Reliable Software-Defined RAN Network Slicing for Mission-Critical 5G Communication Networks

Caner Bektas, Stefan Böcker, Fabian Kurtz, Christian Wietfeld
Communication Networks Institute, TU Dortmund University, Otto-Hahn-Strasse 6, 44227 Dortmund
Email: {caner.bektas, stefan.boecker, fabian.kurtz, christian.wietfeld}@tu-dortmund.de

Abstract—Emerging Industrial Internet of Things (IoT) applications impose challenging and diverse requirements on underlying communication infrastructures. Traditionally, each vertical industry utilizes dedicated communications networks. In contrast, the 5th Generation of Mobile Communication Networks (5G) aims to fulfill Quality of Service (QoS) requirements for various vertical industries. This is achieved by deploying several, virtually dedicated networks, known as slices, on top of a unified, physical communication infrastructure. While various solutions have been presented for Core Network Slicing, the here presented approach aims to provide an end-to-end solution by allocating Radio Access Network (RAN) resources, realizing Slicing for mission-critical applications via a novel scheduler design. A proof of concept is provided by way of detailed empirical evaluations, based on IoT scenarios from the energy sector (i.e., Smart Grids). The proposed RAN Slicing solution is shown to reliably sustain service guarantees of critical applications, while coexisting with non-critical services. Application-dependent, dynamic, inter-slice resource sharing enables an efficient use of available RAN spectrum. Finally, we demonstrate dynamic adaptation of slices to channel quality, ensuring reliable operation of Industrial IoT.

I. INTRODUCTION

Industrial Internet of Things (IoT), as a field with diverse service types and thus, challenging communication requirements, is highly dependent on ultra-reliable Information and Communications Technologies (ICTs). The energy sector represents one such mission-critical vertical industry, increasingly relying on robust communication networks due to the shift towards sustainable energy generation. The volatile feed-in of renewable energy resources and controllable loads (e.g., Electric Vehicles) increasingly endangers energy grid stability. To counteract this development, a rapid transition to Smart Grids, capable of fulfilling continuously more demanding Quality of Service (QoS) requirements, is in progress. Efficient operation of these systems necessitates robust control and monitoring, which in turn depend on reliable ICT. In this context, heterogeneous communication technologies, as well as a variety of network and operator models, are in discussion. A dedicated communication infrastructure tailored to the demands of Smart Grids, is associated with high costs and poses a challenge for utilities, typically lacking network operation expertise. The opposite solution, i.e. harnessing public ICT infrastructures operated by Telecommunication Operators (TelCos), has the benefit of already being pre-deployed and reduced utility efforts. However, such TelCo-operated public networks have the disadvantage of sharing resources with other user groups, endangering the reliable fulfillment of service guarantees, potentially impacting the safe operation of critical infrastructures. A combination of both models is seen as promising solution ([1]). Currently, QoS in shared wireless infrastructures is mainly restricted to basic services like Voice over LTE (VoLTE) in 4G. In contrast, the 5th Generation of Mobile Communication Networks (5G) aims at integrating vertical markets, industries and the IoT. Hence it strives to provide new functionalities for realizing shared communication infrastructure usage and operation. Specifically, 5G introduces a paradigm shift via the so-called Network Slicing ([2] (c.f., Figure 1). It supports establishing multiple logical (i.e., virtual) networks, called slices, on basis of a single, shared (public) physical communication infrastructure. Each slice can be tailored efficiently and flexibly to the specific needs of an application and its corresponding Service Level Agreement (SLA). Thereby, Slicing enables reliable QoS guarantees for diverging services and tenants, through application-dependent, dynamic resource allocation. Hence, Network Slicing facilitates billing dependent on actual resource usage, while being highly efficient in its allocation. This paper focuses on evaluating end-to-end 5G Radio Access Network (RAN) Slicing for the requirement profiles of IoT applications like Smart Grids.

The remainder of this work is structured as follows: First, related work is discussed in Section [II]. Next, Section [III] introduces the proposed, scheduling-based approach to RAN slicing. Our Software-Defined Radio (SDR)- and Software-
Defined Networking (SDN)-based laboratory setup, employed for empirical evaluations, is presented in Section [V]. Subsequently, Section [V] details the analyzed IoT scenarios, as well as measurements results showcasing the proposed RAN Network Slicing solution’s capabilities. Finally, a conclusion and outlook are given in Section [VI].

II. RELATED WORK

Due to the importance of Network Slicing in the context of 5G and the IoT, a variety of related works exist. However, as is shown by Kaloxyllos in [3], the majority of publications either concentrate on the specification of interfaces, architectures and management strategies or aspects of the Core Network (CN). The challenges of RAN slicing, as addressed in this work, are identified in [4]. Efficient, flexible radio spectrum sharing is a key aspect of RAN slicing and studied analytically in [5] and [6]. While the former presents a Markov model for studying e.g. blocking probabilities when providing guaranteed bitrate services, the latter analyzes the slicing of resources based on heterogeneous non-orthogonal multiple access. The authors of [7] introduce and evaluate a distributed algorithm for dividing RAN resources among slices. While insightful results are achieved, the aforementioned works focus on different aspects of RAN slicing. A more similar approach is employed by the author's of [8], which simulate the sharing of radio resources based on earliest deadline first scheduling. An empirical evaluation within a large scale testing environment is conducted by [9]. However, no measurements or technical implementation details are provided. In contrast, this work not only introduces a novel strategy for RAN slicing, but gives an empirical evaluation of its performance.

III. PROPOSED SYSTEM ARCHITECTURE: RAN SLICING BASED ON GUARANTEED CHANNEL ACCESS

In this section, our proposed and implemented RAN Slicing scheduler is described in detail. First, an introduction to basic terminology is conducted, on the example of a Round Robin scheduler. Next, our solution, extending the Round Robin mechanism to support Network Slicing, is presented.

A. Basic Scheduler Terminology and Introduction to the standard Round Robin Scheduler

As our proposed scheduler is based on a Round Robin scheduler, a brief introduction on the behavior of a standard Round Robin scheduler is required. Typically, a standard Round Robin scheduler allows all User Equipments (UEs) to access the channel in an alternating manner. For instance, when two UEs want to send data, in the first Transmission Time Interval (TTI), UE A gets to send first. Following that, in the second TTI, UE B can access the channel first, and so on. The amount of data in Long Term Evolution (LTE) and 5G, that can be sent per TTI, is the so-called Transport Block Size (TBS), which varies with the allocated number of Physical Resource Blocks (PRBs) and used Modulation Coding Scheme (MCS) [10]. This means that, when $N$ UEs in a cell utilize all their available resources (PRBs), the resulting data rate for a UE $x$ is:

$$\text{RoundRobinDR}(x) = \frac{1}{N} \cdot \frac{TBS(x)}{TTI = 1 \text{ms}} [\text{bps}]$$

(1)

This results in an equally distributed data rate among all UEs in a standard Round Robin scheduler.

B. Extension of the Round Robin scheduler to support 5G RAN Slicing

To guarantee data rates, e.g., for mission-critical Smart Grid networks, our proposed and implemented RAN Slicing extends the Round Robin scheduler by providing guaranteed channel access for UEs and network slices, respectively. The operating principle can be described best with an example, as depicted in Figure 2. There, two slices A and B are configured with weights 3 and 2, respectively. In this case, the higher value results in a higher share of the PRBs per second and thus in a (potentially) higher data rate (dependent on channel quality / MCS). According to our scheduling algorithm, slice A gets to
send data first in 3 consecutive TTIs, and after that, slice B can send data first in 2 consecutive TTIs. This results in a repeating cycle of 5 TTIs, which can be seen in Figure 2. There, the left side shows situation 1, in which both slices utilize all their available resource blocks, resulting in the maximum TBS per TTI. In situation 2 (right side), slice A only needs half of the guaranteed PRBs. Our algorithm then proceeds on to the next slice in order of their weight, so that slice B can use all remaining PRBs of slice A. In TTI 3 and 4, slice B utilizes its guaranteed 2 TTIs of PRBs. In summary, our algorithm provides guaranteed data rates for all network slices and this does not lead to unused PRBs (i.e. evaluations in Scenario 1 / Section V-A). Derived from Equation 1, the minimal and maximal data rate of our proposed scheduler can be described analytically. For this, assume one UE per slice and let \( N \) be the number of slices and \( W \) the assigned weight value. The minimal data rate of slice \( x \) can be described as follows:

\[
\text{MinimalSliceDR}(x) = \frac{W(x)}{\sum_{i=1}^{N} W(i)} \cdot \frac{TBS(x)}{TTI = 1\text{ms}} \text{bps} \tag{2}
\]

This equation describes the resulting data rate of each slice \( x \), when every slice is utilizing the maximum available data rate, using all available PRBs per TTI. As every slice gets all the available PRBs in one TTI when it needs to, the only remaining variable is MCS. This means that changing MCS in a slice also requires a change in slice weights to guarantee the same minimal data rate as before. This relation is shown in the empirical evaluations in Scenario 2 (Section V-B).

As for the maximum data rate, the equation is as follows:

\[
\text{MaximalSliceDR}(x) = \frac{TBS(x)}{TTI = 1\text{ms}} \text{bps} \tag{3}
\]

This occurs when slice \( x \) is using all bandwidth available. Assuming MCS 9 as well as 22 PRBs for both slices, a TBS of 3496 Bit \([10]\) results and via Equation 2, we get approx. 2.1 Mbps for slice A and 1.4 Mbps for slice B. Shown data rates are not provided on the application layer, because the overhead is not included (retransmissions and packet headers).

IV. SOFTWARE-DEFINED RADIO / SOFTWARE-DEFINED NETWORKING-BASED LABORATORY SETUP

Figure 3 depicts a schematic diagram of our laboratory setup. Our base station consists of two main parts: An LTE stack based on \( srsLTE \) \([11]\) (Version 18.12), including our RAN Slicing implementation, as well as the SDN-based CN presented in \([12]\) and \([13]\) (using \textit{NextEPC} \([14]\)). The latter is connected to our SDN / Network Function Virtualization (NFV) Node hosting all \textit{iperf} \([15]\) servers, which undertake the data rate and latency measurements. In this setup UEs also contain a full LTE stack (i.e. base station). To create data traffic representative of the role assigned to an UE, e.g. smart meters, \textit{iperf} is used. Radio components are connected via SDRs (\textit{Ettus USRP B210}) through a Radio Frequency (RF) combiner. The measurement setup is time-synchronized via Precision Time Protocol (PTP) (mean clock offset: \(<10\mu s\)) and is based on the \textit{tinyLTE} architecture \([16]\).

V. EMPIRICAL EVALUATIONS BASED ON SCENARIOS USING 700MHz 5G RAN

In the following subsections, two scenarios are presented, which were designed and evaluated in our laboratory setup (c.f. Section IV) to show the working principle as well as different features of our RAN Slicing implementation. All evaluations are conducted empirically using uplink measurements at band 28 (700 MHz) 5G frequencies (channel bandwidth: 5 MHz). Measurements are repeated at least 100 times to achieve results of statistical significance. The number of devices does not correspond to the number of UEs employed. Aggregate data rates of multiple devices within a slice are recreated via a single device emulating the corresponding rate.

A. Scenario 1: Dynamic mission-critical slice prioritization

In Scenario 1, three network slices, as depicted in Figure 4 are configured, using examples from the energy industry: Highly mission-critical \textit{Monitoring and Control} (e.g. Wide Area Monitoring Protection and Control (WAMPAC), green), mission-critical \textit{Sensing} (e.g. Smart Metering, orange), as well as a non-critical \textit{Best Effort} slice representing all other data rate-hungry users (c.f. Enhanced Mobile Broadband (eMBB), blue). Additionally, a hierarchy between the two mission-critical slices is established, putting the \textit{Monitoring and Control} slice slightly higher than the \textit{Sensing} slice. However, both these slices are ranked higher than the \textit{Best Effort} slice. This relation is also reflected in the weight values of the slices, which were configured in correspondence to their criticality (minimal demanded data rate). The other users, which transmit non-critical data traffic such as multimedia streaming, always get the remaining resources not utilized by the mission-critical slices. The weights were chosen using Equation 2 and an empirically evaluated factor of 0.864 (to convert the LTE data rate to the measured end-to-end data rate, due to packet headers and retransmissions), providing the depicted aggregate data rates. Additionally, in the \textit{Monitoring and Control} slice, a distinction between the two states \textit{normal} as well as \textit{critical} is shown, where the latter describes a situation where higher

![Fig. 3. Schematic diagram of the laboratory setup, based on real hardware using SDR and SDN.](image-url)
traffic is demanded due to a failure in the energy grid. It is notable that the weight of the Monitoring and Control slice is chosen according to the critical operation state, as our scheduler does not waste the unused resources in the normal operation. These relations will be analyzed in the empirical evaluations in this section. At the bottom of Figure 4, the measurement sequence is depicted schematically over the measurement time. There, the other users always utilize all available PRBs during the whole measurement. Starting at checkpoint 1, the Sensing slice begins to transmit its 0.7 Mbps of data, whereas devices in the Monitoring and Control slice begin at checkpoint 2 with 0.24 Mbps (normal) and 1.1 Mbps (critical), respectively. The goal of this scenario is to show the hierarchical and dynamic data rate allocation as well as service prioritization of our Network Slicing system in the time domain. Moreover, highly effective inter-slice resource sharing shall be demonstrated. In Figure 5, one sample measurement from the evaluations is presented. On the y-axis, the end-to-end application data rate of each slice is shown in Mbps (colors c.f. Figure 4). The elapsed measurement time, including the two mentioned checkpoints 1 and 2 from Figure 4, is drawn along the x-axis. Additionally, the dashed lines represent the required service data rate needed for each slice, according to their QoS requirements. In this plot, the highly mission-critical Monitoring and Control slice is operated in the normal state. Slice data transmissions are sequenced within three states, beginning with the Best Effort users utilizing all available PRBs (green). This results in the full possible channel data rate of 2.9 Mbps (for given MCS of 9), which shows that unused PRBs of mission-critical slices can effectively be shared with non-critical slices. However, at checkpoint 1, smart meters in the Sensing slice begin to send their required data with a rate of 0.7 Mbps (orange). Our scheduler instantly reacts with a reduction of the low priority slices’ data rate. The same instant reaction can be observed at checkpoint 2, where the highly mission-critical Grid Control and Monitoring slice utilizes its required data rate for normal operation (blue). Comparatively, Figure 6 shows the same results in the critical operation of slice Monitoring and Control (red). There, a considerably higher data rate is required. The reaction of our Network Slicing system can be seen at checkpoint 2: Data rate is deducted from the non-critical users (green) and not from the Sensing slice (orange), as it is higher prioritized. This shows, that our scheduler achieves a hierarchically structured inter-slice resource sharing.

B. Scenario 2: Dynamic slice priority control and channel quality adaptation

Similar to Scenario 1, the second scenario is split into two situations, as depicted in Figure 7. Scenario 2 as a whole contains 2 slices, the first slice being a mission-critical Sensing (Smart Metering) and second one again being a Best Effort slice with other users, with the same attributes as described
in Section V-A. However, the Sensing slice is split into two separate virtual sub-slices to represent the two channel conditions depicted in Figure 7. There, the two situations described above are shown on the left and right, respectively. On the left side, the majority (70%) of Smart Metering devices are at the cell center (MCS 8) and have better channel conditions than the minority (30%) of devices at the cell edge (MCS 6). On the right side, the conditions are exactly reversed, so that the weights of the virtual sub-slices have to be adapted according to the channel quality and the number of devices in the cell. The other users, which are not depicted in Figure 7, experience an MCS of 8 in both situations. The resulting dynamic behavior of our Network Slicing system will be shown in the evaluations of this scenario. For this, the aggregate data traffic of all devices is generated and sent over the RAN, based on Table I. Results obtained from this scenario are shown in Figure 8. The two aforementioned situations from Figure 7 are plotted on the x-axis. The share of available PRBs for each slice is depicted in the y-axis in %. When the majority of devices are placed in the cell center, 44.44% and 33.33% for the cell center and cell edge devices are needed for the reliable data transmission of the sum data rate (1.6 Mbps), respectively. 22.22% of data remains for the Best Effort slice (MCS 8). Please note that the Best Effort users always try to use all available resources, which again proves that our scheduler works as designed, as the mission-critical Sensing slice reliably reaches its required sum data rate (c.f. colored boxes). Moreover, a low median latency of 19.2 ms and 13.8 ms are guaranteed for the Sensing slice, whereas the Best Effort slice experiences high latencies of up to 977 ms (not depicted). As the number of cell center and cell edge devices is reversed (left bar), the Sensing slice still gets its required sum data rate and low latencies. However, 9.73% more resources are required to fulfill these QoS constraints, as our scheduler has to adjust the slice weights, adapting to the worse overall channel conditions. This demonstrated two relations: Firstly, our implemented RAN Network Slicing system reliably adjusts its configuration to fulfill all requirements of mission-critical slices. Secondly, it is not enough to simply reserve PRBs, and thus bandwidth, for a given slice to satisfy service demands, but rather incorporate the location of devices affecting their channel quality (and throughput).

VI. CONCLUSION AND OUTLOOK

In this paper, we presented the design of our scheduler implementing a 5G Network Slicing system in the RAN, developed on a SDR-based full LTE stack. Then, empirical evaluations based on Industrial IoT scenarios were conducted, using examples from the energy sector. Measurement results show that our Network Slicing realization can reliably support the harmless coexistence of different 5G service types, by providing efficient inter-slice resource sharing and mission-critical service prioritization. Moreover, channel conditions are sufficiently considered to provide reliable data rates and latencies for mission-critical slices, even in variable channels. Also, we show that the location, and thus the channel quality, of mission-critical devices has to be taken into account when provisioning slices, as bandwidth (or PRB) allocation cause unstable data rates as channel conditions change.

Additional work can be done in improving our Network Slicing system. For instance, external data, such as weather (e.g., renewable energy) or device density predictions (e.g., derived from crowd sensing), can be used in combination with Machine Learning (ML) to optimize the weights chosen for a given slice. Moreover, higher TBSs and shorter TTIs in future 5G releases can be evaluated, as our approach scales with mobile communication network improvements.

<table>
<thead>
<tr>
<th>Slice</th>
<th># Devices</th>
<th>Sub-Slice Data Rate [Mbps]</th>
<th>Slice Data Rate [Mbps]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart Metering (70%)</td>
<td>220</td>
<td>1.1</td>
<td>1.6</td>
</tr>
<tr>
<td>Smart Metering (30%)</td>
<td>100</td>
<td>0.5</td>
<td>1.6</td>
</tr>
<tr>
<td>Other Users</td>
<td>Unspecified</td>
<td>Max. possible</td>
<td>Max. possible</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENT

This work has been supported by IDEAL project (grant agreement no 03ET7557A) funded by the German Federal Ministry for Economic Affairs and Energy (BMWi), the federal state of Northrhine-Westphalia and the “European Regional Development Fund” (EFRE) 2014–2020 in the course of the OPUS project under grant number EFRE-0800885, as well as by the Ministry of Economic Affairs, Innovation, Digitalisation and Energy of the State of North Rhine-Westphalia (MWIDE NRW) along with the Competence Center 5G.NRW under grant number 005-01903-0047 and the Collaborative Research Center SFB 876 “Providing Information by Resource-Constrained Analysis”, project A4.

REFERENCES


