A network slice resource allocation process in 5G mobile networks

Andrea Fendt, Lars Christoph Schmelz, Wieslawa Wajda, Simon Lohmüller, Bernhard Bauer

Angaben zur Veröffentlichung / Publication details:

Fendt, Andrea, Lars Christoph Schmelz, Wieslawa Wajda, Simon Lohmüller, and Bernhard Bauer. 2019. "A network slice resource allocation process in 5G mobile networks." In *Innovative Mobile and Internet Services in Ubiquitous Computing: Proceedings of the 12th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2018)*, edited by Leonard Barolli, Fatos Xhafa, Nadeem Javaid, and Tomoya Enokido, 695–704. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-93554-6_68.

A Network Slice Resource Allocation Process in 5G Mobile Networks

Andrea Fendt, Lars Christoph Schmelz, Wieslawa Wajda, Simon Lohmüller and Bernhard Bauer

Abstract The fifth generation of mobile networks (5G) is associated with a wide spectrum of novel use cases that introduce a large number of very diverse requirements, regarding for instance throughput, latency, delay, availability and reliability. End-to-end network slicing is seen as a solution that allows to simultaneously accomplish those manifold requirements in isolated slices running on a shared network infrastructure. However, embedding those virtual end-to-end network slices into a common physical network containing wireless as well as wired network elements, while meeting all the different requirements, is still an unsolved problem. In this paper, a vision of an end-to-end network slice resource allocation process will be presented allowing to give fast feedback to a network operator or tenant on the feasibility of embedding new network slices. The associated research challenges will be discussed, especially focusing on the more complex Radio Access Network (RAN) resource allocation.

1 Introduction

The fifth generation of mobile networks (5G) covers a wide variety of novel use cases, such as the Internet of Things (IoT) and the industry of the future (Industry 4.0), requiring massive Machine Type Communication (mMTC). Furthermore, highly safety and security critical use cases, like autonomous driving and vehicular communication, require Ultra-Reliable and Low Latency Communication (URLLC). But also, traditional enhanced Mobile Broadband (eMBB) appli-

Andrea Fendt, Simon Lohmüller and Prof. Dr. Bernhard Bauer
University of Augsburg, Department of Computer Science, Augsburg, e-mail: {andrea.fendt,simon.lohmueller,bauer}@informatik.uni-augsburg.de

Lars Christoph Schmelz and Wieslawa Wajda
Nokia Bell Labs, Munich, Stuttgart, e-mail: {christoph.schmelz,wieslawa.wajda}@nokia-bell-labs.com

cations like HD video streaming and augmented reality must be taken into account. These diverse use cases enforce several radically different requirements on mobile networks. Network slicing is seen as the key concept of future 5G mobile networks, which aims at making the networks flexible enough to cope with those divergent requirements by dissolving the traditional concept of one monolithic mobile network serving all purposes [1]. A network slice is an isolated end-to-end virtual network containing all required resources and network functions to fulfill specific service requirements based on fixed Service Level Agreements (SLAs). Usually, several network slices share the same physical infrastructure. Network slices might be instantiated, modified or terminated dynamically or on short notice [2]. Like any network virtualization, network slicing necessitates a strong decoupling of software-based network functions from their underlying network infrastructure, as proposed by the 5G NORMA project [3]. This motivates enhancements in network programmability, flexibility and network sharing and allows to differentiate between three potentially distinct parties involved in network slice instantiation and management: The network infrastructure provider, which is the owner of the physical network elements, like antenna sites, data lines and computational hardware, the mobile service provider managing virtualized and physical resources and the tenant owning one or several network slices [4].

A customer portal provided by a mobile service provider allowing tenants to easily configure and order new network slices is envisioned. The tenant shall receive instant feedback on the feasibility of his or her request accompanied by a cost estimation for the set up and operation of the requested network slice via the portal. If the network slice request is not fully realizable, changes that would make it feasible are proposed to the tenant.

In the virtual mobile network several network slices may have been instantiated and are already running on the physical or virtualized network, also referred to as the substrate network. Therefore, a concept of how to decide whether an incoming network slice request can be accepted is necessary. To answer this question, the available resources in the network and the required resources of the running network slices as well as of the requested network slice have to be considered. Efficient Virtual Network Embedding (VNE) as well as Wireless Virtual Network Embedding (WiNE) algorithms and heuristics for embedding virtual end-to-end network slices into mixed physical networks with both wired and wireless network elements are required. Furthermore, resource and demand predictions are needed. This concerns the research areas of data analysis and Machine Learning (ML).

It is assumed that the mobile communication network is a self-optimizing mobile network (see [5]). Thus, further network optimization efforts regarding network slice deployment can be omitted. But still network slice resource allocation of the RAN and WiFi hotspot resources in an end-to-end communication system is the most demanding part of the overall VNE task [6], since the backbone of each cell usually provides enough resources for the data transport and therefore is usually not the primary bottleneck. Due to that, the main focus will be on the RAN and WiFi resource allocation, which is then extended on the core and transport network.

2 Related Work

Due to the ongoing work on network virtualization, VNE is getting increasing attention in current research. The VNE problem has been proven to be NP-hard [7], but a lot of publications can be found on heuristics for efficient solutions of the general VNE problem. However, there is a lack of concepts for embedding virtual end-to-end networks, i.e., network slices, onto a common mixed wired and wireless physical network. While the core network embedding is very well researched (see [8] for a recent survey), there are only few publications on wireless network embedding (see for instance [9] and [10]). Moreover, the survey paper [6] of Richard et. al compares several recent proposals on mobile network slice embedding. These approaches have in common that they deal with network slice resource allocation at runtime, i.e., focusing on resource partitioning and sharing as well as on network slice isolation. Network slice isolation refers to the problem of assuring that the slice specification is not violated because of changes in another slice. The authors of [6] criticize that the evaluated algorithms are described very vague and that they are based on assigning a certain number of Physical Resource Blocks (PRBs). This way no performance guarantees can be given. Varying chancel conditions make it hard to determine a suitable number of PRBs. Richard et. al [6] state that higher-level variables, such as a percentage of the total resources, should be used instead of PRBs. Considering these results, this work abstracts from PRBs and directly draws on network performance parameters, like throughput and latency.
Nevertheless, several promising algorithms have been published during the last years. For example, Esposito et. al propose a distributed approach for the NP-hard network slice resource allocation problem in form of a consensus-based auction mechanism [11]. Two possible policy configurations are analyzed. In the Single Allocation Distributed (SAD) slice embedding only one bid on one virtual node can be made per auction round. In contrast to that, the Multiple Allocation Distributed (MAD) slice embedding allows to bid on several virtual nodes simultaneously. While MAD has a lower convergence time, SAD results in a better load balancing and is faster in deciding on the feasibility of a network slice embedding. The virtual links are embedded in the next step, after the virtual nodes have been assigned completely. Yang et. al [12] use a karnaugh-map based heuristic for an efficient virtual network embedding. In a first step, the wireless resources are divided by, e.g., Time Division Multiple Access (TDMA) or Frequency Division Multiple Access (FDMA) into resource blocks. In the second step, the virtual networks are assigned to the resource blocks using a karnaugh-map based concept. The authors show that this is a feasible and efficient way of wireless virtual network embedding. However, in both approaches the actual available throughput provided to a network slice, depending on varying channel conditions, as well as important capabilities, like latency and accessibility are not considered. Moreover, they don't allow to analyze possible resource overbookings and assume steady resource provisioning and utilization. In [13] Vassilaras et. al propose an Integer Linear Programming (ILP) model for finding an optimal solution for a simplified network slice embedding problem. They state that the solution with ILP would take tens of minutes. In or-

der to achieve faster run times, an efficient heuristic targeting at a near-optimal solution is required. However, developing such a heuristic remains an open research question. In addition to that, Vassilaras et. al mainly focus on problems related to network slicing at run time, for instance, the authors shed light on the challenges regarding end-to-end latency constraints, heterogeneous networks, multitenancy and slice fairness as well as the issue of dynamic network slicing and online optimization.
In contrast to that, this paper concentrates on challenges regarding the network slice acceptance and does not analyze questions related to the implementation, e.g., network slice isolation and fairness at runtime. Beyond that, the publications above assume steady resource demands, which is not realistic due to user mobility and fluctuations of network utilization, for instance during rush hour. In this paper a three-step process of how to make a well-founded decision on the feasibility of a network slice embedding will be given. The process estimates resource availability and demands in advance of a potential network slice deployment. Such an end-to-end view of allocating resources of a mobile network to network slices including the backbone and allowing potential overbooking of the substrate network is a novel and unique approach in literature.

3 Concept of Network Slice Resource Allocation

In this section, a process to decide whether or not a new network slice request should be allowed in a given substrate network, with already running network slices, is presented. A flowchart of this process on a high abstraction level is shown in Figure 1. For the sake of a fast and efficient decision making on the acceptability of the network slice request, it is advisable to first of all check the technical feasibility of the network slices request (see step 1). That means to determine whether or not the SLAs of the requested network slice could be fulfilled, while in a first step only considering the network infrastructure requirements. For instance, imagine a network slice requiring a very low latency in a certain area with very low tolerance for packet loss and delay. Such an area must be fully covered with cells of 4G technology or higher. However, if this is not the case the network slice request must be considered as being technically not feasible. In such a case, it is of course important to identify the shortcomings of the substrate networks preventing the network slice acceptance. This might help to propose a mitigated version of the network slice request and its accompanying SLAs to the customer. In addition to that, the identified weak points of the current network infrastructure might also serve as an indicator for mid- and long-term network planning for the infrastructure provider. Otherwise, if the technical feasibility check is successful, the remaining resources of the substrate network have to be compared to the estimated required resources of the new network slice in the second step. Due to the large variety of partially interdependent network resources and the large number of requirements of the SLAs, usually based on different Key Performance Indicators (KPIs) to be monitored dur-

ing network slice operation, this is a pretty complex task. The fact that the RAN resources, e.g., throughput and latency, cannot be constantly provided, since the environment bears a lot of potential interferences that can hardly be predicted, makes this allocation task even more complex. For instance, heavy rainfall, foliage and temperature can have a significant impact on the signal quality resulting in a high effect on the radio resources available within the network. But also, the actual resource utilization of the deployed network slices usually underlies some variabilities or even unforeseen behavior. A very conservative approach to cope with these uncertainties would be assuming the guaranteed SLAs as a guideline to determine the maximum resources a network slice is allowed to occupy. However, for the physical network resource estimation such a lower benchmark does not exist. A possible approach would be to analyze the amount of resources provided over time and set up confidence intervals that could be used as a guideline. With such an approach the SLAs should be fulfilled at nearly any time with a certain confidence. However, modern RAN, especially in the Mobile Broadband (MBB) domain, are not capable of such an overprovisioning, since transmission frequencies are a limited resource and the mobile data traffic is ever increasing. Furthermore, the prices per transferred Mbit of data are falling, while the Operating Expenditures (OPEX) for the network infrastructure tend to increase with rising customer needs. To remain profitable an overbooking of the network resources seems to be unavoidable, but such overbookings must be carefully evaluated in advance (see step 3). Before being able to decide whether or not an overbooking of a certain resources can be accepted, it is essential to carefully evaluate the expected available RAN resources. Historical network performance data as well as multi-seasonal time forecasting might be suitable for this task. Also, the running network slices and their past as well as current resource utilization have to be evaluated. Moreover, predictions of the future demand of the network slices can be made often based on insider knowledge, like for example that a certain autonomous driving network slice has been set up only recently and its utilization of the booked network resources is expected to grow considerably in the next months. All those factors as well as experiences from past network slice allocations have to be taken into account to get a prediction of the remaining resources in the network that is as accurate and reliable as possible. This prediction has to be compared with the predictions for the network slice request made on the basis of the SLAs and service utilization specifications provided by the tenant. This comparison will be made with a VNE approach, which is described in more detail below.

As already explained above, the network slice resource allocation can result in overbooking of especially the RAN resources. The risk of violating SLAs can be subdivided into three main classes according to its impact on the service quality. An SLA violation with only minor impact is defined as a disruption that will hardly be noticed by the end-user of the network slice and therefore only is a Quality of Service (QoS) disturbance. Disruptions of service quality that are likely to be noticed by the end-users, i.e., affecting the Quality of Experience (QoE), are seen as major disturbances, while the third and most serious disruption would be network down times. Mobile service providers should define an upper limit of a probability for such a disruption for each of those three categories of network shortcomings.

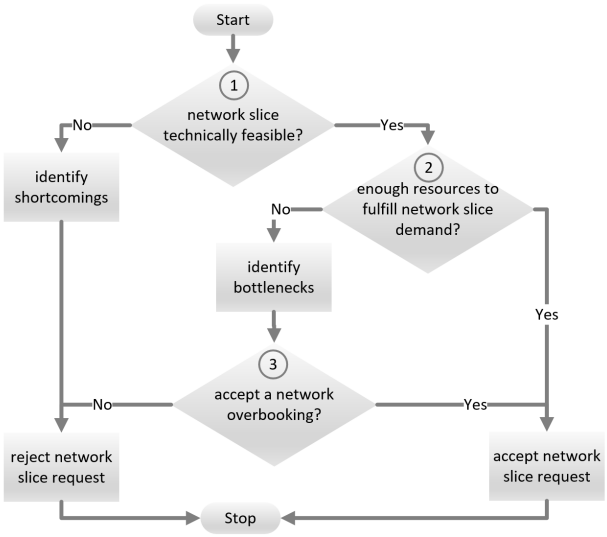

Fig. 1 Process of network slice resource allocation

These boundaries might not only depend on the business goals of the mobile service provider, but also on the SLAs and the penalties that have been agreed on individually for each network slice. Therefore, they might also be network slice specific. To decide on an overbooking issue, the expected probability of violating the SLAs needs to be calculated for each of the tree categories and compared with the maximum tolerated probabilities of violation as defined in the business policies. Obviously, an overbooking can only be accepted when the expected probability of SLA violation in all three categories is less or equal to the maximum tolerated likelihood. When the overbooking bears an acceptable risk of SLA violation, the network slice request can be accepted, otherwise the bottlenecks should be identified in order to support future network planning. Therefore, the resource shortages should be determined, for example too little bandwidth within a certain area.

4 Example of Network Slice Request Analysis

To illustrate the network slice resource allocation process, a quite academic and simplified example shall be given for the RAN part of the network. Imagine that a new MBB slice for the campus of the University of Augsburg shall be added to the existing network. Only one other network slice, also an MBB network slice covering the whole city of Augsburg, is deployed in the network. The network slice request might demand a maximum throughput per User Equipment (UE) of 200 Mbps in Uplink (UL) and an average throughput per UE of 90 Mbps in Downlink (DL) as well as a maximum throughput per UE of 50 Mbps and an average throughput

per UE of 25 Mbps in UL. For the sake of simplicity, a fulfillment rate of 100% is assumed for the campus of the University of Augsburg during daytime and the evening hours (6:00 - 24:00 o'clock), from Mondays to Sundays with up to 1,000 concurrently connected UEs, with an assumed idle time of about 90%. The campus site is covered by three non-overlapping LTE Advanced cells providing a maximum throughput for UL of 300 Mbps and DL of 100 Mbps per UE. Each cell is able to handle up to 2,000 UEs concurrently and provides a total throughput of 3 Gbps in UL and 1.5 Gbps in DL. In order to keep this example as simple and intuitive as possible resource fluctuations are ignored here. The already running MBB city of Augsburg slice, overlapping with the requested university campus slice, is specified as follows: maximum/average throughput for the DL of 50 Mbps/ 45 Mbps and for the UL 25 Mbps/ 20 Mbps with a fulfillment rate of also 100%, 24 hours a day, 7 days a week, with up to 45,000 concurrently connected UEs and the same 90% idle time as in the university campus slice. For checking the technical feasibility of the campus slice request in this simplified example we only have to compare the required and provided maximum UL and DL rates of the three cells installed at the university campus. In the second step the resources have to be evaluated. Therefore, we can focus solely on the three LTE Advanced cells on the university campus. First of all, the actual network utilization of the city of Augsburg MBB network slice at the university campus area during day and evening time have to be analyzed. It is assumed that historical data shows that in average about 2,000 UEs from the city of Augsburg network slice are connected to the campus cells and that the data traffic they produce and their idle times do not differ much from the average of the whole slice. Furthermore, it is assumed that in the future the number of users of the city of Augsburg network slice in this area will decrease by about 50% when the new campus slice is deployed, since most of the students would want to switch to the faster university campus slice. When simply multiplying the number of UEs with their average idle time and their average throughput rate, a total throughput of 4.4 Gbps in DL is expected for the city of Augsburg network slice on the campus area. The tree cells provide an overall throughput for DL of 9 Gbps. However, the new university campus slice requires a total throughput in DL of about 8.8 Gbps, which is much more than the expected available remaining resources of 4.6 Gbps.

In order to answer the question whether or not an overbooking would be acceptable, the probability of violating QoS and QoE for each network slice as well as the expected network down time need to be calculated and compared to the individual business policies of the mobile service provider. A risk estimation as shown in table 1 can be calculated based on confidence intervals for the network slice resource provisioning and utilization. Therefore, the probability distributions of the resource availability as well as of the resource utilization for each resource have to be compared and aggregated per network slice. However, this is quite challenging, since diverse potential disruptions have to be considered and categorized. In this example an overbooking is likely to cause massive violations of QoS and significant QoE violations in the university campus network slice. According to the business policies the requested slice must be rejected.

When thinking about which changes could be made to be able to accept the univer-

Table 1 Risk of SLA Violation and Business Policies

	City Slice		**Campus Slice**	
	Risk	Policy	Risk	Policy
QoS Violation	5%	10%	80%	10%
QoE Violation	1%	1%	60%	3%
Down Time	0.09%	0.1%	0.09%	0.1%

sity campus network slice, even in this very simplified example several alternatives are possible. The most promising options would be to drastically reduce the maximum and average throughput for the DL each UE is allowed to consume, to reduce the required fulfillment rate or to reduce the number of concurrently connected UEs that are admitted.

5 Vision and Challenges of Mobile Network Slice Embedding

The second step of the network slice allocation process, presented in section 3, is the most important step of the decision-making process. In this step, the predicted available network resources of the substrate network have to be allocated on the estimated resources that will be needed by the existing and the new network slice. In order to do this, the physical network resources as well as the network slice resource demands have to be modeled. Both models are represented in form of undirected graphs, consisting of nodes and edges. For the substrate networks the nodes are representing different kinds of network elements, amongst others: UEs, cells, switches, servers and cloud computers. The edges model the data connections between the network nodes. This can be mobile radio connections, transport links as well as core network links. The graphs representing the end-to-end network slices only consist of the end-nodes of the communication links required by the network slice description. In this case, an end-node can be any UE or server communicating via the network slice. For example, in a car-to-x network slices the vehicles, the roadside-infrastructure as well as services providing for example current weather information are the end-nodes of the communication network. However, modeling every single potential end-to-end connection within a network slice won't be possible for the most network slices. Therefore, it is proposed to aggregate mobile UE connections on cell level, i.e., several UEs connected to the same cell and using the same service are aggregated to one end-to-end connection. Obviously, the end-nodes and cell-nodes of the virtual networks, representing the network slices, can be directly mapped onto their according representation in the substrate network. The network slice model would usually not predefine a routing for the data between two end-nodes through intermediate nodes of the substrate network. The mapping of all end-to-end links in the virtual networks on the paths of the substrate network has to be made by a VNE algorithm regarding the resources each link and node element of

the physical network provides and the expected resource utilization of the network slices. The resources involved are the throughput and latency of a communication link, but also CPU power and storage of e.g., cloud and mobile edge computing services. Furthermore, specific RAN capabilities are required, for instance, coverage of a certain area, specific services, like security services, mobility capabilities or guarantees on robustness and availability. Additionally, the duration and time window a network slice is active is an important parameter of the network slice embedding. The models will abstract from resource isolation and sharing which could be done by, e.g., using TDMA or FDMA and considers, for instance, the throughput available in a certain place as a resource to be shared among several tenants instead. Here, the varying available throughput resources depending on the channel quality and the Signal-to-Interference and Noise Ratio (SNIR) an UE experiences, which is due to the distance between the transmitter and receiver as well as obstacles like buildings, hills or vegetation and interferences by other antennas, plays an important role. The second most important resource of the physical network model is the data packet latency. The minimum time a data packet needs to be transported from the sender to the receiver is called the end-to-end latency. It depends on the used technologies and the network infrastructure hardware. The delay is the extra time a packet needs if the channels are crowded and the optimal latency cannot be achieved. However, the end-to-end latency plus delay, i.e., the actual time a data packet needs to be transported from the sender to the receiver, is also usually referred to as latency.

Network slice embedding in an end-to-end mobile network introduces specific challenges. Compared to ordinary VNE instances the nodes and links of the substrate network graph are annotated with a quite large number of different resources, capabilities and further parameters. Since the network resources can have a high variability it is helpful to annotate the graph not only with the average expected values of the resources to be allocated, but with a whole probability distribution of the resource availability. This allows to determine confidence intervals for the availability of the required amount of resources and to give the best possible indication on the probability or, in other words, the risk that resource shortages occur during network operation. This is also useful when analyzing the degree of overbooking and the associated risks in the third step of the network slice resource allocation process. Additionally, some resources are tightly coupled to each other, for instance the packet latency and delay of a mobile communication link is highly dependent on the packet size and the available throughput, since the amount of data that can be sent in parallel depends upon the available bandwidth. It is clear that a delay of zero can only be achieved when there are enough remaining throughput resources for instant data transmission using the maximum feasible bandwidth. In order to enable a correct resource allocation such interdependencies have to be considered in the network resource model. Beyond that, the vision of a customer portal allowing a tenant to configure new network slices and receive instant feedback on the feasibility of his or her request requires a very fast and efficient solution of the network slice embedding and the risk analysis. Consequently, the computational complexity of the network slice embedding heuristic has to scale for large and complex problem instances.

6 Conclusion and Outlook

Network slicing is seen as one of the key features to cope with the diverse requirements arising with the fifth generation of mobile networks. But yet, network slice resource allocation is still an unsolved issue. In this paper, a process for network slice resource allocation that allows to decide whether to accept or to reject an incoming network slice request taking into account the individual business policies and risk tolerance of the mobile service provider is presented. The approach considers network resource overbooking as a way of profitable network operation. Additionally, a vision of an end-to-end network slice resource allocation based on VNE has been presented.
The next steps will be evaluating suitable VNE and WiNE algorithms, methods of network resource estimation and network slice resource utilization as well as developing concepts for the network disruption risk estimation. Future steps will include the identification of bottlenecks and shortcomings of the network and concepts for proposing changes that transform a non-viable network slice request into a feasible one.

References

1. Nokia, "Dynamic end-to-end network slicing for 5G", Espoo, Finland, 2016.
2. 3GPP, "Study on management and orchestration of network slicing for next generation network", TR 28.801 V15.0.0T, 2017.
3. 5G NORMA, `https://5gnorma.5g-ppp.eu`, accessed: April, 22nd, 2018.
4. 5G NORMA, "5G NORMA network architecture - intermediate report", Deliverable D3.2, 2017.
5. S. Hmlinen et al., "LTE self-organizing networks (SON), John Wiley & Sons", Great Britian, 2012.
6. M. Richart et al., "Resource Slicing in Virtual Wireless Networks: A Survey" in IEEE Transactions on Network and Service Management, VOL. 13, NO. 3, 2016.
7. D. Andersen, "Theoretical approaches to node assignment", Computer Science Department, 2002.
8. A. Fischer et al., "Virtual Network Embedding: A Survey" in IEEE Communication Surveys & Tutorials", Vol. 15, No. 4, 2013.
9. R. Riggio, "Scheduling Wireless Virtual Networks Functions" in IEEE Transactions on network and service management, Vol. 13, No. 2, 2016.
10. I. Tsompanidis, A. Zahran and C. Sreenan, "A Utility-based Resource and Network Assignment Framework for Heterogeneous Mobile Networks" in 2015 IEEE Global Communications Conference, San Diego, 2015.
11. F. Esposito, D. Di Paola and I. Matta, "A general distributed approach to slice embedding with guarantees", 2013 IFIP Networking Conference, Brooklyn, NY, 2013, pp. 1-9.
12. M. Yang, Y. Li et al., "Karnaugh-map like online embedding algorithm of wireless virtualization", The 15th International Symposium on Wireless Personal Multimedia Communications, Taipei, 2012, pp. 594-598.
13. S. Vassilaras et al., "The algorithmic aspects of network slicing" in IEEE Communications Magazine, vol. 55, no. 8, pp. 112-119, 2017.