Optimisation of Numerology and Packet Scheduling in 5G Networks

To Slice or not to Slice?

Raftopoulou, Maria; Litjens, Remco

DOI
10.1109/VTC2021-Spring51267.2021.9448814

Publication date
2021

Document Version
Final published version

Published in
2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)

Optimisation of Numerology and Packet Scheduling in 5G Networks: To Slice or not to Slice?

Maria Raftopoulou
Quantum and Computer Engineering
Delft University of Technology
Delft, The Netherlands
M.Raftopoulou@tudelft.nl

Remco Litjens
Department of Networks
TNO
The Hague, The Netherlands
Quantum and Computer Engineering
Delft University of Technology
Delft, The Netherlands
remco.litjens@tno.nl

2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring) | 978-1-7281-8964-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/VTC2021-Spring51267.2021.9448814

***Abstract*—Network slicing has been introduced in 5G networks as an enabling feature for the effective Quality of Service (QoS) provisioning to multiple service classes with distinct performance requirements. When applied in the Radio Access Network (RAN), a class-specific slice is assigned a set of radio resources and can furthermore be optimally configured in terms of the applied numerology and packet scheduler. As both the optimal numerology and the most suitable packet scheduler may be different for e.g. a class of Latency-Constrained (LC) and a class of Throughput-Oriented (TO) services, the potential of slicing is clear. However, the inherent trunking loss incurred when applying slicing with dedicated resources provides an argument against such slicing. In this paper we demonstrate that the performance and traffic handling capacity in an optimally configured non-sliced scenario may exceed that attained when using segregated individually optimised slices. To that end, we use simulations to assess the best-performing numerology and packet scheduler for a sliced scenario with LC and TO services. We then compare the thus optimised sliced scenario with an optimal non-sliced scenario and show that the non-sliced scenario can serve about 20% more traffic than the sliced scenario while satisfying the same class-specific QoS requirements.**

***Index Terms*—5G networks, packet scheduling, flexible numerology, URLLC services, eMBB services, network slicing**

I. Introduction

5G networks are designed to support new services with diverse characteristics and requirements. Specifically, the new services are categorised in three groups: enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC) and massive Machine Type Communications (mMTC) [1]. In the Radio Access Network (RAN), the choice of numerology and packet scheduler has an impact on the Quality of Service (QoS) for each service.

Similarly to 4G networks, 5G networks use Orthogonal Frequency Division Multiple Access (OFDMA). While in 4G networks, the SubCarrier Spacing (SCS) of the Orthogonal Frequency Division Multiplex (OFDM) symbols is fixed to 15 kHz, in 5G networks the SCS can be flexibly configured to 15, 30, 60, 120 or 240 kHz [2]. In the 3GPP standards, the choice of SCS and, correspondingly, the symbol duration, is referred to as the numerology. The introduction of *flexible numerology* allows for a shorter OFDM symbol time and thus for a shorter Transmission Time Interval (TTI) at the cost of a lower number of Physical Resource Blocks (PRBs) for a given carrier bandwidth, which may come at the cost of reduced throughput gains from frequency-domain packet scheduling.

The role of the *packet scheduler* is to assign the available radio resources to the active QoS flows in the network. Considering the diverse requirements across services, it is not trivial to accommodate all services with a single scheduling rule. However, scheduling rules have been designed for 4G networks that distinguish between real-time and non real-time traffic [3], which in combination with a suitably configured numerology can potentially offer higher throughput and lower latency than in 4G networks. Additionally, the concepts of mini-slots and self-contained slots have been introduced in 5G to enhance support of URLLC transmissions.

A novel concept introduced in 5G networks that is designed in specific support of services with different requirements in the same network infrastructure, is *network slicing*. The main principle of network slicing is to create multiple independent virtual networks that share the same network infrastructure while each virtual network serves traffic with a specific Service Level Agreement (SLA). Therefore, in the RAN, each slice can use the packet scheduler and numerology that best serves the intended traffic. However, assigning the radio resources to each slice in such a way that the SLA for each slice is guaranteed is a non-trivial task as the traffic can be very dynamic, the QoS requirements may be very demanding and the amount of available radio resources in the network is limited.

There is a significant amount of work in literature addressing the resource assignment problem in sliced networks. In [4], a two-level scheduler is proposed that uses resource virtualization to perform inter- and intra-slice resource allocation. In [5] an AI-based method is proposed assigning radio resources to the slices based on traffic prediction. Other publications demonstrate how services with different requirements can be served in a non-sliced network. For example, in [6], punctured scheduling is used to multiplex low-latency and mobile broadband traffic. Scheduling in multi-numerology networks is addressed in [7]. Considering the variety of methods available in literature that address the problem of QoS provisioning,

there is no clear indication about the circumstances under which RAN slicing provides the most efficient optimisation. While sliced networks offer the possibility of an optimal configuration per service in terms of packet scheduler and numerology, non-sliced networks enable full flexibility in resource sharing which leads to maximal trunking gains. The purpose of this paper is to provide new insights into the relative merits of slice-optimised numerologies and scheduling on the one hand, and the trunking gains in non-sliced scenarios on the other hand, by comparing an optimised sliced with an optimised non-sliced scenario.

The remainder of the paper is organised as follows. In Section II the concepts of flexible numerology and packet scheduling are discussed. Further, modelling aspects and traffic characteristics are presented in Section III. The simulation results are analysed in Section IV. Finally, conclusions and recommendations for future work are given in Section V.

II. RAN Configuration

This section discusses the concepts of flexible numerology and packet scheduling.

A. Flexible Numerology

In 5G networks the SCS of the OFDM symbols is flexible, whereas in 4G networks the SCS is fixed to 15 kHz. For a given numerology value μ, the SCS and consequently the symbol duration is defined, as shown in Table I. An increase of the SCS, decreases the OFDM symbol duration which leads to a shorter slot duration and thus a shorter TTI. Additionally, a wider SCS reduces the number of PRBs within a given bandwidth compared to a narrower SCS because a PRB comprises twelve subcarriers regardless of the numerology. Low-latency services can thus be supported by higher numerologies as the TTI is shortened, but the consequent reduction in the number of PRBs, for a given carrier bandwidth, compared to lower numerologies can have a negative impact on the throughput as the gains obtained from frequency-domain channel-adaptive scheduling are reduced. This intrinsic trade-off between latency and throughput implies that Latency-Constrained (LC) and Throughput-Oriented (TO) traffic are best served with higher and lower numerologies, respectively.

Apart from the QoS requirements, the choice of numerology is also limited by whether the carrier frequency is in Frequency Range 1 (FR1) ($<$ 6 GHz) or in FR2 ($>$ 6 GHz), as also shown in Table I. Finally, for large cell sizes and harsh propagation environments, lower numerologies are more appropriate as symbols with long durations are more robust to inter-symbol interference [8].

TABLE I
5G numerologies [2].

μ	SCS (kHz)	OFDM symbol duration (μs)	Slot duration (ms)	Frequency range
0	15	71.35	1	FR1
1	30	35.68	0.5	FR1
2	60	17.84	0.25	FR1 and FR2
3	120	8.92	0.125	FR2
4	240	4.46	0.0625	FR2

B. Packet Scheduling

Packet scheduling determines which packets of the active QoS flows will be served at a particular TTI and PRB, where a queue of packets is maintained for each active QoS flow. For each QoS flow i at TTI t and PRB f, the metric $M_{S,i}(t,f)$ is calculated based on the scheduler S. In a network with N active flows, at a given TTI t, the scheduler assigns PRB f to QoS flow i^* which has the highest $M_{S,i}(t,f)$ value:

$$i^* = \underset{1\leq i\leq N}{\operatorname{argmax}} M_{S,i}(t,f)$$

Each TTI, decisions are taken on a per-PRB level, in order to exploit the fact that some flows perform better than others on particular PRBs due to frequency-selective fading and interference, and thus achieve gains from frequency diversity. Also, a packet scheduler can decide to not serve (and hence drop) a Head-Of-Line (HOL) packet if it cannot be delivered within an imposed latency constraint.

We consider a range of packet schedulers. Among these, the *Maximum Rate (MR)* scheduler aims to maximise the system throughput as it considers the instantaneously attainable bit rate $R_i(t,f)$ at TTI t and PRB f of QoS flow i:

$$M_{MR,i}(t,f) = R_i(t,f)$$

The *Proportional Fair (PF)* scheduler also aims to provide high system throughput, with a scheduling metric given by:

$$M_{PF,i}(t,f) = \frac{R_i(t,f)}{\overline{R}_i(t-1)}$$

where $\overline{R}_i(t) = (1-\frac{1}{t_c})\overline{R}_i(t-1) + \frac{1}{t_c}R_i(t-1)$ is the exponentially smoothed experienced bit rate of flow i up to and including TTI t [9]. In contrast to the MR scheduler, the PF scheduler provides a fairer resource distribution among the flows as the selection of flows is also based on the flows' experienced bit rate, in the sense that flows with relatively low experienced bit rates have a higher likelihood of being scheduled. The smoothing parameter t_c is effectively setting the trade-off between resource fairness and system throughput. For a very high value of t_c, the PF scheduler behaves similarly to the MR scheduler [9]. Both the MR and PF schedulers are only considering the channel quality of each flow (in terms of the attainable bit rates) and hence are latency-oblivious, which makes them unsuitable to serve TO traffic.

The *Earliest Deadline First (EDF)* [3] and the *Weighted Earliest Deadline First (W-EDF)* [10] schedulers explicitly aim to deliver packets within their latency constraints by considering the remaining time until the expiry of the imposed deadline:

$$M_{EDF,i}(t) = \frac{1}{\tau_i - W_i(t)}$$

$$M_{W-EDF,i}(t) = \frac{W_i(t)}{\tau_i - W_i(t)}$$

where τ_i denotes the latency constraint of flow i and $W_i(t)$ denotes the HOL packet latency experienced up to TTI t for flow i. The difference between the two schedulers is that the W-EDF scheduler uses the HOL packet latency $W_i(t)$ as a weight. Both schedulers are purely latency-based and hence are appropriate for serving LC traffic. Note, however, that these schedulers have no channel-adaptive component and may consequently be rather resource-inefficient.

The *Modified-Largest Weighted Delay First (M-LWDF)*, the *Exponential Proportional Fair (EXP-PF)*, the *Log-Rule* and the *EXP-Rule* schedulers [3] are based on the PF scheduler but aim to serve both LC and TO flows, featuring both channel-adaptive and latency-oriented aspects. Specifically, LC flows are served with a weighted version of metric $M_{PF,i}(t,f)$ and TO flows are served with metric $M_{PF,i}(t,f)$:

$$M_{S,i}(t,f) = \begin{cases} \phi_S(W_i(t)) M_{PF,i}(t,f) & i \in \text{LC} \\ M_{PF,i}(t,f) & i \in \text{TO} \end{cases}$$

where $\phi_S(W_i(t))$ is the weight function for scheduler $S \in \{M-LWDF, EXP-PF, Log-Rule, EXP-Rule\}$. For the M-LWDF scheduler:

$$\phi_{M-LWDF}(W_i(t)) = a_i W_i(t)$$

where $a_i = -log(\delta_i)/\tau_i$ and $\delta_i \in [0,1]$ is the maximum allowed packet drop rate for flow i. The EXP-PF scheduler tries to guarantee the packet delivery latency by using an exponential function:

$$\phi_{EXP-PF}(W_i(t)) = \exp\left(\frac{a_i W_i(t) - \overline{aW}(t)}{1+\sqrt{\overline{aW}(t)}}\right)$$

with $a_i = 10/\tau_i$ and $\overline{aW}(t) = \frac{1}{N_{LC}} \sum_{i \in LC} a_i W_i(t)$ where N_{LC} is the total number of LC flows. The Log-Rule scheduler tries to guarantee the packet delivery latency based on the logarithmic function which increases more slowly compared to the exponential function:

$$\phi_{Log-Rule}(W_i(t)) = b_i \log(c + a_i W_i)$$

where $a_i = 5/0.99\tau_i$, $b_i = 1/E[M_{PF,i}(t,f)]$, $c = 1.1$ and $E[\cdot]$ denotes the expected value. Finally, the EXP-Rule scheduler combines characteristics of the EXP-PF and Log-Rule schedulers:

$$\phi_{EXP-Rule}(W_i(t)) = b_i \exp\left(\frac{a_i W_i(t)}{c + \sqrt{\frac{1}{N_{LC}} \sum_{i \in LC} W_i(t)}}\right)$$

where $a_i \in [5/0.99\tau_i, 10/0.99\tau_i]$, $b_i = 1/E[M_{PF,i}(t,f)]$ and $c = 1.1$.

III. Modelling

This section describes further modelling aspects such as the network layout, the propagation environment and the traffic model. We further define the used Key Performance Indicators (KPIs).

A. System Model

Although our study has much broader validity, we consider an Industry 4.0-inspired use case with distinct services in a factory hall environment. The modelled factory hall is of dimensions 100 m $\times$ 100 m $\times$ 10 m and an indoor base station (gNB) with an omnidirectional antenna is mounted at the centre of the ceiling [11]. The gNB has a 2 dBi gain and a transmit power of 21 dBm. We assume downlink transmissions to devices that are randomly distributed in space, but at a fixed height of 1.5 m, and have a receiver noise figure of 5 dB.

The propagation environment of the factory is generated with the QuaDRiGa Industrial NLOS model [12]. The model includes among others distance-based path loss, shadowing and Ricean fading. Finally, the factory is assumed to be isolated from other traffic, hence there is no interference.

We assume a 20 MHz wide carrier in the 3.5 GHz band (FR1). The carrier applies Time Division Duplexing (TDD) with a five-slot frame format comprising one uplink and four downlink slots. Devices report to the gNB their downlink channel quality through sub-band Channel Quality Indicator (CQI) reporting with a period of 5 ms. The CQI sub-band size is given in [13]. Based on the CQI reporting, the gNB selects for the downlink transmission the highest attainable Modulation and Coding Scheme (MCS) with an estimated BLock Error Rate (BLER) not exceeding 0.001% and 10% for LC and TO flows, respectively. Modulation schemes up to 64-QAM are supported. MCS-specific BLER-vs-SINR curves have been derived using the Vienna 5G Link Level Simulator [14]. Additionally, the Mutual Information Effective SINR Mapping (MIESM) method is used to map a set of PRB-specific SINRs to a single effective SINR value for the full set of PRBs [15]. An Outer-Loop Link Adaptation (OLLA) scheme is used to modify the mapping of the SINR to an MCS due to imperfections in channel quality reporting, e.g. due to inherent feedback delays [16][17]. Lastly, for the unsuccessful downlink transmissions, the gNB retransmits the lost transport blocks. For LC flows, the transport blocks are only retransmitted if they can still be delivered within their latency constraint.

B. Traffic Model

We distinguish between persistent LC and non-persistent TO flows with traffic models inspired by the Industry 4.0 use cases 'precise cooperative robotic motion control' and 'remote access and maintenance', respectively [18]. We further consider that each flow targets a different device. Specifically, we assume the presence of N_{LC} persistent LC flows generating packets of size X_{LC} bytes with a fixed inter-arrival time of 3 ms. The latency budget for each LC packet is 3 ms. The non-persistent TO flows are generated according to a Poisson process with arrival rate λ_{TO} (in flows/s) and each TO flow is modelled as a deterministic file download of X_{TO} MB.

C. KPI Definitions

Distinct KPIs are defined for the LC and TO flows. For the TO flows the KPI of relevance is the 10th throughput

percentile. The applied target level for this KPI is 10 Mbps.

For the LC flows, the KPI of relevance is the fraction of LC flows experiencing a reliability of at least 99.9%. We define reliability as the fraction of packets per LC flow that are successfully received at the targeted device within the latency budget. The LC packet latency is defined as the time between the packet arrival in the buffer at the gNB and the successful packet reception at the targeted device. The processing latencies at both the gNB and the device are also considered. Fig. 1 shows the measured latency, which concerns the PHY/MAC layers in the user plane, for a case with one retransmission. Parameter *K1* is signalled to the device, via the Physical Downlink Control CHannel (PDCCH), to indicate the time between the reception of the downlink data on the Physical Downlink Shared CHannel (PDSCH) and the transmission of the Hybrid Automatic Repeat Request (HARQ) feedback on the Physical Uplink Control CHannel (PUCCH) [13][19]. The value of *K1* depends a.o. on the device capability and the operational numerology [20]. Additionally, parameter *K3* indicates the time between the reception of the HARQ Negative ACKnowledgement (NACK) on the PUCCH and the retransmission of the downlink data on the PDSCH and its value is up to the gNB implementation [21]. Finally, the processing latency at the gNB is assumed to be one slot for both transmission and reception of data [22].

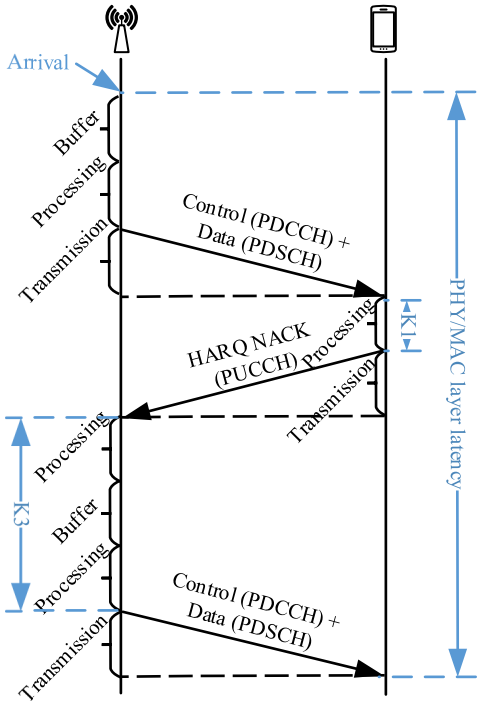

Fig. 1. PHY/MAC layer latency for a downlink transmission with one transport block retransmission.

IV. Numerical Results

This section shows the impact of the packet schedulers and the numerology on the QoS *(i)* for a *sliced scenario* with distinct and isolated LC and TO slices that equally share the radio resources and *(ii)* a *non-sliced scenario* with mixed LC/TO traffic. We then compare the sliced and non-sliced scenarios based on their performance on the QoS targets. For the analysis of both scenarios, dynamic system-level simulations are performed and the results are based on multiple independent simulations with distinct random seeds. Considering the use of the 3.5 GHz carrier frequency, from Table I, numerologies 0, 1 and 2 are used in the experiments. Additionally, the parameters related to the schedulers are set to $t_c = 10$ ms, $\tau_i = 3$ ms, $\delta_i = 10^{-5}$ for LC flows, $\delta_i = 10^{-1}$ for TO flows and $a_i = 7/0.99\tau_i$ for the EXP-Rule scheduler.

A. Sliced Scenario: Latency-Constrained Slice

In the *LC slice*, we evaluate the impact of numerology on the reliability performance when using the M-LWDF scheduler while we vary the number of persistently active flows N_{LC} and the packet size X_{LC}. Fig. 2a shows the fraction of flows that meet the reliability requirement. For numerology 0, the slot duration is 1 ms and the fixed processing latency is 2 ms, according to Section III-C, thus packets can spend a maximum of 1 ms in the buffer given the 3 ms latency budget. Fig. 2a shows that regardless of the offered load, none of the flows can meet the imposed reliability requirement for numerology 0. For numerology 1, the fixed processing latency is reduced to 1 ms, as the slot duration is 0.5 ms, which allows packets to spend up to 2 ms in the buffer. Fig. 2a shows the benefits of increasing the numerology from 0 to 1 as for some loads the required reliability can be achieved by about 95% of the flows. The fixed processing latency is further reduced to 0.5 ms for numerology 2. Moreover, packets can be retransmitted if they are not correctly received at the targeted device, which is now possible because of the shortened slot duration. Due to the retransmissions, the fraction of flows meeting the required reliability is further increased, even reaching up to 100% for cases with 30 active flows and a 100-byte packet size.

A realistic packet size X_{LC} for the considered Industry 4.0 LC use case is 150 bytes [18]. The highest number of flows N_{LC} that can be supported with $X_{LC} = 150$ bytes such that the KPI target is met, is 25 flows. For this scenario, Fig. 2b shows the schedulers' comparison for all three numerologies including 90% confidence intervals for the shown KPI. From Fig. 2b the same observations for the impact of numerology on the KPI hold as discussed for the M-LWDF scheduler. Observe from the results that *the optimal configuration for the LC slice is the M-LWDF scheduler and numerology 2* as it provides the highest fraction of users that meet the KPI target. This is the reason the M-LWDF scheduler was used in the more detailed analysis of the numerology impact on the KPI in Fig. 2a.

The M-LWDF scheduler outperforms the EDF and the W-EDF schedulers as it considers both the latency constraint and the instantaneous bit rate in contrast to the EDF and W-EDF schedulers that are channel-oblivious. Fig. 2b also shows that the EXP-PF and EXP-Rule schedulers yield relatively poor performance as they somehow consider the normalised sum of the HOL latency of all LC flows. Also, their fair design

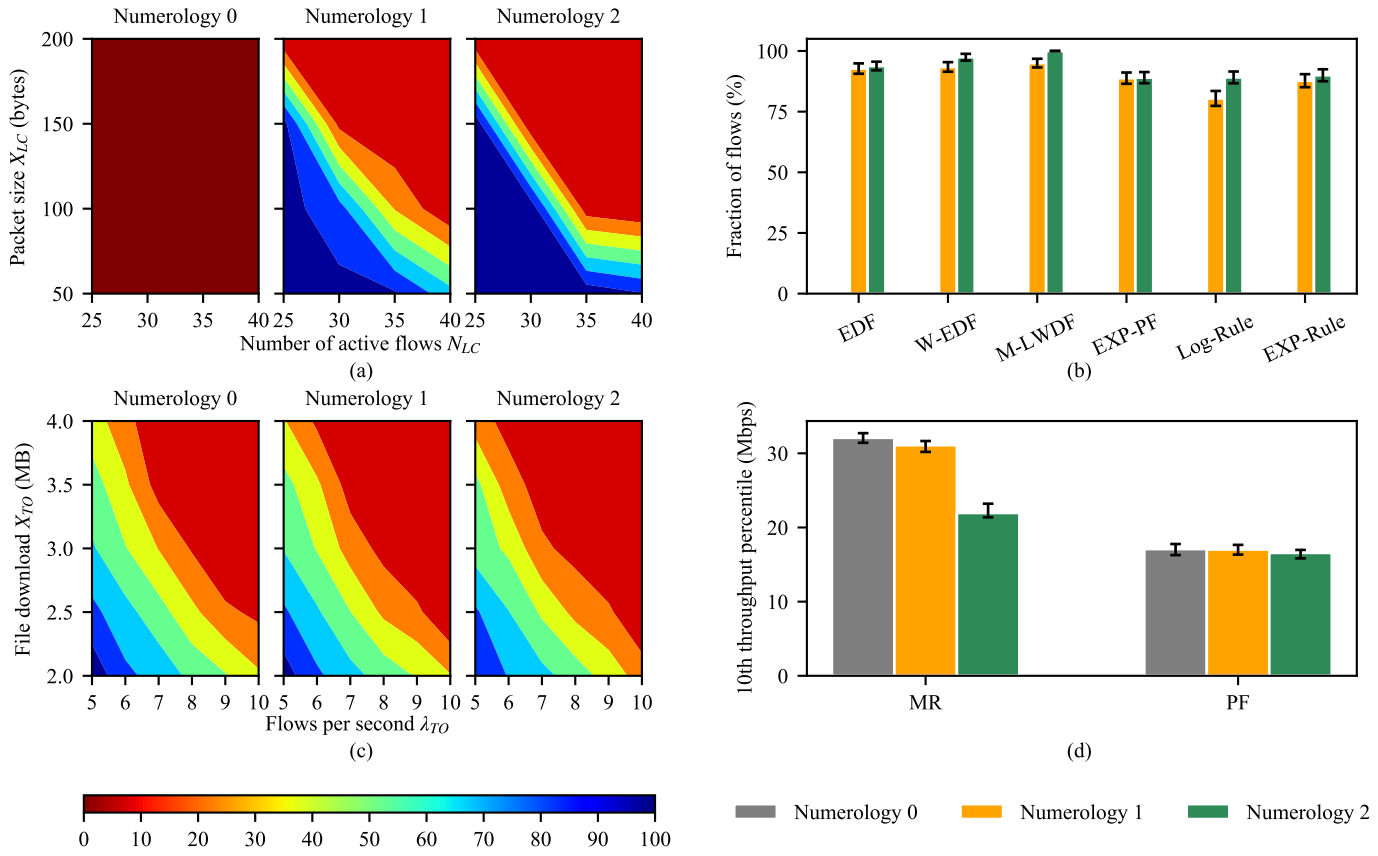

Fig. 2. (a) Fraction of LC flows meeting the reliability requirement for a *LC slice* when the M-LWDF scheduler is used. (b) Fraction of LC flows meeting the reliability requirement for a *LC slice* with 25 LC flows and 150-byte packet sizes. (c) 10th throughput percentile in Mbps for a *TO slice* when the MR scheduler is used. (d) 10th throughput percentile for a *TO slice* with an offered traffic load of 6 TO flows per second and file sizes of 3.5 MB.

limits the gains of retransmissions at numerology 2 as the retransmitted packets are closer to their deadline than packets that are transmitted for the first time. The Log-Rule scheduler performs worse than the EXP-Rule scheduler for numerology 1 due to its logarithmic component, while for numerology 2 they perform similarly as the Log-Rule scheduler benefits more from the retransmissions than the EXP-Rule scheduler.

B. Sliced Scenario: Throughput-Oriented Slice

Equivalently as for the LC slice, in the *TO slice*, we vary the arrival rate λ_{TO} of TO flows and the download file size X_{TO} to evaluate the impact of numerology on the 10th throughput percentile when the MR scheduler is used. Fig. 2c shows the results measured in Mbps and it is observed that the throughput decreases slightly in the numerology due to reduced gains from frequency-domain channel-adaptive scheduling. For example, for a flow arrival rate of $\lambda_{TO} = 8$ flows per second and $X_{TO} = 3$ MB, the throughput decreases from about 13.5 Mbps to about 9.9 Mbps for numerologies 0 and 2, respectively. This effect is however rather modest in the considered scenarios due to the very good propagation conditions and the lack of interference which implies a generally very high channel quality across all PRBs, offering little potential for frequency-domain scheduling.

For a file size of 3.5 MB, which relates to the Industry 4.0 TO use case [18][23], the maximum arrival rate λ_{TO} that satisfies the KPI requirement for numerology 0 is about 6 flows per second, considering discrete integer choices. For this load scenario, Fig. 2d shows the comparison between the MR and PF schedulers for all three numerologies including 90% confidence intervals for the shown KPI. Observe that the MR scheduler is performing better than the PF scheduler and thus *the optimal configuration for the TO slice is the MR scheduler with numerology 0*. The good propagation conditions in combination with the non-persistent nature of the flows allow the MR scheduler to more efficiently use the resources and make the channel more quickly available to the flows that experience lesser good channels. The fairness aspect of the PF scheduler is effectively reducing all transmission rates, resulting in a reduced 10th throughput percentile. Furthermore, the fair design of the PF scheduler prevents the full exploitation of frequency diversity and thus the attained throughput gains for higher numerologies are also relatively modest.

C. Non-Sliced Scenario

For the non-sliced scenario we combine the two maximum class-specific loads found for the sliced scenario, which still satisfy the KPI targets, from the previously considered slices,

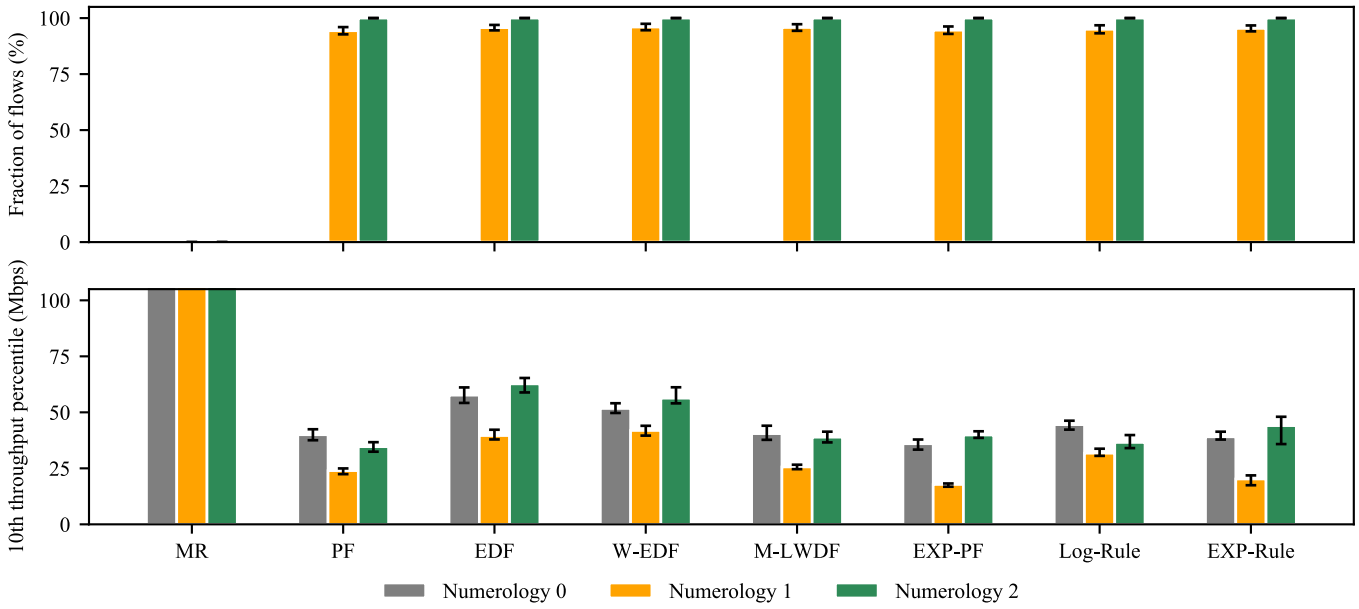

Fig. 3. Fraction of LC flows meeting the reliability requirement and 10th throughput percentile for TO flows in a non-sliced scenario.

i.e. 25 persistent LC flows transmitting 150-byte packets and non-persistent TO flows originating at a rate of 6 flows per second and with file size of 3.5 MB. Also, the full bandwidth is available as it does not have to be split between slices. Fig. 3 shows the fraction of LC flows that meet the 99.9% reliability requirement and the 10th throughput percentile of TO flows for different schedulers and for all three numerologies.

Regarding the impact of numerology on the performance of LC flows, a similar observation as for the sliced scenario holds: with numerology 0 none of the flows meet the reliability requirement due to the high processing latencies regardless of the scheduler, while with numerology 2 all flows meet the reliability requirement due to the possibility of a retransmission.

Moreover, Fig. 3 illustrates that all schedulers designed to support LC flows (EDF, W-EDF, M-LWDF, EXP-PF, Log-Rule and EXP-Rule) are performing similarly and with numerology 2 they all satisfy the KPI target. Because resources are not split over distinct slices, the above-mentioned schedulers can assign more resources to the LC flows compared to the sliced scenario which improves the performance of LC flows. For example, with the M-LWDF scheduler and numerology 2, the average packet latency is 1.12 ms and 0.92 ms for the sliced and non-sliced scenarios, respectively. In other words, packets are transmitted more quickly in the non-sliced scenario compared to the sliced scenario. Additionally, Fig. 3 shows that the MR scheduler is outperformed by all the other schedulers from the perspective of the LC flows, regardless of the choice of numerology. This clearly reveals the unsuitability of the MR scheduler for LC flows. On the other hand, even though the PF scheduler is also not specifically designed to support LC flows, its fair design, in combination with the trunking gains inherent to the non-sliced scenario, make the PF scheduler perform similarly as those schedulers that have been specifically designed to support LC flows.

Fig. 3 also shows the impact of the numerology on the 10th throughput percentile of the TO flows which is different compared to the sliced scenario. The impact of the numerology on the 10th throughput percentile is the net effect that a higher numerology has in terms of *(i)* an increased load, since fewer LC packets are dropped and hence the carried LC traffic load is larger; *(ii)* reduced frequency-diversity gains; and *(iii)* a reduced transfer time of LC packets. When the numerology increases from 0 to 1, there is a significant increase of carried traffic. For example, with the EDF scheduler, the percentage of packets of LC flows that are dropped by the scheduler with numerology 0 and 1 are 22.698% and 0.024%, respectively, resulting in more LC packet transmissions with numerology 1. This traffic increase is the dominant factor and causes the observed drop in the 10th throughput percentile of TO flows. When the numerology increases from 1 to 2, there is a further (yet more modest) traffic increase, as packet retransmissions of LC flows occur, and also a further reduction of the frequency-diversity gains. These effects are however relatively modest compared to the gains due to faster LC packet transmissions, which dominate in causing the observed increase in the 10th throughput percentile.

Regarding the performance of schedulers on the 10th throughput percentile of TO flows, the MR scheduler is performing significantly better than the other schedulers for all three numerologies and also in comparison with the sliced scenario. Specifically, with numerology 0, the 10th throughput percentile increases by about a factor six compared to the

sliced scenario which is primarily due to trunking gains. However, this throughput increase comes at the performance cost of the LC flows, as noted before. The PF, M-LWDF, EXP-PF, Log-Rule and EXP-Rule schedulers perform similarly as they serve TO flows with the same scheduling rule. Their small performance differences are based on the efficiency of each scheduler to serve LC flows. Also, observe that the PF scheduler performs better compared to the sliced scenario due to trunking gains. Further, Fig. 3 shows that the EDF and W-EDF schedulers perform the best after the MR scheduler in terms of the 10th throughput percentile of TO flows. The non-persistent nature of the TO flows allows the two schedulers to efficiently transmit the packets of LC flows and make the channel more quickly available to TO flows. Considering that, only with numerology 2, the KPI for LC flows is met for all schedulers, except for the MR scheduler, *the optimal combination for the non-sliced scenario is given by the EDF scheduler in combination with numerology 2*, noting that (when disregarding the MR scheduler) the EDF scheduler provides the highest 10th throughput percentile for TO flows.

D. Comparison of Scenarios

To quantify the gains from the non-sliced scenario over the sliced scenario, we conducted additional simulations. We gradually increased the aggregate traffic in the non-sliced scenario, up to the level where the non-sliced scenario no longer outperforms the sliced scenario. This study revealed that a load increase of up to 20% can be handled in a non-sliced scenario. Comparing the optimal sliced and the optimal non-sliced scenarios, as defined in the previous subsections, we conclude that the non-sliced scenario performs better than the sliced scenario due to the trunking gains that are bigger than the gains from separately configuring slices.

V. Concluding Remarks

The need to support new services with diverse requirements has introduced the concepts of flexible numerology and network slicing in 5G networks. There is evidence in literature that the QoS requirements for particular services can be efficiently guaranteed with the use of RAN slicing or with novel packet schedulers and/or the use of flexible numerology. In this study, we have compared an optimal sliced scenario with isolated slices and an optimal non-sliced scenario. We showed that the trunking gains obtained from the non-sliced scenario are greater than the gains obtained by separately optimizing the packet scheduler and numerology for each slice. In particular, we show that the non-sliced scenario can serve about 20% more traffic than the sliced scenario while providing the required performance to each service class.

In the current study, we considered that the two slices are isolated and therefore idle resources of a slice cannot be used by another slice which limits the performance of sliced scenarios. As a next step, we will compare the gains of dynamic slicing (with heuristic and machine-learning-based methods) to the gains of non-sliced scenarios. Furthermore, an extension of the study to multi-tenant networks and to networks with even more diverse requirements, more challenging propagation conditions and interference is recommended.

Acknowledgment

This work is part of NExTWORKx, a collaboration between TU Delft and KPN on future telecommunication networks.

References

[1] S. E. Elayoubi, M. Fallgren, P. Spapis, G. Zimmermann, D. Martin-Sacristan, C. Yang, S. Jeux, P. Agyapong, L. Campoy, Y. Qi and S. Singh, "5G service requirements and operational use cases: Analysis and METIS II vision," in *Proceedings of EuCNC '16*, 2016.

[2] 3GPP, TS 38.211, "NR; Physical channels and modulation," v16.3.0, 2020.

[3] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia and P. Camarda "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 2, 2013.

[4] C.-Y. Chang and N. Nikaein, "RAN runtime slicing system for flexible and dynamic service execution environment," *IEEE Access*, vol. 6, 2018.

[5] S. Khatibi and A. Jano, "Elastic slice-aware radio resource management with AI-traffic prediction," in *Proceedings of EuCNC '19*, 2019.

[6] K. I. Pedersen, G. Pocovi, J. Steiner and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proceedings of VTC-Fall '17*, 2017.

[7] A. Akhtar and H. Arslan, "Downlink resource allocation and packet scheduling in multi-numerology wireless systems," in *Proceedings of WCNCW '18*, 2018.

[8] A. A. Zaidi, R. Baldemair, V. Moles-Cases, N. He, K. Werner and A. Cedergre, "OFDM numerology design for 5G new radio to support IoT, eMBB and MBSFN," *IEEE Communications Standards Magazine*, vol. 2, no. 2, 2018.

[9] N. Bechir, M. Nasreddine, A. Mahmoud, H. Walid and M. Sofien, "Novel scheduling algorithm for 3GPP downlink LTE cellular network," *Procedia Computer Science*, vol. 40, 2014.

[10] V. Nair, R. Litjens and H. Zhang, "Optimisation of NB-IoT deployment for smart energy distribution networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, 2019.

[11] 3GPP, TR 38.901, "Study on channel model for frequencies from 0.5 to 100 GHz," v16.1.0, 2020.

[12] S. Jaeckel, L. Raschkowski, K. Borner, L. Thiele, F. Burkhardt and E. Eberlein, "QuaDRiGa - Quasi deterministic radio channel generator: User manual and documentation," v2.2.0, 2019.

[13] 3GPP, TS 38.214, "NR, physical layer procedures for data," v16.3.0, 2020.

[14] S. Pratschner, B. Tahir, L. Marijanovic, M. Mussbah, K. Kirev, R. Nissel, S.Schwarz and M. Rupp, "Versatile mobile communications simulation: The Vienna 5G link level simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, 2018.

[15] K. Sayana, J. Zhuang and K. Stewart, "Link performance abstraction based on mean mutual information per bit (MMIB) for the LLR channel," 2007.

[16] K. I. Pedersen, G. Monghal, I. Z. Kovacs, T. E. Kolding, A. Pokhariyal, F. Frederiksen and P. Mogensen, "Frequency Domain Scheduling for OFDMA with Limited and Noisy Channel Feedback," in *Proceedings of VTC-Fall '07*, 2007.

[17] S. N. Anbalagan, R. Litjens, K. Das, A. Chiumento, P. Havinga and J. L. van den Berg, "A Sensitivity Analysis on the Potential of 5G Channel Quality Prediction," in *Proceedings of VTC-Spring '21*, 2021.

[18] 3GPP, TS 22.104, "Service requirements for cyber-physical control applications in vertical domains," v17.4.0, 2020.

[19] 3GPP, TS 38.213, "NR, physical layer procedures for control," v16.3.0, 2020.

[20] Qualcomm Incorporated, "Summary of DL/UL scheduling and HARQ management," R1-1721652, 3GPP TSG-RAN WG1 Meeting, 2017.

[21] 5G Americas, "New services and applications with 5G ultra-reliable low latency communications," 2018.

[22] Ericsson, "UP latency in NR," R2-1711550, 3GPP TSG-RAN WG2 Meeting, 2017.

[23] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz and J. M. Lopez-Soler, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, 2020.