 1

Assume that transmission is performed using OMA, such that $P_i \in [P_i^{\min}, P_i^{\max}]$ and $\forall j \neq i,\ P_j = 0$. The FTM based objective function $\mathcal{M}_i^O(P_i)\Big|_{\Phi_F}$ can be written as

$$\begin{aligned}\mathcal{M}_i^O(P_i)\Big|_{\Phi_F} &= -V\left(\frac{m_i-(d_i-1)}{m_i}\right)^{\alpha_i}\frac{\tau\mathcal{B}}{q_i}\log_2\left(1+\frac{|h_i|^2P_i}{\mathcal{B}N_0}\right)+X_iP_i+C_i^O \\ &= -\Gamma_i\cdot\log_2\left(1+\frac{|h_i|^2P_i}{\mathcal{B}N_0}\right)+X_iP_i+C_i^O\end{aligned} \tag{A1}$$

where C_i^O is a constant given by

$$C_i^O = V\sum_{j=1}^{N}\left(\frac{m_j-(d_j-1)}{m_j}\right)^{\alpha_i} + \sum_{j=1}^{N}X_j(-\gamma_j) + \Gamma_i\sqrt{\frac{\mathcal{V}_i}{\tau\mathcal{B}}}\frac{Q^{-1}\left(\epsilon^O\right)}{\ln 2} \tag{A2}$$

Note that $\mathcal{M}_i^O(P_i \in \mathcal{P}_i)\Big|_{\Phi_F} \geq 0$ and the reduction ratio in (7) is non-negative for $P_i \in [P_i^{\min}, P_i^{\max}]$. It can be shown that $\mathcal{M}_i^O(P_i)\Big|_{\Phi_F}$ is convex for $P_i \in [P_i^{\min}, P_i^{\max}]$. The global minimizer P_i^* for FTM is given as,

$$P_i^* = \frac{\Gamma_i}{X_i\ln 2} - \frac{\mathcal{B}N_0}{|h_i|^2}. \tag{A3}$$

The CTM based objective function $\mathcal{M}_i^O(P_i)\Big|_{\Phi_C}$ can be written as

$$\mathcal{M}_i^O(P_i)\Big|_{\Phi_C} = V\left(\frac{m_i-(d_i-1)}{m_i}\right)^{\alpha_i}\cdot\left(\mathbb{1}\left\{P_i < P_i^{req}\right\}-1\right)+X_iP_i+C_i^O. \tag{A4}$$

Note that FTM and CTM based objective functions in (A1) and (A4), respectively, are equal for $P_i \in \left\{0, P_i^{req}\right\}$. CTM based objective function in (A4) increases linearly with P_i for $P_i \in [0, P_i^{req})$. Since FTM based objective function in (A1) is convex for $P_i \in [P_i^{\min}, P_i^{\max}]$, we can conclude that $\mathcal{M}_i^O(P_i)\Big|_{\Phi_F} \leq \mathcal{M}_i^O(P_i)\Big|_{\Phi_C}$ for $P_i \in [0, P_i^{\max}]$. Thus, FTM is always a better option in terms of optimizing power assignment for OMA over the all available power region. P_i^O in (15) can be obtained using P_i^* and available power region $P_i \in \mathcal{P}_i$.

B PROOF OF THEOREM 2

It can be shown that $g(\theta, P_j)$ is convex for $\theta \in [P_{(i,j)}^{min}, P_{(i,j)}^{max}]$. The global minimizer θ^* is given as,

$$\theta^* = \frac{\Gamma_i}{X_i\ln 2} - \frac{\mathcal{B}N_0}{|h_i|^2}. \tag{B5}$$

θ^N in (24) can be obtained using θ^* and available total power region $\theta \in \mathcal{P}_{(i,j)}$.

C PROOF OF THEOREM 3

It can be shown that $g(\theta^N, P_j)$ is convex for $\Gamma_j \geq \Gamma_i$. The minimizer P_j^* is given as

$$P_j^* = \begin{cases} \mathcal{B}N_0 \left(\dfrac{\Gamma_j |h_j|^2 + \Gamma_i |h_i|^2}{\left(\Gamma_i - \Gamma_j\right) |h_i|^2 |h_j|^2} \right) & \text{, if } X_i = X_j \\ \max\left(\zeta_1, \zeta_2\right) & \text{, otherwise} \end{cases} \tag{C6}$$

where ζ_1 and ζ_2 are the roots of the polynomial $\mathtt{a} \cdot P_j^2 + \mathtt{b} \cdot P_j + \mathtt{c} = 0$ such that

$$\mathtt{a} = \ln 2 \left(X_j - X_i\right) \tag{C7a}$$

$$\mathtt{b} = \Gamma_i - \Gamma_j + \ln 2 \left(X_j - X_i\right) \mathcal{B}N_0 \left(\frac{1}{|h_i|^2} + \frac{1}{|h_j|^2} \right) \tag{C7b}$$

$$\mathtt{c} = \mathcal{B}N_0 \left(\frac{\Gamma_i}{|h_i|^2} - \frac{\Gamma_j}{|h_j|^2} \right) + \frac{\left(X_j - X_i\right) \ln 2 \left(\mathcal{B}N_0\right)^2}{|h_i|^2 + |h_j|^2} \tag{C7c}$$

Under $\Gamma_j \geq \Gamma_i$ condition, $P'_{j,(i,j)}\Big|_{\Phi_F}$ in (27) can be obtained using P_j^* and available power region $P_j \in \{0\} \cup [P_{j,(i,j)}^{min}, P_j^{\dagger}]$.

D PROOF OF THEOREM 4

Note that $P_j^{req} = P_{j,(i,j)}^{req}$ and $P_i^{req} < P_{i,(i,j)}^{req}$. We have

$$\mathcal{M}_j^O(P_j)\Big|_{\Phi_C} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j)\Big|_{\Phi_C} \text{ when } P_i < P_{i,(i,j)}^{req} \text{ and } P_j = P_{j,(i,j)}^{req} = P_j^{req} \tag{D8}$$

$$\mathcal{M}_i^O(P_i)\Big|_{\Phi_C} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j)\Big|_{\Phi_C} \text{ when } P_i = P_{i,(i,j)}^{req} > P_i^{req} \text{ and } P_j < P_{j,(i,j)}^{req}. \tag{D9}$$

Then, we can conclude that CTM based OMA is better than CTM based NOMA when only one of users' packet is completed in a NOMA pair. Therefore, these cases can be ignored for CTM based NOMA in the Hybrid MA scenario. Moreover, since $\mathcal{M}_{(i,j)}^N(P_i, P_j)\Big|_{\Phi_C}$ linearly increases with P_i and P_j, we can conclude that $\mathcal{M}_{(i,j)}^N(0,0)\Big|_{\Phi_C} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j)\Big|_{\Phi_C}$ when $P_j < P_{j,(i,j)}^{req}$ and $P_i < P_{i,(i,j)}^{req}$. Therefore, we only need to consider binary decision of completely serving both packets or not transmitting them at all for CTM based NOMA in the Hybrid MA scenario. We can determine optimal decision by selecting one of $\{0,0\}$ and $\left\{ P_{i,(i,j)}^{req}, P_{j,(i,j)}^{req} \right\}$ power allocation options for minimizing $\mathcal{M}_{(i,j)}^N(P_i, P_j)\Big|_{\Phi_C}$. The decision in (30)-(31) can be obtained under total power constraint, such that $P_i + P_j \leq P_{(i,j)}^{\max}$.

References

1. P. Popovski, f. Stefanovifá, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp.

5783–5801, 2019.

2. M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and non-terrestrial iot: A survey on 5g advances and the road toward 6g," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1117–1174, Secondquarter 2022.

3. I. Ud Din, M. Guizani, S. Hassan, B.-S. Kim, M. Khurram Khan, M. Atiquzzaman, and S. H. Ahmed, "The internet of things: A review of enabled technologies and future challenges," *IEEE Access*, vol. 7, pp. 7606–7640, 2019.

4. M. E. Tanab and W. Hamouda, "Efficient resource allocation in fast-uplink grant for machine-type communications with noma," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 18 113–18 129, 2022.

5. J. Yao, Q. Zhang, and J. Qin, "Joint decoding in downlink noma systems with finite blocklength transmissions for ultrareliable low-latency tasks," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 705–17 713, Sep. 2022.

6. S. Baek, D. Kim, M. Tesanovic, and A. Agiwal, "3GPP new radio release 16: Evolution of 5G for industrial internet of things," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 41–47, 2021.

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

8. K. T. Phan, P. Huynh, D. N. Nguyen, D. T. Ngo, Y. Hong, and T. Le-Ngoc, "Energy-efficient dual-hop internet of things communications network with delay-outage constraints," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4892–4903, 2021.

9. I.-H. Hou and P. Kumar, "Packets with deadlines: A framework for real-time wireless networks," *Synthesis Lectures on Communication Networks*, vol. 6, 05 2013.

10. S. Lashgari and A. S. Avestimehr, "Timely throughput of heterogeneous wireless networks: Fundamental limits and algorithms," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8414–8433, 2013.

11. G. Maciel Ferreira Silva and T. Abrão, "Multipower-level q-learning algorithm for random access in nonorthogonal multiple access massive machine-type communications systems," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 9, p. e4509, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.4509

12. U. Cisco, "Cisco annual internet report (2018–2023) white paper," *Cisco: San Jose, CA, USA*, 2020.

13. S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-s. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

14. M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.

15. 3GPP, "Study on downlink multiuser superposition transmission (must) for LTE," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 36.859, Jan 2016, version 13.0.0.

16. H. Chamkhia, A. Erbad, A. Al-Ali, A. Mohamed, A. Refaey, and M. Guizani, "PLS performance analysis of a Hybrid NOMA-OMA based IoT system with mobile sensors," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 1419–1424.

17. U. Ghafoor, M. Ali, H. Z. Khan, A. M. Siddiqui, and M. Naeem, "Efficient resource allocation for hybrid nonorthogonal multiple access based heterogeneous networks beyond fifth-generation," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 12, p. e4630, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.4630

18. Q. Wang, H. Chen, C. Zhao, Y. Li, P. Popovski, and B. Vucetic, "Optimizing information freshness via multiuser scheduling with adaptive noma/oma," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1766–1778, 2022.

19. H. Pan, J. Liang, S. C. Liew, V. C. M. Leung, and J. Li, "Timely information update with nonorthogonal multiple access," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4096–4106, 2021.

20. Q. Wang, H. Chen, Y. Li, and B. Vucetic, "Minimizing age of information via hybrid NOMA/OMA," in *IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1753–1758.

21. Q. Abbas, S. A. Hassan, H. K. Qureshi, K. Dev, and H. Jung, "A comprehensive survey on age of information in massive iot networks," *Computer Communications*, vol. 197, pp. 199–213, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366422004066

22. M. Choi, J. Kim, and J. Moon, "Dynamic power allocation and user scheduling for power-efficient and delay-constrained multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4846–4858, Oct 2019.

23. M. Neely, "Stochastic Network Optimization with Application to Communication and Queueing Systems,". Morgan & Claypool, 2010. [Online]. Available: https://ieeexplore.ieee.org/document/6813406

24. M. Zafer and E. Modiano, "Optimal rate control for delay-constrained data transmission over a wireless channel," *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 4020–4039, Sep. 2008.

25. L. Lei, L. You, Q. He, T. X. Vu, S. Chatzinotas, D. Yuan, and B. Ottersten, "Learning-assisted optimization for energy-efficient scheduling in deadline-aware noma systems," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 3, pp. 615–627, Sep. 2019.

26. E. Fountoulakis, N. Pappas, Q. Liao, A. Ephremides, and V. Angelakis, "Dynamic power control for packets with deadlines," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–6.

27. E. Fountoulakis, N. Pappas, and A. Ephremides, "Dynamic power control for time-critical networking with heterogeneous traffic," *ITU Journal on Future and Evolving Technologies*, vol. 2, no. 1, March 2021.

28. P. Popovski, F. Chiariotti, K. Huang, A. E. Kalør, M. Kountouris, N. Pappas, and B. Soret, "A perspective on time toward wireless 6g," *Proceedings of the IEEE*, vol. 110, no. 8, pp. 1116–1146, 2022.

29. J. Wang, X. Jia, Z. Chen, and X. Zhang, "Optimization on information freshness for multi-access users with energy harvesting cognitive radio networks," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 11, p. e4591, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.4591

30. L. Georgiadis, R. Guerin, and A. Parekh, "Optimal multiplexing on a single link: delay and buffer requirements," *IEEE Transactions on Information Theory*, vol. 43, no. 5, pp. 1518–1535, 1997.

31. Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

32. X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4550–4564, 2018.

33. M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2238–2252, 2017.

34. M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 174–180, 2019.

35. R. K. Jain, D.-M. W. Chiu, W. R. Hawe *et al.*, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, vol. 21, 1984.

AUTHOR BIOGRAPHY

Onur Berkay Gamgam. Onur Berkay Gamgam received the B.S. and M.S. degrees from Bilkent University, all in electrical and electronics engineering. He has been also been working as Design Engineer on high speed communication projects since 2011. He is currently pursuing a PhD degree in Bilkent University and his research interests are on the performance analysis of latency constrained communications using dynamic algorithms.

Ezhan Karasan. Prof. Ezhan Karaşan received the B.S. degree from Middle East Technical University, M.S. degree from Bilkent University and Ph.D. degree from Rutgers University, all in electrical engineering. He worked at Bell Laboratories for three years before joining the Department of Electrical and Electronics Engineering at Bilkent in 1998, where he is currently the Vice-Rector. He is the recipient of the TÜBİTAK Young Scientist Award and Mustafa Parlar Foundation Young Scientist Award. He has also received the Distinguished Teaching Award from Bilkent University as well as the IEEE Turkey Section Distinguished Service Award. Dr. Karaşan is a member of the Editorial Board of the Optical Switching and Networking journal.