Slicing a single wireless collision channel among throughput- and timeliness-sensitive services

Israel Leyva-Mayorga, Federico Chiariotti, Čedomir Stefanović, Anders E. Kalør, and Petar Popovski
Department of Electronic Systems, Aalborg University, 9220, Aalborg, Denmark
email: {ilm,fchi,cs,aek,petarp}@es.aau.dk

Abstract*—The fifth generation (5G) of wireless systems has a platform-driven approach, aiming to support heterogeneous connections with very diverse requirements. The shared wireless resources should be sliced in a way that each user perceives that its requirements have been met. Heterogeneity challenges the traditional notion of resource efficiency, as the resource usage has to cater for, e.g., rate maximization for one user and a *timeliness* requirement for another user. This paper treats a model for radio access network (RAN) uplink, where a throughput-demanding broadband user shares wireless resources with an intermittently active user that wants to optimize the timeliness, expressed in terms of latency-reliability or Age of Information (AoI). We evaluate the trade-offs between throughput and timeliness for Orthogonal Multiple Access (OMA) as well as Non-Orthogonal Multiple Access (NOMA) with successive interference cancellation (SIC). We observe that NOMA with SIC, in a conservative scenario with destructive collisions, is just slightly inferior to that of OMA, which indicates that it may offer significant benefits in practical deployments where the capture effect is frequently encountered. On the other hand, finding the optimal configuration of NOMA with SIC depends on the activity pattern of the intermittent user, to which OMA is insensitive.

***Index Terms*—Age of Information (AoI), heterogeneous services, network slicing, Non-Orthogonal Multiple Access (NOMA).**

I. Introduction

Perhaps the main innovation in 5G is its platform approach to connectivity, aiming to support any type of connection through a suitable combination of three generic connection types with vastly different requirements: enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC) [1]. While eMBB aims at delivering enhanced data-rates in comparison to 4G, URLLC usually involves the exchange of small amounts of data at a very low latency (say, 1 ms) and very high reliability (e.g. packet loss rate 10^{-5}). Finally, mMTC aims at supporting the sporadic transmission of small data chunks, but from a vast number (e.g., thousands) of devices in a single cell.

Network slicing refers to the allocation of the network resources among the active services, such that the network is able to provide performance guarantees and meet their widely different requirements [2]. When related to the slicing of the shared wireless resource in a radio access network (RAN) context, we term it wireless slicing. It has been studied in the form of diverse Orthogonal Multiple Access (OMA) and Non-Orthogonal Multiple Access (NOMA) techniques in the presence of multiple users with the same service type [3]–[5], and the trade-offs in achievable data rates for eMBB services are clearly characterized [4], [5]. However, further research is needed on novel slicing mechanisms for heterogeneous services to support the widely diverse traffic patterns and requirements of URLLC, mMTC, and eMBB [6]. In this respect, a common uplink scenario involves two categories of users: broadband and intermittent users. Broadband users transmit data continuously, aiming to maximize the throughput. In contrast, intermittent users transmit short packets sporadically, being mainly interested in the *timeliness* of their data. URLLC and mMTC users fall in the latter category, but the URLLC timeliness requirements are more stringent than for mMTC.

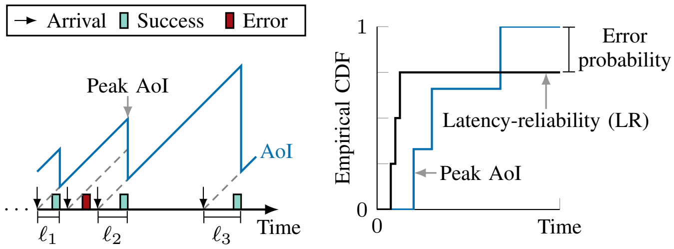

Fig. 1: Exemplary diagram of the AoI and LR KPIs in a period with four packet transmissions. The latency for packets transmitted with errors is set to ∞.

For the intermittent users with timeliness requirements, there are two different Key Performance Indicators (KPIs) that may be relevant. The first is the Age of Information (AoI), which represents the time elapsed since the generation of the last received packet [7], and is useful when the intermittent user sends updates of an ongoing process. In this case, the information freshness is more important than any single update and, thus, packet loss is tolerable, while the update generation pattern is a key factor to ensure a low age [8]. In applications requiring high reliability, the high percentiles of the Peak AoI (PAoI), defined as the maximum AoI achieved immediately before the reception of an update [9], are the most relevant metric. The other KPI is a combination of latency and reliability, useful for URLLC-like applications that send critical data. We exploit a latency-reliability (LR) metric measuring the time between the generation and successful reception of each packet in the form of a Cumulative Distribution Function (CDF) [10], assuming that lost packets have infinite latency, and using high percentiles of the CDF as a KPI. Fig. 1 illustrates the difference between PAoI and LR during an observation period with four packet transmissions.

Non-orthogonal network slicing has been investigated in combination with successive interference cancellation (SIC) in uplink scenarios with heterogeneous services. For instance, [2] investigated the benefits of NOMA and OMA assuming that *(i)* eMBB users are allocated orthogonal resources among them, *(ii)* there is a single URLLC user, and *(iii)* mMTC traffic is Poisson distributed. It was observed that NOMA may offer benefits with respect to (w.r.t.) OMA depending on the rate of the eMBB users and whether the coexisting traffic is URLLC or mMTC. The work was later extended to a multi-cell scenario with strict latency guarantees for URLLC traffic [11]: a single URLLC user per cell was assumed and it was observed that NOMA leads to a greater spectral efficiency w.r.t. OMA. A similar conclusion was drawn by Maatouk *et al.* [3] in an uplink scenario where two intermittent users aim to minimize the AoI. Their results show that a greater spectral efficiency does not directly translate into a lower average AoI.

In this paper, we investigate the trade-offs and performance bounds with orthogonal and non-orthogonal slicing with SIC for the case where a single frequency band is sliced to accommodate one broadband and one intermittent user. To the best of our knowledge, this is the first paper to explore the inherent differences between slicing the radio access for LR- and for AoI-oriented services with additional broadband traffic. In particular, we provide closed-form expressions for the achievable throughput (given in packets per slot) for the broadband user and the timeliness for the intermittent user. Our analysis in this baseline scenario *(i)* illustrates the fundamental performance trade-offs between access methods; *(ii)* can be directly extended to multiple frequency bands and a greater number of users; and *(iii)* provides the basis to explore complex slicing approaches. The scenario assumed in the paper is a conservative one, based on a simple model in which collision are destructive – no involved packet can be directly decoded (e.g., through capture effect). Consequently, the results presented in the paper may be assumed to correspond to lower bounds on the performance of NOMA. We have observed that NOMA greatly outperforms OMA when the intermittent user requires LR, but is not able to maintain a message (i.e., update) queue. Additionally, NOMA can outperform OMA with queuing if the target is extremely low latency or extremely high throughput. On the other hand, the achievable PAoI with OMA is considerably lower than that with NOMA, as OMA is able to tolerate higher arrival rates with a higher reliability. These results highlight the need to adapt the system to the particular timeliness requirement posed by the intermittent user.

In the rest of the paper, Section II presents the general system model, while we derive the distribution of the KPIs for OMA and NOMA in Section III. Section IV discusses simulation results, while Section V concludes the paper and lists possible avenues for future work.

II. System model

We consider a general uplink scenario where multiple users transmit data to a base station (BS). A time-slotted and frequency-division multiple access channel is considered.

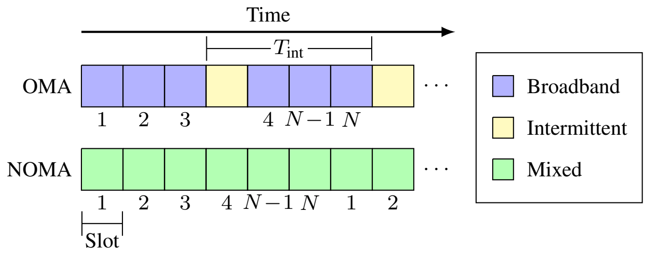

Fig. 2: Frame structure for OMA and NOMA; $K = 4$, $N = 6$.

Our focus is on *one* of the many possible frequency bands (subcarriers) where a broadband and an intermittent user are allocated. The broadband user falls under the full-buffer traffic model, where messages of length of K packets are encoded into N linearly independent coded packets using a packet-level coding scheme. This can be achieved with, e.g., Maximum Distance Separable (MDS) codes and infinite queue of the broadband user. Changing the coding rate K/N can trade off rate for robustness against packet erasures, caused by noise errors, but also collisions with the packets from an intermittent user. The intermittent user generates new messages at each time slot with probability α. Each message fits into a single packet, which the intermittent users transmits just once at the next available slot (without application of a packet-level code). The intermittent user can maintain a queue of up to Q of the generated packets. If a new packet arrives when the queue is full, the oldest packet is discarded and the new packet is kept.

The BS allocates resources to the users. We define $\mathcal{U} = \{1, 2\}$, where user 1 is the broadband user and user 2 is the intermittent user, and assume three different allocations of a particular time slot: 1) Broadband, allocated to user 1; 2) Intermittent, allocated to user 2; and 3) Mixed, allocated to both users. We define the following two access methods. *(1) OMA* Only broadband and intermittent resources are allocated. T_{int} is the period between intermittent resources. The broadband user is not affected by the activity of the intermittent user. *(2) NOMA*: Only mixed resources are allocated. Hence, the frame structure is exclusively determined by K and N. Here the rate and reliability of the broadband user are dependent on the level of activity of the intermittent user. The frame structures for these methods are illustrated in Fig. 2.

Finally, we consider a binary erasure channel (BEC) with collisions (in NOMA case) and constant transmission power at each user. The erasure probabilities for the broadband and intermittent user are denoted as ϵ_1 and ϵ_2, respectively; these depend on the received signal-to-noise ratio (SNR) and the receiver sensitivity. The considered collision model is simple, where the signal-to-interference-plus-noise ratio (SINR) is below the receiver sensitivity every time that users 1 and 2 transmit at the same resource. That is, a packet may only be immediately decoded at a particular slot if only one user transmits. The BS stores collisions in a frame and may able to decode transmissions from the intermittent user *after* decoding the block of K packets from the broadband user. For this, implements SIC to remove the interference from the broadband user. Then, the packet from the intermittent user is decoded (at the expense of increased latency) if the SNR is sufficient.

III. Performance analysis

In this section, we derive the KPIs for the OMA and NOMA systems, for a LR- or PAoI-oriented intermittent user with activation probability α. In the following, we define the multinomial function $\text{Mult}(\mathbf{K}; N, \mathbf{p})$ as

$$\text{Mult}(\mathbf{K}; N, \mathbf{p}) = \frac{N! \prod_{i=1}^{|\mathbf{p}|} p_i^{K_i} (1 - \sum_{i=1}^{|\mathbf{p}|} p_i)^{N - \sum_{i=1}^{|\mathbf{p}|} K_i}}{(N - \sum_{i=1}^{|\mathbf{p}|} K_i)! \prod_{i=1}^{|\mathbf{p}|} K_i!}, \quad (1)$$

where $|\mathbf{p}|$ is the length of vectors $\mathbf{p} = \left[p_1, p_2, \dots, p_{|\mathbf{p}|}\right]$ and $\mathbf{K} = \left[K_1, K_2, \dots, K_{|\mathbf{p}|}\right]$. The binomial function $\text{Bin}(K; N, p)$ is the special case with $|\mathbf{p}| = 1$.

A. LR-oriented OMA

In OMA the intermittent user has a reserved slot every T_{int}. Denoting the decoding success probability of the broadband user as $p_{s,1}$, the expected throughput S_1 is

$$S_1 = p_{s,1} \frac{(T_{\text{int}} - 1)K}{T_{\text{int}} N} \quad (2)$$

$$p_{s,1} = \sum_{r=K}^{N} \text{Bin}(r : N, 1 - \epsilon_1). \quad (3)$$

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 will reduce the throughput.

To derive the latency probability mass function (pmf) for the intermittent user, we must compute the state-transition probabilities $\mathbf{P}^{(0)} = \left[P_{ij}^{(0)}\right]$ of its queue, given as

$$P_{ij}^{(0)} = \begin{cases} \text{Bin}(q_j - q_i + 1; T_{\text{int}}, \alpha) & \text{if } q_i \leq q_j + 1 < Q; \\ \sum_{m=Q-q_i}^{T_{\text{int}}} \text{Bin}(m; T_{\text{int}}, \alpha) & \text{if } q_j = Q - 1, \end{cases} \quad (4)$$

where $q_i, q_j \in \{0, 1, \dots, Q\}$ are the queue states before and after a transmission, respectively. The steady-state distribution of the queue immediately after a transmission $\boldsymbol{\pi}^{(0)} = \left[\pi_0^{(0)}, \pi_1^{(0)}, \dots, \pi_Q^{(0)}\right]$ is the left-eigenvector of $P^{(q)}$ with eigenvalue 1, normalized to sum to 1 to be a valid probability measure. That is,

$$\boldsymbol{\pi}^{(0)}(\mathbf{I} - \mathbf{P}^{(0)}) = 0; \qquad \text{s.t.} \sum_{q=0}^{Q} \pi_q^{(0)} = 1. \quad (5)$$

We calculate the steady-state distribution of the queue occupation n slots after an intermittent user transmission as

$$\pi_q^{(n)} = \begin{cases} \sum_{s=0}^{q} \pi_s^{(0)} \text{Bin}(q - s; n\alpha) & \text{if } q < Q; \\ \sum_{s=0}^{Q} \sum_{m=Q-s}^{n} \pi_s^{(0)} \text{Bin}(m; n, \alpha) & \text{if } q = Q. \end{cases} \quad (6)$$

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 – if the queue is full, the oldest packet (i.e., the first in the queue) is dropped. Let $\mathcal{G}_\ell^{(n)} = \{0, \dots, T_{\text{int}} - n\} \times \{0, \dots, T_{\text{int}}\}^{\ell-1}$ be the set containing the possible numbers of new packets generated by the intermittent user in the next ℓ transmission windows of length T_{int} after the considered packet is generated in the n slots after the last intermittent slot. The probability of each element $\mathbf{g} \in \mathcal{G}_\ell^{(n)}$ is simply:

$$p_{\mathcal{G}_\ell^{(n)}}(\mathbf{g}) = \text{Bin}(g_1; T_{\text{int}} - n, \alpha) \prod_{i=1}^{\ell} \text{Bin}(g_i; T_{\text{int}}, \alpha). \quad (7)$$

At each transmission opportunity, one packet is transmitted, and other packets are dropped if the number of generated packets exceeds the number of remaining places in the queue. For a given generation vector $\mathbf{g} \in \mathcal{G}_\ell^{(n)}$, the considered packet is transmitted at the ℓ-th transmission opportunity, where ℓ is the first index that satisfies

$$\psi_k^{(\mathbf{g},q)} = \delta \left(\sum_{i=1}^{k} \left[q + 1 - Q + \sum_{j=1}^{i} g_j \right]^+ + k - (q+1) \right), \quad (8)$$

where $\delta(x)$ is the delta function, equal to 1 if $x = 0$ and 0 otherwise, and $[x]^+ = \max(x, 0)$. We now define the set

$$\mathcal{S}_\ell^{(n,q)} = \left\{ \mathbf{g} \in \mathcal{G}_\ell^{(n)} : \psi_\ell^{(\mathbf{g},q)} - \sum_{k=1}^{\ell-1} \psi_k^{(\mathbf{g},q)} = 1 \right\}, \quad (9)$$

with the elements $\mathbf{g} \in \mathcal{G}_\ell^{(n)}$ for which the considered packet is transmitted at the ℓ-th opportunity. Since the packet is either transmitted within $q+1$ transmission attempts or discarded, the success probability for the intermittent user is given by

$$p_{s,2} = \sum_{n=1}^{T_{\text{int}}} \sum_{q=0}^{Q} \sum_{\ell=1}^{q+1} \pi_q^{(n-1)} \sum_{\mathbf{g} \in \mathcal{S}_\ell^{(n,q)}} p_{\mathcal{G}_\ell^{(n)}}(\mathbf{g}) \frac{(1 - \epsilon_2)}{T_{\text{int}}}. \quad (10)$$

Knowing that packet generation probability is the same for every slot, we now use (6) to uncondition the latency pmf

$$p_T(\ell T_{\text{int}} - n) = \sum_{n=1}^{T_{\text{int}}} \sum_{q=0}^{Q} \frac{\pi_q^{(n-1)} \sum_{\mathbf{g} \in \mathcal{S}_\ell^{(n, \min(q, Q-1))}} p_{\mathcal{G}_\ell^{(n)}}(\mathbf{g})}{T_{\text{int}} p_{s,2}(n, \min(q, Q-1))}. \quad (11)$$

B. PAoI-oriented OMA

Preemption is the optimal strategy to minimize PAoI. Therefore, we set $Q = 1$ for the PAoI-oriented OMA and packets are always sent at the first available transmission opportunity. Hence, the pmf of the delay of a successful transmission is

$$p_T(t) = \frac{\alpha(1 - \alpha)^t}{1 - (1 - \alpha)^{T_{\text{int}}}}. \quad (12)$$

We now compute the PAoI for the OMA scheme, which is given by the sum of the transmission delay and the inter-transmission time, denoted by Z. Since exactly one slot every T_{int} is reserved for user 2, Z is equal to T_{int} times the number of reserved slots between consecutive transmissions. The latter is a geometric random variable, whose parameter $\xi = (1 - (1 - \alpha)^{T_{\text{int}}})(1 - \epsilon_2)$ is the probability of a successful transmission. Hence, we obtain the pmf of the PAoI Δ as

$$p_\Delta(T_{\text{int}} z + t) = (1 - \xi)^{z-1} \xi p_T(t). \quad (13)$$

C. LR-oriented NOMA

In the NOMA case all slots are mixed and the queue size is irrelevant; any intermittent packet is immediately transmitted and queued at the receiver until SIC is performed. Hence, we set $Q = 1$ and define $p_1 = 1-\epsilon_1$ and $p_2 = \alpha(1-\epsilon_2)$; the latter is the probability that, at each slot after SIC, an intermittent packet is transmitted and received correctly.

As a starting point, we derive the success probability and throughput of user 1, respectively, as

$$p_{s,1} = \sum_{r_2=0}^{N-K} \text{Bin}(r_2; N, \alpha) \sum_{r_1=K}^{N-r_2} \text{Bin}(r_1; N-r_2, p_1), \quad (14)$$

$$S_1 = \frac{K}{N} p_{s,1}. \quad (15)$$

As intermittent packets are transmitted immediately, we simply need to compute the delay T between the slot in which the packet is transmitted and when it is decoded. First, we define F to be the random variable of the slot within the frame where the *first decoding event* (for the intermittent user) occurs. As a result of the simple collision model, the first decoding event can only occur after the broadband user is decoded. Therefore, the pmf of F, whose support is $\{K+1, \dots, N\}$, is given as

$$p_F(f) = \sum_{r_2=1}^{f-K} \sum_{r_1=K}^{f-r_2} \text{Mult}((r_1, r_2); f, (p_1, p_2)) - \sum_{i=1}^{f-1} p_F(i). \quad (16)$$

It is easy to compute the CDF of F, denoted $P_F(f)$. The probability of having at least one successful decoding in a frame is then $P_F(N)$, so we get $p_{s,2} = P_F(N)$. The probability of receiving R more packets from user 2 in a frame, conditioned on the first decoding event F, is

$$p_R(r; f) = \frac{\text{Bin}(k; N, p_2)}{P_F(N)}. \quad (17)$$

We compute the probability of having a decoding event D in a given slot, conditioned on F, as

$$p_D(d; f) = \begin{cases} \sum_{r=0}^{N-d} p_R(r; d) & \text{if } f = d; \\ \sum_{r=0}^{N-f} \frac{p_R(r;f)r}{(N-f)} & \text{if } f > d. \end{cases} \quad (18)$$

Combining (16) and (18) yields

$$p_D(d) = \sum_{f=K+1}^{d} p_F(f) p_D(d; f). \quad (19)$$

In order to compute the pmf of the delay T, we consider three cases. First, the delay is always 0 if the decoding event is not the first in the slot, as SIC can be performed immediately. Second, the first decoding event has delay 0 if the broadband user frame was already decoded before slot d, but no packets from user 2 were successfully received before d. Hence,

$$p_{T|F}(0; d) = \sum_{r_1=K}^{d-1} \frac{\text{Mult}((r_1, 0); d-1, ((1-\alpha)p_1, p_2))p_2}{p_F(d)}. \quad (20)$$

Third, the delay for the first decoding event is larger than 0 if there were intermittent packets before slot d, but the broadband user frame was not decoded yet. In this latter case, we also need to consider that multiple packets might be decoded in the first decoding event. Therefore, we compute the conditional pmf of the delay of the last packet ℓ given that c packets are decoded in the first decoding event as

$$p_{T_\ell|F,C}(t; d, c) = \frac{c(d-t)!(d-c)!}{(d-1)!(d-t-c)!}. \quad (21)$$

Let T_p be the the delay of the previous packets, whose pmf is

$$p_{T_p|F,C,T_\ell}(t; d, c, t_\ell) = \frac{c-1}{d-(t_\ell+1)}, \; t \in \{t_\ell+1, \dots, d\}. \quad (22)$$

The overall delay pmf for a given number of simultaneous decoded packets c is

$$p_{T|F,C}(t; d, c) = p_{T_\ell|F,C}(t; d, c) + \sum_{t_\ell=0}^{t-1} p_{T_p|F,C,T_\ell}(t; d, c, t_\ell). \quad (23)$$

Further, we calculate the probability of decoding C packets simultaneously if the first decoding event is in slot d as

$$p_C(c; d) = \frac{\text{Mult}((K-1, c); d-1, (p_1, p_2))p_1}{p_F(d)}. \quad (24)$$

The latter allows us to uncondition (23) on C and compute the overall delay for the first decoding event in the reliable transmission scenario as

$$p_{T|F}(t; d) = \sum_{c=1}^{d-K} p(T|F, C)(t; d, c) p_C(c; d). \quad (25)$$

Finally, the probability that a received packet is part of the first decoding event is given by

$$p_{F|D}(d) = \sum_{c=1}^{d-K} \sum_{r=0}^{N-d} \frac{p_F(d) p_C(c; d) p_R(r; d) c}{(c+r) p_D(d) P_F(N)}, \quad (26)$$

and the overall delay pmf is

$$p_T(t) = \begin{cases} \sum\limits_{d=K+1}^{N} p_D(d) p_{F|D}(d) p_{T|F}(t; d) & t > 0; \\ \sum\limits_{d=K+1}^{N} p_D(d)(1 - p_{F|D}(d)(1 - p_{T|F}(0; d))) & t = 0. \end{cases} \quad (27)$$

D. PAoI-oriented NOMA

Most of the analysis from the LR-oriented case is valid for the PAoI-oriented case, the only difference being that the only meaningful packet in a decoding event is the most recent one:

$$p_{T|F,C}(t; d, c) = p_{T_\ell|F,C}(t; d, c), \quad (28)$$

where $p_{T_\ell|F,C}(t; d, c)$ is the same as in (21). This affects the probability of a packet being in the first decoding event in the frame, as computation no longer considers previous packets:

$$p_{F|D}(d) = \sum_{r=0}^{N-d} \frac{p_F(d) p_R(r; d)}{(r+1) p_D(d) P_F(N)}. \quad (29)$$

We compute the conditioned delay pmf $p_T(t;d)$ by using the values in (27). First, we compute the probability that a decoding event is the last in the frame as

$$p_L(d)=\sum_{f=K+1}^{d} p_F(f) \sum_{r=1}^{d-f} \frac{p_D(d;f)(d-f-1)!(N-f-r+1)!}{(d-f-r+1)!(N-f-1)!p_D(d)}, \quad (30)$$

Next, let $\bar{L}$ indicate that the decoding event is not the last in the frame, i.e., that there are more decoding events in the frame after slot d. Building on this, the pmf of Z when the decoding event is not the last in the frame

$$p_Z(z;d,\bar{L}) = \frac{(1-\alpha(1-\epsilon_2)^{z-1}\alpha(1-\epsilon_2)}{1-(1-\alpha(1-\epsilon_2)^{N-d}} \quad (31)$$

If the decoding event is the last in the frame, the inter-transmission time Z is formed by three components: the $N-d$ slots until the end of the current frame, e full frames, and the time F until the first decoding event from the beginning of the frame. Hence, we derive the pmf of Z in this case as

$$p_Z((e-1)N+d+f;d,L) = (1-P_N(F))^e p_F(f). \quad (32)$$

Using (31) and (32) to uncondition on L, we get

$$p_Z(z;d) = \begin{cases} p_Z(z;d,\bar{L})(1-p_L(d)) & z < N-d; \\ p_Z(z;d,L)p_L(d) & z \geq N-d. \end{cases} \quad (33)$$

With the pmfs of T and Z from (27) and (33), along with $p_D(d)$ from (19), we can finally compute the pmf of the PAoI by convolving T and Z and unconditioning over D. That is,

$$p_\Delta(\tau) = \sum_{d=K+1}^{N} p_D(d) \sum_{t=0}^{\min(d,\tau)} p_T(t;d) p_Z(\tau-t;d). \quad (34)$$

IV. Evaluation

In this section, we perform the evaluation of the analyzed schemes in terms of the following KPIs. The throughput is the main KPI for the broadband user, assuming that $K > 1$. The KPIs of the intermittent user are: 1) the 90-th percentile of the PAoI $\Delta_{90} = \max_n\{n : \Pr[\Delta < n] < 0.9\}$, and 2) the 90-th percentile of LR $L_{90} = \max_n\{n : \Pr[L < n] < 0.9\}$.

As these KPIs are interlinked, we evaluate their trade-offs for a specific activation probability α and erasure probabilities ϵ_1 and ϵ_2, 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

$$Y = \{(S_1, \tau_2) : (S_1, \tau_2) = f(c), c \in \mathcal{C}\},$$

where S_1 is the throughput of the broadband user and τ is the timeliness of the intermittent user (i.e., L_{90} and Δ_{90}). The *Pareto frontier* $\mathcal{P}$ for a combination of α, ϵ_1, and ϵ_2 is the set

$$\mathcal{P}(Y) = \{(S_1, \tau) \in Y : \nexists (S_1', \tau') \in Y : S_1 < S_1', \tau > \tau'\}. \quad (35)$$

To evaluate the performance of each scheme as a function of α, we set a minimum throughput for the broadband user $S_{1_{\min}}$. Then, for each considered value of α, we obtain the configuration of both OMA and NOMA that results in the minimum timeliness for the intermittent user while maintaining $S_1 \geq S_{1_{\min}}$ for a given combination of ϵ_1 and ϵ_2; we call this the optimal configuration. The parameter settings considered for the evaluation are listed in Table I. All presented results were verified with Monte Carlo simulations, not shown in the plots as the simulated curves mirror the analytical results.

Our first observation is that, for OMA, the values of K and N can be chosen independently of α and T_{int} with the single objective of maximizing the throughput of the broadband user S_1. Afterwards, by carefully choosing T_{int}, the trade-off can be adjusted to ensure $S_1 \geq S_{1_{\min}}$ is achieved with the minimal PAoI or LR. This is due to the orthogonality between the two users, where the settings for user 1 and user 2 do not affect each other. Hereafter, we denote the optimal values for K and N as K^* and N^*. By fixing $K \leq 64$, we got $K^* = 64$, and $N^* = 77$, which we use in the following.

Next, we investigate the performance of heterogeneous access with a PAoI-oriented service. Note that, in OMA case, queuing is not needed at the intermittent user (i.e., $Q = 1$), as newly arrived packets supersede the older ones. Fig. 3 shows the Pareto frontier for S_1 vs. Δ_{90}. As it can be seen, the maximum S_1 is around 0.8 with both OMA and NOMA, with OMA achieving a slightly higher S_1 at the expense of a considerable increase in Δ_{90}. Furthermore, Δ_{90} with OMA is relatively constant until $S_1 \approx 0.75$, after which it sharply increases. Conversely, the increase in Δ_{90} with NOMA is clearly observed from $S_1 \approx 0.5$ and reaches a maximum of $\Delta_{90} \approx 310$; this is lower than the maximum with OMA. In summary, OMA provides a better trade-off than NOMA when S_1 is between 0.6 and 0.8. If the target is a greater or lower S_1, both access methods deliver comparable results.

The performance of heterogeneous access with a LR-

TABLE I: Parameter settings

Parameter	Symbol	Setting
User 1 source block size	K	$\{2,\ldots,64\}$
User 1 coded block size	N	$\geq K$
User 1/User 2 erasure prob.	ϵ_1/ϵ_2	$0.1/0.05$
User 2 activation probability	α	$[10^{-4}, 10^{-1}]$
Intermittent slot period	T_{int}	$\{1,\ldots,64\}$
Max. queue length (OMA)	Q	$\{1,4\}$

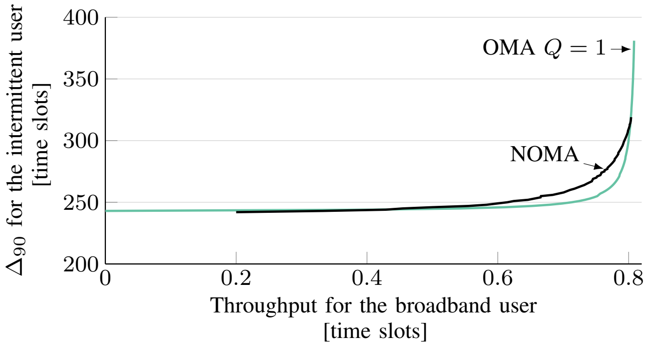

Fig. 3: Pareto frontier for the throughput of user 1 S_1 vs. 90-th percentile of the PAoI of user 2 Δ_{90} for $\alpha = 0.01$.

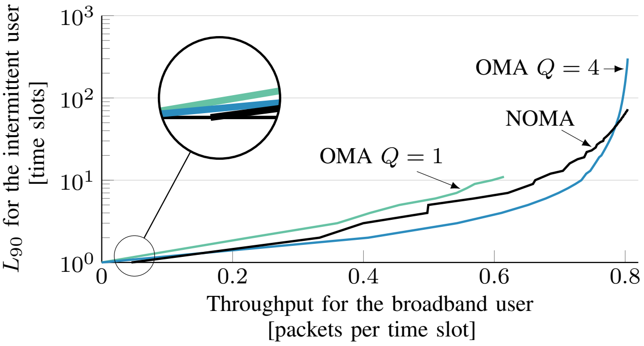

Fig. 4: Pareto frontier for the throughput of user 1 S_1 vs. 90-th percentile of LR for user 2 L_{90} for $\alpha = 0.01$.

oriented service is presented in Fig. 4, showing the Pareto frontier for S_1 vs. L_{90}. In this case, queuing at the intermittent user is needed with OMA to achieve comparable results as with NOMA. This can be clearly observed in Fig. 4: for OMA and $Q = 1$, the maximum S_1 that can be achieved with $L_{90} < \infty$ is around 0.6; and, for $Q = 4$, the achievable S_1 grows to nearly 0.8. Further, OMA and NOMA deliver similar values of L_{90} for low values of S_1, and, while OMA results in a lower L_{90} for $S_1 \in [0.35, 0.75]$, NOMA delivers a lower L_{90} for the maximum achievable S_1.

Finally, Fig. 5 shows the achievable timeliness for user 2 while ensuring $S_1 \geq 0.75$ for a wide range of values of α. There is an asymptote at the points where the requirements cannot be met. As it can be seen, OMA leads to a slightly lower Δ_{90} and L_{90} than NOMA for all values of α. Besides, Fig. 5 exhibits that the configuration of OMA is trivial: select the optimal K^* and N^* for user 1 and, then, select the shortest possible T_{int} to ensure $S_1 \geq 0.75$ and the probability of success for user 2 $p_{s,2} \geq 90$. This occurs because, for OMA, S_1 is not affected by α. For the considered scenario, $T^*_{\text{int}} = 13$ time slots, implying that $p_{s,2} \geq 90$ cannot be guaranteed for $\alpha \gtrsim 1/T^*_{\text{int}} = 0.076$. Conversely, the optimal values of K and N with NOMA increase with α. Furthermore, $S_1 \geq 0.75$ cannot be guaranteed with NOMA for $\alpha > 0.032$.

V. Conclusions

In this paper, we investigated orthogonal and non-orthogonal slicing mechanisms for heterogeneous services in the RAN. Our analyses and results highlighted the different trade-offs and achievable performance of OMA and NOMA. Despite the fact that we assumed a conservative collision model for NOMA, which directly impacts the prospects of packet decoding and, hence, the timeliness of the intermittent users, the presented Pareto frontiers for OMA and NOMA are closely similar. Thus, NOMA is expected to outperform OMA in a scenario with capture. On the other hand, our evaluations showed that finding the optimal configuration with NOMA depends on the activity pattern of the intermittent user, to which OMA is insensitive. Our future work includes the analysis and design of slicing mechanisms in scenarios with multiple broadband and intermittent users, as well as capture.

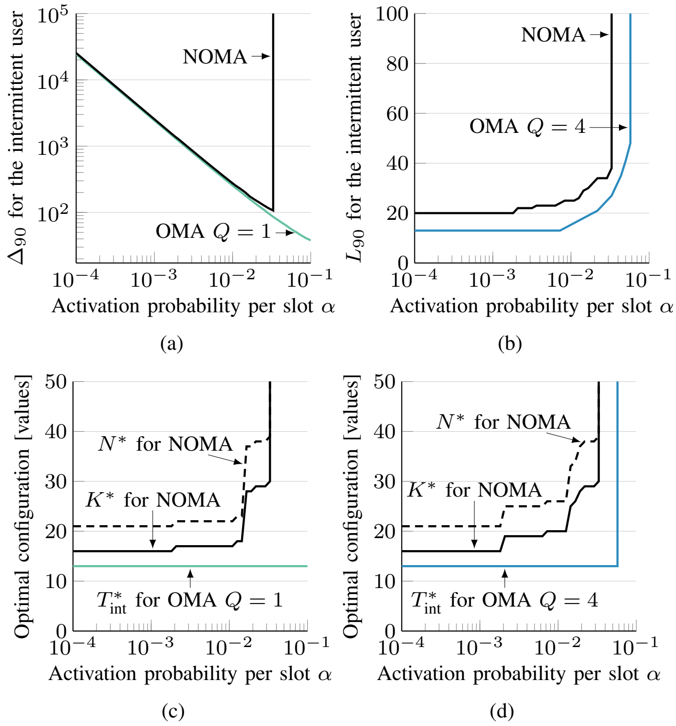

Fig. 5: Achievable timeliness for user 2 while ensuring $S_1 \geq 0.75$: (a) Δ_{90} and (b) L_{90}, and optimal configuration of the access methods for (c) Δ_{90} and (d) L_{90} as a function of α.

References

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

[2] 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. 55 765–55 779, 2018.

[3] A. Maatouk, M. Assaad, and A. Ephremides, "Minimizing the Age of Information: NOMA or OMA?" in *Proc. IEEE INFOCOM Workshops*, vol. 65, no. 8, 2019, pp. 102–108.

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

[5] Z. Wu, K. Lu, C. Jiang, and X. Shao, "Comprehensive study and comparison on 5G NOMA schemes," *IEEE Access*, vol. 6, pp. 18 511–18 519, 2018.

[6] M. Richart, J. Baliosian, J. Serrat, and J. L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462–476, 2016.

[7] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" *Proc. IEEE INFOCOM*, pp. 2731–2735, 2012.

[8] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Non-linear Age of Information in a discrete time queue: Stationary distribution and average performance analysis," in *Proc. IEEE International Conference on Communications (ICC)*, 2020.

[9] M. Costa, M. Codreanu, and A. Ephremides, "Age of information with packet management," *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 1583–1587, 2014.

[10] J. J. Nielsen, R. Liu, and P. Popovski, "Ultra-reliable low latency communication using interface diversity," *IEEE Transactions on Communications*, vol. 66, no. 3, pp. 1322–1334, 2018.

[11] 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. 13 035–13 049, 2019.