DYNAMIC RESOURCE ALLOCATION IN VIRTUALIZED NETWORKS FOR NETWORK SLICING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

CEREN CANPOLAT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

FEBRUARY 2020

Approval of the thesis:

DYNAMIC RESOURCE ALLOCATION IN VIRTUALIZED NETWORKS FOR NETWORK SLICING

submitted by **CEREN CANPOLAT** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İlkay Ulusoy
Head of Department, **Electrical and Electronics Engineering** _____

Prof. Dr. Şenan Ece Güran Schmidt
Supervisor, **Electrical and Electronics Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Gözde Bozdağı Akar
Electrical and Electronics Engineering, METU _____

Prof. Dr. Şenan Ece Güran Schmidt
Electrical and Electronics Engineering, METU _____

Prof. Dr. İbrahim Körpeoğlu
Computer Engineering, Bilkent University _____

Assist. Prof. Dr. Gökhan Güvensen
Electrical and Electronics Engineering, METU _____

Assist. Prof. Dr. Pelin Angın
Computer Engineering, METU _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ceren Canpolat

Signature :

ABSTRACT

DYNAMIC RESOURCE ALLOCATION IN VIRTUALIZED NETWORKS FOR NETWORK SLICING

Canpolat, Ceren
M.S., Department of Electrical and Electronics Engineering
Supervisor: Prof. Dr. Şenan Ece Güran Schmidt

February 2020, 61 pages

The developments of 5G wireless technology, enables serving to various vertical industries through sharing a common infrastructure. To this end, multi-tenancy support of these diverse industries are realized on virtualized networks with the help of network slicing. The introduction of sharing brings many challenges such as QoS satisfaction, fairness and performance isolation among slices. The diversity of these slices mainly lies in their data rate requests and user populations. The slices with high data rate traffic requests are often given large number of resources. That resource allocation approach leaves the slice requests for low data rate slices out of the network. Overcoming all of these challenges with an efficient resource allocation approach is the main concern of this work.

In this thesis, we propose DUCA (Dynamic User Count Aware) network slicing, a novel resource allocation scheme for virtualized radio access networks (RAN). DUCA's resource allocation objective is to serve large number of user requests with high resource utilization results in the presence of diverse slice requirements. To this end, DUCA is formulated with an additional user count parameter so that not only

requested amount of data rates but also user populations of slices can effect resource allocation. DUCA is compared with other resource allocation schemes Dynamic [1] and NVS [2] under different network configurations. Simulation results show that the proposed DUCA outperforms compared resource allocation methods.

Keywords: Network slicing, fifth-generation (5G), resource reservation, resource allocation, radio access networks

ÖZ

AĞ DİLİMLEME İÇİN SANALLAŞTIRILMIŞ AĞLARDA DİNAMİK KAYNAK ATAMA

Canpolat, Ceren
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Şenan Ece Güran Schmidt

Şubat 2020 , 61 sayfa

5G kablosuz ağlar teknolojisindeki gelişmeler, ortak bir altyapıyı paylaşarak çeşitli dikey endüstrilere hizmet vermeye olanak sağlamaktadır. Bu amaçla, sanallaştırılmış ağlarda ağ dilimleme yöntemi yardımıyla farklı endüstrilere çoklu kiracılık desteği verilmektedir. Paylaşımın dahil olmasıyla birlikte, hizmet kalitesi memnuniyeti, dilimler arasında adalet ve performans izolasyonu gibi birçok zorluk oluşmaktadır. Dilimlerin çeşitliliği temel olarak veri hızı taleplerinde ve kullanıcı popülasyonlarındaki farklılıklarda göze çarpmaktadır. Genellikle, yüksek veri hızı isteğine sahip dilimlere çok sayıda kaynak ataması yapılır. Bu kaynak atama yaklaşımı, düşük veri hızı talep eden dilimlerin ağ dışında kalmasıyla sonuçlanır. Bu çalışmanın ana konusu, tüm bu zorlukların etkin bir kaynak atama yaklaşımıyla aşılmasıdır.

Bu tezde, sanallaştırılmış radyo erişim ağları (RAN) için yeni bir kaynak atama yöntemi olan DUCA (Dinamik Kullanıcı Sayısı Farkında) ağ dilimleme yöntemi önerilmektedir. DUCA'nın kaynak atama hedefi, farklı dilim gereksinimlerinin varlığında kaynakların etkili kullanımını sağlayıp çok sayıda kullanıcı talebine hizmet vermektir. Bu amaçla, DUCA ek bir kullanıcı sayısı parametresi ile formüle edilmiştir. Böy-

lece sadece istenen veri hızı miktarı değil, aynı zamanda kullanıcı popülasyonları da kaynak atama kararını etkilemektedir. DUCA, farklı ağ konfigürasyonlarında diğer kaynak atama yöntemleri olan Dynamic [1] ve NVS [2] ile karşılaştırılmıştır. Simülasyon sonuçları, önerilen DUCA'nın diğer kaynak atama yöntemlerinden daha iyi performans elde ettiğini göstermektedir.

Anahtar Kelimeler: Ağ dilimleme, beşinci jenerasyon (5G), kaynak ayırma, kaynak atama, radyo erişim ağları

To my family

ACKNOWLEDGMENTS

I would like to express my greatest gratitude to my supervisor Prof. Dr. Ece Güran Schmidt. Without her outstanding guidance and support, this work would not have been succeeded.

I would also like to thank ASELSAN Inc. for their support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ALGORITHMS

ALGORITHMS

LIST OF ABBREVIATIONS

ABBREVIATIONS	
RAN	Radio Access Network
InP	Infrastructure Provider
SP	Service Provider
SLA	Service Level Agreement
QoS	Quality of Service
3GPP	3G Project Partnership
GWCN	Gateway Core Network
MOCN	Multi-Operator Core Network
MNO	Mobile Network Operator
RB	Resource Block
BS	Base-Station
QoE	Quality of Experience
UeB	User Equipment Broadband
ULL	Ultra Low Latency
MIoT	Massive Internet of Things
HdTV	High Definition Television
MVNO	Mobile Virtual Network Operator
CAPEX	Capital Expenditure
OPEX	Operational Expenditure
SDN	Software Defined Networking
NV	Network Virtualization
5G	Fifth Generation
OFDMA	Orthogonal Frequency Division Multiple Access

CHAPTER 1

INTRODUCTION

The cellular network technology is evolving to provide services to a wide range of applications. With the enormous amount of increase in data traffic and existence of challenging diverse requirements, mobile network operators (MNO) come across with drastic drops in their revenues. Throughout time, many suggestions have been realized to prevent this decline, and the most effective solutions included network sharing. Network sharing brought a new perspective to the current physical infrastructure capabilities. Instead of building new sites, antennas and base-stations (BS); the existing networks are shared such that all tenants became profitable [4]. With minimum physical effort, programmable infrastructures are created with the help of Network Virtualization (NV) [5], [6]. Instead of adding middle-boxes, networks have been virtualized and decoupled the functions run on them from the physical network. Programmable controllers, that are basically developed as an software program to be uploaded on a BS, are added to the networks with the Software Defined Networking (SDN) architecture [7]. With these controllers all information of the network's current state is being traced and allocations of the underlying virtualized networks between active tenants are done accordingly.

Up and coming fifth generation (5G) wireless technology brings many benefits alongside with some controversial issues. 5G is expected to provide major digital conversion that will enable industries request distinct requirements.Through shared virtual networking approach, that requirements will form a *slice* under a common goal. Network slicing which is the key player in the realization of 5G technology will enable network operators to cope with *service for a specific need* requests [8]. Fulfillment of these requests with new admission control strategies are discussed in many works.

At [9], after an heuristic admission control algorithm, the idea to forecast the amount of resources a slice would need and adjusting the values online by tracing its traffic is considered. In another work [10] , dynamic resource sharing method slice-share constrained proportionally fair (SCPF), is compared with static slicing method, which allocates resources according to service level agreement (SLA), and general processor sharing method which allocates resources only to active users.In [11], users are prioritized as well as slices to have an efficient resource allocation and admission control. In [12], licensed and unlicensed bands are being shared by mobile network operators to maximize individual slices' profits. Some of the other recent works on network slicing use the machine learning techniques to achieve most efficient resource allocation in 5G networks as in [13], [14],[15],[16] and [17]. Even though machine learning has the ability to provide accurate resource management by analyzing a large amount of data, it has the disadvantage of having a learning period that could violate delay requirements of slices.

The main contribution of this thesis work is to provide a novel resource allocation scheme that we call Dynamic User Count Aware (DUCA) Network Slicing for multi-sliced radio access networks. Different than previous approaches, DUCA considers not only requested data rates but also the effect of number of user requests. The slices that are considered in the context of this thesis are enhanced mobile broadband (eMBB), high definition TV (HdTV), massive internet of things (MIoT) and ultra low latency (ULL) services. While rate-based resource allocation schemes such as Dynamic [1] and NVS [2] favors high data requested slices (HdTV and eMBB), DUCA provides a fair, profitable and sustainable allocation among all by giving MIoT and ULL services higher reservations by considering the excess in their user populations. This work also provides two methods for intra-slice resource allocation. These methods provide efficient resource allocations among user requests that belongs to same slice by considering the amount of required of resources and user channel gain differences.

The remainder of the thesis is organized as follows. Chapter 2 introduces the necessary background information, system model and important challenges on Radio Access Networks (RAN), along with the comparison of previous resource allocation approaches that uses same model. Chapter 3 states the proposed DUCA method

for inter-slice resource allocation approach and proposes two additional methods for intra-slice resource allocation. In Chapter 4 DUCA's performance is evaluated in a simulated network and simulation results are presented. Performance evaluations are done with five experiments that simulate different network configurations. The results show that DUCA outperforms previous approaches on many performance metrics such as number of served user requests and resource utilization. And finally, Chapter 5 concludes the thesis.

CHAPTER 2

BACKGROUND

2.1 Radio Access Network Sharing

With the growing amount of traffic volume in Mobile Networks, sharing the radio access network (RAN) became quite popular among infrastructure providers (InP) [18]. Network sharing enables operators to significantly and sustainably improve network costs, and ensures them to use network capacity effectively. With an unshared physical platform, decreasing heavy costs included in capital expenditure (CAPEX) and operational expenditure (OPEX) is not possible due to increased number of diverse applications. In addition to cost improvements, sharing also enables creating logical self-contained networks called *network slices* that can be run by an organization who has its own requirements and orchestration way. This opportunity provides higher infrastructure utilization by letting variety of diverse industries enter the game [9]. According to 3GPP mobile broadband standard, there are two ways to share RAN as passive and active. In passive sharing; physical infrastructure is being shared, whereas in active sharing active networks elements are being shared. In active radio access network sharing there are also two architectures that can be used: Gateway Core Network (GWCN) and Multi-Operator Core Network (MOCN) [19]. In GWCN network operators have a common RAN as well as a shared Core Network (CN) nodes.

In the scope of this thesis, we focus on MOCN, where mobile network operators (MNOs) can share RAN with the spectrum-sharing approach and pool their frequencies together and in the meantime they have individual core networks. With the MOCN architecture, the spectrum to be shared needs to be virtualized into func-

tional blocks with well defined external interfaces and functional behaviors in order to achieve flexible resource distribution. [20] refers to resource as a manageable unit defined by a set of attributes or capabilities that can be used to deliver a service. Virtualization enables these resources to serve for a different vertical industry at each scheduling period by decoupling them from the physical network. With the virtualization four important actors come forward: infrastructure provider (InP), mobile virtual network operator (MVNO), tenant/service provider (SP) and end user [20]. In a multi-sliced network, MVNOs lease infrastructure form InPs and enables SPs to allocate the virtualized network. SPs make service level agreements (SLA) with MVNO in order to guarantee their QoS requirements.

The spectrum-sharing model MOCN, enables bandwidth of a base station (BS) to be allocated to each slice in order to meet the SLA requirements. To abstract physical resources of every LTE or WiMAX compatible BS, bandwidth B of BS can be divided into equal sized sub-channels which can be assigned to an end user for a fixed time duration with the help of OFDMA (Orthogonal frequency division multiple access) frame structure [21]. Base stations carry their own resource allocation techniques for each slice in every scheduling round with the virtualization of this slotted OFDMA structure. This virtualization defines atomic resource blocks (RB) that are defined over frequency and time as shown in Fig. 2.1. The RB's are allocated by end users with a periodic schedule of T time slots. During a time slot a sub-channel frequency out of C subchannel frequencies can be allocated. To this end, the total resource to be shared in time and frequency with this granularity of RAN, is $T \cdot C$ RBs.

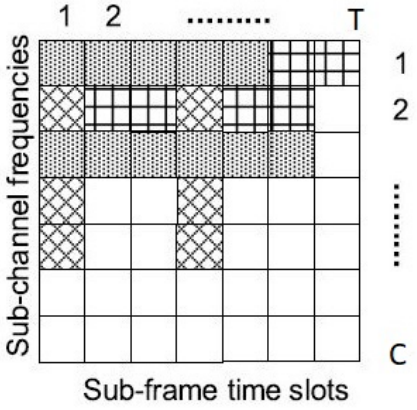

Figure 2.1: Virtualized RAN [1]

To meet the performance requirements such as resource utilization, slice satisfaction, fairness and isolation in an optimal manner, at every scheduling round with the duration T, number of allocated RBs of each slice can be updated by BS.

In addition to bandwidth, base stations (BS) have physical limitations such as *backhaul capacity* and *total transmit power* which should be taken into consideration when resource allocation is being made. The necessary power consumption for occupying a single RB p_{Tx} needs to be adjusted so that total transmit power is sufficient, and sum of simultaneously achieved data rates on the BS needs to be inspected so that backhaul capacity is not exceeded. Another physical limitation between BS and the user is the *channel gain* $h_{\text{u,BS}}$ which is derived using the Free Space Path Loss Model.

$$FSPL_{(\text{dB})} = 20 \cdot log_{10}(d) + 20 \cdot log_{10}(f) - 27.55 \tag{2.1.1}$$

This path loss model depends on the distance d between user and BS and f is the carrier frequency. From the path loss value channel gain between user and BS is calculated as follows [22]:

$$h_{\text{u,BS}} = 10^{-\text{FSPL}/10} \tag{2.1.2}$$

With the parameters explained above, the data rate achieved by an user associated with a single RB of a BS can be calculated as (where N_0 is the noise power density and w is the bandwidth of the resource block)[23]:

$$ar_{\text{u,BS}} = \frac{w}{T} \cdot log_2(1 + \frac{p_{\text{Tx}} \cdot h_{\text{u,BS}}}{N_0 \cdot w}) \tag{2.1.3}$$

For each user if the number of RBs assigned to them increases their achieved data rate increases proportionally.

With the realization of multi-slice existence on a single base station (BS) by RB model, a connected network with multiple BSs can be evaluated in a more efficient way to achieve even better resource allocation. To study this problem in a larger scale, a *Controller* with the global network view should be deployed in compliance with the software defined networking (SDN) principles as in Fig.2.2 [1]. In this hierarchical

model, Controller collects the requests of users and piles them up according to the slice they belong to so that it can decide the total resource demand of each slice on every BS. Controller can effectively calculate the number of RBs to be reserved on each BS with the help of current states of BSs and slices. State of a BS involves resource availability, power capacity and channel gain, while state of a slice involves diverse user requests and user population.

This approach prevents unnecessary reservations on BSs due to lack of information about whole network, and responsibility of a BS narrows down to user allocation of each slice into pre-determined amount of resources. When BSs are informed about the shares each slice gets, they start allocating users that are waiting in the queues of their prefered BSs. While a BS is granting users by allocating enough number of RBs to match their requests, some of the users could be rejected due to insufficient remaining capacity for that slice. In such cases, Controller should monitor the unsatisfied number of customers for each slice and decide a new resource allocation scheme to increase the number of served users.

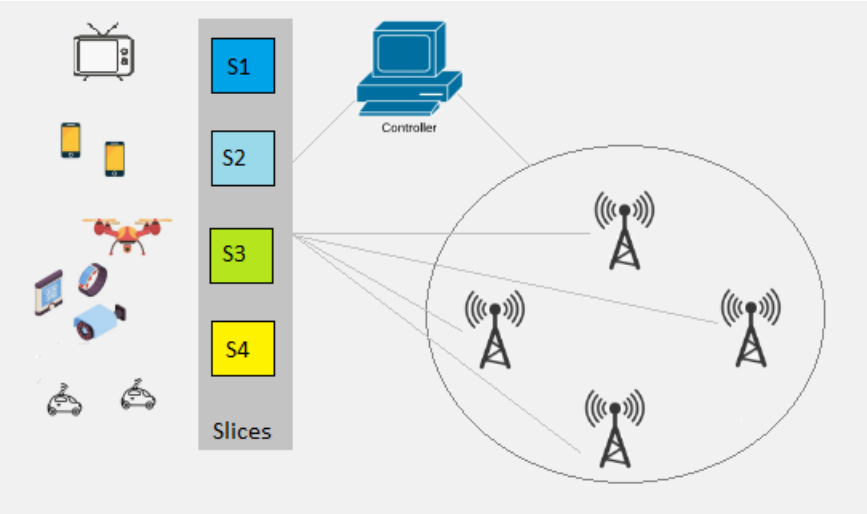

Figure 2.2: System Model

According to this model above, we define the following challenges for network slicing with dynamic resource allocation.

The first challenge is *isolation among the slices*. With the number of slices sharing

limited amount of resources, the performance of a slice should be protected from the demands of other slices. Controller must make sure that a slice does not get affected by the congestion of other slices.

The second challenge is *maximizing resource utilization*. Each slice should utilize their reserved RBs on each BS, so that network can use its physical infrastructure more efficiently. Utilization on every BS should be monitored by Controller, in case unused RBs that are reserved for a slice are detected and re-assignment that does not violate SLA becomes necessary.

The third challenge is *maximizing the number of served users*. With the mobile nature of wireless networks, it is most likely that fluctuations on the user population will occur. The resource allocation problem on the Controller side needs to be re-evaluated dynamically to increase the number of granted users whenever it is possible to enhance slice satisfactions.

2.2 Previous Works

[23] divides the slicing problem into two separate problems, RB allocation and power allocation to realize active RAN sharing with network slicing. To this end, [23] suggests an iterative method by solving two allocation problems separately to find the optimal solution. The result of each separate problem is given to the other problem as an input until both solutions converge to the same output. However, finding the optimal solution with this method is inadequate in terms of calculation time for a real-time application. As a result, a fast heuristic allocation method that gives priority to users with highest channel gain is suggested and uniform power distribution to every RB is assumed. Even though this approach is fast and enables network slicing, it only considers single BS and takes the resource allocation ratios of slices as SLA which is unrealistic in real-life applications.

[2] proposes a static slice scheduler called NVS to maximize base station profit and fulfill demands of slices. This scheduler makes resource provisioning in a weight-based method where weight of a slice is determined with requested data rates of users. In a network with diverse requirements, NVS approach provides a solid solu-

tion in terms of quality of experience (QoE) and strict isolation. However, since every base-station makes the RB reservation without considering the other BSs across network, NVS tends to allocate more resources than needed and causes lower resource utilization.

According to [24], the RAN sharing approaches that focus on only single BS lose the advantage of collecting all resources in a pool. Without considering all base-stations as a connected network, important conditions that affect the decision-making process such as user distributions and channel conditions would be ignored. In order to meet the user demands, choosing the best possible BS first, could lead to higher resource utilization. NetShare framework solves the utilization problem by considering all users on every BS and applies techniques to ensure proportional fairness. NetShare sets minimum and maximum bounds to the amount of assigned resources for each entity in a BS as well as whole network. Algorithm ensures the user's share is in between its pre-determined boundaries to achieve performance isolation.

In another work, [1] proposes a multi base-station resource provisioning approach to achieve the QoS in terms of data rate and delay requirements for mixed traffics in virtualized RAN. To decide these traffic types, [1] collects the improvements of 5G over 4G in terms of higher data rates, low latency, scalability and consistent QoE into four main use cases as: enhanced User Equipment Broadband (UEb), Ultra Low-Latency communications (ULL), Massive Internet-of-Things (MIoT) and High-definition TV (HdTV). All these use cases have diverse performance requirements and traffic characteristics that can be virtualized as network slices on the same physical infrastructure. [1] builds a framework to dynamically customize resources to match individual slice requirements. In the network model, the requests of users have two pieces of information: requested data rate and affordable time delay. For each user, the algorithm sorts the base-stations across the network according to their ability to provide highest data rates per RB. From the sum of requested data rates of users on their preferred BSs *weight* of each slice and from the achievable data rates per RB *rank* of BSs for each slice is formulated. With these two parameters resource shares of each slice at every BS is determined. To make the intra-slice assignments, [1] calculates the height and width of users' requests by deciding the minimum number of time slots and the amount of RBs needed in a single time slot and gives priorities to them according to

their affordable time delay. If the area reserved for a slice on the BS has room for the user's required space "height · width", user is being granted with the shape-based resource allocation algorithm. This admission control method only admits users if their demanded data rate is achievable. Since the shape-based method collects RBs assigned to a user together, resource fragmentation is reduced.

Summary of all approaches can be expressed as in table 2.1. The main contribution that distinguishes each framework seems to be the formulation that enables inter-slice resource allocation. Each technique is supported with a powerful intra-slice allocation method in order to achieve isolation, resource utilization and slice satisfaction. In this thesis, inter-slice resource allocation method called Dynamic and User Count Aware (DUCA) network slicing is proposed and performance comparisons is made with one of the single BS methods [2] and one of multi BS methods [1]. For intra-slice allocation this work uses a "highest channel gain first" algorithm and a 0-1 knapsack solver approach.

Table 2.1: Summary of Different Approaches

	Objectives	**BS model**	**Inter-Slice Allocation Formulation**	**Intra-Slice Allocation Aproach**
LTE [23]	Fairness, isolation	Single	Given by SLA as: $[Share1\ Share2..ShareN]$	Grants users with highest gain first with ensured fairness
NVS [2]	Isolation, customization, resource utilization	Single	Proportional to requested data rates of users Ru: $\sum_{u=1}^{U_{(s)}} Ru \ \ \forall u \in U_{(s)}$	• Gives slice multiple options of schedulers • Lets slice to perform its own scheduling method • Gives flows of the slices a weight and grants highest weight first
NetShare [24]	Isolation, effective distribution, resource utilization	Multi	Proportional to requested resource demand (D) and amount of allocated (A) resources of slices on each BS with minimum (L) and maximum (U) bounds: $\mathrm{D_{(s,k)}} \cdot \log(\mathrm{A_{(s,k)}})$ $\mathrm{L} \leq A \leq U$	Gives options so each slice can have its own user demand calculation policy and then uses the same approach with inter-slice distribution method
Dynamic [1]	Slice satisfaction, isolation, resource utilization	Multi	Multiplication of requested data rate (W) and achievable rate ranking parameter (Rnk) at each BS: $\frac{\mathrm{W_{(s,k)}} \cdot \mathrm{Rnk_{(s,k)}}}{\sum_{\mathrm{s}=1}^{\mathrm{S}}(\mathrm{W_{(s,k)}} \cdot \mathrm{Rnk_{(s,k)}})}$	Shape-Based resource allocation method

CHAPTER 3

DYNAMIC USER COUNT AWARE (DUCA) NETWORK SLICING

This chapter formulates and proposes efficient solutions to the two-level resource allocation problem in virtualized multi-tenant RAN with multiple BSs. First, the inter-slice allocation problem is formulated to consider different requirements of slices in terms of data rate and number of user requests on each BS and Algorithm 1 is suggested as a solution. Second, the intra-slice resource allocation problem is formulated at the BS level and two algorithms; 0-1 knapsack solver and Algorithm 2 are suggested as alternative solutions. Finally, deciding the level of user count effect on resource allocation is discussed.

In the whole system, there are several actors that have direct communication links between each other, as can be seen on Fig. 2.2, that affect resource provisioning process. Each of these actors has different responsibilities for admission control. Every user has a *request* Ru, that contains the transmit rate demand per scheduling frame T. This request reaches to both Controller and BS to become the main input of resource allocation problem. *Channel gain* values are determined at the BSs by taking user distances in the coverage area into account, and reporting back to the controller. With these data not only the achievable transmit rate of users per RB but also required number of RBs per user can be calculated in the controller.

3.1 Inter-Slice Resource Allocation Problem Formulation

If a user is in the coverage area of more than one BS, the Controller chooses the best one with highest achievable data rate to be user's designated BS. The reason behind this assignment is to achieve the best performance possible against physical

obstructions such as channel gain and limited power. After that, the controller uses three pieces of information in the decision formulation.

First, requested transmission rates of users that are chosen for BS k and belong to slice s are added up and normalized with respect to aggregated requested transmission rates of all users on the same BS, to obtain the total requested transmission rate parameter, $ReqRN_{\text{(s,k)}}$, of slice s on BS k:

$$ReqRN_{\text{(s,k)}} = \frac{\sum_{u=1}^{U_{\text{(s,k)}}} Ru}{\sum_{s=1}^{S}\sum_{u=1}^{U_{\text{(s,k)}}} Ru} \quad \forall s \in S \qquad u \in \{1, 2, ..U_{(s,k)}\} \tag{3.1.1}$$

Second, average achievable data rate per RB of slice s at BS k is calculated and normalized with respect to sum of average achievable data rates of all slices on the same BS, to obtain BS ranking parameter $ARN_{(s,k)}$, where $ar_{\text{(u,k)}}$ is defined as in (2.1.3) and U_s denotes the number of users on slice s:

$$ARN_{\text{(s,k)}} = \frac{\frac{1}{U_{\text{(s,k)}}} \cdot \sum_{u=1}^{U_{\text{(s,k)}}} ar_{\text{(u,k)}}}{\sum_{s=1}^{S} \frac{1}{U_{\text{(s,k)}}} \cdot \sum_{u=1}^{U_{\text{(s,k)}}} ar_{\text{(u,k)}}} \quad \forall s \in S \tag{3.1.2}$$

These two expressions in 3.1.1 and 3.1.2 represent the parameters in [1] that are used in its dynamic resource allocation method. Differently, in this thesis, we propose to additionally incorporate the normalized active user count at base-station k for slice s, $Uc_{\text{(s,k)}}$:

$$Uc_{\text{(s,k)}} = \frac{U_{\text{(s,k)}}}{\sum_{s=1}^{S} U_{\text{(s,k)}}} \quad \forall s \in S \tag{3.1.3}$$

Even though, the effect of user count is indirectly involved in parameter $ReqRN$ due to the aggregation of requested rates over users, we observed that without a separate user count parameter, high transmission rate requested slices with small number of users dominate the network by getting larger shares and leaving the slices with low transmission rate requests and high number of users out. To maintain a fairness between these two end, we felt the urge to introduce parameter Uc.

To denote the total amount of radio resources at BS k as in Fig.2.1, parameter $totalRes_{(k)}$

is introduced as:

$$totalRes_{\text{(k)}} = T \cdot C \tag{3.1.4}$$

Finally, with all the parameters resource allocation of slice s at BS k becomes:

$$RSC^{DUCA}_{\text{(s,k)}} = \frac{Uc_{\text{(s,k)}} \cdot ARN_{\text{(s,k)}} \cdot ReqRN_{\text{(s,k)}}}{\sum_{s=1}^{S}(Uc_{\text{(s,k)}} \cdot ARN_{\text{(s,k)}} \cdot ReqRN_{\text{(s,k)}})} \cdot totalRes_{\text{(k)}} \tag{3.1.5}$$

We note that there can be issues with isolation and fairness with this formulation. If a slice's total user count is too high, the congestion on that slice can affect other slices through Uc. Therefore, we add a limiting value $alpha$ to Uc that prevents user count domination in resource allocation. The use of $alpha$ sets a limit to user count that affects Uc and gives flexibility up to that point. For instance, if $alpha$ is equal to 0.3, then user count on BS k that belongs to slice s can only change the value of $Uc_{(s,k)}$ if it is lower than 30 percent of the total number of users on BS k, otherwise it is assumed to have exactly 30 percent of the total users and not more. The method to calculate eq. 3.1.5 is explained in Algorithm 1.

input : userReq (Ru), numberOfUsers (NU), numberOfBS (NBS), numberOfSlices (NS), totalRes, alpha

output: $ARN_{(s,k)}, ReqRN_{(s,k)}, Uc_{(s,k)}$

for $k = 1, \ldots, NBS$ **do**
  for $s = 1, \ldots, NS$ **do**
    calculate $ar_{u,k}$ $\forall u \in U_{\mathrm{s}}$
  end
end

for $s = 1, \ldots, NS$ **do**
  for $u = 1, \ldots, NU$ **do**
    find the $k \in$ {1,...NBS} where $ar_{u,k}$ is max for u
    $U_{(s,k)}$ = $U_{(s,k)}$ + 1
  end
end

for $k = 1, \ldots, NBS$ **do**
  for $s = 1, \ldots, NS$ **do**
    calculate $ARN_{(s,k)}$ by (3.1.2)
    calculate $ReqRN_{(s,k)}$ by (3.1.1)
  end
end

for $k = 1, \ldots, NBS$ **do**
  for $s = 1, \ldots, NS$ **do**
    if $U_{(s,k)} > (\sum_{s=1}^{NS} U_{(s,k)}) \cdot alpha$ **then**
      $U_{(s,k)} = (\sum_{s=1}^{NS} U_{(s,k)}) \cdot alpha$
    end
  end
end

calculate $Uc_{(s,k)}$ by (3.1.3)

Algorithm 1: Controller Resource Allocation Algorithm

3.2 Intra-Slice Resource Allocation Problem Formulation

After the shares of each slice has been decided and reported back to BSs by Controller, BSs start to solve intra-slice allocation problem on their own. In order to obtain the most beneficial distribution of users so that utilization, provided throughput and number of served requests becomes highest, each BS should allocate maximum *C* number of sub-channels on the *T* number of time slots in a scheduling round with the following constraints achieved;

$$min \sum_{t=1}^{T} \sum_{c=1}^{C} x_{t,c}^{s,u} \ \ \forall s, u \in \ (S, U) \tag{3.2.1}$$

where:

$$x_{t,c}^{s,u} = \begin{cases} 1, & \text{if } \mathrm{RB}_{(\mathrm{t,c})} \text{ assigned to user } \textit{(s,u)} \\ 0, & \text{otherwise} \end{cases} \tag{3.2.2}$$

Each RB can be allocated to one user only:

$$\sum_{s=1}^{S} \sum_{u=1}^{U_s} x_{t,c}^{s,u} = 1 \ \ \forall t, c \in \ (T, C) \tag{3.2.3}$$

User requests must be satisfied;

$$\sum_{t=1}^{T} \sum_{c=1}^{C} x_{t,c}^{s,u} \cdot ar_{t,c}^{s,u} > R_u^s \tag{3.2.4}$$

Total amount of BS backhaul limit τ must not be exceeded:

$$\sum_{s=1}^{S} \sum_{u=1}^{U_s} \sum_{t=1}^{T} \sum_{c=1}^{C} x_{t,c}^{s,u} \cdot ar_{t,c}^{s,u} \leq \tau \tag{3.2.5}$$

Controller decided share of a slice at the designated BS must not be exceeded:

$$\sum_{u=1}^{U_s} \sum_{t=1}^{T} \sum_{c=1}^{C} x_{t,c}^{s,u} \leq RSC_{(\mathrm{s,k})} \ \ \forall s \in \ S \tag{3.2.6}$$

We consider two methods to solve the above problem with the maximum number of users granted. The first algorithm is a version of a 0-1 knapsack problem solver that has been discussed in lots of places throughout the literature [25]. In this algorithm, we define *capacity* as the resource allocation output of each slice determined by the Controller, *weight* as the required number of RB of users and, *value* as the user's achieved transmission rate with the needed number of RBs. Here, knapsack algorithm works with the defined capacity to fit users with respect to user parameters weight and value. The main goal is to achieve maximum value with minimum weight.

We propose a second method in this thesis, which lets users with the highest channel gain as defined in Eq. (2.1.2) to have priority over others, so that they can achieve their requested data rates with the minimum possible number of RBs. To this end, we propose the BS resource allocation algorithm 2. In Alg. 2, A is the virtualized RAN matrix where its element $A_{(t,c)}$ is equal to 0 if a user is assigned to the sub-channel c at the time slot t, and UA is the user association matrix that contains the number of RBs assigned to users at their designated BSs. Finally, $cap_{(s,k)}$ is the remaining number of RB capacity of slice s at BS k.

input : $A_{(t,c)}, UA_{(u,k)}, cap_{(s,k)}, userReqs, UserCounts, BSID, numberOfSlices, \tau$

output: $A_{(t,c)}, UA_{(u,k)}, cap_{(s,k)}, \tau$

k= BSID

NS= numberOfSlices

NU = UserCounts

Ru= userReqs

for $t = 1, \ldots, T$ **do**

  for $c = 1, \ldots, C$ **do**

    if $A_{(t,c)} \neq 0$ **then**

      for $s = 1, \ldots, NS$ **do**

        for $u = 1, \ldots, NU$ **do**

          find user u in slice s with highest $h_{(u,k)}$ on BS k

          if $Ru > ar_{u,k} \cdot UA_{(u,k)}$ ***and*** $cap_{(s,k)} > 0$ ***and*** $\tau > 0$ **then**

            $A_{(t,c)}$ = 0

            increase $UA_{(u,k)}$

            decrease $cap_{(s,k)}$

            τ = τ - $ar_{u,k} \cdot UA_{(u,k)}$

          end

        end

      end

    end

  end

end

Algorithm 2: BS Resource Allocation Algorithm

3.3 Determining the user count parameter Alpha

As explained in Section 3.1, the $alpha$ parameter limits the effect of the user count on the allocation. To this end, we propose a heuristic, Algorithm 3, for the Controller to determine the value of $alpha$.

At first, Alg. 3 sets $alpha$ to 0 and obtains the same allocation results with [1], that

has no additional user count awareness. After running one of the BS level resource allocation algorithms (Knapsack or Alg. 2) for each BS, it checks if the number of served users at each slice in every BS is increased compared to current allocation scheme. Then, Alg. 3 increases $alpha$ with a 0.05 step size until it becomes 1 (full user count aware) and chooses the $alpha$ with the highest allocation outcome. The algorithm does not let any granted user to be dropped off since it checks user allocation results of every slice on every BS.

input : oldUserAlloc
output: alpha
alpha= 0
tempAlpha= 0
while $alpha \leq 1$ **do**
  calculate RSC^{DUCA} with $tempAlpha$ via Algorithm 1
  calculate newUserAlloc via Algorithm 2 $\forall$ BS
  if *newUserAlloc* > *oldUserAlloc* **then**
    alpha = tempAlpha
    oldUserAlloc = newUserAlloc
  else
    tempAlpha = tempAlpha + 0.05
  end
end
return alpha;

Algorithm 3: Alpha Decider

3.4 Novelty Of the Proposed Approach

The proposed inter-slice resource allocation method DUCA, provides a novel approach in multi-tenant RAN sharing. DUCA adds the effect of user population to the resource allocation decision as a main contribution where other approaches fail to address. The allocation schemes, such as Dynamic and NVS, that uses only the aggregated user data rate demands of slices and do not consider the impact of number of user requests on respective slices, tend to be unsuccessful in meeting the needs of

the users that request low amount of data. Here we note that such users compose the majority of user population.

NVS considers a single BS approach. The resource allocation is decided with just one parameter: the aggregated data rate demand of every user on the same slice. The proportion of obtained values determines the shares of each slice. With this scheme, resource allocation problem is formulated independent of user channel conditions, therefore slices are more protected from the possibility of an unfair allocation due to unbalanced user distributions around BSs.

Dynamic considers more than one BSs. The resource allocation result differs on each BS and is calculated with two parameters. The first one, similar to NVS, is the aggregated data rate requests of users. However in contrast to NVS, not all the users in BS's coverage is included in summation. Data rate request of a user is only used in its designated BS's calculations. The second parameter is the average of achieved data rates of users on the same slice, if only one RB were allocated to every user. With this approach, Dynamic provides a localized solution to the resource allocation problem for each BS.

DUCA formulates resource allocation problem on multi BS approach with three parameters. The first two parameters are taken from the Dynamic scheme. With the introduction of the third parameter, which contains the limited user count density of each slice, DUCA provides a more fair and efficient resource allocation scheme. Same as Dynamic, user requests are considered in the calculations of their designated BSs, therefore not all the users in a BS's coverage area affect its resource allocation. In addition to accomplishing higher number of granted users, DUCA aims to provide isolation among slices by acknowledging the diversity of user populations in a limited manner.

After DUCA provides an efficient resource allocation scheme, this thesis proposes two intra-slice allocation algorithms to serve the largest number of user requests with high resource utilization results. First, resource allocation problem is considered in a knapsack problem framework and a 0-1 knapsack solver approach is suggested where weight of a user request is defined as the number of necessary RBs and value of a user request defined as the achieved throughput. Second in algorithm 2, priorities are

assigned to users with respect to their channel conditions. Since high data rate values through small amount of RBs are achieved by users with highest channel gain values, this approach provides higher resource utilization results and increased number of granted users by favoring users with minimum physical limitations.

CHAPTER 4

EVALUATION

In order to observe how DUCA improves the challenges introduced in Section 2.1, details of the methods that are compared with DUCA; Dynamic [1] and NVS [2] need to be addressed. The formulations to simulate each method are bench-marked in [1] as follows:

In NVS framework, which has a static slicing approach on a single BS, the fixed share of every slice is computed as:

$$RSC_{(s)}^{NVS} = \frac{\sum_{u=1}^{U_{(s)}} Ru}{\sum_{s=1}^{S} \sum_{u=1}^{U_{(s)}} Ru} \quad \forall s \in S \tag{4.0.1}$$

In Dynamic framework, the same parameters that are formulated in equations 3.1.1 and 3.1.2 are used. [1] calculates the share of slice s on the BS k with eq. 4.0.2 as:

$$RSC_{(s,k)}^{Dynamic} = \frac{ARN_{(s,k)} \cdot ReqRN_{(s,k)}}{\sum_{s=1}^{S}(ARN_{(s,k)} \cdot ReqRN_{(s,k)})} \cdot totalRes_{(k)} \tag{4.0.2}$$

In the next Sections, we compare the inter-slice resource allocation results according to these expressions.

4.1 Performance Comparisons

The algorithms to be compared and simulations are implemented on MATLAB. The simulation scenario of the connected network with multiple BSs is realized with the parameters given in Table 4.1. Except the requested data rates of slices, parameters

in [13] are used for the evaluation. In its simulations, [13] uses a minimum rate demand parameter for each slice, to achieve more diverse requests that are compatible with the abilities of slices, we assumed higher values for requested rates as in Table 4.1. In our simulation model, the network is assumed to be located on an area of $700m \cdot 700m$ with 4 BSs as in [1]. Each BS, whose coverage radius is 300 meters, is placed minimum 250 meters away from the nearest BS, so that the whole area of the network is covered. The user population of every slice is uniformly distributed on the network. The transmitting power of each BS is assumed to be distributed equally among sub-channels.

The slices on the simulations are chosen from [1] and [13] based on their diverse data rate and user population requirements on 5G, as can be seen from Fig. 4.1. Slice 1 is high definition TV (HdTV) and it serves online streaming services to a small number of devices with moderately high demand of data rate. Slice 2, represents the enhanced mobile broadband (eMBB) which has highest data rate requirement among all. Slice 3 is the massive internet of things (MIoT) services with high number of devices and low data rate requirements. Finally, Slice 4 is the ultra low latency (ULL) service with a very little data rate demand and moderately high number of device count.

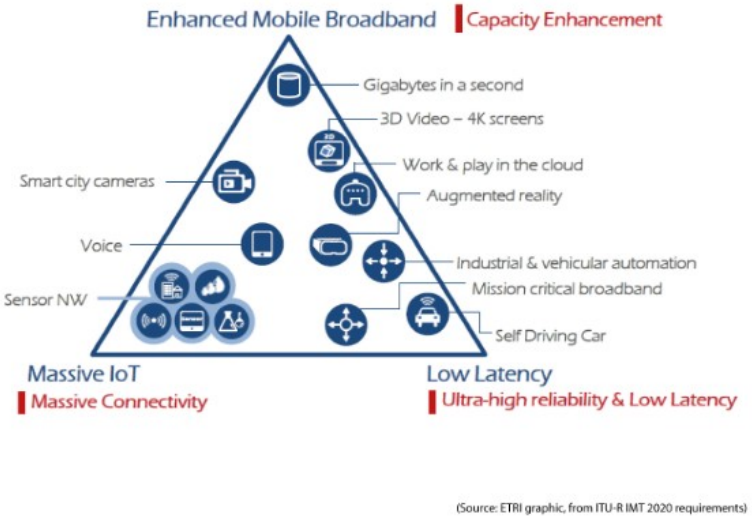

Figure 4.1: Slices of 5G [3]

DUCA aims to achieve better performance through consideration of the user count. To this end, we construct simulation scenarios with different user configurations. We

emphasize that the effect of diversity on user populations is not considered in [1] and [2] as a critical issue.

The first configuration has a realistic and constant number of users that are assigned to each slice in a certain amount of time. The aim of this experiment is to clearly see the effect of DUCA compared to [1] and [2] under stable request traffic.

The second configuration realizes user requests that arrive at the network at random time instants and do not exit. The reasoning behind this experiment is to see the effect of increasing demand, and compare the performance of algorithms on a slowly congesting network.

The third configuration realizes a realistic model where user requests come randomly in time and each of them has a lifetime. For the first two configurations, every request on the network stays permanently, however, this third configuration represents the realistic scenario of RAN where users that are leaving the network after being served enough amount of time is considered.

The forth configuration realizes the effect of mobility when realistic and constant number of mobile users are at the network. The aim of this experiment is to see the performances of DUCA, NVS and Dynamic while user requests change their designated BSs repeatedly.

The fifth and final configuration realizes the effect of additional slices and a large number of user requests being present on the network. The capacity of BSs are also increased to 100 MHz in this configuration to support large user populations. The aim of this experiment is to see the behaviour of DUCA with increased diverse slice requirements.

Table 4.1: System Parameters

Parameters	Values
The number of BS	4
The number of slices	4
System Bandwidth	5MHz
Backhaul capacity	2 Gbps for each BS
Transmitting Power	30 dBm for each BS
Noise power density	-174 dBm/Hz
Requested Rates	HdTV:1000, eMBB:5000, MIoT:100, ULL:10 (kbps)
User Count	HdTV:60, eMBB:11, MIoT:240, ULL:124
User distribution	Uniform
Sub-channels per BS	25
Sub-frame duration	1 ms
Frame length	10 ms

4.1.1 Experiment 1: Constant Number Of Users

In the first experiment with the realistic and constant number of users, the user counts given in the system parameters Table 4.1 are used. The number of users are consistent with real life, since high data rate demanding slices tend to have lower number of users and slices with high number of users usually have low data rate demands [26]. Inter-slice resource reservations are done according to Eqns. 4.0.1, 4.0.2 and 3.1.5 on each BS. The resulting reservation values decided by the controller can be seen in Fig. 4.7.

Before proceeding to the resource allocation comparisons of NVS, Dynamic and DUCA, the two BS level user allocation methods; 0-1 knapsack solver and Algorithm 2 are compared under DUCA reservation. Resource allocation results of each BS can be found on figures 4.2, 4.3, 4.4 and 4.5. As it can be seen clearly, granted number of users are very close to each other in both of the algorithms. Total through-

put achieved on each BS, Fig 4.6, is mostly the same with little differences. In spite of all the similarity, the execution times of both algorithms are far away from each other. 0-1 knapsack solver can reach up to 10 times higher execution times than Algorithm 2 to compute if user count is high (as in the case with Slice 3 on each BS). Even though knapsack approach provides a little higher throughput on some BSs, due to unacceptable execution time, Algorithm 2 will be used on all the experiments from now on.

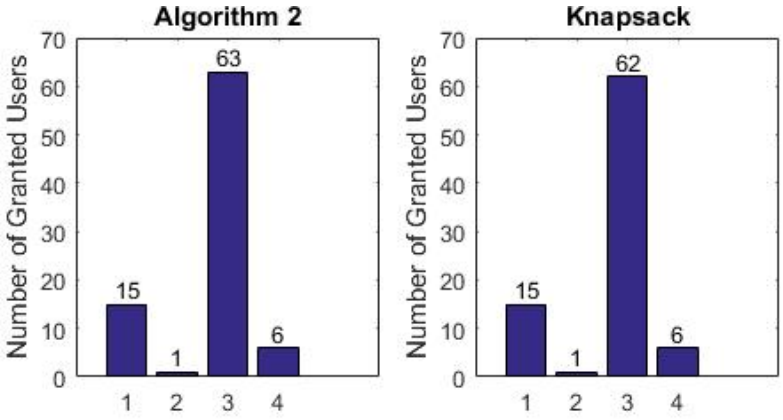

Figure 4.2: Resource Allocation of Slices on BS1

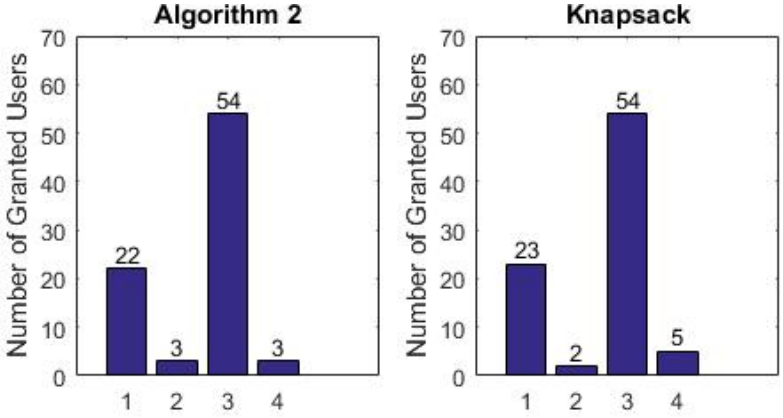

Figure 4.3: Resource Allocation of Slices on BS2

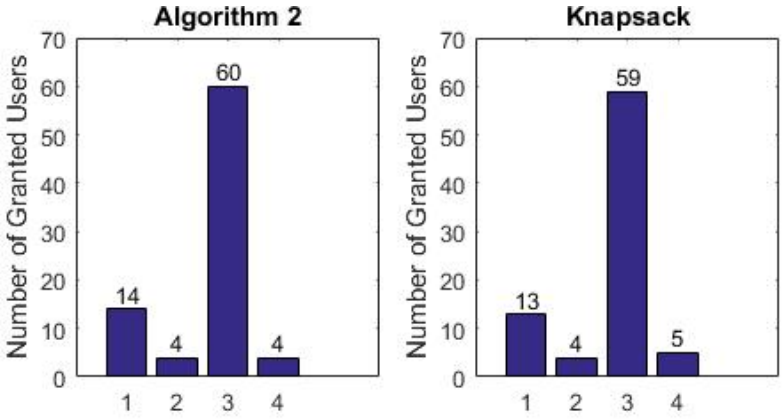

Figure 4.4: Resource Allocation of Slices on BS3

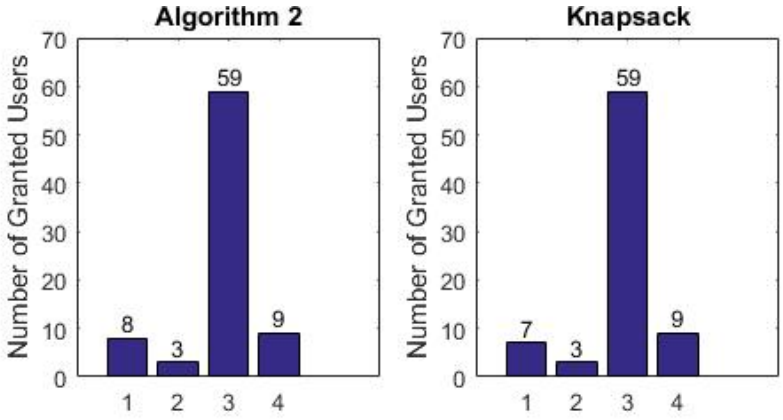

Figure 4.5: Resource Allocation of Slices on BS4

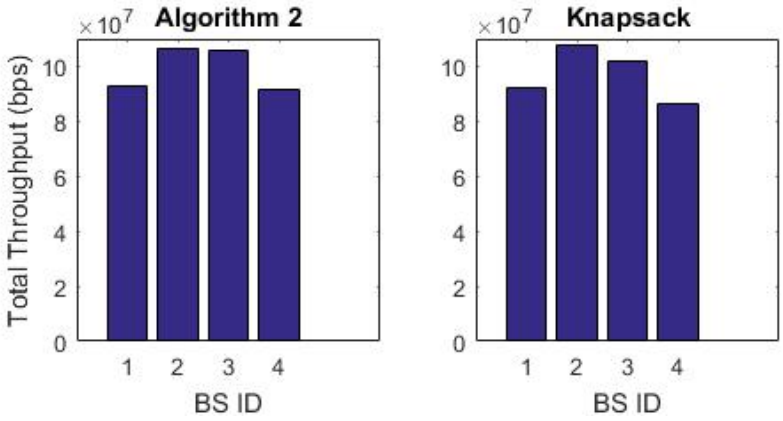

Figure 4.6: Throughput comparison

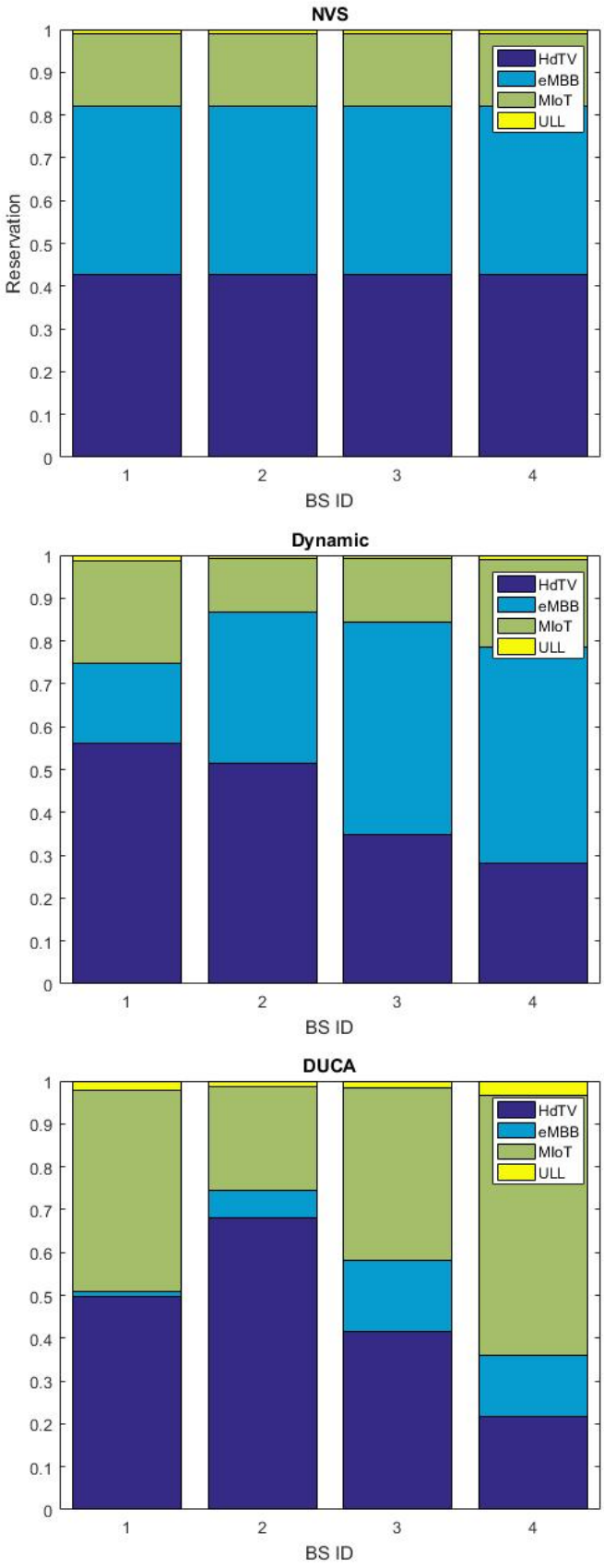

Figure 4.7: Comparison of Reservations

According to reservations calculated with NVS, Dynamic and DUCA as in Fig 4.7, Intra-Slice resource allocations are done to compare the performances of reservations in terms of number of served users and resource utilization. Afterwards, the isolation performance of DUCA is inspected.

The BS level resource allocation results in terms of number of granted users on each BS under NVS, Dynamic and DUCA can be seen on Fig. 4.12. Clearly highest number of served users achieved with DUCA on all of the BSs. NVS and Dynamic allocated lower number of users on the low data rate requested slices ULL and MIoT as expected. In addition to increasing the numbers on ULL and MIoT, DUCA also maintained same level of allocations on HdTV and eMBB.

The resource utilization results in Fig. 4.9 show that DUCA decreases the amount of unused resources by increasing the total number of granted requests compared to NVS and Dynamic. The unused reservations under NVS and Dynamic consists of eMBB and HdTV slices as compliant with our claim about the domination of high demanding slices on the network while other slices starve.

We evaluate the the capability of providing isolation of DUCA by increasing the number of users on MIoT d to 500 in order to create congestion and monitoring the allocation of other slices with respect to time. As can be seen on Fig. 4.10, with the increase of users on MIoT at scheduling round 50, the performance of slices other than MIoT is unaffected. Therefore, DUCA protects the isolation of slices.

The number of served users under DUCA with other possible α values can be seen on Fig. 4.11. Since the value 0.3 provides the highest count, DUCA used it as the user count limitation on the resource allocation of this experiment.

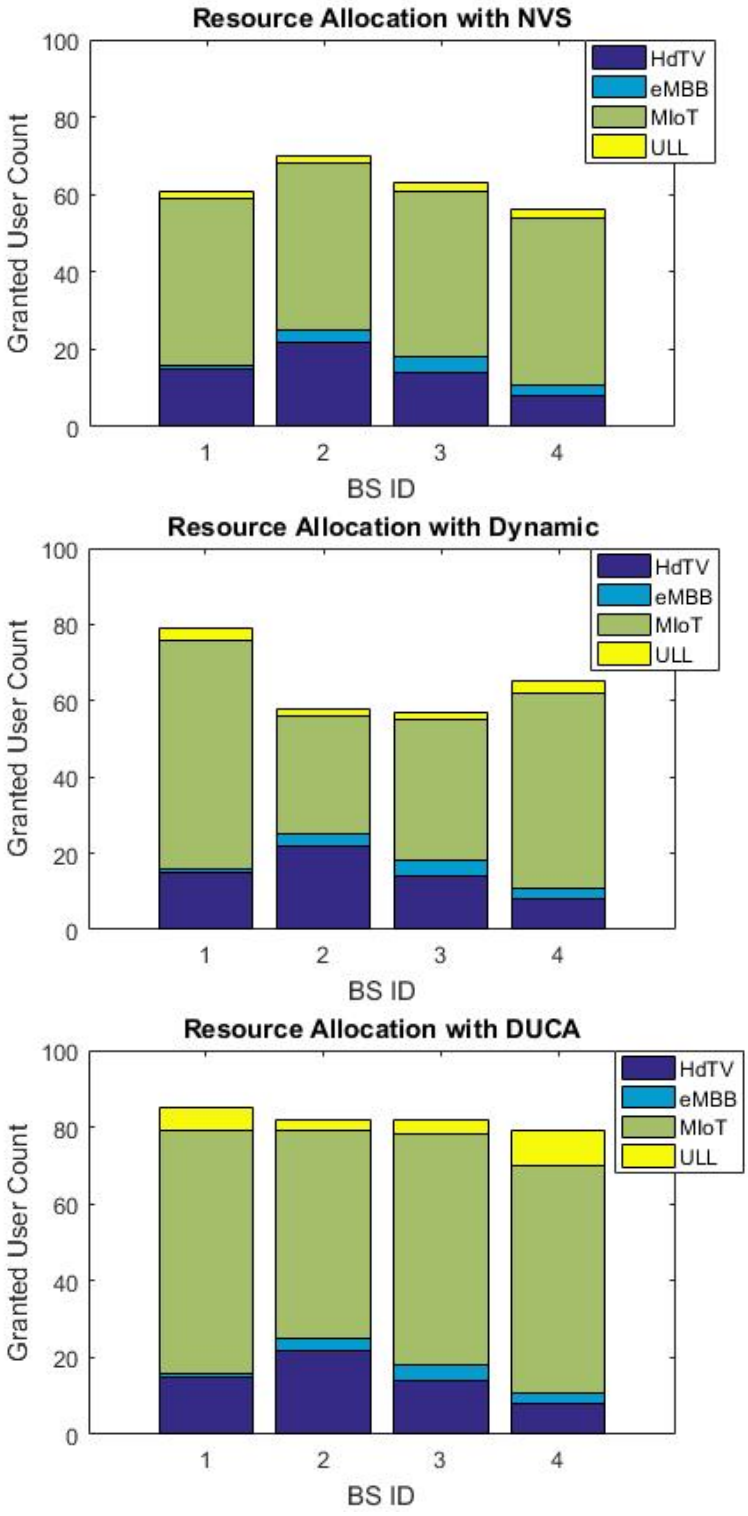

Figure 4.8: Number of Granted Users Comparison

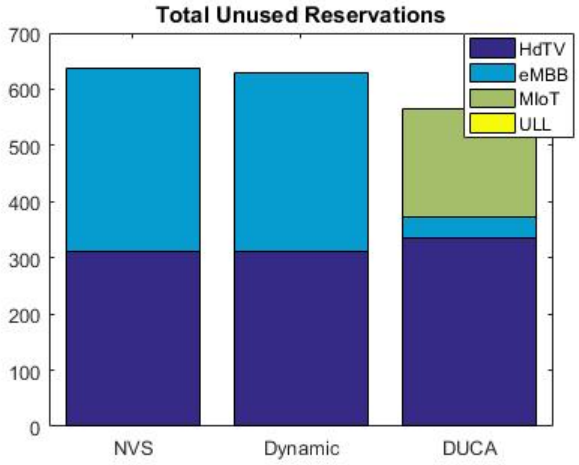

Figure 4.9: Resource Utilization

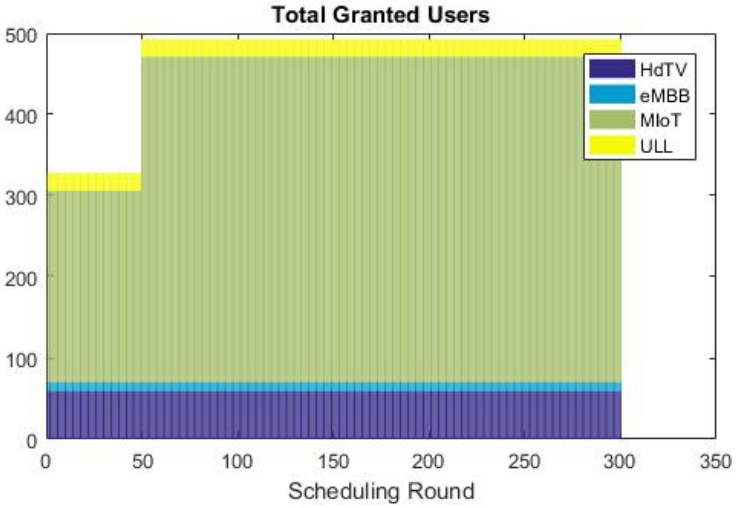

Figure 4.10: Isolation performance

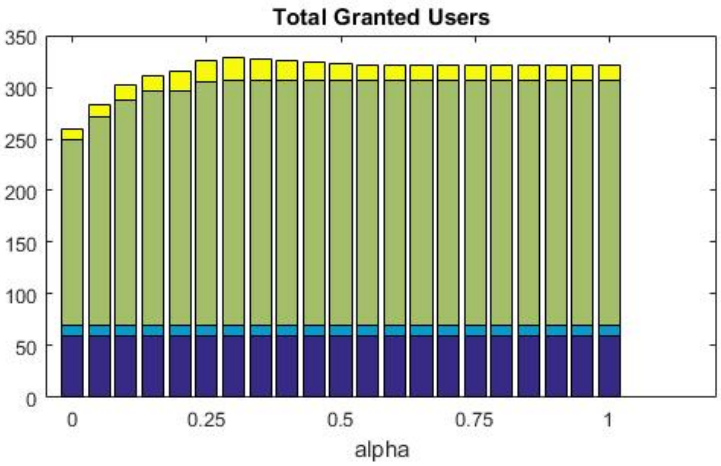

Figure 4.11: Granted User Counts vs alpha

4.1.2 Experiment 2: Poisson Request Arrivals

In the second experiment, we use Poisson distribution to model arrival of random number of user requests to the network. We assume that the users on each slice have Poisson arrival rates with different means where each mean is chosen to be realistic according to respective slices. The initial user state of the network is chosen to be same as the first experiment. The mean of number of users arriving to network during every 50 scheduling round is determined from the initial state as; 4 for HdTV, 1 for eMBB, 12 for MIoT and 8 for ULL.

In the simulation, three identical networks have been modeled where controller of each runs one of the algorithms: DUCA, Dynamic or NVS. In each network, Controller collects all the new requests during a period of 50 scheduling rounds and calculates new reservation values with their respective algorithm. If new reservations provide better allocation results in terms of number of granted users for each slice, controller reports those new shares to the BSs. If new reservations require dropping some users off, controller does not notify BSs and lets them use old reservation values. The simulation has been run for 3000 scheduling rounds.

Total number of served users as can be seen from Fig. 4.12 is significantly higher under DUCA reservation. As we expected, DUCA provides higher reservation values for the slices with lower data requests and large user populations, whereas both NVS and Dynamic favor slices with higher data requests even they have small user populations. As a result, NVS and Dynamic schemes granted higher number of user requests on eMBB, while with DUCA, users on MIoT and ULL slices had higher possibilities to be served as in Fig. 4.13.

Since eMBB users achieve highest throughput numbers, letting users of MIoT instead of eMBB can reduce total achieved throughput on BSs. As can be seen from Fig. 4.14, until the user count becomes 1000, the achieved throughput is similar under all three reservation schemes. After that point, throughput of NVS and Dynamic exceed that of DUCA. However, larger numbers than 1000 users in a network like ours represents a very heavy load as stated in [1], and possibility of it happening frequently is low.

Total amount of unused reservations with respect to slices as in Fig. 4.15 shows

that resource utilization under DUCA is higher than NVS and Dynamic. As network slowly becomes congested the effect of over-reservation on the high data requesting slice, eMBB, can be seen more clearly.

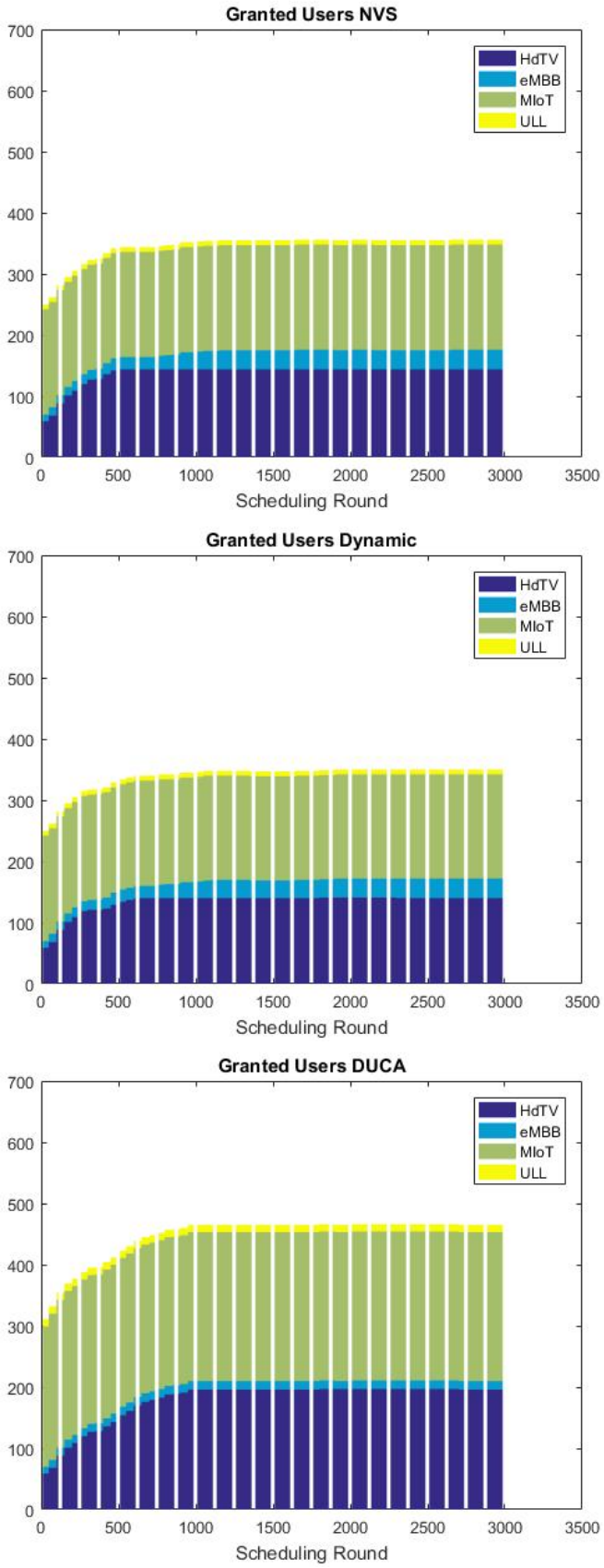

Figure 4.12: Total Granted Users

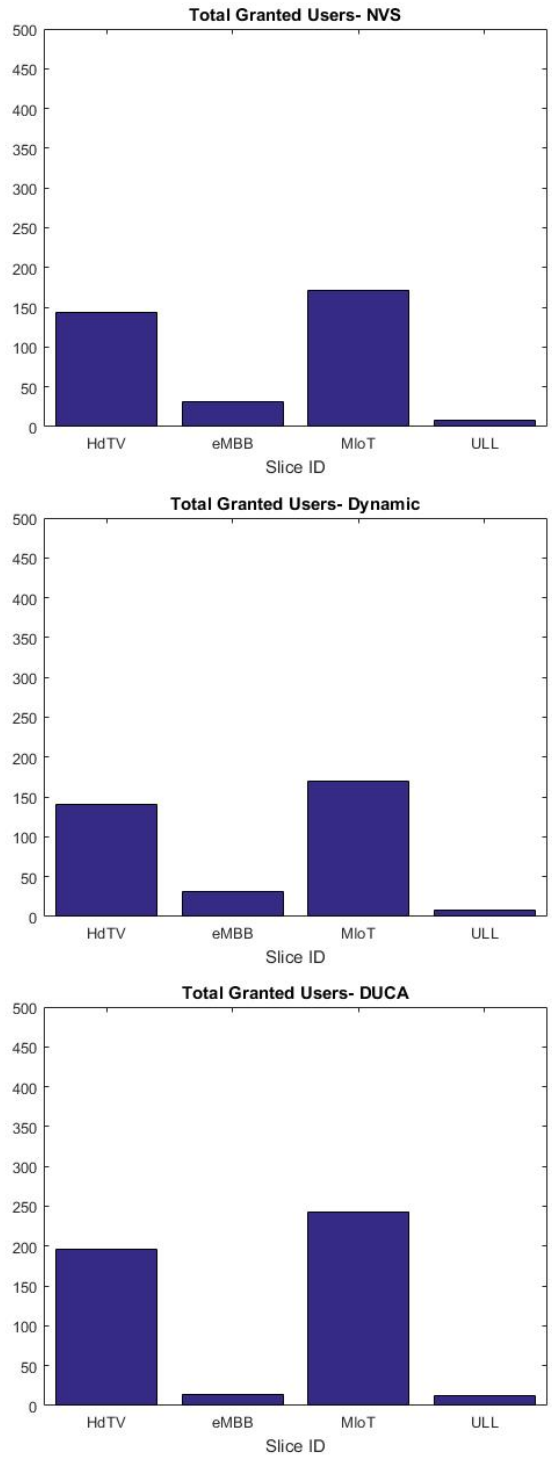

Figure 4.13: Total Granted Users wrt Slices

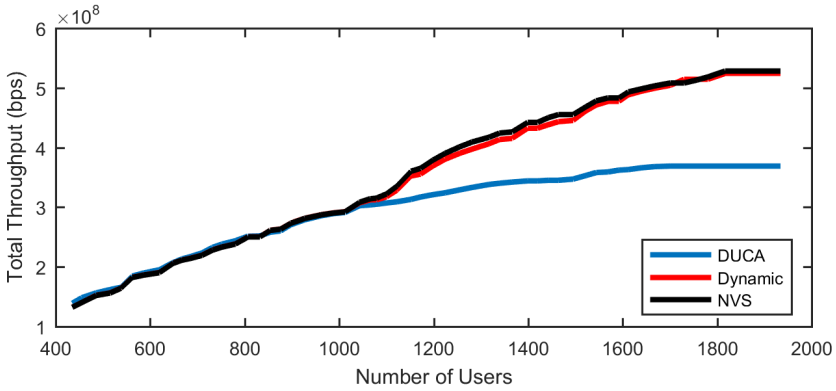

Figure 4.14: Total Throughput

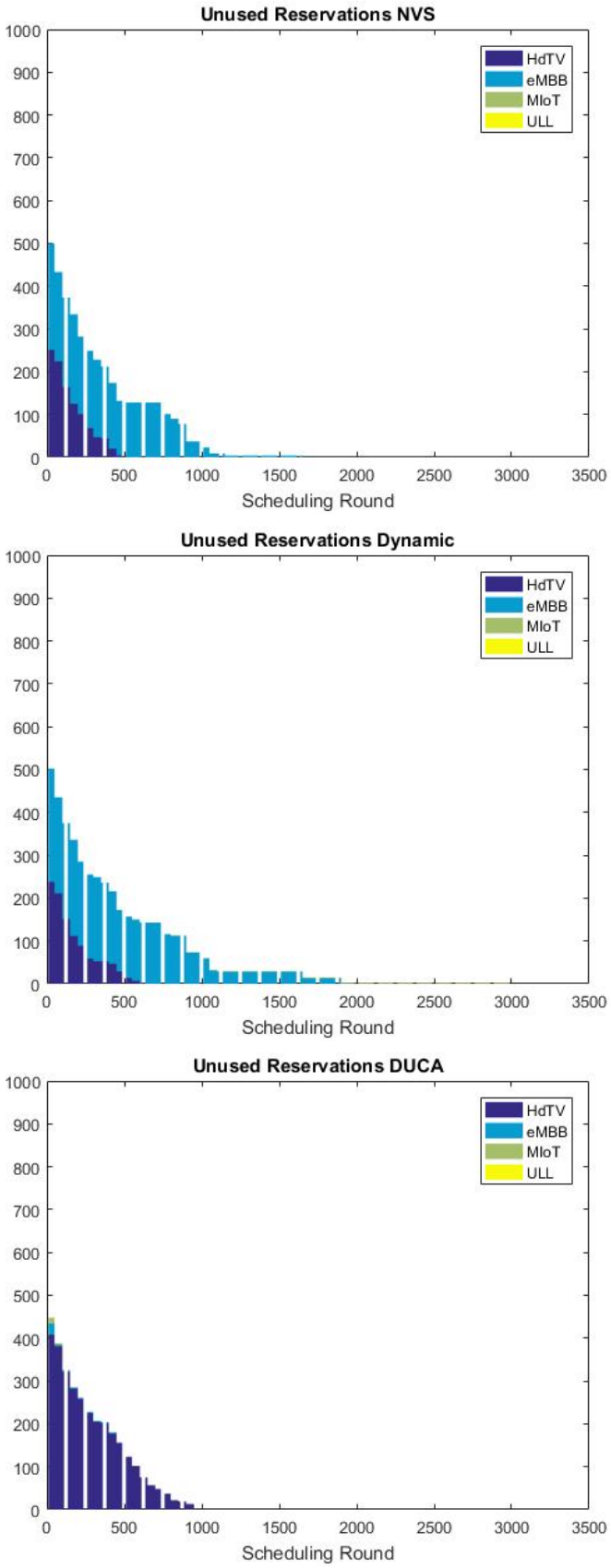

Figure 4.15: Total Unused Reservations

4.1.3 Experiment 3: Poisson Request Arrivals with Exponential Service Time

We model Poisson request arrivals with exponential service times to create a realistic network where user requests come randomly with a lifetime. The initial state of the network is chosen to be the same as Experiment 1. We model the exponential service times using the realistic use cases of slices. As stated in [27], eMBB users tend to stay active for an extended time intervals whereas MIoT users stay active for shorter duration and ULL users stay active for even smaller block-lengths. Since HdTV users require moderately high data rates they are also assumed to be stay active for a long period of time. To this end, we set the mean service time of requests in scheduling rounds (SR) to be 200 SR for HdTV, 200 SR for eMBB, 100 SR for MIoT and 10 SR for ULL in the experiment with 3000 scheduling rounds. Mean arrival rates of slice requests are set to be 15 for HdTV, 3 for eMBB, 60 for MIoT, 100 for ULL in a 50 scheduling-round duration to keep the number of active requests realistic and stable during the experiment. Same as the second experiment, in the simulation three identical networks have been modeled where the Controller of each runs one of the algorithms: DUCA, Dynamic or NVS. In each network, the Controller collects all the new requests during a period of 50 scheduling rounds and calculates new reservation values with their respective algorithm. If new reservations provide better allocation results in terms of number of granted users for each slice, controller reports those new shares to the BSs. If new reservations require dropping some users off, controller does not notify BSs and lets them use old reservation values.

Similar to first two experiment, the number of granted users on the network is highest under DUCA algorithm as can be seen from Fig. 4.16. Since the majority of the requests come from MIoT and ULL users, Dynamic and NVS schemes fail to grant them due to the over-reservation of eMBB and HdTV slices. When we examine the queues of slices, where requests that are waiting to be granted piles up as in Fig. 4.17, DUCA achieves smaller number of queued user requests compared to Dynamic and NVS. All the algorithms fail to meet the demands of ULL requests due to very small amount of data rate demand, and even DUCA could not provide enough reservation values to ULL despite of its high number of users. Total queued users graphic on Fig. 4.17 shows that, high number of MIoT users wait in the queues under NVS and

Dynamic. On the other hand with DUCA, some very small number of eMBB users are queued, and they have been stayed stable and not showed an infeasible rise in the upcoming scheduling rounds.

The total number of user requests that have been served and exited the network is highest with DUCA as in Fig. 4.18. The throughput achieved under all three schemes are closed to each other, therefore giving high reservation values to the slices with low data rate demands, in case they have high number of users, did not cause DUCA to under-perform. To address utilization problem in a realistic network, Fig. 4.20 is used to show that DUCA provides lower numbers of unused reservations compared to NVS and Dynamic and therefore achieves higher resource utilization values. In the configuration where the users come with a Poisson arrival and leave the network after being served with an exponential service time, DUCA seems to respond better to the needs of slices than NVS and Dynamic.

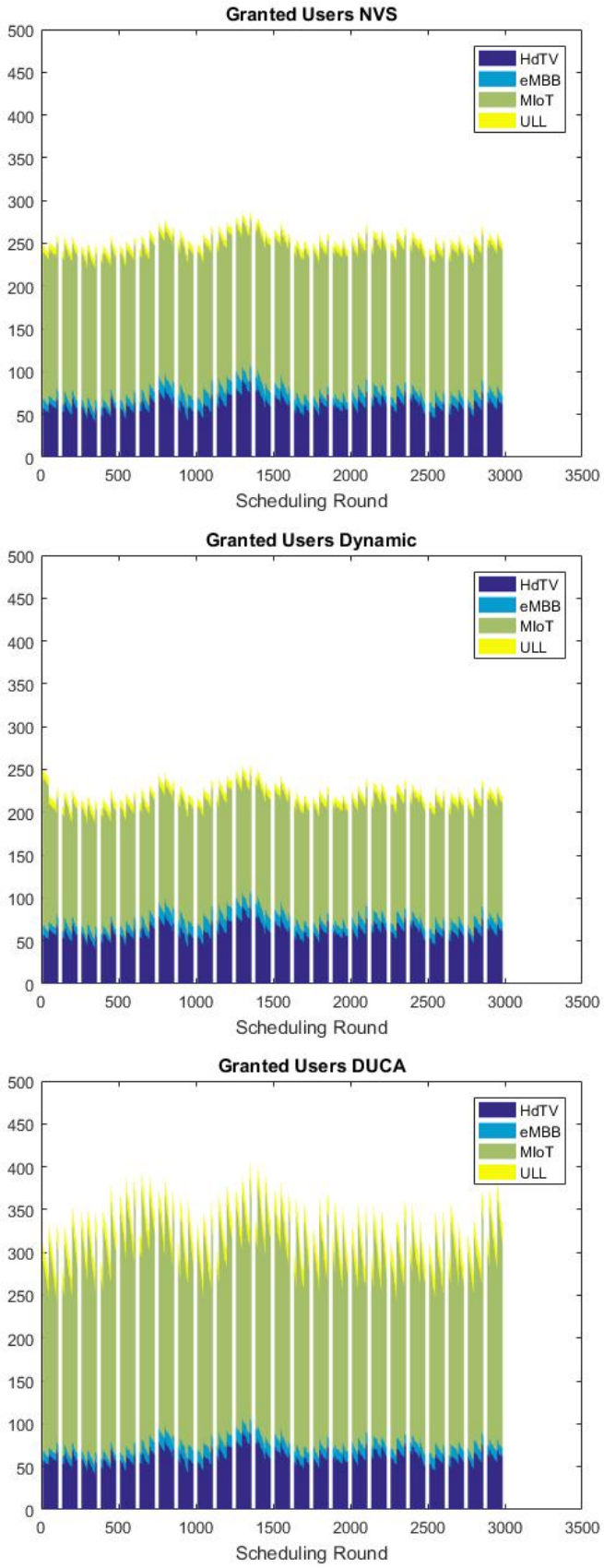

Figure 4.16: Granted Users

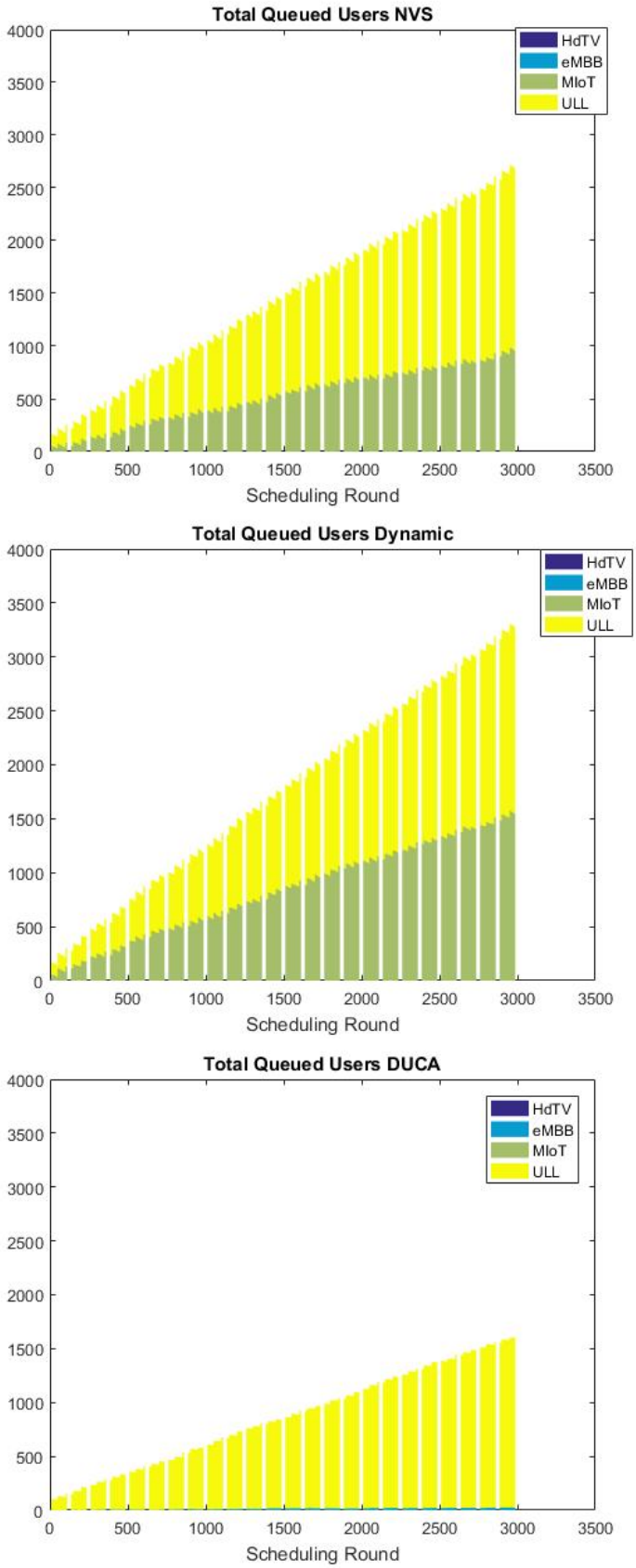

Figure 4.17: Total Queued Users

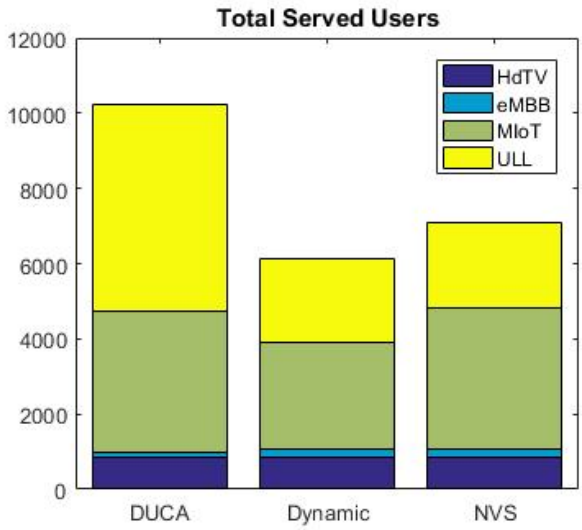

Figure 4.18: Total Served Users

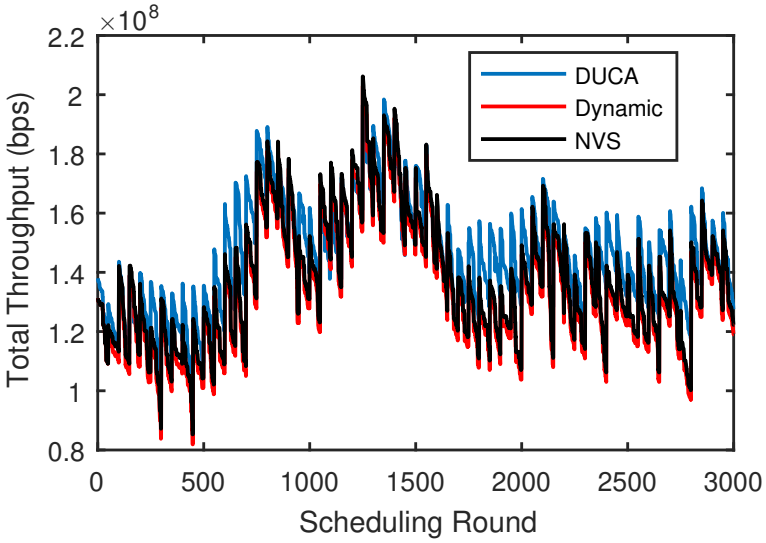

Figure 4.19: Total Throughput

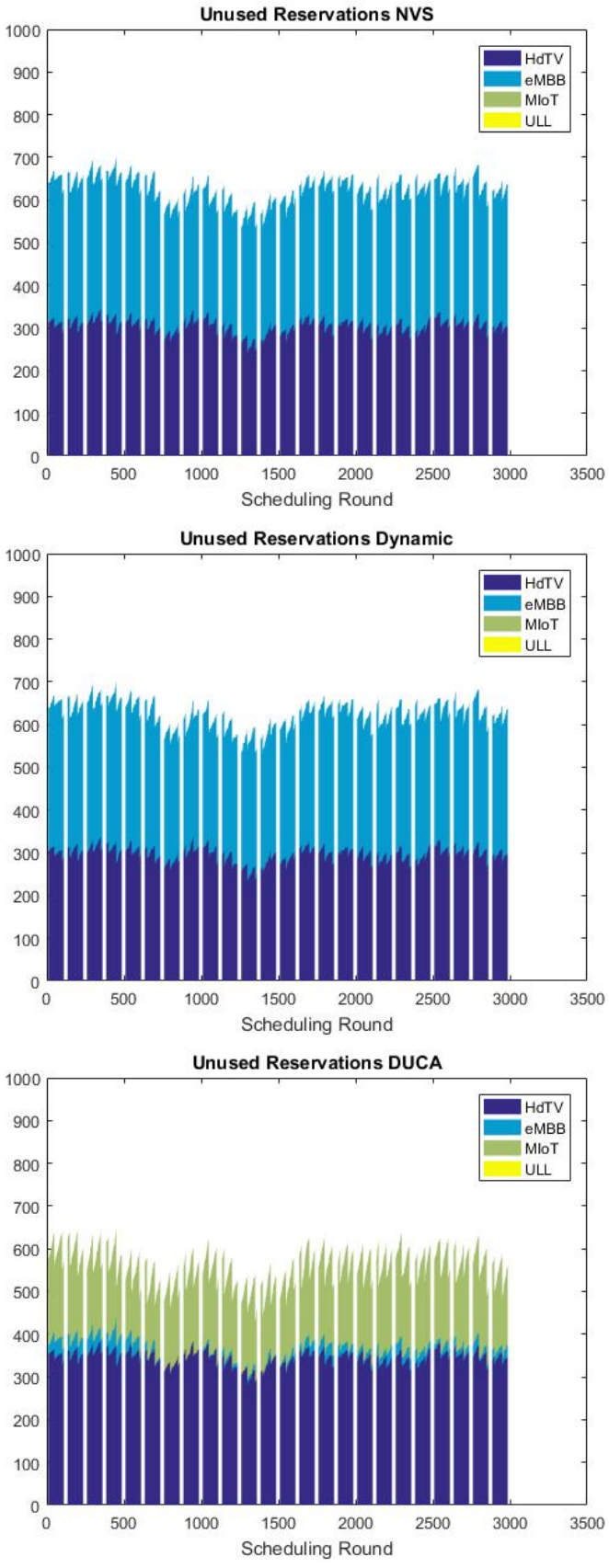

Figure 4.20: Total Unused Reservations

4.1.4 Experiment 4: Mobility

In this fourth experiment, we evaluate the effect of the user mobility. We assume that each user moves according to Random Waypoint Mobility model that has been widely used on mobile networks [28]. In this model, each user pauses for a fixed duration and then selects a random destination in the network area and moves there with a random speed. After a while, user pauses for a fixed duration again before it moves to another random location. This process is repeated for the entire simulation.

We select a constant number of users same as Experiment 1 to be on the network. We assume that the users only leave the network due to mobility. One critical point we ensured is providing a connection to a user that is in motion without any interruption. To this end, If a user is granted on one BS and leaves its coverage area or gets higher channel gain levels on another BS due to mobility, that user gets a priority among user requests on the queue of the new BS. As in the previous experiments, in the simulation three identical networks have been modeled where the Controller of each runs one of the algorithms: DUCA, Dynamic or NVS. If new reservations provide better allocation results in terms of number of granted users for each slice, controller reports those new shares to the BSs. If new reservations require dropping some users off, controller does not notify BSs and lets them use old reservation values. Simulations are executed for 2000 scheduling rounds.

Since the slices with high user populations experience more re-allocations than low user populated slices, the reservations made for them needs to be higher to support mobility of large number of user requests. Therefore, as it can be seen from the granted user numbers on each base-station in figures 4.21, 4.22, 4.23 and 4.24 DUCA outperforms Dynamic and NVS under the effect of mobility. Since total number of requests does not change on the network, NVS could not catch Dynamic and DUCA's performance. Even though Dynamic provides high numbers on some BSs, DUCA provides better ones on every one of them. As a result, total number of served users as can be seen clearly from Fig. 4.25 is highest under DUCA resource allocation scheme.

Resource utilization results show that with NVS and Dynamic high data demanded

users composes most of the unused reservations as in Fig. 4.26. Since with DUCA, number of user requests adjustment provides better allocation results, total unused reservation value is lower than Dynamic and NVS.

From the total achieved throughput values in Fig. 4.27; even though favoring low data demanding slices due to high number of users, DUCA provides slightly higher results than Dynamic and NVS.

To conclude, under the effect of mobility DUCA performs better than Dynamic and NVS in terms of number of served users and resource utilization.

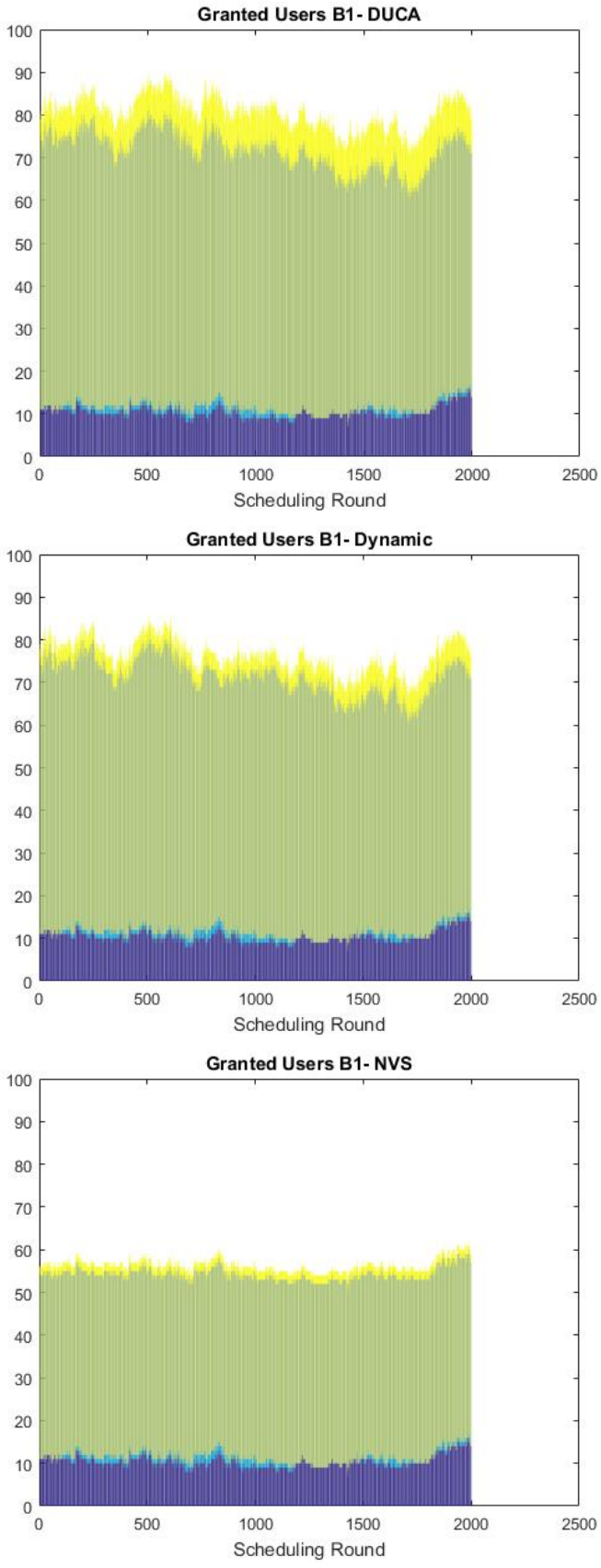

Figure 4.21: BS1 Granted Users under Mobility

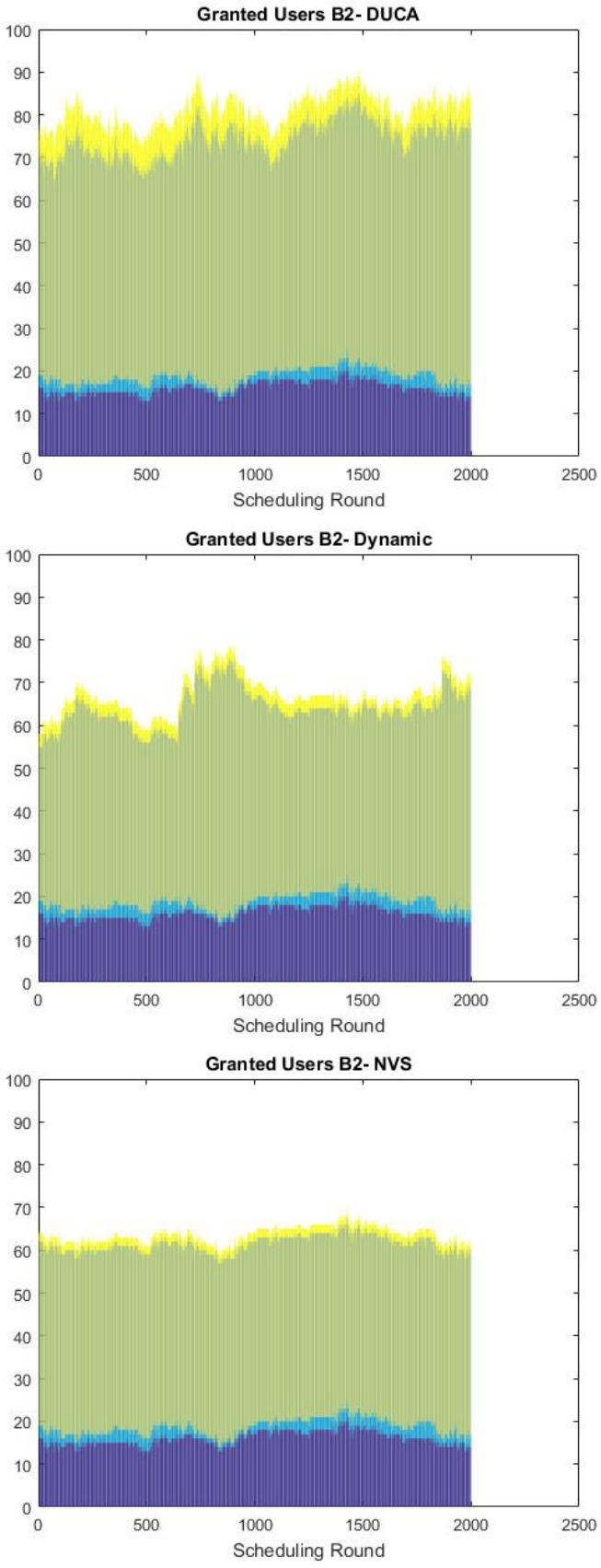

Figure 4.22: BS2 Granted Users under Mobility

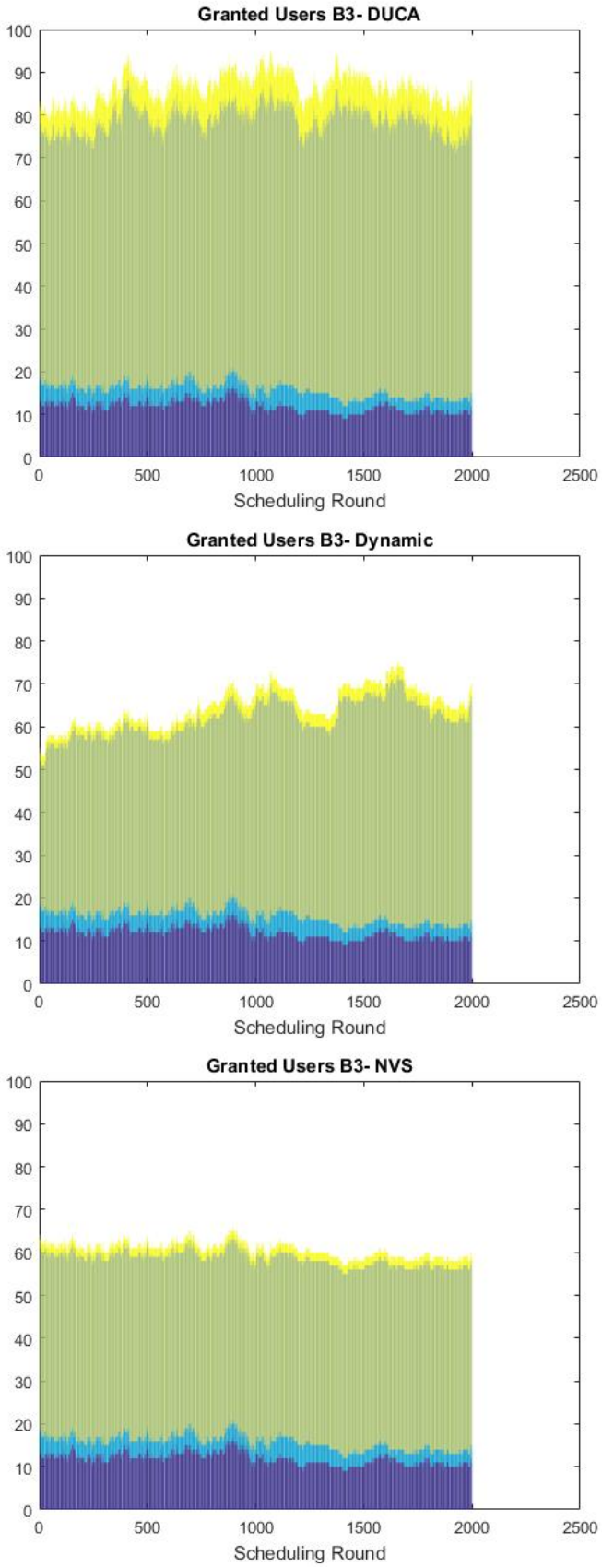

Figure 4.23: BS3 Granted Users under Mobility

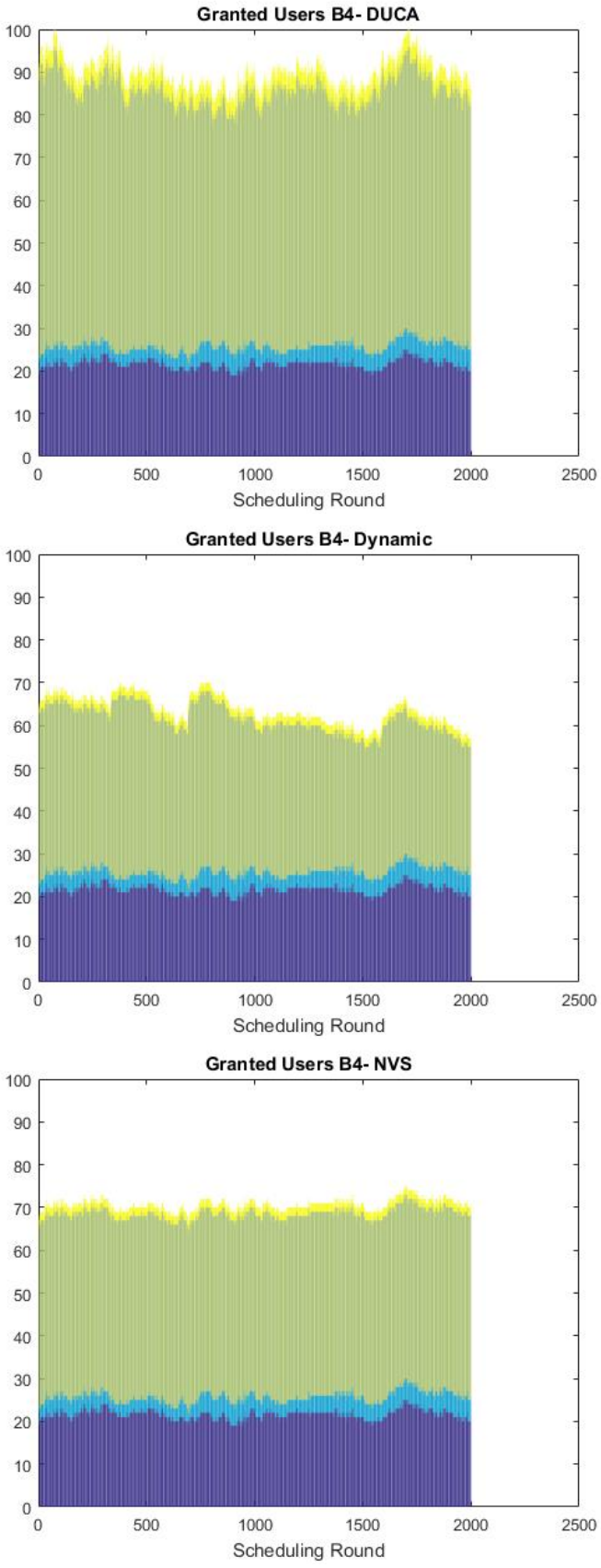

Figure 4.24: BS4 Granted Users under Mobility

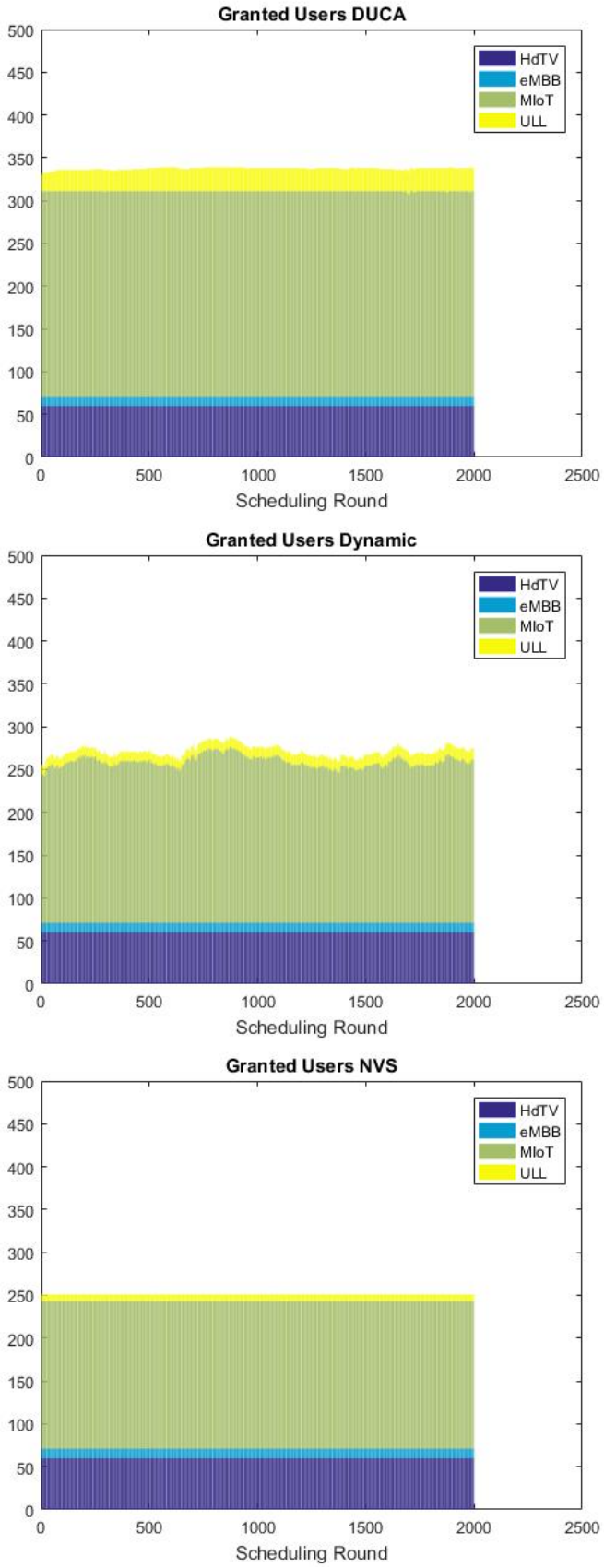

Figure 4.25: Total Granted Users under Mobility

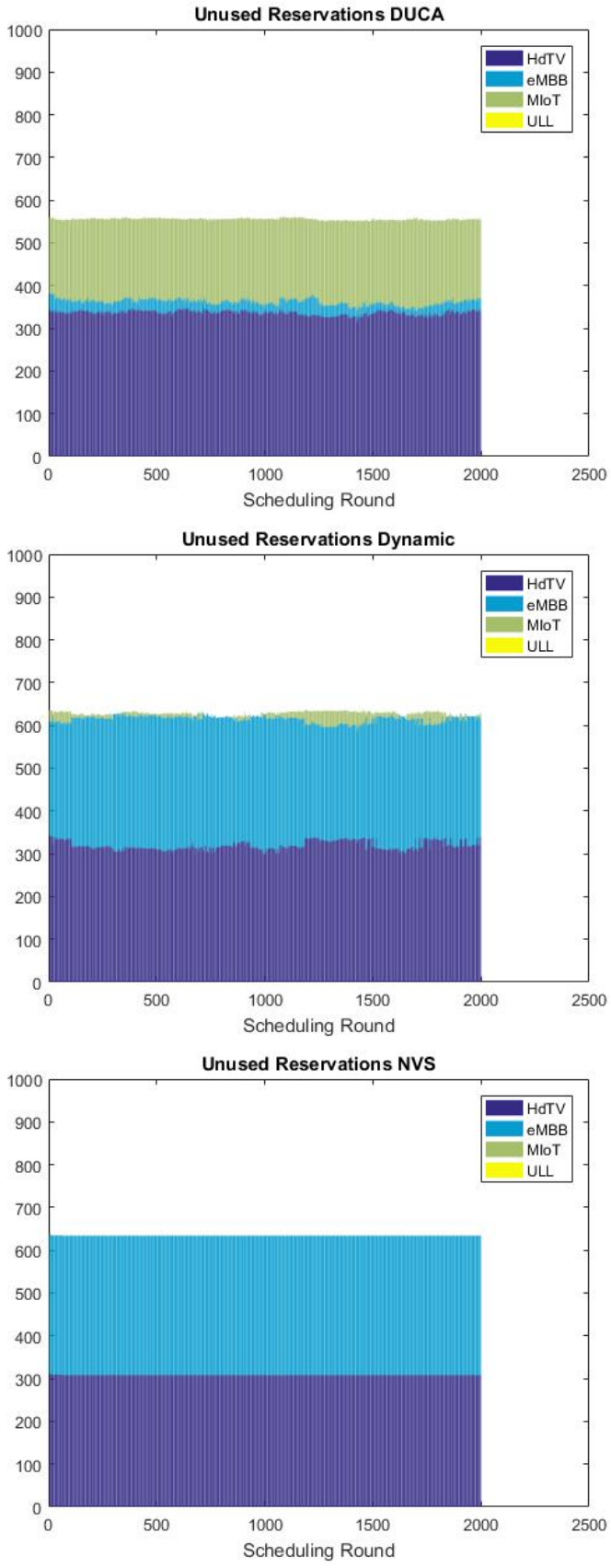

Figure 4.26: Total Unused Reservations under Mobility

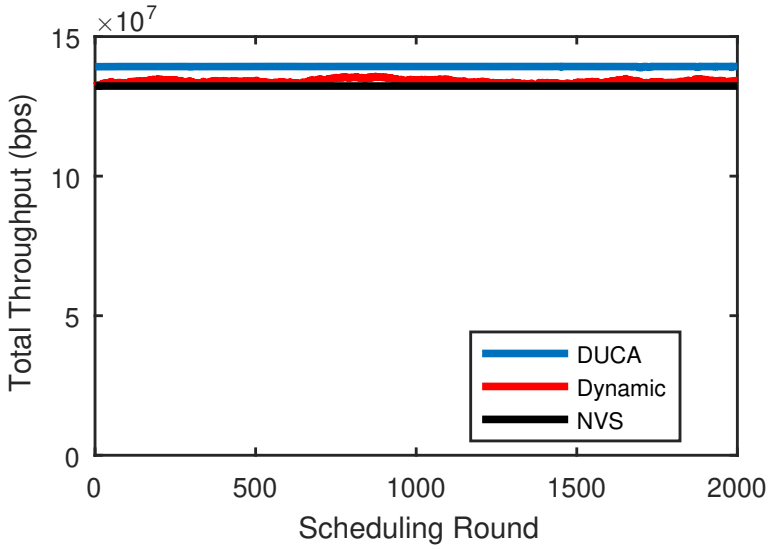

Figure 4.27: Total Throughput under Mobility

4.1.5 Experiment 5: Additional Slices with Increased Capacity

In this final experiment, we added two more slices to network and increased the bandwidth of BSs from 5 Mhz to 100 Mhz. To simulate more realistic demands from new BS bandwidths, user populations on slices are also increased to 700 for HdTV, 400 for eMBB, 2400 for MIoT and 1700 for ULL. The data rate requirements of two additional slices that we call AddS1 and AddS2 are assumed to be 150 kbps and 2 Mbps respectively. In compliance with data rate demands, their user populations are also decided to be 2000 for AddS1 and 500 for AddS2. In this experiment we keep the number of users constant as in Experiment 1.

As can be seen from simulation result in Fig. 4.28 with the presence of new slices number of total granted users is still highest under DUCA. MIoT and AddS1 slices that have higher user populations with small data rate demands are favored more with DUCA compared to Dynamic and NVS.

The number of unused reservations as in Fig. 4.29 is also lower with DUCA. Dynamic and NVS makes over-reservation for high data rate demanded slices eMBB and AddS2. Total throughput under all resource allocation schemes are similar and results are 4.7 Gbps for DUCA, 4.61 Gbps for Dynamic and 4.62 Gbps for NVS.

To conclude, with different network models with increased BS capacity and increased number of slices, DUCA still outperforms both Dynamic and NVS.

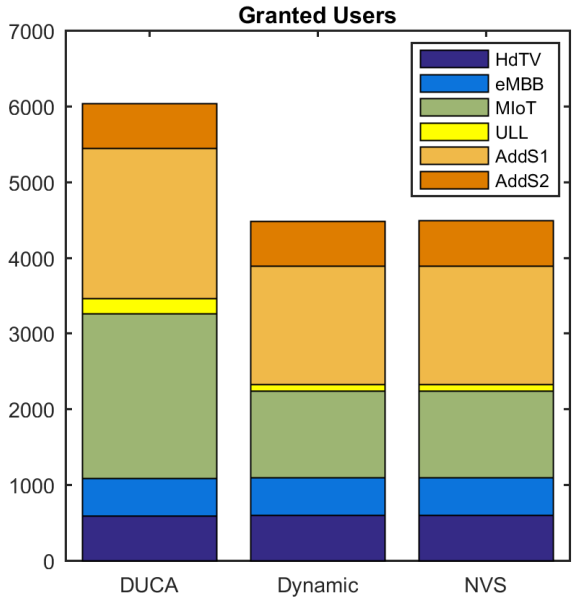

Figure 4.28: Granted Users

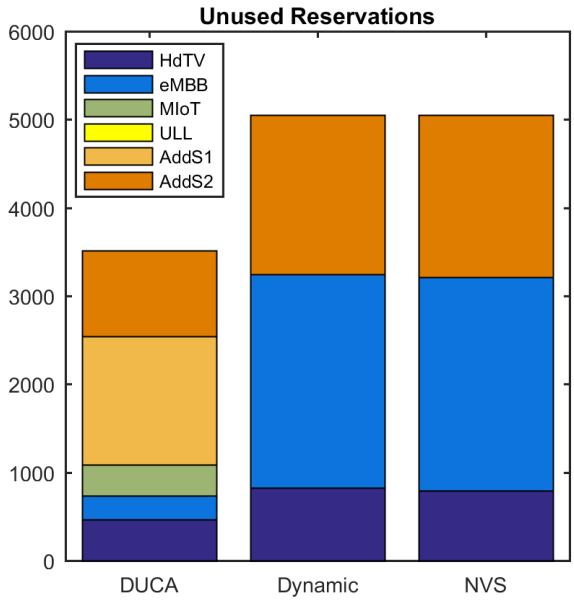

Figure 4.29: Unused Reservations

CHAPTER 5

CONCLUSION

In Radio Access Networks, sharing a virtualized network to serve diverse user demands through network slicing is the common used approach. In order to meet the demands of these slices in an efficient way, this paper proposes DUCA (Dynamic User Count Aware) network slicing scheme as the resource allocation approach. The resource allocation problem is considered for a multi-BS network. The objective of DUCA is to make reservations for each slice on every BS, so that large number of granted user requests with high resource utilization results is achieved on each BS.

Slices with high, moderate, low and very low data rate requested users have been modeled to be on the same network to evaluate the performance of DUCA. Five simulation experiments are conducted to see the performance of DUCA compared to NVS and Dynamic resource allocation schemes. The first experiment included fix number of user requests to stay on the network for each slice. The second experiment is conducted to see the performance of allocation schemes on a slowly congested network. The third experiment is simulated to see the performances on a realistic model where user requests come randomly with a life time. The forth experiment is simulated to see how mobility of fixed number of users affect the performances of allocation schemes.The fifth experiment is conducted to see the effect of additional slices and very high number of users.

On all cases, DUCA outperforms NVS and Dynamic in terms of number of granted users and used resource reservations. This is achieved by favoring slices with high number of users even though they have low data rate requests. In total achieved throughput,in spite of making higher reservations for low data rate requested users, DUCA maintained similar results with NVS and Dynamic and did not under-perform.

In addition to all, DUCA preserves the isolation among slices even a congestion with a drastic increase on the number of user requests on a slice occurs.

To realize the reservations in BSs, two intra-slice allocation methods 0-1 knapsack solver and algorithm 2 is suggested. Even though, they performed similar results in terms of number of granted users and achieved throughput, the difference in execution times put Algorithm 2 forward as the final solution.

To conclude, DUCA provides an efficient resource allocation reservation compared to NVS and Dynamic by introducing user counts of slices as a metric. In a multi-tenant networks with diverse data rate requests and user populations, DUCA is proven to be a fair, profitable and sustainable approach.

REFERENCES

[1] G. O. B. G. L. G. S. K. Xiong, S. S. R. Adolphe, "Dynamic resource provisioning and resource customization for mixed traffics in virtualized radio access network," *IEEE Access*, vol. 7, pp. 115440–115452, 8 2019.

[2] H. Z. R. Kokku, R. Mahindra and S. Rangarajan, "Nvs: a substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Transactions on Networking*, vol. 20, pp. 1333–1346, 10 2012.

[3] B. Murara, "Imt-2020 network high level requirements, how african countries can cope," *IMT-2020*, 2020.

[4] X. C.-P. K. Samdanis and V. Sciancalepore, "From network sharing to multitenancy: The 5g network slice broker," *IEEE Communications Magazine*, vol. 54, pp. 32–39, 7 2016.

[5] N. A. Qiang Duan and M. Toy, "Software-defined network virtualization: An architectural framework for integrating sdn and nfv for service provisioning in future networks," *IEEE Network*, vol. 30, pp. 10–16, 10 2016.

[6] J. G. N. B. F. D. T. R. Mijumbi, J. Serrat and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.

[7] P. E. V. C. E. R. S. A. D. Kreutz, F. M. V. Ramos and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, pp. 14–76, 1 2015.

[8] S. S. Faqir Zarrar Yousaf, Michael Bredel and F. Schneider, "Nfv and sdn—key technology enablers for 5g networks," *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS*, vol. 35, pp. 2468–2478, 11 2017.

[9] K. S. V. Sciancalepore, M. G. X. Costa-Perez, D. Bega, and A. Banchs, "Mobile

traffic forecasting for maximizing 5g network slicing resource utilization," *IEEE INFOCOM 2017*, 05 2017.

[10] G. d. V. S. J. B. J. Zheng, P. Caballero and A. Banchs, "Statistical multiplexing and traffic shaping games for network slicing," *IEEE/ACM Transactions on Networking (TON)*, vol. 26, pp. 1–13, 12 2018.

[11] M. C. M. Jiang and T. Mahmoodi, "Network slicing management prioritization in 5g mobile systems," *European Wireless 2016*, pp. 1–6, 2016.

[12] M. H. Y. Xiao and M. Krunz, "Distributed resource allocation for network slicing over licensed and unlicensed bands," *IEEE JSAC*, vol. 36, pp. 2260–2274, 10 2018.

[13] G. O. B. D. A.-M. G. L. G Sun, K Xiong and W. Jiang, "Autonomous resource provisioning and resource customization for mixed traffics in virtualized radio access network," *IEEE Syst. J.*, vol. 13, pp. 2454–2465, 09 2019.

[14] A. B. V. S. D. Bega, M. Gramaglia and X. Costa-Perez, "A machine learning approach to 5g infrastructure market optimization," *IEEE Transactions on Mobile Computing*, vol. 19, pp. 498–512, 3 2020.

[15] P. M.-J. V. P. Kafle, Y. Fukushima and T. Miyazawa, "Consideration on automation of 5g network slicing with machine learning," *2018 ITU Kaleidoscope: Machine Learning for a 5G Future (ITU K)*, pp. 1–8, 2018.

[16] J. Koo, V. Mendiratta, M. R. Rahman, and A. Walid, "Deep reinforcement learning for network slicing with heterogeneous resource requirements and time varying traffic dynamics," 08 2019.

[17] A. Thantharate, R. Paropkari, V. Walunj, and C. Beard, "Deepslice: A deep learning approach towards an efficient and reliable network slicing in 5g networks," 10 2019.

[18] T. G.-R. M. X. Costa-Perez, J. Swetina and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Communications Magazine*, vol. 51, pp. 27–35, 07 2013.

[19] *3GPP. 3rd generation partnership project; technical specification group services and system aspects; network sharing; architecture and functional description (release 121T).* 2011.

[20] D. L.-J. J. R.-M. J. L. J. Ordonez-Lucena, P. Ameigeiras and J. Folgueira, "Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, pp. 80–87, 05 2017.

[21] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser ofdm system using dynamic subchannel allocation," *2000 IEEE 51st Vehicular Technology Conference Proceedings*, vol. 2, pp. 1085–1089, 2000.

[22] A. S. A. Moubayed and H. Lutfiyya, "Wireless resource virtualization with device-to-device communication underlaying lte network," *IEEE Transactions on Broadcasting*, vol. 61, pp. 734–740, 12 2015.

[23] A. G. Mahmoud I. Kamel, Long Bao Le, "Lte wireless network virtualization: Dynamic slicing via flexible scheduling," *IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pp. 1–5, 09 2014.

[24] H. Z. S. R. Rajesh Mahindra, Mohammad A Khojastepour, "Radio access network sharing in cellular networks," *21st IEEE International Conference on Network Protocols (ICNP)*, pp. 1–10, 2013.

[25] D. N. N. H. M. Weingartner, "Methods for the solution of the multi-dimensional 0/1 knapsack problem," *Operations Research*, vol. 15, 5 1967.

[26] A. L. T. L. M.-P. P. F. Boccardi, R. W. Heath, "Five disruptive technology directions for 5g," *IEEE Commun. Mag.*, vol. 53, pp. 74–80, 2 2014.

[27] O. S. P. Popovski, K. F. Trillingsgaard and G. Durisi, "5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view," *IEEE Access*, vol. 6, 9 2018.

[28] T. S. C. W. A. A. Pramanik, B. Choudhury and J. Mehedi, "Simulative study of random waypoint mobility model for mobile ad hoc networks," *2015 GCCT*, pp. 112–116, 4 2015.