Priority Enabled Grant-Free Access With Dynamic Slot Allocation for Heterogeneous mMTC Traffic in 5G NR Networks

Thilina N. Weerasinghe, Vicente Casares-Giner, *Life Member, IEEE*, Indika A. M. Balapuwaduge, *Member, IEEE*, and Frank Y. Li, *Senior Member, IEEE*

***Abstract*— Although grant-based mechanisms have been a predominant approach for wireless access for years, the additional latency required for initial handshake message exchange and the extra control overhead for packet transmissions have stimulated the emergence of grant-free (GF) transmission. GF access provides a promising mechanism for carrying low and moderate traffic with small data and fits especially well for massive machine type communications (mMTC) applications. Despite a surge of interest in GF access, how to handle heterogeneous mMTC traffic based on GF mechanisms has not been investigated in depth. In this paper, we propose a priority enabled GF access scheme which performs dynamic slot allocation in each 5G new radio subframe to devices with different priority levels on a subframe-by-subframe basis. While high priority traffic has access privilege for slot occupancy, the remaining slots in the same subframe will be allocated to low priority traffic. To evaluate the performance of the proposed scheme, we develop a two-dimensional Markov chain model which integrates these two types of traffic via a pseudo-aggregated process. Furthermore, the model is validated through simulations and the performance of the scheme is evaluated both analytically and by simulations and compared with two other GF access schemes.**

***Index Terms*— Grant-free access, NR numerology, mMTC traffic, dynamic slot allocation, two-dimensional Markov chain, pseudo-aggregated process.**

I. Introduction

SIMULTANEOUS packet transmissions over the same radio resource cause performance deterioration for wireless access due to potential collisions among transmissions from competing devices. In fourth generation (4G) cellular networks, i.e., long term evolution-advanced (LTE-A), this problem was primarily addressed using grant-based (GB) communications. For GB channel access, a device follows a four-step handshake procedure for initial access with an evolved nodeB (eNB) by first transmitting a preamble before it obtains a grant for its data packet transmission. Once access is granted by the eNB, a data packet can be successfully transmitted without collision under ideal channel conditions. The initial preamble transmission, however, is still subject to collision(s) and could require multiple transmissions depending on traffic load and the availability of preamble resources at the eNB.

In LTE-A, the time required for initial four-step handshaking, which occurs prior to a data transmission, is in the order of 15 ms [1]. This is not a major concern since many 4G applications do not have stringent low latency requirements. In emerging fifth generation (5G) networks specified by the 3rd generation partnership project (3GPP), however, a variety of applications necessitate novel approaches for ultra-reliable low latency communications (URLLC) and massive machine type communications (mMTC). For small data transmissions which are common for mMTC traffic, the amount of control overhead required before an actual data transmission in GB schemes is too high with respect to the actual data to be transmitted and the handshake procedure lasts too long [2].

Although GB initial access is still kept as a legacy mechanism in 5G new radio (NR) networks, to perform such a four-step initial access procedure requires extra delay and protocol overhead [3], [4]. As an alternative to reduce overall latency, another category of mechanisms for data transmission, known as grant-free (GF), configured grant, or without grant, has emerged [4], [5]. Different from the GB principle, devices in GF communications transmit their data packets together with (or without using specific) control messages directly to a 5G NR nodeB (gNB) in available GF slots *without requiring the initial access procedure*. In other words, no dedicated preamble transmission for granting access and allocating radio resources is required for GF communications before starting a data packet transmission [3]. The benefits brought by this principle in terms of shortened delay and reduced protocol overhead make GF mechanisms attractive for various applications with URLLC/mMTC requirements and small data packets [1].

For periodic or deterministic traffic, a gNB can allocate dedicated slots to devices for their data transmissions. However, such a mechanism will lead to resource underutilization and

Manuscript received November 12, 2020; accepted January 13, 2021. Date of publication January 25, 2021; date of current version May 18, 2021. This work was supported by the Research Council of Norway through the Center for Research-Based Innovation (SFI) Offshore Mechatronics under Project 237896. For the work of Vicente Casares-Giner, Indika A. M. Balapuwaduge, and Frank Y. Li, the research leading to these results has received funding from the NO Grants 2014–2021, under Project contract no. 42/2021. The work of Vicente Casares-Giner has also been supported by the Spanish mobility program PRX19/00602 and by the project PGC2018-094151-B-I00. The associate editor coordinating the review of this article and approving it for publication was J. Choi. *(Corresponding author: Frank Y. Li.)*

Thilina N. Weerasinghe, Indika A. M. Balapuwaduge, and Frank Y. Li are with the Department of Information and Communication Technology, University of Agder (UiA), 4898 Grimstad, Norway (e-mail: thilina.weerasinghe@uia.no; indika.balapuwaduge@uia.no; frank.li@uia.no).

Vicente Casares-Giner is with the Departamento de Comunicaciones, Universitat Politècnica de València (UPV), 46022 València, Spain (e-mail: vcasares@dcom.upv.es).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2021.3053990.

Digital Object Identifier 10.1109/TCOMM.2021.3053990

long delay when traffic load is low or sporadic which is the case for many mMTC applications. Due to the unpredictability of sporadic traffic arrival patterns, it is beneficial to apply a *random access* protocol for GF data transmissions based on the principles of ALOHA or slotted ALOHA [6]. Furthermore, GF transmissions are generally recommended for small data transmission with a low or moderate level of traffic arrivals [7], [8].

A. Related Work

1) GF Communications: While GF is a more popular terminology favored by the research community, similar mechanisms are commonly referred to as configured grant or without grant in 3GPP specifications [4], [5], [9], [10]. In brief, existing GF based transmission schemes can be classified into four major categories, as summarized below. (i) *GF reactive*: A device needs to send its GF transmission and wait for an acknowledgment (ACK) or a negative ACK (NACK) from the gNB. If no ACK is received within the ACK timeout, or a NACK is received, the same packet will be retransmitted up to a retry limit; (ii) *GF reactive with power boost*: In order to increase successful reception probability, the transmit power of each retransmission could be higher than that of the previous unsuccessful transmission; (iii) *K repetitions without feedback*: A device transmits proactively $K > 1$ replicas of the same data packet across different GF slots in the same subframe [9]; and (iv) *K repetitions with feedback*: Similar to (iii), but it requires feedback from a gNB regarding its transmission status. Accordingly, a device will stop its transmission attempt once an ACK is received. Furthermore, the 3GPP states clearly that *at least an uplink transmission scheme without grant is supported for URLLC* and *an uplink transmission scheme without grant is targeted to be supported for mMTC [5].*

On the other hand, recent academic efforts foresee the feasibility of facilitating multi-packet reception by applying more advanced technologies for instance non-orthogonal multiple access (NOMA) and multiple-input multiple-output (MIMO) to GF transmissions. By treating collisions as interference through successive joint decoding or successive interference cancellation (SIC), [11] derived expressions for outage probability and throughput for GF-NOMA transmissions. In [12], a semi-GF scheme which provides dedicated GB access for one user while facilitating the other users with GF opportunistic access was proposed. Another recent work investigated the suitability of applying non-orthogonal sequences for abbreviating preamble collisions for GF transmissions and concluded that such sequences did not necessarily lead to better performance than the orthogonal ones [13]. In general, GF access exhibits the characteristic of slotted ALOHA-alike access mechanisms as presented below.

2) Slotted, Framed Slotted, and SIC-Enabled Slotted ALOHA for MTC Access: Depending on multi-packet reception is enabled or not, numerous variants of ALOHA-alike protocols, including framed slotted ALOHA (FSA), multi-channel slotted ALOHA, and SIC-enabled slotted ALOHA play an important role for medium access in mMTC [2], [14].

Based on the requirements for mMTC applications and design principles, FSA can be operated with either fixed or flexible frame length [15]. On the other hand, channels in multi-channel slotted ALOHA regard to different kinds of orthogonal resources such as codes or preambles which are used in the same, for instance time slot, during the initial access procedure. Using different orthogonal resources, multiple devices can access to a common channel simultaneously [1]. However, the amount of resources is still limited. For random access of mMTC traffic without multi-packet reception capability, a collision happens if two or more devices select the same preamble for their initial access or transmit their packets simultaneously in the same slot. More recent work intends to resolve collision following the principle of SIC through coded slotted ALOHA, e.g., in the form of frameless ALOHA [16].

Furthermore, priority oriented schemes in FSA have been studied previously. In [17], a pseudo-Bayesian ALOHA algorithm with mixed priorities was proposed. Similar to the pseudo-Bayesian ALOHA scheme presented in [18], the algorithm proposed in [17] allows multiple independent Poisson traffic streams compete for a slot or a batch of slots in a frame each with an assigned transmission probability. Following the idea on resource sharing, an adaptive framed pseudo-Bayesian ALOHA algorithm was proposed in [19].

Considering that the subframe length in NR is constant as 1 ms regardless of the adopted NR numerology [20], we adopt a fixed subframe length for our scheme design. Furthermore, since no dedicated preamble for initiating access and resource allocation is needed for GF transmissions, the access scheme proposed below in Sec. III allows devices transmit their packets directly to the associated gNB in the allocated GF slots. The scheme is designed upon the FSA principle but is based on the NR frame structure to be presented in Subsec. II-A.

B. Contributions

So far, little work has been done considering GF access for heterogeneous mMTC traffic. In this paper, we consider heterogeneous GF traffic arrivals with different reliability and/or latency requirements and propose a novel GF based access and data transmission scheme with dynamic slot allocation (DSA) in each NR subframe. Hereafter, the scheme is referred to as DSA-GF which stands for DSA for GF based access for heterogeneous traffic. Targeting at providing better performance to high priority traffic (HPT), the scheme accommodates the remaining slots in the same subframe to low priority traffic (LPT) so that higher total slot utilization is achieved.

In contrast to most existing work which generally neglected *slot based* GF transmissions and slot utilization, this paper targets at 5G NR numerology as the basis for our scheme design and intends to maximize slot utilization for heterogeneous traffic integrated with priority enabled access. Through dynamic slot allocation, the dependence of two types of GF traffic is handled and modeled through a pseudo-aggregated process where both traffic types share available slots in each subframe and slot allocation to HPT is independent of that of LPT.

In brief, the main contributions of this paper are summarized as follows.

- Based on the NR frame structure, a novel GF based data transmission scheme, DSA-GF, which considers arrivals of heterogeneous GF traffic is proposed. The scheme performs traffic estimation, access control, and dynamic slot allocation on a subframe-by-subframe basis. Based on our scheme, both HPT access privilege and LPT resource preservation are achieved and they are bound together smoothly in each subframe.
- To evaluate the performance of the proposed scheme, a two-dimensional (2D) Markov chain model, in which a pseudo-aggregated process is defined to link two types of GF traffic by considering their coherence for slot allocation in a common subframe, has been developed. For a network with the same configuration, the number of states in our model is much less than what is needed in conventional Markov models.
- Extensive discrete-event based simulations have been performed to validate the preciseness of the developed model and assess the performance of the DSA-GF scheme. Through performance assessment under various HPT/LPT traffic variations and comparison with two other GF schemes, the effectiveness of the proposed scheme is further demonstrated.

In a nutshell, the uniqueness and novelty of our paper are reflected by the fact that this work is anchored at a niche with an intersection among 5G NR numerology, traffic estimation based dynamic slot allocation, proper handling of heterogeneous traffic considering the performance of both HPT and LPT, and pseudo-aggregated 2D Markov chain modeling for heterogeneous traffic. To the best of our knowledge, this is the first attempt which is dedicated for 5G NR numerology based GF transmission with dynamic slot allocation at the subframe level for heterogeneous traffic, combined with a Markov model with a significantly reduced state space bridging both types of traffic together for performance evaluation.

Furthermore, it is worth mentioning that the 3GPP has newly decided to discontinue NOMA as a work-item for 5G NR but leave it as a study-item for beyond 5G [21]. Under such a circumstance, the importance of investigating viable GF schemes based on the existing NR frame structure remains significant and it becomes even an imperative task as such schemes may serve as the basis or at least references for NOMA based GF scheme design.

The remainder of the paper is organized as follows. Sec. II provides preliminaries on NR numerology and presents the network scenario. In Sec. III, the proposed scheme is explained in details. Then we develop a 2D Markov model in Sec. IV to analyze its performance. Thereafter, Sec. V illustrates the numerical results. Finally, the paper is concluded in Sec. VI.

II. Preliminaries, Scenario and Assumptions

This section presents the NR frame structure which forms the basis for our scheme design and outlines the scenario.

A. 5G NR Frame Structure and Numerologies

With 15 kHz as a baseline for subcarriers as used in 4G, 5G NR defines five numerologies based on subcarrier spacing $\Delta f = 2^{\beta} * 15$ kHz, where $\beta = 0, 1, \ldots, 4$ is the numerology index, with different slot duration lengths downwards from 1 ms to 62.5 μs [20], [22]. As depicted in Fig. 1, the per frame duration in NR is still 10 ms, and the same as in LTE-A, one *frame* consists of 10 *subframes* each with 1 ms duration. Moreover, one NR subframe may have one (for $\beta = 0$) or multiple (up to 16) *slots* depending on the value of the numerology index β.

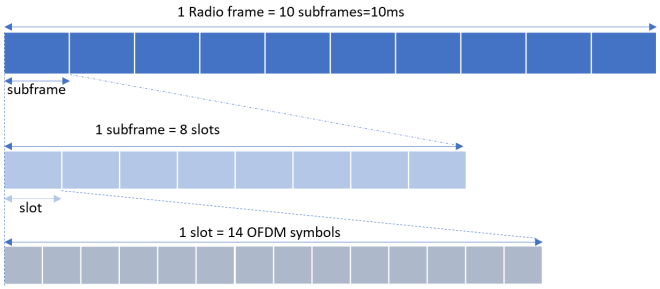

Fig. 1. 5G NR frame structure for numerology $\beta = 3$.

Depending on the size of a packet, one or multiple orthogonal frequency division multiplexing (OFDM) symbols out of the available 14 symbols within a slot can be utilized by GF traffic [9], [10]. Considering that GF transmissions are targeted at small data packets in mMTC networks [24], we assume in this study that a packet with the size of less than 14 OFDM symbols is sufficient for one GF packet transmission. The remaining symbol(s) within the same slot can be allocated to other data traffic (for instance GB transmissions) and control information exchange as NR allows flexible uplink and downlink scheduling *at a symbol level* within one NR slot [20]. As such, *all slots in a subframe* can be utilized for GF data transmissions.

B. Scenario and Traffic Arrivals

Consider a scenario where an NR cell covers a large number of mMTC devices. Although both GF and GB devices may coexist, this study focuses only on GF data transmissions. More specially, GF data transmissions considered in this study are performed in each subframe following the DSA-GF scheme presented in the next section. A device is regarded as *active* if it has one packet ready to transmit. The transmission of a device is regarded as successful if no other device transmits in the same slot and it is confirmed through an ACK message provided at the end of each subframe. If two or more devices transmit in the same slot, a collision occurs and all involved transmissions are considered to be failed. If a device does not obtain a transmission opportunity due to the constraint of the permission probability in the current subframe or its transmission in the current subframe collided, it will try again in the next subframe based on a new permission probability broadcast by the gNB right before the next subframe begins.

Although the total number of mMTC devices covered by a cell could be huge [25], they generate typically sporadic traffic with small packet sizes. Therefore, the number of arrivals per subframe, i.e., within 1 ms, is rather limited. To reflect this, we adopt a combination of number of devices and activation probability as an indicator to represent offered traffic.

Without loss of generality, we consider numerology $\beta = 3$ as an example in most figures and descriptions in this paper. Later on in Subsec. V-F, we further demonstrate the applicability of the scheme to two other numerologies, i.e., $\beta = 2$ and $\beta = 4$ which have 4 and 16 slots per subframe respectively.

Two categories of traffic arrivals are considered, known as HPT and LPT respectively. While HPT requires superior performance, LPT can tolerate longer access delay and higher packet loss. For slot allocation in each subframe, HPT has access privilege over its counterpart, i.e., LPT. In the considered cell covered by one gNB, there are a finite number of HPT and LPT devices, denoted by M_x with $x = 1$ for HPT and $x = 2$ for LPT, respectively. The arrival process for both categories follows a Bernoulli process. That is, each device generates one data packet per subframe with activation probability a_x. This assumption means that each device has at most one packet ready to transmit at each subframe. Furthermore, we assume that the ACK message transmission from the gNB is always successful. No channel impairment is considered in this study and propagation delay is regarded to be negligible compared with access delay.

III. Proposed Transmission Scheme for GF Traffic

The DSA-GF scheme focuses on the NR frame structure and features the flavors of both 4G and 5G access mechanisms such as access class barring and unified access control [1], [23], imposing different permission probabilities to heterogeneous types of traffic. It is operated on a subframe-by-subframe basis. First of all, an observation-based slot allocation algorithm assigns an optimal number of slots to serve HPT transmissions in order to achieve maximum throughput, low access delay and reduced packet loss probability. In the meantime, the algorithm takes into account the performance of LPT through *slot preservation to LPT in order to avoid starvation of LPT*. To do so, the maximal number of slots to be allocated to HPT is restricted to the total number of slots per subframe *minus* one (for $\beta = 2$ and 3) or two (for $\beta = 4$). Then the remaining slots in the same subframe will be allocated to LPT. For a given subframe, the more slots allocated to HPT, the less slots assigned to LPT.

For a given numerology, the total number of slots per subframe, denoted by U, is a constant and it is decided by the NR frame structure presented above. Inspired by the pseudo-Bayesian broadcast algorithm for slotted ALOHA proposed in [18], we develop a novel random access scheme for NR based GF transmissions as presented below. While the algorithm in [18] targeted at slotted ALOHA with a single slot, the protocol designed in this paper is tailored to operations where multiple slots together form one subframe, taking into account the NR frame structure. A list of notations used in this paper and their explanations can be found in Tab. I.

A. Transmission Principles of DSA-GF

At the beginning of each subframe, the gNB provides to all devices through a broadcast message with the permission probability for each type of traffic, denoted as p_x where $x = 1$

TABLE I

Summary of Notations and Descriptions

Notation	Description
x	Traffic type index where $x = 1$ and $x = 2$ represent HPT and LPT respectively
β	Numerology index in the NR frame structure $0 \leq \beta \leq 4$
$\mathbb{N}$ ($\mathbb{Z}$)	The set of natural (integer) numbers
M_x	Number of HPT and LPT devices
a_x	Probability of generating one data packet per subframe for traffic type x, i.e., activation probability
$M_x a_x$	Offered traffic of type x
U	The number of slots per subframe (with its value decided by the adopted NR numerology)
$u_{x,min}$	The min. number of slots allocated to traffic type x
$u_{x,max}$	The max. number of slots allocated to traffic type x
$W_{x,t}$	r.v. of the number of active devices estimated by the gNB for traffic type x at subframe t
$U_{x,t}$	r.v. of the number of slots allocated by the gNB to traffic type x at subframe t
$N_{x,t}$	r.v. of the number of active devices of type x according to the packet arrival and departure processes
$w_{x,t}$	Number of packets of type x estimated by the gNB ready for transmission at subframe t
$E_{x,max}$	The maximum value of $W_{x,t}$ estimated by the gNB
$m_{x,t}$	Number of slots allocated to traffic type x at subframe t ($m_{1,t} + m_{2,t} = U$)
$\overline{m}_{x,t}$	Average number of slots allocated to traffic type x at subframe t
$n_{x,t}$	Number of active devices of type x at subframe t according to the packet arrival and departure processes
(μ, u, i)	Simplified notation for $(w_{x,t}, m_{x,t}, n_{x,t})$
(ν, v, j)	Simplified notation for $(w_{x,t+1}, m_{x,t+1}, n_{x,t+1})$
$P_{\mu,u,i;\nu,v,j}$	The set of transition probabilities from subframe t to subframe $t+1$
$\pi_{\mu,u,i}$	Steady-state probability that, at the beginning of a subframe, the number of active devices estimated by the gNB is μ, the number of allocated slots is u, and the number active terminals is i
$\hat{P}_{u,v}$	The set of transition probabilities from subframe t to subframe $t+1$ for the pseudo-aggregated process
$\hat{\pi}_u$	Steady-state probability of number of slots occupied by HPT
$p_{x,t}$	Permission probability for packet transmissions in subframe t
$h_{x,t}$	Number of unused slots (holes) in subframe t
$s_{x,t}$	Number of successful slots in subframe t
$c_{x,t}$	Number of collided slots in subframe t
$(h, s, c)_{x,t}$	The set of observations at subframe t
$\hat{\lambda}_{x,t}$	The estimated number of *new* devices that have become active during subframe t
$\hat{w}_{x,t+1}$	Number of estimated backlogged devices which will be active in subframe $t+1$
$R^{z,u}_{s_t,c_t}$	Probability that within subframe t, z active devices intend to access over u slots, and the result is s_t successes and c_t collisions.
$D^{i,u}_{s_t,c_t}(p)$	Probability that within subframe t, z out of i active devices transmitted with permission probability p, and the result is s_t successes and c_t collisions
$B^i_z(p_t)$	Probability that z out of i active devices ($0 \leq z \leq i$) transmit in subframe t following a binomial distribution
Ω_x	The set of (h, s, c) values observed in subframe t for traffic type x
γ^{sf}_x	Total throughput per subframe for traffic type x
θ_x	Packet loss probability for traffic type x
γ^{slot}_x	Total throughput per slot for traffic type x
d^{sf}_x	Delay for traffic type x, in number of subframes
Δ_1, Δ_2	The set of all possible values in μ and i, for $\pi_{\mu,i}$ (Δ_1); in μ, u, and i, for $\pi_{\mu,u,i}$ (Δ_2)
$\mathcal{C}$	The set of all possible collisions such that $h_t + s_t + c_t = u$
$E = \{(\mu, i)\}$	The set of states of the Markov chain for HPT

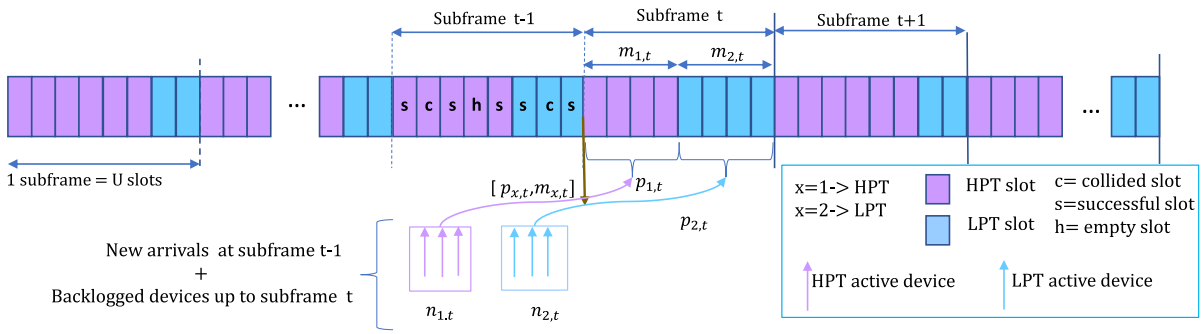

Fig. 2. Illustration of DSA-GF: With priority, the slots are dynamically divided into two groups, one for HPT and the other for LPT.

for HPT and $x = 2$ for LPT respectively. With probability p_x, each *active* device randomly selects one of the allocated slots to type x within the current subframe to transmit its packet. With probability $1 - p_x$, the device postpones its transmission to the next subframe. The permission probability is updated for each subframe based on two ingredients.

For each type, the gNB first observes each slot of the current subframe and counts the number of holes h (a slot that is not occupied by any transmission(s) is referred to as a hole), successes s (a slot with a single packet transmission), and collisions c (a slot with more than one packet transmissions). Then, it proceeds to estimate the number of packets involved in the transmissions of the current subframe.

Second, the gNB estimates the new arrivals of type x during the current subframe, which together with the backlogged devices will attempt to transmit their packets with an updated permission probability in the next subframe. Backlogged devices are those devices that postpone their transmission in the current subframe due to the imposed permission probability *plus* those devices that were involved in collisions. Furthermore, active devices comprise both backlogged devices and new arrivals. In the next subframe, all active devices will attempt to transmit following the permission probability for each traffic type (details are given in the next subsection).

In DSA-GF, new arrivals follow the immediate first transmission (IFT) principle. By IFT, it is meant that any just-arrived packet in the current subframe will be potentially transmitted in the next immediately available subframe according to the updated permission probability provided by the gNB. Upon the successful reception of a packet transmission, immediate feedback is performed. The operation of the DSA-GF scheme is illustrated in Fig. 2.

B. Detailed Access Procedure for Heterogeneous Traffic

Within each subframe, there are a total number of U slots shared by both streams, one from each type of devices. Let m_x denote the number of slots allocated to the HPT ($x = 1$) and LPT ($x = 2$) flows respectively. We have $m_1 + m_2 = U$. The same notations for subscripts apply to other expressions throughout the context. Denote by $u_{1,min}$ ($u_{1,max}$) the minimum (maximum) number of slots that can be allocated to HPT at any subframe, in such a way that $1 \leq u_{1,min} \leq m_1 \leq u_{1,max} < U$. By keeping $u_{1,max} < U$, our scheme reserves *at least one slot per subframe* for LPT so that no starvation happens to LPT regardless of HPT traffic intensity. In what follows, we present how m_x and p_x are updated from subframe to subframe.

During each subframe, the gNB observes what happened in each slot. Let $(h, s, c)_{x,t}$ denote the number of holes, successes and collided slots, respectively, observed for type x during subframe t. Obviously, we have $h_{x,t} + s_{x,t} + c_{x,t} = m_{x,t}$ with $m_{1,t} + m_{2,t} = U$. Furthermore, let $\hat{\lambda}_{x,t}$ be the estimation of new arrivals assessed by the gNB, i.e., the estimated number of devices that have generated a packet during subframe t. Then, the $(m, p)_{x,t} \rightarrow (m, p)_{x,t+1}$ update is performed according to the three steps presented below.

Step 1: Update the estimated number of active devices for HPT and LPT, at the end of subframe t.

- First, for $x = 1, 2$, based on the observations $(h, s, c)_{x,t}$ and the estimated number of active devices at the beginning of subframe t, $w_{x,t}$, the gNB estimates the number of backlogged devices at the end of this subframe t, $\hat{w}_{x,t+1}$. For that purpose, we extend Rivest's pseudo-Bayesian broadcast control algorithm [18] to data transmissions with multiple slots in each subframe so that $\hat{w}_{x,t+1} = w_{x,t} + \frac{c_{x,t}}{e-2} - (h_{x,t} + s_{x,t}) \approx w_{x,t} + 1.3922 \times c_{x,t} - (h_{x,t} + s_{x,t})$. In this expression, $1.3922 \times c_{x,t}$ represents an increment in the estimated number of collided packets which will attempt to transmit again following the rule given in **Step 3**, and $h_{x,t} + s_{x,t}$ represent the idle slots plus successful transmissions that will not retransmit in the next subframe.
- Second, for $x = 1, 2$, the gNB estimates the number of *new arrivals* during subframe t, $\hat{\lambda}_{x,t}$. Considering a network in the steady state where the offered traffic and the carried traffic reach an equilibrium, we set $\hat{\lambda}_{x,t}$ equal to $s_{x,t}$, the number of successes at subframe t. A more elaborated estimator of the arrival process is however out of the focus of this paper.
- Third, the total number of active devices at the end of subframe t ready for transmission at subframe $t+1$ is the sum of $\hat{w}_{x,t+1}$ and $\hat{\lambda}_{x,t}$. Since $w_{x,t+1}$ cannot be negative, we set $w_{x,t+1} = \max(\hat{w}_{x,t+1}, 0) + \hat{\lambda}_{x,t}$. Note that $w_{x,t+1}$ can be any non-negative real number.

Step 2: Update the number of slots to be allocated in the next subframe $t + 1$.

- To give higher priority to HPT, we first allocate $m_{1,t+1} = \max\big(u_{1,min}, \min(\lceil w_{1,t+1} \rceil, u_{1,max})\big)$ and then configure $m_{2,t+1} = U - m_{1,t+1}$. That is, the capacity that is not assigned to HPT will be allocated to LPT. In the above expression, a ceiling function is introduced considering that $m_{x,t+1}$ is an integer number such as $u_{x,min} \leq m_{x,t+1} \leq u_{x,max}$.

Step 3: Update the permission probabilities for subframe $t+1$ for each type of traffic.

- Set $p_{x,t+1} = \dfrac{m_{x,t+1}}{\max(m_{x,t+1}, w_{x,t+1})} = \min\left(\dfrac{m_{x,t+1}}{w_{x,t+1}}, 1\right)$. That is, when the estimated number of active devices, $w_{x,t+1}$, is greater than the number of allocated slots, $m_{x,t+1}$, the assigned permission probability becomes less than 1. Otherwise, it is 1. Note that for each type of traffic *the same permission probability applies to all active devices.*

IV. Discrete-Time Markov Model for DSA-GF

To evaluate the performance of the proposed DSA-GF scheme for heterogeneous GF traffic, we develop a 2D Markov model which integrates HPT and LPT through a pseudo-aggregated process. During each subframe, every device generates one data packet with probability a_x according to a Bernoulli process. For packet buffering, a packet rejection mechanism is adopted meaning that a packet is rejected when it arrives at a device and finds the buffer full [26].

A. Building a Discrete-Time Markov Model

Thanks to the memoryless property of the arrival processes, we can build a discrete-time Markov chain for the presented DSA-GF scheme. For this purpose, let us observe the system at the border of two consecutive subframes, e.g., at the time instant when subframe $t-1$ ends (or subframe t begins), where $t \in \mathbb{Z}$ ($\mathbb{Z}$ is the set of integer numbers). Subframe by subframe, these time instants are regarded as the *transition instants* in the developed Markov model defined by a set of three random variables for each type of traffic. For traffic type x where $x = 1$ (HPT) or $x = 2$ (LPT), let $W_{x,t}$ be the random variable (r.v.) representing the number of active devices estimated by the gNB at transition instant t, $U_{x,t}$ be the r.v representing the number of slots in subframe t allocated to traffic type x, and $N_{x,t}$ be the r.v. representing the actual number of active devices (new arrivals plus backlogged devices) ready for transmission in subframe t.

The transition probabilities of the Markov chain, in a compact format, are as follow.

$$\begin{aligned}&\Pr\big((\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_{t+1}|(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_t,(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_{t-1},\\ &\qquad(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_{t-2},\dots\big)\\ &\quad=\Pr\big((\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_{t+1}|(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_t\big)\end{aligned} \tag{1}$$

where $(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_t$ denotes $\boldsymbol{W}_t = [W_{1,t}, W_{2,t}]$, $\boldsymbol{U}_t = [U_{1,t}, U_{2,t}]$, and $\boldsymbol{N}_t = [N_{1,t}, N_{2,t}]$. Note that $U_{1,t} + U_{2,t} = U_t = U$. In the expressions presented hereafter, we have introduced a compact expression based on (1) with simplified notations $\Pr(W_{x,t} = w_{x,t})$, $\Pr(U_{x,t} = m_{x,t})$, and $\Pr(N_{x,t} = n_{x,t})$ respectively.

It is worth mentioning that the Markov chain defined in (1) entails high complexity. In what follows, we opt a lightweight and consistent procedure which consists of the following three phases. 1) Subsec. IV-B performs the analysis of HPT since its behavior is independent of that of LPT; 2) Subsec. IV-C builds a pseudo-aggregated process which takes into account the correlation or dependence between these two types of traffic for slot allocation in the same subframe; and 3) Subsec. IV-D presents the performance of LPT.

B. The Analysis of High Priority Traffic

1) Modeling the HPT Process: Consider that a total number of M_1 devices generate data packets according to a Bernoulli process with probability a_1. Clearly, a Markov chain can be built at the transition instants as defined above. Using (1) and omitting the random variables related to the notations of LPT, we have the corresponding transition probabilities, i.e.,

$$\begin{aligned}&\Pr\big((W_1,U_1,N_1)_{t+1} = (\nu,v,j)|(W_1,U_1,N_1)_t = (\mu,u,i)\big)\\ &\qquad= P_{\mu,i;\nu,j}.\end{aligned} \tag{2}$$

In (2), the following short notations are used: $(w,m,n)_{1,t} \equiv (\mu,u,i)$ and $(w,m,n)_{1,t+1} \equiv (\nu,v,j)$. For convenience, we restrict the values of $w_{1,t}$ $t \in \mathbb{N}$ to natural numbers (notice that, according to **Step 1** of the DSA-GF scheme, $w_{1,t}$ can be any real number). Such a restriction makes it possible to enumerate or list the states of the Markov chain. Since there exists a deterministic relationship between $u = m_{1,t}$ and $\mu = w_{1,t}$, i.e., $u = \max\big(u_{1,min}, \min(\mu, u_{1,max})\big)$, only two random variables, $W_{1,t}$ and $N_{1,t}$, are sufficient to fully describe this Markov chain. In other words, the Markov chain with three sets of r.v. defined in (1) shrinks to a 2D model. Accordingly, the short notation of $P_{\mu,i;\nu,j}$ in (2) represents the set of corresponding transition probabilities from subframe t to subframe $t+1$. For expression clarity in the rest of this subsection, subscript "1" which is meant for HPT is intentionally omitted unless it is explicitly necessary.

Let us first derive the explicit expressions for $P_{\mu,i;\nu,j}$ starting with the transition $\mu \to \nu$. Based on the observations of slots for traffic type 1 during subframe t, i.e., $(h,s,c)_t$ where $h_t + s_t + c_t = u$, the gNB uses a function $f((h,s,c)_t) = c_t/(e-2) - (h_t + s_t)$ to estimate the number of backlogged devices, i.e., $\hat{w}_{t+1} = \mu + f((h,s,c)_t)$ (see **Step 1**: First presented in Subsec. III-B). After that and following **Step 1**: Second and **Step 1**: Third, the estimated number of new arrivals during subframe t is taken into account, such that the estimated number of devices active at the beginning of subframe $t+1$ is $\nu = \max(\hat{w}_{t+1}, 0) + \hat{\lambda}_t$.

Although $\mu = w_t$ is set to an integer number, in general, neither $\hat{w}_{t+1}$ nor $\hat{\lambda}_t$ is an integer number. As $\nu = w_{t+1}$ is also set to be an integer number, we introduce the 'ceil' operation such that,

$$\begin{aligned}\nu &= \lceil \max(\mu + f((h,s,c)_t), 0) + \hat{\lambda}_t \rceil;\\ v &= \max\big(u_{1,min}, \min(\nu, u_{1,max})\big);\\ p_{t+1} &= \min\left(\frac{v}{\nu}, 1\right).\end{aligned} \tag{3}$$

Note that the updated probability p_{t+1} applies to all active devices at subframe $t+1$ and it is restricted to be a fraction of two integer numbers.

Second, we evaluate the transition probability $i \rightarrow j$ referred to in (2). For that purpose, we consider in the first step the departure process, i.e., for packets that successfully finished their transmissions during the actual subframe t. At the beginning of subframe t, each of the i active devices chooses to transmit with permission probability p_t or to postpone its transmission with probability $1-p_t$, respectively. Then, the probability that z out of i active devices ($0 \leq z \leq i$) transmit in subframe t follows a binomial distribution, $B_z^i(p_t) = \binom{i}{z} p_t^z (1-p_t)^{i-z}$. With $u = m_{1,t}$, let $R_{s_t,c_t}^{z,u}$ denote the probability that z packets (active devices) intend to access over u slots of subframe t resulting in s_t successful transmissions and c_t collided slots. For any packet transmission, each of the z active devices chooses, with equal probability, one of the u slots of subframe t. Jointly considering these two sequential and independent actions, we obtain the probability that within subframe t, s_t out of i active devices succeed in the transmission of its own packet whereas the other $i-s_t$ devices were involved in collisions or deferred their transmissions. Analytically, it is expressed as,

$$D_{s_t,c_t}^{i,u}(p_t) = \sum_{z=s_t}^{i} B_z^i(p_t) R_{s_t,c_t}^{z,u}. \tag{4}$$

In (4), $R_{s_t,c_t}^{z,u}$ can be evaluated using, for instance, the recursions given at [28].

In the second step, we take into account the number of devices that will be active at the transition instant at the end of subframe t. Since the arrival of packets comes from M_1 sources each one with activation probability a_1, the arrival process follows a binomial distribution. Jointly considering the departure and arrival processes, which are independent of each other, we have,

$$P_{\mu,i;\nu,j} = \sum_{(h,s,c)_t \in \Omega_1} D_{s_t,c_t}^{i,u}(p_t) A_{j-i+s_t}^{M_1-i+s_t}(a_1), \tag{5}$$

where $A_{j-i+s}^{M_1-i+s}(a_1)$ follows the binomial distribution, as $A_l^k(a_1) = B_l^k(a_1) = \binom{k}{l} a_1^l (1-a_1)^{k-l}$. The set Ω_1 defined in (5) represents the set of $(h,s,c)_t$ values observed in subframe t that satisfy the following two conditions,

$$\Omega_1 \overset{\text{def}}{\equiv} \begin{cases} u = h_t + s_t + c_t = \max\big(u_{1,min}, \min(\mu, u_{1,max})\big); \\ \nu = \lceil \max(\mu + f((h,s,c)_t), 0) + \hat{\lambda}_t \rceil. \end{cases} \tag{6}$$

Then, the solution in the steady state regime is given by the stochastic row vector $\boldsymbol{\pi}$ ($\boldsymbol{\pi}\mathbf{e} = \mathbf{1}$) which can be obtained from the linear equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ with $\boldsymbol{\pi} = \{\pi_{\mu,i}\}, \mathbf{P} = \{P_{\mu,i;\nu,j}\}$. Here, $\mathbf{e}$ is a column vector of all 1's, $\pi_{\mu,i}$ is the probability that at the start of an arbitrary subframe the number of active devices estimated by the gNB is μ and the actual number of active devices is i.

2) Throughput, Access Delay, and Packet Loss Probability for HPT: Based on $\boldsymbol{\pi}$, we derive below expressions for the performance of HPT in terms of four parameters as defined below.

Firstly, the mean value of the number of successfully transmitted packets *within one subframe*, defined as throughput per subframe, is obtained according to,

$$\begin{aligned} \gamma_1^{sf} &= \sum_{s_t=u_{1,min}}^{u_{1,max}} s_t \sum_{(\mu,i)\in\Delta_1} \pi_{\mu,i} \sum_{c_t \in \mathcal{C}} D_{s_t,c_t}^{i,u}(p_t) \\ &= \sum_{(\mu,i)\in\Delta_1} i p_t \left(1 - \frac{p_t}{u}\right)^{i-1} \pi_{\mu,i}. \end{aligned} \tag{7}$$

In (7), the set Δ_1 shown as $(\mu,i) \in \Delta_1$ contains all possible values in μ and i and the set $\mathcal{C}$ shown as $c_t \in \mathcal{C}$ contains all possible collided slots such that $h_t + s_t + c_t = u$. Observe that the relationship between $\mu = w_t$ and $u = m_t$ is given in (6). The second equality is obtained after some algebraic operations and the details are omitted for the sake of brevity. Instead, a short clue is outlined as follows. Using DSA-GF, the expected number of successful transmissions when i active devices access to a set of u slots with permission probability $p_t = \min(u/\mu, 1)$ is given by,

$$\begin{aligned} &E\big(\text{success} | (N_t = i, U_t = u, p_t = p = \min(u/\mu, 1)\big) \\ &\qquad = ip\left(1 - \frac{p}{u}\right)^{i-1}. \end{aligned} \tag{8}$$

Then, the last equality in (7) is a weighted sum of (8) with probabilities $\pi_{\mu,i}$.

To give further insights on HPT performance in terms of resource utilization, how long a packet has to stay in a buffer, and how likely a packet may get lost, we define three other parameters. The mean value of the number of successfully transmitted packets *within one slot*, i.e., throughput per slot, which represents resource utilization is obtained based on (7),

$$\gamma_1^{slot} = \gamma_1^{sf} / \sum_{(\mu,i)\in\Delta_1} u\pi_{\mu,i}. \tag{9}$$

Thirdly, access delay in this study, d_1^{sf}, is defined as the mean sojourn time a packet stays in a buffer until it is successfully transmitted. Using Little's formula, the average number of customers in our steady state system (which is the mean number of active devices at the beginning of an arbitrary subframe, obtained by $\sum_{(\mu,i)\in\Delta_1} i\pi_{\mu,i}$) equals to d_1^{sf} multiplied by the average successful rate (i.e., the average number of successful transmissions per subframe, γ_1^{sf}). Therefore, we have

$$d_1^{sf} = \sum_{(\mu,i)\in\Delta_1} i\pi_{\mu,i} / \gamma_1^{sf}. \tag{10}$$

The fourth performance parameter, packet loss probability, is defined as the ratio of the rejected, i.e., offered minus carried traffic, to the offered traffic. For HPT, it is expressed as

$$\theta_1 = (M_1 a_1 - \gamma_1^{sf}) / M_1 a_1. \tag{11}$$

C. Linking HPT and LPT With a Pseudo-Aggregated Process

Based on the 2D Markov chain that models the HPT behavior, denoted as X, we construct a tailored pseudo-aggregated process that links the HPT process with the LPT process.

Inspired by the procedure presented in [27], we make a partition of the states of the Markov chain X. Let $E = \{(\mu,i)\}$

be the set of states of our initial Markov chain X, where $(1 \leq \mu \leq E_{1,max},\, 0 \leq i \leq M_1)$. It is understood that $E_{1,max}$ represents the maximum number of active HPT devices that the gNB can estimate. Let us sort the set of states into the following order,

$$\begin{aligned}\mathcal{E}_\mu &= \{(\mu,0),(\mu,1),\ldots,(\mu,M_1)\};\\ \mu &= 1,2,\ldots,u_{1,min}-1,u_{1,min},\\ &\quad u_{1,min}+1,\ldots,u_{1,max}-1,\\ &\quad u_{1,max},u_{1,max}+1,\ldots,E_{1,max}-1,E_{1,max}.\end{aligned}$$

Now, let $\mathfrak{F} = \{\mathcal{F}(u_{1,min}),\ldots,\mathcal{F}(u_{1,max})\}$ be a partition of E such that

$$\begin{aligned}\mathcal{F}(u_{1,min}) &= \mathcal{E}_1 \cup \mathcal{E}_2 \cup \cdots \cup \mathcal{E}_{u_{1,min}};\\ \mathcal{F}(\mu) &= \mathcal{E}_\mu, \quad u_{1,min} < \mu < u_{1,max};\\ \mathcal{F}(u_{1,max}) &= \mathcal{E}_{u_{1,max}} \cup \mathcal{E}_{u_{1,max}+1} \cup \cdots \cup \mathcal{E}_{E_{1,max}}.\end{aligned}$$

Let F be the set of integers $\{u_{1,min},\ \ldots,\ u_{1,max}\}$. Based on the initial Markov chain X with known values on E, we associate the pseudo-aggregated Markov chain Y with potential values on F, defined by: $Y_t = m \iff X_t \in \mathcal{F}(m)$ for all values of $t \in \mathbb{Z}$. Observe that, due to the mapping procedure, the pseudo-aggregated process includes the statistics of the number of slots allocated to HPT devices in the same subframe. Then, the transition probabilities of the pseudo-aggregated Markov chain Y are given as follows,

$$\hat{P}_{u,v} \overset{\text{def}}{=} \sum_{i\in\mathcal{F}(\mu)} \left(\pi_{\mu,i} / \sum_{h\in\mathcal{F}(\mu)} \pi_{\mu,h}\right) \sum_{j\in\mathcal{F}(\nu)} P_{\mu,i;\nu,j}; \quad (12)$$

where $u = \max\big(u_{1,min}, \min(\mu, u_{1,max})\big)$ and $v = \max\big(u_{1,min}, \min(\nu, u_{1,max})\big)$.

Clearly, the probabilities $\hat{P}_{u,v}$ for $u_{1,min} \leq u, v \leq u_{1,max}$ constitute the Markov chain that counts the number of slots per subframe allocated to HPT devices. The Markov chain defined by (12) preserves the mean values (sojourn times in each set of state) of the original process, but in general higher statistical moments of these two processes are different from each other.

By solving the linear equation $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}\hat{\mathbf{P}}$ with $\hat{\boldsymbol{\pi}} = \{\hat{\pi}_u\}$ and $\hat{\mathbf{P}} = \{\hat{P}_{u,v}\}$, the stochastic vector $\hat{\boldsymbol{\pi}}$ ($\hat{\boldsymbol{\pi}}\mathbf{e} = \mathbf{1}$) is obtained. Accordingly, the statistics of the r.v. number of slots allocated per frame for HPT can be easily obtained. This pseudo-aggregated Markov chain provides a link between HPT and LPT. This link will be used to analyze the performance of LPT as presented next.

D. The Analysis of Low Priority Traffic

1) Modeling the LPT Process: The analysis of LPT can be derived in a similar and parallel way as its HPT counterpart. The main difference is that the number of slots per subframe allocated to LPT is dictated by the dynamic behavior of the HPT occurring in the same subframe. A link between both types of traffic is established based on the pseudo-aggregated process defined above, hence simplifying the analysis of LPT. Intuitively, this approach could loose the "synchronization" or the existing coupling between HPT and LPT. However, the rationale behind our analysis lies on the fact that this approach largely captures the behavior of LPT, which utilizes the remaining capacity, i.e., a number of slots in the same subframe that are not allocated to the HPT transmissions.

Correspondingly, in a parallel way to (2) and omitting the r.v. of HPT, we have

$$\begin{aligned}\Pr\big((W_2,U_2,N_2)_{t+1} = (\nu,v,j) | (W_2,U_2,N_2)_t = (\mu,u,i)\big)\\ = P_{\mu,u,j;\nu,v,j}. \quad (13)\end{aligned}$$

In (13), the same notations as in (2) have been introduced, *but now it is referred to LPT*, i.e., $(w,m,n)_{2,t} \equiv (\mu,u,i)$ and $(w,m,n)_{2,t+1} \equiv (\nu,v,j)$. However, the difference between (2) and (13) is that in the LPT case the transitions $w_{2,t} \equiv \mu \to w_{2,t+1} \equiv \nu$ and $m_{2,t} \equiv u \to m_{2,t+1} \equiv v$ evolve independently of each other, whereas the second transition is dictated by the behavior of the HPT process. Accordingly, in contrast to the HPT process which is represented by 2 random variables, 3 random variables are needed to identify the Markov chain of the LPT process.

The evaluation of (13) is similar to the counterpart model of HPT. First, for the transition $(\mu,i) \to (\nu,j)$, we consider the packets that have been successfully transmitted, i.e., the departure process (see (4)).

$$D^{i,u}_{s_t,c_t}(p_t) = \sum_{z=s_t}^{i} B^i_z(p_t) R^{z,u}_{s_t,c_t}, \quad (14)$$

where $R^{z,u}_{s_t,c_t}$ has the same meaning as in (4). Furthermore, in parallel to (5), we have the following expression for LPT.

$$P_{\mu,i;\nu,j} = \sum_{(h,s,c)_t \in \Omega_2} D^{i,u}_{s_t,c_t}(p_t) A^{M_2-i+s_t}_{j-i+s_t}(a_2), \quad (15)$$

where $A^{M_2-i+s}_{j-i+s}(a_2)$ follows the binomial distribution similar to the one presented in (5) but for LPT. In (15), a set, Ω_2, which is defined as $(h,s,c)_t \in \Omega_2$, is the set of values that satisfy the following two conditions,

$$\Omega_2 \overset{\text{def}}{=} \begin{cases} u = h_t + s_t + c_t \\ \quad \neq \max\big(u_{2,min}, \min(\mu, u_{2,max})\big); \\ \nu \quad = \lceil \max(\mu + f((h,s,c)_t), 0) + \hat{\lambda}_t \rceil. \end{cases} \quad (16)$$

To gain clarity in the rest of this paragraph, we have recovered the notations with subscripts in $m_{x,t}$ and $w_{x,t}$ where $x = 1$ for HPT and $x = 2$ for LPT, respectively. Consider that, at subframe t, we have $m_{2,t} = U - m_{1,t}$. In the next subframe $t+1$, the gNB will allocate $m_{2,t+1} = U - m_{1,t+1}$ with probability $\hat{P}_{m_{1,t},m_{1,t+1}}$ given by (12), i.e., by the transition probabilities of the pseudo-aggregated Markov chain. In other words, the number of slots per subframe allocated to LPT by the gNB in the next subframe $t+1$ only depends on the transitions $m_{1,t} \to m_{1,t+1}$ of HPT. We highlight this fact with the inequality of (16). Then, the equivalent expression of (3) for LPT devices becomes,

$$\begin{aligned}\nu &= \lceil \max(w_{2,t} + f((h,s,c)_{2,t}), 0) + \hat{\lambda}_{2,t} \rceil;\\ v &= m_{2,t+1} = U - m_{1,t+1};\\ p_{2,t+1} &= \min\left(\frac{m_{2,t+1}}{w_{2,t+1}}, 1\right) = \min\left(\frac{v}{\nu}, 1\right). \quad (17)\end{aligned}$$

By combining (15) with the transition probabilities (12) of the pseudo-aggregated Markov chain, we obtain the transition

probabilities corresponding to the Markov chain for LPT,

$$P_{\mu,u,i;\nu,v,j} = P_{\mu,i;\nu,j}\hat{P}_{U-u,U-v}. \tag{18}$$

Note that $P_{\mu,i;\nu,j}$ in (18) refers to (15), i.e., it is meant for LPT and it differs from (5) which refers to HPT. Through (18), we claim that 1) the product of both probabilities reflects the 'independence' in the treatment of both types of traffic; and 2) the correlation or dependence between HPT and LPT is taken into account with the transition probabilities (12) of the pseudo-aggregated Markov chain.

The steady state regime for LPT is given by the stochastic row vector $\boldsymbol{\pi}$ ($\boldsymbol{\pi}\mathbf{e} = \mathbf{1}$) derived by solving the linear equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ with $\boldsymbol{\pi} = \{\pi_{\mu,u,i}\}$ and $\mathbf{P} = \{P_{\mu,u,i;\nu,v,j}\}$. Then, $\pi_{\mu,u,i}$ is the steady state probability at the start of an arbitrary subframe of the number of active devices estimated by the gNB being μ, the number of active devices being i, and the slot allocated to the LPT flows being u.

2) Throughput, Access Delay, and Packet Loss Probability for LPT: Similar to the HPT case, we assess the performance of LPT with respect to the same four parameters defined above. In particular, the throughput per subframe for LPT is obtained as follows

$$\begin{aligned}\gamma_2^{sf} &= \sum_{s_t=1}^{u_{2,max}} s_t \sum_{(\mu,u,i)\in\Delta_2} \pi_{\mu,u,i} \sum_{c_t\in\mathcal{C}} D^{i,u}_{s_t,c_t}(p_t) \\ &= \sum_{(\mu,u,i)\in\Delta_2} i p_t \left(1-\frac{p_t}{u}\right)^{i-1} \pi_{\mu,u,i}. \end{aligned} \tag{19}$$

The second equality in (19) is derived in the same way as what is deduced in (7).

In a similar way as for (9), the expression of the throughput per slot for LPT is obtained as

$$\gamma_2^{slot} = \gamma_2^{sf} / \sum_{(\mu,u,i)\in\Delta_2} u\pi_{\mu,,i}. \tag{20}$$

Similar to (10), the access delay for LPT is obtained by

$$d_2^{sf} = \sum_{(\mu,u,i)\in\Delta_2} i\pi_{\mu,i}/\gamma_2^{sf}. \tag{21}$$

Lastly, similar to the expression in (11), the packet loss probability for LPT is defined as

$$\theta_2 = (M_2a_2 - \gamma_2^{sf})/M_2a_2. \tag{22}$$

V. Simulations and Numerical Results

This section presents the numerical results obtained from both the analytical model and discrete-event based simulations. The proposed DSA-GF scheme has been implemented based on a custom-built simulator constructed in MATLAB which mimics the behavior of the scheme. Extensive simulations are performed under various configurations. The results with respect to the four performance parameters defined in Sec. IV, i.e., throughput per subframe/slot (in terms of number of packets per subframe/slot), access delay for the successfully transmitted packets, and packet loss probability, are presented below. Two other GF access schemes, known as *complete sharing* and *GF reactive* (see Subsec. V-E), have also been implemented and the performance of these three schemes is compared with each other therein. The applicability of DSA-GF to two other numerologies is validated in Subsec. V-F.

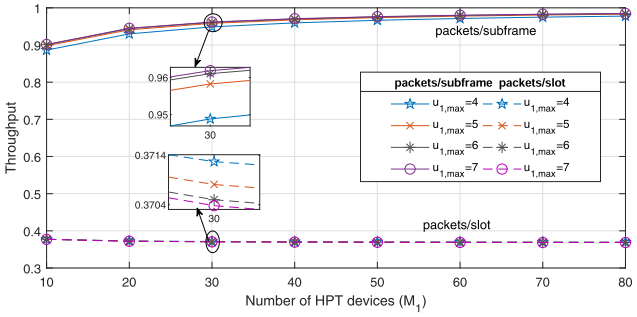

Fig. 3. Throughput of HPT when $M_1a_1 = 1$ and $u_{1,min} = 1$, $u_{1,max} = 4, 5, 6, 7$.

A. Simulation Setup and Model Validation

Consider an NR cell with all three GF schemes, i.e., DSA-GF, complete sharing, and GF reactive, enabled. As mentioned earlier in Subsec. II-B, although the total number of mMTC devices covered by the cell could be large, the number of devices attempting for channel access at a given subframe is considered to be rather limited. In this study, we consider that the device population for each type varies from 10 up to 100, coupled with different activation probabilities. The offered traffic intensities are represented by M_1a_1 and M_2a_2 (in terms of packets per subframe) for HPT and LPT, respectively. Except Subsec. V-F which considers numerology $\beta = 2$ and $\beta = 4$, we adopt $\beta = 3$ for our performance evaluations in all simulations presented below. Note that no matter there are $U = 4, 8$, or 16 slots in each subframe, all of them are available for GF transmissions (discussed in Subsec. II-A). For these simulations, we set $u_{2,min} \geq 1$. That is, $u_{1,max} \leq U - 1$. The other parameters like $u_{1,min}$ are configured in favor of HPT performance with the concrete values shown in each figure caption or the corresponding explanations. For all simulation results presented below, we report the average values obtained from multiple runs of simulations.

The accuracy of the developed Markov model is verified through extensive simulations. Under all network configurations, the analytical and simulation results coincide with each other so tightly that the curves obtained from these two methods are largely overlapping. As such, the accuracy of the developed Markov model is validated. As two examples, we plot separately in Figs. 4 and 5 the curves obtained from both analysis and simulations. For the sake of illustration clarity, we do not plot both sets of results in other figures.

B. HPT Performance With Variable Device Population

As explained earlier, the performance of HPT is independent of that of LPT. Accordingly, we evaluate the performance of HPT by varying the number of HPT devices M_1 and the activation probability a_1 while keeping the offered traffic constant as $M_1a_1 = 1$. Keep in mind that the actual number of slots allocated to HPT per subframe is governed by the DSA-GF scheme where both $u_{1,min}$ and $u_{1,max}$ are tunable parameters but they do not vary on a subframe or frame basis.

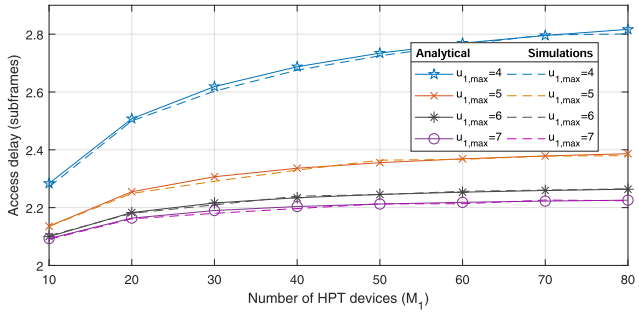

Fig. 4. Access delay of HPT when $M_1a_1 = 1$ and $u_{1,min} = 1, u_{1,max} = 4, 5, 6, 7$.

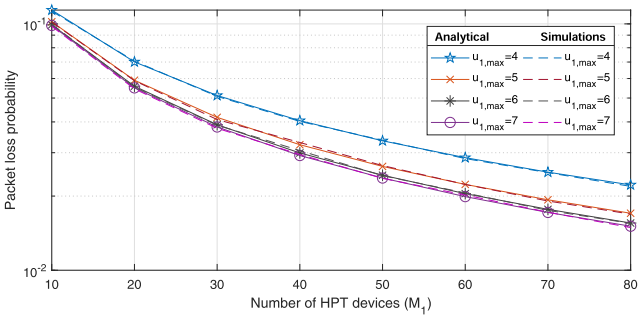

Fig. 5. Packet loss probability of HPT when $M_1a_1 = 1$ and $u_{1,min} = 1, u_{1,max} = 4, 5, 6, 7$.

The performance of HPT in terms of throughput per slot and per subframe is illustrated in Fig. 3 where $u_{1,min} = 1$ and $u_{1,max} = 4, 5, 6,$ or 7 respectively. It is clear that the achieved throughput per slot for these configured $u_{1,min}$ and $u_{1,max}$ values is slightly higher than the maximum throughput for slotted ALOHA, i.e., $1/e \approx 0.3679$, which is obtained with an infinite population. This is because the number of devices in our simulations is finite. For instance, for a fixed value of $M_1 = (10, 20, 30, \ldots, 70, 80)$, the resulting successful probability takes the values as $(0.3874, 0.3774, 0.3741, \ldots, 0.3705, 0.3701)$ respectively, indicating a slightly lower successful probability which approaches the *throughput per slot* for slotted ALOHA as M_1 increases.

On the other hand, we observe that, as M_1 becomes larger, 1) the achieved throughput per subframe increases monotonically towards a maximum value and 2) these values are much higher than that of the throughput per slot. For 1), note first that when a collision occurs, the corresponding packet remains pending to the next subframe. This behavior contributes to an increased number of packets awaiting to be transmitted. Furthermore, the devices that succeeded in the current subframe will also generate with probability a_1 one packet ready for transmission in the next subframe. The net effect is that, when M_1 increases, the mean value of the number of backlogged packets increases slightly and more slots will be allocated to HPT, resulting in thus higher throughput per subframe. For 2), it is because multiple slots within the same subframe are utilized by HPT devices. For example, when $(u_{1,min}, u_{1,max}) = (1, 4), (1, 5), (1, 6),$ or $(1, 7)$ and $M_1 = 30$, there are on average 2.5555, 2.5842, 2.5939, or 2.5969 number of slots allocated to HPT respectively. Indeed, this result is in accordance with the relationship between per subframe and per slot throughput expressed in (9).

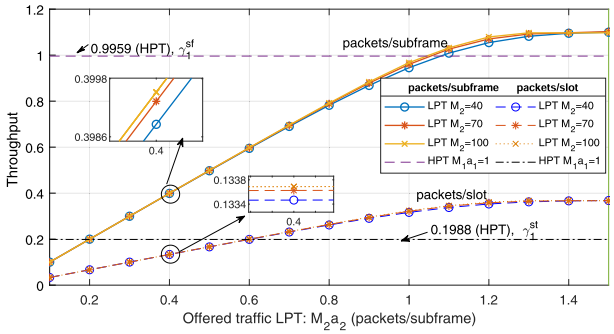

Fig. 6. Throughput per subframe and per slot for LPT under various offered traffic M_2a_2 where $M_1a_1 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

Furthermore, the obtained access delay and packet loss probability performance is depicted in Figs. 4 and 5 respectively. With a larger device population, DSA-GF needs more subframes to accommodate HPT packets, leading to an increasing trend for access delay. On the other hand, the achieved access delay decreases significantly with a larger $u_{1,max}$ due to the fact that more slots are available for HPT. With a larger $u_{1,max}$ value, a competing device obtains a higher probability of selecting a unique slot for successful transmission, resulting in a lower delay. In Fig. 5, it is shown that a larger $u_{1,max}$ leads to a lower loss probability. With a lager number of HPT devices, the activation probability decreases in order to maintain constant offered traffic. Hence, the impact of buffer limitation is reduced. Correspondingly, the packet loss probability decreases with an increasing M_1. Moreover, one may notice a decreasing gap between two adjacent curves in these two figures with a larger $u_{1,max}$. This is because the performance acceleration rate declines when $u_{1,max}$ increases.

C. LPT Performance With Variable Offered Traffic

To evaluate the performance of LPT devices, we vary the offered traffic load by LPT devices M_2a_2 given that $M_1a_1 = 1$ with $M_1 = 100$ and $(u_{1,min}, u_{1,max}) = (5, 7)$. Under such traffic conditions, the average number of slots allocated to HPT is $\overline{m}_{1,t} \approx 5.0115$. Accordingly, LPT obtains $\overline{m}_{2,t} \approx 2.9885$ slots on average. Figs. 6-8 illustrate the performance in terms of the four parameters defined above.

Fig. 6 illustrates the obtained throughput per subframe/slot for LPT as M_2a_2 increases. Initially, the throughput per subframe increases linearly with M_2a_2 and gradually the behavior reaches a stable limit when the network approaches saturation. A similar trend is observed for the behavior of throughput per slot. The reason is as follows. Since our scheme follows the principle of ALOHA, the highest throughput that can be achieved is $\overline{m}_{2,t}/e = 2.9985/2.7183 = 1.1031$. Therefore, as long as $M_2a_2 < 1.1031$, LPT will exhibit a linear throughput response corresponding to the offered LPT traffic load. The more we increase the offered traffic M_2a_2, the closer we are approaching to the theoretical limit. When

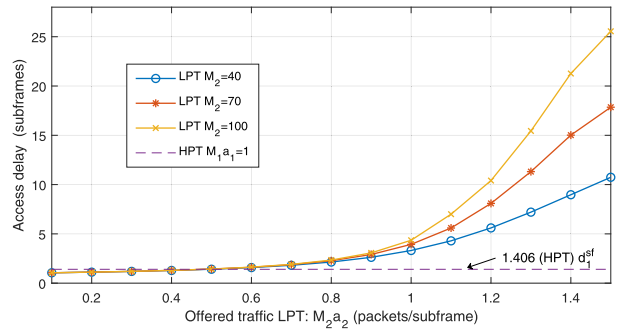

Fig. 7. Access delay for LPT under various offered traffic M_2a_2 where $M_1a_1 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

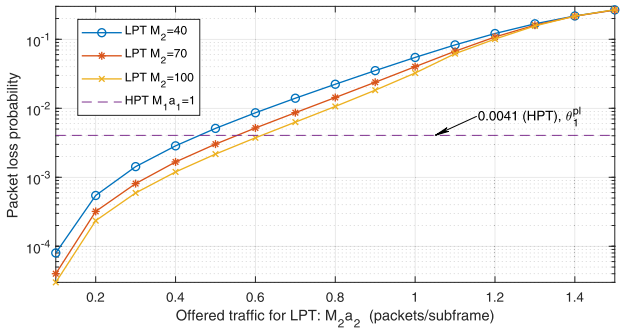

Fig. 8. Packet loss probability for $M_2 = 40, 70, 100$ and various M_2a_2 values in LPT where $M_1a_1 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

M_2a_2 approaches the value of 1.1031, the curve starts to bend and in an asymptotic way it reaches the maximum throughput value.

Fig. 7 reveals the access delay for successful LPT packet transmissions. When the LPT traffic load increases, a higher number of collisions occur, causing packets to wait for a longer period of time in the buffer. Accordingly, the average delay increases. Recall that devices are equipped with a buffer of unit size. When a new packet arrives and finds the buffer full, it is rejected. This is indeed the implementation of the packet rejection mechanism [26]. It causes a higher packet loss probability when the offered LPT traffic increases, as shown in Fig. 8.

Although a loss probability higher than 1% is out of interest, it is worth studying the asymptotic behavior of the loss probability for LPT, i.e., when $a_2 \rightarrow 1$. Under the principle of blocking a new packet when the buffer is occupied, the asymptotic loss probability can be expressed as the fraction $(M_2 - \overline{m}_{2,t}e^{-1})/M_2$. It becomes 0.9725, 0.9843, and 0.9989 for $M_2 = 40, 70$, and 100 devices, respectively. These values match perfectly the results provided by the Markov model. Furthermore, due to the introduction of a single size buffer, the asymptotic behavior of the delay performance can be derived as follows. For a given number of LPT devices M_2, when $a_2 \rightarrow 1$ (which is the condition for saturation), all M_2 buffers are full, each with one packet ready for transmission at the beginning of each subframe. Since the mean number of successful transmissions per subframe is given by $\overline{m}_{2,t}e^{-1}$,

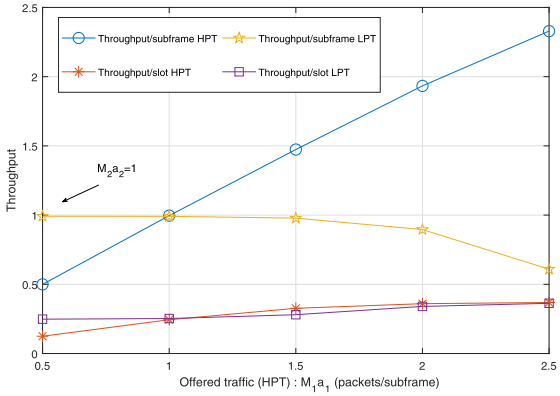

Fig. 9. Throughput per subframe and per slot for HPT/LPT under various offered HPT traffic M_1a_1 where $M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

the mean number of subframes that a given packet has to wait in its buffer is given by $M_2e/\overline{m}_{2,t}$. Following the same illustrative example given above with $M_2 = 40, 70$, and 100 LPT devices, the obtained access delay becomes 36.3832, 63.6706, and 90.9581 subframes, respectively. The same as above, these results are in precise agreement with the ones obtained from the Markov model, as expressed in (21).

D. Impact of Offered HPT Traffic Load on HPT/LPT Performance

To assess the impact of offered HPT traffic load on the performance of both HPT and LPT, we perform two sets of simulations, with a combination of constant or variable traffic loads for HPT or LPT respectively. As already discussed above, the performance of HPT remains constant throughout the whole range of the M_2a_2 variations. In other words, these results confirm that HPT's performance remains intact regardless of the variations of the injected LPT traffic load.

On the other hand, LPT's performance will be dominated by HPT traffic intensity since LPT can only occupy the remaining slots in the same subframe that are not allocated to HPT packets. As shown in Fig. 9, while the HPT throughput per subframe increases linearly as M_1a_1 increases (until the saturation point), the LPT throughput per subframe has to sacrifice its performance. With respect to the performance of DSA-GF in terms of access delay and packet loss probability shown in Figs. 10-11, it is convincing that HPT achieves better performance than LPT does.

E. Performance Comparison With Complete Sharing and GF Reactive

First of all, note that no traffic classification is introduced in these two reference schemes. Before presenting the results, let us outline briefly the principles of these two schemes as follows. 1) Complete sharing works similarly as the proposed scheme. However, the slot allocation and data transmission process in complete sharing does not enable any priorities. Instead of treating HPT and LPT separately, a single class of arrivals will compete for access in all available slots in each subframe. The packet transmission probability is dynamically adjusted on a subframe-by-subframe basis following the same pseudo-Bayesian estimation process. 2) The GF reactive

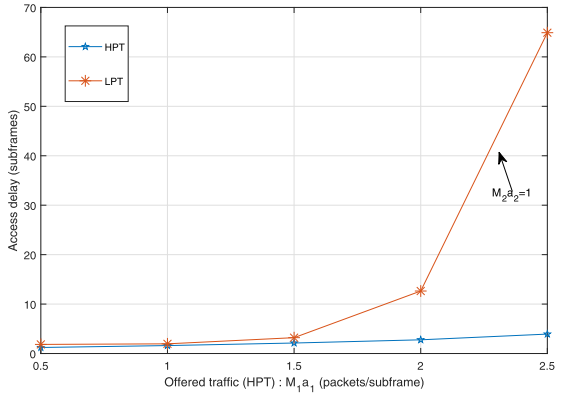

Fig. 10. Access delay for HPT/LPT under various offered HPT traffic M_1a_1 where $M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

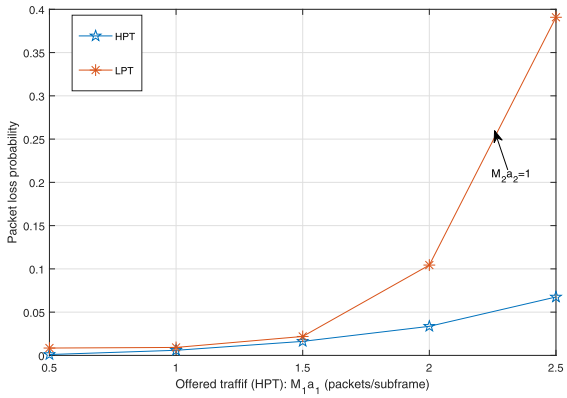

Fig. 11. Packet loss probability for HPT/LPT under various offered HPT traffic M_1a_1 where $M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

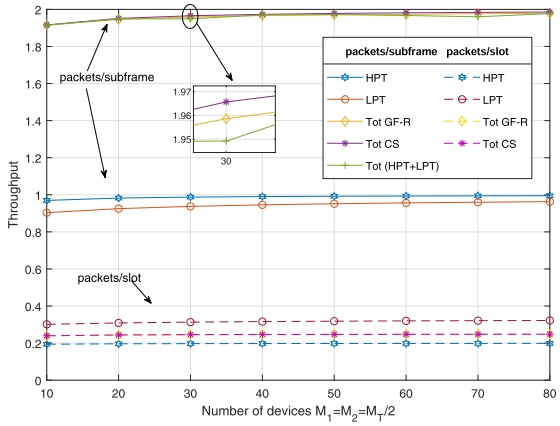

Fig. 12. Throughput comparison with GF reactive and complete sharing ($M_1a_1 = M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$).

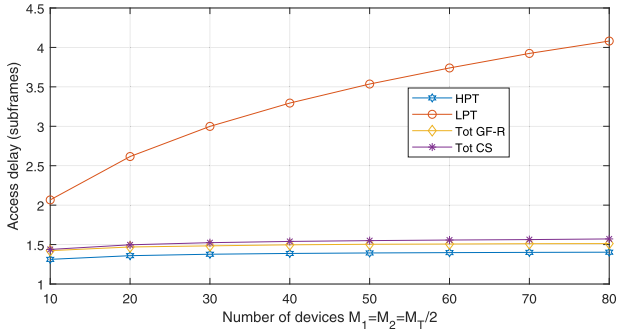

Fig. 13. Access delay comparison with GF reactive and complete sharing ($M_1a_1 = M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$).

scheme discussed in Subsec. I-A. No permission probability exists in this scheme, i.e., a failed transmission attempt shall for sure try again in the next subframe. To avoid the situation that an 'unlucky' packet could attempt to transmit forever, a retry limit of 10 is configured in our simulations for GF reactive. In this study, we do not include any proactive GF scheme due to the consideration that high collision could occur for GF proactive with $K > 1$ since two or more packet replicas from the same device will compete for slot access inside the same subframe in GF proactive schemes.

The numerical results obtained from the three studied schemes are compared in Figs. 12-16 where GF-R and CS in the legends stand for GF reactive and complete sharing, respectively. With respect to the achieved throughput per subframe, the values obtained from all three schemes (for DSA-GF, it is meant for the sum of HPT and LPT throughput) are very close to each other (the curves for throughput per slot for GF-R and CS are indeed overlapping). This is because the offered traffic in all cases is high enough so that the highest slot utilization has been reached. Thanks to the privilege given to HPT with $(u_{1,min}, u_{1,max}) = (5, 7)$, the throughput per subframe for HPT exhibits the highest values, at the cost of reduced LPT throughput.

When it turns to access delay, one may observe a similar trend. That is, HPT achieves the lowest delay across the whole range of device populations, obtained after a small sacrifice of LPT's delay. On the other hand, the reason that GF reactive reaches lower access delay than complete sharing does is that more access opportunities are given to GF reactive devices due to the fact that there is no permission probability as well as the constraint of the retry limit.

Let us now compare the performance in terms of packet loss probability for those three schemes. It is convincing that HPT under the DSA-GF scheme achieves the lowest packet loss probability thanks to its access privilege. This result reveals once again the benefit brought by introducing priority for dynamic slot allocation. On the other hand, when comparing the packet loss probabilities for complete sharing and GF reactive, the results meet our intuition that complete sharing performs better. This is because complete sharing imposes access control via a permission probability when collisions are detected in the previous subframe, thus limiting the number of competing devices in the current subframe. Given that the number of slots in each subframe is fixed, the lesser the number of competing devices, the lower the packet loss.

F. Applicability of DSA-GF to Numerology $\beta = 2$ and $\beta = 4$

Considering that the subframe duration is fixed for all numerologies as 1 ms, we keep *the offered traffic per subframe constant*, however, with different combinations of device populations and activation probabilities. More specifically, for $\beta = 4$, we configure four sets of device populations

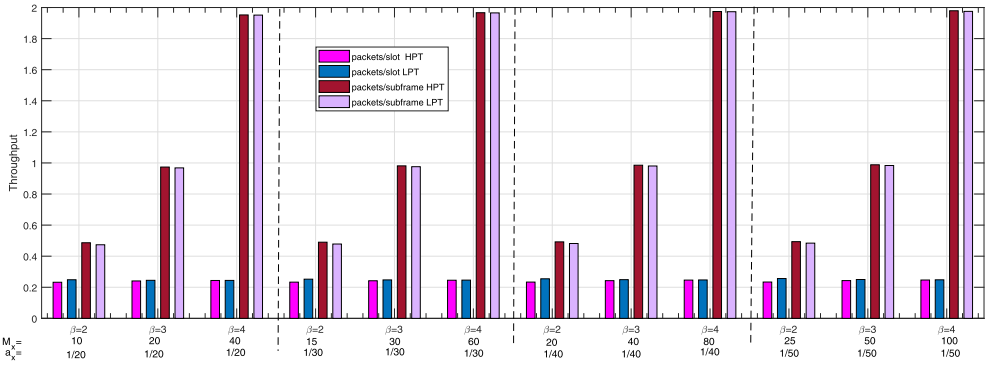

Fig. 14. Applying DSA-GF to three numerologies, $\beta = 2$, 3, and 4 respectively: Throughput per subframe/per slot.

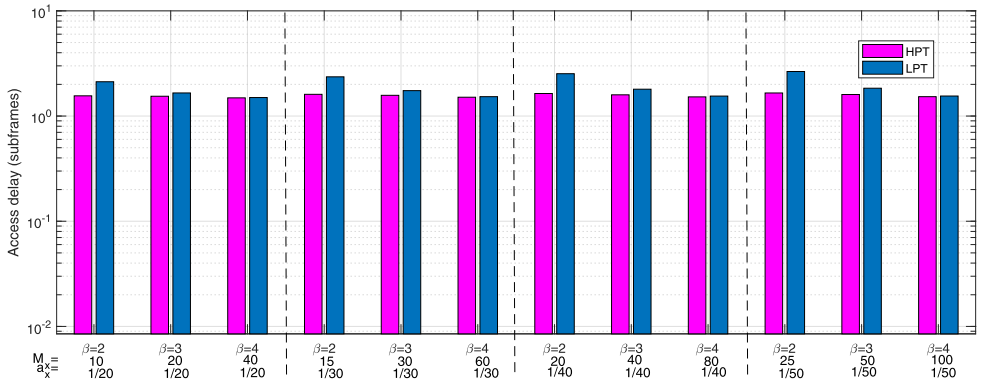

Fig. 15. Applying DSA-GF to three numerologies, $\beta = 2$, 3, and 4 respectively: Access Delay.

as $M_1 = M_2 = 40, 60, 80$, and 100 for HPT and LPT, each set coupled with an activation probability of $a_1 = a_2 = 1/20, 1/30, 1/40$, and 1/50 respectively. In this way, the offered traffic per subframe equals to $M_x a_x = 2$ for each type of traffic, i.e., 2 packets/subframe. For $\beta = 3$ and 2, devices are split into 2 and 4 groups respectively due to resource allocation explained in the next paragraph. Accordingly, we have $M_x a_x = 1$ and 0.5 per subframe for $\beta = 3$ and 2, since there are 2 and 4 parallel subframes respectively. Detailed configurations on M_x and a_x for each numerology can be found in Figs. 14-17.

To accommodate the offered traffic, the gNB can either allocate one subframe for $\beta = 4$, or two parallel subframes over the frequency domain for $\beta = 3$. This configuration is reasonable since the subcarrier spacing in $\beta = 4$ is twice as much as in $\beta = 3$. Following the same logic, there will be four parallel subframes over the frequency domain when $\beta = 2$ is adopted. As such, the total number of slots in all three numerologies is the same as 16 slots, however, grouped into 4, 8, or 16 slots per subframe for $\beta = 2, 3$, or 4 respectively. Accordingly, we configure the tunable parameters as $(u_{1,min}, u_{1,max}) = (2, 3), (4, 6)$, and $(6, 12)$ respectively.

In Figs. 14-17, the performance of the DSA-GF scheme is illustrated as a histogram for the four sets of device population and activation probability configurations respectively. As shown in Fig. 14, the achieved throughput per slot remains almost constant in all three numerologies. On the other hand, the throughput per subframe is doubled when a higher-level numerology is adopted. As expected, the achieved throughput per subframe reaches 0.4865, 0.9736, and 1.9522 for $\beta = 2$, 3, and 4 respectively. This is due to the fact that more slots per subframe are available with a higher-level numerology. When observing the achieved packet loss probability in Fig. 17, it is evident that a higher-level numerology leads to lower packet loss. This is because more slots are aggregated in one subframe as resources for devices to share when a higher-level numerology is adopted. In addition, for a fixed numerology, the probability of packet loss decreases as M_2 increases, since a_2 reduces when $M_2 a_2$ is constant.

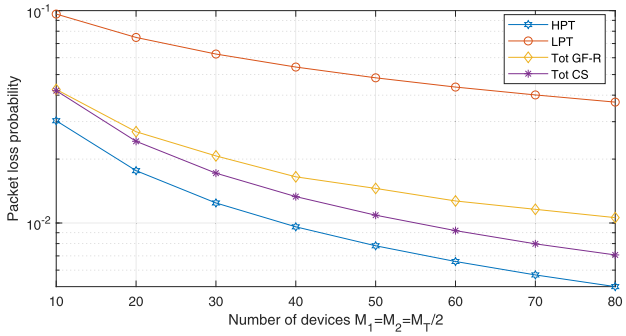

Fig. 16. Packet loss probability comparison with GF reactive and complete sharing ($M_1 a_1 = M_2 a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$).

Finally, let us compare the performance of HPT and LPT. Although the offered HPT and LPT traffic is the same, the achieved throughput per subframe for HPT is slightly higher than that of LPT. As a consequence, better performance

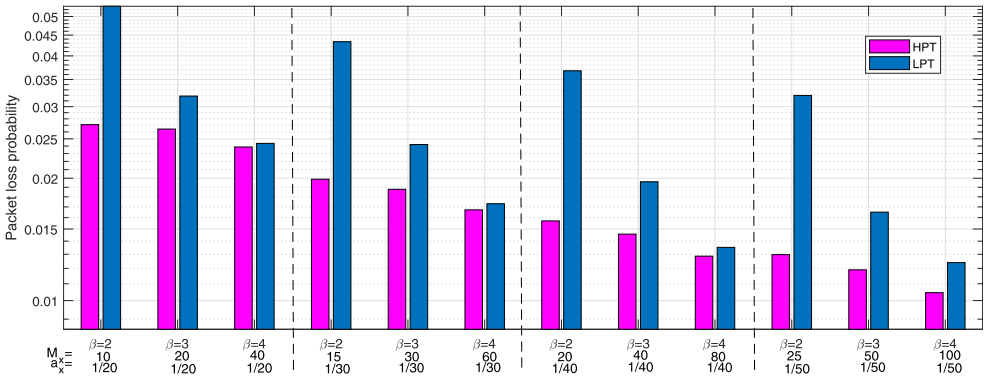

Fig. 17. Applying DSA-GF to three numerologies, $\beta = 2$, 3, and 4 respectively: Packet loss probability.

has been achieved for HPT than for LPT in terms of access delay and packet loss probability. This benefit is brought by the priority enabled DSA-GF adaptive algorithm which performs well in all studied numerologies and network configurations. As shown in Fig. 15, the access delay for HPT is slightly decreasing with a higher-level numerology whereas (much) higher delays are experienced by LPT. The same trend applies to the packet loss probability performance as well, as illustrated in Fig. 17. Obviously, better performance for HPT can be achieved by increasing the values of $u_{1,min}$ and $u_{1,max}$, at the expenses of slight penalties for the performance of LPT.

G. Further Discussions

The DSA-GF scheme considers the distinctive characteristics of two co-existing traffic types in a 5G NR network. Although it is unavoidable to sacrifice the performance of LPT in order to ensure the high performance of HPT, serious access congestion for LPT can be avoided or minimized through proper parameter configurations. In general, there is a tradeoff between the performance of these two traffic classes when deciding the values for $u_{1,min}$ and $u_{1,max}$.

Furthermore, $u_{1,min}$ and $u_{1,max}$ are two configurable parameters. Their values are considered to be pre-configured based on gNB's observations as well as service requirements and do not change over a short term (i.e., neither on a subframe-by-subframe nor on a frame-by-frame basis).

VI. Conclusion and Future Work

This paper presents a priority enabled GF access and data transmission scheme which enables dynamic slot allocation for heterogeneous GF traffic in 5G NR networks. Based on the NR frame structure, the proposed scheme grants access privilege for slot occupancy to high priority traffic based on traffic estimation and the observed transmission status and allocates the remaining slots in each subframe to low priority traffic. While the performance of high priority traffic is guaranteed through proper configuration of relevant parameters, low priority traffic also enjoys satisfactory performance. Furthermore, the precedence of high priority traffic and the dependence between two heterogeneous traffic classes are captured through a Markov model which derives a pseudo-aggregated process to bridge the aforementioned dependency. Through both analysis and simulations, we demonstrate the elegance and effectiveness of the scheme with respect to four performance parameters, i.e., throughput per subframe and per slot, access delay, and packet loss probability, as well as its applicability. To achieve optimal performance, proper parameter tuning is needed based on network setup and traffic conditions. How to adjust $u_{1,min}$ and $u_{1,max}$ configurations periodically, e.g., in the order of seconds, over a long term, or reactively depending on real-time traffic measurements, and how to deal with estimation error are left as our future work.

References

[1] I. Leyva-Mayorga, C. Stefanovic, P. Popovski, V. Pla, and J. Martinez-Bauset, *Random Access for Machine-Type Communications*. Hoboken, NJ, USA: Wiley, 2019.

[2] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, Feb. 2014.

[3] *NR; Physical Layer Procedures for Control*, document TS 38.213 v16.3.0, 3GPP, Sep. 2020.

[4] *NR; Physical Layer Procedures for Data*, document TS 38.214 v16.3.0, 3GPP, Sep. 2020.

[5] *Study on New Radio (NR) Access Technology*, document TS 38.912 v16.0.0, 3GPP, Jul. 2020.

[6] N. Abramson, "The alohanet-surfing for wireless data," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 21–25, Dec. 2009.

[7] *Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC)*, document TR 38.824 R16, v16.0.0, 3GPP, Mar. 2019.

[8] A. T. Abebe and C. G. Kang, "Comprehensive grant-free random access for massive & low latency communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[9] N. H. Mahmood, R. Abreu, R. Bohnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 607–612.

[10] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 182–185, Apr. 2018.

[11] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novel analytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Mar. 2019.

[12] Z. Ding, R. Schober, P. Fan, and H. V. Poor, "Simple Semi-Grant-Free transmission strategies assisted by non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4464–4478, Jun. 2019.

[13] J. Ding, D. Qu, and J. Choi, "Analysis of non-orthogonal sequences for grant-free RA with massive MIMO," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 150–160, Jan. 2020.

[14] E. Casini, R. De Gaudenzi, and O. R. Herrero, "Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access schemefor satellite access packet networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1408–1419, Apr. 2007.

[15] V. Casares-Giner, J. Martinez-Bauset, and C. Portillo, "Performance evaluation of framed slotted ALOHA with reservation packets and succesive interference cancelation for M2M networks," *Comput. Netw.*, vol. 155, pp. 15–30, May 2019.
[16] F. Lazaro, C. Stefanovic, and P. Popovski, "Reliability-latency performance of frameless ALOHA with and without feedback," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6302–6316, Oct. 2020.
[17] J.-F. Frignon and V. C. M. Leung, "A pseudo-Bayesian ALOHA algorithm with mixed priorities," *Wireless Netw.*, vol. 7, no. 1, pp. 55–63, Jan. 2001.
[18] R. Rivest, "Network control by Bayesian broadcast," *IEEE Trans. Inf. Theory*, vol. 33, no. 3, pp. 323–328, May 1987.
[19] M. H. Habaebi, B. M. Ali, and M. R. Mukerjee, "Wireless adaptive framed pseudo-Bayesian ALOHA (AFPBA)," *Int. J. Wireless Inf. Netw.*, vol. 8, no. 1, pp. 49–59, Jan. 2001.
[20] *NR; Physical Channels and Modulation*, document TS38.211 R16, v16.1.0, 3GPP, Mar. 2020.
[21] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 1, no. 4, pp. 179–189, Jan. 2020.
[22] S.-Y. Lien, S.-L. Shieh, Y. Huang, B. Su, Y.-L. Hsu, and H.-Y. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.
[23] *Service Requirements for the 5G System*, document TS 22.261 R18, v18.0.0, 3GPP, Sep. 2020.
[24] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–7.
[25] T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Priority-based initial access for URLLC traffic in massive IoT networks: Schemes and performance analysis," *Comput. Netw.*, vol. 178, Sep. 2020, Art. no. 107360.
[26] B. Doshi and H. Heffes, "Overload performance of several processor queueing disciplines for the M/M/1 queue," *IEEE Trans. Commun.*, vol. 34, no. 6, pp. 538–546, Jun. 1986.
[27] G. Rubino and B. Sericola, "Sojourn times in finite Markov processes," *J. Appl. Probab.*, vol. 26, no. 4, pp. 744–756, Dec. 1989.
[28] V. Casares-Giner, V. Sempere-Payá, and D. Todolí-Ferrandis, "Framed ALOHA protocol with FIFO-blocking and LIFO-push out discipline," *Netw. Protocols Algorithms*, vol. 6, no. 3, pp. 82–102, Aug. 2014.

Thilina N. Weerasinghe received the B.Sc. degree (Hons.) in engineering from the University of Ruhuna, Sri Lanka, in 2009, and the M.Sc. degree in information and communication technology (ICT) from the University of Agder (UiA), Norway, in 2015, where he is currently pursuing the Ph.D. degree with the Department of ICT. He worked in telecommunication related projects in Sri Lanka and Maldives and also worked as a Lecturer with the Department of Electrical and Information Engineering, University of Ruhuna. His current research interests include mobile and wireless communications, including ultra-reliable low latency communication, massive MTC, the Internet of Things, and ad-hoc and sensor network MAC protocols.

Vicente Casares-Giner (Life Member, IEEE) received the telecommunication engineering degree from the Escuela Técnica Superior de Ingenieros en Telecomunicación-Universidad Politécnica de Madrid (ETSIT-UPM) in October 1974 and the Ph.D. degree in telecommunication engineering from the ETSIT-Universidad Politécnica de Catalunya (ETSIT-UPC), Barcelona, in September 1980. He was an Assistant Professor in 1974, an Associate Professor in 1985, and a Full Professor in 1991. During the period of 1974–1983, he worked on problems related to signal processing, image restoration, and propagation aspects of radio-link systems. In the first half of 1984, he was a Visiting Scholar with the Royal Institute of Technology (KTH), Stockholm, Sweden, dealing with digital switching and concurrent programming for Stored Program Control (SPC) telephone systems. From September 1994 until August 1995, he was a Visiting Scholar with WINLAB, Rutgers University, USA, working with random access protocols in wireless networks, wireless resource management, and land mobile trunking systems. In 1990, he worked in traffic and mobility models in several EU projects. Since September 1996, he has been with the Universitat Politècnica de València (UPV), Valencia, Spain. From 2000 to 2010, he had been involved in multiple Spanish national and EU projects. During the first half of 2020, he was a Visiting Professor with the Department of Information and Communication Technology (ICT), University of Agder (UiA), Norway. His research interests include performances evaluation of wireless systems, in particular random access protocols, system capacity and dimensioning, mobility management, cognitive radio, the IoT, and wireless sensor networks. In February 2021, he was promoted to Professor Emeritus. He has served as the General Co-Chair for ISCC 2005 and NGI-2006, and as a TPC Member for many conferences and workshops (Networking 2011, GLOBECOM 2013, ICC 2015, and VTC 2016).

Indika A. M. Balapuwaduge (Member, IEEE) received the B.Sc. degree (Hons.) in engineering from the University of Ruhuna, Sri Lanka, in 2008, and the M.Sc. and Ph.D. degrees in information and communication technology (ICT) from the University of Agder (UiA), Norway, in 2012 and 2016, respectively. His Master's thesis was awarded as the Best Master's thesis in ICT, UiA, in 2012. He spent one year as an Engineer at Huawei Technologies, Sri Lanka, from October 2008 to August 2009, and he worked as a Lecturer with the Department of Electrical and Information Engineering, University of Ruhuna, from August 2009 to August 2010. He is currently a Post-Doctoral Research Fellow with the Department of ICT, UiA. His current research interests include mobile and wireless communications, including cognitive radio networks, ultra-reliable communication, dependability analysis, massive MTC, the Internet of Things, and modeling and performance analysis of modern communications systems and networks.

Frank Y. Li (Senior Member, IEEE) received the Ph.D. degree from the Department of Telematics (now Department of Information Security and Communication Technology), Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2003. He was a Senior Researcher with the UniK-University Graduate Center (now Department of Technology Systems), University of Oslo, Norway, before joining the Department of Information and Communication Technology, University of Agder (UiA), Norway, in August 2007, as an Associate Professor and then a Full Professor. From August 2017 to July 2018, he was a Visiting Professor with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. During the past few years, he has been an active participant in multiple Norwegian and EU research projects. His research interests include MAC mechanisms and routing protocols in 5G and beyond mobile systems and wireless networks, the Internet of Things, mesh and ad hoc networks, wireless sensor networks, D2D communications, cooperative communications, cognitive radio networks, green wireless communications, dependability and reliability in wireless networks, QoS, resource management, and traffic engineering in wired and wireless IP-based networks, and the analysis, simulation, and performance evaluation of communication protocols and networks. He was listed as a Lead Scientist by the European Commission DG RTD Unit A.03- Evaluation and Monitoring of Programmes in November 2007.