

Low-complexity Dynamic Resource Slicing for Mixed Traffics in Virtualized Radio Access Network

Guolin Sun, Sebakara Samuel Rene Adolphe, Daniel Ayepah-Mensah

School of Computer Science and Engineering
University of Electronic Science and Technology of China
guolin.sun@uestc.edu.cn

Abstract— In this paper, we investigate dynamic network slicing strategies with mixed traffics and multiple base stations in virtualized radio access network. Considering versatile user's QoS requirements on delay and rate, low complexity resource slicing algorithms and shape-based heuristic algorithm for user resource customization are devised in order to improve resource utilization and QoS satisfaction. To validate the advantage, a system-level simulation based on the 5G air interface design is conducted. Results show performances of the proposed algorithm outperform existing benchmarks.

Keywords— Network slicing, resource provisioning, resource allocation, user association, 5G;

I. INTRODUCTION

5G networks are envisioned to predispose flexible edge-to-core infrastructure to offer diverse applications. Convergence of software-defined networking and virtualization is a promising approach to support such dynamic networks [1]. Satisfying customers' demands in terms of good quality of service (QoS) is very challenging to service providers for *mixed traffics*. 5G services improvement over 4G can be classified into four main use cases, *i*) Enhanced mobile broadband (eMBB), which is expected to attain the speeds up to 20Gbps per user equipment (UE), such as new RAT, mmWave, massive MIMO, etc. *ii*) Ultra-low latency considering the end-to-end delivery delay, which needs few micro-seconds for the mission-critical IoT services. *iii*) The scenario of massive connectivity probably introduces hundreds of millions of Internet of Things (IoT) devices like static sensors. *iv*) High definition TV and high definition video streaming with elephant flows and critical time delays such as virtual reality. These scenarios when present in the same network can be termed as *mixed traffics*.

Recently, *network virtualization (NV)* has attracted a lot of interest from both academia and industry [2]. NV abstracts resources and services in a way that different services can share the physical hardware. The combination of NV and SDN technologies leads to achieve two complementary goals of 5G, *i*) a service-oriented architecture that is programmable and extensible in terms of infrastructure, network services and mobile applications; *ii*) network slicing that decouples operations of virtual networks on top of physical infrastructure with slice isolation and customized radio resource allocation to users.

Most of the current resource virtualization and provisioning still present various limitations [3]–[5]. First, the minimum and

maximum resource provisioning for multiple slices are based on the generalization of slice requirements only, which leads to the dissatisfaction of user's QoS in case of a limited resource in a network. Second, existing algorithms consider slice resource provisioning aiming to achieve QoS satisfaction of a slice on one intended base station (BS). However, the re-association provisioning is not considered, which results in network inefficiency. Third, most of the existing works assume network slices to be *homogenous* and do not tackle resource slicing specifically for *mixed traffics* in a network as one of the main pillars for users' QoS satisfaction. This degrades the effectiveness of slice resource provisioning.

In order to address resource slicing problem in a dynamic wireless environment, we propose a novel scheme that fairly reserves slice's resources among BSs based on slice's weights and preferences on those BSs. The admitted UEs could be re-associated or rescheduled with BSs which could provide high channel quality with consideration of the UE's priority. Moreover, customization of heterogeneous traffics is a challenging job for physical resource allocation to balance resource utilization and users' QoS. Therefore, the motivation behind this paper also includes customizing logical view of physical resources for intra-slice resource allocation with logically shaped resource units (LSRU), which is defined as a group of adjacent resource blocks (RBs) in terms of time-frequency. Lastly, we decouple our work into dynamic resource provisioning per slice at BS level and shape based intra-slice resource allocation. The motivations behind this work are based on the following:

- In this paper, we propose a generic framework of dynamic slice resource provisioning and customized intra-slice resource allocation in virtualized radio access network for heterogeneous traffics.
- Considering the slice's preference and slice's weight at different BSs, we develop a low-complexity heuristic algorithm to perform fair resource provisioning for the versatile QoS characteristics of *mixed traffics* and slices that provisions for flexible re-association is also included.
- We propose a shape-based resource allocation algorithm for users where shape (group of LSRUs) is defined in terms of transmit rate and time delay requirement. This low-complexity algorithm can be easily implemented in any OFDM-based RAN.

The rest of this paper is organized as follows: Section II presents system models. In Section III, we present the problem formulation and the low-complexity heuristic algorithm. Then,

simulation results are discussed in Section IV. Finally, we make a conclusion on this study in Section V.

II. SYSTEM MODEL

A. Network Model

Let us consider the cellular network of $k \in \{1, 2, \dots, K\}$ as a set of BSs and let $s \in \{1, 2, \dots, S\}$ be a set of slices in a virtual network. Each slice s has a set of UEs, $u \in \{1, 2, \dots, U_s\}$. The slices are assigned priorities due to the nature of the resource provisioning scheme. We denote the priority of a slice by η_s . The *QoS constrained request* is $Rqs(R_s, Td_s)$, where Td_s denotes affordable time delay in terms of the inter-departure time of consecutive packets and R_s is the sum of requested transmit rate R_u of *user* (u), i.e. $R_s = \sum_{u=1}^{U_s} R_u$.

In this paper, we discuss resource virtualization on RB level, instead of time slots and bandwidth [6]. The physical radio resources from infrastructure providers (InPs), represented in a form of bandwidth B , are divided into M RBs in the frequency domain, each of which has bandwidth B_m . A scheduling time frame T is also subdivided into t sub-frames. Each of them has t_l length in the time domain. Hence, the unit of resources is defined as one RB per sub-frame (RB_m^t). Each $BS(k)$ has a total transmit power of P_t and $p_{uk}^{(t,m)}$ denotes the power allocated to *user* (u) (i.e., user u to the $BS(k)$) in the RB m at time t .

We further assume that the base station receives perfect channel gain information from all users belonging to service providers (SPs) and $g_{uk}^{(t,m)}$ is the channel gain for *user* (u) on an RB at time t . The noise spectral density is denoted by N_o . Therefore, data rate associated with each RB_m^t for a given *user* (u) is defined as follows:

$$RB_{uk}^{(t,m)} = \frac{B_m}{T} \log_2 \left(1 + \frac{p_{uk}^{(t,m)} g_{uk}^{(t,m)}}{N_o B_m} \right), \quad (1)$$

In order to guarantee the delay desired by each *user* (u) in slice s , we expect the inter-departure time of consecutive packets for *user* (u) must be less than Td_s to achieve smoothness of traffic flows with a common shape-based resource allocation algorithm for mixed traffics. We divide the time-axis into *scheduling frames*, each consisting of T sub-frames. The sub-frame length is t_l . Thus, n_u^h is the required minimum number of time slots to satisfy its delay demand, as shown in equation (2). n_u^v is the number of RB_m^t needed by *user* (u) in a single time slot. Therefore, we define n_u^h and n_u^v as:

$$n_u^h = \left\lceil \frac{T \times t_l}{Td_s} \right\rceil, \quad (2)$$

$$n_u^v = \left\lceil \frac{R_u^T}{n_u^h \times RB_{uk}^{(t,m)}} \right\rceil, \quad (3)$$

where R_u^T is the user's requested throughput within a single scheduling time frame T . Without loss of generality, the user has to be allocated a sub-channel defined by a LSRU of width n_u^h and height n_u^v . Thus, the achievable data rate r_{um} per this *user* (u) will be calculated as,

$$r_{uk} = n_u^h \times n_u^v \times RB_{uk}^{(t,m)}, \quad (4)$$

To set up a delay model for traffics with queuing model, at first, we make the following assumptions: (1) The inter-arrival times between the packets for *user* (u) are independent and exponentially distributed with mean $1/\lambda_u$ seconds, where λ_u is the packet arriving rate for *user* (u); (2) The lengths of the packets for slices are independent but for *users* (u) are common in slice (s) as L_u bits; The traffic of service aimed at *user* (u) is then exponentially distributed with mean equal to $t_{uk} = L_u / r_{uk} = 1/\gamma_{uk}$, where γ_{uk} denotes the normalized achievable rate with respect to the average packet length. Clearly, the queue of slice for the traffic to *user* (u) is an M/M/1 queue. Accordingly, the average delay experienced by a packet in the queue to *user* (u) is

$$\tau_u = \frac{1}{\frac{1}{t_{uk}} - \lambda_u} = \frac{1}{\gamma_{uk} - \lambda_u} \quad (5)$$

B. Utility Model

We classify user requests status into three main categories; Let $Sat_U_{(s,k)}$ be a set of satisfied users of slice (s) at a $BS(k)$, $Del_U_{(s,k)}$ a set of delayed users of slice (s) at a $BS(k)$, and $Rej_U_{(s,k)}$ a set of rejected users of slice (s) at a $BS(k)$. We can compute the utility metrics: slice's satisfaction ratio β_s and slice's delaying ratio μ_s , respectively as;

$$\beta_s = \sum_{k=1}^K \frac{|Sat_U_{(s,k)}|}{|U_{(s,k)}|}, \quad (6)$$

$$\mu_s = \sum_{k=1}^K \frac{|Del_U_{(s,k)}| + |Rej_U_{(s,k)}|}{|U_{(s,k)}|}, \quad (7)$$

where $|U_{(s,k)}|$ is the user population for slice (s) at $BS(k)$ is the set of users that has identified $BS(k)$ as their preferred BS i.e. the BS with the highest rank by the user. Finally we compute the system level satisfaction ratio (β) and delaying ratio (μ) respectively as follow;

$$\beta = \frac{1}{S} \sum_{s=1}^S \beta_s, \quad (8)$$

$$\mu = \frac{1}{S} \sum_{s=1}^S \mu_s, \quad (9)$$

III. PROBLEM FORMULATION AND ALGORITHM

A. Dynamic slice resource provisioning scheme

By defining Ω_k as the number of occupied RB resources at a $BS(k)$, and Π_k as number of RBs successfully allocated or provisioned on that $BS(k)$, we can define the system's resource efficiency (θ) as follow;

$$\theta = \frac{\sum_{k=1}^K \Omega_k}{\sum_{k=1}^K \Pi_k}, \quad (10)$$

Based on user requests in slice queues, each *user* (u) will identify a potential list of BSs to join. This is done by ranking each BS according to its average *achievable data rate per RB*, $RB_{uk}^{(t,m)}$ and denote the rank of $BS(k)$ by *user* (u) as $\mathcal{R}_{u,k}$. In this paper, we define logical resource as the portion of resources assigned to a slice by a BS. We also aggregate the slice level BS rank ($\bar{r}_{sk}$) which is the average of ranks given to $BS(k)$ by users of a slice (s) as

$$\bar{r}_{sk} = \frac{\sum_{u=1}^{|U_s|} \delta_{u,k}}{|U_s|}, \quad \forall u \in \{1, 2, \dots, U_s\}, \quad (11)$$

The rank based ratio $rnk_rat_{k,s}$ for each slice at a $BS(k)$ is defined as

$$rnk_rat_{k,s} = \frac{\bar{r}_{sk}}{\sum_{s=1}^S \bar{r}_{sk}}, \quad \forall s \in \{1, 2, \dots, S\}, \quad (12)$$

and the slice's weight, ϖ_{sk} as the sum of requested transmit rate per user R_u at $BS(k)$ as

$$\varpi_{sk} = \sum_{u=1}^{|U_{sk}|} R_u, \quad \forall u \in U_{sk}, \quad (13)$$

Each slice depends on the ratio of the weights of other slices sharing that BS ($W_rat_{k,s}$). We formulate $W_rat_{k,s}$ as

$$W_rat_{k,s} = \frac{\varpi_{sk}}{\sum_{s=1}^S \varpi_{sk}}, \quad \forall s \in \{1, 2, \dots, S\}, \quad (14)$$

The overall resource virtualization of our scheme is derived from equations (12) and (14). The resources $Rsc_{k,s}$ required by slice s can be mathematically formulated as

$$Rsc_{k,s} = \frac{(rnk_rat_{k,s} \times W_rat_{k,s})}{\sum_{s=1}^S (rnk_rat_{k,s} \times W_rat_{k,s})} \times \mathbb{L}_k, \quad \forall s \in \{1, 2, \dots, S\}, \quad (15)$$

where $\mathbb{L}_k$ is total resources of $BS(k)$ and $Rsc_{k,s}$ is the amount of resources of a slice(s) at $BS(k)$.

Firstly, all users can be classified into slices based on their requested service type. For each slice, we aggregate the slice level BS rank ($\bar{r}_{sk}$) based on the rank $\delta_{u,k}$ of $BS(k)$ from user (u) by equation (11). We can also aggregate ϖ_{sk} which is slice's weight at a $BS(k)$ according to the user's requested throughput R_u by equation (13). The rank based ratio $rnk_rat_{k,s}$ is derived from the slice level BS rank ($\bar{r}_{sk}$) by equation (12). We can also get the slice's weight based ratio $W_rat_{k,s}$ based on its weight ϖ_{sk} at a $BS(k)$ by equation (14). Thus, for each slice at BS k , we get two parameters ($W_rat_{k,s}, rnk_rat_{k,s}$). Using these parameters, we find the amount of resource to be given to slice(s) at a $BS(k)$ by equation (15). Here we divide the slicing problem into two and solve with a two-step low-complexity heuristic algorithm. In step 1, we initialize all parameters and receive all users' requests in a predefined time window. The requests are classified and the waiting queue updated for each slice. In step 2, we aggregate resource portion to be allocated for each slice at each BS. The user and slice BS preference ratio are determined by the rank of the *average achievable data rate*.

B. Shape-based resource allocation and association control

After allocating resources to each slice at a given BS, the controller performs user association within the slices and resource allocation among BSs. The algorithm considers user-level shaped virtual resources in terms of frequency domain and time domain blocks, which is defined as LSRU. The achievable data rate r_{uk} for user u on sub-channel m can be formulated as follow;

$$r_{uk} = \sum_1^{(n_u^h \times n_u^v) \frac{B_m}{T}} \log_2 \left(1 + \frac{p_{uk}^{(t,m)} \times g_{uk}^{(t,m)}}{N_0 B_m} \right), \quad n_u^h \times t_l \leq T, l \in \{1, 2, \dots, T\} / t_l, \quad (16)$$

Therefore, r_{sk} which denotes the achievable data rate for the whole slice (s) at a certain $BS(k)$, will be computed by;

$$\Gamma_{sk} = \sum_{u=1}^{U_s} \Gamma_{uk} \cdot x_{uk}, \quad (17)$$

where x_{uk} is the association indicator between *user*(u) and base station k . In addition, radio bandwidth on each $BS(k)$ cannot exceed its backhaul bandwidth Ψ_k .

If the remaining RBs at a $BS(k)$ cannot satisfy the QoS requirement in R_s , the controller attempts to combine the remaining resources from other BSs in order to meet the required resources $\Psi_s = n \text{ RBs}$ that is able to satisfy the demand of $Rqst$. The total throughput r_k in front-haul for BS k is

$$\Gamma_k = \sum_{s=1}^S \Gamma_{sk}. \quad (18)$$

As a result, the request Rq_s will be allocated with the appropriate shaped resources from the association x_{uk} and the minimum number of RBs for providing the requested resources against the constrained QoS. Now, we formulate the user association and resource allocation as a NP-complete 0-1 multiple knapsack problem:

$$\min \quad \sum_{k=1}^K \sum_{s=1}^S \sum_{u=1}^{U_s} (n_u^h \times n_u^v) \cdot O_{us} \cdot x_{uk} \quad (19)$$

s.t.

$$x_{uk} \cdot (x_{uk} - 1) = 0 \quad \forall k \in \{1, 2, \dots, K\}, \forall s \in \{1, 2, \dots, S\}, \quad (20)$$

$$\sum_{u=1}^{|U_{sk}|} (n_u^h \times n_u^v) \cdot x_{uk} \leq Rsc_{k,s}, \quad \forall k \in \{1, 2, \dots, K\}, \quad (21)$$

$$\sum_{k=1}^K O_{us} \cdot x_{uk} = 1, \quad \forall u \in \{1, 2, \dots, U_s\}, \quad (22)$$

$$\sum_{s=1}^S \Gamma_{sk} \leq \Psi_k, \quad \forall k \in \{1, 2, \dots, K\}, \quad (23)$$

$$\tau_u = \frac{1}{x_{uk} \cdot \gamma_{uk} - \lambda_u} \leq Td_u, \quad \forall u \in \{1, 2, \dots, U_s\}, \quad (24)$$

$$r_{uk} \geq R_u, \quad \forall u \in \{1, 2, \dots, U_s\}, \quad (25)$$

where x_{uk} is a binary association indicator that takes value 1 if user from slice(s) had been granted the resource from $BS(k)$ and otherwise it takes value 0; n_u^h and n_u^v are the two parameters in the defined shape of LSRU for user u from *QoS constrained request* $Rqs(R_u, Td_u)$ based on the formula (2) and (3); $(n_u^h \times n_u^v)$ is the area of LSRU for user u , also as number of RBs, which is guaranteed non-overlapped for different users by user scheduling and resource allocation. O_{us} is the association indicator to indicate that *user*(u) is associated with *slice*(s); Constraint (21) shows that all users of *slice* (s) to be allocated to the $BS(k)$ must be satisfied by resources of *slice*(s) at $BS(k)$, and (22) enforces that *user*(u) must be allocated a channel from only one BS. Here, we assume strict resource isolation among slices can be achieved on each BS as $Rsc_{k,s}$. Constraint (23) guarantees that backhaul capacity Ψ_k of a serving $BS(k)$, can afford the achievable data rate of its associated users from all slices. Constraint (24) and Constraint (25) can guarantee the *QoS constrained request* $Rqs(R_u, Td_u)$ for *user*(u).

IV. PERFORMANCE EVALUATION

A. Scenario configuration and assumptions

The mobile network scenario is considered based on 4G network standards and is summarized in Table 1. In a given area of 700m*700m, 4 BSs are uniformly scattered with coverage of about 150meters. The distance between two BSs is 120 meters. The Path Loss (PL) model is defined as follow; PL (dB) = 20

$\log_{10}(d) + 20 \log_{10}(f) - 27.55$ with d (in meters) and f (in MHz) denoting user-BS distance and channel frequency respectively. Results are averaged over all users and over 50 different runs with 95% confidence level.

TABLE 1. SYSTEM PARAMETERS AND SETTINGS

Parameters	Settings
The number of BS	4
The number of slice	4
System bandwidth	5MHz
Transmitting power	30dBm
Channel gain	To be generated via path loss model
Sub-frame duration	1ms
Frame length	10ms
Number of sub-channels per BS	25
Noise power per sub-channel	18dBm
The number of users	[445, 508, 571, 635, 698, 761, 825, 888, 951] light load:698, heavy load:1015
The packet arriving rate per user	[UEb: 10, Hdtv:10, MIoT:10, ULL:100](packet per second); ULL: uniform, Others: exponential
The minimum rate demand of slice	[UEb:200,Hdtv:1000, MIoT:150, ULL:10] (kbps)
The maximum delay demand of slice	[UEb:100, Hdtv:120, MIoT:105, ULL:10] (ms)
The packet size	[UEb:400, Hdtv:1000, MIoT:300, ULL:120] (bits)
The user distribution	Uniform

B. Performance analysis on system satisfaction

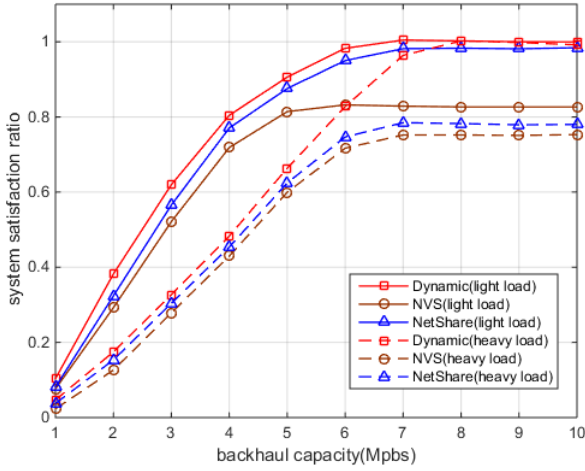


Figure 1. System satisfaction ratio over Backhaul capacity

The proposed algorithm is compared with existing slicing schemes NVS [4] and Netshare [5]. As presented in Figure 1, we can find that the proposed heuristic algorithm achieves the best performance on the system satisfaction ratio, in terms of maximizing the number of served users, and cannot easily be affected by resource limitations.

C. Performance analysis on resource utilization

On the average slice resource utilization ratio, both of two dynamic slice resource allocation schemes outperform static slice resource allocation, NVS, in both light-load and heavy-load scenarios as illustrated in Figure 2. Not only does the proposed scheme dynamically distribute the resource of slices among BSs, but also can control the user association and the proper resource allocation. Under the RMSE rule, the relative

gain on the resource utilization ratio is 14.0% and 1.21% respectively.

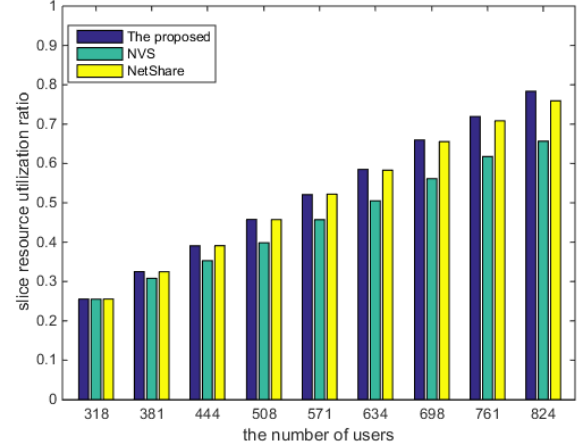


Figure 2. Slice resource utilization ratio over user population

V. CONCLUSION

In this paper, we proposed a novel scheme of slice resource provisions for the mixed traffics in virtualized radio access network, which dynamically provisions and allocates proper portion of resources to a certain slice at each BS. Then a shape-based heuristic algorithm was proposed to handle the intra-slice resource allocation and user association control. Results showed that the proposed one outperforms NVS and Netshare about 16.64% and 6.86%, respectively on the slice satisfaction, while on the resource utilization ratio about 14% and 1.21% respectively in uniform user distribution scenarios.

ACKNOWLEDGMENT

This work is supported by National Natural Science Research Foundation of China, Grant no. 61771098, the Fundamental Research Funds for the Central Universities under grant no. ZYGX2018J049 and the ZTE Innovation Research Fund for Universities Program 2016.

REFERENCES

- [1] Mahsa Derakhshani, Saeedeh Parsaeefard, Tho Le-Ngoc, Alberto Leon-Garcia, "Leveraging synergy of SDWN and multi-layer resource management for 5G networks," *IET Networks*, Early access, Feb. 2018.
- [2] M. Kalil, A. Shami, A. Al-Dweik, "Wireless resources virtualization for Cloud Radio Access Networks (C-RAN)," *In IEEE Journal on Computer Science*, vol.120, pp. 3366-3379, 2015.
- [3] P. Caballero Garces, X. costa Perez, K. Samdanis, A. Banchs, "A cell slicing controller for virtualized multi-tenant mobile networks," in *2015 IEEE 81st Vehicular Technology Conf. (VTC spring)*, pp.1-6, May 2015.
- [4] R. Kokku, R. Mahindra, H. Zhang, S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1333-1346, Oct. 2012.
- [5] R. Mahindra, M. Khojastepour, H. Zhang, S. Rangarajan, "Radio access network sharing in cellular networks," *2013 21st IEEE International Conference on Network Protocols (ICNP)*, pp. 1-10, 2013.
- [6] Richart, Matias, et al. "Resource Slicing in Virtual Wireless Networks: A Survey," *IEEE Transactions on Network & Service Management* vol.13, no.3, pp. 462-476, 2016.