Received 9 March 2022; accepted 31 March 2022. Date of publication 4 April 2022; date of current version 15 April 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3164733

RAN Slicing Performance Tradeoffs: Timing Versus Throughput Requirements

FEDERICO CHIARIOTTI (Member, IEEE), ISRAEL LEYVA-MAYORGA (Member, IEEE), ČEDOMIR STEFANOVIĆ (Senior Member, IEEE), ANDERS E. KALØR (Graduate Student Member, IEEE), AND PETAR POPOVSKI (Fellow, IEEE)

Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark

CORRESPONDING AUTHOR: F. CHIARIOTTI (e-mail: fchi@es.aau.dk)

This work was supported in part by the Huawei as part of the STELLAR project, and in part by the Velux Foundation under the Villum Investigator grant WATER.

ABSTRACT The coexistence of diverse services with heterogeneous requirements is a fundamental feature of 5G. This necessitates efficient *Radio Access Network (RAN) slicing*, defined as sharing of the wireless resources among diverse services while guaranteeing their throughput, timing, and/or reliability requirements. In this paper, we investigate RAN slicing for an uplink scenario in the form of multiple access schemes for two user types: (1) broadband users with throughput requirements and (2) intermittently active users with timing requirements, expressed as either Latency-Reliability (LR) or Peak Age of Information (PAoI). Broadband users transmit data continuously, hence, are allocated non-overlapping parts of the spectrum. We evaluate the trade-offs between the achievable throughput of a broadband user and the timing requirements of an intermittent user under Orthogonal Multiple Access (OMA) and Non-Orthogonal Multiple Access (NOMA), considering capture. Our analysis shows that NOMA, in combination with packet-level coding, is a superior strategy in most cases for both LR and PAoI, achieving a similar LR with only a slight 2% decrease in throughput with respect to the case where an independent channel is allocated to each user. The latter solution leads to the upper bound in performance but requires double the amount of resources than the considered OMA and NOMA schemes. However, there are extreme cases where OMA achieves a slightly greater throughput than NOMA at the expense of an increased PAoI.

INDEX TERMS Age of information (AoI), heterogeneous services, non-orthogonal multiple access (NOMA), reliability, slicing.

I. INTRODUCTION

THE FIFTH generation of mobile networks (5G) aims to support three main service categories: enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC) [1]. eMBB is the direct evolution of the 4G mobile broadband service with higher data rates, along with greater spectral and spatial efficiency. URLLC services, on the other hand, usually involve exchange of small amounts of data, but require latency in the order of a few milliseconds and high reliability guarantees, e.g., a packet loss ratio below 10^{-5}. Finally, mMTC also involve transmissions of small amounts of data per device, but consist of hundreds or thousands of devices in the service area. The main challenge in mMTC is to design access networking mechanisms that maximize the success probability while maintaining an adequate *timing* in data delivery and resource efficiency.

The main strategy for service co-existence adopted by 3GPP [2], [3] is *network slicing*, which refers to the allocation of subsets of the network resources to the active services. The idea is to provide performance guarantees by limiting the mutual impact among services and/or service categories [4]. In general, Radio Access Network (RAN) slicing can be implemented in the form of Orthogonal Multiple Access (OMA) and Non-Orthogonal Multiple Access (NOMA). OMA schemes, such as Frequency Division Multiple Access (FDMA), Time

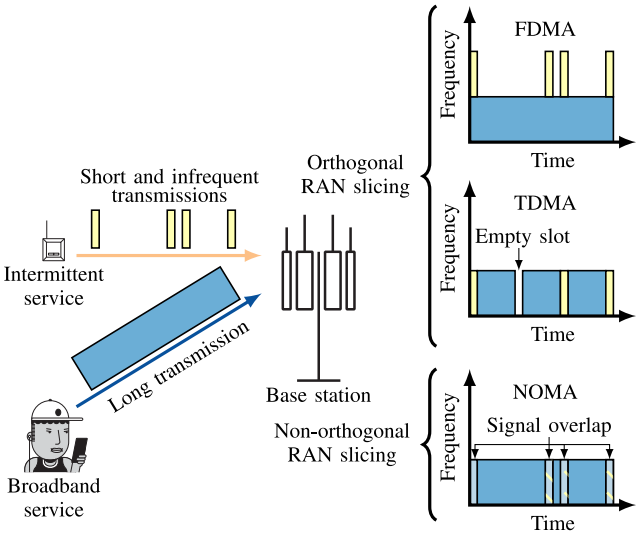


FIGURE 1. Orthogonal (i.e., FDMA and TDMA) and non-orthogonal RAN slicing (i.e., NOMA) between broadband and intermittent services.

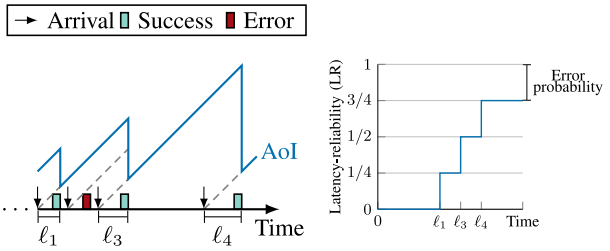


FIGURE 2. Exemplary diagram of the AoI and latency-reliability KPIs in a period with four packet transmissions. The latency ℓ_j for packets transmitted with errors is set to ∞.

Division Multiple Access (TDMA), and Code Division Multiple Access (CDMA), have been extensively studied and implemented in commercial and cellular systems. Moreover, OMA seems to be the approach preferred by 3GPP for 5G and beyond 5G systems, contextualized in the concept of bandwidth part [2]. On the other hand, in NOMA the same time-frequency resources are assigned to multiple services or users. We are considering the uplink case, as the users are not coordinated and compete for resources; in the downlink, the Base Station (BS) can schedule the resources, either orthogonally or by allocating appropriate power in broadcast NOMA schemes. NOMA allows, for example, to increase the number of served users with the available resources and/or the spectral efficiency of the system [5]. To enable communication in shared time-frequency resources, NOMA is usually accompanied by multi-user detection techniques, like separation of the users in the code domain, or in the power domain accompanied with Successive Interference Cancellation (SIC) where the individual signals are in turn decoded and subtracted from the received composite signal [5]–[8].

The difference between OMA and NOMA slicing is illustrated in Fig. 1. Here, it can be seen that TDMA and NOMA achieve higher resource utilization than FDMA when the intermittent service transmits infrequently, while the difference will be less pronounced when the intermittent service transmits frequently.

The performance of OMA and NOMA slicing has been widely studied in the presence of multiple users of the same service type [6], [9]–[11]. For instance, the trade-offs in achievable data rates for eMBB services are characterized in an Additive White Gaussian Noise (AWGN) channel with OMA and power-domain NOMA [10], [11]. On the other hand, performance with heterogeneous services has been studied only by a few works [12]–[15], mostly concentrating on TDMA as the orthogonal alternative to NOMA. However, the apparent trade-off between efficiency and timeliness of the slicing schemes is particularly relevant in this scenario. In our previous work, we derived the performance trade-offs with heterogeneous services with TDMA and NOMA with packet-level coding in a simplified collision channel model, which provides conservative results for NOMA [16], [17]. Some results with capture, obtained by simulation, were provided in [17], which served as one of the main motivations for this study, as these illustrated the potential gains of NOMA. The aim of this work is to provide an extensive and exact analytical treatment of OMA and NOMA slicing in an uplink scenario with two different service types: broadband and intermittent users with throughput and timing requirements, respectively.

We assume that the broadband users transmit data continuously and are primarily interested in achieving a high throughput. In contrast, intermittent users transmit short packets sporadically and are primarily interested in the *timeliness* of their data, expressed through two different Key Performance Indicator (KPI). The first KPI, which reflects flow-level requirements, is Peak Age of Information (PAoI), relevant for users that send updates of an ongoing process in which the freshness of information is the most important objective. PAoI measures the time elapsed since the generation of the last received update until a new update is received [18], and it is therefore determined by the transmission latency, reliability and the update generation pattern. PAoI-focused applications can tolerate individual packet losses, as there are no strict reliability requirements and new updates can supersede old ones. The second KPI, which reflects packet-level requirements, is denoted by *Latency-Reliability (LR)* and captures the probability of delivering individual packets within a given latency threshold [19]. For this, we use the distribution of latency where lost packets are defined to have infinite latency. LR captures, for example, URLLC traffic with strict constraints on the reliability of communication within a maximum latency. Our specific focus will be on computing high percentiles of LR and PAoI, which can be used to design systems with probabilistic reliability guarantees.

The scenario assumed in this paper comprises a single frequency band (i.e., bandwidth part in 5G New Radio (NR)

terminology) sliced in time to accommodate one broadband and one intermittent user.[1] In the case of NOMA, we assume application of SIC (1) in conjunction with the capture effect, such that the colliding packets are immediately resolved, and (2) coupled with the packet-level coding, such that after decoding of the broadband user block, the interference is removed and past packets from the intermittent user can be recovered. We analyze the achievable performance and the inherent trade-offs, providing closed-form expressions for throughput of the broadband user and timing of the intermittent user. The derivations are contextualized for a simple fading-based channel model, however, the elaborated approach is general and easily transferable to other settings.

In particular, the main contributions of this paper are:

- We analyze the joint use of OMA or NOMA with packet-level coding, which can significantly improve the performance of SIC by allowing the receiver to recover undecipherable packets after decoding a data block;
- We analyze the operating regions and trade-offs of OMA and NOMA with a realistic channel model that includes the capture effect;
- We show that the inclusion of capture gives rise to fundamental differences in performance in comparison to the erasure channel-based analyses from our previous work [16], [17];
- We analyze the impact of the wireless conditions of the intermittent user, including distance from the BS and path loss, on the performance of NOMA in terms of timing and throughput;
- We provide design guidelines for selecting the multiple access scheme and its parameters, depending on:
    1) The requirements and features of the different types of services in the system;
    2) The available bandwidth;
    3) The wireless conditions of the intermittent user.

To the best of our knowledge, our work is the first to combine packet-level coding and a channel model that includes the possibility of capture in a rigorous analysis of NOMA systems with heterogeneous traffic, which derives analytical probability mass functions (pmfs) both latency and AoI.

We observe that, while FDMA provides the upper bound in performance, NOMA schemes offer significant benefits w.r.t. OMA when the target KPI for the intermittent user is LR. Specifically, NOMA can achieve similar performance trade-offs as FDMA but with a much higher resource utilization. On the other hand, the potential gains of NOMA w.r.t. TDMA decrease when the target KPI is PAoI, with TDMA outperforming NOMA in extreme cases where throughput is maximized in exchange for a longer PAoI.

The rest of the paper is organized as follows. Section II presents the related work. The system model and KPIs are specified in Section III. We then derive the analytical distributions of those metrics for OMA and NOMA in Section IV and Section V, respectively. Section VI presents simulation results and discussion of the performance of the different access schemes. Finally, Section VII concludes the paper.

1. The scenario is inspired by the latest non-orthogonal multiplexing approaches in the uplink studied by the 3GPP [20].

II. RELATED WORK

Non-orthogonal slicing, in the form of NOMA, offers the possibility of increasing the spectral efficiency and the number of supported users with respect to OMA in exchange for a greater decoding complexity at the receiver to perform SIC [5], [21] or other multi-user detection techniques. Hence, NOMA has been widely studied in the literature in systems with a single service type [5], [6], [9]–[11]. NOMA often assumes user separation in the power domain such that the benefits of SIC can be fully exploited. However, different performance gains have been observed for NOMA in the uplink and in the downlink. In particular, the effect of power control in the uplink can be eclipsed by the channel conditions of the users in combination with imperfect channel state information [21].

A particularly interesting approach towards heterogeneous service coexistence with NOMA is presented in [6], emphasizing the importance of power control in NOMA and formulating resource allocation as a non-cooperative game and as a matching problem. One of the first studies that addresses the coexistence of heterogeneous services in OMA and NOMA was presented in [13], considering different combinations of 5G services in an uplink scenario. Specifically, eMBB users are allocated orthogonal resources between them; these coexist with either one URLLC user or with mMTC traffic, which follows a Poisson distribution. It was observed that NOMA may offer benefits with respect to OMA depending on the rate of the eMBB users and on the type of coexisting traffic, in terms of the achievable rates for eMBB and URLLC traffic and the achievable eMBB rates as a function of the arrival rate of mMTC packets. This work was later extended to a multi-cell scenario with strict latency guarantees for URLLC traffic [22], where it was observed that NOMA leads to a greater spectral efficiency w.r.t. OMA. This same conclusion was drawn by Maatouk *et al.* [9] in an uplink scenario with two users with and one service type. The aim of the latter study was to minimize the average AoI. However, it was also observed that a greater spectral efficiency does not directly translate in a lower average AoI.

In general, power diversity [14] is the most common way to successfully use NOMA with heterogeneous services: if either the eMBB or the mMTC user has a much better channel than the other, both packets can be decoded with SIC, leading to significant performance gains. A more comprehensive approach is to exploit space diversity as well, using a Multiple-Input Multiple-Output (MIMO) system to differentiate between the wireless channels for the two types of services and maximizing their spectral efficiency. This can be further improved in dense and Orthogonal Frequency-Division Multiple Access (OFDMA) networks by selecting

the couples of interfering users to maximize the spatial and channel diversity: by putting users with very different channel statistics together, the SIC recovery probability increases significantly [23]. The same goal can be accomplished by controlling the rate of eMBB users, lowering their decoding threshold depending on the channel of the coupled mMTC user [15]. To the best of our knowledge, this and our own previous works [16], [17] are the first to consider coding, exploiting redundant information to decode collided packets instead of just the properties of the wireless channel.

AoI is a relatively new performance metric, but it has been rapidly adopted due to its relevance in remote control tasks [24]. Most papers in the literature have examined it in the context of queuing theory, often in ideal systems with Markovian service [25], because of the relative simplicity of the analysis, but a few have considered the effect of physical layer issues and medium access schemes on it. Recent works compute the average AoI in Carrier Sense Multiple Access (CSMA) [26], ALOHA [27] and slotted ALOHA [28] networks, considering the impact of the different medium access policies on the age.

Another important missing piece in the AoI literature is the worst-case performance analysis: while studies on average AoI are common, the tail of its distribution is rarely considered [29], limiting the relevance of the existing body of work for reliability-oriented applications. The analytical complexity of deriving the complete distribution of the age is a daunting obstacle; only recently, advances have been made in this line. A recent work [30] uses the Chernoff bound to derive an upper bound of the quantile function of the AoI for two queues in tandem with deterministic arrivals. Using a more analytical approach, the PAoI distribution was computed over a single-hop link with fading and retransmissions in [31]. We also mention the work in [32], where different service classes are defined and the system is modeled as an M/G/1/1 clocking queue with hyperexponential service time. However, in the latter, only the service rate is different among classes. Then, the classes can adapt the arrival rate to minimize the AoI.

III. SYSTEM MODEL

We consider an uplink scenario with a set of users $\mathcal{U}$ transmitting data to a BS through an OFDMA system whose time-frequency resources are divided into time slots and bandwidth parts as in 5G NR [33]. A bandwidth part is defined as a set of contiguous resource blocks in the frequency domain. We consider the case where the users transmit up to one packet per time slot, occupying the whole bandwidth part. Herein, we consider the case where two heterogeneous users must be allocated resources. The options for the BS are 1) allocate the users in the same bandwidth part and define how the resources should be shared among them or 2) allocate a different bandwidth part for each of the users using FDMA.

User 1 is a broadband user following the eMBB model [1]: it is a full-buffer user that maintains an infinite transmission queue. To counteract potential packet losses due to fading and noise, the broadband user implements a packet-level coding scheme, where blocks of K (source) packets are encoded to generate a *frame* of N coded packets. The coded packets are transmitted in the same bandwidth part, one after the other, and have a zero probability of linear dependence, which can be achieved, e.g., with Maximum Distance Separable (MDS) codes. Hence, the block of packets is decoded when any K packets from the same frame are received without errors.

User 2 is an intermittent user that, with (a relatively low) probability α, may generate a short packet at each slot. Each generated packet is transmitted just once at the next available time slot. User 2 maintains up to Q of the generated packets in a transmission queue. We denote by $q_t \leq Q$ the length of the queue at time slot t. If a new packet is generated when $q_t = Q$, user 2 discards the oldest buffered packet and adds the newly generated one at the end of the queue.

When both users are allocated to the same bandwidth part, the BS must allocate the time slots that are available for the transmission of each of the two users. For this, we define the resource allocation set $\mathcal{A}_t \subseteq \mathcal{U}$ as the subset of users that can access the bandwidth part at slot t. We define the following three types of slot allocations.

1) *Broadband:* The slot is reserved for the broadband user. Hence, $\mathcal{A}_t = 1$.
2) *Intermittent:* The slot is reserved for the intermittent user and may use it if it has one or more packets in its queue. Hence, $\mathcal{A}_t = 2$.
3) *Mixed:* Both users are allowed to access the slot, implying that the signals will overlap if the intermittent user transmits. Hence, $\mathcal{A}_t = \{1, 2\}$.

We define the following different access schemes.

1) *TDMA:* Both users are allocated resources in a single bandwidth part with separate broadband and intermittent slots. Specifically, the intermittent user has a reserved intermittent slot once every T_{int} slots, while the rest of the slots are reserved for the broadband user. As such, this is a non-orthogonal slicing in the frequency domain but orthogonal in the time domain where there is no interference among the users.
2) *NOMA:* Both users are allocated resources in a single bandwidth part with only mixed slots. Hence, the intermittent user may transmit at any slot. The two users interfere any time the intermittent user transmits, but the packets can be recovered through SIC by decoding one of the signals immediately or at a later time slot (as explained in Section III-A).
3) *FDMA:* The users are allocated resources in different and non-overlapping bandwidth parts. Hence, one of the bandwidth parts contains only broadband slots and the other bandwidth part contains only intermittent slots.

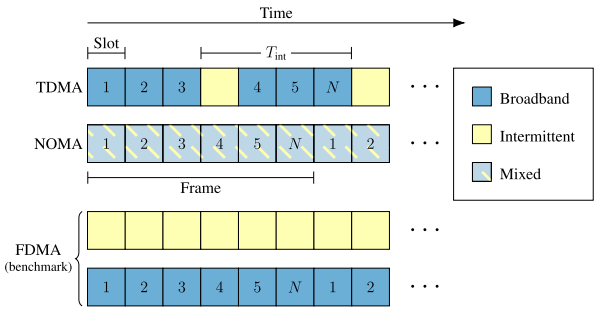


FIGURE 3. Frame structure for the TDMA, NOMA, and FDMA schemes with $K = 4$ and $N = 6$.

The frame structures for these access schemes are illustrated in Fig. 3. FDMA can only take place when there are two bandwidth parts available, where the bandwidth part allocated to the intermittent user is likely to be under-utilized, as α is relatively small. Hence, this approach results in a low resource efficiency and is used here only as a benchmark in which the performance of the users is fully independent of each other.

A. CHANNEL MODEL

We consider a block fading channel, where the received signal by the BS at time slot t is given as:

$$y_t = \sum_{u\in\{1,2\}} h_{u,t}a_{u,t}x_{u,t} + z_t \tag{1}$$

where $h_{u,t} \in \mathbb{C}$ is the random fading coefficient for user u; z_t is the circularly-symmetric Gaussian noise with power σ^2; and $a_{u,t}$ is an activity indicator, equal to 1 if the user is active in slot t and 0 otherwise. A user is active at time t if and only if $u \in \mathcal{A}_t$ and if its packet queue $q_{u,t}$ is not empty:

$$a_{u,t} = I(u \in \mathcal{A}_t)I(q_{u,t} > 0), \tag{2}$$

where $I(\cdot)$ is the indicator function, equal to 1 if the condition is true and 0 otherwise. Let $P_u \le P_{\max}$ be the selected (i.e., fixed) transmission power for user u, where $P_{\max}$ is the maximum transmission power. The Signal-to-Noise Ratio (SNR) of user u is given as:

$$\text{SNR}_{u,t} = \frac{|h_{u,t}|^2 P_u a_{u,t}}{\sigma^2} = \frac{|h'_{u,t}|^2 P_u a_{u,t}}{\ell_u \sigma^2}, \tag{3}$$

where ℓ_u is the constant large-scale fading, including path loss, and $|h'_{u,t}|$ is the envelope of the channel coefficient due to fast fading. The path loss is a function of the distance of user u to the BS r_u, the carrier frequency f_c, and a path loss exponent η. We assume the standard path loss model:

$$\ell_u = \frac{(4\pi f_c)^2 r_u^\eta}{c^2}, \tag{4}$$

where c is the speed of light.

The expected SNR for a transmission by user u is:

$$\overline{\text{SNR}}_u = \frac{\mathbb{E}\left[|h_{u,t}|^2\right]P_u}{\sigma^2} = \frac{\mathbb{E}\left[|h'_{u,t}|^2\right]P_u}{\ell_u \sigma^2}, \tag{5}$$

By using the standard assumption of treating the interfering signal v as AWGN noise, the Signal-to-Interference-plus-Noise Ratio (SINR) for user u in the considered scenario is:

$$\text{SINR}_{u,t} = \frac{|h_{u,t}|^2 P_u a_{u,t}}{\sigma^2 + |h_{v,t}|^2 P_v a_{v,t}} = \frac{\text{SNR}_{u,t}}{1 + \text{SNR}_{v,t}}. \tag{6}$$

B. RECEPTION MODEL

Let X be the Random Variable (RV) of the number of packets from user 1 that belong to the same block and are received without errors. The success probability of user 1, denoted as $p_{s,1}$, is defined as the probability of receiving K or more packets out of the N that comprise the block:

$$p_{s,1} \triangleq \Pr[X \ge K|N]. \tag{7}$$

We define γ_u as the threshold in the SINR to decode a packet transmitted by user u. In practice, the threshold is mainly a function of the modulation and coding scheme and the receiver sensitivity. In the following, we consider the case in which the fading envelope $|h_{u,t}|$ is Rayleigh distributed and define the erasure probabilities for the two users.

Erasure probability for the broadband user: The BS has collected sufficient Channel State Information (CSI) about the broadband user so that the appropriate transmission power $P_1 \le P_{\max}$, block length, and data rate (i.e., modulation and coding) to achieve a target erasure probability ε_1 are signaled back to the broadband user. Thus, user 1 transmits with power:

$$P_1 \le P_{\max} \ : \ \Pr\left[\text{SNR}_{1,t} < \gamma_1\right] = \varepsilon_1. \tag{8}$$

Erasure probability for the intermittent user: Due to the infrequent transmissions, the CSI of this user at the BS is insufficient to perform a precise selection of parameters as done for the broadband user. Instead, the user always transmits at $P_2 = P_{\max}$ and its erasure probability ε_2 is determined by its path loss ℓ_2 and by γ_2. Hence, the erasure probability for user 2 is calculated from (5) as:

$$\varepsilon_2 = \Pr\left[\text{SNR}_{2,t} < \gamma_2\right] = 1 - e^{\frac{-\gamma_2}{\overline{\text{SNR}}_2}} = 1 - e^{\frac{-\gamma_2 \ell_2 \sigma^2}{P_u}}, \tag{9}$$

since $\mathbb{E}[|h'_{u,t}|^2] = 1$ for unitary Rayleigh fading.

Next, let $o_u \in \{\mathcal{I}, \mathcal{E}, \mathcal{R}\}$ denote the outcome of user u's signal. The possible outcomes are described in the following for the case where SIC is performed only with the signals received within the same time slot (i.e., intra-slot SIC).

- $\mathcal{I}$: The signal of interest is decoded within the same time slot. Either 1) the signal of interest has sufficient SINR to be immediately decoded or 2) the signal from the other user is immediately decoded, its interference removed through SIC and, then, the signal of interest is decoded.
- $\mathcal{E}$: The signal has insufficient SNR to be decoded.
- $\mathcal{R}$: The signal of interest has sufficient SNR but cannot be decoded within the same time slot. This occurs when the signals of both users overlap and these have

insufficient SINR. However, the signal of interest can be decoded after the interference from the other is removed.

Further, the ordered pairs (o_1, o_2) describe possible outcomes for the signals of both users when these overlap as:

- $(\mathcal{I}, \mathcal{I})$: The signal with the highest SINR is decoded and its interference is immediately removed through SIC. Then, the second signal is decoded.
- $(\mathcal{I}, \mathcal{E})$ and $(\mathcal{E}, \mathcal{I})$: The signal with the higher SINR is decoded and its interference is immediately removed through SIC. However, the second signal cannot be decoded due to the impact of noise, i.e., a low SNR.
- $(\mathcal{E}, \mathcal{E})$: The SNR of both signals is insufficient and, thus, neither can be decoded even if the interference from the other user were removed.
- $(\mathcal{R}, \mathcal{E})$: The signal from user 2 has insufficient SNR, while the signal from user 2 has a sufficient SNR but insufficient SINR. Since the system cannot remove the interference from user 2 without decoding it first, both packets remain undecoded.
- $(\cdot, \mathcal{R})$: In this case, none of the signals can be immediately recovered but the signal from user 2 could be decoded if the interference from user 1 is removed via SIC after decoding the block of user 1. Therefore, this outcome includes the cases $(\mathcal{E}, \mathcal{R})$ and $(\mathcal{R}, \mathcal{R})$.

Note that the cases $(\mathcal{I}, \mathcal{R})$ and $(\mathcal{R}, \mathcal{I})$ are not feasible, as outcome $\mathcal{I}$ indicates that a signal is immediately decoded and that its interference to the other signal is removed. Hence, the other signal is either decoded after intra-slot SIC (i.e., $\mathcal{I}$) or not decoded (i.e., $\mathcal{E}$).

Throughout the rest of the paper, we assume that the interference from the decoded signals of the users can be perfectly removed with SIC. Under this assumption, the closed-form expressions for the probabilities of these outcomes, denoted as $\pi_{o_1 o_2}$, are provided in Appendix A for a Rayleigh fading channel. Nevertheless, these probabilities can be calculated assuming imperfect SIC and for a different fast fading model so they can be directly as input to our analytical model.

C. KEY PERFORMANCE INDICATORS

The broadband user (user 1) is interested in maximizing its throughput S under the constraint that the desired reliability $p_{s,1}$ must be greater than $1 - \varepsilon_1$. Note that increasing the reliability of the broadband user entails a reduction in the coding rate K/N. The definition of the throughput is not given in bits per second, but normalized for the slot: in an ideal scenario, user 1 would have a throughput equal to 1, corresponding to 1 successful packet transmitted for each slot. As $N - K$ packets in each block contain redundant information, reducing the coding rate correspondingly reduces the amount of information transmitted in each new block.

The intermittent user (user 2) is interested in the timeliness of its data, i.e., either LR or PAoI, where we select quantile ρ as the main KPIs. Let T and Δ be the RVs of LR and PAoI, respectively. Then, the quantile ρ of LR is defined as:

$$T_\rho := \min_n \{n : \Pr[T \leq n] > \rho\} \tag{10}$$

and the quantile ρ of the PAoI Δ_ρ is defined analogously. The latter allows us to evaluate the tail distribution of the PAoI in a general scenario and can be used to compare the performance with different values of α [31]. The use of a generic quantile allows us to tune the required reliability to the application. However, the value of the LR T_ρ is infinite if the selected quantile $\rho > 1 - \gamma_2$, as packets are not retransmitted and the error probability for an interference-free transmission is γ_2.

Since S and the timeliness of the intermittent user are interlinked, we evaluate their trade-offs for a specific activation probability α and erasure probabilities ε_u, via the *Pareto frontier* defined in the following.

Definition 1: Let $\mathcal{C}$ be the set of feasible configurations for a specific access method and $f : \mathcal{C} \to \mathbb{R}^2$. Next, let:

$$Y = \left\{ \left(S, \tau_\rho\right) : \left(S, \tau_\rho\right) = f(c), c \in \mathcal{C} \right\}, \tag{11}$$

where S is the throughput of user 1 and τ_ρ is the timeliness of user 2, and $\tau_\rho \in \{T_\rho, \Delta_\rho\}$. The *Pareto frontier* is the set:

$$\mathcal{P}_\rho(Y) = \left\{ (S, \tau_\rho) \in Y : \forall \left(S', \tau'_\rho\right) \in Y : S > S' \vee \tau_\rho < \tau'_\rho \right\}. \tag{12}$$

Besides obtaining the Pareto frontiers, we evaluate the schemes by setting a minimum requirement for S, the throughput of user 1. The optimal configuration of an access method is then defined as the combination of parameters that minimizes the timing, either expressed as LR or PAoI, while maintaining S above the minimum required.

Table 1 summarizes the relevant notation introduced in this section. To simplify the analytical expressions in the rest of the paper, we define the binomial pmf $\mathrm{Bin}(K; N, p)$ as:

$$\mathrm{Bin}(K; N, p) = \begin{cases} \binom{N}{K} p^K (1-p)^{N-K}, & \text{if } K \leq N; \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

and the multinomial pmf $\mathrm{Mult}(\mathbf{K}; N, \mathbf{p})$ as:

$$\mathrm{Mult}(\mathbf{K}; N, \mathbf{p}) = \frac{N! \prod_{i=1}^{|\mathbf{p}|} p_i^{K_i} \left(1 - \sum_{i=1}^{|\mathbf{p}|} p_i\right)^{N - \sum_{i=1}^{|\mathbf{p}|} K_i}}{\left(N - \sum_{i=1}^{|\mathbf{p}|} K_i\right)! \prod_{i=1}^{|\mathbf{p}|} K_i!}, \tag{14}$$

where $|\mathbf{p}|$ is the length of vector $\mathbf{p}$. As for the binomial case, the probability is 0 if $\sum_{i=1}^{|\mathbf{K}|} K_i > N$. Finally, we denote $\delta(x)$ as the delta function, which is equal to 1 if $x = 0$ and 0 otherwise, and $[x]^+ = \max(x, 0)$.

IV. PERFORMANCE WITH TDMA

Here we derive the KPIs for the TDMA system, for a LR- or PAoI-oriented intermittent user. For LR, the maximum length of the intermittent user's queue is assumed to be fixed to some $Q \geq 1$. On the other hand, for PAoI, the maximum length of the intermittent user's queue is set to

TABLE 1. Notation summary.

Symbol	Description	Symbol	Description
$\mathcal{U}=\{1,2\}$	Set of users; $u=1$ is the broadband user	P_u	Transmission power of user u
	and $u=2$ is the intermittent user	$\overline{\text{SNR}}_u$	Expected SNR for user u
K	Size of the source block for user 1	$\text{SNR}_{u,t}$	SNR of user u at slot t
N	Size of the coded block for user 1	$\text{SINR}_{u,t}$	SINR of user u at slot t
Q	Maximum queue length for user 2	ε_u	Erasure probability of user u
T_{int}	Period between slots allocated to user 2 in TDMA	$o_u \in \{\mathcal{I},\mathcal{R},\mathcal{E}\}$	Outcome for user u when signals overlap
$t \in \mathbb{Z}$	Time slot index	(o_1,o_2)	Outcome when signals overlap
q_t	Length of the queue for user 2 at t	$\pi_{o_1 o_2}$	Probability of outcome (o_1,o_2)
$\mathcal{A}_t \subseteq \mathcal{U}$	Allocation of time slot t	$p_{s,u}$	Success probability of user u
$a_{u,t}$	Activity indicator for user u at t; 1 if active	S	Throughput of user 1
$h_{u,t}$	Fading envelope for user u at t	T	RV of LR for user 2
σ^2	Noise power	Δ	RV of PAoI for user 2
ℓ_u	Path loss of user u	T_ρ, Δ_ρ	Quantile ρ of LR and PAoI
r	Distance between user 2 and the BS	$\delta(x)$	Delta function, equal to 1 if $x=0$ and 0 otherwise

$Q=1$. This is because transmitting the newest packet is the optimal strategy to minimize PAoI but packet retransmissions are not allowed.

In the assumed TDMA system, the broadband user has frames of N slots, each of which contains K data packets and $N-K$ redundancy packets, while the intermittent user has one reserved slot every T_{int}. The success probability for user 1 is easy to compute:

$$p_{s,1} = \sum_{m=K}^{N} \text{Bin}(m; N, 1-\varepsilon_1). \tag{15}$$

The expected throughput of user 1 is:

$$S = p_{s,1}\frac{(T_{\text{int}}-1)K}{T_{\text{int}}N}. \tag{16}$$

That is, the throughput measures the rate of innovative (i.e., non-redundant) packets received at the BS from user 1 per time slot. As the broadband user can only use $T_{\text{int}}-1$ slots for each T_{int}, setting up more frequent transmission opportunities for the intermittent user reduces the throughput.

A. LATENCY-RELIABILITY (LR)

In order to derive the pmf of the LR for the intermittent user, without loss of generality, we take the origin of time to be a slot in which a transmission occurs. We define a Markov chain representing the state of the queue q_t for the intermittent user, i.e., the number of packets in the queue at time t. The transition matrix of the chain is $\mathbf{M}^{(1)}$, whose elements $M^{(1)}_{ij}$ represent the probability of transitioning from state i to state j in the queue of the intermittent user at the end of such slot [34]. The elements $M^{(1)}_{ij}$ are obtained as:

$$M^{(1)}_{ij} = \begin{cases} 0 & \text{if } j < i-1; \\ \text{Bin}(j-i+1; T_{\text{int}}, \alpha) & \text{if } i-1 \le j < Q; \\ \sum_{m=Q-i+1}^{T_{\text{int}}} \text{Bin}(m; T_{\text{int}}, \alpha) & \text{if } j = Q. \end{cases} \tag{17}$$

Let $\boldsymbol{\varphi}^{(1)} = [\varphi^{(1)}_0, \varphi^{(1)}_1, \ldots, \varphi^{(1)}_Q]$ be the steady-state distribution vector of the queue immediately after a transmission. From the transition matrix computed in (17), we can easily derive $\boldsymbol{\varphi}^{(1)}$ as the left-eigenvector of $\mathbf{M}^{(1)}$ with eigenvalue 1, normalized to sum to 1 to be a valid probability metric:

$$\boldsymbol{\varphi}^{(1)}(\mathbf{I}-\mathbf{M}^{(1)}) = \mathbf{0} \wedge \sum_{q=0}^{Q} \varphi^{(1)}_q = 1. \tag{18}$$

It is easy to derive the steady-state distribution of the queue q_n (i.e., n slots after a transmission) as:

$$\varphi^{(n)}_q = \begin{cases} \sum_{s=0}^{q} \varphi^{(1)}_s \text{Bin}(q-s; n\alpha) & \text{if } q < Q; \\ \sum_{s=0}^{Q} \sum_{m=Q-s}^{n-1} \varphi^{(1)}_s \text{Bin}(m; n-1, \alpha) & \text{if } q = Q, \end{cases} \tag{19}$$

where $\varphi^{(1)}_q$ is the q-th element of vector $\boldsymbol{\varphi}^{(1)}$. If a packet is queued behind q others, it will be transmitted at the $(q+1)$-th opportunity, unless new arrivals make the system drop some of the packets ahead of it in the queue: we remind the reader that, if the queue is full, the oldest packet (i.e., the first in the queue) is dropped. Let $g_i \in \{0, 1, \ldots, T_{\text{int}}\}$ for $i \ge 1$ be the number of packets generated by user 2 between the i-th and $(i+1)$-th intermittent slots after the current one. Further, let g_0 be the number of packets generated between the current time slot and the next intermittent slot. We define:

$$\mathcal{G}^{(n)}_\ell = \left\{\left[g_0 \in \{0, 1, \ldots, T_{\text{int}}-n\}, g_1, \ldots, g_\ell\right]\right\} \tag{20}$$

to be the set of possible vectors for the number of packets generated by user 1 given that there are $T_{\text{int}}-n$ slots until the next intermittent slot.

The probability of occurrence of each element $\mathbf{g} \in \mathcal{G}^{(n)}_\ell$ is:

$$p_{\text{gen}}(\mathbf{g}; \ell, n) = \text{Bin}(g_0; T_{\text{int}}-n, \alpha) \prod_{i=1}^{\ell} \text{Bin}(g_i; T_{\text{int}}, \alpha). \tag{21}$$

Each vector $\mathbf{g} \in \mathcal{G}^{(n)}_\ell$ represents a possible sequence of events over the next ℓ transmission slots: we can compute the queue dynamics for a given $\mathbf{g}$ to determine if and when the considered packet will be transmitted.

At each intermittent slot, up to one packet is transmitted and, hence, removed from the queue. Other packets are removed if the number of generated packets exceeds the number of remaining spaces in the queue. If the considered packet has q others ahead of it in the queue, we can then give the condition $\psi_\ell^{(\mathbf{g},q)}$, which is 1 if the packet is transmitted at or before the ℓ-th transmission opportunity:

$$\psi_\ell^{(\mathbf{g},q)} = \delta\left(\sum_{i=1}^{\ell}\left[q+1-Q+\sum_{j=1}^{i} g_j\right]^+ + \ell - (q+1)\right). \tag{22}$$

In order for the packet to be transmitted at the ℓ-th opportunity, we then have$\psi_\ell^{(\mathbf{g},q)} = 1$ and $\psi_k^{(\mathbf{g},q)} = 0$, $\forall k < \ell$.

We can then define the set $\mathcal{S}_\ell^{(n,q)}$, which includes all the vectors $\mathbf{g} \in \mathcal{G}_\ell^{(n)}$ for which the considered packet is transmitted at the ℓ-th opportunity:

$$\mathcal{S}_\ell^{(n,q)} = \left\{\mathbf{g} \in \mathcal{G}_\ell^{(n)} : \psi_\ell^{(\mathbf{g},q)} - \sum_{k=1}^{\ell-1}\psi_k^{(\mathbf{g},q)} = 1\right\}. \tag{23}$$

With a small abuse of notation, we define $\mathcal{S}_\ell^{(n,Q)} = \mathcal{S}_\ell^{(n,Q-1)}$, as one packet in the queue is always discarded if the new packet finds a full queue. Since the packet is either transmitted within $q+1$ transmission attempts or discarded, the conditioned success probability $p_{s,2}(n,q;T_{\text{int}})$ for the intermittent user is simply given by:

$$p_{s,2}(n,q) = \sum_{\ell=1}^{\min(Q,q+1)} \sum_{\mathbf{g}\in\mathcal{S}_\ell^{(n,q)}} p_{\text{gen}}(\mathbf{g};\ell,n)(1-\varepsilon_2). \tag{24}$$

We can then remove the condition on the success probability by applying the law of total probability:

$$p_{s,2} = \sum_{n=1}^{T_{\text{int}}}\sum_{q=0}^{Q}\frac{\varphi_q^{(n-1)} p_{s,2}(n,q)}{T_{\text{int}}}. \tag{25}$$

Finally, we compute the latency pmf $p_T(t)$, considering the fact that it takes 1 slot to transmit the packet:

$$p_T(t) = p_{s,2}\sum_{n=1}^{T_{\text{int}}}\sum_{q=0}^{Q}\varphi_q^{(n)}\sum_{\mathbf{g}\in\mathcal{S}_\ell^{(n,q)}}\frac{\delta(\text{mod}(t+n-1,T_{\text{int}})}{T_{\text{int}}p_{s,2}(n,q)} \times p_{\text{gen}}\left(\mathbf{g};\left\lfloor\frac{t+n-1}{T_{\text{int}}}\right\rfloor,n\right), \tag{26}$$

where $\text{mod}(m,n)$ is the integer modulo function. The computational complexity of the LR pmf is $\mathcal{O}(Q^3 T_{\text{int}}^{Q+2})$, and consequently, computing the Cumulative Distribution Function (CDF) is $\mathcal{O}(Q^4 T_{\text{int}}^{Q+3})$.

B. PEAK AGE OF INFORMATION

In the PAoI-oriented case, the pmf is given by the sum of the waiting time W between the instant in which a new packet is generated and the slot in which it is transmitted and the inter-update interval Z between consecutive successful transmissions [24].

Since $Q = 1$, the generated packets are always sent at the first available transmission opportunity. The pmf of the waiting time W for a successful transmission is given by:

$$p_W(w) = \frac{\alpha(1-\alpha)^{w-1}}{1-(1-\alpha)^{T_{\text{int}}}},\ w \in \{1,\ldots,T_{\text{int}}\}. \tag{27}$$

The probability of having a successful update in a given intermittent slot is given by:

$$\xi = (1-(1-\alpha)^{T_{\text{int}}})(1-\varepsilon_2). \tag{28}$$

We can then compute the pmf of Z. Since exactly one slot every T_{int} is reserved for the intermittent user, Z is T_{int} times the number of reserved slots between consecutive successful transmissions. This is a geometric random variable with parameter ξ, whose pmf is then:

$$p_Z(z) = (1-\xi)^{\frac{z}{T_{\text{int}}}-1}\xi\,\delta(\text{mod}(z,T_{\text{int}})). \tag{29}$$

The pmf of the PAoI is:

$$p_\Delta(t+1) = p_Z(t-\text{mod}(t,T_{\text{int}}))p_W(1+\text{mod}(t,T_{\text{int}})). \tag{30}$$

The computational complexity of the PAoI calculation is then $\mathcal{O}(1)$ for the pmf, and $\mathcal{O}(t)$ for computing the CDF up to t.

V. PERFORMANCE WITH NOMA

We now derive the distributions of the KPIs in the NOMA case, in which the broadband user has frames of N slots, all of which are mixed, i.e., allocated both to the intermittent and broadband user.

First, we define $p_{r,1}$ as the probability that a packet from the broadband user is received correctly in a given slot, which is given by:

$$p_{r,1} = ((1-\alpha)(1-\varepsilon_1) + \alpha(\pi_{\mathcal{II}} + \pi_{\mathcal{IE}})). \tag{31}$$

The probability that the block from the broadband user is decoded in the d-th slot of the frame, denoted as $p_D(d)$, is then given by:

$$p_D(d) = \begin{cases} p_{r,1}\text{Bin}(K-1;d-1,p_{r,1}), & \text{if } d \geq K; \\ 0, & \text{otherwise.}\end{cases} \tag{32}$$

The CDF of the decoding instant D, $P_D(d)$, is given by:

$$P_D(d) = \sum_{m=K}^{d}\text{Bin}(m;d,p_{r,1}). \tag{33}$$

We then simply have $p_{s,1} = P_D(N)$. The average throughput for the broadband user is

$$S = \frac{KP_D(N)}{N}. \tag{34}$$

The computational complexity of the throughput calculation is then $\mathcal{O}(N-K)$.

A. LATENCY-RELIABILITY (LR)

We now analyze the latency distribution for the intermittent user. All the intermittent packets transmitted after decoding slot d – once the block from the broadband user has been decoded – can be either decoded immediately or lost with probability $\varepsilon_2 = \pi_{\mathcal{IE}} + \pi_{\mathcal{EE}} + \pi_{\mathcal{RE}}$. On the other hand, if the intermittent user packet is sent before the decoding slot d, it is decoded instantly with probability $\pi_{\mathcal{II}} + \pi_{\mathcal{EI}}$, while it can be decoded after SIC with probability $\pi_{\cdot\mathcal{R}}$.

We can then count the number of collisions C_b before the decoding slot, I_b of which are resolved immediately by SIC (events $(\mathcal{I},\mathcal{I})$ and $(\mathcal{E},\mathcal{I})$), while V_b are buffered for future decoding (event $(\cdot,\mathcal{R})$). If the broadband user is decoded in slot d, we have the following joint pmf $p_{C_b,I_b,V_b|D}(c_b,i_b,v_b|d)$:

$$\begin{aligned} p_{C_b,I_b,V_b|D}(c_b,i_b,v_b|d) = &\sum_{\ell=K-d+c_b}^{\min(c_b,K-1)} \mathrm{Bin}(c_b; d-1, \alpha) \\ &\times \frac{p_{r,1}}{p_D(d)} \sum_{m=0}^{\min(i_b,\ell)} \mathrm{Bin}(K-1-\ell; d-1-c_b, 1-\varepsilon_1) \\ &\times \mathrm{Mult}([m, i_b-m, v_b, \ell-m]; c_b, [\pi_{\mathcal{II}}, \pi_{\mathcal{EI}}, \pi_{\cdot\mathcal{R}}, \pi_{\mathcal{IE}}]). \end{aligned} \tag{35}$$

We then simply take the four cases for packets from the intermittent user (transmitted before slot d, in slot d, after slot d, or in lost frames), and compute $p_{s,2}$. We compute $p_{C_d,I_d}(c_d,i_d)$, the probability that a packet from user 2 is sent and correctly decoded in the same slot as the broadband user block decoding:

$$p_{C_d,I_d}(c_d,i_d) = \begin{cases} \frac{\alpha\pi_{\mathcal{II}}}{p_{r,1}}, & c_d=1, i_d=1; \\ \frac{\alpha\pi_{\mathcal{IE}}}{p_{r,1}}, & c_d=1, i_d=0; \\ \frac{(1-\alpha)(1-\varepsilon_1)}{p_{r,1}}, & c_d=0, i_d=0; \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

We then give the probability of having C_a packets after the decoding of the broadband user block in slot d, I_a of which are correctly received:

$$\begin{aligned} p_{C_a,I_a|D}(c_a,i_a|d) = &\frac{\mathrm{Bin}(i_a; c_a, \pi_{\mathcal{II}} + \pi_{\mathcal{EI}} + \pi_{\cdot\mathcal{R}})}{p_D(d)} \\ &\times \mathrm{Bin}(c_a; N-d, \alpha). \end{aligned} \tag{37}$$

We note that during and after the decoding slot there are no buffered packets from the intermittent user, as SIC can be performed immediately: the broadband user's block has already been decoded, and no new information from it will arrive in future slots. Finally, we can consider the case in which the broadband user frame is not decoded: in this case, the only intermittent packets that are decoded are immediate captures. We can then compute the probability $p_{C_z,I_z|\tilde{D}}(c_z,i_z)$:

$$\begin{aligned} p_{C_z,I_z|\tilde{D}}(c_z,i_z) = &\sum_{c=0}^{\min(K-1,N-c_z)} \mathrm{Bin}(c; N-c_z, 1-\varepsilon_1) \\ &\times \sum_{\ell=0}^{\min(c_z,K-1-c)} \mathrm{Bin}(c_z; N, \alpha) \\ &\times \sum_{m=0}^{\min(\ell,i_z)} \frac{\mathrm{Mult}([m, i_z-m, \ell-m]; c_z, [\pi_{\mathcal{II}}, \pi_{\mathcal{EI}}, \pi_{\mathcal{IE}}])}{1-p_{s,1}}. \end{aligned} \tag{38}$$

We now know that all packets transmitted by the intermittent user at or after the decoding of the broadband block, or in frames for which the broadband block is not decoded, are either lost or decoded immediately. To compute the latency distribution, we then only need to distinguish the case in which a packet transmitted before d is decoded instantly or after SIC. The probability of a packet from the intermittent user being decoded instantly is then $p_T(1)$:

$$\begin{aligned} p_T(1) = &\left(1-p_{s,1}\right) \sum_{c_z=1}^{N} \sum_{i_z=0}^{c_z} \frac{i_z p_{C_z,I_z|\tilde{D}}(c_z,i_z)}{(1-\mathrm{Bin}(0;N,\alpha))c_z} \\ &+ \sum_{d=K}^{N} \sum_{c_b=0}^{d-1} \sum_{c_d=0}^{1} \sum_{c_a=0}^{N-d} \sum_{i_b=0}^{c_b} \sum_{i_d=0}^{c_d} \sum_{i_a=0}^{c_a} \sum_{v_b=0}^{c_b-i_b} p_D(d) \\ &\times \frac{(i_b+i_d+i_a) p_{C_d,I_d}(c_d,i_d) p_{C_a,I_a|D}(c_a,i_a|d)}{(1-\mathrm{Bin}(0;N,\alpha))(c_b+c_d+c_a)} \\ &\times p_{C_b,I_b,V_b|D}(c_b,i_b,v_b|d). \end{aligned} \tag{39}$$

As the delay from any packet decoded after SIC is distributed uniformly between 2 and $d+1$, we can easily compute $p_T(t)$:

$$\begin{aligned} p_T(t) = &\sum_{d=\min(K,t-1)}^{N} p_D(d) \sum_{c_b=0}^{d-1} \sum_{c_d=0}^{1} \sum_{c_a=0}^{N-d} \sum_{i_b=0}^{c_b} \sum_{v_b=0}^{c_b-i_b} \frac{v_b}{d} \\ &\times \sum_{i_d=0}^{c_d} \sum_{i_a=0}^{c_a} \frac{p_{C_d,I_d}(c_d,i_d) p_{C_a,I_a|D}(c_a,i_a|d)}{(c_b+c_d+c_a)(1-\mathrm{Bin}(0;N,\alpha))} \\ &\times p_{C_b,I_b,V_b|D}(c_b,i_b,v_b|d), \quad t \in \{2,\ldots,N\}. \end{aligned} \tag{40}$$

The combination of (39) and (40) is the latency-reliability pmf for the intermittent user. We then have:

$$p_{s,2} = \sum_{t=1}^{N} p_T(t). \tag{41}$$

The computational complexity of the LR pmf calculation is $\mathcal{O}(N(N-K)^3)$ and of CDF is $\mathcal{O}(N^4(N-K)^3)$.

B. PEAK AGE OF INFORMATION

In order to derive the pmf of the PAoI, we first need to compute some auxiliary values. First, we derive the probability that the first decoded packet from the intermittent user in a frame is decoded in slot f, denoted as $p_F(f)$ and given in (42), shown at the bottom of the next page. This result is given by the previously computed probabilities, and considers all possible outcomes for both users.

It is then easy to get ξ, the probability of decoding a new intermittent packet in a frame:

$$\xi = \sum_{f=1}^{N} p_F(f). \tag{43}$$

The pmf of the number of slots Y from the frame start until the first decoded packet from the intermittent user is:

$$p_Y(y) = (1-\xi)^{\lfloor \frac{y}{N} \rfloor} p_F(\text{mod}(y, N)). \quad (44)$$

We now consider the probability $p_U(x)$ of receiving an update from the intermittent user, i.e., a packet with newer information than the one already available, in slot x. We have the following pmf, conditioned on the decoding slot d of the broadband block. First, we consider the case in which $d < x$:

$$p_{U|D}(x|d) = \sum_{c_b=1}^{d-1} \sum_{i_b=1}^{c_b} \sum_{v_b=0}^{c_b-i_b} \frac{i_b p_{C_b,I_b,V_b|D}(c_b, i_b, v_b|d)}{d-1}. \quad (45)$$

Next, for $d > x$,

$$p_{U|D}(x|d) = \sum_{c_a=1}^{N-d} \sum_{i_a=1}^{c_a} \frac{i_a}{N-d} p_{C_a,I_a|D}(c_a, i_a|d). \quad (46)$$

Finally, the case for $d = x$ is more complex: in the previous cases, all packets that were successfully decoded were also newer than any previously decoded ones, as their delay was 1. In this case, we have to consider the fact that some packets that were buffered and can only be decoded after getting the full broadband user block and performing SIC might be older than a packet that was already decoded through immediate capture. We then have:

$$\begin{aligned} p_{U|D}(d|d) = &\sum_{c_b=1}^{d-1} \sum_{i_b=0}^{c_b-1} \sum_{v_b=1}^{c_b-i_b} \sum_{m=i_b+v_b}^{d-1} \sum_{c_d=0}^{1} \frac{v_b p_{C_d,I_d}(c_d, 0)}{d-1} \\ &\times \mathcal{H}_{m-1,d-1}(i_b+v_b-1, i_b+v_b-1) \\ &\times p_{C_b,I_b,V_b(c_b,i_b,v_b|d)|D} + p_{C_d,I_d}(1,1), \end{aligned} \quad (47)$$

where $\mathcal{H}_{M,N}(m,n)$ is the hypergeometric distribution, whose pmf is given by:

$$\mathcal{H}_{M,N}(m,n) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}. \quad (48)$$

We also consider the probability $p_{U|\tilde{D}}(x)$, i.e., the probability of receiving an update in slot x if the broadband user block is not decoded (in this case, all packets are valid updates, as their delay is 1):

$$p_{U|\tilde{D}}(x) = \sum_{c_z=1}^{N} \sum_{i_z=1}^{c_z} \frac{i_z p_{C_z,I_z|\tilde{D}}(c_z, i_z)}{N}. \quad (49)$$

By applying the law of total probability, we obtain $p_U(x)$

$$p_U(x) = \sum_{d=K}^{N} p_D(d) p_{U|D}(x|d) + \big(1-p_{s,1}\big) p_{U|\tilde{D}}(x). \quad (50)$$

We now compute the probability that a given update is the last in the frame, given that the decoding happens in slot d, denoted as $p_{L|D}(\ell|d)$. Again, we distinguish three cases, starting from $\ell < d$:

$$\begin{aligned} p_{L|D}(\ell|d) = &\sum_{c_b=1}^{d-1} \sum_{i_b=1}^{\ell} \sum_{v_b=0}^{\ell-i_b} \sum_{c_d=0}^{1} \sum_{c_a=0}^{N-d} \frac{i_b p_{C_d,I_d}(c_d, 0)}{d-1} \\ &\times p_{C_b,I_b,V_b|D}(c_b, i_b, v_b|d) p_{C_a,I_a|D}(c_a, 0|d) \\ &\times \frac{\mathcal{H}_{\ell-1,d-2}(v_b+i_b-1, v_b+i_b-1)}{p_{U|D}(\ell|d)}. \end{aligned} \quad (51)$$

If $\ell = d$, we have:

$$p_{L|D}(d|d) = \sum_{c_a=0}^{N-d} p_{C_a,I_a|D}(c_a, 0|d). \quad (52)$$

Finally, if $\ell > d$ the probability is:

$$\begin{aligned} p_{L|D}(\ell|d) = &\sum_{c_a=1}^{\ell-d} \sum_{i_a=1}^{c_a} \frac{i_a p_{C_a,I_a|D}(c_a, i_a|d)}{(N-d) p_{U|D}(\ell|d)} \\ &\times \mathcal{H}_{\ell-d-1,N-d-1}(i_a-1, i_a-1). \end{aligned} \quad (53)$$

The probability that an update in slot ℓ is the last in the frame, given that the broadband user frame is lost, $p_{L|\tilde{D}}(\ell)$, is:

$$p_{L|\tilde{D}}(\ell) = \sum_{c_z=1}^{\ell} \sum_{i_z=1}^{c_z} \frac{p_{C_z,I_z}(c_z, i_z) i_z \mathcal{H}_{N-\ell,N-1}(0, i_z-1)}{N p_{U|\tilde{D}}(\ell)}. \quad (54)$$

Combining the expressions derived above, we get:

$$p_L(\ell) = \sum_{d=K}^{N} p_D(d) p_{L|D}(\ell|d) + (1-p_{s,1}) p_{L|\tilde{D}}(\ell). \quad (55)$$

$$\begin{aligned} p_F(f) = &\sum_{a=0}^{f-1} \alpha\big(\pi_{\mathcal{I},\mathcal{I}} + \pi_{\mathcal{E},\mathcal{I}}\big) \sum_{\ell=0}^{\min(K-2,a)} \text{Bin}(a; f-1, \alpha) \\ &\times \sum_{m=0}^{\min(K-1-\ell, f-a-1)} \text{Mult}([0,0,\ell]; a, [\pi_{\mathcal{I}\mathcal{I}}, \pi_{\mathcal{E}\mathcal{I}}, \pi_{\mathcal{I}\mathcal{E}}]) \text{Bin}(m; f-a-1, 1-\varepsilon_1) \\ &+ p_D(f) \sum_{c_b=0}^{f-1} p_{C_b,I_b,V_b|D}(c_b, 0, v_b|f) p_{C_d,I_d}(1,1) + \sum_{d=K}^{f-1} p_D(d) \sum_{c_b=0}^{d} p_{C_b,I_b,V_b|D}(c_b, 0, 0|d) \\ &\times \sum_{c_d=0}^{1} p_{C_d,I_d}(c_d, 0) \text{Bin}(0; f-d-1, \alpha(\pi_{\mathcal{I}\mathcal{I}} + \pi_{\mathcal{E}\mathcal{I}} + \pi_{\cdot\mathcal{R}})) \alpha(\pi_{\mathcal{I}\mathcal{I}} + \pi_{\mathcal{E}\mathcal{I}} + \pi_{\cdot\mathcal{R}}). \end{aligned} \quad (42)$$

If the update is not the last in the frame, we can compute the conditioned pmf $p_{Z|U,D,\tilde{L}}(z|x,d)$ of the inter-update interval Z. We first consider the case in which $z+x<d$:

$$p_{Z|U,D,\tilde{L}}(z|x,d)=\sum_{c_b=2}^{d-1}\sum_{i_b=2}^{c_b}\frac{i_b(i_b-1)\mathcal{H}_{z-1,d-3}(0,i_b-2)}{p_{U|D}(x|d)} \times\sum_{v_b=0}^{c_b-i_b}\frac{p_{C_b,I_b,V_b|D}(c_b,i_b,v_b|d)}{(d-1)(d-2)(1-p_{L|D}(x|d))},\ x+z<d. \quad (56)$$

In this case, the only possibility to have another update after z is to have two immediate captures in slots x and $x+z$, without any immediate captures in between. Further, for $x>d$,

$$p_{Z|U,D,\tilde{L}}(z|x,d)=\sum_{c_a=2}^{N-d-z+1}\sum_{i_a=2}^{c_a}\frac{p_{C_a,I_a|D}(c_a,i_a|d)}{(N-d)(N-d-1)} \times\frac{i_a(i_a-1)\mathcal{H}_{z-1,N-d-2}(0,i_a-2)}{(1-p_{L|D}(x|d))p_{U|D}(x|d)},\ x>d\wedge x+z\le N. \quad (57)$$

Next, for $x=d$:

$$p_{Z|U,D,\tilde{L}}(z|d,d)=\sum_{c_a=1}^{N-d-z+1}\sum_{i_a=1}^{c_a}\frac{i_a p_{C_a,I_a|D}(c_a,i_a|d)}{(N-d)(1-p_{L|D}(d|d))} \times\mathcal{H}_{z-1,N-d-1}(0,i_a-1),\ d+z\le N. \quad (58)$$

We then consider the case that $x+z=d$:

$$p_{Z|U,D,\tilde{L}}(d-x|x,d)=\sum_{c_b=1}^{d-1}\sum_{i_b=1}^{c_b}\sum_{v_b=0}^{c_b-i_b}p_{C_b,I_b,V_b|D}(c_b,i_b,v_b|d) \times\frac{\mathcal{H}_{x-1,d-2}(i_b-1,i_b-1)}{(d-1)p_{U|D}(x|d)(1-p_{L|D}(x|d))} \times\left(p_{C_d,I_D}(1,1)+\mathbb{1}(v_b-1)\sum_{c_d=0}^{1}p_{C_d,I_d}(c_d,0) \times\left(1-\mathcal{H}_{x-i_b,d-i_b-1}(v_b,v_b)\right)\right), \quad (59)$$

where $\mathbb{1}(x)$ is the step function, equal to 1 if $x\ge 0$ and 0 otherwise. Finally, we can derive $p_{Z|U,\tilde{D},\tilde{L}}(z|x)$, if the broadband user frame is not decoded:

$$p_{Z|U,\tilde{D},\tilde{L}}(z|x)=\sum_{c_z=2}^{N}\sum_{i_z=2}^{c_z}\frac{\mathcal{H}_{z-1,N-2}(0,i_z-2)}{N(N-1)(1-p_{L|\tilde{D}}(x)p_{U|\tilde{D}}(x)} \times p_{C_z,I_z}(c_z,i_z),\ x+z\le N. \quad (60)$$

We now compute the pmf of the inter-update interval Z if the next packet is in the same frame:

$$p_Z(z|x)=\sum_{d=K}^{N}p_{Z|U,D,\tilde{L}}(z|x,d)p_D(d)+(1-p_{s,1}) \times p_{Z|U,\tilde{D},\tilde{L}}(z|x),\ z\le N-x. \quad (61)$$

On the other hand, if $z>N-x$, we have:

$$p_Z(z|x)=p_L(x)(1-\xi)^{\left\lfloor\frac{z-(N-x)}{N}\right\rfloor} \times p_F(\mathrm{mod}(z-(N-x),N)),\forall z>N-x. \quad (62)$$

The decoding delay W component of PAoI applies only if the update transmitted before the decoding slot d and decoded with SIC only after the decoding of the broadband user block. We then give $p_{W|U,D}(w|x,d)$, the pmf of W for an update in the same slot d which the broadband user block is decoded in:

$$p_{W|U,D}(w|d,d)=\frac{1}{p_{U|D}(x|d)}\Bigg(p_{C_d,I_d}(1,1)\delta(w-1)+\sum_{c_b=0}^{d-1} \times\sum_{v_b=1}^{\min(c_b,d-w+1)}\mathcal{H}_{w-2,d-2}(0,v_b-1) \times\sum_{i_b=0}^{\min(c_b,d-w+1)-v_b}\mathcal{H}_{w-2,d-v_b-1}(0,i_b) \times\frac{p_{C_b,I_b,V_b|D}(c_b,i_b,v_b|d)}{(d-1)}\sum_{c_d=0}^{1}p_{C_d,I_d}(c_d,0)\Bigg),\ x=d. \quad (63)$$

In all other cases, the packet is captured instantaneously, and we simply have:

$$p_{W|U,D}(w|x,d)=\delta(w-1),\ x\ne d. \quad (64)$$

If the broadband user frame is not decoded, the decoding delay is always 1, as the only updates are due to immediate capture:

$$p_{W|U,\tilde{D}}(w|x)=\delta(w-1). \quad (65)$$

By applying the law of total probability, we get:

$$p_{W|U}(w|x)=\sum_{d=K}^{N}p_D(d)p_{W|U,D}(w|x,d) +\left(1-p_{s,2}\right)p_{W|U,\tilde{D}}(w|x). \quad (66)$$

Finally, we get the PAoI as the convolution of Z and W and removing the condition on U:

$$p_\Delta(t)=\sum_{x=1}^{N}p_U(x)\sum_{w=1}^{\min(x,t-1)}p_{W|U}(w|x)p_{Z|U}(t-w|x). \quad (67)$$

The computational complexity of the PAoI pmf calculation is $\mathcal{O}(Nt+(N-K)N^4)$. Computing its CDF up to t requires $\mathcal{O}(Nt^2+(N-K)N^4)$ steps, considering an efficient implementation that uses memory to avoid computing the same value multiple times.

TABLE 2. Parameter settings.

Parameter	Symbol	Setting	Parameter	symbol	Setting
Coded block length for user 1	N	$\{2, 3, \ldots, 32\}$	Source block length for user 1	K	$< N$
Erasure probability of user 1	ε_1	0.1	Transmission power of user 2	P_2	23 dBm
Activation probability for user 2	α	$\{0.01, 0.05, 0.1\}$	Period between intermittent slots in TDMA	T_{int}	$\{1, 2, \ldots, 40\}$
SINR threshold to decode a packet	$\gamma_1 = \gamma_2$	3 dB	Noise power	σ^2	-127.216 dBm
Distance from user 2 to the BS	r	$\{50, 100, \ldots, 400\}$ m	Carrier frequency	f_c	2 GHz
Path loss exponent	η	$\{2.6, 3\}$	Queue length in TDMA	Q	4 packets
Quantile	ρ	0.9			

VI. EVALUATION

We assume that user 1 (the broadband user) selects its transmission power to achieve $\varepsilon_1 = 0.1$. On the other hand, user 2 (the intermittent user) transmits infrequently, and thus cannot get up-to-date information on the channel state. The best possible strategy for it is then to always transmit at maximum power; in this case, ε_2 depends on its distance from the BS r and the erasure probability ε_2 is minimized.

For performance evaluation, we set parameters that represent a typical 5G urban scenario [1]. Namely, the carrier frequency is 2 GHz, the path loss exponent is $\eta \in \{2.6, 3\}$ dB, the noise power σ^2 is determined by the noise temperature and the subcarrier spacing, set to a typical $\Delta f = 15$ kHz, plus a noise figure of 5 dB. The resulting noise power and other relevant parameter settings are listed in Table 2. For simplicity's sake, the SINR thresholds for decoding both users are set to the same value $\gamma_1 = \gamma_2 = 3$ dB. As a reference, the SINR threshold when calculating the maximum coverage in 5G is 0 dB [1]. Fig. 4 show the area plots for the probability of the outcomes when both users transmit in the same slot for $\eta \in \{2.6, 3\}$. The figure shows that a high reliability for the intermittent user is only achievable when it is close to the base station, particularly when $\eta = 3$. On the other hand, recovering packets after decoding the broadband user block is crucial, as case $(\cdot, \mathcal{R})$ occurs with a relatively high probability for both values of η and is critical to achieve high reliability for the intermittent user.

FDMA is selected as a benchmark for TDMA and NOMA. In FDMA, each user is allocated to a different frequency band with an equal bandwidth to the one considered for TDMA and NOMA. Because of this, FDMA achieves the upper bound in performance at the expense of consuming twice the amount of bandwidth resources than TDMA and NOMA. The expressions to calculate the throughput, LR, and PAoI with FDMA are provided in Appendix B.

An essential aspect of our analysis is to identify the values of K and N that maximize the throughput S of user 1. These can be selected independently of user 2's parameters for TDMA and FDMA and, hence, represent the optimal configuration for user 1 with these schemes.

Note that implementing a longer coded block size N would grant a greater throughput, bounded by $1 - \varepsilon_1$ for $N \to \infty$, but would also lead to a longer decoding latency and complexity. Hence, we limit the value of $N \leq 32$ to achieve an adequate balance between S and decoding latency and

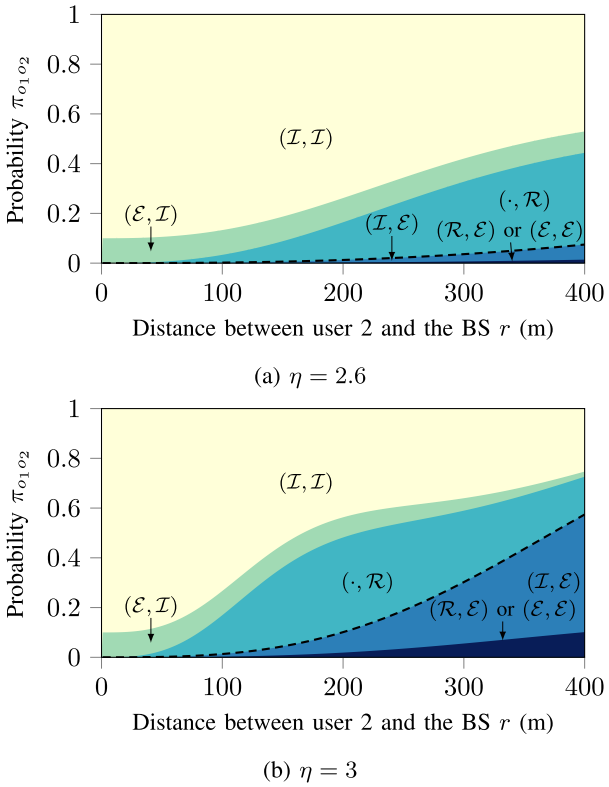
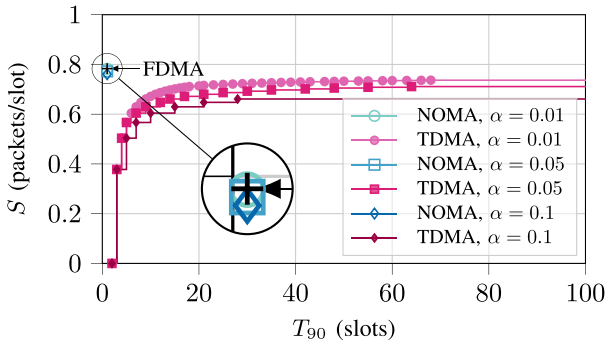


FIGURE 4. Area plot for the probabilities of the different outcomes (o_1, o_2) when the signals of both users collide for (a) $\eta = 2.6$ and (b) $\eta = 3$. The dashed line indicates the value of ε_2.

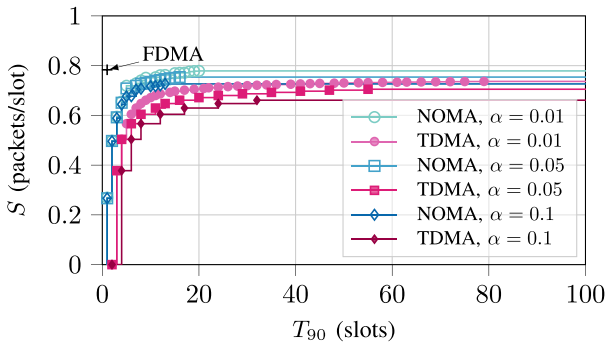
complexity. By restricting $N \leq 32$, the optimal configuration for user 1 for both TDMA and FDMA is $K = 26$ and $N = 32$, which leads to $p_{s,1} = 0.964$. With this configuration, FDMA achieves a throughput of $S = 0.7833$ for all cases, as user 1 operates in a separate channel from user 2 and, hence, there is no trade-off between S and the KPI of user 2. On the other hand, the optimal configuration for TDMA and NOMA depends on the desired performance trade-off and, hence, these are given at the end of this section. The performance trade-offs were also evaluated by Monte Carlo simulation, performing the check for 10^6 blocks. In all cases, the empirical CDF matched the theoretical curve perfectly: for this reason, the following graphs only report the theoretical results. The computational complexity of the theoretical calculations for each configuration is reported in Table 3.

TABLE 3. Big $\mathcal{O}$ computational complexity of the theoretical calculation.

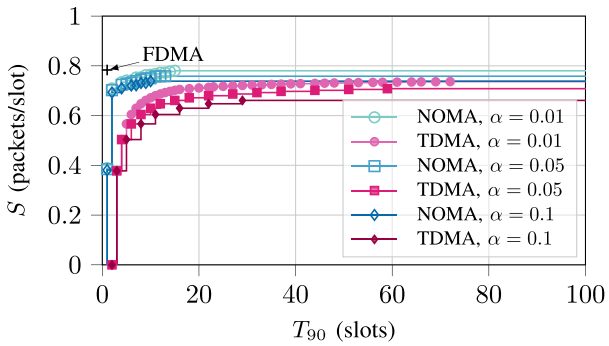
Value	TDMA	NOMA	FDMA
Throughput	1	$N-K$	1
LR pmf	$Q^3 T_{\text{int}}^{Q+2}$	$(N(N-K))^3$	$Nt+(N-K)N^4$
LR CDF	$Q^4 T_{\text{int}}^{Q+3}$	$N^4(N-K)^3$	$Nt^2+(N-K)N^4$
PAoI pmf	1	1	1
PAoI CDF	t	1	t



(a) $r = 150$ m, $\eta = 2.6$.



(b) $r = 150$ m, $\eta = 3$.



(c) $r = 250$ m, $\eta = 2.6$.

FIGURE 5. Pareto frontiers for latency-reliability versus throughput with TDMA and NOMA, with different values of α. The cross marks indicate the performance with FDMA (benchmark).

A. PARETO ANALYSIS

We first present the Pareto frontier for throughput of user 1 S and timing of user 2, for LR T_{90} or PAoI Δ_{90}, which describes the best achievable trade-offs between these KPIs. We consider three different distances (50, 150, and 250 m) for the intermittent user, with three different activation probabilities. It is easy to see in Fig. 5 that NOMA easily outperforms TDMA in terms of LR and throughput in all scenarios.

Furthermore, Fig. 5(a) shows that $T_{90} = 1$ can be achieved with NOMA if the distance and path loss allow to immediately decode more than 90% of the packets from user 2 due to capture and the use of SIC in the same slot. In these cases, the throughput with NOMA is only up to 2% lower than with FDMA. Therefore, NOMA is the most efficient scheme in these cases, as it achieves a similar performance to FDMA but with half the resources: one bandwidth part instead of two.

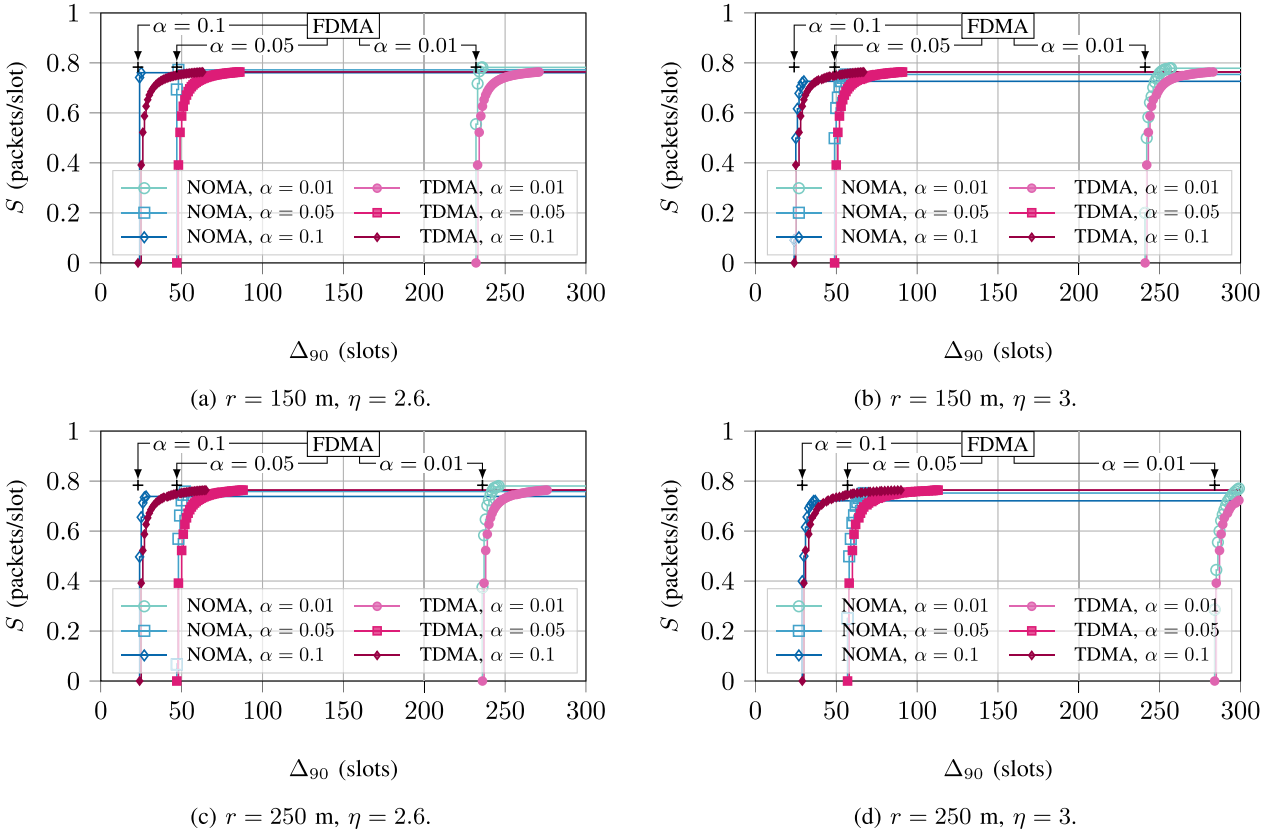
On the other hand, there is a strict trade-off between LR and throughput for all cases with TDMA, as the only way to reduce the latency is to decrease the period between intermittent slots T_{int}, which decreases the amount of resources assigned to the broadband user. The same trade-off appears with NOMA for the cases where $\pi_{\mathcal{I},\mathcal{I}} + \pi_{\mathcal{E},\mathcal{I}} < 0.9$ due to an increase in path loss, as shown in Fig. 5(b)-(c). In these cases, reducing the latency also requires reducing the efficiency of the code. However, the Pareto frontier for NOMA is always above and to the left of the curve for the equivalent scenario with TDMA, showing that NOMA can achieve better performance in both metrics. The Pareto frontiers for $r = 250$ m and path loss exponent $\eta = 3$ are not shown, as in this case it is impossible for the intermittent user to deliver 90% of packets at any latency (i.e., $T_{90} = \infty$).

NOMA also achieves a lower PAoI with high throughput for short distances, as shown in Fig. 6(a). As for the latency, the Pareto frontier increases abruptly and reaches its maximum $S \approx 0.78$, which is close to the one achieved with FDMA of 0.7833. This occurs at exactly or only a few time slots later than the minimum Δ_{90}. Thus, the resource efficiency of NOMA is much greater than that of FDMA and achieves similar trade-offs for the age as well.

On the other hand, for $r \geq 150$ m, TDMA achieves a higher throughput than NOMA at the expense of an increase in PAoI. This is expected, as greater values of T_{int} increase S but also Δ. Specifically, as described in Appendix B, the throughput with TDMA for $T_{\text{int}} \to \infty$ is equal to that with FDMA.

However, the activation rate α has the greatest impact on the PAoI. Hence, the choice of access method is of secondary importance to minimize PAoI as long as the parameters are chosen correctly and its impact decreases with α. This is because the interval time between consecutive packets with low values of α can be so long that reducing the latency for each individual packet has only a minor effect on the PAoI.

In conclusion, NOMA significantly outperforms TDMA in most scenarios, both for LR and PAoI, and can achieve almost the same performance as FDMA using half of the bandwidth. If the propagation conditions for the intermittent user are particularly bad and capture is relatively rare, TDMA can achieve a slightly better throughput when higher values



(a) $r = 150$ m, $\eta = 2.6$.

(b) $r = 150$ m, $\eta = 3$.

(c) $r = 250$ m, $\eta = 2.6$.

(d) $r = 250$ m, $\eta = 3$.

FIGURE 6. Pareto frontiers for PAoI versus throughput with TDMA and NOMA, with different values of α. The cross marks indicate the performance with FDMA (benchmark).

of the PAoI are acceptable, but the difference is minimal, and NOMA is always better for LR.

B. DISTANCE ANALYSIS

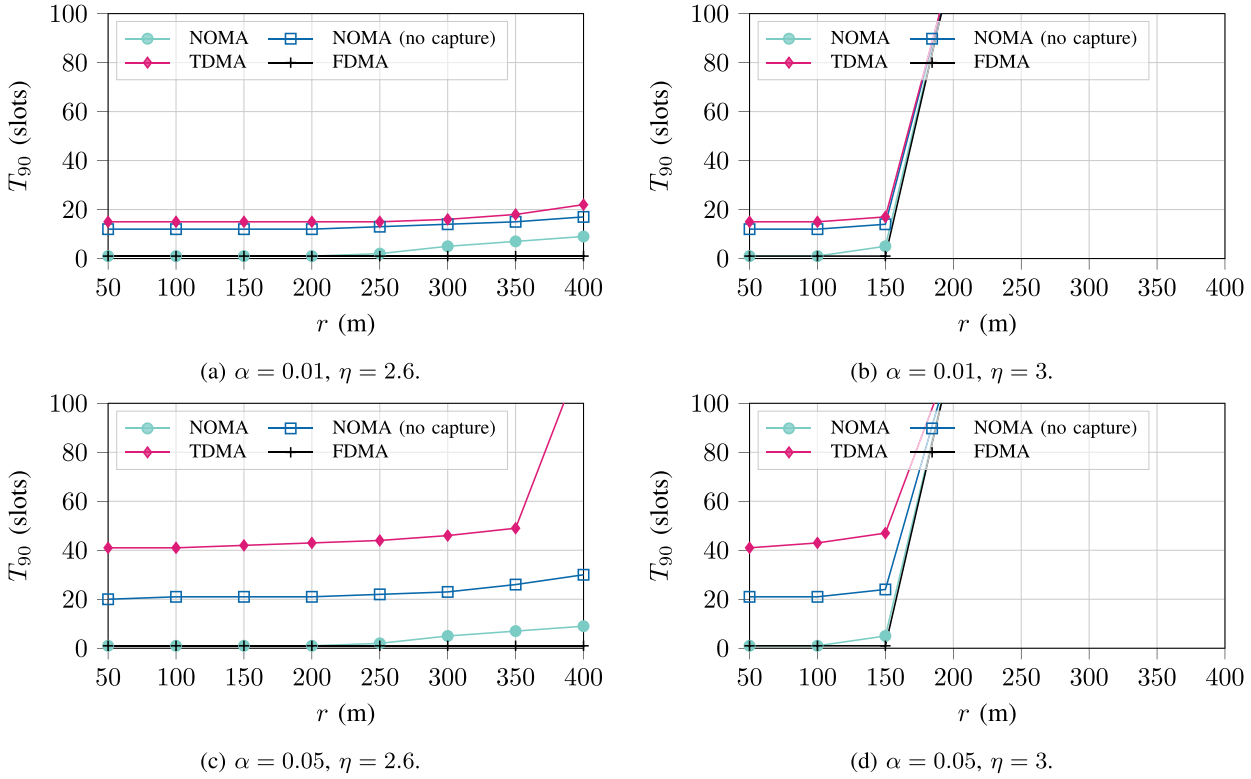
We now investigate the performance of the schemes as a function of the distance r between user 2 and the base station. In this case, we also consider the case for NOMA with fully destructive interference and, hence, no capture, which was investigated in our previous work [17]. In this later case, we have $\pi_{\cdot,\mathcal{R}} = 1 - \varepsilon_2$ and $\pi_{\mathcal{E},\mathcal{E}} = \varepsilon_2$ for any slot in which the two users collide, eliminating the possibility of instantaneous SIC. This scenario is naturally a lower bound for NOMA's performance, as removing the possibility of capture makes the scheme perform significantly worse.

Fig. 7 shows the performance of the schemes in terms of the minimum LR T_{90} that can be achieved while fulfilling a relatively high throughput requirement $S \geq 0.7$ for $\alpha = 0.01, 0.05$. In general, NOMA can outperform TDMA in most cases, but it is interesting to observe the behavior of the schemes when α is high. In these cases, we notice a performance drop for both TDMA, which has to allocate more slots to the intermittent user, and NOMA without capture, which has to increase the robustness of the packet-level code to protect the transmission from the additional intermittent user packets. On the other hand, capture allows NOMA to be more robust to the increased activation probability, maintaining a performance that is close to FDMA. In fact, while not shown in the figures, NOMA and FDMA are the only schemes that can achieve $S \geq 0.7$ with $\alpha = 0.1$, while the other schemes do not achieve the required throughput for any configuration.

On the other hand, we can confirm the trend that we observed in Fig. 6 for PAoI at different distances, as Fig. 8 shows that NOMA achieves a slightly lower Δ_{90} than TDMA. However, capture is essential for the NOMA scheme with higher values of α: without it, it performs slightly worse than TDMA for $\alpha = 0.05$, and it never reaches the required throughput for $\alpha = 0.1$. Finally, it can be seen that NOMA achieves similar values of Δ_{90} than FDMA for (1) most values of r with $\eta = 2.6$ and (2) short distances $r \leq 150$ with $\eta = 3$. This demonstrates that, in the cases where the system can benefit from capture and SIC, NOMA is nearly equivalent to FDMA in terms of performance, even when the latter utilizes twice the amount of resources. These cases occur, for example, when pairing the broadband user with an intermittent user located near the BS that achieves a high mean SNR.

C. PARAMETER ANALYSIS

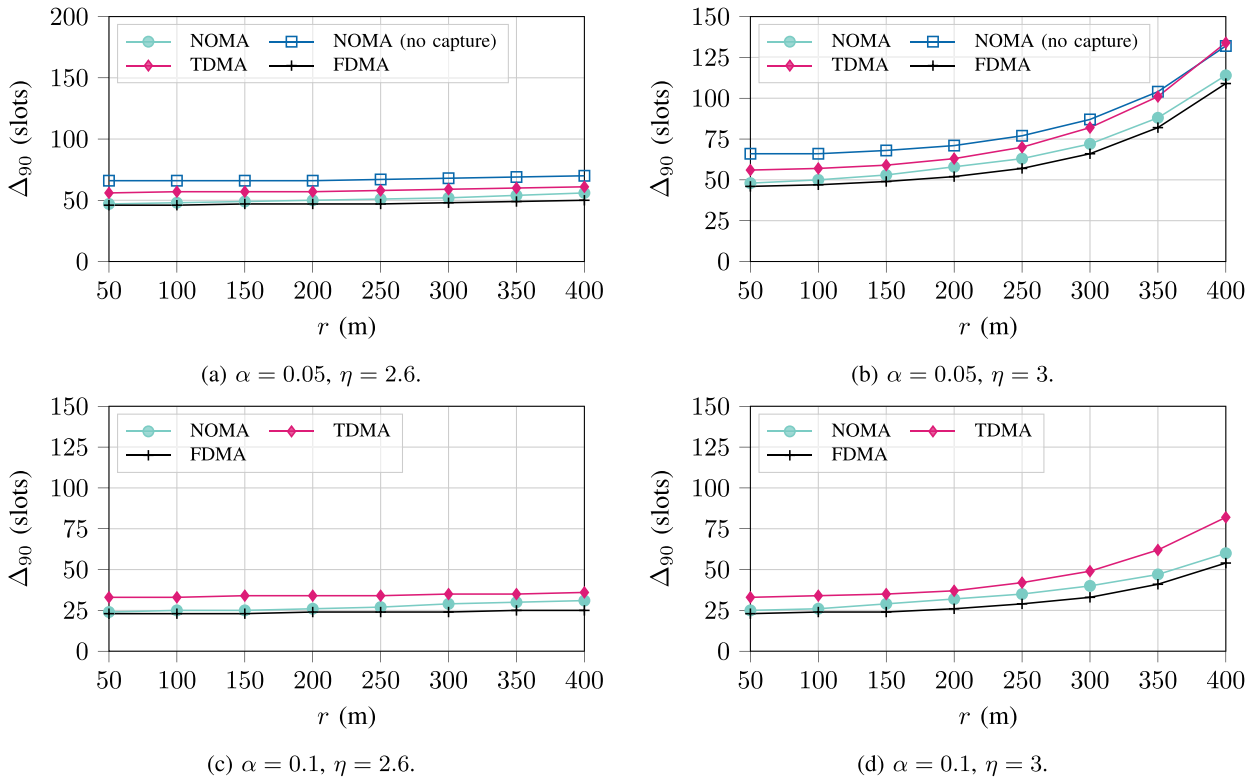
We conclude by investigating the optimal configurations for the schemes as a function of the distance of user 2, under the constraint $S \geq 0.7$. Fig. 9 shows the optimal values of K and N for NOMA and T_{int} for TDMA, for both LR-oriented and



(a) $\alpha = 0.01$, $\eta = 2.6$. (b) $\alpha = 0.01$, $\eta = 3$.

(c) $\alpha = 0.05$, $\eta = 2.6$. (d) $\alpha = 0.05$, $\eta = 3$.

FIGURE 7. Minimum LR with $S \geq 0.7$ as a function of the distance between user 2 and the BS for different values of α.

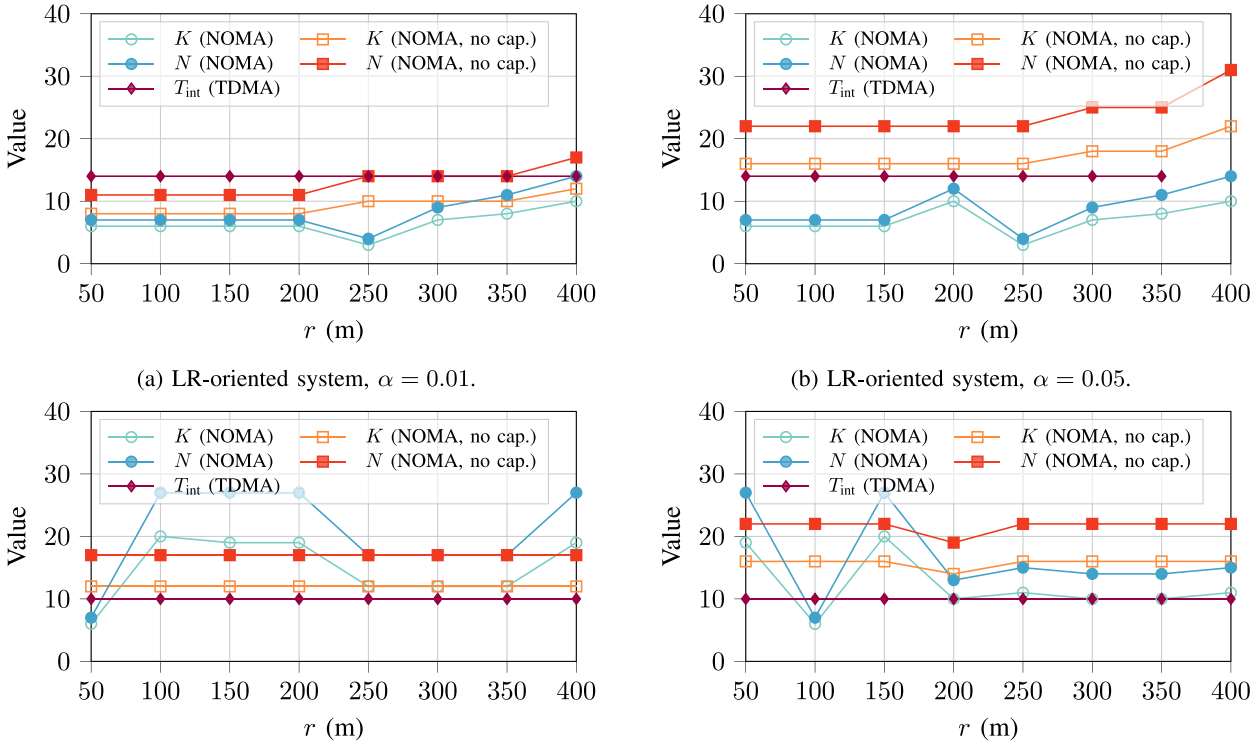


(a) $\alpha = 0.05$, $\eta = 2.6$. (b) $\alpha = 0.05$, $\eta = 3$.

(c) $\alpha = 0.1$, $\eta = 2.6$. (d) $\alpha = 0.1$, $\eta = 3$.

FIGURE 8. Minimum PAoI with $S \geq 0.7$ as a function of the distance between user 2 and the BS for different values of α.

PAoI-oriented systems with $\eta = 2.6$. The optimal values of K and N were computed by direct enumeration: we computed the values of S, T_{90}, and Δ_{90} for all configurations with $N \in \{2, \ldots, 32\}$ and $K < N$, then found the combination of setting that resulted in the minimum value of T_{90} or Δ_{90} while having a throughput $S \geq 0.7$.



(a) LR-oriented system, $\alpha = 0.01$.

(b) LR-oriented system, $\alpha = 0.05$.

(c) PAoI-oriented system, $\alpha = 0.01$.

(d) PAoI-oriented system, $\alpha = 0.05$.

FIGURE 9. Optimal settings for the three schemes with $S \geq 0.7$ as a function of the distance between user 2, with $\eta = 2.6$.

Fig. 9(a)-(b), which represent LR-oriented systems, show that the value of T_{int} is always 14, independently from the distance. On the other hand, LR-oriented NOMA systems tend to slightly increase both K and N as the distance increases. This occurs because the capture probability decreases as the distance from user 2 to the BS increases. Increasing N and K then increases the robustness of the codes to errors in the transmission. This also implies that, when the capture probability is high, the NOMA system can significantly reduce the frame size, which reduces the latency, even for intermittent user packets that need to wait for SIC.

On the other hand, if PAoI is the main objective, Fig. 9(c)-(d) show a different picture: the value of K and N for NOMA without capture is almost constant, as is the value of T_{int} for TDMA, while the best possible values of K and N for NOMA are higher at some distances and lower for others. This phenomenon is likely due to the interplay between the different outcome probabilities and their effects on the PAoI.

VII. CONCLUSION

In this paper, we evaluated orthogonal and non-orthogonal slicing for heterogeneous services, namely, broadband and intermittent, in the uplink RAN. Our model considered power control and packet-level coding for the broadband user and the use of SIC at the BS. Our analyses and results highlighted the achievable performance of TDMA and NOMA schemes when compared to a benchmark FDMA scheme utilizing double the bandwidth. In addition, we observed stark differences in terms of achievable trade-offs, impact of the inter-arrival times, and optimal configuration of the access schemes between the cases where the intermittent user aims to minimize either LR and PAoI. Hence, our results highlight the importance of the considered performance indicator for the intermittent user and of its wireless conditions, which must be taken into account for an efficient user pairing in NOMA.

In particular, our results showed that, with the considered schemes, the difference between NOMA and FDMA is negligible if the intermittent user has a sufficiently high mean SINR as a result of a relatively low path loss. Since NOMA utilized half of the resources of FDMA (which represents an upper bound for achievable performance with a single bandwidth part), it achieved the best balance between resource efficiency and performance when the intermittent user aims to minimize LR. Furthermore, even NOMA without capture achieved a better performance than TDMA in terms of LR.

Furthermore, NOMA achieved better trade-offs between throughput and LR than TDMA in every studied scenario. In particular, TDMA only showed a superior performance when aiming for the highest throughput possible in exchange for a longer PAoI. However, the differences in PAoI were considerably smaller than those for the LR cases, especially for short distances from user 2 to the BS. Hence, TDMA

may be preferred in the cases where the intermittent user is close to the BS due to its simplicity.

Finally, is it important to note that, since the slicing is performed independently for each bandwidth part, our model and analyses can be easily extended to the case with multiple users and multiple bandwidth parts. This is the case with multiple broadband users, each with its own bandwidth part that can be shared with up to one of the intermittent users. Further, the FDMA scheme could be used to allocate multiple intermittent users in the same bandwidth part. However, the complexity of this scenario, which would necessarily consider both the access model among the multiple intermittent users and the effects of concurrent transmissions by different groups of users, with many different possible outcomes in terms of decoded, retrievable, and erased packets, makes the analysis extremely cumbersome.

APPENDIX A
PROBABILITIES FOR THE DIFFERENT OUTCOMES WITH OVERLAPPING SIGNALS

Herein, we provide the closed-form expressions for the probabilities of the possible outcomes when both signals overlap under Rayleigh fading and after intra-slot SIC is performed. These outcomes were described in Section III-A. The probability of outcome (o_1, o_2) is denoted as π_{o1o2}.

$$\begin{aligned}\pi_{\mathcal{II}} &= \Pr\left[\frac{\mathrm{SNR}_{1,t}}{1+\mathrm{SNR}_{2,t}} \geq \gamma_1 \ \wedge \ \mathrm{SNR}_{2,t} \geq \gamma_2\right] \\ &\quad + \Pr\left[\frac{\mathrm{SNR}_{2,t}}{1+\mathrm{SNR}_{1,t}} \geq \gamma_2 \ \wedge \ \mathrm{SNR}_{1,t} \geq \gamma_1\right] \\ &\quad - \Pr\left[\frac{\mathrm{SNR}_{2,t}}{1+\mathrm{SNR}_{1,t}} \geq \gamma_2 \ \wedge \ \frac{\mathrm{SNR}_{1,t}}{1+\mathrm{SNR}_{2,t}} \geq \gamma_1\right] \\ &= \overline{\mathrm{SNR}_2} e^{\frac{-\gamma_1}{\overline{\mathrm{SNR}_1}}} \left(\frac{e^{\frac{-\gamma_1(1+\gamma_1)}{\overline{\mathrm{SNR}_2}(1-\gamma_1\gamma_2)}}}{\overline{\mathrm{SNR}_2}+\gamma_1\overline{\mathrm{SNR}_1}} - \frac{e^{\frac{-\gamma_2(1+\gamma_1)}{\overline{\mathrm{SNR}_2}(1-\gamma_1\gamma_2)}}}{\overline{\mathrm{SNR}_2}+\gamma_2\overline{\mathrm{SNR}_1}}\right) \\ &\quad + \frac{\overline{\mathrm{SNR}_2}}{\gamma_2\overline{\mathrm{SNR}_1}+\overline{\mathrm{SNR}_2}} e^{\frac{-\gamma_2}{\overline{\mathrm{SNR}_2}}} e^{-\gamma_1\left(\frac{1}{\overline{\mathrm{SNR}_1}}+\frac{\gamma_2}{\overline{\mathrm{SNR}_2}}\right)} \\ &\quad + \frac{\overline{\mathrm{SNR}_1}}{\overline{\mathrm{SNR}_1}+\gamma_1\overline{\mathrm{SNR}_2}} e^{\frac{-\gamma_1}{\overline{\mathrm{SNR}_1}}} e^{-\gamma_2\left(\frac{\gamma_1}{\overline{\mathrm{SNR}_1}}+\frac{1}{\overline{\mathrm{SNR}_2}}\right)} \qquad (68)\end{aligned}$$

If $\gamma_1 > 1$ and $\gamma_2 > 1$, the two events in which each of the two packets is decodable before SIC are mutually exclusive, as the two users cannot both have SINRs higher than 1. In this case, we can simplify the calculation by removing the third term, which represents their intersection.

$$\begin{aligned}\pi_{\mathcal{IE}} &= \Pr\left[\frac{\mathrm{SNR}_{1,t}}{1+\mathrm{SNR}_{2,t}} \geq \gamma_1 \ \wedge \ \mathrm{SNR}_{2,t} < \gamma_2\right] \\ &= \frac{\overline{\mathrm{SNR}_1} e^{\frac{-\gamma_1}{\overline{\mathrm{SNR}_1}}}}{\overline{\mathrm{SNR}_1}+\gamma_1\overline{\mathrm{SNR}_2}} \left(1 - e^{-\gamma_2\left(\frac{\gamma_1}{\overline{\mathrm{SNR}_1}}+\frac{1}{\overline{\mathrm{SNR}_2}}\right)}\right). \qquad (69)\end{aligned}$$

$$\begin{aligned}\pi_{\mathcal{EI}} &= \Pr\left[\frac{\mathrm{SNR}_{2,t}}{1+\mathrm{SNR}_{1,t}} \geq \gamma_2 \ \wedge \ \mathrm{SNR}_{1,t} < \gamma_1\right] \\ &= \frac{\overline{\mathrm{SNR}_2} e^{\frac{-\gamma_2}{\overline{\mathrm{SNR}_2}}}}{\gamma_2\overline{\mathrm{SNR}_1}+\overline{\mathrm{SNR}_2}} \left(1 - e^{-\gamma_1\left(\frac{1}{\overline{\mathrm{SNR}_1}}+\frac{\gamma_2}{\overline{\mathrm{SNR}_2}}\right)}\right). \qquad (70)\end{aligned}$$

$$\begin{aligned}\pi_{\mathcal{EE}} &= \Pr\left[\mathrm{SNR}_{2,t} < \gamma_2 \ \wedge \ \mathrm{SNR}_{1,t} < \gamma_1\right] \\ &= \left(1 - e^{\frac{-\gamma_1}{\overline{\mathrm{SNR}_1}}}\right)\left(1 - e^{\frac{-\gamma_2}{\overline{\mathrm{SNR}_2}}}\right). \qquad (71)\end{aligned}$$

$$\begin{aligned}\pi_{\mathcal{RE}} &= \Pr\left[\gamma_1 \leq \mathrm{SNR}_{1,t} < \gamma_1(1+\mathrm{SNR}_{2,t}) \right. \\ &\qquad \left. \wedge \ \mathrm{SNR}_{2,t} < \gamma_2\right] \\ &= \left(\frac{\overline{\mathrm{SNR}_1}}{\overline{\mathrm{SNR}_1}+\gamma_1\overline{\mathrm{SNR}_2}} \left(e^{-\gamma_2\left(\frac{\gamma_1}{\overline{\mathrm{SNR}_1}}+\frac{1}{\overline{\mathrm{SNR}_2}}\right)} - 1\right)\right. \\ &\qquad \left. + \left(1 - e^{\frac{-\gamma_2}{\overline{\mathrm{SNR}_2}}}\right)\right) e^{\frac{-\gamma_1}{\overline{\mathrm{SNR}_1}}} \qquad (72)\end{aligned}$$

$$\pi_{\cdot\mathcal{R}} = 1 - \pi_{\mathcal{II}} - \pi_{\mathcal{IE}} - \pi_{\mathcal{EI}} - \pi_{\mathcal{EE}} - \pi_{\mathcal{RE}}. \qquad (73)$$

APPENDIX B
BENCHMARK: PERFORMANCE WITH FDMA

In case of FDMA, each the two users are occupying a dedicated frequency band, and their KPIs are independent. The success probability for user 1 is equal to that in TDMA, given by (15). The throughput of user 1 can be computed by setting $T_{\mathrm{int}} \to \infty$ in (16), which gives:

$$S = \frac{K p_{s,1}}{N}. \qquad (74)$$

For user 2, the latency for all successfully decoded packets is 1. Further, $p_{s,2} = 1 - \varepsilon_2$ and the pmf of LR is simply $p_L(t) = \delta(t-1)$.

The PAoI for user 2 can be obtained as the latency $T = 1$ plus the inter-decoding time Z when setting $T_{\mathrm{int}} = 1$ in (28) and (29). Hence, it is simply a function of the inter-arrival time and ε_2. Namely,

$$p_\Delta(t) = (1 - \alpha(1-\varepsilon_2))^{t-2}\alpha(1-\varepsilon_2), \ \ t \geq 2. \qquad (75)$$

REFERENCES

[1] "5G: Study on scenarios and requirements for next generation access technologies," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 38.913 V16.0.0, Jul. 2020.

[2] "Release 15 description," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 21.915 V15.0.0, Sep. 2019.

[3] "Release 16 description," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 21.916 V0.6.0, Sep. 2020.

[4] P. Rost *et al.*, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, May 2017.

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

[6] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Netw.*, vol. 31, no. 4, pp. 8–14, Jul./Aug. 2017.

[7] S. M. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

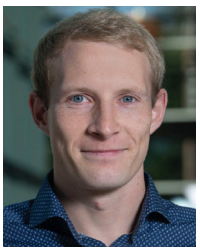
[8] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.

[9] A. Maatouk, M. Assaad, and A. Ephremides, "Minimizing the age of information: NOMA or OMA?" in *Proc. IEEE INFOCOM Workshops*, vol. 65, 2019, pp. 102–108.
[10] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.
[11] Z. Wu, K. Lu, C. Jiang, and X. Shao, "Comprehensive study and comparison on 5G NOMA schemes," *IEEE Access*, vol. 6, pp. 18511–18519, 2018.
[12] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 3, pp. 462–476, Sep. 2016.
[13] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, no. 8, pp. 55765–55779, 2018.
[14] M. Kamel, W. Hamouda, and A. Youssef, "Uplink performance of NOMA-based combined HTC and MTC in ultradense networks," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7319–7333, Aug. 2020.
[15] G. Sreya, S. Saigadha, P. D. Mankar, G. Das, and H. S. Dhillon, "Adaptive rate NOMA for cellular IoT networks," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 478–482, Mar. 2022.
[16] F. Chiariotti, I. Leyva-Mayorga, Č. Stefanović, A. E. Kalør, and P. Popovski, "Spectrum slicing for multiple access channels with heterogeneous services," *Entropy*, vol. 23, no. 6, p. 686, May 2021.
[17] I. Leyva-Mayorga, F. Chiariotti, Č. Stefanović, A. E. Kalør, and P. Popovski, "Slicing a single wireless collision channel among throughput-and timeliness-sensitive services," in *Proc. IEEE Int. Commun. Conf. (ICC)*, Jun. 2021, pp. 1–6.
[18] M. Costa, M. Codreanu, and A. Ephremides, "Age of information with packet management," in *Proc. ISIT*, 2014, pp. 1583–1587.
[19] J. J. Nielsen, R. Liu, and P. Popovski, "Ultra-reliable low latency communication using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322–1334, Mar. 2018.
[20] "Study on non-orthogonal multiple access (NOMA) for NR," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 38.812 V16.0.0, Dec. 2018.
[21] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.
[22] R. Kassab, O. Simeone, P. Popovski, and T. Islam, "Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures," *IEEE Access*, vol. 7, pp. 13035–13049, 2019.
[23] M. Elbayoumi, W. Hamouda, and A. Youssef, "A hybrid NOMA/OMA scheme for MTC in ultra-dense networks," in *Proc. Global Commun. Conf. (GLOBECOM)*, 2020, pp. 1–6.
[24] S. Kaul, R. D. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, 2012, pp. 2731–2735.
[25] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age of information performance of multiaccess strategies with packet management," *J. Commun. Netw.*, vol. 21, no. 3, pp. 244–255, Jun. 2019.
[26] A. Maatouk, M. Assaad, and A. Ephremides, "On the age of information in a CSMA environment," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 818–831, Apr. 2020.
[27] R. D. Yates and S. K. Kaul, "Age of Information in uncoordinated unslotted updating," Feb. 2020, *arXiv:2002.02026*.
[28] X. Chen, K. Gatsis, H. Hassani, and S. S. Bidokhti, "Age of information in random access channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1770–1775.
[29] R. D. Yates, "The age of information in networks: Moments, distributions, and sampling," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5712–5728, Sep. 2020.
[30] J. P. Champati, H. Al-Zubaidy, and J. Gross, "Statistical guarantee optimization for AoI in single-hop and two-hop FCFS systems with periodic arrivals," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 365–381, Jan. 2021.
[31] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 721–734, Apr. 2019.
[32] R. D. Yates, J. Zhong, and W. Zhang, "Updates with multiple service classes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1017–1021.
[33] "NR; Physical channels and modulation," 3GPP, Sophia Antipolis, France, 3GPP Rep. TS 38.211 V16.2.0, Mar. 2020.
[34] L. Kleinrock and R. Gail, *Queueing Systems: Problems and Solutions*. Hoboken, NJ, USA: Wiley, 1996.

FEDERICO CHIARIOTTI (Member, IEEE) received the bachelor's and master's degrees (*cum laude*) in telecommunication engineering and the Ph.D. degree in information engineering from the University of Padova, Italy, in 2013, 2015, and 2019, respectively. He is currently an Assistant Professor with the Department of Electronic Systems, Aalborg University, Denmark. He has authored over 50 peer-reviewed papers on wireless networks and the use of artificial intelligence techniques to improve their performance. His current research interests include semantic communications, protocol design, age of information, bike sharing system optimization, and adaptive video streaming. He was a recipient of the Best Paper Award at several conferences, including the IEEE INFOCOM 2020 WCNEE Workshop.

ISRAEL LEYVA-MAYORGA (Member, IEEE) received the B.Sc. degree in telematics engineering and the M.Sc. degree (Hons.) in mobile computing systems from the Instituto Politecnico Nacional (IPN), Mexico, in 2012 and 2014, respectively, and the Ph.D. degree (*cum laude* and extraordinary prize) in telecommunications from the Universitat Politecnica de Valencia, Spain, in 2018, where he was a Visiting Researcher with the Department of Communications in 2014 and with the Deutsche Telekom Chair of Communication Networks, Technische Universitat Dresden, Germany, in 2018. He is currently an Assistant Professor with the Connectivity Section (CNT), Department of Electronic Systems, Aalborg University (AAU), Denmark, where he served as a Postdoctoral Researcher from January 2019 to July 2021. His research interests include beyond-5G, and 6G networks, satellite communications, and random and multiple access protocols. He is an Associate Editor of IEEE WIRELESS COMMUNICATIONS LETTERS, a Board Member of one6G, and a representative of AAU in 6G IA SNS.

ČEDOMIR STEFANOVIĆ (Senior Member, IEEE) received the Diploma Ing., Mr.-Ing., and Ph.D. degrees from the University of Novi Sad, Serbia. He is currently a Professor with the Department of Electronic Systems, Aalborg University, where he leads Edge Computing and Networking Group. He is a Principal Researcher on a number of European projects related to IoT, 5G, and mission-critical communications. He has coauthored more than 100 peer-reviewed publications. His research interests include communication theory and wireless communications. He serves as an Editor for the IEEE INTERNET OF THINGS JOURNAL.



ANDERS E. KALØR (Graduate Student Member, IEEE) received the B.Sc. degree in computer engineering and the M.Sc. degree in networks and distributed systems from Aalborg University, Denmark, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree in wireless communications and networking with Connectivity Section. In 2017, he was a visiting student with Robert Bosch, Germany, and with King's College London, U.K., in 2020. His research interests include communication theory, MAC layer design for wireless systems, and networking.

PETAR POPOVSKI (Fellow, IEEE) received the Dipl.-Ing and M.Sc. degrees in communication engineering from the University of Sts. Cyril and Methodius in Skopje and the Ph.D. degree from Aalborg University in 2005. He is a Professor with Aalborg University, where he heads the section on Connectivity and a Visiting Excellence Chair with the University of Bremen. He authored the book *Wireless Connectivity: An Intuitive and Fundamental Guide* (Wiley, 2020). His research interests are in the area of wireless communication and communication theory. He received the ERC Consolidator Grant in 2015, the Danish Elite Researcher Award in 2016, the IEEE Fred W. Ellersick Prize in 2016, the IEEE Stephen O. Rice Prize in 2018, the Technical Achievement Award from the IEEE Technical Committee on Smart Grid Communications in 2019, the Danish Telecommunication Prize in 2020, and the Villum Investigator Grant in 2021. He was a Member-at-Large at the Board of Governors in IEEE Communication Society from 2019 to 2021. He is currently the Editor-in-Chief of IEEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He also serves as the Vice-Chair for the IEEE Communication Theory Technical Committee and the Steering Committee for IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. He was the General Chair of IEEE SmartGridComm 2018 and IEEE Communication Theory Workshop 2019.