Providing Response Times Guarantees for Mixed-Criticality Network Slicing in 5G

Andrea Nota, Selma Saidi, Dennis Overbeck, Fabian Kurtz and Christian Wietfeld
TU Dortmund University, Dortmund, Germany
Email: *name.surname*@tu-dortmund.de

***Abstract*—Mission critical applications in domains such as Industry 4.0, autonomous vehicles or Smart Grids are increasingly dependent on flexible, yet highly reliable communication systems. In this context, Fifth Generation of mobile Communication Networks (5G) promises to support mixed-criticality applications on a single unified physical communication network. This is achieved by a novel approach known as network slicing, that promises to fulfil diverging requirements while providing strict separation between network tenants. We focus in this work on hard performance guarantees by formalizing an analytical method for bounding response times in mixed-criticality 5G network slicing. We reduce pessimism considering models on workload variations.**

***Index Terms*—5G network slicing, Formal performance analysis**

I. Motivation and Related Work

The Fifth Generation of mobile Communication Networks (5G) is currently an established technology foreseen for use in fields like industry automation and Vehicle-to-Everything (V2X) communication, where stringent timing and reliability requirements need to be met. A classification of use cases was introduced in 5G New Radio (NR) grouping different classes of services and their requirements under massive Machine Type Communication (mMTC), Ultra-reliable and Low Latency Communication (uRLLC) including critical MTC, and Enhanced Mobile Broadband (eMBB), see Fig 1.

Integrating these different service types into a single physical communication network is a significant challenge. Network slicing [1] [2] has been introduced as a key enabler in 5G for integrating these different service types into a single physical communication network (see 3GPP TS 28.530). Therefore, the network operator needs to carefully explore the opportunities for allocating network resources to network slices in a flexible way [3]: the resources that are needed from the higher criticality slice, must be subtracted from other lower criticality ones that are temporally slowed due to the insufficient resources. Furthermore, the process of scheduling requests and granting them can be eliminated using the so-called Configured Grant (CG) scheduling [4], planned to be integrated into 5G. CG scheduling allows to reserve fixed resources thereby eliminating the need for latency-inducing scheduling operations per packet.

However, reasoning about timing performance of radio access networks resources in order to provide guarantees is very challenging since wireless based communication is highly dynamic in terms of inter-arrival times of requests and also wireless channels conditions [5]. Main existing approaches in this field like [6] [7] rely on providing statistical or empirical delay and loss probability. In this paper, we provide an alternative analytical approach based on the busy-window approach [8] to reason about formal bounds on response times of mixed criticality slices in 5G considering configured grant scheduling. Arbitration between different slices is scheduled based on static-priority preemptive policy. Variations in packet sizes and their activation are upper bounded considering workload arrival functions [9] to reduce the pessimism of the worst-case analysis. We demonstrate the results considering data from realistic scenarios for Smart Grid and Electric Vehicle Charging applications.

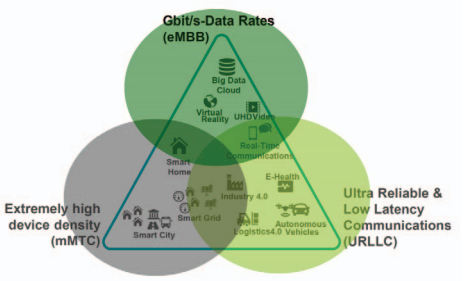

Fig. 1: Overview of 5G applications with different requirements.

II. Mixed-Criticality 5G Network Slicing

A. Background

a) Radio Access Resources in 5G: Wireless communication resources are radio frequency waves transmitted using subcarriers which can be multiplexed considering multiple frequency and time domains following the classical Orthogonal Frequency-Division Multiple Access (OFDMA). A Resource Element (RE) is therefore the smallest time-frequency resource (i.e., one OFDM symbol) which consists of one subcarrier modulated over time. A Resource Block (RB) is a group of subcarriers contiguous in frequency over symbol in time. For the sake of simplicity, we consider in the rest of the paper that RBs are the basic unit that can be allocated to a given application as considered in [4].

b) Network Slicing and QoS Support in 5G: Network slicing views resources in 5G as a grid of multiple RBs, each block is two dimensional and corresponds to an allocation in the radio frequency and time domains, see Fig 2. Based on the size of transmissions, the priority of the slice and the modulation scheme, the network slicing scheduler allocates the required number of RBs to be used by a given application.

978-3-9819263-6-1/DATE22/©2022 EDAA

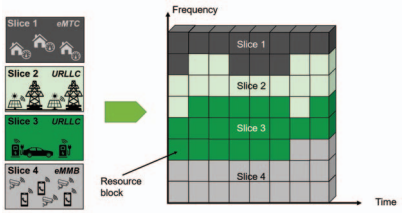

Fig. 2: Example of Resource Allocation of 5G Slices in Frequency and Time Domains.

B. System Model and Assumptions

Our starting point is a system with a set of mixed criticality slices $S = \{s_1, ...s_l\}$ simultaneously requesting resources.

1) Applications Characterization:

Definition 1 (Slice): A slice $s_i = (\rho_i, N_{UEi}, I_i)$ is an application or a use case defined with a given (latency-sensitive) importance or criticality. Based on criticality, a slice is assigned a *statically fixed* priority level ρ_i. Every slice s_i has a number N_{UEi} of User Equipments (UEs) executing the same application and performing data transmissions. I_i is the aggregated sequence of all transmissions performed by UEs within a slice i. Note that all UEs belonging to the same slice have the same priority ρ_i

Definition 2 (Data Streams): Data streams $I = \{e_1, ..., e_n\}$ are traces defined as a sequence of events. Every event $e_k = (t_k, w_k)$ is a transmission request defined as the time t_k where the request is activated and a workload w_k which corresponds to the size of the packets to be transmitted by request e_k. Note that for each slice i, all data stream requests inherit the priority level from their corresponding slice.

The behavior of data streams in 5G is highly dynamic. We use event models, standardly used to model task activations in real-time analysis methods like real-time calculus or compositional performance analysis [8], to bound the arrival time of data requests. We consider additionally workload arrival functions [9] to bound variations in packet sizes and therefore transmissions workloads.

Definition 3 (Data Transmissions Event Models): Event models are used to characterize for every slice i the arrival of data transmissions. It is defined using the function $\eta_i(\Delta t)$ which denotes for every slice i the maximum number of transmissions issued within a time window Δt. The inverse function $\delta_i(n)$ denotes the minimum time interval between the first and the last transmission in any sequence of n transmissions from slice i. Given a data stream I, event arrival functions are extracted by looking at the smallest sliding window I_k of $k \in [2..n]$ subsequent events in the trace used to derive the minimum time interval between the first and the k^{th} data transmissions requests.

$$\delta_i(k) = \min_{\forall I_k} \; e_k(t) - e_1(t) \tag{1}$$

Definition 4 (Data Transmissions Workload Models): Workload models are used to characterize workload arrival when events do not have the same execution time or the same data size. Let $\alpha_i^w(\Delta t)$ capture for every slice i the maximum accumulated workload w_i in terms of transmissions sizes during a time window of size Δt, see Fig 3. This corresponds to the sum of workloads from all transmissions η_i received during Δt.

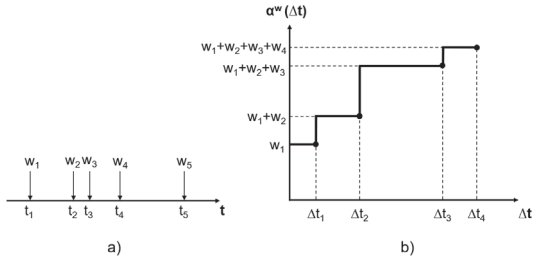

Fig. 3: Data streams illustrated as a) traces and corresponding b) workload arrival functions, $\alpha_i^W(\Delta t_2) = w_1 + w_2$ is the total amount of data that can be transmitted considering 2 events at t_1 and t_2.

$$\alpha_i^W(\Delta t) = \sum_{k=0}^{\eta_i(\Delta t)} w_k \tag{2}$$

2) Resources Characterization and Allocation:

Definition 5 (5G Resource Grid): A 5G resource grid RG_{5G} is defined as a matrix of $n \times m$ Resource Blocks (RBs). The total number of RBs in the resource grid is limited by the channel bandwidth and subcarrier spacing (or numerology) as defined in the 5G specifications [10].

Definition 6 (Network Slicing Allocation): Allocation in network slicing is performed using a radio resource scheduler, and defines for every slice a mapping $m : S \longrightarrow RG_{5G}$, where $m(s_i)$ is the set of resource blocks from the 5G resource grid assigned to slice s_i.

Definition 7 (Configured Grants): Number of resources blocks for each slice i are allocated in terms of Configured Grants $CG_i = |m(s_i)|$. The number of reserved RBs corresponds to the necessary amount of resources which allow to transmit the maximum size packet of slice i within one Transfer Time Interval (TTI) (i.e., without backlog).

III. Response Time Analysis

A. Basic Slice Execution Time

Definition 8 (Slice Bandwidth): The number of bits that can be transmitted within a resource block is not fixed. It depends on multiple factors, such as the modulation scheme, subcarrier spacing configuration, and the number of OFDMA symbols. This calculation is standarized following the 3GPP 39.214 (chapter 5.1.3.2)[1]. We abstract this calculation considering the function $b(CG)$ where configuration parameters are fixed and only the number of reserved RBs as CG can vary.

Definition 9 (Resources Load): Let R^W be the workload in terms of number of bits that can be transmitted by a number of resource blocks CG during a time Δt and considering the bandwidth b. This can be derived as follows:

$$R^W(\Delta t) = \Delta t \times b(CG) \tag{3}$$

[1]https://5g-tools.com/5g-nr-tbs-transport-block-size-calculator/

Lemma 1 (Basic Slice Latency Bounds): Let $C_i(\Delta t)$ be the basic execution time for every slice when the slice is executed in isolation (i.e., when maximum RBs can be allocated without any interference from other slices). It can be defined as follows,

1) Maximum slice demand (highly pessimistic approach), that is the worst-case demand based on maximum packet size

$$C_i(\Delta t) = \frac{\eta_i(\Delta t) \times \forall_{k=0}^{\eta_i(\Delta t)} \; max(w_k)}{b(CG_i)} \tag{4}$$

2) Current slice demand (less pessimistic approach), that is the worst-case slice demand captured as accumulated maximum load by the workload event curves.

$$C_i(\Delta t) = \frac{\alpha_i^W(\Delta t)}{b(CG_i)} \tag{5}$$

Proof: The proof is derived directly considering the ratio between the maximum workload and the bandwidth to achieve a bound on the maximum execution time. ■

B. Derivation of Blocking Time

Every slice i has initially an amount of allocated resource blocks RBs as configuration grants CGs. However, when multiple slices are active at the same time, the scheduler allocates RBs first to higher priority slices, thereby leading to a reduced bandwidth available for lower priority and best effort slices.

Lemma 2: Let Γ_i be the blocking time[2] which q requests from slice i experience in a time window Δt, due to transmissions from slices with higher priority can be bounded by:

$$\Gamma_i(\Delta t) = \Sigma_{\forall j \in hp(i)} \; \eta_j(\Delta t) \times \frac{R_{ij}^W(C_j)}{b(CG_i - CG_{ij})} \tag{6}$$

where, $hp(i)$ is the set of transmissions from slices with higher priority than slice i, $\eta_j(\Delta t)$ denote the maximum number of requests from slice j within an interval of Δt, $R_{ij}^W = C_j \times b(CG_{ij})$ is the workload (i.e., number of bits) that can be transmitted considering the difference in resource blocks between slice i and slice j, that is $CG_{ij} = |m(s_i) \bigcap m(s_j)|$.

Proof: In the presence of higher priority slice j, the effect of interference concerns resource blocks CG_{ij} that cannot be used by lower priority slice i. It results in decreased bandwidth $b(CG_i - CG_{ij})$ available to lower criticality slice i to progress on sending data during the entire execution C_j of higher priority slices. The $R_{ij}^W = C_j \times b(CG_{ij})$ describes the number of bits that could not be transmitted due to RBs allocated to higher priority slices and that still need to be transmitted. This occurs whenever a higher priority task is activated therefore accounting for $\eta_j(\Delta t)$ to bound the maximum amount of activations from slice j during a time interval Δt. ■

C. Worst-Case Response Time

Definition 10 (Busy Window): The maximum q-event busy windows $\omega_i^+(q)$ of a slice i describes the maximum time interval required to complete q consecutive transmissions considering network slicing and mixed criticality.

[2]Note that blocking time in this case is the slow down time due to reduced bandwidth for lower priority slices in the presence of higher priority ones.

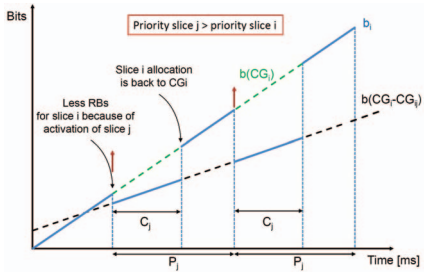

Fig. 4: Change of the available RBs and therefore bandwidth for a slice i given the activation of higher priority slice j.

Theorem 1: The worst-case time necessary to conduct q transmissions from slice i is bounded by:

$$\omega_i^+(q) \leq \gamma_i + C_i(\delta(q)) + \Gamma_i(\omega_i^+(q)) \tag{7}$$

where, $\gamma_i = TTI$ is the blocking time due to lower priority slices to perform the reallocation. TTI represents the minimum time separating two frames on the transport layer. $C_i(\delta(q))$ is the execution time of slice i in isolation and $\Gamma_i(\omega_i^+(q))$ is blocking time due to the delay caused by higher priority tasks acquiring CG_{ij} from slice i.

Proof: The total busy window of q consecutive transmissions from slice i is constructed considering the worst-case execution time when the slice is executed in isolation and additional worst-case blocking time considering maximum latency resulting from reallocation of RBs from lower priority slices to higher priority ones and maximum arrival of requests from higher priority slices. ■

Note that $\omega_i^+(q)$ appears on both sides of Eq. 7 which constitutes a fixed-point computation which can be solved iteratively starting with $\omega_i^+(q) = C_i(\delta(q))$.

IV. Evaluation Results

We consider the following 5G slices ordered by decreasing priority: Smart Grid (SG) slice (uRLLC)[3], Electric Vehicle (EV) Charging slice (uRLLC)[4] and Best Effort (BE) slice (eMBB). We consider data from real-world applications. Table I shows the number of User Equipments (UEs) performing multiple data packet transmissions with different Arrival Times (AT) and the amount of resources (RBs) allocated as CG. We consider a channel bandwidth of 20 MHz that corresponds to 106 available RBs for each slot, subcarrier spacing of 15 kHz, modulation order of 8 and a packet TTI of 1 ms. Since the sum of RBs required by each slice exceeds the size of the 5G resource grid, the execution of lower priority slices will be slowed whenever higher priority ones are activated.

Figure 5 summarizes the WCRT of the each slice in every configuration. Formal timing analysis is implemented using the pyCPA tool [8] for the worst-case response time computation. In the highest priority slice Smart Grid, the WCRT of each packet is independent from the amount of the load of other slices and is stable to 1 ms. Contrarily, the values in the EV

[3]https://www.nrel.gov/grid/solar-power-data.html
[4]https://new-poi.chargecloud.de/bonn (January 2020)

	Smart Grid			EV Charging		
	UEs	AT [ms]	RBs	UEs	AT [ms]	RBs
Conf 1	6	2	37	5	2	68
Conf 2	7	2	43	5	2	68
Conf 3	8	2	49	5	2	68
Conf 4	6	2	37	6	2	82
Conf 5	7	2	43	6	2	82
Conf 6	8	2	49	6	2	82
Conf 7	6	2	37	7	2	96
Conf 8	7	2	43	7	2	96
Conf 9	8	2	49	7	2	96
Conf 10	12	3	73	5	2	68
Conf 11	12	3	73	6	3	82
Conf 12	12	3	73	7	3	96
Conf 13	12	4	73	8	3	106
Conf 14	13	4	79	7	4	96
Conf 15	14	5	85	7	4	96

TABLE I: Values of the main parameters used for analyzing the WCRT in multiple configurations.

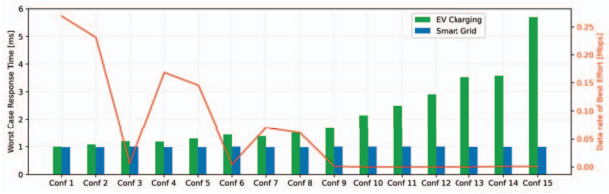

Fig. 5: WCRT for transmissions in the Smart Grid (high priority) and Electrical Vehicle (EV) (low priority). Data rate of Best Effort (BE) slice. When the channel is overloaded (config 3, 6 and ≥ 9), the data rate of the BE is minimum.

slice follow a different trend. Starting from left, EV does not show any delay since the total RBs required is 105 (37 + 68). As we move to the right part of the plot, the load increases and less RBs are available for the EV slice, thereby increasing the WCRT of EV transmissions. The right y axis of Figure 5 depicts the throughput of BE consuming the remaining bandwidth without guarantees. The throughput of BE drastically drops as the demand from SG and EV is increasing. Drastic performance drops can be observed in different configurations based on different distribution of users and packet sizes between SG and EV.

Later we demonstrate that the workload arrival function $\alpha^W(\Delta t)$ reduces the pessimism of our approach which considers that all packets have maximum size. To demonstrate that, we vary the data packet size and their distribution in the SG slice, see Table II, and consider their effect on the WCRT of EV. In particular, every data packet, except 1 that is fixed to the maximum value, has now the minimum size. Thus, when the number of consecutive packets of maximum size (now forced to 1) is lower than number of activation n of the highest priority slice (SG) in the busy window of the EV slice, the $\alpha^W(\Delta t)$ will show a lower value compared to the scenario where all n packets have the maximum size. Consequently, the WCRT of a EV slice packet will decrease compared to considering always n consecutive packets of maximum size (without $\alpha^W(\Delta t)$). Figure 6 confirms, for multiple configurations, that $\alpha^W(\Delta t)$ increases the accuracy of the WCRT computation.

V. Conclusion

We provide in this paper a first attempt in formalizing bounds on the response times for mixed criticality applications in network slicing, based on configured grants for resource blocks allocation and static-priority preemptive scheduling.

	Smart Grid (SG)					EV Charging			
	Min packet size [Bytes]	Max packet size [Bytes]	UEs	AT [ms]	RBs	UEs	AT [ms]	RBs	n
Conf 1	200	600	5	2	71	6	4	82	2
Conf 2	200	800	4	3	76	7	4	96	4
Conf 3	200	1000	3	2	71	6	5	82	2
Conf 4	200	1200	3	3	86	5	5	68	5

TABLE II: Characterization of each simulation setup. In Smart Grid the number of packets of maximum size is set to 1. n is the number of activation of the highest priority slice (SG) in the busy window of the EV slice obtained throught simulation.

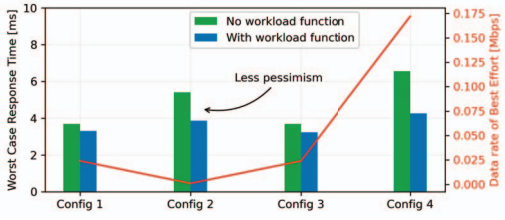

Fig. 6: Worst Case Response Time for the packets in the EV Charging slice and data rate of the BE slice considering the workload arrival function $\alpha^W(\Delta t)$. The pessimism depends on the rate between the number of packets of max size and the total number of consecutive transmissions of the highest priority task in the busy window.

Using workload arrival functions, we tackled one aspect of dynamic execution in wireless communication networks. Future work will deal with providing models for dynamic variation in network resources and channel conditions.

Acknowledgement

This work has been supported by the Federal Ministry of Education and Research (BMBF) in the course of the project *6GEM* under the funding reference 16KISK038 and by the Federal Ministry for Economic Affairs and Energy (BMWi) in the course of the project *5Gain* under the funding reference 03EI6018C.

References

[1] Q. Chen, X. Wang, and Y. Lv, "An overview of 5G network slicing architecture," *AIP Conference Proceedings*, vol. 1967, no. 1, p. 020004, 2018. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/1.5038976

[2] L. U. Khan, I. Yaqoob, N. H. Tran, Z. Han, and C. S. Hong, "Network slicing: Recent advances, taxonomy, requirements, and open research challenges," *IEEE Access*, vol. 8, pp. 36 009–36 028, 2020.

[3] R. Trivisonno, R. Guerzoni, I. Vaishnavi, and A. Frimpong, "Network resource management and qos in sdn-enabled 5G systems," in *2015 IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–7.

[4] C. Bektas, D. Overbeck, and C. Wietfeld, "SAMUS: Slice-aware machine learning-based ultra-reliable scheduling," in *2021 IEEE International Conference on Communications (ICC)*, Montreal, Canada, jun 2021.

[5] C. Arendt, S. Böcker, and C. Wietfeld, "Data-driven model-predictive communication for resource-efficient iot networks," in *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, 2020, pp. 1–6.

[6] S. Wang, R. Nathuji, R. Bettati, and W. Zhao, "Providing statistical delay guarantees in wireless networks," in *24th International Conference on Distributed Computing Systems, 2004. Proceedings.*, 2004, pp. 48–55.

[7] A. Dailianas and A. Bovopoulos, "Real-time admission control algorithms with delay and loss guarantees in atm networks," *Computer Communications*, vol. 19, no. 3, pp. 169–179, 1996, algorithms for ATM networks. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0140366496010535

[8] J. Diemer, J. Rox, and R. Ernst, "Compositional Performance Analysis in Python with pyCPA," in *WATERS 2012*, 2012. [Online]. Available: http://retis.sssup.it/waters2012/accepted/102_Final_paper.pdf

[9] M. Neukirchner, P. Axer, T. Michaels, and R. Ernst, "Monitoring of workload arrival functions for mixed-criticality systems," in *2013 IEEE 34th Real-Time Systems Symposium*, 2013, pp. 88–96.

[10] G. T. 23.501, "System Architecture for the 5G System (Release 16)," Dec. 2019.