Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations

Pedersen, Klaus I.; Pocovi, Guillermo; Steiner, Jens; Maeder, Andreas

Published in:
I E E E Communications Magazine

DOI (link to publication from Publisher):
10.1109/MCOM.2017.1700517

Creative Commons License
Unspecified

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Pedersen, K. I., Pocovi, G., Steiner, J., & Maeder, A. (2018). Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations. *I E E E Communications Magazine*, *56*(3), 210-217. https://doi.org/10.1109/MCOM.2017.1700517

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

ACCEPTED FROM OPEN CALL

Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations

Klaus Pedersen, Guillermo Pocovi, Jens Steiner, and Andreas Maeder

The authors present a holistic overview of the agile multi-user scheduling functionality in 5G. An E2E perspective is given, including the enhanced QoS architecture that comes with 5G, and the large number of scheduling related options from the new access stratum sub-layer, MAC, and PHY layer. A survey of the 5G design agreements from the recently concluded 5G Study in 3GPP is presented.

Abstract

In this article, we present a holistic overview of the agile multi-user scheduling functionality in 5G. An E2E perspective is given, including the enhanced QoS architecture that comes with 5G, and the large number of scheduling related options from the new access stratum sub-layer, MAC, and PHY layer. A survey of the 5G design agreements from the recently concluded 5G Study in 3GPP is presented, and it is explained how to best utilize all these new degrees of freedom to arrive at an agile scheduling design that offers superior E2E performance for a variety of services with highly diverse QoS requirements. Enhancements to ensure efficient implementation of the 5G scheduler for different network architectures are outlined. Finally, state-of-the-art system level performance results are presented, showing the ability to efficiently multiplex services with highly diverse QoS requirements.

Introduction

An impressive amount of research related to the upcoming 5G has been published in recent years; as an example, see the survey in [1]. This has formed a solid foundation for progressing also with the 3GPP standardization of 5G, which has recently achieved an important milestone with the completion of the 5G New Radio (NR) Study, as captured in the technical reports [2] and [3]. Although 3GPP has adopted the name NR, we will use the term "5G" throughout this article. The 5G system is set to deliver superior performance for three main service categories: enhanced mobile broadband (eMBB), ultra-reliable low latency communication (URLLC), and massive machine type of communication (mMTC). In achieving the 5G targets, the packet scheduler plays an important role, both in terms of fulfilling the End-to-End (E2E) Quality-of-Service (QoS) performance targets for each session, as well as in efficiently multiplexing and orchestrating a large number of sessions with highly diverse QoS requirements in one unified system. Specifically, efficient scheduling of URLLC traffic represents a challenging problem, as URLLC is associated with a strict latency target of only 1 ms from the time a packet is delivered to Layer 3/2 in the 5G Radio Access Network (RAN) until it is successfully received, with an outage probability of only 10^{-5}. In addition to the multi-service dimension, the 5G design is also set to scale to a variety of different network implementations. Those ambitious requirements for 5G will help meet the growing demands for future mobile broadband, as well as enable a plethora of new applications, e.g. industrial wireless control, autonomous vehicles (cars, drones, and so on), high quality virtual reality, and remote healthcare.

In this article, we first present a survey of the most important design decisions made in 3GPP that relates to the 5G scheduler design, and in particular to the E2E service delivery capabilities. We explain the rationales behind those design choices, and offer additional insight into how to most efficiently utilize and benefit from the large degrees of freedom that the new 5G design is set to provide. The new QoS architecture for 5G is presented, highlighting the possibilities of enhanced high-layer scheduling functionality at a new access stratum sub-layer that works in harmony with the advanced Medium Access Control (MAC) layer scheduler, which sits closer to the radio interface, often referred to as the radio scheduler. The 5G radio scheduler comes with many new innovations, especially enabled by the flexible physical layer design. We aim at explaining how the radio scheduler can take advantage of the enhanced 5G physical layer design. Examples of system level performance are presented for two different use cases to illustrate how the 5G scheduler offers improvements, and how those translate to improved E2E performance. To conclude the study, the different enhancements that contribute to the flexible and responsive 5G scheduler design are summarized in Table 1 at the end of this article. In line with the 3GPP terminology, we refer to terminals as user equipment (UE) and a base station as a "gNB" (fifth generation Node-B).

QoS Control and Protocol Framework

The 5G design includes a new QoS service architecture (as compared to LTE), and several enhancements to the protocol stack [2]. This is illustrated in Fig. 1, where the QoS architecture is pictured on the left and the user plane protocol stack on the right. The non-access stratum (NAS) filters the data packets in the UE and the 5G core network (CN) to associate the data packets with QoS flows. One or more QoS flows are associated to an E2E session, which is capable of trans-

Digital Object Identifier:
10.1109/MCOM.2017.1700517

The authors are with Nokia Bell Labs.

0163-6804/18/$25.00 © 2018 IEEE

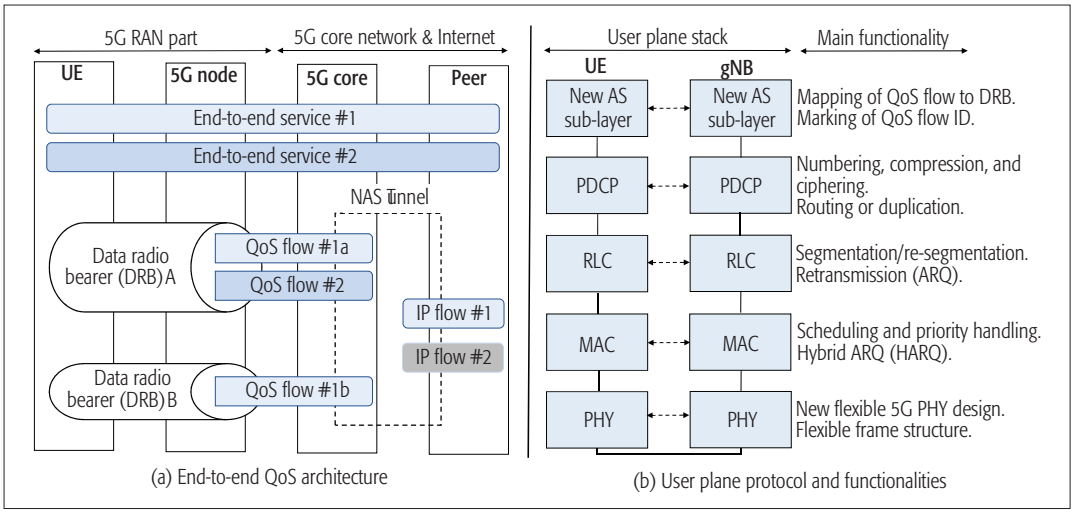


Figure 1. QoS architecture, user plane protocol stack, and related functionalities.

The mapping of an E2E session to QoS flows, and DRBs can be updated dynamically. This kind of flexibility presents opportunities for applying state-of-the-art higher-layer scheduling policies that differentiate application flows, via the mapping to DRBs, as well as adaptation of DRB requirements for the radio scheduler.

porting IP, Ethernet, or unstructured datagrams (the latter, e.g., for raw machine-type-communication data). For each UE, at least one packet session is established. The access stratum (AS) mapping in the UE and the 5G RAN associates the QoS flows with the data radio bearers (DRBs). This mapping is based on 5G QoS class indices (5QI) in the transport header of the packets, and on corresponding QoS parameters, which are signaled via CN interface when a packet session is established. As illustrated in Fig. 1, one or more QoS flow(s) can be mapped to a DRB. Hence, the 5G CN and RAN ensures the QoS in harmony by intelligent mapping of QoS flows and DRBs, essentially constituting a two-step mapping of E2E session flows (e.g., IP-flows) to QoS flows and subsequently to DRBs.

In the 5G RAN at least one default DRB is established for each UE when a new E2E packet session is created. As illustrated in Fig. 1, an E2E packet session may be mapped to two different QoS flows and DRBs to facilitate cases where the E2E packet session contains data flows with two different sets of QoS requirements, such as, e.g., a website with embedded high-definition live streaming video. The 5G RAN may choose to, e.g., map a guaranteed bit rate (GBR), or multiple GBR flows to the same DRB. The mapping of an E2E session to QoS flows, and DRBs can be updated dynamically. This kind of flexibility presents opportunities for applying state-of-the-art higher-layer scheduling policies that differentiate application flows, via the mapping to DRBs, as well as adaptation of DRB requirements for the radio scheduler. The latter mechanisms are also sometimes referred to as higher-layer application-aware scheduling [4], or advanced Quality of Experience (QoE) management [5].

On the terminal side, the concept of reflective QoS eliminates the need to use dedicated flow filters signaled by the network to match traffic to QoS flows. This was one of the main reasons why in LTE, IP traffic was always mapped to default DRBs. In reflective QoS, the terminal derives the mapping of uplink traffic to QoS flows by correlating the corresponding downlink traffic and its attributes, e.g., in Transport Control Protocol (TCP) flows.

On the 5G radio interface, the packet treatment is defined separately for each DRB. Different DRBs may be established for QoS flows requiring different packet forwarding treatment (e.g., associated with different requirements such as latency budget, packet loss rate tolerance, GBR). As will be described in greater detail later, the MAC-level scheduler aims at fulfilling the requirements for the users' DRBs, as well as to prioritize accordingly if the system reaches congestion where requirements for all users cannot be simultaneously fulfilled.

The user-plane protocol stack for 5G is illustrated in the right part of Fig. 1. Here, a new AS sub-layer (with Service Data Application Protocol) is included that is responsible for the aforementioned mapping of QoS flows to DRB and the related marking in uplink. The proposed QoE Manager in [5] can be implemented in this sub-layer. As an example, for the use case of YouTube streaming, the QoE manager in the AS sub-layer may adaptively monitor and adjust the mapping of QoS flow to DRB, adjusting, e.g., the GBR and latency budget associated with the DRB to guide the lower-layer radio scheduler, and ensure a positive end-user experience where the playout of the video starts quickly and runs smoothly without any re-buffering events (for more details we refer to [5]). For the majority of cases, it is envisioned that all traffic for a UE is mapped to a single (or few) DRB, while the QoE manager at the AS sub-layer takes care of differentiation, e.g., by modifying packet priorities, while only seldom modifying the QoS parameters of the DRB that the MAC scheduler shall fulfill. Thereby, the QoE manager (aka application-layer scheduler) is operated in harmony with the lower-layer radio scheduler to avoid the well known double responsibility conflict problem from control theory, that is, avoiding the scenario in which the higher-layer and lower-layer schedulers in the worst case make colliding decisions that result in undesirable behaviors.

The packet data convergence protocol (PDCP) layer for 5G inherits the fundamentals from LTE, but also brings valuable enhancements [2]. Among those, PDCP packet duplication is supported as a means to improve the end-user

The MAC scheduler works by dynamically allocating radio transmission resources (transport blocks) on a per-user basis for downlink and uplink transmissions, separately. The objective of the scheduler is to fulfill the QoS service targets for all the DRBs of the served UEs.

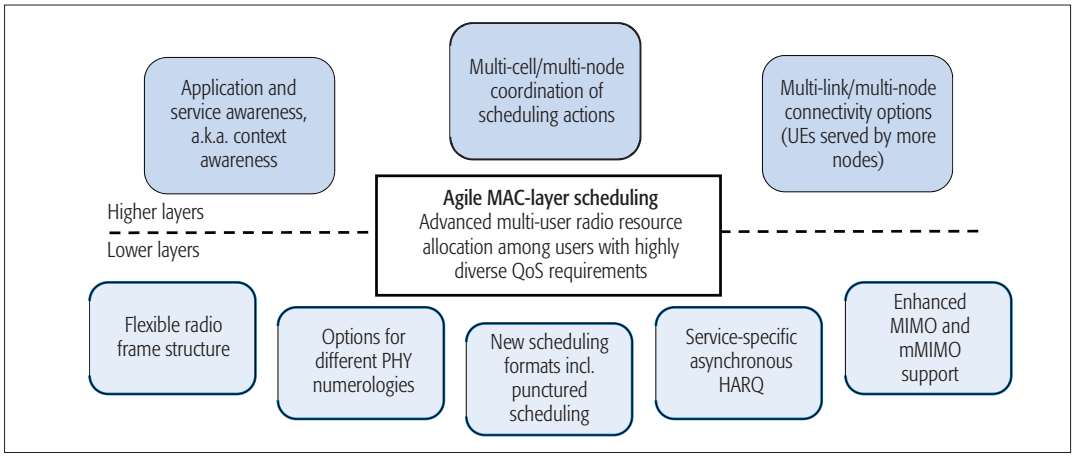


Figure 2. High-level overview of MAC dynamic scheduler interfaces and options.

packet reception reliability, thus being one of the enablers for reaching the reliability part of the 5G URLLC requirement. This means that if a UE is configured with, e.g., carrier aggregation or multi-node connectivity [1, 4], the same PDCP packet can be duplicated and sent via different transmission paths, thereby reducing the probability of losing packets. Furthermore, PDCP is responsible for packet re-ordering in case the lower layers do not deliver in-sequence. The Radio Link Control (RLC) includes segmentation and Automatic request repeat (ARQ), while the Medium Access Control (MAC) is the home of the agile radio-layer packet scheduler and the Hybrid ARQ (HARQ) functionality. A large set of PHYsical (PHY) enhancements are coming with 5G [3], which offers significant degrees of freedom for the multi-user, multi-service capable radio scheduler (discussed in more detail in later sections). On a further note, the concept of network slicing is also supported, where different types of traffic could be handled by separate slices. The network may realize different slices by mapping data to different QoS flows/DRBs based on slice-specific policies, and by scheduling. UEs should be able to aid information related to slice selection, if it has been provided by the NAS (for more information on slicing, see [2, 4]). As will be further described in the rest of the article, the combined benefits of the new QoS architecture and enhanced protocol flexibility and features result in improved E2E performance.

MAC Scheduler Overview

A high-level overview of the 5G MAC scheduler functionality is pictured in Fig. 2. The MAC scheduler is the controlling entity for multi-user radio resource allocations, which are subject to several constraints but also many options for efficiently serving the different terminals. The enlarged number of options for the 5G MAC scheduler, as compared to LTE, naturally offers performance improvements, but also presents a non-trivial problem of how to best utilize those degrees of freedom in an efficient manner. The MAC scheduler works by dynamically allocating radio transmission resources (transport blocks) on a per-user basis for downlink and uplink transmissions, separately. The objective of the scheduler is to fulfill the QoS service targets for all the DRBs of the served UEs. This is illustrated in Fig. 2, where the application and service awareness is provided by the higher layers as discussed in the previous section. Furthermore, the scheduler has to support multi-cell connectivity mode [1, 2], where UEs are configured to be simultaneously served by multiple nodes (and cells). Additionally, there may be other multi-cell coordination constraints, e.g., if enforcing inter-cell interference coordination between neighboring cells where certain radio resources are dynamically muted, and hence not available for dynamic scheduling of users. At the MAC sub-layer, enhanced service-specific HARQ enhancements are included [6]. The 5G PHY layer offers a large set of new options for the MAC scheduler, which enable significant improvements for efficiently multiplexing users with highly diverse service requirements.

Figure 3 presents further information on multiplexing of users on the PHY layer and the related MAC sublayer functionality. 5G comes with a new flexible structure, consisting of 10 ms radio frames and 1 ms subframes. The subframes are constructed of slot building blocks of seven OFDM symbols. For FDD cases, the slots are naturally all downlink (for the downlink band), and all uplink (for the uplink band), while for TDD cases the slots can also be bidirectional (starting with downlink transmission followed by uplink transmission). To support operation in different frequency bands, the PHY numerology is configurable, building on the same base subcarrier spacing (SCS) of 15 kHz as used in LTE. The SCS can scale from the base value by a factor 2^N, where $N \in [0, 1, 2, 3, 4, 5]$. For 15 kHz SCS ($N = 0$), the slot duration is 0.5 ms, while it equals 0.25 ms for 30 kHz ($N = 1$). Furthermore, mini-slots of 1-3 OFDM symbols are defined as well [3]. The smallest time-domain scheduling resolution for the MAC scheduler is a mini-slot, but it is also possible to schedule users on slot resolution, or on resolution of multiple slots (aka slot aggregation). This essentially means that dynamic scheduling with different transmission time interval (TTI) sizes is supported. The latter enables the MAC scheduler to more efficiently match the radio resource allocations for different users in coherence with the radio condition, QoS requirements, and cell load conditions [7–9]. The short TTI size is needed for URLLC use cases [10], but not restricted to such traffic. In the frequency domain, the minimum scheduling resolution is

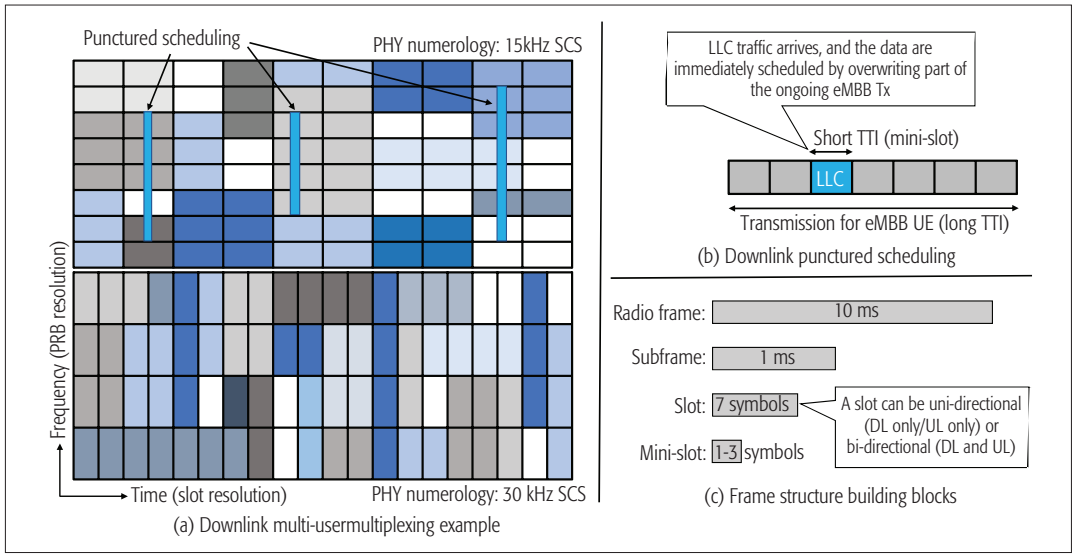


Figure 3. Resource allocation framework: a) downlink multi-user mux; b) punctured scheduling; c) frame structure building blocks.

Each scheduling allocation (downlink and uplink) is announced to the UE via a PHY downlink control channel carrying the scheduling grant. The downlink control channel is flexibly time-frequency multiplexed with the other downlink PHY channels, and can be mapped contiguously or non-contiguously in the frequency domain. This constitutes a highly flexible design.

one physical resource block of 12 subcarriers, corresponding to 180 kHz for 15 kHz SCS, 360 kHz for 30 kHz, and so forth [3].

Figure 3a shows how different users are multiplexed in the downlink on an FDD carrier (different colors represent transmissions to different users). As can be seen from this example, the majority of the users are multiplexed on slot resolution. Users can be dynamically scheduled with a TTI size of one slot, or multiple slots. For the example in Fig. 3a, the carrier is configured to allow frequency domain multiplexing of two different PHY numerologies, namely 15 kHz (upper part) and 30 kHz (lower part). The MAC scheduler can freely decide how to schedule its different users on the carriers (i.e., on which PHY numerology, with which TTI sizes, and so on), and it is not visible to the RLC layer how this is done. It is, however, possible to enforce some restrictions via higher-layer control signaling (radio resource control (RRC)) to schedule data from certain DRBs only on a given PHY numerology, and a certain TTI size. Each scheduling allocation (downlink and uplink) is announced to the UE via a PHY downlink control channel carrying the scheduling grant. The downlink control channel is flexibly time-frequency multiplexed with the other downlink PHY channels, and can be mapped contiguously or non-contiguously in the frequency domain. This constitutes a highly flexible design, where the relative downlink control channel overhead can take values from sub-one-percentage values (if, e.g., scheduling a few users with long TTI size) and up to tens of percentages if scheduling a larger number of users with very short TTI sizes [7, 9]. The design, therefore, overcomes the control channel blocking problems from LTE (and LTE-Advanced) as reported in [11, 12]. As studied in [7–9], these advantages are achieved by migrating toward a user-centric design with in-resource control channel signaling, as compared to the predominantly cell-centric LTE design. Another advantage brought by the more flexible 5G downlink control channel design is the support of UEs that only operate on a fraction of the carrier bandwidth (e.g., narrowband MTC devices). As is further described in [3], resource allocation for UEs not capable of supporting the full carrier bandwidth is derived based on a two-step frequency-domain assignment process.

Figure 3 also illustrates the principle of punctured scheduling for efficient expedition of Low Latency Communication (LLC) traffic. Efficient scheduling of LLC is rather challenging, as such traffic is typically bursty (random nature) and requires immediate scheduling with short TTIs to fulfill the corresponding latency budget [10]. Instead of pre-reserving radio resources for LLC traffic bursts (that may or may not come), it is proposed to use punctured scheduling, which is inherited from the preemptive scheduling ideas known from real-time scheduling in computer networks. The basic principle is as follows [13]. Traffic such as eMBB is scheduled on all the available radio resources (whenever there is sufficient offered traffic). Once an LLC packet arrives at the gNB, the MAC scheduler immediately transmits it to the designated terminal by overwriting part of an ongoing scheduled transmission, using mini-slot transmission, as illustrated in both Fig. 3a and 3b. This has the advantage that the LLC payload is transmitted immediately without waiting for ongoing scheduled transmissions to be completed, and without the need for pre-reserving radio resources for LLC traffic. The price of puncturing is for the user whose parts of its transmission are overwritten. To minimize the impact on the users that experience the puncturing, related recovery mechanisms are introduced [13]. Those include physical layer control signaling from the gNB to indicate to the victim terminal that part of its transmission has been punctured. This enables the terminal to take this effect into account when decoding the transmission, that is, it knows that part of the transmission is corrupted. Moreover, options for smart HARQ retransmission options are considered, where the damaged part of the punctured transmission is first retransmitted. The benefits of those options are further illustrated in the section on Performance Results.

As illustrated in Fig. 2, the 5G PHY also offers enhanced antenna techniques, i.e., multiple-in-

The combination of the flexible scheduling timing and asynchronous HARQ is an important enabler that paves the way for supporting cases with decoupled downlink/uplink cell associations, where downlink transmissions are scheduled to the UE from one cell, while uplink transmissions are toward a different cell.

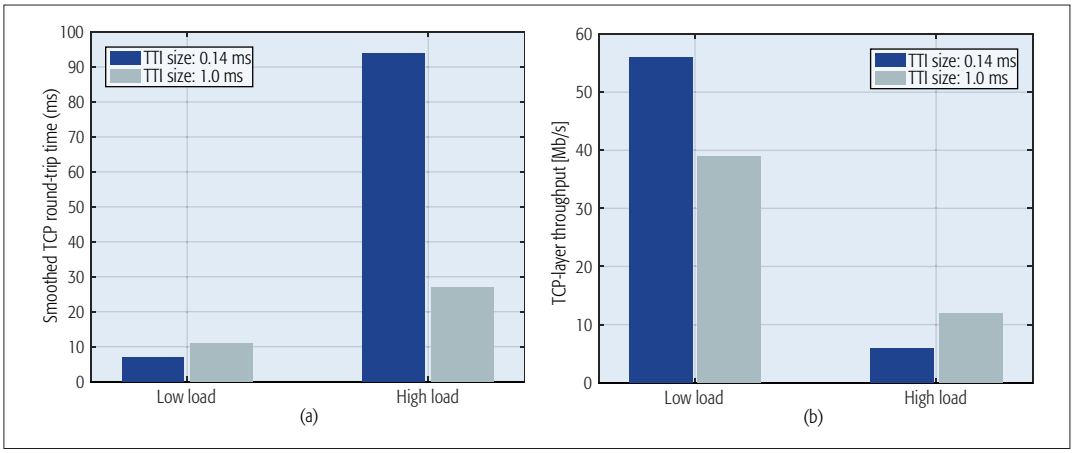


Figure 4. Performance of eMBB with different TTI sizes: a) smoothed median TCP packet round trip time; b) median TCP-layer end-user throughput.

put-multiple-output (MIMO) schemes [1, 3]. For cases with Single-User (SU) MIMO, this makes it possible to schedule up to eight parallel streams of data to one UE on the same PHY resources. Similarly, enhanced Multi-User (MU) MIMO is supported, where streams toward different users can be scheduled on the same PHY resources. This includes massive MIMO (mMIMO) enhancements, where users can be simultaneously scheduled on different beams, allowing flexible support for implementations with digital beamforming, analog beamforming, and hybrids of those two options. For cases with analog beamforming, the MAC scheduler is typically restricted to only apply time-domain multiplexing between users within each beam, although options for frequency domain multiplexing are not excluded.

Flexibility for Different Network Implementations

The 5G scheduler is designed to be applicable for different network implementations [1, 2]. This includes distributed network implementations with separate schedulers implemented in each gNB per cell, as well as more advanced centralized or semi-centralized radio access network solutions [4]. The latter includes cases with a centralized Cloud Edge entity connected via a midhaul interface to a Front End Unit (FEU), which may have RF integrated or may connect to a Remote Radio Head (RRH) via a fronthaul connection [1, 2]. For such advanced network architectures, the implementation of the PDCP and RLC sub-layers is possibly located in the Cloud Edge, while the MAC is distributed over Cloud Edge and FEU, and the PHY is distributed over the FEU and RRH. Given the possible ranges of processing latencies at the different network units, as well as communication latencies over the midhaul and fronthaul interface, the MAC scheduling and the HARQ loop timing of 5G needs to be equally flexible. Thus, in comparison to the strict hardcoded scheduling timing of LTE, 5G offers a much more flexible configuration. The timing between the downlink scheduling and the actual data transmission is indicated as part of the scheduling grant (i.e., on the downlink PHY control channel). The same applies for the timing of uplink data transmissions. The timing relation between the data channel reception, and the time where a corresponding HARQ feedback (positive or negative acknowledgement) shall be sent is also flexibly indicated and configurable. Furthermore, asynchronous HARQ is adopted for both link directions, giving the network full flexibility for deciding when to schedule HARQ retransmissions. See, for instance, the study in [6] where the HARQ round trip timing is studied for cases with different fronthaul latencies. The combination of the flexible scheduling timing and asynchronous HARQ is an important enabler that paves the way for supporting cases with decoupled downlink/uplink cell associations, where downlink transmissions are scheduled to the UE from one cell, while uplink transmissions are toward a different cell [14].

Moreover, as compared to the LTE design, the RLC concatenation is replaced with MAC multiplexing, which allows pre-generation and interleaving of PDCP/RLC/MAC headers. This basically means that the time-consuming generation of RLC Packet Data Units (PDUs) for each new scheduled transport block (i.e., scheduling instant) as done for LTE is avoided. This makes 5G more efficient and flexible, allowing the RLC and MAC entities to, for instance, be implemented on different network elements. See more details in [2].

Performance Results

Performance results from extensive system-level simulations, following the 3GPP 5G simulation guidelines, are presented in the following to illustrate the benefits of some of the 5G scheduling enhancements. Results are presented for a standard three-sector macro scenario, operating at 2 GHz with a 10 MHz carrier bandwidth, assuming 2x2 SU-MIMO and the base PHY numerology (15 kHz SCS). We first present downlink eMBB performance results for file download over TCP, using the well known Reno model [15], hence illustrating E2E eMBB performance. A 2 ms CN delay is assumed from the client to the 5G RAN. Traffic is arriving according to a homogenous Poisson point process, and users are leaving the system when a download of a 500 kB payload is

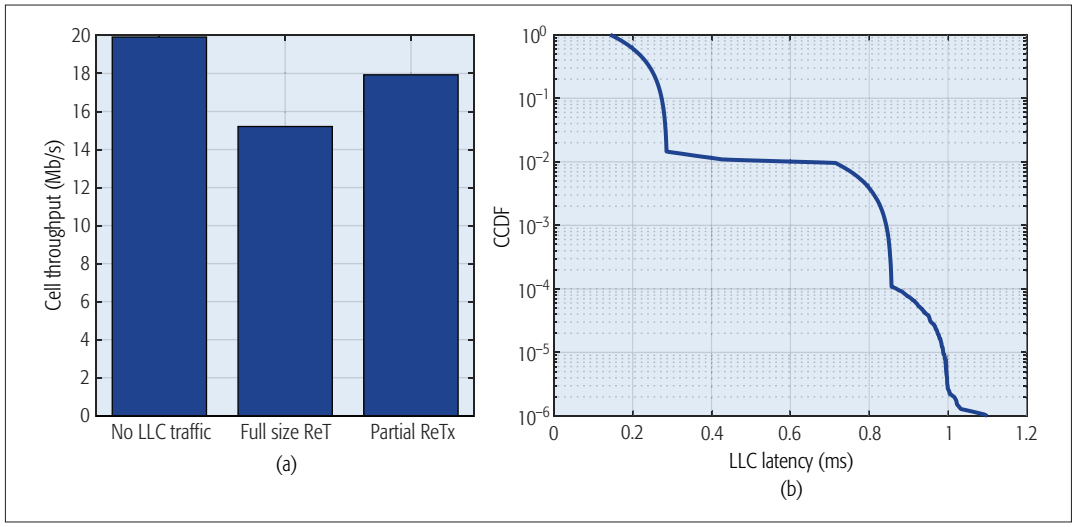


Figure 5. Performance of LLC/eMBB with punctured scheduling: a) average cell throughput; b) ccdf of LLC payload latency.

The 5G system design, and particularly the scheduler related mechanisms at the different layers, presents opportunities for improved E2E performance, capabilities for more efficiently multiplexing users with highly diverse QoS requirements, and flexibility for different network implementations. System-level performance results confirm that the new scheduling functionalities offer promising benefits.

completed. Both cases with low and high offered loads are considered, corresponding to average offered load of 2 Mbps/cell and 16 Mbps/cell, respectively. RLC acknowledged mode is assumed. Fig. 4 shows the performance for short (0.14 ms) and long (1 ms) TTI sizes, considering both the case with low offered traffic and high offered traffic. It is assumed that the HARQ RTT equals 4 TTIs, including the effects of gNB and UE processing times, time for sending ACK/NACK, and so on [9]. One of the reported performance metrics is the smoothed round trip time (RTT) of TCP packets in line with the definition in RFC6298. It is observed that the best performance is achieved for the short TTI at the low offered load. This is due to the lower air interface latency that helps to quickly overcome the slow start TCP phase. The higher PHY control channel overhead from operating with short TTIs is not a problem at the low offered load. However, at the high offered load case, the best performance is clearly observed for the case with the long TTI. This is due to the fact that using longer TTIs results in higher average spectral efficiency. If operating with the short TTI size (at high offered load), excessive queuing delays are observed at the gNB due to the lower spectral efficiency because of higher PHY control channel overhead. Thus, the results in Fig. 4 clearly show the benefit of being able to dynamically adjust the TTI size. See the study in [9] for additional insight.

Next, we present downlink performance for a mixture of eMBB and low latency communication (LLC) type of traffic. In this example, there are on average five active eMBB users per macro-cell, performing a download with 500 kB file size, using TCP. As soon as one of eMBB users finishes its file download, the user is removed and a new one is generated at a random location. In addition, there are on average 10 LLC users per cell, where small latency critical payloads of 50 Bytes are sporadically generated according to a homogeneous Poisson point process, arriving in the gNB. As this scenario corresponds to a fully loaded network, eMBB users are scheduled with a TTI size of 1 ms, using all available PRBs. Hence, no radio resources are reserved for potentially coming LLC traffic. Instead, punctured scheduling is applied whenever LLC payloads appear in the gNB. The LLC payloads are immediately scheduled on arrival with mini-slot resolution (0.14 ms TTI size), overwriting part of the ongoing eMBB scheduled transmissions as also illustrated in Fig. 3a and 3b. Due to the urgency of the LLC traffic, we assume RLC transparent mode, and a low initial Block Error Rate (BLER) of only 1 percent for such transmissions to avoid too many HARQ retransmissions. The average cell throughput is illustrated in Fig. 5a, where the performance is shown for cases with/without LLC traffic. For the cases with LLC traffic, the offered load is such that approximately 12 percent of radio resources are used for LLC. Two sets of results are shown for the case with LLC traffic: one for the case where the full transport block is retransmitted for failed eMBB HARQ transmissions, and a case where only the damaged part of the eMBB transmission that has been subject to puncturing is retransmitted (labelled as partial retransmission in Fig. 5a). As observed from Fig. 5a, the latter option is clearly the most promising solution, as fewer radio resources for HARQ retransmissions of eMBB transmissions that have suffered from puncturing are used. However, the cost of using this approach is a slightly larger latency for the eMBB users, as the probability of triggering a second HARQ retransmission is higher, as compared to the case where the first HARQ retransmission includes the full transport block. Fig. 5b shows the complementary cumulative distribution function (ccdf) of latency of LLC traffic. The latency is measured from the time when the LLC payload arrives at the gNB until it is correctly received by the UE. The ccdf shows that even under the considered full load conditions, the performance of the LLC traffic fulfills the challenging URLLC target of 1 ms latency with an outage probability of only 10-5 (i.e., one out of 100,000 LLC payloads exceeds the 1 ms latency target). Hence, the punctured scheduling scheme fulfills its purpose, i.e., being able to efficiently schedule the LLC traffic in line with its challenging latency and reliability constraints, while still having efficient scheduling of eMBB traffic without the need for pre-reservation

Functionality	Summary	Benefit
New end-to-end QoS architecture	User data packets are mapped to QoS flows at the UE and CN. UE and RAN maps the QoSs flows to DRBs. DRBs carry QoS flow(s) over the radio interface. QoS differentiation inside NG3 connection is based on packet based QoS flows. Mapping relationship between sessions and DRB is 1 to N and between QoS flows and DRBs N to N.	Improved end-to-end QoS control and orchestration.
Packet duplication	Duplication solution for CA and multi-node connectivity cases can use PDCP duplication, so duplicated PDCP packets are sent over different carriers. Supported for both link directions.	Improved RAN reliability.
DRB mapping to PHY	Data from a DRB can be mapped to one or more lower layer PHY numerologies and TTI sizes. It is transparent to the RLC which PHY / TTI is used. The DRB to lower layer PHY mapping can, however, be reconfigured via higher layer RRC reconfigurations.	Full flexibility for optimization of per data flow.
MAC layer concatenation	Replacing RLC concatenation with MAC Multiplexing allows pre-generating and interleaving PDCP/RLC/MAC headers with the respective data blocks. Thereby overcoming the time-consuming on-the-fly generation of RLC packet data units (PDUs) for each new scheduling grant as done for LTE.	Optimized PDU generation, offering higher degrees of freedom for network implementations where e.g. the RLC and MAC is implemented on different hardware units.
Flexible scheduling timing	Timing between DL scheduling grant and corresponding DL data transmission is indicated as part of the scheduling grant (PHY control channel). Timing between UL scheduling assignment and corresponding UL data transmission is indicated as part of the scheduling grant (PHY control channel). Timing between DL data reception and corresponding HARQ ACK/NACK is indicated as part of the scheduling grant (PHY control channel).	Flexible timing for different network implementations, e.g. cloud RAN with different fronthaul latencies and processing time capabilities.
HARQ characteristics	Asynchronous HARQ for both link directions. Support for specific HARQ enhancements such as automatic retransmissions (low latency) and multi-bit HARQ feedback to enable variable block HARQ retransmissions (mainly relevant for large transport block size eMBB transmissions).	Increased timing and scheduling flexibility. Optimized resource efficiency for retransmissions.
Control channel flexibility	The control channel carrying the scheduling grant (NR-PDCCH) can be flexible time-frequency multiplexed with the other downlink PHY channels. Resource allocation for data transmission for a UE not capable of supporting the full carrier bandwidth can be derived based on a two-step frequency-domain assignment process.	Scalable solution, where known problems of control channel blocking from LTE are circumvented.
Variable TTI sizes	Dynamic scheduling with variable TTI sizes is supported. The TTI size can equal one mini-slot, a slot, or multiple slots. The time-duration of mini-slots and slots depends on the chosen PHY numerology. The slot length equals 0.5 ms for 15 kHz SCS, 0.25 for 30 kHz SCS, and so forth.	Reduced latency, and increased flexibility for scheduling in coherence with the users' QoS requirements and RAN conditions.
Punctured/preemptive scheduling	Allows to quickly schedule an urgent latency critical payload with a short TTI that overwrites another ongoing downlink scheduling transmission. The concept includes efficient recovery mechanisms, where penalty for the victim UE that experiences overwriting of its transmission is minimized.	Efficient downlink scheduling of sporadic low latency traffic without reserving transmission resources in advance.
PHY numerologies	Configurable PHY numerology with base subcarrier spacing (SCS) of 15 kHz (as in LTE), which can be scaled with 2^N, where $N \in [0, 1, 2, 3, 4, 5]$ for the first 5G NR specs. A cell can be configured to have multiplexing of different PHY numerologies (requires appropriate guard intervals between those).	Scalability to larger frequency ranges and different deployments.
MIMO/beamforming	SU-MIMO: Support for at least up to eight streams. Can schedule new transmissions and HARQ retransmission on different streams.	MU-MIMO: Can schedule up to N users on the same time-frequency resources.

Table 1. Summary of the agile 5G multi-service scheduling related functionalities.

of radio resources for sporadic LLC traffic. Further radio resource management considerations for punctured scheduling are presented in [13].

Summary

In this article we presented an extensive survey of the broad family of packet scheduling related improvements that comes with the new 5G system. Those enhancements and their related benefits are summarized in Table 1. In short, a new end-to-end QoS architecture is envisioned that offers improved opportunities for application-layer scheduling functionality to ensure satisfactory QoE. The latter works in harmony with the lower-layer agile MAC scheduler. The MAC scheduler comes with a large number of options, primarily offered by the highly flexible PHY design of the 5G New Radio; including scheduling with dynamic TTI sizes, flexible timing, different PHY numerologies, new paradigms such as punctured scheduling, and so on. In conclusion, the 5G system design, and particularly the scheduler related mechanisms at the different layers, presents opportunities for improved E2E performance, capabilities for more efficiently multiplexing users with highly diverse QoS requirements, and flexibility for different network implementations. System-level performance results confirm that the new scheduling functionalities offer promising benefits.

Acknowledgments:

Part of this work has been performed within the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union. The views expressed in this contribution are those of the authors and do not necessarily represent the project.

References

[1] C. Sexton *et al.*, "5G: Adaptable Networks Enabled by Versatile Radio Access Technologies," *IEEE Commun. Surveys Tutorials*, vol. 19, no. 2, 2017, pp. 688–720.

[2] 3GPP Technical Report (TR) 38.804, "Study on New Radio Access Technology; Radio Interface Protocol Aspects (Release 14)," Mar. 2017.

[3] 3GPP Technical Report (TR) 38.802: "Study on New Radio (NR) Access Technology Physical Layer Aspects," Mar. 2017.

[4] A. Maeder *et al.*, "A Scalable and Flexible Radio Access Network Architecture for Fifth Generation Mobile Networks," *IEEE Commun. Mag.*, vol. 54, no. 11, Nov. 2016, pp. 16–23.

[5] B. Héder, P. Szilágyi, and C. Vulkán, "Dynamic and Adaptive QoE Management for OTT Application Sessions in LTE," *IEEE Proc. Int'l. Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, Sept. 2016.

[6] S. Khosravirad *et al.*, "Enhanced HARQ Design for 5G Wide Area Technology," *IEEE Proc. VTC-2016-Spring*, 5G Air Interface Workshop, May 2016.

[7] K. I. Pedersen *et al.*, "A Flexible 5G Frame Structure Design for Frequency-Division Duplex Cases," *IEEE Commun. Mag.*, vol. 54, no. 3, Mar. 2016, pp. 53–59.

[8] Q. Liao *et al.*, "Resource Scheduling for Mixed Traffic Types with Scalable TTI in Dynamic TDD Systems," *IEEE Proc. Globecom*, Dec. 2016.

[9] K. I. Pedersen *et al.*, "System Level Analysis of Dynamic User-Centric Scheduling for a Flexible 5G Design," *IEEE Proc. Globecom*, Dec. 2016.

[10] G. Pocovi *et al.*, "MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks," *IEEE Proc. ICC* (workshop), June 2017.

[11] D. Laselva *et al.*, "On the Impact of Realistic Control Channel Constraints on QoS Provisioning in UTRAN LTE," *IEEE Proc. VTC 2009 Fall*, Sept. 2009.

[12] A. K. Talukdar, "Performance Evaluation of the Enhanced Physical Downlink Control Channel in a LTE Network," *IEEE Proc. PIMRC*, Sept. 2013, pp. 987–91.

[13] K. I. Pedersen, G. Pocovi, and J. Steiner, "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband," *IEEE Proc. Vehicular Technology Conf.*, Sept. 2017.

[14] F. Boccardi *et al.*, "Why to Decouple the Uplink and Downlink in Cellular Networks and How to Do It," *IEEE Commun. Mag.*, vol. 54, no. 3, Mar. 2016, pp. 110–17.

[15] J. Padhye *et al.*, "Modeling TCP Reno Performance: A Simple Model and Its Empirical Validation," *IEEE/ACM Trans. Networking*, vol. 8, no. 2, Apr. 2000, pp. 133–45.

Biographies

KLAUS PEDERSEN (klaus.pedersen@nokia-bell-labs.com) received the M.Sc. degree in electrical engineering and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. He is currently leading the Nokia Bell Labs research team in Aalborg, and he is a part-time professor at Aalborg University in the Wireless Communications Network (WCN) section. He is the author/co-author of approximately 160 peer-reviewed publications on a wide range of topics, as well as an inventor on several patents. His current work is related to 5G New Radio, including radio resource management aspects, and the continued Long Term Evolution (LTE) and its future development, with special emphasis on mechanisms that offer improved end-to-end (E2E) performance delivery. He is currently part of the EU funded research project ONE5G that focuses on E2E-aware optimizations and advancements for the network edge of 5G New Radio. He is an IEEE Senior Member.

GUILLERMO POCOVI (guillermo.pocovi@nokia-bell-labs.com) received his M.Sc. degree in telecommunications engineering from Universitat Politècnica de Catalunya in 2014, and his Ph.D. from Aalborg University, Denmark, in 2017. He is currently an industrial post doc at Nokia Bell Labs Aalborg, partly sponsored by the Danish Innovation Fund. His research interests are mainly related to ultra-reliable and low latency communications for current wireless networks and upcoming 5G New Radio.

JENS STEINER (jens.steiner@nokia-bell-labs.com) received his M.Sc. degree in electrical engineering from Aalborg University, Denmark, in 1996, with a specialty in software engineering. Since 1996 he has been working for different companies mainly in the telecommunications sector. Since 2005 he has been working at Nokia, Aalborg, first as an external consultant, and subsequently as a permanent member of staff. At Nokia Bell Labs, he is currently involved in 5G radio access network system-level simulator research and development. He also contributes to radio research beyond software development. He is part of the EU funded research project ONE5G that focuses on E2E-aware optimizations and advancements for the network edge of 5G New Radio.

ANDREAS MAEDER (andreas.maeder@nokia-bell-labs.com) received his Ph.D. in 2008 from the University of Wuerzburg, Germany. He is currently affiliated with Nokia Bell Labs, where he is coordinating the standardization and research work on 5G RAN architecture and protocols. He has contributed since 2008 on next generation mobile networks, including air interface design, system architecture, virtualization, and cloudification of mobile network functional components. He was actively contributing as a delegate to the standardization of 3GPP RAN and system architecture, as well as to IEEE 802.16. He is the author of numerous standard contributions, conference papers, journal articles, book chapters, and more than 50 patents and patent applications.