IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 31, 2020, accepted April 11, 2020, date of publication April 20, 2020, date of current version May 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2988710

Latency-Aware Dynamic Resource Allocation Scheme for Multi-Tier 5G Network: A Network Slicing-Multitenancy Scenario

SUNDAY OLADAYO OLADEJO, (Graduate Student Member, IEEE), AND OLABISI EMMANUEL FALOWO, (Senior Member, IEEE)
Department of Electrical Engineering, University of Cape Town, Rondebosch 7700, South Africa

Corresponding author: Sunday Oladayo Oladejo (oldsun002@myuct.ac.za)

This work was supported in part by the National Research Foundation, South Africa and Telkom, South Africa.

ABSTRACT In 5G slice networks, the multi-tenant, multi-tier heterogeneous network will be critical in meeting the quality of service (QoS) requirement of the different slice use cases and in reduction of the capital expenditure (CAPEX) and operational expenditure (OPEX) of mobile network operators. Hence, 5G slice networks should be as flexible as possible to accommodate different network dynamics such as user location and distribution, different slice use case QoS requirements, cell load, intra-cluster interference, delay bound, packet loss probability, and service level agreement (SLA) of mobile virtual network operators (MVNO). Motivated by this condition, this paper addresses a latency-aware dynamic resource allocation problem for 5G slice networks in a multi-tenant, multi-tier heterogeneous environment, for efficient radio resource management. The latency-aware dynamic resource allocation problem is formulated as a maximum utility optimisation problem. The optimisation problem is transformed and the hierarchical decomposition technique is adopted to reduce the complexities in solving the optimisation problem. Furthermore, we propose a genetic algorithm (GA) intelligent latency-aware resource allocation scheme (GI-LARE). We compare GI-LARE with the static slicing (SS) resource allocation; the spatial branch and bound-based scheme; and, an optimal resource allocation algorithm (ORA) via Monte Carlo simulation. Our findings reveal that GI-LARE outperformed these other schemes.

INDEX TERMS Network slicing, multi-tier, multi-tenancy, resource allocation.

I. INTRODUCTION

A. BACKGROUND

Network slicing (NS) refers to the abstraction of the physical infrastructure and resources of a mobile network into logical networks, which operate as autonomous entities or networks. NS is envisioned to play a critical role in the full implementation of IMT-2020 networks widely regarded as the fifth generation (5G) mobile networks. 5G networks will be pivotal in the Industry 4.0 revolution; hence, 5G networks will support diverse verticals and services to reshape the way we live, transact businesses, and conduct human-machine relationship [1].

Despite the positive economic impact of 5G NS, realising effective NS schemes requires financial commitment [2], [3] by key industry players such as the infrastructure providers (InPs), mobile virtual network operators (MVNOs), backhaul operators (BO), service providers (SP), and over-the-top players (OTP). To make 5G networks profitable (i.e. by the reduction of the capital expenditure (CAPEX) and operating expenditure (OPEX) many business models [4], [5] have been proposed, which revolve around multi-tenancy. In this work, we adapt the models in [4], [5] to address multi-tenancy in 5G NS as depicted in Fig. 1. Here, Fig. 1 depicts a two-stage hierarchical business model for NS in a multi-tenancy scenario where the InP leases out virtual network resources to different MVNOs.

In this paper, we address the diverse service requirement via three main slice use cases [6], [7]: (1) the enhanced mobile broadband (eMBB); (2) the massive machine-type communications (mMTC); and, (3) the ultra-reliable low-latency communications (URLLC) slice use cases. The eMBB use case is bandwidth-crunching and supports applications such as high definition (HD) video streaming and virtual reality (VR).

The associate editor coordinating the review of this manuscript and approving it for publication was Kwok L. Chung.

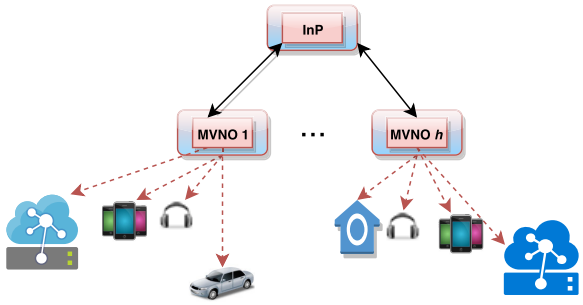


FIGURE 1. InP-MVNO model.

The mMTC use case is latency-dependent with intermittent small-size data payloads. It supports applications such as e-health and internet of things (IoT) devices. The URLLC use case supports applications and services that require very low-latency small-size payload transmissions with extremely high reliability such as autonomous driving and vehicle-to-everything (V2X).

In a multi-tenant multi-tier heterogeneous 5G slice network, addressing the different quality of service (QoS) requirements of different verticals and several services could be an uphill task. Moreover, owing to the stochastic characteristics of the mobile network environment and the dynamic allocation of network resources (such as bandwidth and power) between the InP and the MVNOs; the respective MVNOs and the numerous slice users could be very challenging. Besides, unlike the widely investigated 2-tier heterogeneous network environment in the study of 5G NS, we examine the concept of NS deployed or implemented in a hierarchical multi-tier clustered heterogeneous multi-tenant network. Furthermore, unlike most works in the literature, we focus on both the latency and received data rate QoS requirement of three slice use cases.

B. RELATED WORK

There are a number of architectures and solutions that have been proposed for NS. The authors in [8], [9] and [10] presented maximum capacity, profit-aware, and energy-efficient resource allocation schemes for NS in a multi-tenancy scenario. Slice priorities and bandwidth-power cost were considered; however, a static resource scheme between the InP and MVNOs was adapted. Besides, the latency constraints requirement of the respective slices was not considered. Also, a single-tier homogeneous network was considered, which does not entirely depict a 5G network and its complexities. The authors in [11] proposed an incentive scheme for slice cooperation based on the D2D communication in a multi-tenant 5G network for achieving maximum system utility. The authors did not address the multi-tier and multi-slice peculiarities of 5G networks. Moreover, the latency-aware requirements of the slice use case, such as the URLLC, were not considered.

To meet the latency requirements of the cloud radio access network (C-RAN), the authors in [12] proposed a queuing delay model for front haul network dimensioning in 5G networks. Kingman's exponential law of congestion was adopted by the authors to estimate the delay on the front-haul.

In [13], a maximum-revenue resource allocation optimisation problem was formulated for a virtual network in a 2-tier heterogeneous network. In solving this problem, the authors pre-allocated radio resources to the respective base stations or access point. In [14], a dynamic resource sharing scheme for a single-tier homogeneous C-RANs multi-tenancy was proposed. A network utility maximisation problem was formulated while considering the tenants' priorities. Although the proposed two-step sub-optimal approach improved the network utility, users were not categorised based on their slice requirements. The authors in [15] presented a dynamic radio resource slicing scheme for a 2-tier heterogeneous wireless network. An alternating concave search algorithm was designed to solve the maximum network utility optimisation problem. The 2-tier heterogeneous network, due to its simplistic model, may not fully represent a 5G network environment with its many tiers of access networks in order to meet the ever-rising user demands. Besides, the authors did not address the concept of multi-tenancy, which is a critical requirement for CAPEX and OPEX reduction in 5G networks.

In [16], [17], [18], and [19], the authors considered a dynamic allocation of radio resources in a network slicing scenario. An auction game-based algorithm was proposed for efficient resource allocation between the InP and MVNOs. Additionally, the authors did not consider the challenge of latency constraints in the resource allocation scheme. Although the authors considered multi-tenancy, the multi-tier and multi-slice features of the 5G network were not taken into consideration in their studies.

In [20], an efficient RAN slicing strategy for a heterogeneous network with eMBB and V2X services was investigated. The authors proposed an off-line reinforcement learning scheme which allocates radio resources to the eMBB and V2X slice user case with the sole aim of maximising network resource utilisation. However, the latency requirements of the V2X use case were not considered. Besides, the small scale fading factors that significantly affect fast-moving devices and vehicles were not included in their model. In addition, a single-tier network was considered, which does not entirely depict a 5G network which is envisioned in [21] to be a multi-tenant multi-tier network.

In [22], the authors discussed the different approaches to realise URLLC use cases for V2X communications. The authors adopted the large deviation theoretical (effective bandwidth or capacity) framework of the MAC layer approach.

In [23], the authors proposed a cooperative communications scheme based on the average bit error probability (ABEP) to enhance the performance of IoT communication systems. The scheme relies on the capabilities of the

back-propagation neural network to predict the ABEP performance of the investigated system.

In [24], the authors investigated a slice-aware admission scheme for multi-tenant radio access networks, which supports guaranteed eMBB and mission-critical services. A Markovian model was proposed to characterise resource sharing in a multi-tenant network slicing environment. However, in addition to considering only the single-cell scenario, authors did not address the latency requirement of the mission-critical use case.

Furthermore, in [25], the authors examined dynamic resource allocation in a virtualised network slicing environment. A dynamic resource allocation scheme based on deep reinforcement learning was proposed to address the challenge, as mentioned earlier. Nevertheless, they did not address the multi-tenancy scenario and its challenges. Moreover, it was not shown how the average delay utility of the delay constrained slice was guaranteed or ensured.

In [26], the authors proposed a dynamic network slicing and resource allocation scheme for video streaming and IoT applications, which is based on the Lyapunov Optimisation in a single cell scenario. However, the URLLC use case, which is highly latency-dependent and requires extreme-reliability, was not considered. In addition, multi-tenancy, which is a critical feature in NS, was not considered.

The authors in [27] showed that low error rates and low latencies are attainable and practicable over an air interface. Moreover, the authors emphasised the importance of channel error rates and short transmission intervals in achieving low latency.

The authors in [28] considered the challenge of latency in the allocation of resources to users in a multi-access edge computing network. A virtual network function placement assignment algorithm based on the polynomial-time combinatorial algorithm was proposed to guarantee user satisfaction. However, a multi-slice multi-tier 5G network, in which slice users require different latency thresholds, was not considered. Besides this point, the authors also did not consider the peculiarities of multi-tenancy in their problem formulation.

In [29], the authors studied network slicing resource allocation challenges in vehicular networks. The eMBB and URLLC slice-use cases were considered in the proposed scheme, which is based on the effective capacity theory. However, the vehicular network was not studied in the context of a multi-tier multi-tenant network. The vehicular network cannot exist in isolation [30], [31] because of its interaction with other slice users in other tiers such as macro, pico and femto tiers. In addition, dynamic resource allocation was not considered.

In [32], the authors proposed a dynamic resource allocation scheme for eMBB and URLLC slice-use cases. The proposed scheme is based on optimal power control for latency-aware resource allocation. The dynamic allocation of the bandwidth, which is a scare resource, was not addressed. The authors did not consider the peculiarities of multi-tenant multi-tier in their problem formulation.

In [33], the authors addressed the challenge of slice users' quality of experience and resource allocation in a vehicular network. The authors partitioned vehicles into multiple logical networks based on a network slicing clustering algorithm. A multi-tier network which reflects one of the 5G features was not considered. Moreover, static partitioning of radio resources was adopted rather than dynamic resource allocation of resources which can easily adapt to traffic variation.

In satisfying the diverse demand requirements of the respective slice use cases, the authors in [34] proposed an on-demand cooperation scheme among multi-tenants in a network slicing scenario. The proposed framework was centred on complex network theory to obtain the topology related information of networks for efficient resource management. However, the latency requirement of the slice use cases was not addressed.

Different from the above-mentioned works, in the present paper we investigate the latency-aware dynamic resource allocation problem in a multi-tier clustered heterogeneous network for multi-tenancy network slicing.

C. CONTRIBUTIONS

The main contributions of the present paper are summarised as follows:

1) We consider radio resource allocation concerning the three broad slice use cases, namely the eMBB, mMTC, and URLLC respectively, in a multi-tier multi-tenant 5G slice network. A latency-aware dynamic resource allocation scheme is developed as an optimisation framework to maximise the total utility of the network. This framework efficiently allocates radio resources to the different slice use cases by considering the data rate and latency requirements of the respective slice use cases. In meeting the slice user QoS requirements, the network bandwidth is sliced by taking into consideration the users' location and distribution, slice use case QoS requirements, cell load, tier load, intra-cluster interference, delay bound, packet loss probability, and the service level agreement of the respective MVNOs (i.e. tenants).
2) As stated, the latency-aware dynamic resource allocation problem is formulated as a maximum utility optimisation problem. In solving the maximum utility optimisation problem, we transform and decompose the main problem via hierarchical decomposition [35] to reduce the complexity of the main problem. Consequently, we optimally associate slice users with the different tiers in the clustered multi-tier multi-tenant network. We exploit the matching game theory to optimally associate slice users to the respective access points. The selection, crossover, mutation and elitism processes of the Genetic Algorithm are adapted to solve the transformed maximum utility problem.

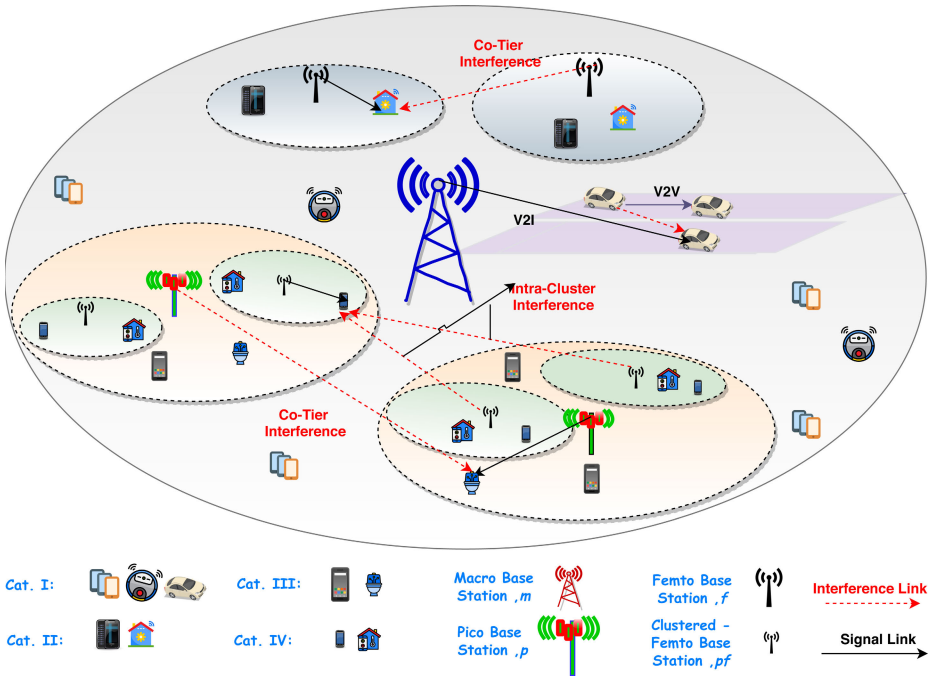


FIGURE 2. System model.

3) Through extensive Monte-Carlo simulations, we demonstrate the performance of the proposed latency-aware dynamic resource allocation framework in a clustered multi-tier multi-tenant network. We also compare the proposed GI-LARE with three other schemes, namely: a static slicing scheme (SS) [36], a spatial branch and branch scheme (sBB) [37], and an optimal resource allocation algorithm (ORA) [38].

D. ORGANISATION

We organised the remainder of this paper as follows. In Section II, we give a detailed explanation of the system model. In Section III, the latency-aware and dynamic resource model is discussed. The latency-aware dynamic radio resource allocation problem is formulated in Section IV. In Section V, the proposed solutions are discussed in detail. To this end, we discussed the computational complexities of the proposed algorithms in Section VI. Simulation results are shown and discussed in Section VII. Finally, we draw the conclusion of this paper in Section VIII. For convenience, the notations used in this paper are summarised in Table 1.

II. SYSTEM MODEL

In this section, a multi-tier multi-tenant heterogeneous network system model is presented. Table 1 shows the main notations to be used in the following sections.

A. GENERAL MODEL

We describe the system model considered in this paper, as depicted in Fig. 2. The considered scenario assumes a clustered multi-tier heterogeneous network whose physical resources are owned by an InP. The InP provides services to a set of MVNOs $\mathcal{H} = \{h|h \in \mathcal{N}, 1 \leq h \leq |\mathcal{H}|\}$. Each MVNO, $h \in \mathcal{H}$ is uniquely independent of each other; that is, $h \neq h'$, and h has its own set of network slice use cases, $\mathcal{S}_h$, it offers to its slice users. However, $\mathcal{S}_h = \{\mathcal{E} \cup \mathcal{M} \cup \mathcal{R}\}$, in which $\mathcal{E}$ denotes the eMBB slice user case, $\mathcal{M}$ indicates the mMTC slice use case and $\mathcal{R}$ stands for the URLLC. Subsection II-B gives a detailed explanation of the slice use case specifications. The multi-tier network comprises femtocells, picocells, clustered femtocells, a macrocell and a device-to-device (D2D) based V2X communication layers. The set of unclustered femtocells located in the coverage of the macrocell only is numbered as $\mathcal{F} = \{f|f \in \mathcal{N}, 1 \leq f \leq |\mathcal{F}|\}$, while the set of picocells is denoted as $\mathcal{P} = \{p|p \in \mathcal{N}, 1 \leq p \leq |\mathcal{P}|\}$. Owing to the relatively large radius of p compared to f, we consider that there are femtocells in the coverage of a p. These femtocells we call clustered femtocells, such that a set of clustered femtocell in the coverage of picocell is numbered as $\mathcal{PF} = \{pf|pf \in \mathcal{N}, 1 \leq pf \leq |\mathcal{PF}|\}$. The coverage area of a p helps to create a cluster area for a set of clustered femtocells, pf. It is essential to state that the users of an MVNO, h, are categorised according to their requested slice use case $\{\mathcal{E} \cup \mathcal{M} \cup \mathcal{R}\}$ and geographical position in the multi-tier network.

B. SLICE USER CATEGORISATION

Considering the slice use-case requested by QoS requirement and its geographical location of the users, we categorise users into four:

TABLE 1. List of noations.

Symbol	Description
h,$\mathcal{H}$	MVNO, Set of MVNO
$\mathcal{S}_h$	Set of Slices offered by a MVNO
$\mathcal{E}$, $\mathcal{M}$, $\mathcal{R}$	eMBB, mMTC, URLLC Slice
m, f, p, pf	Macro, femto, pico and clustered femtocells
$\mathcal{F}$, $\mathcal{P}$, $\mathcal{PF}$	Set of femtocell, picocell, and clustered femtocell
$\mathcal{E}_{h,m}$, $\mathcal{E}_{h,f}$, $\mathcal{E}_{h,p}$, $\mathcal{E}_{h,pf}$	Set of eMBB users in the macro, femto, pico and clustered femto tiers
$\mathcal{M}_{h,m}$, $\mathcal{M}_{h,f}$, $\mathcal{M}_{h,p}$, $\mathcal{M}_{h,pf}$	Set of mMTC users in the macro, femto, pico and clustered femto tiers
$\mathcal{R}_{h,m}$	Set of URLLC Slices
$d_{i,j,h}$	Distance of user i in MVNO h from access point j
$\rho_{i,j,h}$, $\rho_{r,m,h}$	Path loss of the link between user i and access point
$\gamma_{i,j,h}$	Spectrum efficiency
$\Gamma_{i,j,h}$	Channel Gain
$\psi_{j,h}$, ψ_w	Tx. power of an access point, V2V car in Tx mode
T_w, T_r	Effective antenna height of a V2V car in Tx mode, V2I in Rx mode
X_c	Carrier frequency in GHz
$\mathcal{W}$	Set of paired vehicles in V2V communication
$\alpha_{r,m}$, $\alpha_{w,r}$	Small scale fading components of a $\mathcal{R}_{h,m}$ slice user engage in a V2V and V2V communication.
$\lambda_{i,h}$	Packet arrival rate
D_{max}	Maximum delay bound
μ	Delay-bound violation probability threshold
$\theta_{i,h}$	QoS Exponent
$L_{i,h}$	Packet size
$\mathcal{B}$	Total Bandwidth of the network
$\beta_{t,h}$	Network Slice ratio per tier t for each MVNO h
$\varphi_{i,p,h}$	user slice ratio
P_m, P_e, P_c	Probability of mutation, elitism, crossover
$\boldsymbol{A}$	Population of chromosomes in CGA
y	Number of chromosomes
$\|t\|$	Number of tiers
$\boldsymbol{U}$	Fitness vector of $\boldsymbol{A}$
g	Maximum number of iterations

1) **Cat. I:** $\mathcal{E}_{h,m}$ is the set of eMBB users belonging to MVNO h, attached to the macro-tier. In addition, the set of users requesting mMTC slice belonging to MVNO h and attached to the macro base station, m, in the macro-tier, is denoted by $\mathcal{M}_{h,m}$. Furthermore, the set of vehicles pre-installed with Subscriber Identity Module (SIM) of MVNO h requesting URLLC services is denoted by $\mathcal{R}_{m,h}$.
2) **Cat. II:** The set of MVNO h eMBB users in the coverage of a femtocell, $f \in \mathcal{F}$, is denoted as $\mathcal{E}_{h,f}$ and similarly, $\mathcal{M}_{h,f}$ for the set of mMTC slice users in the coverage of a femtocell $f \in \mathcal{F}$.
3) **Cat. III:** For the set of slice users belonging to MVNO h, in the coverage area of a picocell $p \in \mathcal{P}$, however which do not fall under the coverage of a clustered femtocell, pf, is denoted as $\mathcal{E}_{h,p}$. Likewise, $\mathcal{M}_{h,p}$ denotes the set of mMTC slice users which are under the coverage area of $p \in \mathcal{P}$ and not under the coverage area of a clustered femto cell pf.
4) **Cat. IV:** Similar to the other categories, $\mathcal{E}_{h,pf}$ denotes the set of MVNO h users requesting eMBB slice in the coverage of a clustered femtocell $pf \in \mathcal{PF}$. For MVNO h, users requesting mMTC slice in a clustered femtocell pf, its set is denoted by $\mathcal{M}_{h,pf}$. The total number of cluster $c \in \mathcal{C}$ is the same as $|\mathcal{P}|$.

C. V2X COMMUNICATION MODEL

The set of URLLC users and devices, $\mathcal{R}_{m,h}$, is modelled using the D2D-based V2X communication. In this work, we assume that V2X communication is based on the Cellular-V2X (C-V2X) rather than the Dedicated Short Range Communications (DSRC). Our assumption is due to the growing popularity of C-V2X in the vehicle-communications and manufacturing industry and other reasons in [22], [39], and [40].

We assume the V2X communication-enabled cars are under the coverage of the macro base station alone to minimise the handover signalling. The V2X layer comprises a set of vehicles (i.e. URLLC slice users engaged in Vehicle-to-Network (V2N)) $\mathcal{R} = \{r|r \in \mathcal{N}, 1 \leq r \leq |\mathcal{R}|\}$ connected to the macro base station requesting for the URLLC slice. In addition to the V2X layer, the set of paired vehicles that engage in Vehicle-to-Vehicle (V2V) communications using the PC5 sidelink is numbered as $\mathcal{W} = \{w|w \in \mathcal{N}, 1 \leq w \leq |\mathcal{W}|\}$.

D. CHANNEL MODEL

Specifically, our paper draws on the downlink of multi-tier heterogeneous networks based on the link layer model given in [41], [42] and mobility characteristic of slice users in modelling the channel. We categorise the channel modelling into two; (i) Static and moderately mobile Slice users and (ii) highly mobile slice users. Without loss of generality, we assume mMTC and eMBB slice users are in the first category and the URLLC users in the latter.

1) STATIC SLICE USERS

We consider a slice user i_h with a path loss given as [43]:

$$\rho_{i,j,h} = \begin{cases} 30 + 35\log(d_{i,j,h}), & \forall\, i_h \in \{\mathcal{E}_{m,h}, \mathcal{M}_{m,h}\}, j = m \\ 35 + 35\log(d_{i,j,h}), & \forall\, i_h \in \{\mathcal{E}_{p,h}, \mathcal{M}_{p,h}\}, j = p \\ 40 + 35\log(d_{i,j,h}), & \\ & \forall\, i_h \in \{\mathcal{E}_{f,h}, \mathcal{E}_{pf,h}, \mathcal{M}_{f,h}, \mathcal{M}_{pf,h}\} \end{cases} \tag{1}$$

where $d_{i,j,h}$ denotes the distance of the slice user, i_h, belonging to MVNO, h from an access point, $j \in \{m, f, p, pf\}$. The spectrum efficiency of a user $i_h \in \{\mathcal{E}_{m,h}, \mathcal{M}_{m,h}\}$ is expressed as:

$$\gamma_{i,j,h} = \log_2\left(1 + \frac{\psi_{j,h}\Gamma_{i,j,h}}{\sigma^2}\right), \quad \forall j = m \tag{2}$$

where $\psi_{j,h}$ is the transmit power and $\Gamma_{i,j,h}$ denotes the channel gain associated with a user i_h which belongs to MVNO h and an access point in tier $j \in \{m, f, p, pf\}$. Similarly, for a user $i_h \in \{\mathcal{E}_{f,h}, \mathcal{M}_{f,h}\}$, its spectrum efficiency is given as:

$$\gamma_{i,j,h} = \log_2\left(1 + \frac{\psi_{j,h}\Gamma_{i,j,h}}{\sigma^2 + \sum\limits_{\substack{k\in\{\mathcal{F}\}\\ j\neq k}} \psi_{k,h}\Gamma_{i,k,h}}\right) \tag{3}$$

Likewise, for a user $i_h \in \{\mathcal{E}_{p,h}, \mathcal{M}_{p,h}\}$, its spectral efficiency is given as:

$$\gamma_{i,j,h} = \log_2\left(1 + \frac{\psi_{j,h}\Gamma_{i,j,h}}{\sigma^2 + \sum\limits_{\substack{k\in\{\mathcal{P}\}\\ j\neq k}} \psi_{k,h}\Gamma_{i,k,h}}\right) \tag{4}$$

where the terms $\sum\limits_{\substack{k\in\{\mathcal{F}\}\\ j\neq k}} \psi_{k,h}\Gamma_{i,k,h}$ and $\sum\limits_{\substack{k\in\{\mathcal{P}\}\\ j\neq k}} \psi_{k,h}\Gamma_{i,k,h}$ in (3) and (4) denote the co-tier interference associated with the femto and pico tiers.

For a user $i_h \in \{\mathcal{E}_{pf,h}, \mathcal{M}_{pf,h}\}$ who subscribes to the services of MVNO, h, its spectrum efficiency is given as:

$$\gamma_{i,j,h} = \log_2\left(1 + \frac{\psi_{j\in c,h}\,\Gamma_{i,j\in c,h}}{\sigma^2 + \sum\limits_{\substack{p'\in c'\\ c\neq c'}} \sum\limits_{\substack{k\in\{\mathcal{P}\mathcal{F}'\}\\ j\neq k\\ \mathcal{P}\mathcal{F}\neq\mathcal{P}\mathcal{F}'}} \psi_{k,h}\Gamma_{i,k,h}}\right) \tag{5}$$

where $\psi_{j,h}$ is the transmit power of the access point in tier $j \in \{m, f, p, pf\}$. $\Gamma_{i,j,h}$ denotes the channel gain associated with a user i_h which belongs to MVNO h and an access point in tier $j \in \{m, f, p, pf\}$. The double-summation term in (5) is the inter-cluster interference with respect to the clustered femtocells in the coverage of the picocells.

2) HIGHLY MOBILE SLICE USERS

Without loss of generality, we assume that the URLLC users are based on V2X communication [22]. V2X communication is characterised by highly mobile users or vehicles in this case. Unlike the static or moderately mobile slice users, we include the small-scale fast fading component in the channel model in addition to the large scale factors. For a URLLC slice user, $r \in \mathcal{R}$, (engaged in V2N communications otherwise known as V2I as shown in Fig. 2) the path loss (i.e. the large scale fading) is expressed as [44]:

$$\rho_{r,m,h} = 128.1 + 37.6\log(d_{r,m,h}) \tag{6}$$

where $d_{r,m,h}$ is the distance between the URLLC slice user r and the macrocell m. However, for a vehicle in transmit mode in the V2V set, $\mathcal{W}$, its path loss model is dependent on its respective distance from the URLLC slice user-vehicle and it is given [45] as:

$$\rho_{w,m,h} = \begin{cases} 40 + 22.7\log(d_{r,w,h}) \\ \quad + 20\log\left(X_c\right), \quad d_{r,w,h} \leq d_{thres} \\ 9.45 + 40\log(d_{r,w,h}) - 17.3\log(T_w) \\ \quad - 17.3\log(T_r) + 2.7\log(X_c), \quad d_{thres} \leq d_{r,w,h} \end{cases} \tag{7}$$

where $d_{r,w,h}$ is the distance between the URLLC slice user r engaged in V2N and a car in transmit mode in the V2V set. X_c, T_w and T_r denote the carrier frequency in GHz, effective antenna height of the transmit vehicle, w, in V2V communication and effective antenna height of the receive vehicle, r, requesting for URLLC slice engaged in V2N. The threshold distance, d_{thres}, is given as [45]:

$$d_{thres} = \frac{4(T_w)(T_r)X_c}{speed\ of\ light} \tag{8}$$

Hence, for URLLC slice users, that is for vehicles engaged in V2N communication, the spectrum efficiency is given as:

$$\gamma_{r,m,h} = \log_2\left(1 + \frac{\psi_{m,h}\Gamma_{r,m,h}|\alpha_{r,m}|^2}{\sigma^2 + \sum\limits_{\substack{w\in\{\mathcal{W}\}\\ w\neq r}} \psi_w\Gamma_{w,r,h}|\alpha_{w,r}|^2}\right) \tag{9}$$

where $\psi_{m,h}$ and ψ_w denote the transmit powers of the macro base station in the macro-tier and the transmitting vehicle w in the V2V set, $\mathcal{W}$. Here, $\alpha_{r,m}$ and $\alpha_{w,r}$ denote the small-scale fading component. We assume that the small-scale fast fading component is independent and identically distributed (i.i.d) as $\mathcal{CN}(0, 1)$.

III. LATENCY-AWARE AND DYNAMIC RESOURCE MODEL

In this section, we explain the latency and dynamic resource allocation model. First, we discuss the Latency and Delay Model and later the dynamic resource model.

A. LATENCY AND DELAY MODEL

In order to guarantee the service rate of the mMTC and URLLC slice users within a latency threshold and a transmission delay bound, we consider the link-layer model and apply the effective capacity theory in [41], [42] which is based on the theory of large deviations. We employ the link-layer model owing to its ease of translating QoS metrics such as delay bounds and packet loss probability into guarantees; simple implementation process and high accuracy. The effective capacity of a slice use case is the maximum arrival rate it can accommodate to guarantee a QoS requirement which is specified by a QoS exponent. The effective capacity, being a robust statistical approach for QoS analysis, employs the QoS triplets of packets arrival rate $\lambda_{i,h}$, a maximum delay bound D_{max} and, a delay-bound violation probability threshold μ. The stochastic behaviour of the mMTC and URLLC slice

user can be modelled by their effective capacity, which is expressed as [41]:

$$\phi(\theta_{i,h}) = \lim_{t\to\infty} \frac{1}{t}\frac{1}{\theta_{i,h}} \log \mathbb{E}\left[e^{\theta_{i,h}Q^t_{i,h}}\right] \tag{10}$$

where $\theta_{i,h}$ is the QoS exponent, and $Q^t_{i,h}$ is the source data (i.e. packet arrivals) over a time interval of $[0, t)$. In this work, we assume a Poisson traffic process with an arrival rate of $\lambda_{i,h}$ packets/s. In computing large deviations, we apply the Moment Generating Function of a Poisson process $Q^t_{i,h}$ with an arrival rate of $\lambda_{i,h}$, which is given as [46]:

$$M_{Q_{i,h}}(\theta_{i,h}) = e^{\lambda_{i,h}(e^{\theta_{i,h}}-1)} \tag{11}$$

Substituting (11) into (10), therefore, (10) can be rewritten as:

$$\phi(\theta_{i,h}) = \frac{1}{t}\frac{1}{\theta_{i,h}} \log e^{\lambda_{i,h}t(e^{\theta_{i,h}}-1)} \tag{12}$$

We simplify (12) and can be expressed as:

$$\phi(\theta_{i,h}) = \frac{\lambda_{i,h}}{\theta_{i,h}}\left(e^{\theta_{i,h}} - 1\right) \tag{13}$$

To ensure that the delay QoS requirement is met, the delay violation probability should always be less than a given threshold of μ such that:

$$\Pr\{D(\infty) \geq D_{max}\} \leq \mu \tag{14}$$

D_{max} and $D(\infty)$ are the maximum delay-bound of a slice a use case (mMTC and URLLC) and the steady-state delay of a slice use case. Expression (14) is approximately equal to:

$$\Pr\{D(\infty) \geq D_{max}\} \approx e^{-\theta_{i,h}\lambda_{i,h}D_{max}} \tag{15}$$

However, we denote the packet size $L_{i,h}$ and hence the minimum achievable rate for a bounded delay violation probability of slice user (i.e. $i_h \in \mathcal{M}_{m,h}, \mathcal{M}_{p,h}, \mathcal{M}_{f,h}, \mathcal{M}_{pf,h}, \mathcal{R}_{m,h}$) is given as:

$$\vartheta_h^{thres} = -\frac{L_{i,h}\log(\mu)}{D_{max}\log_e(1 - \frac{\log(\mu)}{D_{max}\lambda_{i,h}})} \tag{16}$$

The proof of ϑ_h^{thres} is provided in Appendix A.

B. DYNAMIC RESOURCE ALLOCATION MODEL

In this work, the network resources of the clustered multi-tier multi-tenant heterogeneous network are pooled and virtualised to a cloud server by the InP and then allocated to the respective MVNOs contracted to it. The bandwidth allocated to each MVNO h in each tier t is given as:

$$\sum_{t\in\{m,\mathcal{F},\mathcal{P},\mathcal{PF}\}} \beta_{t,h}\mathcal{B} \tag{17}$$

where $\mathcal{B}$ is the total bandwidth of the network and $\beta_{t,h}$ is the network slice ratio of the MVNO h, in tier t. For the entire network, the sum network slice ratio is given as:

$$\sum_{h\in\mathcal{H}} \sum_{t\in\{m,\mathcal{F},\mathcal{P},\mathcal{PF}\}} \beta_{t,h} = 1 \tag{18}$$

The slice network ratio being dynamic is a function of slice user distribution and location, cell load characteristics, user slice use case QoS requirement; BS-User association, Interference and slice user mobility characteristics. For a user with a user-slice ratio, $\varphi_{i,j,m}$, which is dependent on the above mentioned factors, its logarithmic utility is given as:

$$\log\left(\vartheta_{i,j,h}\right) = \log\left(\mathcal{B}\,\beta_{t,h}\,\varphi_{i,j,h}\,\gamma_{i,j,h}\right) \tag{19}$$

Then, the question arises, how can network resources be dynamically allocated to MVNO and slice users while guaranteeing the slice use case QoS requirement?

IV. PROBLEM FORMULATION

In this section, the problem of latency-aware requirement and dynamic allocation of radio resources in a multi-tier multi-tenant heterogeneous 5G Network in a network slicing scenario is examined. In order to fully maximise the capacity of the network, we formulate a joint user-association InP-MVNO resource allocation problem in (20), as shown at the bottom of the next page.

The utility of each MVNO is the summation of the utility (or rate) of the four categories of slice users explained in Section II (B) and therefore, the network's sum utility is the maximisation of the several MVNOs utility which is given in (19). As shown in (20), the constraints C1, C5, C7 and C9 ensure that the minimum achievable data rate for eMBB slice users is guaranteed in all tiers. In addition, constraints C2, C4, C6, and C8 ensure that the minimum achievable rate of the latency-aware mMTC slice users is guaranteed in all tiers. For the URLLC users, constraint C3 ensures that the latency-aware received data rate is above the minimum threshold. Constraints C10, C11 and C12 impose the slice user-access point association constraints; a slice user can only be associated with one access point at a point in time. Constraint C10 ensures that a category II slice user is either associated with a femtocell f, or the macrocell m. Besides, C11 imposes the constraint that a category III slice user is either associated with a picocell p, or the macrocell m; while constraint C12 is to ensure that a category IV slice user is associated with a clustered femtocell pf, or its closest picocell p, or the macrocell m. Constraints C13-C15 are the relaxation of constraints C10-C12. From the foregoing, constraints C16-C19 in general highlight the bandwidth allocation requirement of the individual slice users in each of the tiers. The constraints for the bandwidth user-slice ratio for each user in the different tiers are given in C20-C23. The user-slice ratio is a fractional allocation indicator which must be between 0 and 1 i.e. a positive fractional value.

V. PROPOSED SOLUTION

In this section, we present the detailed description of the proposed solution to the latency-aware dynamic resource allocation problem in a multi-tier multi-tenant heterogeneous network stated in (20). First, we simplify (20) by transforming

the objective function into a tractable expression. The transformed expression is a summation of the utilities of the MVNOs in the 5G multi-tier network.This is done by considering each term of the objective function in (20) and transforming as follows in (21), as shown at the bottom of the next page. In order to solve (21), each term is expressed to fully capture its essence.

Herein, $\sum_{h\in\mathcal{H}} \vartheta_m \varphi_{i,m,h}$ denotes the aggregate utility of slice users associated to the macrocell and is given in (22), as shown at the bottom of the next page.

$$
\begin{aligned}
\max \sum_{h\in\mathcal{H}} \Bigg[& \sum_{i\in(\mathcal{E}_{m,h}\cup\mathcal{M}_{m,h}\cup\mathcal{R}_{m,h})} \log\left(\vartheta_{i,m,h}\right) + \sum_{f\in\mathcal{F}} \sum_{i\in(\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h})} \sum_{j\in\{m,f\}} \delta_{i,j,h}\ \log(\vartheta_{i,j,h}) \\
& + \sum_{p\in\mathcal{P}} \sum_{i\in(\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h})} \sum_{j\in\{m,\,p\}} \delta'_{i,j,h}\ \log\left(\vartheta_{i,m,h}\right) + \sum_{p\in\mathcal{P}} \sum_{pf\in\mathcal{PF}} \sum_{i\in(\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h})} \sum_{j\in\{m,\,p,\,pf\}} \delta''_{i,j,h}\ \log\left(\vartheta_{i,m,h}\right) \Bigg] \\
\text{s.t. C1}: \ & \vartheta_{i,m,h} \geq \lambda_{h,\mathcal{E}_m} L_{h,\mathcal{E}_m} \quad \forall\, i \in \mathcal{E}_{m,h},\ \forall\ h \in \mathcal{H} \\
\text{C2}: \ & \vartheta_{i,m,h} \geq \vartheta_h^{thres} \quad \forall\, i \in \mathcal{M}_{m,h},\ \forall\ h \in \mathcal{H} \\
\text{C3}: \ & \vartheta_{i,m,h} \geq \vartheta_h^{th} \quad \forall\, i \in \mathcal{R}_{m,h},\ \forall\ h \in \mathcal{H} \\
\text{C4}: \ & \delta_{i,j,h}\left[\vartheta_{i,m,h} - \vartheta_h^{thres}\right] \geq 0 \quad \forall\, i \in \mathcal{M}_{f,h},\ j \in \{m,f\},\ \forall\ h \in \mathcal{H} \\
\text{C5}: \ & \delta_{i,j,h}\left[\vartheta_{i,m,h} - \lambda_{h,\mathcal{E}_m} L_{h,\mathcal{E}_m}\right] \geq 0 \quad \forall\, i \in \mathcal{E}_{f,h},\ j \in \{m,f\},\ \forall\ h \in \mathcal{H} \\
\text{C6}: \ & \delta'_{i,j,h}\left[\vartheta_{i,m,h} - \vartheta_h^{thres}\right] \geq 0 \quad \forall\, i \in \mathcal{M}_{p,h},\ j \in \{m,p\},\ \forall\ h \in \mathcal{H} \\
\text{C7}: \ & \delta'_{i,j,h}\left[\vartheta_{i,m,h} - \lambda_{h,\mathcal{E}_p} L_{h,\mathcal{E}_p}\right] \geq \quad \forall\, i \in \mathcal{E}_{p,h},\ j \in \{m,p\},\ \forall\ h \in \mathcal{H} \\
\text{C8}: \ & \delta''_{i,j,h}\left[\vartheta_{i,m,h} - \vartheta_h^{thres}\right] \geq 0 \quad \forall\, i \in \mathcal{M}_{pf,h},\ j \in \{m,\ p,\ pf\},\ \forall\ h \in \mathcal{H} \\
\text{C9}: \ & \delta''_{i,j,h}\left[\vartheta_{i,m,h} - \lambda_{h,\mathcal{E}_{pf}} L_{h,\mathcal{E}_{pf}}\right] \geq 0 \quad \forall\, i \in \mathcal{E}_{pf,h},\ j \in \{m,\ p,\ pf\},\ \forall\ h \in \mathcal{H} \\
\text{C10}: \ & \sum_{j\in\{m,f\}} \delta_{i,j,h} = 1 \quad \forall\, i \in (\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h}) \\
\text{C11}: \ & \sum_{j\in\{m,\,p\}} \delta'_{i,j,h} = 1 \quad \forall\, i \in (\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h}) \\
\text{C12}: \ & \sum_{j\in\{m,\,p,\,pf\}} \delta''_{i,j,h} = 1 \quad \forall\, i \in (\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h}) \\
\text{C13}: \ & \delta_{i,j,h} \in \{0,1\} \quad \forall\, i \in (\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h}),\ j \in \{m,f\},\ \forall\ h \in \mathcal{H} \\
\text{C14}: \ & \delta'_{i,j,h} \in \{0,1\} \quad \forall\, i \in (\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h}),\ j \in \{m,\ p\},\ \forall\ h \in \mathcal{H} \\
\text{C15}: \ & \delta''_{i,j,h} \in \{0,1\} \quad \forall\, i \in (\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h}),\ j \in \{m,\ p,\ pf\},\ \forall\ h \in \mathcal{H} \\
\text{C16}: \ & \sum_{i\in(\mathcal{E}_{m,h}\cup\mathcal{M}_{m,h}\cup\mathcal{R}_{m,h})} \varphi_{i,m,h} + \sum_{f\in\mathcal{F}} \sum_{i\in(\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h})} \delta_{i,j,h}\varphi_{i,m,h} + \sum_{p\in\mathcal{P}} \sum_{i\in(\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h})} \delta'_{i,j,h}\varphi_{i,m,h} \\
& \quad + \sum_{p\in\mathcal{P}} \sum_{pf\in\mathcal{PF}} \sum_{i\in(\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h})} \delta''_{i,j,h}\varphi_{i,m,h} = 1 \\
\text{C17}: \ & \sum_{f\in\mathcal{F}} \sum_{i\in(\mathcal{E}_f\cup\mathcal{M}_f)} \delta_{i,f,h}\ \varphi_{i,f,h} = 1 \quad \forall\ h \in \mathcal{H} \\
\text{C18}: \ & \sum_{p\in\mathcal{P}} \sum_{i\in(\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h})} \delta'_{i,p,h}\ \varphi_{i,p,h} + \sum_{p\in\mathcal{P}} \sum_{pf\in\mathcal{PF}} \sum_{i\in(\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h})} \delta''_{i,p,h}\varphi_{i,p,h} = 1 \quad \forall\ h \in \mathcal{H} \\
\text{C19}: \ & \sum_{p\in\mathcal{P}} \sum_{pf\in\mathcal{PF}} \sum_{i\in(\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h})} \delta''_{i,pf,h}\varphi_{i,pf,h} = 1 \quad \forall\ h \in \mathcal{H} \\
\text{C20}: \ & \varphi_{i,m,h} \in (0,1) \quad i \in (\mathcal{E}_{m,h}\cup\mathcal{M}_{m,h}\cup\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h}\cup\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h}\cup\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h}) \\
\text{C21}: \ & \varphi_{i,f,h} \in (0,1) \quad i \in (\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h}) \\
\text{C22}: \ & \varphi_{i,p,h} \in (0,1) \quad i \in (\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h}\cup\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h}) \\
\text{C23}: \ & \varphi_{i,pf,h} \in (0,1) \quad i \in (\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h})
\end{aligned} \tag{20}
$$

The aggregate utility of slice users associated with the femtocell is given by:

$$\sum_{h\in\mathcal{H}}\sum_{f\in\mathcal{F}}\vartheta_f\varphi_{i,f,h} = \overbrace{\sum_{h\in\mathcal{H}}\sum_{f\in\mathcal{F}}\sum_{i\in\overline{\mathcal{E}'_f}}\log\left(\mathcal{B}\,\beta_{f,h}\,\varphi_{i,f,h}\,\gamma_{i,f,h}\right)}^{\text{net utility from users associated with } f \text{ but only located within } f}$$
$$\overline{\mathcal{E}'_f}=\{l\in\mathcal{E}_f\cup\mathcal{M}_f\,|\delta_{l,f,h}=1\} \tag{23}$$

Likewise from (21), $\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{F}}\vartheta_p\varphi_{i,p,h}$ which denotes the aggregate utility of slice users associated to the picocell is given as:

$$\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{F}}\vartheta_p\varphi_{i,p,h} = \overbrace{\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\sum_{i\in\overline{\mathcal{E}'_p}}\log\left(\mathcal{B}\,\beta_{p,h}\,\varphi_{i,p,h}\,\gamma_{i,p,h}\right)}^{\text{net utility from users associated with } p \text{ but only located within } p}$$
$$\overline{\mathcal{E}'_p}=\{q\in\mathcal{E}_p\cup\mathcal{M}_p\,|\delta'_{q,p,h}=1\}$$
$$+\overbrace{\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\sum_{pf\in\mathcal{PF}}\sum_{i\in\overline{\mathcal{E}''_p}}\Big[}^{\text{net utility from users associated with } p \text{ but within } pf}$$
$$\overline{\mathcal{E}''_p}=\{r\in\mathcal{E}_{pf}\cup\mathcal{M}_{pf}\,|\delta''_{r,p,h}=1\}$$
$$\log\left(\mathcal{B}\,\beta_{p,h}\,\varphi_{i,p,h}\,\gamma_{i,p,h}\right)\Big] \tag{24}$$

The net utility from all slice users associated with the clustered femtocells is denoted by $\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\sum_{pf\in\mathcal{PF}}\vartheta_{pf}\varphi_{i,pf,h}$ and given as:

$$\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\sum_{pf\in\mathcal{PF}}\vartheta_{pf}\varphi_{i,pf,h} = \overbrace{\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\sum_{pf\in\mathcal{PF}}\sum_{i\in\overline{\mathcal{E}''_p}}\Big[}^{\text{net utility from users associated with pf and only within } pf}$$
$$\overline{\mathcal{E}''_p}=\{r\in\mathcal{E}_{pf}\cup\mathcal{M}_{pf}\,|\delta''_{r,pf,h}=1\}$$
$$\log\left(\mathcal{B}\,\beta_{p,h}\,\varphi_{r,pf,h}\,\gamma_{r,pf,h}\right)\Big] \tag{25}$$

To further transform (22) - (25), the following Lemma is quite important.

Lemma 1: Given that $f = a \times b$. *Hence* $\log_z(f) = \log_z(a \times b)$. *Therefore,*

$$\log_z(a\times b)=\log_z(a)+\log_z(b) \tag{26}$$

By **Lemma 1**, the logarithmic expression in (22) can be expressed as:

$$\log\left(\mathcal{B}\,\beta_{m,h}\,\varphi_{i,m,h}\,\gamma_{i,m,h}\right) = \log\left(\mathcal{B}\,\beta_{m,h}\,\gamma_{i,m,h}\right)+\log\left(\varphi_{i,m,h}\right) \tag{27}$$

Similarly by **Lemma** 1, for (23);

$$\log\left(\mathcal{B}\,\beta_{f,h}\,\varphi_{i,f,h}\,\gamma_{i,f,h}\right) = \log\left(\mathcal{B}\,\beta_{f,h}\,\gamma_{i,f,h}\right)+\log\left(\varphi_{i,f,h}\right) \tag{28}$$

Likewise for (24),

$$\log\left(\mathcal{B}\,\beta_{p,h}\,\varphi_{i,p,h}\,\gamma_{i,p,h}\right) = \log\left(\mathcal{B}\,\beta_{p,h}\,\gamma_{i,p,h}\right)+\log\left(\varphi_{i,p,h}\right) \tag{29}$$

By **Lemma** 1, the logarithmic term in (25) can be simplified as:

$$\log\left(\mathcal{B}\,\beta_{pf,h}\,\varphi_{i,pf,h}\,\gamma_{i,pf,h}\right) = \log\left(\mathcal{B}\,\beta_{pf,h}\,\gamma_{i,pf,h}\right)+\log\left(\varphi_{i,pf,h}\right) \tag{30}$$

$$\sum_{h\in\mathcal{H}}\vartheta_m\varphi_{i,m,h}+\sum_{h\in\mathcal{H}}\sum_{f\in\mathcal{F}}\vartheta_f\varphi_{i,f,h}+\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\vartheta_p\varphi_{i,p,h}+\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\sum_{pf\in\mathcal{PF}}\vartheta_{pf}\varphi_{i,pf,h} \tag{21}$$

$$\sum_{h\in\mathcal{H}}\vartheta_m\varphi_{i,m,h} = \overbrace{\sum_{h\in\mathcal{H}}\sum_{i\in\mathcal{E}_{m,h}\cup\mathcal{M}_{m,h}\cup\mathcal{R}_{m,h}}\log\left(\mathcal{B}\beta_{m,h}\,\varphi_{i,m,h}\,\gamma_{i,m,h}\right)}^{\text{net utility from users within } m \text{ only}} + \overbrace{\sum_{h\in\mathcal{H}}\sum_{f\in\mathcal{F}}\sum_{i\in\mathcal{E}'_f}\log\left(\mathcal{B}\,\beta_{m,h}\,\varphi_{i,m,h}\,\gamma_{i,m,h}\right)}^{\text{net utility from users associated with } m \text{ but located within } f}$$
$$\mathcal{E}'_f=\{l\in\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h}|\delta_{l,m,h}=1\}$$
$$+\overbrace{\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\sum_{i\in\mathcal{E}'_p}\log\left(\mathcal{B}\,\beta_{m,h}\,\varphi_{i,m,h}\,\gamma_{i,m,h}\right)}^{\text{utility from users associated with } m \text{ but within } p}$$
$$\mathcal{E}'_p=\{q\in\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h}|\delta'_{q,m,h}=1\}$$
$$+\overbrace{\sum_{h\in\mathcal{H}}\sum_{p\in\mathcal{P}}\sum_{pf\in\mathcal{PF}}\sum_{i\in\mathcal{E}'_{pf}}\left[\log\left(\mathcal{B}\,\beta_{m,h}\,\varphi_{i,m,h}\,\gamma_{i,m,h}\right)\right]}^{\text{net utility from users associated with } m \text{ but within } pf} \tag{22}$$
$$\mathcal{E}'_{pf}=\{r\in\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h}|\delta''_{r,m,h}=1\}$$

The expression for $\varphi_{i,j,h}$ is presented in Appendix B. With (27) - (30), the optimisation problem in (20) is solved with $\beta_{t,h}$ being the decision variable. The hierarchical decomposition method [35] is adapted in solving (20) and the base station-slice user association is solved first in order to reduce the complexity of solving (20).

A. THE BASE STATION-SLICE USER ASSOCIATION

In the multi-tier heterogeneous network, the base station-slice user association is formulated as an integer programming problem [47], [48]. It is given as:

$$\max_{\delta} \sum_{h\in\mathcal{H}}\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}} \delta_{i,j,h}$$
$$\text{subject to } C24: \sum_{j} \delta_{i,j,h} \leq 1; \quad \forall h,\ \forall i,\ j \in \{f, p, pf\}$$
$$C25: \delta_{i,j,h} \in \{0, 1\}; \quad \forall h,\ \forall i,\ j \in \{f, p, pf\} \quad (31)$$

The base station-slice user association optimisation problem in (31) is adapted to the respective tiers taking into consideration the index of the association indicator for each tier. Constraint C24 is to ensure that the slice user can only be associated with one base station or access point. Constraint C25 is to ensure that the base station-slice user association indicator is Boolean. In this work, a maximum SINR matching algorithm is developed to solve the sub-problem in (31). The many-to-one matching [49] concept is adapted owing to its practical applications to heterogeneous wireless networks. Fig. 3 depicts the base station-slice user matching game for the multi-tier heterogeneous network and the different categories of the slice users.

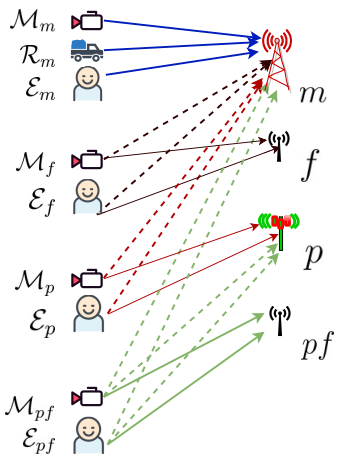


FIGURE 3. Base station-slice user matching game.

Consequently, we develop Algorithm 2 to solve (31) following the matching concept in Fig. 3

B. CONTINUOUS GENETIC ALGORITHM

We solve the transformed dynamic resource allocation problem in a multi-tenant multi-tier network in network slice scenario via the Continuous Genetic Algorithm (CGA). We adapt the Genetic Algorithm (GA) in solving the maximisation problem (20) owing to its robustness and effectiveness in finding the global optimal solutions compared to most heuristic algorithms [50]. Consequently, the GA can handle all kinds of optimisation problems and any constraints, such as linear and non-linear. In particular, the CGA, widely acknowledged for its high precision in representing solutions without extra-long strings of the chromosomes, hence its low computational complexity, less storage requirement and faster speeds [51], [52].

It is a stochastic search algorithm which is based on the principle of natural selection, biological reproduction and genetics [53], [54]. It starts with an initial randomly generated set of solutions otherwise called the population. The population satisfies the boundary conditions of the optimisation at hand. Each individual in the population is called a chromosome. A chromosome is a standard representation of solutions, otherwise called genes [55].

The GA determines the fitness of each chromosome in the population via the objective function in (20). In this work, the chromosomes with the best fitness values are selected via the roulette wheel technique, thus creating the different set of pairs referred to as parents for a crossover which is significantly governed by the probability of crossover P_c. The crossover process results in new chromosomes, otherwise called children. In order to mimic the process of natural

Algorithm 1 Latency-Aware Dynamic Resource Allocation

for $h \leftarrow 1$ to $|\mathcal{H}|$ **do**
  for $\mathcal{S}_h \leftarrow$ to $\{\mathcal{E} \cup \mathcal{M} \cup \mathcal{R}\}$ **do**
    for $i \in \mathcal{U}_{h,s}$[1] **do**
      optimally associate to an access point (Alg. 2) and (31)
      determine user cat. using Subsection II-B
    end for
  end for
  for $t \leftarrow \{m,\ \mathcal{F},\ \mathcal{P},\ \mathcal{PF}\}$ **do**
    for $k \leftarrow \{m,\ p,\ pf,\ f\}$ **do**
      for $i \in \mathcal{U}_{h,s}$ **do**
        determine $\gamma_{i,k,h}$ (2)-(5), (9)
        determine ϑ_h^{thres} (10) - (16)
      end for
      determine the cell load characteristics (22)-(25)
    end for
    optimally determine $\beta_{t,h}$ (17)-(19)
  end for
  for $\mathcal{S}_h \leftarrow$ to $\{\mathcal{E} \cup \mathcal{M} \cup \mathcal{R}\}$ **do**
    for $i \in \mathcal{U}_{h,s}$ **do**
      dynamically allocate radio resources (20), (Alg. 3)
    end for
  end for
end for

[1] $\mathcal{U}_{h,s} = (\mathcal{E}_{h,m} \cup \mathcal{E}_{h,f} \cup \mathcal{E}_{h,p} \cup \mathcal{E}_{h,pf} \cup \mathcal{M}_{h,m} \cup \mathcal{M}_{h,f} \cup \mathcal{M}_{h,p} \cup \mathcal{M}_{h,pf} \cup \mathcal{R}_{h,m})$

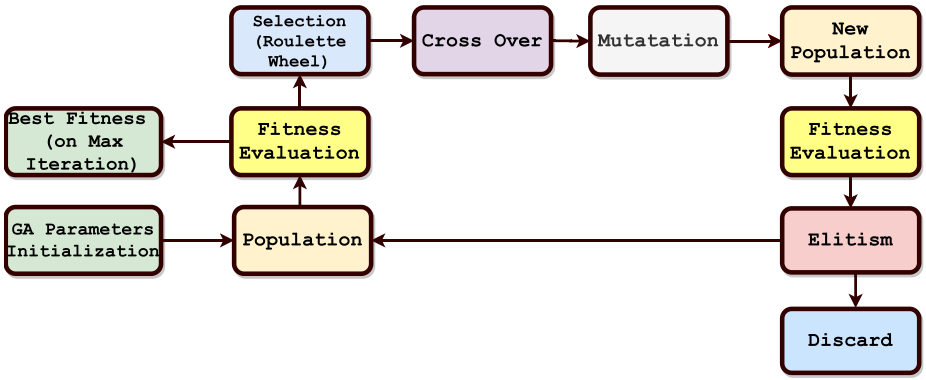


FIGURE 4. The GA process.

Algorithm 2 Base Station-Slice User Association

Input: $\psi_{j,h}$, $d_{i,j,h}$

if i is under the coverage of a femtocell **then**
  if femtocell is clustered **then**
    Calculate: $\gamma_{i,m,h}$, $\gamma_{i,p,h}$, $\gamma_{i,pf,h}$ (2), (4), (5)
    if $\gamma_{i,pf,h} \geq (\gamma_{i,m,h}$ & $\gamma_{i,p,h})$ **then**
      $\delta''_{i,pf,h} = 1$
    else
      if $\gamma_{i,p,h} \geq (\gamma_{i,m,h}$ & $\gamma_{i,pf,h})$ **then**
        $\delta''_{i,p,h} = 1$
      else
        $\delta''_{i,m,h} = 1$
      end if
    end if
    break;
  else
    Calculate $\gamma_{i,m,h}$, $\gamma_{i,f,h}$ (2), (3)
    if $\gamma_{i,f,h} \geq \gamma_{i,m,h}$ **then**
      $\delta_{i,f,h} = 1$
    else
      $\delta_{i,m,h} = 1$
    end if
    break;
  end if
else
  Calculate $\gamma_{i,p,h}$, $\gamma_{i,m,h}$ (4), (5)
  if $\gamma_{i,p,h} \geq \gamma_{i,m,h}$ **then**
    $\delta'_{i,p,h} = 1$
  else
    $\delta'_{i,m,h} = 1$
  end if
end if

reproduction, the genes of the children are mutated at birth, giving rise to a new population. The fitness of the new population is evaluated, and by means of elitism a small fraction of the best individuals from the old population are retained in the new population, and the others are discarded. This process is illustrated in Fig. 4 and corresponding parameter values given in Table 2.

Algorithm 3 CGA-Based Radio Resource Allocation

Input: P_m, P_c, P_e, g, y, t, H

number of genes, $n = |t| \cdot |H|$
Initialise the random population $\boldsymbol{A} = y \times n$

$$\boldsymbol{A} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{y1} & \beta_{y2} & \cdots & \beta_{yn} \end{bmatrix}$$

while iteration $\leq$ g **do**
  iteration = iteration + 1;
  Evaluate the fitness U_y of each chromosome in $\boldsymbol{A}$

$$\boldsymbol{U} = \begin{bmatrix} U_1 & U_2 & \cdots & U_y \end{bmatrix}$$

  Normalise the fitness vector $\boldsymbol{U}$

$$\hat{\boldsymbol{U}} = \frac{\boldsymbol{U}}{\|\boldsymbol{U}\|}$$

  Sort $\hat{\boldsymbol{U}}$ in descending order

$$[\sim, \text{index}] = \text{Sort}\,(\hat{\boldsymbol{U}}, \text{'descend'})$$

  Sort the population $\boldsymbol{A}$ according to index

$$\boldsymbol{A} = \boldsymbol{A}(\text{index:})$$

  Select the Chromosome in $\boldsymbol{A}$ wrt. the sorted fitness $\hat{\boldsymbol{U}}$ using the roulette wheel section.
  The probability of a chromosome k being selected P_k is given as

$$P_k = \frac{\hat{\boldsymbol{U}}}{\sum_{z=1}^{y} \hat{\boldsymbol{U}}_z}$$

  Carry out crossover using the P_c
  Mutate the genes of chromosomes with P_m
  Perform elitism on the initial population wrt. P_e
  Select population
end while

The pseudocode of the CGA-based radio resource allocation algorithm is shown in Algorithm 3. The algorithm follows the procedure of the GA model in Fig. 4.

TABLE 2. GA parameters and values.

Parameter	Value
Probability of Elitism (Pe)	0.2
Probability of Crossover (Pc)	0.95
Probability of Mutation (Pm)	0.9
Max. No. of Iteration (g)	30
No. of Chromosomes (y)	700

Algorithm 4 Spatial BnB-Based Radio Resource Allocation

Initialise the upper bound, ω^{ub}, of **(20)**
Set the list of region $\mathcal{G}$ to a single domain
Use the least lower bound rule to choose a subregion $\mathcal{A} \in \mathcal{G}$
while $\mathcal{G} \neq \emptyset$ **do**
  if $\omega^{\mathcal{A},lb} \geq \omega^{ub} - \pi$ **then**
    Delete $\mathcal{A}$ from $\mathcal{G}$
  else
    if $\omega^{\mathcal{A},ub} > \omega^{ub}$ **then**
      Partition A into subregions $\mathcal{A}_{left}$ *and* $\mathcal{A}_{right}$
    else
      $\omega^{ub} = \omega^{\mathcal{A},ub}$
      Delete all subregions in $\mathcal{G}$
      if $\omega^{\mathcal{A},ub} - \omega^{\mathcal{A},lb} \leq \pi$ **then**
        Delete $\mathcal{A}$ from $\mathcal{G}$
      end if
    end if
  end if
end while
if $\omega^{ub} = \infty$ **then**
  problem is infeasible
else
  ω^{ub} is the global optimal of the solution
end if

C. SPATIAL BRANCH AND BOUND ALGORITHM

A spatial branch and bound (sBB) method [37] is adapted to solve the maximisation problem in (20) in order to verify the optimality of the CGA-based results. The sBB algorithm gives a globally optimal solution, and it is shown in Algorithm 4. The sBB is a widely used deterministic search algorithm to solve the optimisation problem owing to its exact solutions [56]. The sBB iteratively searches the solution space of the defined problem. The wide range of solutions in the search space forms a hierarchical tree taking into consideration the upper and lower bounds of the solutions. These sets of feasible solutions are evaluated with respect to the objective function. If the evaluated solution does not result in a better solution than the current best solution, then it is discarded; however, if it is a better solution, the current best solution is discarded while the evaluated solution becomes the current best solution. This procedure is repeated until an optimal solution is discovered [57], [58]. The pseudo code of the sBB algorithm is illustrated in Algorithm 4.

VI. COMPLEXITY ANALYSIS

We examine the algorithms discussed in Section V. Herein, we focus on the time complexity of the algorithms. The time complexity is employed to determine the worst-case running time of an algorithm. Furthermore, we employ the big Omicron (big-$\mathcal{O}$) in our characterisation of the algorithms. The big-$\mathcal{O}$ notation gives a theoretical measure of the upper bound or worst-case scenario of the growth rate concerning the execution time (or memory) of an algorithm or a function. A detailed explanation of the big-$\mathcal{O}$ is given in [59], [60]. First, we examine the time complexity of the CGA and we further our analysis of the sBB algorithm. Additionally, we discuss the Latency-aware dynamic resource algorithm and, finally, the ORA resource allocation algorithm.

A. COMPUTATIONAL COMPLEXITY OF THE CGA

The computational complexity of the GA and other evolutionary meta-heuristics are quite difficult to determine owing to their stochastic behaviour. However, the big-$\mathcal{O}$ notation of the CGA adopted in this paper is given by $\mathcal{O}(g(y \cdot n(t, |\mathcal{H}|)))$, where g denotes the number of generation, y represents the number of chromosomes, and n denotes the size of the genes in a chromosome, which in this paper is a function of both the number of the tiers t in the network and the number of MVNOs $|\mathcal{H}|$.

B. COMPUTATIONAL COMPLEXITY OF THE sBB

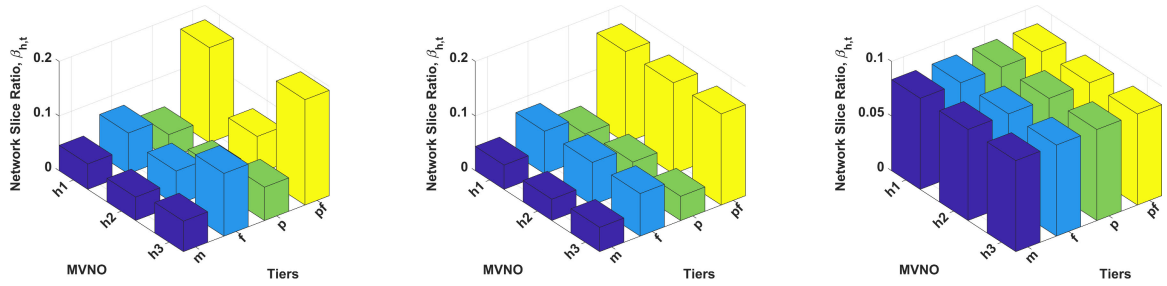
The time complexity of the sBB method depends on the size of the search tree. However, it is pertinent to note that the time complexity does not include the time for executing the branching rule or inserting nodes in the queue. sBB decomposes non-linear or non-convex objective functions symbolically and recursively into simple operations by applying simple operations [61] such as linear over- and underestimators given in [62]. Furthermore, an integer linear programming is NP-hard, hence optimal solutions would mostly require exponential upper bound (i.e. worst case) run time in tandem with input size [63], [64]. Therefore, the time complexity of the sBB is given by $\mathcal{O}(2^{t \cdot |\mathcal{H}|})$.

C. COMPUTATIONAL COMPLEXITY OF THE LATENCY-AWARE DYNAMIC RESOURCE ALLOCATION

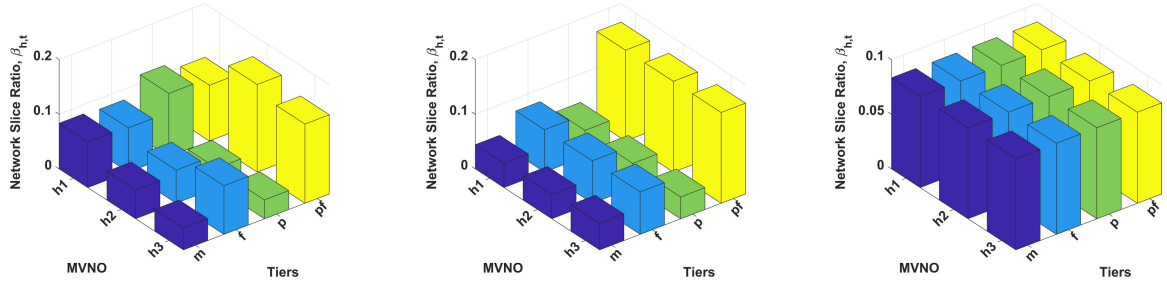
Furthermore, we examine the time complexity of the latency-aware dynamic resource allocation in a multi-tier multi-tenant network. It is given by $\mathcal{O}(|\mathcal{H}| \cdot t \cdot |\mathcal{S}_h|)$. The time complexity of the ORA is given in [38] as $\mathcal{O}(KM \log(\frac{1}{\epsilon}) + M^3)$, where M is the number of slice users and we have adapted K to be 1 to fit into the multi-tier multi-tenant slice network.

VII. NUMERICAL RESULTS

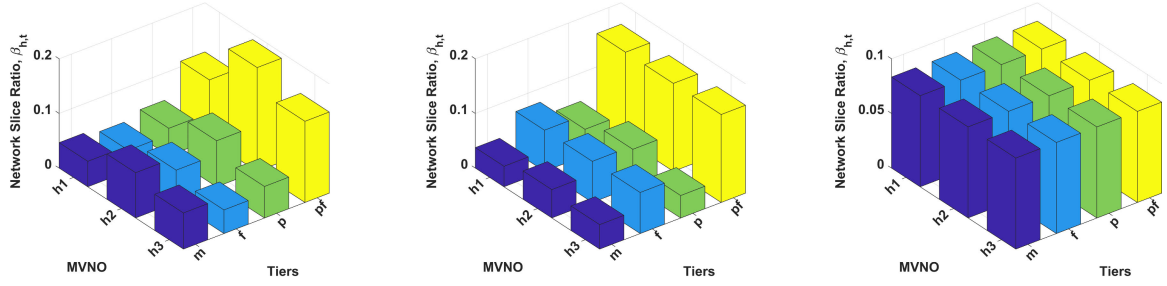
In this section, the performance of the proposed genetic algorithm (GA) intelligent latency-aware resource allocation scheme (GI-LARE) is evaluated via Monte Carlo based computer simulations in a Matlab environment. We considered a multi-tier multi-tenant network of MVNOs operating in



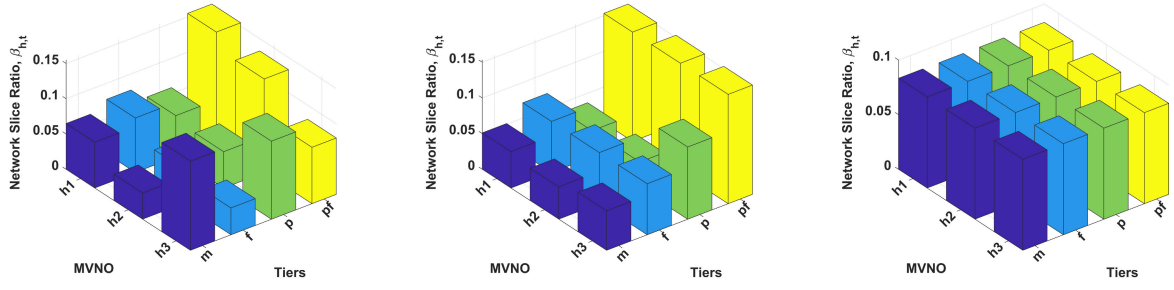
(a) GI-LARE scheme with a user density of 4 (b) sBB-based scheme with a user density of 4 (c) SS with a user density of 4



(d) GI-LARE scheme with a user density of 5 (e) sBB-based scheme with a user density of 5 (f) SS with a user density of 5



(g) GI-LARE scheme with a user density of 6 (h) sBB-based scheme with a user density of 6 (i) SS with a user density of 6



(j) GI-LARE scheme with a user density of 7 (k) sBB-based scheme with a user density of 7 (l) SS with a user density of 7

FIGURE 5. **Impact of the slice user density on the Network Slicing ratio $\beta_{t,h}$.**

an area of interest of 950m radius. The macro station was placed at the centre and surrounded with femtocells and picocells, which had a coverage radius of 50m and 250m, respectively. The multi-tier network consisted of 7 femtocells, 4 picocells and 5 clustered femtocells per picocells. Furthermore, the transmit powers budget of 15dBm, 30dBm, 36.9dBm, and 40dBm for V2V transmit mode, femtocells, picocells and the macrocell was considered. The different categories of slice users were randomly distributed across the different access points in all the tiers with a data packet arrival rate of 5 packet/s, 20 packet/s, and 20 packet/s for mMTC, eMBB and URLLC slice use cases with packet size of 1000bits, 9000bits, and 500bits, respectively. In addition, for URLLC slice users, we assumed the vehicles were moving at a velocity of 60Km/hr on a 4-lane highway with a lane width of 4m. For each simulation, 10000 iterations were generated and then averaged to obtain a numerical result.

A. IMPACT OF THE SLICE USER DENSITY

First, we evaluate the performance of the proposed algorithm with different network parameters. With an assumed maximum delay bound of 100ms and a maximum delay bound violation of 0.001, we investigate the impact of varying the slice user density on the network slice ratio per tier for each MVNO respectively and also its impact on the total network utility. In Fig. (5), the impact of the slice user density on the network slice ratio, $\beta_{h,t}$ is studied. We consider the different densities of slice users in the range of 4 to 7 for the respective slice categories in the different tiers for the 3 MVNOs. The proposed GI-LARE is compared with a Static resource scheme and also an exact solution from the sBB scheme. In Fig. 5(a), Fig. 5(b) and Fig. 5(c), we investigate the performance of the algorithms with a slice user density of 4. Fig. 5(d) to Fig. 5(f) show the results of the GI-LARE, sBB-based and SS Schemes for a user density of 5. Similarly, Fig. 5(g) to Fig. 5(i) show the results of the respective schemes when the user density is set to 6. Finally, Fig. 5(j) to Fig. 5(l) show the results of the GI-LARE, sBB-based and SS schemes with user density of 7. We observe that unlike the SS scheme, the GI-LARE and the sBB-based schemes dynamically respond to the variation of the user density in the respective tiers and MVNOs. The dynamic scheme ensures fairness in the different tiers while at the same time maximising the network utility.

From the foregoing, we investigate the impact of the user density on the network utility. Fig. 6 shows the impact of the user density on the total network utility. Similar to Fig. 5, the user density is set between 3 and 10 slice users. It is observed that as the user density increases, the total network utility increases owing to the increase in the utilisation of network resources. The GI-LARE scheme outperforms the SS (Static scheme) by an average of 25%, however its performance is almost the same as sBB-based scheme which gives a global optimum. Besides, the GI-LARE scheme outperforms the ORA when adopted to the multi-tier multi-tenant slice network.

B. IMPACT OF THE NETWORK BANDWIDTH

Fig. 7 and Fig. 8 present the effect of the total bandwidth of the network on network utility. In Fig. 7, we vary the total network bandwidth from 200MHz to 700MHz and set the user density at 5 users and the delay bound at 1ms. From Fig. 7, we can observe that the network utility increases as the total bandwidth of the network increases. This is owing to the fact that the utility increases as more resources are available to the network slice users. Similarly, in Fig. 8, the delay bound is set at 10ms and a user density of 5 users. Similar to Fig. 7, we observe that the network utility increases as the total bandwidth of the network increases. However, with a relax delay bound constraint of 10ms, the network utility is quite higher than that of the 1ms but with a compromise on the QoE. In both Fig. 7 and Fig. 8, the GI-LARE outperforms the SS and ORA schemes.

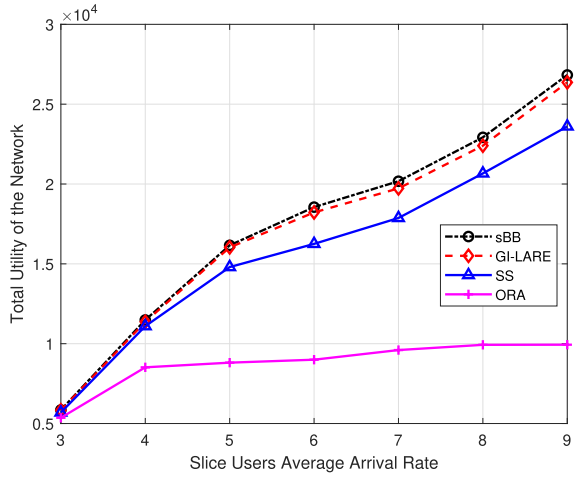


FIGURE 6. Effect of the slice user density on the total utility of the network.

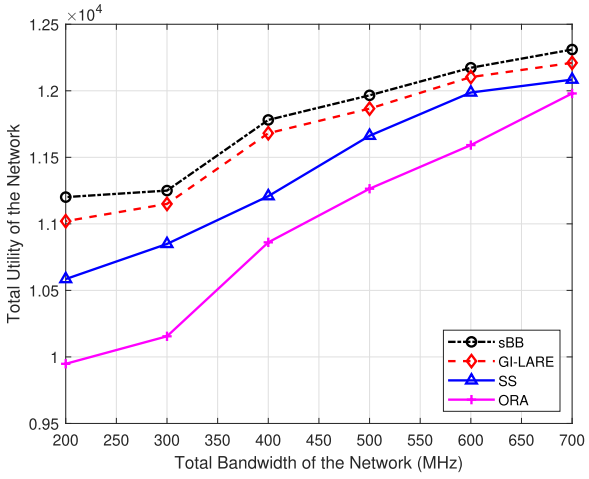


FIGURE 7. Impact of the total bandwidth on the total utility of the network at 1ms delay bound.

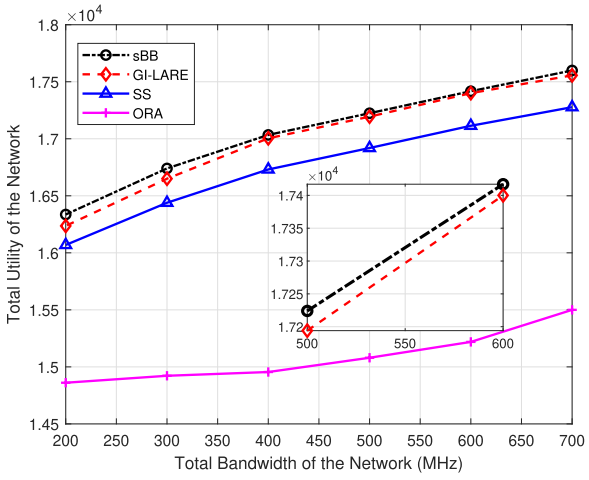


FIGURE 8. Impact of the total bandwidth on the total utility of the network at 10ms delay bound.

C. IMPACT OF THE DELAY BOUND

In Fig. 9 and Fig. 10, we show the impact of the delay bound on the network utility and effective bandwidth. With a user

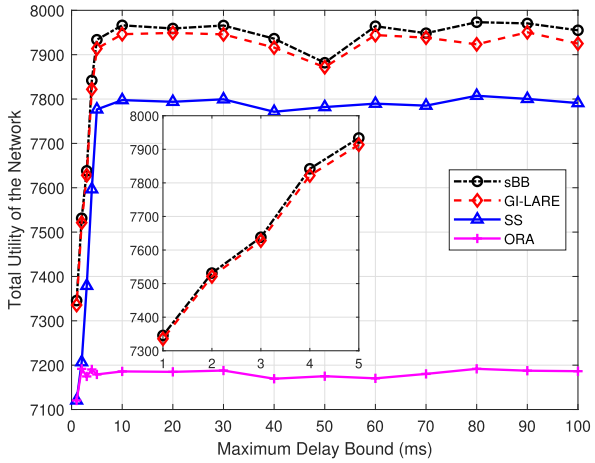


FIGURE 9. Effect of the delay bound on the total utility of the network.

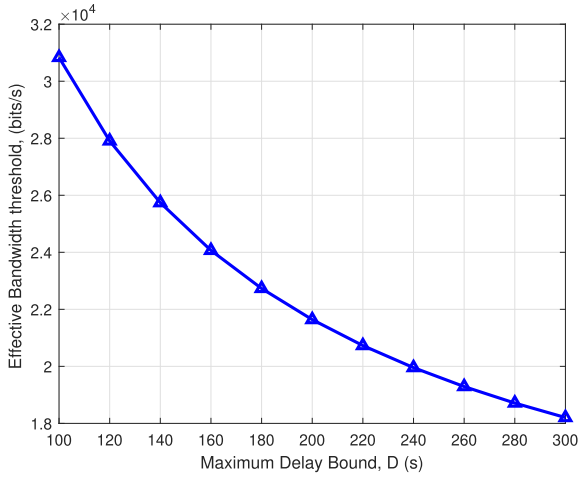


FIGURE 10. Effect of the maximum delay bound on the effective bandwidth threshold.

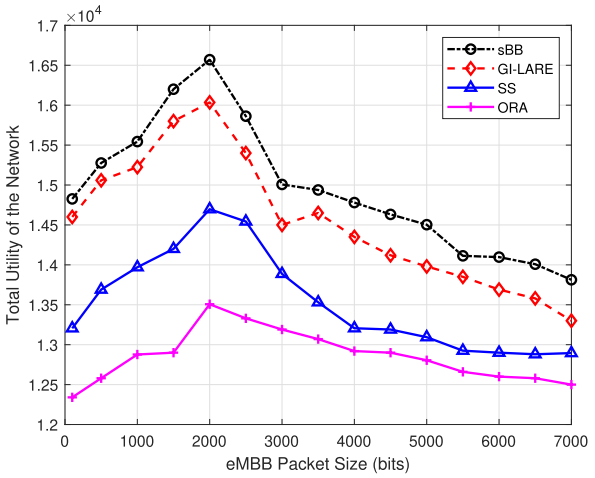


FIGURE 11. Impact of the eMBB data packet size on the total network utility at 1ms delay bound.

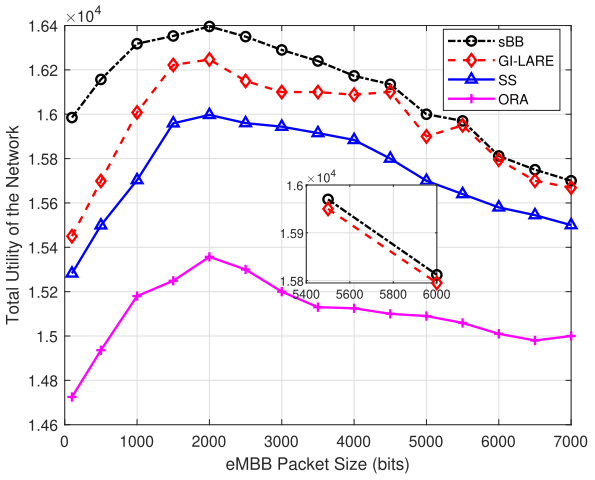


FIGURE 12. Impact of the eMBB data packet size on the total network utility at 100ms delay bound.

density of 2 users and a network bandwidth of 100MHz, in Fig. 9, we present the impact of the delay bound on the network utility. Similar to Fig. 7 and Fig. 8, there is a rapid increase in network utility for a delay bound relaxation from 1ms to 10ms; however, with a limited network resource, the utility remains constant despite the increase in the delay bound. Fig. 10 shows the effect of the delay bound on the effective bandwidth. As seen in constraints C2, C3, C5 and C6, the effective bandwidth greatly affects the received rate of the mMTC and URLLC slice users. We observe that as the maximum delay bound increases, the threshold decreases which is in tandem with (16).

D. IMPACT OF THE PACKET SIZE

Fig. 11 and Fig. 12 present the impact of the eMBB data packet size on the network utility. In Fig. 11, with a delay bound of 1ms, a user density of 5 users and a total network bandwidth of 200MHz, it can be observed that as the packet size increases, the net utility increases to about 2000 bits and then a dip occurs in the net utility. This can be ascribed to the bandwidth and power limitation of the network. However, the GI-LARE scheme outperforms the SS and ORA resource allocation schemes. Similar to Fig. 11, in Fig. 12, we further study the impact of the packet size on the network utility with delay bound and user density parameters set at 100ms and 5 users and follows the same trend with Fig. 11. However, the network utility fared better at a delay bound of 100ms than at a delay bound of 10ms.

E. IMPACT OF THE PACKET LOSS PROBABILITY

Fig. 13 shows the impact of the packet loss probability on the network utility. Typically, the packet loss includes loss due to errors in the network, buffer overflows and late arrivals of packets. We vary the packet loss probability range from 10^{-5} to 10^{-1}, with a user density of 5 users; a delay bound of 10ms; and a bandwidth of 200MHz. Although the network utility increases with lower packet loss probability, the GI-LARE outperforms the SS and ORA resource allocation

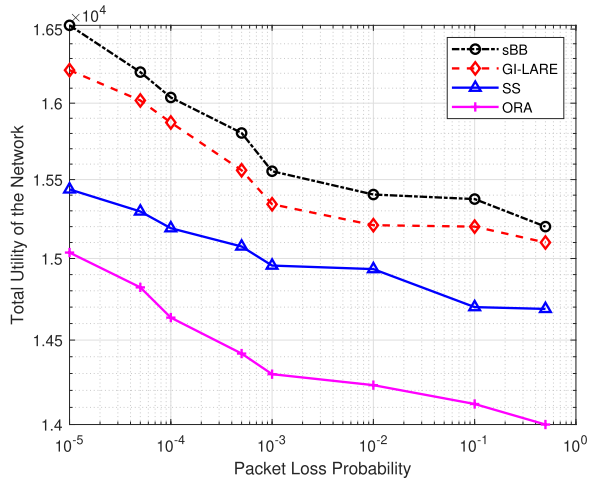


FIGURE 13. Impact of the Packet loss probability on the total network utility.

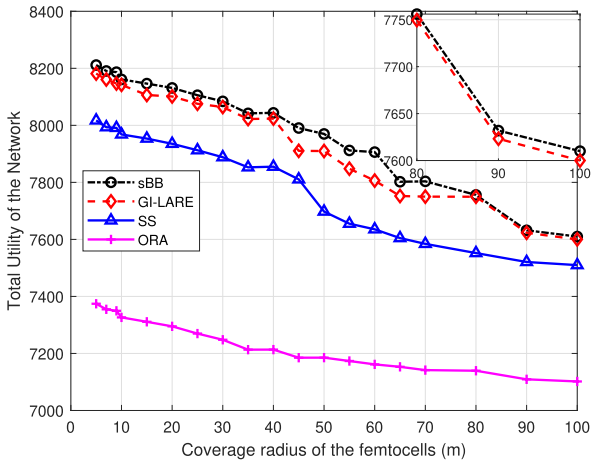


FIGURE 14. Effect of the Coverage radius of the femtocells on the network utility.

schemes. The lower the packet loss probability, the higher the probability that the packets are received. Consequently, we observe from Fig. 13 that the performance of the network can be improved by ensuring the packet loss probability value is low. Moreover, we observe that at a packet loss probability value of greater than 1, the network revenue and ultimately the performance of the network significantly degrades.

F. IMPACT OF THE COVERAGE RADIUS

Fig. 14 and Fig. 15 present the impact of the coverage radius of the femtocells and picocells on the network utility. We vary the coverage radius range from 10m to 100m and set the delay bound at 10ms, with a network bandwidth of 100MHz. We observe that the network utility increases when the coverage radius reduces. This is owing to better channel conditions of the respective slice users. Similar to Fig. 14, in Fig. 15, we examine the impact of the coverage radius of the 4 picocells on the network utility. We vary the coverage radius range

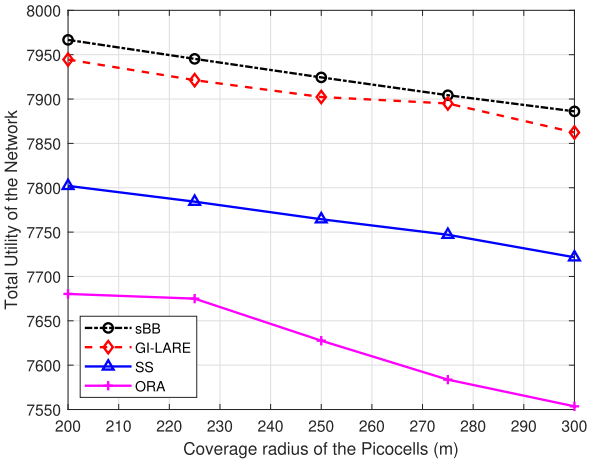


FIGURE 15. Effect of the Coverage radius of the picocells on the network utility.

from 200m to 300m and set the delay bound of 10ms. In the same trend with the femtocells, we observe that the network utility increases when the coverage radius reduces. However, it is not as significant that of the femtocells, as a result of the closeness of the femtocells to the slice users.

VIII. CONCLUSION

In this paper, we have proposed a genetic algorithm (GA) intelligent latency-aware resource allocation scheme (GI-LARE) that explicitly considered the latency and data rate constraints slices in a multi-tenant, multi-tier heterogeneous network. The optimisation problem was transformed and solved via the hierarchical decomposition method. Slice users were associated with base stations in different tiers by the concept of matching game theory, and the latency-aware dynamic resource allocation problem is solved using GI-LARE. Using the Monte Carlo simulation, GI-LARE was compared with the sBB-based, static slicing resource allocation (SS) and optimal resource allocation (ORA) schemes under different scenarios. Our dynamic GI-LARE scheme is shown to have outperformed the SS approach.

With the successes in the field of machine learning (ML) and, by extension, deep learning (DL) and generative adversarial network (GAN), together with the increasing influence of big data in mobile networks, the challenge of latency-aware dynamic resource allocation in a multi-tenant multi-tier network could be approached from the ML perspective. Our future work would address the dynamic resource allocation problem in a multi-tier, multi-tenant network slicing by adopting the concept of GAN.

APPENDIXES

APPENDIX A

Combining (14) and (15), we have:

$$e^{-\theta_{i,h}\lambda_{i,h}D_{max}} \leq \mu \tag{32}$$

Taking the logarithms of both sides of (32), this yields:

$$-\theta_{i,h}\lambda_{i,h}D_{max} = \log_e \mu \tag{33}$$

where $\lambda_{i,h}$ can also be said to be the minimum achievable rate in packet/s of slice i (i.e. $\mathcal{M}_{m,h}$, $\mathcal{M}_{p,h}$, $\mathcal{M}_{f,h}$, $\mathcal{M}_{pf,h}$, $\mathcal{R}_{m,h}$). From (33), we express $\lambda_{i,h}$ as:

$$\lambda_{i,h} = \frac{-\log_e \mu}{\theta_{i,h}D_{max}} \tag{34}$$

Based on the effective bandwidth theory, the delay-bound violation probability threshold can be guaranteed if and only if the effective bandwidth is equal to the minimum achievable rate. Therefore, from (13) and (34), we have:

$$\frac{-\log_e \mu}{\theta_{i,h}D_{max}} = \frac{\lambda_{i,h}}{\theta_{i,h}}\left(e^{\theta_{i,h}} - 1\right) \tag{35}$$

Therefore, $e^{\theta_{i,h}}$ can be expressed as:

$$e^{\theta_{i,h}} = 1 - \frac{\log_e \mu}{\lambda_{i,h}D_{max}} \tag{36}$$

Consequently from (36), $\theta_{i,h}$ is given as:

$$\theta_{i,h} = \log_e\left(1 - \frac{\log_e \mu}{\lambda_{i,h}D_{max}}\right) \tag{37}$$

Note the unit of $\lambda_{i,h}$ in (34) is packet/s and it can be transformed to bit/s by multiplying (34) by the packet size $L_{i,h}$. From the foregoing, by substituting (37) into (34), we now have the minimum achievable rate bounded by the delay violation probability, ϑ_h^{thres}, for a user which is given as:

$$\vartheta_h^{thres} = -\frac{L_{i,h}\log(\mu)}{D_{max}\log_e(1 - \frac{\log(\mu)}{D_{max}\lambda_{i,h}})} \tag{38}$$

APPENDIX B

The slice user ratio is a function of the aggregate number of slice users associated to an access point in a tier. Taking a holistic look at (22), constraints C16, and C20, $\varphi_{i,m,h}$ is given as:

$$\varphi_{i,m,h} = \frac{1}{\left[\begin{array}{c}|\mathcal{E}_{m,h}| + |\mathcal{M}_{m,h}| + |\mathcal{R}_{m,h}| + \sum\limits_{f\in\mathcal{F}}\sum\limits_{l\in\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h}}\delta_{l,m,h} \\ + \sum\limits_{p\in\mathcal{P}}\sum\limits_{q\in\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h}}\delta'_{q,m,h} \\ + \sum\limits_{p\in\mathcal{P}}\sum\limits_{pf\in\mathcal{PF}}\sum\limits_{r\in\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h}}\delta''_{r,m,h}\end{array}\right]} \tag{39}$$

For $\varphi_{i,f,h}$, taking into consideration (23), C17 and C21, it can be expressed as:

$$\varphi_{i,f,h} = \frac{1}{\left[\sum\limits_{f\in\mathcal{F}}\sum\limits_{l\in\mathcal{E}_{f,h}\cup\mathcal{M}_{f,h}}\delta_{l,f,h}\right]} \tag{40}$$

Similarly, for $\varphi_{i,p,h}$, taking into consideration (24), C18 and C22, it can be expressed as:

$$\varphi_{i,p,h} = \frac{1}{\left[\sum\limits_{p\in\mathcal{P}}\sum\limits_{q\in\mathcal{E}_{p,h}\cup\mathcal{M}_{p,h}}\delta'_{q,p,h} + \sum\limits_{p\in\mathcal{P}}\sum\limits_{pf\in\mathcal{PF}}\sum\limits_{r\in\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h}}\delta''_{r,p,h}\right]} \tag{41}$$

The user slice ratio in the clustered femtocells, $\varphi_{i,pf,h}$ looking at (25), C19 and C23, is expressed as:

$$\varphi_{i,pf,h} = \frac{1}{\left[\sum\limits_{p\in\mathcal{P}}\sum\limits_{pf\in\mathcal{PF}}\sum\limits_{r\in\mathcal{E}_{pf,h}\cup\mathcal{M}_{pf,h}}\delta''_{r,f,h}\right]} \tag{42}$$

REFERENCES

[1] N.-N. Dao, Y. Lee, S. Cho, E. Kim, K.-S. Chung, and C. Keum, "Multi-tier multi-access edge computing: The role for the fourth industrial revolution," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2017, pp. 1280–1282.

[2] J. Rendon Schneir, A. Ajibulu, K. Konstantinou, J. Bradford, G. Zimmermann, H. Droste, and R. Canto, "A business case for 5G mobile broadband in a dense urban area," *Telecommun. Policy*, vol. 43, no. 7, Aug. 2019, Art. no. 101813.

[3] E. J. Oughton and Z. Frias, "The cost, coverage and rollout implications of 5G infrastructure in britain," *Telecommun. Policy*, vol. 42, no. 8, pp. 636–652, Sep. 2018.

[4] A. Belbekkouche, M. M. Hasan, and A. Karmouch, "Resource discovery and allocation in network virtualization," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 1114–1128, 4th Quart., 2012.

[5] S. Kuklinski, L. Tomaszewski, K. Kozlowski, and S. Pietrzyk, "Business models of network slicing," in *Proc. 9th Int. Conf. Netw. Future (NOF)*, Nov. 2018, pp. 39–43.

[6] NGMN. (2016). *Description of Network Slicing Concept*. Accessed: Jul. 18, 2019. [Online]. Available: https://www.ngmn.org/fileadmin/user_upload/160113_Network_Slicing_v1_0.pdf

[7] ITU-R. (2017). *Minimum Requirements Related to Technical Performance For IMT-2020 Radio Interface(s)*. Accessed: Jul. 18, 2019. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf

[8] S. O. Oladejo and O. E. Falowo, "5G network slicing: A multi-tenancy scenario," in *Proc. Global Wireless Summit (GWS)*, Oct. 2017, pp. 88–92.

[9] S. O. Oladejo and O. E. Falowo, "Profit-aware resource allocation for 5G sliced networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2018, pp. 9–43.

[10] S. O. Oladejo and O. E. Falowo, "An energy-efficient resource allocation scheme for 5G slice networks," in *Proc. Southern Afr. Telecommun. Netw. Appl. Conf.*, Sep. 2019, pp. 1–4.

[11] Q. Sun, L. Tian, Y. Zhou, J. Shi, and Z. Zhang, "Incentive scheme for slice cooperation based on d2d communication in 5g networks," *China Commun.*, vol. 17, no. 1, pp. 28–41, Jan. 2020.

[12] G. O. Pérez, J. A. Hernández, and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 10, no. 6, pp. 573–581, Jun. 2018.

[13] C. Liang, F. R. Yu, H. Yao, and Z. Han, "Virtual resource allocation in information-centric wireless networks with virtualization," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9902–9914, Dec. 2016.

[14] Z. Jian, W. Muqing, M. Ruiqiang, and W. Xiusheng, "Dynamic resource sharing scheme across network slicing for multi-tenant C-RANs," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, Aug. 2018, pp. 172–177.

[15] Q. Ye, W. Zhuang, S. Zhang, A.-L. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, Oct. 2018.

[16] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Perez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6419–6432, Oct. 2018.

[17] J. Zheng, P. Caballero, G. de Veciana, S. J. Baek, and A. Banchs, "Statistical multiplexing and traffic shaping games for network slicing," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2528–2541, Dec. 2018.

[18] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Perez, "Network slicing games: Enabling customization in multi-tenant mobile networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 662–675, Apr. 2019.

[19] J. Zheng, P. Caballero, G. de Veciana, S. J. Baek, and A. Banchs, "Statistical multiplexing and traffic shaping games for network slicing," in *Proc. 15th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2017, pp. 1–8.

[20] H. D. R. Albonda and J. Perez-Romero, "An efficient RAN slicing strategy for a heterogeneous network with eMBB and V2X services," *IEEE Access*, vol. 7, pp. 44771–44782, 2019.

[21] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A comprehensive survey of RAN architectures toward 5G mobile communication system," *IEEE Access*, vol. 7, pp. 70371–70421, 2019.

[22] H. Yang, K. Zheng, K. Zhang, J. Mei, and Y. Qian, "Ultra-reliable and low-latency communications for connected vehicles: Challenges and solutions," 2017, *arXiv:1712.00537*. [Online]. Available: http://arxiv.org/abs/1712.00537

[23] L. Xu, J. Wang, H. Wang, T. Aaron Gulliver, and K. N. Le, "BP neural network-based ABEP performance prediction for mobile Internet of Things communication systems," *Neural Comput. Appl.*, pp. 1–17, Dec. 2019, doi: 10.1007/s00521-019-04604-z.

[24] I. Vila, O. Sallent, A. Umbert, and J. Perez-Romero, "An analytical model for multi-tenant radio access networks supporting guaranteed bit rate services," *IEEE Access*, vol. 7, pp. 57651–57662, 2019.

[25] G. Sun, Z. T. Gebrekidan, G. O. Boateng, D. Ayepah-Mensah, and W. Jiang, "Dynamic reservation and deep reinforcement learning based autonomous resource slicing for virtualized radio access networks," *IEEE Access*, vol. 7, pp. 45758–45772, 2019.

[26] J. Kwak, J. Moon, H.-W. Lee, and L. B. Le, "Dynamic network slicing and resource allocation for heterogeneous wireless services," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–5.

[27] N. A. Johansson, Y.-P.-E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 1184–1189.

[28] D. Harris, J. Naor, and D. Raz, "Latency aware placement in multi-access edge computing," in *Proc. 4th IEEE Conf. Netw. Softwarization Workshops (NetSoft)*, Jun. 2018, pp. 132–140.

[29] S. Zhang, H. Luo, J. Li, W. Shi, and X. Shen, "Hierarchical soft slicing to meet multi-dimensional QoS demand in cache-enabled vehicular networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2150–2162, Mar. 2020.

[30] A. Alnasser, H. Sun, and J. Jiang, "Cyber security challenges and solutions for V2X communications: A survey," *Comput. Netw.*, vol. 151, pp. 52–67, Mar. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128618306157

[31] A. Ghosal and M. Conti, "Security issues and challenges in V2X: A survey," *Comput. Netw.*, vol. 169, Mar. 2020, Art. no. 107093. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128619305857

[32] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu, and M. Kadoch, "Dynamic resource allocation with RAN slicing and scheduling for uRLLC and eMBB hybrid services," *IEEE Access*, vol. 8, pp. 34538–34551, 2020.

[33] H. Khan, S. Samarakoon, and M. Bennis, "Enhancing video streaming in vehicular networks via resource slicing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3513–3522, Apr. 2020.

[34] W. Guan, X. Wen, L. Wang, and Z. Lu, "On-demand cooperation among multiple infrastructure networks for multi-tenant slicing: A complex network perspective," *IEEE Access*, vol. 6, pp. 78689–78699, 2018.

[35] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[36] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2013.

[37] L. Yu, E. Karipidis, and E. G. Larsson, "Coordinated scheduling and beamforming for multicell spectrum sharing networks using branch & bound," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 819–823.

[38] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for high-reliability low-latency vehicular communications with packet retransmission," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6219–6230, Jul. 2019.

[39] M. Gonzalez-Martin, M. Sepulcre, R. Molina-Masegosa, and J. Gozalvez, "Analytical models of the performance of C-V2X mode 4 vehicular communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1155–1166, Feb. 2019.

[40] G. Naik, B. Choudhury, and J.-M. Park, "IEEE 802.11 bd & 5G NR V2X: Evolution of radio access technologies for V2X communications," *IEEE Access*, vol. 7, pp. 70169–70184, 2019.

[41] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 630–643, May 2003.

[42] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," in *Proc. 1st Int. Conf. Broadband Netw.*, Oct. 2004, pp. 527–536.

[43] S. O. Oladejo and O. E. Falowo, "Latency-aware dynamic resource allocation scheme for 5G heterogeneous network: A network slicing-multitenancy scenario," in *Proc. Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Oct. 2019, pp. 1–7.

[44] L. Liang, G. Y. Li, and W. Xu, "Resource allocation for D2D-enabled vehicular communications," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3186–3197, Jul. 2017.

[45] NGMN. (2007). *Winner II Channel Models, Standard IST-4-027756 WINNER II D1.1.2 v1.2*. Accessed: May 18, 2019. [Online]. Available: https://www.cept.org/files/8339/winner2%20-%20final%20report.pdf

[46] J. Gravner, "Lecture notes for introductory probability," Dept. Math., Univ. California, Davis, Davis, CA, USA, Dec. 2017.

[47] Z. Mlika, M. Goonewardena, W. Ajib, and H. Elbiaze, "User–base-station association in HetSNets: Complexity and efficient algorithms," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1484–1495, Feb. 2017.

[48] Z. Mlika, E. Driouch, and W. Ajib, "User association under SINR constraints in HetNets: Upper bound and NP-hardness," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1672–1675, Aug. 2018.

[49] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52–59, May 2015.

[50] A. Jafari, T. Khalili, E. Babaei, and A. Bidram, "A hybrid optimization technique using exchange market and genetic algorithms," *IEEE Access*, vol. 8, pp. 2417–2427, 2020.

[51] S. Mirjalili, *Evolutionary Algorithms and Neural Networks*. Reading, MA, USA: Springer, 2018.

[52] R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*, 2nd ed. Hoboken, NJ, USA: Wiley, 2004.

[53] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. New Delhi, India: Dorling Kindersley, 2008.

[54] L. Jiacheng and L. Lei, "A hybrid genetic algorithm based on information entropy and game theory," *IEEE Access*, vol. 8, pp. 36602–36611, 2020.

[55] B. Han, J. Lianghai, and H. D. Schotten, "Slice as an evolutionary service: Genetic optimization for inter-slice resource management in 5G networks," *IEEE Access*, vol. 6, pp. 33137–33147, 2018.

[56] L. Liberti, *Introduction to Global Optimization*. École Polytechnique, Palaiseau, France, 2008.

[57] P. M. Castro, "Spatial branch-and-bound algorithm for MIQCPs featuring multiparametric disaggregation," *Optim. Methods Softw.*, vol. 32, no. 4, pp. 719–737, Jul. 2017, doi: 10.1080/10556788.2016.1264397.

[58] P. Kirst, O. Stein, and P. Steuermann, "Deterministic upper bounds for spatial branch-and-bound methods in global minimization with nonconvex constraints," *TOP*, vol. 23, no. 2, pp. 591–616, Jul. 2015, doi: 10.1007/s11750-015-0387-7.

[59] D. E. Knuth, "Big omicron and big omega and big theta," *ACM SIGACT News*, vol. 8, no. 2, pp. 18–24, Apr. 1976.

[60] P. Black, "Big-O notation, dictionary of algorithms and data structures," US Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep., 2008.

[61] E. M. B. Smith and C. C. Pantelides, "A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs," *Comput. Chem. Eng.*, vol. 23, nos. 4–5, pp. 457–478, May 1999.

[62] G. P. McCormick, "Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems," *Math. Program.*, vol. 10, no. 1, pp. 147–175, 1976.
[63] M. R. Gary and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA, USA: Freeman, 1979.
[64] D. S. Johnson, "The NP-completeness column: An ongoing guide," *J. Algorithms*, vol. 7, no. 4, pp. 584–601, Dec. 1986.

SUNDAY OLADAYO OLADEJO (Graduate Student Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from the Federal University of Technology, Akure, Nigeria, in 2004, and the M.Eng. degree in communication engineering from the Federal University of Technology, Minna, Nigeria, in 2016. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Cape Town, South Africa.

From 2007 to 2017, he was a Senior Core Network Engineer with Glo-Mobile, Nigeria. His research interests include radio resource management in wireless networks and artificial intelligence.

OLABISI EMMANUEL FALOWO (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Cape Town, South Africa, in 2008.

He is currently an Associate Professor with the University of Cape Town. He has published over 100 technical articles in peer-reviewed conference proceedings and journals. His primary research interest includes radio resource management in heterogeneous wireless networks.