This document is a preprint of: B. khodapanah, A. Awada, I. Viering et al., "Fulfillment of Service Level Agreements via Slice-Aware Radio Resource Management in 5G Networks," in Proceedings of IEEE Vehicular Technology Conference (VTC Spring 2018), Porto, Portugal, Jun 2018. DOI:10.1109/VTCSpring.2018.8417711

Fulfillment of Service Level Agreements via Slice-Aware Radio Resource Management in 5G Networks

Behnam Khodapanah*, Ahmad Awada†, Ingo Viering‡, David Öhmann*, Meryem Simsek*, Gerhard P. Fettweis*
*Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Germany
Email:{behnam.khodapanah, david.oehmann, meryem.simsek, 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

***Abstract*—In the context of 5G mobile networks, several new use cases with various requirements with respect to throughput, latency, coverage, etc., should be addressed. To avoid deployment of separate networks for each of the use cases, the concept of network slicing has been introduced, where several logical networks share a single physical network. However, the accommodation of networks with diverse requirements in a single physical network is a new challenge. In this work, we study the effects of a mapping layer, which supervises the network over a service area and manages the allocation of radio resources to slices to guarantee their target service requirements. To do so, we propose an adaptation algorithm based on minimizing deviations from slice requirements. The results show that by utilizing the mapping layer, the resources can be shared efficiently and fairly and the deviations of Key Performance Indicators (KPIs) from the Service Level Agreement (SLA) targets are reduced compared to distributed control methods that are typically used in legacy and current cellular systems.**

***Index Terms*—Network Slicing, Radio Resource Management, Service Level Agreement, KPI, Slice Isolation, 5G**

I. Introduction

Fifth generation mobile networks (5G) should be able to serve multiple service classes with different requirements such as eMBB (enhanced Mobile Broadband), mMTC (massive Machine Type Communications) and URLLC (Ultra-Reliable and Low Latency Communications) [1]. It is essential that individual network deployment for each of the use case is avoided, since this will result in an impractical solution where the network is largely segmented. Therefore, a cost-efficient, flexible and scalable solution is required to be adoptable for different services and use cases [2]. To meet the mentioned necessities, the concept of network slicing has been introduced where resources of a single physical network is shared by multiple end-to-end logical networks, i.e., network slices [1]. Resources such as spectrum resources (time and frequency), Radio Access Technologies (RATs), multi-connectivity admission, network functions utilization, etc., are shared dynamically between different slices.

According to [3], in a sliced network, network operators realize the network slices instances to provide the required network characteristics to different services, pertaining to third-party tenants. Therefore, there will be Service Level Agreements (SLAs) between network operator and the tenants to declare the requirements of a particular service and the operator should fulfill these SLAs via instantiating appropriate network slices. Requirements of the service instances are specified in terms of Key Performance Indicators (KPIs), such as throughput, latency, availability, coverage, etc.

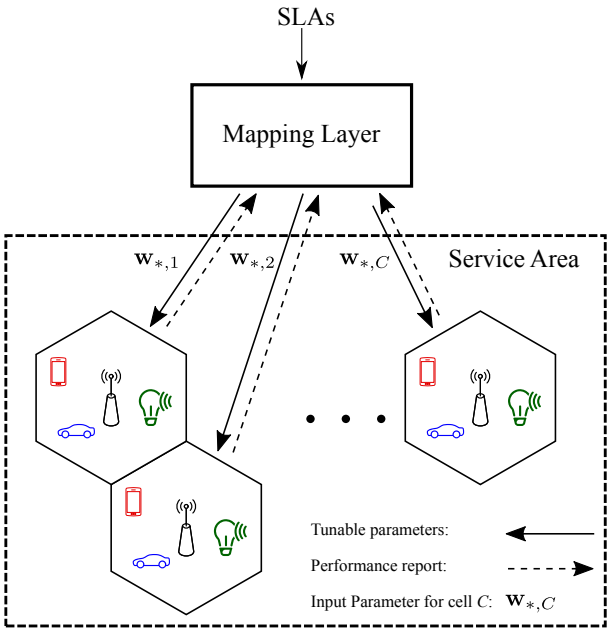

Fig. 1: Mapping layer visualization.

An important aspect of network slicing is to guarantee that the slices operate independently, i.e., the performance, congestion, failure, etc., in one slice will not negatively influence the performance of other slices which are sharing the resources. This means that providing protection from other slices is one of the necessities for network slicing [4]. A primitive strategy to ensure traffic isolation is to dedicate the resources statically to each slice. However, this method is not resource efficient since no multiplexing gains are achieved. Therefore, it is anticipated that the resources are shared dynamically between slices. The goal is to devise a strategy that can keep the KPIs

of different slices near the target KPIs (declared in the SLAs of respective services), while maintaining the isolation of slices with dynamically shared resources.

The virtualization of network functions in wireless systems has been studied for a while in the context of Long-Term Evolution (LTE), e.g., [5], [6]. Additionally, several works have proposed solutions in the context of scheduling with Quality of Service (QoS) [7]. Multi-QoS scheduling has been investigated in [8], [9]. However, authors in [4] have instead argued that existing QoS mechanisms do not support the application of specific policies to a group of users in the network. Therefore, it is envisioned that the radio access network (RAN) needs to be slice-aware to support coexistence of slices with different requirements. Besides, as mentioned before, the slice isolation plays an important role in the operation of parallel services in future mobile networks, which is not easily feasible with current QoS mechanisms.

In this work we study the slice-aware Radio Resource Management (RRM), where a network entity called *mapping layer* keeps track of the KPIs of different slices and according to the SLAs tunes a weighting parameter of the Packet Scheduler (PS) so that the SLA targets for slices are fulfilled [10]. As visualized in Fig. 1, the mapping layer keeps track of the load and performance (in terms of KPIs) of different slices in different cells and outputs the proper tunable parameters to assign radio resource shares to the specific slices. Besides, this entity should be able to decide which slices in cells have higher priority, since in situations where the network is congested (demand is higher than the available resources), the mapping layer should be able to protect the slices that are not introducing excessive demand. We propose a slice-aware adaptation algorithm for the mapping layer, which tries to minimize the deviations from the target KPIs for slices in a service area by assigning the slice weighting parameter.

This paper is structured as follows. After describing the system model in Section II, we introduce an adaptation algorithm in Section III that maps the SLA fulfillment problem into an optimization problem. Next, In Section IV, we introduce different adaptation schemes to examine the adaptive resource allocation algorithm in different network settings. In Section V, we present different scenarios in order to evaluate the proposed adaptation schemes and finally in Section VI we conclude this paper.

II. System Model

We consider a mobile cellular network with slices $s = 1, 2, \cdots, S$ and cells $c = 1, 2, \cdots, C$. The service area is defined as an area with a multitude of adjacent cells, which serve users from different slices. The number of users from slice s in cell c is denoted as $N_{s,c}$ and the total number of users of slice s in the service area is denoted as $N_{s,*} = \sum_{c=1}^{C} N_{s,c}$.

A. Traffic Load

Slice SLAs should not merely assert target KPIs because it is also important to consider the amount of traffic that each slice will impose on the service area. Considering the full buffer traffic model, the traffic load for slice s is defined as

$$L_s = \frac{N_{s,*}}{a} \quad [\text{users}/\text{km}^2], \tag{1}$$

where a is the service area (in km^2). If the traffic load of a slice in the entire service area or in some parts of it is more than the anticipated traffic load defined in the SLA, the network may not be able to fulfill the target KPIs for all of the slices. This situation is what we call network congestion, and the network needs to prioritize other slices that are not overloading.

B. Packet Scheduler (PS)

In this work, a resource fair scheduler with prioritization is used. A conventional resource fair scheduler gives the same fair share of resources to every user. To be able to prioritize different slices, a weight parameter is assigned to each slice and its users in each cell and the scheduler assigns their resource share based on the weights. Here we define the weight matrix as

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,C} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ w_{S,1} & w_{S,2} & \cdots & w_{S,C} \end{bmatrix} = [\mathbf{w}_{*,1}, \mathbf{w}_{*,2}, \cdots, \mathbf{w}_{*,C}], \tag{2}$$

where $\mathbf{w}_{*,c}$ is $S \times 1$ vector consisting of the weights of all S slices in cell c. In case a single weight vector is used for all cells, then $\mathbf{w} = \mathbf{w}_{*,c}$ for $c = 1, 2, \cdots, C$ is used. The resource share of user $i = 1, 2, \cdots, N_{s,c}$ from slice s in cell c is defined as

$$r^i_{s,c}(\mathbf{w}_{*,c}) = \frac{w_{s,c}}{\sum_{s'=1}^{S} N_{s',c} \cdot w_{s',c}}, \tag{3}$$

where the weights are normalized so that $\sum_{s=1}^{S} w_{s,c} = 1$ and all the weights should be positive. The throughput of each user is calculated based on its resource share and its Signal-to-Interference-plus-Noise-Ratio (SINR) as

$$T^i_{s,c}(\mathbf{w}_{*,c}) = r^i_{s,c}(\mathbf{w}_{*,c}) \cdot n \cdot B \cdot \log_2(1 + \gamma^i_{s,c}), \tag{4}$$

where $\gamma^i_{s,c}$ is user i's SINR, n is the number of Physical Resource Blocks (PRB) and B is the bandwidth of each PRB. Note that here we use Shannon's capacity to map the SINR to rate, thereby, neglecting the detailed implementation of the LTE physical layer subroutines.

C. Definition of Key Performance Indicators

There are numerous KPIs to be considered in an SLA, such as latency, coverage, energy efficiency, etc. [11]. However, since in this work we are focusing on PS and the related KPIs, we assume that the SLA consists of two KPIs, namely average throughput and minimum throughput. The considered KPIs can well represent the behavior of the PS in a cellular network. Conventionally fifth-percentile throughput is considered as a KPI for worst-case users in cellular networks; however, we

focus on the minimum throughput as it can be viewed as stricter version of the fifth-percentile throughput (e.g. for certain applications like URLLC, where 99.999% reliability is important).

To calculate the average throughput of a slice over the service area, we average over all the users in all the cells, i.e.,

$$A_s(\mathbf{W}) = \left(\sum_{c=1}^{C} \sum_{i=1}^{N_{s,c}} T^i_{s,c}(\mathbf{w}_{*,c}) \right) / N_{s,*}. \quad (5)$$

The minimum throughput of a slice over the service area, can be expressed as

$$M_s(\mathbf{W}) = \min_{c \in \{1,2,\cdots,C\}} \left(\min_{i \in \{1,2,\cdots,N_{s,c}\}} T^i_{s,c}(\mathbf{w}_{*,c}) \right), \quad (6)$$

where we take the minimum throughput of each cell and then the minimum over all cells. According to the SLA, slices have targets for different KPIs. Each slice s has a target average throughput $\hat{A}_s$ and a target minimum throughput $\hat{M}_s$. Besides, traffic load of slice L_s and the traffic load limit described in the SLA is denoted as $\hat{L}_s$.

III. Mapping Layer Adaptation Algorithm

A. Optimization Approach

For a scalable and flexible adaptation algorithm that outputs the tunable parameters to the PS, an optimization problem is formulated. The idea is to use cost functions for all different KPIs of slices and associate cost values for deviations from target KPI defined in the SLA. Let $\Phi^A_s(\mathbf{X})$ and $\Phi^M_s(\mathbf{X})$ be the cost function related to the average throughput and minimum throughput of slice s, respectively. The parameter $\mathbf{X}$ can be defined as the matrix of weights, i.e., $\mathbf{W}$, cell specific weights vector of $\mathbf{w}_{*,c}$ or a weight vector of $\mathbf{w}$ for the service area (further specification in Section IV). Next, we define the total cost as the sum of costs of all slices, i.e.,

$$\Phi(\mathbf{X}) = \sum_{s=1}^{S} \Phi^A_s(\mathbf{X}) + \sum_{s=1}^{S} \Phi^M_s(\mathbf{X}). \quad (7)$$

The goal is to minimize the cost function of Eq. (7), which in turn is equivalent to search for minimum deviations from the target KPIs or in other words, a better fulfillment of the SLA. This optimization problem is defined as

$$\min_{\mathbf{X}} \Phi(\mathbf{X}) \quad \text{s.t. constraints on } \mathbf{X}. \quad (8)$$

The constraints here are that the weights in each cell should be positive and normalized (refer to Section II-B).

B. Definition of Cost Functions

For each KPI, the following criteria have been considered in the definition of cost functions.

- Cost functions should suit the nature of the KPI of a service. For instance, for the average throughput $\hat{A}_s$ and minimum throughput $\hat{M}_s$ KPIs, no cost should be associated for values above the target throughput. However, if delay is the KPI, costs should be associated for high values of delay.
- Cost functions should be normalized so that the total cost is not dominated by a single cost function of one slice.
- Cost functions should be able to accommodate slice priorities. As stated in Section II-A, the slice which is overloading, i.e., exceeding the defined traffic load limit, should be given less priority.

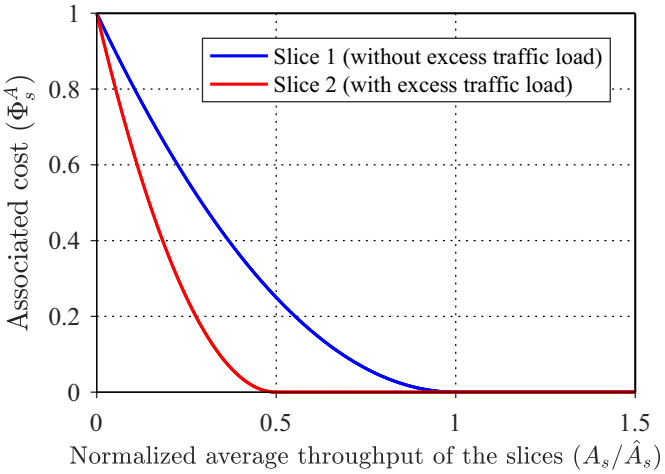

Fig. 2: Cost functions related to the average throughput. Both slices have target throughput of $\hat{A}_1 = \hat{A}_2 = 1$, but the load index for Slice 1 is $l_1 = 1$ and for Slice 2 is $l_2 = 2$.

The cost function for average and minimum throughput are defined as

$$\Phi^A_s(\mathbf{W}) = \left(1 - \frac{A_s(\mathbf{W})}{\hat{A}_s} \cdot l_s\right)^2 \cdot H\left(1 - \frac{A_s(\mathbf{W})}{\hat{A}_s} \cdot l_s\right) \quad (9)$$

$$\Phi^M_s(\mathbf{W}) = \left(1 - \frac{M_s(\mathbf{W})}{\hat{M}_s} \cdot l_s\right)^2 \cdot H\left(1 - \frac{M_s(\mathbf{W})}{\hat{M}_s} \cdot l_s\right), \quad (10)$$

where $H(\cdot)$ is a step function and l_s is overload index, which is defined as

$$l_s = \max\left(\frac{L_s}{\hat{L}_s}, 1\right). \quad (11)$$

As we can see in Fig. 2, the defined cost functions match the KPIs, they are normalized and we can prioritize different slices based on the load limit defined in the SLA and the instantaneous load that the slice is imposing on the network.

IV. Mapping Layer Adaptation Schemes

To apply the adaptation of slice weights based on the optimization approach of Section III, different schemes have been investigated. Adaptation can be done in a distributed manner, where each cell individually decides on the best slice weights. This is a similar approach to the current QoS mechanisms. Alternatively, the mapping layer can output weights to the cells in a centralized manner. In the following four schemes of adaptation are introduced.

TABLE I: Simulation Parameters

Spatial distribution of users	Uniform / Gaussian
Propagation model	Free-space path loss with Log-normal shadowing
Shadowing std. dev. [dB]	8
Traffic	Constant and infinite queue
Number of PRBs	50
Total bandwidth [MHz]	10
Number of cells	7
Cell radius [km]	1
Number of simulation runs	2500
Mean of 2D Gaussian user distribution [km]	$(5 \cdot 3/4, 3\sqrt{3} \cdot 1/4)$
Std. dev. of 2D Gaussian user distribution [km]	$(5 \cdot 1/4, 3\sqrt{3} \cdot 1/4)$

- *Scheme I: No Adaptation*
  All users from all slices have a constant slice weight. Therefore, no adaptation is considered. This scheme serves as a benchmark for resource management with equal weights for all users and is used merely for comparison.
- *Scheme II: Cell-wise Adaptation*
  Each cell aims to achieve the goals of its SLA internally without any information exchange with other cells and has only information about the users that it is serving. This means that it is unaware of the neighboring cells' performance. This scheme can be viewed as an example of a distributed adaptation approach, which resembles the existing QoS mechanisms.
- *Scheme III: Mapping Layer with Global Slice Weights*
  The mapping layer tries to ensure the SLA fulfillment for the whole network. It collects statistics about traffic load and KPIs from all cells in the service area. Then, each slice will be given a certain weight for the whole network. Therefore, the slice weights in different cells are identical. This is a centralized mapping scheme without consideration of cell specific conditions.
- *Scheme IV: Mapping Layer with Local Slice Weights*
  Similar to the Scheme III, the mapping layer aims to fulfill the SLA requirements, given that it has sufficient information about the users' performance in the service area. With this scheme, however, weights can be different for each cell. Therefore, the mapping layer has more freedom in choosing the slice weights.

The optimization problem is solved for each of the schemes introduced. However, each scheme has different knowledge about the performance of users of different slices. In Scheme II, the cell is only aware of the users that it is serving but in Schemes III and IV the mapping layer is aware of the whole service area. Besides, each scheme has different degrees of freedom in tuning the parameters. In Scheme II, the optimizer outputs a vector of weights $\mathbf{w}_{*,c}$ for slices within each cell.

TABLE II: Simulation Scenarios

Scenario	s	$\hat{A}_s$ [$\frac{\text{Mbit}}{\text{s}}$]	$\hat{M}_s$ [$\frac{\text{kbit}}{\text{s}}$]	$\hat{L}_s$ [$\frac{\text{users}}{\text{km}^2}$]	L_s [$\frac{\text{users}}{\text{km}^2}$]	UE dist.
A	1	1.0	135	7.27	7.27	Uniform
	2	1.0	135	7.27	$4-13$	Uniform
B	1	1.0	135	7.27	7.27	Uniform
	2	1.0	135	7.27	$4-13$	Gaussian
C	1	1.0	135	3.63	3.63	Uniform
	2	1.0	135	3.63	3.63	Uniform
	3	1.0	0	3.63	3.63	Uniform
	4	1.0	0	3.63	$1.5-6$	Gaussian

Scheme III also outputs a vector of weights $\mathbf{w}$ but this vector is not for individual cells; it is identical for all cells. Finally, Scheme IV outputs a matrix $\mathbf{W}$, where each slice in each cell has an individual weight. To solve the optimization problem, the interior point method is used as a numerical method [12].

V. Simulation Results

To assess the performance of different schemes, a cellular network with a typical 3GPP LTE setup is investigated. Furthermore, we assume that the users of slices can be uniformly or non-uniformly distributed within the service area. A 2D Gaussian PDF (Probability Density Function) is used for simulating a user hot-spot of one slice in an area (similar to [13]). The simulation parameters are summarized in Table I.

In the following, the performance of the different adaptation schemes are evaluated in three scenarios. The settings of the three scenarios are summarized in Table II. To obtain the target values, first we assume that 1 [Mbit/s] is the target average throughput. By running a simulation without slicing, we determine the maximum traffic load to be 14.54 [users/km^2] (320 users in 7 cells) and the minimum throughput to be 135 [kbit/s]. In Scenarios A and B, we have two slices, each with half traffic load of the simulation, i.e. 7.27 [users/km^2] (160 users) and in Scenario C we have four slices, each with quarter traffic load, i.e. 3.63 [users/km^2] (80 users). It is desired that the proposed adaptation algorithm isolates the traffic of different slices. Therefore, by keeping the load of particular slices at their limit and varying the load of another slice, we observe whether the proposed algorithms can ensure that the slices are isolated. Besides, the effect of non-uniform distribution of users on the protection of slices is studied as well.

For Scenario A, Slice 1 offers the maximum load limit defined in the SLA. Since it is desired to share resources fairly, the variations from Slice 2 should not negatively affect the performance of Slice 1. In Fig. 3 curves are drawn for Schemes I, II, III and IV for each Slice (e.g. I, S1 represents the performance of Scheme I for Slice 1). Here, we can observe the effects of variations in Slice 2 on the performance of Slice 1. When Slice 2 offers less traffic than the limit declared in the

This document is a preprint of: B. khodapanah, A. Awada, I. Viering et al., "Fulfillment of Service Level Agreements via Slice-Aware Radio Resource Management in 5G Networks," in Proceedings of IEEE Vehicular Technology Conference (VTC Spring 2018), Porto, Portugal, Jun 2018. DOI:10.1109/VTCSpring.2018.8417711

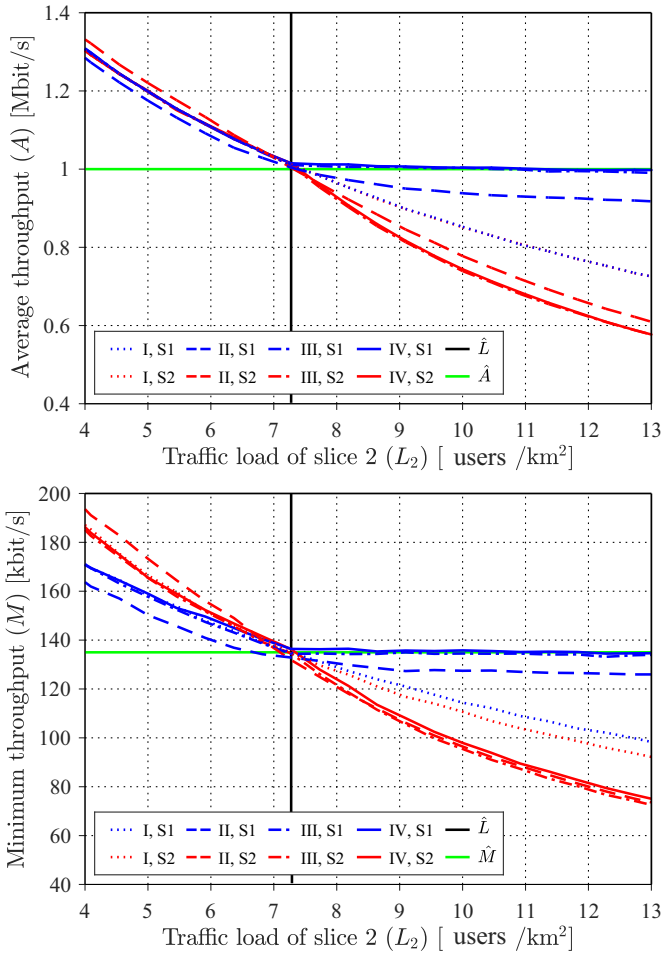

Fig. 3: Effects of load variation in Scenario A.

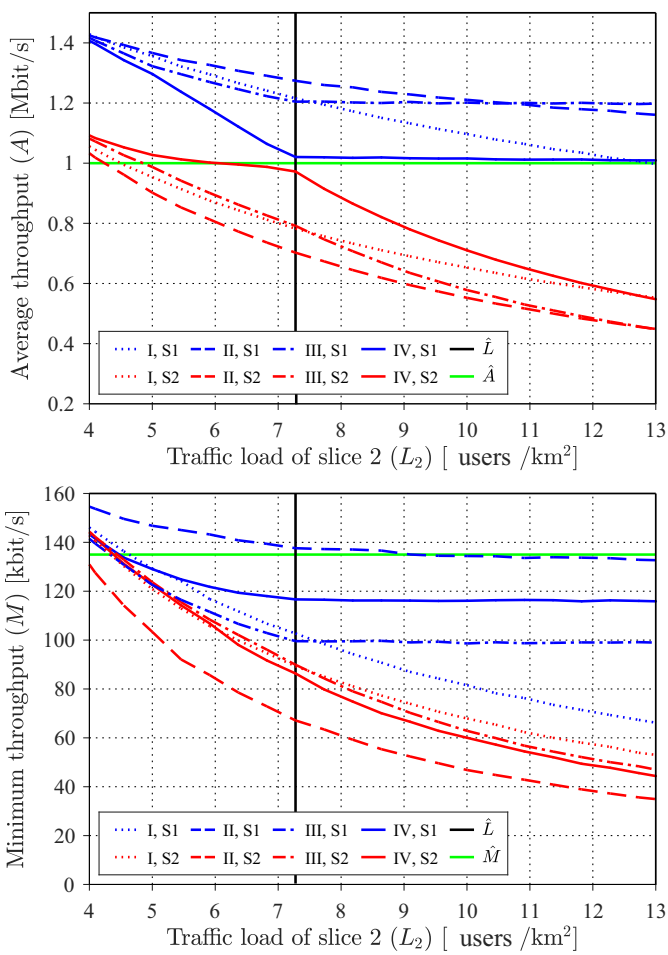

Fig. 4: Effects of load variation in Scenario B.

SLA, the excess resources are shared by both slices equally and this holds true for all Schemes. However, when further increasing the load of Slice 2 above the load limit, Schemes I and II are unable to protect the traffic of Slice 1 and we see degradation in both Slices. On the other hand, Schemes III and IV can manage to protect Slice 1 from degradation caused by overload of Slice 2. Note that Schemes III and IV have similar performance here; this is because the user distribution is uniform and global weights are similar to local weights. A similar effect is observed for the minimum throughput KPI. This shows that the adaptation algorithm with mapping layer can achieve protection for Slice 1 in all the KPIs.

Next, Scenario B is considered which is similar to Scenario A. The difference is that the users from Slice 2 are now distributed non-uniformly. This implies that Slice 2 with the same load as in Scenario A, will cause congestion in some cells. As we can see in Fig. 4, Schemes I, II and III will assign more resources to Slice 1, where Slice 2 is suffering too much. Although in Scheme I the resource share is equal for both slices, the users of Slice 2 suffer much more than Slice 1 and that is because such users are mostly in the cells that are congested. Scheme II is blind to the load situation in neighboring cells and cannot appropriately adapt the slice weights. Scheme III, on the other hand, has knowledge about the load in the service area but since it can provide only one weight per slice in all cells, it lacks the required freedom in choosing the slice weights. Only Scheme IV can properly protect Slice 1 from overloading and non-uniform distribution of Slice 2, while keeping the performance of Slice 2 as close as possible to its target. This is because in Scheme IV we have the benefits of both Scheme II and Scheme III simultaneously, i.e., full knowledge about service area plus individually tunable weights.

In Fig. 4, however, we see that Slice 1 also lags behind its target minimum throughput. This is because in the definition of the cost function, disparity in users' positions has not been accounted for. For instance, if traffic load of both slices is equal to the traffic load limit, i.e., $L_1 = L_2 = \hat{L}_1 = \hat{L}_1$, despite the non-uniform distribution of Slice 2 users, the mapping layer does not prioritize Slice 1 (they have similar costs). However, since satisfying Slice 1 is more cost-efficient, Slice 1 is favoured over Slice 2.

Finally, in Scenario C, we focus on the performance of Schemes III and IV only, since Scheme I and II have already

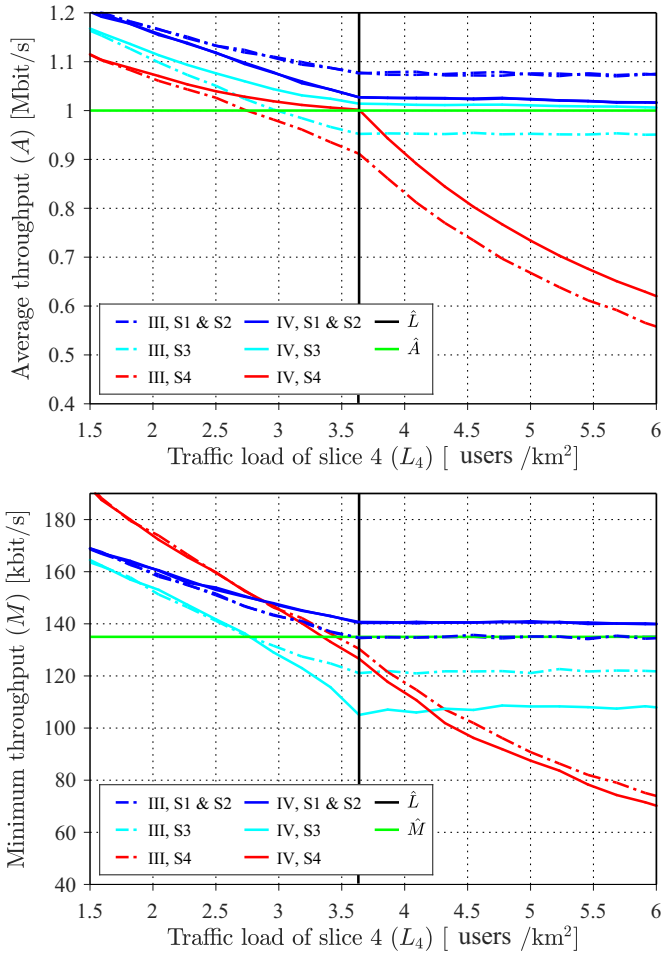

Fig. 5: Effects of load variation in Scenario C.

shown to have inferior performance. In this Scenario, we assume 4 slices, where all of them have similar average throughput targets. However, we assume that the minimum constraint has been removed for Slices 3 and 4. As we can see in Fig. 5, again the Scheme IV has better performance in protecting righteous slices in both average throughput and minimum throughput. Additionally, the relaxed constraints on minimum throughput, can be used by the adaptation algorithm to improve the KPIs that are important (according to SLA). In Fig. 5, we can see that the minimum throughput for Slice 3 from Scheme IV is relatively low, but in turn, the average throughput of Slice 3 can be maintained.

From the different scenarios, we can conclude that Scheme II cannot provide proper slice weights even in case of uniform distribution of users. Scheme III shows similar performance as Scheme IV for the uniformly distributed users, however, the non-uniform distribution can severely affect its performance. On the other Scheme IV is able to adapt to non-uniform user distributions. Besides, as we presented in Scenario C, the adaptation algorithm will take advantage of relaxed KPI requirements to improve the performance of the system in terms of smaller deviations from the target KPI.

VI. Conclusion

In this work, we evaluated a network entity, called the mapping layer, which has the responsibility of adapting the weights of the packet scheduler in a multi-cell environment so that the network-wide SLAs of slices are not violated. The proposed adaptation algorithm ensures minimum deviation from the target KPIs for all slices, by minimizing the cost functions. Additionally, to protect righteous slices from congestion caused by other slices, a prioritization mechanism has been introduced. Furthermore, different possible schemes that utilize the adaptation algorithm in different configurations are defined. Simulation results confirm that the adaptation algorithm is able to maintain traffic protection while dynamically sharing the resources and keeping the KPIs of slices near the target defined in an SLA. Also, it was shown that in the scheme where the mapping layer has knowledge about the whole service area and can influence the slice weights in each cell, the adaptation algorithm performs best and can cope with non-uniform user distributions. The algorithm and future extensions are well suited to enable a slice-aware RRM in future 5G networks.

References

[1] NGMN Alliance, "NGMN 5G White Paper," Tech. Rep., Feb. 2015.

[2] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, "Network slicing to enable scalability and flexibility in 5g mobile networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, May 2017.

[3] NGMN Alliance, "Description of Network Slicing Concept," Tech. Rep., 1 2016.

[4] 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.

[5] 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.

[6] L. Zhao, M. Li, Y. Zaki, A. Timm-Giel, and C. Grg, "Lte virtualization: From theoretical gain to practical solution," in *2011 23rd International Teletraffic Congress (ITC)*, Sept 2011, pp. 71–78.

[7] A. Asadi and V. Mancuso, "A survey on opportunistic scheduling in wireless communications," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1671–1688, Fourth 2013.

[8] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-qos-aware fair scheduling for lte," in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, May 2011, pp. 1–5.

[9] T. Guo and R. Arnott, "Active lte ran sharing with partial resource reservation," in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, Sept 2013, pp. 1–5.

[10] METIS-II, "Deliverable D2.4 Final Overall 5G RAN Design," 3rd Generation Partnership Project (3GPP), TR, 6 17.

[11] 3GPP, "Study on Scenarios and Requirements for Next Generation Access Technologies," 3rd Generation Partnership Project (3GPP), TR 38.913, 10 2016.

[12] R. H. Byrd, J. C. Gilbert, and J. Nocedal, "A trust region method based on interior point techniques for nonlinear programming," *Mathematical Programming*, vol. 89, no. 1, pp. 149–185, Nov 2000. [Online]. Available: https://doi.org/10.1007/PL00011391

[13] A. A. Khalek, L. Al-Kanj, Z. Dawy, and G. Turkiyyah, "Optimization models and algorithms for joint uplink/downlink umts radio network planning with sir-based power control," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 4, pp. 1612–1625, May 2011.