Architecture," in IEEE Communications Magazine, vol. 55, no. 2, pp. 70-78, February 2017.
doi: 10.1109/MCOM.2017.1600435CM

Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture

Philipp Schulz, Maximilian Matthé, Henrik Klessig, Meryem Simsek, and Gerhard Fettweis, Technische Universität Dresden, Germany
Junaid Ansari and Shehzad Ali Ashraf, Ericsson Research, Aachen, Germany
Bjoern Almeroth, RadioOpt GmbH, Dresden, Germany
Jens Voigt and Ines Riedel, Amdocs, Dresden, Germany
Andre Puschmann and Andreas Mitschele-Thiel, Technische Universität Ilmenau, Germany
Michael Müller, IVM gGmbH, Oberlungwitz, Germany
Thomas Elste, IMMS GmbH, Ilmenau, Germany
Marcus Windisch, Freedelity GmbH, Dresden, Germany

1 Abstract

Next generation mobile networks not only envision on enhancing the traditional mobile broadband (MBB) use case but also aim at meeting the requirements of new use cases, such as the Internet of Things (IoT). This article focuses on latency critical IoT applications and analyzes their requirements. We discuss the design challenges and propose solutions for the radio interface and network architecture to fulfill these requirements which mainly benefit from flexibility and service-centric approaches. The article also discusses new business opportunities through IoT connectivity enabled by future networks.

2 Introduction

Besides the traditional MBB, the development of 5G networks is driven by IoT connectivity. Therefore, in addition to the classical MBB traffic demands of high throughput and capacity, new requirements of achieving low latency and high reliability for many IoT use cases are very important. In the context of new 5G use cases, IoT applications have been categorized into two classes, namely massive machine-type communications (mMTC) and ultra-reliable low-latency communications (URLLC). The former consists of large number low cost devices with high requirements on scalability and increased battery lifetime. In contrast, URLLC requirements relate to the mission critical applications, where uninterrupted and robust exchange of data is of the foremost importance.

In this article we focus on the latency critical IoT use cases, which are being investigated in the collaborative research project *fast wireless*[1]. We have comprehensively analyzed such use cases and distilled their requirements. Our measurement results of the 4G network motivate the need of new design concepts on radio interface and network architecture in order to meet the demands of the latency critical IoT applications. In the context of the radio interface design, we discuss the latency enhancements on both the medium access control (MAC) and physical (PHY) layers. In addition, we present concepts on service-centric architecture of 5G networks. Virtualization in 5G network leads to flexible design that enables it to shift the computing power to the edge of the network and hence, reduce the latency. It also facilitates analyzing and managing the network in a service-centric fashion. This virtualization approach in the 5G network architecture allows seamless transitions between technologies or operators. Hence, service-centric management and operation disclose novel business

[1] http://de.fast-zwanzig20.de/basisvorhaben/fast-wireless/

models, from which not only network operators and service providers but also IoT customers can profit.

3 Latency critical IoT Use Cases and Requirements

We consider five important use cases of latency critical IoT applications and characterize them based on several different requirements as summarized in Table 1.

3.A Factory Automation

Factory automation applications are typically characterized by real-time control of machines and systems in fast production and manufacturing lines, where machine parts are in motion within a limited space (e.g., a factory hall). Examples of such applications include high speed assembly, packaging, palletizing, etc. Factory automation applications are generally considered to be highly challenging in terms of latency and reliability demands, which also vary among different applications as given in Table 1. The reliability requirements for factory automation applications are typically 10^{-9} packet loss rate (PLR) while the latency requirements vary from 250 µs to 10 ms.

3.B Process Automation

Process automation includes applications for monitoring and diagnostics of industrial elements and processes like heating, cooling, mixing, stirring, and pumping procedures, etc. The measured values for these applications change relatively slowly. Therefore, the latency requirements for such services range from 50 ms to 100 ms with affordable PLR of up to 10^{-3}. The coverage area is often quite large (e.g., a power plant) and typically comprises of multiple buildings and outdoor sites.

3.C Smart Grids

Smart grid applications have relatively less stringent requirements on latency and reliability compared to factory automation applications, i.e., latency and PLR requirements of up to 20 ms and 10^{-6}, respectively. However, the communication range needs to be much longer, i.e., up to a few kilometers.

3.D Intelligent Transport Systems

Autonomous driving and the optimization of road traffic create new challenges on communications. Requirements result from different intelligent transport systems (ITS) use cases such as autonomous driving, road safety, and traffic efficiency services [1].

Road safety includes warning other road devices about collisions or dangerous situations. Autonomous driving additionally requires coordination of actions, for instance, to perform overtaking or platooning. Therefore, communication systems have to operate with communication ranges of up to 500 m and latency of less than 50 ms while ensuring a high reliability. However, periodic traffic consisting of small packet sizes generated at the rate of 10 Hz leads to data rates of only 2 kbps per device.

Traffic efficiency services aim to control traffic flows. In an urban environment, these include information on the status of traffic lights and local traffic situation to accordingly allow adapting vehicle velocities at intersections. These services require a wireless infrastructure with communication ranges of up to 2 km and high reliability, but relaxed end-to-end (E2E) latency of less than 100 ms.

3.E Professional Audio

The majority of today's professional audio links is built based on conventional analog transmission techniques in dedicated licensed frequency bands in the VHF and UHF range. Compared to digital transmission, analog transmission is spectrally inefficient and requires an extensive frequency planning. Hence, it is important to treat professional audio as a part of the future 5G IoT ecosystem,

as well. Professional audio applications, such as live concerts, also demand for extremely low latency in the transmission links. It has been observed that trained musicians find latency exceeding 4 ms between sound generation (singers voice or instrument) and tonal perception (by means of monitor speakers or in-ear-monitoring) as disturbing and thus, unacceptable. In a typical stage setup, the total round-trip latency budget of 4 ms is divided into three parts: the wireless link to the central mixing desk, the tonal processing in the mixer (typically 2 ms), and the wireless link back to the musician. Each of the two wireless links must therefore add not more than 1 ms latency while providing sufficient transmission reliability.

Table 1. Communication requirements of latency critical IoT applications [1 - 3]. Please note that Update Time *only applies to the periodic traffic. The application use cases may also include sporadic or event based traffic but the traffic arrival distributions are not mentioned in the table.*

	Use Case	Latency [ms]	Reliability [PLR]	Update time [ms]	Data size [bytes]	Device density [devices/m² or devices/plant or devices/km²]	Communication range [m]	Mobility [km/h]
A	**Factory automation**	0.25 to 10	10^{-9}	0.5 to 50	10 to 300	0.33 to 3 devices/m²	50 to 100	< 30
A1	Manufacturing cell	5	10^{-9}	50	< 16	0.33 to 3 devices/m²	50 to 100	< 30
A2	Machine tools	0.25	10^{-9}	0.5	50	0.33 to 3 devices/m²	50 to 100	< 30
A3	Printing machines	1	10^{-9}	2	30	0.33 to 3 devices/m²	50 to 100	< 30
A4	Packaging machines	2.5	10^{-9}	5	15	0.33 to 3 devices/m²	50 to 100	< 30
B	**Process automation**	50 to 100	10^{-4} to 10^{-3}	100 to 5000	40 to 100	10000 devices/plant	100 to 500	< 5
C	**Smart grids**	3 to 20	10^{-6}	10 to 100	80 to 1000	10 to 2000 devices/km²	A few m to km	0
D	**ITS**							
D1	Road safety urban	10 to 100	10^{-3} to 10^{-5}	100	< 500	3000 /km²	500	< 100
D2	Road safety highway	10 to 100	10^{-3} to 10^{-5}	100	< 500	500 /km²	2000	< 500
D3	Urban intersection	< 100	10^{-5}	1000	1M / car	3000/km²	200	< 50
D4	Traffic efficiency	< 100	10^{-3}	1000	1k	3000/km²	2000	< 500
E	**Professional audio**	2	10^{-6}	0.01 to 0.5	3 to 1000	up to 1/m²	100	< 5

4 Latency Measurements for Current 4G Networks

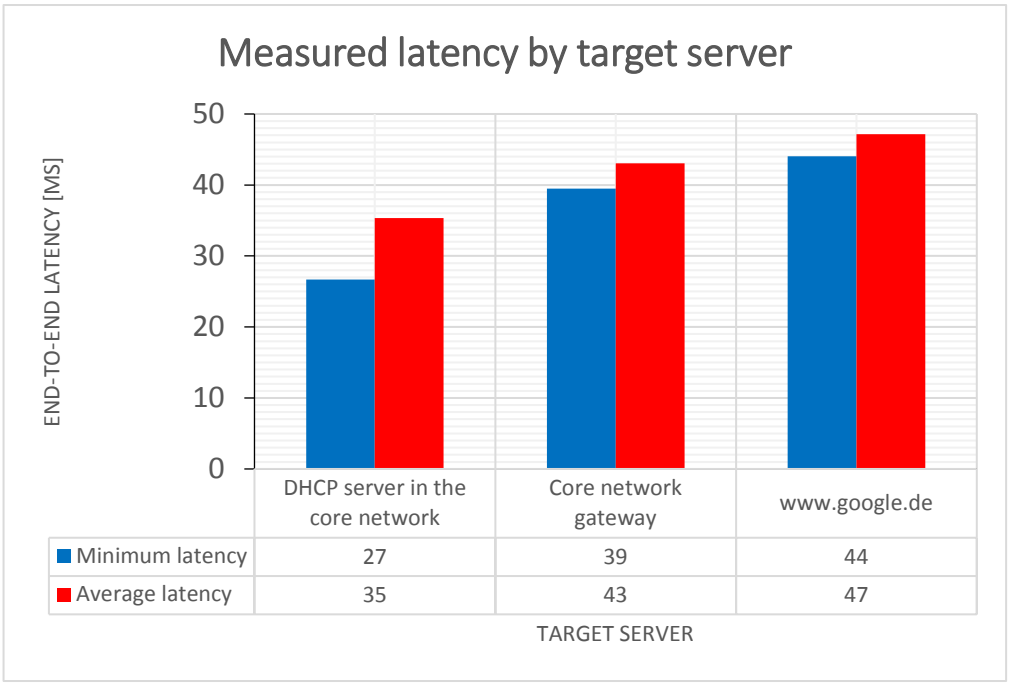
Dedicated E2E latency measurements in public cellular networks, such as LTE, have disclosed two key limitations, namely the distance to the target device and the number of active devices per cell as shown in Figure 1. Therefore, both limitations should be considered in future 5G networks to enable the low-latency use cases discussed above.

First, we analyze the impact of the physical or virtual distance on the minimum achievable latency in an LTE network. As shown in Figure 1(a), E2E latency increases with a larger distance between the two endpoints. Moreover, a considerable portion of the overall latency budget is spent in the core network of the operator. For example, a minimum of 39 ms is necessary to contact the gateway of the core network towards the Internet and only additional 5 ms is needed to receive the reply from the

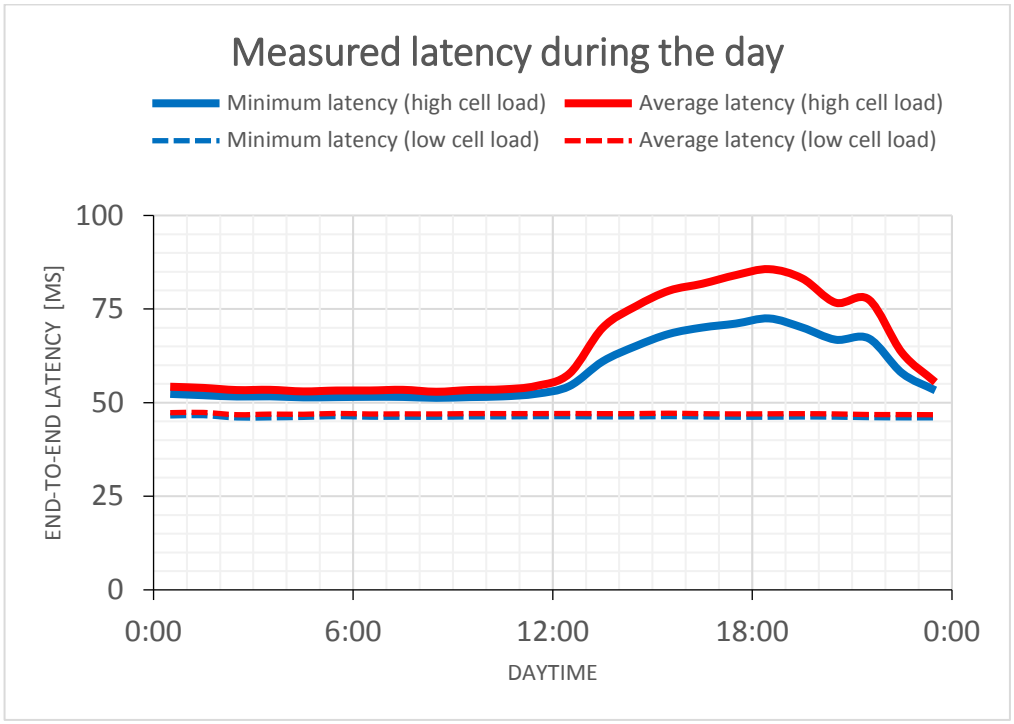
Google server. This means that about 90 percent of the overall E2E latency originates from the cellular network. A second observation, made from Figure 1(b), is that the number of active devices per cell affects the achievable latency in cellular networks. In the measurement setup, two LTE cells are compared based on their minimum and mean latency during daytime. The highly frequented LTE cell (high cell load) shows increased latencies, i.e., the mean latency increases from 50 ms to 85 ms during the afternoon. This observation correlates with the increase in the number of active devices in the measured cell (local market place) during this time. For comparison, the reference cell with a low cell load is located in a residential area and shows an almost constant latency during the entire day.

These observations motivate the need for a carefully designed network architecture for latency critical IoT applications and worthwhile to study the impact of placing the application close to the edge of the cellular network (cf. Section 6.B). Furthermore, the fundamental impact of a high number of active devices per cell on the E2E latency needs to be considered carefully (cf. Section 6.A).

The presented E2E latency measurements have been performed using conventional user equipment, i.e. an Android smartphone, connected to the public LTE network. The latencies are captured using the standardized ICMP procedure (Layer 3 ping). Performing such dedicated ping tests is the first choice when measuring the latency of the communication link in the current systems. Nowadays, this latency measurement technique is widely used and gives valuable insights into the network performance. However, this method only gives snapshots of the actual link latency and may not represent the 'true' latency of the communication link for a dedicated application during communication. In this regard, new solutions which enable monitoring of latency critical IoT applications need to be established. In addition, upcoming low-latency systems demand for new methods to measure the latency at various levels inside the considered system and not only rely on the IP layer latency, for instance, measuring the scheduling latency of the operating system. In order to do this, timing information for events, function calls, interrupts, etc. need to be observed and assessed. Hence, the 5G network architecture enabling latency critical IoT applications has to provide interfaces to monitor the related key performance indicators (KPIs). This will allow the end-user and the operator to analyze the quality of the provided service.



(a)



(b)

Figure 1. (a) Measured minimum and average end-to-end latency (ICMP ping) for different target servers with an increasing distance from left to right. (b) Measured minimum and average end-to-end latency (ICMP ping) in a cell with low load (residential area) and a cell with high load (crowded market place with many active users). Both measurements have been made in a dense urban environment (Dresden, Germany, city center) for a low-mobility scenario with a proprietary application running on an Android smartphone.

5 Concepts of Radio Interface

Mission critical IoT applications require low transmission latency and high reliability as described in Section 3. In 3GPP, transmission time interval (TTI) is defined as the time required for the

transmission of the smallest decodable data. Considering the default TTI size of 1 ms, i.e., 14 OFDM (orthogonal frequency-division multiplexing) symbols, LTE Release 8-13 does not efficiently utilize the available radio resources for small data sizes as in mission critical IoT applications. This is primarily because the granularity of resources that can be allocated to a single device in LTE is too coarse, resulting in parts of the allocated resource being wasted. TTI duration also impacts the achievable user plane latency. Therefore, changes are required in the current radio interface design to provide low user plane latency. In this section, we describe relevant enhancements for MAC and PHY layers to achieve low latency communication for IoT applications that are not fulfilled by any of the existing wireless technology standards.

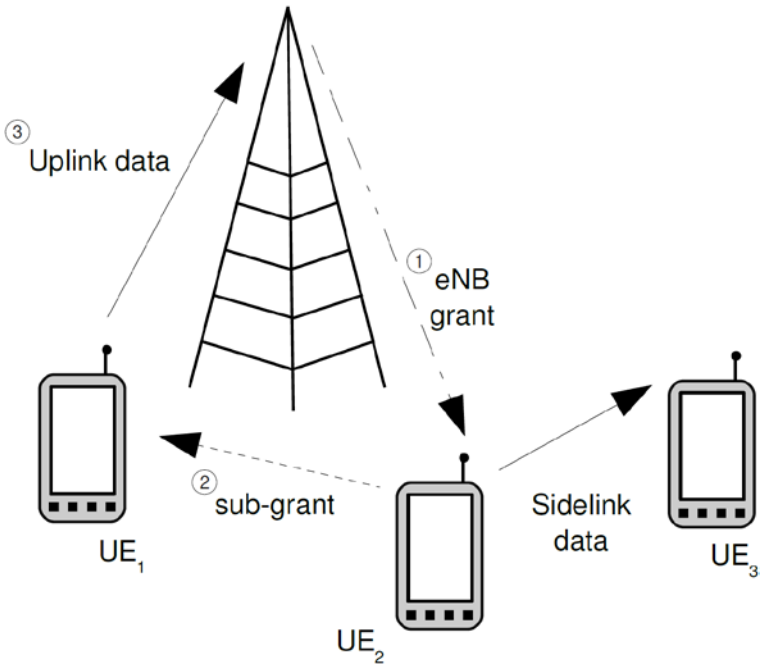
5.A Resource Delegation Scheme

Future releases of LTE will support network assisted device-to-device (D2D) communication without directly involving the base station (BS) in data exchange between devices. The D2D communication paradigm not only allows reducing the communication latency between devices but also provides a possible solution to increase the resource utilization in case of IoT applications. To achieve the latter, we propose a solution to delegate resources that are not needed by a device to another device that still has data to transmit. Partially or fully unused scheduling grants that were originally assigned in a dynamic or semi-persistent manner could be granted to other devices in their vicinity by leveraging from D2D communication and thus, increasing the overall cell throughput. However, special care needs to be taken during D2D discovery to avoid additional access delays.

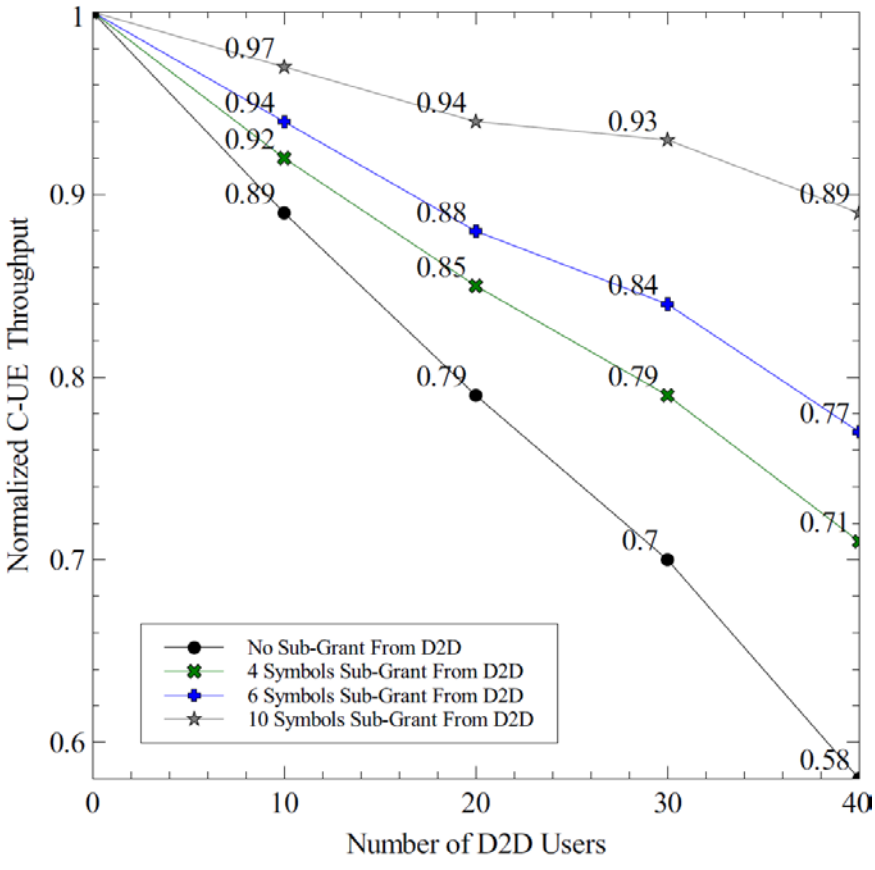
In order to allow the above mentioned secondary reuse of resources, the current LTE radio resource management (RRM) schemes need to be modified [4]. This can be achieved by splitting the RRM into a two-layer hierarchy. Thereby, the first level is managed entirely through the BS, similar to current LTE networks, whereas the second level is managed by the devices themselves. Figure 2(a) illustrates the approach. In particular, in the proposed RRM scheme devices are allowed to further delegate unused resources to other nearby devices, which we refer to as 'sub-granting'. We propose that a device which has unused resources available, i.e., the sub-grant provider, uses a small portion of its original grant to indicate to another device, i.e., the sub-grant beneficiary, to use the remaining portion of the grant. In order to convey the required sub-grant information, we propose to use an in-band signaling mechanism, e.g., to use one OFDM symbol of a sub-frame for signaling. To minimize the overall latency of the scheme we propose to carry out the selection of sub-granting candidate pairs prior to the actual communication. For example, the BS could convey relevant information along with ordinary scheduling grants to potential sub-grant providers.

Our simulation results on this scheme show an uplink performance enhancement of 3-31 percent, depending on the number of users and sub-grant size (cf. Figure 2(b)).

This document is a preprint of: P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S.A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste and M. Windisch, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," in IEEE Communications Magazine (ComMag), vol. 55, no. 2, Feb 2017. DOI:10.1109/MCOM.2017.1600435CM



(a)



(b)

Figure 2. (a) UE_2 delegates its unused uplink resources to UE_1. (b) Simulations show that in the investigated scenario with, e.g., 20 devices delegating 4 OFDM symbols per sub-frame to some other devices, uplink throughput can be increased by 6%. More general, the gain increases with the number of devices and the size of the sub-grant in this setting. Details may be found in [4].

5.B Fast Uplink Access

In LTE, a BS centrally coordinates channel access and RRM. The BS is able to efficiently carry out downlink transmissions as it itself manages the medium access. However, uplink transmissions using the default dynamic scheduling scheme impart extra signaling overhead, which leads to undesirable communication delays. As illustrated in Figure 3, according to the LTE dynamic scheduling scheme, a device needs to send a scheduling request (SR) to the BS when data needs to be transmitted in uplink. The BS correspondingly allocates radio resources for the requested traffic and notifies them using a scheduling grant (SG). After receiving SG, the device is able to send its data in the assigned resources. With the default TTI size of 1 ms and the default SR periodicity of 10 ms, the average latency becomes 12.5 ms.

In LTE Rel. 13, the concept of fast uplink access has been proposed [5] which we advocate for the investigated latency critical applications (cf. Table 1). In fast uplink access, the explicit signaling overhead of SR and SG is eliminated. Fast uplink access is based on semi-persistent scheduling [6], where resources are assigned to devices on a prior basis. Data arriving at a device can directly be transmitted on the pre-allocated resources. When there is no data, devices do not need to send out the padding information. Using the default LTE TTI of 1 ms, fast uplink access can reduce the average communication latency to 4.5 ms, which is a significant improvement compared to the LTE dynamic scheduling. One slight drawback is a lower capacity due to pre-allocation of resources. The PHY layer design features of short TTI (cf. Section 5.C) can be complementary applied for further minimizing the overall communication delay.

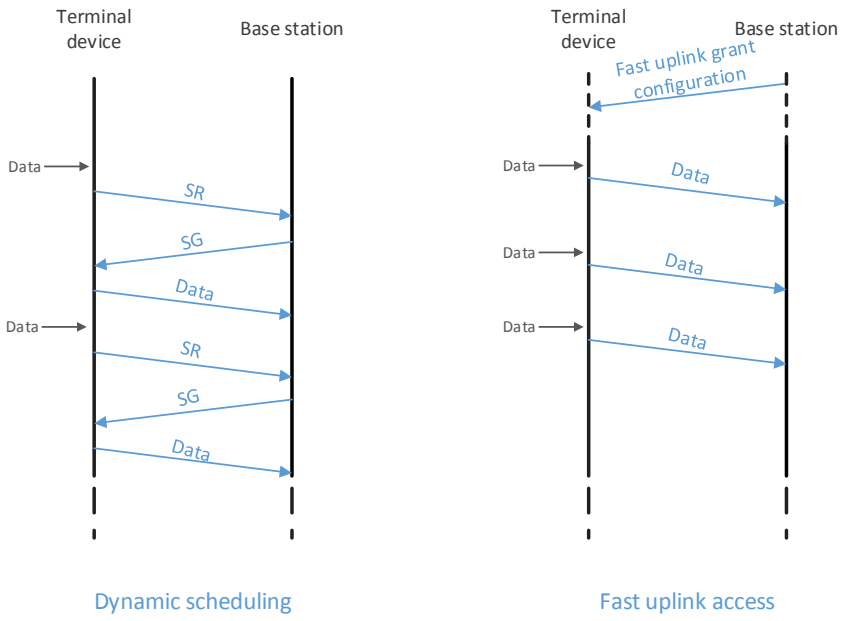


Figure 3 Message sequence chart for the dynamic scheduling and fast uplink access schemes. Dynamic scheduling imparts extra signaling delays due to the exchange of scheduling request (SR) and scheduling grant (SG) messages after data has arrived at the device. In fast uplink access, a base station configures the uplink resources in advance and after the data arrives, it can be directly transmitted without explicit SR/SG exchange.

5.C TTI Shortening

In Section 5.A, a resource delegation scheme was presented to increase the resource utilization for IoT use cases. Alternatively, TTI shortening, which not only allows low transmission latency but also increases the resource utilization, is being investigated in 3GPP. Short TTI durations of 0.5 ms (7 OFDM symbols) and 72 µs (1 OFDM symbol) are being considered for LTE Rel. 13-14 [7]. A shorter TTI duration also implies faster processing time needed for demodulation and decoding of data. While reducing the TTI duration to 1 OFDM symbol imparts substantial signaling and control overhead, we believe that a

short TTI of 2 OFDM symbols is highly relevant for several latency critical IoT applications as shown in Table 1 and can fulfill the latency requirements of most use cases. For instance, the average latency of using a TTI of 14 OFDM symbols along with the dynamic scheduling scheme for uplink and downlink transmission of 12.5 ms and 7.5 ms (cf. Section 5.B), can be reduced to 1.8 ms and 1 ms, respectively, by restricting the TTI to only 2 OFDM symbols.

However, TTI shortening concept of Rel. 13-14 is restricted by backward compatibility which may lead to sub-optimum design for latency critical IoT applications. Most of the latency critical IoT deployments require relatively small coverage area compared to the LTE macro deployments. Therefore, LTE-based scaled numerology is being proposed for new radio interface design of 5G [8]. Accordingly, LTE subcarrier spacing is either increased or decreased by an integer factor, which equally shortens or lengthens the OFDM symbol and cyclic prefix (CP) durations, respectively. However, a channel-dependent CP is required for robustness against inter symbol interference (ISI) regardless of the subcarrier spacing. Especially for small-sized latency critical data with large subcarrier spacing (cf. Table 1) the CP overhead becomes significant. It is desirable to have the least subcarrier spacing that suffices the requirements of robustness against phase noise, Doppler spread and latency without imparting unnecessarily large CP overhead.

5.D Waveform design

Waveform design of LTE can be enhanced to fulfill the requirements of 5G IoT use cases, where relaxed synchronization, efficiently supporting very small packet transmissions (cf. Table 1), spectral confinement, time localization and very low power consumption are of key importance [9]. Therefore, 5G waveform is to be chosen keeping in view the relevant KPIs for a particular use case.

One viable option for 5G waveform design is configuring specific aspects of OFDM in order to meet the requirements of IoT use cases. Filtering or windowing adjacent bands (F/W-OFDM) leads to spectral confinement, allowing relaxed synchronization and facilitating asynchronous transmission of spectrally adjacent systems. Moreover, it increases the overall spectral efficiency by narrowing necessary guard bands.

In contrast, several new 5G waveform proposals challenge the orthogonality constraint of OFDM towards allowing relaxed synchronization and achieving spectral confinement [10]. For example, in Generalized Frequency Division Multiplexing (GFDM) several short symbols are protected by a single CP, which keeps spectral efficiency even with very short symbols without compromising on the ISI robustness in long channels. In addition, wide subcarriers provide robustness against high Doppler spreads and subcarrier-based filtering allows flexible spectral confinements.

Additionally, high peak-to-average power ratio (PAPR) is a common problem in multicarrier waveforms [9], which needs to be mitigated in order to achieve high power amplifier efficiency. Several techniques for PAPR reduction exist [11], but these typically reduce the overall spectral efficiency especially in narrow-band allocations. Alternatively, allocating a single wide subcarrier to one device completely avoids the PAPR problem, but as a downside allows only low data rates per device.

6 Concepts on Network Architecture

Ultra-low latency cannot be achieved by improving only the radio interface design. Future 5G network will be based on software defined networking (SDN) and network function virtualization (NFV) enabling a flexible and scalable architecture that can be adjusted to the needs of several use cases, which run concurrently on the same infrastructure. Therefore, use-case-specific network slices [12] are introduced, which comprise appropriate subsets of network resources and settings. In particular, latency critical IoT use cases can benefit from local computational power provided by applications running in the mobile edge cloud (MEC) since this reduces the physical and virtual communication

distance (cf. Section 4). This section introduces a mathematical tool to analyze such flexible network architecture and also provides insights on how IoT applications can benefit from such network architecture.

6.A Flow-Level Modeling as a Tool to Derive Device-Centric KPIs

In addition to lower OSI-layer effects (cf. Section 5), there exists another large impact on latency due to sharing of (radio) resources among a vast number of IoT devices. This impact is governed to a great extent by the spatial distribution of devices, their individual data rates, and their demand for network and radio resources. In particular, devices at the cell-edge reduce network performance substantially due to their significantly lower spectral efficiency. Hence, device-centric considerations are required to guarantee a minimum latency for all devices in the network.

Flow-level models [13] are based on queuing theory and constitute useful tools to model and analyze the aforementioned effects on device-centric KPIs. Similar to the well-known SDN protocol OpenFlow, data traffic is investigated on flow-level rather than on Internet protocol (IP) level. A data flow aggregates all information belonging to a transmitted object, which can be a sensor or control signal or a periodic message, depending on the IoT service type. Latency in this approach is then understood as *sojourn time*, i.e., the time between the arrival of the information at the sender until it is fully transmitted. Thus, flow-level models give a macroscopic view on the network that allows deriving statistics of device-centric KPIs, such as the distribution of sojourn times, blocking probabilities, and statistics on the fulfillment of service requirements etc.

As SDN/NFV and MECs are becoming increasingly important, investigations target at evaluating flow performance at components at the edge of the network. Accurate modeling helps understanding the underlying processes and evaluating existing and new concepts. Furthermore, it builds the foundation for the design of optimization algorithms, which can act on two different levels. Firstly, there will be data and resource management within one network slice, i.e., for its dedicated network elements, resource elements, and functionalities, servicing a certain application typically characterized by a particular traffic type. In addition, appropriate traffic or service models mimicking, on a macroscopic level, the applications at hand play a crucial role for an appropriate algorithm design and performance evaluation. Secondly on a higher level, the resources have to be allocated to each slice by a network orchestrator, assuring the coexistence of the different applications on the same infrastructure. In the SDN/NFV context, resources are understood in a more general sense, comprising network elements, functionalities or even RATs, and thus may require more general allocation. Examples on how flow-level modeling can be applied on SDN/NFV architectures may be found in [14].

6.B Service-Centric Analytics, Management and Orchestration

State-of-the-art network analytics and management and orchestration (MANO) are almost exclusively deployed in cellular networks. They generally comprise multiple 3GPP-compatible Radio Access Network (RAN) technology generations, typically incorporate a variety of equipment vendors, and go over into parts of the Evolved Packet Core (EPC), or IP Multimedia Subsystem (IMS). This includes already virtualized versions of EPC and IMS (vEPC and vIMS).

Figure 4 depicts the exemplary use case of a future integrated factory automation (cf. Section 3.A) such as a car manufacturing plant: IoT-wearing components are manufactured at different production sites and need to be transferred between sites while being monitored at all times. In addition, the IoT components are processed by wirelessly connected sensor-actuator systems inside the production site. A heterogeneity of communication requirements including low latency parts all along this production cycle is needed. This implies that services could run over a variety of radio technologies including 3GPP's cellular IoT technologies such as the future 5G RAN, Extended Coverage GSM (EC-GSM), LTE CAT-M, or narrowband IoT (NB-IoT), but also multiple Wi-Fi flavors and non-cellular IoT

technologies such as WirelessHART, ISA 100.11a, Bluetooth, Long Range WAN (LoRaWAN), and SIGFOX. Especially 5G RAN network slicing reveals the necessary shift from the conventional separation of core network and RAN towards a network architecture which evolves the virtualization concept into parts of the RAN (virtualized RAN - vRAN). Initiatives such as the IEEE Next Generation Fronthaul Interface (NGFI), the Small Cell Forum, and 3GPP drive the specification of the required new fronthaul interfaces in-between. In addition, vRAN orchestration includes a dynamic and real-time capacity management, which can benifit from flow-level analysis (cf. Section 6.A), to follow capacity demands and traffic pattern over time and space over the various aforementioned air interface technologies as well as service-specific parameterization or vendor-agnostic orchestration of vRAN parts, at least in cellular vRAN implementations.

IoT services customers such as the exemplary car manufacturer will have several underlying cellular operators as well as other or own proprietary network services under contract. Interoperability between these different networks and network technologies will be a key requirement. To support this, a future cellular network architecture should further comprise a common core network hereby enabling the revolutionary change towards service-centric analytics and MANO as well as at least a common Authentication, Authorization, and Accounting.

Considering this, a service-centric MANO will have to go far beyond pure cellular network MANO to support a holistic view on the service on hundreds of thousands of things per service. Such a service-centric MANO will provide a vendor-agnostic view on the entire heterogeneous network and will have to cope with different life cycles of things in addition to the orchestration of the virtualized network infrastructure: The exemplary IoT customer car manufacturer may as well deploy IoT services such as ITS-related services (cf. Section 3.D) after the production, when the car is at the car dealer and especially when the car is deployed by the end customer. Usually network technologies develop faster than the life span of end consumer products. Consequently, a service-centric IoT SP has to maintain an excellent IoT service while the underlying networks change.

Summarizing, a service-centric MANO can rather be seen as an ecosystem of its own instead of just a new technology. Agility requirements suggest pure software solutions based on analytics and a full range of data science technologies as well as organically interfacing with a SDN/NFV network architecture as wide as possible. Such a design would help to significantly reduce costs and to target new frontiers of integrated operational automation and agile introduction of new services in order to reduce time-to-market. A service-centric MANO platform needs to have capabilities which are independent from services and lifecycles as illustrated before. Furthermore, an interleaving between the network and the real-time analytics could also boost further metadata-like driven business opportunities for the infrastructure owners and service providers leveraging the flow of information appearing all along the process of providing *connections* for an improved user experience and network efficiency.

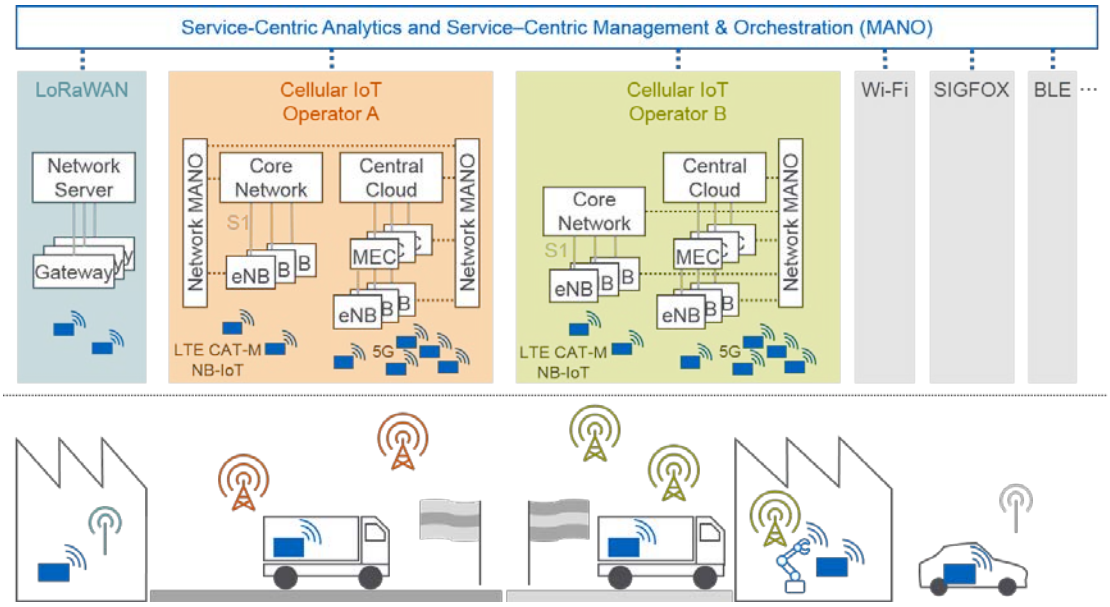


Figure 4: Service-centric analytics and service-centric management and orchestration at an exemplary IoT customer automobile manufacturer. IoT services during the production as well as over the entire life span of the end consumer product will run over a variety of radio access technologies.

7 Summary

Enabling latency critical IoT applications is one of the key targets for 5G. This article presents a comprehensive analysis of important low latency IoT use cases and their requirements on the underlying communication system. Analyzing current LTE PHY and MAC technologies and undertaking higher layer latency measurements reveal that the requirements can only be met by introducing new radio interface design and novel network architecture concepts. In particular, we have described resource delegation schemes for D2D communication, new waveform candidates, fast uplink access schemes and TTI shortening techniques. Moreover, a flexible network architecture incorporating SDN and NFV concepts will be able to adapt to different service requirements, where applications become less dependent upon RATs or operators and follow a service-centric perspective. However, such a perspective demands concepts of network (self-) optimization and network orchestration, too, since the gained flexibility naturally comes along with increased complexity arising from the vast number of IoT use cases and their additional optimization constraints. One promising approach to efficiently control the increased complexity of service-centric management is through flow-level models. In particular, the ability to describe networks, which serve different types of data traffic with diverse requirements, on a large scale analytically can help designing and analyzing SDN and NFV functionalities effectively.

From a service-centric perspective, MANO becomes increasingly important for IoT applications. Hundreds of thousands of devices will be connected through different RATs during their life cycles. Due to the prevalent environmental conditions, the application requirements, and development and changes in the underlying network, the best suitable RAT may vary over time. Therefore, interoperability between various RATs or even operators has to be guaranteed. Consequently, service-centric MANO has to provide a holistic, vendor-agnostic view on the entire network.

8 Acknowledgement

The work presented in this paper was partly sponsored by the Federal Ministry of Education and Research within the program "Twenty20 - Partnership for Innovation" - "fast wireless".

9 References

[1] 5G PPP, "5G Automotive Vision," Whitepaper, 2015.

[2] ETSI TR 102 889-2 v1.1.1, "Electromagnetic compatibility and Radio spectrum Matters (ERM); System Reference Document;Short Range Devices (SRD); Part 2: Technical characteristics for SRD equipment for wireless industrial applications using technologies different from Ultra-Wide Band," Aug. 2011.

[3] A. Frotzscher *et al.*, "Requirements and current solutions of wireless communication in industrial automation," in *Proceedings of the IEEE International Conference on Communications Workshops (ICC)*, Sydney, NSW, 2014, pp. 67-72.

[4] D. M. Soleymani *et al.*, „A Hierarchical Radio Resource Management Scheme for Next Generation Cellular Networks,“ in *Proceedings of the IEEE WCNC Workshop on Device to Device communications for 5G Networks*, Doha, Qatar, 2016, pp. 416-420.

[5] 3GPP TR 36.881 v0.6.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Study on latency reduction techniques for LTE (Release 13)," Feb. 2016.

[6] D. Jiang *et al.*, "Principle and Performance of Semi-Persistent Scheduling for VoIP in LTE System," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*, Shanghai, 2007.

[7] Ericsson contribution to 3GPP TSG RAN WG1 Meeting #84, "System level evaluation results for TTI shortening techniques," Tech. Rep. R1-161167, St Julian’s, Malta, Feb. 2016.

[8] Ericsson contribution to 3GPP TSG RAN WG1 Meeting #84bis, "Numerology for NR," Tech. Rep. R1-163227, Busan, South Korea, Apr. 2016.

[9] A. Zaidi *et al.*, "A Preliminary Study on Waveform Candidates for 5G Mobile Radio Communications Above 6 GHz," in *Proceedings of the workshop on 5G New Air Interface in conjunction with IEEE VTC Spring*, Nanjing, China, 2016.

[10] G. Wunder *et al.*, "5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Commun. Mag.,* vol. 52, no. 2, 2014, pp. 97-105.

[11] Y. Rahmatallah and S. Mohan, "Peak-To-Average Power Ratio Reduction in OFDM Systems: A Survey And Taxonomy," *IEEE Communications Surveys & Tutorials,* vol. 15, no. 4, 2013, pp. 1567 - 1592.

[12] NGMN Alliance, “5G White Paper,“ Whitepaper, 2015.

[13] J. W. Roberts, "Traffic Theory and the Internet," *IEEE Commun. Mag.,* vol. 39, no. 1, 2001, pp. 94-99.

[14] K. Mahmood *et al.*, "Modelling of OpenFlow-based software-defined networks: the multiple node case," *IET Networks,* vol. 4, no. 5, 2015, pp. 278 - 284.

This document is a preprint of: P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S.A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste and M. Windisch, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," in IEEE Communications Magazine (ComMag), vol. 55, no. 2, Feb 2017. DOI:10.1109/MCOM.2017.1600435CM

10 Biographies

PHILIPP SCHULZ (philipp.schulz@ifn.et.tu-dresden.de), born in 1990, studied Mathematics at TU Dresden, where he received his M.Sc. degree in 2014. There he also worked as a research assistant in the field of numerical mathematics, modeling, and simulation. In July 2015 he joined the Vodafone Chair Mobile Communications Systems at TU Dresden and became a member of the system-level group. Now his research focuses on flow-level modeling and the application of queuing theory on communications systems.

MAXIMILIAN MATTHÉ (maximilian.matthe@ifn.et.tu-dresden.de) received his Dipl.-Ing. degree in Electrical Engineering from TU Dresden in 2013. During his studies he focused on mobile communication systems and communication theory. In his Diploma Thesis he concentrated on waveform design for flexible multicarrier systems. Since 2013 he pursues his Ph.D. in the Vodafone Chair Mobile Communication Systems at TU Dresden. Maximilian's research focuses on the design and evaluation of MIMO architectures for future networks.

HENRIK KLESSIG (henrik.klessig@ifn.et.tu-dresden.de) received his M.Sc. and Ph.D. degrees from the Department of Electrical and Computer Engineering of TU Dresden, Germany, in 2012 and 2016, respectively. In 2011, he visited Alcatel-Lucent, Bell Labs, Germany, where he was engaged in LTE base station power modeling and energy-efficient resource management. Currently, Henrik continues as a PostDoc at the Vodafone Chair at TU Dresden. His research interests include data traffic modeling, SON, ultra-low latency communications, and the Tactile Internet.

MERYEM SIMSEK (meryem.simsek@ifn.et.tu-dresden.de) is a group leader at the TU Dresden since 2014. She won the IEEE Communications Society Fred W. Ellersick Prize in 2015. Since June 2015, she is chairing the IEEE ComSoc Tactile Internet TSC and has initiated the IEEE P1918.1 working group. She joined ICSI Berkeley in October 2016. Her main research interests include fifth generation (5G) wireless systems, wireless network design and optimization, and the Tactile Internet and its applications.

This document is a preprint of: P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S.A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste and M. Windisch, “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture,” in IEEE Communications Magazine (ComMag), vol. 55, no. 2, Feb 2017. DOI:10.1109/MCOM.2017.1600435CM

GERHARD FETTWEIS (fettweis@ifn.et.tu-dresden.de, F’09) earned his Ph.D. under H. Meyr at RWTH Aachen. After one year at IBM Research, San Jose, CA, he moved to TCSI Inc., Berkeley. Since 1994 he is Vodafone Chair Professor at TU Dresden, Germany, with 20 companies sponsoring his research on wireless transmission and chip design. He coordinates 2 DFG centers at TU Dresden (cfaed and HAEC), is member of the German academy acatech, and has spun-out eleven start-ups.

JUNAID ANSARI (junaid.ansari@ericsson.com) is associated with Ericsson Research and contributing to 5G standardization. Previously, he worked as postdoctoral researcher and research assistant at the Institute for Networked Systems at RWTH Aachen University. He received his Ph.D. (2012) and M.Sc. (2006) degrees from RWTH Aachen University. He has actively contributed to several collaborative national and European Union funded research projects. His research interests include embedded intelligence and system architecture design for the next generation of wireless networks.

SHEHZAD ALI ASHRAF (shehzad.ali.ashraf@ericsson.com) is an experienced researcher at Ericsson Research, which he joined in 2013. He holds an M.Sc. in electrical engineering from RWTH Aachen University. Since joining Ericsson, he has been deeply involved in European Union and German government funded research projects related to the development of 5G concepts. Currently, he is also involved in 3GPP standardization. His research interests include 4G and 5G radio access technologies, and machine-type communications.

BJOERN ALMEROTH (bjoern.almeroth@radioopt.com) is a data scientist at RadioOpt GmbH. He received his Ph.D. in Electrical Engineering in 2015 from TU Dresden, Germany. During his studies, he investigated the subject of analog-to-digital conversion in the context of multi-band signal reception. His current research focus is on the topic of Quality-of-Service and Quality-of-Experience monitoring in today’s and upcoming 5G networks using crowd-sourced mobile customer experience measurements.

This document is a preprint of: P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S.A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste and M. Windisch, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," in IEEE Communications Magazine (ComMag), vol. 55, no. 2, Feb 2017. DOI:10.1109/MCOM.2017.1600435CM

JENS VOIGT (jens.voigt@amdocs.com) is currently with Amdocs Network Solutions in Dresden, Germany. His professional background includes radio access network analytics and optimization. He is a technology passionate, which he has proven in long-term university collaboration, successfully managed research projects, and product innovation. He holds a diploma (1995) and a doctoral (2001) degree in electrical and computer engineering from TU Dresden and is the co-author of 40+ scientific publications and multiple international patent families.

INES RIEDEL (ines.riedel@amdocs.com) received her Dipl-Ing. (2006) and Dr.-Ing. (2014) degrees from TU Dresden, Germany. During internships at Sony, Germany and Telefónica R&D, Spain, she was involved in software defined radio developments. From 2006 to 2014 she worked as research associate and senior research associate at the Vodafone Chair at TU Dresden and joined Amdocs in 2014. Her research interests include radio access network analytics and optimization and future network technologies.

ANDRE PUSCHMANN (andre.puschmann@softwareradiosystems.com) received his Dipl.-Inf. and Ph.D. degrees in computer engineering from Technische Universität Ilmenau, Germany, in 2009 and 2015, respectively. He is now with the CONNECT Centre for Future Networks and Communications in Dublin, Ireland and also a senior engineer within Software Radio Systems Ltd. His research focuses on lower layer radio protocols and resource management strategies for safety-critical applications, including vehicular networks and machine-to-machine communication.

ANDREAS MITSCHELE-THIEL (andreas.mitschele-thiel@tu-ilmenau.de) is a full professor and head of the Integrated Communication Systems group at the Ilmenau University of Technology, Germany. He received a M.S. in Computer and Information Science from The Ohio State University (1989) and a Doctoral (1994) and Habilitation degree (2000) in Engineering from the University of Erlangen. He has held various positions in the telecommunication industry including Lucent Bell Labs and Alcatel. His research focuses on the engineering of telecommunication systems.

This document is a preprint of: P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S.A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste and M. Windisch, “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture,” in IEEE Communications Magazine (ComMag), vol. 55, no. 2, Feb 2017. DOI:10.1109/MCOM.2017.1600435CM

MICHAEL MÜLLER (michael.mueller@ivm-sachsen.de) is the head of research of the Institut für Vernetzte Mobilität gGmbH (IVM) and was the former chief of research and development of the MUGLER AG. His research interests are (mobile) communication networks and interconnected mobility. He has a long research experience in physics and wireless networks as well as in the interdisciplinary study of the mobility sector.

THOMAS ELSTE (thomas.elste@imms.de) has studied computer science at the Ilmenau University of Technology. Since 2005 he has been working at the Institut for Microelectronics and Mechatronics Systems GmbH (IMMS) in Ilmenau, Germany at the department System Design where his work is focused on kernel level programming and application development for embedded systems running linux, realtime operating systems and realtime network technologies for automation applications.

MARCUS WINDISCH (marcus.windisch@freedelity.com) is co-founder and CEO of Freedelity, a leading specialist for low-latency wireless communications systems with focus on professional audio applications. After master studies at the University of Madison, Wisconsin, he received his Dipl.-Ing. degree (2002) and his Ph.D. (2007) in Electrical Engineering from TU Dresden. He co-founded two start-up companies and holds several patents.