Latency-Based 5G RAN Slicing Descriptor to Support Deterministic Industry 4.0 Applications

Jan García-Morales, M. Carmen Lucas-Estañ, and Javier Gozalvez
UWICORE Laboratory, Universidad Miguel Hernández de Elche (UMH), Elche 03202, Spain
Email: {jan.garcia, m.lucas, j.gozalvez}@umh.es

***Abstract*—5G networks can support the development of the Industry 4.0. To this aim, 5G must be able to guarantee the deterministic latency requirements that characterize many industrial applications. This objective can be achieved using network slicing, a novel 5G paradigm that exploits the softwarization of networks to create different logical instances of the network over a common network infrastructure. Each instance is configured to support specific applications. Slicing can be applied at the Core Network or at the Radio Access Network (RAN). This study focuses on RAN slicing since the RAN typical accounts for a large part of the end-to-end delay. RAN slicing splits (and configures) resources at the RAN level between the slices in order to adequately serve nodes with a particular profile. This includes identifying the necessary radio resources per slice. To date, most proposals define slices in terms of the number of required radio resources. While this descriptor can account for bandwidth or rate requirements, it does not adequately reflect the latency requirements characteristic of many Industry 4.0 applications. This paper proposes a novel latency-based RAN slice descriptor and demonstrates that the new descriptor improves the capacity of RAN slicing to meet the latency requirements of Industry 4.0 applications with deterministic periodic traffic.**

Keywords—RAN slicing, Industry 4.0, 5G, slice creation, latency, deterministic.

I. INTRODUCTION

The digitalization of factories will allow creating smarter and reconfigurable manufacturing environments for safer, more adaptive and zero-defect production [1]. Progressing towards this vision requires communication networks capable to sustain data-intensive services while ubiquitously guaranteeing low latency and reliable connections. This includes mobile connections with robots, vehicles or workers among others. In this context, 5G has been identified as a key technology enabler for the digitalization of factories and the development of the Industry 4.0 or Factories of the Future (FoF). The 5G Alliance for Connected Industries and Automation (5G-ACIA) and the 3GPP have defined and classified Industry 4.0 use cases that could be supported by 5G [2]. This includes use cases such as factory control, monitoring, process automation and maintenance. Industry 4.0 use cases can have very distinct communication requirements and are usually classified into three different traffic classes [2]: deterministic periodic, deterministic aperiodic and non-deterministic (periodic or aperiodic). According to [2], deterministic periodic traffic is the most common industry traffic class, and relates to use cases such as motion control, control to control communication, mobile robot communication, process automation, and augmented reality among others. Deterministic periodic communication stands for periodic communication with stringent requirements on timeliness of the transmission. For example, motion control requests a maximum of 2 ms latency, control-to-control applications have latency requirements equal to 4 ms, and factory automation requests a maximum latency between 0.25 and 10 ms.

5G networks are being designed to include unprecedented network flexibility to guarantee the diverse Quality of Service (QoS) requirements needed to support vertical industries, including manufacturing. Such flexibility is achieved through the softwarization of networks and the introduction of network slicing [3], [4]. Network slicing is a novel 5G paradigm aimed to simultaneously support various services with different requirements over a common physical network infrastructure. Network slicing exploits the softwarization of networks to create different logical partitions or instances (properly isolated) over the common network infrastructure. A slice is formed by a set of network functions, radio access technology settings and resources (including computing, storage, networking and radio resources) that are tailored to support specific applications. Network Slicing can be applied at the Core Network (CN) or at the Radio Access Network (RAN). RAN slicing is key for Industry 4.0 applications that require low and deterministic latency since the RAN typical accounts for a large part of the end-to-end service delay [5]. This paper focuses then on the design of RAN slicing solutions capable to support the latency requirements of Industry 4.0.

RAN slicing is in charge of splitting and configuring resources at the RAN level between the slices. This includes identifying the necessary radio technology, communication mode and radio resources per slice [6]. Each slice must be defined and configured to adequately serve nodes with a particular QoS profile. Correctly defining the slices is hence critical to ensure that RAN slicing can successfully serve nodes with distinct QoS profiles. To date, most proposals define slices in terms of required number of radio resources [6]–[8]. This slice descriptor can account for the bandwidth or data rate requirements of services, but not for the latency requirements that are critical for Industry 4.0 applications with deterministic traffic. In this context, this paper proposes a novel slice descriptor that takes into account latency requirements. The impact of the proposed latency-based descriptor is illustrated in this paper for deterministic periodic traffic. To this aim, we compare the performance achieved when slices are defined in terms of only the number of radio resources and when they are defined considering the number of radio resources and the latency-based descriptor. The results demonstrate that our proposal significantly improves the capacity of RAN slicing to meet the latency requirements of Industry 4.0 applications.

978-1-7281-0302-0/19/$31.00 ©2019 IEEE

II. Related Work

RAN slices have been mostly defined to date in terms of the number of radio resources necessary for a slice to adequately serve its nodes [7]–[10]. Kokku et al. designed and implemented in [7] a Network Virtualization Substrate (NVS) for the effective virtualization of wireless resources in cellular networks. The study shows that slices can be defined in terms of bandwidth or resources, and introduces a slice creation scheduler that allows for their simultaneous existence. [8] extends the NVS concept to design effective RAN slicing solutions that operate with resources from multiple base stations (BSs). [9] introduces the concept of a Slice Broker that offers slices as a service (SlaaS). The broker initially reserves an amount of resources per slice based on the request from services. It then monitors the traffic per slice and augments the allocation of resources if necessary. This reactive approach can ultimately assign the exact amount of resources needed but incurs in some delays that might not be tolerable by time-critical services such as those found in Industry 4.0. A proactive approach is proposed in [10] where RAN slicing is considered to support haptic communications. The proposal computes the size of slices periodically and uses a dynamic queuing system to allocate resources to nodes in order to meet latency requirements. However, such latency requirements are not considered in the process to create the slices.

Recent studies propose creating slices considering bit rate requirements when considering different types of traffic and service. For example, [6] considers a combination of resource-oriented and rate-oriented parameters that limit the number and characteristics of the resources per slice. Resource-oriented parameters can include for example occupation levels of the radio resources. Rate-oriented parameters can include limits on the aggregate bit rate. [11] considers Guaranteed Bit Rate (GBR) services and computes the amount of resources necessary per slice based on the aggregate GBR requirements. [12] proposes slices for elastic and inelastic traffic. Inelastic nodes require a certain fixed throughput demand which needs to be satisfied at all times while elastic nodes only require that the expected average throughput over long time scales is above a certain threshold. The proposal in [12] can achieve certain delay guarantees to inelastic traffic by ensuring that the throughput demand is guaranteed at all times. However, delay requirements are not embedded directly in the slice creation process and hence these requirements cannot be fully guaranteed. To the authors' knowledge, none of the existing studies directly consider delay or latency requirements when creating slices. This complicates the capacity of RAN slicing to adequately guarantee the stringent latency requirements that characterize time-critical Industry 4.0 applications. To overcome this limitation, we propose a novel latency-based slice descriptor and we demonstrate its utility to support time-critical Industry 4.0 applications.

III. Slice Descriptors

This study proposes to define slices with two descriptors. The first one is the most commonly used to date, and is based on the amount of resources needed to satisfy the bandwidth or rate requirements of the supported services. This descriptor is referred to as the Slice Size. The second descriptor is proposed in this paper, and is a novel latency-based descriptor that accounts for latency requirements of the supported services. This descriptor is referred to as the Slice Shape. This section derives analytical expressions for both descriptors in the case of deterministic periodic traffic.

A. Slice Size

Without loss of generality, we assume the Long-Term Evolution (LTE) radio interface. In LTE, a wideband channel is divided into sub-frames and Resource Blocks (RBs). The duration of sub-frames is 1ms and is equal to the Transmission Time Interval (TTI). A RB is the smallest unit of frequency resources that can be allocated to a node. Each RB is 180kHz wide in frequency and consists of 12 adjacent subcarriers of 15kHz. In the time domain, each RB occupies a full TTI. An LTE channel can then be represented as time/frequency resource grid map where the unit is an RB. The RAN slicing scheme must allocate RBs per slice.

We consider services with deterministic periodic traffic that generate packets every T_p seconds (T_p is the transmission period). We define the Slice Size as the amount of resources within the transmission period that must be reserved for a slice to satisfy the rate required by a service. The packets must be received before a deadline D_s. The data rate (in bps) required by a node u to transmit a payload of L_u bits before D_s is:

$$R_u = \frac{L_u}{D_s} \tag{1}$$

Following [7], we define R_u^{eff} as the effective transmission rate of node u or throughput that node u will experience per assigned RB. This throughput is a function of the experienced signal-to-interference-plus-noise ratio (SINR) and the reliability required by the application that is here represented in terms of the Block Error Rate (BLER):

$$R_u^{\text{eff}}(\gamma_u) = \frac{T_c(\gamma_u)}{D_s}(1 - BLER) \tag{2}$$

where γ_u is the SINR experienced by the node u on a RB, and $T_c(\gamma_u)$ represents the transport block size (TBS in bits) that can be transmitted over a RB. This TBS depends on the modulation and coding scheme (MCS) selected to transmit the data. MCSs with higher error correction capabilities can transmit a smaller TBS over a higher number of RBs but can operate over lower SINR levels. The MCS is selected based on the experienced SINR and the BLER required to deliver the data block $T_c(\gamma_u)$. We select the MCS with larger TBS size that guarantees the BLER for the experienced SINR. The MCS is selected using the lookup table specified in [13] for the Extended Pedestrian A model (EPA). This lookup table maps the SINR to the MCS necessary to guarantee a target BLER. Using this lookup table, we obtain the value of $T_c(\gamma_u)$ for the experienced SINR γ_u by node u. We compute then the amount of RBs required by u to transmit L_u bits before D_s as ($\lceil x \rceil$ denotes the ceil operator):

$$J_u(\gamma_u) = \left\lceil \frac{R_u}{R_u^{\text{eff}}(\gamma_u)} \right\rceil \tag{3}$$

A slice should serve a group of nodes with similar QoS requirements. The Slice Size (K_s) or total number of radio resources required by a slice s to serve M nodes during a transmission period is then:

$$K_s = \sum_{u=1}^{M} J_u(\gamma_u) \quad (4)$$

The SINR level used to compute K_s should then not be an instantaneous SINR level since it does not adequately reflect the SINR that nodes can experience during the complete transmission period. The node measures its experienced SINR every 1ms and stores the SINR values of the last second. We then use the 25th-percentile of the measured SINR values.

B. Slice Shape

The Slice Shape descriptor identifies the number of TTIs within the transmission period over which the K_s resources have to be reserved to meet the latency requirements of the service to be supported. For deterministic periodic traffic, we must guarantee that all K_s resources are available between the time a new packet is generated (this time is known for periodic traffic) and the latency deadline D_s. The Slice Shape descriptor identifies the TTIs over which the K_s resources should be reserved. Fig. 1 illustrates an example of the size and shape of a slice for deterministic periodic traffic. The grid represents the RBs in time and frequency domains. The example represents the case where a node requires a slice size of four RBs in each transmission period. The slice size must be contained within a time window of length D_s from the start of every transmission at $t_0 + kT_p$ where t_0 is the time at which the first transmission starts and k is a non-negative integer. Fig. 1 represents an example with different slice shapes in each transmission period, but where both shapes reserve K_s radio resources within D_s from the start of each transmission.

We define $L_{t,s}$ as the amount of radio resources allocated to slice s in TTI t. To meet the latency deadline D_s, the slice must be created so that:

$$\sum_{t=t_i}^{t_i+D_s-1} L_{t,s} = K_s \quad (5)$$

where $t_i \in \{t_0, t_0 + T_p, t_0 + 2T_p, \ldots\}$ and D_s is expressed as an integer number of TTIs.

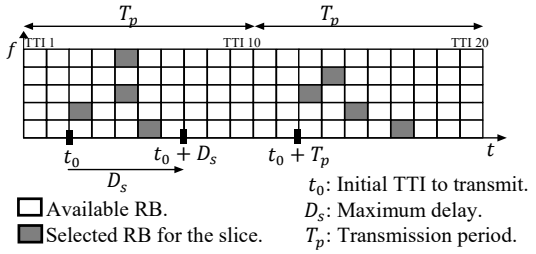

Fig. 1. Slice size and shape for deterministic periodic traffic.

IV. Evaluation

This section compares the performance achieved when slices are defined only in terms of size (referred to as benchmark) and when they are defined in terms of size and shape (referred to as proposal). In the first case, the number of RBs per slice based on eq. (4) must be guaranteed while the application is being executed, i.e. during T_p in the case of deterministic period traffic. The set of possible slices that guarantee K_s RBs in T_p is equal to:

$$N_L = \binom{N_b T_p}{K_s} \quad (6)$$

where N_b is the number of RBs per TTI and $\binom{N_b T_p}{K_s}$ represents the number of combinations of K_s RBs within a resource grid including $N_b T_p$ RBs.

Our proposal defines slices considering K_s based on eq. (4) and the latency condition expressed in eq. (5). We then only consider out of all possible N_L solutions in eq. (6) those that include the K_s RBs in a time window starting at t_i and with length D_s. The set of possible solutions for our proposal is then:

$$N_P = \binom{N_b D_s}{K_s} \quad (7)$$

where $\binom{N_b D_s}{K_s}$ represents the number of combinations of K_s RBs within a resource grid of $N_b D_s$ RBs.

Our proposal guarantees that all N_P possible slices satisfy the latency requirements expressed in eq. (5). This is not always the case for the N_L possible slices obtained considering only the slice size descriptor. Fig. 2 depicts the percentage of possible slices that would satisfy the latency requirements for all served nodes with deterministic periodic traffic. This percentage is depicted as a function of K_s for the benchmark and proposal options. Results are shown for various latency deadlines D_s expressed as a function of T_p. Fig. 2 shows that defining slices considering only the size descriptor is not a viable approach to guarantee that nodes with deterministic periodic traffic meet their latency requirements. This can though be guaranteed with our proposed approach that also takes into account the novel latency-based slice descriptor when creating slices. Fig. 2 also shows that the capacity of the benchmark approach to define slices that meet the latency requirements decreases when the resources demanded per slice (K_s) increases. This is the case because the difference between N_L and N_P augments when K_s increases. The capacity of the benchmark approach to define slices that meet the latency requirements also decreases faster when these requirements become more demanding (i.e. when D_s decreases). On the other hand, the benchmark solution can approach the performance achieved with our latency-based proposal when the latency requirements are relaxed and D_s tends to T_p.

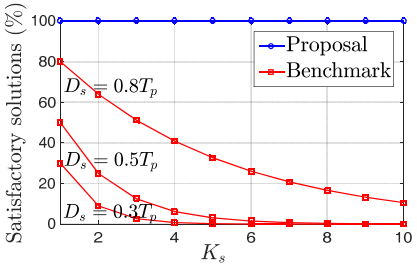

Fig. 2. Percentage of slices that satisfy the latency requirements.

The performance of the benchmark and proposed approaches is also compared in terms of the *Successful Transmission Ratio (STR)*. This metric quantifies the percentage of successful transmissions. A transmission is considered successful if it meets the QoS requirements demanded by the application. These requirements are defined by the number of resources demanded by the application (K_s) and the latency deadline D_s. We first estimate the *STR* metric

for the benchmark approach. This approach can identify N_L possible slices that include the K_s resources demanded by an application. Out of these K_s resources, n RBs could be located within a time window of length D_s and the remaining (K_s-n) RBs within a time window of length T_p-D_s, where $n \in \{0,1,2,\dots,K_s\}$. Eq. (6) can then be expressed as:

$$N_L = \binom{N_b T_p}{K_s} = \sum_{n=0}^{K_s} \binom{N_b D_s}{n} \binom{N_b T_p - N_b D_s}{K_s - n} \quad (8)$$

where $\binom{N_b D_s}{n}\binom{N_b T_p - N_b D_s}{K_s - n}$ represents the number of possible slices that have n RBs within the time window defined by D_s. The average *STR* for the benchmark approach is estimated as:

$$STR_{|\text{benchmark}} = \frac{1}{N_L} \sum_{n=0}^{K_s} \binom{N_b D_s}{n} \binom{N_b T_p - N_b D_s}{K_s - n} \tau(n) \quad (9)$$

where $\tau(n)$ is the *STR* for a given slice defined with the benchmark approach when there are n RBs within the time window of length D_s.

The average *STR* metric with our proposal can be obtained from (9) with $T_p = D_s$ and $n = K_s$. It can then be expressed as:

$$STR_{|\text{proposal}} = \frac{1}{N_P} \binom{N_b D_s}{K_s} \tau(K_s) = \tau(K_s). \quad (10)$$

The values of $\tau(n)$ and $\tau(K_s)$ in eq. (9) and eq. (10) are obtained though Monte-Carlo simulations. To this aim, we model a scenario where three micro cells (with 120m radius) are deployed to cover a factory. The communications channel model accounts for the path loss, small scale fading and shadowing. Nodes are uniformly distributed in the plant and transmit packets of 20 bytes with a target BLER of 10^{-5}. We consider 100 RBs per TTI (i.e. $N_b = 100$).

Fig. 3 depicts the average *STR* for a scenario with *M*=50 nodes. The figure shows that our proposal outperforms the benchmark approach under all conditions. This shows that defining RAN slices considering also the novel latency-based slice descriptor is a better strategy to satisfy the QoS demands of applications with deterministic periodic traffic. Our proposal also achieves an *STR* performance that is independent of the D_s/T_p ratio. On the other hand, the performance of the benchmark approach strongly depends on this ratio and can only approximate the performance achieved with our proposal when the latency requirements are relaxed and D_s tends to T_p.

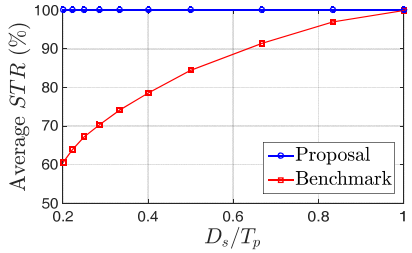

Fig. 3. Average *STR* as a function of D_s/T_p.

V. Conclusions

This paper has presented a novel latency-based slice descriptor to define RAN slices that can support the QoS requirements of Industry 4.0 applications. The study has demonstrated that the proposed descriptor significantly improves the capacity of 5G to satisfy the latency requirements that characterize industrial applications with deterministic periodic traffic. The authors are working to extend the proposal to other types of traffic, in particular non-deterministic and deterministic aperiodic traffic. Supporting deterministic aperiodic traffic is the most important challenge since we cannot anticipate when data must be transmitted but its transmission is bounded by a deterministic latency requirement and we cannot overprovision slices given the scarcity and cost of the cellular spectrum. The work will also be extended to develop novel RAN Slicing provisioning algorithms that can embed the latency-based slice descriptor when partitioning radio resources between slices.

Acknowledgments

This work has been funded by the European Commission through the FoF-RIA Project AUTOWARE: Wireless Autonomous, Reliable and Resilient Production Operation Architecture for Cognitive Manufacturing (No. 723909), the Spanish Ministry of Economy, Industry, and Competitiveness, AEI, and FEDER funds (TEC2017-88612-R).

References

[1] European Factories of the Future Association (EFFRA), "Factories 4.0 and Beyond," September 2016.

[2] 3GPP TR 22.804 V16.2.0, "Study on Communication for Automation in Vertical Domains (Release 16)," December 2018.

[3] 3GPP TS 22.261 V16.8.0, "Service requirements for the 5G system; Stage 1 (Release 16)," June 2019.

[4] 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, June 2017.

[5] 3GPP TR 36.881 V14.0.0, "Study on latency reduction techniques for LTE (Release 14)," June 2016.

[6] R. Ferrus, *et al.*, "On 5G radio access network slicing: Radio Interface Protocol Features and Configuration," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 184–192, May 2018.

[7] R. Kokku, *et al.*, "NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1333–1346, October 2012.

[8] R. Kokku, *et al.*, "CellSlice: Cellular wireless resource slicing for active RAN sharing," in *Proc. of the Fifth IEEE International Conference on Communication Systems and Networks (COMSNETS)*, Bangalore, January 2013, pp. 1–10.

[9] V. Sciancalepore, *et al.*, "Slice as a service (SlaaS) optimal IoT slice resources orchestration," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Singapore, December 2017, pp. 1–7.

[10] A. Aijaz, "Hap-Slice: A Radio Resource Slicing Framework for 5G Networks With Haptic Communications," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2285–2296, September 2018.

[11] I. Vilà, *et al.*, "Guaranteed Bit Rate Traffic Prioritisation and Isolation in Multi-tenant Radio Access Networks," in *Proc. of the 23rd IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Barcelona, September 2018, pp. 1–6.

[12] D. Bega, *et al.*, "A Machine Learning approach to 5G Infrastructure Market optimization," *IEEE Transactions on Mobile Computing*, Early Access, February 2019.

[13] W.-B. Yang, W.-B. Yang, and M. Souryal, "LTE physical layer performance analysis," US Department of Commerce, National Institute of Standards and Technology (NISTIR), 2014.