Radio Resource Management in context of Network Slicing: What is Missing in Existing Mechanisms?

Behnam Khodapanah*, Ahmad Awada†, Ingo Viering‡, Jobin Francis*, Meryem Simsek§, Gerhard P. Fettweis*
*Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Germany Email:{behnam.khodapanah, jobin.francis, gerhard.fettweis}@tu-dresden.de
†Nokia Bell Labs, Munich, Germany; Email: ahmad.awada@nokia-bell-labs.com
‡Nomor Research GmbH, Munich, Germany; Email: viering@nomor.de
§International Computer Science Institute, Berkeley, USA; Email: simsek@icsi.berkeley.edu

Abstract—Fifth generation (5G) of mobile networks are expected to serve multiple heterogeneous use cases. These use cases are extremely diverse in terms of service requirements and bundling them in a single monolithic network is a challenge. Network slicing is identified as one of the main enablers of 5G systems, where multiple logical End-to-End (E2E) networks share the resources of a single physical network. Radio Resource Management (RRM) in a sliced network should be able to simultaneously fulfill the required services of slices, dynamically share the network and assure the independence of slices so that slices cannot affect each other negatively. In this paper, we study the existing mechanisms that provide similar features in legacy networks and demonstrate the contributions and shortcomings of such existing RRM mechanisms in a sliced network. Thereafter, we argue the need for a new entity that will complement the existing RRM mechanisms to be slice-aware. With the aid of system-level simulations, we compare different slicing schemes and illustrate the drawbacks of legacy networks in fulfilling the objectives of a fully sliced network. Moreover, we illustrate the capabilities of the slice-aware RRM in steering the network's Key Performance Indicators (KPIs).

Index Terms—Network Slicing, Slice Management, Radio Resource Management, Dynamic Resource Sharing, Slice Isolation, 5G

I. Introduction

The fifth generation (5G) cellular networks are envisioned to simultaneously support multiple heterogeneous services, which include enhanced mobile broadband, mission-critical services, and massive connectivity to enable Internet-of-Things (IoT) [1]. Deploying separate service-specific network infrastructures is clearly not feasible. On the other hand, the demanding and diverse nature of the requirements of these services make their coexistence in a single network challenging.

Network slicing is a flexible and scalable solution for efficiently sharing the resources of network infrastructure among multiple services and enables a smooth integration to the new services. In a sliced network, several End-to-End (E2E) logical networks, referred to as slices, share the resources of a single physical network. This improves the resource and energy efficiency by exploiting statistical multiplexing gains. Further, the network operator needs to deploy only the functionalities required to enable the service, which allows for cost savings. Network slicing enables the tenants of the network to specify their service requirements in a Service Level Agreement (SLA) and the network operator should instantiate the appropriate network slice to meet these SLAs [2]. At the same time, since the slices share the same physical infrastructure, they must be protected from each other so that the dynamics of one slice do not adversely affect other slices.

Since the slices are conceived to be E2E networks, both the core network (CN) and the Radio Access Network (RAN) needs to be sliced [2]. Software-Defined Network (SDN) and Network Function Virtualization (NFV) enable the slicing of the CN. The architectural design aspects of the instantiation and deployment of network functions are studied in [3]. In the RAN, slicing involves sharing the physical radio resources such as base stations, bandwidth, and transmission time interval. Thus, radio resource management (RRM) is critical in ensuring that the SLAs are met while ensuring isolation between the slices.

To a certain degree, the mentioned objectives have been separately addressed in the previous generations of the mobile networks. For example, the Long-Term Evolution (LTE) systems utilize the Quality-of-Service (QoS) class indicators that guarantee certain service requirements to a class of users [4]. In mobile virtual network operators (MVNOs), the tenants share the radio resources via a fixed resource sharing agreement. However, it is unclear whether the RRM mechanisms in the existing cellular networks are sufficient for a fully sliced network.

In this paper, we first review the existing RRM mechanisms in the context of network slicing. We focus on two main functionalities of the RRM mechanism, namely, Admission Control (AC) and Packet Scheduler (PS) as they determine the number of users in the network and the amount of radio resources allocated to each user, respectively. Additionally, we review the current approaches to network sharing and point out the difference between dynamic and static sharing. We argue why the current RRM mechanisms in the legacy networks are not adequate in achieving the objectives of a fully sliced network. Having reviewed the current capabilities and observing the deficiencies of legacy architectures, the concept of SLA mapping layer is studied [5]. This entity enables dynamic sharing of the resource and by properly controlling the long-term behavior of AC and PS the slice protection can be achieved. This new entity complements the existing

978-1-5386-7646-2/19/$31.00 ©2019 IEEE

mechanisms such that network slicing can be adopted in RAN. With the help of system-level simulations, we illustrate the weaknesses of existing mechanisms and show why the proposed entity is capable of making the RAN slice-aware. Besides, the impact of different control parameters of the AC and PS on the Key Performance Indicators (KPIs) of different slices are studied. These results show that the SLA mapping layer is capable of controlling the KPIs so that it matches the targets defined in the SLA. Finally, we conclude the paper and discuss the outlook.

II. Existing and Missing RRM Mechanisms

In this section, we review the existing RRM mechanisms to determine if they fit to the general objectives of RRM in a sliced network. Objectives similar to network slicing have been under investigation for years in separate fields of study. Firstly, we review the QoS-aware systems that guarantee certain services to the users to determine whether they are able to fulfill the SLAs of slices. By reviewing QoSaware PS and AC, we identify the crucial mechanisms for enforcing the service guarantees of the users in the network. Thereafter, we review the previous research on the paradigm of sharing the mobile network with multiple tenants. We point out the similarities and differences between such shared networks and sliced networks. Then, we discuss why these existing mechanisms are not fully adequate in achieving the RRM objectives of a sliced network. We also review the suggestions from the literature in adapting them to network slicing. Finally, to overcome the shortcomings of previously mentioned approaches, we introduce a network entity, which was introduced in our previous work [5], that steers the RRM in the direction of network slicing objectives.

A. QoS-Aware Systems

AC and PS are vital components in RRM that enable the network to serve its users to meet their QoS requirements. The design of these components will have a great impact on the overall performance of the network. Operators can opt between different designs choices, based on their targets and objective. Yet, wireless medium is highly dynamic and is shared between multiple users. Because of this dynamism and interference from other users, unlike the wired medium, guaranteeing QoS via the wireless medium has been proven to be a challenging task [6].

Today’s state of the art QoS-aware PSs are designed to maximize the utility, which attempts to guarantee QoS for the maximum possible number of users. [7]. LTE standardization has introduced several QoS Class Identifiers (QCIs) to support different requirements, such as Guaranteed-Bit-Rate (GBR) and non-Guaranteed-Bit-Rate (non-GBR) services. QoS-unaware PSs allocate the Physical Resource Blocks (PRBs) based on different buffer status, CQI, transmission queues, allocation history, etc. A QoS-aware PSs also considers QoS priorities in resource allocation decision making. Such priorities steer the scheduling so that the required services (e.g. throughput, delay, VoIP support, etc.) are provided.

QoS-awareness in today’s systems guarantees the services to the users individually, i.e., the service fulfillment is confined to a single user in a certain location and time and cannot guarantee the SLA fulfillment for the slices. In sliced networks, the service requirements of a collection of users, belonging to a certain slice, should be provisioned over a longer timescale and over a multitude of cells. This indicates that the RRM mechanisms should reflect not only the requirements of each user but also the overall requirement of an entire slice. Hence, a simple QoS-aware scheduler inadequate to meet SLAs in a sliced network. To be applicable to the sliced-network, a two-level scheduling has been proposed in [8] for virtualizing and sharing the radio resources. Such hierarchical methods not only consider users’ QoS requirements but also the slices’ requirements. However, it is not clear whether inter-slice influences are under control, when a slice is introducing too much traffic.

The design of AC scheme that is suitable for sliced networks is proposed [9], where the authors discuss cutoff thresholds for AC with the presence of multiple service classes. In [10], the authors have introduced thinning schemes to enable the AC to introduce prioritization between multiple users with different service requirements. In [3] and [11], two-level AC has been proposed that can increase the Quality-of-Experience (QoE) and overall efficiency of the systems. The hierarchical nature of these proposals helps the system to prioritize slices and increase the spectral efficiency. Nonetheless, fulfillment of SLAs is not the focus of these works and increasing the QoE and spectral efficiency are not the main objectives in sliced networks.

B. Network Sharing

The concept of sharing the mobile network between multiple virtual tenants had already gained attention in the legacy LTE systems. The cost savings is the main drive for sharing the network infrastructure and radio resources. This network virtualization can be realized at different layers and degrees, from virtualizing the CN to physical layer at base stations [7]. However, in most of such architectures, the radio resources are shared via static agreements, which inhibits flexibility and lowers the spectral efficiency.

Several suggestions have been made in the context of virtualized networks to enable dynamic sharing. Authors in [12] propose efficient schemes that aim to maximize the utilization of resources. In [13], the authors have offered a multi-operator scheduling that allows a dynamic sharing of radio resources and have shown that dynamic resource guarantees can improve the spectral efficiency and utilization of the network.

The network sharing proposals in legacy network architectures, in the future generations of mobile networks, the SLA will demand services directly and only provisioning the resources, either statically or dynamically, will not be sufficient [2]. Besides, in the design of slice-aware RRM, the protection mechanism for slices is a vital feature of the network.

C. SLA Mapping Layer

Although previously mentioned concepts in the legacy mobile networks separately reflect the objectives of network slicing, it is not clear how all of them can be structured together to form a fully sliced network. To overcome the mentioned drawbacks of the existing mechanisms, we introduce a new network entity that orchestrates the RRM of the sliced network in long-term and over a multitude of cells. This management/orchestration is performed via tuning the slicespecific weights and thresholds in the PS and AC, respectively. In each cell, the PS weights determine the priorities of the users of different slices and the AC thresholds define the maximum resources that can be allocated to a slice. We call this entity SLA mapping layer since it is responsible to enforce the targets of the SLAs in the network. By monitoring the performance of the slices and the traffic loads that they impose on the network, SLA mapping layer steers the RRM so that the slices that do not overload have their SLAs fulfilled.

Moreover, since this entity orchestrates the RRM over the whole network, slice-specific control parameters are tuned for each cell separately. This property assures that based on the SLAs, spatial heterogeneities of slices are also under control. For instance, if there is a hotspot of users of a certain slice in the network, the SLA mapping layer can prioritize that slice locally and compensate other slices elsewhere.

It is envisioned that this new entity receives regular updates about the slices' performance and their deviations from the SLA targets. Based on the traffic load conditions and the SLAs, it determines if any of the slices are lagging behind their SLA targets and dynamically allocates the radio resources to rectify it. Further, the SLA mapping layer protects the slices that do not introduce load into the network higher than what was agreed upon instantiating the slice. This process should be executed iteratively, as the traffic conditions along with many other factors are non-stationary random processes and the network should be able to adapt itself to new conditions. The functionality of this entity as a slice manager is that based on the feedback from the network, SLAs and the traffic conditions, it outputs proper control parameters so that the SLAs are fulfilled. Examples of such entity can be found in our other works in [5].

III. System Model for Network with Multiple Slices

To illustrate the impacts of different choices of RRM mechanisms in a sliced network, we present a system model of mobile network with multiple slices. We consider a mobile cellular network with $c = 1,2,...,C$ cells and let S be the set of all slices in the network. Total number of slices in the network is denoted as S. This is equal to the cardinality of set S, i.e. |S| = S. Moreover, we assume that the arrival process of the users of slice s is a Poisson-distributed random variable with an arrival rate of λ_s. Each of these slices can be chosen from different slice types. In Section III-A, three different slice types are introduced. The users belonging to these slices arrive in the network at a random time and location, attempt to download a file and leave the network (traffic model similar to [14]). The processes of admitting and serving these users are carried out by the AC and the PS. In Sections III-B and III-C, we propose a PS and AC that can accommodate slicespecific priorities and thresholds. Finally, different approaches of slicing the RAN, i.e., existing approaches as well as SLA mapping layer, are described in Section III-D.

A. Slices with Diverse Requirements

We consider a network with three slice types that have different service requirements.

- *Best Effort (BE)*

  Users of this slice do not have any rigid throughput requirements. Applications like web browsing can be considered as a BE service. Based on the users' channel conditions and PS's decision, the throughput is determined. However, it is assumed that the long-term average values are declared in the SLAs as targets that the network should achieve. Average throughput (T_{BE}) and fifthpercentile throughput (F_{BE}) are considered as KPIs for this slice. Moreover, dropping mechanism has been also considered for the BE users and it acts as a contention control mechanism. The network drops the users that are in the network for more than a certain time (θ_D). This mechanism helps to control the number of users that are connected to the network. Therefore, another KPI for this slice type is the dropping rate (D_{BE}). Alternatively, we can use $1 - D_{\mathrm{BE}}$ rate as the KPI. With this definition, increase in all of the KPIs is desirable. We assume that all of the BE users are admitted to the network, i.e., there is no admission control.

- *Constant Bit-Rate (CBR)*

  The network must serve the users of the CBR slice with a constant throughput, irrespective of the required resources it takes to fulfill it. As an example of a CBR slice, we could refer to the Voice-over-IP (VoIP) as it requires a fixed throughput to deliver seamless service. This implies that if the AC has granted connection to a CBR user, the network has to serve it at a constant rate, regardless of its channel conditions. We consider the average admission rate (A_{CBR}) of the users of this slice as the KPI.

- *Minimum Bit-Rate (MBR)*

  These users have a minimum throughput that the network will provide them if they are admitted. Video streaming is a good example of such services because the video codecs usually require a minimum bit-rate to be able to stream the video with the lowest quality. These users are similar to BE users, as their throughput is determined by their channel conditions and PS's decisions. However, some users of this slice might need more resources than what PS provides them initially. In that cases, the minimum resource required to satisfy the MBR should be given to those users. Average throughput (T_{MBR}), fifth-percentile throughput (F_{MBR}) and average admission rate (A_{MBR}) are the KPIs of these slices.

These slice types are the templates that the slices in the network can be instantiated from. We define SBE, SCBR and SMBR to be the sets of all BE, CBR and MBR slices, respectively.

B. Packet Scheduler

To model the PS operation, we start by modeling the users' throughput. Based on Shannon's capacity formula, the throughput of users $i = 1,2,\cdots,N_{s,c}$ from slice S_s in cell c is defined as

[redacted], (1)

where[redacted] is the resource share of the user[redacted] is user i's Signal-to-Interference-plus-Noise-Ratio (SINR) and B is the total bandwidth.

The CBR users have a constant guaranteed throughput which is given in the SLA ($\bar{G}_s$). Therefore, the resource share needed to fulfill the throughput for every user belonging to slice s in SCBR is given by

[redacted]. (2)

Since the network is obligated to serve the admitted CBR users, they will take their share of resources first. The total required resources of all users of all CBR slices is

$$R_{\mathrm{CBR},c} = \sum_{s\in S_{\mathrm{CBR}}} \sum_{i=1}^{N_{s,c}} r_{s,ci}\,, \qquad (3)$$

and the rest of the resources $(1 - R_{\mathrm{CBR},c})$ will be shared between the MBR and BE users. A resource-fair scheduler with prioritization is proposed to model the scheduling of MBR and BE users. A legacy resource-fair scheduler distributes an equal amount of resources to each user. In cell c, a weight vector of $\mathrm{w}_{*,c} = [w_{1,c}, w_{2,c}, \cdots, w_{|S_{\mathrm{BE}} \cup S_{\mathrm{MBR}}|,c}]$ is defined, which enables the prioritization of different slices. $S_{\mathrm{BE}} \cup S_{\mathrm{MBR}}$ constitutes all the BE and MBR slices. The resource share of users $i = 1,2,\cdots, N_{s,c}$ for every user belonging to slice s in $S_{\mathrm{BE}} \cup S_{\mathrm{MBR}}$ in cell c is defined as

[redacted]

$s^0 \in S_{\mathrm{BE}}$ $s^{00} \in S_{\mathrm{MBR}}$

Eq. (4) does not guarantee that the users belonging to S_{MBR} will achieve their minimum bit-rate. To simultaneously use Eq. (4) and fulfill the MBR requirement, an iterative scheduling is proposed. In the beginning, the resources are shared based on Eq. (4). The minimum resources will be determined and assigned to the MBR users that have lower throughput than their minimum bit-rate (similar to Eq. (2)). The collective resource consumption of the users of S_{CBR} is

$$\breve{R}_{\mathrm{MBR},c} = \sum_{s^{00}\in S_{\mathrm{MBR}}} \sum_{i=1}^{\breve{N}_{s00,c}} r_{si00,c}, \qquad (5)$$

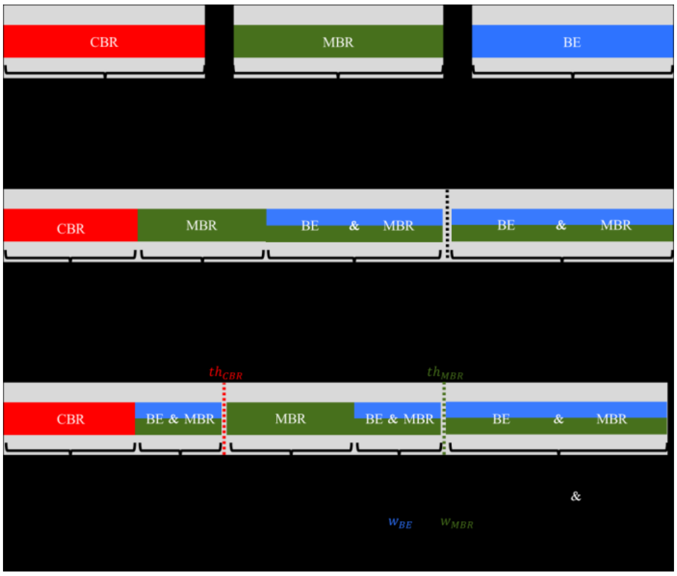

Fig. 1: Assignment of radio resources (in frequency) by different slicing schemes.

where $\breve{N}_{s00,c}$ is the number of users that have received this special treatment. The resource share of every user belonging to slice s in $\in S_{\mathrm{BE}} \cup S_{\mathrm{MBR}}$ in cell c will be

$$r_{s,ci}(\mathrm{w}_{*,c}) = \frac{N w_{s,c s0,c} \cdot\cdot (1 w_{s-0,c} R +_{\mathrm{CBR},c} P - \breve{R} N_{\mathrm{MBR}}'' {}_{s00,c,c}) \cdot w_{s00,c}}{\sum_{s^0\in S_{\mathrm{BE}}} \quad \sum_{s^{00}\in S_{\mathrm{MBR}}}}, \qquad (6)$$

where $\ddot{N}_{s00,c} = N_{s00,c} - \breve{N}_{s00,c}$ is the number of MBR users of slice s^{00} whose assigned resources are sufficient and the minimum bit-rate is achieved for them. Since it is not certain that the Eq. (6) can satisfy all the users of S_{MBR} in a single shot, we use this equation iteratively until all of these users have achieved their minimum bit-rate.

C. Admission Control

Regarding the AC, a rather straightforward approach to implement slice-awareness is to employ resource thresholds for the CBR and MBR users in each cell. The AC will grant admission to the incoming for every user belonging to slice s in $S_{\mathrm{CBR}} \cup S_{\mathrm{MBR}}$ only if the sum of the resources required by a slice's users do not exceed a certain threshold, i.e.

$$\begin{cases} \text{If } R_{s,c} \le th_{s,c} & \text{grant admission} \\ \text{If } R_{s,c} > th_{s,c} & \text{deny admission} \end{cases}, \qquad (7)$$

where $th_{s,c}$ is the resource threshold for slice s in cell [redacted] N c andis the minimum resources that is required to satisfy the users belonging to slice s in $S_{\mathrm{CBR}} \cup S_{\mathrm{MBR}}$. Note that we assume that all of the users of BE slices are admitted. Instead, these users will be dropped if they linger in the network for more than θ_D.

D. Slicing Schemes

To study the impact of RRM mechanisms in a sliced network, we consider the following approaches to slicing a

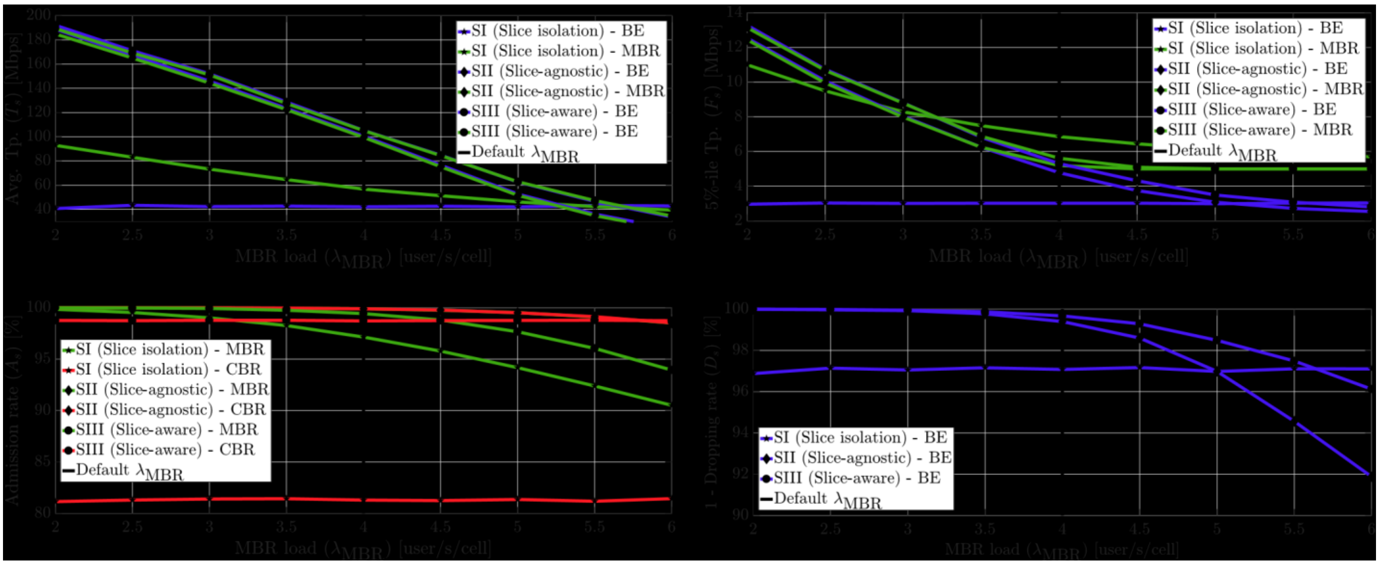

Fig. 2: Impact of load variations on the KPIs in different slicing schemes.

network, which cover the spectrum of slicing options. Fig. 1 depicts how the resources are allocated based on different slicing schemes.

- *Scheme I : Slice isolation*

  In this scheme, the resources are dedicated to slices and no sharing is possible between them. Because of its low complexity in implementation and management, this scheme has attracted much attention in the architectures that allow the slicing only in CN. Network sharing schemes that employ fixed resource sharing are examples of these architectures.

- *Scheme II : Slice-agnostic*

  The other extreme scheme would be to fully share the network with all the slices, without slice-specific control parameters. This scheme resembles the types of networks that have heterogeneous traffic. In such networks, the promises are directly made to the users. Hence, the collective behavior and performance of the users over the whole network is not considered. One example of this scheme is the LTE system that utilizes QoS-bearer mechanisms to promise certain services to the users directly.

- *Scheme III : Slice-aware*

  Finally, we assess a scheme, where the slices are sharing the radio resources but slice-specific control parameters are present to make sure that the SLAs are fulfilled for all of the slices and that the slices do not negatively affect each other's performance. SLA mapping layer is responsible for tuning these slice-specific weights and thresholds in PS and AC, respectively.

TABLE I: Simulation parameters

File size	16 [Mb]
Antenna Model	Omni-directional

Simulation duration	1 [hours]
Simulation realization	15
Drop time threshold (θ_D)	8 [sec]
Carrier frequency	2 [GHz]
Downlink transmit power	45 [dBm]
Noise power density	-174 [dBm/Hz]
Propagation model	Free-space path loss + Log-normal shadowing
Interference	Full interference from all cells
Total bandwidth (B)	90 [MHz]
Number of serving cells	7
Number of surrounding cells	12
Cell radius	1 [km]
Shadowing std. dev.	8 [dB]
Load of CBR (λ_{CBR})	3 [users/s/cell]
Load of BE (λ_{BE})	10 [users/s/cell]
Default load of MBR (λ_{MBR})	4 [users/s/cell]
Guaranteed constant bit-rate ($\bar{G}_{\text{CBR}}$)	5 [Mbps]
Guaranteed minimum bit-rate ($\bar{G}_{\text{MBR}}$)	5 [Mbps]
SI: Static resources of hCBR, BE, MBRi slices	h17, 60, 23i [%]
SII: AC threshold (th_{all}) for both CBR and MBR slices	0.66 [%]
SIII: Default control parameters hth_{CBR},th_{MBR},w_{BE},w_{MBR}i	h0.33, 0.33, 0.5, 0.5i [%]

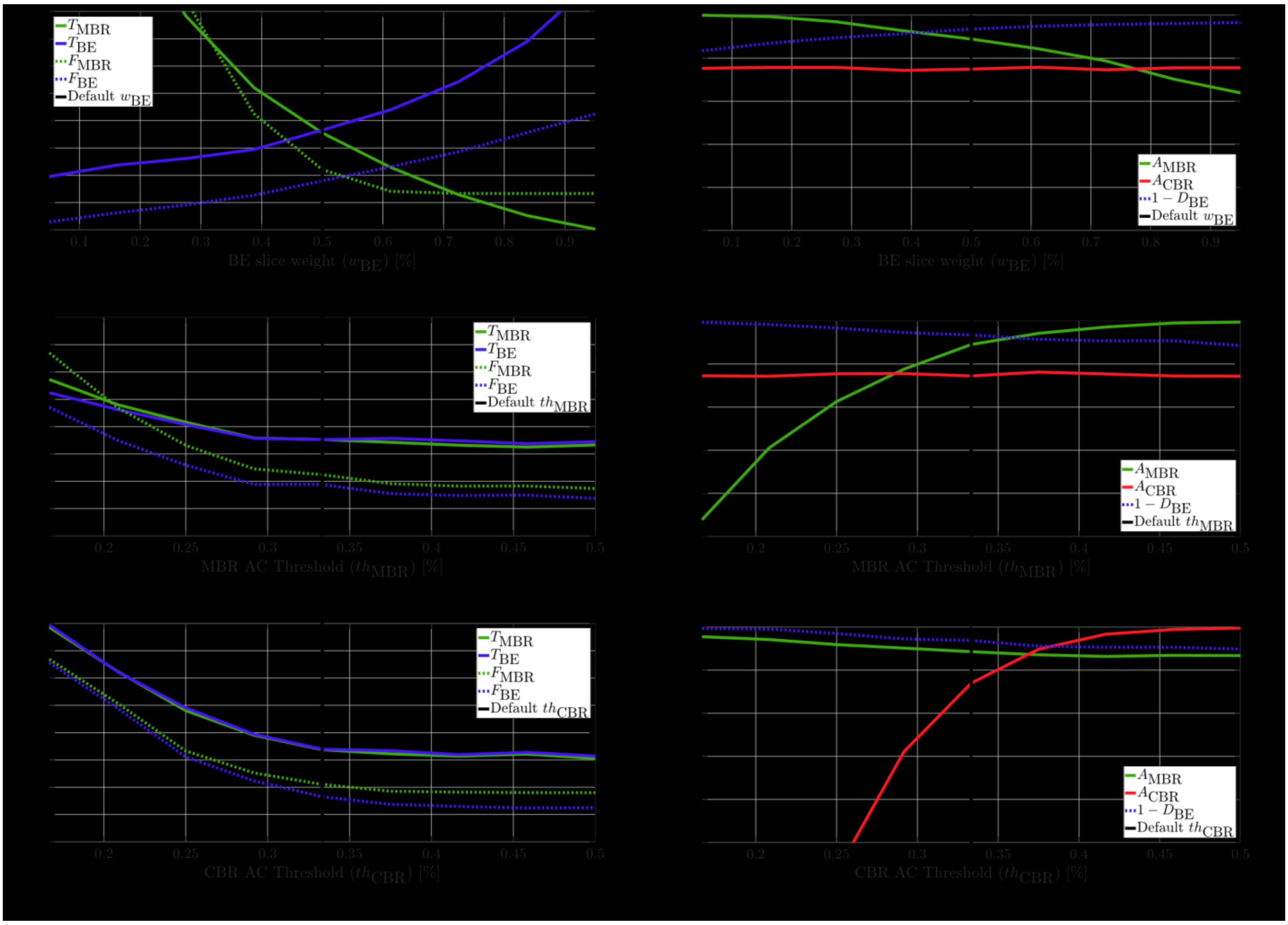

Fig. 3: Impact of control parameters of the AC and PS on the KPIs of the slices.

IV. System-Level Simulations and Evaluation

In this section, we first present the simulation setup. Then, we demonstrate the fundamental differences between different slicing schemes and show why we need slice specific control parameters in the RRM to fulfill the objectives of network slicing. Afterward, the effects of different control parameters on the KPIs of different slices is studied.

A. Simulation Setup

We assume that there is one instance from each of the slice types (BE, MBR and CBR) present in the network. Therefore, three slices with different requirements have SLAs with the network operator and the users of different slices appear in the network according to a Poisson arrival process, requesting a file of size 16 Mb. The arrival rate determines the traffic demand of a slice. In the simulations, we increase the traffic load of a slice (the MBR slice in this study) and observe the inter-slice influences under different RRM slicing schemes. Thereafter, we focus on the slice-aware RRM and note the effect of sweeping the control parameters on the KPIs of different slices. Details of the simulation setup are summarized in Table I.

B. Evaluations

Fig. 2 illustrates the effect of load variations of the MBR slice on the KPIs of all of the slices under different slicing schemes. Starting from Scheme I (Slice Isolation), we see that in all the KPIs the inter-slice influences are not existent. This is clear since the resources are completely separated and the increase in the MBR load will only affect the MBR's slice only. However, the lack of multiplexing gains causes the KPI's to be lower than of the other schemes. This effect will be more prominent if there are more slices available in the network. Moving to Scheme II (slice-agnostic) and Scheme III (slice-aware), we observe that at the default load conditions ($\lambda_{MBR} = 4$), all of the KPIs have identical performances. However, as the load of the MBR slice increases, the interslice influences of the Scheme II are more pronounced. This is because, unlike Scheme III , the slices are ignored in the RRM and no slice-specific AC or PS are present. Note that, in this paper, we have kept the control parameters of the Scheme III in a default value. However, these slice-specific control parameters should be tuned in the network to adapt to the conditions of the network.

To see the effects of tuning the control parameters of the AC and PS on the KPIs of the slices, Fig. 3 is presented. These parameters are available to the network management in SIII and other schemes are not flexible enough to allow control over the KPIs. Starting from the slice weights in the PS, Fig. 3a and Fig. 3b demonstrate the effect of prioritizing one slice over the other in the PS. As the slice weight for the BE slice (w_{BE}) increases, the average throughput (T_{BE}) and fifth-percentile (F_{BE}) throughput of this slice increase noticeably. However, the same KPIs for the MBR slice decrease. Notice that the F_{MBR} never decreases below 5 [Mbps] since this is the minimum bit-rate and the scheduler guarantees it to all the MBR users. Moreover, the A_{MBR} decreases and $1 - D_{\mathrm{BE}}$ increases slightly as the w_{BE} increases. This is because the MBR users, compared to the default w_{BE}, have lower throughput as they are allocated lower resources. Hence, these users stay in the network for a longer time, which in turn, causes more MBR users to be blocked. On the other hand, the BE users enjoy higher throughput and the probability that they can leave the network before $\theta_D = 8$ [sec] increases. Finally, note that the CBR slice is not affected by this control parameter, since it is not controlled by PS, and receives its required resources regardless.

Fig. 3c - 3f demonstrate the effect of changing the AC thresholds for the MBR and CBR slices. Starting from AC threshold for the MBR slice (th_{MBR}), the increase in this parameter ensures the increase in the A_{MBR}. This will increase the number of active MBR users and thereby, increases the load of the network. This causes more competition for the resources. This behavior can be observed in Fig. 3c, where average throughput and fifth-percentile throughput decrease. Moreover, the dropping of the BE users increases slightly as they have lower throughput. Note that this decrease has a floor, because after some point (approximately at $th_{\mathrm{MBR}} = 40$ [%]) all of the users of MBR slices will be admitted to the network. Moving to the th_{CBR}, we observe a similar behaviour, where more CBR users will cause a decrease in T_{BE}, F_{BE}, T_{MBR} and F_{MBR}. Moreover, since the users of MBR and BE slices have lower throughput (compared to the default th_{CBR}) the A_{MBR} and $1 - D_{\mathrm{BE}}$ decreases slightly.

In Figs. 3a - 3f, we have shown that the KPIs of the slices can be adjusted by changing the control parameters. It is the task of the SLA mapping layer to find the appropriate control parameters so that the SLAs of all of the slices are fulfilled. In our previous work [5], we have shown examples of such an entity in the network. It should be further added that the relationship between these KPIs and the control parameters is an analytically intractable function, which has to be estimated in an online fashion.

V. Conclusion and Outlook

In this paper, we have studied different approaches of slicing the RAN. To this end, we have defined a system model for a network with multiple slices that have fundamentally different requirements. We showed that considering the QoS-awareness or network sharing concepts are not sufficient by themselves and the slice-aware RRM is a crucial aspect of RAN slicing. By studying the impact of load variation in different schemes, we have concluded that the slice-aware scheme can simultaneously benefit from multiplexing gains and control the inter-slice influences. We have further illustrated the The impact of slice-specific control parameters of the AC and PS on the KPIs. These control parameters can be fine tuned by an intelligent slice management algorithm. The goal of such algorithm would be the fulfillment of SLAs of different slices in the network. To cope with the dynamic changes of a real cellular network, such an algorithm should iteratively adapt the control parameter.

Acknowledgments

We would also like to thank Dr. Andre Noll Barreto (from Barkhausen Institute, Dresden, Germany) for his constructive comments and assistance that greatly improved the manuscript.

References

[1] I. da Silva, G. Mildh, A. Kaloxylos, P. Spapis, E. Buracchini, A. Trogolo, G. Zimmermann, and N. Bayer, “Impact of network slicing on 5G Radio Access Networks,” in *2016 European Conference on Networks and Communications (EuCNC)*, June 2016, pp. 153–157.

[2] M. Jiang, M. Condoluci, and T. Mahmoodi, “Network slicing management amp; prioritization in 5G mobile systems,” in *European Wireless 2016; 22th European Wireless Conference*, May 2016, pp. 1–6.

[3] NGMN Alliance, “NGMN 5G White Paper,” Tech. Rep., Feb. 2015.

[4] 3GPP, “Policy and charging control architecture,” 3rd Generation Partnership Project (3GPP), TR 23.203, Mar. 2018.

[5] B. Khodapanah, A. Awada, I. Viering, D. Oehmann, M. Simsek, and G. P. Fettweis, “Fulfillment of Service Level Agreements via SliceAware Radio Resource Management in 5G Networks,” in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, June 2018, pp. 1–6.

[6] M. Richart, J. Baliosian, J. Serrat, and J. Gorricho, “Resource Slicing in Virtual Wireless Networks: A Survey,” *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462–476, Sept 2016.

[7] C. Liang and F. R. Yu, “Wireless Network Virtualization: A Survey, Some Research Issues and Challenges,” *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 358–380, Firstquarter 2015.

[8] M. I. Kamel, L. B. Le, and A. Girard, “LTE Wireless Network Virtualization: Dynamic Slicing via Flexible Scheduling,” in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Sept 2014, pp. 1–5.

[9] B. Li, C. Lin, and S. T. Chanson, “Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks,” *Wireless Networks*, vol. 4, no. 4, pp. 279–290, Jun 1998. [Online]. Available: https://doi.org/10.1023/A:1019116424411

[10] Y. Fang, “Thinning schemes for call admission control in wireless networks,” *IEEE Transactions on Computers*, vol. 52, no. 5, pp. 685–687, May 2003.

[11] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí, “Self-optimized admission control for multitenant radio access networks,” in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–5.

[12] A. Ksentini and N. Nikaein, “Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction,” *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102–108, 2017.

[13] I. Malanchini, S. Valentin, and O. Aydin, “Generalized resource sharing for multiple operators in cellular wireless networks,” in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug 2014, pp. 803–808.

[14] 3GPP, “Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects,” 3rd Generation Partnership Project (3GPP), TR 23.203, Mar. 2017.