This paper has been accepted for presentation at IEEE GLOBECOM 2021, 07–11 December 2021.

Learning-Based Multiplexing of Grant-Based and Grant-Free Heterogeneous Services with Short Packets

Duc-Dung Tran*, Shree Krishna Sharma*, Symeon Chatzinotas*, and Isaac Woungang†
*Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg
†Department of Computer Science, Ryerson University, Toronto, Canada
Email: {duc.tran, shree.sharma, Symeon.Chatzinotas}@uni.lu, iwoungan@ryerson.ca

***Abstract*—In this paper, we investigate the multiplexing of grant-based (GB) and grant-free (GF) device transmissions in an uplink heterogeneous network (HetNet), namely GB-GF HetNet, where the devices transmit their information using low-rate short data packets. Specifically, GB devices are granted unique time-slots for their transmissions. In contrast, GF devices can randomly select time-slots to transmit their messages utilizing the GF non-orthogonal multiple access (NOMA), which has emerged as a promising enabler for massive access and reducing access latency. However, random access (RA) in the GF NOMA can cause collisions and severe interference, leading to system performance degradation. To overcome this issue, we propose a multiple access (MA) protocol based on reinforcement learning for effective RA slots allocation. The proposed learning method aims to guarantee that the GF devices do not cause any collisions to the GB devices and the number of GF devices choosing the same time-slot does not exceed a predetermined threshold to reduce the interference. In addition, based on the results of the RA slots allocation using the proposed method, we derive the approximate closed-form expressions of the average decoding error probability (ADEP) for all devices to characterize the system performance. Our results presented in terms of access efficiency (AE), collision probability (CP), and overall ADEP (OADEP), show that our proposed method can ensure a smooth operation of the GB and GF devices within the same network while significantly minimizing the collision and interference among the device transmissions in the GB-GF HetNet.**

***Index Terms*—Grant-free NOMA, heterogeneous networks, Q-learning, short-packet communications.**

I. Introduction

With the dramatic increase of wireless devices, such as internet of things (IoT) and machine type communications (MTC) devices, the fifth generation (5G) and beyond wireless networks are required to support massive access over a limited radio spectrum [1]. Furthermore, they are expected to support heterogeneous services, including enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine type communications (mMTC), with different quality-of-service (QoS) requirements [1]. These major challenges have led to the demand for efficient multiplexing and multiple access (MA) technologies. In this regard, grant-free (GF) MA scheme and its coexistence with the conventional grant-based (GB) scheme have emerged as promising enablers for massive access and multiplexing of heterogeneous services in future wireless networks [2–4]. One typical example of GB-GF multiplexing is the eMBB/URLLC multiplexing [3, 4], where the GB access can provide an eMBB service to maximize the throughput, whereas the GF access can be used by the URLLC service to fulfill the strict latency requirement. In the GB access, the devices exchange the scheduling requests with the base station (BS) to be granted different resource blocks for their transmissions [2]. However, with the proliferation of wireless devices, this technique comes at the cost of tremendous signaling overhead and computational resource consumption. In contrast, the GF access enables the devices to select the resource blocks independently and transmit the data directly, hence reducing the signaling overhead and the random access (RA) latency [2]. Nevertheless, it may lead to high collision because multiple devices can select the same resource block. To mitigate this congestion problem, the GF non-orthogonal multiple access (NOMA) could be a promising solution by allowing the devices to select the same resource block at the risk of harmful interference [5].

Congestion control (CC) is a fundamental mechanism for the implementation of heterogeneous networks (HetNets), such as GB-GF HetNet, and the GF access technique. In the recent years, reinforcement learning (RL)-based smart CC method has drawn an important attention [1, 6–11]. RL is a type of machine learning technique, which can enable the agents/devices to interact with the environment in order to learn efficient strategies that maximize the long-term system performance [12]. The most typical RL algorithm is the Q-learning (QL) algorithm, which can be implemented at the user equipment even without an operating model of the environment [12]. Given this context, there have been some works studying CC methods based on QL [6–11], which are briefly summarized as follows.

The work in [6] investigated a QL algorithm to dynamically allocate the resource blocks for devices to mitigate the collision in GF orthogonal multiple access (OMA) systems, where one resource block is used by at most one device. To further improve the resource block access efficiency and reduce the collision, the studies in [7–9] examined the GF NOMA scheme for QL-based congestion control in different scenarios. However, the above works [6–9] only considered homogeneous networks and most of these works did not analyze the system performance with short-packet commu-

nications (SPC). Note that SPC is an enabling paradigm to reduce the latency in 5G and beyond applications [13]. The works in [10, 11] investigated different QL-based congestion avoidance approaches to reduce the collision and improve the system performance for HetNets. Nevertheless, these studies [10, 11] did not consider the SPC-based transmission process and the GB-GF multiplexing.

In this paper, we investigate an SPC-based GB-GF HetNet, where the GB and GF devices try to access the time-slots of a common wireless medium, and a QL-based MA scheme is utilized to support the devices to select the best time-slots for their transmissions. The main contributions of this paper are briefly summarized as follows: i) we propose a QL-based MA protocol for optimal allocation of time-slots to the devices in order to avoid the collision between the GB and GF devices, and reduce the interference of GF devices selecting the same time-slot; ii) we derive approximate closed-form expressions for the average decoding error probability (ADEP) of all devices to characterize the system performance based on the time-slot allocation result achieved by using the proposed QL-based method; iii) we perform the performance evaluation of the considered GB-GF HetNet in terms of access efficiency (AE), collision probability (CP), and overall ADEP (OADEP).

The remainder of the paper is organized as follows. Section II presents the system model in detail. Section III describes the proposed QL-based MA protocol and performance analysis of the GB-GF HetNet with SPC. Section IV presents the obtained numerical results. Finally, Section V concludes this paper.

II. System Model

We consider the GB-GF multiplexing in a time-slotted uplink HetNet, as depicted in Fig. 1. The network consists of one BS, M GB devices, and N GF devices. In this setting, both the GB and GF devices communicate with the BS by attempting to access T available time-slots of a shared wireless medium. Specifically, we investigate the following scenarios: (i) the GB devices communicate with the BS over different granted specific time-slots, where each device uses only one time-slot; (ii) the GF devices randomly select the time-slots for their transmissions. Note that the GF devices are not allowed to use the time-slots granted to the GB devices. In this regard, they compete for $T_{gf} = T - M$ remaining time-slots. To improve the spectrum access efficiency and the number of active GF devices, we assume that the GF devices can select the same time-slots by using the GF NOMA scheme.

Let $\hat{N}_t$ denote the number of GF devices using the same time-slot t. Thus, time-slot t is utilized for the OMA transmission when $\hat{N}_t = 1$, otherwise, it is used for the NOMA transmission, i.e., $\hat{N}_t > 1$. For $\hat{N}_t = 1$, only device n transmits its short packets to the BS in time-slot t, hence, the received signal-to-noise ratio (SNR) of device n can be calculated as

$$\gamma_{n,t} = \gamma_0 |h_{n,t}|^2, \tag{1}$$

where $h_{n,t}$ is the channel coefficient of the link from the GF device n to the BS; $\gamma_0 = P_0/\sigma^2$ is the average transmit SNR;

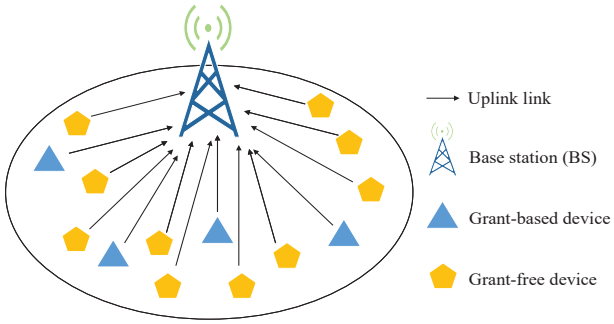

Fig. 1. Illustration of a GB-GF uplink HetNet with SPC.

P_0 is the peak transmit power for each device; and σ^2 is the variance of the additive white Gaussian noise (AWGN).

For $\hat{N}_t > 1$, multiple devices can choose the same time-slot t. In this paper, we investigate a scenario where the devices in time-slot t have different QoS requirements and communicate with the BS by using the QoS-based NOMA (e.g., [14] and the references therein). Specifically, we assume that the device set $\{1, \ldots, \hat{N}_t\}$ is ordered in the descending priority level and the devices with higher priority are decoded earlier at the BS [15]. Given this context, the power allocation coefficients satisfy the condition $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{\hat{N}_t}$ and $\sum_{i=1}^{\hat{N}_t} \alpha_i = 1$. Following the NOMA principle, the BS uses the successive interference cancellation (SIC) to detect any multi-user data. Specifically, it decodes the message of a device by treating the messages of devices with lower priority as noise [14, 15]. Accordingly, the received signal-to-interference-plus-noise ratio (SINR) of device n is expressed as

$$\gamma_{n,t} = \frac{\alpha_n \gamma_0 |h_{n,t}|^2}{I_{n,t} + \hat{I}_{n,t} + 1}, \tag{2}$$

where $\hat{I}_{n,t} = \sum_{j=n+1}^{\hat{N}_t} \alpha_j \gamma_0 |h_{j,t}|^2$ is the interference of device n; $I_{n,t} = \sum_{i=1}^{n-1} \eta \alpha_i \gamma_0 |h_{i,t}|^2$ and $0 \leq \eta \leq 1$ are the residual interference component and the level of residual interference at device n caused by imperfect SIC (ISIC), respectively.

III. Proposed Efficient Time-Slot Allocation and Performance Analysis with SPC

With random access (RA) nature, the GF devices can select the forbidden time-slots which are granted to the GB devices, leading to a collision. Furthermore, too many GF devices can select the same time-slot, resulting in severe interference and significant performance degradation. To overcome these drawbacks, we investigate the application of QL to enable the GF devices to select the best time-slots for their transmissions with the purpose of ensuring the following two requirements. Firstly, the GF devices do not cause any collisions to the GB devices when they are operating in the same network. Secondly, the number of GF devices choosing the same time-slot is limited to avoid high interference. In this paper, we consider a scenario where at most two GF devices can select

the same time-slot for a two-user NOMA implementation, which is widely used in the NOMA literature [8, 14, 15].

A. QL-based Efficient RA for GB-GF HetNet

QL is one of the most popular RL algorithms, where an agent interacts with the surrounding environment to perform a task in a sequence of time-steps $\{1, \ldots, u, \ldots, U\}$ with the best strategies by learning from previous experience [12]. At time-step u, the agent takes an action $a_u \in \mathcal{A}$ to move from the current sate $s_u \in \mathcal{S}$ to the next state s_{u+1}, and receive a respective reward r_{u+1}. To apply the QL algorithm into the considered GB-GF HetNet, we define the action, observation, state, and reward function (RF) in the following way.

The action taken by a GF device n at time-step u is $a_{n,u} \in \mathcal{A} = \{1, 2, \ldots, T\}$, where the device n selects a time-slot for its transmission in time-step u. We define the observation after taking the action $a_{n,u}$ by $o_{n,u} = \{\mathbb{S}, \mathbb{C}, \mathbb{F}\}$, where, $\mathbb{S}$ depicts a successful transmission, i.e., the device n does not cause a collision to a GB device and the number of GF devices selecting the same time-slot does not exceed two; $\mathbb{C}$ represents a collision, where the GF device n selects a time-slot granted to a GB device; and $\mathbb{F}$ indicates a failed transmission, where the GF device n does not cause a collision to a GB device, but it selects an overloaded time-slot which is used by more than two GF devices. Thus, the difference between the observations $\mathbb{C}$ and $\mathbb{F}$ is that $\mathbb{C}$ refers to the collision between a GF device and a GB device, whereas $\mathbb{F}$ represents the severe contention between GF devices. The network state of the GF device n at time-step u is defined as $s_{n,u} = \{a_{n,u}, o_{n,u}\}$. For the RF of device n, we consider two different definitions, namely binary RF (BRF), which is used in the existing literature [6, 7], and observation-based RF (ORF), i.e.,

$$r_{n,u+1} = \begin{cases} 1, & \text{if } o_{n,u} = \mathbb{S} \\ p_v, & \text{if } o_{n,u} = \mathbb{C} \\ -1, & \text{if } o_{n,u} = \mathbb{F} \end{cases}, \tag{3}$$

where p_v is the penalty, $p_v = -1$ for BRF, and $p_v < -1$ for ORF. In the BRF, the RF receives two different values, where $r_{n,u+1} = 1$ if $o_{n,u} = \mathbb{S}$ and $r_{n,u+1} = -1$ otherwise. In contrast, the proposed ORF uses a three-value RF according to three different observations as indicated in (3). Unlike the BRF, the ORF considers that the penalty for the case $o_{n,u} = \mathbb{C}$ is higher than that for the case $o_{n,u} = \mathbb{F}$ to ensure that the GF devices are not allowed to cause any collisions to the GB devices, but they can interfere with each other.

To depict the relationship between the agents (GF devices) and the environment, the agents build an action-value function, namely Q-function. Let $Q_u(n, a_{n,u})$ be the Q-value of the GF device n at time-step u with the action $a_{n,u}$. After taking the action $a_{n,u}$, the new Q-value $Q_{u+1}(n, a_{n,u})$ is updated based on an iterative procedure as follows [12]:

$$\begin{aligned} Q_{u+1}(n, a_{n,u}) = (1-\beta)\, Q_u(n, a_{n,u}) \\ + \beta \left[r_{n,u+1} + \gamma \max_{a \in \mathcal{A}} Q_u(n, a) \right], \end{aligned} \tag{4}$$

where $0 \leq \beta \leq 1$ is the learning rate and $0 \leq \gamma \leq 1$ is the discount factor. The action $a_{n,u}$ can be determined by the ε-greedy policy as [12]

$$a_{n,u} = \begin{cases} \text{random action}, & \text{probability } \varepsilon \\ \arg\max_{a \in A} \{Q_u(n, a)\}, & \text{probability } 1-\varepsilon \end{cases}, \tag{5}$$

where the GF device n can select an action randomly with probability ε to fully explore the action space. This probability decreases with the increase in the learning time, i.e., $\varepsilon_{u+1} = \delta \varepsilon_u$, where $0 \leq \delta \leq 1$ is the exploration decay coefficient.

The proposed time-slot allocation algorithm for GB-GF HetNet is shown in Algorithm 1. Specifically, the Q-table of the GF device n is first initialized as a $1 \times T$ array of zeros. This device selects a time-slot for its transmission based on (5). It then updates the respective Q-value by using (4) with the BRF and the ORF defined in (3). After updating the Q-value, it performs another time-slot selection for its next transmission. This learning process continues until the convergence is observed, where each GF device finds the best time-slot for its transmission.

B. Performance Analysis of GB-GF HetNet with SPC

Based on the outcomes of the QL-based time-slot allocation algorithm presented in Section III-A, we analyze the performance of the investigated SPC-based GB-GF HetNet in terms of the ADEP. For quasi-static fading channels, the ADEP at the BS to decode the message of a device is approximated as $\phi \approx \int_0^\infty Q\left(\frac{\log_2(1+\gamma) - b/B}{\sqrt{V(x)/B}}\right) f_\gamma(x)\, dx$ [13], where $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ is the Gaussian Q-function, γ is the SNR/SINR, $f_\gamma(x)$ is the probability density function (PDF) of γ, b is the number of information bits, B is the blocklength, and $V(\gamma) = (\log_2 e)^2 \left[1 - \frac{1}{(1+\gamma)^2}\right]$ is the channel dispersion. Note that the approximation of ϕ is achieved when B is sufficiently large, i.e., $B \geq 100$. It is difficult to directly derive ϕ, hence, similar to [14, 15], we approximate ϕ as follows:

$$\phi \approx \chi \sqrt{B} \int_v^\mu F_\gamma(x)\, dx, \tag{6}$$

where $\chi = \sqrt{\frac{1}{2\pi\left(2^{\frac{2b}{B}} - 1\right)}}$, $v = \kappa - \frac{1}{2\chi\sqrt{B}}$, $\mu = \kappa + \frac{1}{2\chi\sqrt{B}}$, $\kappa = 2^{\frac{b}{B}} - 1$, and $F_\gamma(x)$ is the cumulative distribution function (CDF) of γ. In this paper, we consider a scenario where the channels between the devices and the BS follow a Rayleigh distribution. The ADEPs of devices when $\hat{N}_t = 1$ and $\hat{N}_t = 2$ are respectively derived in Propositions 1 and 2 as follows:

Proposition 1. *Under Rayleigh quasi-static fading channels, the ADEPs of GB device m using time-slot t_1 and GF device n using time-slot t_2 in case $\hat{N}_{t_2} = 1$ are given by*

$$\phi_{l,\hat{t}} \approx 1 - \chi\sqrt{B}\lambda_{l,\hat{t}}\gamma_0 \left(e^{-\frac{v}{\lambda_{l,\hat{t}}\gamma_0}} - e^{-\frac{\mu}{\lambda_{l,\hat{t}}\gamma_0}} \right), \tag{7}$$

Algorithm 1: QL-based Time-Slot Allocation and Performance Analysis of the SPC-based GB-GF HetNet.

Data : M, N, T, γ_0, b, B, α_n, β, γ, number of iterations for learning process U.

Result: Q-table for N devices $Q(n)$ and OADEP $\hat{\phi}$.

(1) QL-based Time-Slot Allocation for GB-GF HetNet

Initialize $Q(n,a)$ $(1 \le n \le N, a \in \mathcal{A})$, $\varepsilon \leftarrow 1$, $\delta = 0.95$, $u \leftarrow 1$;
while $u \le U$ **do**
  Device n selects an action $a_{n,u}$ using (5);
  Take action $a_{n,u}$, observe $o_{n,u}$, and get the reward according to (3);
  Update Q-value according to (4);
  Update ε: $\varepsilon_{u+1} \leftarrow \delta\varepsilon_u$;
  $u \leftarrow u+1$;
end
Return $Q(n)$;

(2) Performance Analysis of GB-GF HetNet

Get time-slots granted to the GB devices $\mathbb{A}_{GB}$ and the best action for each GF device:
$a_n \leftarrow \arg\max_{a\in\mathcal{A}} Q(n,a)$;
if *No collision* $\left(a_n \notin \mathbb{A}_{GB} \text{ and } \hat{N}_{a_n} \le 2\right)$ **then**
  Calculate ADEPs of GB and GF devices in case $\hat{N}_{a_n} = 1$ based on (7);
  Calculate ADEPs of GF devices in case $\hat{N}_{a_n} = 2$ based on (8) and (9) ;
else
  ADEPs of collided GB and GF devices are equal to one;
end
Calculate OADEP:
$\hat{\phi} \leftarrow \left(\sum_{m=1}^{M}\phi_m + \sum_{n=1}^{N}\phi_n\right)/(M+N)$;
Return $\hat{\phi}$;

where $l \in \{m,n\}$, $\hat{t} \in \{t_1,t_2\}$, $\lambda_{l,\hat{t}} = \mathbb{E}\{|h_{l,\hat{t}}|^2\}$, *and* $\mathbb{E}\{\cdot\}$ *is the expectation operation.*

Proof: See Appendix A. ∎

Proposition 2. *Under Rayleigh quasi-static fading channels, the ADEPs of GF devices n_1 and n_2 using time-slot t_3 in case $\hat{N}_{t_3} = 2$ are, respectively, given by*

$$\phi_{n_1,t_3} \approx 1 - \chi\sqrt{B}a_{1,t_3}e^{b_{1,t_3}a_{1,t_3}}\Xi_{1,t_3}, \tag{8}$$

and

$$\phi_{n_2,t_3} = \phi_{n_1,t_3} + (1-\phi_{n_1,t_3})\,\hat{\phi}_{n_2,t_3}, \tag{9}$$

where

$$\hat{\phi}_{n_2,t_3} \approx \begin{cases} 1 - \frac{\chi\sqrt{B}}{b_{2,t_3}}\left(e^{-b_{2,t_3}v} - e^{-b_{2,t_3}\mu}\right), & \eta = 0 \\ 1 - \chi\sqrt{B}a_{2,t_3}e^{b_{2,t_3}a_{2,t_3}}\Xi_{2,t_3}, & 0 < \eta \le 1 \end{cases},$$

$a_{1,t_3} = \frac{\alpha_{n_1}\lambda_{n_1,t_3}}{\alpha_{n_2}\lambda_{n_2,t_3}}$, $b_{1,t_3} = \frac{1}{\alpha_{n_1}\gamma_0\lambda_{n_1,t_3}}$, $a_{2,t_3} = \frac{\alpha_{n_2}\lambda_{n_2,t_3}}{\eta\alpha_{n_1}\lambda_{n_1,t_3}}$, $b_{2,t_3} = \frac{1}{\alpha_{n_2}\gamma_0\lambda_{n_2,t_3}}$, $\Xi_{k,t_3} = \mathrm{Ei}\left(\psi^{\mu}_{\mathrm{k,t_3}}\right) - \mathrm{Ei}\left(\psi^{\mathrm{v}}_{\mathrm{k,t_3}}\right)$, $\psi^{p}_{k,t_3} = -b_{k,t_3}p - b_{k,t_3}a_{k,t_3}$, $k \in \{1;2\}$, $p \in \{v,\mu\}$, $\lambda_{n_k,t_3} = \mathbb{E}\{|h_{n_k,t_3}|^2\}$, *and* $\mathrm{Ei}(\mathrm{x}) = -\int_{-\mathrm{x}}^{\infty}\frac{\mathrm{e}^{-\mathrm{t}}}{\mathrm{t}}\mathrm{dt}$ *is the exponential integral function.*

Proof: See Appendix B. ∎

Note that for the cases $o_{n,u} = \mathbb{C}$ and $o_{n,u} = \mathbb{F}$, the transmissions are unsuccessful. In this regard, we assume that the ADEPs of the collided devices are equal to 1. The details of the performance analysis of the SPC-based GB-GF HetNet are given in Algorithm 1.

To evaluate the proposed method for effective time-slot allocation and system performance analysis, we use three performance metrics, namely access efficiency (AE), collision probability (CP), and OADEP. Specifically, AE is defined as the ratio of the number of successful transmissions (NoS) to the number of available time-slots (T), i.e., $\mathrm{AE_{GB}} = \mathrm{NoS_{GB}/T}$ for GB devices and $\mathrm{AE_{GF}} = \mathrm{NoS_{GF}/T}$ for GF devices. CP is defined as the ratio of the number of collided devices (NoC) to the number of devices, i.e., $\mathrm{CP_{GB}} = \mathrm{NoC_{GB}/(M+N)}$ for GB devices and $\mathrm{CP_{GF}} = \mathrm{NoC_{GF}/(M+N)}$ for GF devices. OADEP is calculated as $\hat{\phi} = \left(\sum_{m=1}^{M}\phi_m + \sum_{n=1}^{N}\phi_n\right)/(M+N)$, where ϕ_m and ϕ_n are the ADEP of GB device m and GF device n, respectively.

IV. Numerical Results

This section provides the numerical results to analyze the performance of the considered SPC-based GB-GF HetNet in terms of AE, CP, and OADEP. For the proposed QL algorithm, we set the parameters as in [6, 7]: the learning rate $\beta = 0.1$, the discount factor $\gamma = 0.5$, and the penalty value $p_v = -10$ for the ORF. In addition, the predetermined simulation parameters are set as follows unless otherwise stated: $U = 2000$, $T = 50$, $M = 20$, $T_{gf} = T - M = 30$, $\gamma_0 = 20$ (dB), $\eta = 0.1$, $\alpha_{n_1} = 0.7$, $\alpha_{n_2} = 0.3$. All devices are assumed to have the same number of information bits $b = 80$ and blocklength $B = 100$. Furthermore, each GB device is granted a specific time-slot within each frame of T time-slots for its transmission in a repetitive manner from frame to frame.

Fig. 2 plots the AE of GB and GF devices versus the number of GF devices (N) with the fixed number of GB devices (M). We investigate three access methods, the proposed QL approach with the ORF in (3) (ORF-QL), the proposed QL method with the BRF in (3) (BRF-QL), and the random selection (RS) scheme. In the RS scheme, each GF device randomly selects a time-slot without the use of QL.

Fig. 2 shows that the AE of the GB and GF devices can be significantly improved by using QL. Furthermore, the proposed ORF-QL and BRF-QL methods can guarantee the best AE for GB and GF devices when $N \le 2T_{gf}$. Herein, with fixed value of M, each GB device is always granted to a separate conflict-free time-slot, resulting in the unchanged AE of GB devices. Meanwhile, the GF devices find the best time-slots for their transmissions, hence, the AE of these devices increases with the increase in N. However, when $N > 2T_{gf}$, the BRF-QL method cannot ensure a conflict-free environment

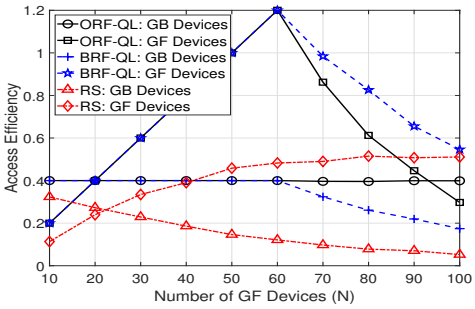

Fig. 2. AE versus number of GF devices (N) with different access methods.

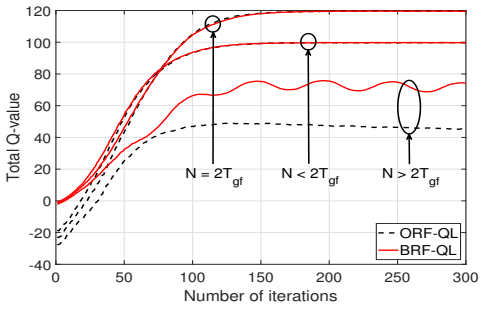

Fig. 4. Total Q-value of the considered QL-based access methods.

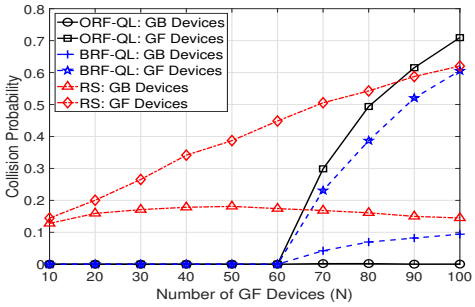

Fig. 3. CP versus N with different access methods.

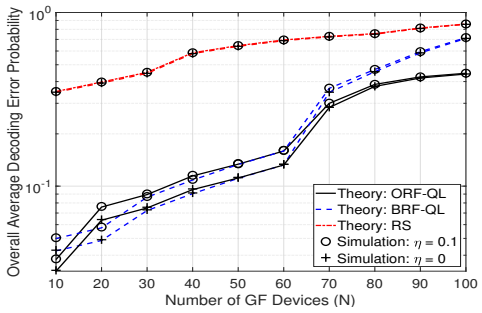

Fig. 5. OADEP versus N with different access methods.

to the GB devices, which is an important requirement of the GB-GF HetNet. This leads to the AE reduction of GB devices. In contrast, the ORF-QL method still brings the best AE for these devices when $N > 2T_{gf}$. Another observation from this figure is that when $N > 2T_{gf}$, the AE of the GF devices decreases for both ORF-QL and BRF-QL methods due to a rapidly increase in the collision rate. Nevertheless, the AE for the GF devices achieved by the ORF-QL method is lower than that achieved by the BRF-QL method in this region. This is because the ORF-QL method can provide the GB devices with conflict-free links when $N > 2T_{gf}$, hence, the number of available RA slots for the GF devices decreases, resulting in the lower AE as compared to BRF-QL.

Fig. 3 depicts the CP of the GB and GF devices versus N for different access methods. This figure shows that by using the ORF-QL method, the CP for GB devices becomes significantly low and almost unchanged with the increase in N, whereas, its value for the GF devices is considerably small when $N \leq 2T_{gf}$ and increases when $N > 2T_{gf}$ due to the higher collision rate. In contrast, the BRF-QL method can only ensure a low CP similar to the ORF-QL method when $N \leq 2T_{gf}$, but it increases the CP for both the GB and GF devices when $N > 2T_{gf}$. Therefore, the BRF-QL method cannot ensure the independent operation of the GB and GF devices when N becomes higher. As the worst case, using the RS method results in the higher CP for both the GB and GF devices as compared to the ORF-QL and BRF-QL approaches due to its higher collision rate. Thus, the proposed ORF-QL method outperforms the BRF-QL and RS methods in guaranteeing the coexistence of GB and GF devices in a GB-GF HetNet.

Fig. 4 provides the convergence analysis of the proposed QL methods (i.e., ORF-QL and BRF-QL) by considering the variation of the total Q-value versus the number of iterations. We investigate the following three cases: $N < 2T_{gf}$, $N = 2T_{gf}$, and $N > 2T_{gf}$, where $N \in \{50; 60; 70\}$ and $T_{gf} = T - M = 30$. This figure indicates that the ORF-QL and BRF-QL methods have the similar convergence when $N \leq 2T_{gf}$. However, BRF-QL obtains the higher total Q-value leading to the better performance for GF devices when $N > 2T_{gf}$ as compared to ORF-QL. This confirms the results achieved in Figs. 2 and 3, where BRF-QL outperforms ORF-QL in terms of the AE and CP for the GF devices when $N > 2T_{gf}$, however, it cannot guarantee a good coexistence of the GB and GF devices like in the ORF-QL method.

After training the GF devices to select the best time-slots for their transmissions by using the proposed QL method, we analyze the system performance in terms of OADEP as shown in Fig. 5. Herein, we evaluate the effect of the residual interference level caused by ISIC (η) on the system performance. Specifically, the increase in the value of η leads to the growth of OADEP due to the higher interference between the GF devices. In addition, the higher value of the OADEP can be observed when N increases for the considered access methods. This is because the increase in N results in the higher collision rate, leading to the larger error probability. Furthermore, this figure indicates that using QL (i.e., ORF-QL and BRF-QL methods) achieves the better performance (i.e., lower OADEP) than the case without using QL (i.e., RS scheme). Moreover, the OADEPs achieved by using the proposed ORF-QL and BRF-QL methods are similar in the region of $N \leq 2T_{gf}$, but ORF-QL outperforms BRF-QL when $N > 2T_{gf}$.

V. Conclusion

In this paper, we have proposed a QL-based MA method for the coexistence of GB and GF devices in an SPC-based GB-GF uplink HetNet. The proposed method enables the GF devices to select the best RA slots such that these devices do not collide with the GB devices and the number of GF devices choosing the same time-slot using the GF NOMA does not exceed a predetermined threshold to reduce the interference. Based on the result of the RA slot allocation using the proposed learning method, we have characterized the system performance by deriving the approximate closed-form expressions of the ADEP for all the devices. The achieved results have shown that among the proposed QL methods, ORF-QL outperforms BRF-QL in improving the system performance while also guaranteeing the harmonious coexistence between the GB and GF devices in a GB-GF HetNet.

Acknowledgment

This work was supported in part by ERC-funded project Agnostic under Grant 742648.

Appendix A
Proof of Proposition 1

Based on (1), we have the SNR of device l in time-slot $\hat{t}$ $\left(l \in \{m, n\}\right.$ and $\left.\hat{t} \in \{t_1, t_2\}\right)$ is $\gamma_{l,t} = \gamma_0 \left|h_{l,\hat{t}}\right|^2$. To derive $\phi_{l,\hat{t}}$ based on (6), we first need to calculate the CDF of $\gamma_{l,t}$ which is expressed as $F_{\gamma_{l,\hat{t}}}(x) = 1 - e^{-\frac{x}{\lambda_{l,\hat{t}}\gamma_0}}$. From this formula and (6), $\phi_{l,\hat{t}}$ has the following form

$$\phi_{l,\hat{t}} \approx \chi\sqrt{B}\int_{v}^{\mu}\left(1 - e^{-\frac{x}{\lambda_{l,\hat{t}}\gamma_0}}\right)dx. \tag{10}$$

And the final expression of $\phi_{l,\hat{t}}$ is achieved as in (7).

Appendix B
Proof of Proposition 2

With two GF devices using the same time-slot t_3, i.e., $N_{\hat{t}_3} = 2$, based on (2), we have the SINRs of devices n_1 and n_2 are as follows: $\gamma_{n_1,t_3} = \frac{\alpha_{n_1}\gamma_0\left|h_{n_1,t_3}\right|^2}{\alpha_{n_2}\gamma_0\left|h_{n_2,t_3}\right|^2+1}$ and $\gamma_{n_2,t_3} = \frac{\alpha_{n_2}\gamma_0\left|h_{n_2,t_3}\right|^2}{\eta\alpha_{n_1}\gamma_0\left|h_{n_1,t_3}\right|^2+1}$. The CDF of γ_{n_1,t_3} and γ_{n_2,t_3} are, respectively, calculated as

$$\begin{aligned} &F_{\gamma_{n_1,t_3}}(x) \\ &= \int_0^{\infty} F_{\left|h_{n_1,t_3}\right|^2}\left(\frac{\beta_{n_2}xy}{\beta_{n_1}} + \frac{x}{\beta_{n_1}\gamma_0}\right) f_{\left|h_{n_2,t_3}\right|^2}(y)\,dy \\ &= 1 - \frac{a_{1,t_3}e^{-b_{1,t_3}x}}{x + a_{1,t_3}}, \end{aligned} \tag{11}$$

and

$$F_{\gamma_{n_2,t_3}}(x) = \begin{cases} 1 - e^{-b_{2,t_3}x}, & \eta = 0 \\ 1 - \frac{a_{2,t_3}e^{-b_{2,t_3}x}}{x+a_{2,t_3}}, & 0 < \eta \leq 1 \end{cases}, \tag{12}$$

where $f_{|h_n|^2}(x) = \frac{1}{\lambda_n}e^{-\frac{x}{\lambda_n}}$ and $F_{|h_n|^2}(x) = 1 - e^{-\frac{x}{\lambda_n}}$ are the PDF and CDF of channel gain $|h_n|^2$, respectively. From (6) and (11), the ADEP of device n_1 is given by $\phi_{n_1,t_3} \approx \chi\sqrt{B}\int_{v}^{\mu}\left(1 - \frac{a_{1,t_3}e^{-b_{1,t_3}x}}{x+a_{1,t_3}}\right)dx$. With the aid of [16, Eq. (3.352.2)], the final expression of ϕ_{n_1,t_3} is obtained in (8).

For device n_2, the BS first needs to decode the message of device n_1 and remove this component from its observation by using ISIC before detecting the message of device n_2. Given this context, the ADEP of device n_2 is expressed as in (9), where $\hat{\phi}_{n_2,t_3}$ is derived by using (6) and (12) as $\hat{\phi}_{n_2,t_3} \approx \chi\sqrt{B}\int_{v}^{\mu}\left(1 - e^{-b_{2,t_3}x}\right)dx$ when $\eta = 0$ and $\hat{\phi}_{n_2,t_3} \approx \chi\sqrt{B}\int_{v}^{\mu}\left(1 - \frac{a_{2,t_3}e^{-b_{2,t_3}x}}{x+a_{2,t_3}}\right)dx$ when $0 < \eta \leq 1$. By utilizing [16, Eq. (3.352.2)], $\hat{\phi}_{n_2,t_3}$ can be achieved as in (9).

References

[1] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, Firstquarter 2020.

[2] S.-Y. Lien *et al.*, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.

[3] R. Abreu *et al.*, "On the multiplexing of broadband traffic and grant-free ultra-reliable communication in uplink," in *IEEE Veh. Technol. Conf. (VTC-Spring)*, Kuala Lumpur, Malaysia, Apr. 2019, pp. 1–6.

[4] I. Gerasin, A. Krasilov, and E. Khorov, "Flexible multiplexing of grant-free URLLC and eMBB in uplink," in *IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, London, UK, Aug. 2020, pp. 1–6.

[5] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novela nalytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Nov. 2018.

[6] S. K. Sharma and X. Wang, "Collaborative distributed Q-learning for RACH congestion minimization in cellular IoT networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 600–603, Apr. 2019.

[7] M. V. da Silva, R. D. Souza, H. Alves, and T. Abrão, "A NOMA-based Q-learning random access method for machine type communications," *IEEE Wireless Commun. Lett.*, Jun. 2020, Early Access.

[8] S. Han *et al.*, "Energy-efficient short packet communications for uplink NOMA-based massive MTC networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12 066–12 078, Dec. 2019.

[9] D.-D. Tran, S. K. Sharma, and S. Chatzinotas, "BLER-based adaptive Q-learning for efficient random access in NOMA-based mMTC networks," in *IEEE Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, to appear.

[10] W. Li *et al.*, "SmartCC: A reinforcement learning approach for multipath TCP congestion control in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2621–2633, Nov. 2019.

[11] F. Jameel *et al.*, "Reinforcement learning for scalable and reliable power allocation in SDN-based backscatter heterogeneous network," in *IEEE INFOCOM*, Toronto, ON, Canada, Jul. 2020, pp. 1069–1074.

[12] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[13] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[14] D.-D. Tran, S. K. Sharma, S. Chatzainotas, I. Woungang, and B. Ottersten, "Short-packet communications for MIMO NOMA systems over Nakagami-m fading: BLER and minimum blocklength analysis," *IEEE Trans. Veh. Technol.*, pp. 1–16, Mar. 2021, Early Access.

[15] H. Liu, N. I. Miridakis, T. A. Tsiftsis, K. J. Kim, and K. S. Kwak, "Coordinated uplink transmission for cooperative NOMA systems," in *IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[16] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, Mar. 2007.