3GPP NR: the standard for 5G cellular networks

Michele Polese, Marco Giordani, Michele Zorzi

Abstract The next generation of cellular networks (5G) will be characterized by ultra-high data rates, ultra-low latency and support for a massive number of connections. The 3rd Generation Partnership Project (3GPP) has recently standardized the specifications for NR, the new Radio Access Network (RAN) designed to match 5G requirements thanks to the use of the millimeter wave (mmWave) spectrum, massive Multiple Input Multiple Output (MIMO), a flexible design of the air interface and novel deployment paradigms. This chapter reviews the main novelties that NR introduces with respect to 4G cellular networks, with a focus on how they can be used to provide unprecedented performance in 5G deployments.

1 Introduction

According to the International Telecommunication Union (ITU), the 5th generation (5G) of cellular networks will need to address the traffic demands and new use cases of the digital society beyond 2020 [17]. In particular, 5G networks should support: (i) a user experienced rate of at least 100 Mbit/s, with a peak data rate in ideal conditions of 20 Gbit/s, and three times higher spectral efficiency with respect to 4G; (ii) ultra-low latency, i.e., 1 ms round-trip over the air; (iii) support for mobility, with communications at up to 500 km/h; (iv) an area capacity of 10 Mbit/s/m^2 with up to 10^6 connections per km^2; and (v) a 100x increase in energy efficiency with respect to 4G networks.

Long Term Evolution (LTE) is the set of specifications that the 3rd Generation Partnership Project (3GPP) has introduced in 2009 and evolved since then to satisfy the current 4G requirements. The evolutions of LTE will match some of the next generation requirements in specific deployment scenarios [12], but they will not be

Michele Polese, Marco Giordani, Michele Zorzi
Department of Information Engineering, University of Padova, Via G. Gradenigo, 6/B, 35131 Padova, e-mail: {polesemi,giordani,zorzi}@dei.unipd.it

able to effectively address all the 5G use cases. For example, LTE operates with a maximum of 20 MHz per carrier, thus limiting the achievable data rate, and has a rigid frame structure that makes it difficult to reduce the round-trip latency below 1 ms. Moreover, LTE has not been designed to account for energy efficiency (e.g., pilot signals are always-on) and to support a massive number of connections (even though this is targeted by the recent Narrow Band IoT (NB-IoT) evolution).

In order to overcome the limitation of LTE networks, the 3GPP has recently defined a new Radio Access Technology (RAT), i.e., 3GPP NR[1], that introduces novel designs and technologies that will comply with the 5G requirements. NR exploits a new spectrum, i.e., the millimeter wave (mmWave) band, and features the support for new techniques such as massive Multiple Input Multiple Output (MIMO), flexibility in terms of frame structure, to target different use cases, and multiple deployment options for the Radio Access Network (RAN). Moreover, a new core network design (i.e., 5G Core (5GC)) has been introduced to offer network slicing and virtualization, and different deployment options and inter-networking with LTE have been specified.

NR has been standardized by 3GPP with a first set of specifications[2] (Release 15) in December 2017 and a complete one published in June 2018. Release 16 for NR is expected to be completed in December 2019, and will be composed of a set of specifications that match the ITU 5G requirements previously described [3].

This chapter is organized as follows. In Sec. 2 we will describe the main characteristics of the NR air interface design, describing why it is flexible and lean and how low latency is achieved. Then, in Sec. 3 we will introduce mmWave communications and describe the procedures integrated in NR for the support of the mmWave bands. In Sec. 4, we will present massive MIMO techniques, focusing on both below- and above-6 GHz NR use cases. In Sec. 5 we provide insights on the new RAN and core deployment options and in Sec. 6 we conclude the chapter.

2 A flexible and lean design

The main characteristic of the NR physical layer is its flexibility: the standard, indeed, provides a general technology framework designed to address the different and, in some cases, conflicting 5G requirements [17] and to be forward compatible, so that it can accommodate future applications and use cases.

Both LTE and NR use the Orthogonal Frequency Division Multiplexing (OFDM) modulation, which divides the available *time* resources in frames of 10 ms with subframes of 1 ms, and *frequency* resources in subcarriers with spacing Δf. Moreover, subframes are further divided in slots and symbols, where the combination

[1] While NR was originally meant as the acronym for "New Radio" [2], according to the latest 3GPP specifications [8] it has lost its original meaning and it now refers to the 5G Radio Access Network.

[2] The specifications for NR are in the Technical Specification (TS) of 3GPP *38 series*, together with Technical Reports (TRs) that contain related studies. Other relevant RAN specifications can be found in the 36 (LTE) and 37 (LTE-NR inter-networking) series.

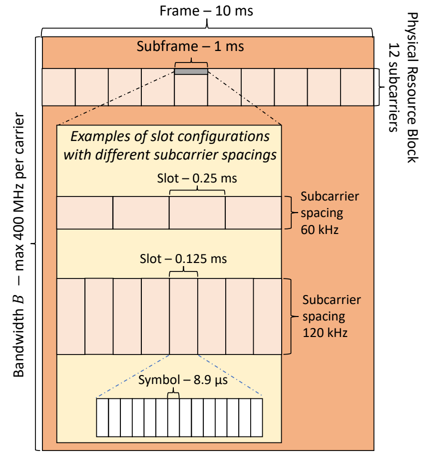

Fig. 1 Frame structure configuration for the physical layer of 3GPP NR

of a single OFDM symbol and a single subcarrier constitutes the smallest physical resource in NR. While with LTE the symbol duration and the subcarrier spacing are fixed, with NR it is possible to configure different OFDM *numerologies*[3] on a subframe basis, i.e., every subframe is self-contained and can be characterized by a different numerology [6]. This makes it possible to address different 5G use cases with a single RAT: for example, a shorter OFDM symbol duration, combined with a higher subcarrier spacing, can be used for high-data-rate and low-latency traffic, while lower subcarrier spacing can be used for low-frequency narrowband communications for machine-generated traffic [22]. Fig. 1 illustrates an example of NR frame structure with two different possible subcarrier spacings.

Another main NR novelty with respect to LTE is the support for ultra-low latency communications [13], to target the sub-1 ms round-trip latency requirement of 5G. First of all, the usage of larger subcarrier spacings and shorter symbols has the potential to reduce the transmission time with respect to the basic LTE frame structure. Moreover, control information related to modulation and resource allocation can be added at the beginning of data packets, allowing the devices to start decoding as soon as they start receiving data [6]. This also translates into tighter processing constraints in 5G NR devices, which must be able to process a received packet in a few hundreds of microseconds (the actual constraints depend on the subcarrier spacing, as discussed in [12]). Another consequence is that the devices will be able to transmit the Hybrid Automatic Repeat reQuest (HARQ) acknowledgement after just one slot, making it possible to reduce the round-trip latency below 1 ms.[4] Moreover, latency-sensitive data does not need to wait for a new slot to be transmitted, but

[3] The term numerology refers to a set of parameters for the OFDM waveform, such as subcarrier spacing and symbol duration [30].

[4] In LTE (Release 14), the round-trip latency was fixed to 3 ms [28].

the base station may decide to transmit it as soon as possible using *mini-slots*, i.e., groups of at least 2 OFDM symbols that can be allocated to a data transmission and do not need to be aligned with the beginning of a standard slot [9].

Finally, in order to increase the flexibility and the energy efficiency of the RAN, NR limits the number of always-on reference signals, thereby configuring them to match the deployment scenario and increase the energy efficiency [21]. Moreover, the self-contained subframe and the minimization of always-on signals make the NR design forward-compatible, i.e., they enable the evolution of the NR RAT to support unforeseen use cases with novel technologies and solutions without compromising the support for legacy devices [9].

3 A new spectrum frontier: millimeter waves in 3GPP NR

5G cellular systems introduce unprecedented requirements in terms of data rate, latency, link resilience, and end-to-end reliability, which go beyond what existing mobile technologies can support. In this perspective, the mmWave spectrum – roughly above 10 GHz[5] – has rapidly emerged as an enabler of the 5G performance demands in micro and picocellular networks [25]. These frequencies, combined with high-order modulation, offer much more bandwidth than 4G/LTE systems operating in the congested bands below 6 GHz, and initial capacity estimates have suggested that networks operating at mmWaves can offer orders of magnitude higher bit-rates than legacy cellular networks. Moreover, mmWave systems operate through highly directional communications which tend to isolate the users and deliver reduced interference. Additionally, inherent security and privacy is also improved because of blockage and of the short-range transmissions which are typically established.

Motivated by the above introduction, NR will boost the 5G performance by supporting, for the first time, frequencies up to 52.6 GHz in Release 15, including therefore mmWave bands [1]. Nevertheless, communication at mmWaves introduces new challenges for the whole protocol stack, which may have a significant impact on the overall end-to-end system performance. First, signals propagating in the mmWave spectrum suffer from severe path loss and susceptibility to shadowing, thereby preventing long-range omnidirectional transmissions. Second, mmWave links are highly sensitive to blockage and have ever more stringent requirements on electronic components, size, and power consumption. Third, directionality requires precise beam alignment at the transmitter and the receiver and implies increased control overhead. In order to overcome these limitations, the NR specifications include new Physical (PHY) and Medium Access Control (MAC) layer operations to support directional communications, which are collectively referred to as *beam management* according to the 3GPP terminology [15, 8]. In particular, NR networks must provide a mechanism by which User Equipments (UEs) and Next Generation Node Base Stations

[5] Although strictly speaking mmWave bands include frequencies between 30 and 300 GHz, industry has loosely defined it to include any frequency above 10 GHz.

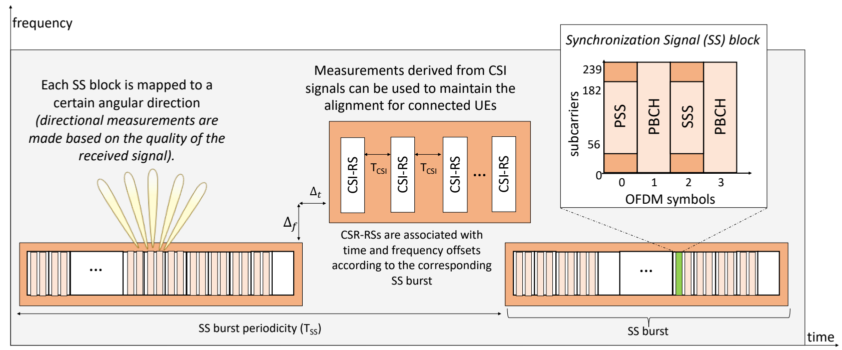

Fig. 2 Beam management structure in NR systems. SS blocks and CSI-RSs are used for beam measurements in idle and connected modes.

(gNBs)[6] regularly identify the optimal directional beams to interconnect at any given time.

The following four beam management operations are defined:

- *Beam sweeping*, i.e., exhaustively scanning a spatial area with a set of beams transmitted and received according to pre-specified intervals and directions.
- *Beam measurement*, i.e., measuring the quality, e.g., in terms of received power (RSRP) or Signal to Interference plus Noise Ratio (SINR), of the received beamformed signals [7].
- *Beam determination*, i.e., selecting the optimal beam (or set of beams) for establishing directional (and fully beamformed) communications.
- *Beam reporting*, i.e., the procedure through which the nodes feed back to the RAN information on the quality of the beamformed signals and on the decision made in the previous phase.

For idle users, beam management is fundamental to design a directional initial access strategy, which allows the mobile terminals to establish a physical link connection when first accessing the network [16]. In this case, beam management operations rely on a directional version of the 3GPP LTE synchronization signals called Synchronization Signal (SS) block, i.e., a group of 4 OFDM symbols in time and 240 subcarriers in frequency with the Primary Synchronization Signal (PSS), the Secondary Synchronization Signal (SSS) and the Physical Broadcast Channel (PBCH). Each SS block is mapped to a certain angular direction so that directional measurements can be made based on the quality of the received signal. To reduce the impact of SS transmissions, up to 64 SS blocks can also be grouped into the first 5 ms of an SS bursts [6], as illustrated in Fig. 2.

For users in connected mode, as the dynamics of the mmWave channel imply that the directional path to any cell can deteriorate rapidly, beam management is required to maintain precise alignment of the transmitter and receiver beams as the

[6] gNB is the 3GPP terminology for a base station.

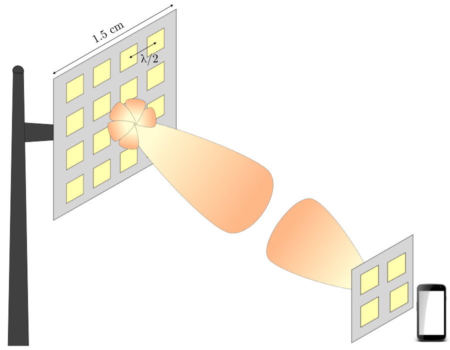

Fig. 3 Illustration of an UPA MIMO array. At the gNB side, the array (which has dimension of roughly 1.5 cm ×1.5 cm) is comprised of 4×4 elements, at the UE side the array has 2×2 elements. The antenna element radiation pattern is modeled as a patch antenna element with horizontal and vertical spacing equal to $\lambda/2$.

UEs move, an operation that is defined as tracking [23]. In this case, besides SS blocks, Channel State Information - Reference Signals (CSI-RSs) can also be used for beam measurement operations.

The beam management performance for both idle and connected UEs is a function of several parameters, including beamwidth, frame structure, SS burst and CSI-RS periodicity, and gNB density: the trade-off involves network reactiveness, system overhead and measurement accuracy. In general, better performance can be guaranteed considering *Non-Standalone (NSA)* deployments [5] (which is part of the 3GPP Rel. 15 standard specifications), in which NR gNBs use LTE as a radio overlay for control plane management [15].

4 Massive MIMO: a core component of NR systems

While the combination of extreme cell densification, increased system bandwidth, and more flexible spectrum usage (e.g., by resource sharing) represents a feasible and sustainable solution to meet 5G performance requirements, MIMO techniques have also emerged in modern wireless networks to improve reliability and spectral efficiency. The main concept is to use multiple transmit and receive antennas to exploit multipath propagation. Among the possible antenna array designs, the most suitable approach is the use of Uniform Planar Arrays (UPAs) where the antenna elements are evenly spaced on a two-dimensional plane and a 3D beam can be synthesized by adapting both azimuth and elevation planes, as illustrated in Fig. 3.

Depending on the channel proprieties, at the PHY layer MIMO systems can be configured for [29, 26]:

- *Spatial diversity*, i.e., sufficiently separated antennas are used to transmit redundant versions of the same message over multiple paths. The quasi-independent fading characteristics of the channel are thereby exploited to make links more robust and decrease the outage probability.

- *Beamforming*, i.e., multiple antenna elements are adaptively phased to form a concentrated beam pattern towards a specific direction. Beamforming provides significant array gains, thereby guaranteeing increased SNR (since propagation path loss is mitigated) and reduced co-channel interference (resulting from the spatial selectivity of the directional antenna).
- *Spatial multiplexing*, i.e., an outgoing signal is split into multiple independent streams which are transmitted simultaneously and in parallel on the same channel through different antennas. Throughput gains can be achieved, provided that Channel State Information (CSI) is available.

Moreover, multi-user MIMO (MU-MIMO) can be enabled through *Spatial Division Multiple Access (SDMA)*, in which the multipath proprieties of the channel are used to multiplex users in the spatial dimension while operating in the same time-frequency resource.

Typical MIMO installations use relatively few (i.e., less than 10) antennas, and the corresponding improvement in spectral efficiency has been relatively modest [27]. When combined with mmWave propagation, instead, the full potential of the MIMO paradigm can be truly unleashed. In fact, the physical size of antennas at mmWave frequencies is so small that it becomes practical to build large antenna arrays (e.g., with 100 or more elements), thereby scaling up the network performance by possibly orders of magnitude compared to state of the art MIMO implementations. The concept of using a number of antennas in network nodes which is much higher than the number of users is usually referred to as *massive MIMO* [18]. The promise of these benefits has elevated massive MIMO to a central position in NR, with a foreseen role of providing high-capacity and almost ubiquitous coverage in ultra-dense deployments [8]. For mmWave transmissions, massive MIMO is mainly used for beamforming while, at sub-6 GHz, it provides *channel hardening*, i.e., the combined usage of a massive number of antennas decreases the channel variability by averaging the small-scale fading [11].

However, massive MIMO comes with its own set of challenges, mainly related to:

- *hardware impairments:* massive MIMO systems exploit channel reciprocity to estimate the channel responses on the uplink and use such information for both uplink and downlink transmissions. Since the transceiver hardware is generally not reciprocal, calibration is needed to exploit the channel reciprocity in practice.
- *energy-consumption vs. flexibility trade-off:* while it is desirable to design *digital* beamforming architectures (which enable the transceiver to generate beams in multiple directions at the same time), they may suffer from increased energy consumption with respect to an *analog* strategy (which, in turn, has little flexibility since the transceiver can only beamform in one direction at a time).
- *CSI acquisition*: dynamic environments impose a finite coherence interval during which CSI must be acquired and utilized. As a consequence, there is a finite number of orthogonal pilot sequences that can be assigned to the network terminals.

Reuse of such pilots may result in *pilot contamination* and coherent interference, which cause performance degradation.

For NR, support for massive MIMO is introduced by using high-resolution CSI feedback and uplink Sounding Reference Signals (SRSs) targeting the utilization of channel reciprocity (e.g., twelve orthogonal demodulation reference signals are specified for multi-user MIMO transmission operations). Additionally, NR focuses on the support of *distributed MIMO*, through which the NR devices can receive multiple independent Physical Downlink Shared Channels (PDSCHs) per slot to enable simultaneous transmissions from multiple points to the same receiver.

5 Towards a disaggregated and virtualized network

As mentioned in Sec. 2, NR has been designed with flexibility in mind, in order to address the different 5G use cases. This has an impact also on the possible cellular network deployment architectures [4, 10], which follow two recent emerging technology trends: disaggregation and Cloud RAN (C-RAN) [20], and virtualization [10]. Fig. 4 represents the main novelties in terms of architectures and deployment options for NR and 5GC.

The LTE RAN and the associated core networks (Evolved Packet Core (EPC)) are characterized by the deployment of standalone pieces of equipment and servers, e.g., the evolved Node Bases (eNBs), and the core elements such as the Packet Gateways (PGWs) and Mobility Management Entities (MMEs). With NR, instead, the gNB can be split into separate physical units, i.e., the Distributed Unit (DU), which contains the lower layers of the protocol stack and is deployed in the field, and the Centralized Unit (CU) incorporating complete stack functionalities, which can be co-located with the DU or hosted in a data center facility, according to the C-RAN paradigm. As discussed in [19], this allows network operators to deploy the 5G RAN according to the use cases they want to serve, e.g., an ultra-dense small cell deployment with low utilization but high peak rate can rely on the C-RAN CU/DU split to maximize the multiplexing and enable a centralized control of the RAN, while a rural low-density deployment for the support of Internet of Things (IoT)

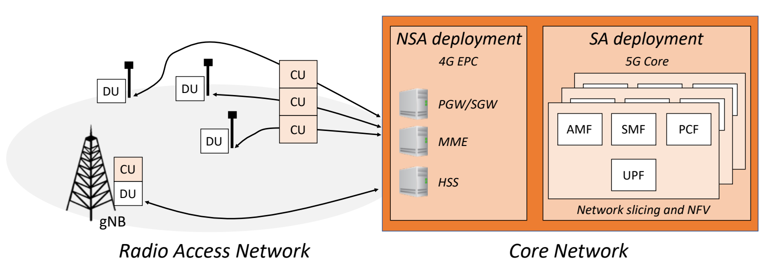

Fig. 4 Representation of different deployment and operation modes for the 5G NR RAN and the 5GC.

applications can feature complete gNB nodes. Moreover, in order to smooth the transition between the different network generations and reuse the widely deployed LTE and EPC infrastructure, the NR specifications foresee a NSA deployment, in which NR gNBs are connected to the EPC, possibly with a Dual Connectivity (DC) setup aided by LTE [24]. The other option is a standalone (SA) deployment, in which both the RAN and the core network respect the 5G specifications.

Finally, the 5G core network has been redesigned with respect to the 4G core following a service-based approach [10]: the 5G core is composed of multiple network functions, that provide mobility, authentication and routing support, that can be dynamically instantiated in data centers according to the load and traffic demands of the network. For example, while in LTE/EPC networks the control plane for the mobility of the user was handled by a single server (e.g., the MME), with the 5GC multiple network functions concur to offer the same set of services, but can be deployed in different data center locations and quickly turned off and on to decrease resource utilization. Moreover, the 5GC supports network slicing [14], i.e., the resources of the network can be split to serve different portions of traffic, that have different Quality of Service (QoS) requirements (e.g., IoT and mobile broadband traffic). The service-based 5GC architecture is an important enabler of network slicing in 5G, given that network functions can be provisioned dynamically to serve new network slices without the need to use separate servers, as would happen with the EPC.

6 Conclusions

In this chapter we described the main novelties that the 3GPP has specified for 5G NR, focusing on how they are used to match the 5G performance requirements. We showed that 3GPP NR is a flexible technology framework, which can be tuned to enable a wide range of 5G scenarios: it exploits a novel portion of the spectrum, to increase the throughput, a frame structure that can provide ultra-low latency, massive MIMO and several deployment architectures.

References

1. 3GPP, “Study on New Radio (NR) Access Technology - Physical Layer Aspects - Release 14,” TR 38.802, 2017.
2. ——, “TR 38.913, Study on Scenarios and Requirements for Next Generation Access Technologies, V14.1.0,” 2017.
3. ——, “WF on Work plan of Self Evaluation SI,” RP-172101, 3GPP TSG RAN77, Sapporo, Japan, September 2017.
4. ——, “NG-RAN; Architecture description - Release 15,” TS 38.401, V15.3.0, 2018.
5. ——, “NR - Multi-connectivity - Overall description (Stage 2),” TS 37.340, 2018.
6. ——, “NR - Physical channels and modulation - Release 15,” TS 38.211, V15.0.0, 2018.
7. ——, “NR - Physical layer measurements - Release 15,” TS 38.215, 2018.
8. ——, “NR and NG-RAN Overall Description - Release 15,” TS 38.300, 2018.
9. ——, “Study on New Radio (NR) access technology,” TR 38.912, V15.0.0, 2018.
10. ——, “System Architecture for the 5G System; Stage 2 - Release 15,” TS 23.501, V15.3.0, 2018.

11. E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114–123, February 2016.
12. E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2018.
13. R. Ford, M. Zhang, M. Mezzavilla, S. Dutta, S. Rangan, and M. Zorzi, "Achieving Ultra-Low Latency in 5G Millimeter Wave Cellular Networks," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 196–203, March 2017.
14. X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, May 2017.
15. M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Communications Surveys & Tutorials*, 2018.
16. M. Giordani, M. Mezzavilla, and M. Zorzi, "Initial access in 5G mmWave cellular networks," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 40–47, November 2016.
17. ITU-R, "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond," Recommendation ITU-R M.2083, September 2015.
18. E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.
19. A. Maeder, A. Ali, A. Bedekar, A. F. Cattoni, D. Chandramouli, S. Chandrashekar, L. Du, M. Hesse, C. Sartori, and S. Turtinen, "A scalable and flexible radio access network architecture for fifth generation mobile networks," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 16–23, November 2016.
20. N. Makris, C. Zarafetas, P. Basaras, T. Korakis, N. Nikaein, and L. Tassiulas, "Cloud-Based Convergence of Heterogeneous RANs in 5G Disaggregated Architectures," in *IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.
21. S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The New 5G Radio Access Technology," *IEEE Communications Standards Magazine*, vol. 1, no. 4, pp. 24–30, Dec 2017.
22. N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "5G New Radio Numerologies and their Impact on the End-To-End Latency," in *IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Sept 2018, pp. 1–6.
23. M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved Handover Through Dual Connectivity in 5G mmWave Mobile Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2069–2084, Sept 2017.
24. M. Polese, M. Mezzavilla, and M. Zorzi, "Performance Comparison of Dual Connectivity and Hard Handover for LTE-5G Tight Integration," in *Proceedings of the 9th EAI International Conference on Simulation Tools and Techniques*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2016, pp. 118–123.
25. T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.
26. S. Sun, T. S. Rappaport, R. W. Heath, A. Nix, and S. Rangan, "MIMO for millimeter-wave wireless communications: beamforming, spatial multiplexing, or both?" *IEEE Communications Magazine*, vol. 52, no. 12, pp. 110–121, December 2014.
27. A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: the next wireless revolution?" *IEEE Communications Magazine*, vol. 52, no. 9, pp. 56–62, September 2014.
28. K. Takeda, L. H. Wang, and S. Nagata, "Latency Reduction toward 5G," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 2–4, June 2017.
29. D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.
30. A. A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, and I. D. Silva, "Waveform and Numerology to Support 5G Services and Requirements," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 90–98, November 2016.