Received: 25 December 2022 | Revised: 2 May 2023 | Accepted: 30 May 2023

DOI: 10.1002/ett.4814

RESEARCH ARTICLE

WILEY

Satisfying strict deadlines for cellular Internet of Things through hybrid multiple access

Onur Berkay Gamgam | Ezhan Karasan

Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey

Correspondence
Onur Berkay Gamgam, Electrical and Electronics Engineering, Bilkent University, Bilkent, Ankara 06800, Turkey.
Email: onur.gamgam@bilkent.edu.tr

Abstract

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 are dropped from the buffer. 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. Moreover, we propose a flexible packet dropping mechanism called Early Packet Dropping (EPD) to detect likely to become expired packets and drop them immediately. Numerical results show that Hybrid MA improves the timely throughput compared to conventional OMA by up to 46% and on average by more than 21%. With EPD, these timely throughput gains improve to 53% and 24.5%, respectively.

1 | INTRODUCTION

Emerging communication ecosystems that facilitate massive machine-to-machine and machine-to-human connectivity create the need for ultra-reliable and low-latency communications (URLLC) services. The cellular Internet of Things (IoT) is a framework for conceptualizing such massive connectivity while addressing fundamental challenges such as the ever-increasing number of interconnected devices, latency constraints, and high-throughput demands.[1-3] For instance, Machine Type Communication- (MTC) based cellular IoT applications focus on the delivery of small-sized packets between a significantly large number of devices with latency, reliability, and connectivity constraints.[4] Reliable communication for IoT with small-sized packets requires short blocklength transmission, which further increases the problem complexity.[5] Moreover, a recent standardization study[6] on 5G communications emphasizes the extremely stringent latency and reliability requirements for applications such as motion control, mobile automation, and electric power grids, with latency budgets of 0.5, 1, and 10 ms, respectively. Such challenging URLLC constraints in IoT applications motivate the research on latency-constrained communications under 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 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 is expected to increase to 29.3 billion in 2023.[12] The limitations of widely used OMA schemes introduce new challenges in terms of increasing the 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 using the same spectral resource.[14] Yet, the advantage of NOMA in terms of system capacity depends on the diversity of user channel conditions and the 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 Third Generation Partnership Project (3GPP).[15-17]

The focus of this study is to maximize the timely throughput using Hybrid MA for cellular IoT applications. Although the potential of Hybrid MA is widely studied in the scope of information freshness,[18-21] 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

State of the art on latency-constrained communication often optimizes average latency, operates under strict deadline constraints, aims to maintain information freshness, and optimizes packet block-length. Moreover, topics of NOMA, hybrid NOMA, and short packet communication are studied in the state of the art as key solutions for low-latency IoT applications for beyond-5G cellular networks.[2]

In the first type of study, the aim is to optimize average latency by minimizing the average queue length, based on Little's theorem.[7,8,22] In Reference 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 study aims to meet strict deadlines for the problem of latency-constrained communications. In Reference 24, the authors introduce *BT-Problem* for sending B bits of data within a T duration of a deadline while minimizing the power utilization, which is solved using a continuous time model. In Reference 25, timely data transmission under deadline constraints using NOMA is considered, and the high computational complexity of scheduling and resource allocation tasks is addressed with deep learning techniques. In References 26 and 27, Fountoulakis et al considered the packet drop rate minimization with a limited power budget. Fixed sized packet arrivals are served in a packet per slot manner with an OMA transmission scheme through a wireless medium modeled as a 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 the *drift-plus-penalty*[23] technique for the maximization of timely throughput under average power constraints. In Reference 28, heterogeneous latency-constrained communications under the finite blocklength (FBL) regime are addressed with hybrid NOMA in which multiple short and urgent packets are transmitted nonorthogonally along with a large nonurgent packet. The objective is to minimize the transmission energy while considering packet deadlines. In Reference 29, the time-critical nature of MTC for industrial automation applications is addressed with hybrid NOMA/Frequency Division Multiple Access (FDMA) under FBL regime. The objective is to maximize total revenue over the network under diverse requirements. In Reference 30, latency-critical communication is modeled as the reliability of the one-shot transmission, and a hybrid NOMA/TDMA-based solution is proposed. In Reference 31, NOMA is considered for uplink transmissions in FBL regime with delay constraints, which is modeled as transmitting a finite number of packets within a given number of slots with high probability. In Reference 25, strict deadlines are served using NOMA with a learning-assisted solution approach, but FBL regime is not considered.

The third type of study is concerned with satisfying information freshness, for which, Age of Information (AoI) is used as a general performance metric.[21] In Reference 32, AoI is considered for the increasing connectivity in massive MTC applications, which demand diverse latency requirements. In References 18 and 33, the potential of NOMA is investigated for information freshness along with increased system connectivity. In Reference 19, NOMA is considered for the task of timely information updates. In Reference 20, Hybrid MA is considered for AoI. In Reference 34, hybrid MA with NOMA and OMA under FBL is considered for optimizing AoI in cellular networks. The hybrid NOMA/OMA policy is based on Lyapunov stochastic optimization. In Reference 35, hybrid NOMA/OMA is considered to serve time-sensitive and throughput-sensitive devices in IoT networks under the FBL regime. The effect of AoI and throughput requirements on NOMA user grouping is studied. In Reference 36, NOMA is considered to minimize average AoI, and it is compared with the conventional OMA scheme. In Reference 20, hybrid NOMA/OMA is considered for minimizing the weighted sum of expected AoI. In Reference 19, NOMA is considered for timely information update to reduce AoI compared to conventional OMA techniques. Results show that although NOMA could increase packet error rate, it reduces AoI compared to OMA.

The fourth type of study is about minimizing packet block-length for reducing transmission latency. In Reference 37, two-user downlink NOMA is introduced into short packet communication for achieving low-latency. Effective throughput is increased subject to a FBL constraint for reducing latency. Transmission rate and power allocation for two-user NOMA under FBL are optimized. In Reference 38, latency is modeled as packet blocklength, and the transmission energy is minimized under strictly heterogeneous latency requirements using Hybrid NOMA/TDMA under FBL regime. Various interference mitigation schemes are studied due to the possible infeasibility of conventional successive interference cancellation techniques under heterogeneous receiver conditions.

The studies on latency-constrained communications with hybrid NOMA in FBL regime mainly consider delay metrics such as average latency,[22] AoI,[19,20,34-36] packet block-length,[37,38] and packet deadline.[28,31] Thus, there is a lack of study on maximizing the timely throughput for given strict deadlines using hybrid NOMA in the FBL regime. To the best of our knowledge, this is the first study focusing on this gap in the literature.

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 that allocates user power in real-time to satisfy time average power constraints while maximizing the timely throughput. We process packets in the first-in-first-out (FIFO) manner. Therefore, only the head of the queue packet is considered for transmission. Other packets in the queue wait until the packets ahead of them are either fully transmitted or dropped. Since packets are dropped due to deadline expiration,[9] detection of likely to be expired packets before the expiration of their deadline may further increase the timely throughput. With this motivation, we consider flexible packet dropping schemes to deal with this situation.

The main contributions of this study are as follows:

- This study contributes to the gap in the literature on timely throughput maximization with hybrid MA using OMA and NOMA.
- 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 modeling the wireless medium as a fading channel. Moreover, we consider short packet communication with 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 the remaining deadline diminishes in order 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 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.
- In order to further improve the performance of the proposed dynamic algorithm, we propose a flexible packet dropping scheme called Early Packet Dropping (EPD) for detecting likely to be expired packets and removing them from the system before their deadline expiration to eliminate unnecessary packet fragment transmissions.

Our key numerical results show that the Hybrid MA outperforms OMA-only based systems by increasing the timely throughput up to 46% and on the average by more than 21% while satisfying time average power constraints. Proposed flexible packet drop mechanism EPD further improves the timely throughput with Hybrid MA compared to OMA-only by up to 53% and on the average by more than 24.5%. In delay-constrained wired systems, Earliest Deadline First (EDF) is the optimal scheduling algorithm.[39] 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 nonearliest deadline packet of a user with a strong channel condition can be eligible for transmission in order to minimize the overall packet drop rate, instead of the earliest deadline packet of another user with a weak channel condition. RDPPA controls the trade-off between power allocations and Channel-Queue State Information (CQSI) in the system to minimize the 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 the optimal power assignment for Hybrid MA. Then, the proposed EPD method is presented. Finally, the numerical results are presented with an elaborate analysis of system parameters' effects. The abbreviations used in this paper are presented in Table 1 The list of symbol and notation is presented in Table 2.

TABLE 1 List of abbreviations.

Abbreviation	Definition
IoT	Internet of Things
URLLC	Ultra-reliable and low-latency communications
MA	Multiple access
OMA	Orthogonal multiple access
NOMA	Non-orthogonal multiple access
M2M	Machine-to-machine
MTC	Machine type communication
3GPP	Third generation partnership project
AoI	Age of information
FIFO	First-in first-out
FBL	Finite blocklength
RDPPA	Remaining deadline-based parametric prioritization approach
EPD	Early packet dropping
EDF	Earliest deadline first
CQSI	Channel queue state information
AP	Access point
BLER	Block error rate
SNR	Signal to noise ratio
RV	Random variable
i.i.d.	Independent and identically distributed
TM	Transmission mode
CTM	Complete transmission mode
FTM	Fragmented transmission mode
SIC	Successive interference cancellation
FI	Fairness index
s-OMA	Soft-OMA
H-MA	Hybrid-MA
h-OMA	Hard-OMA
p-OMA	P_{inf}-OMA

TABLE 2 List of Symbol and Notation.

Symbol	Description	Symbol	Description
$\mathbb{Z}$	The set of whole numbers.	$\mathbb{Z}_+$	The set of positive whole numbers.
$\mathcal{N}$	The set of users	N	Number of users
t	Slot	τ	Slot duration in seconds
$\mathcal{B}$	Transmission bandwidth	η	Signal-to-noise ratio (SNR)
$\mathcal{V}_i$	Channel dispersion of user i	$Q^{-1}(.)$	inverse Q-function.
P_0	Power budget for transmitter	N_0	Noise spectral density
$P_i(t)$	Power allocation for user i	$\mathbf{P}(t)$	Vector of $P_i(t), \forall i \in \mathcal{N}$
$\overline{P_i}$	Average power consumption of user i	$\overline{P}$	Overall average power consumption
γ_i	Average power constraint for user i	R	Radius of the circular coverage area
r_i	Distance between the transmitter and user i	h_i	Channel gain of user i
$m_i \in \mathbb{Z}_+$	Deadline for packets of user i	π_i	Packet arrival probability of user i
Λ	Set of available packet sizes in bits	$\lambda_i(t)$	Arrival rate for user i
$Q_i(t)$	Queue backlog of user i	$d_i(t)$	Number of slots left before expiration for user i
$q_i(t)$	Number of data bits left for packet of user i	$\mu_i(t)$	Departure rate for user i
$D_i(t)$	Number of dropped bits for user i	$E_i(t)$	Number of dropped packets by EPD for user i
$\overline{D_i}$	Packet drop rate for user i	$\overline{D}$	Overall packet drop rate
$\phi(t)$	Occupied TM for a packet	Φ_C	Complete TM
Φ_F	Fragmented TM	$\varphi(t)$	Queue reduction ratio
$\alpha_i(t)$	RDPPA parameter for user i	$\mathcal{F}_i(.)$	Proposed cost function
$\overline{\mathcal{F}_i}$	Time average of $\mathcal{F}_i$	$X_i(t)$	Virtual queue of user i
$\mathbf{X}(t)$	Vector of $X_i(t), \forall i \in \mathcal{N}$	$L(\mathbf{X}(t))$	Quadratic Lyapunov Function
$\Delta(\mathbf{X}(t))$	Conditional Lyapunov Drift	V	Weight parameter of *drift-plus-penalty*
$\mathcal{M}(\mathbf{P}(t))$	Objective function	ϵ_O	BLER for OMA
$\mathcal{P}_i$	OMA power region for user i	P_i^O	Optimal OMA power for user i
ϵ_N	BLER for NOMA	$\{i,j\}$	NOMA user pair, such that $\|h_i\|^2 \leq \|h_j\|^2$
$\mathcal{P}_{(i,j)}$	NOMA total power region for pair $\{i,j\}$	θ	Auxiliary variable representing total NOMA power
$P^N_{i,(i,j)}$	Optimal NOMA power of user i of pair $\{i,j\}$	$P^N_{j,(i,j)}$	Optimal NOMA power of user i of pair $\{i,j\}$
S	EPD slot count	$\hat{\mu}_i(S)$	EPD transmission rate estimate

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 the 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[40] proposed a framework for tightly approximating transmission rate $R^*(\eta)$ in the FBL regime for blocklength $\tau\mathcal{B}$ and block error rate (BLER) ϵ, as follows:

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

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. Note that, (1) holds for Additive White Gaussian Noise (AWGN) channels. We can apply this rate calculation for downlink NOMA transmission by properly deriving SNR for NOMA user pair considering Successive Interference Cancellation (SIC).[41] Assume that AP communicates with a user pair using NOMA. The channel gains of strong and weak users are denoted as h_s and h_w, such that $|h_s|^2 > |h_w|^2$. Moreover, the allocated power by AP for strong and weak users are denoted as P_s and P_w. The strong user removes interference due to the weak user's message using SIC, then decodes its own message. Thus, SNR of the strong user is $\eta_s^N = \frac{P_s|h_s|^2}{N_0}$, where N_0 is the noise spectral density. The weak user treats the signal of the strong user as part of the noise and decodes its own message. Thus, the SNR of the weak user is $\eta_w^N = \frac{P_w|h_w|^2}{P_s|h_w|^2+N_0}$. Then, NOMA transmission rate in FBL regime can be calculated for strong and weak users as $R^*(\eta_s^N)$ and $R^*(\eta_w^N)$ using (1), respectively.

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 a slot t. We consider continuous power levels such that $0 \le P_i(t) \le P_0$ 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 a time t, respectively. Let $\overline{P_i}$ be the average power utilization overall 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 $\overline{P_i} \le \gamma_i$. 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. The 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.[37,42]

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 a user i at a slot t. Let $u_i(t) \sim U[\Lambda]$ be the bit count of the arriving packet for the user i in the tth 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 tth slot is denoted as $\lambda_i(t) \triangleq a_i(t)\cdot u_i(t)$, in bits per slot. The queue backlog for a user i in the tth slot is denoted as $Q_i(t)$, where $Q_i(0) = 0$. The packets in a queue are processed in the FIFO 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 tth slot for the packet at the head of the queue i.

The departure process represents the transmitted number of bits in a slot. Let $\Psi_i(t) = j$ be the user paired with the 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 a user i in the tth 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 due to either EPD mechanism or deadline violation. At first, EPD mechanism detects packets to be dropped. Let $E_i(t)$ be the number of packets of user i dropped at slot t by EPD mechanism. $Q_i(t)$, $q_i(t)$ and $d_i(t)$ are updated according to $E_i(t)$. EPD mechanism is presented in Section 5. Secondly, the drop event occurs due to deadline violation for the packet at the head of the queue i when $d_i(t) = 1$ and $\mu_i(t) < q_i(t)$. The number of dropped bits for user i due to deadline violation 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 the number of dropped bits. $\overline{D_i}$ represents the packet drop rate due to deadline violation for a user i and $\overline{D}$ is the overall average drop rate.

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

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

$$\overline{D} \triangleq \frac{1}{N}\sum_{i=1}^{N}\left(\lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{1}\{D_i(\tau) > 0\} + E_i(\tau)\right), \qquad \forall i \in \mathcal{N}. \tag{5}$$

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:

$$\min_{\mathbf{P}(t)} \quad \sum_{i=1}^{N} \overline{D_i} \tag{6a}$$

$$\text{s.t.} \quad \overline{P_i} \leq \gamma_i, \forall i \in \mathcal{N} \tag{6b}$$

$$\mathbf{P}(t) \in \mathcal{P}^H(t). \tag{6c}$$

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 the 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 the deadline expiration. Let $\phi(t) \in \{\Phi_C, \Phi_F\}$ be the occupied TM at a time t, where Φ_C and Φ_F represent CTM and FTM, respectively.

We define a metric called queue reduction ratio denoted by $\varphi(t, \phi(t))$ to represent the effect of occupied TM on the head of the queue packet. In CTM, the target is to completely transmit the packet in a slot. Thus, the queue reduction ratio represents whether the head of the queue packet belonging to user i is completely transmitted in the slot with $\mu_i(t)$ or not by taking the values $\varphi(.) = 0$ or 1, respectively. In CTM, the target is to transmit a fragment of the packet in a slot. Thus, the queue reduction ratio represents the remaining ratio of the packet after transmission such that, $\varphi(.) = 0$ represents completely transmitting the packet. $\varphi(.) = 1$ represents no transmission, and transmission of a fragment is linearly represented between these two extremes, that is, $0 < \varphi(.) < 1$. $\varphi(t, \phi(t))$ can be expressed as follows:

$$\varphi(t, \phi(t)) \triangleq \begin{cases} \mathbb{1}\{q_i(t) - \mu_i(t) > 0\} & , \text{if } \phi(t) = \Phi_C \\ (q_i(t) - \mu_i(t))/q_i(t) & , \text{if } \phi(t) = \Phi_F \end{cases}. \tag{7}$$

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 (6a). In Reference 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 the expiration of the packet at the head of the queue. In this paper, we propose a novel function $\mathcal{F}_i(t, \alpha_i(t), \phi(t))$ as:

$$\mathcal{F}_i(t, \alpha_i(t), \phi(t)) \triangleq \left(\frac{m_i - (d_i(t) - 1)}{m_i} \right)^{\alpha_i(t)} \varphi(t, \phi(t)), \tag{8}$$

where $\alpha_i(t)$ is a nonnegative exponent parameter proposed to adjust the importance of the remaining deadline in the objective function for user i at slot t. The $f_i(t)$ In Reference 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 reduce. 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 toward 1 for $\alpha_i(t) < 1$ and $\alpha_i(t) > 1$ cases, respectively. Therefore, $\mathcal{F}_i$ provides information that helps us to predict the future consequences of our actions. Let $\overline{\mathcal{F}_i}$ be the time average of $\mathcal{F}_i$. We define the following new problem:

$$\min_{\mathbf{P}(t)} \quad \sum_{i=1}^{N} \overline{\mathcal{F}_i} \triangleq \sum_{i=1}^{N} \left(\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathcal{F}_i \right) \tag{9a}$$

$$\text{s.t.} \quad \overline{P_i} \le \gamma_i, \forall i \in \mathcal{N} \tag{9b}$$

$$\mathbf{P}(t) \in \mathcal{P}^H(t). \tag{9c}$$

The difference between the problem definitions for packet drop rate minimization in (6) and (9) is about how the cost is introduced to the objective function. In the problem definition (6), the cost is introduced to the objective function only when a packet is dropped. In the problem definition (9), the cost is introduced to the objective function according to $\mathcal{F}_i$, whose future values are determined according to the power allocation decision on the correct slot and the remaining deadline of the packet. When a packet is dropped, the introduced cost values are equal in (6) and (9). Otherwise, the introduced cost is zero in (6), whereas, the introduced cost with $\mathcal{F}_i$ in (9) is designed to quantify the consequences of our actions on the current slot. In order to solve (6), the slot-by-slot decisions need to be made while considering their effects in the infinite horizon. To solve (9), per-slot decisions can be made by the dynamic algorithm based on Lyapunov stochastic optimization.[23] Since the solution of the per-slot decisions based problem definition in (9) is simpler than that of the infinite horizon-based problem definition in (6), we focus on solving (9).

The problem definition presented in (9) is a minimization problem with constraints in the form of time averages, which can be solved using the *drift-plus-penalty* technique.[23] The time average constraints in (9b) are transformed into virtual-queues ($X_i(t), \forall i \in \mathcal{N}$) in (10), where arrivals are $P_i(t)$ and respective service rates are γ_i. The problem becomes a queue stability problem with a penalty metric. The strong stability of these virtual queues guarantees the respective time average constraints.

$$X_i(t+1) \triangleq \max[X_i(t) - \gamma_i, 0] + P_i(t). \tag{10}$$

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:

$$\Delta(\mathbf{X}(t)) \triangleq \mathrm{E}[L(\mathbf{X}(t+1)) - L(\mathbf{X}(t)) | \mathbf{X}(t)]. \tag{11}$$

At every time slot t, the problem in (9) is solved by determining $\mathbf{P}(t)$ in order to minimize the *drift-plus-penalty* expression,[23] $\Delta(\mathbf{X}(t)) + V\sum_{i=1}^{N} \mathrm{E}[\mathcal{F}_i | \mathbf{X}(t)]$, 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 CQSI and determine $\mathbf{P}(t)$ by solving the following *drift-plus-penalty* problem at each time slot t:

$$\min_{\mathbf{P}(t)} \mathcal{M}(\mathbf{P}(t)) = \sum_{i=1}^{N} (V\mathcal{F}_i + X_i(t)(P_i(t) - \gamma_i)) \tag{12a}$$

$$\text{s.t.} \quad \mathbf{P}(t) \in \mathcal{P}^H(t). \tag{12b}$$

The Slater condition indicates that all constraints can be satisfied with a slackness value $\delta > 0$ under a decision policy,[23,43] such that $limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} P_i(\tau) \le (\gamma_i - \delta), \ \forall i \in \mathcal{N}$. The Slater condition also ensures the strong stability of virtual queues.[23] The *drift-plus-penalty* algorithm gives an $O(\epsilon)$ approximation to optimality under the Slater condition.[23,43] Thus, the proper selection of the weight parameter V in order to satisfy time average constraints while minimizing the penalty is important for ensuring optimality with an $O(\epsilon)$ approximation, too.

Flow chart of the proposed dynamic power allocation algorithm is presented in Figure 1. The algorithm starts with the initialization of the local variables. In step 1, the algorithm waits for the beginning of the next slot t and then obtains new packet arrivals and observes the channel state. In step 2, EPD, which is presented in Section 5, is applied to drop packets that are detected as to-be-expired. In step 3, we calculate the optimal power assignments for candidate OMA and NOMA transmissions. Due to our single transmitter model, there are N possible OMA transmissions and $\binom{N}{2}$ possible NOMA transmissions. Thus, the scheduling complexity is $O(N)$ for only-OMA and $O(N^2)$ for only-NOMA and hybrid-MA

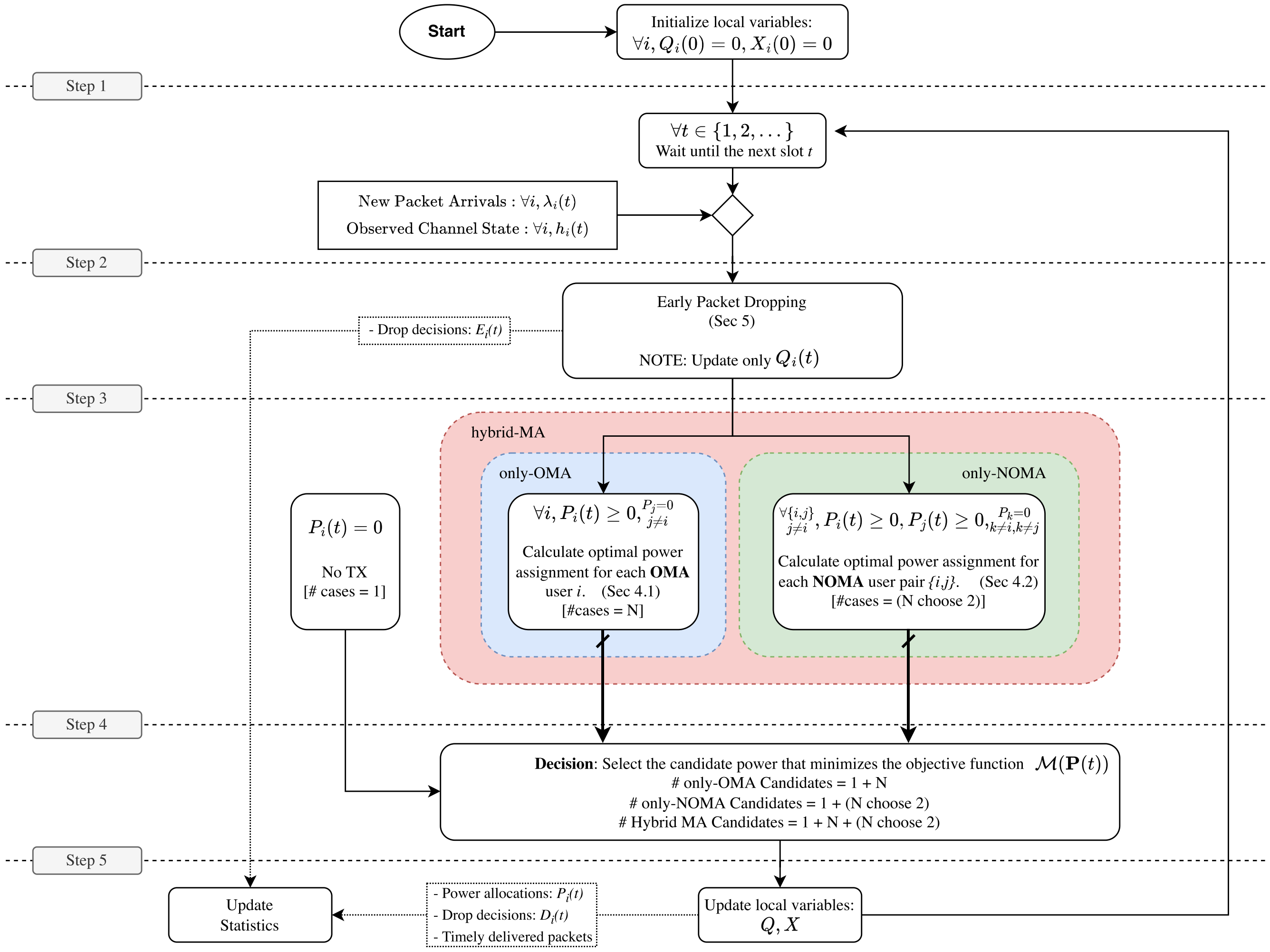
Start
Initialize local variables:
$\forall i, Q_i(0) = 0, X_i(0) = 0$
Step 1
$\forall t \in \{1, 2, \dots\}$
Wait until the next slot t
New Packet Arrivals : $\forall i, \lambda_i(t)$
Observed Channel State : $\forall i, h_i(t)$
Step 2
Early Packet Dropping
(Sec 5)
NOTE: Update only $Q_i(t)$
- Drop decisions: $E_i(t)$
Step 3
hybrid-MA
only-OMA
$\forall i, P_i(t) \geq 0, {}^{P_j=0}_{j \neq i}$
Calculate optimal power assignment for each **OMA** user i. (Sec 4.1)
[#cases = N]
only-NOMA
$\forall \{i,j\}_{j \neq i}, P_i(t) \geq 0, P_j(t) \geq 0, {}^{P_k=0}_{k \neq i, k \neq j}$
Calculate optimal power assignment for each **NOMA** user pair $\{i,j\}$. (Sec 4.2)
[#cases = (N choose 2)]
$P_i(t) = 0$
No TX
[# cases = 1]
Step 4
Decision: Select the candidate power that minimizes the objective function $\mathcal{M}(\mathbf{P}(t))$
# only-OMA Candidates = 1 + N
# only-NOMA Candidates = 1 + (N choose 2)
# Hybrid MA Candidates = 1 + N + (N choose 2)
Step 5
Update
Statistics
- Power allocations: $P_i(t)$
- Drop decisions: $D_i(t)$
- Timely delivered packets
Update local variables:
Q, X

FIGURE 1 Flow chart of the proposed algorithm.

cases. In step 4, the optimization metric in (12) is exhaustively minimized. One of the minimizing candidates is selected randomly for the sake of fairness among users. In each slot, hybrid-MA dynamically switches between OMA and NOMA by selecting the best candidate among OMA and NOMA transmissions with derived optimal power assignments. In step 5, local variables and statistics are updated according to the given decision. Finally, the algorithm goes to step 1 to wait for the beginning of the next slot.

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 Reference 40, departure rate $\mu_i(P_i)$† for a user i using OMA with FBL $\tau\mathcal{B}$ and BLER ϵ_O is given by:

$$\mu_i(P_i) = \tau\mathcal{B}\log_2\left(1 + \frac{|h_i|^2 P_i}{\mathcal{B}N_0}\right) - \sqrt{\tau\mathcal{B}\mathcal{V}_i} \cdot \frac{Q^{-1}(\epsilon_O)}{\ln 2}, \tag{13}$$

where N_0 is the noise power spectral density. Note that, we obtain a lower bound on $\mu_i(P_i)$ with the approximation of $\mathcal{V}_i \approx 1$, since $\mathcal{V}_i \leq 1$.[44] Thus, the lower bound can be used in the solution for satisfying strict deadlines. Moreover, for URLLC which requires high reliability and low latency, SNR should be sufficiently high,[44] and hence the approximation of $\mathcal{V}_i \approx 1$ is accurate in this operating region. 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 $\mu_i(P_i^{\text{req}}) = q_i$ and $\mu_i(P_i^{\text{one}}) = 1$. 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:

$$P_i^{\min} = P_i^{\text{req}} \mathbb{1}\{d_i = 1\} + P_i^{\text{one}} \mathbb{1}\{d_i > 1\} \tag{14}$$

Let $P_i^{\max} \triangleq \min(P_i^{\text{req}}, P_0)$ 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 $d_i = 1$. Thus, we avoid the waste of power for transmitting a fraction of a to-be-expired packet with OMA by constraining $\mathcal{P}_i$ properly. 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:

$$P_i^O = \arg\min_{P_i} \mathcal{M}_i^O(P_i) \triangleq \underset{P_j=0}{\mathcal{M}(\mathbf{P})}, \tag{15a}$$

$$\text{s.t.} \quad P_i \in \mathcal{P}_i, \ \forall j \neq i, \tag{15b}$$

where P_i^O is the optimal power value for user i under OMA transmission scheme. The solution to 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:*

$$P_i^O = \begin{cases} P_i' & \text{, if } \mathcal{M}_i^O(0)\Big|_{\Phi_F} > \mathcal{M}_i^O(P_i')\Big|_{\Phi_F} \\ 0 & \text{, otherwise} \end{cases}, \tag{16}$$

where

$$P_i' = \begin{cases} P_i^{\max} & \text{, if } P_i^{\max} \leq P_i^* \\ P_i^* & \text{, if } P_i^{\min} \leq P_i^* < P_i^{\max} \\ P_i^{\min} & \text{, if } P_i^* < P_i^{\min} \end{cases} \tag{17a}$$

$$P_i^* = \Gamma_i/(X_i \ln 2) - (\mathcal{B}N_0)/|h_i|^2 \tag{17b}$$

$$\Gamma_i = V(1 - (d_i - 1)/m_i)^{\alpha_i}(\tau\mathcal{B})/q_i. \tag{17c}$$

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.[45] Therefore, allocated power values are only constrained by with total power budget: $P_i + P_j \leq P_0$. Based on Reference 40, 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:

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

$$\mu_{j,(i,j)} = \tau B \log_2\left(1 + \frac{|h_j|^2 P_j}{BN_0}\right) - \sqrt{\tau\mathcal{B}\mathcal{V}_j}\frac{Q^{-1}(\epsilon_N)}{\ln 2}, \tag{19}$$

where the channel dispersion parameters are approximated as $\mathcal{V}_i \approx 1$ and $\mathcal{V}_j \approx 1$, which is valid in the high SNR regime. Similarly, we obtain a lower bound on $\mu_{i,(i,j)}$ and $\mu_{j,(i,j)}$ with these approximations.[44] The lower bounds can be used in the solution for satisfying strict deadlines with NOMA under FBL regime. Let $\{P^{\text{req}}_{j,(i,j)}, P^{\text{one}}_{j,(i,j)}\}$ and $\{P^{\text{req}}_{i,(i,j)}, P^{\text{one}}_{i,(i,j)}\}$ 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 $P^{\min}_{k,(i,j)}$, $k \in \{i, j\}$ be the minimum required power to successfully perform a NOMA transmission under deadline constraint, as follows:

$$P^{\min}_{k,(i,j)} = P^{\text{req}}_{k,(i,j)} \mathbb{1}\{d_k = 1\} + P^{\text{one}}_{k,(i,j)} \mathbb{1}\{d_k > 1\} \tag{20}$$

Thus, we have $\mu_{j,(i,j)}(P^{\text{req}}_{j,(i,j)}) = q_j$, $\mu_{j,(i,j)}(P^{\text{one}}_{j,(i,j)}) = 1$, $\mu_{i,(i,j)}(P^{\text{req}}_{i,(i,j)}, P^{\min}_{j,(i,j)}) = q_i$, and $\mu_{i,(i,j)}(P^{\text{one}}_{i,(i,j)}, P^{\min}_{j,(i,j)}) = 1$. Let $P^{\max}_{(i,j)} \triangleq \min(P_0, (P^{\text{req}}_{j,(i,j)} + P^{\text{req}}_{i,(i,j)}))$, $P^{\min}_{(i,j)} \triangleq (P^{\min}_{j,(i,j)} + P^{\min}_{i,(i,j)})$, and $\mathcal{P}_{(i,j)}$ be the available total power region for a NOMA transmission of user pair $\{i, j\}$. $\mathcal{P}_{(i,j)} = \{0\} \cup [P^{\min}_{(i,j)}, P^{\max}_{(i,j)}]$ when $P^{\min}_{(i,j)} \leq P^{\max}_{(i,j)}$. Otherwise, $\mathcal{P}_{(i,j)} = \{0\}$, since the respective packet will definitely be dropped when $d_i = 1$ or $d_j = 1$. Thus, we avoid waste of power for transmitting a fraction of a to-be-expired packet with NOMA by constraining $\mathcal{P}_{(i,j)}$ properly. Let $\mathcal{M}^N_{(i,j)}(P_i, P_j) = \mathcal{M}(\mathbf{P})$ 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, the optimization problem of power assignment for FTM based NOMA transmission of user pair $\{i, j\}$ can be expressed as:

$$\{P^N_{i,(i,j)}\Big|_{\Phi_F}, P^N_{j,(i,j)}\Big|_{\Phi_F}\} = \arg\min_{P_i, P_j} \mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F} \tag{21a}$$

$$s.t.\ P_i + P_j \in \mathcal{P}_{(i,j)}, \tag{21b}$$

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:

$$\begin{aligned}\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F} &= -\Gamma_i \log_2\left(1 + \frac{|h_i|^2 P_i}{|h_i|^2 P_j + \mathcal{B}N_0}\right) \\ &\quad - \Gamma_j \log_2\left(1 + \frac{|h_j|^2 P_j}{\mathcal{B}N_0}\right) + X_i P_i + X_j P_j + C^N_{(i,j)},\end{aligned} \tag{22}$$

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 (8) is nonnegative for $P_k \in [P^{\min}_{k,(i,j)}, P^{\text{req}}_{k,(i,j)}]$, $k \in \{i, j\}$. Since $\mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_F}$ in (22) is not convex, the optimization problem in (21) is not convex either. In order to solve this problem, an auxiliary variable $\theta = P_i + P_j$ is introduced and the solution process of the problem is divided into two consecutive subproblems. At first, the optimal value of θ is calculated. Then, the optimal value of P_j is calculated for given θ. The first subproblem is to find the optimal value of θ as follows:

$$\theta^N = \arg\min_{\theta} g(\theta, P_j) \tag{23a}$$

$$s.t.\ \theta \in [P^{\min}_{(i,j)}, P^{\max}_{(i,j)}], \tag{23b}$$

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

$$\begin{aligned}g(\theta, P_j) &= -\Gamma_i \log_2\left(\frac{|h_i|^2\theta + \mathcal{B}N_0}{|h_i|^2 P_j + \mathcal{B}N_0}\right) + X_i\theta \\ &\quad - \Gamma_j \log_2\left(1 + \frac{|h_j|^2 P_j}{\mathcal{B}N_0}\right) + \left(X_j - X_i\right) P_j + C^N_{(i,j)}.\end{aligned} \tag{24}$$

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

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

$$\theta^N = \begin{cases} P^{\max}_{(i,j)} & \text{, if } P^{\min}_{(i,j)} \leq P^{\max}_{(i,j)} \leq \theta^* \\ \theta^* & \text{, if } P^{\min}_{(i,j)} \leq \theta^* \leq P^{\max}_{(i,j)} \\ P^{\min}_{(i,j)} & \text{, if } \theta^* \leq P^{\min}_{(i,j)} \leq P^{\max}_{(i,j)} \end{cases}, \tag{25}$$

where

$$\theta^* = \frac{\Gamma_i}{X_i \cdot \ln 2} - \frac{\mathcal{B} \cdot N_0}{|h_i|^2}. \tag{26}$$

Proof. *Proof can be found in Appendix B.* ■

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

$$P'_{j,(i,j)}\Big|_{\Phi_F} = \arg\min_{P_j} g(\theta^N, P_j) \tag{27a}$$

$$s.t.\ P_i + P_j = \theta^N, \tag{27b}$$

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

Theorem 3. *Suppose that $\Gamma_j \geq \Gamma_i$. Then, the 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:*

$$P'_{j,(i,j)}\Big|_{\Phi_F} = \begin{cases} P_j^{\min} & \text{, if } \frac{dg(P_j)}{dP_j}\Big|_{P_j=P_j^{\min}} \geq 0 \\ P_j^{\dagger} & \text{, if } \frac{dg(P_j)}{dP_j}\Big|_{P_j=P_j^{\dagger}} \leq 0 \\ P_j^{\ddagger} & \text{, otherwise} \end{cases}, \tag{28}$$

where $P_j^{\dagger} = \min((\theta - P^{\min}_{i,(i,j)}), P^{req}_{j,(i,j)})$, $\frac{dg(P_j)}{dP_j}\Big|_{P_j=P_j^*} = 0$ *and* $P_j^{\ddagger} = \min(P_j^*, P_j^{\dagger})$.

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:

$$\begin{aligned} \mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_C} &= X_i P_i + X_j P_j + C^N_{(i,j)} \\ &+ \sum_{k \in \{i,j\}} \left(\frac{\Gamma_k q_k}{\tau \mathcal{B}} \right) \left(\mathbb{1}\left\{ P_k < P^{\text{req}}_{k,(i,j)} \right\} - 1 \right). \end{aligned} \tag{29}$$

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 $\{P^{\text{req}}_{i,(i,j)}, P^{\text{req}}_{j,(i,j)}\}$. 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 function in (22) 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 Reference 26. The optimization problem of power assignment for CTM-based NOMA transmission can be expressed as:

$$\arg\min_{P_i, P_j} \mathcal{M}^N_{(i,j)}(P_i, P_j)\Big|_{\Phi_C} \tag{30a}$$

$$s.t.\ P_i + P_j \in \mathcal{P}_{(i,j)}, \tag{30b}$$

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 (30) 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 $(P^{req}_{i,(i,j)} + P^{req}_{j,(i,j)} \leq P^{\max}_{(i,j)})$,*

$$\{P'_{i,(i,j)}\big|_{\Phi_C}, P'_{j,(i,j)}\big|_{\Phi_C}\} = \{P^{req}_{i,(i,j)}, P^{req}_{j,(i,j)}\}, \tag{31}$$

otherwise,

$$\{P'_{i,(i,j)}\big|_{\Phi_C}, P'_{j,(i,j)}\big|_{\Phi_C}\} = \{0, 0\} \tag{32}$$

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:

$$\{P^N_{i,(i,j)}, P^N_{j,(i,j)}\} = \begin{cases} \{P'_{i,(i,j)}, P'_{j,(i,j)}\} & \text{, if } \mathcal{M}^N(0,0) > \mathcal{M}^N(P'_{i,(i,j)}, P'_{j,(i,j)}) \\ \{0, 0\} & \text{, otherwise} \end{cases}, \tag{33}$$

where $\{P'_{i,(i,j)}, P'_{j,(i,j)}\}$ is as follows:

$$\{P'_{i,(i,j)}, P'_{j,(i,j)}\} = \begin{cases} \{(\theta^N - P'_{j,(i,j)}\big|_{\Phi_F}), P'_{j,(i,j)}\big|_{\Phi_F}\} & \text{, if}(\Gamma_j \geq \Gamma_i) \\ \{P'_{i,(i,j)}\big|_{\Phi_C}, P'_{j,(i,j)}\big|_{\Phi_C}\} & \text{, otherwise.} \end{cases} \tag{34}$$

In (34), we have $\theta^N \in [P^{\min}_{(i,j)}, P^{\max}_{(i,j)}]$, as presented in Theorem 3. Therefore, the resultant $\{P'_{i,(i,j)}, P'_{j,(i,j)}\}$ in (34) cannot be directly used as the final power allocation decision. In (33), we compare $\mathcal{M}^N(P'_{i,(i,j)}, P'_{j,(i,j)})$ with the $\theta^N = 0$ case, which is denoted as $\mathcal{M}^N(0,0)$. Thus, we fully cover $\mathcal{P}_{(i,j)}$ in (33) to make our final decision for $\{P^N_{i,(i,j)}, P^N_{j,(i,j)}\}$, which is the power allocation decision under NOMA transmission scheme.

5 | EARLY PACKET DROPPING MECHANISM

The purpose of the EPD algorithm is to detect the packets that are impossible to be transmitted completely by the end of their deadlines in the best-case scenario. We define the best-case scenario for a packet as dedicating the transmitter with a full power budget of P_0 only for the OMA transmission of the packet throughout its remaining slots until expiration. We consider the full power budget of P_0 to achieve the maximum transmission rate for the best-case scenario. Since the user channel gain alters in each slot, we need to estimate the OMA transmission rate for future slots. Then, we derive the number of slots required to transmit the packet completely using the estimated OMA transmission rate. We consider a packet as likely to become expired if the derived slot count is greater than the remaining deadline of the packet. EPD aims to detect packets that are likely to become expired due to deadline violation in the best-case scenario and drops them immediately.

We estimate OMA transmission rate of the future slots by averaging the actual OMA transmission rates regarding a number of most recently observed user channel gain values. Let $S \in \mathbb{Z}$ be the number of slots over which the EPD mechanism estimates the OMA transmission rate. $S \geq 0$ indicates the set of user channel states in the last $S + 1$ slots, including the current slot. Thus, $S = 0$ indicates only the currently observed user channel state. Finally, $S < 0$ indicates the case where EPD mechanism is inactivated. In this case, we set the estimated OMA transmission rate as ∞, so that, EPD does not drop any packet, and we have $E_i(t) = 0, \forall i, t$ for $S < 0$ case. Let $\hat{\mu}_i(S)$ be the estimated OMA transmission rate for user i which is expressed as follows;

Algorithm 1. Pseudocode for epd mechanism

```
Inputs: $Q_i(t), \forall i \in \mathcal{N}, S$
Outputs: $Q_i(t), E_i(t) \forall i \in \mathcal{N}$
Local variables: $\{q', d'\}$
Operation:
for each user $i$ do
    Initialize $E_i(t) = 0$
    Calculate $\hat{\mu}_i(S)$
    for each packet in $Q_i(t)$ starting from head of the queue do
        $q' \leftarrow$ Bit size of the head of the queue packet from $Q_i(t)$.
        $d' \leftarrow$ Remaining deadline of the head of queue packet from $Q_i(t)$.
        if $\lceil \frac{q'}{\hat{\mu}_i(S)} \rceil > d'$ then
            Drop the head of the queue packet.
            $Q_i(t) \leftarrow Q_i(t) - q'$
            Move on to the next packet.
        else
            Do not drop the packet. Move on to the next user.
            break
        end if
    end for
end for
return $Q_i(t), E_i(t), \forall i \in \mathcal{N}$
```

$$\hat{\mu}_i(S) = \begin{cases} \frac{1}{S+1} \sum_{s=0}^{S} \mu_i(P_0) \Big|_{h_i(t-s)} & , S \geq 0 \\ \infty & , S < 0. \end{cases} \tag{35}$$

where $\mu_i(.)$ is the OMA transmission rate in the FBL regime expressed in (13). The pseudo-code for the EPD mechanism is presented in Algorithm 1.

6 | 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 SIC , NOMA related target BLER is set as $\epsilon_N = 5 \cdot 10^{-6}$, so that overall system target BLER is ensured.[46] A cellular model which is considered by Choi et al[22] in a study for satisfying latency constraints under energy efficiency objective in IoT networks with Hybrid MA, is used in this paper. A channel model for IoT applications in urban areas[22,47] is used in this study. The default values of all parameters used in obtaining the numerical results are given in Table 3, 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)[48] 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, the 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 toward π. Therefore, in order to clearly analyze the impact of other parameters on the system, we select $N = 5$ in Table 3. The problem becomes intractable for low values of m due to the 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 3. The algorithms proposed in this paper are as follows. FTM-based OMA is indicated as soft, others as hard.

- soft-OMA (s-OMA): Optimal power assignment for OMA with FTM is considered and Theorem 1 is used. This algorithm performs RDPPA.

TABLE 3 Simulation parameters and default values.[22,46,47]

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	$35.3 + 37.6 \log_{10}(r_i)$ dB
Fast fading component	$CN(0, 1)$
N_0 (Noise power spectral density)	−174 dBm/Hz
ϵ_O (OMA target BLER)	10^{-5}
ϵ_N (NOMA target BLER)	$5 \cdot 10^{-6}$

- Hybrid-MA (H-MA): Optimal power assignment for Hybrid MA is considered. The content of "Hybrid" consists of the NOMA and OMA using (33) and Theorem 1, respectively. This algorithm performs RDPPA.
- NOMA: Optimal power assignment for only NOMA is considered, using (33).

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

- hard-OMA[26] (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 Reference 26 is used.
- EDF[39]: It uniformly selects one of the users with the 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 an 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 order to analyze the potential of EPD mechanism, firstly the results without EPD are presented in Section 6.1. In this case, EPD mechanism in step 2 of the flow chart of the algorithm given in Figure 1 is ignored by setting $S < 0$, thus $E_i(t) = 0, \forall i, t$ in this case. The system parameters α and V are optimized in this setting. Secondly, EPD will be applied along with the optimized system parameters α and V. The purpose is to present the benefit provided by EPD for improving the timely throughput. The results with the EPD mechanism are presented in Section 6.2.

6.1 | Dynamic power allocation without EPD

In Figures 2,3, and 4, $\overline{D}$ 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 Figures 2,3, and 4 show the consistent relation between α and $\overline{D}$ under different traffic levels. Moreover, $\overline{D}$ converges at around $V = 100$ and increasing V toward 1000 does not make much difference on $\overline{D}$. Therefore, $V = 100$ is considered as the most suitable choice in terms of the tradeoff between $\overline{D}$ and γ. In Figure 2, p-OMA can fully

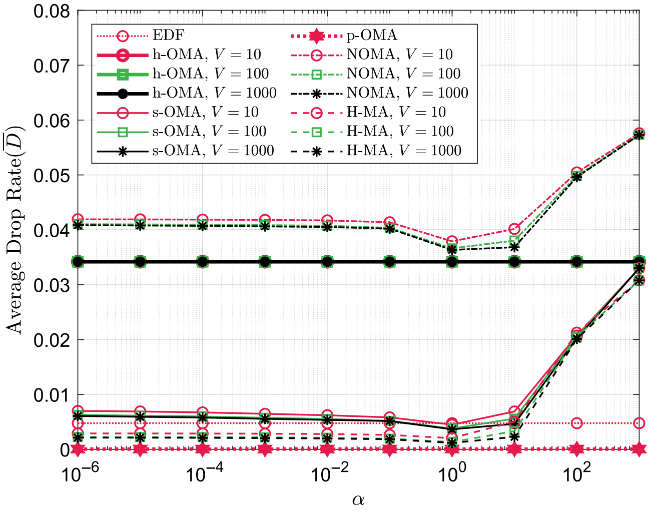

FIGURE 2 Average drop rate as a function of α with $\pi = 0.1$.

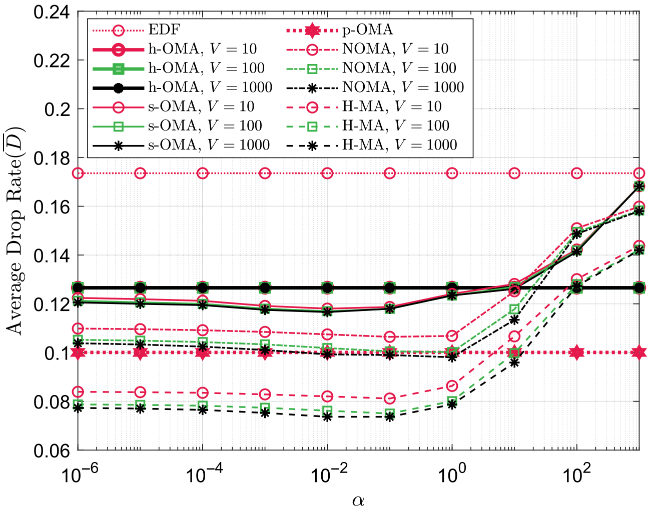

FIGURE 3 Average drop rate as a function of α with $\pi = 0.3$.

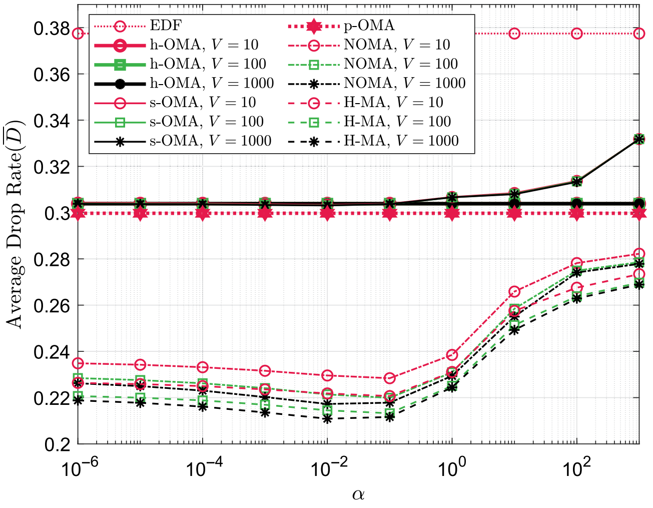

FIGURE 4 Average drop rate as a function of α with $\pi = 0.5$.

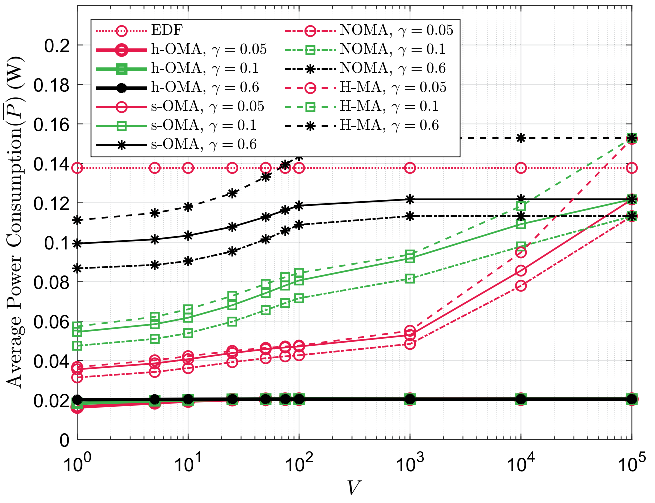

FIGURE 5 Average power consumption as a function of γ with $\pi = 0.1$.

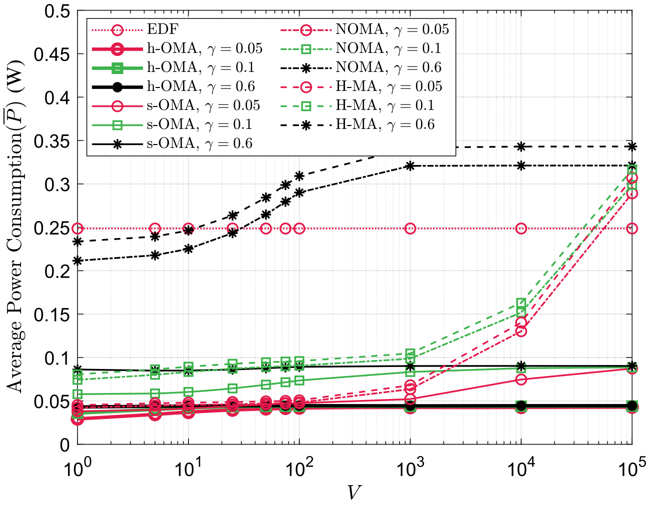

FIGURE 6 Average power consumption as a function of γ with $\pi = 0.3$.

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 a 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 Figure 3 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 Figure 4 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 Figures 5,6, and 7, $\overline{P}$ with different values of γ and $\pi \in \{0.1, 0.3, 0.5\}$ 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

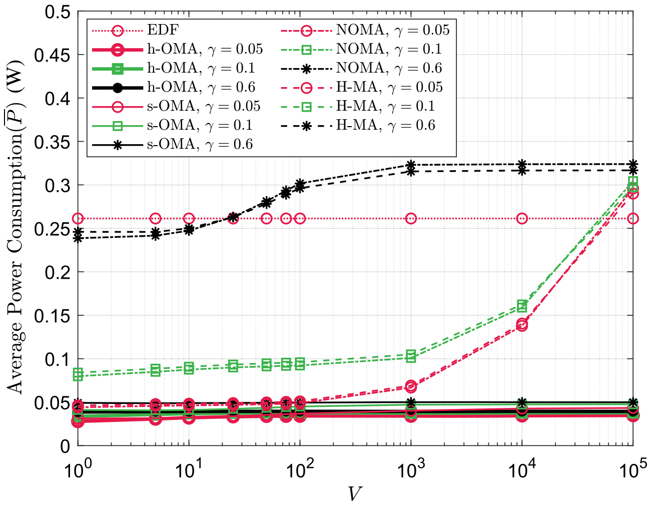

FIGURE 7 Average power consumption as a function of γ with $\pi = 0.5$.

TABLE 4 In hybrid-MA and Soft-OMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

	π		
γ	0.1	0.3	0.5
0.05	3.63%	15.92%	36.42%
0.1	3.50%	18.59%	40.71%
0.6	3.47%	23.38%	45.79%

TABLE 5 In hybrid-MA and NOMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

	π		
γ	0.1	0.3	0.5
0.05	66.72%	13.03%	4.25%
0.1	65.42%	13.04%	3.76%
0.6	64.42%	12.71%	2.40%

power utilization constraint and NOMA consumes slightly less power. Significantly high values of V force the algorithms to ignore the provided average power constraint in Figures 5,6, and 7. For all the arrival rates $\pi \in \{0.1, 0.3, 0.5\}$, H-MA, NOMA and s-OMA start to violate the provided average power constraint as V increases from 100, which is observed also for $\pi = 0.3$ in Figure 6. Based on these observations, it can be concluded that $V = 100$ is suitable for balancing the tradeoff between $\overline{D}$ and $\overline{P}$.

Another important role of the V parameter is about the optimality of the *drift-plus-penalty* algorithm. As presented in Section 3, the *drift-plus-penalty* algorithm closely approximates optimality under the Slater condition,[23,43] which indicates the satisfaction of all time average constraints. Therefore, proper selection of V is also important for balancing the tradeoff between optimality and timely throughput maximization. In Figures 5,6, and 7, $\overline{P}$ satisfies the average power constraints γ with $V = 100$ and diverse traffic levels. Thus, we can conclude that $V = 100$ is suitable for closely achieving the highest timely throughput performance while satisfying optimality.

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 4 and 5, respectively. In Table 4, the increase in 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

TABLE 6 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	0.00481	0.00378	0.00312
s-OMA	0.89435	0.90751	0.92444	0.00496	0.00374	0.00238
NOMA	0.87862	0.88264	0.90010	0.00612	0.00612	0.00579
H-MA	0.85498	0.86147	0.88792	0.01171	0.01183	0.01087

TABLE 7 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	0.01310	0.01311	0.01311
s-OMA	0.71412	0.72200	0.86293	0.01528	0.01556	0.01543
NOMA	0.92265	0.92813	0.93506	0.00178	0.00165	0.00154
H-MA	0.77739	0.79863	0.82947	0.01901	0.01907	0.01803

TABLE 8 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	0.00054	0.00026	0.00012
s-OMA	0.97026	0.97949	0.98592	0.00054	0.00027	0.00012
NOMA	0.91289	0.91341	0.92093	0.0049273	0.00499	0.00478
H-MA	0.90627	0.90890	0.92179	0.00636	0.00626	0.00569

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, the increase in 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 6, the 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 packet dropping rate compared to NOMA-based H-MA and NOMA algorithms, since their average *FI* is higher and the variance of *FI* is lower. The same relation is observed also for $\pi = 0.1$ and $\pi = 0.5$ traffic levels in Tables 7 and 8, respectively. Thus, we can conclude that, NOMA-based H-MA significantly increases overall performance at the cost of a slight decrease in *FI*.

In Figure 8, the average incomplete bit-rates ($\overline{I}$) with different values of N under arrival rate $\pi = 0.3$ are plotted. $\overline{I}$ 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. $\overline{I}$ 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 values of 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 the baseline algorithms, they increase $\overline{I}$ due to the dynamic nature of the opportunistic scheduling.

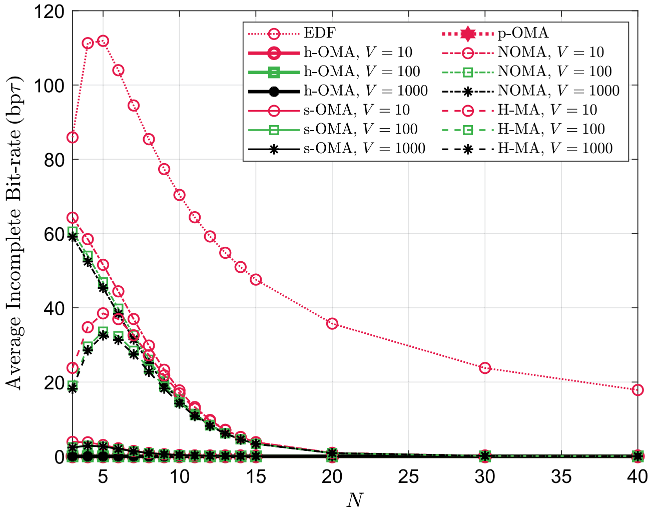

FIGURE 8 Average incomplete bit-rate as a function of N with $\pi = 0.3$.

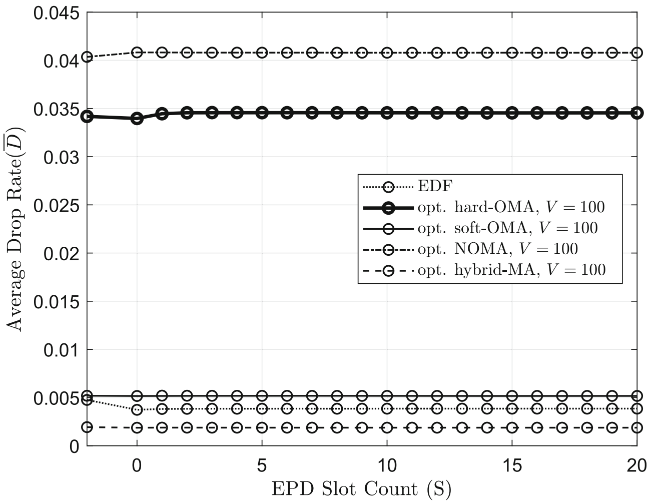

FIGURE 9 Average drop rate with $\pi = 0.1$ as a function of S in early packet dropping mechanism.

6.2 | Dynamic power allocation with EPD

In this section, the EPD mechanism is introduced to the system with $\alpha = 0.1$ and $V = 100$, which are the values optimized in Section 6.1. In Figures 9–11, $\overline{D}$ results are plotted as a function of EPD slot count S under $\pi \in \{0.1, 0.3, 0.5\}$, respectively. The set of available EPD slot count values is selected as $S \in \{-2, 0, 1, \ldots, 20\}$. Note that, $S = -2$ refers to no EPD mechanism and $S \geq 0$ indicates an active EPD mechanism as presented in Section 5. The first thing to notice in Figure 9 is the difference in $\overline{D}$ between $S = -2$ and $S \geq 0$ cases. As EPD slot count increases from $S = -2$ toward $S = 0$, $\overline{D}$ increases for NOMA, whereas it either decreases or does not change for the other algorithms. The reason for that is, the wrong decisions made by the EPD mechanism at this low traffic level affect the overall timely throughput. Moreover, since NOMA can operate with exactly two users, the wrong drop decisions made by EPD may block NOMA due to a lack of packets suitable for NOMA pairing. The second thing to notice is the algorithms' sensitivity with respect to S for $S \geq 0$. NOMA, s-OMA and H-MA are not sensitive to S in this low traffic level. The only algorithms that apparently perform the best $\overline{D}$ for $S = 0$ are EDF and h-OMA, which are algorithms based on CTM. The reason is, CTM- and OMA-based algorithms can compensate for the wrong decisions of EPD by aiming to transmit a packet fully in a slot. Therefore, the probability that EPD wrongly detects an incomplete packet as likely to-be-expired reduces. These observations indicate

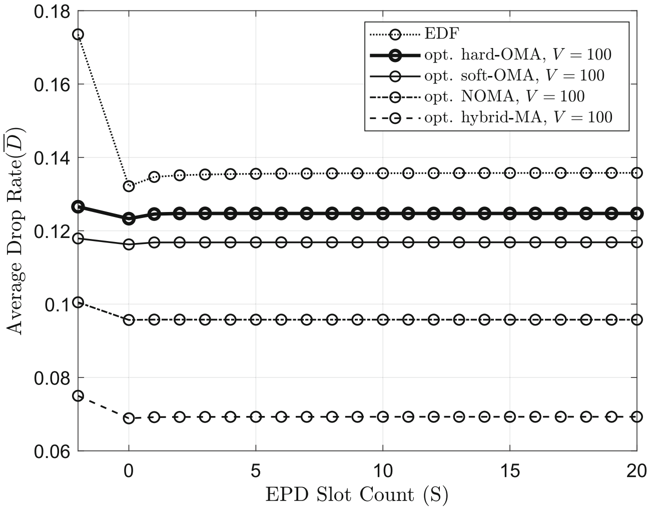

FIGURE 10 Average drop rate with $\pi = 0.3$ as a function of S in early packet dropping mechanism.

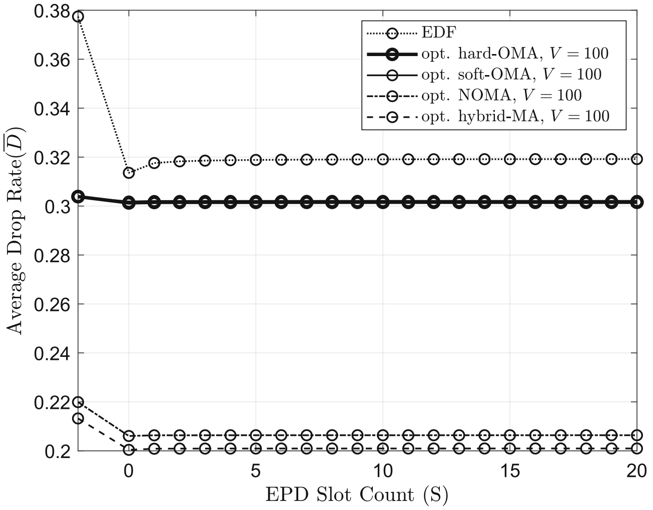

FIGURE 11 Average drop rate with $\pi = 0.5$ as a function of S in early packet dropping mechanism.

that EPD is not effective at low traffic levels, especially for the algorithms utilizing FTM and NOMA. Meanwhile, CTM- and OMA-based algorithms perform slightly better with EPD even at low traffic levels.

As the traffic level increases toward $\pi = 0.3$ and $\pi = 0.5$ as shown in Figures 10 and 11, timely throughput performances of all the algorithms increase with S. The reason is, high traffic levels increase the number of user packets left in the queues with the EPD mechanism, so that the probability of an algorithm finding a proper packet to serve on a slot increases. In this case, the timely throughput increases with the more effective utilization of the available network resources. Secondly, we observe that the algorithms' sensitivity with respect to S increases at high traffic levels. Interestingly, the best $\overline{D}$ performance for most of the algorithms, including H-MA, is observed with $S = 0$ at $\pi = 0.3$ and $\pi = 0.5$ traffic levels. Note that, $S = 0$ refers to considering only the channel gain observed in the current slot. Thus, there is no averaging over a range of slots. In the $S = 0$ case, the variance of the $\hat{\mu}_i(S)$ over the slots is expected to be at the highest among all $S \geq 0$ cases. For instance, estimated OMA transmission rate $\hat{\mu}_i(S)$ might result in a small value due to a deep fading encountered by user i. In such a case, the respective packets of user i might be wrongly detected as likely to-be-expired and be dropped by the EPD mechanism. Although the channel gain of user i might increase in the next slot, there will be no chance for the dropped packets. Therefore, EPD might decide to drop a packet just because of a very low channel gain observed in a slot. In the low traffic scenario of $\pi = 0.1$, we observed that EPD reduces the timely throughput performance in such a situation, especially for algorithms utilizing FTM and NOMA. However, the results at high traffic levels of $\pi = 0.3$ and $\pi = 0.5$ show that EPD improves the timely throughput by increasing the chance of transmission for

TABLE 9 In hybrid-MA and s-OMA comparison, percentage of increase in timely throughput for different π values and early packet dropping (EPD) utilization cases for under $\gamma = 0.6$, $\alpha = 0.1$ and $V = 100$.

	π		
γ	**0.1**	**0.3**	**0.5**
0.05	3.94%	19.50%	42.04%
0.1	3.74%	22.25%	46.63%
0.6	3.63%	27.21%	52.78%

Note: In all cases, s-OMA is operated without EPD.

TABLE 10 In hybrid-MA and NOMA comparison, percentage of increase in timely throughput for different π values and early packet dropping (EPD) utilization cases under $\gamma = 0.6$, $\alpha = 0.1$ and $V = 100$.

	π		
γ	**0.1**	**0.3**	**0.5**
0.05	67.22%	16.52%	8.54%
0.1	65.80%	16.53%	8.13%
0.6	64.68%	16.21%	7.31%

Note: In all cases, NOMA is operated without EPD.

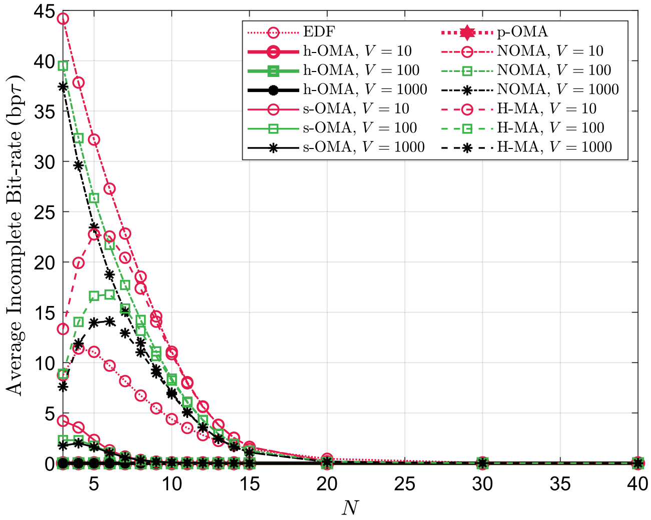

FIGURE 12 Average incomplete bit-rate as a function of N with $\pi = 0.3$ and early packet dropping utilization.

packets waiting in the queue. This happens due to early dropping of packets that are likely to become expired. Thus, we can conclude that EPD further improves the timely throughput as the traffic level increases.

The percentage of increase in timely throughput achieved by H-MA with EPD compared to s-OMA and NOMA without EPD for different π and γ values under the optimized $\alpha = 0.1$ and $V = 100$ values are presented in Tables 9 and 10, respectively. In Table 9, H-MA with EPD increases timely throughput up to 53% compared to s-OMA at the high traffic level of $\pi = 0.5$. Moreover, H-MA with EPD achieves improvement in timely throughput compared to s-OMA on the average by nearly 24.5% while satisfying average power constraints for all arrival rates. However, the effect of EPD on the improvement achieved with H-MA compared to s-OMA depends on the arriving traffic level: while there is no improvement at low traffic rates, the timely throughput improves as the traffic rate increases. The improvement in the timely throughput with EPD when H-MA is compared with NOMA is clear at all traffic rates. Finally, improvement achieved in timely throughput at various traffic levels using H-MA with EPD compared to s-OMA and NOMA in Tables 5 and 10 show that H-MA with EPD is robust under varying traffic loads.

$\bar{I}$ under traffic level of $\pi = 0.3$ with EPD is presented in Figure 12. The only parametric difference between Figures 8 and 12 is the utilization of the EPD mechanism. These two figures reveal that EPD significantly reduces $\bar{I}$. Thus, we can conclude that EPD reduces the wasted network resources for incomplete packets. The proposed dynamic power allocation algorithm leverages the achieved reduction in the wasted power to further improve the timely throughput. These results indicate that flexible packet dropping approaches such as EPD have a significant potential for improving the timely throughput.

7 | CONCLUSION

In this study, we address the problem of latency-constrained communications with strict deadlines under average power constraint using OMA- and NOMA-based Hybrid MA. We developed a dynamic algorithm which assigns user power in real-time to minimize the packet drop rate under average power constraints. Numerical results indicate that Hybrid MA increases the timely throughput compared to the OMA-only case by up to 46% and on average by more than 21% while satisfying average power constraints. The increase in the timely throughput with Hybrid MA is further improved up to 53% and on average by 24.5% using the proposed flexible packet dropping mechanism EPD. 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 channel states. EPD is a simple but effective example for illustrating the potential of flexible packet dropping mechanisms on the timely throughput. In order to further increase the timely throughput performances of the proposed s-OMA and H-MA, advanced flexible packet dropping mechanisms to reduce incomplete transmissions can be studied as a future work. Moreover, the optimal power allocation with exact channel dispersion $\mathcal{V}(\eta)$ in the FBL regime is subject to further research.

DATA AVAILABILITY STATEMENT

We did not use “data” in our study.

ENDNOTES

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

[†] 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.

ORCID

Onur Berkay Gamgam https://orcid.org/0000-0003-4660-9252

REFERENCES

1. Petar P, Čedomir S, Nielsen Jimmy J, et al. Wireless access in ultra-reliable low-latency communication (URLLC). *IEEE Trans Commun*. 2019;67(8):5783-5801.
2. Mojtaba V, Amin A, Khosravirad Saeed R, et al. Cellular, wide-area, and non-terrestrial IoT: a survey on 5G advances and the road toward 6G. *IEEE Commun Surv Tutor*. 2022;24(2):1117-1174.
3. Din Ikram U, Mohsen G, Suhaidi H, et al. The internet of things: a review of enabled technologies and future challenges. *IEEE Access*. 2018;7:7606-7640.
4. Manal ET, Walaa H. Efficient resource allocation in fast-uplink grant for machine-type communications with NOMA. *IEEE Internet Things J*. 2022;9(18):18113-18129.
5. Junteng Y, Qi Z, Jiayin Q. Joint decoding in downlink NOMA systems with finite blocklength transmissions for ultrareliable low-latency tasks. *IEEE Internet Things J*. 2022;9(18):17705-17713.
6. Sangkyu B, Donggun K, Milos T, Anil A. 3GPP new radio release 16: evolution of 5G for industrial internet of things. *IEEE Commun Mag*. 2021;59(1):41-47.
7. Changwei Z, Xinghua S, Jun Z, Xianbin W, Shi J, Hongbo Z. Throughput optimization with delay guarantee for massive random access of M2M communications in industrial IoT. *IEEE Internet Things J*. 2019;6(6):10077-10092.
8. Phan Khoa T, Phat H, Nguyen Diep N, Ngo Duy T, Yi H, Tho L-N. Energy-efficient dual-hop internet of things communications network with delay-outage constraints. *IEEE Trans Ind Inform*. 2020;17(7):4892-4903.
9. Hou I-H, Kumar PR. Packets with deadlines: a framework for real-time wireless networks. *Syn Lect Commun Netw*. 2013;6(1):1-116.
10. Sina L, Salman AA. Timely throughput of heterogeneous wireless networks: fundamental limits and algorithms. *IEEE Trans Inform Theory*. 2013;59(12):8414-8433.
11. Giovanni MFS, Taufik A. Multipower-level Q-learning algorithm for random access in nonorthogonal multiple access massive machine-type communications systems. *Trans Emerg Telecommun Technol*. 2022;33(9):e4509.
12. Cisco. *Cisco Annual Internet Report (2018–2023) White Paper*. *Cisco*; 2020;10(1):1-35.

13. Riazul ISM, Nurilla A, Dobre Octavia A, Kyung-Sup K. Power-domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges. *IEEE Commun Surv Tutor*. 2016;19(2):721-742.
14. Mahyar S, Mischa D, Johnson SJ. Massive non-orthogonal multiple access for cellular IoT: potentials and limitations. *IEEE Commun Mag*. 2017;55(9):55-61.
15. 3GPP. Study on downlink multiuser superposition transmission (MUST) for LTE. Version 13.0.0. 2016.
16. Chamkhia H, Erbad A, Al-Ali A, Mohamed A, Refaey A, Guizani M. PLS performance analysis of a hybrid NOMA-OMA based IoT system with mobile sensors. Paper presented at: IEEE Wireless Communications and Networking Conference (WCNC). Austin, TX; 2022:1419-1424.
17. Umar G, Mudassar A, Zubair KH, Masood SA, Muhammad N. Efficient resource allocation for hybrid nonorthogonal multiple access based heterogeneous networks beyond fifth-generation. *Trans Emerg Telecommun Technol*. 2022;33(12):e4630.
18. Qian W, He C, Changhong Z, Yonghui L, Petar P, Branka V. Optimizing information freshness via multiuser scheduling with adaptive NOMA/OMA. *IEEE Trans Wirel Commun*. 2021;21(3):1766-1778.
19. Haoyuan P, Jiaxin L, Chang LS, Leung Victor CM, Jianqiang L. Timely information update with nonorthogonal multiple access. *IEEE Trans Ind Inform*. 2020;17(6):4096-4106.
20. Wang Q, Chen H, Li Y, Vucetic B. Minimizing age of information via hybrid NOMA/OMA. Paper presented at: IEEE International Symposium on Information Theory (ISIT). Los Angeles, CA; 2020:1753-1758.
21. Qamar A, Ali HS, Khaliq QH, Kapal D, Haejoon J. A comprehensive survey on age of information in massive IoT networks. *Comput Commun*. 2022;197:199-213.
22. Minseok C, Joongheon K, Jaekyun M. Dynamic power allocation and user scheduling for power-efficient and delay-constrained multiple access networks. *IEEE Trans Wirel Commun*. 2019;18(10):4846-4858.
23. Neely MJ. Stochastic network optimization with application to communication and queueing systems. *Syn Lect Commun Netw*. 2010;3(1):1-211.
24. Murtaza Z, Eytan M. Optimal rate control for delay-constrained data transmission over a wireless channel. *IEEE Trans Inform Theory*. 2008;54(9):4020-4039.
25. Lei L, Lei Y, Qing H, et al. Learning-assisted optimization for energy-efficient scheduling in deadline-aware NOMA systems. *IEEE Trans Green Commun Netw*. 2019;3(3):615-627.
26. Fountoulakis E, Pappas N, Liao Q, Ephremides A, Angelakis V. Dynamic power control for packets with deadlines. Paper presented at: IEEE Global Communications Conference (GLOBECOM). Abu Dhabi, UAE; 2018:1-6.
27. Fountoulakis E, Pappas N, Ephremides A. Dynamic power control for time-critical networking with heterogeneous traffic. *arXiv preprint arXiv:2011.04448*. 2020.
28. Yanqing X, Chao S, Donghong C, Gang Z. Latency constrained non-orthogonal packets scheduling with finite blocklength codes. *IEEE Trans Vehic Technol*. 2020;69(10):12312-12316.
29. Lyu L, Chen C, Cheng N, Guan X, Shen X. NOMA-assisted small-packet transmissions in Mission-critical MTCs for industrial automation. Paper presented at: IEEE Global Communications Conference (GLOBECOM). Abu Dhabi, UAE; 2018:1-6.
30. Yao Z, Yuan Xiaopeng H, Yulin WT, Cenk GM, Anke S. Low-latency hybrid NOMA-TDMA: QoS-driven design framework. *IEEE Trans Wirel Commun*. 2022;22:1. doi:10.1109/TWC.2022.3215450
31. Jinho C. Opportunistic NOMA for uplink short-message delivery with a delay constraint. *IEEE Trans Wirel Commun*. 2020;19(6):3727-3737.
32. Petar P, Federico C, Kaibin H, et al. A perspective on time toward wireless 6G. *Proc IEEE*. 2022;110(8):1116-1146.
33. Junyan W, Xiangdong J, Zhi C, Xin Z. Optimization on information freshness for multi-access users with energy harvesting cognitive radio networks. *Trans Emerg Telecommun Technol*. 2022;33(11):e4591.
34. Guo C, Wu S, Deng Z, Jiao J, Zhang N, Zhang Q. Age-optimal power allocation policies for NOMA and hybrid NOMA/OMA Systems. Paper presented at: IEEE International Conference on Communications (ICC). Montreal, Canada; 2021.
35. Wang C, Zhang R, Tan J, Jiao B. Hybrid NOMA user grouping for short packet communications in IoT network with different types of devices. Paper presented at: IEEE International Conference on Communications (ICC) Workshops. Seoul, Korea; 2022:228-234.
36. Maatouk A, Assaad M, Ephremides A. Minimizing the age of information: NOMA or OMA? Paper presented at: IEEE Conference on Computer Communications (INFOCOM) Workshops. Paris, France; 2019:102-108.
37. Xiaofang S, Shihao Y, Nan Y, Zhiguo D, Chao S, Zhangdui Z. Short-packet downlink transmission with non-orthogonal multiple access. *IEEE Trans Wirel Commun*. 2018;17(7):4550-4564.
38. Yanqing X, Chao S, Tsung-Hui C, Shih-Chun L, Yajun Z, Gang Z. Transmission energy minimization for heterogeneous low-latency NOMA downlink. *IEEE Trans Wirel Commun*. 2019;19(2):1054-1069.
39. Leonidas G, Roch G, Abhay P. Optimal multiplexing on a single link: delay and buffer requirements. *IEEE Trans Inform Theory*. 1997;43(5):1518-1535.
40. Yury P, Vincent PH, Sergio V. Channel coding rate in the finite blocklength regime. *IEEE Trans Inform Theory*. 2010;56(5):2307-2359.
41. Muhammad A, Leila M, Sonia A. NOMA versus OMA in finite blocklength regime: link-layer rate performance. *IEEE Trans Vehic Technol*. 2020;69(12):16253-16257.
42. Mahyar S, Massimo C, Mischa D, Johnson SJ. On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT. *IEEE J Select Areas Commun*. 2017;35(10):2238-2252.
43. Georgiadis L, Neely MJ, Tassiulas L, et al. *Resource Allocation and Cross-Layer Control in Wireless Networks*. Now Publishers; 2006;1-144.
44. Chunhui L, Changyang S, Nan Y, Quek Tony QS. Secure transmission rate of short packets with queueing delay requirement. *IEEE Trans Wirel Commun*. 2021;21(1):203-218.

45. Mojtaba V, Robert S, Zhiguo D, Vincent PH. Non-orthogonal multiple access: common myths and critical questions. *IEEE Wirel Commun*. 2019;26(5):174-180.
46. Yunnuo X, Yijie M, Onur D, Bruno C. Rate-splitting multiple access with finite blocklength for short-packet and low-latency downlink communications. *IEEE Trans Vehic Technol*. 2022;71(11):12333-12337.
47. She C, Yang C. Ensuring the quality-of-service of tactile internet. Paper presented at: IEEE 83rd Vehicular Technology Conference (VTC Spring). Nanjing, China; 2016:1-5.
48. Jain Rajendra K, Chiu Dah-Ming W, Hawe WR. *A Quantitative Measure of Fairness and Discrimination*. Eastern Research Laboratory, Digital Equipment Corporation; 1984:21.

How to cite this article: Gamgam OB, Karasan E. Satisfying strict deadlines for cellular Internet of Things through hybrid multiple access. *Trans Emerging Tel Tech*. 2023;e4814. doi: 10.1002/ett.4814

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 (8) is nonnegative 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

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

Note that FTM- and CTM-based objective functions in (A1) and (A4), respectively, are equal for $P_i \in \left\{0, P_i^{\text{req}}\right\}$. CTM based objective function in (A4) increases linearly with P_i for $P_i \in [0, P_i^{\text{req}})$. 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 (16) can be obtained using P_i^* and available power region $P_i \in \mathcal{P}_i$.

APPENDIX 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,

$$\theta^* = \frac{\Gamma_i}{X_i\ln 2}-\frac{\mathcal{B}N_0}{|h_i|^2}. \tag{B1}$$

θ^N in (25) can be obtained using θ^* and available total power region $\theta \in \mathcal{P}_{(i,j)}$.

APPENDIX 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

$$P_j^* = \begin{cases} \mathcal{B}N_0 \left(\frac{\Gamma_j |h_j|^2 + \Gamma_i |h_i|^2}{(\Gamma_i - \Gamma_j)|h_i|^2 |h_j|^2} \right) & \text{, if } X_i = X_j \\ \max(\zeta_1, \zeta_2) & \text{, otherwise} \end{cases}, \tag{C1}$$

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

$$\mathtt{a} = \ln 2 \left(X_j - X_i \right) \tag{C2a}$$

$$\mathtt{b} = \Gamma_i - \Gamma_j + \ln 2 \left(X_j - X_i \right) \mathcal{B}N_0 \left(\frac{1}{|h_i|^2} + \frac{1}{|h_j|^2} \right) \tag{C2b}$$

$$\mathtt{c} = \mathcal{B}N_0 \left(\frac{\Gamma_i}{|h_i|^2} - \frac{\Gamma_j}{|h_j|^2} \right) + \frac{\left(X_j - X_i \right) \ln 2 (\mathcal{B}N_0)^2}{|h_i|^2 + |h_j|^2}. \tag{C2c}$$

Under $\Gamma_j \geq \Gamma_i$ condition, $P'_{j,(i,j)} \Big|_{\Phi_F}$ in (28) can be obtained using P_j^* and available power region $P_j \in \{0\} \cup [P_{j,(i,j)}^{\min}, P_j^{\dagger}]$.

APPENDIX D. PROOF OF THEOREM 4

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

$$\mathcal{M}_j^O(P_j) \Big|_{\Phi_C} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j) \Big|_{\Phi_C} \text{ when } P_i < P_{i,(i,j)}^{\text{req}} \text{ and } P_j = P_{j,(i,j)}^{\text{req}} = P_j^{\text{req}}. \tag{D1}$$

$$\mathcal{M}_i^O(P_i) \Big|_{\Phi_C} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j) \Big|_{\Phi_C} \text{ when } P_i = P_{i,(i,j)}^{\text{req}} > P_i^{\text{req}} \text{ and } P_j < P_{j,(i,j)}^{\text{req}}. \tag{D2}$$

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 $P_j < P_{j,(i,j)}^{\text{req}}$ and $P_i < P_{i,(i,j)}^{\text{req}}$. Therefore, we only need to consider the 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 the optimal decision by selecting one of $\{0, 0\}$ and $\left\{ P_{i,(i,j)}^{\text{req}}, P_{j,(i,j)}^{\text{req}} \right\}$ power allocation options for minimizing $\mathcal{M}_{(i,j)}^N(P_i, P_j) \Big|_{\Phi_C}$. The decision in (31) and (32) can be obtained under total power constraint, such that $P_i + P_j \leq P_{(i,j)}^{\max}$.

AUTHOR BIOGRAPHIES

Onur Berkay Gamgam received the BS and MS 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 received the BS degree from Middle East Technical University, MS degree from Bilkent University and PhD degree from Rutgers University, all in electrical engineering. He worked at Bell Laboratories for 3 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.