Priority-Based Downlink Wireless Resource Provisioning for Radio Access Network Slicing

Abdullah R. Hossain, *Student Member, IEEE*, and Nirwan Ansari, *Fellow, IEEE*

***Abstract*—This paper examines network slicing within the radio access network which employs an orthogonal frequency division multiple access system for downlink communications. Its radio resources are allocated among slices configured for specific 5G use cases as well as other non-5G services. This work sets out to achieve optimal resource provisioning and power allocation within the context of priority slicing where slices are assigned priority levels by the base station. A mixed-integer non-linear programming problem is formulated to maximize the throughput of a best effort slice while satisfying the constraints of the high priority 5G slices. The problem is simplified and relaxed into a convex optimization problem. Three scenarios under various conditions are presented to serve as benchmarks. The impact of the chosen schedulers on the allocations, intra-slice, and inter-slice contentions is also examined. The results highlight that the wireless network can satisfy the quality-of-service requirements of opposing levels of priorities of traffic while simultaneously mitigating both inter-slice and intra-slice contentions under conditions mirroring that of practical wireless network deployments. Furthermore, it is proven that the convex approximation problem serves as a reasonable approximation of the original MINLP problem.**

***Index Terms*—5G, wireless resource allocation, priority, radio access network, inter-slice contention, intra-slice contention, scheduler, network slicing.**

I. Introduction

THE explosive growth of mobile devices, automation, and inter-connectivity is quickly revealing the inadequacy of the current networking infrastructure; it is not suited to handle such an increase in traffic and demand [1]. Moreover, the fifth generation (5G) use cases, broadly defined as enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable low latency communications (URLLC), have divergent requirements [2]. The current one-size-fits-all approach succumbs to the challenge of fully supporting such services including vehicle-to-everything (V2X) communications which often require strict minimal latency, edge computing, security, and extreme robustness [3]–[4]. Furthermore, the sheer volume of such critical V2X traffic is undoubtedly anticipated to overwhelm the contemporary networks [5]–[6]. Consequently, recent research has proposed and examined numerous approaches to address the above limitations; among such propositions is network slicing.

Network slicing research within the core network has been relatively well studied and thus the focus has shifted to the radio access network (RAN); the research in this area has been multi-dimensional [7]. Several works have explored novel frameworks for traffic and load balancing for software-defined network (SDN) enabled RANs while others have sought to enhance the energy efficiency of the RANs; both via green and hybrid energy sources [8]–[9]. Many works within the RAN slicing literature have focused on slice orchestration, management, architecture, and control layer aspects [10]–[13]. Others have investigated slicing costs for customers, mobile network operators (MNOs), admission control, energy efficiency, and combined eMBB and URLLC slicing [14]–[19]. Ultimately, the overarching objective of such works, in the grand scheme of things, is to achieve secure end-to-end (E2E) network slicing.

To that end, works in this area have studied the impact of artificial intelligence (AI) based slicing and resource allocations [20]–[22], latency-constrained slicing [23], and network function virtualization (NFV) deployments and slicing [24]–[26]. While there has been consideration for prioritized traffic in fronthauling in the passive optical networks (PONs) to support the RANs [27], to the best of our knowledge however, there has not been ample consideration for wireless RAN slicing from the context of priority slicing where certain services or traffic are given higher priority than others. A network which does not assign priorities to network traffic essentially is rigid; it lacks flexibility and malleability in traffic management, routing, resource allocation, etc. Treating all traffic types equally is a hallmark of the one-size-fits-all approach which essentially limits the maximum potential of 5G networks. Thus, the need for priority slicing is quite essential to maximize the robustness of a network especially one which is expected to support applications with extremely heterogenous quality of service (QoS) requirements. In this work, priority slicing is investigated where both 5G and non-5G traffic are multiplexed together and contend with each other for network resources (Fig. 1).

The contributions of this work are briefly summarized as follows:

- We formulate a joint power and bandwidth allocation problem which maximizes the aggregate throughput of a specific slice.
- We implement a prioritized scheduling scheme for the slices where the mentioned 5G use cases are considered to

Manuscript received December 18, 2020; revised April 27, 2021; accepted July 5, 2021. Date of publication July 9, 2021; date of current version September 17, 2021. The review of this article was coordinated by Dr. Ai-Chun Pang. *(Corresponding author: Abdullah R. Hossain.)*

The authors are with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07017 USA (e-mail: arh24@njit.edu; nirwan.ansari@njit.edu).

Digital Object Identifier 10.1109/TVT.2021.3095901

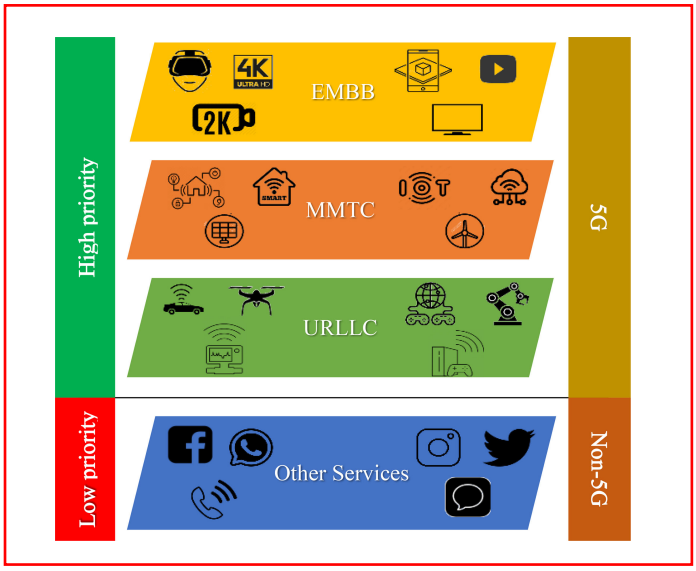

Fig. 1. Hierarchical representation of priority-based resource allocation in RAN slicing.

Fig. 2. User equipment classification for the base station schedulers (slices are not shown).

be high priority traffic while all others are designated as low priority. This is accomplished by utilizing multiple schedulers simultaneously to provision power and resources for the users across the slices as per their slice associations; such a feature can be exploited in a sliced network.

- We originally formulate a mixed integer mixed-integer non-linear programming (MINLP) problem which considers the frequency dependent channel conditions of all the users when carrying out the allocations. The MINLP problem is then relaxed and transformed into a convex problem (partially by disregarding the frequency dependent channel conditions) to reduce the computational load on the schedulers and allocation algorithms.
- We discuss the extensive simulation results of multiple scenarios which are reflective of the real-life conditions associated with the 5G RAN deployments. These scenarios serve as useful performance benchmarking tools. We also examine in detail how the specific schedulers impact the optimal allocations.
- We also demonstrate and prove that the convex approximation is very near optimal by solving both the MINLP and convex problems. We also note that despite us disregarding the frequency dependent channel conditions of each user (uniform gain experienced by each user), the results were reasonably accurate.
- Lastly, the complexity of the problem is discussed and proven in the Appendix.

The rest of the paper is organized as follows. Section II presents the system model and original MINLP problem formulation which is subsequently simplified and transformed into a convex optimization problem. Section III presents three benchmark scenarios and extensively discusses the simulation results as well as the unique impact of the schedulers employed in each slice by the BS. It also concludes with a comparison of the original problem solutions with the approximate solutions and demonstrates the reasonable accuracy of the relaxed convex problem. Section IV presents the concluding remarks of this work and presents future avenues of exploration. Finally, the Appendix discusses and proves the complexity of the original MINLP problem formulated in this work.

II. System Model and Problem Formulation

Consider a single cell which is utilizing an orthogonal frequency division multiple access (OFDMA) scheme for downlink communications (Fig. 2). The total system bandwidth is divided into N resource blocks (RBs); each is B kHz wide and can be utilized by only one user at most. The set of RBs is $\mathcal{N} = \{\mathrm{R}_1, \mathrm{R}_2, \ldots, R_Z\}$. The RAN is partitioned into four distinct slices; the slices are denoted by the set $\mathcal{S} = \{\mathrm{VR}, \mathrm{VS}, \mathrm{IoT}, \mathrm{BE}\}$. The first three slices are dedicated to virtual reality (VR), video streaming (VS), and Internet-of-Things (IoT) which are considered to fall under the eMBB and mMTC use cases of 5G; as such, they are assigned a high priority. The fourth slice is a non-5G general purpose slice assigned a best effort (BE)/low priority. The user set is U which is the union of subsets U_s where s indexes the slices. Each user is assumed to be associated with a single slice at most and cannot move between slices.

For clarity's sake, we note that the non-5G traffic (the BE slice) is considered to be any traffic that is not generated or received by devices that are not necessarily IoT sensors, autonomous driving vehicles, industrial manufacturing sensors, unmanned-aerial-vehicles (UAVs), remotely operated medical equipment, and so forth. Such users can engage in occasional light internet browsing, text messaging, voice over IP (VoIP) calls, or simply be in transit between one BS coverage area to another. Such users can be collectively admitted by the MNOs into a general-purpose, low priority slice.

A. Problem Formulation

The objective is to maximize the aggregate throughput of the BE slice while meeting the minimal requirements of the 5G

TABLE I
DEFINITION OF NOTATIONS

Notation	Definition
$a_{n,u}$	Binary variable for the assignment of resource block n for user u.
B	Resource block bandwidth.
D_s	Data rate of slice s.
$g_{n,u}$	Channel gain on resource block n for user u.
N	Total amount of resource blocks.
N_0	Noise power spectral density.
$\eta_{n,u}$	Noise on resource block n for user u.
n_u	Number of resource blocks allocated to user u.
P	Maximum base station transmission power.
p_u	Base station transmission power allocated for user u.
$p_{n,u}$	Base station transmission power allocated on resource block n for user u.
P_n	Maximum base station transmission power on resource block n.
y_u	Average channel gain to noise ratio for user u.

high priority slices. Considering the system model, the problem is formulated as follows;

$$\max_{p_{n,u},\, a_{n,u}} \sum_{u\in\mathcal{U}_4}\sum_{n\in\mathcal{N}} a_{n,u}\, B\log\left(1+\frac{g_{n,u}\, p_{n,u}}{B\eta_{n,u}}\right) \quad (1)$$

$$\text{s.t.} \sum_{n\in\mathcal{N}} a_{n,u}\, B\log\left(1+\frac{g_{n,u}\, p_{n,u}}{B\eta_{n,u}}\right) \geq Ds, \forall u\in\mathcal{U}s, \forall \text{s}\ \in\mathcal{S}, \quad (2)$$

$$a_{n,u}\, p_{n,u} \leq P_n, \forall n\in\mathcal{N}, \forall u\in\mathcal{U},) \quad (3)$$

$$a_{n,u}\in\{0,\ 1\}, \forall n\in\mathcal{N}, \forall u\in\mathcal{U}, \quad (4)$$

$$\sum_{u\in\mathcal{U}} a_{n,u} \leq\ 1, \forall n\in\mathcal{N}. \quad (5)$$

$a_{n,u}$ is a binary variable which indicates if RB n is allocated to user u. $g_{n,u}$, $p_{n,u}$, and $\mathcal{N}\mathrm{n,u}$ are the frequency dependent channel gain, base station (BS) transmission power, and noise for user u on RB n, respectively. Pn is the maximum transmission power of the BS on RB n while D_s denotes the data rate associated with slice s. A complete list of notations is provided in Table I.

Equation (1) maximizes the aggregate throughput of the BE slice and is subject to the constraints in (2)–(5). Equation (2) ensures that all the users in their respective slices are provisioned sufficient resources to meet their minimum guaranteed data rates. Equation (3) imposes that the allocated power to user u on any of its allocated RBs does not exceed the maximum power constraint for that RB. Equation (4) imposes that $a_{n,u}$ is a binary variable indicative of whether or not RB n is allocated to user u. Equation (5) guarantees that no more than one user is associated with a RB.

The rationale behind the objective in this problem scenario is as follows: the maximization of the BE slice throughput helps to mitigate the resource starvation problem caused by the users of the high priority slices i.e., inter-slice contention. Maximizing the total network throughput, which is typically the objective in most works, does not necessarily guarantee a reasonable allocation for the BE slice users; such an overall maximization can be just as well achieved via the throughput maximization of other higher priority slices. Furthermore, the number of users within the non-5G slice in certain restricted cases or coverage areas (i.e., rural areas) may significantly exceed that of the 5G slices. This can be attributed to much of their typical daily usage habits which do not necessarily burden the BS greatly. Thus, the stated objective function in Eq. (1) not only mitigates the inter-slice resource starvation problem that such users may face due to the 5G traffic, but it also increases the number of non-5G users served. This will be demonstrated in the various results discussed in Sections III-B and III-C.

We also note that while the problem defined in Eqs. (1) to (5) is formulated from a perspective of throughput enhancement only, 5G use cases can however prioritize different requirements. For certain use cases, a higher premium is placed on latency rather than throughput; in other cases, energy efficiency (power consumption minimization) is more of a concern than is latency (such as for IoT applications). We are examining this problem from a relatively simpler perspective to demonstrate slicing over different classes of traffic which does not necessarily require for heterogenous requirements to be considered; for a more robust scenario, heterogenous constraints can be included to address other parameters associated with various service types. In bandwidth limited scenarios, spectral efficiency is perhaps the most important metric of network performance. Furthermore, depending on the direction of transmission (uplink vs downlink), different objectives and requirements should be considered. Hence, multiple objective optimization problems can be formulated to simultaneously cater to different service types. Such heterogeneous objectives and service requirements are being considered in a later work.

B. Problem Simplification and Transformation

The problem defined in Eqs. (1) to (5) increases in complexity with respect to the scale of users and resource blocks since it must comb through each RB and decide whether to allocate it to a user or not. This process is then repeated for each user within the network. Evidently, this is computationally taxing and time-consuming. Moreover, this is a well-known multi-user, multi-subcarrier and power allocation problem which belongs to a class of optimization problems known as MINLP. Arriving at the optimal solution for such a problem is known to be NP-hard and therefore may not be achievable in real-time [28]. The proof of the complexity of a problem of this form is relegated to the Appendix.

To simplify the problem, it is assumed that each user experiences a uniform channel gain over all the RBs. Effectively, this decouples a user's channel condition from each specific RB (hence u and n are no longer coupled together). Hence, instead of the frequency dependent channel gain denoted as $\mathrm{gu,n}$, we have an average channel gain to noise ratio denoted $\mathrm{yu} = \mathrm{gu}/\mathrm{BN}_0$ where N_0 is the noise spectral density. Consequently, the BS will transmit with equal power over all the allocated RBs for a single user. While this single assumption

significantly simplifies the problem, the problem nonetheless remains non-convex. Hence, the integrality constraint of the set of RBs, $\mathcal{N}$, is relaxed so that the problem is transformed into,

$$\max_{p_u,\, n_u} \sum_{u \in \mathcal{U}_4} B n_u \log\left(1+\frac{y_u p_u}{n_u}\right) \tag{6}$$

$$\text{s.t. } D_s - B n_u \log\left(1+\frac{y_u p_u}{n_u}\right) \leq 0, \forall u \in \mathcal{U}_s, \forall s \in \mathcal{S}, \tag{7}$$

$$\sum_{u \in \mathcal{U}} p_u - P = 0, p_u \geq 0, \tag{8}$$

$$\sum_{u \in \mathcal{U}} n_u - N = 0, \tag{9}$$

$$0 \leq N_u \leq N, \forall u \in \mathcal{U}. \tag{10}$$

The objective function in Eq. (6) maximizes the aggregate throughput of the BE slice and is subject to the constraints in Eqs. (7) to (10). Note that the objective function and constraints are no longer coupled to the RBs because of the uniform gain assumption made earlier. p_u, and n_u are BS transmission power and number of resource blocks for user u, correspondingly; P is the maximum BS transmission power. Equation (2) is now transformed to (7). Equation (3) is turned into (8). Finally, Eqs. (4) and (5) have been transformed into (9) because the integrality constraint of the RBs is relaxed as specified by Eq. (10). Hence, the transformed problem is no longer deciding bandwidth allocations on a *per-RB basis* for each user but is rather allocating groups of RBs altogether to each user. In other words, the problem decides *how many* as opposed to *which* RBs to allocate to user u. Such a problem is solvable by off-the-shelf solvers such as CVX or CPLEX.

The average gain assumption for this problem simplification does not fully reflect the true channel conditions experienced by users within a wireless network. Channel conditions are frequency dependent and thus, a BS must consider the channel quality over all the RBs in the transmission band. Undoubtedly, this greatly increases the power and resource allocation complexity. This assumption is only made to assist us in reducing the computational burden and to relax the problem into a convex one. However, when we solve the original MINLP problem in Section III-D, we take the frequency dependent conditions into consideration and dispense with the simplistic assumption discussed earlier.

We now prove the convexity of the problem defined in Eqs. (6) to (10) as follows: assume the general form of (6) is *u (a, b) = a log (1 + b/a)*. By proving that a function of this form is *concave* given that a and b are positive, the inversion of the function is thus convex. The Hessian matrix of *u (a, b)*, H, is calculated as follows:

$$H = \begin{bmatrix} \frac{\partial^2 \mathrm{u}\ (\mathrm{a},\ \mathrm{b})}{\partial \mathrm{a}^2} & \frac{\partial^2 \mathrm{u}\ (\mathrm{a},\ \mathrm{b})}{\partial \mathrm{a} \partial \mathrm{b}} \\ \frac{\partial^2 \mathrm{u}\ (\mathrm{a},\ \mathrm{b})}{\partial \mathrm{b} \partial \mathrm{a}} & \frac{\partial^2 \mathrm{u}\ (\mathrm{a},\ \mathrm{b})}{\partial \mathrm{b}^2} \end{bmatrix}$$

TABLE II
POWER AND RESOURCE ALLOCATIONS PER SLICE

RAN Slice	Bandwidth (%)	Power (Watts)	Throughput
Virtual reality	19.7687	8.7908	6.8 Gbps
Video streaming	1.4946	0.6925	500 Mbps
Internet of Things	0.0015	.0007	520 Kbps
Best effort	78.7352	30.5161	10 Gbps
Total	**100**	**40**	**17.3 Gbps**

$$= \begin{bmatrix} \frac{-\mathrm{b}^2}{\mathrm{a}^3\left(1+\frac{\mathrm{b}}{\mathrm{a}}\right)^2 \ln 2} & \frac{\mathrm{b}}{\mathrm{a}^2\left(1+\frac{\mathrm{b}}{\mathrm{a}}\right)^2 \ln 2} \\ \frac{\mathrm{b}}{\mathrm{a}^2\left(1+\frac{\mathrm{b}}{\mathrm{a}}\right)^2 \ln 2} & \frac{-1}{\mathrm{a}\left(1+\frac{\mathrm{b}}{\mathrm{a}}\right)^2 \ln 2} \end{bmatrix}.$$

Since $\det(H) = 0$ and $\frac{\partial^2 \mathrm{u}\ (\mathrm{a},\ \mathrm{b})}{\partial \mathrm{a}^2}, \frac{\partial^2 \mathrm{u}\ (\mathrm{a},\ \mathrm{b})}{\partial \mathrm{b}^2} < 0$, H is negative semi-definite. Therefore, *u (a, b)* is a concave function, and its inversion is a convex function as written in (6). Equation (7) is also convex due to the convexity of (6) while (8) and (9) are linear functions. Equation (10) serves to relax the integer domain of a decision variable. Thus, the problem defined in (6) to (10) is a convex optimization problem.

III. NUMERICAL RESULTS AND DISCUSSIONS

In this section, three different cases are considered and discussed. The baseline case is characterized by a fixed number of users within each slice; such a scenario allows one to observe the initial power and bandwidth allocations under a fixed network load. The second case involves varying the load of the network by admitting users into a slice. The final case involves varying the total BS transmission power. This is especially useful in observing the resulting performance of heterogeneous networks (HetNets) and 5G small cells (for cell densification). After presenting the results, we briefly discuss and compare the network performance when solving the MINLP and convex problems.

A. Fixed Number of Users in Slices

In this first scenario where the maximum BS transmission power and number of users within the network are fixed, the total number of users in the network was set to 80 with equal distribution among the slices. The total network bandwidth consists of a 100 RBs; each is 180 kHz wide. The maximum BS transmission power is set to 40 W and the data rates for the VR, VS, and IoT slices are set to 340 Mbps [29], 25 Mbps [30], and 26 kbps [31], respectively. The BE slice is offered an arbitrary data rate of 5 Mbps to protect its users against intra-slice resource starvation. To fully load the BS, all users within the network are assumed to be scheduled simultaneously.

Table II lists the power and resource allocations along with the resultant aggregate throughput of each slice. All the users' requirements were met for each slice. Visibly, most of the power and bandwidth were allocated to the BE slice (approximately 79 percent). The total power consumed by the IoT slice (0.7 mW) is almost negligible and this agrees with the general minimal power requirements of the IoT sensor applications and devices. The total network throughput was approximately 17.3 Gbps and over half of that total throughput resulted from the BE slice alone which also has the highest throughput of all the slices. The

joint power and bandwidth allocation took approximately 7.26 seconds to complete for this scale of users. Prior to discussing the subsequent scenarios, a few remarks about the chosen schedulers are in order.

1) Impact of Schedulers on RAN Performance: To maximize the aggregate throughput of a particular slice, the optimization problem applies the Blind-Equal Throughput (BET) and Proportional Fair (PF) schedulers for the 5G and non-5G slices, respectively. Evidently, one of the benefits of network slicing exploited herein is the ability to employ various schedulers for different slices as needed. The BET scheduler provisions resources such that all the 5G slice users are provided the minimum equal throughput guaranteed by their respective slices.

The PF scheduler, which is applied to the BE slice, attempts to maintain a balance between fairness as well as slice throughput maximization. It allocates resources to facilitate a minimum data rate to all the BE slice users but maximizes the slice throughput by provisioning the remnant resources to the slice users with the best channel conditions. Thus, the allocations are initially done with an inverse relation to the channel gains of all the users until their minimum requirements are met. Upon the satisfaction of such requirements, the allocation is then done with a proportional relation to the channel gains of the users with the best channel conditions.

The two schedulers that are employed in this problem are specified implicitly within the formulations themselves in Section II. Note that in constraint (7), a minimum data rate, $D_{s,}$ for each slice s is provisioned but the throughput is maximized in (6) for a single slice (denoted by the user subset $\mathcal{U}4$). This forces users in all the other slices to be provisioned their minimum guaranteed requirements via the BET scheduler. The BE slice, however, allows provisioning for *at least* the minimum guaranteed requirement but the maximization in (6) allows for the BE slice users with the best channel conditions to maximize their throughputs and in turn, the slice throughput as well. It should be stated that while in realistic scenarios, there are hardware and software limitations on a user's maximum throughput, as well as throughput limitations imposed by MNOs depending on a user's subscription plan, such details are not considered in this work.

B. Increasing Network Load by User Admission Into Slices

The network performance was assessed under higher loads by increasing the number of users in the VR slice while keeping the number of users in all the other slices fixed. Under these conditions, the throughputs of the VR and BE slices, resource allocations, and the solver run-time were observed. In Fig. 3, the aggregate throughput of the BE slice was observed to deteriorate under heavier VR slice loads by nearly a fourth. The resources that were initially allocated to the BE slice were diverted to the VR slice to support its increasing load. As the 5G slices are of a higher priority, it is expected that the BS will divert resources from the low priority BE slice.

Note that although not shown in Fig. 3, all the other 5G slices from the first scenario are still burdening the network. The resources that the BE slice once enjoyed with only 20 users in the VR slice dropped by 15 percent with the onset of more

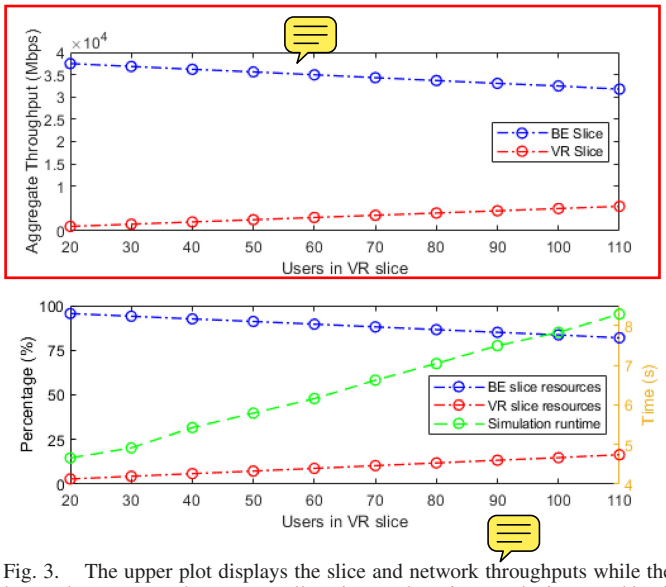

Fig. 3. The upper plot displays the slice and network throughputs while the lower plot expresses the resource allocations and runtimes under increased load.

VR slice users. It should be noted that while it does not alter the problem, the VR slice bit rate in this and the next case, was set to 50 Mbps [29]. The solver took increasingly a longer time to arrive at its optimal decisions as the total number of users increased; the run-time nearly doubled from 4.6 seconds to 8.4 seconds. Bearing in mind that this is the execution time of the simplified (and relaxed) problem, this only further serves to highlight the complexity of the original MINLP problem formulated in Section II-A.

It is critical to mention at this juncture that if needed, the BS will only divert resources from the BE slice users that are enjoying *better* than their minimal slice requirements. This will be the case anytime the overall network load increases (by admission of users in any slice). However, the users of the BE slice that are already at the minimum threshold will not be affected by the additional load because siphoning resources from them would result in a violation of their QoS requirements. If the load of any of the 5G slices increases, this will result in increased inter-slice contention. Increasing the load within the BE slice however exacerbates intra-slice contention.

C. Varying Maximum Base Station Transmission Power

To assess the performance of BSs with lower power budgets (as is required of small cells within HetNets and 5G cell densification), the total transmission power budget was varied while the number of users was kept fixed. Both the total network and BE slice throughputs were obtained under said conditions. At lower BS transmission powers, the aggregate BE slice throughput suffered since power was scarce and the high priority slice users were given precedence in the allocation. However, as the BS transmission power budget increased, the BE slice throughput increased since additional power was available for provisioning for the BE slice users after satisfying the 5G slice users. Fig. 4 demonstrates this scenario.

While meeting the guaranteed data rates for each of the users in each of the slices, the aggregate throughputs of the

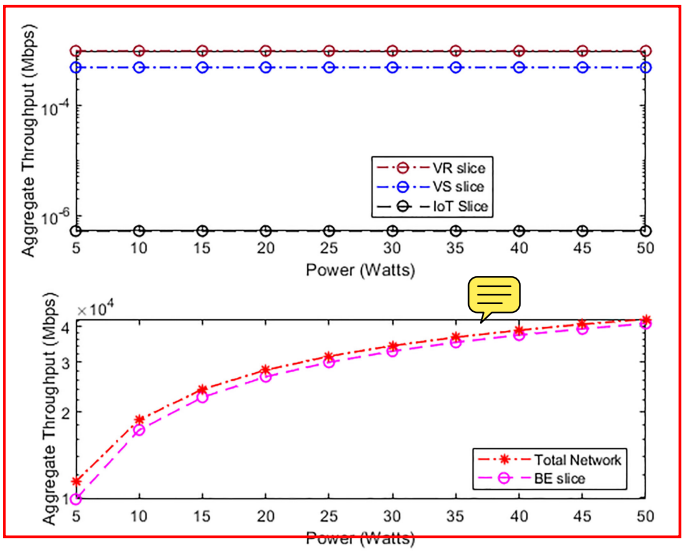

Fig. 4. Network performance with varying transmission power corresponding to different types of BSs.

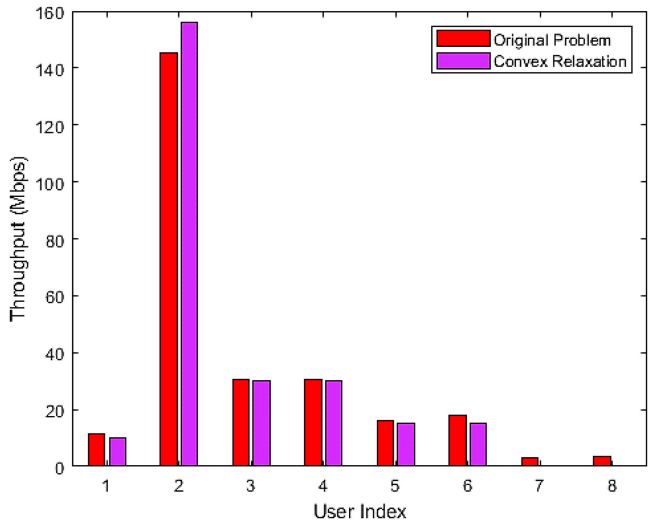

Fig. 5. Network performance comparison between the original and convex approximation problems.

VR, VS, and IoT slices are 1000 Mbps, 500 Mbps, and 520 kbps, respectively. It is useful to note that in Fig. 4, the gap between the aggregate throughputs of the BE slice and the total network reduces as the BS transmission power budget increases. At the minimum BS power, the BE slice throughput lags the total network throughput by approximately 1500 Mbps. At the maximum BS transmission power however, the throughput gap is reduced. Thus, the BE slice portion of the throughput increases and makes up a bigger portion of the total network load as the BS increases in transmission power. Resultantly, the ratio between the BE slice and total network throughputs should asymptotically approach unity with higher availability of both BS transmission power and bandwidth.

It is understood that below a certain amount of maximum BS transmission power or wireless bandwidth, the solution is infeasible. Likewise, the solution becomes infeasible above a certain network load. This is because it is simply not possible to satisfy the requirements of each user in each slice with a finite amount of bandwidth and power, even after siphoning resources from lower priority slices. This is often a limitation expected to be particularly faced by low-power BSs in HetNets. The issue is further exacerbated by cell edge users that force the BS to transmit to them with much higher power and bandwidth, thus reducing the amount of power and/or bandwidth available for the rest of the users in the network. This reduces such users' performances along with the overall slice and network throughputs.

Through the simulation results of the problem scenarios presented, it was amply demonstrated that despite the severity of the network load or scarcity of BS transmission power strength, the network slices were able to meet the guaranteed service requirements of both the high and low priority traffic. The wireless network exhibited that it properly handled resource contention among users. Overall, the RAN slicing scheme maintained a reasonable balance between fairness for all its users and throughput maximization. This is undoubtedly one of the aims of MNOs to be achieved in real-life wireless networks that are typically associated with various service level agreements (SLAs).

D. Network Performance Comparison

In this section, we compare the network performances between the original MINLP and the convex approximation problems formulated in Sections II-A and II-B, respectively (particularly for the first scenario discussed, i.e., fixed network load). We solve the original problem for a small-scale network and observe that the convex relaxation of the MINLP problem still affords us a reasonably close approximation of the true network behavior. The small-scale network consists of eight users that are equally distributed among the slices. The first two users are in the BE slice, while the remaining users are equally distributed among the rest of the 5G slices. For simplicity, the minimum required data rates of the three 5G slices (in the order presented in Section III-A) are 30 Mbps, 15 Mbps, and 20 kbps, respectively. The BE slice requires at least a 10 Mbps link speed.

In Fig. 5, we observe that the convex approximation is well near the optimal solution. The last two users' throughputs are minimal (for the convex approximation) as they are associated with the IoT slice; hence, their throughputs are not visible at the scale which Fig. 5 is utilizing. The slight performance gap experienced by each of the users can be easily explained by the fact that in the convex approximation problem, the wireless bandwidth is relaxed from an integral domain to that which is continuous. This affords the BS a higher allocation flexibility. Furthermore, the transmission power limit for each individual RB per user is eliminated for problem tractability in the convex problem.

In Fig. 6, we examine the power and bandwidth allocations. Once more, because the power limit per RB is removed, we see that in the convex approximation, much more power is allocated

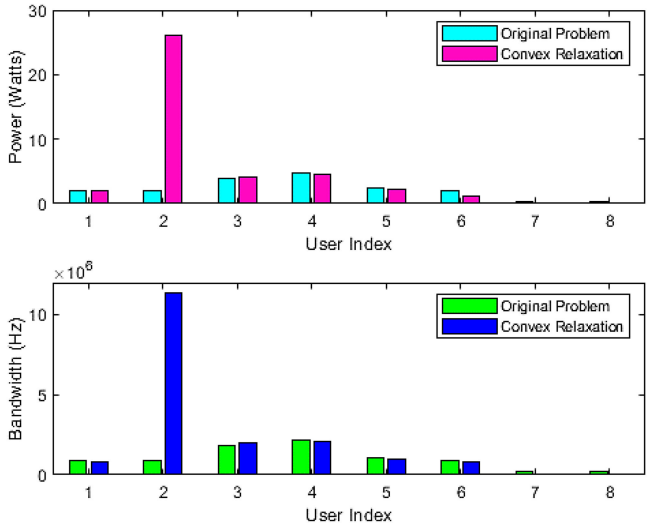

Fig. 6. Power and bandwidth allocation comparison between the original and convex approximation problems.

to the BE users as opposed to in the original problem. Because of the power limitation imposed on each RB for each user in the original problem, the BS cannot allocate extra power for any RB even when there is unallocated power available; hence we see a noticeable gap for the second user's power allocation. If the power cap in the original problem were to be bypassed for the BE slice users of the network, the second user's performance would have inched closer to the approximated one in Fig. 5.

Through both Figs. 5 and 6, we can justify our assumption of uniform channel gain for the users in the convex approximation problem. The power and resource allocations in both cases are reasonably close to each other, i.e., the accuracy of the relaxed problem is satisfactory enough to present us with a quickly attainable solution with a significantly lower computational load. Of course, as the system bandwidth increases, this accuracy degrades because the algorithm is operating at a lower resolution by ignoring too many frequency dependent conditions (per each RB). The lower the network bandwidth is, the more accurate the convex approximation problem becomes because it averages the channel conditions over a lesser amount of RBs (i.e., there is less disregarding of frequency dependent conditions).

IV. Conclusion

In this work, we have presented a priority-based throughput maximization problem of a non-5G slice while satisfying the 5G slices' constraints. The impact of three important scenarios on the network and the rationale behind examining them were discussed. Moreover, the implications of the selected schedulers upon intra-slice and inter-slice contentions as well as the throughputs were addressed in detail. The problem was originally formulated as a MINLP problem which was simplified and subsequently relaxed into a convex optimization problem. The results demonstrated that despite the varying SLAs (slice priorities), the RAN was able to simultaneously uphold QoS satisfaction, fairness, and throughput maximization. It was also shown that the convex approximation problem provides a near optimal solution despite the simplifying assumption of uniform channel gain for the network users. All in all, upon striking a reasonable balance between fairness and throughput maximization, the QoS requirements of all the slice users were comfortably met. For future research, employing the concept of 5G new radio numerologies will be explored for RAN slicing.

Appendix

We prove that the throughput maximization problem of Section II-A is NP-hard. As per the computational complexity theory [32]–[35], if an optimization problem is proven to be NP-hard, the goal would be to determine optimal solutions which are either approximate or local solutions instead of exact or global ones. The conventional procedure to prove whether an optimization problem is indeed NP-hard is to examine its corresponding feasibility or decision problem. The feasibility problem is always easier than its corresponding optimization problem; the feasibility problem does not require an actual global minimizer or maximizer to be found. If such a feasibility problem is shown to be NP-hard, then the corresponding optimization problem is likewise NP-hard.

For the sake of generality, assume there exists an optimization problem A whose corresponding feasibility problem is B. To prove that A is NP-hard, it must be proven that B is NP-hard. B can be shown to be NP-hard as follows: 1) select an appropriate decision problem C, which is already established to be NP-complete, 2) transform any instance of C into an instance of B, and 3) prove that the instance of C is satisfied when the instance of B is feasible.

We denote the problem formulated in Section II-A as A and its feasibility problem as B. Note that B is presented in a more compact form than that of A by removing the binary variable constraint by stipulating (13) as the orthogonality constraint; it is indeed the same problem but presented in a reduced form. B then becomes,

$$\sum_{n\in\mathcal{N}} B\log\left(1+\frac{g_{n,u}p_{n,u}}{B\eta_{n,u}}\right)\geq D_s, u\in\mathcal{U}s,\ s\ \in\mathcal{S}, \tag{11}$$

$$0\leq p_{n,u}\leq P_n,\ u\in\mathcal{U}_s,\ n\in\mathcal{N},\ s\ \in\mathcal{S}, \tag{12}$$

$$p_{n,u}\ p_{n,v}=0, \forall u\neq v, u, v\in\mathcal{U}_s,\ n\in\mathcal{N},\ s\ \in\mathcal{S}. \tag{13}$$

C is chosen to be the 3D Matching Problem (3DMP) which is an NP-complete problem. If it is assumed that $\mathcal{J}$, $\mathcal{K}$, and $\mathcal{L}$ are three sets such that $|\mathcal{J}|=|\mathcal{K}|=|\mathcal{L}|=Z$ and $\mathcal{T}\in\mathcal{J}\times\mathcal{K}\times\mathcal{L}$, then the 3DMP asks the following question: is there any $\mathcal{M}\in\mathcal{T}$ such that for any pair of triples (j_1, k_1, l_1), $(j_2, k_2, l_2)\in\mathcal{M}$, $j_1\neq j_2$, $k_1\neq k_2$, and $l_1\neq l_2$ where $\mathcal{J}=\{j\,|\,(j,k,l)\in\mathcal{M}\}$, $\mathcal{K}=\{k\,|\,(j,k,l)\in\mathcal{M}\}$, and $\mathcal{L}=\{l\,|\,(j,k,l)\in\mathcal{M}\}$?

The subsequent step is to transform C into an instance of B. If this transformation into an instance of B proves to be feasible, then B is NP-hard. One instance of the 3DMP that can be transformed into B is the case when there is twice the amount of RBs as there are users. In this case, examine the following: $\mathcal{J}=\{1_j, 2_j, \ldots, U_j\}$, $\mathcal{K}=\{1_k, 2_k, \ldots, U_k\}$, $\mathcal{L}=\{1_l, 2_l, \ldots, U_l\}$,

and $\mathcal{T} = \{(a_j, b_k, c_l) \mid a_j \in \mathcal{J}, b_k \in \mathcal{K}, c_l \in \mathcal{L}\} \subseteq \mathcal{J} \times \mathcal{K} \times \mathcal{L}$. Assume that the user set, $\mathcal{U} = \mathcal{J}$ and that $\mathcal{N} = \mathcal{K} \cup \mathcal{L}$. We then define the following sets: $\mathcal{H}_1 = \{(a_j, b_k) \mid (a_j, b_k, c_l) \in \mathcal{T}\}$ and $\mathcal{H}_2 = \{(a_j, c_l) \mid (a_j, b_k, c_l) \in \mathcal{T}\}$.

Assume that for (11), the RB bandwidth is 1, and that for all s, $D_s = 3$. Finally, consider the following:

$$Pn = \begin{cases} 3, \text{ if } n \in \mathcal{K}, \\ 2, \text{ if } n \in \mathcal{L}, \end{cases} \tag{14}$$

$$\mathcal{N}_{n,u} = \begin{cases} 1, \text{ if } (\mathrm{n},\ \mathrm{u}) \in \mathcal{H}1, \\ 2, \text{ if } (\mathrm{n},\ \mathrm{u}) \in \mathcal{H}2, \\ 3, \text{ if } (\mathrm{n},\ \mathrm{u}) \notin \mathcal{H}1 \cup \mathcal{H}2, \end{cases} \tag{15}$$

$$g_{n_1,u} = g_{n_2,u} = \begin{cases} 1, \text{ if } (\mathrm{u},\ \mathrm{n1},\ \mathrm{n2}) \in \mathcal{T}, \\ 0.25, \text{ if } (\mathrm{u},\ \mathrm{n1},\ \mathrm{n2}) \notin \mathcal{T}. \end{cases} \tag{16}$$

Then, (11)–(13) becomes as follows due to (14)–(16) and the bandwidth and data rate assumptions made earlier:

$$\sum_{n \in \mathcal{N}} \log\left(1 + \frac{g_{n,u}\, p_{n,u}}{\eta_{n,u}}\right) \geq 3,\ u \in \mathcal{U}, \tag{17}$$

$$0 \leq p_{n,u} \leq 3,\ u \in \mathcal{U},\ n \in \mathcal{K}, \tag{18}$$

$$0 \leq p_{n,u} \leq 2,\ u \in \mathcal{U},\ n \in \mathcal{L}, \tag{19}$$

$$p_{n,u} p_{n,v} = 0, \forall u \neq v, u, v \in \mathcal{U},\ n \in \mathcal{N}. \tag{20}$$

The final step is to evaluate the transformed instance C into B in (17)–(20) with respect to (14)–(16) and show if it is feasible. If it indeed is feasible, then B is proven to be NP-hard and thus, A too is NP-hard. If $\{(a_j, b_k, c_l)\}$ satisfies the 3DMP question stated, then a suitable power allocation for the instance of B is,

$$\mathrm{p_{n,a_j}} \begin{cases} 3, \text{if } n = bk, \\ 2, \text{ if } n = cl, \\ 0, \text{ if } n \neq \text{bk or cl}. \end{cases} \tag{21}$$

Thus, for any user $a_j = 1_j, 2_j, \ldots, U_j$, we have for two RBs assigned to that user:

$$\log\left(1 + \frac{1(3)}{1}\right) + \log\left(1 + \frac{1(2)}{1}\right) = 3. \tag{22}$$

Equation (22) proves that indeed the instance of problem B transformed from C is feasible; thus, B is proven to be NP-hard and consequently, A is also NP-hard.

REFERENCES

[1] X. Sun and N. Ansari, "EdgeIoT: Mobile edge computing for the Internet of Things," *IEEE Commun. Mag.*, Special Issue on Internet of Things, vol. 54, no. 12, pp. 22–29, Dec. 2016.

[2] B. Chatras, U. S. Tsang Kwong, and N. Bihannic, "NFV enabling network slicing for 5G," in *Proc. 20th Conf. Innovations Clouds, Internet Netw.*, Paris, 2017, pp. 219–225, doi: 10.1109/ICIN.2017.7899415.

[3] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5G Network slicing for vehicle-to-everything services," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 38–45, Dec. 2017.

[4] H. Ullah, N. Gopalakrishnan Nair, A. Moore, C. Nugent, P. Muschamp, and M. Cuevas, "5G Communication: An overview of vehicle-to-everything, drones, and healthcare use-cases," *IEEE Access*, vol. 7, pp. 37251–37268, 2019, doi: 10.1109/ACCESS.2019.2905347.

[5] M. Afaq, J. Iqbal, T. Ahmed, I. Islam, M. Khan, and M. S. Khan, "Towards 5G network slicing for vehicular ad-hoc networks: An end-to-end approach," *Comput. Commun.*, vol. 149, pp. 252–258, Jan. 2020, doi: 10.1016/j.comcom.2019.10.018.

[6] S. A. A. Shah, E. Ahmed, M. Imran, and S. Zeadally, "5G for Vehicular communications," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 111–117, Jan. 2018.

[7] D. Sattar and A. Matrawy, "Optimal slice allocation in 5G core networks," *IEEE Netw. Lett.*, vol. 1, no. 2, pp. 48–51, Jun. 2019.

[8] T. Han and N. Ansari, "A traffic load balancing framework for Software-defined radio access networks powered by hybrid energy sources," *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 1038–1051, Apr. 2016.

[9] Q. Liu, T. Han, N. Ansari, and G. Wu, "On designing energy-efficient heterogeneous cloud radio access networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 721–734, Sep. 2018.

[10] B. Khodapanah, A. Awada, I. Viering, A. N. Barreto, M. Simsek, and G. Fettweis, "Slice management in radio access network via iterative adaptation," in *Proc. ICC 2019 - 2019 IEEE Int. Conf. Commun.*, Shanghai, China, 2019, pp. 1–7, doi: 10.1109/ICC.2019.8761376.

[11] I. da Silva *et al.*, "Impact of network slicing on 5G radio access networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Athens, 2016, pp. 153–157, doi: 10.1109/EuCNC.2016.7561023.

[12] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G radio access network slicing: Radio interface protocol features and configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, May. 2018.

[13] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On the automation of RAN slicing provisioning: Solution framework and applicability examples," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 167, Jun. 2019.

[14] G. Wang, G. Feng, S. Qin, R. Wen, and S. Sun, "Optimizing network slice dimensioning via resource pricing," *IEEE Access*, vol. 7, pp. 30331–30343, Mar. 2019.

[15] G. Wang, G. Feng, W. Tan, S. Qin, R. Wen, and S. Sun, "Resource allocation for network slices in 5G with network resource pricing," in *Proc. GLOBECOM 2017 - 2017 IEEE Glob. Commun. Conf.*, Singapore, 2017, pp. 1–6, doi: 10.1109/GLOCOM.2017.8254074.

[16] M. O. Ojijo and O. E. Falowo, "A survey on slice admission control strategies and optimization schemes in 5G network," *IEEE Access*, vol. 8, pp. 14977–14990, 2020, doi: 10.1109/ACCESS.2020.2967626.

[17] T. Dang and M. Peng, "Delay-aware radio resource allocation optimization for network slicing in fog radio access networks," in *Proc. IEEE Int. Conf. Commun. Workshops*, KS City, MO, 2018, pp. 1–6, doi: 10.1109/ICCW.2018.8403717.

[18] Q. Liu, T. Han, and N. Ansari, "Energy-efficient on-demand resource provisioning in cloud radio access networks," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 1142–1151, Dec. 2019.

[19] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.

[20] Q. Liu, T. Han, and N. Ansari, "Learning-assisted secure end-to-end network slicing for cyber-physical systems," *IEEE Netw.*, vol. 34, no. 3, pp. 37–43, May/Jun. 2020.

[21] F. Tang, Y. Zhou, and N. Kato, "Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G hetnet," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2773–2782, Dec. 2020.

[22] F. Tang, Z. Md. Fadlullah, B. Mao, and N. Kato, "An intelligent traffic load prediction based adaptive channel assignment algorithm in SDN-IoT: A deep learning approach," in *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5141–5154, Dec. 2018.

[23] L. Zanzi and V. Sciancalepore, "On guaranteeing end-to-end network slice latency constraints in 5G networks," in *Proc. 15th Int. Symp. Wireless Commun. Syst.*, Lisbon, 2018, pp. 1–6, doi: 10.1109/ISWCS.2018.8491249.

[24] I. Afolabi, M. Bagaa, T. Taleb, and H. Flinck, "End-to-end network slicing enabled through network function virtualization," in *Proc. IEEE Conf. Standards for Commun. Netw.*, Helsinki, 2017, pp. 30–35, doi: 10.1109/CSCN.2017.8088594.

[25] J. Liu, N. Kato, O. Akashi, and A. Takahara, "Reliability evaluation for NFV deployment of future mobile broadband networks," *IEEE Wireless Commun. Mag.*, vol. 23, no. 3, pp. 90–96, Jan. 2016.

[26] Q. Duan, N. Ansari, and M. Toy, "Software-Defined network virtualization – An architectural framework for integrating SDN and NFV for service provisioning in future networks," *IEEE Netw.*, vol. 30, no. 5, pp. 10–16, Sep./Oct. 2016.

[27] A. D. Hossain and A. R. Hossain, "A distributed control framework for TDM-PON based 5G mobile fronthaul," *IEEE Access*, vol. 7, pp. 162102–162114, 2019, doi: 10.1109/ACCESS.2019.2951581.

[28] S. M. Almalfouh and G. L. Stuber, "Uplink resource allocation in cognitive radio networks with imperfect spectrum sensing," in *Proc. IEEE 72nd Veh. Technol. Conf. - Fall*, Ottawa, ON, 2010, pp. 1–6, doi: 10.1109/VETECF.2010.5594305.
[29] K. Qin and M. Zarri, "Network slicing use case requirements," GSMA future networks programme. London, England, p. 22, Jul. 2018. [Online]. Available: https://www.gsma.com/futurenetworks/wp-content/uploads/2018/07/Network-Slicing-Use-Case-Requirements-fixed.pdf
[30] W. Pan and G. Cheng, "QoE assessment of encrypted YouTube adaptive streaming for energy saving in smart cities," *IEEE Access*, vol. 6, pp. 25142–25156, 2018, doi: 10.1109/ACCESS.2018.2811416.
[31] R. Lobras, "*What is the Difference in Data Throughput Between LTE-M/NB-IoT and 3G Or 4G?"*. Accessed: Apr./Mar., 2020, [Online]. Available: https://www.gsma.com/iot/resources/what-is-the-difference-indata-throughput-between-lte-m-nb-iot-and-3g-or-4g/
[32] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. SF, USA: W. H. Freeman, 1979.
[33] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Englewood Cliffs, NJ: Prentice-Hall, Inc. 1998.
[34] C. H. Papadimitriou, *Computational Complexity*. Mass., USA: Reading, MA: Addison-Wesley, 1994.
[35] V. V. Vazirani, *Approximation Algorithms*. Berlin, Germany: Springer-Verlag, 2001.

Abdullah R. Hossain (Student Member, IEEE) received the B.E. and the M.S. degrees in electrical engineering from The City College of New York, City University of New York, New York City, NY, USA, in 2017 and 2019, respectively. He is currently working toward the Ph.D. degree with the New Jersey Institute of Technology, Newark, NJ, USA. He has authored or coauthored seven publications and three book chapters. His research interests include free space and fiber optics, optimization of optical networking, wireless communications, UAV networks, and accreditation and assessment.

Nirwan Ansari (Fellow, IEEE) received the B.S.E.E. (summa cum laude with a perfect GPA) from the New Jersey Institute of Technology (NJIT), NJ, USA, the M.S.E.E. from the University of Michigan, Ann Arbour, MI, USA, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA. He is currently a Distinguished Professor of electrical and computer engineering with NJIT. He is also a Fellow of National Academy of Inventors. He has authored *Green Mobile Networks: A Networking Perspective* (Wiley-IEEE, 2017) with T. Han, and coauthored two other books. He has also authored or coauthored more than 600 technical publications. His current research interests include green communications and networking, cloud computing, drone-assisted networking, and various aspects of broadband networks. He has guest-edited a number of special issues covering various emerging topics in communications and networking. He was on the Editorial or Advisory Board of more than ten journals, including as an Associate Editor-in-Chief of the *IEEE Wireless Communications Magazine*. He was elected to the IEEE Communications Society (ComSoc) Board of Governors as a member-at-large, has chaired some ComSoc technical and steering committees, is currently the Director of ComSoc Educational Services Board, is in many committees such as the IEEE Fellow Committee, and is actively organizing numerous IEEE International Conferences/Symposia/Workshops. He is frequently invited to deliver keynote addresses, distinguished lectures, tutorials, and invited talks. Some of his recognitions include various excellence in teaching awards, a few best paper awards, the NCE Excellence in Research Award, various ComSoc TC Technical Recognition awards, the NJ Inventors Hall of Fame Inventor of the Year Award, the Thomas Alva Edison Patent Award, the Purdue University Outstanding Electrical and Computer Engineering Award, NCE 100 Medal, NJIT Excellence in Research Prize and Medal, and designation as a COMSOC Distinguished Lecturer. He was granted more than 40 U.S. patents.