5G agile scheduler for low latency communication

5G radio resource management for multi-service support

White paper

One of the key performance benefits of forthcoming 5G networks is ultra-low latency of as little as 1 ms. The capability will be crucial to support new uses such as collaborative virtual reality, autonomous vehicle control and synchronized robot control as part of Industry 4.0.

In this white paper, we look at a new Quality of Service (QoS) architecture that supports application-layer scheduling controlled by an agile packet scheduler. A wide variety of scheduling based on dynamic frame sizes, flexible timing, different numerologies and new paradigms, such as preemptive scheduling, supports different network implementations that can meet diverse and challenging reliability and latency targets.

Contents

Executive summary

With 5G networks just around the corner, expectations are high. 5G will support extreme data rates for mobile broadband exceeding 10 Gbps with up to 10,000 times higher capacity than today's best networks. It will also support Internet of Things (IoT) connectivity with low cost modules and long battery life, while also meeting the high reliability and low latency needs of critical communications.

The same network infrastructure will support smartphones, tablets, virtual reality connections, personal health devices, critical remote control or automotive connectivity. This white paper focuses on the radio network innovations that facilitate Ultra-Reliable Low Latency Communication (URLLC) in 5G networks.

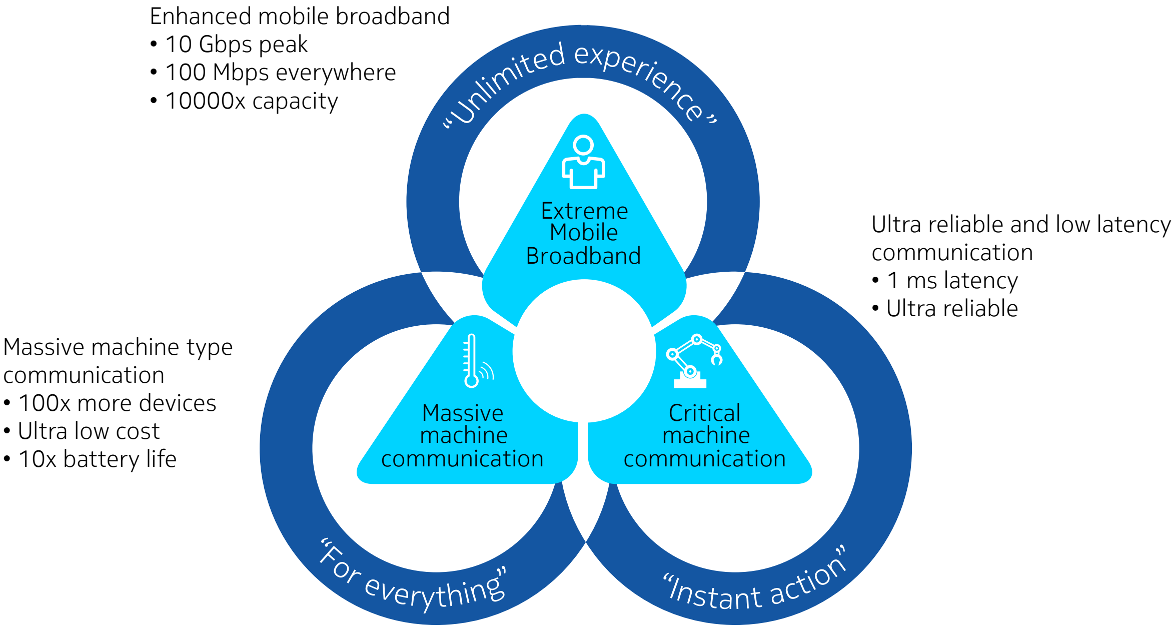
Systems for the management of 5G radio resources, in particular the packet scheduler, are important in meeting the end-to-end Quality-of-Service (QoS) performance targets for each session. The packet scheduler is also responsible for efficiently managing many sessions with highly diverse QoS requirements in one unified system.

Efficient scheduling of URLLC traffic is particularly challenging. There is 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 until it is successfully received, with an outage probability of only 0.001 percent.

This paper describes a new end-to-end QoS architecture that offers improved opportunities for application-layer scheduling functionality that works with lower-layer agile packet scheduler. The scheduler offers many options made possible by the highly flexible physical layer design of the new 5G radio. These include scheduling with dynamic transmission time intervals, flexible timing, different numerologies and new paradigms such as preemptive scheduling.

The 5G system design, and particularly the scheduler related mechanisms at the different layers, offer opportunities for improved end-to-end performance, more efficient multiplexing of users with highly diverse QoS requirements and the flexibility to suit different network implementations. System-level performance results confirm that the new scheduling functions offer promising benefits.

Figure 1. 5G performance targets



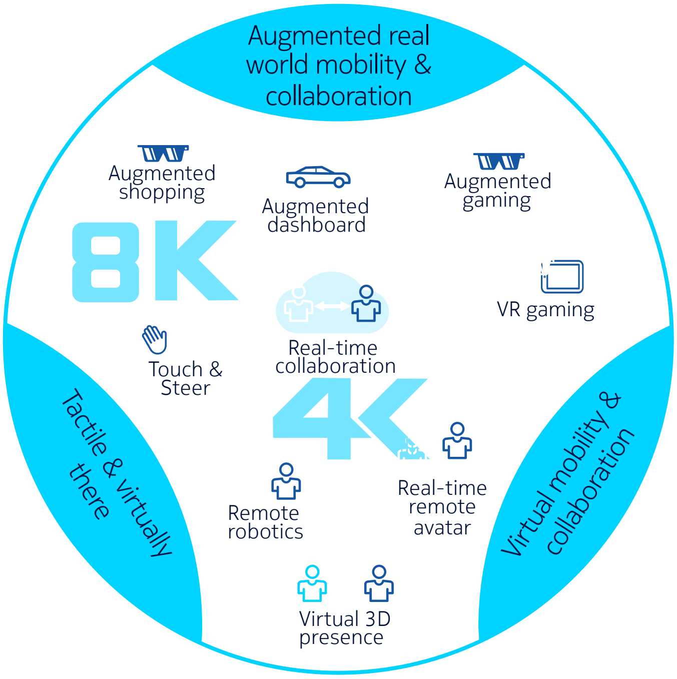
Ultra-Reliable Low Latency Communication (URLLC)

Mobile networks that offer ultra-high reliability combined with low latency open up potentially lucrative new business opportunities for the industry, arising from new applications that simply will not work properly if network delays are too high. Real-time functionality demands the lowest possible latency in the network, while reliability assures users that they can depend on their communications even in life-threatening situations. Example applications include augmented reality, virtual reality, power network smart grids, vehicular control and factory automation. Nokia has demonstrated several 5G ultra-reliability, low latency use cases, including:

- Autonomous vehicles communicating and steering
- Collaboration in a Virtual Reality environment focused on a training/ education task.
- Industry 4.0 featuring fast and synchronized collaboration of robots
- Reliable high performance low latency multicast provides new viewing experiences for stadium visitors. Multiple live video channels from cameras around a stadium can be viewed by all stadium visitors simultaneously. Visitors can switch between the real-time video channels and can recommend a channel to another device

Current commercial mobile networks cannot support highly reliable communication with tight latency constraints, such as 5G at 99.999 percent reliability within a delay of 1 ms for small packet sizes. New solutions are required to support this. This paper illustrates how 5G networks will be able to meet these challenging reliability and latency targets.

Figure 2. Example use cases for ultra-reliable and low latency communication



Quality of Service Architecture

Compared to LTE, 5G design includes a new QoS service architecture, as well as several enhancements to the protocol stack. These include dynamic QoS mapping, reflective uplink QoS and flexible Layer 2 and Layer 1 solutions.

The Service Data Application Protocol (SDAP) filters the data packets in the mobile and the 5G core network to associate the data packets with QoS flows. The access stratum (AS) mapping in the device and the 5G radio network associates the QoS flows with the data radio bearers (DRBs). This mapping is based on 5G QoS class indices in the transport header of the packets, as well as on corresponding QoS parameters. End-to-end packet sessions may be mapped to two different QoS flows and DRBs for cases where the end-to-end packet session contains data flows with two different sets of QoS requirements. An example is a website with embedded high-definition live streaming video.

Mapping of end-to-end session to QoS flows and DRBs can be updated as required. This kind of flexibility allows the application of 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 referred to as higher-layer application-aware scheduling.

On the device side, the concept of reflective QoS in 5G eliminates the use of dedicated flow filters signaled by the network to match traffic to QoS flows. In reflective QoS, the device derives the mapping of uplink traffic to QoS flows by correlating the corresponding downlink traffic and its attributes, for example in Transport Control Protocol (TCP) flows. On the 5G radio interface, the packet treatment is defined separately for each DRB. The packet scheduler aims to meet the requirements for the users' DRBs, as well as prioritizing accordingly if the system becomes congested and not all users' requirements can be met simultaneously.

Figure 3. A new QoS architecture

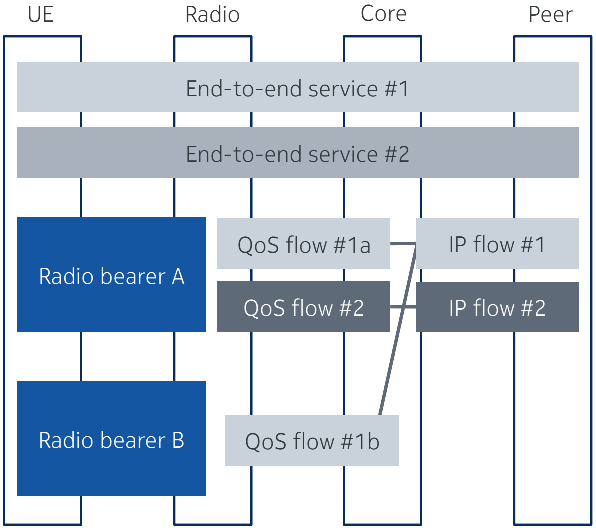


Figure 4. User plane protocol

SDAP sub-layer	Mapping of Quality of Service (QoS) flow to radio bearer
Packet Data Convergence Protocol (PDCP)	Numbering, compression and ciphering
Radio Link Control (RLC)	Segmentation and retransmissions
Media Access Control (MAC)	Scheduling, priority handling and hybrid Automatic Repeat Request (ARQ)
Physical layer	New flexible physical layer

Flexible frame structure supporting slicing

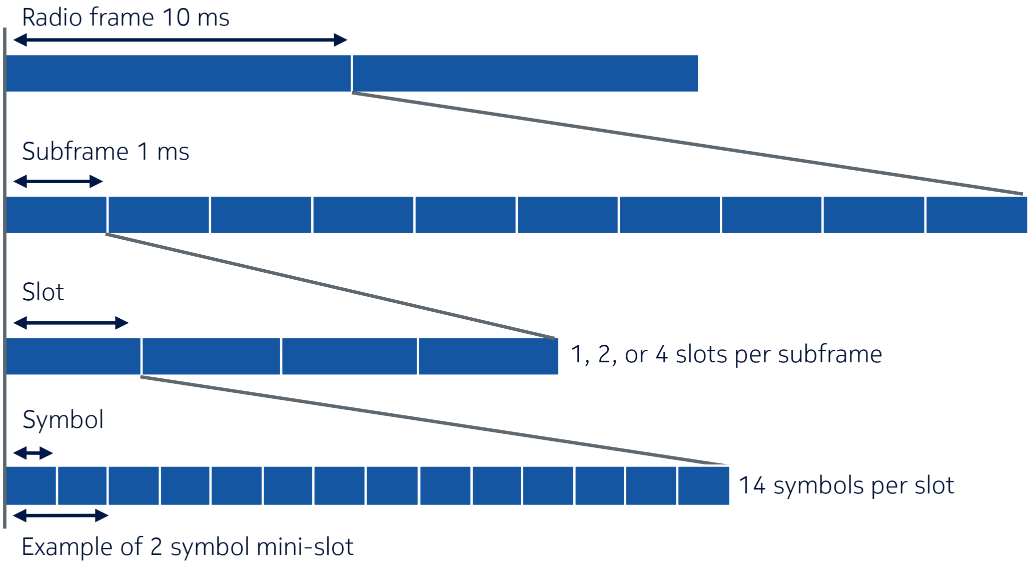
5G radio can be deployed in a variety of ways and make use of different spectrum allocations. The subcarrier spacing can be selected according to the bandwidth. Narrowband deployments use narrowband subcarrier spacing. 5G at below 1 GHz band typically uses 15 kHz subcarrier spacing, while 5G at 3.5 GHz is configured with 30 kHz subcarrier spacing and mmWaves with 120 kHz. The scheduling interval gets shorter as subcarrier spacing increases.

The 5G frame structure can support several service requirements simultaneously. The radio frame is 10 ms in length, with the sub-frame being 1 ms long. The number of slots per sub-frame depends on the total subcarrier spacing and can be 1, 2 or 4. The slot length equals 14 symbols. The typical scheduling interval is one slot. It is also possible to schedule data using a mini-slot with a length of typically 2, 4 or 7 symbols.

Table 1. 5G numerology is designed for flexible deployment options

Subcarrier spacing (kHz)	15	30	60	120
Spectrum	<6 GHz	<6 GHz	<6...>20	>20 GHz
Max bandwidth (MHz)	50	100	200	400
Symbol duration (us)	66.7	33.3	16.7	8.33
Nominal cyclic prefix (us)	4.7	2.3	1.2	0.59
Slot length (ms)	1.0	0.5	0.25	0.125

Figure 5. 5G frame structure supporting low latency communication



Agile packet scheduler

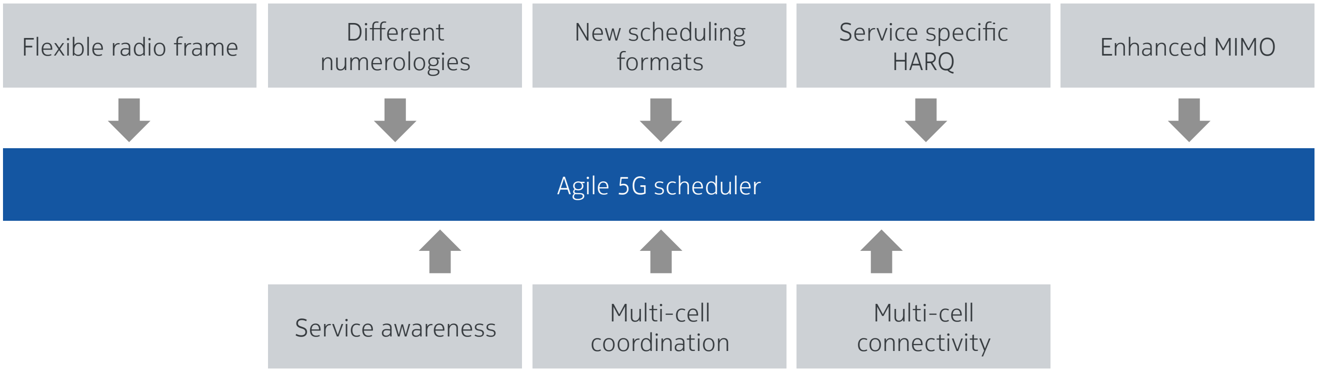
The packet scheduler is responsible for enforcing QoS differentiation in the radio interface.

The MAC scheduler is the controlling entity for multi-user radio resource allocations, which is subject to several constraints but also many options for serving the different devices efficiently. The greater number of options for the 5G MAC scheduler offers better performance, but also presents the problem of how to make best use of it. The MAC scheduler works by allocating radio transmission resources for downlink and uplink transmissions separately for each user.

The scheduler aims at fulfilling the QoS service targets for all the DRBs of the served users. The scheduler must support multi-cell connectivity mode, where devices are configured to be simultaneously served by multiple cells. There may also be other multi-cell coordination constraints, including inter-cell interference coordination between neighboring cells where certain radio resources are muted as required and hence not available for scheduling of users.

At the MAC sub-layer, enhanced service-specific Hybrid Automatic Repeat Request (HARQ) enhancements are included. The 5G physical layer offers many new options for the MAC scheduler to support large numbers of users with differing service needs.

Figure 6. Agile 5G packet scheduler



The smallest time-domain scheduling resolution for the MAC scheduler is mini-slot, but it is also possible to schedule users on slot resolution, or on resolution of multiple slots. Thus, dynamic scheduling with different transmission time interval sizes is supported. This enables the MAC scheduler to match the radio resource allocations more efficiently for different users in coherence with the radio condition, QoS requirements, and cell load conditions. The short transmission time is needed for URLLC use cases, but is not restricted to such traffic. In the frequency domain, the minimum scheduling resolution is one physical resource block of 12 subcarriers, corresponding to 180 kHz for 15 kHz subcarrier spacing and 360 kHz for 30 kHz, and so forth.

Figure 7 shows how different users are multiplexed in the downlink on an FDD carrier. Different colors represent transmissions to different users. Most users are multiplexed on slot resolution. The MAC scheduler can freely decide how to schedule its different users on the carriers, and the RLC layer is not aware of how this is done. However, it is possible to enforce some restrictions via higher-layer control signaling to schedule data from certain DRBs only on a given physical numerology and a particular transmission size. Each scheduling allocation is sent to the UE via a downlink control channel carrying the scheduling grant. The downlink control channel is flexibly time-frequency multiplexed with the other downlink physical channels and can be mapped contiguously or non-contiguously in the frequency domain.

This is a highly flexible design, where the relative downlink control channel overhead can take values from below 1% if, for example, scheduling few users with long transmission times, up to tens of percentages if scheduling a larger number of users with very short transmission times. The design therefore overcomes control channel blocking problems.

These advantages are achieved by migrating towards a user-centric design, as compared to the predominantly cell-centric LTE design. Another advantage brought by the more flexible 5G downlink control channel design is the ability to support devices that only operate on a fraction of the carrier bandwidth, such as narrowband IoT devices.

Figure 7. Preemptive scheduling in downlink multi-user mux

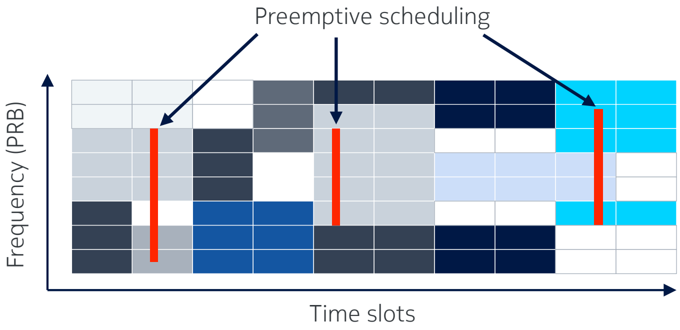


Figure 8. Preemptive scheduling with short TTI

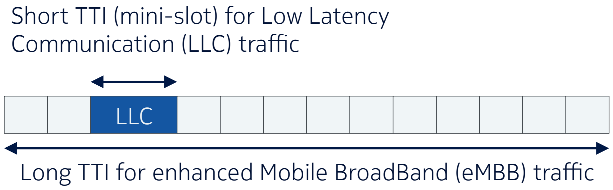

Short TTI (mini-slot) for Low Latency Communication (LLC) traffic
LLC
Long TTI for enhanced Mobile BroadBand (eMBB) traffic

Figure 8 illustrates the principle of preemptive scheduling for efficient handling of Low Latency Communication (LLC) traffic. Efficient scheduling of LLC is challenging as such traffic is typically bursty and random and requires immediate scheduling with a short transmission time to fulfil the corresponding latency budget. Instead of pre-reserving radio resources for LLC traffic bursts, it is proposed to use preemptive scheduling.

The basic principle is as follows: mobile broadband traffic is scheduled on all the available radio resource. Once a LLC packet arrives at the base station, the MAC scheduler immediately transmits it to the designated device by overwriting part of a current scheduled transmission, using mini-slot transmission. This has the advantage that the LLC payload is transmitted immediately without waiting for current scheduled transmissions to be completed and without the need for pre-reserving radio resources for LLC traffic.

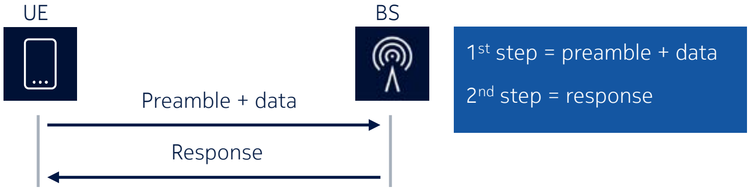
The price of preemptive scheduling is borne by the user whose transmission is partly overwritten. Related recovery mechanisms are introduced to minimize the impact on the user. These include an indication to the victim device that part of its transmission has been preempted. This enables the device to take this effect into account when decoding the transmission, i.e. it knows that part of the transmission is corrupted. Moreover, options for smart HARQ retransmission are considered, where the damaged part of the preempted transmission is first retransmitted.

The discussion above focused on the downlink low latency transmission. 5G also provides tools for low latency transmission in the uplink by using pre-scheduled transmission and contention based access. Data transmission in the mobile radio is typically controlled by the base station packet scheduler to maximize efficiency, but the scheduling adds delay in the uplink transmission.

It is possible to get around this delay by pre-allocating resources for the devices in the uplink, an approach called pre-scheduled transmission. Contention based access refers to the uplink transmission where the device autonomously sends data without any specific allocation or grant from the network.

This approach minimizes signaling, which offers benefits to the latency and device power consumption. A preamble might be used as reference signal, or transmitted in a separate resource. The device can be identified based on time and frequency resources and the reference signal parameter.

Figure 9. Contention based access for low latency in 5G uplink



Reliability and latency results

The flexible 5G design, together with agile scheduler, allows tough performance requirements to be met. Results from extensive system-level simulations are presented in Figures 10 and 11 to illustrate the benefits of some of the 5G scheduling enhancements. We first show mobile broadband performance results for file download over TCP. Figures 10 and 11 show the performance for short (mini-slot) and long (1 ms) transmission times, considering both low offered traffic and high offered traffic.

One of the reported performance metrics is the smoothed round trip time of TCP packets. It is observed that the best performance is achieved for the short transmission time 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 control channel overhead from operating with a short transmission time is not a problem at the low offered load.

However, with the high offered load, the best performance is observed with the long transmission time. This is because using longer transmission time results in higher average spectral efficiency with lower control channel overhead. The results clearly show the benefit of being able to adjust the transmission time as required.

Figure 10. Smoothed median TCP packet round trip time for eMBB file download

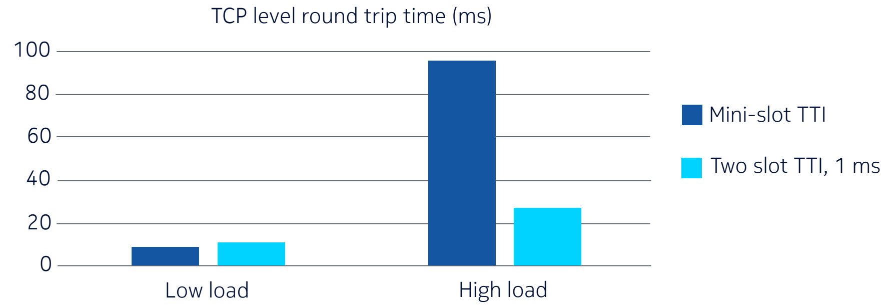
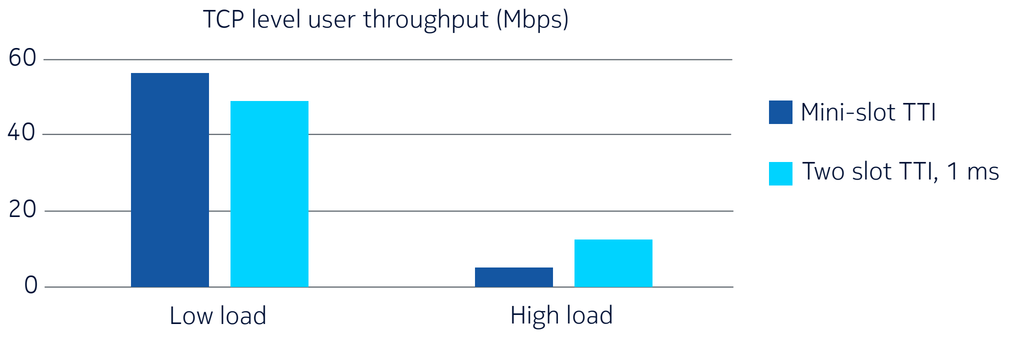


Figure 11. Median TCP level user throughput for eMBB file download



We also need to consider downlink performance for traffic that is a mixture of mobile broadband and low latency communication (LLC). In this example, there are on average five active mobile broadband users per macrocell, downloading 500 kB files using TCP. As soon as one of the mobile broadband users finishes their 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 low latency critical payloads of 50 bytes are sporadically generated according to a Poisson process.

As this scenario corresponds to a fully loaded network, the mobile broadband users are scheduled with a transmission time of 1 ms, using all available resources. Hence, no radio resources are reserved for potentially incoming LLC traffic. Instead, preemptive scheduling is applied whenever LLC payloads arrive. These latency critical payloads are immediately scheduled on arrival with mini-slot resolution, overwriting part of the ongoing mobile broadband transmissions.

Due to the urgency of the LLC traffic, we assume RLC transparent mode and a low initial Block Error Rate of only 1 percent for such transmissions to avoid too many HARQ retransmissions. The average cell throughput is illustrated in Figure 12, where the performance is shown for cases with and 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 mobile broadband HARQ transmissions and a case where only the damaged part of the mobile broadband transmission that has been subject to preemption is retransmitted (labelled as partial retransmission in Figure 12).

The latter option is the most promising solution, as fewer radio resources are used for HARQ retransmissions of mobile broadband transmissions that have suffered from puncturing . However, the cost of using this approach is a slightly longer latency for the mobile broadband users, as the probability of triggering a second HARQ retransmission is higher, compared to when the first HARQ retransmission includes the full transport block.

Figure 13 shows the complementary cumulative distribution function of latency of LLC traffic. The latency is measured from when the LLC payload arrives at the base station until it is correctly received by the devices. This includes the aggregated latency from scheduling, frame alignment, transmission time and tx/rx processing delays. Even under the considered full load conditions, the performance for the LLC traffic fulfills the challenging URLLC target of 1 ms latency with an outage of only 0.001 percent (=10^{-5}). Hence, the preemptive scheduling scheme is able to schedule the LLC traffic efficiently in line with its challenging latency and reliability constraints. At the same time, it offers efficient scheduling of mobile broadband traffic without the need for pre-reservation of radio resources for sporadic LLC traffic.

Figure 12. Average cell capacity for cases with LLC/eMBB and preemptive scheduling

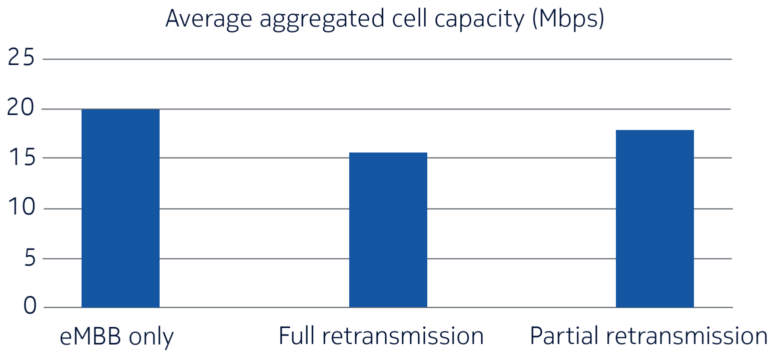
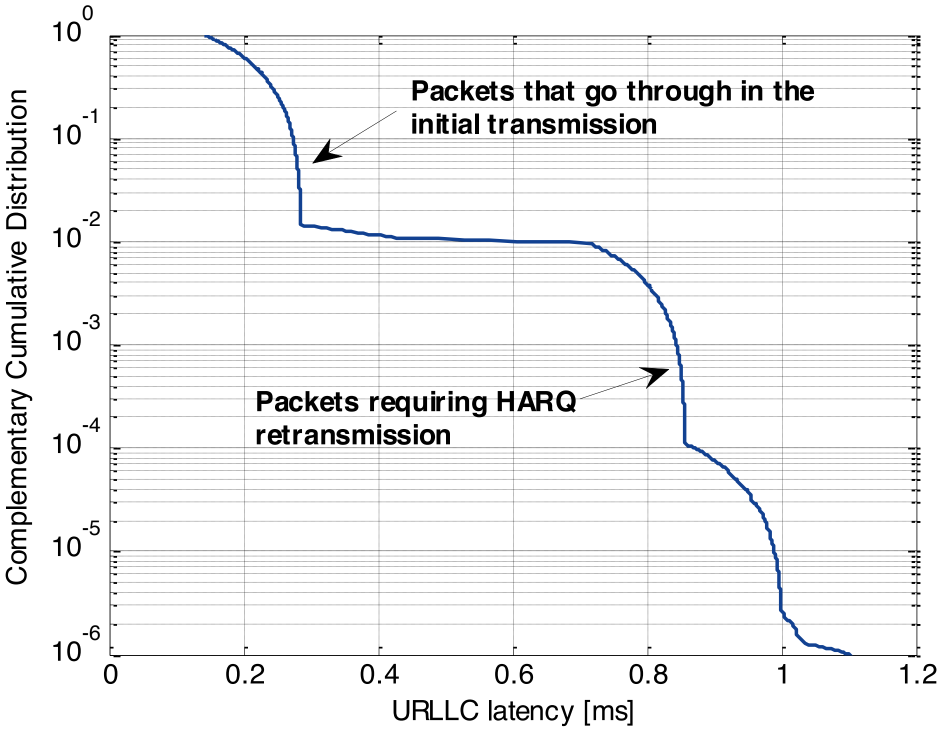


Figure 13. Complementary cumulative distribution function of the LLC packet latency



Further reading

"5G for Mission Critical Communication", Nokia white paper http://info.networks.nokia.com/5GforMissionCriticalCommunication_01.LP.html

"Dynamic end-to-end network slicing for 5G", Nokia white paper https://resources.ext.nokia.com/asset/200339

Abbreviations

ARQ	Automatic Repeat Request
AS	Access Stratum
DRB	Data Radio Bearer
eMBB	Enhanced Mobile Broadband
HARQ	Hybrid ARQ
IoT	Internet of Things
LLC	Low Latency Communication
LTE	Long Term Evolution
MAC	Media Access Control
MIMO	Multiple Input Multiple Output
NAS	Non-Access stratum
OFDM	Orthogonal Frequency Division Multiplexing
PDCP	Packet Data Convergence Protocol
QoS	Quality of Service
RLC	Radio Link Control
TCP	Transport Control Protocol
URLLC	Ultra Reliable Low Latency Communication

About Nokia

We create the technology to connect the world. Powered by the research and innovation of Nokia Bell Labs, we serve communications service providers, governments, large enterprises and consumers, with the industry's most complete, end-to-end portfolio of products, services and licensing.

From the enabling infrastructure for 5G and the Internet of Things, to emerging applications in digital health, we are shaping the future of technology to transform the human experience. networks.nokia.com

Nokia Oyj
Karaportti 3
FI-02610 Espoo, Finland
Tel. +358 (0) 10 44 88 000

Document code: SR1712020226EN (March)