NETWORK ARCHITECTURES FOR DEMANDING 5G PERFORMANCE REQUIREMENTS

Tailored Toward Specific Needs of Efficiency and Flexibility

Philipp Schulz, Albrecht Wolf, Gerhard P. Fettweis, Abubaker Matovu Waswa, Dariush Mohammad Soleymani, Andreas Mitschele-Thiel, Torsten Dudda, Markus Dod, Marco Rehme, Jens Voigt, Ines Riedel, Tushar Wankhede, Walter P. Nitzold, and Bjoern Almeroth

The fifth generation (5G) of mobile networks is envisioned to support new applications having demanding requirements, such as low latency and high reliability, which is the focus of this article along with enhanced traditional mobile broadband and massive sensing. Different approaches have already been proposed to achieve low latency while guaranteeing high reliability. However, the challenge of efficient resource utilization remains. In this article, concepts for a flexible and low-latency-enabling mobile network architecture are presented, along with strategies for staying efficient. The work is put in perspective with respect to ongoing standardization activities. Finally, future visions for network management architectures and 5G's impact on economic aspects are discussed.

Digital Object Identifier 10.1109/MVT.2019.2904185
Date of publication: 18 April 2019

Envisioned 5G Applications and the Current Situation

With 5G, a variety of new applications is anticipated. In addition to enhancing traditional mobile broadband (eMBB), which demands increasing throughput and capacity, and massive machine-type communications (mMTC), mainly driven by the support of extremely high numbers of devices and low energy consumption, there are mission-critical applications that require ultrareliable low-latency communications (URLLC). The latter

1556-6072/19©2019IEEE

domain is investigated within the collaborative research project "fast wireless" [16], the outcomes of which are presented here.

In our prior work [1], promising 5G applications along with their requirements were provided, and concepts relating to the physical (PHY) and media access control (MAC) layer as well as to network architecture were presented. Applications from factory automation were identified as the most demanding use cases, with end-to-end (E2E) latency requirements down to 0.25 ms and packet loss rates (PLRs) of 10^{-9} as the reliability constraint. Less stringent requirements were found for intelligent transport systems (ITSs), with latency and reliability demands as low as 10 ms and 10^{-5} PLR, respectively. Furthermore, requirements may be formulated as guaranteed latency, i.e., data need to be received within a latency bound up to a very low error margin. For instance, the International Telecommunication Union [2] requires 5G to deliver a 32-B packet within 1 ms for a residual error of 10^{-5}.

Our studies in [1] also revealed that current mobile networks (4G) are far from meeting those stringent requirements, providing E2E round-trip latency of only around 40 ms under low load conditions. Reviewing the number of mobile subscriptions and the mobile traffic for eMBB suggests that both metrics grow continuously, confronting networks with increasing load conditions [3]. Simultaneously, advancements in hard- and software for handsets, infrastructure in the mobile operator core network, and connectivity to target Internet servers lead to unceasing improvements in existing 4G networks. By analyzing crowd-sourced data from millions of handsets, we observed an improvement in E2E latency and throughput of about 10–20% from 2017 to 2018. However, these marginal tweaks do not satisfy the requirements postulated previously. Solutions to fulfill the ambitious requirements include network densification and flexible network architectures to tackle the coming challenges [4].

Based on that, this article focuses on a network architecture to meet the challenging URLLC requirements while retaining efficiency. Starting from the state of the art by recapping ongoing standardization activities, the envisioned architecture is presented with respect to exemplary applications. Thereafter, the focus shifts to technical analysis, covering the radio interface, multiconnectivity (MC), latency modeling, and device-to-device (D2D) communications. Finally, we consider future visions for network management and the expected economic impact 5G will have on the market.

Standardization Activities

The 5G standardization in the 3rd Generation Partnership Project (3GPP) concluded with the major Release 15 [5]. It is the first release for the newly developed radio access technology (RAT) New Radio (NR). Moreover, for LTE, several features to enable 5G use cases have been specified. The new standard supports integration of both technologies in multiple variants, i.e., LTE base stations (eNBs) interworking with NR base stations (gNBs), with the Evolved Universal Terrestrial Radio Access core network and 5G core network (5GC), respectively. In such solutions, the user equipment (UE) connects simultaneously via different carriers with two base stations (eNB or gNB), which is denoted as *dual connectivity* (*DC*) or *MC*.

The protocol architectures for radio access in LTE and NR largely coincide with and consist of the PHY, MAC, radio link control (RLC), and packet data convergence protocol (PDCP) as well as the service data adaption protocol for quality of service (QoS) flow-handling from 5GC for NR. To support URLLC, new features and enhancements have been introduced for both RATs.

Shorter Transmission Time Intervals

To achieve low latency, a feature called *short transmission time interval* (*TTI*) in LTE and *minislot* in NR has been introduced. It allows scheduling and transmission with a granularity finer than a 1-ms transmission slot length, i.e., allowing subdivision of the transmission slot into up to six short TTIs. Thereby, waiting times for the next TTI, the TTI length itself, and round-trip times for potential retransmissions were significantly reduced. Furthermore, in NR, a flexible orthogonal frequency-division multiplexing (OFDM) numerology, i.e., subcarrier spacing (SCS), provides another potential for latency reduction because higher SCSs (15–240 kHz) translate to shorter slot lengths (1–0.0625 ms). Moreover, faster processing time requirements have been specified for those TTI lengths, i.e., the UE must encode uplink (UL) and decode downlink (DL) transmissions faster.

Resource Preallocation with Configured Scheduling

Configured scheduling provides a long-lasting periodic UL grant to the UE before potential UL transmissions occur. This approach avoids the otherwise introduced latency resulting from requesting and waiting for UL resources. In LTE, this feature is concerned with enhancement of the semipersistent scheduling framework.

MC

MC concepts generally refer to system architectures, where users are simultaneously connected via multiple wireless links. For URLLC, it is most desirable to transmit information over independent channels in a single time slot. Thus, spatial and frequency channel separation are suitable to combat channel dispersion, e.g., fading.

Release 15 enables MC by DC or carrier aggregation (CA) with duplication on the PDCP layer, such that packets are transmitted via two distinct frequency carriers,

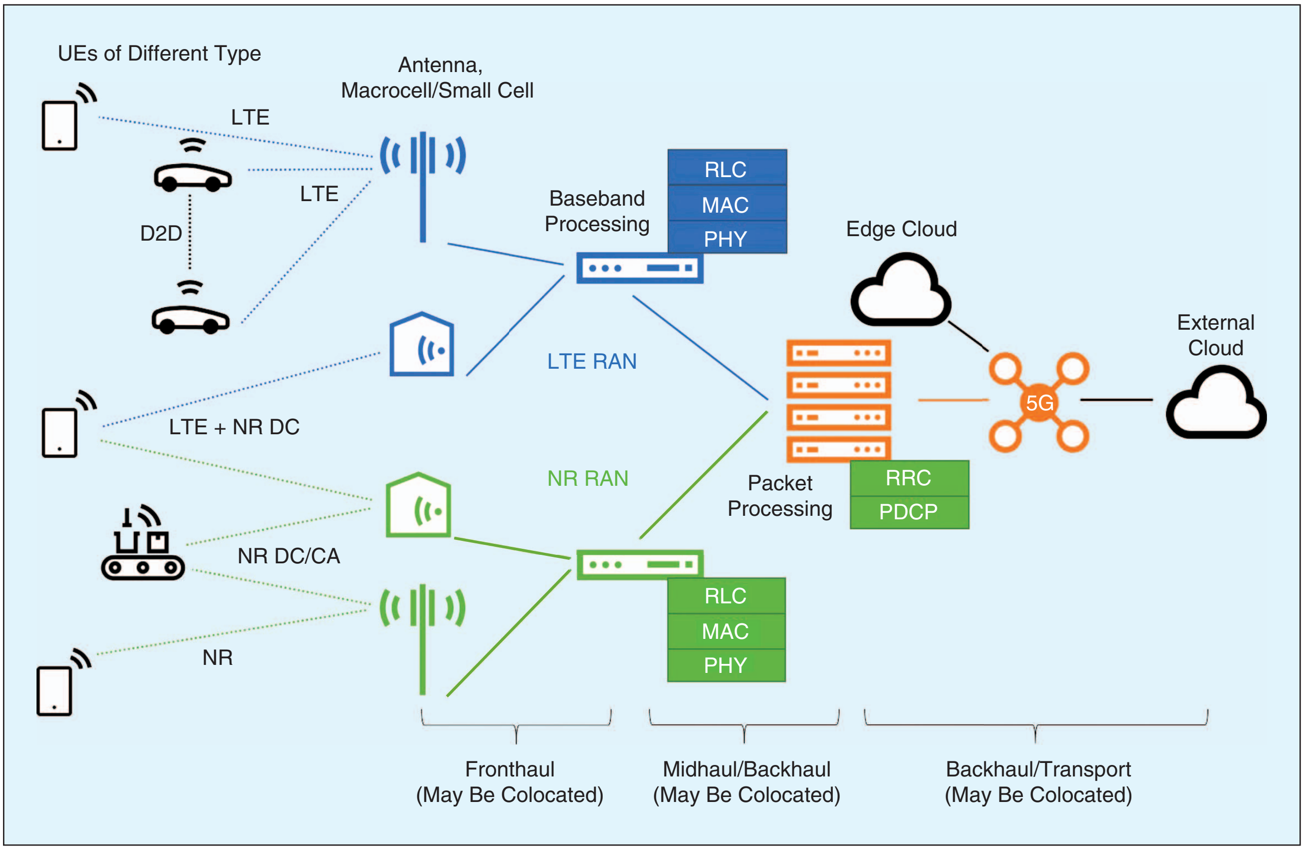

FIGURE 1 An illustration of a flexible 5G architecture for different applications.

thus providing robustness against fading. By employing the PDCP layer, hybrid automatic repeat request (HARQ) and RLC ARQ retransmissions are still possible individually per carrier. This way, reliability, i.e., robustness against temporary outages on the radio link (something HARQ and RLC ARQ on only a single link could not mitigate), can be greatly improved.

A Flexible System Architecture

Example applications from ITSs and factory automation present entirely different requirements in terms of latency and reliability than do eMBB use cases, for which scalability and efficiency are of higher importance. For instance, at a road intersection, safety-critical sensor data will be shared among cars and roadside units, but in-car entertainment or pedestrians with eMBB services may also be present, both with varying demands. Serving these applications with a single system calls for a flexible and adaptive architecture. The 5G standards allow such flexibility, and this section provides appropriate configurations for the aforementioned use cases.

The radio access network (RAN), consisting of LTE and NR RAN, is illustrated in Figure 1 as antenna sites for macrocell or small cell, connected via fronthaul to the baseband for PHY and MAC processing and connected via mid-/backhaul to packet processing, i.e., PDCP and radio-resource-control processing. 5G RAN is then connected to the core network and application, depicted in Figure 1 as the cloud. Although logically separable, these functions may be colocated. This way, the 5G network allows low-latency applications to be served by integrating these functions more closely together; e.g., it is instrumental to deliver a packet with a 1-ms latency requirement as fast as possible over the RAN, and colocation of RAN protocol functions thereby enables quick reactions, such as retransmissions, to meet latency and reliability targets. In contrast, centralization of these network functions can lead to efficiency/cost advantages when low latency is not required.

UE may be served by a single RAT, either LTE or NR, but may also be served within a DC protocol architecture by both technologies simultaneously. For eMBB, higher throughputs can be achieved by aggregating these resources. Therein, PDCP serves as router for packets between the technologies. Other use cases, e.g., factory automation, can benefit from DC in terms of reliability, i.e., with PDCP duplication (see the “Standardization Activities” section). The tradeoff between these two modes is investigated in the “Rate–Reliability Tradeoff for MC” section. DC may also be realized within one RAT. Furthermore, PDCP duplication can also reuse the CA scheme, where transmissions may be more synchronized, if

Table 1 *The achievable one-way latency with NR (see [6]).*

Latency (ms)	*HARQ*	*LTE Release 14 FDD*	*NR 15-kHz SCS FDD*				*NR 30-kHz SCS FDD*				*NR 120-kHz SCS FDD*			*NR 120-kHz SCS TDD*		
		14 os	14os	7os	4os	2os	14os	7os	4os	2os	14os	7os	4os	14os	7os	4os
DL	First Tx	4.0	2.4	1.4	1.0	0.71	1.2	0.71	0.50	0.36	0.41	0.29	0.23	0.54	0.35	0.27
	One re-Tx	12.0	5.4	2.9	1.9	1.4	2.7	1.5	0.93	0.71	0.79	0.60	0.48	1.0	0.72	0.55
UL (SR)	First Tx	12.0	4.5	2.5	1.6	1.4	2.3	1.3	0.82	0.68	0.67	0.54	0.46	0.92	0.67	0.53
	One re-Tx	20.0	8.4	4.4	2.7	2.1	4.2	2.2	1.4	1.1	1.2	0.91	0.73	1.5	1.1	0.84
UL (CS)	First Tx	4.0	2.4	1.4	1.0	0.71	1.2	0.71	0.50	0.36	0.41	0.29	0.23	0.54	0.35	0.27
	One re-Tx	12.0	5.4	2.9	1.9	1.4	2.7	1.5	0.93	0.71	0.79	0.60	0.48	1.0	0.72	0.55

CS: configured scheduling; first Tx: latency of the initial transmission; one re-Tx: latency of one retransmission; SR: scheduling request.

Values below 1 ms are highlighted in green.

carriers are terminated by the same baseband (i.e., MAC). Hence, it becomes obvious how appealing the 5G RAN architecture options are, i.e., CA duplication (split in baseband) for URLLC applications requiring strictly synchronized duplicate packet transmissions as well as DC duplication (split in packet processing) for applications requiring ultrahigh reliability (robust against carrier outages) with a moderate latency target.

Another communication mode is D2D, also illustrated in Figure 1. D2D is useful in ITSs with typically high densities of UE (i.e., cars) within a local area. A low-latency D2D communication among vehicles is necessary for cooperative road safety applications. There, cars depend on fast and reliable communication, characterized by short packets and frequent transmissions, to share maneuver intensions and warnings regarding near-traffic situations (e.g., accidents). The D2D communication may be complemented by sharing and combining information from vehicles with mobile edge computing (MEC) located close to the eNB/gNB. Individual-vehicle sensor data can be aggregated by MEC to provide an overview of an entire intersection, point out blind spots, or provide direct warnings for affected vehicles.

Along with D2D and the local information exchange via MEC, communication with the network for global information exchange is typically required as well. Communication with a traffic management center or an Internet service is necessary for applications related to traffic efficiency and comfort. This requires wide network coverage rather than low latency or high throughput. Therein, the infrastructure may also assist D2D scheduling.

The presented architecture is capable of covering all use case requirements we discuss. We foresee that it can handle different services by multiple network slices, i.e., logical networks using different parameters, resources, or technologies, tailored to application-specific needs.

Technical Aspects and Evaluation

The presented architecture promises a flexibly configurable infrastructure, depending on the application and corresponding requirements. Therefore, it is important to know the achievable performance through different configurations and technologies. This section provides numbers to explain the previously discussed concepts and introduces approaches for reliability and latency modeling.

Achievable Latency With the NR Interface

In the following, NR and short TTI for the radio network to support URLLC requirements are evaluated in terms of guaranteed latency (see the "Overview of the Current Situation" section).

For mobile communication system design, generally, the tradeoff between latency, reliability, and resource efficiency is crucial: robust transmissions can be achieved by spending more resources (e.g., bandwidth), whereas it may be more efficient to allow higher error margins, which can be corrected with retransmissions if necessary, resulting in higher latency. Table 1 compares the achievable latency for an LTE baseline to NR system configurations [OFDM SCS, slot/minislot length in number of OFDM symbols (os), and frequency-division duplex (FDD)/time-division duplex (TDD)] and shows a potential additional latency of one HARQ retransmission compared to a successful initial transmission. UL transmissions without prescheduling (see the previous section) require additional latency for requesting resources. Further latency components are described in [6]. Table 1 does not include reliability in terms of achieving those latency numbers; this depends on the link quality and chosen encoding and is analyzed in the following section.

To evaluate the impact of shorter transmission lengths for minislots and compare the theoretical values of Table

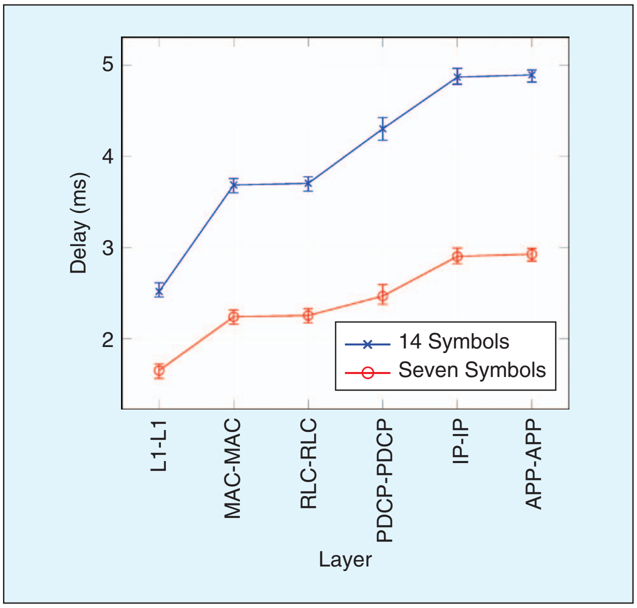

FIGURE 2 The latencies measured from layer to layer in a DL transmission for 15-kHz SCS. Higher-layer latencies include latencies of lower layers.

1 with empirical results, E2E latency measurements were conducted. This includes latency additions of higher layers, up to the application layer. Theoretical investigations claim a latency reduction of up to 50% [7]. For the measurements, a combination of a basic LTE PHY layer implementation on a National Instruments software-defined radio platform [8], capable of running LTE and 5G NR-like symbol configurations such as seven symbol minislots, was combined with the higher layers from the network simulator ns-3. The measurement gives insight into the contribution of different layers to the overall latency. Figure 2 shows the results of a single DL transmission with 15-kHz SCS for normal transmission length and a minislot length of seven symbols.

Overall, a clear latency reduction can be observed. The base latency of the PHY layer with 2.5 ms for 14 symbols can be reduced by 35%, to 1.6 ms. Subsequent layers add different amounts of latency. Although, e.g., RLC latency is negligible, the Internet Protocol (IP) layer adds 0.5 ms of latency irrespective of the transmission length. An overall latency reduction of 45% is shown and verifies the theoretical thresholds. The results also verify the calculated possible low latencies in Table 1 in the DL configuration of 15-kHz SCS. The measured 2.5 ms for 14 symbols and 1.6 ms for seven symbols correspond closely to the theoretical values of first transmission.

Rate–Reliability Tradeoff for MC

MC concepts, which have been mainly applied for enhancing data rates (multiplexing), are also suitable to improve transmission reliability (diversity). Both objectives are desirable, but contradictory, in most wireless networks. In general, improving data rates usually decreases transmission reliability, and vice versa.

With PDCP packet duplication, for example, reliability can be greatly improved. Given a provided residual error probability of 10^{-3} for each individual link, duplicating a packet over both links and assuming they are uncorrelated (which is likely when aggregating different frequency carriers), a residual error probability below 10^{-6} is achievable. However, this halves resource efficiency and thus the achievable data rate. In [9], an analytical framework was established to describe this tradeoff based on the interrelation between outage probability (reliability) and system throughput (data rate), which is denoted as *rate–reliability tradeoff* (*RRT*). The RRT goes beyond current standardization and is a first important step toward a flexible wireless network that can quickly adapt to user requirements. The tradeoff was established by time-sharing between a multiplexing mode and diversity mode, i.e., both modes are in operation for a certain time interval. In the multiplexing mode, different information is transmitted via multiple links, leading to high system throughputs but high outage probabilities. In contrast, transmitting the same information in the diversity mode leads to an opposite result.

The time-sharing of both modes is a powerful tool to adjust the system configuration accordingly to the user requirements, as illustrated in the following. Let us assume an MC system with three uncorrelated links (e.g., separated carrier frequencies), a bandwidth of 20 MHz per link, and an average system transmit signal-to-noise ratio (SNR) of 30 dB equally allocated to all links. Furthermore, we assume a channel code rate of 1/2 and binary phase-shift keying, 4-quadrature amplitude modulation (QAM), and 16-QAM as the modulation scheme, yielding a spectral efficiency of $R_c = \{0.5, 1, 2\}$. For a detailed explanation of the system model, we refer to [9]. We first discuss the RRT for a spectral efficiency of 1/2, as depicted by the solid blue line with star markers in Figure 3. At full diversity, an outage probability below 10^{-9} can be reached at the cost of a low system throughput of 10 Mb/s. In contrast, at full multiplexing, a high system throughput of 30 Mb/s is achievable for a tolerated outage probability above 10^{-3}. By time-sharing between both modes, every pair on the solid line is achievable, thus allowing adaptation to manifold requirements. If the spectral efficiency is increased, the RRT curves shift toward higher outage probabilities and higher system throughputs. Moreover, evaluating both corner cases (full multiplexing and full diversity) for any spectral efficiency value provides the dashed lines, which define the feasible configurations.

Even though MC is suitable for achieving high reliability, the highly dynamic radio and traffic environment must be considered as well. It may happen that duplicates on one path cannot be delivered within the required latency bound, i.e., receiving the duplicate too

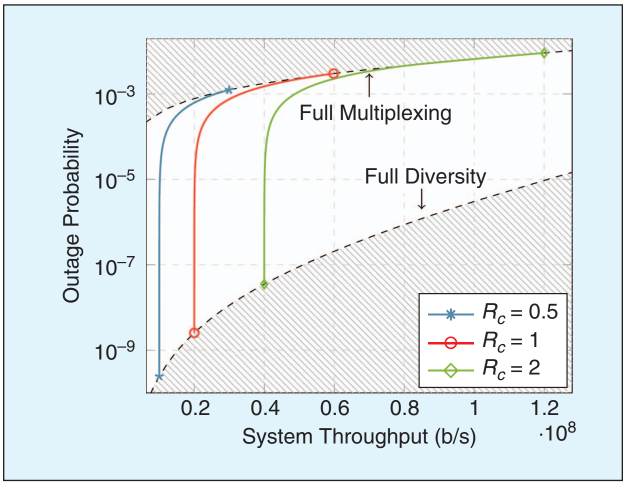

FIGURE 3 The RRT with three links, an average system transmit SNR of 30 dB, a bandwidth of 20 MHz, and different spectral efficiencies. Infeasible regions are depicted as shaded areas.

late is not beneficial. If the estimated latency is too high for the duplicate transmission, it may be discarded, saving resources. Here, latency modeling, presented in the “Latency Modeling” section. could be a foundation for exploiting this potential.

Efficiency Through Resource Delegation in D2D Communication

In this section, we evaluate efficiency aspects of different schemes for D2D communication while fulfilling the URLLC requirements (see the “Overview of the Current Situation” section). The proximity service in D2D communication can support extremely high data rates and low delays and, thus, makes D2D communication a promising technique for URLLC. Moreover, concepts such as configured scheduling (see the “Standardization Activities” section) may result in wastage of reserved resources due to the coarse granularity of subframes in LTE and small payload of URLLC applications.

Short TTI and fast delegation schemes, i.e., subgranting, are two concepts for reducing radio resources wastage [1]. However, even for the shortest TTI duration, efficiency is deteriorated by 25–45% [10]. This is primarily because MAC adds padding information when the oncoming traffic payload does not fill the scheduled subframe and cannot wait for further data from the upper layers due to latency constraints. To address this drawback, inspired by short TTI and subgranting schemes, we propose a flexible subframe combining the advantages of both schemes [10]. In this approach, the transmitter in D2D communication avoids adding the padding information by choosing the appropriate subframe length and then signals the subframe length in the subgranting signaling. The D2D receiver is able to decode the transmitted data indicated in the subgranting signaling while these data are received. Consequently, the overhead and transmission time are reduced while utilization is increased, by delegation of unused resources to device-to-infrastructure (D2I) users.

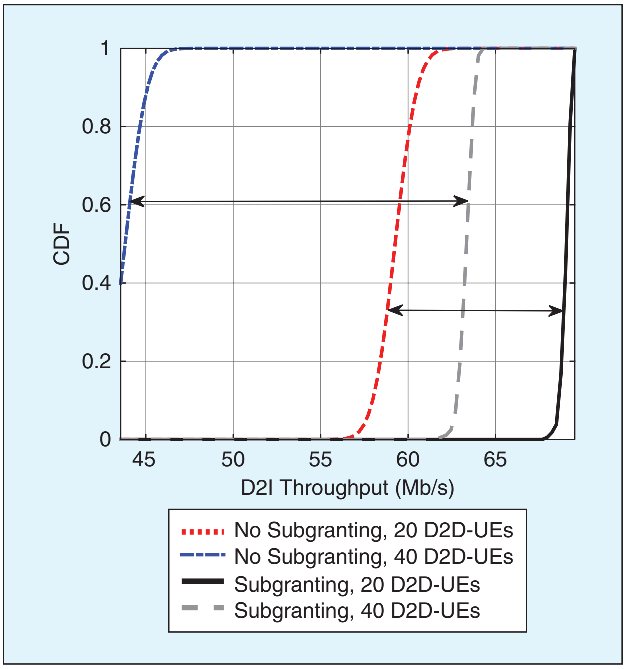

FIGURE 4 An illustration of user throughput in a scenario with 20 and 40 D2D users delegating the allocated but unused UL resources to a D2I beneficiary user in the vicinity. CDF: cumulative distribution function.

Our observation shows that overhead and transmission time decrease by at least 25% and 0.1 ms, respectively, using the flexible subframe [10]. Moreover, the user throughput increases by 17% with subgranting for 20 D2D users (see Figure 4) because the D2I user can reutilize the unused resources of D2D users. In a scenario with 40 D2D users, where fewer resources for D2I users are available, subgranting increases the D2I user throughput by 46%.

Latency Modeling

The desired network flexibility appears promising but also generates challenges. It requires sophisticated network management because resources and parameters ideally should be chosen autonomously. This network management should optimize the tradeoff between required performance (e.g., latency and reliability) and network capacity, as illustrated in Figure 5 [11].

The red triangle shows that current networks are configured to maximize their capacity at the expense of latency (milliseconds) and reliability (PLR). However, the stringent latency and reliability demands for URLLC (see the “Overview of the Current Situation” section) adversely affect the capacity (e.g., a 70% capacity drop as depicted by the blue triangle in Figure 5). Network management should thus enhance the tradeoffs among these three competing objectives for URLLC

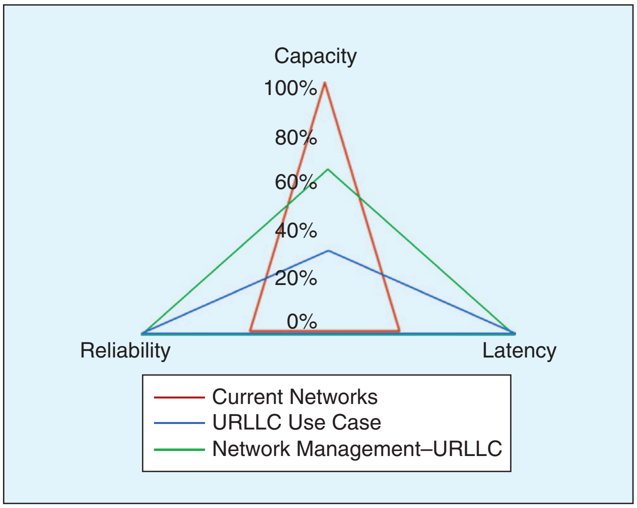

Figure 5 The tradeoffs among key network requirements.

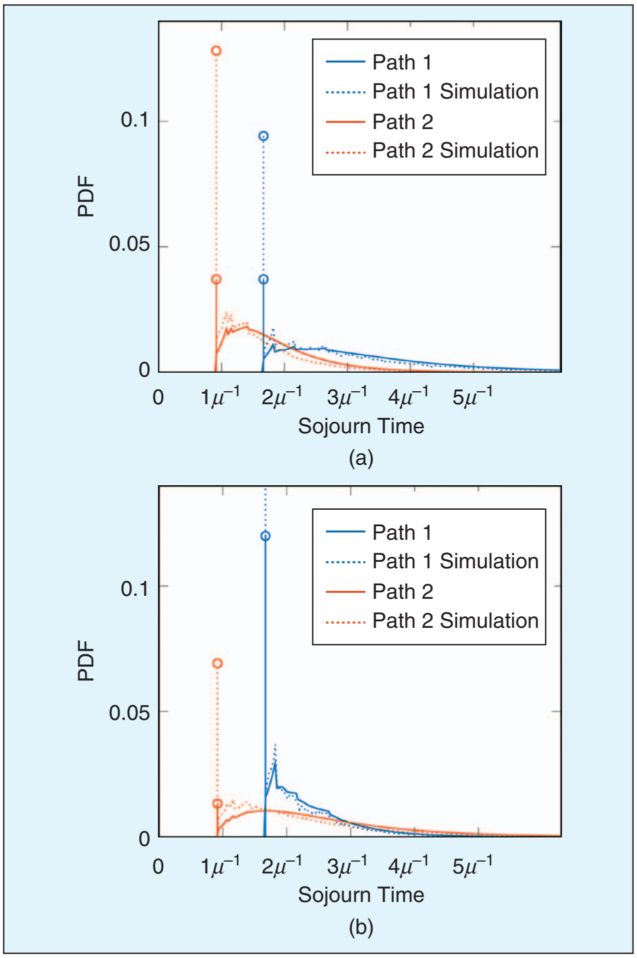

Figure 6 The latency distribution from queueing models compared to that from simulation results. Paths 1 and 2 refer to an LTE and an NR branch, respectively. (a) The load is equally distributed between both RATs. (b) NR carries twice the traffic.

applications by minimizing the capacity drop (e.g., 30%) while meeting the performance demands as shown by the green triangle.

Appropriate mathematical network models can support such network management. They allow fast network performance evaluation and can provide bounds on latency and reliability. Moreover, such models help in choosing or optimizing parameters for an efficient network. Our recently developed evaluation framework based on queuing theory and queuing networks approximates the E2E latency distribution in RANs. Here, the model is applied to the architecture in Figure 1. Each node is modeled as a queuing system, mainly characterized by the arrival and service time distribution. For this study, Poisson arrivals are assumed, whereas the service is set to be deterministic, reflecting short URLLC packets. With μ being the service rate of an LTE node, the rate of the NR nodes is set to $2\,\mu$ for 30-kHz SCS, modeling shorter TTIs. For a fair comparison, i.e., based on the same throughput, twice the number of packets will be forwarded to the NR nodes. The rate of the packet processing unit is accordingly dimensioned as $3\,\mu$. Any additional technical delays, e.g., packet processing, could be incorporated but are set to zero for simplicity.

The probability density functions (PDF) of the E2E latency are shown for the LTE branch and the NR branch in Figure 6(a). Simulation results are added for model validation. An approximation error occurs, mainly because the model is intended to be applied on more complex architectures, exploiting Kleinrock's independency approximation [12] for dense networks. With regard to both paths having the same load, it can be observed that the distribution tail benefits from shorter packages. In Figure 6(b), the traffic amount routed through the NR branch was doubled again, which results in reshaping the PDFs. Whereas the NR distribution tail suffers, the LTE path profits, especially in the higher percentiles. By degrading lower-percentile performance, the right tails can be improved to achieve guaranteed latency.

Open RAN Network Management Architecture

The flexible system architecture and concepts discussed in the previous sections are well suited to theoretically meet the envisioned requirements. However, successfully managing such networks remains a challenge. Recent studies [13] reveal that the achievable air interface latency depends considerably on the cell load and interference situation and so suggest optimizing the air interface efficiency, e.g., by beamforming or intercell interference coordination. Furthermore, in these studies, URLLC targets could not be met for higher mobility or larger packet sizes. To maintain high efficiency, this suggests deploying flexible network configurations, e.g., by separating network slices for URLLC services when and where needed.

Considering this, network management needs to operate on a more service-based architecture, enabling software-defined programmability for network automation and real-time management of QoS flows through E2E network slices. To cope with diverging QoS demands, including URLLC, the network management architecture requires a new software-based foundation and open, standardized interfaces, now being targeted by new industry initiatives including and beyond the 3GPP. See [14] for a recent overview.

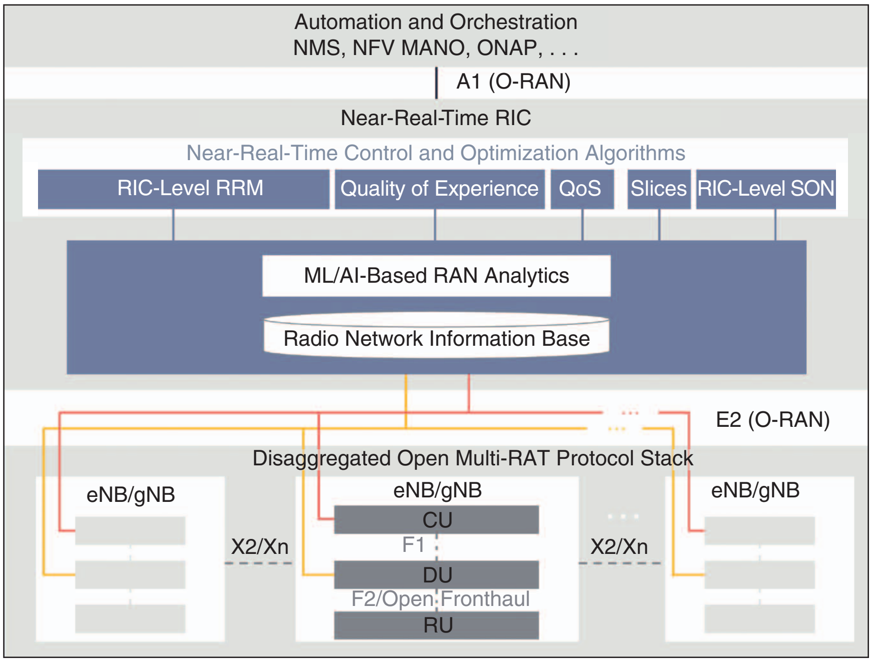
Automation and Orchestration
NMS, NFV MANO, ONAP, . . .
A1 (O-RAN)
Near-Real-Time RIC
Near-Real-Time Control and Optimization Algorithms
RIC-Level RRM
Quality of Experience
QoS
Slices
RIC-Level SON
ML/AI-Based RAN Analytics
Radio Network Information Base
E2 (O-RAN)
Disaggregated Open Multi-RAT Protocol Stack
eNB/gNB
eNB/gNB
eNB/gNB
CU
F1
DU
F2/Open Fronthaul
RU
X2/Xn
X2/Xn

FIGURE 7 The O-RAN network management architecture and open interfaces.

Figure 7 depicts the envisioned network management architecture consisting of three layers. The top layer comprises the automation and orchestration for network management systems (NMSs) as well as network function virtualization management and orchestration (NFV MANO). This layer is expected to operate on a non-real-time cycle with reaction times longer than 500 ms. Initiatives such as Linux Foundation's Open Network Automation Platform (ONAP) [17] are defining a framework for the design and operation of network functions such as network design, control, policies, inventory, configuration, and non-real-time intelligent controlling of network functions as well as network services orchestration.

The near-real-time RAN intelligent controller (RIC) constitutes the middle layer and is dedicated to a completely new concept driven by an industry initiative for open RAN (O-RAN), the O-RAN Alliance e.V. [18], founded by former Cloud-RAN and extensible RAN organizations in 2018 [15]. As a purely software-based concept, it allows for a virtualized implementation of the RIC on white box hardware. The RIC includes a radio network information base and the next generation of machine learning (ML)- and artificial intelligence (AI)-based RAN analytics as well as several control and optimization algorithms with near-real-time reaction times between 10 and 500 ms.

Having these entities north of the new open southbound (E2) interface allows exploiting the potential of combining legacy single-cell radio resource management (RRM) and conventional self-organizing network (SON) functionalities with RIC-level control and optimization algorithms at the network edge. The intelligent RAN analytics operate on a cluster of cells that are grouped considering network parameters (e.g., transmission power and connectivity) and real environment conditions (e.g., radio conditions and traffic characteristics). Thus, RIC-level algorithms are able to make intelligent control and optimization decisions and thus better fulfill application requirements (e.g., latency and reliability) by controlling multiple mutually interacting cells. Currently, O-RAN is standardizing the required new northbound (A1) and E2 interfaces.

The third layer includes implementations of potentially multiple RAT protocol stacks (see the "Standardization Activities" section). These may be deployed in several disaggregated entities, supplementing the 3GPP standards: a possibly virtualized central unit (CU), the distributed unit (DU), and the radio unit (RU), the latter including the antenna (array). To support low E2E latency, any of the current initiatives for an O-RAN architecture and open interfaces needs to enable stringent internal latency management and prioritization of low-latency services.

Impact on Evolving Value Chains

The desired flexibility, scalability, and efficiency are expected to bring fundamental changes to markets. It seems likely that 5G will transform well-established value chains. By evolving new business models, this affects not only vertical industries (e.g., manufacturing, transportation, and energy) as customers of the telecommunication sector but also the telecommunication industry itself. As indicated in Figure 8, new market roles can evolve due to more complex value creation. Alongside traditional roles, e.g., mobile network operators (MNOs), mobile virtual network enablers (MVNEs), mobile virtual network operators (MVNOs), subscribers, vendors, and regulators, new intermediaries and more specialized roles for specific tasks or supporting services are expected to emerge. These include, among others, network orchestration and slice management,

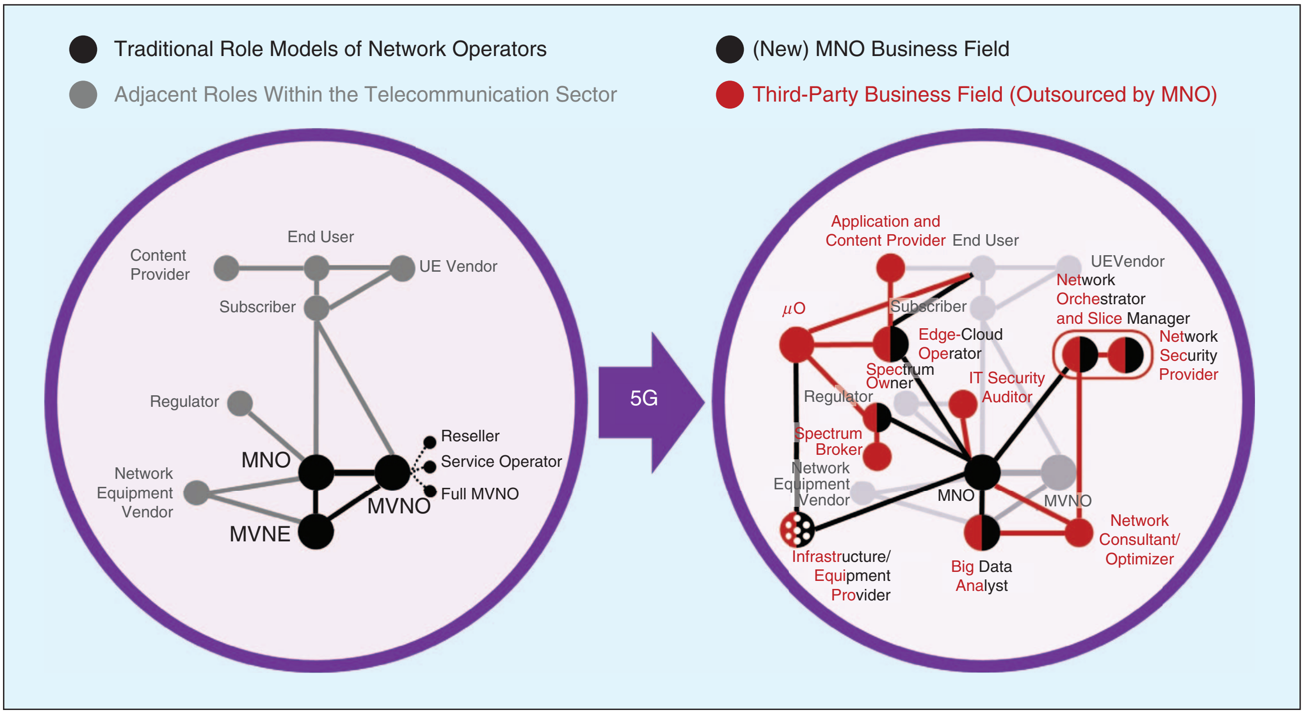
● Traditional Role Models of Network Operators
● Adjacent Roles Within the Telecommunication Sector
● (New) MNO Business Field
● Third-Party Business Field (Outsourced by MNO)
End User
Content Provider
UE Vendor
Subscriber
Regulator
Reseller
Service Operator
Full MVNO
MNO
MVNO
Network Equipment Vendor
MVNE
5G
Application and Content Provider
End User
UEVendor
Network Orchestrator and Slice Manager
µO
Subscriber
Edge-Cloud Operator
Network Security Provider
Spectrum Owner
IT Security Auditor
Regulator
Spectrum Broker
Network Equipment Vendor
MNO
MVNO
Network Consultant/ Optimizer
Infrastructure/ Equipment Provider
Big Data Analyst

FIGURE 8 The evolving value chain and market roles with 5G.

local cloud services, spectrum licensing, IT security in virtualized networks, various analysis and consulting services, and operating local subnetworks as a micro-operator (µO).

Whether traditional MNOs or third parties will carry out these roles and which mutual relations will emerge cannot yet be assessed with reasonable certainty. Multiple market configurations are plausible, ranging from barely changed to highly individualized and fragmented constellations. In the example of a µO, subnetworks for the last mile and locally tailored communication services may be operated either by new industry-specific providers for similar customers or by the site owners themselves (e.g., manufacturing, utility, or rail companies), especially in case of highly specific URLLC and mMTC applications such as factory automation and ITS. Future regulation schemes, efficiency requirements, investment needs, and migration scenarios will heavily influence the outcome for market configurations.

Summary

Envisioned 5G use cases pose requirements on future mobile networks that are challenging and as diverse as the anticipated applications. This variety calls for a flexible system architecture that can be tailored toward specific needs, as demonstrated in this article; these can be realized by establishing network slices. However, for optimal slice parametrization, an understanding of possible configurations along with their expected performance is necessary. This article addressed this issue by evaluating features of the 5G radio interface and providing insights into modeling approaches. Thereby, the focus is on latency, reliability, and efficiency.

However, even with this theoretical knowledge, challenges for actual network management remain. Thus, an O-RAN network management architecture, driven by industrial initiatives, was also presented. Finally, the new network architecture may involve fundamental changes in well-established market structures by creating new market roles and business models.

Acknowledgments

This work was partly sponsored by the Federal Ministry of Education and Research within the program Twenty20–Partnership for Innovation–“fast wireless.”

Author Information

Philipp Schulz (philipp.schulz2@tu-dresden.de) is a member of the system-level group in the Vodafone Chair Mobile Communications Systems at the Technische Universität Dresden, Germany. His research focuses on flow-level modeling and the application of queuing theory to communications systems. He is a Student Member of the IEEE.

Albrecht Wolf (albrecht.wolf@tu-dresden.de) is currently pursuing a Ph.D. degree with the Vodafone Chair Mobile Communication Systems at the Technische Universität Dresden, Germany. His research interests include network information theory and cooperative wireless communications.

Gerhard P. Fettweis (fettweis@tu-dresden.de) is Vodafone Chair Professor at the Technische Universität Dresden, Germany. His research focuses on wireless transmission and chip design for wireless/Internet of Things platforms. He is a Fellow of the IEEE.

Abubaker Matovu Waswa (abubakermatovu.waswa@tu-ilmenau.de) is a project associate in the Integrated Communication Systems Group at the Ilmenau University of Technology, Germany. His current research focuses on device-to-device communication spectrum/interference management in the licensed cellular band.

Dariush Mohammad Soleymani (Dariush.soleymani@tu-ilmenau.de) is a research associate in the Integrated Communication Systems Group at the Ilmenau University of Technology, Germany. His current research focuses on radio resource allocation for device-to-device 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. His research focuses on the engineering of telecommunication systems. He is a Member of the IEEE.

Torsten Dudda (torsten.dudda@ericsson.com) is a master researcher at Ericsson in Aachen, Germany. His current research focuses on enhancing 5G New Radio to enable critical machine-type communication use cases, such as the Industrial Internet of Things.

Markus Dod (markus.dod@mugler.de) is the head of the Software Development Department of Mugler AG, Oberlungwitz, Germany. His research focuses on communication technologies for intelligent transport systems and mathematical models for message propagation in ad hoc networks.

Marco Rehme (marco.rehme@wirtschaft.tu-chemnitz.de) is a research assistant in the Faculty of Economics and Business Administration of the Technische Universität Chemnitz, Germany. His research areas are value chain analyses and business model development for intelligent transport systems and the suitability of communication technologies for associated applications.

Jens Voigt (jens.voigt@amdocs.com) is currently with Amdocs Open Network in Dresden, Germany. His research interests include radio access network (RAN) analytics and optimization, massive MIMO, and Open RAN. He is a member of the IEEE Future Networks Community.

Ines Riedel (ines.riedel@amdocs.com) is a member of the Amdocs Open Network product management team in Dresden, Germany. Her research interests include radio access network analytics and optimization as well as the implications of future network technologies. She is a member of the IEEE Future Networks Initiative.

Tushar Wankhede (tusharashok.wankhede@elektrobit.com) is a software engineer with Elektrobit Automotive GmbH in Erlangen, Germany. His current research interest is in vehicular communication systems.

Walter P. Nitzold (walter.nitzold@ni.com) is a research engineer at National Instruments Dresden GmbH, Germany. His current research interests include radio access technology, interworking technologies, 5G coding schemes, and field-programmable gate array-based software-defined radio prototyping.

Bjoern Almeroth (bjoern.almeroth2@vodafone.com) is a senior data analyst and senior data architect at Vodafone Group Services GmbH, Dresden, Germany. His current research focus is on quality-of-service and quality-of-experience monitoring in today's and upcoming 5G networks using crowdsourced mobile customer experience measurements.

References

[1] P. Schulz et al., "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, 2017.

[2] International Telecommunication Union, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," ITU, Geneva, Switzerland, Rep. ITU-R M.2410-0, 2017.

[3] Ericsson. (2018). Ericsson mobility report. Telefonaktiebolaget L. M. Ericsson, Stockholm, Sweden. [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-june-2018.pdf

[4] Next Generation Mobile Networks Alliance, "NGMN 5G white paper," NGMN, Frankfurt, Germany, White Paper, 2015.

[5] 3rd Generation Partnership Project, "3GPP, TS 38.300 NR, NR; NR and NG-RAN overall description; stage 2 (Release 15)," 2018. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/38_series/38.300/38300-f30.zip

[6] 3rd Generation Partnership Project, "IMT-2020 self-evaluation: UP latency in NR, Ericsson, 3GPP TSG-RAN WG2#103," 2018. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R2-103--18797.htm

[7] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.

[8] National Instruments, "LabVIEW communications LTE application framework," 2019. [Online]. Available: http://www.ni.com/download/lte-application-framework-2.2/7740/en/

[9] A. Wolf, P. Schulz, D. Öhmann, M. Dörpinghaus, and G. Fettweis, "Rate–reliability tradeoff for multi-connectivity," in *Proc. IEEE Wireless Communications and Networking Conf. (WCNC)*, Barcelona, Spain, 2018, pp. 1–6.

[10] D. M. Soleymani, J. Mückenheim, M. Harounabadi, A. M. Waswa, Z. Shaik, and A. Mitschele-Thiel, "Implementation aspects of hierarchical radio resource management scheme for overlay D2D," in *Proc. 9th Int. Congr. Ultra-Modern Telecommunications and Control Systems (ICUMT)*, Munich, Germany, 2017, pp. 154–161. doi: 10.1109/ICUMT.2017.8255149.

[11] E. Roth-Mandutz, A. M. Waswa, and A. Mitschele-Thiel, "Capacity optimization for ultra-reliable low-latency communication in 5G—The SON perspective," in *Proc. 13th Int. Conf. Network and Service Management (CNSM)*, 2017, pp. 1–8. doi: 10.23919/CNSM.2017.8255978.

[12] D. P. Bertsekas and R. G. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1992.

[13] Next Generation Mobile Networks Alliance. (2018). 5G extreme requirements: Radio access network solution. NGMN, Frankfurt, Germany. [Online]. Available: https://www.ngmn.org/publications/all-downloads.html?tx_news_pi1%5Bnews%5D=706&cHash=7cdec049f36a7adc0638794b442154a7

[14] Next Generation Mobile Networks Alliance. (2018). NGMN overview on 5G RAN functional decomposition. NGMN, Frankfurt, Germany. [Online]. Available: https://www.ngmn.org/publications/all-downloads.html?tx_news_pi1%5Bnews%5D=687&cHash=dfd1f0d794cf0dd3cdc226174f4c19ee

[15] O-RAN Alliance, "Towards an open and smart RAN," O-RAN Alliance, Alfter, Germany, White Paper, 2018.

[16] P. Schulz, "Fast wireless," Technische Universität Dresden, Germany, 2009. Accessed on: Apr. 1, 2019. [Online]. Available: http://de.fast-zwanzig20.de/basisvorhaben/fast-wireless

[17] Open Network Automation Platform. Accessed on: Apr. 1, 2019. [Online]. Available: www.onap.org

[18] O-RAN Alliance. Accessed on: Accessed on: Apr. 1, 2019. [Online]. Available: www.o-ran.org

VT